Analyses statistisques de données

Gilles Le Chenadec

ENSTA Bretagne

21 août 2018

Objectifs du cours

J'ai des données ...qu'est-ce que je peux faire ? :-(

- Passer des données à un modèle mathématique qu'on exploitera pour faire parler les données !
- Exemple de la pièce biaisée ou pas (6/10 par rapport à 4/10 ou 60/100 par rapport à 4/10)
- Passer de l'intuition (impression subjective) à des données objectives
- A partir de ces données, établir une stratégie pour votre étude, votre entreprise (aide à la décision)

Plan du cours

Séquencement du cours

1	lun. 27 août 18	C3 (10:20-11:15)	CM	Amphi/TD	Statistiques descriptives univariées
2	lun. 27 août 18	C4 (11:20-12:15)	TD	Amphi/TD	Statistiques descriptives univariées
3	lun. 3 septembre 18	C3 (10:20-11:15)	TP	Info	Statistiques descriptives univariées
4	lun. 3 septembre 18	C4 (11:20-12:15)	TP	Info	Statistiques descriptives univariées
5	jeu. 6 septembre 18	C3 (10:20-11:15)	CM	Amphi/TD	Statistiques descriptives bivariées
6	jeu. 6 septembre 18	C4 (11:20-12:15)	TD	Amphi/TD	Statistiques descriptives bivariées
7	lun. 10 septembre 18	C3 (10:20-11:15)	TP	Info	Statistiques descriptives bivariées
8	lun. 10 septembre 18	C4 (11:20-12:15)	TP	Info	Statistiques descriptives bivariées
9	jeu. 13 septembre 18	C3 (10:20-11:15)	CM	Amphi/TD	Estimation paramétrique
10	jeu. 13 septembre 18	C4 (11:20-12:15)	TD	Amphi/TD	Estimation paramétrique
11	lun. 17 septembre 18	C3 (10:20-11:15)	TP	Info	Estimation paramétrique
12	lun. 17 septembre 18	C4 (11:20-12:15)	TP	Info	Estimation paramétrique
13	jeu. 20 septembre 18	C3 (10:20-11:15)	CM	Amphi/TD	Tests paramétriques
14	jeu. 20 septembre 18	C4 (11:20-12:15)	TD	Amphi/TD	Tests paramétriques
15	lun. 24 septembre 18	C3 (10:20-11:15)	TP	Info	Tests paramétriques
16	lun. 24 septembre 18	C4 (11:20-12:15)	TP	Info	Tests paramétriques
17	jeu. 27 septembre 18	C3 (10:20-11:15)	CM	Amphi/TD	Tests non-paramétriques
18	jeu. 27 septembre 18	C4 (11:20-12:15)	TD	Amphi/TD	Tests non-paramétriques
19	lun. 1 octobre 18	C3 (10:20-11:15)	TP	Info	Tests non-paramétriques
20	lun. 1 octobre 18	C4 (11:20-12:15)	TP	Info	Tests non-paramétriques
21	jeu. 4 octobre 18	C3 (10:20-11:15)	CM	Amphi/TD	Arbre de décisions
22	jeu. 4 octobre 18	C4 (11:20-12:15)	TD	Amphi/TD	Arbre de décisions
23	lun. 8 octobre 18	C3 (10:20-11:15)	TP	Info	Arbre de décisions
24	lun. 8 octobre 18	C4 (11:20-12:15)	TP	Info	Arbre de décisions
25	jeu. 18 octobre 18	C3 (10:20-11:15)	CE	Examen	Contrôle
26	jeu. 18 octobre 18	C4 (11:20-12:15)	CE	Examen	Contrôle

Etudions l'hélicoptère en papier

Introduction

- Mettre au point un hélicoptère en papier qui, lancé d'une certaine hauteur, atteint le sol en un temps de vol maximal.
- Cet hélicoptère est
  - construit à partir d'une demi-feuille A4, découpée et pliée selon le schéma ci-dessous
  - Cet hélicoptère est caractérisé par ses dimensions Ha, Hc, Hp, La et Lp
  - ainsi que par d'autres paramètres possibles : L'épaisseur et le grainage du papier, La présence d'un stabilisateur en bas du pied (un trombone), Le sens de la pliure des ailes, La coupe éventuelle des coins du corps pour augmenter l'aérodynamisme.

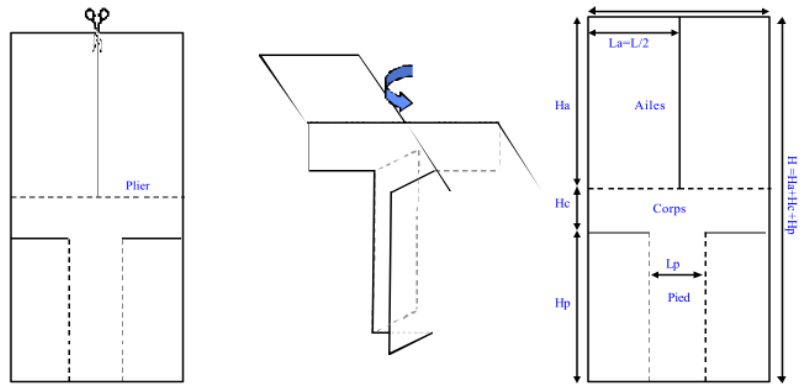


Figure – (Université Catholique de Louvain, 2010)

C'est quoi l'analyse statistique ?

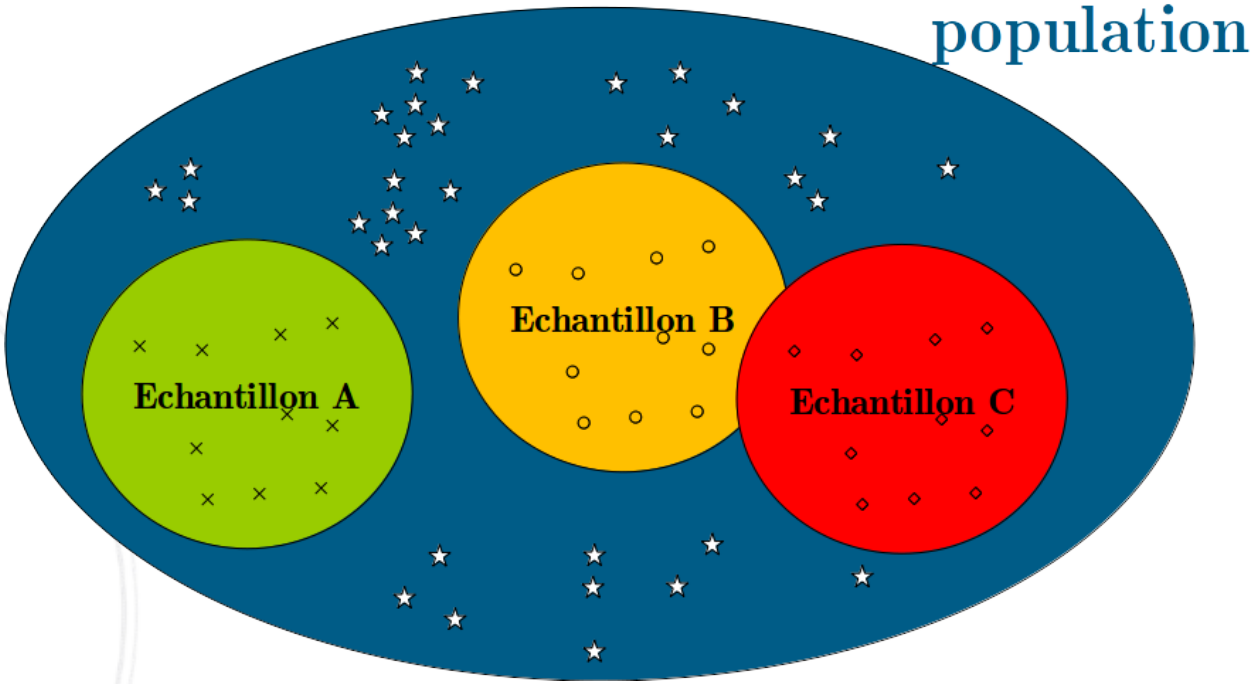
Introduction

- **Je construits**
  - 10 hélicoptères en papier, je mesure un temps de vol moyen de 8.3 s.
  - 100 hélicoptères en papier, je mesure un temps de vol moyen de 8.7 s.
  - 10 hélicoptères en papier, avec un trombone accroché au pied, je mesure un temps de vol moyen de 8.1 s.
  - 10 autres hélicoptères en papier, avec un corps de 4 cm, je mesure un temps de vol moyen de 9 s.
- **Questions posées par le statisticien**
  - Comment a été mesuré le temps de vol ? Y-a-t-il un biais ?
  - Sont ils construits de la même manière, ...
  - Peut-on comparer le temps de deux expériences ?
  - Est-ce que je peux dire que les temps de vol sont identiques ?
  - Quand la différence de temps de vol devient significative ?
  - A partir de combien d'essai ?
  - Quand est-ce que je peux généraliser à toute la population des hélicos en papier ?
  - Quel est le risque que je prends de me tromper ?

La population et l'échantillon

Introduction

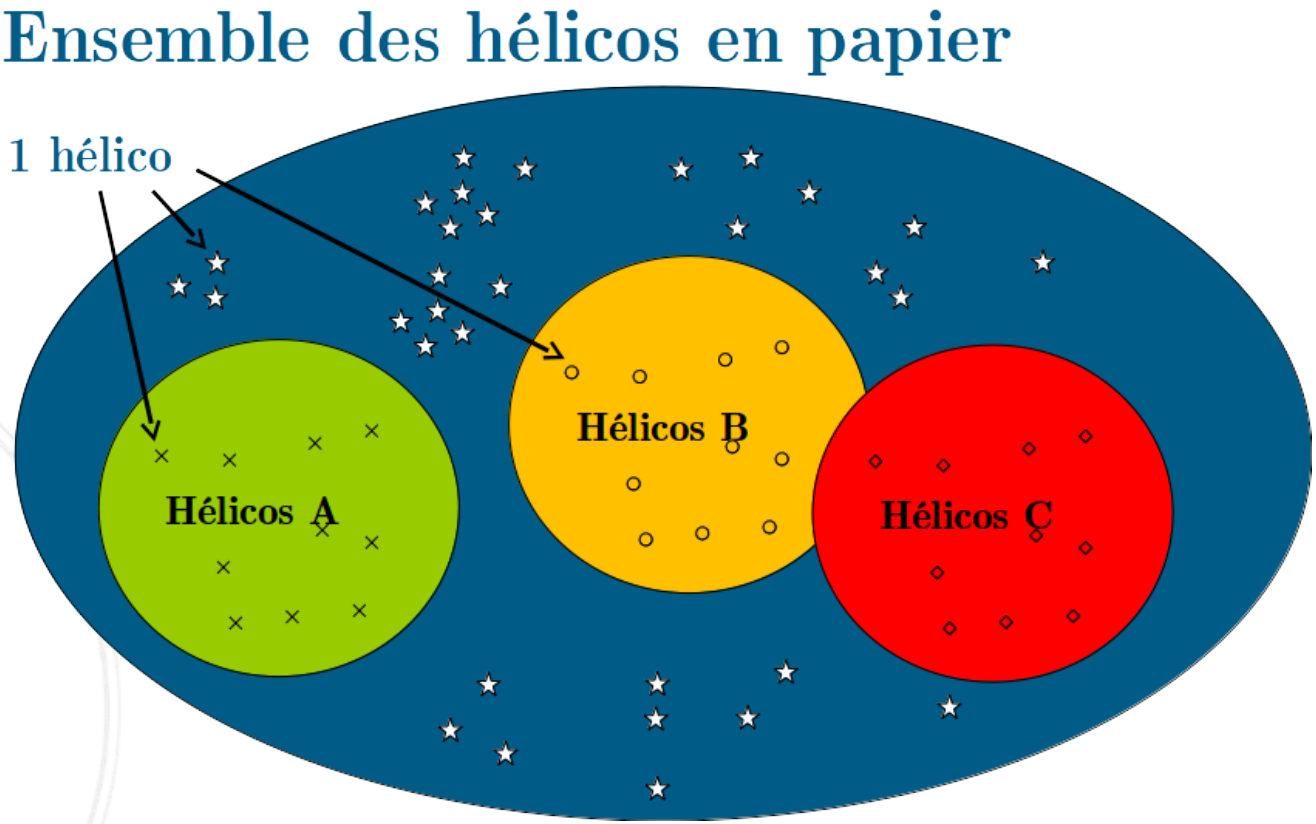
- Un des concepts de base en statistiques est l'échantillonnage (sampling)
- Dans la plupart des problèmes statistiques, un nombre de mesures ou données appelé l'**échantillon** est tiré d'un beaucoup plus grand (voire infini) ensemble de mesures ou données appelé la **population**.



La population et l'échantillon

Introduction

- Pour le problème de l'hélico en papier, nous espérons que les mesures prises correspondent à tous les hélico du monde !
- Parmi la population et l'échantillon, qu'est-ce qui est le plus intéressant ?



Parmi la population et l'échantillon, qu'est-ce qui est le plus intéressant ?

- Dans la plupart des cas, nous nous intéresserons principalement à la population.
- mais comme il est difficile d'enregistrer le temps de vol de tous les hélicoptères possibles,
- nous tenterons de **décrire ou prédire** le comportement de la population sur la base des informations obtenues à partir de cet échantillon (représentatif de la population !)

Quand on est face à un ensemble de mesures (échantillon ou population), le statisticien procéde en le décrivant, en représentant et en le résumant. La branche des statistiques qui le permet est appelé statistiques descriptives. Il peut être trop cher, trop "time-consuming" ou impossible d'avoir les mesures correspondant à la population complète. Il s'agit alors d'exploiter un échantillon de cette population et de tenter de répondre aux questions que vous vous posez sur la population.

Definitions

- Les **statistiques descriptives** représentent un ensemble de techniques utilisées pour résumer et décrire les importantes caractéristiques d'un ensemble de mesures.
- Les **statistiques inferentielles** représentent un ensemble de techniques utilisées pour réaliser des inferences sur les caracteristiques d'une population à partir des informations contenues dans un échantillon tiré de cette population.

Première partie I

Eléments d'analyse statistique descriptive

Une fois que vous avez des données, comment peut-on les afficher, les mettre en valeur dans une forme claire et compréhensive ? On doit d'abord définir les différents types de données ou variables qu'on rencontre dans la vie courante.

Definition

L'**individu ou unité statistique** est un élément de l'échantillon ou de la population.

Definition

On appelle **variable** toute valeur attachée aux individus de la population ou d'un échantillon.

Variables et données : exemples

Eléments d'analyse statistique descriptive

Echantillon de 32 mesures du temps de vol d'hélicoptères

Numéro du lancer	Helico	H	L	H_a	H_c	H_p	L_a	L_p	Coupé	Temps
1	A	29,7	10,5	7	2,7	12	5,25	3,5	NON	1,97
2	A	29,7	10,5	7	2,7	12	5,25	3,5	NON	2,03
3	A	29,7	10,5	7	2,7	12	5,25	3,5	NON	2,1
4	A	29,7	10,5	7	2,7	12	5,25	3,5	NON	1,86
5	A	29,7	10,5	7	2,7	12	5,25	3,5	NON	2
6	A	29,7	10,5	7	2,7	12	5,25	3,5	NON	2,14
7	A	29,7	10,5	7	2,7	12	5,25	3,5	NON	2,07
8	A	29,7	10,5	7	2,7	12	5,25	3,5	NON	2,19
9	B	29,7	10,5	15	2,7	12	5,25	3,5	NON	2,52
10	B	29,7	10,5	15	2,7	12	5,25	3,5	NON	3,01
11	B	29,7	10,5	15	2,7	12	5,25	3,5	NON	2,69
12	B	29,7	10,5	15	2,7	12	5,25	3,5	NON	2,71
13	B	29,7	10,5	15	2,7	12	5,25	3,5	NON	2,93
14	B	29,7	10,5	15	2,7	12	5,25	3,5	NON	2,83
15	B	29,7	10,5	15	2,7	12	5,25	3,5	NON	2,7
16	B	29,7	10,5	15	2,7	12	5,25	3,5	NON	3,11
17	C	29,7	10,5	7	2,7	2	5,25	3,5	NON	2,52
18	C	29,7	10,5	7	2,7	2	5,25	3,5	NON	2,25
19	C	29,7	10,5	7	2,7	2	5,25	3,5	NON	2,47
20	C	29,7	10,5	7	2,7	2	5,25	3,5	NON	2,44
21	C	29,7	10,5	7	2,7	2	5,25	3,5	NON	2,44
22	C	29,7	10,5	7	2,7	2	5,25	3,5	NON	2,45
23	C	29,7	10,5	7	2,7	2	5,25	3,5	NON	2,36
24	C	29,7	10,5	7	2,7	2	5,25	3,5	NON	2,27
25	D	29,7	10,5	7	2,7	12	5,25	3,5	OUI	1,87
26	D	29,7	10,5	7	2,7	12	5,25	3,5	OUI	2,12
27	D	29,7	10,5	7	2,7	12	5,25	3,5	OUI	1,84
28	D	29,7	10,5	7	2,7	12	5,25	3,5	OUI	2,08
29	D	29,7	10,5	7	2,7	12	5,25	3,5	OUI	1,98
30	D	29,7	10,5	7	2,7	12	5,25	3,5	OUI	1,98
31	D	29,7	10,5	7	2,7	12	5,25	3,5	OUI	1,78
32	D	29,7	10,5	7	2,7	12	5,25	3,5	OUI	1,88

Figure – Mesures du temps de vol d'un hélico

Definition

Les **données univariées** sont composées d'**une seule variable** pour chaque individu statistique de l'échantillon ou de la population.

Definition

Les **données bivariées** sont composées de **deux variables** pour chaque individu statistique de l'échantillon ou de la population.
Les **données multivariées** sont composées de **plus de deux variables** pour chaque individu statistique de l'échantillon ou de la population.

Types de variables : qualitatives vs quantitatives

Les variables peuvent être classées en deux types : qualitatives ou quantitatives.

Definition

Les **variables qualitatives ou catégorielles** mesurent l'appartenance à une catégorie donnée. Elles peuvent être **nominales** ou **ordinales**

Exemples : sexe, catégories socio-profesonnelles, age = moins de 20 ans ; de 20 ans à 39 ans ; de 40 ans à 59 ans, 60 ans et plus

Definition

Les **variables quantitatives** mesurent une quantité numérique. Elles peuvent être **discrètes** ou **continues**.

Exemples : prix, taille, etc.

Après la collecte de données, elles peuvent être exploitées et résumées pour révéler les informations suivantes :

- Quelles valeurs des variables ont été mesurées ?
- Combien de fois chaque valeur a été observée ?

Dans cette optique, une table statistique peut être construite afin de représenter graphiquement les données. On parle de **distribution des données** (sous entendu en fonction des valeurs qu'elles prennent !).
Le type de graphique utilisé dépend du type de la variable mesurée....

Representations d'une variable catégorielle

Eléments d'analyse statistique descriptive

Considérons la variable Hélico :

- Quel est son type ?
- Qu'est-ce qu'on peut dire ?
- Comment la représenter ?

Lancer	Hélico
1	A
2	A
3	A
4	A
5	A
6	A
7	A
8	A
9	B
10	B
11	B
12	B
13	B
14	B
15	B
16	B
17	C
18	C
19	C
20	C
21	C
22	C
23	C
24	C
25	D
26	D
27	D
28	D
29	D
30	D
31	D
32	D

Quand la variable est qualitative ou catégorielle, la **table statistique ou table des effectifs** est une liste de catégories à laquelle on associe une mesure de "combien de fois la catégorie" a été rencontrée. Trois mesures différentes existent :

Definition

- la **fréquence ou effectif** : le nombre de mesures de chaque catégorie
- la **fréquence relative** : la proportion de mesures de chaque catégorie
- le **pourcentage** de mesures de chaque catégorie

Soit n le nombre total de mesures dans un ensemble,

$$\text{Relative frequency} = \frac{\text{Frequency}}{n} \qquad \text{Percent} = 100 \times \text{Relative frequency}$$

Représentations d'une variable catégorielle : remarques

- La somme des :
  - fréquences est n
  - fréquences relatives est 1
  - pourcentages est 100 %
- Les catégories doivent être choisies de telle sorte que
  - une mesure appartiendra à une et une seule catégorie
  - chaque mesure peut être assigné à une catégorie

Construction de la table statistique

Representations d'une variable catégorielle

Données brutes

Lancer	Hélico
1	A
2	A
3	A
4	A
5	A
6	A
7	A
8	A
9	B
10	B
11	B
12	B
13	B
14	B
15	B
16	B
17	C
18	C
19	C
20	C
21	C
22	C
23	C
24	C
25	D
26	D
27	D
28	D
29	D
30	D
31	D
32	D

Table statistique

Modalité du type d'hélico x_j	A	B	C	D
Effectifs n_j	8	8	8	8

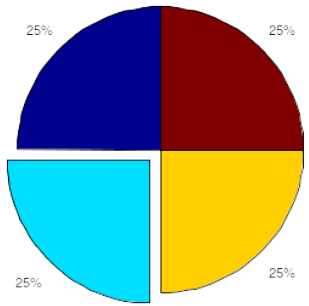
Representations d'une variable catégorielle

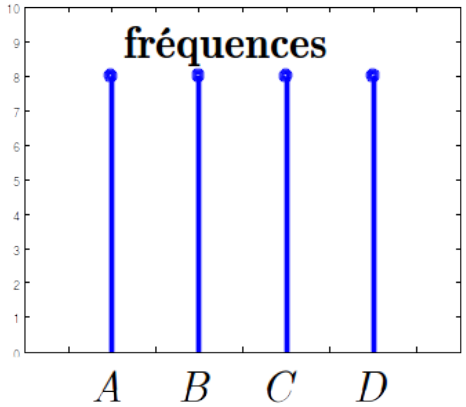
Une fois que les données ont été catégorisées et résumées dans une table statistique, on peut utiliser soit la représentation en bâtons ou en secteurs.

Definition

- La **représentation en secteurs** est le graphique circulaire ou camembert des pourcentages.
- La **représentation en bâton** est la représentation ou l'axe des abscisses représente les catégories et l'axe des ordonnées la mesure de combien de fois la catégorie est mesurée.

Catégorie	A	B	C	D
Fréquence	8	8	8	8





Representations d'une variable catégorielle : remarque

L'impact visuel des deux représentations n'est pas le même

- La **représentation en secteurs** est utilisée pour représenter les relations des différentes catégories
- La **représentation en bâton** est utilisée pour mettre en avant la valeur de chaque catégorie.

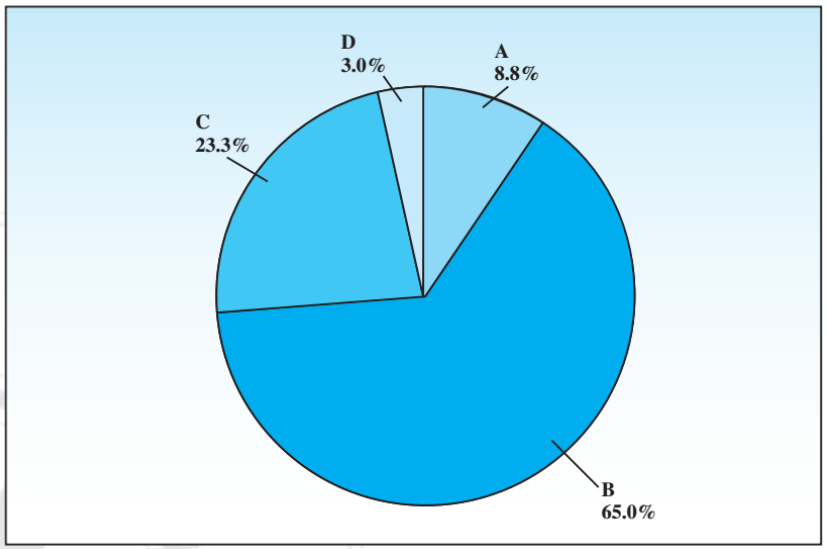


Figure – extrait de Mendenhall et al, 2010

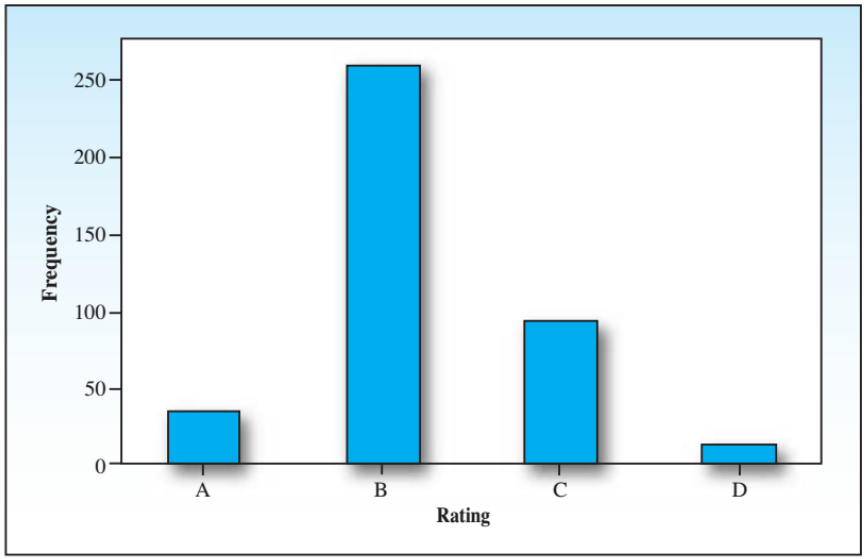
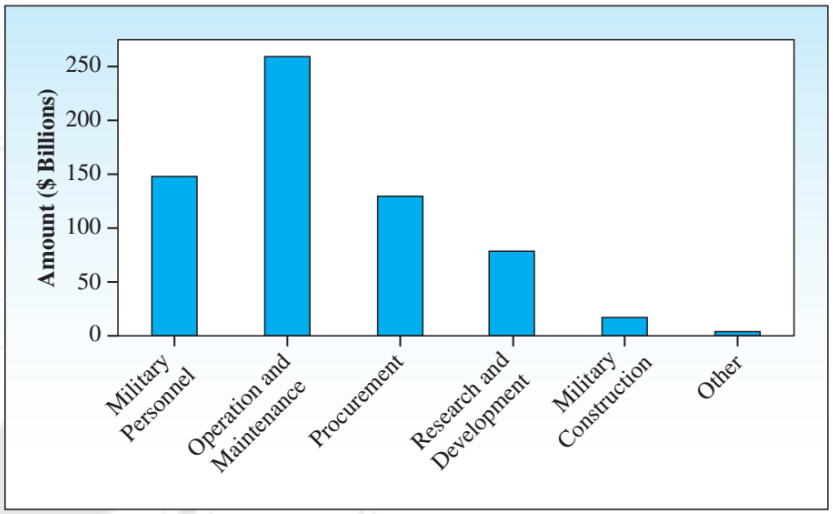


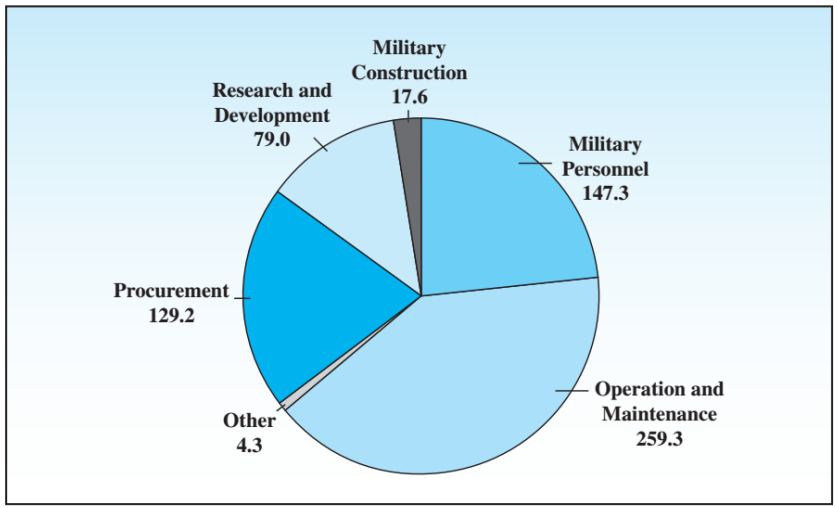
Figure – extrait de Mendenhall et al, 2010

Representations d'une variable quantitative

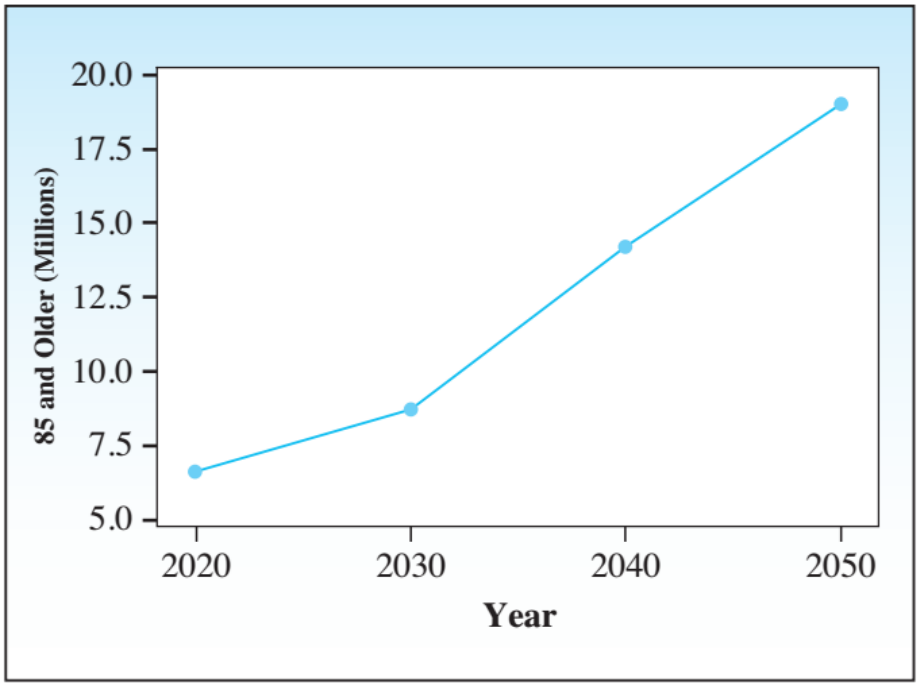
Diagramme en bâtons ou en secteurs

Lorsque les informations sont collectées pour une variable quantitative dans différents segments de la population ou pour différentes catégories, on peut alors utiliser le **diagramme en bâtons ou en secteurs**.
Exemple des milliards de dollars dépensés en 2009 par le département de la défense américain dans diverses catégories.





Quand la variable quantitative a été collectée en fonction du temps, les données constituent **une série temporelle i.e un signal**. On peut alors représenter cette série en fonction du temps pour discerner un **motif ou une tendance particulière.**



Representations d'une variable quantitative

Considérons la variable Temps de vol :

- Quel est son type ?
- Qu'est-ce qu'on peut dire ?
- Comment la représenter ?

N	temps
1	1,97
2	2,25
3	2,09
4	1,89
5	1,92
6	2,04
7	1,96
8	1,83
9	1,8
10	2,16
11	2,07
12	1,87
13	1,96
14	2,23
15	2,14
16	1,92
17	1,89
18	2,03
19	2,09
20	1,66
21	2,1
22	1,87
23	1,94
24	2,04
25	1,81
26	2,09
27	2
28	2,28
29	1,94
30	2,05
31	2,16
32	2,03
33	2,32
34	2,29
35	1,79
36	1,9
37	**2,01**

Representations d'une variable quantitative

Histogramme des fréquences relatives

Definition

L'**histogramme des fréquences relatives** pour une variable quantitative est une représentation en barres pour laquelle la hauteur représente "combien de fois" les mesures tombent dans un intervalle de données. Les intervalles sont tracés sur l'axe horizontal. Ce processus aboutit à la création d'une **table statistique** (comme pour une variable qualitative !).

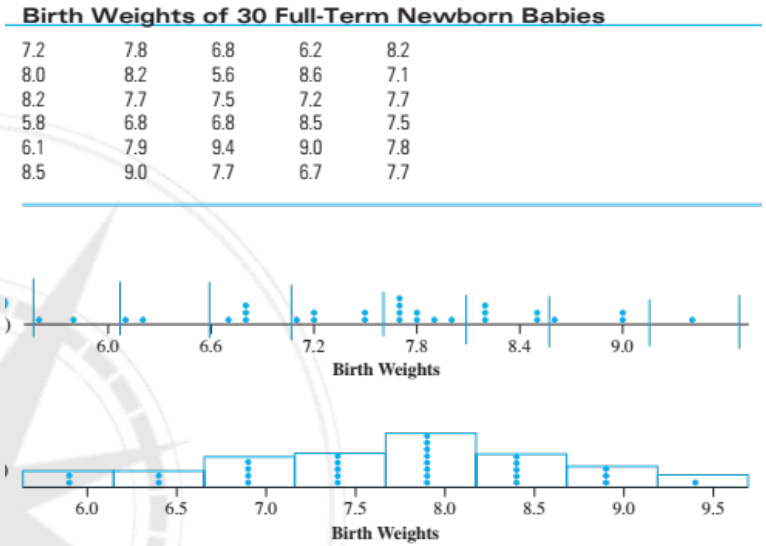


Figure – Construction de l'histogramme

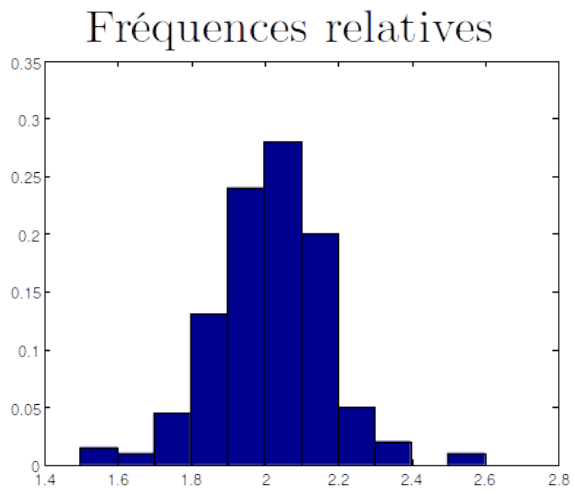


Figure – Temps de vol des hélicos

Procédure :

- Fixer le nombre d'intervalles
- Déterminer l'amplitude des intervalles ((max - min)/nb d'intervalle)
- Affecter les valeurs à chaque intervalle en utilisant la règle de l'inclusions par la gauche i.e [x0,x1)

Remarque sur le nombre d'intervalles

- certains auteurs préconisent entre 5 et 12
- plus on a de données, plus le nombre d'intervalles doit être important
- compromis entre le nombre de données et la qualité de la forme de l'histogramme
- des règles empiriques existent :
  - Sturge : $1 + 3.3 \log n$
  - Yule : $2.5 \sqrt[4]{n}$

Interpréter une distribution avec un oeil critique

Vous avez produit un graphique, qu'est ce que vous devez faire pour essayer de décrire les données ?

Point méthodo

- Vérifiez l'axe horizontal et l'axe vertical pour être certain de ce qui a été mesuré et affiché.
- Examinez la localisation de la distribution des données. Où sur l'axe horizontal est le centre de la distribution ? Si vous comparez deux distributions, ont elles le même centre ?
- Examinez la forme de la distribution. Est-ce que la distribution a un pic ? Si oui, il s'agit de la catégorie/intervalle le/la plus fréquente. Y-a-t-il plus d'un pic ? Y-a-t-il approximativement un nombre égal de mesures du côté gauche ou droit d'un pic ?
- Recherchez des mesures extrèmes (outliers). C'est à dire y-a-t-il des mesures beaucoup plus grandes ou petites que toutes les autres ? De quoi ces valeurs sont elles représentatives ?

Interpréter une distribution avec un oeil critique

Les distributions sont très souvent décrites par leur forme.

Definition

Une distribution est :

- **symétrique** si les cotés gauche et droit de la distribution forment deux images mirroirs.
- **asymétrique à gauche (à droite)** si une plus grande proportion des données se situe à gauche (à droite).
- **unimodale** si elle n'a qu'un pic ;
- **bimodale** si elle en a deux ; révélant le mélange de deux populations dans les données (une autre variable explique ces deux pics, laquelle ?)

Décrire des données avec des mesures numériques

Dans cette partie, les variables sont supposées être quantitatives !

Les distributions sont très souvent décrites par leur forme mais les définitions vues précédemment sont subjectives (i.e. cette distribution est plus asymétrique que l'autre...) Pour résoudre ce problème, on calcule des valeurs numériques caractérisant la distribution : son centre (sa position), sa variabilité autour de son centre (sa dispersion) et sa forme.

Mesures de la position : moyenne

Décrire des données avec des mesures numériques

Il s'agit de la mesure $\bar{x}$ qui positionne le centre de la distribution sur l'axe horizontal.

Definition

Soit n est le nombre de données dans l'échantillon, la moyenne (arithmétique) est défini par :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Remarques

- la moyenne de la population est générale notée μ
- dans le cas d'une variable quantitative dans une table statistique (x_i valeur discrète/centre d'intervalle et n_i effectif/fréquence) la moyenne devient :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} n_i x_i$$

Definition

La médiane m est un nombre tel qu'au moins la moitié de l'effectif de la série X soit inférieur ou égal à m et au moins la moitié de l'effectif de la série soit supérieur ou égal à m.

- Trier les n données X par ordre croissant
- Déterminer m en fonction de la parité de n :
  - Si n impair, $n = 2p + 1$, alors $m = X(p + 1)$
  - Si n pair, $n = 2p$, toute valeur de l'intervalle $[X(p), X(p + 1)]$ est une médiane

Example

- 2 5 6 9 11
- 2 5 6 9 11 27

Mesures de la position : médiane pour une variable quantitative continue groupée par intervalles

Soit X continue regroupée en classes selon une subdivision $(x_0, ..., x_p)$ de $[\min(X), \max(X)[$, $I_i = [x_{i-1}, x_i[$ le i-ième intervalle de la subdivision (ou i-ème classe) .

Procédure

- Construire le tableau des effectifs cumulés $N_i = \sum_{k=1}^{i} n_k$
- Déterminer l'intervalle $I_m = [x_{m-1}, x_m]$ tel que l'effectif cumulé N_i contienne l'individu médian $N/2$
- n_m, a_m : effectif et amplitude de I_m
- N_{m-1} : effectif cumulé de I_{m-1}

La médiane m d'une variable continue X satisfait à l'égalité suivante :

$$m = x_{m-1} + \frac{\frac{N}{2} - N_{m-1}}{n_m} a_m$$

Mesures de la position : médiane pour une variable quantitative continue groupée par intervalles

Décrire des données avec des mesures numériques

Example (Médiane des $n = 200$ temps de vol)

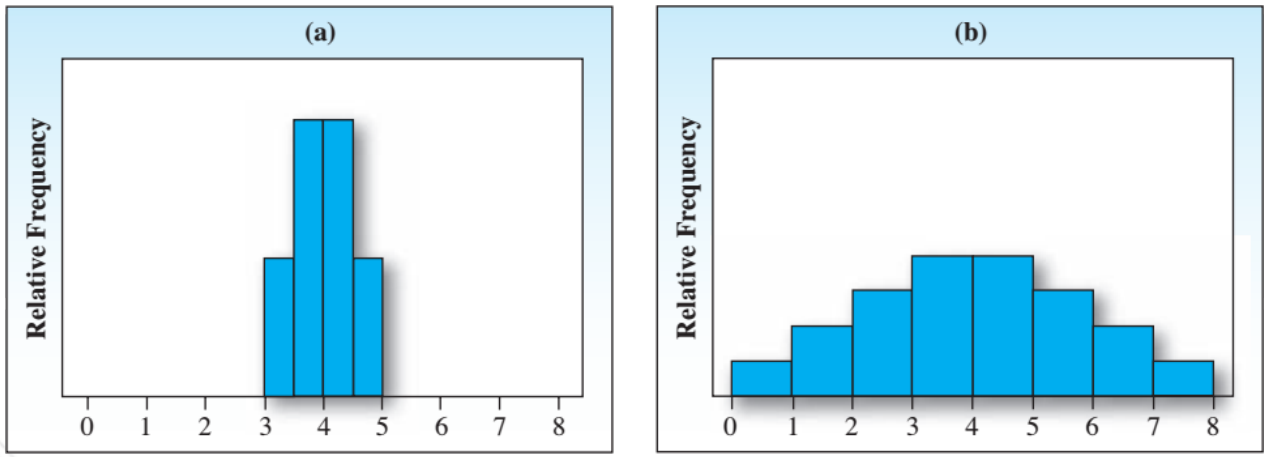
Intervalle I_j	[1.5,1.6[	[1.6,1.7[	[1.7,1.8[	[1.8,1.9[	[1.9,2.0[	[2.0,2.1[	[2.1,2.2[	[2.2,2.3[	[2.3,2.4[	[2.4,2.5[	[2.5,2.6[
Centre c_j	1.55	1.65	1.75	1.85	1.95	2.05	2.15	2.25	2.35	2.45	2.55
Effectifs n_j	3	2	9	26	48	56	40	10	4	1	1
Effectifs cumulés N_j	3	5	14	40	88	144	184	194	198	199	200
Fréquences f_j	0.015	0.010	0.045	0.130	0.240	0.280	0.200	0.050	0.020	0.0050	0.0050
Fréquences cumulées	0.015	0.025	0.070	0.200	0.440	0.720	0.920	0.970	0.990	0.995	1.000

- individu médian $n/2 = 100$
- Intervalle $I_m = [2.0, 2.1[$
- $n_m = 56$; $a_m = 0.1$
- $N_{m-1} = 88$
- $m = 2.0 + 0.1\frac{100-88}{56} = 2.02$

Mesures de variabilité/dispersion : intro

Décrire des données avec des mesures numériques

Deux ensembles de données peuvent avoir le même centre mais avoir des représentations différentes car elles se dispersent différement par rapport à leur centre.



- Même centre
- (gauche) mesure de 3 à 5 (droite) de 0 à 8

La variabilité ou la dispersion est une caractéristique importante des données.

Definition

La **variance empirique** d'un échantillon de n mesures est donnée par :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2$$

Remarques

- la variance théorique d'une population sera notée σ
- Dans le cas d'un calcul à partir d'une table statistique

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}n_i\,(x_i - \bar{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n}n_i x_i^2\right) - \bar{x}^2$$

Definition

L'**écart-type** est la racine carrée de la variance :

$$s = \sqrt{s^2}$$

Mesures de la dispersion : quantile pour une variable quantitative continue groupée par intervalles

Soit X continue regroupée en classes selon une subdivision $(x_0, ..., x_p)$ de $[\min(X), \max(X)[$, $I_i = [x_{i-1}, x_i[$ le i-ième intervalle de la subdivision (ou i-ème classe) .

Procédure d'interpolation pour le calcul du quantile j/q

- Construire le tableau des effectifs cumulés $N_i = \sum_{k=1}^{i} n_k$
- Déterminer l'intervalle $I_q = [x_{q-1}, x_q]$ tel que l'effectif cumulé N_i contienne l'individu jN/q
- n_q, a_q : effectif et amplitude de I_q
- N_{q-1} : effectif cumulé de I_{q-1}

La quantile j/q d'une variable continue X satisfait à l'égalité suivante :

$$m = x_{q-1} + \frac{\frac{jN}{q} - N_{q-1}}{n_q} a_q$$

Mesures de la position : quantile pour une variable quantitative continue groupée par intervalles

Décrire des données avec des mesures numériques

Example (Médiane des $n = 200$ temps de vol, Q1 ? Q3 ?)

Intervalle I_j	[1.5,1.6[	[1.6,1.7[	[1.7,1.8[	[1.8,1.9[	[1.9,2.0[	[2.0,2.1[	[2.1,2.2[	[2.2,2.3[	[2.3,2.4[	[2.4,2.5[	[2.5,2.6[
Centre c_j	1.55	1.65	1.75	1.85	1.95	2.05	2.15	2.25	2.35	2.45	2.55
Effectifs n_j	3	2	9	26	48	56	40	10	4	1	1
Effectifs cumulés N_j	3	5	14	40	88	144	184	194	198	199	200
Fréquences f_j	0.015	0.010	0.045	0.130	0.240	0.280	0.200	0.050	0.020	0.0050	0.0050
Fréquences cumulées	0.015	0.025	0.070	0.200	0.440	0.720	0.920	0.970	0.990	0.995	1.000

- $j = 1$; $q = 4$
- individu 50 ;
- Intervalle $I_q = [1.9, 2.0[$;
- $n_q = 48$; $a_q = 0.1$;
- $N_{q-1} = 40$;
- $Q1 = 1.9 + 0.1\frac{50-40}{48} = 1.92$

- $j = 3$; $q = 4$
- individu 150 ;
- Intervalle $I_q = [2.1, 2.2[$
- $n_q = 40$; $a_q = 0.1$;
- $N_{q-1} = 144$
- $Q3 = 2.1 + 0.1\frac{150-144}{40} = 2.115$

Definition

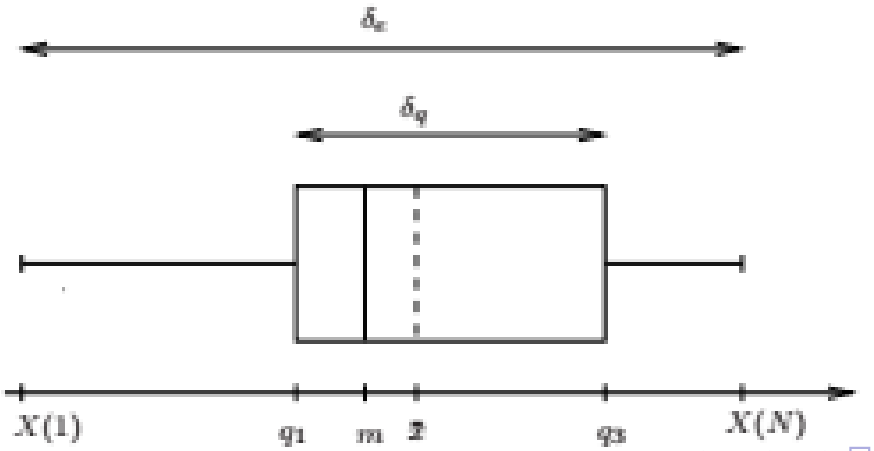
- L'**étendue** est définie par $X(max) - X(min)$
- L'**écart interquartile** est définie par $Q3 - Q1$

Représentation de la dispersion par une boîte à moustaches

Definition

La **boîte à moustaches** (box plot en anglais), est une représentation graphique mettant en exergue certains des paramètres de position et de dispersion précédents. Elle permet de visualiser sommairement la répartition des données de la série statistique et les données "extrêmes" (outliers). Elle est constituée :

- d'un diagramme en boîte : un rectangle de bornes $Q1$, $Q3$ coupé au niveau de la médiane m et éventuellement également au niveau de la moyenne $\bar{x}$;
- de moustaches : deux segments joignant $Q1$ à X(1) pour l'un et $Q3$ à X(N) pour l'autre.



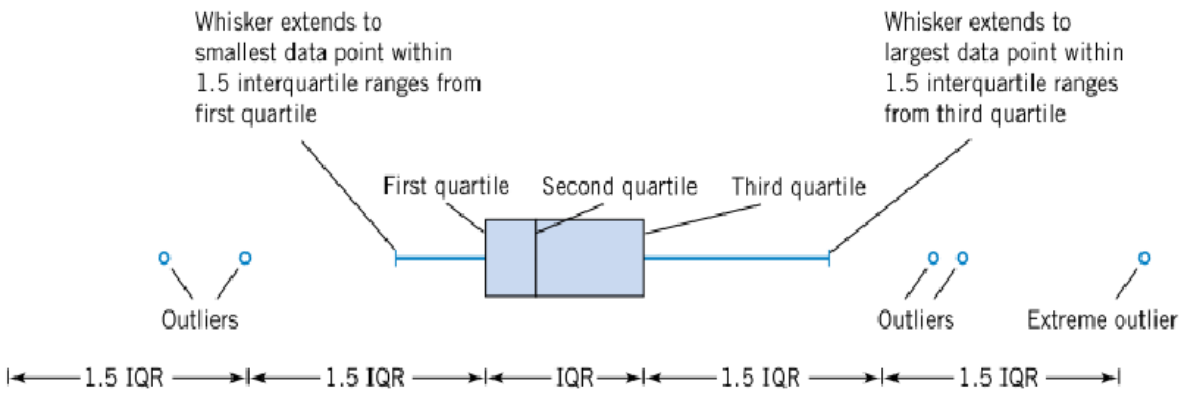
Représentation de la dispersion par une boîte à moustaches

Remarque

Des variantes sur les bornes des moustaches existents. Plutôt que de les faire aller jusqu'à X(1), X(N), on peut :

- , les faire aller jusqu'aux premier et neuvième décile,
- ou aux cinquième et 95-ième centiles
- ou aux premier et 99-ième centiles, etc.
- ou $25 - 1.5.(Q3 - Q1)$ et à $75 + 1.5.(Q3 - Q1)$

On appelle alors **données extrêmes ou outliers** les données à l'extérieur des moustaches.



Mesures liées à la forme : asymétrie

Décrire des données avec des mesures numériques

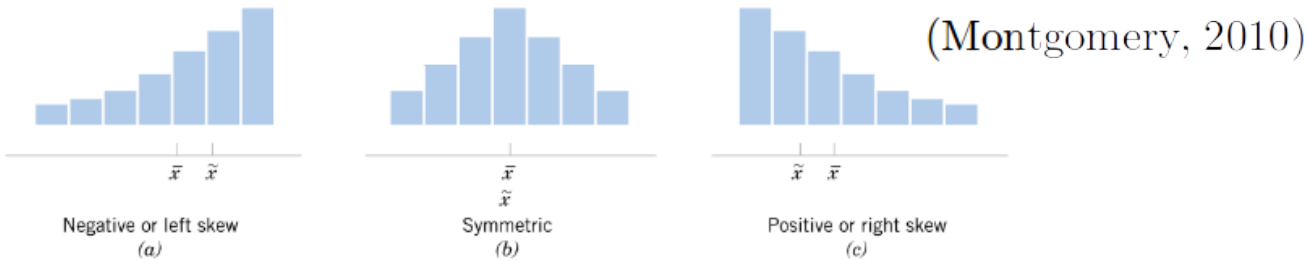
Definition

Le coefficient d'asymétrie de Fisher est la quantité suivante :

$$\gamma_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$$

Ce coefficient peut prendre des valeurs positives, négatives ou nulles :

- si la distribution est symétrique, il est nul
- si la distribution est allongée à gauche (les grandes valeurs sont plus fréquentes que les petites), il est négatif,
- si la distribution est allongée à droite (les petites valeurs sont plus fréquentes que les grandes), il est positif.



(Montgomery, 2010)

Definition

Le coefficient d'aplatissement de Fisher ou coefficient de Yule est la quantité suivante :

$$\gamma_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4} - 3$$

Une distribution est dite :

- mésokurtique si $\gamma_2 \simeq 0$ (la loi normale).
- leptokurtique si $\gamma_2 > 0$: histogramme plus pointu et queues plus longues (distribution moins aplatie que la distribution normale).
- platykurtique si $\gamma_2 < 0$: histogramme plus arrondi et queues plus courtes (distribution plus aplatie que la distribution normale).