Statistiquement Vôtre

Lettre d'information publiée par le groupe "Formation à la Statistique " de l'ASU (Association pour la statistique et ses utilisations)

numéro 2 septembre 1993

Éditorial

Annie Morin, ENSAE

Voici avec un certain retard le deuxième numéro de *Statistiquement Vôtre*. Le bon accueil du numéro un, le nombre important de demandes d'abonnement et le courrier que nous avons reçus nous encouragent à continuer. Comme l'an dernier, cette lettre sera diffusée par l'intermédiaire des IREMs mais ceux qui en ont fait la demande la recevront directement. Nous espérons faire paraître deux numéros par an en régime stationnaire mais n'oubliez pas que cette lettre est la vôtre et que vous êtes invités à y publier des articles, des activités, des dessins. Je rappelle néanmoins que ***Statistiquement Vôtre*** est destinée en priorité aux enseignants du secondaire quelle que soit leur discipline.

Dans ce numéro, vous trouverez une activité sur les pourcentages, un article sur les sondages et un dossier sur la notion de "centre" en statistique. Nous attendons vos commentaires et vos suggestions pour le futur et nous vous souhaitons une bonne lecture.

Reproduit

Sommaire

Pourquoi faire des enquêtes par sondage?

Alain Gély,
INSEE—Observatoire Économique Régional de Bretagne

Quand on parle de statistiques, on imagine des grands tableaux de chiffres construits après avoir rempli des centaines de milliers, voire des millions de questionnaires. On pense aussi à l'exploitation des fichiers administratifs, comme ceux de l'état civil, des impôts ou de la sécurité sociale où on dispose là aussi, d'informations sur chaque individu de la population étudiée. Ces recensements, ces dépouillements de fichiers, ne sont pourtant pas les seuls moyens d'obtenir des informations chiffrées. Les statisticiens disposent en effet d'un autre outil popularisé par les analyses électorales et par les mesures d'audience télé (l'audimat). Cet outil, c'est le sondage, qui connaît par ailleurs de multiples applications moins célèbres, du contrôle de la qualité à l'enquête Emploi de l'INSEE, en passant par les études de marketing et... le dépouillement du recensement de la population.

Qu'est-ce qu'un sondage?

C'est une méthode qui, par une utilisation raisonnée du hasard, permet de tirer des enseignements sur une population en n'observant qu'une fraction de celle-ci. La théorie des sondages est une branche des mathématiques appliquées, qui n'a rien de magique, ni de scandaleux. Elle permet de déterminer comment choisir des échantillons, et dans quelle mesure ces échantillons seront ou non significatifs c'est-à-dire comment déduire, à partir de l'échantillon choisi des lois générales et quelle erreur risque-t-on de commettre? Il y a deux grands types de sondages.

Tout d'abord, les **sondages aléatoires,** pour lesquels on dispose d'une base de sondage. Cette base de sondage peut être par exemple une liste complète et à jour, des habitants d'une région ou encore l'ensemble des électeurs à un scrutin donné.

Deuxième type de sondage : les **enquêtes par quotas** pour lesquelles on dispose d'informations générales sur la population mais pas d'une liste complète. Il va donc s'agir dans ce type d'enquêtes de constituer un "modèle réduit" de la population. Ainsi, pour une enquête sur les intentions de vote, on essaiera d'avoir dans l'échantillon une proportion de jeunes, d'ouvriers etc., à peu près conforme à leur part dans l'électoral global, part connue grâce au recensement de la population.

Un triomphe des sondages

Dans les années 30, à l'approche des élections du président des Etats-Unis, un grand

journal populaire avait organisé une consultation auprès de ses lecteurs pour leur demander comment ils allaient voter. Il y avait eu quatre millions de réponses mais au bout un résultat faux. Gallup qui de son côté avait interrogé deux mille personnes choisies au hasard, avait su désigner le vainqueur. Coup de chance? Miracle? Nullement. Il s'avère que les quatre millions de lecteurs du journal ne votaient pas comme les autres Américains. En revanche, les deux mille sondés de Gallup constituaient un échantillon significatif parce qu'ils avaient été choisis selon une méthode aléatoire.

Par sa connaissance des lois du hasard et de la théorie des sondages, Gallup avait calculé des fourchettes (dits "intervalles de confiance") pour les résultats et évalué le risque de se tromper.

Dans quels cas les sondages sont-ils efficaces?

Il y a des cas où l'utilisation des sondages est le seul moyen envisageable pour obtenir des informations. Par exemple, en contrôle de qualité industrielle, lorsque la vérification des produits entraîne leur destruction, il faut se contenter de prélever un échantillon et tenter d'en déduire si l'ensemble du lot est acceptable ou non. Le sondage sera également préférable dans le cas d'investigations très coûteuses. Ou encore si la complexité du sujet exige des enquêteurs très bien formés, il vaut mieux se contenter d'un petit échantillon correctement observé plutôt que d'un gros échantillon truffé d'erreurs d'observation. Enfin, malgré la puissance des ordinateurs, le dépouillement et l'exploitation d'une grosse enquête statistique peuvent être fort longs. Un sondage s'impose si l'on désire des résultats rapides. Peut-on dire dans ce cas que les sondages sont la panacée toujours préférable à un recensement? La réponse est non.

Regnier

Reprenons en effet l'exemple du sondage préélectoral. Si l'élection s'était effectivement jouée à quelques voix près, Gallup aurait été incapable de fournir le résultat à l'avance. De même, si un mouvement d'opinion se produit dans les tous derniers jours avant une élection, un sondage effectué deux semaines auparavant peut aboutir à un résultat erroné. Les sondages sont d'un intérêt limité lorsqu'il s'agit d'étudier des populations peu nombreuses. Par exemple, dans le cas du scrutin majoritaire en vigueur en France pour les élections, il faut pratiquement un échantillon de la même taille pour obtenir des résultats significatifs sur une circonscription que sur l'ensemble du pays. Enfin pour construire des échantillons aléatoires, il faut des bases de sondage, c'est-à-dire des listes complètes, à jour et accessibles de la population étudiée. Ce n'est pas toujours possible. Une façon de s'en tirer consiste à utiliser la méthode de sondages par quotas.

Fichiers, enquêtes exhaustives, sondages : trois moyens de connaissance complémentaires

En fait, il ne faut pas opposer les recensements, les dépouillements de fichiers administratifs et les investigations auprès d'échantillons. La pratique de l'INSEE avec le recensement de la population en fournit une excellente illustration. Le recensement est donc une enquête exhaustive et qui doit nécessairement l'être pour déterminer sans "erreur aléatoire" la population légale de chaque commune. Mais on désire en tirer d'autres résultats rapides et de bonne qualité pour un coût acceptable par le contribuable. Cette préoccupation conduit l'INSEE à dépouiller certaines questions complexes par un "sondage au quart", n'exploitant donc complètement qu'un questionnaire sur quatre. A défaut, il faudrait recruter quatre fois plus de personnes qu'on ne pourrait pas nécessairement former toutes de manière convenable, ou ...attendre les résultats quatre fois plus longtemps, à supposer que les ordinateurs ne souffrent pas d'indigestions devant de telles masses de données à traiter. Par la suite, le recensement sert de base de sondage principale aux enquêtes par sondage de l'INSEE sur l'emploi, le logement etc.. Mais comme le recensement n'a lieu que tous les sept ou huit ans, il faut recourir à des fichiers administratifs pour mettre à jour les listes des logements issues du recensement précédent, et disposer ainsi d'une base de sondage efficace. Recensements, fichiers et sondages constituent donc trois moyens souvent complémentaires plus que concurrents, de mesurer des phénomènes démographiques, économiques et sociaux.

Regalien

Les sondages sont-ils manipulés?

La loi des grands nombres permet , dans le cas de sondages aléatoires, de calculer des fourchettes. Il y a peu de risques d'en sortir mais ce risque n'est pas nul. Très précisément, quand on communique une fourchette, on ne devrait pas dire "le chiffre vrai se situe entre 44 et 48% mais "il y a 95% de chances que le pourcentage vrai se situe entre 44 et 48%.

Il y a là un problème de présentation sur lequel les professionnels ne devraient pas transiger. Le manque de rigueur dans la présentation des résultats peut nourrir un légitime réflexe de méfiance.

Mais le problème principal est ailleurs. Il réside moins dans les techniques d'échantillonnage que dans le conception même des questionnaires et dans l'interprétation tendancieuse ou erronée des résultats.

Les spécialistes des questionnaires d'enquête savent en effet très bien que la formulation des questions, l'ordre et même le contexte dans lequel elles sont posées influencent les réponses. Il ne faut pas condamner les sondages à cause d'utilisations incompétentes ou

malhonnêtes qui peuvent en être faites. Mais il faut être conscient qu'il y a de multiples causes d'erreur d'observation ou d'interprétation ou de manipulation pour ceux qui souhaitent utiliser les sondages afin d'influencer l'opinion plutôt que pour l'observer et la comprendre.

Regnier

PRÉSENTATION DE VIDÉOSTAT :
DIDACTICIEL INTERACTIF D'INTRODUCTION AUX STATISTIQUES.

Didacticiel réalisé par les départements statistiques de l'**ACTA**, de l'**ENITA** de Bordeaux, de l'**INA-PG**, de l'**Institut de l'Élevage**, du **SCEES**, du **CFPPA** de Pixérécourt.

OBJECTIFS ; PRÉ-REQUIS ; MÉTHODES :

★ C'est une introduction, aux statistiques descriptives, aux probabilités, et à l'inférence, conventionnelles et parfois un peu moins.
★ Il a été conçu et réalisé par une équipe d'enseignents et de praticiens des statistiques.
★ Il ne nécessite pas de pré-requis en statistique mais un niveau 2nde scientifique pour faire les initiations, et un niveau terminale scientifique pour utiliser l'ensemble du didacticiel.
★ Il procède de la pédagogie par l'exemple, la formalisation étant présentée en fin de modules.
★ Il est interactif pendant les cessions, et des exercices corrigés sont proposés en fin de modules.
★ Il peut être utilisé comme support de cours avec des élèves, ou bien comme outils d'autoformation pour les enseignants. Les exemples présentés sont originaux, les analyses et commentaires pratiques sont très fournis.

CONTENU :

★ Il sera composé de 7 modules indépendants, représentant entre 30 et 60 heures de formation :
1-Graphisme et statistiques descriptives. **2**-Probabilités. **3**-Régression linéaire. **4**-Estimation et échantillonnage, **5**-Analyse d'un problème et analyse de la variance, **6**-Lois usuelles. **7**-Tests d'hypothèses. La version actuelle comprend les modules 1, 2, 3, 4 et 5. les modules 6 et 7 seront disponibles au début du 2ième semestre 1993.
★ Chaque module est composé de nombreuses étapes qui peuvent constituer des points d'arrêt et de reprise des cessions.

MATÉRIEL ; LICENCE :

★ Il nécessite au minimum un I386SX-VGA-couleur-souris.
★ Il est en licence mixte au Ministère de l'Agriculture pour 1500FTTC au CNERTA à DIJON ; (à l'étude pour l'E.N.).
★ Il est commercialisé par 3P Informatique ; 4 rue R. Barthélémy, 92120 MONTROUGE ; (1)40 92 08 07.
★ Tous renseignements complémentaires et démonstrations peuvent être demandés auprès des personnes suivantes :
J.P. Desécures ; ACTA, 149, rue de Bercy 75595 PARIS CEDEX 12 ; (1)40 04 50 00
C. Lopez ; Institut de l'Elevage, 149, rue de Bercy 75595 PARIS CEDEX 12 ; (1)40 04 52 69
H. Raymondaud ; CFPPA de Pixérécourt, BP 10, 54220 MALZEVILLE ; 83 21 65 22.

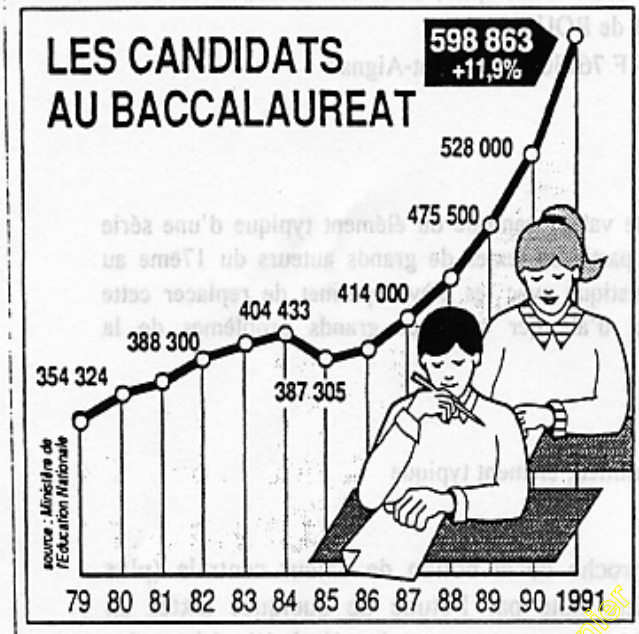
Nombre de candidats au bac

Contenu : augmentation et pourcentages,
critique d'un graphique paru dans la presse.

Niveau : 3ème

Durée : moins d'une heure

Fiche élève



a) Comment ont été calculés les 11,9% ? Qu'en pensez-vous ?

b) Quelle est en pourcentage, la variation du nombre de candidats entre :
89 et 90 ?
90 et 91 ?
89 et 91 ?

c) Faites une estimation du nombre de candidats en 1988

d) Faites une estimation du nombre de candidats en 1992.

Regnier

Commentaires :

a) et b) Une bonne occasion pour présenter ou revoir les principes de calculs avec les pourcentages :

- augmenter de 5%, c'est multiplier par 1,05
- 2 pourcentages successifs ne s'ajoutent pas
- un truc de calcul rapide :

$$\frac{598\,863}{528\,000} = 1,\boxed{13}42$$ donc l'augmentation est de 13,42 %

c) Utilisation de la moyenne et des arrondis. Les élèves calculent la différence 89 - 87, divisent par 2 et ajoutent à 87.

d) 3 modes de calculs sont proposés :

1) l'augmentation 91 - 92 est la même qu'en 90 - 91 : 598 863 - 528 000 = 70 863
598 863 + 70 863 = 669 726 env. 670 000

2) Le pourcentage reste le même : 598 863 x 1,1342... = 679 230 env. 680 000

3) Le pourcentage continue à progresser : la valeur retenue est 15%
598 863 x 1,15 = 688 692 env. 690 000

Rendez - vous en 92 pour comparer avec la réalité.

En attendant, on peut tester la méthode en l'appliquant à l'évolution 90 - 91.

APPROCHE HISTORIQUE DE LA NOTION DE VALEUR CENTRALE D'UNE SÉRIE STATISTIQUE

PICHARD Jean François
Laboratoire Analyse et Modèles Stochastiques, URA CNRS D1378 et IREM de ROUEN
Université de ROUEN - F 76130 Mont-Saint-Aignan

Résumé :

On propose ici une approche de la notion de valeur centrale ou élément typique d'une série statistique à une ou plusieurs dimensions à partir de textes de grands auteurs du 17ème au 20ème siècle. Cette façon d'aborder la Statistique avec les élèves permet de replacer cette science dans une perspective historique et d'aborder quelques grands problèmes de la Statistique et du calcul des probabilités.

Mots-clés :

histoire de la statistique, moyenne, médiane, milieu, élément typique

Une introduction ou une autre approche de la notion de valeur centrale (plus généralement d'élément typique) peut être faite par l'étude de quelques textes en français (ou en anglais) au lycée soit dans le cursus normal, soit à l'intérieur des modules qui devraient se mettre en place prochainement dans le cadre de la rénovation des lycées, ou en DEUG comme sujet de mémoire sur projet.

Reginier

Ce type d'approche, à faire au lycée en liaison avec le professeur de français et/ou d'anglais, permet de confronter les élèves à la lecture d'un texte scientifique proche de la recherche de leur époque (jusqu'au début du 20ème siècle) et de les mettre en contact avec quelques grands problèmes du calcul des probabilités et de la statistique.

Dans cette optique, nous indiquons différents textes d'un abord relativement facile.

L'espérance mathématique et la moyenne arithmétique

Une première ébauche de la notion se trouve chez Blaise Pascal : dans la lettre à Fermat du 29 juillet 1654 et le Traité du triangle arithmétique [13] à partir de laquelle les continuateurs ont tiré les règles du calcul des espérances.

Les premiers textes disponibles traitant du début du calcul des probabilités et utilisables à un niveau relativement élémentaire sont ceux de :

- de Montmort, 1713 en français [12],
- de Moivre, 1711 et 1718 en anglais [11],
- Lacroix, 1816 en français [9], qui est un traité déjà élaboré,
- Laplace, [10a] (le début) et son Essai philosophique [10d] qui est d'un abord assez facile.

Un autre problème intéressant, qui peut être étudié en mémoire au niveau du DEUG, est celui de la ruine du joueur, traité dans les ouvrages cités ci-dessus.

La médiane

Fermat vers 1630 dans [7, p.l36j pose le problème d'un point de vue géométrique : chercher un point du plan dont la somme des distances à 3 points donnés est minimum. La solution est exposée par Torricelli en 1646, d'où le nom de point de Torricelli du triangle formé de ces 3 points. Ce problème est ensuite généralisé à 3 points pondérés, puis à n points (cf. par exemple [5]). La première mention en français du milieu de probabilité (la médiane) est faite par Laplace dans son Mémoire de 1774 [10b].

Les "milieux"

Le problème de la recherche d'un milieu d'observations en théorie des erreurs va amener D. Bernoulli [1] et Laplace [10b,c,a] à étudier différentes lois des erreurs et différents estimateurs. Laplace indique qu'on peut définir autant de "milieu" d'un nombre quelconque d'observations qu'on s'impose de conditions. En particulier, citant D. Bernoulli, Laplace considère [10c] la valeur la plus probable (le mode) et écrit que le milieu doit être recherché par une condition de minimum : "..si au lieu de considérer le minimum des carrés des erreurs, on considérait le minimum d'autres puissances des erreurs ou même de toute fonction des erreurs...". Ce type d'approche sera exploité par la suite par M. Fréchet dans [8].

Reglier

L'espérance morale

Cette notion est introduite par D. Bernoulli en 1730 (dans un mémoire en latin) comme fondement de la théorie de l'utilité. Dans son application au problème de Saint-Petersbourg, on a un texte de Buffon avec un essai de validation par expérimentation dans [3]. L'exposé de ce problème et la tentative de résolution par l'espérance morale calculée comme moyenne généralisée se trouve aussi dans Laplace [9a,d]. Néanmoins, la solution de ce problème passe par une définition opératoire d'un jeu équitable et la démonstration est assez délicate, même au niveau DEUG (cf. Feller [6]).

La moyenne dans un espace à plusieurs dimensions et les éléments typiques

La série statistique ou la variable aléatoire à plusieurs dimensions est (supposée être) à valeurs dans un espace vectoriel. La moyenne est alors calculée en utilisant la structure vectorielle. Quetelet applique cela à la définition d'un "homme moyen" [14], théorie qui soulève rapidement des objections, en particulier de Cournot [4], et aussi de manière plus vive par Bertrand [2].

Fréchet, après avoir étendu la notion de moyenne comme intégrale sur des espaces vectoriels plus généraux, va reprendre l'approche indiquée par Laplace (cf. ci-dessus) et définir par une condition de minimum des éléments typiques dans un espace métrique [8]. Une présentation très simple est faite dans [8b]

Références

[1] Daniel Bernoulli, *Dijudicatio maxime probabilis...,* 1777. traduit en anglais *"The most probable choice between several discrepant observations."* dans E S Pearson and M. Kendall, *Studies in the history of Statistics and Probability* Griffin & Co, London, 1970.

[2] J. Bertrand, *Calcul des probabilités,* éd. Gauthier-Villars, Paris, 1889.

[3] G.L. Buffon. *Essai d'arithmétique morale,* 1777, inclus dans J. Binet et J Roger *Un autre Buffon.* éd. Hermann, Paris, 1977. '

[4] A. Cournot. *Exposition de la théorie des chances et des probabilités,* Paris, 1843.

[5] J. Feldman. G. I.agneau, B. Matalon, *Moyenne. Milieu. Centre. Histoires et Usages,* éd. Ecole des Hautes Études en Sciences Sociales, Paris, 1991.

[6] J W. Feller, *An Introduction to Probability Theory,* Wiley, New York, 1950.

[7] P. de Fermat. *Méthode pour la recherche du maximum et du minimum* dans Oeuvres t.3, trad. P. Tannery et C. Henry, éd. Gauthiers-Villars, Paris, 1896. '

[8] M. Fréchet,

(a) *Généralités sur les probabilités - Éléments aléatoires,* éd. Gauthier-Villars Paris. 2ème . 1950,

(b) Palais de la Découverte, Paris, 1949.

[9] F. Lacroix, *Traité élémentaire des probabilités,* Paris. 1816.

[10| **P.S.** de Laplace, Oeuvres complètes, éd. Gauthiers-Villars, Paris de 1878 à 1905

(a) *Théorie analytique des* probabilités Paris, 1820, dans Oeuvres t 7 1886

(b) *Mémoire sur la probabilité des causes par les événements,* 1774, Oeuvres t 8, 1891, p.41-47,

(c) *Mémoire sur les probabilités,* 1781, Oeuvres, t 9. 1893, surtout p 476-482

(d) *Essai philosophique sur les probabilités,* Sème éd. 1825, éd. C. Bourgois, 1986.

[11] A. de Moivre,

(a) *De mensura sortis.* 1711, traduit en anglais par B. MacClintock • *"on the Measurement of Chance".* International Statistical Review (1984). n 237-262 précédé de commentaires par A. Hald,

(b) *The Doctrine of Chances,* 1ère éd. 1718, 3ème éd. 1756, réédité par Cass London, 1967.

[12] P. Rémond de Montmort, *Essay d'analyse des jeux de hazard,* 2ème éd Paris 1713, disponible à la Bibliothèque Nationale.

[13] B. Pascal : *Traité du triangle arithmétique* (vers 1654). Oeuvres, éd. du Seuil Paris 1963 (et dans La Pléïade).

[14] A. Quetelet, *Sur l'homme et le développement de ses facultés ou Essai de physique sociale.* Paris, 1835 et Bruxelles. 1836.

Register

Éléments typiques d'une série statistique par analogie avec des "centres" en physique.

Pichard Jean-François.

Il est fréquent dans l'enseignement, en particulier dans celui des sciences expérimentales, d'utiliser des analogies pour présenter un phénomène nouveau en se référant à un phénomène différent étudié antérieurement. De plus, l'intérêt des analogies n'est pas uniquement didactique car très souvent l'analogie intervient dans la démarche du chercheur.

La notion de centre ou de milieu et celle de barycentre, vues en géométrie, sont insuffisantes pour préparer les élèves à accepter la diversité d'éléments typiques qu'on peut associer à une série statistique suivant les propriétés que l'on veut mettre en évidence. On va utiliser des analogies avec la mécanique pour illustrer la notion d'éléments typiques d'une série statistique, ce qui permet de plus de donner un sens concret à cette notion. L'avantage d'une analogie avec la Physique est que les éléments ainsi introduits ont une existence expérimentale.

Quelques "centres" en physique.

Régnier

Considérons une barre de 1 cm^2 de section obtenue en soudant bout à bout un barreau de cuivre et un barreau d'aluminium de 1 mètre de longueur chacune. L'aluminium et le cuivre ayant des densités différentes (2,7 g/cm^3 et 8,9 g/cm^3), suivant la propriété d'équilibre ou de mouvement étudiée, on peut trouver un point de la barre qui joue le rôle de "centre de la barre".

- Si la barre est étudiée en tant que volume, la section à la jonction des deux barreaux partage la barre en deux parties de même volume. Le point centre de cette section est le "centre de poussée" en hydraulique. La forme de l'objet étant régulière, ce point est aussi le centre géométrique de la barre.
- La section située dans le cuivre à 34,8 cm de la jonction partage la barre en deux parties de même masse ; le centre de cette section est le "centre des masses" en mécanique.
- La section située dans le cuivre à 27,7 cm de la jonction est celle où, en suspendant la barre, il y aura équilibre : c'est le "centre de gravité" ou barycentre de la barre.

Ainsi, sauf dans le cas d'une barre homogène, il n'y a pas un point unique de la barre qui puisse jouer le rôle de centre pour toutes les propriétés énoncées, mais plusieurs centres possibles qui sont déterminés par la propriété considérée.

Donner un point central pour la barre décrite ci-dessus est une information utile dans certains problèmes de physique, mais bien insuffisante dans la plupart des cas. Si l'on étudie

cette barre en mouvement, une autre information au moins est nécessaire pour décrire le comportement de la barre, par exemple son inertie par rapport au centre de gravité.

Analogie avec la statistique.

Les séries statistiques étudiées dans le secondaire (collège et lycée) sont discrètes ; on va donc construire un modèle physique qui soit le plus proche possible de ce cas. Imaginons un système mécanique constitué de n boules de faibles dimensions – on les supposera ponctuelles - de masses m_1, m_2,...,m_n, fixées sur un axe rectiligne de masse négligeable. Pour repérer la position des boules sur l'axe, on choisit une échelle sur cet axe. Soient x_1, x_2,...,x_n les abscisses des différentes boules, qu'on supposera ordonnées par valeurs croissantes. Ce système possède différentes propriétés qui apparaissent au cours de recherche d'équilibre ou de mouvements. En Statistique, une série quantitative { (x_i, p_i)}, où x_i est la valeur observée du caractère et p_i le poids que l'on attribue à cette observation, est souvent assimilée à un tel système. Dans cette analogie, chaque individu i possède un poids p_i identifié à la masse m_i, et la valeur x_i du caractère étudié est représentée par un point d'abscisse x_i sur un axe gradué.

Dans le système mécanique, un point de l'axe joue physiquement un rôle privilégié : le centre de gravité ou barycentre du système. Ce point G admet une abscisse g définie par la formule :

$$(\sum_i m_i)g = \sum_i m_i x_i$$

Si l'unité de mesure des masses est telle que la somme de toutes les masses vaut 1, cette valeur g est définie par :

$$g = \sum_i m_i x_i$$

Ce point moyen G n'est pas nécessairement confondu avec l'un des points du système, mais sa connaissance permet la localisation du système. Sa position, définie à l'aide des masses et des abscisses est indépendante du choix de la graduation de l'axe et des unités de mesure.

Dans le cadre Statistique, on considère la valeur $\bar{x}$ qui correspond à l'abscisse du centre de gravité ou barycentre des points et qui est définie par :

$$\bar{x} = \sum_i p_i x_i \text{ avec } \sum_i p_i = 1$$

Cette valeur calculée à partir des termes (x_i, p_i) est nommée moyenne arithmétique des x_i pondérée par les p_i, ou plus simplement moyenne des x_i. Dans le cas d'une série statistique

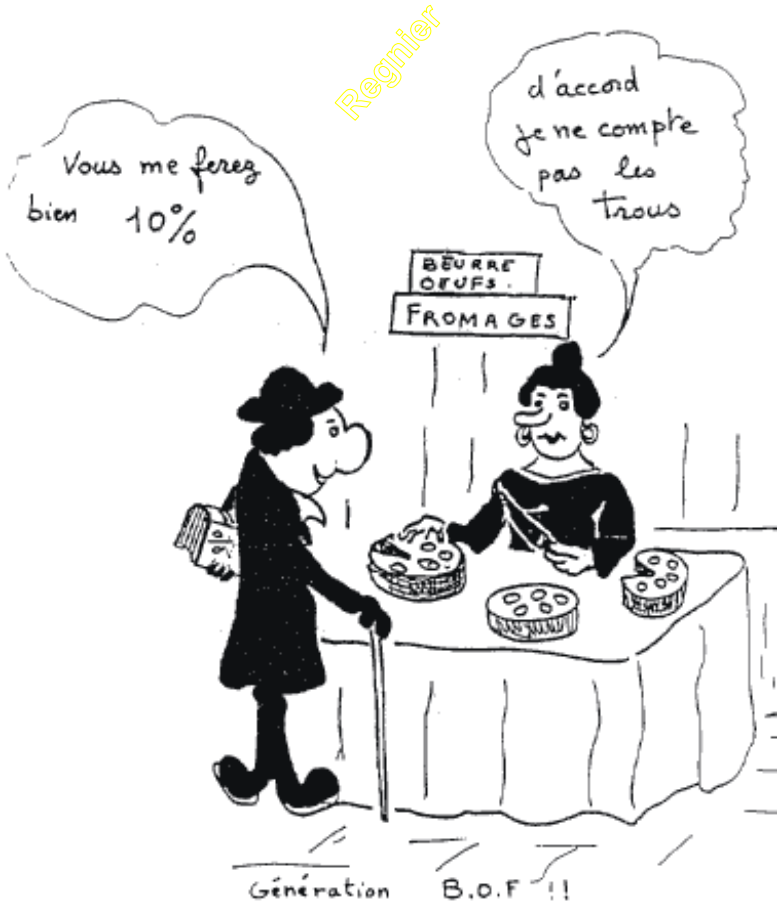
équipondérée, tous les individus ont le même poids p_i et la moyenne arithmétique est obtenue par la formule :

$$\bar{x} = \frac{1}{n}\sum_i x_i$$

Par analogie avec l'exemple donné en haut, dans le cas d'une série statistique numérique $\{(x_i, p_i)\}$., on peut déterminer plusieurs centres ou indicateurs de position qui rendent compte de certaines propriétés :

- Le milieu de l'intervalle entre la plus petite valeur x_1 et la plus grande valeur x_n. l'étendue de la série, est une valeur centrale qui correspond au "centre de poussée" défini ci-dessus.
- Toute valeur séparant la série en deux parties de même poids est une valeur centrale, appelée médiane, qui correspond au "centre des masses".
- La valeur moyenne arithmétique correspond au barycentre comme on l'a déjà vu. L'inertie par rapport au centre de gravité a comme analogue en Statistique la variance de la variable.

Ce dessin est l'oeuvre de Louis Vincent (Irem de Rennes) et l'activité de la page suivante est extraite d'une publication de l'Irem de Rennes intitulée "Statistique au collège", octobre 1992.



Cherchez le centre !

Lejeune Michel

Professeur à l'ENSAE

Considérons le problème pratique suivant. Quelle est la localisation optimale d'un dépôt de carburant à partir duquel on alimentera *n* stations service échelonnées sur une autoroute? On notera $x_1, x_2, \ldots, x_n$, les abscisses des stations représentées schématiquement sur la droite réelle avec une origine choisie arbitrairement. Supposons, avec bon sens, que le coût d'approvisionnement soit proportionnel au nombre de kilomètres parcourus. La localisation doit être choisie à l'abscisse *c* qui minimise la quantité :

$$\sum_{i=1}^{n} |x_i - c|$$

La dérivée de chaque terme par rapport à c vaut +1 ou -1 selon que *x,* est supérieur ou inférieur à c. Ainsi l'abscisse optimale recherchée est la **médiane** de la série numérique des x_i ou plus précisément **une** médiane si *n* est pair car il n'y a pas, alors, unicité de solution. De la même façon l'abscisse **moyenne** est la bonne solution si le coût est proportionnel au carré des kilomètres parcourus puisqu'elle minimise :

$$\sum_{i=1}^{n} (x_i - c)^2$$

Gageons que ces résultats qui permettent de distinguer entre médiane et moyenne ne sont pas vraiment intuitifs. Cet exemple illustre le choix d'une notion de **centralité d'une série statistique.** Cette notion qui constitue le résumé descriptif le plus radical d'une série de nombres est utile entre autres pour comparer succinctement plusieurs groupes, échantillons ou populations, en ne retenant de chacun que la valeur la plus typique et en ignorant les disparités internes. On pourra, par exemple, cartographier les départements français selon l'âge moyen de ses résidents ou selon leur revenu médian.

Si l'on accepte l'idée inspirée de la géométrie que le centre est le point globalement le plus proche des points de la série on voit que la seule part d'arbitraire résulte dans la façon d'agréger les écarts ce qui revient au choix d'une norme dans R^n. Si X désigne le vecteur de composantes $x_1, x_2, \ldots, x,$ il reste alors à déterminer le nombre c qui minimise :

$$|| X - c\mathbf{1} ||$$

Une fois cette norme choisie la mesure globale des disparités ou **mesure de dispersion** est définie par cette même distance d'ensemble des points à leur centre. Ainsi à la moyenne m correspond naturellement l'écart-type :

$$\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2\right]^{\frac{1}{2}}$$

et à la médiane M la déviation absolue moyenne à la médiane (DAM) :

Regnier

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - M|$$

Tout autre choix de dispersion n'est pas cohérent et peut conduire à des aberrations, par exemple la possibilité théorique que la dispersion obtenue en agrégeant deux groupes soit inférieure à la somme (pondérée par les effectifs) des dispersions de chaque groupe. La mesure globale en norme L^P qui généralise les précédentes est :

$$\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^p\right]^{\frac{1}{p}}$$

où l'on voit que plus p est grand plus les valeurs extrêmes sont prépondérantes. A la limite quand p tend vers l'infini le centre devient $\frac{1}{2}$[max{x_i}+min{x_i}] et la dispersion est égale à un facteur constant près à l'étendue [max{x_i}- min{x_i}] de la série. La sensibilité plus ou moins forte aux valeurs extrêmes fournit un élément de réponse à la question du choix entre moyenne et médiane. Si l'on craint la présence de telles valeurs ayant une influence trop forte sur la mesure de dispersion et donc de centralité, ou si, par nature, la répartition des valeurs peut comporter une forte asymétrie (par exemple les revenus sont bornés à gauche mais sans "limite" à droite ...) la médiane aura pour vertu d'être plus représentative de l'ensemble. Cependant la moyenne conserve ses droits quand elle exprime aussi un total rapporté au nombre d'individus. Ainsi le revenu moyen a une interprétation évidente, mais est-ce aussi clair pour l'âge moyen ?

La justification de tel ou tel choix peut venir encore de considérations probabilistes. Ainsi l'écart-type s'est imposé très tôt comme mesure de dispersion avec, pour arrière-fonds , la répartition selon le modèle de Gauss-Laplace.

Pour montrer l'influence d'une répartition sur le choix de distance nous terminerons par un exemple réel, entendu dans un jeu radiophonique. Deux candidats doivent estimer au plus proche le nombre de disques vendus pour une chanson donnée. L'un répond dix mille l'autre répond deux millions. La réponse vraie étant un million c'est le premier candidat qui a été déclaré gagnant. Ici l'éventail des réponses couvre des ordres de grandeurs différents et les écarts seraient appréciés avec plus de pertinence sur une échelle logarithmique. Mais allez convaincre l'animateur et les candidats...

Regnier

Le comité de rédaction est constitué des membres du groupe "Formation à la statistique".
Tout courrier est à envoyer à l'adresse :

ASU a/s ISUP,
Groupe Formation,
Tour 45-55, E2,
4, place Jussieu,
75252 Paris Cedex 05

Les auteurs sont responsables du contenu de leurs articles.

Regnier