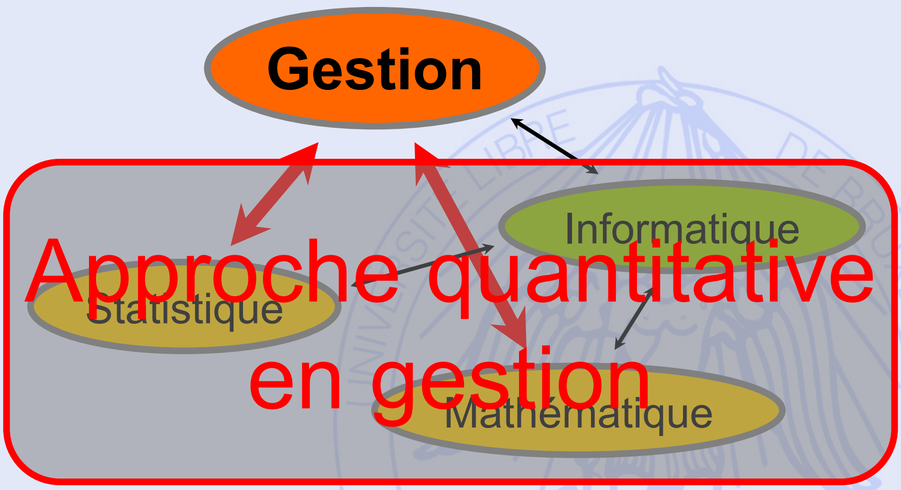
STATISTIQUE DESCRIPTIVE ET ELEMENTS DE PROBABILITES

Bertrand Mareschal
bmaresc@ulb.ac.be

http://homepages.ulb.ac.be/~bmaresc/stat-s-102.html



ULB UNIVERSITÉ LIBRE DE BRUXELLES

Statistique descriptive et élements de probabilités

- Notes de cours
  - Copie des transparents
    http://homepages.ulb.ac.be/~bmaresc/stat-s-102.html
  - Exercices : voir assistants
  - Ouvrages de référence optionnels
    - Notes de cours et exercices PUB
    - Eléments de Statistique (J.J. Droesbeke, Editions de l'ULB)
    - Autres livres…
- Examen écrit sans notes (janvier)
  - Partie théorique (QCM + question ouverte)
  - Partie pratique (questions ouvertes)

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Plan du cours

1. Introduction
2. Statistique descriptive - séries univariées
3. Calcul des probabilités
4. Arbres de décision
5. Variables aléatoires et lois de probabilité
6. Statistique descriptive - séries bivariées
7. Méthodes de prévision

ULB UNIVERSITÉ LIBRE DE BRUXELLES

1. Introduction

- Contexte
- Historique
- Prise de décision
- Aide à la décision
- Modélisation
- Principaux outils
- Statistique

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Contexte

- Augmentation de la taille et de la complexité des organisations.
- Division du travail, spécialisation, décentralisation des responsabilités et de la gestion.
- Nouveaux problèmes liés à la spécialisation :
  - Plus grande autonomie des départements au sein des organisations,
  - Manque de coordination,
  - Objectifs conflictuels,
  - Difficulté d'allouer des ressources limitées aux départements d'une façon globalement optimale.

ULB UNIVERSITÉ LIBRE DE BRUXELLES

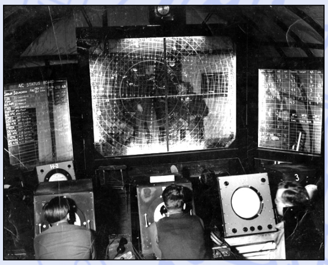
ULB UNIVERSITÉ LIBRE DE BRUXELLES

Historique
2ème guerre mondiale

- Allocation de ressources limitées aux opérations militaires.
- Idée : approche scientifique (UK - USA).
- "Research on Operations" par des équipes multidisciplinaires de scientifiques (Cf. "Blackett's Circus", UK).
- Grand succès : amélioration de l'efficacité des opérations militaires complexes
  - déploiement des radars en Angleterre,
  - determination de la taille des convois,
  - logistique …

Déploiement des radars

Déploiement des radars

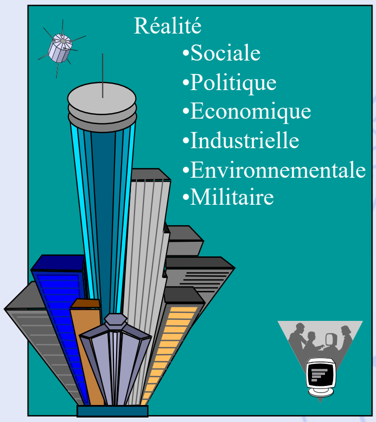


Protection des convois

Historique
Après-guerre

- Succès des applications militaires.
- Intérêt marqué des entreprises pour la RO.
- Applications civiles, d'abord dans les grandes entreprises industrielles :
  - Ex: industrie pétrolière - programmation linéaire pour la gestion de la production
- Plus tard, resultats utilisés (à moindre coût) par des organisations plus petites.
- Facteur clé : développement de l'informatique.

Prise de Décision

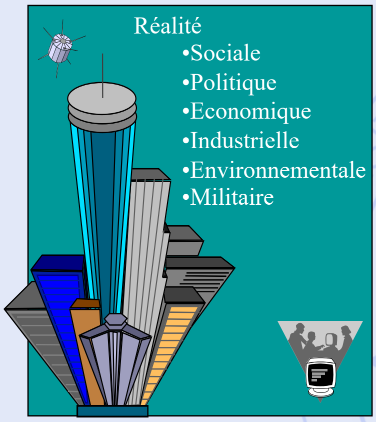


- Décrire la Réalité,
- Comprendre la Réalité,
- Gérer la Réalité.

2 Approches :

- Approche Qualitative,
- <u>Approche Quantitative</u>.

Aide à la Décision

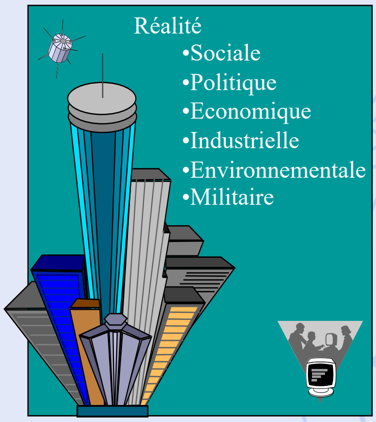


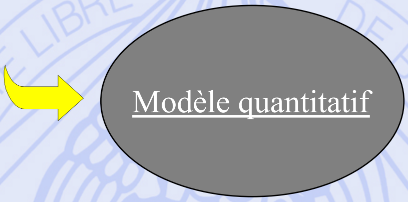
<u>Modèle quantitatif</u>

- Décisions possibles ?
- Comment les comparer ?
- Préférences, Objectifs ?

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Aide à la Décision





- Approximation de la réalité !

➢ Aide à la décision.

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Quelques techniques

- Statistique
- Programmation mathématique (optimisation)
- Aide à la décision de type multicritère
- Simulation
- Gestion de projets : PERT/CPM
- Gestion des stocks et de la production
- Réseaux (transport)
- Fiabilité des équipements

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Statistique ?

- Techniques et procédures pour la collecte, la description, l'analyse et l'interprétation de données.
- Transformer des données en information.
- Données, numériques ou non :
  - démographie, emploi, ventes, production, stocks, index des prix, intentions de votes, habitudes de consommation, …

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Sondages

Grand Baromètre: le Vlaams Belang devient le plus grand parti de Flandre

RTL INFO avec Antonio Solimando, publié le 13 septembre 2019 à 19h00, mis à jour à 20h41

...

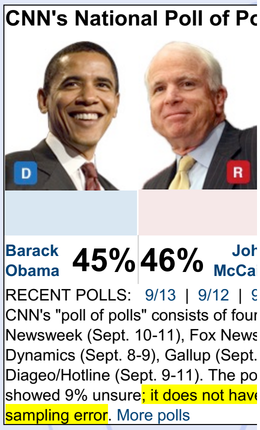
Le Vlaams Belang devient le plus grand parti de Flandre devant la N-VA. C'est l'enseignement majeur de notre Grand Baromètre de rentrée ce vendredi soir. Si on devait voter ce dimanche, le Vlaams Belang récolterait **24,9%** des voix, soit une hausse de 6,2 points par rapport à son score lors des élections du 26 mai.

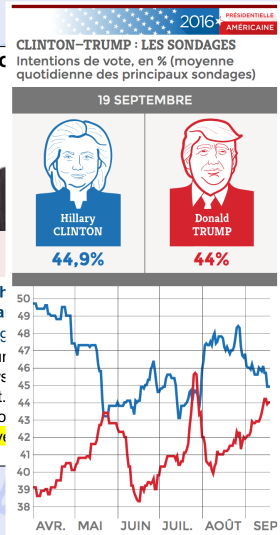
La N-VA, elle, perdrait près de 3 points (-2,8) avec un score de **22,7%**. Attention, il y a toutefois une nuance à faire: l'écart entre les deux partis est faible, sous les marges d'erreurs. La première place du parti d'extrême droite est donc surtout symbolique.

*Les résultats se basent sur une vague de **2540** répondants, formant des échantillons représentatifs des Belges de 18 ans et plus à raison de **992** en Wallonie, 1000 en Flandre et **548** dans les 19 communes de la Région Bruxelles-Capitale, réalisée du 2 au 10 septembre 2019. Les interviews ont eu lieu en ligne. La marge d'erreur maximale, pour un pourcentage de 50% et un taux de confiance de 95% est de **+-3,1** en Wallonie et en Flandre et de **+-4,2** à Bruxelles. Affiliations: ESOMAR, Consumer Understanding Belgium.*

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Sondages





ULB UNIVERSITÉ LIBRE DE BRUXELLES

Sondages

Elections aux USA : les sondages se suivent et ne se ressemblent pas

RTL.be – 18/09/2004

A moins deux de deux mois des élections présidentielles aux Etats-Unis, les sondages d'opinion sensés donner une vision nette des intentions de vote des Américains se multiplient. Aujourd'hui, une enquête de l'institut Gallup, réalisée entre le 13 et le 15 septembre, donne 55 pour cent d'intentions de vote à l'actuel président républicain, contre 42 pour cent à son adversaire démocrate. Ce résultat marque, souligne Gallup, un "progrès significatif" de George W. Bush par rapport à un autre sondage réalisé juste avant le début de la convention républicaine le 30 août, qui lui donnait une avance de trois points (50 pc, contre 47 pc à John Kerry).

Ces conclusions contredisent deux autres sondages publiés hier et qui donnaient jeudi les deux candidats au coude à coude. Pew a sondé un échantillon des électeurs en deux entre le 8 et le 14 septembre. Ses résultats, explique l'institut, montrent un spectaculaire resserrement de l'écart entre MM. Bush et Kerry. Une nette majorité des électeurs inscrits (52 pc contre 40 pc) se disait prêts à voter Bush entre le 8 et le 10 septembre, quelques jours seulement après la fin de la convention républicaine. Mais quelques jours plus tard, entre le 11 et le 14 septembre, MM. Bush et Kerry étaient à égalité avec 46 pc des voix chacun.

Une troisième étude publiée par l'édition en ligne du Wall Street Journal accordait 48 pc des intentions de vote à M. Kerry, contre 47 pc pour M. Bush. Plusieurs autres sondages publiés depuis la fin de la convention républicaine le 2 septembre accordaient une confortable avance au président Bush, allant jusqu'à 11 points.

Exemple 1 - Pepsi

In the last few years, colleges and universities have signed exclusivity agreements with a variety of private companies. These agreements bind the university to sell that company's products exclusively on the campus. Many of the agreements involve food and beverage firms.

A large university with a total enrollment of about 50,000 students has offered Pepsi-Cola an exclusivity agreement that would give Pepsi exclusive rights to sell its products at all university facilities for the next year and an option for future years. In return, the university would receive 35% of the on-campus revenues and an additional lump sum of $200,000 per year. Pepsi has been given 2 weeks to respond.

The management at Pepsi quickly reviews what they know. The market for soft drinks is measured in terms of 12-ounce cans. Pepsi currently sells an average of 22,000 cans per week (over the 40 weeks of the year that the university operates). The cans sell for an average of 75 cents each. The costs including labor amount to 20 cents per can. Pepsi is unsure of its market share but suspects it is considerably less than 50%. A quick analysis reveals that if its current market share were 25%, then, with an exclusivity agreement, Pepsi would sell 88,000 (22,000 is 25% of 88,000) cans per week or 3,520,000 cans per year. The gross revenue would be computed as follows:

$$3{,}520{,}000 \times \$.75/\text{can} = \$2{,}640{,}000$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Exemple 1 - Pepsi

This figure must be multiplied by 65% because the university would rake in 35% of the gross. Thus,

$$65\% \times \$2{,}640{,}000 = \$1{,}716{,}000$$

The total cost of 20 cents per can (or $704,000) and the annual payment to the university of $200,000 are subtracted to obtain the net profit:

$$\text{Net profit} = \$1{,}716{,}000 - \$704{,}000 - \$200{,}000 = \$812{,}000$$

Pepsi's current annual profit is

$$40 \text{ weeks} \times 22{,}000 \text{ cans/week} \times \$.55 = \$484{,}000$$

If the current market share is 25%, the potential gain from the agreement is

$$\$812{,}000 - \$484{,}000 = \$328{,}000$$

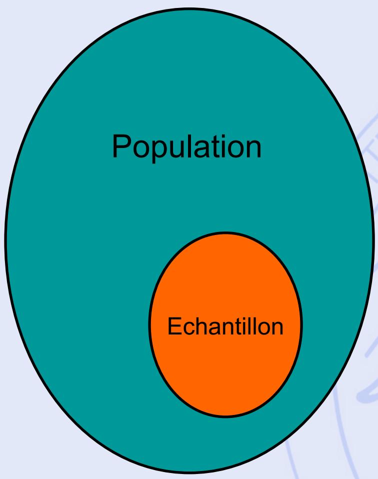
The only problem with this analysis is that Pepsi does not know how many soft drinks are sold weekly at the university. Coke is not likely to supply Pepsi with information about its sales, which together with Pepsi's line of products constitute virtually the entire market.

Pepsi assigned a recently hired university graduate to survey the university's students to supply the missing information. Accordingly, she organizes a survey that asks 500 students to keep track of the number of soft drinks they purchase in the next 7 days.

Statistique - 2 composantes

- Statistique descriptive
  - Organiser, résumer, représenter un ensemble de données.
  - Représentations graphiques.
  - Mesures statistiques (moyenne,…).
- Inférence statistique
  - Tirer des conclusions pour une population à partir d'un échantillon prélevé dans cette population.
  - Mesurer et prendre en compte l'erreur d'échantillonnage.

Terminologie



- Population
  - Ensemble des éléments étudiés. (individus, pays, entreprises, …)
- Echantillon
  - Sous-ensemble de la population pour lequel on récolte des observations et qui peut être analysé.

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Terminologie

- Variable
  - Caractéristique mesurable de la population. (chiffre d'affaire d'une entreprise, temps consacré par un individu à l'utilisation d'une base de données, …)
- Paramètre
  - Quantité numérique qui résume un aspect de la population. (valeur moyenne d'une variable)
- Statistique
  - Quantité numérique qui résume un aspect d'un échantillon. (valeur moyenne observée sur un échantillon)

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Types de données

- Nominales
  - Catégories
  - Ex: sexe, état-civil, nationalité, secteur économique, …
  - Dénombrement
- Ordinales
  - Catégories ordonnées
  - Ex: mentions (Aj., réussi, S, D, GD, LPGD), tailles (S,M,L,XL), …
  - Dénombrements, comparaison
- Numériques
  - Nombres
  - Ex: cotes, quantités vendues, revenus, …
  - Dénombrement, comparaison, différence, rapport, moyenne, …

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Plan du cours

ULB UNIVERSITÉ LIBRE DE BRUXELLES

2. Stat. Descriptive - 1 dim

- **Objectif** : Résumer les caractéristiques d'un (grand) ensemble de données. Mettre en évidence les points importants.
- **Tableaux** : Première étape. Tri et regroupement des observations. Distribution de fréquences.
- **Graphiques** : Visualisation des données. Lignes, barres, histogrammes, diagrammes en secteurs, …
- **Mesures** : Synthèse des données en quelques grandeurs représentatives.
  - Quel est l'ordre de grandeur des valeurs observées ?
    Paramètres de position (moyenne, médiane, …)
  - Y a-t-il de grands écarts entre les valeurs observées ?
    Paramètres de dispersion (variance, écart-type, …)

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Données brutes

$$\{x_1, x_2, \cdots, x_n\} = \{x_i; i = 1, \cdots, n\}$$

- Données telles que recueillies
  - n observations d'une variable x : série statistique univariée
  - Exemple 2 : Ages de 100 employés d'une entreprise (échantillon…)

60	39	23	30	29	26	29	41	40	32
63	22	32	52	46	35	25	28	33	33
20	25	42	34	29	43	41	31	30	36
58	21	24	55	51	28	18	40	44	38
32	21	30	31	25	49	31	26	33	36
43	34	35	22	33	38	34	34	33	34
23	26	57	23	26	36	39	31	35	34
34	51	40	50	35	45	28	36	32	39
26	48	17	45	45	25	25	30	36	30
43	25	27	21	53	25	38	33	37	33

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Série ordonnée

- Si x est ordinale ou quantitative

$$\left\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\right\} \text{ avec } x_{(i)} \leq x_{(j)} \text{ si } i \leq j$$

17	18	20	21	21	21	22	22	23	23
23	24	25	25	25	25	25	25	25	26
26	26	26	26	27	28	28	28	29	29
29	30	30	30	30	30	31	31	31	31
32	32	32	32	33	33	33	33	33	33
33	34	34	34	34	34	34	34	35	35
35	35	36	36	36	36	36	37	38	38
38	39	39	39	40	40	40	41	41	42
43	43	43	44	45	45	45	46	48	49
50	51	51	52	53	55	57	58	60	63

Distribution observée

Age	Effectif
15-19	2
20-24	10
25-29	19
30-34	27
35-39	16
40-44	10
45-49	6
50-54	5
55-59	3
60-64	2
Total	100

- Données regroupées en classes (intervalles).
- Effectif d'une classe = nombre d'observations dans cette classe.

Exemple 3

- Nb de voitures dans 10 familles de même taille.

i	1	2	3	4	5	6	7	8	9	10
x_i	2	0	1	1	2	3	1	1	1	0

$\boxed{n=10}$

- Série observée :

$$\{x_i\} = \{2,0,1,1,2,3,1,1,1,0\}$$

- Série ordonnée :

$$\{x_{(i)}\} = \{0,0,1,1,1,1,1,2,2,3\}$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Exemple 3

- Distribution observée :

Valeurs observées

Effectifs associés

x_j	0	1	2	3
n_j	2	5	2	1

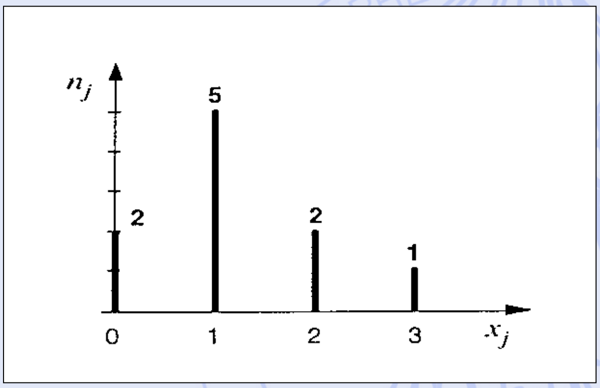
$$j = 1,2,3,4 \quad J = 4 = \text{ nb de valeurs distinctes}$$

$$n_1 + n_2 + \ldots + n_J = \sum_{j=1}^{J} n_j = n$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Représentations graphiques

- Diagramme en bâtons : exemple 3

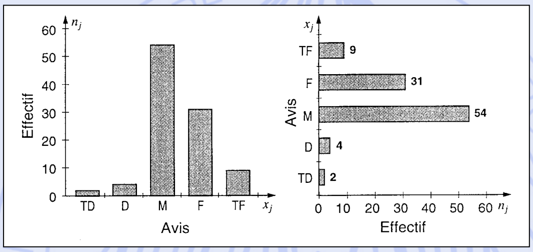


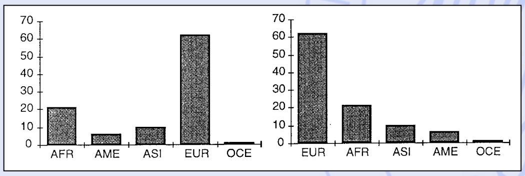
ULB UNIVERSITÉ LIBRE DE BRUXELLES

Représentations graphiques

- Diagramme en barres : pour variables qualitatives

Cas ordinal
(avis pédagogiques)



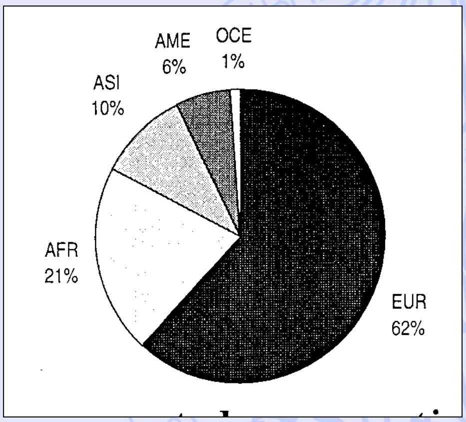


Cas nominal
(continents)

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Représentations graphiques

- Diagramme en secteurs (tarte, camembert):



ULB UNIVERSITÉ LIBRE DE BRUXELLES

Fréquences

- Définition : $f_j = \frac{n_j}{n} \quad j = 1, 2, \ldots, J$
- Exemple 3 : distribution de fréquences

x_j	n_j	f_j
0	2	0.2
1	5	0.5
2	2	0.2
3	1	0.1

$$n = 10$$

$$\sum_{j=1}^{J} f_j = 1$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Effectifs et fréquences cumulés

- Effectifs cumulés :

nombre d'observations $\leq x_j$

$$N_j = n_1 + n_2 + \ldots + n_j$$

- Fréquences cumulées :

pourcentage d'observations $\leq x_j$

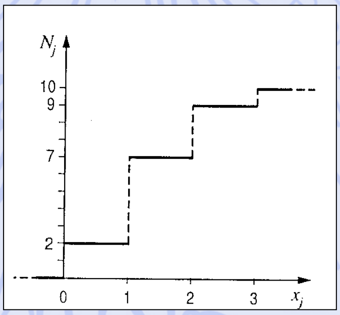
$$F_j = \frac{N_j}{n}$$

Effectifs et fréquences cumulés

- Exemple 3 :

x_j	n_j	f_j	N_j	F_j
0	2	0.2	2	0.2
1	5	0.5	7	0.7
2	2	0.2	9	0.9
3	1	0.1	10	1.0

Courbe cumulative



Effectifs et fréquences cumulés à droite

- Effectifs cumulés à droite :

nombre d'observations $\geq x_j$

$$N_j^* = n_j + n_{j+1} + \ldots + n_J$$

- Fréquences cumulées à droite :

pourcentage d'observations $\geq x_j$

$$F_j^* = \frac{N_j^*}{n}$$

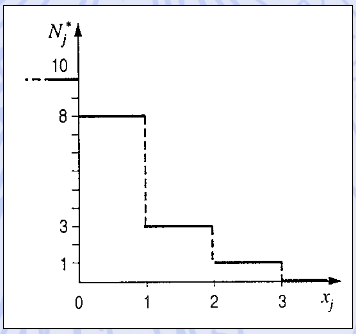
ULB UNIVERSITÉ LIBRE DE BRUXELLES

Effectifs et fréquences cumulés à droite

- Exemple 3 :

x_j	n_j	f_j	N^*_j	F^*_j
0	2	0.2	10	1.0
1	5	0.5	8	0.8
2	2	0.2	3	0.3
3	1	0.1	1	0.1

Courbe cumulative à droite



ULB UNIVERSITÉ LIBRE DE BRUXELLES

Distribution groupée

- Cf. exemple 2 :
  - Grand nombre de valeurs distinctes,
  - Petits effectifs.
- Regroupements des valeurs en classes

Exemple 4

- Série de 175 tailles

176	175	185	176	190	163	185
166	152	178	172	178	190	176
177	160	183	186	182	154	172
150	168	174	180	173	174	157
181	171	181	192	188	171	176
169	184	168	179	185	189	165
175	170	176	184	189	176	171
173	189	177	170	165	182	174
184	178	161	150	175	171	194
162	187	166	170	176	183	175
169	198	185	171	184	177	166
171	180	158	177	155	184	195
193	176	192	169	181	178	180
154	155	196	175	175	190	191
170	165	151	191	177	153	180
155	177	187	180	169	198	156
186	164	178	187	183	175	183
163	182	177	183	170	169	182
184	170	190	164	186	164	160
176	175	175	185	169	170	174
180	173	162	161	196	181	181
188	182	181	177	179	177	179
178	172	163	165	198	167	180
169	184	172	167	151	183	161
177	175	168	174	197	179	179

150	160	165	169	172	175	177	178	181	184	187	192
150	161	165	170	172	175	177	179	181	184	187	193
151	161	166	170	172	175	177	179	181	184	187	194
151	161	166	170	173	175	177	179	181	184	188	195
152	162	166	170	173	175	177	179	182	184	188	196
153	162	167	170	173	175	177	179	182	184	189	196
154	163	167	170	174	176	177	180	182	184	189	197
154	163	168	170	174	176	177	180	182	185	189	198
155	163	168	171	174	176	177	180	182	185	190	198
155	164	168	171	174	176	177	180	183	185	190	198
155	164	169	171	174	176	178	180	183	185	190	
156	164	169	171	175	176	178	180	183	185	190	
157	165	169	171	175	176	178	180	183	186	191	
158	165	169	171	175	176	178	181	183	186	191	
160	165	169	172	175	176	178	181	183	186	192	

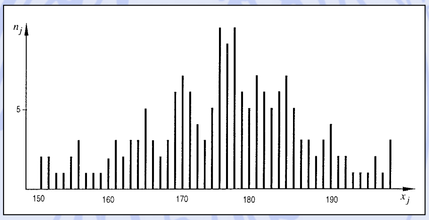


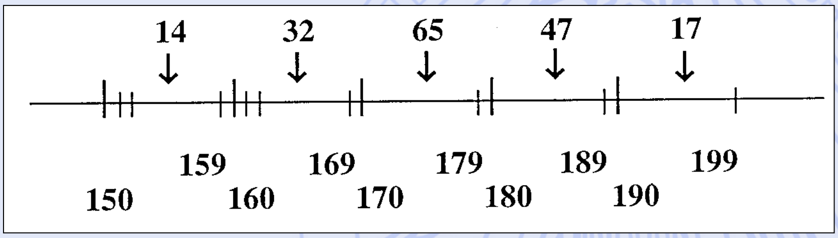
Diagramme en tiges et feuilles

- « Stem and leaf display »

15	04525810514376
16	69253908548162389415759939475601
17	67531068751086705325846787526290017574835657094161785079626145499
18	14640849702245315716040735285941369234135003210
19	38260210687008451

Groupement en classes

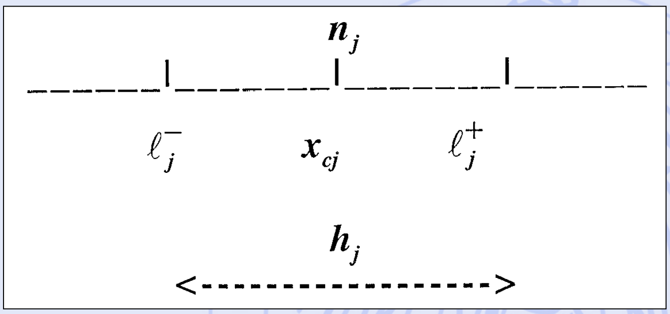
- Intervalles disjoints
- Recouvrir toutes les valeurs observées



ULB UNIVERSITÉ LIBRE DE BRUXELLES

Groupement en classes

- J classes : $j = 1, 2, \ldots, J$



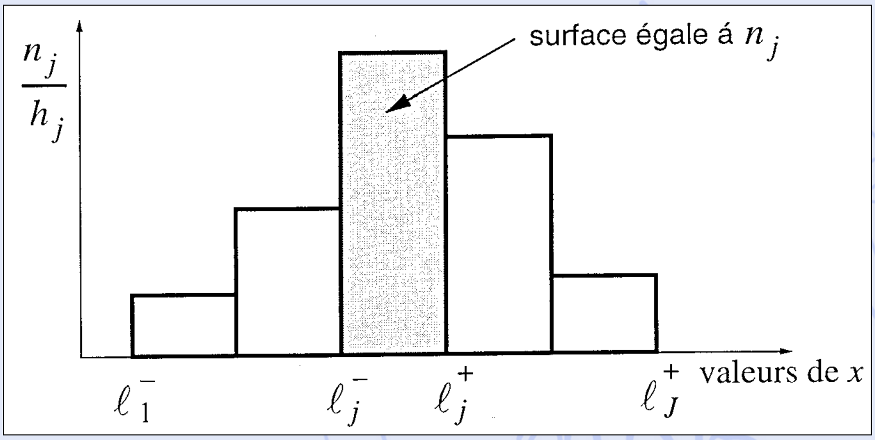
- J = ? Règle de Sturges :

$$J \approx 1 + \frac{10}{3} \log_{10} n$$

Distribution groupée

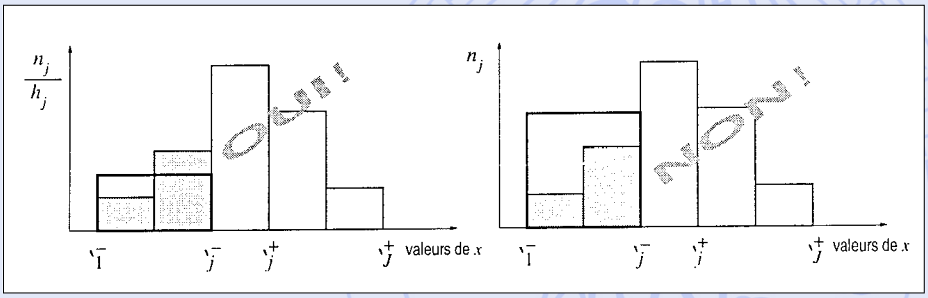
$(\ell_j^- ; \ell_j^+)$	x_{cj}	n_j	f_j	N_j	F_j	N_j^*	F_j^*
149.5-159.5	154.5	14	0.080	14	0.080	175	1
159.5-169.5	164.5	32	0.183	46	0.263	161	0.920
169.5-179.5	174.5	65	0.371	111	0.634	129	0.737
179.5-189.5	184.5	47	0.269	158	0.903	64	0.366
189.5-199.5	194.5	17	0.097	175	1	17	0.097
		n=175	1				

Histogramme des effectifs



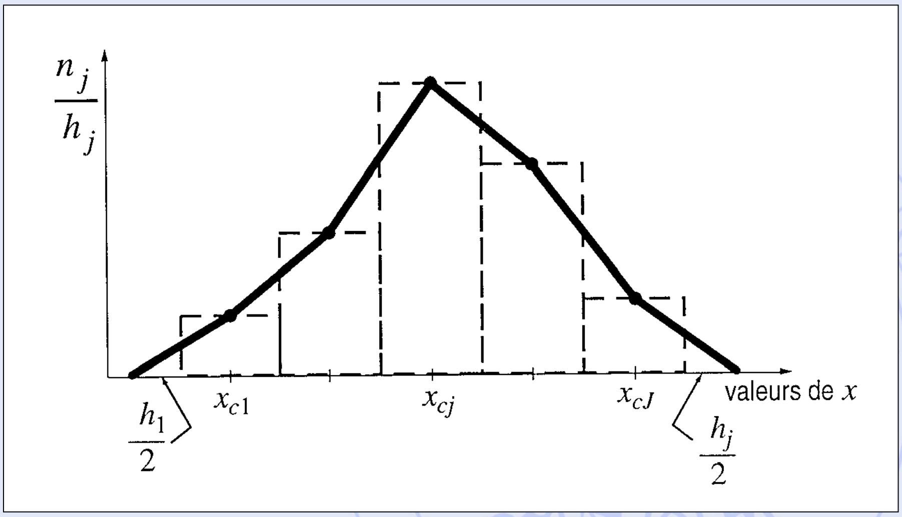
Histogramme des effectifs

- Attention à l'unité sur l'axe vertical !
- Cas de classes de longueurs différentes.



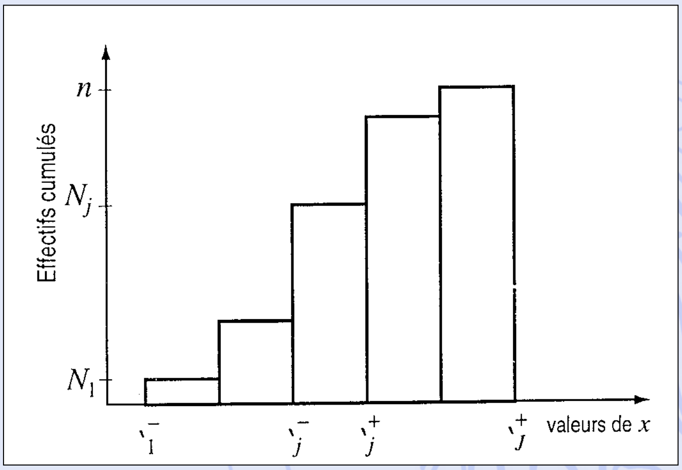
ULB UNIVERSITÉ LIBRE DE BRUXELLES

Polygone des effectifs



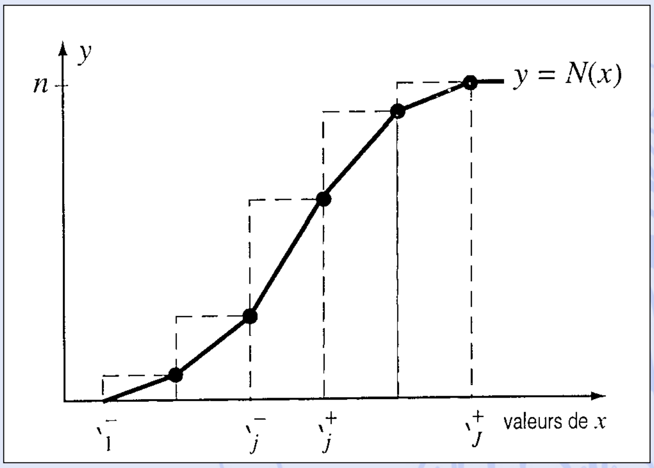
ULB UNIVERSITÉ LIBRE DE BRUXELLES

Histogramme des effectifs cumulés



ULB UNIVERSITÉ LIBRE DE BRUXELLES

Courbe cumulative



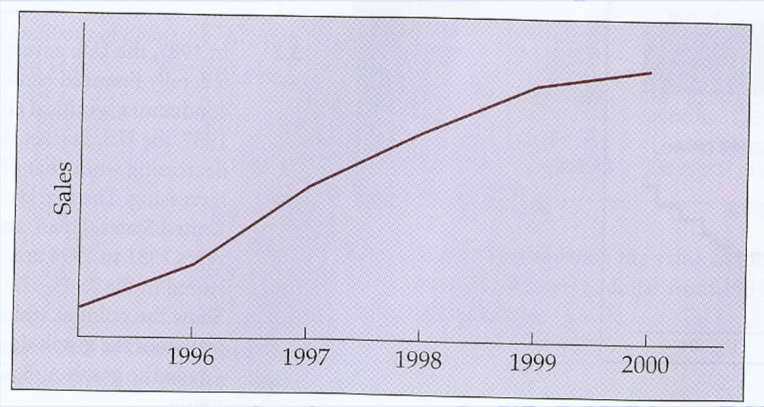
ULB UNIVERSITÉ LIBRE DE BRUXELLES

Utilisation de graphiques

- Inclure informations utiles.
- Eviter éléments inutiles !
- Préférer la simplicité à la sophistication.
- Choix des unités et des axes.
- …

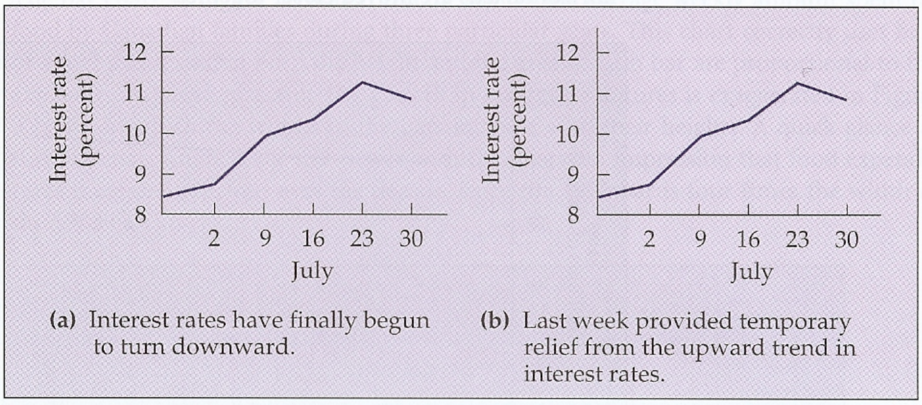
ULB UNIVERSITÉ LIBRE DE BRUXELLES

Graphiques !



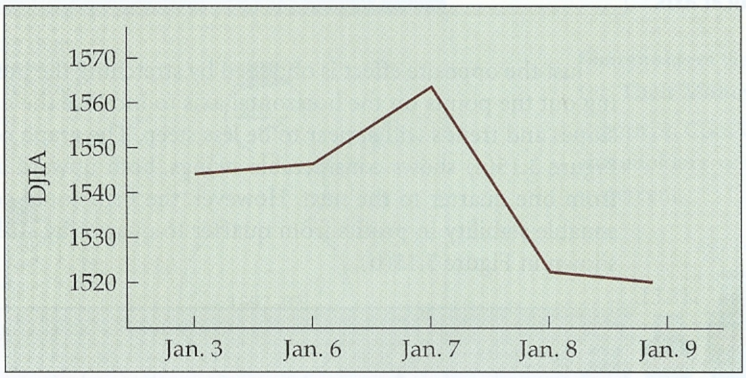
ULB UNIVERSITÉ LIBRE DE BRUXELLES

Graphiques !

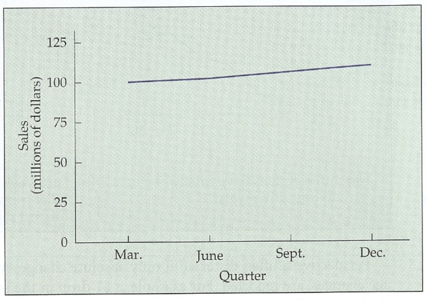


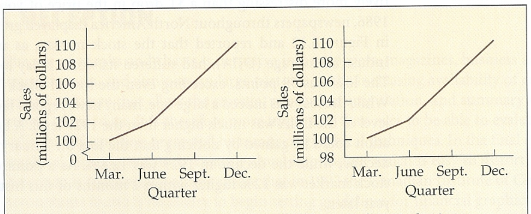
ULB UNIVERSITÉ LIBRE DE BRUXELLES

Graphiques !



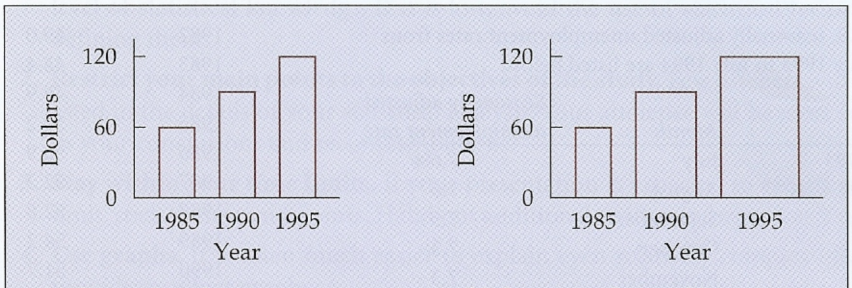
Graphiques !





(a) Break in vertical axis (b) Truncated vertical axis

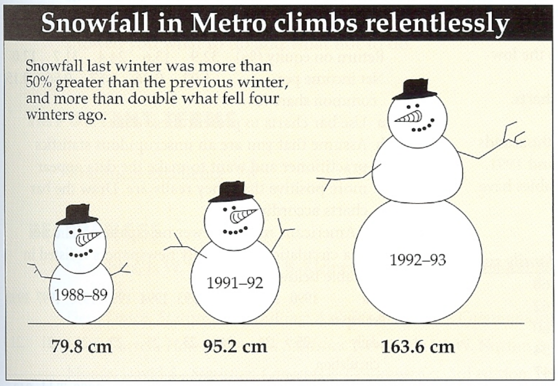
Graphiques !

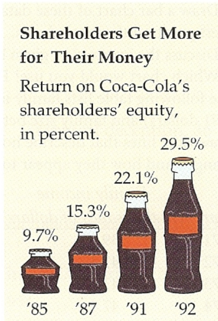


(a) Correct bar chart (b) Increasing bar widths to create distortion

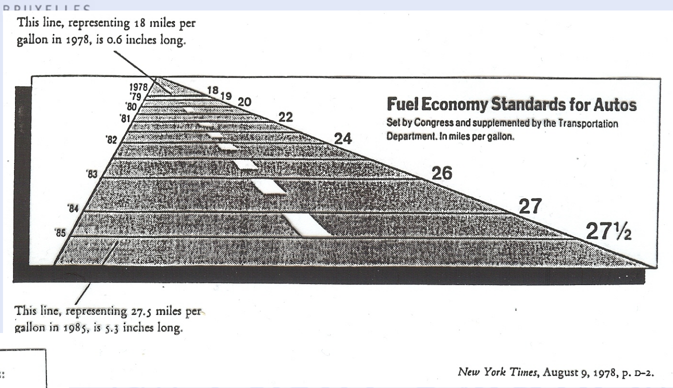
ULB UNIVERSITÉ LIBRE DE BRUXELLES

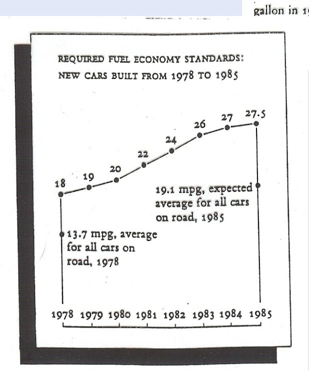
Graphiques !

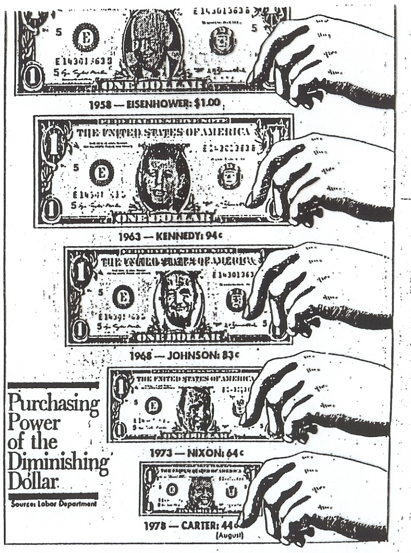




ULB UNIVERSITÉ LIBRE DE BRUXELLES







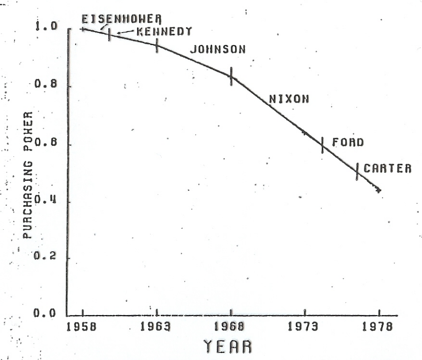


Figure 9. An example of how to goose up the effect by squaring the eyeball (© 1978, The Washington Post).

Figure 10. The data in Figure 9 as an unadorned line chart (from Wainer, 1980).

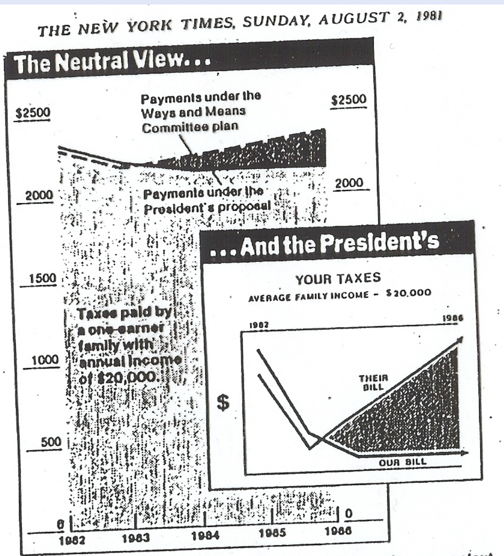


Figure 11. The White House showing neither scale nor context (© 1981, The New York Times, reprinted with permission).

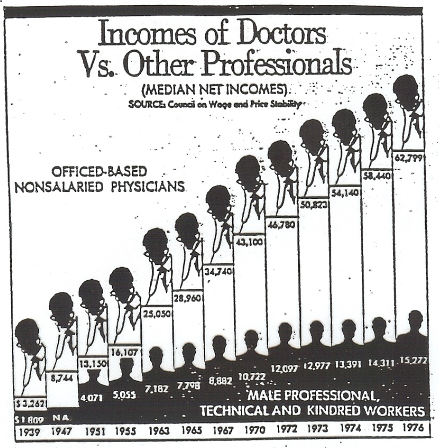


Figure 13. Changing scale in mid-axis to make exponential growth linear (© The Washington Post).

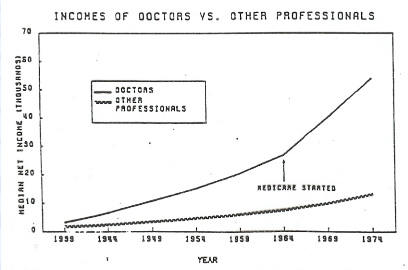


Figure 14. Data from Figure 13 redone with linear scale (from Wainer 1980).

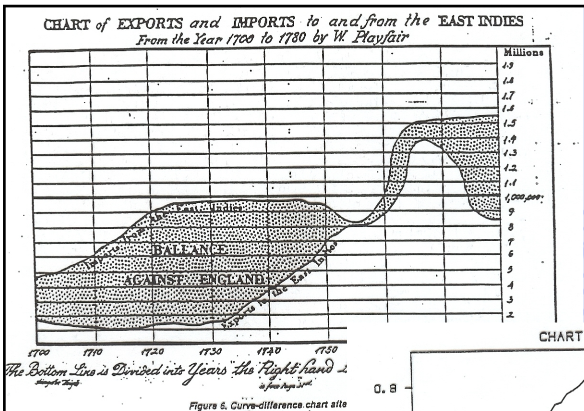


Figure 6. Curve-difference chart afte

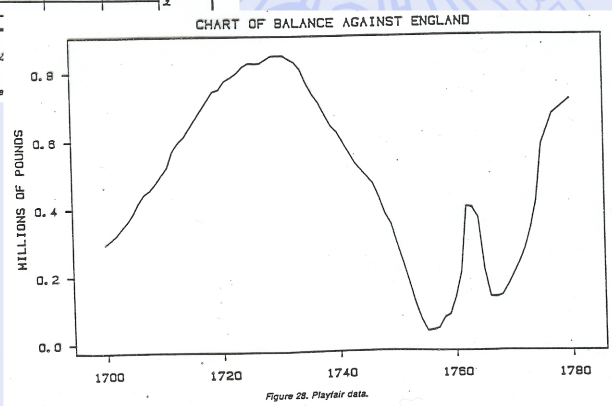


Figure 28. Playfair data.

ULB UNIVERSITÉ LIBRE DE BRUXELLES

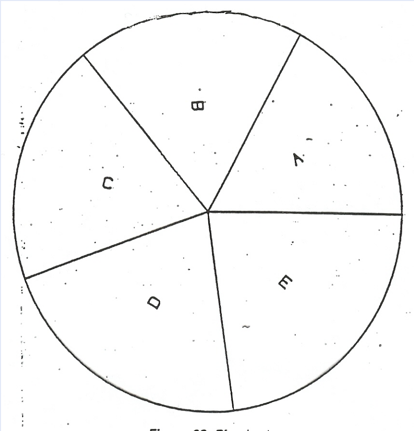


Figure 22. Pie chart.

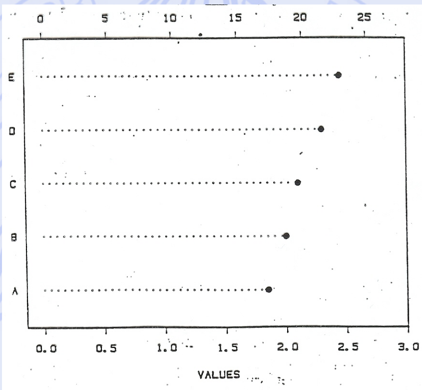


Figure 23. Dot chart.

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Statistique du jour

Davis challenges Schwarzenegger to debate

Issa endorses actor-turned-politician

CNN.com, Friday, September 26, 2003 Posted: 8:14 PM EDT (0014 GMT)

LOS ANGELES, California (CNN) -- Charging that Arnold Schwarzenegger is misleading voters in the recall campaign, Democratic Gov. Gray Davis on Friday challenged the actor-turned-politician to a debate.

...

The latest poll in the recall race, by the Public Policy Institute of California, showed Bustamante with support from 28 percent of likely voters, Schwarzenegger with 26 percent and McClintock with 14 percent. With a margin of error of plus-or-minus 3 percentage points, Bustamante and Schwarzenegger were locked in a statistical tie.

...

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Paramètres d'une série

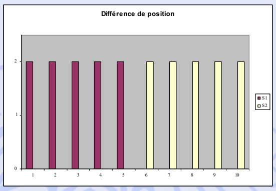
- Résumer la série en quelques grandeurs-clé !
- 3 catégories de paramètres :
  - Position
    - Valeur typique ? Valeur « centrale » ?
    - Moyenne, médiane, mode, …
  - Dispersion
    - Ecarts entre valeurs observées ?
    - Etendue, écart interquartile, variance, écart-type, …
  - Forme
    - Symétrie ? Asymétrie ?

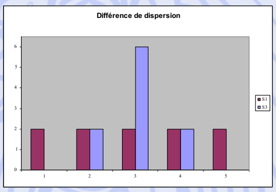
ULB UNIVERSITÉ LIBRE DE BRUXELLES

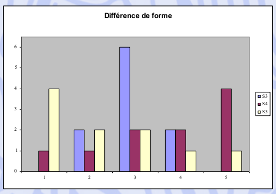
Exemples

i	*S1*	*S2*	*S3*	*S4*	*S5*
1	1	6	2	3	3
2	2	7	3	2	4
3	2	7	3	4	2
4	1	6	2	1	5
5	3	8	3	3	3
6	4	9	3	5	1
7	5	10	4	5	1
8	3	8	3	4	2
9	4	9	3	5	1
10	5	10	4	5	1

Moyenne	3,00	8,00	3,00	3,70	2,30
Médiane	3,00	8,00	3,00	4,00	2,00
Mode	1,00	6,00	3,00	5,00	1,00
Variance	2,00	2,00	0,40	1,81	1,81
Ecart-type	1,41	1,41	0,63	1,35	1,35
Skewness	0,00	0,00	0,00	-0,80	0,80
Kurtosis	-1,33	-1,33	0,08	-0,38	-0,38







Paramètres de position

- Objectif : déterminer une valeur centrale
- ❶ Moyenne arithmétique
  - Exemple 3 :

$$\{x_i\} = \{2,0,1,1,2,3,1,1,1,0\}$$

$$\bar{x} = \frac{2+0+1+1+2+3+1+1+1+0}{10} = 1.2$$

  - Définition :

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

Moyenne

- Propriétés de la moyenne arithmétique
  - Uniquement pour données numériques !
  - Fortement influencée par les valeurs extrêmes :

$$\{x_i\} = \{2,0,1,1,2,30,1,1,1,0\}$$

$$\bar{x} = \frac{2+0+1+1+2+30+1+1+1+0}{10} = 3.9$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Moyenne

- Propriétés de la moyenne arithmétique
  - Agrégation de 2 séries :

$$[n_1 \quad \bar{x}_1] \quad [n_2 \quad \bar{x}_2]$$
$$\Downarrow$$
$$n = n_1 + n_2 \qquad \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n}$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Moyenne

- Moyenne arithmétique pondérée

$$\begin{array}{cccc} x_1 & x_2 & \dots & x_m \\ \downarrow & \downarrow & & \downarrow \\ w_1 & w_2 & \dots & w_m \end{array} \qquad \begin{array}{c} w_k \geq 0 \\ \sum_{k=1}^{m} w_k = 1 \end{array}$$

$$\bar{x}_w = \sum_{k=1}^{m} w_k x_k$$

Moyenne

- Pour une distribution observée :

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j = \sum_{j=1}^{J} f_j x_j$$

- Exemple 3 :

x_j	n_j	f_j
0	2	0.2
1	5	0.5
2	2	0.2
3	1	0.1

$$\begin{aligned}\bar{x} &= \frac{1}{10}(2\times 0+5\times 1+2\times 2+1\times 3) \\ &= 0,2\times 0+0,5\times 1+0,2\times 2+0,1\times 3 \\ &= 1,2\end{aligned}$$

Moyenne

- Pour une distribution groupée :

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{J} n_j x_{cj}$$

$(\ell_j^- ; \ell_j^+)$	x_{cj}	n_j	f_j	N_j	F_j	N_j^*	F_j^*
149.5-159.5	154.5	14	0.080	14	0.080	175	1
159.5-169.5	164.5	32	0.183	46	0.263	161	0.920
169.5-179.5	174.5	65	0.371	111	0.634	129	0.737
179.5-189.5	184.5	47	0.269	158	0.903	64	0.366
189.5-199.5	194.5	17	0.097	175	1	17	0.097
		n=175	1				

$$\bar{x} = \frac{14\times 154,5+32\times 164,5+\ldots+17\times 194,5}{175}$$

☞ Approximation !

Paramètres de position

- ❷ Médiane : $x_{1/2}$
  Valeur telle que le nombre d'observations qui la précèdent est égal au nombre d'observations qui la suivent, dans la série ordonnée.
- Exemple : $\{x_{(i)}\} = \{1,3,7,8,9,13,17\}$

$$\rightarrow x_{1/2} = 8$$

Médiane

- Pour n impair : $x_{1/2} = x_{\left(\frac{n+1}{2}\right)}$
- Pour n pair : (convention) $x_{1/2} = \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$

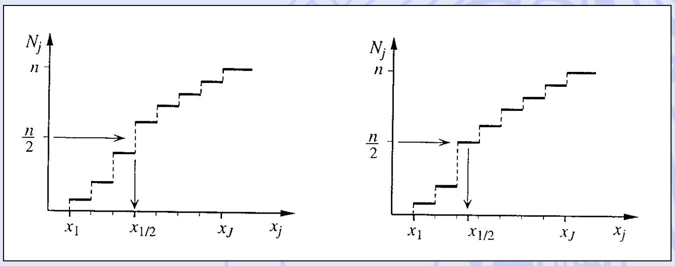
$$\{x_{(i)}\} = \{1,3,7,8,9,13\} \rightarrow x_{1/2} = \frac{7+8}{2} = 7,5$$

- Avantage de la médiane : pas influencée par les valeurs extrêmes.

Médiane

- Pour une distribution observée :

$$x_{1/2} \text{ telle que } N(x_{1/2}) = N^*(x_{1/2})$$



$$N(x_{1/2}) = N^*(x_{1/2}) \geq \frac{n}{2}$$

Médiane

- Pour une distribution groupée :

$$x_{1/2} \text{ solution de } N(x_{1/2}) = N^*(x_{1/2}) = \frac{n}{2}$$

  - Classe médiane : (l_m^-, l_m^+)
  - Valeur approchée de la médiane :

$$x_{1/2} = l_m^- + h_m \frac{\frac{n}{2} - N_{m-1}}{n_m}$$

Médiane

- Exemple 4 :

$(\ell_j^- ; \ell_j^+)$	x_{cj}	n_j	f_j	N_j	F_j	N_j^*	F_j^*
149.5-159.5	154.5	14	0.080	14	0.080	175	1
159.5-169.5	164.5	32	0.183	46	0.263	161	0.920
169.5-179.5	174.5	65	0.371	111	0.634	129	0.737
179.5-189.5	184.5	47	0.269	158	0.903	64	0.366
189.5-199.5	194.5	17	0.097	175	1	17	0.097
		n=175	1				

$$\frac{n}{2} = \frac{175}{2} = 87{,}5$$

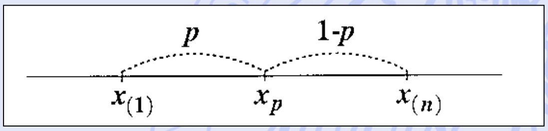
$$(l_m^-, l_m^+) = (169{,}5 \ ; \ 179{,}5)$$

$$x_{1/2} = 169{,}5 + 10 \times \frac{87{,}5 - 46}{65} = 175{,}88$$

Paramètres de position

- ❸ Quantiles (percentiles) :

quantile d'ordre p : $x_p \qquad p \in (0,1)$



$$x_p \quad \text{tel que} \quad \begin{cases} N(x_p) \geq np \\ N^*(x_p) \geq n(1-p) \end{cases}$$

Quantiles

- Quartiles :

$$p = \frac{1}{4} \qquad p = \frac{2}{4} \qquad p = \frac{3}{4}$$

$$\downarrow \qquad \downarrow \qquad \downarrow$$

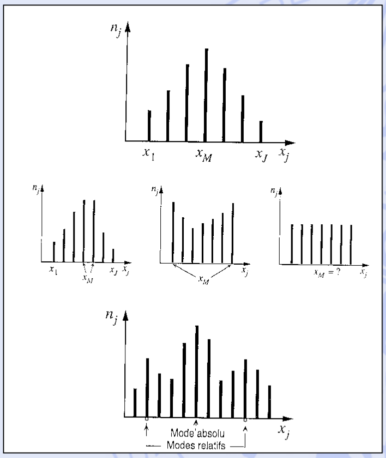
$$Q_1 = x_{1/4} \quad Q_2 = x_{1/2} \quad Q_3 = x_{3/4}$$

- Déciles :

$$p = \frac{1}{10} \quad p = \frac{2}{10} \quad \cdots \quad p = \frac{9}{10}$$

Paramètres de position

- ❹ Mode : Valeur la plus fréquente.



ULB UNIVERSITÉ LIBRE DE BRUXELLES

Paramètres de position

- Autres paramètres :
  - Moyenne tronquée d'ordre 1 : $\bar{x}_{t2} = \frac{1}{n-2}\sum_{i=2}^{n-1} x_{(i)}$
  - Moyenne harmonique : $H = n \Big/ \sum_{i=1}^{n} \frac{1}{x_i}$
  - Moyenne géométrique : $G = \sqrt[n]{\prod_{i=1}^{n} x_i}$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Moyenne géométrique

- pour des ratios, des taux, des pourcentages (valeurs positives uniquement).

Exemple : ventes d'une entreprise sur 5 ans :

Année	1	2	3	4	5
Ventes	100	110	132	151,8	170,02
Variation		+10%	+20%	+15%	+12%

→ Augmentation moyenne annuelle des ventes ?

$\bar{x} = 14,25\%$

$\frac{(10+20+15+12)}{4}$

Moyenne géométrique

$= 14,19\%$

$\sqrt[4]{1,10 \times 1,20 \times 1,15 \times 1,12}$

$100 \times (1,1425)^4 = 170,38$

$100 \times (1,1419)^4 = 170,02$

Moments

- Moment d'ordre r par rapport à c :

$$m_r(c) = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^r$$

- Moment par rapport à l'origine ($c = 0$) :

$$m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r$$

- Moment centré :

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r$$

Paramètres de dispersion

- Etendue :

$$E = x_{(n)} - x_{(1)}$$

- Ecart interquartile :

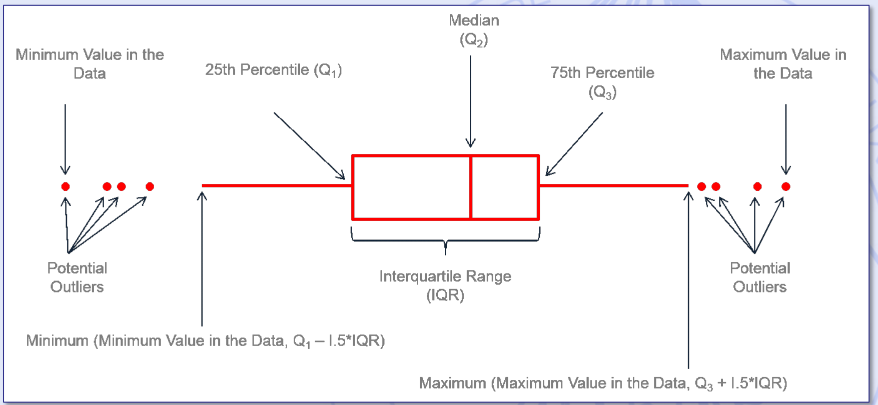
$$E_Q = x_{3/4} - x_{1/4}$$

- Ecart interdécile :

$$E_D = x_{9/10} - x_{1/10}$$

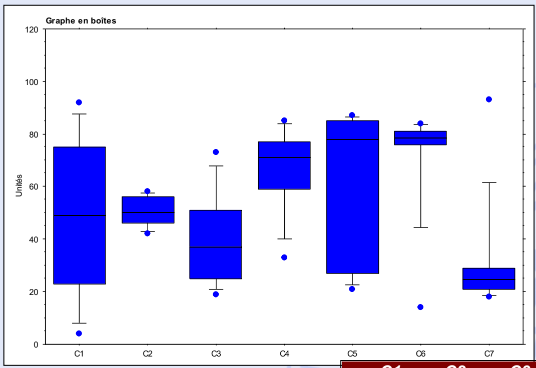
Paramètres de dispersion

- Box-Plot :



leansigmacorporation.com

Box-Plots



C1	C2	C3	C4	C5	C6	C7
4	42	19	33	21	14	18
12	44	23	47	24	75	19
23	46	25	59	27	76	21
35	47	27	67	29	77	23
46	49	31	69	77	78	24
52	51	43	73	79	79	25
67	54	48	75	83	80	27
75	56	51	77	85	81	29
83	57	63	83	86	83	30
92	58	73	85	87	84	93

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Paramètres de dispersion

- Ecart moyen absolu :

$$e_m = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \overline{x}\right|$$

- Ecart médian absolu :

$$e_m^* = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - x_{1/2}\right|$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Paramètres de dispersion

- Variance :

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

- Ecart-type :

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

- Coefficient de variation :

$$CV = \frac{s}{\overline{x}}$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Variance

- Agrégation de deux séries

$$[n_1 \quad \bar{x}_1, s_1^2] \quad [n_2 \quad \bar{x}_2, s_2^2]$$

$$\Downarrow$$

$$n = n_1 + n_2$$

$$s^2 = \underbrace{\frac{n_1 s_1^2 + n_2 s_2^2}{n}}_{\text{Variance dans les « groupes »}} + \underbrace{\frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{n}}_{\text{Variance entre les « groupes »}}$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Variance

- Calcul :

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

- Pour une distribution observée :

$$s^2 = \frac{1}{n}\sum_{j=1}^{J} n_j (x_i - \bar{x})^2 = \frac{1}{n}\sum_{j=1}^{J} n_j x_j^2 - \bar{x}^2$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Variance

- Exemple 3 :

$$
\begin{aligned}
s^2 &= \frac{1}{10}\left(2\times(0-1{,}2)^2+5\times(1-1{,}2)^2+2\times(2-1{,}2)^2+1\times(3-1{,}2)^2\right)\\
&= \frac{1}{10}\left(2\times 0^2+5\times 1^2+2\times 2^2+1\times 3^2\right)-1{,}2^2\\
&= 0{,}76
\end{aligned}
$$

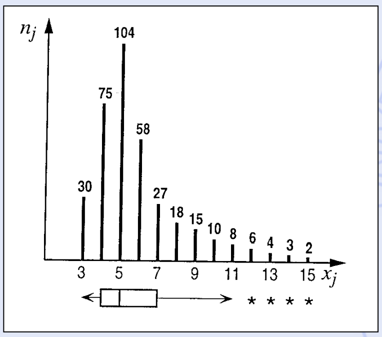
x_j	n_j	f_j
0	2	0.2
1	5	0.5
2	2	0.2
3	1	0.1

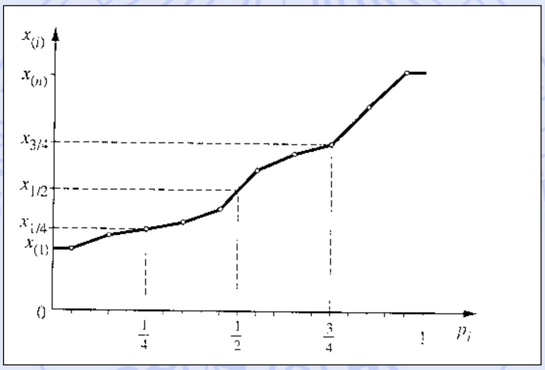
$$s=\sqrt{0{,}76}=0{,}87$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Paramètres de forme

- Box-plot
- Graphique des quantiles



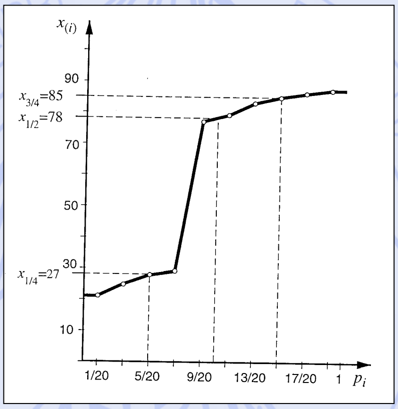


ULB UNIVERSITÉ LIBRE DE BRUXELLES

Graphique des quantiles

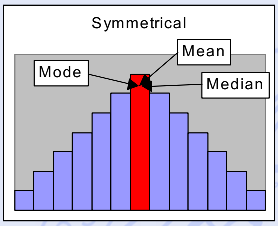
- Exemple C5 :

C1	C2	C3	C4	C5	C6	C7
4	42	19	33	21	14	18
12	44	23	47	24	75	19
23	46	25	59	27	76	21
35	47	27	67	29	77	23
46	49	31	69	77	78	24
52	51	43	73	79	79	25
67	54	48	75	83	80	27
75	56	51	77	85	81	29
83	57	63	83	86	83	30
92	58	73	85	87	84	93

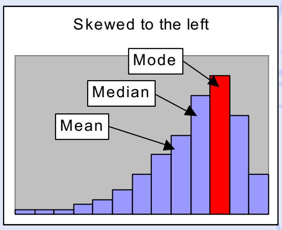


ULB UNIVERSITÉ LIBRE DE BRUXELLES

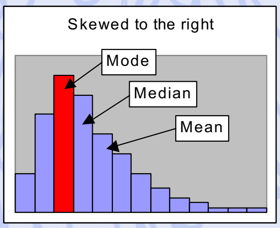
Asymétrie



Symétrie



Asymétrie à droite



Asymétrie à gauche

Coefficients d’asymétrie

- Moment centré d’ordre 3 :

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3$$

asymétrie à gauche	$m_3 > 0$
symétrie	$m_3 = 0$
asymétrie à droite	$m_3 < 0$

- Coefficient de Fisher (« skewness ») :

$$g_1 = m_3 / s^3$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Coefficients d’asymétrie

- Coefficient empirique de Pearson :

$$S_k = \frac{\overline{x} - x_M}{s}$$

- Coefficient empirique de Yule et Kendall :

$$Y_K = \frac{x_{1/4} + x_{3/4} - 2x_{1/2}}{x_{3/4} - x_{1/4}}$$

ULB UNIVERSITÉ LIBRE DE BRUXELLES

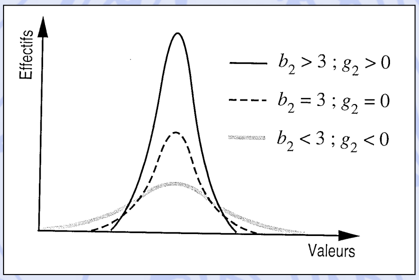
Coefficients d'aplatissement

- Coefficient de Pearson :

$$b_2 = \frac{m_4}{s^4} \quad \text{où} \quad m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

- Coefficient de Fisher (Kurtosis) :

$$g_2 = b_2 - 3$$

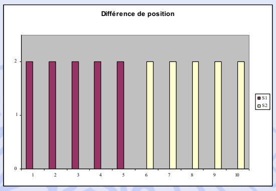


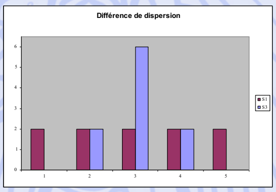
ULB UNIVERSITÉ LIBRE DE BRUXELLES

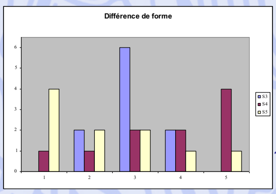
Exemples

i	*S1*	*S2*	*S3*	*S4*	*S5*
1	1	6	2	3	3
2	2	7	3	2	4
3	2	7	3	4	2
4	1	6	2	1	5
5	3	8	3	3	3
6	4	9	3	5	1
7	5	10	4	5	1
8	3	8	3	4	2
9	4	9	3	5	1
10	5	10	4	5	1

Moyenne	3,00	8,00	3,00	3,70	2,30
Médiane	3,00	8,00	3,00	4,00	2,00
Mode	~~1,00~~	~~6,00~~	3,00	5,00	1,00
Variance	2,00	2,00	0,40	1,81	1,81
Ecart-type	1,41	1,41	0,63	1,35	1,35
Skewness	0,00	0,00	0,00	-0,80	0,80
Kurtosis	-1,33	-1,33	0,08	-0,38	-0,38







Transformations de variables

- Transformation linéaire :
  - Changement d'origine et d'unité

$$x \to u = \frac{x - x_0}{d} \qquad x_0 \in R_0 \quad d \in R_0^+$$

  - Exemple :

$$t_{^\circ C} = \frac{t_{^\circ F} - 32}{1,8}$$

Transformations de variables

- Influence sur les paramètres de position :

$$x_i \to u_i = \frac{x_i - x_0}{d}$$

$$\bar{u} = \frac{\bar{x} - x_0}{d} \Leftrightarrow \bar{x} = x_0 + \bar{u}d$$

- Même transformation pour les autres paramètres de position :

$$u_{1/2} \quad u_M \quad u_p \quad \dots$$

Transformations de variables

- Influence sur les paramètres de dispersion :
  - Changement d'origine : sur CV seulement,
  - Changement d'unité : sur tous sauf CV. $(E \quad E_Q \quad E_D \quad s^2 \quad s)$
- Exemple :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s_u^2 = \frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})^2$$

$$\rightarrow s_u^2 = \frac{s_x^2}{d^2}$$

Transformations de variables

- Influence sur les paramètres de forme :
  - Pas d'influence sur $g_1 \quad S_k \quad Y_K \quad b_2 \quad g_2$
- Autres transformations :
  - Transformation logarithmique : $y = \log_{10} x$
  - Famille de transformations de Tukey :

$$\ldots, \frac{1}{x^2}, \frac{1}{x}, \frac{1}{\sqrt{x}}, \log x, \sqrt{x}, x, x^2, x^3, \ldots$$

  - Différence (données chronologiques) :

$$\nabla x_t = x_t - x_{t-1}$$