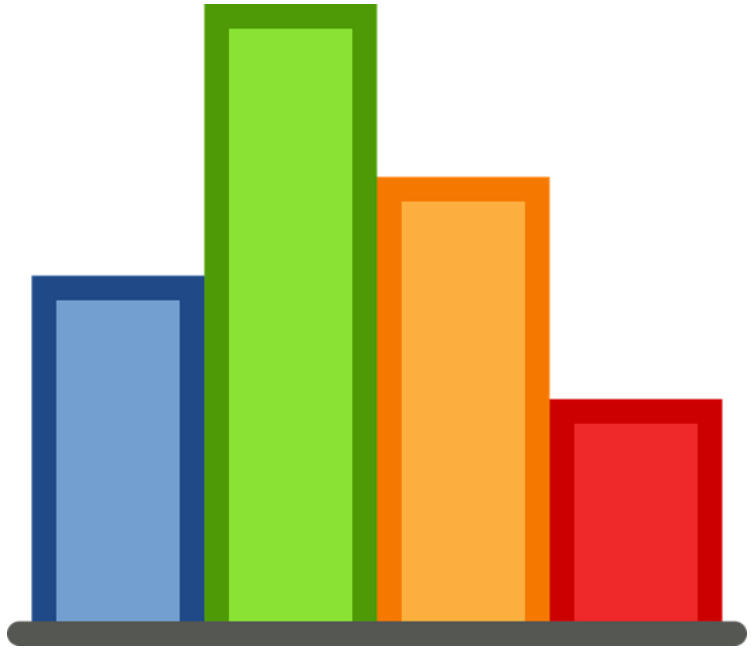


Statistique descriptive

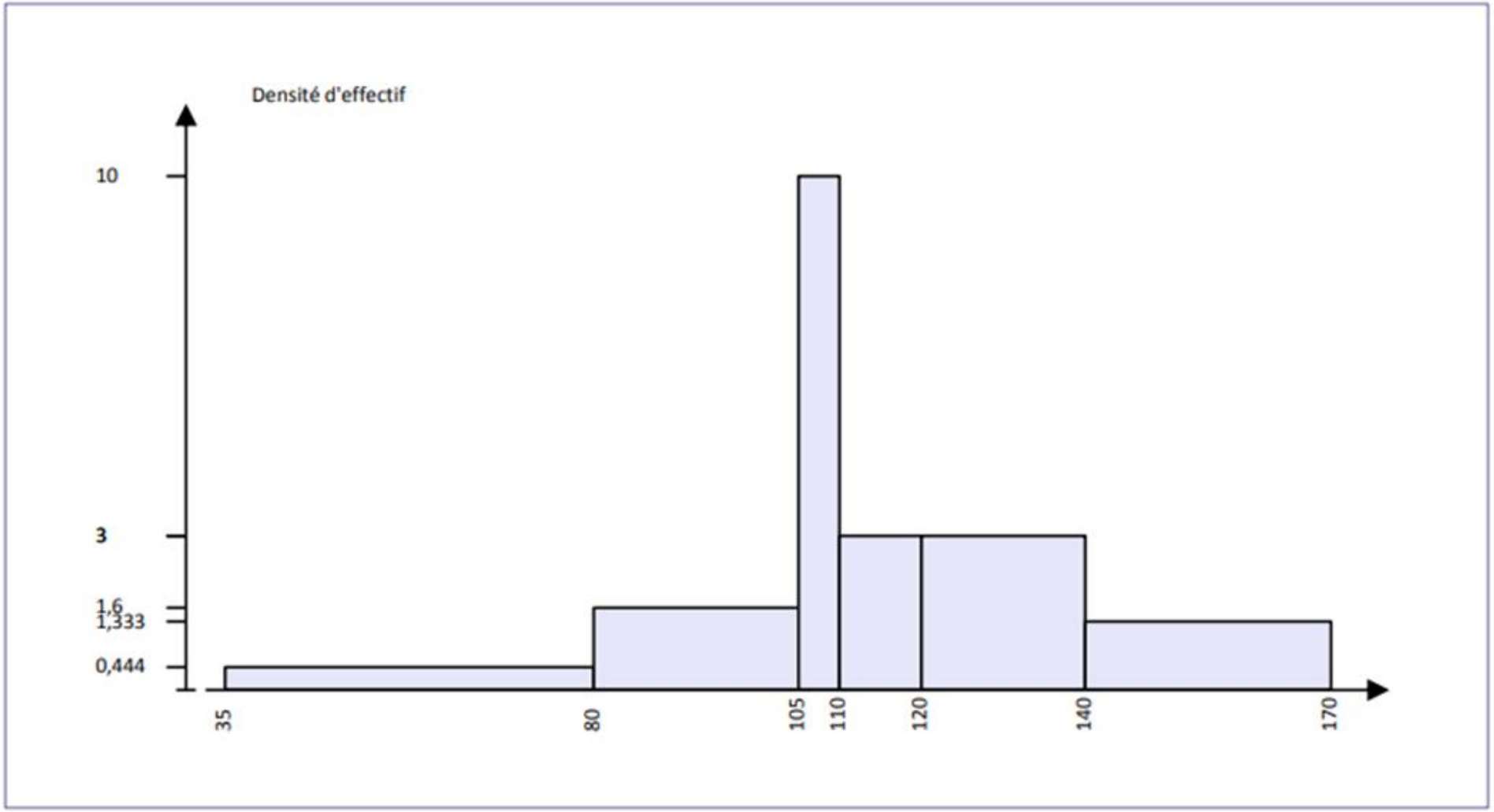
Gérard Barmarin

Cours 3

Exercices 3.6

Ex 3.6

Voici l'histogramme représentant la distribution sur leur superficie (en ares) de 240 terrains agricoles vendus au premier trimestre 2013 à Cracovie.



Ex 3.6

On demande :

- Que sont les individus distribués ?
- Combien y-a-t-il d'individus observés ?
- Combien y-a-t-il d'observations ?
- Pouvez-vous estimer le nombre de valeurs observées différentes ?
- Quelle est la variable étudiée ?
- Pouvez-vous donner la liste de toutes les observations (en utilisant des pointillés si nécessaire) ?
- Quelle est la superficie moyenne de tous ces terrains ?
- Quelle est la variance de cette distribution ?
- Quel est l'écart type ?
- Quel est l'écart absolu moyen ?
- Quelle est l'étendue de cet ensemble d'observations ?
- Quel est la classe modale et le mode de cette distribution ?
- Faites le graphe des effectifs cumulés de cette distribution

Ex 3.6

- Y-a-t-il une valeur telle que ¼ des observations sont en-dessous de cette valeur ? Si oui comment s'appelle-t-elle et pouvez-vous la calculer ?
- Calculer la médiane
- A quoi correspond la médiane ?
- Quelle est la valeur inter-quartile ?
- Quelle est la valeur quantile d'ordre 0,675 ?
- Quelle est la proportion de terrain plus petit que 83,5 ares ?
- Quelle est la proportion de terrains dont la superficie est comprise entre 83,5 et 137 ares.

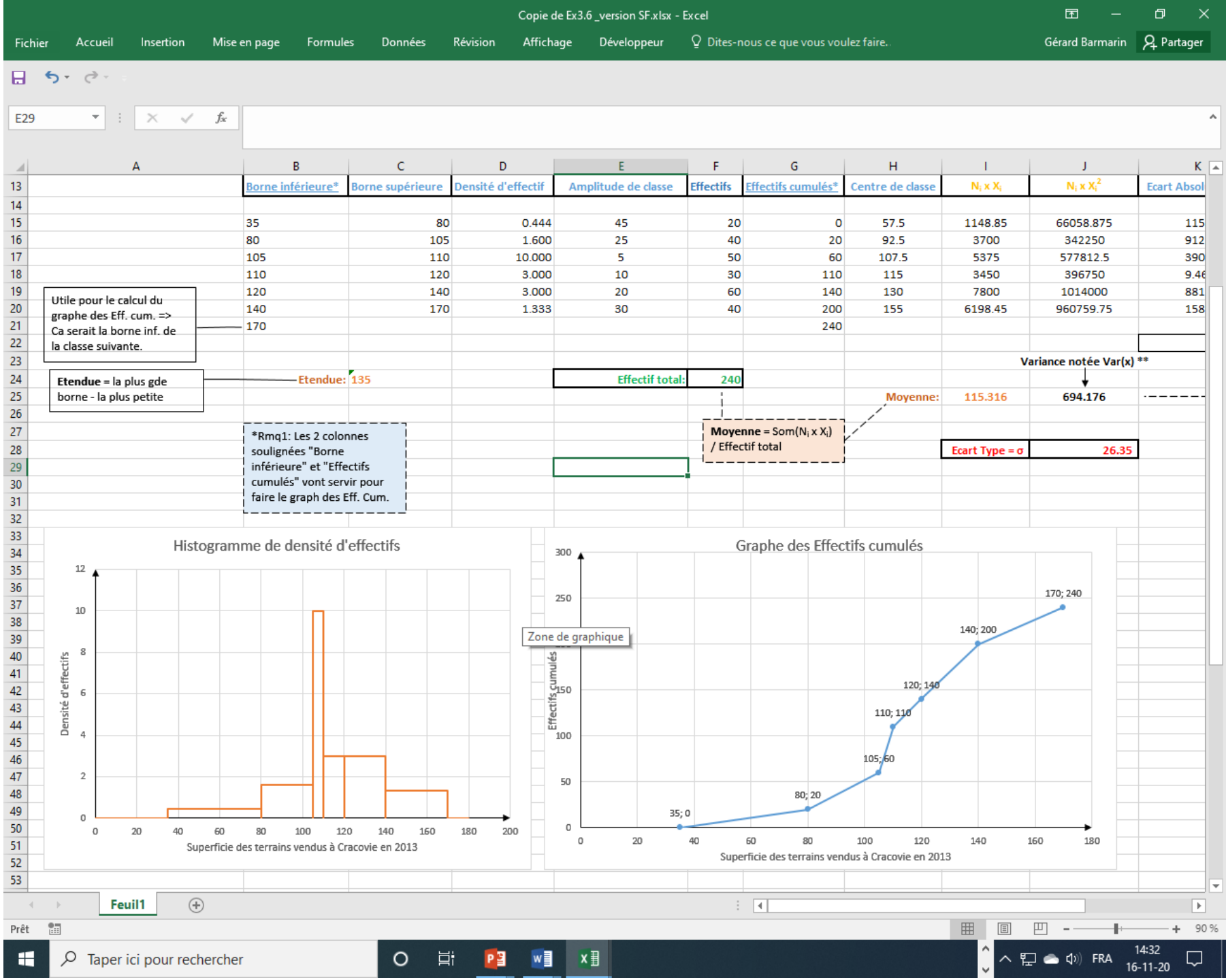
Correction exercice 3.6

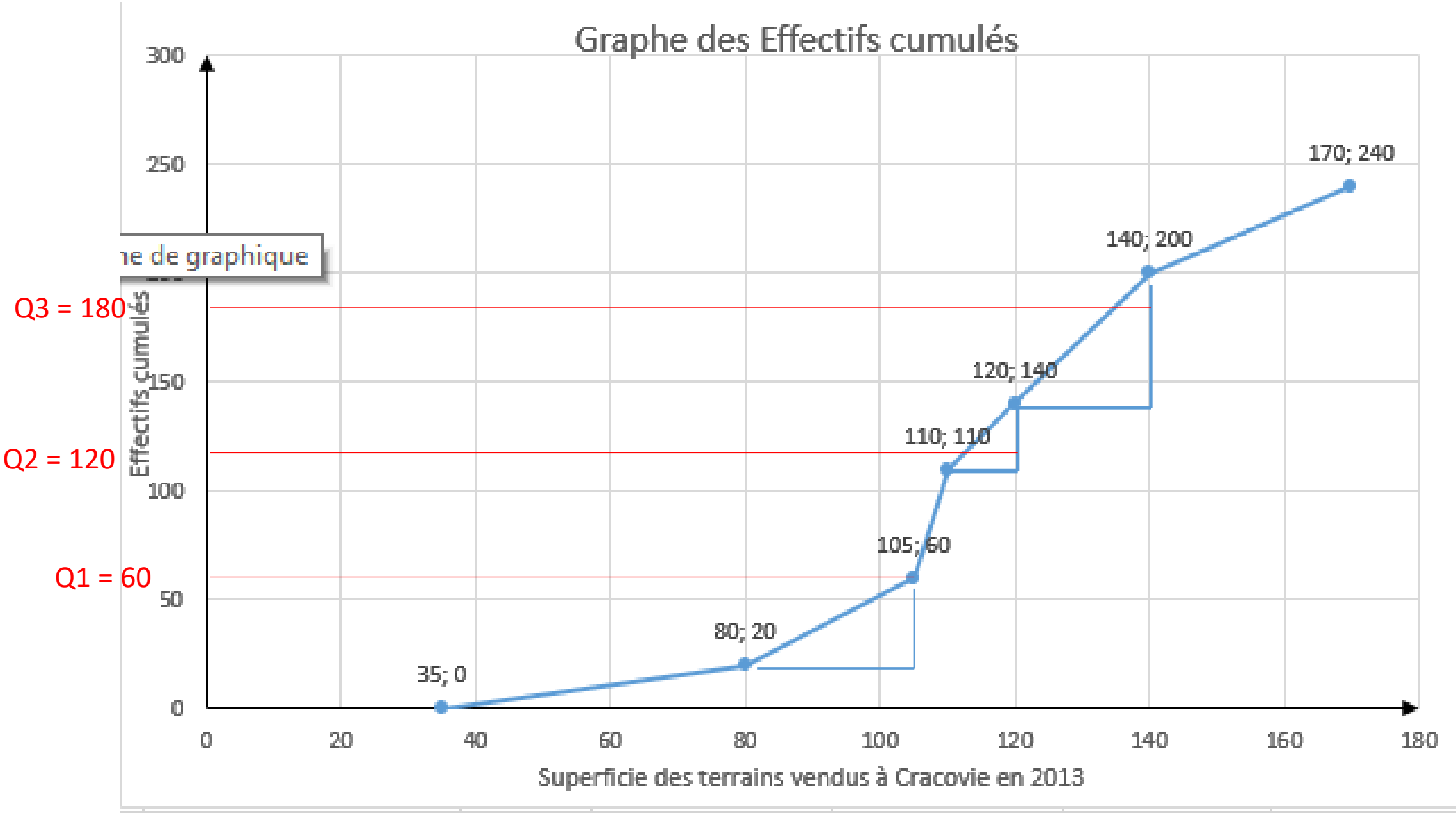
Effectif: 240
Étendue: 135 ares
Moyenne: 115,313 ares
Variance : 694,318
Ecart type: 26,35 ares
Ecart absolu moyen: 20,573 ares

Classe modale:
[105,110[
Mode: 107,5 ares

Q1 = 105 ares
Q2 = 113,333 ares
Q3 = 133,333 ares

Intervalle inter Q: 28,333 ares

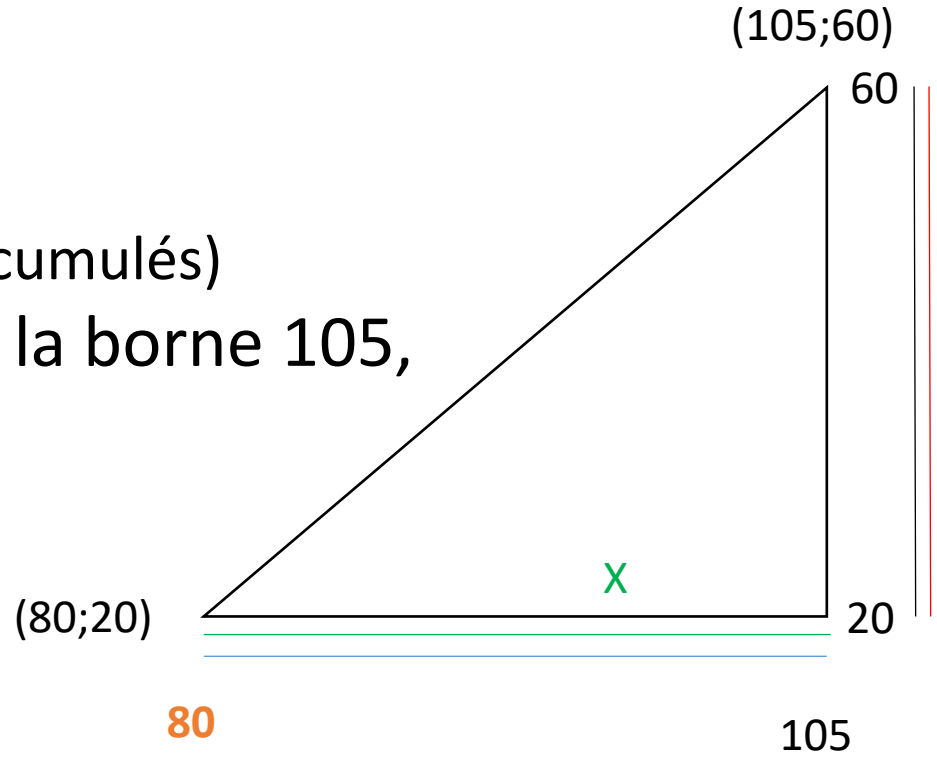



Graphe des Effectifs cumulés
ne de graphique
Effectifs cumulés
300
250
150
100
50
0
0
20
40
60
80
100
120
140
160
180
Superficie des terrains vendus à Cracovie en 2013
35; 0
80; 20
105; 60
110; 110
120; 140
140; 200
170; 240
Q3 = 180
Q2 = 120
Q1 = 60

Calcul de Q1

240 x 0,25 = 60ème valeur (Q1 = 25% des effectifs)

La 60ème valeur est dans la classe [80, 105[(cfr effectifs cumulés)
Et correspond exactement à l'effectif cumulé de la borne 105,
Donc la superficie correspondant
à la 60ème valeur est 105 ares
sans avoir à faire de calculs!



60 – 20 = 40

105 – 80 = 25

60 – 20 = 40

Donc Q1 = 105 ares

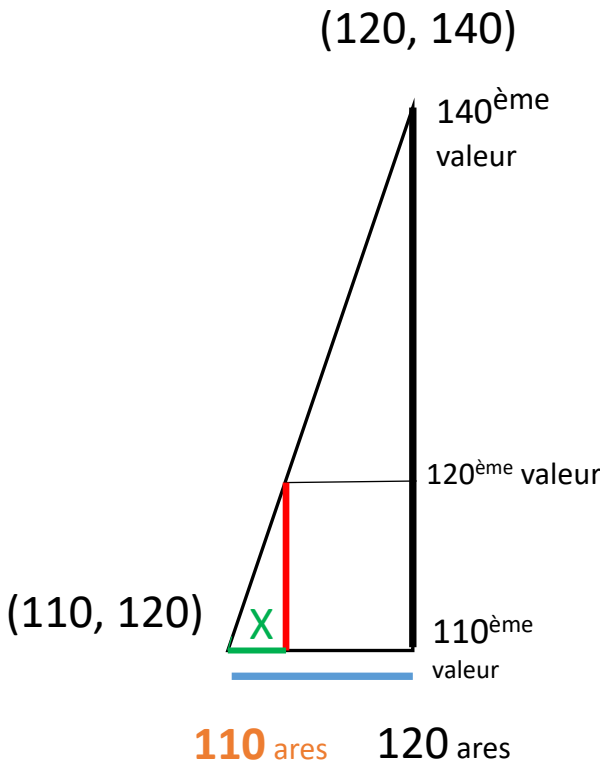
Calcul de Q2 (Médiane)

240 x 0,50 = 120ème valeur (Q2 = 50% des effectifs)

La 120ème valeur est dans la classe [110, 120[(cfr effectifs cumulés)

$$\frac{X}{10} = \frac{10}{30} \qquad \text{donc } X = \frac{10 \times 10}{30} = 3{,}333$$

Donc Q2 = 110 ares + 3,333 ares = 113,333 ares



140 – 110 = 30

120 – 110 = 10

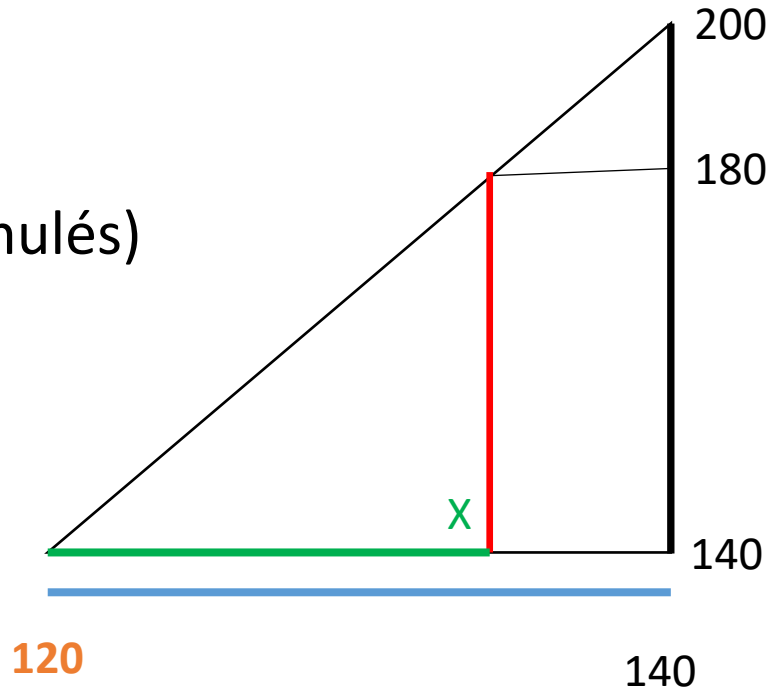
120 – 110 = 10

Calcul de Q3

240 x 0,75 = 180ème valeur (Q3 = 75% des effectifs)

La 180ème valeur est dans la classe [120, 140[(cfr effectifs cumulés)

$$\frac{X}{40} = \frac{20}{60} \qquad \text{donc } X = \frac{20 \times 40}{60} = 13{,}333$$



200 – 140 = 60

140 – 120 = 20

180 – 140 = 40

Donc Q3 = 120 ares + 13,333 = 133,333 ares

Correction exercice 3.7

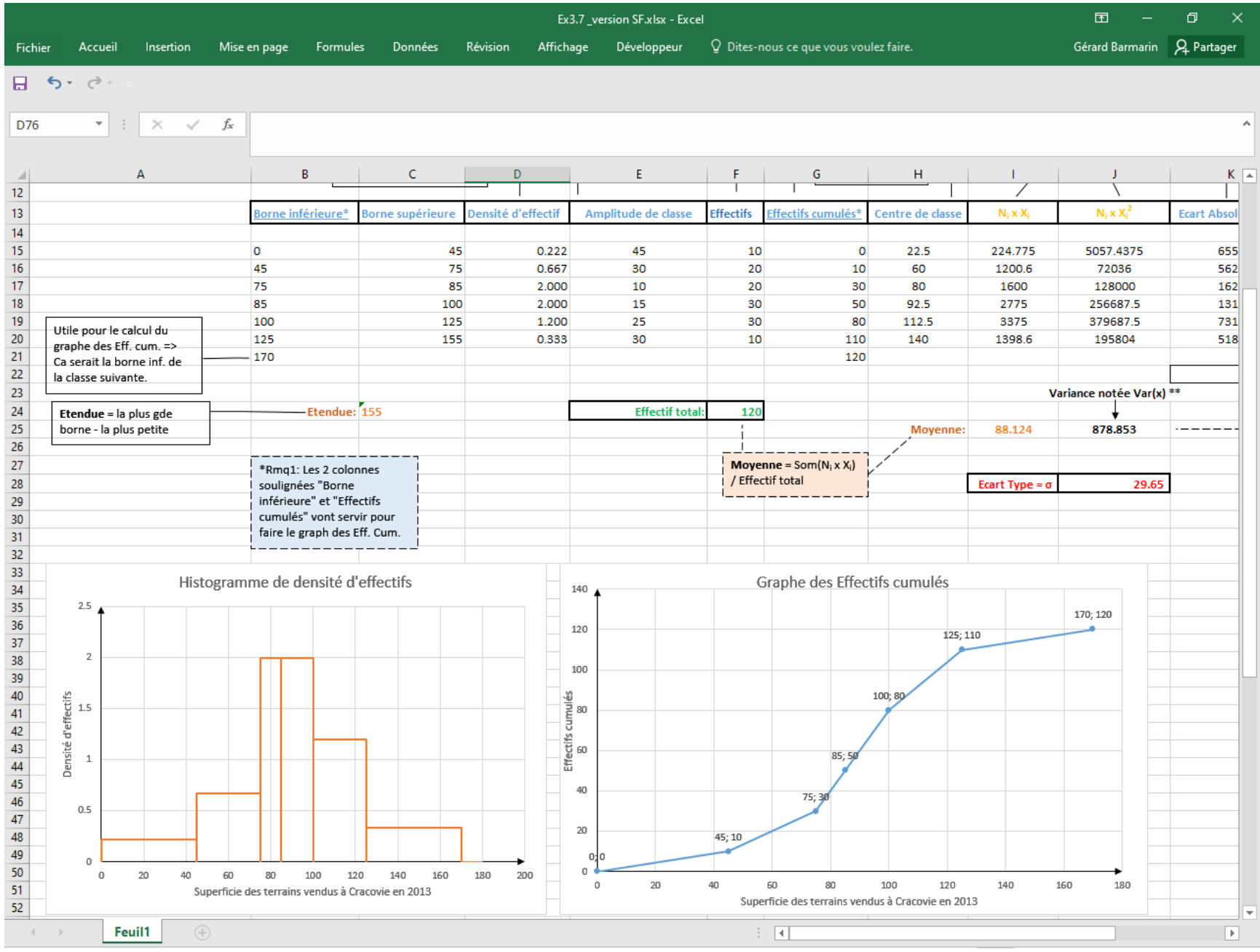
En utilisant la feuille Excel précédente:

Effectif: 120
Étendue: 155 ares
Moyenne: 88,125 ares
Variance 879,297
Ecart type: 29,663 ares
Ecart absolu moyen: 23,021 ares

Classe modale:
[75,85[et [85,100[
Mode: 87,5 ares= (100+75)/2

Q1 = 75 ares
Q2 = 90 ares
Q3 = 108,33 ares

Intervalle IQ: 108,33 -75 = 33,33 ares

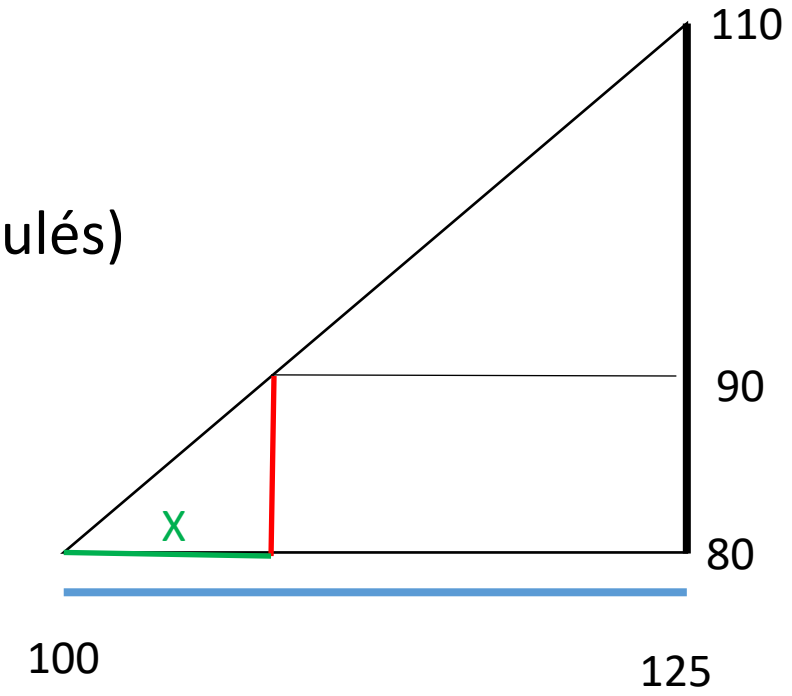


Calcul de Q3

120 x 0,75 = 90ème valeur (Q3 = 75% des effectifs)

La 90ème valeur est dans la classe [100, 125[(cfr effectifs cumulés)

$$\frac{X}{10} = \frac{25}{30} \qquad \text{donc } X = \frac{25 \times 10}{30} = 8{,}33$$



110 – 80 = 30

125 – 100 = 25

90 – 80 = 10

Donc Q3 = 100 ares + 8,33 = 108,33 ares

Correction exercice 3.7:

Quelle est la proportion des observations à plus d'un écart-type de la moyenne?

Moyenne: 88,125 ares
Ecart type: 29,663 ares
Moyenne + 1 écart type: 88,125 + 29,663 = 117,79 ares
Moyenne – 1 écart type: 88,125 – 29,663 = 58,49 ares

On va chercher le nombre de terrains qui ont une superficie < 117,79 ares
Ensuite on cherche le nombre de terrains dont la superficie est < 58,49 ares

La différence correspond au nombre de terrain dont la superficie est comprise entre 117,79 et 58,49 ares et il suffira de diviser par l'effectif total pour obtenir la proportion correspondante! 100% - cette proportion donnera le nombre de terrain hors de ces limites.

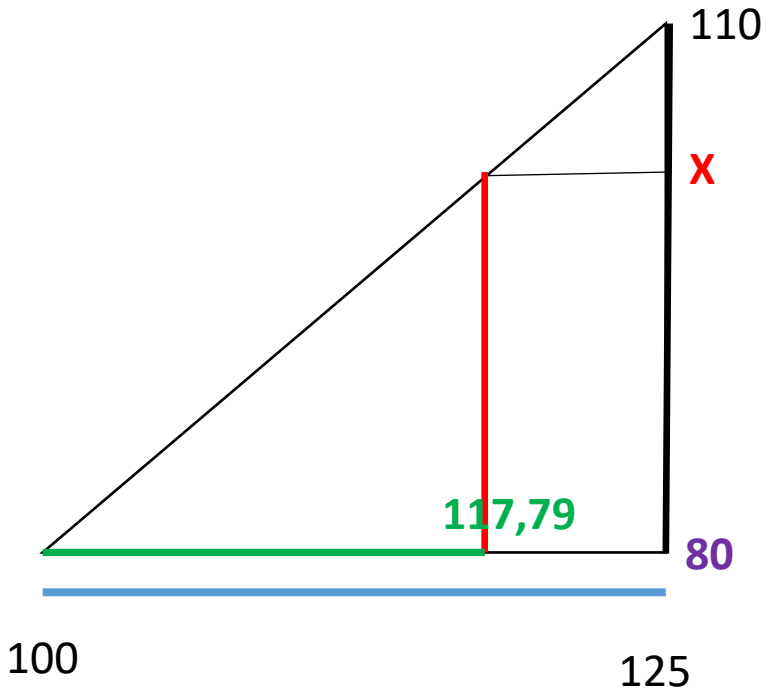
Calcul du nombre de terrains < 117,79 ares

117,79 ares est dans la classe [100, 125[

$$\frac{17{,}79}{X} = \frac{25}{30} \qquad \text{donc } X = \frac{30 \times 17{,}79}{25} = 21{,}35$$

Donc le nombre de terrain =

80 + 21,35 = 101,35



110 – 80 = 30

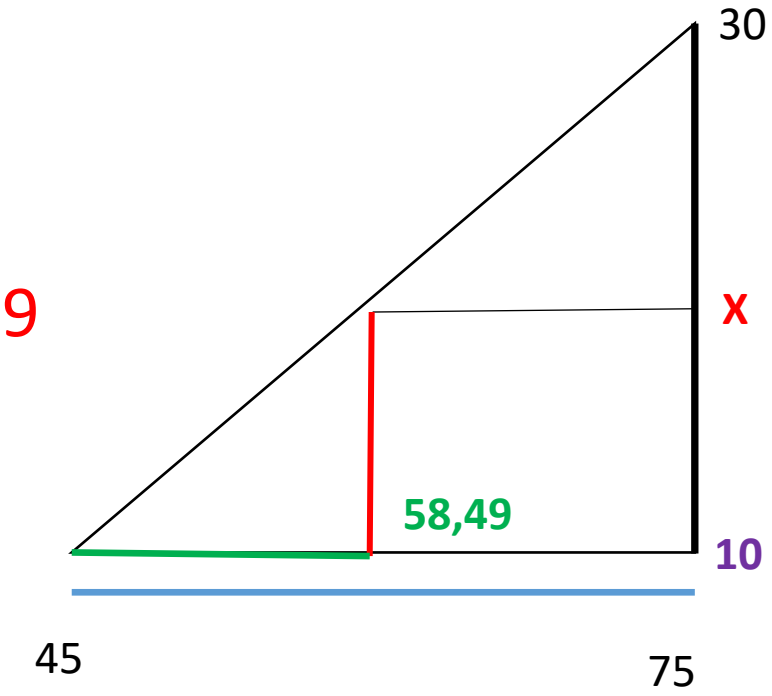
125 – 100 = 25

117,79 -100 = 17,79

Calcul du nombre de terrains < 58,49 ares

58,49 ares est dans la classe [45, 75[

$$\frac{13,49}{X} = \frac{30}{20} \qquad \text{donc } X = \frac{20 \times 13,49}{30} = 8,993 = 9$$



Donc le nombre de terrain =
10 + 9 = 19

30 – 10 = 20

75 – 45 = 30

58,49 - 45 = 13,49

101,35 – 19 = 82,35 terrains
Et 82,35 /120 = 68,62% des terrains à moins d'un Ecart Type
et donc 100 – 68,62 = 31,38% sont à plus d'un écart type.

Complément sur les Indicateurs de centralité et de dispersion

Comment une transformation linéaire de la variable observée X affecte-t-elle les caractéristiques de dispersion de la distribution sur X ?

L'idée à garder à l'esprit est la suivante : toutes les caractéristiques de dispersion se calculent à partir des écarts entre les observations, ou des écarts entre les observations et leur moyenne : il suffit donc de voir comment les transformations mentionnées affectent ces écarts ! Or une translation (+b) ne change pas les écarts!

Et donc :

- **Variance (a.X + b) = a^2. Variance(X)** le +b n'affecte pas la Variance…
- **Ecart type de (a.X + b) = a . Ecart type de X** ni l'écart type
- **Étendue de (aX + b) = a . étendue de X** ni l'étendue
- **Écart absolu moyen de (aX + b) = a . Écart absolu moyen de X** ni l'EAM

seule la multiplication affecte ces indices

Autre propriété:

Comment une transformation linéaire de la variable observée X affecte-t-elle les caractéristiques de localisation de la distribution sur X ?

L'idée à garder à l'esprit est la suivante : toutes les valeurs sont décalées (translation de la valeur +b) et multipliées par a! Or les variables de localisation sont censées représenter … le centre de la distribution Et donc :

Moyenne(aX+b) = a.Moyenne(X) +b

(c'est vrai pour toutes les caractéristiques de localisation, donc aussi pour la valeur modale et la médiane)

La cote Z ou note Z – Variable centrée réduite

Définition

Soit une observation prenant la valeur x et une série de valeurs dont la moyenne est notée μ et l'écart-type est noté σ.
La cote z de l'observation en question est simplement donnée par

$$Z = (x – \mu) / \sigma$$

Par cette opération, on « standardise » le résultat de l'observation.

- On la centre en lui retirant la moyenne de la série.
- On la réduit en la divisant par l'écart-type.

La cote Z ou note Z – Variable centrée réduite

A quoi ça sert?

En quelques mots, la cote z permet d'évaluer la qualité de performance d'une observation au sein d'un groupe.

Elle permet de déterminer si la valeur de cette observation se trouve plutôt dans la moyenne ou est plutôt extrêmement faible ou extrêmement élevée par rapport à la moyenne.

La cote Z ou note Z – Variable centrée réduite

A quoi ça sert?

Partons d'un exemple. Imaginons les résultats à deux tests différents de deux groupes de 10 étudiants provenant d'écoles différentes.

Ecole 1 : (12,12,12,13,14,14,15,15,15,16)

Ecole 2 : (7, 9,10,10,10,11,14,14,17,18)

Considérons deux étudiants, un de l'école 1 et un de l'école 2. Tous les deux ont une note de 14. Peut-on dire que leur performance est de même qualité ?
Probablement pas.
A première vue, un score de 14 est plus difficile à atteindre dans l'école 2 que dans l'école 1. **La cote z permet de formaliser cela.**

La cote Z ou note Z – Variable centrée réduite

A quoi ça sert?

En calculant la note $Z = (x – \mu) / \sigma$ nous standardisons le résultat de l'observation.

- On la **centre** en lui retirant la moyenne de la série. Dans notre exemple, ceci va annuler le fait que les notes sont en moyenne plus élevées dans l'école 1 que dans l'école 2.
- On la **réduit** en la divisant par l'écart-type. Dans notre exemple, ceci va annuler le fait que les notes sont beaucoup plus variables dans l'école 2 que dans l'école 1.

La cote z permet dès lors de déterminer si la performance de l'observation est normale ou exceptionnelle. Dans notre exemple, la cote Z va nous donner un nombre, comparable entre les deux écoles, nous permettant de juger si une note de 14 est une bonne performance ou pas. Dans notre exemple, une note de 14 amène une cote Z de 0.143 dans l'école 1 et de 0.587 dans l'école 2.

La cote Z ou note Z – Variable centrée réduite

Ecole 1 : (12,12,12,13,14,14,15,15,15,16) Moyenne: 13,8 σ = 1,40

Ecole 2 : (7, 9,10,10,10,11,14,14,17,18) Moyenne: 12 σ = 3,41

Observation : 14

Note Z école 1: 0,143 = (14 – 13,8) / 1,40

Note Z école 2: 0,587 = (14 – 12) / 3,41

Dans notre exemple, les deux côtes Z sont positives.
Ceci signifie qu'un 14 est un résultat au-dessus de la moyenne dans les deux écoles. De plus, la cote Z est supérieure dans l'école 2. Ceci signifie qu'un 14 est une meilleure performance dans l'école 2 que dans l'école 1!

La cote Z ou note Z – Variable centrée réduite

La cote z possède plusieurs propriétés intéressantes

- Elle ne dépend pas d'une unité de mesure. Dans notre exemple, si nous avions multiplié les notes de chaque étudiant par 2, le numérateur et le dénominateur de Z auraient été multipliés par 2 et le résultat aurait été le même.
- Si la cote Z est proche de 0, il s'agit d'une performance très proche de la moyenne (si la cote Z est exactement égale à 0, il s'agit d'une performance exactement égale à la moyenne).
- Si la cote Z est grande dans les positifs Z>>0, il s'agit d'une bonne performance, meilleure que la moyenne
- Si la cote Z est grande dans les négatifs Z<<0, il s'agit d'une mauvaise performance, moins bonne que la moyenne

La cote Z ou note Z – Variable centrée réduite: Lien avec la loi Normale

Il s'agit d'un cas particulier où la distribution de la série de valeurs se rapproche d'une distribution normale

- D'un point de vue informel, cela signifie que l'histogramme construit sur base de vos données ressemble à une courbe en cloche relativement symétrique.

Dans ce cas, la cote Z bénéficie d'interprétations supplémentaires.

- Environ 68% des valeurs ont une cote Z entre -1 et 1. Plus précisément, environ 16% des observations ont une cote Z > 1 et environ 16% des observations ont une cote Z < -1.
- Environ 95% des valeurs ont une cote Z entre -2 et 2. Plus précisément, environ 2,5% des observations ont une cote Z > 2 et environ 2,5% des observations ont une cote Z < -2.
- Environ 99.7% des valeurs ont une cote Z entre -3 et 3. Plus précisément, environ 0.15% des observations ont une cote Z > 3 et environ 0.15% des observations ont une cote Z < -3. Obtenir une cote Z de 3 est donc un résultat particulièrement exceptionnel.

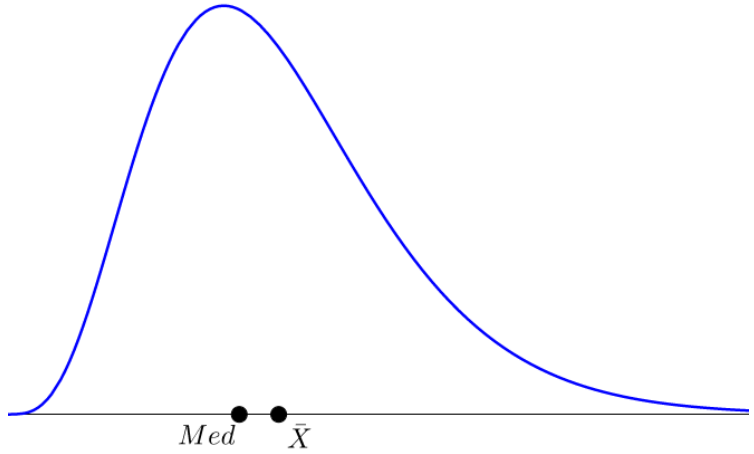
(38% pour +/-0,5 et 86% pour +/- 1,5)

Mesures d'asymétrie & Indicateur d'asymétrie

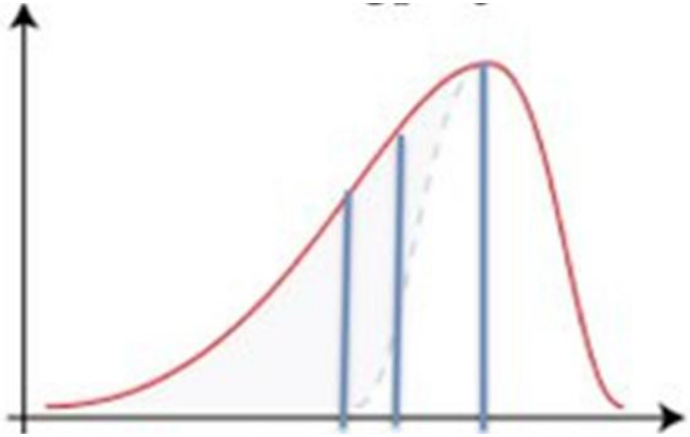
Comment définir un indicateur d'asymétrie ?

On aimerait bien qu'il soit :

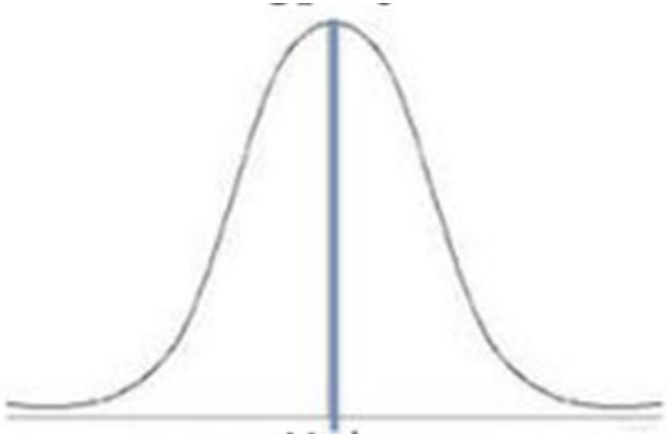
- nul si la distribution est symétrique
- Proportionnel à l 'asymétrie: au plus grand, au plus asymétrique
- Par exemple de signe différent si l'asymétrie est orientée plutôt à gauche de la moyenne, ou à droite



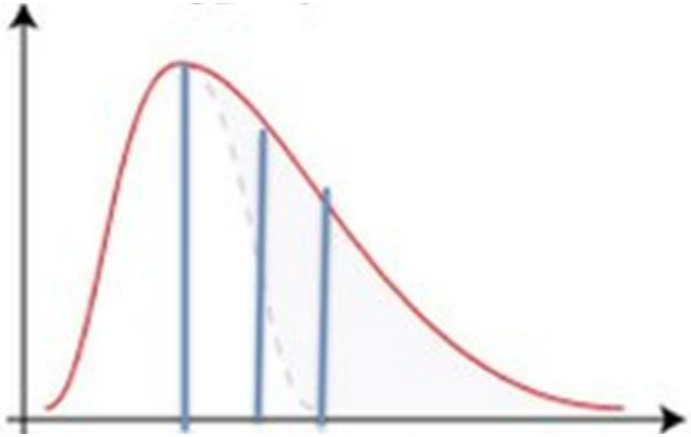
Types d’asymétries



Courbe étalée à gauche ou oblique à droite



Courbe symétrique par rapport à la moyenne



Courbe étalée à droite ou oblique à gauche

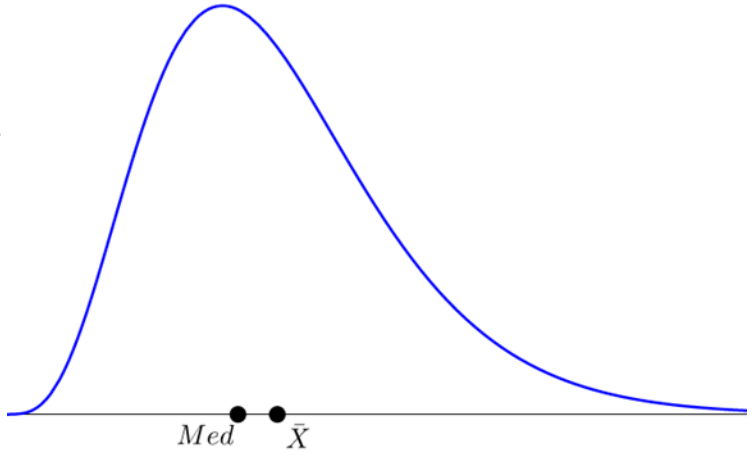
Commençons par déterminer s'il y a présence ou non d'une asymétrie ?

1ère méthode

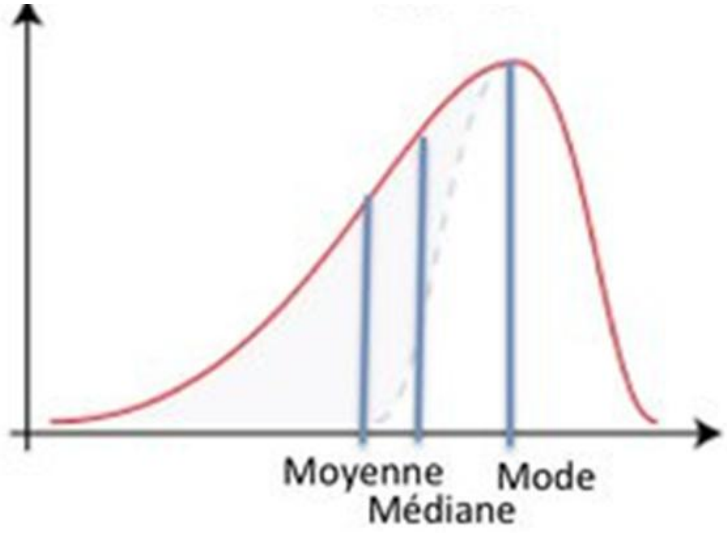
Pour se faire une première idée, on a souvent tendance à regarder la position de la médiane par rapport à la moyenne.

En effet, si Médiane < Moyenne, on s'attend à ce que la distribution soit plutôt oblique à gauche, c'est-à-dire étalée à droite (on parle alors d'asymétrie positive).

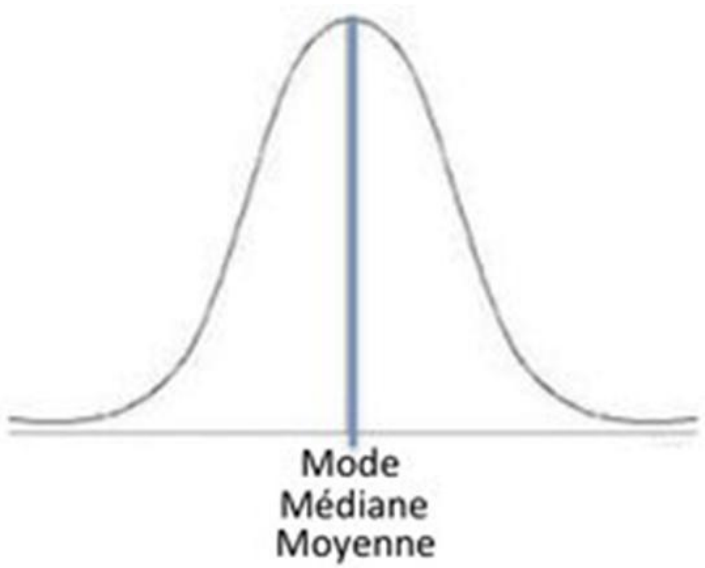
Si, au contraire, Médiane > Moyenne, on s'attend à ce que la distribution soit plutôt oblique à droite et étalée à gauche (on parle alors d'asymétrie négative).



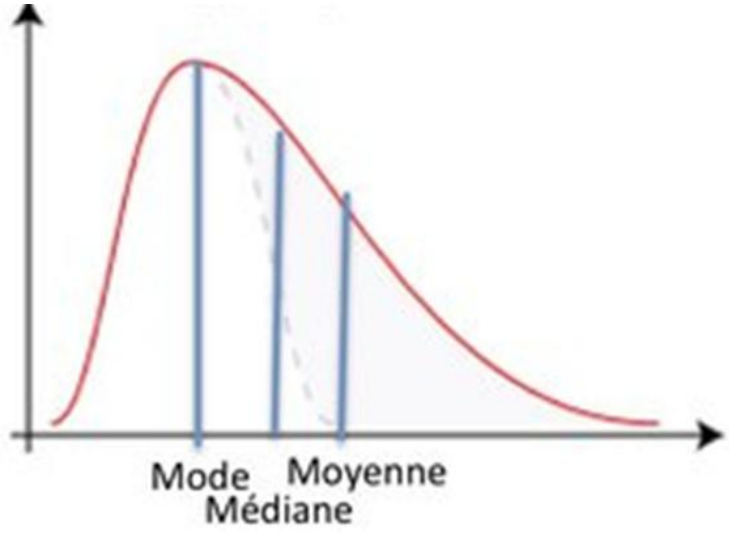
Relation Moyenne, Mode, Médiane et asymétrie



Asymétrie en j

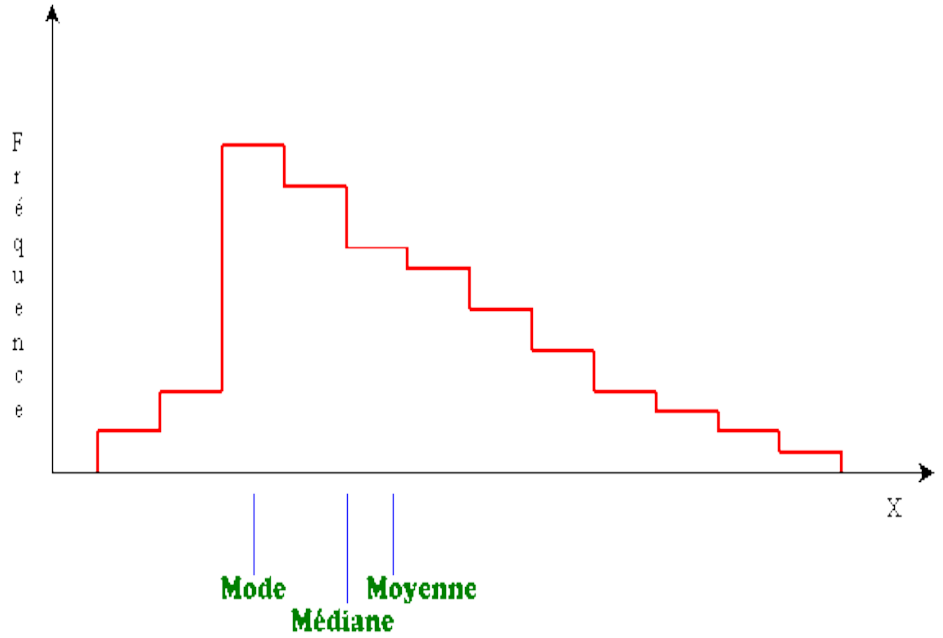


Courbe normale

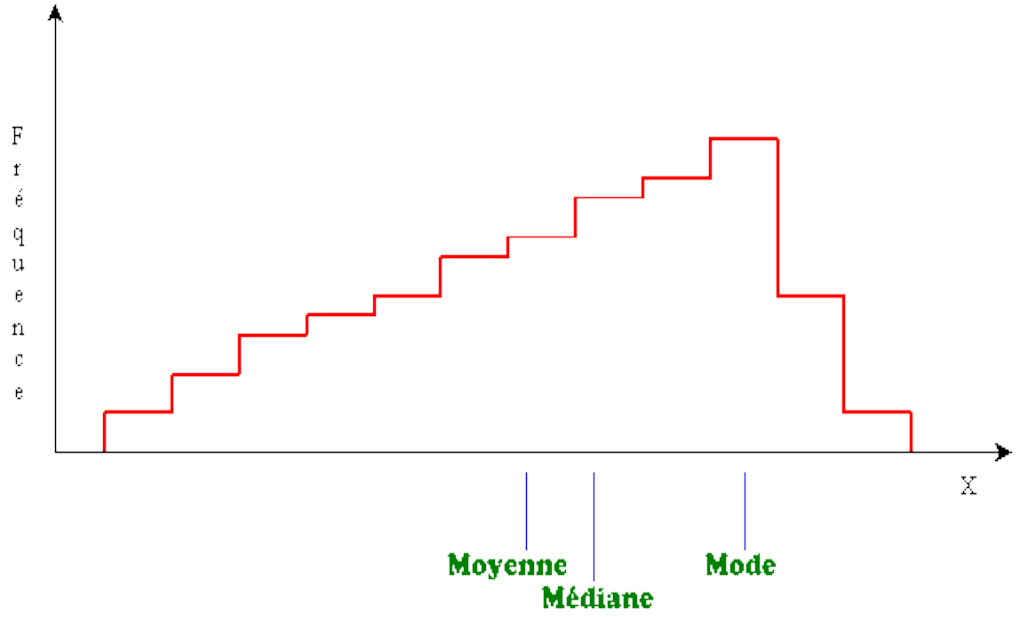


Asymétrie en i

Distribution asymétrique



A. Distribution étalée à droite:



B. Distribution étalée à gauche:

Si les valeurs extrêmes sont modifiées, la médiane ne change pas car elle n'est pas sensible aux valeurs extrêmes.
Par contre la moyenne change car elle tient compte de toutes les valeurs.
-> une ou quelques valeurs extrêmes peuvent rendre notre indicateur moins bon…

2ème méthode: Le coefficient de Yule

C’est le coefficient le plus intuitif… ;-)

$$S_Y = \frac{(Q3 - Med) - (Med - Q1)}{(Q3 - Med) + (Med - Q1)}$$

$$S_Y = \frac{(Q3 - Med) - (Med - Q1)}{(Q3 - Med) + (Med - Q1)}$$

Rappel: Médiane = Q_2 et $Q_1 < Q_2 < Q_3$;

donc la quantité au dénominateur de Sy sera toujours positive.

Le numérateur s'annule quand

$(Q_3 - Med) - (Med - Q_1) = 0$ soit : $(Q_3 - Med) = (Med - Q_1)$

qui est vrai si : $Med = (Q_1 + Q_3) / 2$

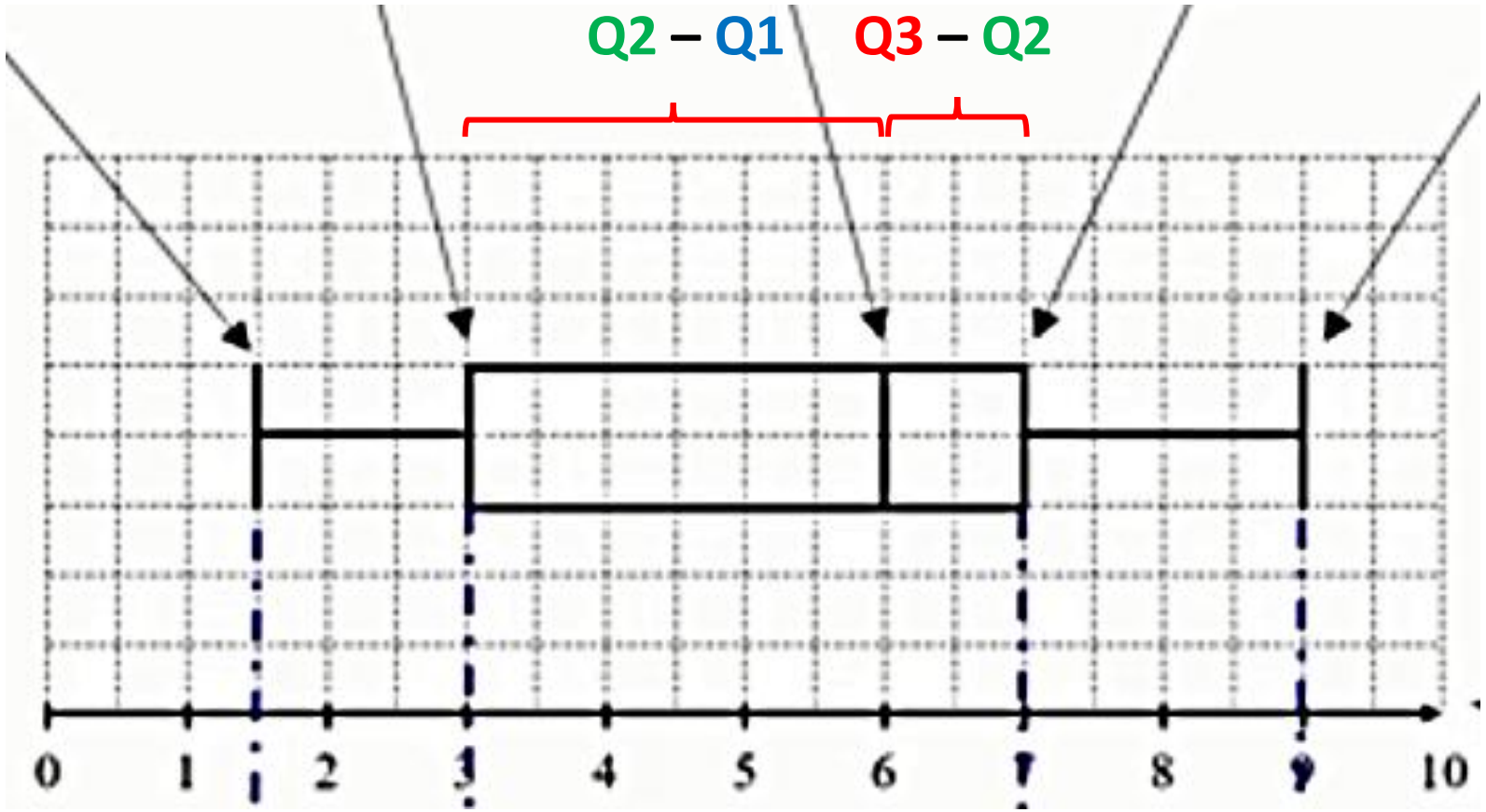
Ce qui signifie qu'il y a autant d'observations en dessous de Q1 qu'au-dessus de Q3.

La distribution est alors parfaitement symétrique.

Si $(Q_3 - Med) - (Med - Q_1) < 0$ Alors $(Q_3 - Med) < (Med - Q_1)$

qui est vrai si : $Med > (Q_1 + Q_3) / 2$

Cela signifie plus d'observations vers les fortes valeurs donc la distribution est étalée à gauche



(Q3 – Q2) + (Q2 – Q1)

= Q3 – Q1

Le coefficient de Yule mesure la dissymétrie de la position de la médiane (Q2) par rapport à Q3 et Q1 proportionnellement à l'écart interquartile (Q3-Q1)

$$S_Y = \frac{(Q3 - Med) - (Med - Q1)}{(Q3 - Med) + (Med - Q1)}$$

Le coefficient de Yule

On peut donc résumer :

- **Si Sy = 0 alors la distribution est symétrique**
- **Si Sy < 0 alors la distribution est étalée à gauche (oblique à droite)**
- **Si Sy >0 alors la distribution est étalée à droite (oblique à gauche)**

$$S_Y = \frac{(Q3 - Med) - (Med - Q1)}{(Q3 - Med) + (Med - Q1)}$$

Exemple pour l'exercice 3.6

Coefficient de Yule:

$$S_Y = \frac{(Q3 - Med) - (Med - Q1)}{(Q3 - Med) + (Med - Q1)}$$

$$\frac{(133{,}337 - 113{,}33) - (113{,}333 - 105)}{(133{,}337 - 113{,}33) + (113{,}333 - 105)} = 0{,}411$$

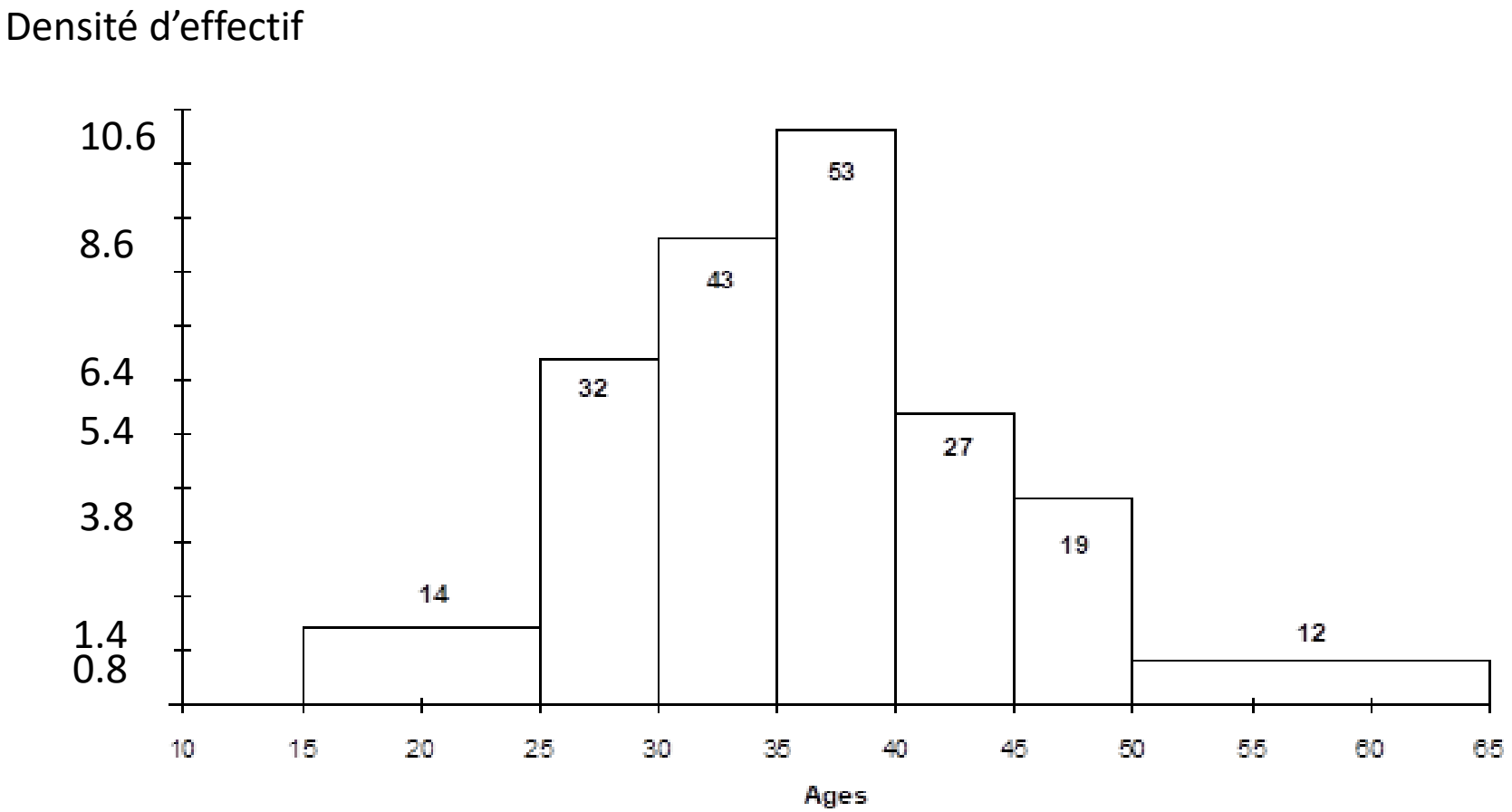
Donc presque symétrique, très légèrement étalée vers la droite

Comparaison Médiane / Moyenne

113,333 ares < 115,313 ares Médiane < Moyenne mais pas de beaucoup

Donc très légèrement étalée vers la droite (oblique à gauche)

2ème Exemple: que dire de la symétrie de la distribution dans cet exercice?



Histogramme de la distribution de l'âge des employés d'une entreprise correspondant au tableau ci-dessous

[15,25[	[25,30[	[30,35[	[35,40[	[40,45[	[45,50[	[50,65[
14	32	43	53	27	19	12

Effectif total: 200 (14 + 32 + 43 + 53 + 27 + 19 + 12 = 200)

Q1: 200 x 0,25 = 50 -> classe [30, 35[
Q2 = Médiane : 200 x 0,5 = 100 -> classe [35, 40[
Q3 : 200 x 0,75 = 150 -> classe [40, 45[

$$Q_1 = 30 + \frac{5 \text{ x } 4}{43} = 30{,}46$$

$$Q_2 = 35 + \frac{5 \times 11}{53} = 36{,}04$$

$$Q_3 = 40 + \frac{5 \text{ x } 8}{27} = 41{,}48$$

Étendue: 64 – 15 = 49 ans

Classe modale: [35-40[

Valeur modale: 37,5 ans

Médiane (à partir du graphe cumulé): +/-36 ans

$$\bar{x} = \frac{\sum n_i x_i}{\sum n_i} = \frac{7285}{200} = 36{,}43 \text{ ans}$$

Q1 = 30,46

Médiane = Q2 = 36,04

Q3 = 41,48

Coefficient de Yule:

$$S_Y = \frac{(Q3 - Med) - (Med - Q1)}{(Q3 - Med) + (Med - Q1)}$$

$$Sy = \frac{(41{,}48 - 36{,}04) - (36{,}04 - 30{,}46)}{(41{,}48 - 36{,}04) + (36{,}04 - 30{,}46)} = -0{,}014$$

Donc pratiquement symétrique, très, très légèrement étalée vers la gauche

Comparaison Médiane / Moyenne

36,04 < 36,43 Médiane = Moyenne

Donc pratiquement symétrique, très, très légèrement étalée vers la droite

Boîte à moustache

(Box-plot ou boîte de Tukey)

Boîte à moustaches ou Box-plot

La boîte à moustaches est une représentation qui permet de présenter graphiquement les principaux paramètres d'une distribution : étendue (modifiée ou non), premier quartile, médiane (second quartile), troisième quartile, intervalle interquartile. Et de détecter sous certaines hypothèses la présence de données aberrantes.

En statistique, une **donnée aberrante** (ou horsain, en anglais *outlier*) est une valeur ou une observation qui est très « distante » des autres observations effectuées sur le même phénomène, c'est-à-dire qu'elle contraste grandement avec les valeurs « normalement » mesurées.

C'est pourquoi on rajoute un trait à + et – 1,5 fois la valeur interquartile de part et d'autre de Q1 et Q3 (moustaches) plutôt que d'utiliser l'étendue stricte comme extrémités des moustaches.

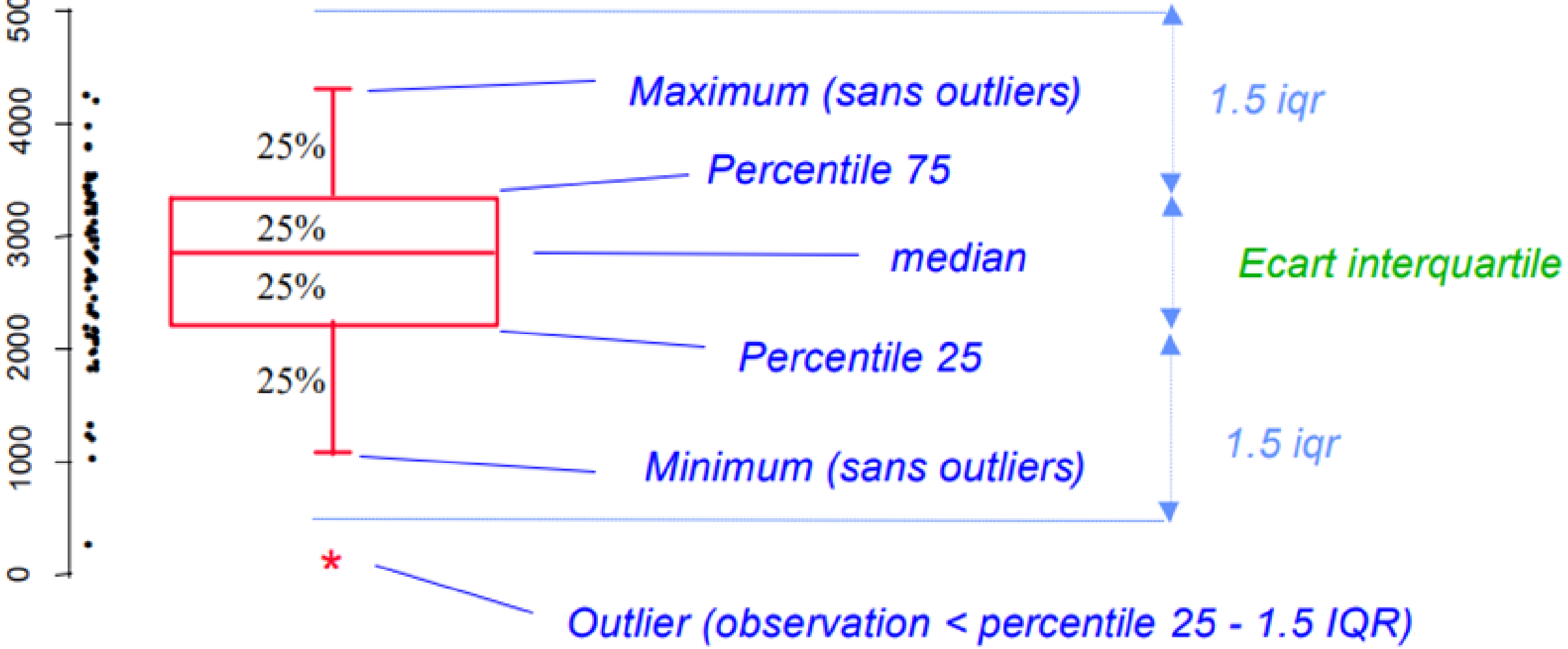
Boîte à moustaches ou Box-plot

Pourquoi utilise-t-on 1,5 pour les extrémités des moustaches ?

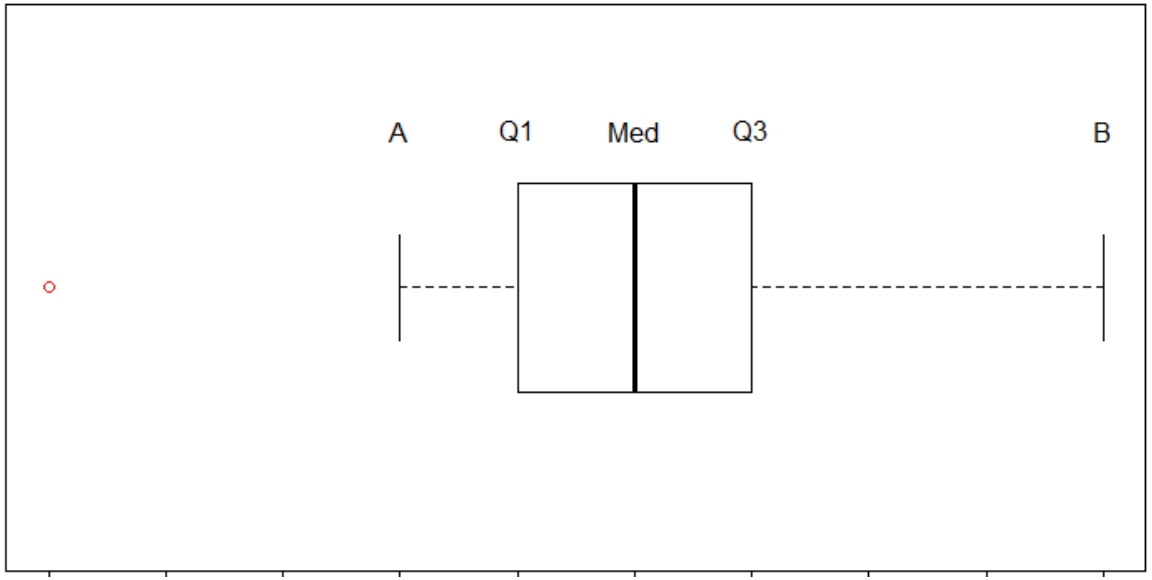
Cela vient du fait que le modèle est basé sur la distribution d'une loi normale. Si une variable suit une loi normale, alors l'intervalle entre les extrémités des moustaches devrait contenir 99,3 % des observations, c'est-à-dire que l'on devrait trouver 0.7 % d'observations en dehors de cet intervalle, que l'on considère alors comme des valeurs aberrantes.

Pour le créateur de cette représentation, John Tukey, 1,5 était un bon compromis pour observer assez de points aberrants, sans pour autant en être débordé.

Voilà ce que cela donne:



La boîte à moustache est le plus souvent présentée à l'horizontale comme ceci :



Comme vous pouvez le constater sur cet exemple, le point rouge est un point potentiellement aberrant à étudier, car il est situé en dehors du rectangle et des moustaches sous l'hypothèse de normalité. A est la valeur la plus petite observée mais supérieure à Q1 – 1,5 x (Q3 – Q1) et B est la valeur observée la plus grande mais inférieure à Q3 + 1,5 x (Q3 – Q1)

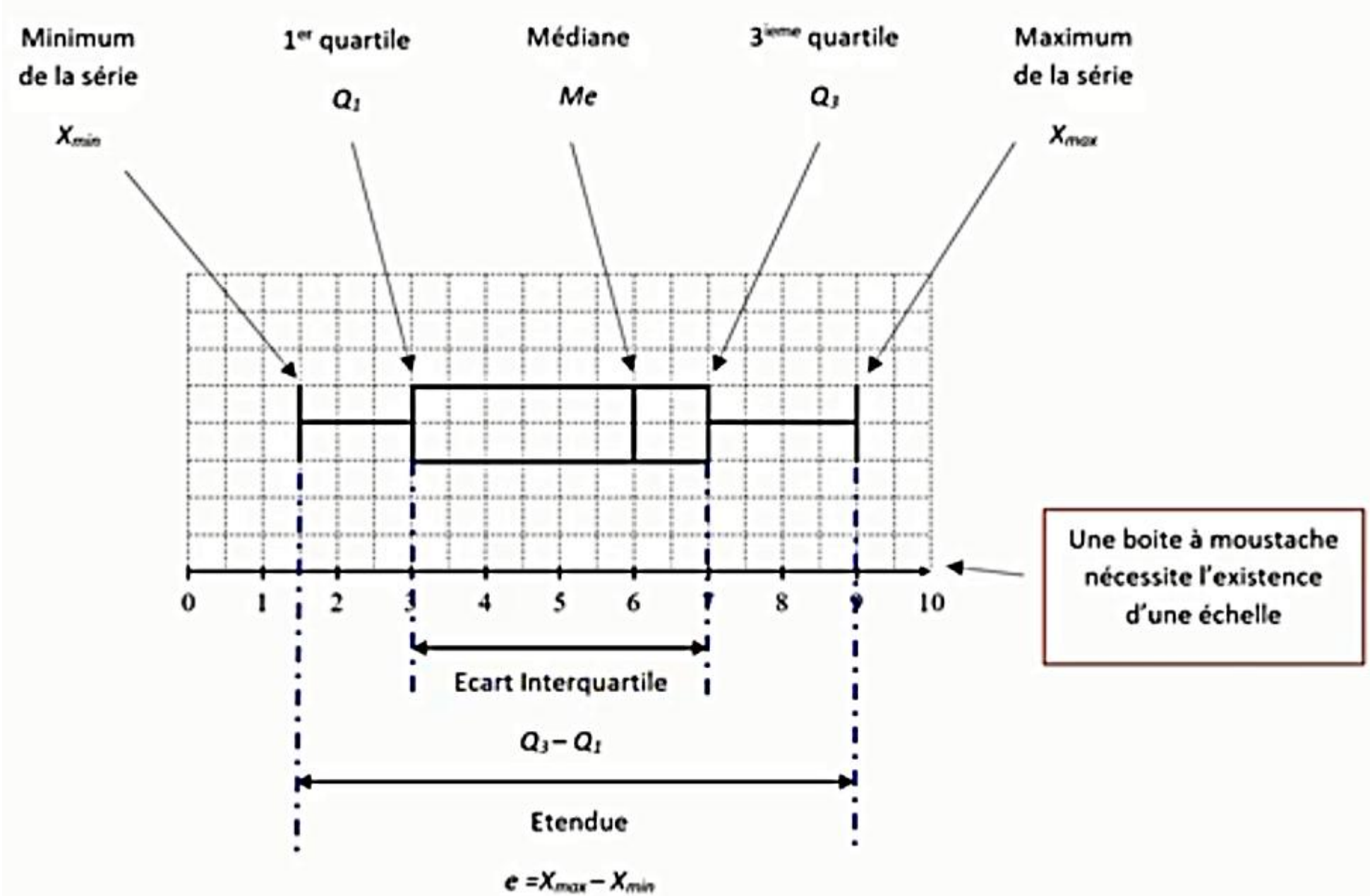
Construction de la boîte à moustaches :

- On construit une boîte rectangulaire entre Q1 et Q3 (donc de longueur égale à l'intervalle interquartile EIQ)
- On détermine ensuite la longueur des moustaches :
  l'extrémité de la moustache inférieure A est la plus petite valeur x_i telle que $x_i >= Q_1 – 1,5 \times EIQ$
- L'extrémité de la moustache supérieure B est la plus grande valeur x_i telle que $x_i <= Q_3 + 1,5 \times EIQ$

En résumé

La valeur centrale du graphique est la médiane (il existe autant de valeurs supérieures qu'inférieures à cette valeur dans l'échantillon, 50/50).
Les bords du rectangle sont les quartiles (Pour le bord inférieur, un quart des observations ont des valeurs plus petites et trois quart ont des valeurs plus grandes (25/75), le bord supérieur suit le même raisonnement (donc 75/25).
Les extrémités des moustaches sont calculées en utilisant 1.5 fois l'espace interquartile (la distance entre le 1er et le 3ème quartile) à partir des bords du rectangle.
On peut remarquer que 50% des observations se trouvent à l'intérieur de la boîte.

Les valeurs à l'extérieur des moustaches sont représentées par des points. On ne peut pas dire que si une observation est à l'extérieur des moustaches alors elle est une valeur aberrante. Par contre, cela indique qu'il faut étudier plus en détail cette observation.

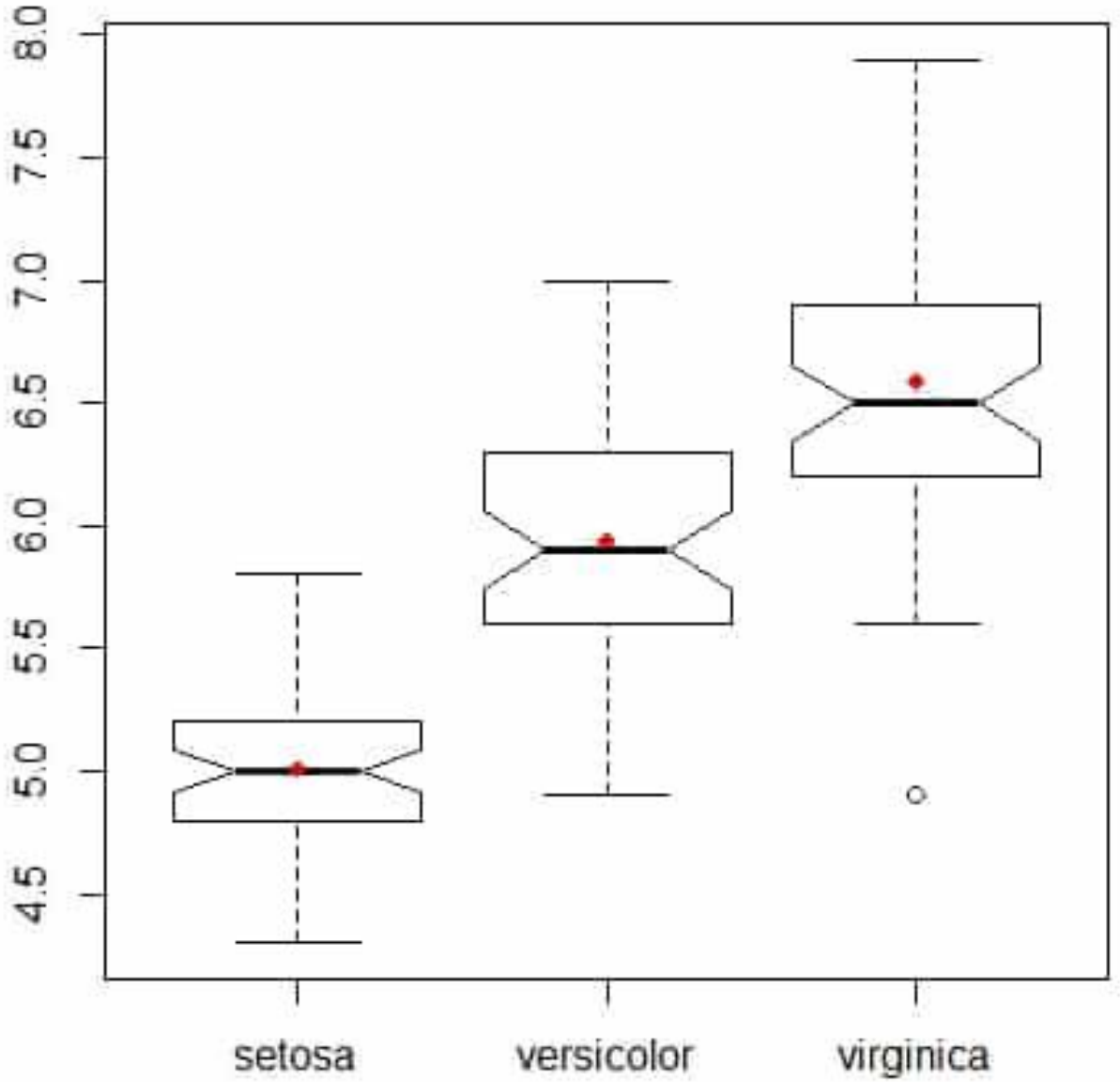

Minimum de la série
X_{min}
1er quartile
Q_1
Médiane
Me
3ième quartile
Q_3
Maximum de la série
X_{max}
0 1 2 3 4 5 6 7 8 9 10
Une boite à moustache nécessite l'existence d'une échelle
Ecart Interquartile
$Q_3 - Q_1$
Etendue
$e = X_{max} - X_{min}$

Variantes

On voit parfois apparaître des boîtes à moustaches avec des formes différentes ou des signes supplémentaires, en voici quelques-uns :

- Une croix rouge dans la boîte : lorsqu'une croix rouge apparaît dans le box-plot, il s'agit toujours d'une représentation de la moyenne sur l'échantillon étudié.
- Des boîtes ayant des largeurs variables : il arrive souvent que les boîtes à moustache lors d'une comparaison n'aient pas la même taille (en largeur), il ne s'agit pas d'une simple transformation esthétique, la largeur est alors proportionnelle à la taille de l'échantillon. Ceci est spécialement intéressant dans le cas de comparaison de groupes d'observations pour lesquelles la taille des groupes n'est pas homogène.
- Des boîtes avec une largeur qui se resserre autour de la médiane (notched) : Cette représentation permet de visualiser un intervalle de confiance à 95% autour de la médiane. Les points où la boîte se resserre représentent les bornes de cet intervalle. On le calcule avec la formule suivante :
  médiane +/- 1.57 * (Q3 - Q1)/racine(N) avec N taille de l'échantillon.

Exemple:



Utilité:

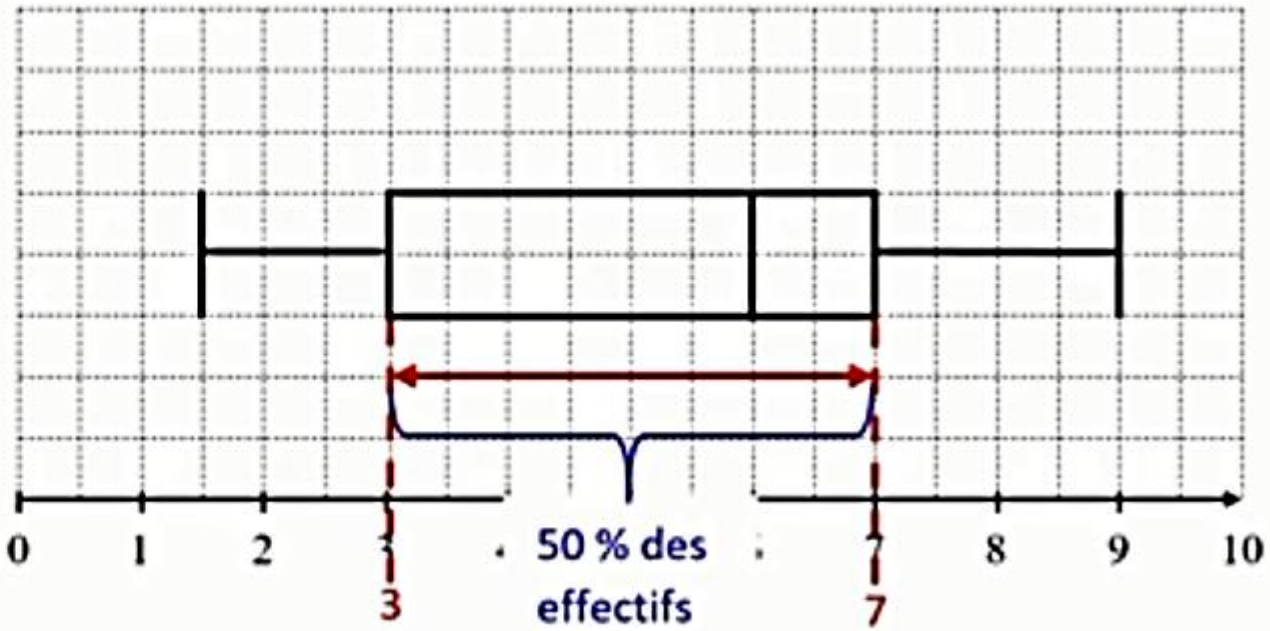
Les boîtes à moustache permettent de facilement comparer la distribution d'une variable quantitative sur plusieurs populations différentes.

(cfr. Diapo précédente et suivante)

Exemple:

Un professeur a recensé les notes obtenues par ses élèves lors du dernier contrôle et présente les résultats sous la forme de la boite à moustache ci-contre.

Quelles informations peut-on extraire de cette représentation graphique ?



	Note (/10)
Minimum X_{min}	1,5
1er quartile Q_1	3
Médiane *Med*	6
3e quartile Q_3	7
Maximum X_{max}	9
Etendue *e*	9 – 1,5 = 7,5
$Q_3 - Q_1$	7 – 3 = 4

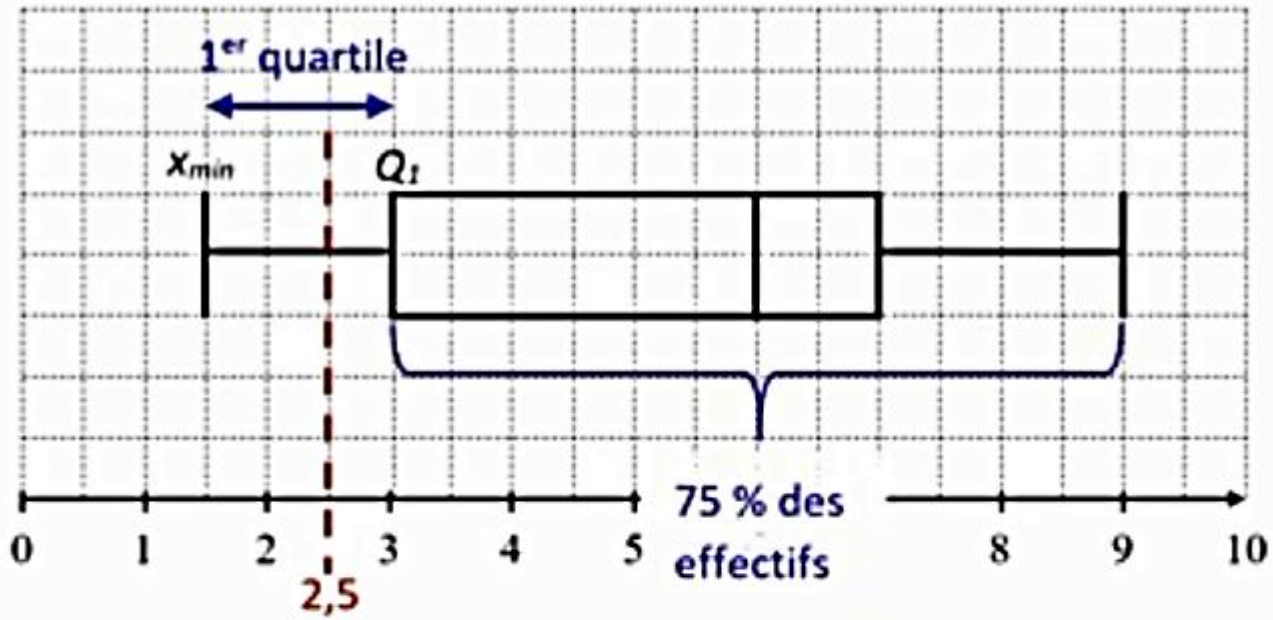
- 25 % des élèves ont eu moins de 3/10
- La moitié des élèves ont eu moins de 6/10
- 75 % des élèves ont eu moins de 7/10
- $Q_3 - Q_1 = 7 - 3 = 4$

50 % des élèves se tient dans un intervalle de 4 points.

https://www.youtube.com/watch?v=VUEAoPBQgJI

Un professeur a recensé les notes obtenues par ses élèves lors du dernier contrôle et présente les résultats sous la forme de la boite à moustache ci-contre.

Quelles informations peut-on extraire de cette représentation graphique ?



Notes obtenues :

Pierre : 2,5 / 10

Marie : 8,5 / 10

Jeanne: 4,5 / 10

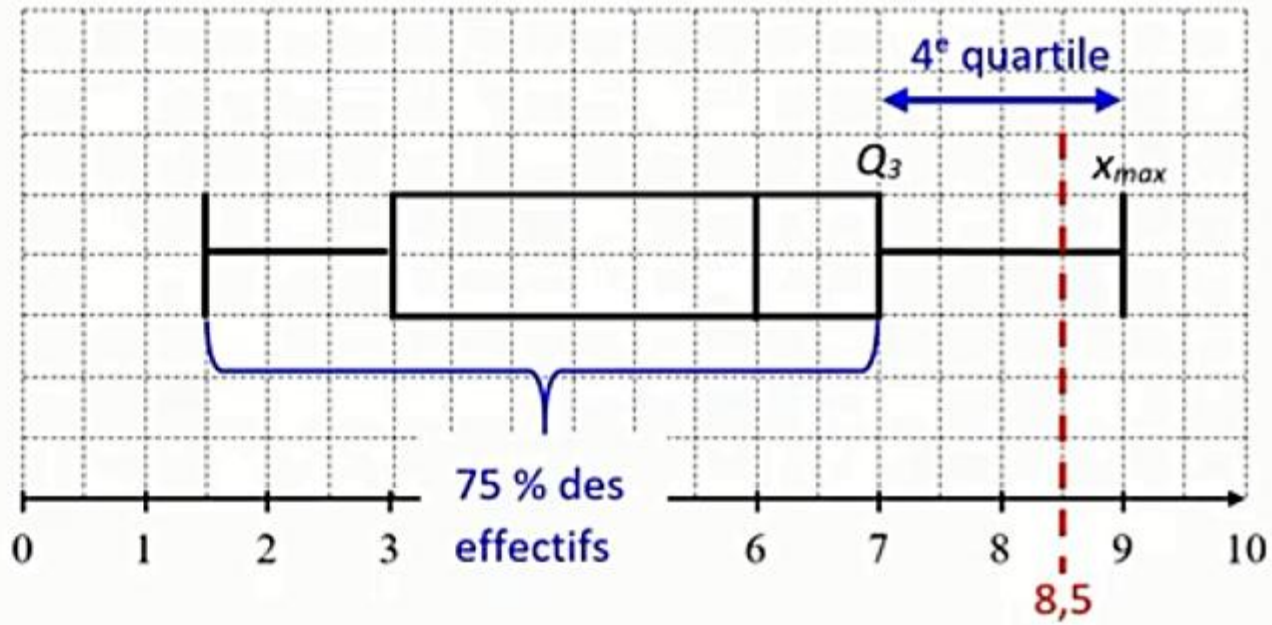
Paul : 6,5 / 10

- Pierre se situe dans le premier quartile.

Au moins 75 % des élèves ont obtenu une note supérieure à la sienne

Un professeur a recensé les notes obtenues par ses élèves lors du dernier contrôle et présente les résultats sous la forme de la boite à moustache ci-contre.

Quelles informations peut-on extraire de cette représentation graphique ?



<u>Notes obtenues</u> :

Pierre : 2,5 / 10

Marie : 8,5 / 10

Jeanne: 4,5 / 10

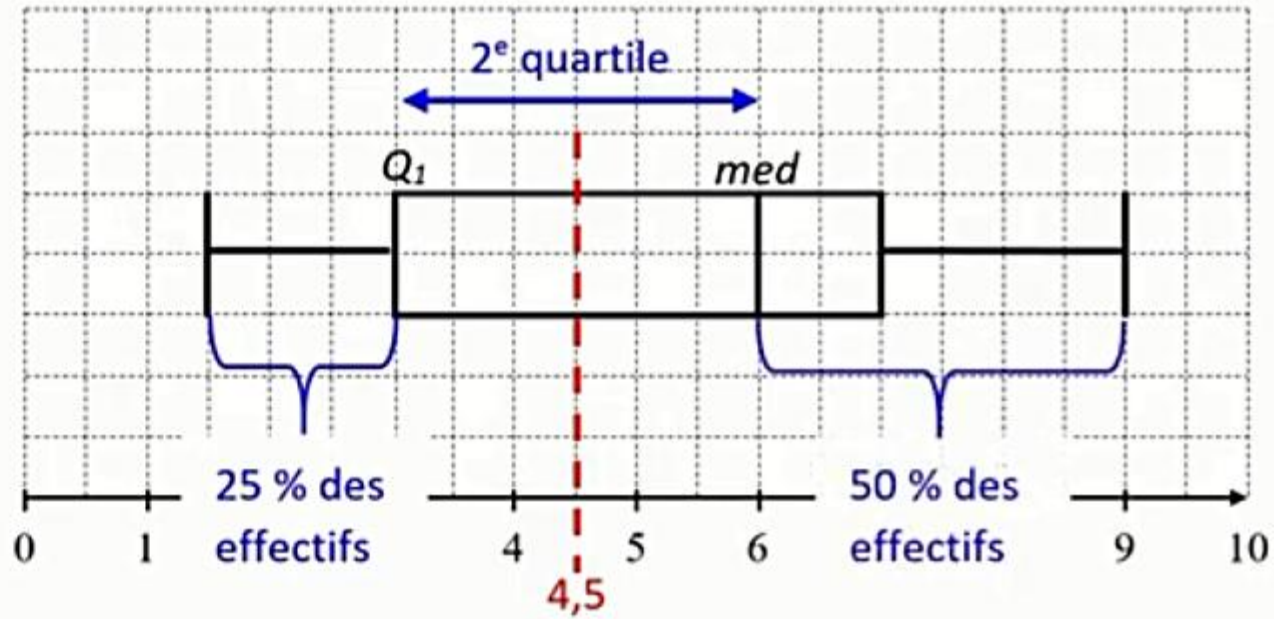
Paul : 6,5 / 10

- Marie se situe dans le quatrième quartile.

Au moins 75 % des élèves ont obtenu une note inférieure à la sienne

Un professeur a recensé les notes obtenues par ses élèves lors du dernier contrôle et présente les résultats sous la forme de la boite à moustache ci-contre.

Quelles informations peut-on extraire de cette représentation graphique ?



Notes obtenues :

Pierre : 2,5 / 10

Marie : 8,5 / 10

Jeanne: 4,5 / 10

Paul : 6,5 / 10

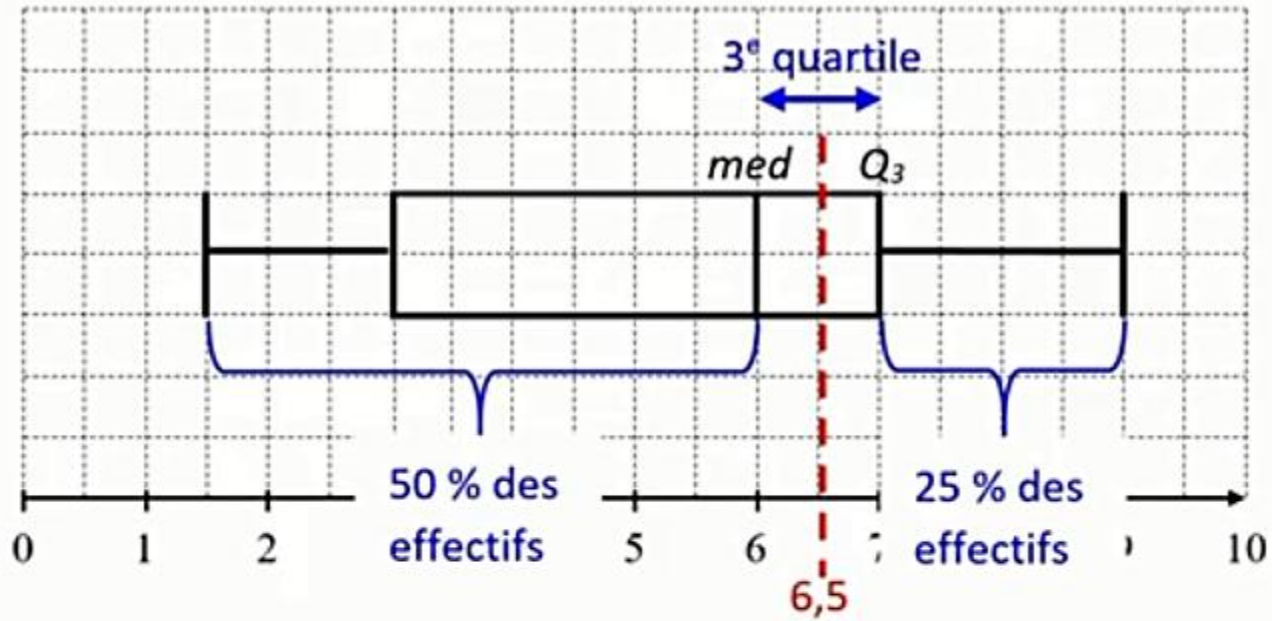
- Jeanne se situe dans le deuxième quartile.

Au moins 25 % des élèves ont obtenu une note inférieure à la sienne

Au moins 50 % des élèves ont obtenu une note supérieure à la sienne

Un professeur a recensé les notes obtenues par ses élèves lors du dernier contrôle et présente les résultats sous la forme de la boite à moustache ci-contre.

Quelles informations peut-on extraire de cette représentation graphique ?



<u>Notes obtenues</u> :

Pierre : 2,5 / 10

Marie : 8,5 / 10

Jeanne: 4,5 / 10

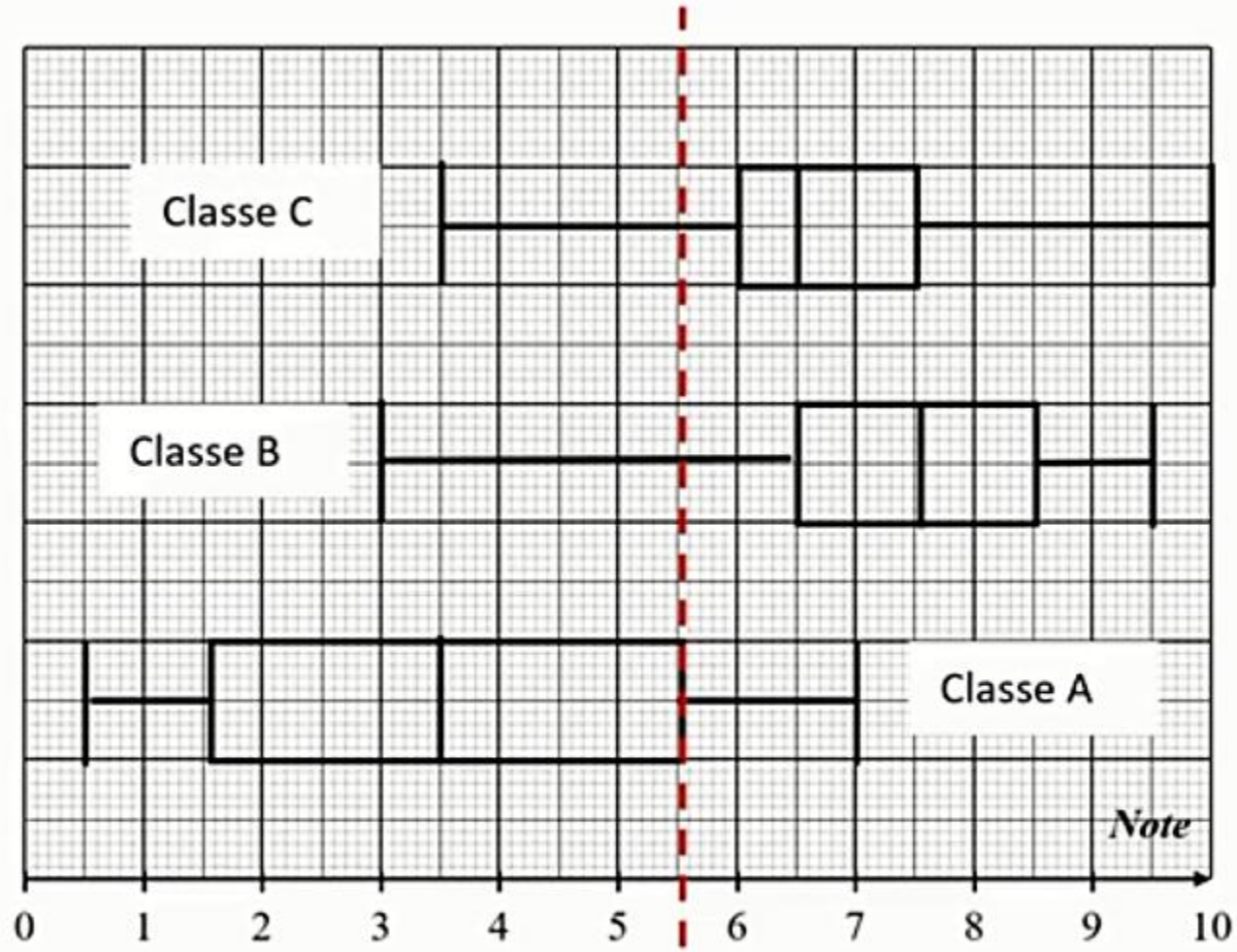
Paul : 6,5 / 10

- Paul se situe dans le troisième quartile.

Au moins 50 % des élèves ont obtenu une note inférieure à la sienne

Au moins 25 % des élèves ont obtenu une note supérieure à la sienne

Trois classes d'un lycée ont composé en même temps sur le même sujet de mathématiques. Les résultats obtenus par ces trois classes sont compilés dans le diagramme ci-contre :



- La plus basse note est de 0,5/10

  Elle a été obtenue dans la classe A

- La plus haute note est de 10/10

  Elle a été obtenue dans la classe C

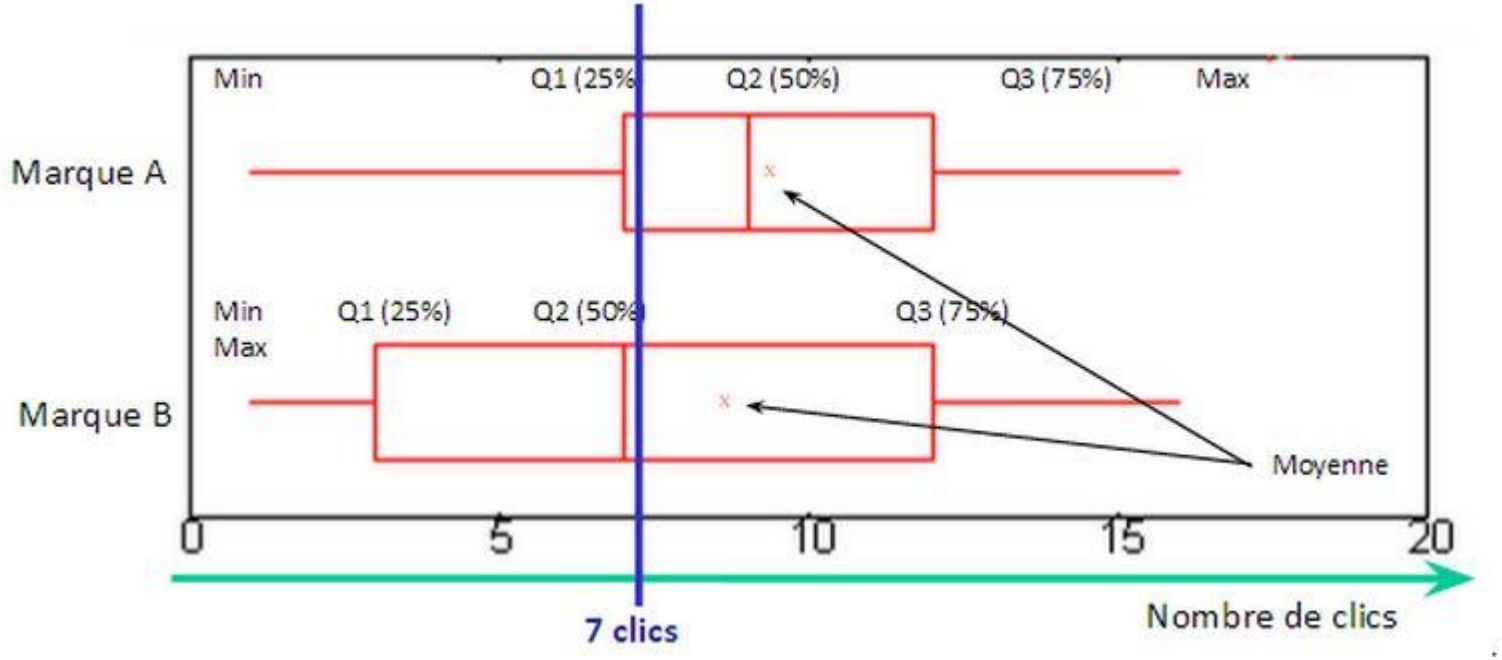
- C'est la classe B qui a globalement obtenu les meilleurs résultats

- La classe C est la classe la plus homogène

- Les résultats de la classe A sont nettement en retrait par rapport à ceux des classes B et C

Exemple :

Comparer le nombre de clics sur les campagnes publicitaires en fonction de la marque



- Interprétation : 50% des campagnes de B ont généré moins de 7 clics, alors que 75% des campagnes de A en ont produit plus de 7.
- Ici, la marque B a tendance à avoir un nombre moins important de clics que la marque A.

Les liens intéressants

Des descriptions des boîtes à moustaches peuvent être trouvées sur beaucoup de sites web. En voici quelques unes :

- https://www.stat4decision.com/fr/le-box-plot-ou-la-fameuse-boite-a-moustache/
- Cours de l'ITSE
- Une présentation intéressante lors des Semin-R du MNHN sur les box-plots avec R

Les références

S'il fallait n'en citer qu'une, ça serait :

John W. Tukey (1977). Exploratory Data Analysis. Addison-Wesley.

Pour la version notched, on peut voir :

John M. Chambers (1983). Graphical methods for data analysis. Wadsworth International Group.

Exercices pour la semaine prochaine:

- Faire les boîtes à moustaches des exercices 3.6 et 3.7 et de l’exercice de la semaine passée basé sur la distribution de l’âge des employés d’une entreprise