Cours 1 – Notions d'échantillonnage aléatoire.

Eya ZOUGAR [1]

Institut National des Sciences appliqués-INSA

Génie mathématiques GM3
Thursday 19th January, 2023

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN NORMANDIE

[1] Basé sur le cours de Bruno PORTIER

Contents

1. Introduction
1.1. L'objectif du cours

L'objectif de ce cours est:

- d'introduire la notion d'échantillonnage aléatoire
- d'appréhender la notion de fluctuations d'échantillonnage
- de différencier la statistique descriptive de la statistique inférentielle.

1.2. Problématique

On considère une population de N individus:

• On s'intéresse à une caractéristique de cette population, inconnue à priori.
• Cette caractéristique est une fonction d'un ou plusieurs caractères présents sur chaque individu.
• Il y aura deux types de caractéristique:

- Type quantitatif (taille, poids, salaire, etc ...) et/ou
- Type qualitatif (couleur des yeux, sexe, nationalité, etc ...).

Exemples:

1. On peut s'intéresser à la note moyenne dans l'examen de statistique pour un classe de GM3.
2. On peut s'intéresser à la taille moyenne des individus dans l'Europe.

La question qui se pose, alors: Comment déterminer ces valeurs ?

il y a deux possibilités:

- effectuer un recensement (enquête exhaustive)
  1. On obtient la note moyenne dans l'examen de statistique, en déterminant la note de chaque étudiant du classe de GM3.
  2. On obtient la taille moyenne, en mesurant la taille de chaque individu composant la population.
- procéder à un échantillonnage de la population, c'est à dire faire seulement l'étude sur une partie de la population (enquête partielle).
  1. On fait l'etude sur un groupe du classe de GM3.
  2. On fait l'etude sur un ensemble des individus de la population.

1.3. Le recensement

On mesure tous les individus de la population et on en déduit la valeur de la caractéristique considérée.

C'est idéalement la meilleure solution (solution déterministe)!

Cependant, une telle solution peut s'avérer impossible pour plusieurs raisons :

1. La population est infinie, ou presque.
2. N est trop grand pour que l'étude statistique soit réalisable en un temps et avec un coût raisonnables.
3. La mesure peut prendre tellement de temps que la caractéristique étudiée peut avoir changée au cours du temps.
4. La mesure détruit l'individu. C'est le cas lorsque l'on s'intéresse par exemple, à la durée de vie des ampoules dans un lot; si l'on teste toutes les ampoules pour obtenir la durée de vie moyenne de toutes les ampoules du lot, on n'aura plus d'ampoules à vendre...

1.4. L'échantillonnage
1.4.1. Le principe

Quand on ne peut pas effectuer un recensement, l'autre solution consiste à mesurer une partie seulement de la population.

On prélève donc un échantillon de cette population et on extrapole le résultat obtenu sur l'échantillon à l'ensemble de la population.

C'est l'approche du statisticien: il n'est pas nécessaire de manger tout le boeuf pour se rendre compte que la viande n'était pas tendre...

Cette approche offre une série d'avantages, mais présente aussi des inconvénients.

1.4.2. Les avantages de l'échantillonnage

Par rapport au recensement, l'échantillonnage offre un certain nombre d'avantages :

- Le coût global de l'échantillonnage est en général plus réduit que le coût global d'un recensement ;
- L'étude sur l'échantillon est plus rapide que l'étude sur toute la population, surtout lorsque la caractéristique étudiée présente des modifications assez importantes au cours du temps ;
- Les erreurs d'observations sont plus réduites que dans l'enquête exhaustive ;
- Enfin dans certaines situations particulières, l'échantillonnage est la seule solution possible, c'est le cas lorsque la mesure est destructive.

1.4.3. Les inconvénients de l'échantillonnage.

Néanmoins, l'échantillonnage présente des inconvénients :

- On est amené à n'étudier qu'une partie de la population et si l'échantillon n'est pas **représentatif** de la population, l'extrapolation des résultats à la population sera plus ou moins juste ; une des difficultés qui se présentent est de savoir comment définir l'échantillon pour qu'il soit représentatif de la population.
- Il faut disposer d'une liste de tous les individus (ou d'un moyen d'accès à n'importe quel individu dans la population), chacun ayant la même chance que les autres d'être sélectionné pour faire partie de l'échantillon.
- Les résultats obtenus peuvent varier d'un échantillon à l'autre.

1.5. Statistique descriptive versus Statistique inférentielle.

La statistique descriptive ne s'intéresse qu'à la sous-population formée par l'échantillon avec comme objectif de décrire et/ou résumer l'information contenue dans l'échantillon, à l'aide de tableaux, graphiques et indicateurs.

La statistique inférentielle s'intéresse à la population dont est issu l'échantillon avec comme objectif d'inférer, à partir des seules caractéristiques de l'échantillon, des propriétés plus générales concernant la population.

La différence fondamentale entre la statistique descriptive et la statistique inférentielle est que l'on va supposer que les données sont en fait des réalisations de variables aléatoires (étape de modélisation).

La statistique inférentielle va donc s'appuyer sur la théorie des probabilités.

2. Définitions
2.1. Population et individus

Définissons le vocabulaire utilisé dans ce contexte.

Pour un statisticien,

- La population est l'ensemble quasi exhaustif des individus ayant quelque chose en commun permettant de définir l'appartenance à la population et pour lesquels on étudie un ou plusieurs caractères.
- Un individu est donc un élément de la population. Il possède un ou plusieurs caractères que nous pouvons mesurer.

2.2. Echantillon.

Pour un statisticien,

- L'échantillon est un sous ensemble de la population étudiée, sur lequel on effectue une série de mesures sur le ou les caractères étudiés.
- En général, ces caractères sont appelés variables. Ces variables peuvent être de type quantitatif (taille, poids, salaire, etc ...) et/ou qualitatif (couleur des yeux, sexe, etc ...).

L'opération qui permet de sélectionner de façon organisée les éléments de l'échantillon s'appelle **l'échantillonnage**.

Il existe plusieurs méthodes d'échantillonnage.

3. Construction d'un échantillon.
3.1. Méthode d'échantillonnage

On présente dans cette partie quelques méthodes d'échantillonnage usuelles.

Une méthode d'échantillonnage est une méthode permettant de prélever un échantillon d'individus au sein d'une population, de manière à reproduire une (sous-)population aussi représentative que possible de la population entière.

Cette notion de représentativité est très importante, car un échantillon représente plus ou moins bien la population de référence et donc les conclusions que l'on pourra tirer d'une étude basée sur un échantillon seront... plus ou moins justes !

On notera que le choix de la méthode dépend du problème considéré.

3.2. Echantillonnage aléatoire simple

Un échantillonnage est dit aléatoire si tous les individus de la population ont la même chance de faire partie de l'échantillon; il est dit simple si les prélèvements des individus sont réalisés indépendamment les uns des autres.

- Dans une population infinie, on prélève des individus au hasard : tous les individus ont la même probabilité d'être prélevés, et ils le sont indépendamment les uns des autres.
- Dans une population finie, on effectue un tirage aléatoire sans remise,
  ce qui permet de traiter les populations finies comme des populations infinies.

C'est la méthode d'échantillonnage la plus simple et la plus usuelle.

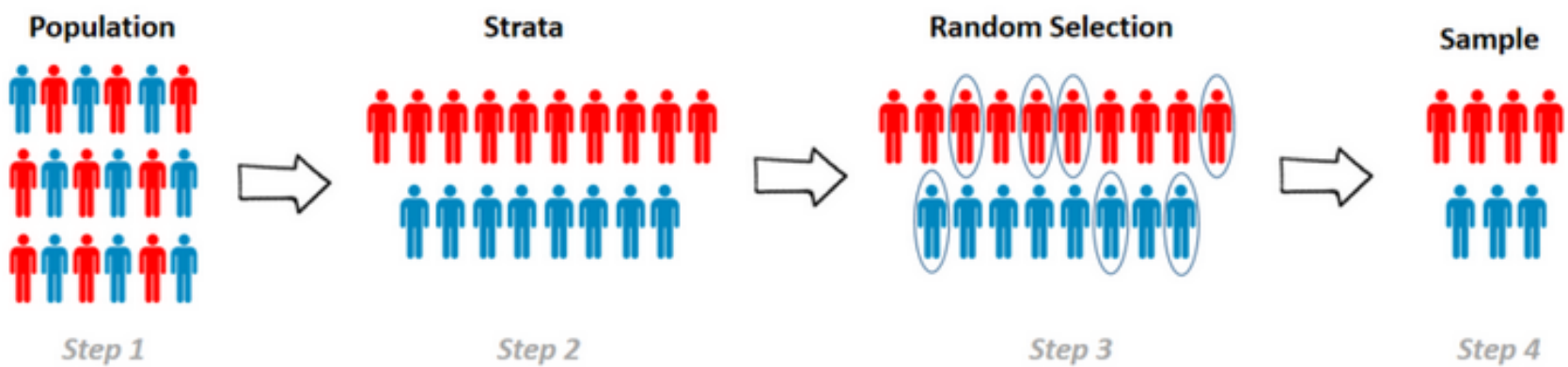
3.2. Echantillonnage aléatoire simple

Avec cette première technique « échantillonnage aléatoire simple »:
Ici laisser faire le hasard consiste à ne rien décider.

3.3. Quelques variantes.

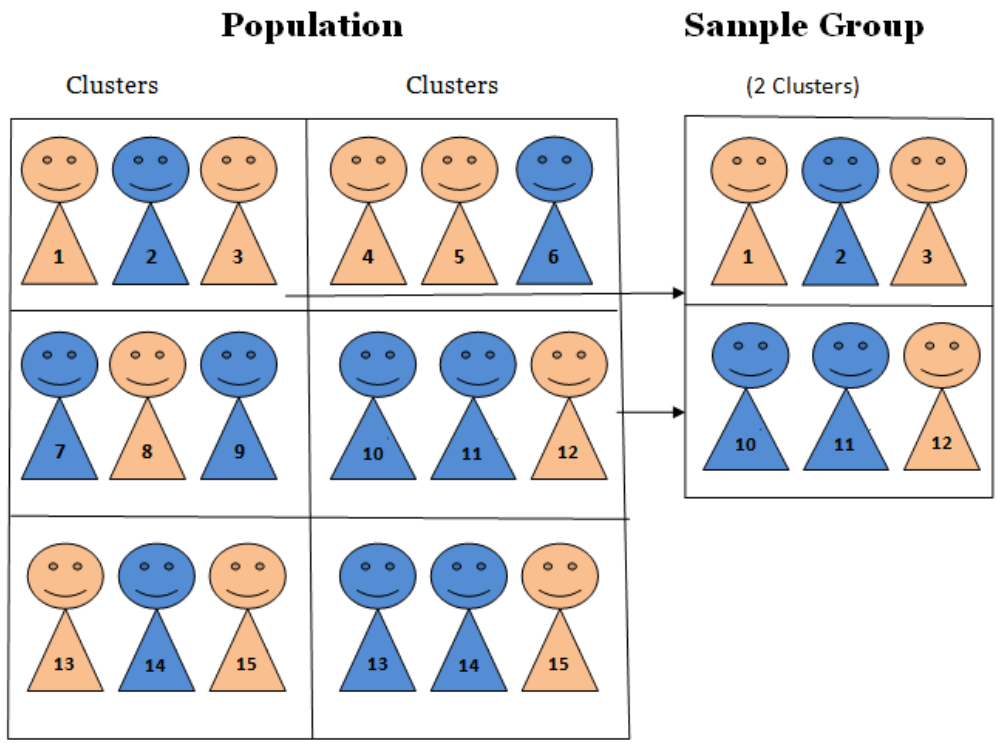
- **Echantillonnage aléatoire stratifié**

Cette méthode présuppose que la population soit stratifiée, i.e. constituée de sous-populations homogènes, les strates (stratification par tranche d'age, sexe, etc ...). Dans chaque strate, on fait un échantillonnage aléatoire simple, de taille proportionnelle à la taille de la strate dans la population (échantillon représentatif). Les individus de la population n'ont pas tous la même probabilité d'être tirés. Nécessite une homogénéité des strates. Augmente la précision des estimations.



- **Echantillonnage par grappes**

  on tire au hasard des grappes ou familles d'individus, et on examine tous les individus de la grappe (on tire au hasard des immeubles dans une ville, puis on interroge tous les habitants de l'immeuble). La méthode est d'autant meilleure que les grappes se ressemblent et que les individus d'une même grappe sont différents, contrairement aux strates.



4. Notions de fluctuations d'échantillonnage.
4.1. Introduction.

"L'esprit statistique naît lorsqu'on prend conscience de la fluctuation d'échantillonnage"

d'après

> Jean-Claude Régnier, "Formation de l'esprit statistique et raisonnement statistique. Que peut-on attendre de la didactique de la statistique ?", Catherine Houdement; Corine Castela (Dir), Séminaire National de Didactique des Mathématiques, Jan 2005, Paris, France. IREM de Paris 7 / Association pour la Recherche en Didactique des Mathématiques, 1, pp.13-38, (2005).

La notion de fluctuation d'échantillonnage est une notion très importante en Statistique.

4.2. Définition.

Lorsque on étudie un caractère sur plusieurs échantillons de même taille d'une même population, on peut observer que les résultats ne sont pas identiques selon les échantillons ; ce phénomène s'appelle fluctuation d'échantillonnage.

$\Longrightarrow$ Par conséquent, comme un échantillon représentera plus ou moins bien la population de référence, les conclusions que l'on pourra tirer d'une étude basée sur un échantillon seront plus ou moins justes, et surtout elles varieront d'un échantillon à l'autre .

4.3. Exemple : Lancer une pièce non truquée.

On lance une pièce bien équilibrée (donc, la probabilité d'obtention des événements **Pile** et **Face** sont égales à $p = 0,5$) 100 fois successivement:

- Pour une $1^{\text{ère}}$ série de 100 lancers, on obtient 54 fois **Pile**, soit une fréquence $f = \dfrac{54}{100} = 0,54$.
- Pour une $2^{\text{ème}}$ série de 100 lancers, on obtient 41 fois **Pile**, soit une fréquence $f = \dfrac{41}{100} = 0,41$.
- Pour une $3^{\text{ème}}$ .....

Dans l'exemple précédent, on sait que même si le nombre de succès varie d'une expérience à l'autre, il sera rare (c'est-à-dire la probabilité sera faible) d'avoir une fréquence de **Pile** très faible ou très grande.

5. Distribution d'échantillonnage.
5.1. Variable d'échantillonnage.

Considérons une population de N individus.

On s'intéresse à un caractère X de ces individus
(par exemple, la taille).

On prélève dans cette population un échantillon de taille n par la méthode d'échantillonnage aléatoire simple.
Pour chaque individu tiré, on mesure son caractère.
On note $x_1, x_2, \ldots, x_n$ les valeurs mesurées.

On s'intéresse à une caractéristique de la population notée S (par exemple, la taille moyenne de la population).

A partir des mesures effectuées sur l'échantillon tiré, on peut construire une approximation, notée s, de S.

Cette approximation est une fonction des valeurs $x_1, x_2, \ldots, x_n$, c'est à dire que l'on peut écrire

$$s = f(x_1, x_2, \ldots, x_n)$$

- *Exp.* : $s = \frac{1}{n} \sum_{i=1}^{n} x_i$ (La moyenne)

On appelle variable d'échantillonnage toute fonction de l'échantillon aléatoire simple qui dépend de $(x_1, x_2, \ldots, x_n)$. Dans note cas, s est une variable d'échantillonnage.

5.2. Distribution d'échantillonnage

Comme il est possible de construire en tout $\mathbf{C}_N^n$ échantillons de taille n dans une population de taille N, on peut construire en tout et pour tout $\mathbf{C}_N^n$ approximations de la caractéristique S.

Il peut-être intéressant d'étudier la répartition de ces approximations en fonction de la taille de n l'échantillon.

On étudie alors ce qu'on appelle la distribution d'échantillonnage. Bien entendu, cette étude ne pourra pas être exhaustive et on se contentera alors de tirer un nombre limité d'échantillons.

5.3. Illustration sur un exemple
5.3.1. Le problème

On considère une population de $N = 1\,000\,000$ individus.
On s'intéresse au poids moyen de la population

On suppose que le poids moyen de la population calculé par recensement est égal à 75 kg.

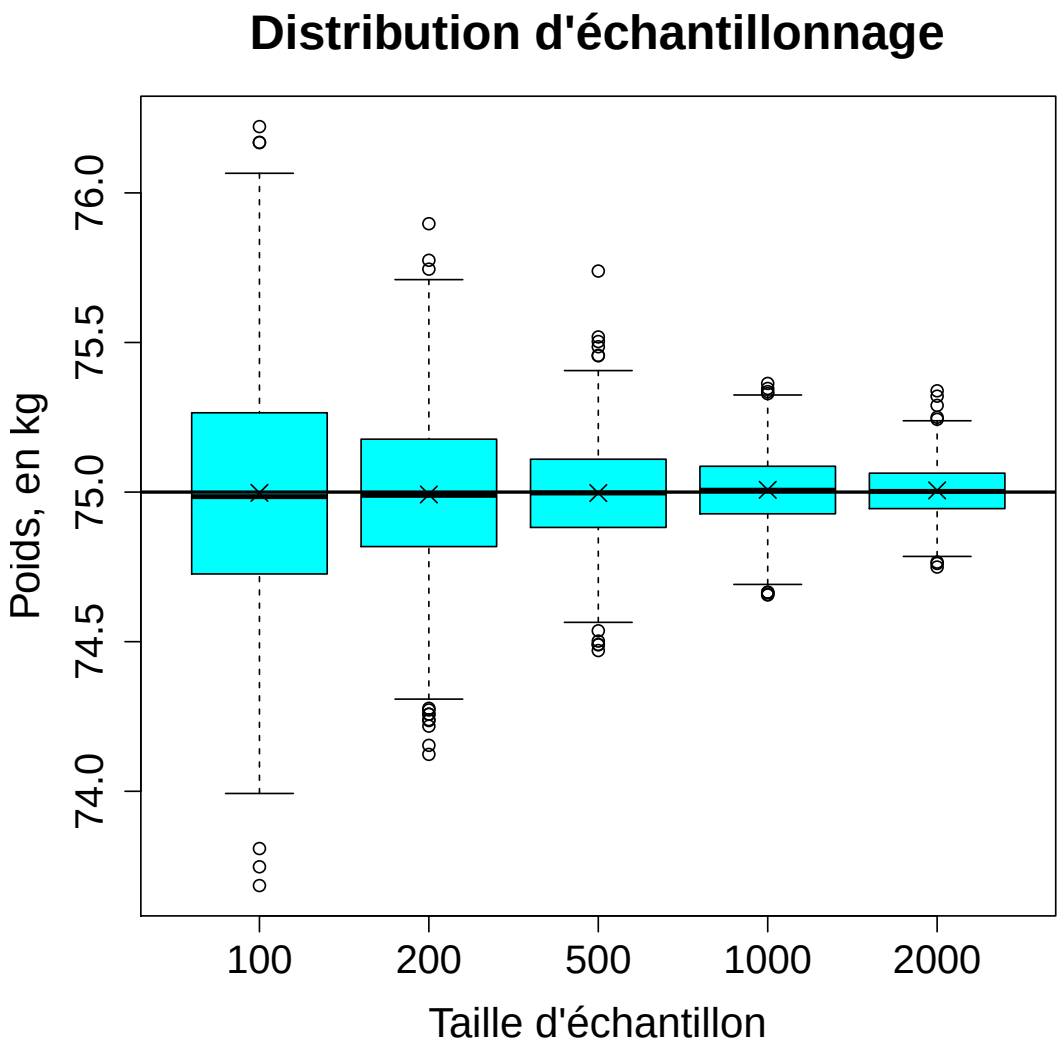
On procéde maintenant par Echantillonnage:

1. On prélève de cette population 1 000 échantillons de taille n, et pour chaque échantillon, on calcule le poids moyen des individus composant l'échantillon (en pesant chacun d'entre eux).
2. On fait varier n pour étudier l'effet de la taille de l'échantillon sur les fluctuations d'échantillonnage.
3. Nous disposons ainsi, pour chaque valeur de n, de 1 000 valeurs du poids moyen censés représenter celui de la population.

Nous allons représenter la distribution d'échantillonnage à l'aide de boîtes à moustaches (se reporter à l'annexe pour la définition), plus adaptées ici que l'histogramme à des fins de comparaison.

5.3.2. Résultat.

Le graphique ci-dessous présente les boites à moustaches des 1 000 poids moyens obtenus, pour différentes tailles d'échantillon n.



5.3.3. Commentaires.

On peut observer, sur le graphique, que :

- L'écart entre les résultats obtenus auprès d'un échantillon et ce que nous apprendrait un recensement de la population varie de moins en moins avec l'augmentation de la taille de l'échantillon (notion de variabilité) ;
- Plus la taille de l'échantillon est grande plus l'erreur d'échantillonnage diminue.

5.4. Intervalle de fluctuations.

Lorsque l'on connait la distribution de la variable d'échantillonnage, il est possible de trouver la taille d'échantillon n nécessaire pour approcher avec une précision donnée la caractéristique S de la population.

Lorsque la caractéristique est une proportion p, on peut démontrer que la variable d'échantillonnage associée appartient à l'intervalle

$$\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$$

avec un niveau de confiance d'au moins 95%.

Il est alors possible de définir la taille de l'échantillon nécessaire pour une précision voulue.

Par exemple, pour approcher la proportion p, avec une marge d'erreur inférieure à 0.1, il faut prendre un échantillon d'au moins 100 individus.

6. Conclusion.

Dès qu'on travaille sur un jeu de données réel, on ne dispose en général que d'un seul échantillon.

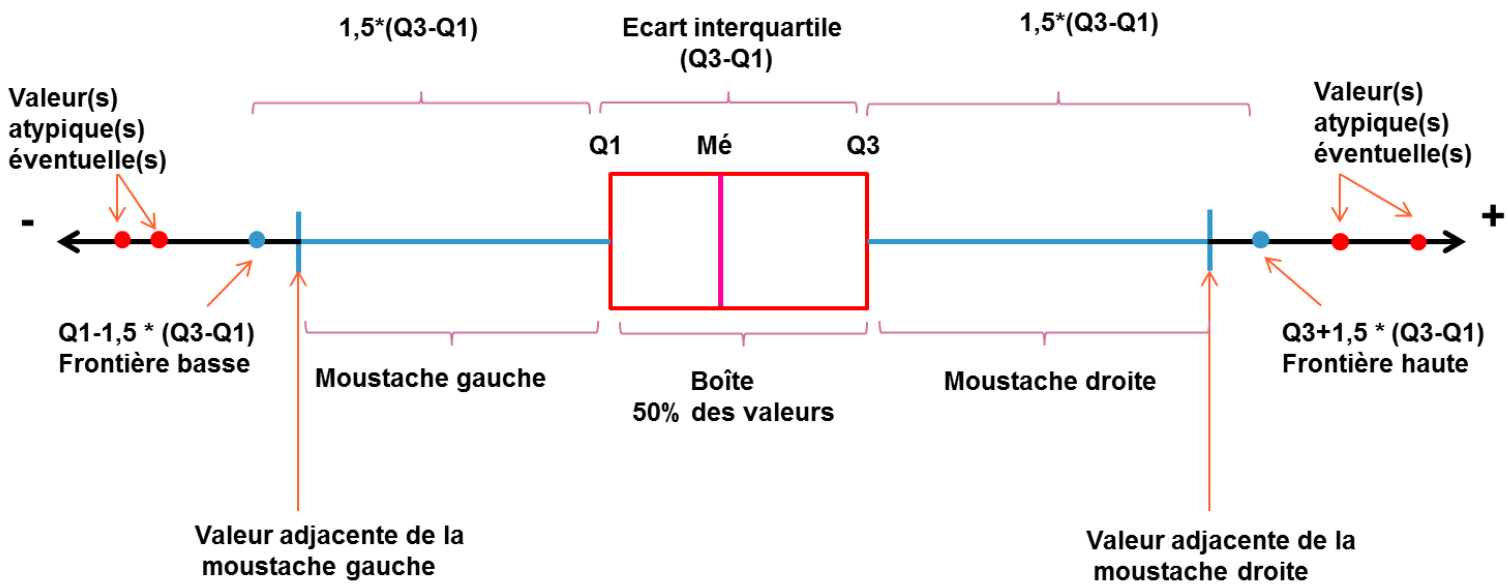
Il est possible de décrire ces données grâce aux outils de statistique descriptive.

La statistique inférentielle va permettre, en considérant que les données sont en fait des réalisations de variables aléatoires, grâce à la théorie des probabilités, d'inférer (induire, extrapoler) à la population tout entière, les résultats obtenus sur l'échantillon, avec des niveaux de confiance, ou des risques de se tromper.

Il faut être conscient que dans certains cas, on sera amené à fournir une information fausse, mais celle-ci sera accompagné d'un niveau de confiance ou fournie avec un risque de se tromper, fixé à l'avance.

7. Annexe : boite à moustaches.

On peut résumer la distribution d'une variable quantitative à l'aide de ce qu'on appelle une boîte à moustaches.
La construction de cette boîte est basée sur les indicateurs de position de la série : min, Q_1, médiane, Q_3 et max.
Remarque : On peut y rajouter avec profit la moyenne.



La définition des moustaches varie d'un logiciel à l'autre.