1
Mathématiques et à peu près

G. Th. GUILBAUD
(Ecole Pratique des Hautes Etudes, Paris)

PREMIERE PARTIE:
Y a-t-il une mathématique de l'à peu près ?

A peu près
Cueillir des mots
Anthologie
Pifométrie
Tolérance
Topologies
Méprises
Confiance
Aléa

DEUXIEME PARTIE:
Mathématique pour (ou à propos de)
la Statistique Descriptive

Trier ou classer
Compter
Normer
Mesures
La cumulative
Exercices
Quantilage
Moyennes et moments
Tchebychev
Les lois
Pareto

PREMIERE PARTIE : Y a-t-il une mathématique de l'à peu près ?

À PEU PRÈS

Dans la Préface de son "Calcul Infinitésimal" (Paris, Hermann, Collection Méthodes, 1968), Jean DIEUDONNE dénonce vigoureusement ceux qui

> *"permettent (ou même encouragent) le laisser-aller, le vague et l'à peu près".*

Tout le monde, et pas seulement les spécialistes, lui accordera qu'il n'y a pas de place, en mathématiques, pour l'à peu près.

Mais, à la page précédente, le même Auteur nous dit:

> *"pour acquérir le sens de l'Analyse... il faut avoir appris à distinguer ce qui est grand de ce qui est petit, ce qui est prépondérant de ce qui est négligeable.*
> *En d'autre termes, le Calcul Infinitésimal... est l'apprentissage du maniement des inégalités bien plus que des égalités, et on pourrait le résumer en trois mots:*
> *MAJORER , MINORER , APPROCHER . "*

Si l'on peut dire qu'il n'y a pas de place pour le vague et l'à peu près dans la pratique mathématicienne, il faut aussi affirmer que la tâche principale de la mathématique est d'apprendre à parler (si l'on ose risquer cette alliance de mots), à parler avec rigueur des approximations.

Je vous propose quelques réflexions sur ce thème: partir de l'à peu près qui est partout présent dans notre vie et notre langage, et comprendre comment les constructions mathématiques relèvent le défi, l'apparente contradiction entre rigueur et approximation.

CUEILLIR DES MOTS

C'est un exercice utile, pour commencer, que de repérer, dans nos manières de parler, ou dans des textes écrits, divers emplois de la modalité:

"presque" ou "à peu près"

On note évidemment : d'abord la variété considérable des expressions.

Voici un début, en vrac :

j'en ai pour cinq minutes / attendez-moi deux secondes / je reviens de suite / des prix très étudiés / à quatre pas d'ici / c'est tout à fait imminent / pratiquement pas / pour ainsi dire cent pour cent / moins que rien / largement cent mille / dans cent sept ans / huit jours facile ! / cents façons de dire / ça va chercher dans les quinze cents / la moyenne se situe aux environs de / ça nous entraînerait un peu loin / malgré quelques résistances isolées / dans la grande majorité des cas / une toute petite minorité / compter sur les doigts d'une seule main / il n'y a pas des kilomètres / au bas mot / c'est rien à côté d'hier / les quatre cents coups / il habite au diable / il s'en est fallu d'un cheveu / à peu près grand comme ça, enfin, pas tout à fait, mais, quand même... / infiniment supérieur aux besoins / à vue de nez / deux ou trois, pas plus / les jeunes de quinze - seize ans / pas épais / pas tellement / c'est quand même rarissime /

(trente six façons de dire : presque)

On peut aussi, et c'est plus intéressant, noter des contextes, des phrases, ou même des textes plus longs.

On réfléchira à ceci: il est "*à peu près impossible*" de parler, sans faire appel à quelques-uns de ces signes qui marquent pour notre interlocuteur qu'il ne faut pas prendre à la lettre ce que nous disons; car la vie elle-même est remplie de vague et de flou.

A titre d'échantillon, voici une douzaine de citations, prises un peu n'importe comment (je ne dis pas au hasard, mais je dis: sans malice) dans un plus grand stock.

ANTHOLOGIE

(R: radio, TV: télévision, J: journal, P: publicité,
B: entendu au bistrot, ...)

... n'importe quel sondage admet une marge d'erreur d'au moins un point ... (R)

... nous vous offrons gratuitement deux cents recettes à préparer vous-même, en un clin d'oeil et pour trois fois rien ... (P)

... une foule innombrable: à combien se chiffrait-elle quand s'ouvrit le meeting ? à cent mille ? davantage ! Il était difficile de l'évaluer. Ce qui est certain, c'est que l'immense hall fut rapidement comble et que des milliers, des dizaines de milliers de personnes sans doute, ne purent y pénétrer ... (J)

... et là, vous avez un câble, à peu près gros comme mon pouce, enfin ... pas tout à fait, mais quand même plutôt costaud ... (B)

... Mais combien ça coûte, une vache ? (Le Marchand de bestiaux): Oh ! là ! là ! mon pauvre, sur mille vaches, il y a mille prix !—Mais dites un prix moyen.—Une vache ? ça peut valoir cinq cent ou mille ou davantage ... (R)

... Les températures seront en légère hausse, il fera assez beau sur le Massif Central ... (R)

... il est impossible de citer les quelque soixante titres, publiés dans différentes maisons d'édition, qui jalonnent une longue carrière littéraire; voici cependant les principaux ouvrages ... (J)

... On peut dire qu'en France, l'écart des revenus est excessif; l'objectif majeur, au cours des prochaines années, doit être de resserrer cet écart. En particulier il faudrait revaloriser le Smic plus rapidement que la moyenne des salaires ... (TV)

... si on prend la moyenne des chiffres, ceux des organisateurs et ceux de la police, on peut estimer à vingt mille le nombre des manifestants ... (R)

... il y a douze candidats; on avait cru, un moment, qu'il y en aurait vingt-cinq ou trente: si bien que douze paraît un petit chiffre; c'est quand même beaucoup ... (J)

... cinq ou six explosions retentirent avant l'arrivée des pompiers, la pharmacie voisine fut partiellement détruite, ses stocks endommagés à 80% ... (J)

PIFOMETRIE

Il ne faudrait pas s'imaginer que l'on abandonne le langage approximatif quand on aborde le discours scientifique.

Voici (encore à titre d'échantillon sans prétention) ce que dit un géographe :

"la population du Dahomey est estimée à 2 100 000 habitants dont 50% au moins, soit sensiblement plus d'un million de personnes, sont rassemblées le long de quelque cent kilomètres de côtes, sur une profondeur de 90 à 100 km, alors que la seconde moitié de la population dispose d'un territoire immense d'une superficie supérieure à cent mille km^2 ...

le domaine Ouemenou couvre environ un millier de km^2 sur lequel vit une population de l'ordre de cent cinquante mille habitants... le total annuel moyen des chutes de pluie est remarquable par sa faiblesse puisqu'il se situe entre 1 100 et 1 200 mm ... "

Et voici un astronome:

"l'orbite d'un satellite artificiel est, en première approximation, une ellipse; les orbites sont liées entre elles par la troisième loi de Kepler, ici pratiquement rigoureuse, les masses des satellites étant négligeables devant celle du corps central ..."

Si personne ne devait s'en formaliser (mais les gens sont susceptibles !) j'aimerais désigner ce premier niveau du discours scientifique, encore très proche du langage de tous les jours, par l'étiquette bien connue:

le Pifomètre.

Ici on "arrondit", on évite de "couper les cheveux en quatre", on dit "égale environ tant" et on l'écrit:

$\pi \simeq 3,14$ ou $3,1416$

TOLERANCE

La tolérance dont je veux parler, c'est celle des fabrications mécaniques:
Ajustement des pièces lisses interchangeables (système I.S.O.)
Diamètre nominal de 30 à 50 :

Alésage	H7	min = 0 ,	max = 25
Arbre	libre (e8)	min = - 89 ,	max = - 50
	tournant (f7)	- 50 ,	- 25
	glissant (g6)	- 25 ,	- 9
	glissant juste (h6)	- 16 ,	0
	légèrement dur (j6)	- 5 ,	+ 11
	bloqué (m6)	+ 9 ,	+ 25

(série simplifiée des tolérances usuelles exprimées en microns)

Il semble que la première apparition de cette façon de parler, soit, bien avant le développement de l'industrie moderne et de ses exigences de standardisation, les décisions concernant les monnaies d'or: le poids est imposé par l'autorité, mais on accepte une petite différence appelée "remedium".

Pour l'enseignement mathématique d'aujourd'hui c'est

ENCADRER

(c'est-à-dire à la fois majorer et minorer)

$$3,1415 < \pi < 3,1416$$

Après le pifomètre, c'est là le second niveau de langage dans la conquête de l'à peu près: on accepte l'incertitude, et on en donne une expression précise.

Il convient d'apprendre à encadrer, et surtout à calculer avec des données qui ne sont données que par encadrement (quand j'étais petit, on appelait cela: le calcul d'erreurs, expression un peu bizarre, puisqu'il n'y a pas d'erreur, il n'y a que des incertitudes).

TOPOLOGIES

Il y a un troisième niveau, plus spécifiquement mathématique. C'est la grande affaire, ce pourrait (presque !) être le fil conducteur pour une histoire des mathématiques.

Sans vouloir faire vraiment de l'histoire, célébrons au moins le célèbre ARCHIMEDE:

Il a écrit un petit traité "sur la mesure du cercle" (on peut aller voir ce que c'est, ça vaut la peine, dans: ITARD et DEDRON, Mathématiques et mathématiciens, Paris, Magnard, 1959, pages 406 et suivantes. Lecture très recommandée: trois étoiles).

Non seulement Archimède encadre le nombre que nous appelons maintenant π :

moins que 3 et 1/7

plus que 3 et 10/71

mais en même temps il nous fournit la méthode pour faire mieux.

Car on approxime le cercle par des polygones, successivement de 6 , 12 , 24 , 48 , 96 côtés. On pourrait continuer.

On mettra des siècles à bien voir ce qui s'est passé. Aujourd'hui toute cette aventure est intégrée à notre mentalité; il arrive qu"on ne s'en étonne plus.

Pour le dire en peu de mots:

premier niveau: à "peu" près

second: à "tant" près

troisième: à "aussi peu" près que vous voudrez.

Il fallait codifier les algorithmes (illimités): environ deux mille ans après, Leibniz sera fier d'exhiber cette autre "quadrature" :

$$\frac{\pi}{8} = \frac{1}{1\times3} + \frac{1}{5\times7} + \frac{1}{9\times11} + \frac{1}{13\times15} + \text{etc.}$$

(tout est dit, ou plutôt sous-entendu, au moyen d'un "etc" !)
C'était il y a trois cents ans.

Séries, Produits infinis, Fractions continues, Intégrales, tout cela si considérable qu'on risque d'oublier qu'il s'agit toujours et partout d' "approximations".

Savoir définir et manier proprement la notion de limite, ce n'est pas rien; aujourd'hui ça s'appelle: Topologie Générale.

MÉPRISES

Pour encadrer il faut donner un intervalle; on peut fournir les deux bornes:

$$a < x < b$$

Mais on peut aussi, et c'est souvent plus commode, indiquer le point milieu:

$$c = (a + b) / 2$$

et compléter l'information par l'amplitude: $2h = b - a$.

Alors, au lieu de dire que

x est compris entre a et b ,

on dira que

x est "égal à c, à plus ou moins h près" .

On a souvent alors pas de scrupules à écrire:

$$x = c \ (\pm h)$$

ce qui veut dire :

c plus ou moins quelque chose pas plus grand que h .

Il n'y aurait aucun inconvénient à instituer de telles conventions d'écriture, si la logique de l'encadrement (ou de la Tolérance) était la seule logique de l'approximation.
Mais ce n'est pas le cas. Et les méprises sont à craindre.

Les renseignements qui suivent sont tirés de l'Annuaire du Bureau des Longitudes pour l'An 1975.
Vers 1940, on estimait que la vitesse de la lumière (dans le vide) était:

$$c = 299\ 774\ (\pm 5) \quad \text{km par seconde.}$$

Des déterminations plus récentes et concordantes ont amené l'Union Radio Scientifique Internationale (U.R.S.I., assemblée générale de 1957) à adopter la valeur

$$c = 299\ 792{,}5\ (\pm 0{,}4) \quad \text{km/s}$$

Depuis 1970, de grands progrès ont été faits: Emerson a obtenu :

$$c = 299\ 792\ 456{,}2\ (\pm 1{,}1) \quad \text{m/s}$$

Nous ne sommes plus, malgré les apparences, dans la logique de l'encadrement: on ne cherche pas à nous assurer que le nombre inconnu est sûrement situé entre telle et telle borne. On nous donne une valeur (officielle et provisoire) et on nous dit l' "erreur à craindre". Laquelle résulte d'estimations assez complexes (moyennes, variances, écart-types, etc.)

C'est que, deux mille ans après Archimède, des hommes qui le connaissaient bien et l'admiraient beaucoup, Pascal et Fermat, ont inventé une nouvelle manière d'affronter l'incertitude: le Calcul des Probabilités.

CONFIANCE

Il ne faut pas confondre les intervalles de tolérance (j'affirme que tel nombre est entre tant et tant) et les intervalles de confiance (il y a tant de chances que le nombre soit dans tel intervalle, mais je n'exclus pas qu'il puisse être en dehors).

Remarque : naguère encore, tant qu'elle était réservée aux artilleurs, la "fourchette" était un intervalle de confiance; depuis que les sondeurs s'en sont emparés, le sens de ce terme est beaucoup moins clair: il faut craindre la méprise, et que le bon public croie à la certitude.

Nous sommes ici au quatrième niveau; et si tout le monde admet que l'apprentissage du langage du troisième niveau (limites, séries, intégrales, et la suite) demande un effort, il ne faut pas dissimuler ce qui nous attend.

Par quoi commencer ? Peut-être pas par jouer aux dés et se plonger dans les délicieux exercices de combinatoire. Mais plutôt par une notion centrale: l'aléa.

Il s'agira bien entendu, pour commencer, des aléas numériques (d'aucuns disent: variable aléatoire).

ALEA

Soit à dire quelque chose d'un nombre (réel ou entier, par exemple) qui n'est pas (encore) complètement connu.

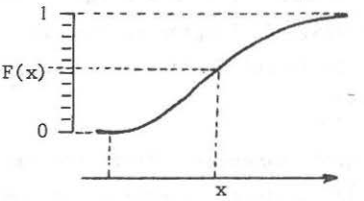
Premier niveau: on donne une valeur approchée, ou comme on dit un "ordre de grandeur".

Second niveau: on donne un intervalle qui encadre;

Troisième niveau: on fournit un algorithme qui donne une précision arbitraire;

Quatrième niveau enfin: pour chaque intervalle possible on dit la probabilité que le nombre y soit.

Une image commode :



F(x) = probabilité que le nombre inconnu soit inférieur à x.

F(b) - F(a) = probabilité qu'il soit inférieur à b et non supérieur à a, c'est-à-dire dans: $[a \leqslant \ldots < b[$

C'est un objet mathématique connu: une mesure sur une partie de la droite réelle.

Ce n'est pas une question très élémentaire: elle ne l'est pas en effet, dans l'état actuel de notre enseignement. Mais la droite réelle non plus, et ça n'empêche pas de chercher des moyens convenables d'y introduire les débutants.

D'où ma proposition: commencer par la statistique descriptive qui peut être une bonne occasion de didactique mathématique.

Et c'est ce que signifie ma seconde partie.

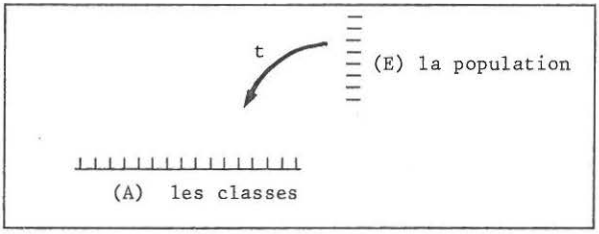
DEUXIEME PARTIE : Mathématique pour (ou à propos de) la Statistique Descriptive

---

TRIER ou CLASSER

Soit l'intitulé: "application d'un ensemble dans un autre ensemble"; s'il s'agit de parler à des débutants, la coutume est d'illustrer le discours par des dessins: graphes fléchés, graphes cartésiens (ces deux sortes de diagrammes conviennent aussi à toute autre relation, pas forcément fonctionnelle).

Il n'est pas mauvais non plus de penser au "TRI": un sac de courrier, un paquet de fiches (ou ce que vous voudrez imaginer), sera l'ensemble de départ. A l'arrivée un ensemble de boîtes ou de casiers.



Un tel tri peut représenter une application

$$t : E \longrightarrow A$$

On peut trier selon l'âge, ou le sexe, ou la catégorie socio-professionnelle, ou le salaire, etc. ou des objets selon le poids, la taille, la couleur, la nature du métal, etc. . Les exemples, réalistes ou non, ne manquent pas.

Caractère fonctionnel: dans chaque cas, chaque élément de E a <u>son</u> affectation (sa destination, son adresse): chaque chose a sa place, comme dit le proverbe.

<u>Rappel</u>

L'ensemble des applications de E dans A est noté A^E (notation exponentielle) et l'on a :

$$\text{card}\,(A^E) = (\text{card}\,A)^{(\text{card}\,E)}$$

(première formule utile de la combinatoire, première rencontre avec la fonction exponentielle) .

Si le compte rendu du Tri se fait par écrit, c'est un "état" (en langage administratif): c'est une liste nominative.
Comparer avec la présentation usuelle des Tables de fonctions numériques.

Pour le travail statistique, on va passer à l'anonymat.

Remarque :

Il y a une autre façon de classer, celle qui connaît seulement une relation d'équivalence, et qui aboutit à une partition de l'ensemble de départ.
On pourrait en parler: mais c'est un autre sujet.

COMPTER

Reprenons l'image: ensemble de départ ou ensemble des objets à ranger (E), ensemble d'arrivée ou ensemble des cases où l'on range (A), une application ($E \longrightarrow A$) ou façon de ranger.

Supposons le rangement effectué (par exemple à partir d'une enquête): on passe alors en revue les cases et l'on examine le contenu de chacune. On commence par compter le nombre de fiches que contient chaque case. Première démarche statistique.

Remarque :

Dans l'enseignement mathématique élémentaire, on se contente parfois d'une première étape: on n'utilise pas le nombre à proprement parler, mais seulement les deux catégories logiques:

- au moins un (un ou davantage)
- pas plus d'un (un ou zéro)

ce qui fait trois classes:

cases vides
cases avec un seul objet
cases avec plusieurs objets.

Et l'on expose la nomenclature classique:

"Injection . Surjection . Bijection" !

Voilà ce qu'on fait quand on ne sait pas encore compter (mais seulement distinguer: un, rien, ou plusieurs).

Quand on sait compter, pour chaque case, on dira

l ' EFFECTIF

(ou Cardinal, ou Fréquence - mais le terme de fréquence est ambigu: fréquence relative, fréquence absolue - d'où la préférence de la plupart des statisticiens pour le mot: effectif, d'origine militaire).

On passe ainsi de

t : $E \longrightarrow A$ (c'est le tri)

à d : $A \longrightarrow \underset{\sim}{N}$ (c'est la Distribution statistique)

($\underset{\sim}{N}$ désigne l'ensemble des Naturels)

Ce passage, (transformation ou foncteur) :

$$St : (t : E \longrightarrow A) \longrightarrow (d : A \longrightarrow \underset{\sim}{N})$$

est très important.

On peut s'amuser un instant :

$$d(a) = \text{card } [t^{-1}(a)]$$

$$d... = \text{card } t^{-1} ...$$

$$St(t) = \text{Card } (t^{-1})$$

Rappel :

Si les effectifs des cases sont

$$n_a , n_b , n_c , \ldots , n_k$$

avec

$$A = \{ a, b, c, \ldots, k\}$$

le nombre des applications qui correspondent à cette distribution est :

$$(\Sigma n_i)! / \Pi (n_i!)$$

Seconde formule utile de la combinatoire, où l'on rencontre la fonction factorielle. Avec la factorielle et l'exponentielle, on va déjà assez loin.

NORMER

La Fréquence (certains disent: fréquence relative, d'autres disent: fréquence, tout court) est le quotient de l'effectif d'une classe par l'effectif total. Il s'agit d'une procédure très courante:

Classes	Effectifs	Fréquences
A	1 721	0,93 %
B	2 413	1,29 %
C	4 342	2,34 %
...	...	...
G	33 251	17,88 %
H	29 211	15,71 %
...	...	...
Total	185 951	100 %

Deux remarques :

1/ La distribution en fréquence n'est plus une application dans $\underset{\sim}{N}$, ensemble des naturels, mais dans $\underset{\sim}{Q}$, ensemble des rationnels. Et l'on sera souvent conduit à "arrondir": ci-dessus, par exemple, on a deux distributions, que l'on souhaitait homothétiques, l'une avec un total de 185 951 , l'autre avec 10 000 . Il n'y a pas de solution exacte. Quelle est la meilleure approximation ? C'est le problème, bien connu par ses implications politiques, de la "Représentation proportionnelle", ou: "homothéties approchées de deux distributions en nombres entiers". Beaucoup trop à dire pour que je m'y arrête (mais je regrette).

2/ Si deux distributions sont homothétiques, elles ne sont pas "équivalentes" en toutes circonstances. La pratique effrénée des "sondages" d'opinion (toujours libellés en pourcentages) risque ici d'induire en tentation. Prudence !

MESURES

Si l'on connaît l'effectif (ou la fréquence) pour chacun des éléments

$$a \ , \ b \ , \ c \ , \ \ldots \ , \ k$$

de l'ensemble A, on le connaîtra aussi pour toute partie de A (par addition).

C'est une extension, naturelle et utile, de l'application

$$d : A \longrightarrow \underset{\sim}{N}$$

à une application

$$f : 2^A \longrightarrow \underset{\sim}{N}$$

(2^A désigne l'ensemble des parties de l'ensemble A).

Désignons par f cette application:

f(X) = effectif (ou fréquence) de X, c'est-à-dire
cardinal de l'ensemble des éléments de E qui vont en X

On n'oubliera pas alors:

$$f(P) + f(Q) = f(P \cap Q) + f(P \cup Q)$$

$$f(\phi) = 0$$

$$f(P) \geqslant 0$$

relations caractéristiques d'une "mesure" (au sens des mathématiciens).

LA CUMULATIVE

Quand l'ensemble A est ordonné (ou ordonnable) et qu'il risque d'être infini, on emploie une procédure de description particulière. C'est le cas, par exemple, de l'âge, du salaire, etc. Disons pour couvrir toute sortes de circonstances qu'on trie des objets selon leur "taille": et on va les ranger "par rang de taille".

On donnera cette information:

$F(x)$ = Nombre des individus dont la taille est inférieure à x

On en déduit:

$\left[\text{Nombre des individus dont la taille est comprise dans l'intervalle } (a \leqslant x < b)\right] = F(b) - F(a)$

Petite chicane traditionnelle (voir: Calot, cours de statistique descriptive, Paris, Dunod, page 21) :

On pourrait définir aussi bien par

" dont la taille n'est pas supérieure à x "

ou bien

" dont la taille est supérieure ... "

ou enfin

" dont la taille n'est pas inférieure ... " .

Quatre variantes: le choix est conventionnel.

La figuration graphique d'une telle fonction $F(x)$ est ce qu'on nomme la (courbe) "cumulative". On ne soulignera jamais assez son importance, pratique et théorique.

Dans les usages statistiques, ce sera toujours une fonction en escalier (on néglige trop les fonctions en escalier, ce sont de bien braves fonctions, bien serviables).

Bien entendu, on peut dessiner la courbe des effectifs cumulés, mais aussi celle des fréquences cumulées (c'est la même, seule l'échelle change).

EXERCICES

Classer par rang de taille est un exercice profitable, s'il s'accompagne de réflexions adéquates.

On a l'embarras du choix: il suffit de bien définir la population E soumise à l'enquête, et le caractère x, ou taille, qu'on veut étudier.

Pêle-mêle :

- *ranger des choses périssables selon leur durée de vie (des animaux, des hommes ou des produits industriels)*
- *des textes selon leur longueur (nombre des mots, ou des phrases, ou des lettres)*
- *des entreprises selon le chiffre d'affaires*
- *des agglomérations urbaines selon leur population*
- *des boîtes d'allumettes de la SEITA selon le nombre d'allumettes*
- *des pays selon leur superficie, ou leur population*
- *des fleuves selon leur longueur*
- *des pommes de terre selon leur poids*

etc... etc... etc...

Faire l'enquête, dépouiller, présenter des tableaux clairs, et faire des graphiques.

Même sans dessiner, on peut constituer une représentation matérielle de la courbe cumulative: si le caractère est figuré par une longueur et chaque individu de la population par un bâton, on les rangera côte à côte:

> *" Je me souviens encore parfaitement du jour où, enfant, je partis avec mon père ramasser cent feuilles de saule choisies au hasard; après avoir éliminé les feuilles abîmées, nous pûmes en rapporter quatre vingt neuf intactes, que nous ordonnâmes par grandeur décroissante. Mon père traça la courbe passant par les sommets et me dit: voici la courbe de Quételet, qui te permet de voir que les individus moyens sont les plus nombreux".*
>
> (B.L. Van der Waerden, Statistique mathématique, Paris, Dunod, 1967, page 66).

QUANTILAGE

Je pense qu'il n'est pas mauvais de se constituer un herbier de distributions statistiques représentées par leur Cumulative. Ne serait-ce que pour apprendre à les décrire, et à les comparer.

Ce qu'on recherche dans une méthode de description, c'est de donner les informations dans un ordre raisonnable: d'abord les grandes lignes, les grosses masses, l'allure générale, puis des détails. C'est l'idée des "approximations successives": on doit décrire, mais on peut être interrompu à tout moment; il faut commencer par le plus important; si on a le temps, on précisera. Facile à dire; mais il n'y a pas de solution universelle.

Parmi les procédures les plus employées et les moins savantes, il convient de faire une place à la procédure de "quantilage" qui consiste à donner des "quantiles". Un quantile, c'est tout simplement un point de la cumulative.

Prenons une table de survie (par exemple, Annuaire du Bureau des Longitudes, 1975, page 692). Nombre de survivants à l'âge a pour 100 000 nés vivants:

A 71 ans, il reste 51 994 survivants, et 49 183 à 72 ans.

La MEDIANE est donc entre 71 et 72 (pour le sexe masculin). Elle est entre 79 et 80 pour l'autre sexe (et pour la population ici décrite).

Après la médiane viendront, par exemple, les quartiles:

75 000 vers 60 ans , 25 000 vers 80 ans (sexe masculin)

ou bien, si cela paraît plus intéressant, les déciles extrêmes:

90 000 vers 45 ans et 10 000 vers 86 ans .

Et l'on peut continuer, la grille de description devenant de plus en plus fine.

MOYENNES ET MOMENTS

Il ne faut pas le cacher: cette autre méthode de description d'une distribution, c'est du calcul intégral.

Sur le diagramme cumulatif, on peut lire la moyenne comme une aire (reprenez la manip' de Papa Van der Waerden).

Pour définir une moyenne (car il y en a autant qu'on veut), il faut commencer par attribuer une valeur numérique à la taille (ce peut être celle qui est usuelle: par exemple le nombre d'années pour la durée de vie, le nombre de centimètres pour la longueur, etc.) mais ce peut être, plus généralement, toute application

$$v : A \longrightarrow \underset{\sim}{R}$$

qui à la taille x fait correspondre une valeur v(x).

On notera que, jusqu'ici, A était seulement ordonné, sans autre structure.

On pourra écrire la

"moyenne des v(x) "

comme faisait Stieltjes :

$$S\ v(x)\ .\ d\ F(x)$$

(le symbole S de la somme sera tantôt écrit $\int$ et tantôt Σ) .

Une seule moyenne (la moyenne "ordinaire" si vous voulez), ce n'est jamais qu'un seul renseignement, ce n'est pas beaucoup.

On procédera alors comme pour le quantilage (mais on ne dit pas le "moyennage" !) en instituant une séquence d'informations: une batterie de fonctions v(x) astucieusement choisies.

Le plus courant: les polynomes, c'est-à-dire la base:

$$1\ ,\ x\ ,\ x^2\ ,\ x^3\ ,\ x^4\ ,\ \text{etc.}$$

D'où : la moyenne ordinaire, la variance et le reste (cf. Calot, cours de Statistique Descriptive, pages 66-67).
Ce sont, dans la tradition, les MOMENTS.

TCHEBYCHEV

La description par les moments offre des commodités mathématiques considérables: mais elle est éloignée de l'intuition immédiate.

Alors qu'il est facile de se rendre compte, pour deux distributions qui auraient même médiane et mêmes quartiles (par exemple), en quoi elles se ressemblent et en quoi elles peuvent différer, c'est beaucoup moins immédiat si les deux distributions ont, par exemple, les mêmes trois premiers moments. Il y a bien une théorie (Stieltjes y a beaucoup travaillé), mais elle reste au fond des armoires.

En d'autres termes, il n'est pas très facile de voir en quoi la description par les moments est une description par "approximations successives".

Il est par contre un peu plus aisé d'apercevoir le lien entre la description par quantiles et celle par moments.

Le premier pas est fort probablement la démonstration de la célèbre inégalité de Markov:

" Si toutes les valeurs de x sont positives, la moyenne (ordinaire) ne peut être inférieure au quart du dernier quartile, ni au dixième du dernier décile, etc... "

ou encore: $\left\{ \begin{array}{l} F(q) = 1 - u \\ F(0) = 0 \end{array} \right\}$ implique: moyenne $\geqslant u.q$

Si l'on applique cela non plus à $\underline{x}$ mais à son carré, on retrouve le très célèbre lemme de Bienaymé et Tchebychev.

La démonstration est rapide (on peut s'aider d'une figure): on sait que le quart, ou le dixième, plus généralement la fraction (u) de la population est située au-delà de la taille (q): si la taille est toujours un nombre positif, il est clair qu'on diminue la moyenne en donnant la taille (q) à tous ceux qui sont plus grands, et la taille zéro à ceux qui sont plus petits.

Et on peut comprendre comment la connaissance de certains quantiles permettrait d'encadrer les moments.

LES LOIS

Il ne suffit pas de collectionner des distributions empiriques: il faut aussi stocker des distributions artificielles (dites "mathématiques" ou "théoriques"). La première raison, pour ce faire, est de trouver ainsi une troisième façon de décrire: la description globale ou schématique (on dit aussi: trouver la ou les Lois).

Donnons deux exemples, parmi les plus simples.

L'historien Vovelle cite ("Mourir autrefois", collection Archives, Gallimard-Julliard, 1974, Page 65) un manuel populaire de piété du début du XVIIIème siècle, intitulé: "Pensez-y bien"; entre autres considérations destinées à nous faire réfléchir et penser à la mort, on lit ceci:

> "en prenant un certain nombre d'hommes, à quelque âge qu'on voudra, il y en aura plus de morts, à vingt ans de là, qu'il n'en restera de vivants ..."

Résumé saisissant d'une table de survie. Traduisons:

" $F(x+20)$ est inférieur à la moitié de $F(x)$ " .

On aura reconnu la référence à la Loi Exponentielle, c'est-à-dire à la distribution telle que le rapport $F(x+a) / F(x)$ soit constant.

Ou encore:

$$F(x) = A^x \qquad \text{(la constante A est plus petite que 1)}$$

La diminution de moitié au bout de vingt ans correspond à un taux annuel de 3,5 % (environ).

La distribution exponentielle est l'une des plus utiles à connaître; on la retrouve partout.
Pour ce qui est de la vie humaine, son approximation est assez médiocre.
Vers 1825, Gompertz a proposé quelque chose de mieux:

$$F(x) = A^{B^x}$$

(l'exposant est lui-même une exponentielle)
et, plus tard, Makeham:

$$F(x) = A^{B^x} \cdot C^x$$

PARETO

En France, en 1958, il y avait 4 300 entreprises faisant un chiffre d'affaires supérieur à 1 milliard, 280 dépassait 10 milliards,

132 plus de 20

89 plus de 30

54 plus de 40

Comment décrire sommairement cette "concentration" ?
La méthode préconisée par Pareto en 1895 consiste à parler un langage "logarithmique", ou, ce qui revient au même, à utiliser une échelle en progression géométrique.

Ici par exemple on dira: les entreprises 40 fois plus grosses (de 1 à 40 milliards) sont 80 fois moins nombreuses (de 4 300 à 54). Comme pour l'exemple précédent, demandons-nous ce qui arrive si cette "loi" est la même tout au long de l'échelle.

$$F(40.x) = F(x) \;/\; 80$$

Cette fois, la solution générale n'est plus l'exponentielle mais la fonction puissance

$$F(x) = C^{te} \,.\, x^{k}$$

(F étant nécessairement décroissant, l'exposant k est négatif).

Ici, il faut prendre : $k = -1{,}2$, et l'on obtient :

calculé	observé
4 300	4 300
272	280
118	132
72	89
54	54

Les distributions d'allure parétienne sont fréquentes: concentration industrielle, concentration urbaine, dispersion des salaires, statistiques lexicales, etc...