Scénario 4

REPRÉSENTATIONS D'UNE SÉRIE DE NOMBRES

1 Présentation générale

1.1 Objectifs :

Objectif général : réaliser un résumé d'une série de nombres.

Objectifs intermédiaires :

- Justifier le choix de paramètres résumant une série de nombres : médianes, écart interquartile, moyenne, variance.
- Réaliser une représentation graphique appropriée.
- Identifier des données anormales ou extérieures.

1.2 Les étapes

1. Exemple introductif et présentation du branchage (stem and leaf) :
   Poids relatifs de 62 femmes hypertensives américaines
2. Présentation d'un jeu de données, et outils connus par les stagiaires :
   Ages des 29 techniciens de recherche du Centre INRA Antilles-Guyane.
3. Les solutions numériques peu synthétiques :
   Classement, effectifs, effectifs par classe et fréquences.
4. Les graphiques associés :
   Le diagramme en bâtons et l'histogramme
5. Les solutions numériques synthétiques : les paramètres de position.
   La moyenne et la médiane.
6. Les solutions numériques synthétiques : les paramètres de dispersion. La variance et l'écart-type, l'écart interquartile, et les quantiles.
7. Les représentations semi-graphiques :
   La boîte à pattes (boxplot), et le branchage (stem and leaf) vu au début.
8. Le tableau récapitulatif et des exercices complémentaires :
   Rappel de tous les outils vus, en les plaçant dans un ordre de valeur synthétique croissante.
9. Evaluation globale.

1.3 Le matériel

Le matériel du formateur : les transparents

Le matériel des stagiaires : une calculette

1.4 La forme

Recherche des méthodes connues par les stagiaires. Un exemple traité avec chacune des méthodes par le formateur et le groupe. Un exemple analogue traité par les stagiaires à chaque étape sous forme d'exercices. Des exemples complémentaires présentés sur transparents.

1.5 Evaluation

Test d'évaluation général, en fin de scénario : sous forme de graphiques à mettre en correspondance, à compléter.

1.6 Prérequis

- Savoir utiliser une calculette
- Savoir faire une moyenne
- Avoir des notions sur les fonctions : x^2 et $\sqrt{x}$

1.7 Durée

3 heures

2 Les étapes

2.1 Etape 1 : Exemple introductif et présentation du branchage : (stem and leaf)

On dispose des poids relatifs de 62 femmes hypertensives américaines. La valeur indiquée est le rapport du poids réel de la femme, sur le poids normal d'une femme de même taille. (Nemenyi et al., 1977).
Les données :

117 89 107 102 95 132 163 123 103 144 108 89 118 123 92 140 102
112 115 123 130 102 119 119 114 119 100 106 115 145 115 108 122 113
153 116 81 85 122 120 113 140 117 119 121 130 114 107 125 94 96
123 91 120 125 125 103 133 102 131 147

On annonce que l'on va leur dicter les données

L'ordre de dictée n'a pas d'importance, par contre on veut avoir une idée de leur répartition et qu'à la fin elles soient grossièrement ordonnées.

On leur demande donc de noter les données "à la volée", en les réordonnant grossièrement au fur et à mesure.

Commencer à dicter 10 à 20 données, puis regarder leur papier pour voir comment ils font.

Tout ceci pour amener simplement à l'idée de la présentation sous forme du branchage (ou stem and leaf).

8	9 9 1 5	(c'est-à-dire 89, 89, 81 et 85 % du poids normal)
9	5 2 4 6 1	(95, 92, 94, 96 et 91)
10	7 2 3 8 2 2 0 6 8 7 3 2	
11	7 8 2 5 9 9 4 9 5 5 3 6 3 7 9 4 8	
12	3 3 3 2 2 0 1 5 3 0 5 5	
13	2 0 0 3 1	
14	4 0 5 0 7	
15	3	
16	3	

On obtient un branchage où les données de chaque branche ne sont pas ordonnées, si on note au fur et à mesure les feuilles sur les branches.

Réfléchir ensuite au choix des lignes pour des données différentes : taux de cholestérol :

227 215 305 255 227 270 218 244 320 198 334 250 175 278 284 241 383
238 295 219 213 253 354 239 345 174 213 304 245 228 267 190 346 189
245 280 281 271 223 250 274 213 317 259 307 245 226 248 188 271 449
210 327 142 225 237 194 254 189 322 225 363

Les deux premiers chiffres (on décide d'ignorer le 3ème) vont de 13 à 33 un choix raisonnable est de faire environ 6 lignes pour ces 62 individus d'où le branchage :

1 pour 10 à 14
1 pour 15 à 19 ...

1	4
1	9 7 7 9 8 8 9 8
2	2 1 2 1 4 4 3 1 1 3 1 4 2 4 2 1 4 2 4 1 2 3 2
2	5 7 5 7 8 9 5 6 8 8 7 5 7 7 5 7
3	0 2 3 4 0 4 1 0 2 2
3	8 5 6
4	4
4	

2.2 Etape 2 : Présentation d'un jeu de données, et outils utilisés par les stagiaires :

On dispose des âges des techniciens de recherche du Centre INRA Antilles-Guyane.

Transparent 1 : Le paquet de données est présenté sans ordre.

37 44 45 61 48 52 44 44 22 44 22 24 39 40 45
43 45 30 39 49 49 55 41 42 42 48 26 41 43

L'effectif est égal à 29; délibérement choisi impair pour un calcul simple de la médiane.

Question aux stagiaires :

- Que faire de ces données pour en tirer des informations synthétiques?
- Comment décrire ces données?

On note les réponses des stagiaires sur le tableau, en différenciant les solutions numériques et les solutions graphiques, en les ordonnant suivant un ordre synthétique croissant.

Même si toutes les solutions ne sont pas citées, on ne complète pas le tableau ainsi obtenu; ce qui manque sera rajouté au fur et à mesure. Par contre si des éléments sont cités, que l'on a pas prévu de faire, il suffira, quand on élaborera avec les stagiaires le tableau récapitulatif final de synthèse, d'en dire deux ou trois mots, en précisant que ces éléments sont peu utilisés; ce sera le cas par exemple, pour le "mode".

Le but est d'obtenir au tableau dès le départ, un début de canevas du tableau général final; tableau sur lequel les stagiaires pourront toujours raccrocher les idées au fur et à mesure qu'elle seront introduites. Le tableau sera par exemple :

ordre synthétique croissant	solutions numériques	solutions graphiques	solutions semi-graphiques
	classement		
	rang		
	effectifs par âge	diagramme en bâtons	
	effectifs par classe	histogramme	branchage
	fréquences		
	paramètres de position et de dispersion :		boîte à pattes
	minimum, maximum, médiane, moyenne		
	étendue, écart inter-quartile variance, écart-type		

Il ne sera présenté sur transparent qu'à l'étape finale.

2.3 Etape 3 : Les solutions numériques peu synthétiques.

On passe en revue les différentes possibilités retenues, en utilisant les données toujours présentées sur le transparent 1.

Classement des données en ordre croissant :

Le classement est marqué au tableau :

22 22 24 26 30 37 39 39 40 41 41 42 42 43 43
44 44 44 44 45 45 45 48 48 49 49 52 55 61

Il permet d'identifier les extremas (minimum et maximum), l'étendue (maximum - minimum), des regroupements éventuels, voire des données "aberrantes".

Le branchage : ou "stem and leaf"

Transparent 2 : Présentation du branchage pour les âges des agents du Centre INRA Antilles-Guyane.

Explication de son mode de construction
Différents choix possibles pour les branches
On a ordonné les "feuilles", dans chaque branche.

On pourrait lui rajouter une colonne à gauche pour mettre les effectifs de chaque branche.

Toutes les données sont présentes, et on a une idée de la forme de la distribution.

Effectifs par âge :

A partir des types d'âges récupérés dans les données classées, on construit et on remplie avec les stagiaires, le tableau suivant :

âge	24	26	30	37	39	40	41	42	43	44	45	48	49	52	55	61
effectif	2	1	1	1	1	2	1	2	2	2	4	3	2	2	1	1

Transparent 3 : sur les classes

Il permet de voir la plus ou moins bonne représentation des différents âges, en constatant que certains ont des effectifs nuls.
C'est la préparation au diagramme en bâtons
On a perdu la notion de valeur individuelle.

Effectifs par classe d'âge :

De même, à partir des effectifs par âge, on présente la possibilité de construire des classes d'âges. On construit par exemple le tableau suivant avec eux :

classe d'âge	20-29	30-39	40-49	50-59	60-69
effectif	4	4	18	2	1

On construit des classes, de telle façon qu'elles contiennent en général au moins quelques éléments.
C'est une description encore plus synthétique que la précédente, qui prépare à la construction de l'histogramme.

Au lieu de considérer les effectifs, on peut aussi utiliser les fréquences associées à chaque âge ou classe d'âges, connaissant l'effectif total, qui ici vaut 29.

On est donc amené à construire le tableau suivant :

classe d'âge	20-29	30-39	40-49	50-59	60-69
effectif	4	4	18	2	1
fréquence	4/29	4/29	18/29	2/29	1/29
soient	0.14	0.14	0.62	0.07	003

ou avec l'effectif total, on a :
fréquence (classe i) = effectif (classe i) / effectif total

On remarque que : la somme des fréquences, pour toutes les classes vaut 1.

2.4 Etape 4 : Les graphiques associés à ces premières solutions numériques.

Le diagramme en bâtons :

Transparent 4 :

Il s'applique à des valeurs continues, découpées en classes.
Le branchage est mieux que l'histogramme, car il présente toutes les données, et donne aussi un idée de la forme de la distribution.
Problèmes que l'on retrouve aussi mais de façon moins cruciale avec le branchage :

- choix du nombre de classes

  Le nombre de classes est choisi de telle façon que l'aspect de l'histogramme soit représentatif de la distribution des âges (aspect régulier), sans en gommer les particularités intéressantes (on perd nécessairement de l'information quand on regroupe en classes)

  Pour le formateur, on précise qu'il existe trois formules générales pouvant être utilisées pour donner un nombre correct de classes, à partir de l'effectif total (n) :

  1. $10 \times log^{10} n$

  2. $2 \times \sqrt{n}$

  3. $1 + log_2 n$

  Dans le logiciel S, c'est la troisième qui est utilisée pour un histogramme.

  Cela donne par exemple, pour différents effectifs, les nombres de classes correspondants :

n	formule 1	formule 2	formule 3
16	12	8	5
32	15	11	6
64	18	16	7
128	21	23	8
256	24	32	9
512	27	45	10

- choix entre amplitudes des classes égales ou différentes; ce qui revient à rendre l'effectif proportionnel à la hauteur si les amplitudes des classes sont égales; dans le cas contraire, on ne peut plus avoir d'échelle verticale, mais plutôt une surface unité de référence; il est donc moins facile à interpréter.
- porter en ordonnées les effectifs ou les fréquences; si on a un seul jeu de données, passer des effectifs aux fréquences ne revient qu'à faire un changement d'échelle, sans modifier l'allure de l'histogramme; par contre si on veut comparer plusieurs histogrammes il faut utiliser les fréquences.
- choix des bornes des classes : on a choisi ici $a \leq x < b$ c'est-à-dire $x \in [a, b[$ sur tous les histogrammes.

◇ Pause-Evaluation :

Distribuer les feuilles stagiaires 1 et 2.

Appliquer ces méthodes de description à une autre série de données :
"l'âge des ingénieurs du Centre INRA Antilles-Guyane".

Solution sur les transparents S1 et S2.

On demande aux stagiaires :

- Le classement des données,
- Les effectifs par âge,
- Le diagramme en bâtons,
- Les effectifs par classe d'âge, d'amplitude 5 ans puis 10 ans,
- Les histogrammes correspondants.

Fin-Evaluation ◇

2.5 Etape 5 : Les solutions numériques synthétiques : les paramètres de position.

La moyenne :

- à introduire au tableau sur le jeu de données des âges des techniciens INRA d'Antilles-Guyane
- C'est la somme de toutes les données, divisée par l'effectif total.
- On introduit le signe de sommation avec un indice, pour la formule de calcul de la moyenne; on doit donc écrire tous les termes du calcul au tableau, afin de mieux expliquer ce signe.

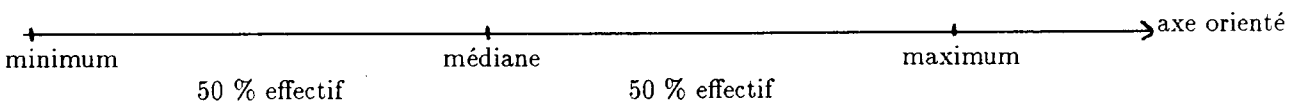
$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- Pour le premier jeu de données, la moyenne est égale à 41.5

La médiane :

- à introduire aussi au tableau sur le même jeu de données.
- C'est la valeur qui sépare les données ordonnées, en deux parties d'effectifs égaux.
- Un problème se pose si le nombre d'éléments est pair :

S'il y a $n = 2p + 1$ termes dans la série, celui de rang $p + 1$ est la médiane, et s'il y a $n = 2p$ termes dans la série, toute valeur comprise entre celle de rang p et celle de rang $p + 1$ peut convenir comme médiane. On convient de prendre comme valeur de la médiane la valeur qui correspond à la moyenne arithmétique des termes de rang p et de rang $p + 1$.



- Pour le premier jeu de données, la médiane est égale à 43

Différence entre la médiane et la moyenne :

Transparent 6 : On dispose de 5 "individus", sur lesquels on a mesuré la valeur d'une variable X ; on dispose donc de 5 valeurs, que l'on peut représenter graphiquement par 5 points sur un axe.

$$X_i \text{ pour l'individu } i, \text{ avec } 1 \leq i \leq 5$$

On a vu que l'on pouvait résumer ces valeurs par un paramètre de position, qui peut être par exemple, la moyenne ou la médiane.
On montre ensuite le base du transparent, où l'on a "tiré" le point "X_4" vers la droite, les autres points ne bougeant pas : la médiane est toujours la même, alors que la moyenne s'est déplacée de l'autre côté de "X_1". On met ainsi en évidence une plus grande sensibilité de la moyenne aux valeurs extrêmes, que la médiane.

◇ Pause-Evaluation, exercices :

1. Faire calculer les paramètres de position du deuxième jeu de données (les ingénieurs INRA Antilles-Guyane) :

   - moyenne : 39.7
   - médiane : 37

Pour le calcul de la médiane, la difficulté vient du nombre pair de données.

Fin Evaluation ◇

2.6 Etape 6 : Les solutions numériques synthétiques : les paramètres de dispersion.

Transparent 7 : Ce transparent présente les histogrammes de fréquences des jeux de données 1 et 2, qui ont même moyenne et même médiane, avec des limites de classe différentes, ne permettant pas de les comparer.

On met l'accent sur le fait que pour pouvoir comparer deux distributions sous forme d'histogrammes, il faut que ce soient des histogrammes de fréquences, avec les mêmes limites de classes.

Transparent 8 : Présentation des distributions des mêmes jeux de données 1 et 2, qui ont même médiane et même moyenne, mais des dispersions différentes, avec des limites de classes identiques.

Les distributions sont représentées sous la forme d'histogrammes de fréquences, avec des limites de classes identiques, afin de permettre la comparaison des deux distributions.

On constate que sur les histogrammes des jeux de données 1 et 2, qui ont même moyenne et même médiane (médiane = moyenne = 40), la forme des distributions est assez différente; les paramètres de position semblent donc insuffisants pour décrire suffisamment les distributions.

Résumé par une seule valeur $\Longrightarrow$ perte d'information importante.

On introduit donc un paramètre qui caractérise la dispersion des données autour de la valeur de tendance centrale (moyenne ou médiane).

La variance et l'écart-type :

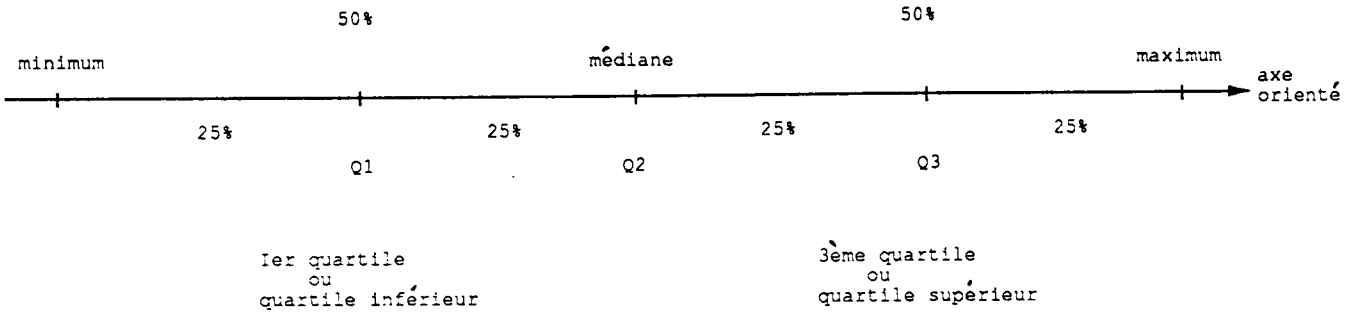
Présentation de la variance et de l'écart-type à partir de transparents, en utilisant des surfaces.

L'écart inter-quartile :

On présente l'insuffisance de l'étendue comme paramètre de dispersion, car trop sensible aux valeurs extrêmes (maximum et minimum); on cherche donc un écart moins sensible à ces données extrêmes.

Faire la présentation suivante au tableau, ou à l'aide du transparent suivant :

Transparent 13 : Présentation de l'écart inter-quartile.
On pense redécouper chaque intervalle en deux parties ayant le même nombre d'éléments, soient :



Remarque : Q2 ou deuxième quartile est aussi la médiane.

On présente les quartiles, puis l'écart inter-quartile.

Si IQR = Intervalle Inter-Quartile (Inter-Quatrile Range), on a :

$$IQR = Q3 - Q1$$

L'IQR est lié à la médiane.

Les quantiles :

Transparent 14 : Permet si on le désire, à partir de la présentation précédente de raccrocher la notion de quantiles
On appelle quantiles d'ordre k, les valeurs caractéristiques :
$q_{1/k}, q_{2/k}, \ldots, q_{k-1/k}$ qui divisent la série statistique en k parties égales.
$q_{i/k}$, le ième quantile est la valeur de la série statistique telle que une proportion p/k des observations lui soient inférieures.
Donc $q_{i/k}$ est la valeur qui est supérieure à 100(i/k)% des valeurs classées dans l'ordre croissant.

Exemples : pour les quartiles, on a donc $k = 4$, avec Q1 supérieur à 25% des données, Q2 à 50% et Q3 à 75%.
On définit aussi de la même façon les déciles ($k = 10$), les centiles $k = 100, \ldots$

2.7 Etape 7 : Une représentation semi-graphique :

La boîte à pattes : ou "boxplot" ou "boite à moustaches"

La présentation suivante est à faire au tableau, ou à l'aide du transparent suivant :

Transparent 15 : Présentation du boxplot

On vient de voir qu'à partir des valeurs observées, ordonnées sur un axe, on pouvait isoler des valeurs résumant la distribution :

min, Q1, médiane, Q3, max.

On explique comment construire la "boîte", puis les "pattes", à partir du même axe ordonné.
L'extrémité des intervalles (1.5 * IQR) est toujours ramenée à la valeur observée la plus proche, dans la direction de la "boite".
Ces extrémités d'intervalles, qui représentent donc toujours des valeurs réelles observées, sont appelées les "valeurs adjacentes".
Les valeurs qui sont au-delà de ces intervalles, sont représentées individuellement par une étoile, symbolisant des valeurs plus ou moins "exceptionnelles" de la distribution.
On peut aussi faire figurer la position de la moyenne par un signe plus.

Remarque : Pour les valeurs adjacentes, on prend presque toujours le coefficient "1.5" car il résulte d'une pratique empirique, basée sur l'expérience; les observations qui sortent de ces (1.5 * IQR) sont vraiment très éloignés de la distribution centrale.

Transparent 16 : Présentation du boxplot des âges des agents du centre INRA Antilles-Guyane.

C'est un graphique synthétique utilisant :

- la médiane
- les quartiles Q1 et Q3
- le maximum et le minimum
- les valeurs adjacentes, déterminées à partir de l'IQR soient :

maximum {minimum , (Q1 - 1.5 * IQR)}
ou
minimum {(Q3 + 1.5 * IQR, maximum }

Il permet de visualiser de façon synthétique la forme d'une distribution, de situer un paramètre de position et un paramètre de dispersion, de se faire une idée sur sa symétrie ou son assymétrie, de mettre en évidence des valeurs "suspectes", ...

La boite à pattes peut être complétée avec un branchage.

Transparent 17 : Présentation des deux boxplots des âges du Centre INRA Antilles-Guyane, pour les techniciens et les ingénieurs.

Le boxplot permet de comparer facilement différentes distributions.

Transparent 18 : Présentation de différents boxplots, avec différentes formes possibles.

On montre :

- la médiane comme paramètre de position,
- la comparaison des dispersions,
- la symétrie ou la non symétrie,
- la présence de points "exceptionnels",
- ...

De par ses limites (pour la bimodalité par exemple, qui fera l'objet d'un exercice complémentaire), c'est donc un outil qui doit être complété par d'autres outils : le branchage par exemple.

Transparent 19 : Récapitulatif comparé des outils graphiques vus sur l'exemple des techniciens INRA Antilles-Guyane.

◇ Pause-Evaluation, exercices :

Feuille stagiaire 3

Faire calculer les paramètres de dispersion du deuxième jeu de données : (ingénieur INRA Antilles-Guyane)

* variance : 110.5
* écart-type : 10.5
* quartile inférieur : 30.5
* quartile supérieur : 48.5
* écart interquartile : 18

Faire construire le branchage et la boite à pattes correspondantes.

Fin Evaluation ◇

2.8 Etape 8 : Présenter le tableau récapitulatif final de tous les outils abordés dans le scénario

Transparent 20 : Résumé des outils

Exercices complémentaires :

Feuille stagiaire 4 : Détermination de la médiane d'une série, à partir de sa représentation en diagramme en bâtons.

Solution : médiane = 10.

Feuille stagiaire 5 : Commentaires sur une série bimodale, présentée sous forme d'un histogramme.

Les données : la malaria a été inoculée à des souris et on s'intéresse à leur durée de vie.

Questions :

- l'histogramme est-il bien fait?
  (réponse : oui)
- faire une boite à pattes grossière
  (on peut ensuite montrer le transparent S6 qui présente la boite à pattes de ces données)
- si on n'a que la boite à pattes; est-ce un bon résumé? Pourquoi?
  (réponse : non, et faire exprimer la notion de bimodalité, soit ici l'existence de deux populations de souris qui répondent différemment)
- la moyenne, la médiane, l'écart-type, l'IQR sont-ils de bons paramètres résumés, dans ce cas?
  (réponse : non, pour les mêmes raisons)

Dans ce cas le boxplot est inadapté $\Longrightarrow$ d'où la nécessité d'utiliser plusieurs outils complémentaires.

2.9 Etape 9 : Evaluation générale finale

Feuilles évaluation 2, 3

On présente différents petits jeux de données sur un axe ordonné, différents histogrammes, stem and leaf et boxplots; au stagiaire de les associer correctement.

Feuille évaluation 4 :

Détermination des valeurs caractéristiques d'un boxplot, à partir d'un jeu de données ordonnées.

Feuilles évaluation 5 et 6 :

Construire deux boxplots à partir de leur stem and leaf; indiquer un caractère qui différencie les deux séries de données.
(réponse : IQR ou écart-type, soit un paramètre de dispersion)

Feuille évaluation 7 :

Sur un axe ordonné on a représenté toutes les valeurs d'un jeu de données sauf une; par contre on a représenté :

- la position de la moyenne, tenant compte de cette valeur absente; déterminer la valeur de cette donnée non représentée?
- la position de la médiane, tenant compte de cette valeur absente; déterminer la valeur manquante de ce deuxième jeu de données.

Feuille évaluation :

Question : Pour décrire les revenus d'une population d'un pays, on peut donner la médiane ou la moyenne des salaires mensuels :

	moyenne	médiane
pays 1	20 000	8 000
pays 2	10 000	10 000
pays 3	11 800	9 000

Quel autre paramètre voulez-vous connaître avant d'émigrer?
Suggérer des distributions.

Feuilles formateur 1,2,3,4,5 : corrections des exercices précédents.

Fin Evaluation

2.10 Mots-clé

Branchage (stem and leaf)
Boîte à pattes (boxplot)
Distribution
Diagramme en bâtons
Ecart inter-quartile (IQR ou Inter-Quartile Range)
Ecart-type (standard deviation)
Histogramme
Médiane
Moyenne
Paramètre de position
Paramètre de dispersion
Quantile
Quartile
Symétrie
Variance

2.11 Liste du vocabulaire

Boîte à pattes
Branchage
Diagramme en bâtons
Ecart-type
Histogramme
Médiane
Moyenne
Quartile
Variance

2.12 Références bibliographiques

NEMENYI P. , DIXON S.K., WHITE Jr. N.B., HEDSTROM M.L. (1977) *Statistics from scratch.* Holden-day, San Francisco.

TUKEY J.W., HOAGLIN D.C., MOSTELLER F. (1983) *Understanding robust and exploratory data analysis.* Wiley.

3 Annexes

3.1 Documents stagiaires

Présentation du scénario
Feuilles stagiaires 1 à 5
Feuilles évaluation 1 à 8

3.2 Documents formateurs

Transparents réduits 1 à 9
Transparents réduits + commentaires 10 à 12
Transparents réduits 13 à 20
Transparents solutions des exercices S1,S2,S3 et S5
Solution de l'évaluation finale : feuilles formateur 1 à 5

3.3 Les transparents