DELPHES * STAT

guide: THEORIQUE

auteurs: Jean-Paul Dufresne
Charles Foucault

No:CD6746 0108

Recherche effectuée au Collège de
Valleyfield grâce à une subvention
de la Direction générale de
l'enseignement collégial du
Ministère de l'Education dans
le cadre du programme de subvention,
à l'innovation pédagogique.

Dépôt légal deuxième trimestre 1985
Bibliothèque nationale du Québec
ISBN 2-550-08330-X

DELPHES * STAT

guide: THEORIQUE

auteurs: Jean-Paul Dufresne
Charles Foucault

No:CD6746 0108

Ce logiciel...«est l'image physique, le symbole palpable de l'Oracle obscur, parfois brutal, à d'autres moments ironique, ou alors se dérobant, refusant de répondre...Symbole ardent de cette étrange divinité à la fois véridique et fourbe, de ce dieu-prophète qui, sans jamais dire tout ce qu'il semble dire, pose aux humains torturés de doute l'éternelle question: quels sont les secrets de l'avenir?»

(G. TOUDOUZE, La Grèce au visage d'énigme, Paris, Berger)

graphisme: Service Audio-Visuel

REMERCIEMENTS

Nous tenons à remercier toutes les personnes qui, de près ou de loin, ont contribué à la création de ce logiciel. S'il est impossible de les nommer toutes, il faut cependant mentionner trois noms:

Soulignons d'abord l'appui considérable que nous a apporté M. Pierre Harrison tout au long des deux années du projet. Il nous a continuellement témoigné son intérêt, sa foi indéfectible dans notre entreprise. Les critiques pour lesquelles nous l'avons sollicité, en plus de celles qu'il nous a spontanément livrées, ont toujours été inspirées par un jugement solide et fiable. Nous lui sommes reconnaissants d'avoir conservé son sens de l'humour et la capacité de s'émerveiller devant nos emportements.

Il faut parler aussi du travail quasi vertueux de M. Gaétan Clément, professeur de littérature au Collège de Valleyfield qui a bien voulu corriger les textes des trois guides. Sans la qualité de son travail, la documentation n'aurait jamais atteint le standard souhaité. Les nombreuses heures que nous avons passées ensemble et qui ont parfois soumis sa maîtrise de la langue à rude épreuve nous ont permis de découvrir un confrère de la plus haute compétence, d'un grand raffinement dans ses analyses et d'une grande délicatesse dans ses interventions. Il nous a permis de vivre une expérience très enrichissante.

Nous voulons enfin signaler la contribution de Mme Joanne Brouard sans qui nos brouillons seraient demeurés brouillons. Dans les circonstances, nous pouvons presque parler de miracle, ou alors nous avions affaire à une fée. Sa capacité de produire un travail de très haute qualité dans un délai aussi court et son initiative dans la prise de nombreuses décisions quant à la gestion du texte nous ont permis de constater qu'elle a été non seulement une secrétaire exceptionnelle mais, plus encore, une partenaire essentielle.

A Chantal
Christian
Nicolas
et Anne-Claude

«Ne crains pas d'avancer lentement
crains seulement de t'arrêter».

Sagesse chinoise.

GUIDE THÉORIQUE .

CHAPITRE 0

INTRODUCTION

0.1 INNOVATION PÉDAGOGIQUE.

0.2 ÉDUCATION ET VIE.

0.3 INTUITION ET APPRENTISSAGE.

0.4 LA CALCULATRICE.

0.5 PLACE DE DUFORO DANS L'ENSEIGNEMENT.

0.6 LES OBJECTIFS.

0.7 LA STRUCTURE DU GUIDE THÉORIQUE.

0.I INNOVATION PÉDAGOGIQUE.

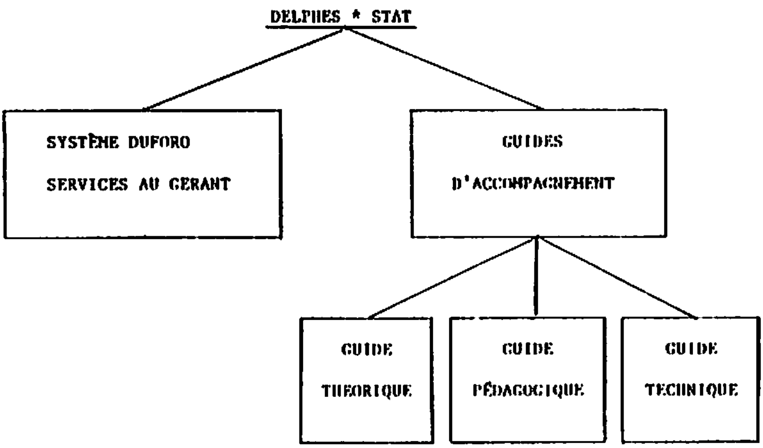
«Un chien qui se remue vaut mieux qu'un lion accroupi» PROVERBE ARABE.

La mise en chantier et la production d'un DIDACTICIEL visant à fournir un support à l'enseignement constituent une entreprise qui tient à la fois de la folie pure et de l'inconscience, et cela à plusieurs égards: d'abord, bien sûr, parce qu'elle exige une incursion dans un domaine nouveau, parsemé d'inconnus, sans beaucoup de références utiles qui puissent servir de guide; aussi, parce qu'elle exige une somme de travail considérable, voire incroyable, due en partie au fait que l'ordinateur est un «chronophage» chronique; et enfin, parce qu'elle nous force à mettre de l'avant (et à les justifier) des idées nouvelles, à secouer l'inertie d'un milieu (l'enseignement) que nous savons être conservateur de nature.

Nous aurions pu, en effet, choisir de rester dans la quiétude tranquille de nos classes, de continuer à enseigner, à donner honnêtement le meilleur de nous-mêmes, à fournir un enseignement de qualité à partir de méthodes traditionnelles, à critiquer les manuels qu'on utilise et qui ne nous satisfont jamais et à espérer finalement jouir d'une retraite aussi anticipée que méritée. Mais ce n'est pas notre choix; nous avons préféré progresser, innover, foncer. Nous avons voulu donner un nouveau souffle aux méthodes d'enseignement en les adaptant à un monde en pleine évolution.

Nous avons parcouru un long chemin depuis le tout début de notre réflexion, mais nous ne sommes toujours pas satisfaits: notre aptitude à trouver les meilleures solutions est forcément limitée devant l'infinité de nos ambitions; des problèmes surgissent qui nous contraignent à un certain piétinement et ralentissent le rythme souhaité. Mais notre progrès est quand même réel. Nous avons même la conviction d'être engagés sur une «autoroute» où, sans faire d'excès de vitesse, nous avons choisi la «voie réservée au dépassement».

Le texte qui suit est partie intégrante de notre recherche et il reste susceptible d'amélioration en tout temps. Il n'a pas la prétention de se substituer à un manuel de référence d'un cours de statistiques; ce n'est pas son but et il n'est pas complet en lui-même. Il sert plutôt de texte d'accompagnement et il doit être considéré comme l'une des parties du logiciel DELPHES * STAT, chacune étant indispensable.



Ce texte, à l'instar du professeur, se présente donc comme un guide, au vrai sens du mot. Il vise à faire comprendre la théorie qui est préalable à l'utilisation du logiciel.

Ce guide présente la théorie en privilégiant l'approche intuitive (lorsque cela est possible) et en utilisant une terminologie qui lui est propre. Contrairement à un manuel scolaire, il ne vise pas un large public mais essentiellement les futurs utilisateurs de Duforo. Il ne contient donc pas de sections que les lecteurs peuvent omettre de lire ni de notions qui peuvent être abordées «au choix»; il ne présente, à notre avis, que l'essentiel.

0.2 ÉDUCATION ET VIE .

«Le sage vient chercher de la lumière et le fou lui en donne.»
RABELAIS.

L'éducation vise d'abord le développement intégral de la personne. Commençant dès les premiers instants de la vie consciente, elle se poursuit pendant toute la durée de la vie de sorte qu'elle se fait partout et à chaque jour. Mais pour les besoins d'un apprentissage plus systématisé la société a cru bon d'institutionnaliser l'éducation par l'implantation d'écoles et la formation de professeurs. Ceux-ci sont des intervenants parmi d'autres qui ont à coeur d'assurer aux jeunes une préparation adéquate, non seulement au marché du travail, mais aussi et surtout à la vie car on ne développe pas ses facultés dans l'unique but de «gagner sa vie»; encore faut-il la vivre.

Pour sa part, la société doit faire tout ce qui est raisonnablement possible pour favoriser l'éducation. N'oublions pas que l'éducation met en relation des personnes en constante évolution et que, mis en situation d'apprentissage, chacun doit fournir sa juste part d'efforts. Aussi ne faut-il pas tout attendre des moyens mis en oeuvre par la société: la volonté, l'habileté et le talent des jeunes restent des facteurs déterminants dans la poursuite du développement personnel. De plus, pour progresser, l'étudiant doit accepter de travailler, de peiner, de souffrir même. (Le dictionnaire est le seul endroit où «succès» vient avant «travail»).

C'est pourquoi le professeur doit être un stimulateur et un catalyseur. Il fait le pont entre la connaissance et l'étudiant qui veut connaître. Il doit susciter en lui, le goût, la soif de savoir davantage et de mieux connaître le monde qui l'entoure. En ce sens, l'éducation doit puiser sa source dans l'expérience quotidienne. Car l'enracinement dans la réalité vécue suscite généralement l'intérêt et l'attention des étudiants.

Il est illusoire de penser que l'industrie modifierait ses procédés de fabrication à la seule fin de les ajuster à l'enseignement. Il est plus réaliste, au contraire, de fournir à l'étudiant des connaissances de bases et de favoriser, chez lui, les différentes formes d'apprentissages que requiert le monde du travail. L'étudiant ainsi formé saura bien, de lui-même, s'adapter aux différentes situations.

«Donne-moi un poisson et je mangerai un jour; montre-moi à pêcher et je mangerai toujours!

PROVERBE CHINOIS.

0.3 INTUITION ET APPRENTISSAGE.

Des anthropologues ont étudié des systèmes de nombres de centaines de tribus aborigènes. Les YACOS de l'Amazonie brésilienne arrêtent de compter à trois puisque le mot qui signifie «trois», dans leur dialecte, est «POETTARRARORINCOAROAC» (Albert Sukoff).

Cet exemple montre à souhait que l'on retient moins bien dans la mesure où l'on complexifie davantage. Aussi faut-il arrêter de complexifier la connaissance, comme si l'on voulait se rendre indispensable à sa transmission. Il est inutile d'utiliser des termes recherchés, de longues explications, d'inextricables détours pour faire comprendre des choses déjà simples. Du même coup, sachons qu'une notion bien comprise n'a pas besoin du recours à la mémoire. Au besoin, lorsqu'une notion a été oubliée, il est possible d'y revenir en remontant à la source du raisonnement. Une notion sans lien est comme un chien sans laisse: difficile à retenir!

En définitive ce n'est pas la forme des mots qui importe mais leur signification.

Le texte qui suit adopte une démarche intuitive; mais l'intuition a ses limites (il suffit de consulter les noms utilisés par les paléontologues pour s'en convaincre).

Cependant, l'intuition nous ayant mis sur la piste, la compréhension n'en sera que plus facile. Mais l'intuition ne suffit pas: il reste que chaque notion doit être expliquée en toute rigueur pour une compréhension complète.

0.4 LA CALCULATRICE.

«Il y a plus de choses dans un jardin que ce qu'on y a semé.» PROVERBE YOUGOSLAVE.

Depuis quelques années, l'utilisation de la calculatrice s'est largement répandue dans les cours de mathématiques. Et c'est heureux! Nous avons perdu assez de temps par le passé dans des calculs laborieux qui ne nous enseignaient en définitive que la patience, pour que nous adaptions et intégrions ce merveilleux outil à notre pédagogie.

Il faut cependant se méfier de la pseudo-connaissance engendrée par une mauvaise utilisation. Trop d'étudiants du collégial en ignorent les rudiments. IL y a donc lieu de consacrer une période de temps à leur apprendre les caractéristiques et les possibilités de leur calculatrice.

Et même lorsque cet ultime objectif aura été atteint, il faudra éviter de se prendre au piège de la mesure à tout prix. On est porté à croire en effet qu'il faut tout mesurer dans notre société moderne. Pourtant les arbres poussent sans qu'on les mesure. Les oiseaux ne grossissent pas démesurément même s'ils n'ont pas de balance électronique pour mesurer leur nourriture. On entretient souvent, dans notre monde «électronifié», la fausse impression qu'on ne possède bien que ce que l'on mesure!... Replaçons plutôt les choses dans leur juste perspective. Ne confondons pas les causes et les effets. Ce n'est pas parce que nous savons évaluer les chances qu'a un événement de se produire qu'il nous sera automatiquement favorable. Mais notre décision, éclairée par cette évaluation, a des chances, elle, d'être meilleure.

0.5 PLACE DE DUFORO DANS L'ENSEIGNEMENT .

DUFORO n'entre pas dans la catégorie des didacticiels destinés à faire de l'enseignement assisté de l'ordinateur (EAO). DUFORO ne fait pas d'enseignement; il soutient l'enseignement; il intervient à la suite d'un cours théorique. En ce sens, DUFORO ne vise pas à remplacer le professeur, mais bien à l'aider à mieux accomplir sa tâche.

DUFORO n'est pas une fin en soi, mais une étape dans le long processus d'éducation. Il n'est pas l'aboutissement du processus d'acquisition des connaissances, mais plutôt un complément, un outil, une intervention souhaitée et souhaitable, l'un des maillons de la chaîne éducative.

DUFORO n'est pas non plus l'un de ces didacticiels qui reproduisent d'une manière plus compliquée ce que le crayon avait déjà fait avec simplicité sur du papier. Car, il faut bien l'avouer, le marché québécois est encore inondé de ces «logiciels-gadgets» où l'ordinateur n'est qu'un prétexte. Ces derniers sont aussi futiles qu'inadaptés au monde de l'Education.

DUFORO est d'abord et avant tout une réponse adéquate à l'utilisation intelligente et efficace de l'ordinateur en éducation. C'est là son objectif principal et nous croyons l'avoir atteint.

0.6 LES OBJECTIFS DU GUIDE THEORIQUE .

Nous avons regroupé sous le titre «GUIDE THÉORIQUE» toute la matière qui doit être abordée en classe avant que l'étudiant n'entre en relation avec DUFORO. Le présent texte vise essentiellement trois catégories de personnes.

A) LES PROFESSEURS-GÉRANTS

Ce sont ceux qui utilisent DUFORO avec un groupe d'étudiants et qui dispenseront le cours théorique. Ils doivent donc prendre connaissance de la matière abordée par le logiciel, de l'approche utilisée par ceux qui l'ont conçu et finalement des termes utilisés dans le logiciel.

B) LES USAGERS

Ce sont ceux qui entrent en relation avec le système DUFORO et qui doivent d'abord bénéficier d'un cours magistral. Le document qui suit peut servir de texte de base et, à ce titre, pourrait leur être distribué.

C) LES PROFESSEURS DE STATISTIQUES

Nous pensons (à tort ou à raison) que notre approche de la Statistique Descriptive pourrait susciter l'intérêt de tout professeur de statistiques, qu'il utilise ou non DUFORO.

A ce titre, il peut servir de référence à tout professeur qui:

- aborde les statistiques descriptives;
- cherche un document couvrant les statistiques descriptives sans truquage;
- veut le compléter en y ajoutant des séries de problèmes de son choix.

0.7 LA STRUCTURE DU GUIDE THÉORIQUE.

On trouvera dans les chapitres qui suivent:

Chapitre 1 - Le regroupement des observations.

Chapitre 2 - Les représentations graphiques.

Chapitre 3 - Les paramètres descriptifs.

Enfin, il faudra se référer au guide pédagogique pour savoir comment ces éléments théoriques ont été incorporés dans DUFORO.

CHAPITRE I

LE REGROUPEMENT DES OBSERVATIONS .

1.0 INTRODUCTION .

1.1 PRÉSENTATION DES OBSERVATIONS .

- Qualitatives .
- Quantitatives .
- Tableau des observations brutes .

1.2 LES PHÉNOMÈNES ET LEUR MESURE .

- Cas discret .
- Cas continu .
- Les instruments de mesure :
  - . par comparaison .
  - . par évaluation mécanique .
  - . par évaluation électronique .
  - . règles à suivre .

1.3 LE DÉNOMBREMENT .

- Quelques exemples avec dénouement plus ou moins heureux.

1.4 LA TABLE DE DISTRIBUTION DE FRÉQUENCES .

a) Les limites .

b) Les bornes .

c) Le cas hybride .

d) Les classes concentrées .

e) Les classes ouvertes .

1.5 LES DÉFINITIONS .

- Limites .

- Bornes .

- Longueur des classes .

- Centre des classes .

- Etendue des données .

1.6 PROCÉDURE DE FABRICATION D'UNE DISTRIBUTION DE FRÉQUENCES .

- Phénomènes qualitatifs .

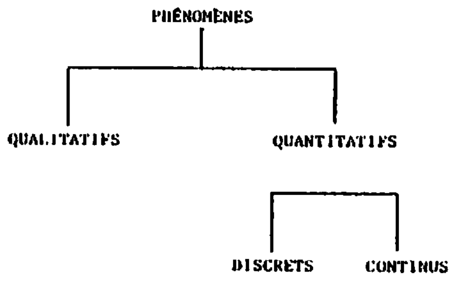
- Phénomènes quantitatifs .

CHAPITRE 1

1.0 INTRODUCTION.

Dans ce chapitre nous apprendrons à présenter et à utiliser des observations recueillies dans le but de mieux connaître un phénomène. En effet sans une bonne connaissance de leur utilisation, l'accumulation d'informations - si abondantes soient-elles - ne saurait être utile pour comprendre la réalité qui nous intéresse.

Nous ferons d'abord le point sur les phénomènes que nous rencontrons en STATISTIQUES.



Nous apprendrons ensuite à faire l'inventaire des observations (dénombrement) et à les résumer en les regroupant sous forme de table de fréquences. Finalement, après avoir défini les termes utilisés, nous étudierons la procédure à suivre dans la fabrication d'une distribution de fréquences.

1.1 PRÉSENTATION DES OBSERVATIONS.

«La crapaude dit à son crapaud:

- Regarde, notre fille ressemble à une perle enfilée sur un fil de soie.
- Certes, répondit le crapaud, est-ce que nous pouvions mettre au monde moins que cela?»

Proverbe arabe .

Les statistiques ont principalement pour tâche de manipuler des observations qui résultent, heureusement, de mesures plus objectives que celles qui concluent à la beauté de cette jeune crapaude. Ainsi l'étude d'un phénomène exige la cueillette d'informations (interroger des individus, consulter des documents, mesurer des objets, observer des événements). Une telle recherche conduit inévitablement à une quantité plus ou moins importante d'observations.

Ces observations peuvent être de deux ordres:

Ordre qualitatif.

Nous dirons qu'il s'agit d'ordre qualitatif si l'observation correspond à une caractéristique qui relève de l'identité même de l'événement. Par exemple, la couleur d'une automobile, la marque (de commerce) d'un appareil électrique, la langue maternelle d'un individu ou la couleur des yeux de son chat.

__Ordre quantitatif__ .

Nous dirons qu'il s'agit d'ordre quantitatif si l'observation est identifiée par un nombre résultant d'une mesure, d'un calcul ou d'un comptage. Par exemple:

- l'âge des étudiants dans un groupe.
- le nombre de passagers d'un autobus scolaire.
- le revenu annuel des Québécois.

REMARQUE:

Il existe des situations «pseudo-quantitatives». Par exemple, si on jette un dé un certain nombre de fois et qu'on note le résultat:

> Il en résulte des observations qui vont de 1 à 6. Ces nombres ne proviennent d'aucune mesure, ni calcul, ni comptage. On a plutôt choisi d'identifier chaque face d'un dé à l'aide d'un chiffre et de le représenter par un ensemble de points. On est donc en présence d'un phénomène d'ordre qualitatif, mais représenté par un nombre. A la place des chiffres, on aurait pu identifier les faces du dé par les lettres A, B, C, D, E, et F ou par les couleurs Blanc, Rouge, Bleu, Noir, Vert et Jaune et continuer à utiliser le dé de la même façon qu'auparavant. Ceci confirme la nature qualitative du phénomène.

Une fois les informations recueillies, il faut les communiquer, les véhiculer. Les observations sont généralement présentées sous forme de tableau (rectangle composé de rangées et de colonnes), selon l'ordre chronologique de leur apparition. Elles constituent ce qu'on appelle le tableau des observations brutes.

Voyons quelques exemples:

<u>Exemple 1</u>:

Un agent de voyage demande à chaque membre de son personnel d'indiquer le nombre d'automobiles dont dispose (en propriété) sa famille immédiate. Voici les réponses obtenues:

1	0	2	1	1
3	0	1	1	1

Ces observations (données) sont <u>quantitatives</u>.

Exemple 2:

Une enquête auprès d'étudiants cherche à déterminer s'ils sont fumeurs (O) ou non-fumeurs (N). On obtient les résultats suivants:

O	N	N	N	N	O	N	N	O	N
N	N	N	O	O	N	N	N	N	O
O	O	N	N	N	N	O	N	N	O

Ces observations sont qualitatives.

Exemple 3:

Un professeur, après avoir soumis ses étudiants à un examen de statistiques, obtient les résultats suivants:

50	38	50	60	22	78
60	70	75	60	70	50
79	15	60	51	40	60
75	78	89	31	50	99
50	38	50	60	22	78
40	60	50	98	88	70

Ces observations sont quantitatives et présentées sous la forme d'un tableau de 6 rangées et de 6 colonnes.

<u>Exemple 4</u>:

Voici les vitesses en (km/h) détectées par un radar, installé dans un secteur scolaire entre 10 h et 11 h, un lundi matin:

36	41	23	15	18	49	65
44	39	30	31	42	17	12
40	62	79	40	53	13	72
26	41	16	23	29	36	33
30	41	40	19	63	48	39
63	14	13	18	29	28	33
41	43	16	61	19	21	29
56	32	30	31	36		

Ces observations sont <u>quantitatives</u>.

Elles sont présentées sous la forme d'un tableau de 8 rangées et de 7 colonnes (dont les deux dernières cases sont vides).

Il est évident que, dans ces exemples, quelqu'un a dû recueillir les observations et les écrire sous forme de tableau. D'ailleurs, l'observateur devra souvent intervenir, non seulement pour noter, mais aussi pour mesurer le phénomène étudié.

1.2 LES PHÉNOMÈNES ET LEUR MESURE

«En y mettant toutes ses forces,

même la souris pourrait dévorer le chat.»

Proverbe arabe

Dans notre vie quotidienne on mesure deux sortes de phénomènes quantitatifs:

Phénomène discret:

C'est celui qui permet à la mesure (la variable) de ne prendre que certaines valeurs numériques d'un intervalle donné. Par exemple: le nombre de passagers dans un autobus, le nombre de députés dans un gouvernement, le nombre d'enfants dans une famille, le nombre de logements d'un immeuble, les pointures de souliers d'un magasin à rayons, le salaire hebdomadaire des employés d'une usine ou le nombre de filles dans un groupe.

<u>Phénomène continu:</u>

C'est celui qui permet à la mesure de prendre, en principe, n'importe quelle valeur numérique d'un intervalle. Par exemple: la distance parcourue par un javelot lancé par un athlète, le temps d'arrêt d'une automobile qui file à 60 km/h, la distance séparant les villes dans une province, le poids des personnes inscrites à un programme de conditionnement physique ou l'angle mesuré entre les directions respectives de deux véhicules.

Lorsqu'un observateur mesure un phénomène discret, il lui suffit de compter le nombre d'observations et de les enregistrer. Mais lorsqu'il doit mesurer un phénomène continu, il doit alors utiliser un instrument de mesure et faire une lecture.

Compte tenu de l'instrument utilisé et de son degré de précision, la mesure que nous prenons d'un phénomène continu est elle-même discrète mais cela n'empêche pas le phénomène (et la variable qui le représente) d'être continu.

Quant aux appareils qui servent à mesurer ces phénomènes continus, on peut les classer en deux catégories selon leur façon de transmettre les résultats.

1- « APPAREILS PROCÉDANT PAR COMPARAISON » .

Ces instruments sont «gradués» et demandent une appréciation visuelle. Par exemple:

- Un mètre qui sert à mesurer une distance. On doit «comparer» le mètre et la distance de façon à lui faire correspondre la plus proche des subdivisions inscrite sur l'outil de mesure, le mètre.
- Une balance à plateau destinée à peser un objet. On doit comparer le poids de l'objet et des unités de poids jusqu'à ce qu'on ait obtenu un équilibre satisfaisant.
- Une «tasse à mesurer», utilisée dans la préparation d'une recette de cuisine.
- Une jauge à pression avec cadran et aiguille destinée à gonfler un pneu à une pression déterminée.

En général, les instruments de ce type peuvent conduire à des erreurs de lecture et donc de mesure. Leur «fiabilité» est limitée. On peut considérer cependant que l'erreur commise n'est pas supérieure à la moitié de la plus petite subdivision inscrite sur l'instrument. Ainsi on peut prévoir que:

Mesure réelle	Précision de l'instrument	Lecture
0,43021	0,1	0,4
7,174953	0,001	7,175
141,621462	1	142

Et à l'inverse:

Lecture	Précision de l'instrument	Mesure réelle
14,7	0,1	[14,65 ; 14,75]
1,982	0,001	[1,98515 ; 1,9825]
187,36	0,01	[187,355 ; 187,365]

REMARQUES:

1) Dans la deuxième situation, déduire la précision est toujours possible car la lecture est connue. Une mesure de 7,926184 m avec un mètre qui n'indique que les millimètres (0,001 m) serait pour le moins farfelue (à moins que l'on soit doté d'un oeil bionique!). A l'opposé, si l'instrument utilisé permet de connaître trois chiffres après le point, il faudra garder les zéros lorsque l'on aura 14,000.

2) La mesure fournie par l'instrument ne correspond pas automatiquement à la mesure réelle. C'est pourquoi il faut la coincer dans un intervalle. Le phénomène étant continu, sa mesure exacte ne nous est connue qu'avec la précision de l'instrument. Il est certain que l'erreur (écart entre la réalité et la mesure prise) causée par l'instrument ne dépasse jamais la moitié de la plus petite subdivision présente sur la graduation de l'instrument. On dit qu'elle n'est pas supérieure à une demi-unité de précision.

Observation	Erreur maximale	Réalité
14,1	0,05	14,1 $\pm$ 0,05
319,47	0,005	319,47 $\pm$ 0,005

2- « APPAREILS PROCÉDANT PAR ÉVALUATION »

Ces instruments disposent d'un mécanisme interne qui leur permet d'évaluer le phénomène à mesurer et de faire part du résultat par affichage. Il y a deux catégories d'affichage:

A) Affichage mécanique:

Ce sont les appareils dont l'affichage dépend d'un déplacement mécanique. (Par exemple: dans une automobile l'indicateur de kilomètres, situé sous le compteur de vitesse, dépend d'une force physique qui fait tourner la roulette indiquant les chiffres de 0 à 9). Ceux-ci affichent la progression au fur et à mesure que se produit le changement. Leur précision dépend du fabricant mais un tel appareil, s'il est de qualité, ne présente jamais une erreur d'affichage supérieure à une demi-unité de précision.

B) <u>Affichage électronique</u>:

Ce sont les appareils dont l'affichage dépend d'une procédure électronique (Exemple: montre à affichage numérique, balance électronique...).

Ces appareils utilisent une procédure interne qui leur permet «d'apprécier» la dernière décimale affichée, à l'aide d'une décimale supplémentaire qu'ils n'affichent pas mais dont ils tiennent compte pour modifier, à la hausse ou à la baisse, la mesure à afficher. Cette modification relève de la règle suivante:

0 1 2 3 4 ⟶ à la baisse

5 6 7 8 9 ⟶ à la hausse

Par exemple:

Mesure interne	Précision	Affichage
1,794	0,01	1,79
4,37	0,1	4,4
141,0135	0,001	141,014

Dans ce cas l'erreur de mesure est aussi maximisée par la moitié de l'unité de précision affichée.

RÈGLE GÉNÉRALE:

Tout instrument de mesure, s'il est utilisé correctement, nous assure d'une erreur (de mesure) non supérieure ($\leqslant$) à la moitié de l'unité de précision affichée.

COROLLAIRE:

Lorsque des nombres résultant d'une mesure sont fournis, il faut considérer que le dernier chiffre (de ce nombre) est exact à $\pm$ 0,5.

Par exemple, si le poids des colis déposés à la réception, un lundi matin, pour fin d'expédition par la poste, est de:

1,47 kg   0,30 kg   7.76 kg

On peut (et on doit) considérer que ces nombres ont une précision de 0,005.

Dans l'utilisation de ces observations pour fins de calculs, il ne faut pas croire qu'il est possible, par un quelconque tour de magie, d'augmenter cette précision. La moyenne de ces nombres peut par exemple se calculer ainsi:

$$\frac{1,47 + 0,30 + 7,76}{3} = \frac{9,53}{3} = 3,176666667$$

Mais ceci ne signifie aucunement que le poids moyen de ces colis est connu avec une précision de 0,000000001.

Le résultat des calculs ne peut jamais dépasser en précision celle de l'instrument de mesure dont on s'est servi. L'avènement de la calculatrice dans nos habitudes peut nous avoir transmis l'illusion d'une précision qui serait plus grande qu'elle ne l'est en réalité. Mais il faut savoir interpréter les résultats avec réalisme. Le statisticien utilisera le résultat de ses calculs tout en gardant en mémoire le contexte du problème qui lui a été soumis. A ce titre, les nombres ne sont qu'un instrument dont il se sert avec jugement. C'est là une situation bien différente de celle de l'étudiant qui est en situation d'apprentissage et pour qui le résultat des calculs est une fin en soi.

Etablissons donc les règles suivantes:

RÈGLE 1

Les nombres fournis comme observations sont précis à la moitié de la dernière décimale (Par exemple: 1,76 est précis à 0,005).

RÈGLE 2

En cours de calcul, on utilise toute la précision permise par la calculatrice. (Par exemple, pour ajouter 1/3 on additionne « 1 divisé par 3 »).

RÈGLE 3

Les calculs étant terminés, on transcrit, comme réponse finale, deux décimales de plus que celles qui nous ont été fournies dans la donnée du problème.

L'établissement de cette règle s'inspire des éléments suivants:

1- Il faut éviter la transcription inutile de toutes les décimales affichées par la calculatrice.

2- Il faut conserver une séquence de chiffres qui permet de comparer des résultats entre eux.

3- Les observations fournies dans la donnée d'un problème sont toujours mesurées avec une décimale supplémentaire: le besoin de précision exige qu'on apprécie le dernier chiffre.

4- Lorsque nous effectuons des calculs, nous manipulons des nombres de façon à en extraire des informations: nous les transformons! Nous sommes en quelque sorte des producteurs dont la matière première est constituée de nombres.

Une acierie extrait son métal du minerai, un moulin transforme le bois en pâtes et papiers: toute transformation exige des modifications au produit de base.

En ce sens, les STATISTIQUES transforment les nombres. Et à ce titre, il serait décevant d'effectuer des calculs sans réussir à «extraire» des nombres une précision supplémentaire.

1.3 LE DÉNOMBREMENT

Lors de la description d'un phénomène étudié, il est relativement facile d'énumérer toutes les observations lorsqu'elles sont en nombre restreint. Tel n'est pas le cas lorsque la quantité d'observations augmente. Il devient alors impérieux d'en simplifier la présentation.

Dans le cas qui nous préoccupe, l'ordre chronologique selon lequel les observations ont été obtenues n'ayant plus d'importance (numérique) après qu'elles ont été enregistrées, on peut se permettre de les disposer autrement. Pour ce faire, il suffit d'identifier chacun des différents résultats et de compter le nombre de fois qu'il apparaît dans la séquence des observations. On effectue de cette façon un inventaire, qu'il convient d'appeler le DÉNOMBREMENT. «Dénombrer» consiste donc à assigner un nombre (de fois) aux différentes observations apparaissant dans le tableau des données brutes.

Exemple 5:

Reprenons les observations de l'Exemple 1 quant au nombre d'automobiles par famille:

1	0	2	1	1
3	0	1	1	1

Observations	Apparitions	Nombre total d'apparitions
0	//	2
1	~~////~~ /	6
2	/	1
3	/	1

Nous présentons la situation à l'aide de la <u>table</u> (deux colonnes) <u>de dénombrement</u> qui suit:

Nombre d'automobiles	Nombre d'individus
0	2
1	6
2	1
3	1

Cette table représente parfaitement la réalité. Il serait inopportun, ici, de vouloir simplifier davantage la présentation. Il y aurait perte d'information sans ajouter à la clarté.

On peut noter qu'à partir de la table de dénombrement on peut reconstituer dans son entier le tableau des observations, ce qui signifie que nous n'avons perdu aucune information.

Exemple 6:

Reprenons les observations de l'Exemple 2 qui traite des fumeurs:

O	N	N	N	N	O	N	N	O	N
N	N	N	O	O	N	N	N	N	O
O	O	N	N	N	N	O	N	N	O

Observations	Apparitions	Nombre d'apparitions
O	~~////~~ ~~////~~	10
N	~~////~~ ~~////~~ ~~////~~ ~~////~~	20

Ce dénombrement nous permet de présenter le phénomène étudié de la façon suivante:

Etat de fumeur	Nombre d'étudiants
O	10
N	20

Encore une fois le dénombrement a eu pour effet de clarifier suffisamment la situation.

<u>Exemple 7</u>: Considérons les résultats de l'exemple 3

50	38	50	60	22	78
60	70	75	60	70	50
79	15	60	51	40	60
75	78	89	31	50	99
40	60	50	98	88	70

on obtient le dénombrement suivant:

Observations	Apparitions	Nombre d'apparitions
15	/	1
22	/	1
31	/	1
38	/	1
40	//	2
50	~~////~~	5
60	~~////~~ /	6
70	///	3
75	//	2
78	//	2
79	/	1
81	/	1
88	/	1
89	/	1
98	/	1
99	/	1

Cela nous amène à simplifier la situation par la table suivante:

Résultats	Nombre d'étudiants
15	1
22	1
31	1
38	1
40	2
50	5
60	6
70	3
75	2
78	2
79	1
81	1
88	1
89	1
98	1
99	1

Même en admettant que cette table de dénombrement a eu pour effet de simplifier (légèrement?) la présentation des observations, il serait exagéré d'affirmer que la situation est «claire comme de l'eau de roche».

Nous trouverions sûrement des alliés qui, comme nous, souhaitent faire un pas de plus vers la «simplification».

Exemple 8:

En reprenant les données de l'exemple 4, nous pouvons obtenir le dénombrement (plus laborieux) suivant:

36	41	23	15	18	49	65
44	39	30	31	42	17	12
40	62	79	40	53	13	72
26	41	16	23	29	36	33
30	41	40	19	63	48	19
63	14	13	18	29	28	33
41	43	16	61	19	21	29
56	32	30	31	36		

Observations	Apparitions	Nombre d'apparitions
12	/	1
13	//	2
14	/	1
15	/	1
16	//	2
17	/	1
18	//	2
19	//	2
21	/	1
23	//	2
26	/	1
28	/	1
29	///	3
30	///	3
31	//	2
32	/	1
33	//	2
36	///	3
39	//	2
40	///	3
41	////	4
42	/	1
43	/	1
44	/	1
48	/	1
49	/	1
53	/	1
56	/	1
61	/	1
62	/	1
63	//	2
65	/	1
72	/	1
79	/	1
		54

Cela nous permet de présenter la situation par:

Vitesse en kmh	Nombre d'automobiles
12	1
13	2
14	1
15	1
16	2
17	1
18	2
19	2
21	1
23	2
26	1
28	1
29	3
30	3
31	2
32	1
33	2
36	3
39	2
40	3
41	4
42	1
43	1
44	1
48	1
49	1
53	1
56	1
61	1
62	1
63	2
65	1
72	1
79	1
	54

Ici, il nous apparaît évident que le dénombrement ne produit pas toujours le même effet. A tout le moins, nous pouvons affirmer avec certitude que la simplicité n'est pas toujours présente. Si nous n'avions comme outil que le dénombrement, nous aurions probablement à l'esprit ce proverbe arabe:

«Quand le loup enseigne aux oies leurs prières, il les croque pour ses honoraires».

1.4 DISTRIBUTION DE FRÉQUENCES

«Qui tient la poêle par la queue,

il la tourne comme il veut».

Proverbe français

a) LES LIMITES

Afin de simplifier davantage et d'apporter plus de clarté encore, de nouveaux outils se présentent.

Exemple 9

Reprenons la table de dénombrement de l'exemple 7:

Résultats	Nombre d'étudiants
15	1
22	1
31	1
38	1
40	2
50	5
60	6
70	3
75	2
78	2
79	1
81	1
88	1
89	1
98	1
99	1
	30

Si on examine ce qui gêne dans cette situation, c'est principalement la grande variété des observations. Précisément 10 d'entre elles ne sont pas répétées (sans compter celles qui le sont). En fait, les observations et le nombre de fois qu'elles apparaissent totalisent 32 nombres à retenir.

Essayons donc d'imaginer un «regroupement» possible. A cette fin, nous pourrions former des «tiroirs» qui nous faciliteraient la tâche.

Rappelons-nous qu'il s'agit de résultats à un examen. Ici le contexte peut nous inciter à choisir une forme de regroupement bien particulière. Il s'agit de regrouper les observations par «dizaines». On obtient ainsi:

Résultats	Nombre d'étudiants
10 à 19	1
20 à 29	1
30 à 39	2
40 à 49	2
50 à 59	5
60 à 69	6
70 à 79	8
80 à 89	3
90 à 99	2
	30

Un tel regroupement simplifie la présentation puisque, pour la retenir, il suffit de savoir que:

1) Chaque «tiroir» est une dizaine;
2) Que le plus petit commence à 10 (inutile de retenir que le plus grand est 99 puisqu'il s'agit d'un examen de statistiques);
3) Les nombres associés sont respectivement: 1, 1, 2, 2, 5, 6, 8, 3, 2.

Cela fait donc 11 éléments à retenir plutôt que 32, ce qui est assurément plus simple. Il est cependant impossible, en se basant uniquement sur cette nouvelle table, de reconstruire le tableau des données brutes car il y a eu perte d'information. La simplification ne se fait qu'à ce prix.

Cette forme de regroupement des données s'appelle une table (deux colonnes) de distribution de fréquences. Chaque «tiroir» défini dans la colonne de gauche s'appelle une «classe» et à chaque classe est associée une fréquence (le nombre d'éléments du tableau qui entrent dans la classe).

Ainsi, la 5ième classe est celle qui «peut» accueillir tous les résultats allant de 50 à 59 (inclusivement) et qui de fait en contient 5 (50, 50, 50, 50, 50). On dit alors que la 5ième classe a une «fréquence» de 5. La 7ième classe va de 70 à 79 et a une fréquence de 8 (elle contient les éléments 70, 70, 70, 75, 75, 78, 78 et 79).

Chaque classe est définie à partir de deux nombres. Prenons dans notre exemple la classe 30 à 39. On sait que, dans le tableau, les données brutes étaient fournies à l'entier près. (On doit supposer qu'elles ont été mesurées à moins une demi-unité (0,5) de précision). Cette classe représente les résultats possibles 30, 31, 32, 33, 34, 35, 36, 37, 38 et 39. A chaque fois que l'un deux se présentera, la fréquence de la classe augmentera de 1. On a donc choisi de définir la classe à l'aide du plus petit et du plus grand nombre pouvant entrer dans cette classe. Ce sont les LIMITES de la classe.

30 est appelé la limite inférieure de la 3e classe

39 est appelé la limite supérieure de la 3e classe

(N.B.: Ici les limites ont le même nombre de décimales que les données).

La définition qu'on a donnée de la classe se suffit à elle-même; la classe existe indépendamment du fait qu'il y ait ou pas un nombre de 30 dans la liste des observations. Ainsi, les limites sont des nombres théoriques mais dont la précision est calquée sur celle des données brutes.

b) LES BORNES

Alignons les limites sur un axe et identifions chaque classe:

10 19 20 29 30 39 40 49 50 59 60 69 70 79 80 89 90 99

——[ ]-[ ]-[ ]-[ ]-[ ]-[ ]-[ ]-[ ]-[ ]——

1 2 3 4 5 6 7 8 9

Chaque classe semble être un morceau détachable, séparé, sans lien avec les autres. Pourtant, si on devait enlever un morceau, il ne pourrait reprendre une autre place que la sienne, à son retour. Ainsi on ne peut interchanger deux classes entre elles. Cela dénote donc une certaine relation d'ordre. Et s'il existe un ordre, il s'agit de toute évidence de celui qui existe déjà dans les Réels. Cette relation «d'ordre» se traduit de la façon suivante:

On crée un état de dépendance entre les classes en les «rattachant» les unes aux autres à l'aide des points-milieux de l'espace qui les sépare:

10 19 20 29 30 39 40 49 50 59 60 69 70 79 80 89 90 99

[——]·[——]·[——]·[——]·[——]·[——]·[——]·[——]·[——]

1 2 3 4 5 6 7 8 9

19,5 29,5 39,5 49,5 59,5 69,5 79,5 89,5

Ainsi, plutôt que de parler de la classe 30 à 39, on parle de la classe 29,5 à 39,5. (Il s'agit toujours de la 31ème classe, elle contient toujours les résultats 30, 31, 32, 33, 34, 35, 36, 37, 38 et 39 et elle possède la même fréquence).

Deux problèmes se posent:

Premier problème

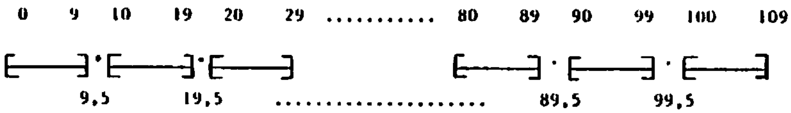
Un même élément appartient à deux classes. (Exemple: 29,5 appartient à la 21ème et à la 3 1ème classe).

Solution: C'est un faux problème puisque ce nombre ne peut apparaître dans le tableau des données brutes qui sont fournies en nombres entiers.

Deuxième problème

Les classes des extrémités (la première et la deuxième) ne possèdent qu'un élément servant à les définir. (19,5 pour la première et 89,5 pour la dernière).

Solution: Il suffit de créer deux classes fictives et d'ajouter les deux éléments manquants.

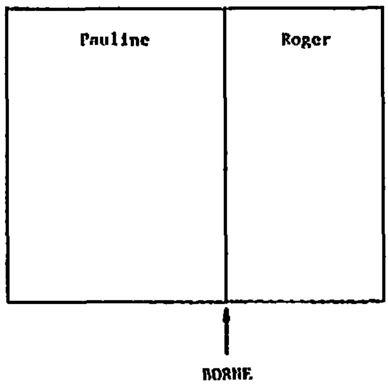


Ainsi on peut redéfinir la distribution de fréquences par:

Résultats	Fréquences
9,5 à 19,5	1
19,5 à 29,5	1
29,5 à 39,5	2
39,5 à 49,5	2
49,5 à 59,5	5
59,5 à 69,5	6
69,5 à 79,5	8
79,5 à 89,5	3
89,5 à 99,5	2

Les nouveaux éléments servant à définir les classes s'appellent des BORNES. Le sens qu'on assigne au mot BORNE est celui qu'il possède dans notre vie quotidienne. Il suffit de penser aux bornes qui séparent deux terrains adjacents.

Il existe, entre les deux terrains, une ligne appelée BORNE qui appartient aux deux propriétaires à la fois.



De même, 39,5 est la borne inférieure de la 4ième classe mais c'est aussi la borne supérieure de la 3ième classe.

De même, 39,5 est la borne inférieure de la 4ième classe mais c'est aussi la borne supérieure de la 3ième classe.

De fait, nous connaissons les éléments suivants:

Classe	Limites Inf.	Limites Sup.	Bornes Inf.	Bornes Sup.	Fréquences
1	10	19	9,5	19,5	1
2	20	29	19,5	29,5	1
3	30	39	29,5	39,5	2
4	40	49	39,5	49,5	2
5	50	59	49,5	59,5	5
6	60	69	59,5	69,5	6
7	70	79	69,5	79,5	8
8	80	89	79,5	89,5	3
9	90	99	89,5	99,5	2

N.B.: Habituellement on définit la distribution de fréquences à l'aide ou des limites ou des bornes mais non à l'aide des deux à la fois. De plus, il est bon de savoir que les notations suivantes sont équivalentes:

a) Quant aux limites:

" 10 à 19 " ou " $[10 \; ; \; 19]$ " ou " $10 \leqslant x \leqslant 19$ "

b) Quant aux bornes:

" 9,5 à 19,5 " ou " $[9,5 ; 19,5]$ " ou " $9,5 \leq x \leq 19,5$ "

On comprendra que, d'une personne à l'autre et l'habitude aidant, on puisse privilégier l'une ou l'autre des notations.

c) CAS HYBRIDE

En pratique, il arrive aussi qu'on définisse les classes à l'aide d'un intervalle semi-ouvert. Ainsi la distribution précédente serait:

Classes	Fréquences
10 mais inférieur à 20	1
20 mais inférieur à 30	1
30 mais inférieur à 40	2
40 mais inférieur à 50	2
50 mais inférieur à 60	5
60 mais inférieur à 70	6
70 mais inférieur à 80	8
80 mais inférieur à 90	3
90 mais inférieur à 100	2

Ceci permet dès le départ d'avoir une répartition des classes sans interruption qui satisfait d'une certaine façon le besoin de continuité. Mais cette façon de définir la distribution de fréquences n'est pas sans causer certains problèmes théoriques:

1- Chaque classe est définie à l'aide d'un élément qui ne lui appartient pas (celui de droite);

2- Sur l'axe des Réels, le point-milieu entre 10 et 19 est 14,5 alors que celui entre 10 et 20 est 15, ce qui a pour effet de déplacer le «centre» vers la droite. Or l'objectif initial étant de représenter les nombres 10, 11, 12, 13, 14, 15, 16, 17 18, 19, leur «centre» se situe donc à 14,5 et non pas à 15.

La prudence s'impose donc dans cette façon de définir une distribution de fréquences. Il faut pouvoir en tout temps, établir la correspondance entre ce «cas hybride» et les cas précédents. Les éléments de gauche correspondent aux limites inférieures des classes mais ceux de droite n'appartiennent pas aux classes qu'ils définissent. Il faut pouvoir en déduire la limite supérieure qui précède en soustrayant une unité de précision (20 - 1 = 19) et pouvoir retrouver la borne qui se situe entre les deux $(\frac{19 + 20}{2} = 19,5)$; de telle sorte que 19 est la limite supérieure de la première classe (les observations nous sont fournies à l'unité près) et que 19,5 est la borne supérieure.

Le centre de la classe doit être calculé à l'aide de deux éléments (bornes ou limites) de même nature, appartenant en propre à la classe. Ainsi, le centre de la 2ième classe " 20 mais inférieur à 30 " est

$$\frac{20 + 29}{2} = 24{,}5 \text{ ou } \frac{19{,}5 + 29{,}5}{2} = 24{,}5.$$

N.B.: Les notations suivantes sont équivalentes:

" 10 mais inférieur à 20 " ou " $[10 ; 20[$ " ou " $10 \leq x < 20$ "

d) LES CLASSES CONCENTRÉES

Il existe une autre forme de distribution de fréquences qui est utilisée lorsqu'il n'y a qu'un seul élément dans une classe. Tel est parfois le cas lorsqu'on étudie un phénomène discret.

Exemple 10:

Le nombre de personnes par véhicule qui passe à une intersection est noté; on obtient alors la distribution de fréquences suivante:

Nombre de passagers	Nombre d'automobiles
1	8
2	2
3	4
4	3
5	1
6	2

La situation est présentée d'une façon concentrée puisqu'au lieu d'utiliser deux nombres (les limites ou les bornes) pour définir une classe, on en utilise un seul.

Cependant, on pourrait, au besoin, en tirer les renseignements suivants:

Classes	Limites Inf.	Limites Sup.	Bornes Inf.	Bornes Sup.	Fréquences
1	1	1	- 0,5	1,5	8
2	2	2	1,5	2,5	2
3	3	3	2,5	3,5	4
4	4	4	3,5	4,5	3
5	5	5	4,5	5,5	1
6	6	6	5,5	6,5	2

N.B.: 1- Une table de dénombrement pourrait aussi constituer une distribution de fréquences (classes concentrées).

2- Dans le cas d'une distribution de fréquences définie à l'aide des classes concentrées, il n'y a aucune perte d'information.

E) LES CLASSES OUVERTES

Dans certaines situations, il arrive qu'on doive utiliser des «classes ouvertes» (à la première classe ou à la dernière classe ou aux deux à la fois). Par exemple:

Revenu annuel	nombre de personnes
9 999$ ou moins	5
10 000$ à 14 999$	51
15 000$ à 19 999$	24
20 000$ à 24 999$	15
25 000$ ou plus	5

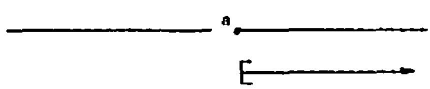
REMARQUES:

1- Les salaires sont connus à 1$ près.

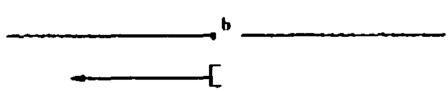
2- Seules les classes des extrémités peuvent être «ouvertes», au besoin. A ce niveau, la notion de classe ouverte se distingue de la notion «d'intervalle ouvert». Par exemple 0 ; 4 est un intervalle ouvert. Il est ouvert parce qu'il n'est pas fermé! (N'appelez surtout pas La Palice...) C'est à dire qu'il suffirait de lui ajouter les nombres des extrémités 0 et 4 pour le fermer. Cela est comparable à un local d'étude. On peut parler d'un local ouvert si la porte est ouverte ou d'un local fermé si la porte est fermée. Au contraire, une classe ouverte est comme un local qui a perdu sa porte.

On peut comparer une classe ouverte à une partie des nombres Réels plutôt qu'à un intervalle. Par exemple:

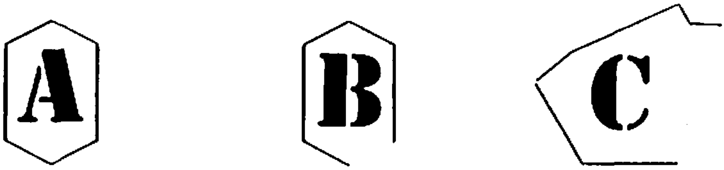
l'intervalle $x \geqslant a$



ou $x < b$



On pourrait aussi associer cette notion de classe ouverte à celle de polygone ouvert ou fermé : A est un hexagone fermé (clôturé) parce qu'il délimite une partie finie du plan. Au contraire, B est un polygone ouvert: il a perdu l'un de ses côtés.



C est aussi un polygone ouvert. Mais on ne peut affirmer avec certitude qu'il a perdu un côté: il pourrait en avoir perdu plusieurs!

Retenons qu'une classe ouverte est nécessairement située à l'une des extrémités de la distribution de fréquences.

3- On pourrait remplacer les limites par les bornes et être en présence de classes ouvertes.

4- La limite (borne) inférieure de la première classe n'existe pas; non plus que la limite (borne) supérieure de la dernière.

5- Cette façon de définir une distribution de fréquences n'est utilisée que lorsque le contexte l'exige. Elle présente de sérieux inconvénients puisque nous ne pouvons pas l'utiliser comme point de départ pour des calculs ultérieurs, ni en faire une représentation graphique. C'est là l'ultime simplification, qu'elle nous satisfasse ou pas. En pratique, lorsque cela est possible, la difficulté peut être contournée par l'élimination (si elles n'ont pas une importance relative trop grande) des valeurs qui s'éloignent exagérément des autres et qui nous ont obligés à faire des classes ouvertes.

1.5 LES DÉFINITIONS

Peu importe la façon qu'on a choisie de définir la distribution de fréquences, les définitions suivantes sont valides:

1- <u>Les LIMITES</u>:

La limite inférieure d'une classe est la plus petite valeur numérique pouvant entrer dans cette classe, compte tenu du degré de précision connu. La limite supérieure est, quant à elle, la plus grande.

2- <u>Les BORNES</u>:

Les bornes correspondent à des points-milieux entre deux limites. Ainsi la borne supérieure d'une classe est le point-milieu entre la limite supérieure de la classe et la limite inférieure de la classe suivante. De même, la borne inférieure d'une classe est le point-milieu entre la limite inférieure de la classe et la limite supérieure de la classe précédente. Pour calculer la première et la dernière borne, on procède comme si une classe précédait (ou suivait).

3- <u>La LONGUEUR</u>:

La longueur (ou largeur ou amplitude) d'une classe est la distance qui sépare ses bornes (et non ses limites).

N.B.:

a) Cette distance se retrouve aussi entre deux limites inférieures successives ou deux limites supérieures ou deux bornes (inférieures ou supérieures) successives.

b) Dans le cas des «classes ouvertes», la première et la dernière classe ont une longueur qui n'a pas de mesure (longueur infinie).

c) Dans tout les autres cas, nous avons choisi de former des classes de même longueur. Ce n'est pas l'effet du hasard. On s'y conformera dans la mesure du possible puisque cela simplifie de beaucoup la manipulation des données ainsi que la représentation graphique (qu'on verra plus loin).

4- Le CENTRE:

Le centre d'une classe correspond au point-milieu entre les bornes (ou les limites) de la classe.

N.B.

a) On peut calculer le centre d'une classe indifféremment à à partir des limites ou des bornes. Par exemple: la classe " 10 à 19 " a son centre à

$\frac{10 + 19}{2}$ = 14,5 si on utilise les limites et à

$\frac{9,5 + 19,5}{2}$ = 14,5 si on utilise les bornes.

(Il faut s'assurer de ne pas confondre les bornes et les limites ou d'utiliser un élément extérieur à la classe pour en calculer le centre).

b) Une classe ouverte n'a pas de centre.

c) Le centre de la classe " 10 mais inférieur à 20 " est bien 14,5 et non 15, puisque 20 n'appartient pas à cette classe et ne peut en aucun cas être utilisé pour calculer le centre;

d) Dans des classes de même longueur, la distance qui sépare les centres est la même que celle qui sépare les limites inférieures ou supérieures entre elles, c'est-à-dire que les centres des classes sont distants entre eux d'une longueur de classe.

5- <u>L'ÉTENDUE</u> des données observées:

C'est la distance qui sépare la plus grande de la plus petite valeur numérique du tableau des données brutes.

Reprenons toutes ces notions à l'aide de l'exemple suivant.

On a mesuré la consommation de gazoline en litres/100 km d'un groupe d'automobiles:

Consommation	Nombre d'automobiles
2,0 à 5,9	3
6,0 à 9,9	12
10,0 à 13,9	7
14,0 à 17,9	4
18,0 à 21,9	2
22,0 à 25,9	2

N.B.: Les données observées sont connues à 0,1 près.

On peut alors en tirer les informations suivantes:

Classes	Limites Inf.	Limites Sup.	Bornes Inf.	Bornes Sup.	Centre	Fréquences
1	2,0	5,9	1,95	5,95	3,95	3
2	6,0	9,9	5,95	9,95	7,95	12
3	10,0	13,9	9,95	13,95	11,95	7
4	14,0	17,9	13,95	17,95	15,95	4
5	18,0	21,9	17,95	21,95	19,95	2
6	22,0	25,9	21,95	25,95	23,95	2

La longueur des classes est de 4 unités.

On remarque ici que les bornes ont une décimale de plus que les limites. De fait, les bornes sont toujours plus précises que les limites et contiennent un chiffre supplémentaire de précision.

Cette distribution de fréquences a été présentée à l'aide de ses limites. Elle aurait pu aussi être présentée de la façon suivante:

Avec les bornes:

Consommation	Fréquences
1,95 à 5,95	3
5,95 à 9,95	12
9,95 à 13,95	7
13,95 à 17,95	4
17,95 à 21,95	2
21,95 à 25,95	2

Ou, cas hybride:

Consommation	Fréquences
2,0 mais inférieur à 6,0	3
6,0 mais inférieur à 10,0	12
10,0 mais inférieur à 14,0	7
14,0 mais inférieur à 18,0	4
18,0 mais inférieur à 22,0	2
22,0 mais inférieur à 26,0	2

2.6 PROCÉDURE DE FABRICATION

Lorsque l'on recueille des observations et qu'on doit fabriquer soi-même une distribution de fréquences, deux cas se présentent:

A) Observations qualitatives

Lorsque le phénomène étudié est d'ordre qualitatif, la table de dénombrement suffit généralement.

B) Observations quantitatives

Après avoir recueilli les observations sous forme de tableau (qu'on appelle le tableau des données brutes), il convient que l'on suive la démarche suivante:

1) Ordonner (si possible) les nombres, du plus petit au plus grand;

2) Faire le dénombrement des observations. Lorsqu'il y a plusieurs répétitions, la table de dénombrement simplifie de beaucoup la présentation des observations; lorsqu'il y a peu de répétitions, cette étape peut être omise à la condition que l'on ait fait 1);

3) Calculer l'étendue;

4) Décider du nombre de classes à former. Souvent, en pratique, le contexte du phénomène étudié va forcer notre choix. Cependant, lorsqu'il n'y a aucune indication de ce genre, il est recommandé de prendre le nombre de classes déterminé par la formule suivante:

$$k = [\![1,5 + 3,3 \quad \text{Log } n]\!]$$

ou k est le nombre idéal de classes.

$[\![\;]\!]$ est la fonction qui fait correspondre au résultat intérieur sa partie entière.

Log est le logarithme à base 10.

n est le nombre d'éléments dans le tableau des observations brutes.

5) Choisir une longueur de classe qui tienne compte des éléments suivants:

- Les classes mises «bout à bout» doivent couvrir complètement l'étendue, de sorte qu'aucune observation n'est laissée pour compte. A ce titre on peut dire que chaque donnée doit pouvoir être classée.

- Longueur approximative = $\dfrac{\text{Etendue}}{\text{Nombre de classes}}$

  si le nombre de classes est un nombre idéal, on obtiendra une longueur approximative «idéale». Il ne faut pas trop s'en éloigner si on veut conserver une image claire des observations. Certaines contraintes pratiques peuvent cependant inciter à s'en éloigner. Par exemple, si la longueur idéale calculée est de 88,5 on peut choisir 100, l'éloignement étant:

$$\frac{(100 - 88,5) \times 100}{88,5} \% = 13\%$$

  Le fait de choisir 100 permet d'obtenir une longueur plus commode (la centaine), sans que l'éloignement ne soit excessif (13%).

Dans le cas où l'on obtient une longueur idéale de 5,8 on ne peut pas choisir 10 même si cela semble plus pratique car cela représente un éloignement exagéré de

$$\frac{(10 - 5{,}8 \times 100}{5{,}8}\ \% = 72\ \%$$

En pratique, l'éloignement ne devrait jamais dépasser 25% à moins que l'on soit en présence d'une situation particulière.

6) Choisir un point de départ qui nous facilite la tâche le plus possible puisqu'il faudra non seulement classer chaque observation mais aussi utiliser des classes comme référence pour expliquer et décrire le phénomène étudié. Ainsi, il est plus facile de parler de la classe 10 à 19 que de faire référence à la classe 12 à 21, car dans le premier cas on se réfère à une «dizaine».
Il en est de même pour la classe 100 à 149 plutôt que 104 à 153 puisque la première correspond à la première moitié de la centaine qui va de 100 à 199. Lorsqu'aucune contrainte d'ordre pratique n'intervient, on doit suivre la règle suivante:

$$\boxed{(K.L) - E = \text{Surplus}}$$

où K = le nombre de classes

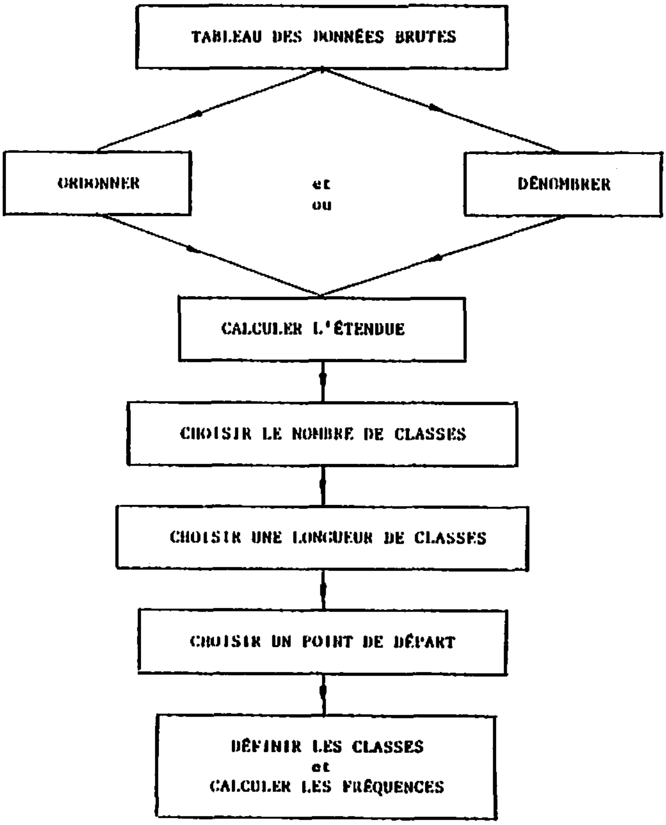
L = la longueur des classes

E = l'étendue des données

On doit veiller à répartir le plus également possible ce surplus entre le début et la fin de la distribution de fréquences;

7) Définir les classes et calculer la fréquence de chacune.

Cette procédure peut être représentée par le diagramme suivant:



Appliquons cette procédure dans l'exemple suivant:

Le concessionnaire DILTEN de votre localité a enregistré le nombre de kilomètres parcourus par chacune de ses automobiles durant la dernière semaine. Il a ainsi observé:

Les DOYODA:	243 422 313 114 612 549.
les TATOSUN:	182 247.
Les DADA:	403 690 192 207.
Les BONTIYAC:	104 390 607.
Les GRAISSELEUR:	179 440 372 590 683.
Les CHEVRE-AU-LAIT:	277 309 394 407 646 509.
et enfin les FORTE:	666 599 407 551

1) Les nombres ordonnés sont:

104	114	179	182	192
207	243	247	277	309
313	372	390	394	403
407	407	422	440	509
549	551	590	599	607
612	646	666	683	690

2) Compte tenu du peu de répétitions, il n'y a pas lieu de faire le dénombrement. Il est à remarquer cependant que, lorsqu'on ne dispose d'aucun moyen (électronique) pour ordonner les données, on fait plutôt un dénombrement, ce qui conduit sensiblement au même résultat.

3) Etendue = 690 - 104 = 586

4) $k = 1,5 + 3,3 \text{ Log } n = \left[\!\left[1,5 + 3,3 \text{ Log } 30 \right]\!\right] = \left[\!\left[6,37 \right]\!\right] = 6$

5) Longueur approximative $= \dfrac{586}{6} = 97,66$

On choisit de former des classes de longueur 100.

6) Point de départ.

Il serait intéressant de partir à 100 de façon à ce qu'une classe corresponde à une centaine.

7) On obtient à l'aide des limites:

Nombre de kilomètres	Nombre d'automobiles
100 à 199	5
200 à 299	4
300 à 399	5
400 à 499	5
500 à 599	5
600 à 699	6

Ceci donnerait, à l'aide des bornes:

Nombre de kilomètres	Nombre d'automobiles
99,5 à 199,5	5
199,5 à 299,5	4
299,5 à 399,5	5
399,5 à 499,5	5
499,5 à 599,5	5
599,5 à 699,5	6

Les centres des classes sont respectivement:

149,5 249,5 349,5 449,5 549,5 et 649,5

On obtiendrait, à l'aide du cas hybride:

Nombre de kilomètres	Nombre d'automobiles
100 x 200	5
200 x 300	4
300 x 400	5
400 x 500	5
500 x 600	5
600 x 700	6

Remarque: les centres des classes sont les mêmes que précédemment.

Considérons un dernier exemple:

Voici le nombre d'actions de la compagnie québécoise de financement Prête-Moi Encore (P.M.E.) transigés quotidiennement entre le 1er mars et le 30 avril 1983 à la Bourse de Montréal:

700	500	900	0	300
1 500	100	600	1 000	100
2 000	1 200	100	1 000	2 000
0	1 500	50	800	1 200
500	800	2 000	300	2 000
100	500	1 000	2 000	1 150
1 500	0	700	100	1 000
2 000	2 000	1 000	500	50

1) Les nombres ordonnés sont:

0	0	0	50	50
100	100	100	100	300
300	500	500	500	500
600	700	800	800	900
1 000	1 000	1 000	1 000	1 000
1 000	1 000	1 150	1 200	1 200
1 500	1 500	1 500	2 000	2 000
2 000	2 000	2 000	2 000	2 000

2) La table de dénombrement:

Nombre d'actions	Nombre de jours
0	3
50	2
100	4
300	2
500	4
600	1
700	1
800	2
900	1
1 000	7
1 150	1
1 200	2
1 500	3
2 000	7
	40

3) L'étendue = 2 000 - 0 = 2 000

4) Le nombre idéal de classes $k = [\![1,5 + 3,3 \text{ Log } 40]\!] = [\![6,79]\!] = 6$

5) La longueur approximative = $\frac{2\ 000}{6}$ = 333,33

On choisit de former des classes de 350 unités de longueur.

6) Le point de départ est imposé, puisqu'il est impossible d'obtenir un nombre négatif d'actions transigées. On choisit donc " 0 ".

7) On obtient donc:

A) A l'aide des limites:

Nombre d'actions	Nombre de jours
0 à 349	11
350 à 699	5
700 à 1 049	11
1 050 à 1 399	3
1 400 à 1 749	3
1 750 à 2 099	7
	40

B) A l'aide des bornes:

Nombre d'actions	Nombre de jours
-0,5 à 349,5	11
349,5 à 699,5	5
699,5 à 1 049,5	11
1 049,5 à 1 399,5	3
1 399,5 à 1 749,5	3
1 749,5 à 2 099,5	7
	40

Le centre des classes est respectivement:

174,5 524,5 874,5 1 244,5 1 574,5 et 1 924,5

C) Cas hybride:

Nombre d'actions	Nombre de jours
$0 \leq x < 350$	11
$350 \leq x < 700$	5
$700 \leq x < 1050$	11
$1050 \leq x < 1400$	3
$1400 \leq x < 1750$	3
$1750 \leq x < 2100$	7
	40

CHAPITRE 2

LES REPRÉSENTATIONS GRAPHIQUES

2.0 INTRODUCTION

2.1 LES DIAGRAMMES

A) Le diagramme circulaire

B) Le diagramme en bâtonnets

C) Les autres diagrammes

2.2 LES HISTOGRAMMES

2.3 LE POLYGONE DES FRÉQUENCES

2.4 CAS PARTICULIER: Cas d'inégalité des longueurs de classes.

2.5 DIAGRAMMES À UTILISER SELON LES PHÉNOMÈNES ÉTUDIÉS.

CHAPITRE 2

2.0 INTRODUCTION

«Le trop comme le trop peu nuit à tout».

Proverbe allemand.

Après avoir recueilli, présenté, regroupé (au besoin) les observations, nous souhaitons souvent en avoir une vue d'ensemble. C'est comme si, après avoir réuni sa famille pour une occasion exceptionnelle, quelqu'un voulait en faire une photographie. Nous apprendrons donc à construire une représentation graphique d'un ensemble de données qui convienne au phénomène étudié.

Ces représentations graphiques sont cependant comme des photos aériennes: elles permettent une bonne vue d'ensemble mais dissimulent les détails. Il ne faut donc pas trop exiger d'elles. Il serait vain de vouloir photographier sa famille en même temps que 100 000 autres personnes qui assistent au même événement et d'espérer distinguer le visage de chacun d'eux. On comprendra ainsi qu'il importe de n'y faire apparaître que l'essentiel.

2.1 LES DIAGRAMMES

Les diagrammes sont des représentations graphiques reliées à l'étude des phénomènes qualitatifs.

A) LE DIAGRAMME CIRCULAIRE

Le diagramme circulaire est un cercle subdivisé en plusieurs secteurs (pointes de tarte), chacun représentant une valeur possible du phénomène étudié. Chaque partie du cercle correspond à une valeur de la variable (qualitative) et doit donc être proportionnelle à son importance relative.

Exemple:

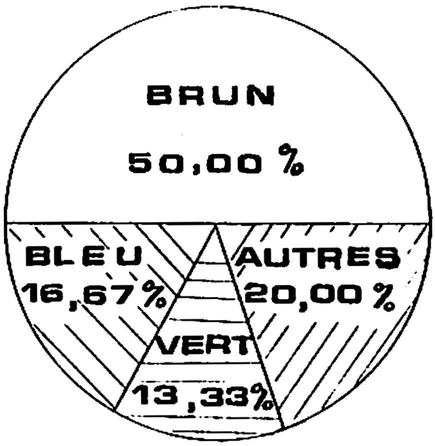
Répartition de la COULEUR des yeux
dans un groupe de 30 étudiants.

Couleur	Nombre d'étudiants	Fréquence relative	Proportion du cercle
BRUN	15	15/30 = 50,00%	180°
BLEU	5	5/30 = 16,67%	60°
VERT	4	4/30 = 13,33%	48°
AUTRES	6	6/30 = 20,00%	72°
	30		360°

Note: Les deux premières colonnes du tableau qui précède constituent une table de dénombrement (car aucune information n'a été perdue). C'est aussi une distribution de fréquences constituée de quatre classes.

Ceci donne lieu au diagramme suivant:

COULEUR DES YEUX DE 30 ÉTUDIANTS.

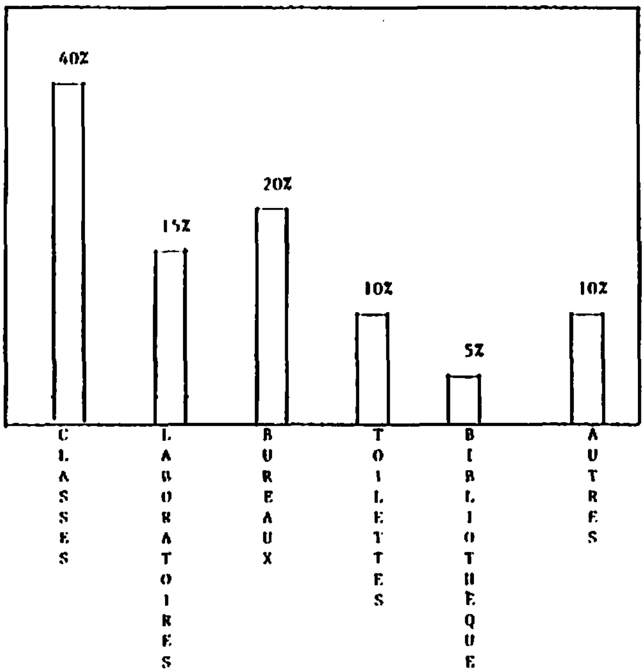


B) LE DIAGRAMME EN BÂTONNETS

Le diagramme en bâtonnets est un ensemble de traits verticaux représentant chacun une valeur possible de la variable (qualitative ou quantitative discrète). La hauteur de chaque trait est proportionnelle à l'importance relative de la valeur de la variable.

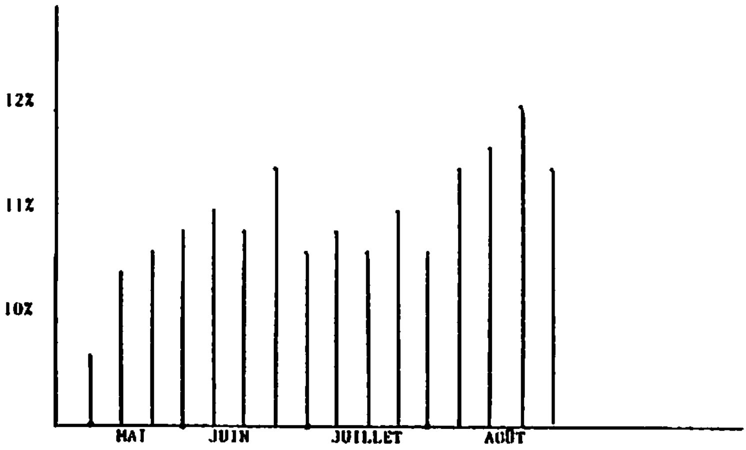
Exemple 2:

SURFACE OCCUPÉE PAR LES LOCAUX DU COLLÈGE.



Exemple 2:

VARIATION DU TAUX D'ESCOMPTE DE LA BANQUE DU CANADA



C) LES AUTRES DIAGRAMMES

Ce sont des diagrammes dont les traits verticaux sont remplacés par des objets qui font plus image.

Exemples:

PHÉNOMÈNES	«OBJETS»
Exportation d'or par continent.	Pièces d'or (empilées les unes sur les autres).
Production de pétrole par pays de l'OPEP.	Barils (superposés).
Revenus familiaux par région du Québec	Colonne de $.

A titre de complément, il suffit de consulter les revues et les journaux spécialisés qui présentent régulièrement des reportages sur des sondages d'opinion publique ou des enquêtes de toutes sortes. Leurs auteurs rivalisent d'habileté (certains avec plus de succès que d'autres) pour attirer l'attention du lecteur. Il faut admettre que ceux qui utilisent la couleur ont un net avantage sur les autres. Si une image vaut mille mots, une image-couleur en vaut des millions!

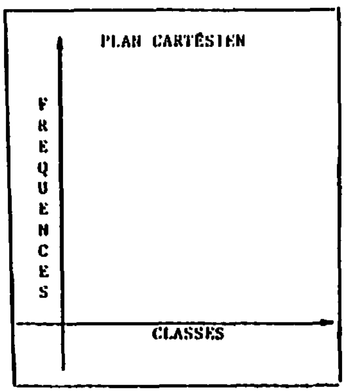
2.2 LES HISTOGRAMMES

Les histogrammes sont des représentations graphiques reliées à l'étude des phénomènes quantitatifs. Ils sont construits à partir d'une distribution de fréquences et du plan cartésien. Comme nous le savons déjà, une table de fréquences est composée de deux colonnes. Le plan cartésien, quant à lui, utilise deux axes.

Il semble donc naturel de faire en sorte que chacun des axes représente une colonne de la table des fréquences.

DISTRIBUTION DE FRÉQUENCES

Classes	Fréquences



1- <u>L'axe vertical</u>

L'axe vertical représente les fréquences. A cette fin, l'identification de la fréquence maximale permet un choix d'échelle qui lui convient. Le graphique qui sera construit se veut une image, une photographie du phénomène étudié. Il est donc illogique de choisir une graduation trop petite ou trop grande. Généralement, lorsque le graphique fait partie d'un manuscrit, il occupe la moitié d'une page. Il est normal, cependant, qu'on ne respecte pas cette proportion dans les livres puisque des appareils spécialisés produisent une aussi bonne clarté dans un plus petit format.

2- <u>L'axe horizontal</u>

L'axe horizontal représente, quant à lui, les classes. Si les données ne prennent que des valeurs positives, seul le premier quadrant servira. Rappelons qu'en mathématiques les axes sont des reproductions de l'ensemble des Réels; cela comporte certaines exigences dont la continuité. Or, au niveau des classes, seules les bornes respectent la continuité. En effet, les BORNES mises «bout à bout» permettent de couvrir toute l'étendue des observations sans interruption. Conséquemment, il faudra faire apparaître les bornes (et seulement celles-ci) sur l'axe horizontal.

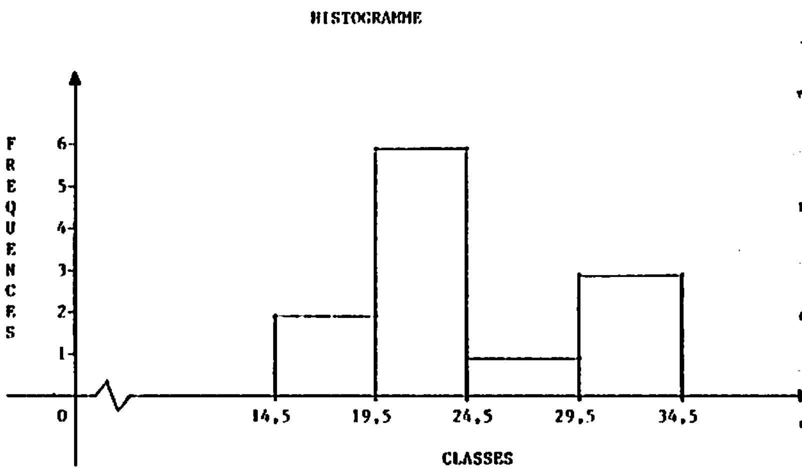
Ainsi, un histogramme sera composé d'un ensemble de rectangles juxtaposés les uns aux autres selon l'ordre des Réels. Chaque rectangle aura la même largeur (correspondant à la longueur des classes) et se verra attribuer une hauteur proportionnelle à la fréquence de la classe qu'il représente.

Examinons la situation suivante:

a) <u>Distribution de fréquences</u>

Classes	Fréquences
15 à 19	2
20 à 24	6
25 à 29	1
30 à 34	3

b) <u>Représentation graphique</u>



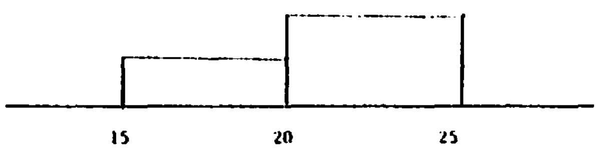
1- Puisqu'on utilise un système d'axes, il faut indiquer la direction positive par des pointes de flèches.

2- Il est important d'identifier chacun des axes. A cette fin, on attribuera à chacun d'eux le titre de la colonne qu'il représente. Ici nous n'avons pas de renseignements spécifiques sur le phénomène étudié. C'est pourquoi les axes portent les titres généraux de «Classes» et de «Fréquences». Il en est de même pour le titre «HISTOGRAMME» qui deviendra spécifique dans une situation donnée.

3- Sur l'axe horizontal nous n'avons fait apparaître que les bornes et ce, pour les raisons suivantes:

a) Respecter la continuité. En effet, les bornes sont les seuls éléments qui respectent la continuité tout en représentant correctement les classes.

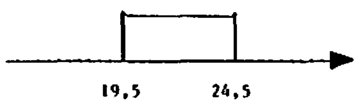
Assurément, la classe «15 mais inférieur à 20» juxtaposée à «20 mais inférieur à 25», et représentée de la façon suivante:



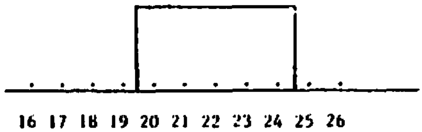
créerait une situation de continuité mais ne pourrait servir à représenter correctement les classes puisque, dans le graphique, l'élément «20» appartient aux deux rectangles alors que, dans la table, il n'appartient qu'à la classe de droite: «20 mais inférieur à 25».

b) La distribution de fréquences est fournie à l'aide des limites. Puisque les limites sont des nombres entiers, on peut déduire que les observations qui ont servi à construire cette table étaient, elles aussi, connues à l'unité près. Il n'y a donc aucun inconvénient à ce que le nombre « 19,5 » appartienne à deux classes puisque ce nombre n'apparaît jamais dans le tableau des données brutes. On pourrait objecter que les nombres qui apparaissent sur l'axe horizontal (les bornes) ne correspondent pas à la réalité (des entiers).
Rappelons-nous qu'il s'agit ici d'une «photo aérienne» et ne soyons pas trop exigeants. Il ne faut quand même pas exiger qu'une photo de la «Baie des Chaleurs» sente le poisson! De toute façon, pour connaître tous les nombres appartenant à chaque classe, on peut, au besoin, les reconstituer en passant des bornes aux limites.

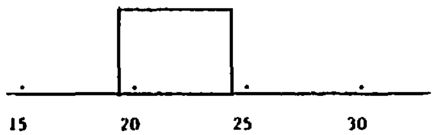
Par souci de clarté et de précision, nous préférons la situation suivante:



à celle-ci



et à celle-là



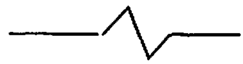
4- Les classes ayant même longueur, une mesure appropriée a été choisie de façon à ce qu'elle puisse être répétée 6 fois sur l'axe horizontal:

a) Une fois pour chacune des quatre classes.

b) Une fois de chaque côté de l'histogramme de façon à:

i) laisser « respirer » le graphique;

ii) prévoir un espace pour tracer, au besoin, le polygone de fréquences (à venir).

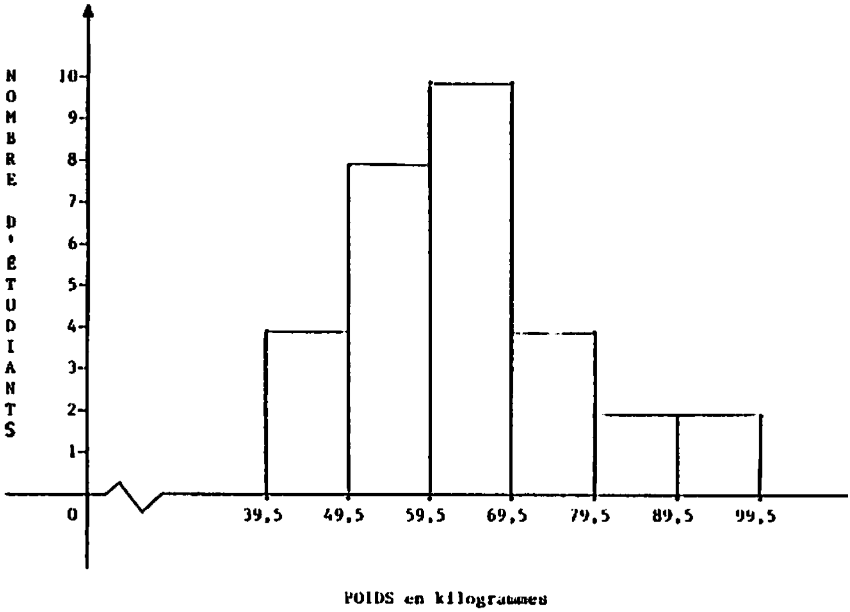
5- Près de l'origine, le symbole «  » indique une « brisure » dans l'axe horizontal. Cela signifie qu'une partie inutilisée du plan cartésien a été retranchée. Il ne faut donc pas s'attendre à ce que l'échelle choisie sur l'axe horizontal soit respectée entre 0 et 14,5. On doit cependant s'y conformer à partir de 14,5.

Examinons l'exemple suivant:

le poids en kilogrammes de 30 étudiants
d'un groupe de Sciences de la santé

Poids	Nombre d'étudiants
$40 \leq X < 50$	4
$50 \leq X < 60$	8
$60 \leq X < 70$	10
$70 \leq X < 80$	4
$80 \leq X < 90$	2
$90 \leq X < 100$	2

RÉPARTITION DE 30 ÉTUDIANTS SELON LEUR POIDS.



Remarques:

1- Le titre et les axes ont été ajustés au phénomène étudié (on a aussi précisé les unités de poids).

2- Même si la distribution de fréquences est fournie à l'aide du «cas hybride», les seules valeurs numériques qui apparaissent sur l'axe horizontal sont les bornes.

3- L'histogramme est utile lorsqu'une «vue d'ensemble» du phénomène est désirée. Il fournit un aperçu général de la situation, mais pas davantage. Lorsque l'on souhaitera avoir des renseignements qui n'apparaissent pas sur l'histogramme, on se donnera d'autres outils. Il ne faut pas, en effet trop exiger de l'histogramme.

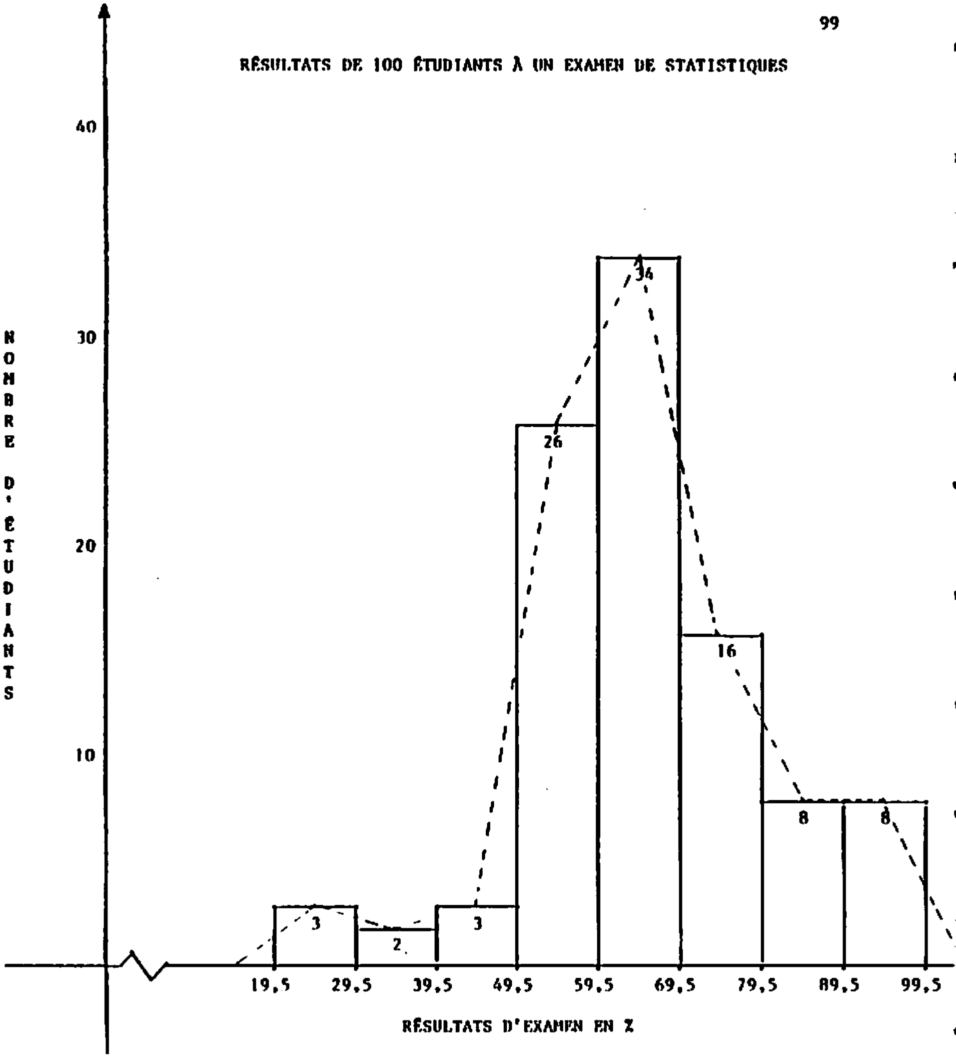
2.3 LE POLYGONE DE FRÉQUENCES

Pour suivre l'évolution du phénomène d'une classe à l'autre, on a souvent recours à un «artifice» visuel qui nous permet de reléguer au second plan les détails de l'histogramme. Il s'agit alors de joindre, par des segments de droites, le centre des bases supérieures des rectangles de l'histogramme. De plus, pour éviter que cette courbe ne reste suspendue indéfiniment, on ajoute à chaque extrémité de l'histogramme deux classes (fictives) de fréquence nulle et on en identifie le centre.

On rejoint ainsi, avec deux autres segments de droites, l'axe horizontal. Lorsque cette courbe, que nous appelons «polygone de fréquences», est tracée à l'aide d'une couleur différente (ou en pointillés si on n'a qu'une couleur à sa disposition), on peut laisser l'histogramme comme toile de fond: les deux graphiques sont alors superposés. Ainsi l'oeil pourra s'attacher à l'un ou à l'autre, au besoin.

Examinons l'exemple suivant qui présente les résultats de 100 étudiants à un examen de Statistiques.

RÉSULTATS DE 100 ÉTUDIANTS À UN EXAMEN DE STATISTIQUES



Remarques:

1- Pour dessiner le polygone de fréquences il est inutile de tracer une courbe régulière. Ce serait aller au-delà de la précision dont nous disposons. Les courbes régulières existent, mais non comme polygones de fréquences. Il suffira donc de relier le centre des bases supérieures à l'aide d'une règle.

2- Un nombre apparaît dans chaque rectangle, ce qui n'est pas une règle générale. Cette apparition est justifiée par les raisons suivantes:

   a) En l'absence d'une table de distribution de fréquences, ces nombres sont le seul moyen de connaître les fréquences exactes.

   b) Les fréquences n'auraient pu être déduites avec exactitude à l'aide de l'échelle en ordonnée.

   c) Il faut veiller à ce que l'apparition de ces nombres n'ait pas pour effet d'embrouiller le graphique.

2.4 CAS PARTICULIER

On observera que nous n'avons pas abordé le cas où la longueur des classes n'était pas la même. En ce cas, l'histogramme ne saurait être construit à partir des fréquences absolues. On comprendra qu'une classe qui est trois fois plus longue que les autres devrait contenir trois fois plus d'observations sans pour cela que le graphique ne lui accorde une hauteur trois fois plus élevée. (Ce qui serait le cas si on portait la fréquence en ordonnée.) Il faudrait alors se donner une mesure qui tienne compte, en même temps, de la fréquence et de la longueur de chaque classe. Cette mesure s'appelle la densité et se calcule avec la formule:

$$di = \frac{fi}{Li}$$

ou di = densité de la i ième classe

fi = fréquence de la i ième classe

Li = longueur de la i ième classe

Ainsi, deux récipients, disposés l'un à côté de l'autre sous la pluie et ayant la même grandeur d'ouverture, recueilleront la même quantité d'eau. Mais si l'un des deux présente une ouverture trois fois plus grande que l'autre, il recueillera, en toute évidence, trois fois plus d'eau. Il en est de même pour les classes qui sont plus larges que les autres. Mais la densité vient rétablir la disparité qui semblait exister de telle sorte qu'une classe trois fois plus large qu'une autre se verra attribuer une densité égale à la condition d'avoir une fréquence trois fois plus élevée.

La construction d'un graphique exigerait alors qu'on porte en ordonnée la densité plutôt que la fréquence. Le graphique ainsi formé n'en continuerait pas moins à porter le nom de «HISTOGRAMME».

En général et dans la mesure du possible, il est préférable d'utiliser des classes de longueur égale.

3.5 REPRÉSENTATON GRAPHIQUE À UTILISER SELON LES PHÉNOMÈNES ÉTUDIÉS

Voici une vue d'ensemble des différentes situations qui peuvent se présenter:

- PHÉNOMÈNE
  - QUALITATIF
    - DÉNOMBREMENT
      - DIAGRAMME
  - QUANTITATIF
    - DISCRET
      - DÉNOMBREMENT
        - DISTRIBUTION DE FRÉQUENCES
          - DIAGRAMME OU HISTOGRAMME
            - POLYGONE DE FRÉQUENCES
    - CONTINU
      - DÉNOMBREMENT
        - DISTRIBUTION DE FRÉQUENCES
          - HISTOGRAMME
            - POLYGONE DE FRÉQUENCES

CHAPITRE 3

LES PARAMÈTRES DESCRIPTIFS

3.0 INTRODUCTION

3.1 LES PARAMÈTRES DE POSITION

- A) LA MOYENNE
  - i) arithmétique
  - ii) pondérée
  - iii) des observations regroupées
- B) LE MODE
  - i) des observations brutes
  - ii) des observations regroupées
- C) LA MÉDIANE

3.2 LES PARAMÈTRES DE DISPERSION

- A) L'ÉTENDUE
- B) L'ÉCART-MOYEN
- C) LA VARIANCE
  - i) des observations brutes
  - ii) des observations regroupées
- D) L'ÉCART-TYPE
- E) LE COEFFICIENT DE VARIATION

CHAPITRE 3

LES PARAMÈTRES DESCRIPTIFS

3.0 INTRODUCTION

«Il utilise les statistiques comme l'ivrogne, les lampadaires: pour s'appuyer plutôt que pour s'éclairer.»

Andrew Lang

Comme nous l'avons déjà dit, la distribution de fréquences et sa représentation graphique sont des outils dont la «portée» est limitée. Il arrive souvent en effet que, pour décrire un ensemble de nombres, on doive faire appel à des outils plus perfectionnés. Rappelons que l'objectif est de décrire un phénomène.

Si l'on vous demandait de décrire la voiture de vos parents, vous feriez sans doute appel à certaines «qualités» visant à pouvoir l'identifier, la distinguer des autres (familiale, 1984, rouge métallique, de marque Mercury, modèle Marquis, etc.).

De même, pour décrire un ensemble de nombres, le statisticien cherchera certaines «qualités intrinsèques» de l'ensemble des nombres, que nous appellerons «PARAMÈTRES». Pour pouvoir qualifier une série de nombres, il doit répondre essentiellement à deux questions.

<u>Première question:</u>

Quelle est l'ordre de grandeur des nombres? Autrement dit, dans quelle partie des Réels se situent les nombres?

La façon idéale de répondre à cette question est d'identifier la zone centrale autour de laquelle se dispersent les nombres. Car il ne faut pas croire que la position d'un groupe de nombres peut être facilement indiquée à l'aide d'un seul d'entre eux. Nous devrons, à l'aide des données, calculer certaines valeurs que nous appellerons des «PARAMÈTRES». Pour décrire clairement la position des nombres, il faudra souvent en calculer plus d'un. Chacun d'eux nous renseignera sur la place qu'occupe le noyau central de la série statistique (ensemble de nombres). Nous les nommerons des paramètres de position (parfois aussi appelés paramètres de tendance centrale).

Nous cherchons, par exemple, à identifier le centre de la série suivante:

$$\begin{vmatrix} 0 & 1 & 4 & 5 & 5 & 5 & 5 & 5 & 6 & 9 & 10 \end{vmatrix}$$

qui semble évident

ou de $\begin{vmatrix} 1,1 & 1,7 & 1,9 & 6,0 & 72,3 & 1\,984,4 \end{vmatrix}$

qui apparaît moins évident.

<u>Deuxième question:</u>

Comment sont répartis les nombres autour du centre?

Car le fait de «situer les nombres» ne suffit pas pour distinguer deux ensembles différents dont les centres sont situés dans la même région. Nous ferons donc appel à des paramètres de dispersion qui nous renseigneront sur le comportement respectif des nombres entre eux, sur leur étalement à l'intérieur de la série.

Il nous serviront, par exemple, à distinguer entre elles les séries suivantes:

$$\begin{vmatrix} -4 & 0 & 10 & 10 & 10 & 20 & 24 \end{vmatrix}$$

$$\begin{vmatrix} 9 & 9 & 10 & 10 & 10 & 11 & 11 \end{vmatrix}$$

3.1 LES PARAMÈTRES DE POSITION

«Quand tu secoues un arbre, regarde aussi où tombent les fruits.»

Proverbe d'Afrique noire

A) LA MOYENNE

1) LA MOYENNE ARITHMÉTIQUE

Imaginons que deux poids (P1 et P2) équivalents (deux unités de poids par exemple) sont placés sur une balance à pivot et cherchons le point d'équilibre, comme à la figure 1.

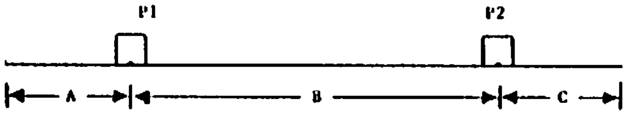


Figure 1

Les poids étant disposés tel que présenté à la figure 1, nous pouvons identifier trois « zones » distinctes (A B et C). Tentons de placer un pivot sous la barre, de façon à établir un équilibre. Par intuition, nous savons que ce pivot ne saurait être placé dans les zones A ou C. Il faudra naturellement diriger nos recherches vers la zone centrale B. (Figure 2)

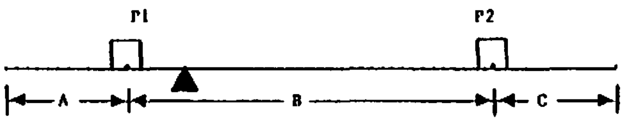


Figure 2

Tentons une première expérience d'équilibre et plaçons le pivot tel que présenté à la figure 2. L'équilibre ne se fera sûrement pas ainsi, puisque les deux poids ont même importance.

Pour y arriver, il faudra déplacer le pivot vers la droite. Effectuons (lentement) un tel déplacement et observons les résultats. Il arrivera un moment où l'équilibre sera parfait. Arrêtons-nous à cet instant et fixons le pivot de telle sorte que nous obtenions la situation illustrée à la figure 3.

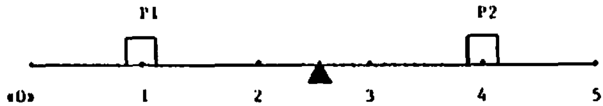


Figure 3

Mesurons la distance des éléments présents à partir d'un point «O». A l'aide de la figure 3, nous pouvons observer que:

P1 est à 1 unité de «O»;

P2 est à 4 unités de «O»;

le point d'équilibre est à 2,5 unités de «O».

Conclusion:

Il y a équilibre parce que les deux poids (équivalents) sont placés à égale distance de part et d'autre du point d'équilibre. (Ou parce qu'un pivot a été placé au centre de la distance qui les sépare).

Résumons la situation d'équilibre entre deux poids équivalents par le diagramme suivant (figure 3B)

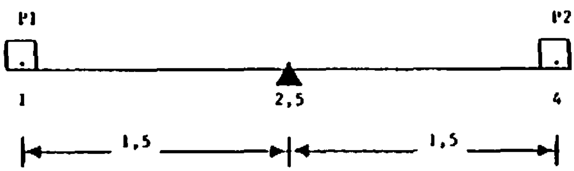


Figure 3B

Examinons maintenant le comportement de trois (unités de) poids (équivalents).

Reprenons le même système d'équilibre et tentons d'ajouter un troisième poids comme en figure 4.

Figure 4

Si cette unité de poids est placée en 2,5 (point d'équilibre du système), l'équilibre est maintenu (figure 4).

Par contre, si ce poids est déplacé à «2», l'équilibre est rompu (Figure 5).

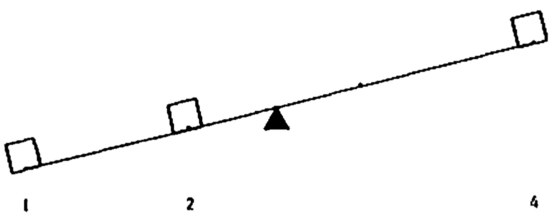


Figure 5

En définitive, ce qui influence l'équilibre ici ce n'est pas le poids (puisqu'il est toujours le même), mais sa «position» dans le système. Il déséquilibre le système d'autant plus qu'il s'éloigne du point d'équilibre.

Pour créer un nouvel équilibre, il faudra déplacer le pivot de la figure 5 vers la gauche. L'équilibre sera rétabli au moment où la «distance totale» des poids, mesurée à partir du pivot, sera équivalente de chaque côté du point d'équilibre.

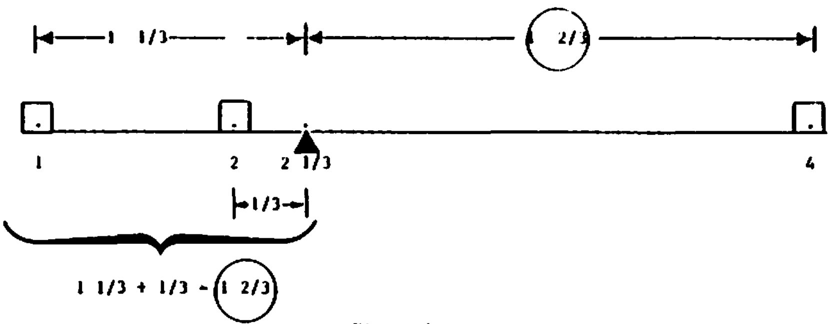


Figure 6

Transposons cette situation dans l'ensemble des Réels et cherchons un «point d'équilibre» entre deux nombres. Par exemple: comme à la figure 7, cherchons le point d'équilibre entre les nombres 2 et 14

Figure 7

Le point d'équilibre entre deux nombres (2 et 14) correspond au point-milieu entre eux $\left(\frac{2 \quad + \quad 14}{2}\right)$

et s'appelle la moyenne arithmétique (ici elle vaut 8).

Le point d'équilibre de trois nombres X1 X2 et X3 correspond encore à la moyenne arithmétique et se calcule par la formule suivante:

$$\bar{X} = \frac{X1 + X2 + X3}{3}$$

où $\bar{X}$ (lire X barre) est la notation utilisée pour la moyenne et X1, X2 et X3 sont les nombres.

Le point d'équilibre de n nombres

$$\overline{X} = \frac{X1 + X2 + X3 + \ldots + Xn}{n} = \sum \frac{Xi}{n}$$

où $\overline{X}$ est la <u>MOYENNE ARITHMÉTIQUE</u>

les Xi sont les nombres

et $\sum$ est un opérateur qui indique de faire la somme.

Puisque le point d'équilibre est situé au «centre» des nombres, c'est un premier paramètre qui peut renseigner sur la position de l'ensemble des nombres. Il faut cependant être prudent dans l'utilisation qu'on fait de ce paramètre car il est fortement influencé par des valeurs extrêmes. (Le point d'équilibre est très sensible aux poids qui sont très éloignés de lui).

11) LA MOYENNE PONDÉRÉE

Nous avons examiné le comportement d'un point d'équilibre lorsque les poids sont équivalents; qu'arriverait-il si ce n'était pas le cas, si certains avaient plus d'importance que d'autres? Examinons cette situation en créant d'abord un équilibre à l'aide de deux poids équivalents (Figure 8)

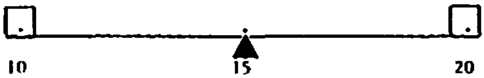


Figure 8

puis en doublant le premier poids (Figure 9).

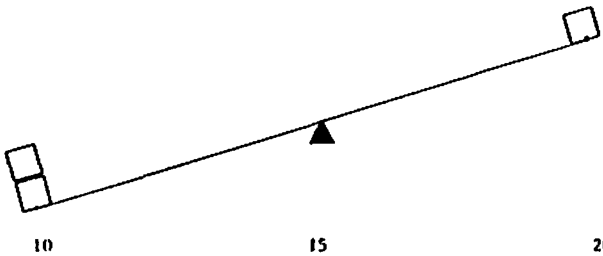


Figure 9

L'équilibre est rompu et, pour le rétablir, il faudra déplacer le pivot vers la gauche et plus exactement à $\frac{10 + 10 + 20}{3} = 13\frac{1}{3}$ unités (Figure 10).

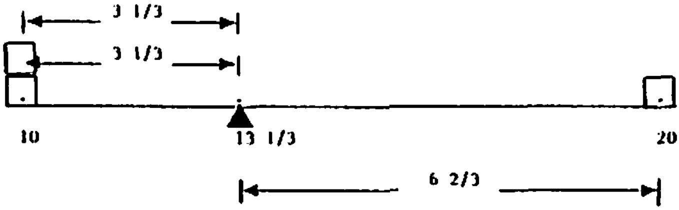


Figure 10

Remarques:

- Puisque $\frac{10 + 20}{2} = 15$, on peut dire que 15 est le point-milieu entre 10 et 20.
- Mais 15 n'est pas le point d'équilibre de 10 et 20 lorsque 10 est deux fois plus important que 20.
- Bien que 15 soit la moyenne arithmétique de 10 et 20, 13 1/3 en est la moyenne pondérée (c'est-à-dire qui tient compte de l'importance de chacun).
- Lorsque les nombres n'ont pas la même importance, la seule moyenne qui soit représentative est la moyenne pondérée.

Mais quel facteur détermine l'importance qu'il faut accorder à chaque élément? C'est le contexte qui dicte l'importance de chacun des nombres.

<u>Exemple 1</u>:

Considérons l'exemple suivant. Des étudiants de collège I doivent inscrire le nombre d'enfants que compte leur famille. Voici les observations recueillies:

Nombre d'enfants	Nombre de familles
1	5
2	8
3	10
4	5
5	2
	30

Quel est le nombre moyen d'enfants par famille?

a) Calculons la moyenne arithmétique du nombre d'enfants:

$$\frac{1 + 2 + 3 + 4 + 5}{5} = \frac{15}{5} = 3 \text{ enfants.}$$

Pouvons-nous pour autant affirmer que le nombre moyen d'enfants par famille (dans ce groupe) est 3?

Pour connaître le nombre moyen d'enfants par famille, il faut connaître deux choses:

1- Le nombre total d'enfants

2- Le nombre total de familles.

A partir de la solution précédente, pouvons-nous affirmer que:

1- Le nombre total d'enfants est 15?

2- Le nombre total de familles est 5?

Bien sûr que non! Alors pourquoi diviser 15 par 5?

Examinons la solution suivante:

b) Reconstituons les observations qui nous ont menés à la table de dénombrement (puisqu'aucune information n'a été perdue).

1	1	1	1	1	2
2	2	2	2	2	2
2	3	3	3	3	3
3	3	3	3	3	4
4	4	4	4	5	5

Il y a 30 familles (30 étudiants de familles différentes). Pour trouver le nombre total d'enfants, nous pouvons additionner les trente nombres ou, plus simplement, utiliser ce merveilleux opérateur qu'est la multiplication:

$$(1 \cdot 5) + (2 \cdot 8) + (3 \cdot 10) + (4 \cdot 5) + (5 \cdot 2) = 81$$

de telle sorte que $\bar{X} = \frac{81}{30} = 2{,}70$ enfants

Ce résultat correspond à une moyenne peut-être moins «commode», moins «vivante» (à cause du ,70) mais elle n'en demeure pas moins la seule vraie moyenne dans la situation présente. <u>C'est la MOYENNE PONDÉRÉE qui se définit comme suit:</u>

$$\bar{X} = \frac{\sum Xi.pi}{\sum pi}$$

où $\bar{X}$ est la <u>MOYENNE PONDÉRÉE</u>

les Xi sont les observations

pi est l'importance à accorder

à chacun des Xi

Il faut donc user de prudence dans le calcul d'une moyenne et choisir d'effectuer le calcul approprié.

Exemple 2:

Si, dans un groupe, les 29 étudiants ont en commun \$60 d'argent en poche et que le professeur a \$40, quel est l'avoir moyen par individu dans ce groupe?

Première réponse:

$$\bar{X} = \frac{60 + 40}{2} = 50,00\ \$$$

Pouvons-nous pour autant affirmer que l'avoir moyen de ce groupe est de 50,00 \$?

Examinons une deuxième réponse possible.

Pour connaître l'avoir moyen par individu il faut connaître:

1- L'avoir total des individus:

\$60 pour l'ensemble des 29 étudiants

+ \$40 pour le professeur

\$100

2- Le nombre total d'individus (29 + 1 = 30)

donc: $\overline{X} = \frac{100}{30} = 3{,}33\$$

Ce qui est loin de 50,00$.

Laquelle des deux réponses est exacte? Rappelons que nous sommes à la recherche de l'avoir moyen « par individu ». Le fait de diviser par deux nous renseigne sur l'avoir moyen de deux groupes. Si chaque groupe n'a pas la même importance, il faut en tenir compte. C'est ce que fait la deuxième réponse et c'est la seule solution acceptable.

En conclusion, on utilise la moyenne arithmétique lorsque tous les nombres ont une même importance et la moyenne pondérée dans tous les autres cas.

III) OBSERVATIONS REGROUPÉES

Lorsque les observations ont été regroupées sous forme de distribution de fréquences et donc qu'il s'est perdu une certaine partie de l'information initiale, il peut arriver que l'ensemble des données brutes ne soit plus accessible (ce qui est de plus en plus rare, cependant, avec l'arrivée des calculatrices et des micro-ordinateurs) et que l'on doive quand même calculer la moyenne. Ne connaissant plus la valeur numérique exacte de chacune des données, il nous faut énoncer deux hypothèses:

HYPOTHÈSE NUMÉRO 1
Dans chaque classe, les observations sont réparties d'une façon homogène.

Ouvrons ici une longue parenthèse et examinons les sens qu'il faut attribuer au mot: HOMOGÈNE.

Que connaissons-nous d'homogène dans notre vie quotidienne? Le lait «homogénéisé». Et en quoi pouvons-nous dire qu'il est homogène? Le gras qu'il contient est réparti d'une façon homogène. S'il contient 2% de matières grasses, le fait d'avoir été homogénéisé assure que chaque partie contient 2% de gras. Proportionnellement, il y a 2 parties de gras pour 100 parties de lait ou une partie de gras pour 50 parties de lait. C'est donc dire que si on prélève, dans un litre de lait, l'équivalent de 50 molécules, on est assuré que celle-ci contienne exactement une molécule de gras. Bien sûr, le fractionnement s'arrête à 50 parties. La molécule est ici la plus petite unité de fractionnement permettant de parler d'homogénéité. Dans l'expression «lait homogène», il y a donc du lait dont le gras a été uniformément réparti.

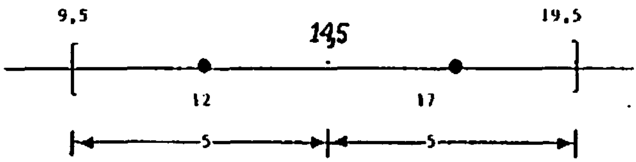
Dans l'expression «classe homogène», il y a au départ une classe dont on a uniformément réparti les observations. Il s'agit de déterminer quelle partie de la classe on doit prélever pour s'assurer la présence d'une observation.

Considérons la distribution de fréquences suivante:

Classes	Fréquences
10 à 19	2
20 à 29	5
30 à 39	10

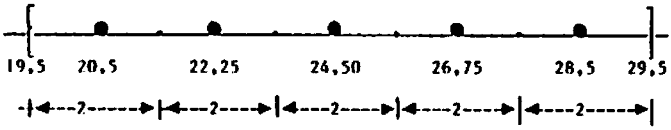
A) Isolons la classe 10 à 19 avec une fréquence de 2. «Ouvrons» cette classe et tentons de répartir ses deux éléments d'une façon homogène:

Puisque nous disposons de 19,5 - 9,5 = 10 unités de longueur, plaçons les deux observations à 12 et à 17 unités. De cette façon, nous obtenons:



Qu'y a-t-il d'homogène dans cette classe? Les deux éléments (12 et 17) font une partition de la classe en deux segments de 5 unités. Chacun des segments contient, en son centre, une observation. Quelle est la proportion d'observations dans cette classe homogène? Une observation pour chaque segment de 5 unités (une partie pour 5). Cette classe est homogène en ce sens que: quel que soit le segment (intervalle semi-ouvert) de 5 unités qu'on choisisse dans la classe, on est assuré qu'il contient une observation.

B) Isolons la classe 20 à 29 qui contient 5 observations. Pour répartir d'une façon homogène ces 5 observations, il faut diviser la classe en 5 parties de $\frac{10}{5} = 2$ unités. On obtient



En quoi cette classe est-elle homogène? Quelle que soit la portion (intervalle semi-ouvert) de 2 unités choisie, on est assuré d'y trouver une observation.

Qu'arrive-t-il si on choisit une portion inférieure à 2 unités? Nous pouvons répondre à cette question en posant une autre question: Que se passe-t-il si on choisit d'examiner une seule molécule de lait homogénéisé plutôt que 50? Peut-on être assuré d'y trouver 2% de gras? Bien sûr que non, car il y a une limite inférieure à la plus petite partie qu'on peut examiner. Pour le lait «une partie par 50» nous contraint à prendre <u>au moins 50 particules élémentaires</u>. Dans une classe, la plus petite partie est déterminée par:

$\frac{L}{f}$ où L est la longueur
f est la fréquence.

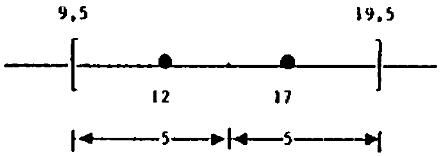
Ainsi, dans la classe «30 à 39», la plus petite partie nous assurant la présence d'un élément est:

$$\frac{10}{10} = 1 \text{ unité.}$$

Cette classe est donc homogène car on retrouve une observation par unité de longueur.

Cette première hypothèse nous rassure quant à la position des nombres puisque le point d'équilibre d'une classe homogène est toujours situé en son centre.

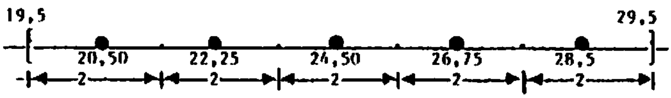
En effet, dans la classe homogène « 10 à 19 » de fréquence 2:



le point d'équilibre est situé à $\frac{12 + 17}{2} = 14,5$.

Or, 14,5 est aussi le centre de la classe $\left(\frac{10 + 19}{2}\right)$

De même, dans la classe homogène «20 à 29»



le point d'équilibre est $\frac{20,5 + 22,5 + 24,5 + 26,5 + 28,5}{5} = 24,5$

et 24,5 est aussi le centre de la classe $\left(\frac{20 + 29}{2} = 24,5\right)$

C'est évidemment une conséquence de l'hypothèse d'homogénéité.

<u>HYPOTHÈSE NUMÉRO 2</u>

On peut attribuer à chaque observation

la valeur numérique du centre de sa classe.

Cette hypothèse vise à satisfaire nos besoins quant au calcul des paramètres de dispersion tout en nous assurant, en moyenne, d'une erreur minimale.

Examinons la classe «14 à 18» avec une fréquence de 1:

Classe	Valeurs possibles	Centre
14 à 18	14 15 16 17 18	16

Valeur attribuée à l'observation	Réalité	Erreur
14	14	0
14	15	1
14	16	2
14	17	3
14	18	4

Erreur totale = 10

Erreur moyenne = 2

15	14	1
15	15	0
15	16	1
15	17	2
15	18	3
	Erreur totale =	7
	Erreur moyenne =	1,4

16	14	2
16	15	1
16	16	0
16	17	1
16	18	2
	Erreur totale =	6
	Erreur moyenne =	1,2

17	14	3
17	15	2
17	16	1
17	17	0
17	18	1
	Erreur totale =	7
	Erreur moyenne =	1,4

18	14	4
18	15	3
18	16	2
18	17	1
18	18	0
	Erreur totale =	10
	Erreur moyenne =	2

PRINCIPE: Si l'erreur est inévitable, tentons de la minimiser le plus possible. Or, comme nous ne connaissons plus la valeur numérique de l'observation, nous lui attribuerons la valeur qui entraîne la plus petite erreur (en moyenne). Dans le tableau qui précède, la plus petite erreur (moyenne) vaut 1,2 et correspond à l'utilisation de 16 comme valeur attribuée. Mais 16 est le centre de la classe. Or, ce n'est pas par hasard qu'il correspond à la plus petite erreur possible (en moyenne): cela est dû à sa position centrale dans la classe «14 à 18».

Ainsi, en utilisant les deux hypothèses qui précèdent, on obtient:

Classe	Centre	Fréquences	Observations	Somme
1	C1	f1	$\underbrace{C1\ \ C1\ \ldots\ldots\ C1}_{f1\ \text{fois}}$	C1.f1
2	C2	f2	$\underbrace{C2\ \ C2\ \ldots\ldots\ C2}_{f2\ \text{fois}}$	C2.f2
3	C3	f3	$\underbrace{C3\ \ C3\ \ldots\ldots\ C3}_{f3\ \text{fois}}$	C3.f3
:	:	:	:	:
:	:	:	:	:
:	:	:	:	:
i	Ci	fi	$\underbrace{Ci\ \ Ci\ \ldots\ldots\ Ci}_{fi\ \text{fois}}$	Ci.fi

Pour connaître la moyenne il faut connaître:

1- La somme des observations: $\sum Ci.fi$

2- Le nombre total d'observations: $\sum fi$

de telle sorte que:

$$\bar{X} = \frac{\sum Ci.fi}{\sum fi}$$

où $\bar{X}$ est la MOYENNE des données regroupées

Ci est le centre de la i ième classe;

fi est la fréquence de la i ième classe.

Remarquons que ce n'est là qu'une application de la formule de la moyenne pondérée où les fréquences de classes servent de paramètres d'importance.

<u>Exemple 3</u>:

Une enquête nous apprend que, dans la région, les usines peuvent être classifiées selon le nombre d'employés, de la façon suivante:

Nombre d'employés	Nombre d'usines
10 à 19	64
20 à 29	21
30 à 39	10
40 à 49	5

Quel est le nombre moyen d'employés par usine?

Pour le trouver il faut connaître deux choses:

a) le nombre total d'employés;

b) le nombre total d'usines.

Puisque nous ne connaissons pas exactement le nombre d'employés dans chaque usine, nous devrons procéder de la façon suivante:

Classe	Ci Centre	fi Fréquence	Ci . fi
1	14,5	64	928
2	24,5	21	514,5
3	34,5	10	345
4	44,5	5	222,5
		100	2 010

$$\overline{X} = \frac{\sum Ci.fi}{\sum fi} = \frac{2\ 010}{100} = 20,10$$

Une telle façon de procéder (en tenant pour acquis que les règles de regroupement des observations ont été suivies) nous assure d'une erreur (écart entre la moyenne calculée et la moyenne des données brutes) qui est (en moyenne) inférieure à 5%.

B) LE MODE

1) OBSERVATIONS BRUTES

Dans la poursuite de notre objectif qui consiste à décrire l'ensemble des nombres observés en identifiant son centre, nous tenterons d'en identifier (un ou plusieurs) qui attirent l'attention, qui se distinguent des autres. Or, toutes choses étant égales par ailleurs, (grosseur de caractère, couleur, relief), comment distinguer un nombre des autres?

Par son originalité?

Par exemple, imaginons un seul nombre positif mêlé à des nombres négatifs ou un seul nombre fractionnaire mêlé à des nombres entiers.

ou Par son abondance?

Par exemple, le nombre 10 apparaît 4 fois alors que les autres n'apparaissent qu'une seule fois.

Bien que l'originalité d'un nombre puisse attirer l'attention de notre oeil, elle ne peut retenir celle du statisticien. Son abondance, au contraire, permettra à ce nombre de prendre position au centre des données.

Lorsqu'une chanson est à la mode on l'entend souvent. Par analogie, on dira d'un nombre qu'il représente <u>le mode</u> lorsque:

1- Il apparaît <u>plus souvent</u> que les autres.

2- C'est celui qui apparaît <u>le plus souvent</u>.

Cela suppose évidemment qu'il y a des répétitions. Et si plusieurs nombres sont répétés, nous choisirons celui qui est répété le plus souvent.

Par exemple, dans les séries statistiques (ensemble de nombres) suivantes:

- $\left\{ 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 2 \quad 7 \quad 3 \right\}$;

le mode est 1

- $\left\{ 4 \quad 4 \quad 7 \quad 10 \quad 11 \quad 15 \quad 15 \quad 16 \quad 19 \right\}$:

le mode est double; il est représenté par 4 et 15.

Attention: il n'existe pas pour autant deux modes puisque le mode est une caractéristique, une qualité, une manifestation des nombres; et en ce sens, il n'y a jamais plus d'un mode, comme il ne peut exister deux moyennes (point d'équilibre).

$$- \left\{ 1 \quad 1 \quad 4 \quad 4 \quad 7 \quad 7 \quad 9 \quad 9 \right\} :$$

Dans cette série, le mode n'existe pas, puisqu'aucun nombre ne se distingue des autres.

$$- \left\{ 1 \quad 2 \quad 7 \quad 7 \quad 14 \quad 20 \quad 22 \quad 22 \quad 29 \quad 30 \quad 30 \quad 47 \quad 99 \right\} :$$

Le mode est triple; il est représenté par 7, 22 et 30.

Compte tenu que le fait d'apparaître plus souvent que les autres a pour effet d'attirer la moyenne (point d'équilibre) vers lui, le mode a de fortes chances de se situer dans la zone centrale des données. En ce sens, c'est, au même titre que la moyenne, un paramètre de position. Il est d'autant plus représentatif du centre qu'il se distingue des autres (et inversement). Dans certaines situations il est le seul paramètre de position qui a un sens. Par exemple, dans un magasin de chaussures, le fait de savoir que le mode est «6» est beaucoup plus utile que le fait de calculer une moyenne de «6,87». Le propriétaire peut réorienter ses achats ou ses ventes à rabais, selon le cas, parce qu'il connaît le mode (et sait l'interpréter).

Dans certaines autres situations, il ne signifie absolument rien. C'est le cas lorsqu'il se distingue peu des autres et se trouve loin du centre.

Par exemple dans:

$$\begin{vmatrix} 0 & 0 & 1 & 2 & 3 & 4 & 5 & 10 \end{vmatrix}$$

où $\bar{X} = \frac{25}{8} = 3,13$

MODE = 0 ; ce qui est loin du centre.

ii) OBSERVATIONS REGROUPÉES

Evidemment, en regroupant les observations sous forme de distribution de fréquences, il a fallu sacrifier une certaine somme d'informations (sauf dans le cas des classes concentrées). Ne connaissant plus l'identité de chaque nombre, il nous est impossible de reconnaître celui (ou ceux) qui se distingue(nt) le plus des autres. Rappelons-nous cependant que nous avons contourné ce problème d'identification pour calculer la moyenne des données regroupées. Nous avons, en effet, émis l'hypothèse que chaque nombre pouvait se voir attribuer la valeur du centre de sa classe. Ainsi, dans l'exemple:

Classe	Fréquence
0 à 2	10
3 à 5	15
6 à 8	10
9 à 11	5

On peut reconstituer une série «hypothétique» de nombres à partir des centres 1, 4, 7, 10.

Il nous faut donc agir comme si nous avions le tableau suivant:

1	1	1	1	1	1	1	1	1	1
4	4	4	4	4	4	4	4	4	4
4	4	4	4	4	7	7	7	7	7
7	7	7	7	7	10	10	10	10	10

De telle sorte que nous puissions identifier le mode comme étant 4 (qui apparaît 15 fois, donc plus souvent que les autres).

Pour identifier le mode, il n'est pas nécessaire de récrire à chaque fois un tableau hypothétique d'observations brutes, il suffit de comprendre le sens de l'hypothèse émise. Il suffit enfin de repérer la fréquence la plus élevée (qu'on appelle la fréquence modale) et d'identifier le centre de la classe correspondante (qu'on appelle classe modale). Ainsi, dans l'exemple qui précède, la fréquence modale est 15; la classe modale est la deuxième; le mode des observations regroupées est 4. Attention: la fréquence modale n'est pas celle qui apparaît le plus souvent (10) mais bien la plus grande (15).

Exemple 3:

Classe	Fréquence
5 mais inférieur à 10	10
10 mais inférieur à 15	10
15 mais inférieur à 20	10
20 mais inférieur à 25	10

Le mode des données regroupées n'existe pas, pas plus que la fréquence modale ou la classe modale.

Exemple 4:

Classe	Fréquence
4999,5 ou moins	4
4999,5 à 9999,5	16
9999,5 à 14 999,5	10
14 999,5 ou plus	25

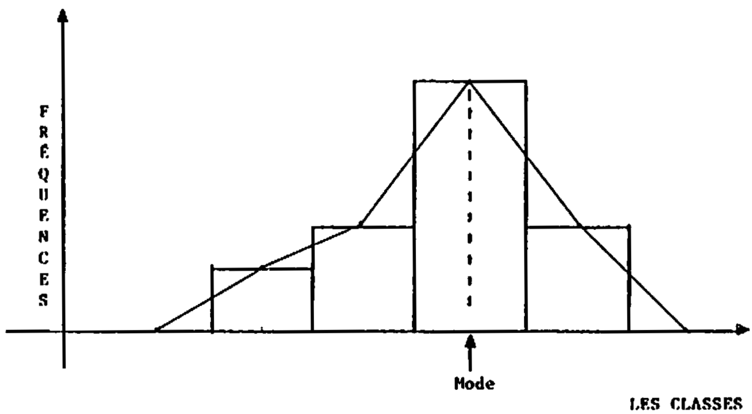
La fréquence modale est 25.

La classe modale est la quatrième: « 14 999,5 ou plus ».

Le mode n'existe pas (puisque c'est une classe ouverte).

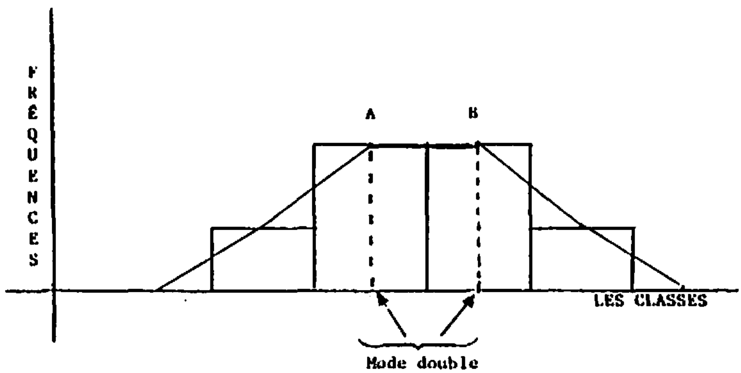
Remarquons que le mode des données regroupées prend la valeur de l'abcisse du point le plus élevé du polygone des fréquences.

Exemple 5:



C'est aussi celui qui (visuellement) attire le plus l'attention.

Exemple 6:



ATTENTION: Puisque tous les points situés sur le segment AB sont à la même hauteur et que celle-ci est la plus élevée, on pourrait croire que le mode peut être représenté par tous les points correspondant à ceux de AB. Mais il n'en est rien puisque, comme nous l'avons déjà dit, le polygone de fréquences ne constitue pas une courbe régulière mais plutôt une suite de segments de droites. Donc, ici, le mode correspond aux centres des deuxième et troisième classes.

C) LA MÉDIANE

Compte tenu que les deux premiers paramètres de position ne décrivent pas parfaitement toutes les situations, nous aurons recours à un troisième qui ne présente pas les défauts des premiers. Rappelons notre objectif qui est de «situer» les nombres, d'identifier le centre de l'ensemble des valeurs numériques.

Prenons la série statistique suivante:

$$\left| \quad 7 \quad 12 \quad 1 \quad 13 \quad 10 \quad \right|$$

Dans un premier temps ordonnons la série:

$$\left| \quad 1 \quad 7 \quad 10 \quad 12 \quad 13 \quad \right| .$$

Eliminons ensuite, tour à tour, le premier et le dernier nombre puisqu'ils sont situés aux extrémités, ce qui est contraire à notre objectif (qui est d'identifier le centre). La première paire éliminée est « 1 et 13 »; il reste donc:

$$\left\{ \boxed{1} \quad 7 \quad 10 \quad 12 \quad \boxed{13} \right\} .$$

La seconde paire est « 7 et 12 ». Il reste

$$\left\{ \boxed{1} \quad \boxed{7} \quad 10 \quad \boxed{12} \quad \boxed{13} \right\} .$$

En ce sens, le nombre restant «10» est au centre des données (indépendamment de la distance qui les sépare puisque la moyenne en tient déjà compte).

De même dans la série

$$\left\{ \boxed{10 \quad 20 \quad 37} \quad 41 \quad \boxed{44 \quad 50 \quad 59} \right\}$$

41 est au «milieu» des nombres ordonnés. Ce paramètre s'appelle la «MÉDIANE» et présuppose qu'on a déjà ordonné les nombres. On remarque, dans les deux séries qui précèdent, qu'il y a autant de nombres à gauche de la médiane qu'il y en a à droite.

Examinons l'exemple suivant:

0 2 3 4 9 14

et utilisons la même procédure d'élimination:

$$\left\{ \boxed{0} \quad \boxed{2} \quad \boxed{3} \quad \boxed{4} \quad \boxed{9} \quad \boxed{14} \right\}.$$

Le partage est très équitable mais il ne reste aucun élément pour identifier le centre. Qu'à cela ne tienne! Il suffit d'en déterminer un qui soit situé entre les éléments de la dernière paire de nombres éliminés (c'est-à-dire entre 3 et 4). Par exemple, si nous choisissons d'identifier la médiane avec le nombre 3,1 il est exact de dire qu'il y a autant de nombres qui précèdent (0, 2 et 3) qu'il y en a qui suivent (4, 9 et 14). Donc 3,1 est une médiane. Oui, mais il en est de même pour 3,9 ou 3,64 ou 3,1497 ou . De fait, il en existe une infinité. Tout nombre appartenant à l'intervalle 3 ; 4 peut donc être appelé médiane. Rappelons cependant que notre objectif est d'identifier le centre des nombres.

Si le centre correspond à un intervalle, l'objectif ne sera vraiment atteint qu'en identifiant le centre de cet intervalle. Donc, dans le cas qui nous intéresse, nous calculons $\frac{3 + 4}{2} = 3,5$. Par conséquent, 3,5 est identifié comme étant la médiane (la seule).

Ainsi, deux cas se présentent:

a) <u>La série est composée d'un nombre impair d'observations:</u>

Il suffit alors d'identifier la valeur centrale de la série ordonnée.

$$\left\{ X1 \quad X2 \quad X3 \ldots Xn \right\}$$

MÉDIANE = $X_{\frac{n+1}{2}}$

où n = nombre (impair) d'observations dans la série.

On procède généralement en 3 étapes:

1- Ordonner les nombres;

2- Calculer le rang de la médiane ($\frac{n+1}{2}$)

3- Identifier la valeur numérique correspondant à ce rang.

Par exemple:

$$\left\{ 14 \quad 21 \quad 29 \quad 9 \quad 13 \right\}$$

(ATTENTION: la médiane ne vaut pas 29)

1- La série ordonnée est

$$\left\{ 9 \quad 13 \quad 14 \quad 21 \quad 29 \right\}$$

2- Il y a 5 nombres; donc la médiane

est située au $\frac{(5 + 1)}{2}$ = 3e rang

3- Médiane = X3 = 14

b) <u>La série est composée d'un nombre pair d'observations:</u>

En ce cas, la médiane correspond au point-milieu entre deux valeurs de la série ordonnée:

$$\left\{ X1 \quad X2 \quad X3 \ \ldots \ Xn \right\}$$

$$\boxed{\text{MÉDIANE} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} \quad \text{où n est pair.}}$$

Il faut donc:

1- Ordonner les nombres;

2- Calculer le rang des observations qui serviront au calcul de la médiane:

$$\left(\frac{n}{2} \quad \text{et} \quad \frac{n}{2} + 1 \right)$$

3- Identifier la valeur numérique correspondant à ces rangs:

$$X_{\frac{n}{2}} \quad \text{et} \quad X_{\frac{n}{2}+1}$$

4- Calculer le point-milieu.

Par exemple, pour trouver la médiane de la série

$$\left\{ 4 \quad 24 \quad 20 \quad 3 \quad 1 \quad 17 \quad 29 \quad 39 \quad 41 \quad 17 \quad 4 \quad 16 \right\}.$$

Il faut procéder de la façon suivante:

1- <u>La série ordonnée:</u>

$$\left\{ 1 \quad 3 \quad 4 \quad 4 \quad 16 \quad 17 \quad 17 \quad 20 \quad 24 \quad 29 \quad 39 \quad 41 \right\}$$

<u>Le rang:</u>

2- $\frac{12}{2} = 6 \qquad \frac{12}{2} + 1 = 7$

3- <u>Les nombres:</u>

$X6 = 17$ et $X7 = 17$

$$\text{MÉDIANE} = \frac{X6 + X7}{2} = \frac{17 + 17}{2} = 17$$

REMARQUES:

1- La médiane ne correspond pas toujours à l'un des éléments de la série.

2- Le calcul du point-milieu entre deux éléments, ne résulte pas toujours en un élément étranger à la série.

Dans la plupart des cas, nous aurons besoin de calculer plusieurs paramètres de position pour décrire adéquatement la situation. Généralement, la moyenne sera un excellent paramètre. Lorsque certaines valeurs (isolées) s'éloignent trop des autres, la moyenne perd de sa représentativité tandis que la médiane en gagne (puisqu'elle n'est absolument pas influencée par les valeurs extrêmes). Dans certaines autres situations (proportionnellement plus rares) le mode sera plus représentatif. Enfin rappelons que les paramètres de position mettent en évidence une région de l'axe des Réels qui correspond au centre, au coeur des observations.

3.2 PARAMÈTRES DE DISPERSION

«Si tu te fâches contre l'obscurité, tu finiras par te faire mal aux pieds.»

Proverbe africain

Si les paramètres de position répondent à la question «OÙ», les paramètres de dispersion répondent, quant à eux, à la question du «COMMENT». Après avoir identifié une «zone grise» au coeur des nombres, il nous reste à mesurer les liens qui unissent ces nombres entre eux. Nous le ferons à l'aide de paramètres de dispersion.

A) L'ÉTENDUE

Le premier paramètre qui nous renseigne sur la dispersion des nombres est l'étendue. Nous avons déjà défini celle-ci comme étant la différence entre la plus grande et la plus petite des observations. Elle sert en quelque sorte à limiter l'intervalle qui accueille les autres observations. L'étendue, seule, n'est pas un bon paramètre. Il faut l'associer à d'autres mesures pour en tirer une signification. Par exemple, ce n'est pas parce qu'un tiroir est grand qu'il contient beaucoup d'objets. Mais il peut en contenir beaucoup. Ce dont nous sommes sûrs c'est que lorsque l'étendue est relativement petite, (compte tenu de l'ordre de grandeur des nombres) il n'y a pas de place pour une grande dispersion.

Mais là encore, il est impossible d'évaluer l'ordre de grandeur de l'étendue uniquement à l'aide de la mesure qu'elle fournit. Ainsi, la mesure «10 cm» peut être petite ou grande selon qu'on la compare à des kilomètres ou à des microns (0,000001 m). L'étendue seule est comme une lampe de poche sans pile: elle ne peut nous éclairer!

B) L'ÉCART-MOYEN

Examinons maintenant l'ensemble des observations et tentons d'en mesurer la dispersion, l'étalement, l'éloignement des unes par rapport aux autres. Or, comme il est difficile de prendre des mesures quand tous les éléments varient à tout moment, fixons un point de repère à partir duquel il sera possible de mesurer et donc de comparer des espacements. L'essentiel est de savoir «où» fixer ce point. Mais nous avons déjà répondu à cette question par l'entremise des paramètres de position. Situons donc ce repère au centre des données observées et mesurons l'éloignement à partir de ce centre. La moyenne étant généralement le meilleur paramètre de position, accordons-lui la préférence et inscrivons-la sur un axe (Figure 1).

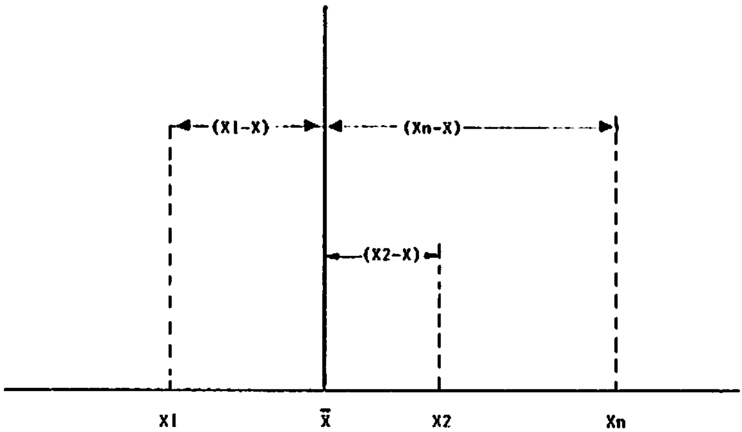


Figure 1

Ensuite, mesurons l'écart qui sépare chaque observation (Xi) de la moyenne (X). Ceci se fait par:

$X1 - \bar{X}$ pour la première observation;

$X2 - \bar{X}$ pour la deuxième observaton;

:

:

$Xi - \bar{X}$ pour la i ième observation;

:

:

$Xn - \bar{X}$ pour la dernière observation;

d'une façon générale, on obtient $(Xi - \bar{X})$.

Chaque mesure représente un espacement entre l'observation et la moyenne. Considérons la somme de ces mesures: $\sum (Xi - \overline{X})$.

REMARQUES:

1- Certaines mesures sont positives (celles dont la valeur numérique est au-delà de la moyenne) et d'autres négatives (celles dont la valeur précède la moyenne). La somme ne peut donc pas représenter l'éloignement total puisque certaines mesures s'annuleraient.

2- Il est même certain que ces valeurs s'annuleront complètement.

3- Il suffit, pour nous en convaincre, de nous rappeler comment la moyenne a été choisie: c'est un point d'équilibre pour lequel la distance totale des poids, mesurée à partir du «pivot», est équivalente de chaque côté de ce point.

Il faut donc agir de façon à ce que ces mesures ne s'annulent plus. L'un des moyens mathématiques que nous pouvons utiliser pour faire disparaître le négatif est d'utiliser la valeur absolue du résultat. Ainsi:

$$| Xi - \overline{X} |$$

représente un éloignement, un écart entre Xi et $\overline{X}$, qui n'est jamais négatif.

Il est donc «possible» d'utiliser $Xi - \bar{X}$ pour mesurer la somme des écarts à la moyenne. C'est cependant un nombre disproportionné par rapport aux observations brutes. Aussi, est-il plus facile de faire le lien avec le phénomène (son ordre de grandeur) en prenant la moyenne de ces écarts: ceci nous permet de définir l'ÉCART-MOYEN par

$$E.M. = \frac{1}{n} \sum \left| Xi - \bar{X} \right|$$

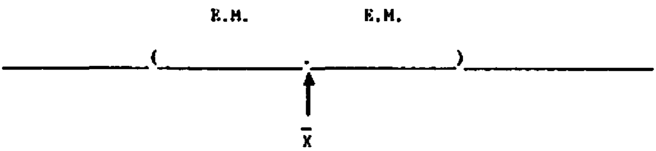
où E.M. représente L'ÉCART-MOYEN;

les Xi sont les observations;

$\bar{X}$ est la moyenne de ces observations;

n est le nombre d'observations.

Ce paramètre correspond à une mesure qui fait image: c'est, en moyenne, l'écartement des nombres de part et d'autre du point d'équilibre ($\bar{X}$).



Plus cet écart-moyen est grand, plus la dispersion est grande (les nombres sont plus étalés, plus isolés les uns des autres). Au contraire, lorsque ce paramètre est petit, les nombres ne sont pas très éloignés les uns des autres, la dispersion est faible, les données sont concentrées.

C) LA VARIANCE

1) OBSERVATIONS BRUTES:

Mais en dépit de ses nombreuses qualités, ce magnifique paramètre qu'est l'écart-moyen est rarement utilisé. La raison en est fort simple: la présence de la valeur absolue rend son utilisation difficile. Sa manipulation exige des précautions dont le statisticien refuse de s'embarrasser à ce stade. Il convient donc de revenir au problème rencontré avant d'introduire la valeur absolue. L'objectif était de faire disparaître le signe négatif de l'expression « $Xi - \bar{X}$ ». Or la valeur absolue n'est pas le seul outil mathématique qui possède ce «mystérieux pouvoir». On peut atteindre le même objectif en élevant au carré. En effet, l'expression $(Xi - \bar{X})^2$ n'est jamais négative. Considérons alors la somme des carrés: $(Xi - \bar{X})^2$. C'est encore une fois une mesure disproportionnée. Prenons-en la moyenne:

$$\frac{1}{n} \sum (Xi - \bar{X})^2 .$$

N.B.: Ce résultat représente effectivement la moyenne du carré de l'écart entre les observations et leur moyenne, puisque par définition de la moyenne nous connaissons:

- la somme de tous les carrés: $\sum (Xi - \bar{X})^2$
- le nombre total de carrés: $\sum n$.

Pour plus de commodité nous nommerons ce paramètre: VARIANCE.

Notons que les unités de la variance sont le carré des unités originales. De plus, même si la variance est moins disproportionnée que la somme des carrés, elle n'établit pas encore une parfaite correspondance. Cependant ce paramètre présente suffisamment de qualités pour nous inciter à l'utiliser.

Mais, auparavant, ouvrons une longue parenthèse. Les observations recueillies par un statisticien peuvent provenir de deux sources:

a) d'une population:

C'est l'ensemble de tous les éléments (personnes, objets ou nombres) concernés par le phénomène étudié. Par exemple, dans l'étude des salaires des travailleurs de «General Motors», la population est constituée par l'ensemble de tous les salariés travaillant chez «General Motors».

b) d'un échantillon:

C'est une partie seulement d'une population. Par exemple, les étudiants qui suivent un cours de statistiques constituent un échantillon de tous les étudiants.

N.B.: Un groupe d'éléments ne constitue pas en soi une population ou un échantillon. Cela dépend toujours de notre point d'intérêt. D'ailleurs, un même groupe peut constituer une population ou un échantillon selon le point de vue adopté. Par exemple, le groupe de travailleurs de «General Motors» qui constituaient tantôt une population deviendra un échantillon dans l'étude des salaires des travailleurs de l'automobile.

Ainsi, lorsque nos observations proviennent d'une population et que nous calculons des paramètres, nous connaissons exactement leur valeur numérique. Cependant, lorsque nos observations proviennent d'un échantillon, nous ne connaissons avec exactitude que les paramètres de notre échantillon.

Or, notre objectif est toujours de connaître la population. Si nous procédons à partir d'échantillons c'est souvent à cause des contraintes de temps ou d'argent. Parfois, aussi, c'est la seule façon de procéder. Par exemple, on imagine mal une industrie qui examinerait toute sa production pour fins de contrôle de la qualité. Cela est d'autant plus inimaginable que la plupart des éléments qui servent au contrôle de la qualité font l'objet d'expertises visant à les détruire pour en connaître la valeur! (Câble d'acier qu'on étire jusqu'à ce qu'il se rompe, ampoule électrique qu'on allume jusqu'à ce qu'elle cesse d'éclairer, téléviseur qu'on branche jusqu'à ce qu'il soit défectueux, soulier de cuir qu'on plie jusqu'à ce qu'il fende, balles de calibre 22 qu'on tire pour en étudier la trajectoire, etc.).

Notre objectif demeure toujours celui de connaître les paramètres de la population. Si les paramètres d'un échantillon sont connus, ils peuvent servir, à certaines conditions, à évaluer ceux de la population. Cette opération s'appelle faire de l'ESTIMATION. Or, lorsque la variance, telle que définie par $\frac{1}{n}\sum(Xi - \bar{X})^2$, sert à estimer la variance de la population, elle ne répond pas à tous les critères d'un bon estimateur. Entre autres, elle ne prend pas, en moyenne, d'un échantillon à l'autre, la même valeur que la variance de la population. Les statisticiens disent alors qu'elle possède un «BIAIS». Pour corriger cette «anomalie», ils utilisent «n - 1 » au lieu de «n» dans la formule. (Ce n'est pas l'effet du hasard, mais bien le résultat de «savants» calculs).

Ainsi $\frac{1}{n-1}\sum(Xi - \bar{X})^2$ ne possède pas le biais dont nous avons parlé. Résumons:

(1) $S^2 = \frac{1}{n} (Xi - \bar{X})^2$ est la VARIANCE d'une population

(2) $s^2 = \frac{1}{n-1} (Xi - \bar{X})^2$ est la VARIANCE d'un échantillon

où les Xi sont les observations,

$\bar{X}$ leur moyenne,

et n le nombre d'observations.

N.B.: S^2 et s^2 sont les symboles utilisés pour représenter la variance. Le «carré» est facile à justifier (et à retenir) puisqu'il provient d'une somme de carrés. Quant à l'utilisation de la majuscule ou de la minuscule, la première sert à mesurer une population et la deuxième mesure un échantillon. On associe donc un «grand» symbole à une population et un «petit» symbole à un échantillon.

Nous pouvons noter (s'il est permis de faire un peu de mathématiques!) que:

$$\sum (Xi - \bar{X})^2 = nS^2 \quad \text{(tirée de (1))}$$

et $$\sum (Xi - \bar{X})^2 = (n-1)\ s^2 \quad \text{(tirée de (2))}$$

donc: $nS^2 = (n-1)\ s^2$

ou (3) $$S^2 = \frac{(n-1)}{n}\ s^2$$

En pratique, la formule la plus utilisée, est la (2) avec «s» puisqu'il est rare de connaître toute une population. On utilisera donc la formule avec (n-1), à moins qu'il ne soit précisé (ou que ce soit clair) que nous avons affaire à une population. Une bonne calculatrice donne cependant accès aux deux calculs.

II) OBSERVATIONS REGROUPÉES:

Lorsque les observations ont été regroupées et que nous n'avons plus accès à l'ensemble des données brutes, il faut quand même trouver le moyen de calculer la variance. Il est à noter cependant qu'avec l'apparition des calculatrices à fonctions statistiques cette situation est de plus en plus rare.

Rappelons-nous l'hypothèse que nous avons dû avancer pour calculer les paramètres de position à partir des renseignements fournis par la distribution de fréquences: chaque élément peut se voir attribuer la valeur du centre de la classe. Nous pouvons ainsi, sous cette hypothèse, calculer la variance comme si nous connaissions toutes les observations.

Classes	Fréquences
Li1 à Ls1	f1
Li2 à Ls2	f2
Li3 à Ls3	f3
:	:
:	:
:	:
Lik à Lsk	fk

où les Li sont les limites inférieures,

les Ls sont les limites supérieures.

Appelons C_i le centre de la ième classe et calculons

$$\bar{X} = \frac{1}{n} \sum C_i \cdot f_i \qquad \text{(tel que déjà vu.)}$$

Ensuite, si nous voulons calculer la VARIANCE, il faudra calculer $(X_i - \bar{X})^2$. Mais $(X_i - \bar{X})^2$ vaut $(C_i - \bar{X})^2$. Cette quantité $(C_i - \bar{X})^2$ vaut $(C_1 - \bar{X})^2$ pour tous les nombres appartenant à la première classe. Or, il y en a f_1. Donc, ce carré $(C_1 - \bar{X})^2$ apparaît f_1 fois; en les additionnant on obtient: $(C_1 - \bar{X})^2 + (C_1 - \bar{X})^2 + \ldots\ldots + (C_1 - \bar{X})^2$ ce qui représente $(C_1 - \bar{X})^2 \cdot f_1$ pour la contribution des éléments de la première classe à la somme des carrés. La contribution de la deuxième classe sera donc de $(C_2 - \bar{X})^2 \cdot f_2$ de telle sorte que:

Classe	Centre	Fréquence	Contribution à la somme des carrés
1	C_1	f_1	$(C_1 - \bar{X})^2 \cdot f_1$
2	C_2	f_2	$(C_2 - \bar{X})^2 \cdot f_2$
3	C_3	f_3	$(C_3 - \bar{X})^2 \cdot f_3$
:	:	:	:
:	:	:	:
:	:	:	
i	C_i	f_i	$(C_i - \bar{X})^2 \cdot f_i$
:	:	:	:
k	C_k	f_k	$(C_k - \bar{X})^2 \cdot f_k$

Ainsi nous pouvons définir:

$$s^2 = \frac{1}{n-1} \sum (Ci - \bar{X})^2 . fi$$

comme étant <u>la VARIANCE des observations regroupées</u>.

où Ci est le centre de la i ième classe,

fi est la fréquence de la i ième classe,

$\bar{X}$ est la moyenne des observations regroupées,

et $n = \sum fi$

N.B.: Une définition analogue serait obtenue pour S^2 en remplaçant (n-1) par n, mais cette valeur numérique (S^2) ne pourra jamais représenter la variance (exacte) de la population étant donné que ce processus nous oblige à recourir à des valeurs hypothétiques et que, sans donner des résultats très différents, il nous amène quand même à commettre une erreur; donc S^2 ne pourra jamais correspondre exactement à la VARIANCE de la population. C'est pourquoi un livre sérieux de statistiques descriptives ne donne que la définition de s^2 dans le cas des données regroupées. C'est ce que nous allons faire.

D) L'ECART-TYPE

Bien que la variance nous renseigne adéquatement sur la dispersion, elle incommode au plan des unités. En effet, si les observations sont en dollars ($), la variance est donnée en dollars carrés ($\$^2$); si elles sont en cm, la variance est en cm^2; si les observations portent sur un nombre d'enfants, la variance est en «enfants carrés» ce qui n'est pas, vous en conviendrez, trop commode (surtout lors de l'accouchement...!).

Afin de revenir à des unités plus «commodes», utilisons la racine carrée de la variance et donnons-lui le nom d'ECART-TYPE.

$S = \sqrt{S^2}$ (identité algébrique qu'on pourrait qualifier de «Tautologie»!)

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

où s et S sont les symboles de l'Ecart-type.

REMARQUE: L'écart-type s'obtient toujours après la variance puisqu'il en est la racine carrée (sa partie positive).

Examinons l'exemple suivant:

Classe	Fréquence	Centre	Ci. fi	$(Ci - \bar{X})$	$(Ci - \bar{X})$ fi
0 à 2	2	1	2	36	72
3 à 5	1	4	4	9	9
6 à 8	2	7	14	0	0
9 à 11	5	10	50	9	45
	10		$\frac{70}{10} = 7$		126

$$s^2 = \frac{1}{n-1} \sum (Ci - X)^2 \cdot fi$$

$$s^2 = \frac{1}{9} \cdot (126) = 14$$

$$\text{et } s = \sqrt{s^2} = \sqrt{14} = 3{,}74$$

E) LE COEFFICIENT DE VARIATION

Bien qu'on connaisse la valeur numérique d'un écart-type, il est parfois difficile de dire si cela correspond à une grande ou à une petite dispersion. Il existe cependant un paramètre qui sert à porter un tel jugement; il s'appelle <u>LE COEFFICIENT DE VARIATION</u> (C.V.):

$$\boxed{C.V. = \frac{s}{\bar{X}} \cdot 100 \text{ et est exprimé en } \%.}$$

Il a l'avantage de relier l'écart-type à l'ordre de grandeur des nombres en présence (puisque la moyenne est un paramètre de position).

Par exemple si nous avons:

$s = 4$ cm et $\bar{X} = 10$ cm.

nous obtenons: $C.V. = \left(\frac{4}{10} \cdot 100\right) \% = 40 \%$

alors que si $s = 4$ et $\bar{X} = 100$ cm.

nous obtenons: $C.V. = \left(\frac{4}{100} \cdot 100\right) \% = 4 \%$

Donc, pour un même écart-type, nous obtenons un coefficient de variation différent.

REMARQUE: En présence d'une population, nous pourrions calculer le coefficient de variation de la population:

$$C.V. = \frac{\text{Ecart-type}}{\text{moyenne}} \cdot 100.\%$$

Notons que le symbole $\bar{X}$ est réservé à la moyenne d'un échantillon. Le symbole qui le remplace dans une population est « μ » (prononcer mû)

Par ailleurs, le symbole utilisé pour représenter l'écart-type d'une population est σ.

Généralement, un coefficient inférieur à 15% indique une petite variation alors que plus le coefficient s'écarte de 15% plus la variation est grande.