La Statistique : connaissance et maîtrise de l'aléatoire

Jeanne Fine
Professeure, IUFM Midi-Pyrénées
IMT, UMR5219, Université de Toulouse

Résumé

Qu'est-ce que la Statistique ? Quel est son objet ? Quelles sont ses méthodes ? Les statistiques font partie de notre quotidien mais ***la Statistique***, en tant que discipline mathématique, est encore peu connue. La faire mieux connaître est l'objectif de cette présentation. Ce texte est le support de la première conférence d'un cycle organisé par l'IUFM Midi-Pyrénées, en 2000, à l'occasion de l'année mondiale des mathématiques, promue par l'Union Internationale des Mathématiciens avec le soutien de l'UNESCO. Il a été légèrement modifié pour être intégré dans le dossier pédagogique distribué lors de la session sur l'enseignement secondaire, proposée dans le cadre du deuxième colloque francophone international sur l'enseignement de la statistique, Université Libre de Bruxelles, 8-10 septembre 2010.

1. Introduction

Une tentative de définition de la Statistique pourrait être la suivante : ***la Statistique*** est une branche des mathématiques regroupant les méthodes de recueil, de traitement et d'interprétation de données afin de décrire et de prévoir les phénomènes les plus divers ; ce peut être pour explorer et mieux connaître les goûts, les préférences, les attitudes ou les comportements d'une certaine population humaine ou animale, ce peut être pour confirmer ou infirmer une hypothèse de recherche. La Statistique est la discipline de référence pour la recherche scientifique basée sur l'expérimentation. On distinguera ***la Statistique*** des ***statistiques*** qui sont les données elles-mêmes, les statistiques de l'emploi, par exemple.

La Statistique, telle qu'elle est définie ci-dessus, repose sur les Probabilités et date des années 1900. Elle est parfois appelée Statistique mathématique pour l'opposer à la Statistique publique dont la mission est l'enregistrement des statistiques de l'État : population, richesse, emploi, consommation, … Cette première mission a donné son nom à la discipline : statistique vient du latin *status* qui signifie État.

Les Probabilités ont pour objet d'étude les phénomènes ***aléatoires*** (du latin *alea* : coup de dés), c'est-à-dire, les phénomènes dont l'issue n'est pas connue à l'avance de façon certaine, dont l'issue est due au ***hasard*** (de l'arabe *azzahr* : jeu de dés) ; par exemple, le lancer d'un dé ou d'une pièce de monnaie, le tirage au loto ou le temps qu'il fera demain.

Le recueil des données est une étape essentielle de la Statistique qui implique de définir de façon précise :

- l'objectif de l'étude permettant d'aborder le phénomène d'intérêt,
- la ***population*** sur laquelle porte l'étude,
- les caractères, appelés ***variables***, que l'on relève sur les ***individus*** (ou unités statistiques) de la population et qui sont sensés mesurer le phénomène d'intérêt.

Exemples de population et de variables :
- un ensemble d'étudiants (la population) et le sexe, l'année d'étude, la filière d'étude et les notes obtenues à différentes épreuves (les variables),
- un ensemble d'entreprises (la population) et le secteur d'activité, le chiffre d'affaires d'une année donnée et le nombre de salariés à une date donnée (trois variables, les individus sont ici les entreprises).

Les variables peuvent être ***catégorielles*** (sexe, année d'étude, filière d'étude pour le premier exemple, secteur d'activité pour le deuxième exemple) ou ***réelles*** (les notes pour le premier exemple, le chiffre d'affaires et le nombre de salariés pour le deuxième exemple).

Lorsque l'information est recueillie auprès de la population tout entière, on parle d'enquête exhaustive ou de ***recensement***. Bien souvent, on recueille l'information sur un sous-ensemble de la population, appelé ***échantillon***. On parle alors d'enquête partielle ou de ***sondage***. Le sondage est dit aléatoire lorsque le tirage de l'échantillon respecte des règles permettant de contrôler les probabilités d'inclusion des individus dans l'échantillon. On utilise alors des ***tables de nombres au hasard*** ou des ***générateurs de nombres au hasard*** implantés sur les calculatrices et les ordinateurs dont on reparlera en conclusion.

Les données peuvent être également collectées selon un protocole expérimental contrôlé appelé ***plan d'expériences***. Que ce soit par sondage sur un échantillon d'individus de taille n ou lors d'une suite de n expériences, si X désigne la variable d'intérêt, la liste des observations $(x_1, \dots, x_n)$ est appelée aussi ***échantillon*** (des valeurs observées).

On peut alors distinguer dans la Statistique, trois parties :

- la ***Statistique Descriptive***, dont l'objectif est de synthétiser sous forme de tableaux condensés, de graphiques et d'indices numériques, l'information recueillie (sur la population, sur un échantillon ou sur une suite d'expériences),

et deux parties qui s'appuient sur le calcul des ***Probabilités*** :

- la ***Statistique Inférentielle*** dont l'objectif est d'inférer à la population des résultats observés sur un échantillon aléatoire et
- la ***Modélisation Aléatoire*** (ou stochastique ou probabiliste ou statistique) dont l'objectif est d'expliquer et de prévoir ; on suppose, par exemple, que le phénomène peut être mis en équations à des erreurs aléatoires près.

Nous allons reprendre les quatre parties présentées ci-dessus (Statistique Descriptive, Probabilités, Statistique Inférentielle et Modélisation Aléatoire) en les replaçant dans un contexte historique, en donnant, à partir d'exemples, quelques éléments clés pour comprendre le contenu des ces différentes parties, enfin en précisant les limites de ce qui est présenté et les extensions possibles.

2. La Statistique Descriptive

2.1. Historique

Les recensements de populations et de richesses remontent à l'époque sumérienne, 5000 à 2000 ans avant notre ère. On trouve des preuves de leur existence en Mésopotamie, 3000 ans avant notre ère. Ils étaient régulièrement organisés en Egypte, à partir de 2900 ans avant notre ère, dans un but fiscal, avec l'obligation de déclarer ses revenus sous peine de mort ! Toutes les civilisations importantes, Japon, Rome, Inca, Inde, se sont dotées d'un système administratif fort, entretenu par la connaissance quantitative des richesses de l'Etat.

En France, le XIVème siècle voit le début des enregistrements des actes d'état civil ; ils sont généralisés au XVIème siècle (registre des feux).

Les progrès fondamentaux de la Statistique vont apparaître lors de la seconde moitié de XVIIème siècle. C'est l'époque de l'école anglaise d'arithmétique politique, guidée par le souci de quantification et la recherche de constantes de comportement permettant des estimations et des prévisions : nombre d'individus par feux, nombre d'enfants par femme, …

La méthode du multiplicateur est utilisée pour estimer la population à partir du recensement des feux. Vauban préconise l'utilisation d'échantillons de terres arables dans chaque province pour estimer au mieux les capacités agricoles. Au XVIIIème, des calculs d'assurances sont réalisés à partir de tables de mortalité.

Le XIXème siècle voit le retour des recensements : en 1801 sont réalisés simultanément les recensements des populations de l'Angleterre, Danemark, France et Norvège. L'astronome et statisticien belge Adolphe Quételet organise, en 1841, le service central de statistique de Belgique qui servira de modèle pour d'autres pays, et, en 1853, le premier congrès international de Statistique.

La science modèle de l'époque est la physique classique, qui explique les phénomènes naturels par des lois déterministes, tandis que le comportement humain semble purement individuel et imprévisible. Il introduit le concept de « physique sociale » et étudie l'influence de variables, telles que le sexe, l'âge, l'éducation, le climat, sur les comportements les plus divers. Les régularités trouvées dans les taux de maladies mentales, de crimes, de suicides et de prostitution perturbent les tenants de la doctrine du libre arbitre et de la responsabilité individuelle. Quételet, tout en étant un farouche défenseur des recensements par rapport aux relevés sur échantillon, est sans doute le premier à voir que la Statistique pouvait être fondée sur les Probabilités dont le développement se déroulait parallèlement et indépendamment de la Statistique.

2.2. Les concepts de base et quelques méthodes

Nous avons vu en introduction que l'objectif de la Statistique Descriptive est de synthétiser sous forme de tableaux condensés, de graphiques et d'indices numériques, l'information recueillie (sur la population, sur un échantillon ou sur une suite d'expériences). Cette partie étant enseignée dans les classes du secondaire, nous en détaillons les éléments clés.

Une *série statistique* de taille n est la liste des observations d'une variable (catégorielle ou réelle) sur les n individus de la population ou de l'échantillon ou encore sur les n expériences réalisées. Par exemple, on lance une pièce de monnaie trente fois de suite et on note X la variable qui à chaque lancer associe P si on obtient pile et F si on obtient face. La

liste des observations des 30 lancers est : P, P, F, P, F, P, F, P, P, P, P, P, P, P, F, F, P, P, F, F, F, F, F, P, P, F, F, P, P, P série statistique de taille 30.

a) Distributions d'effectifs et de fréquences, diagrammes associés

La première étape consiste à ordonner les observations quand elles sont numériques et à regrouper les observations identiques. La ***distribution d'effectifs*** de la variable est la donnée, pour chaque valeur distincte de la variable numérique ou pour chaque modalité de la variable catégorielle, de l'***effectif*** associé (c'est-à-dire du nombre d'observations correspondant). Dans notre exemple {(P, 18), (F, 12)} est la distribution d'effectifs de la variable *X*. L'effectif est appelé aussi fréquence absolue.

On appelle ***fréquence*** (ou fréquence relative) le rapport de l'effectif sur l'effectif total *n*. Dans notre exemple {(P, 0.6), (F, 0.4)} est la ***distribution de fréquences*** de la variable *X*. Les deux distributions sont souvent présentées dans un tableau (sous-entendu d'effectifs et de fréquences).

Parmi les variables catégorielles, on distingue :

- les variables catégorielles ***nominales*** lorsque les modalités de la variable peuvent être listées dans un ordre quelconque et
- les variables catégorielles ***ordinales*** lorsque les modalités de la variable peuvent être ordonnées, par exemple, une évaluation de la satisfaction d'une clientèle sur quatre niveaux : « très insatisfaisant », « insatisfaisant », « satisfaisant », « très satisfaisant »

Parmi les variables réelles, on distingue :

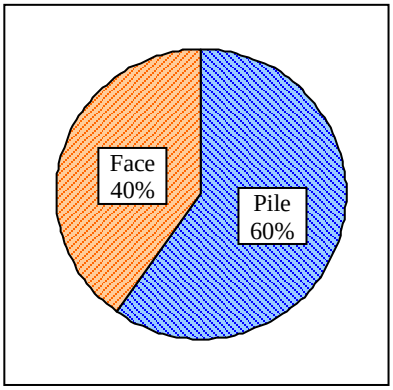
- les variables réelles ***discrètes*** prenant un nombre restreint de valeurs, en général des valeurs entières, par exemple « le nombre de pièces défectueuses par jour » d'un contrôle de production d'un atelier,
- les variables réelles ***continues*** prenant potentiellement toute valeur d'un intervalle de R, par exemple « la durée de réparation en minutes » d'un atelier de fabrication. Les valeurs de la variable sont alors regroupées en ***classes*** (intervalles deux à deux disjoints dont la réunion recouvre l'ensemble des valeurs de la variable). On perd donc un peu d'information ; par convention, on suppose que, dans chaque classe, les observations sont uniformément réparties.

A chaque type de variable correspond un diagramme d'effectifs (et de fréquences) approprié : ***diagramme en secteurs*** circulaire ou semi-circulaire et ***diagramme en barres*** pour les variables catégorielles (cf. ci-après les diagrammes « lancer d'une pièce de monnaie » et « enquête de satisfaction »), ***diagramme en bâtons*** pour les variables réelles discrètes (cf. le diagramme « contrôle de fabrication »), ***histogramme*** pour les variables réelles continues dont les valeurs ont été regroupées en classes (cf. les diagrammes « durée de réparation » pour des classes d'***amplitudes*** égales et « salaire mensuel net des salariés d'une entreprise » pour des classes d'amplitudes inégales). Dans tous les cas, les éléments graphiques représentant les modalités, les valeurs ou les classes de la variable doivent être de mesures proportionnelles aux effectifs et aux fréquences (mesures d'angle des secteurs, mesures de hauteur des barres ou des bâtons, mesures d'aire des rectangles élevés au dessus des classes de l'histogramme). Ces diagrammes, lorsqu'ils sont bien construits, caractérisent les distributions d'effectifs et/ou de fréquences des variables.

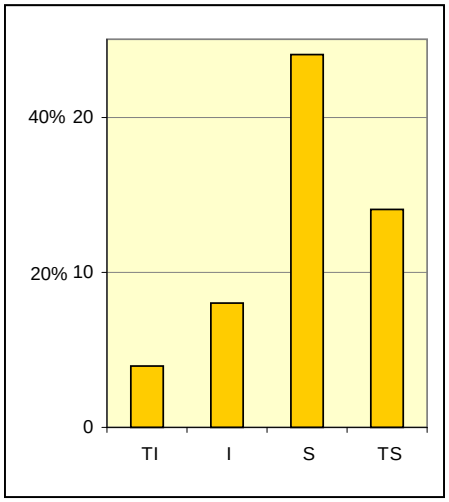
Les diagrammes en secteurs ou en barres peuvent être utilisés pour représenter la répartition d'une « masse totale » d'une variable réelle positive en diverses catégories, par

exemple, la répartition du budget d'une commune (exprimé en euros) selon les divers postes de dépenses ou la répartition du volume global de pêche (exprimé en tonnes) d'un port breton un jour donné selon différentes catégories de pêche. Il ne s'agit plus d'un diagramme d'effectifs ou de fréquences. Pour davantage de visibilité, même pour des variables réelles discrètes, il est possible de remplacer les « bâtons » par des « barres » ou même par des intervalles (passage du discret au continu). Une propriété essentielle de ces diagrammes est que la graduation de l'axe vertical commence au zéro afin de respecter la proportionnalité.

Lancer d'une pièce de monnaie
Diagramme en secteurs



Enquête de satisfaction
Diagramme en barres

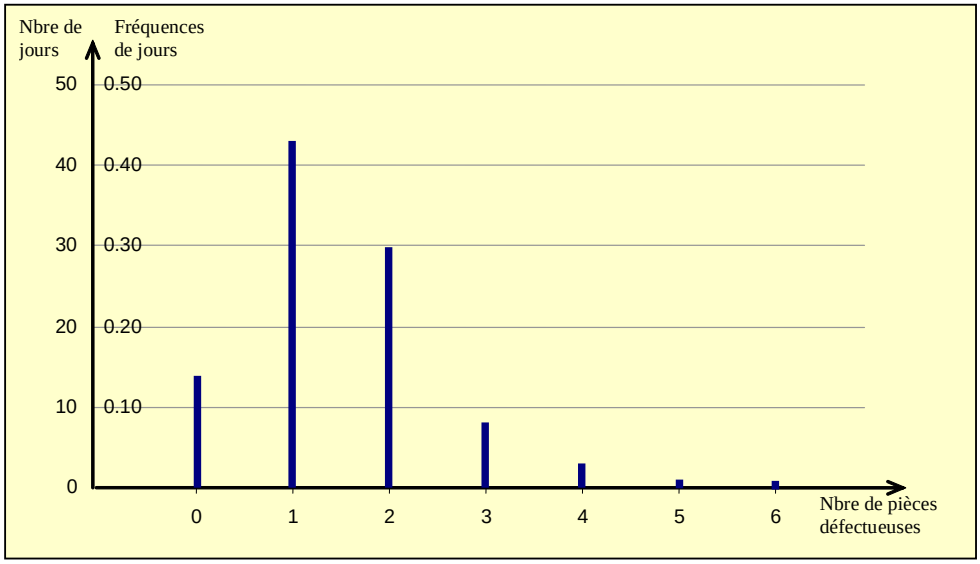


Contrôle de fabrication.

Nombre de pièces défectueuses dans la production journalière d'un atelier, observation réalisée sur 100 jours de production :

nombre de pièces défectueuses	0	1	2	3	4	5	6
nombre de jours	14	43	30	8	3	1	1

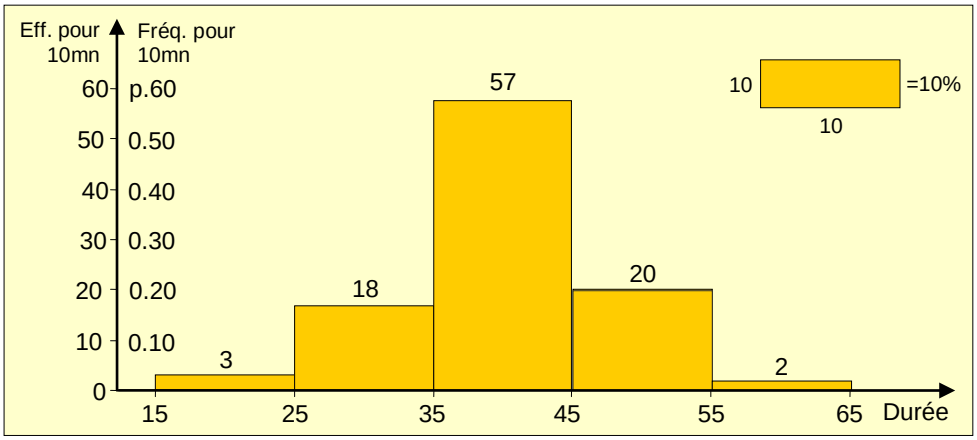
Représentation graphique des effectifs et des fréquences de la variable « Nombre de pièces défectueuses »



Durée de réparation.
Durée en minutes de 100 réparations réalisées dans un atelier.

Durée	Nombre de réparations
[15 ; 25[	3
[25 ; 35[	18
[35 ; 45[	57
[45 ; 55[	20
[55 ; 65[	2

Histogramme des effectifs et des fréquences de la variable « durée »

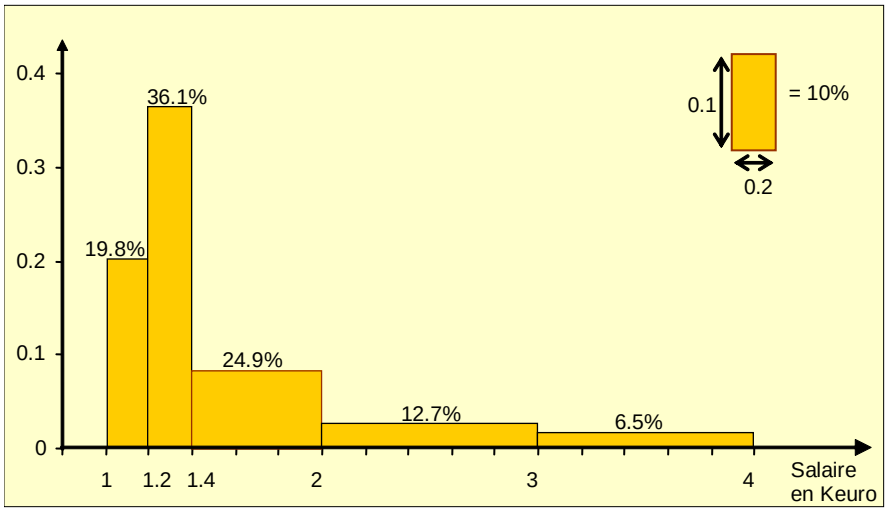


Salaire mensuel net des salariés d'une entreprise

Distributions d'effectifs et de fréquences du salaire mensuel net

Salaire en K€	Effectif	Fréquence
[1.0 ; 1.2[	198	0.198
[1.2 ; 1.4[	361	0.361
[1.4 ; 2.0[	249	0.249
[2.0 ; 3.0[	127	0.127
[3.0 ; 4.0[	65	0.065
Total	1000	1

Histogramme des effectifs et des fréquences du salaire mensuel net



Les fréquences sont proportionnelles aux aires des surfaces hachurées. L'aire totale est égale à 1 (100% des salariés). On notera que pour les représentations par des histogrammes, il est nécessaire de donner une indication de l'unité d'aire utilisée afin d'interpréter le diagramme sans ambiguïté.

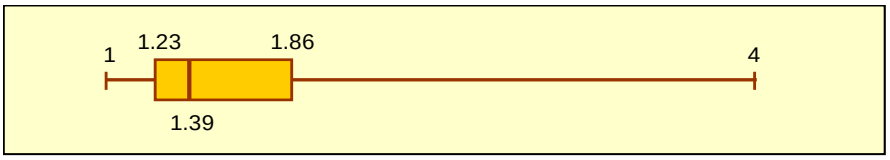
b) Résumés numériques, moyenne, variance, écart-type, médiane, quartiles

Il peut être utile de résumer une distribution d'effectifs ou de fréquences par un ou plusieurs indices numériques. L'indice numérique de ***tendance centrale*** le plus utilisé est ***la moyenne*** arithmétique $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ qui s'écrit $\bar{x} = \frac{1}{n}\sum n_k x_k = \sum f_k x_k$ lorsqu'on dispose de la distribution d'effectifs ou de fréquences (*k* est alors l'indice des valeurs distinctes de la variable). On utilise ***la variance*** (moyenne des carrés des écarts à la moyenne) et ***l'écart-type*** (racine carrée de la variance) pour mesurer la ***dispersion*** des valeurs de la variable par rapport à la moyenne. Moyenne et écart-type sont mesurés dans la même unité de mesure que la variable. Dans le cas où la distribution est symétrique (la « durée de réparation » par exemple) et peut être approchée par la célèbre « courbe en cloche » de la loi normale dont on reparlera plus loin, ces indices caractérisent la distribution de fréquences.

En revanche, pour la distribution du « salaire mensuel net » qui est dissymétrique, le salaire moyen (égal à 1 656 €) est décentré vers les fortes valeurs de la variable. Un meilleur résumé de la distribution sera donné par ***les quartiles***. Les quartiles sont les niveaux de salaire qui partagent la population des 1000 salariés en quatre sous-populations d'effectifs égaux. On trouve ici que le ***premier quartile*** est égal à 1 228 €, le deuxième quartile (ou ***médiane***) à 1 388 € et le ***troisième quartile*** à 1 860 €.

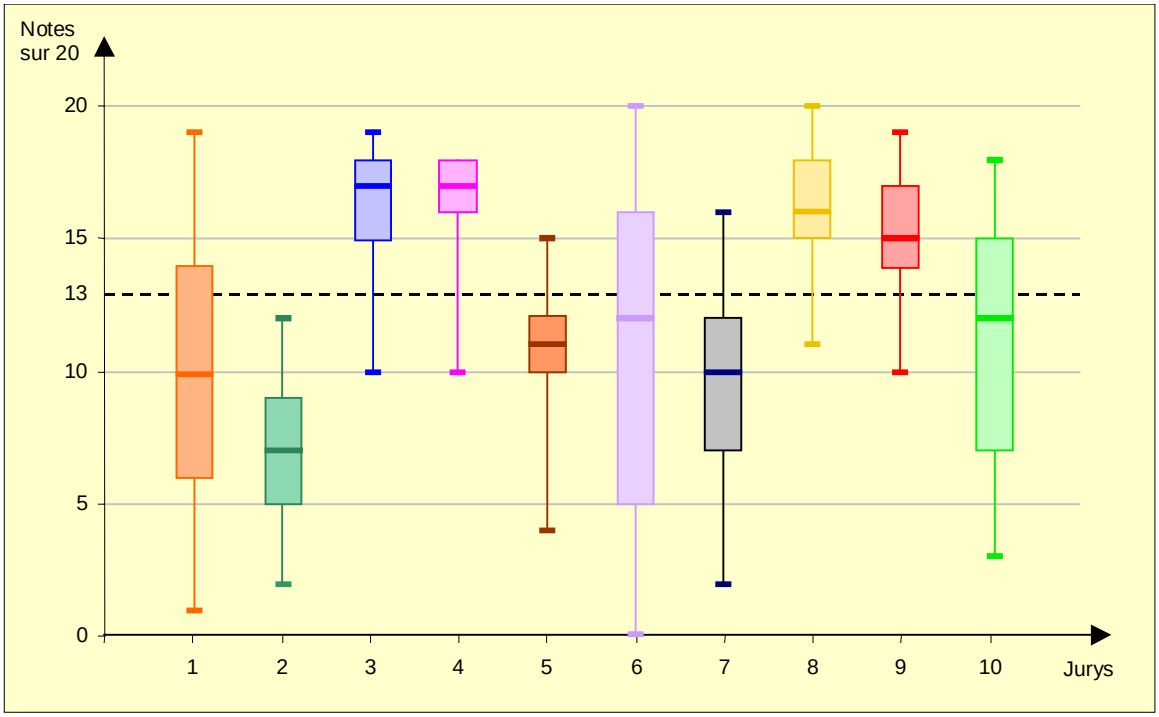
Le ***diagramme en boîte*** (« box plot » en anglais) est une représentation graphique de la distribution des effectifs et des fréquences à partir des quartiles. La boîte est limitée par les premier et troisième quartiles et le trait intérieur à la boîte correspond à la médiane, enfin les extrémités correspondent au ***minimum*** et au ***maximum*** de la variable.

Diagramme en boîte du salaire mensuel en K€



L'avantage des diagrammes en boîtes est de pouvoir comparer plusieurs distributions. Prenons un exemple : 800 étudiants présentent leur candidature pour leur admission dans une école. La direction de l'école répartit de façon aléatoire les 800 candidats selon 10 jurys de 80 étudiants pour un entretien. Sur les 800 candidats, les 400 ayant la meilleure note à l'entretien seront admis. La médiane des 800 notes étant 13, seront admis les étudiants dont la note à l'entretien est supérieure à 13. On présente ci-après les distributions des notes à l'entretien des candidats pour chacun des 10 jurys. On remarque que, si l'on ne "redresse" pas les notes, l'admission dans cette école dépendra très fortement du jury avec lequel le candidat aura passé son entretien.

Distribution des notes d'entretien données par 10 jurys pour environ 80 candidats chacun

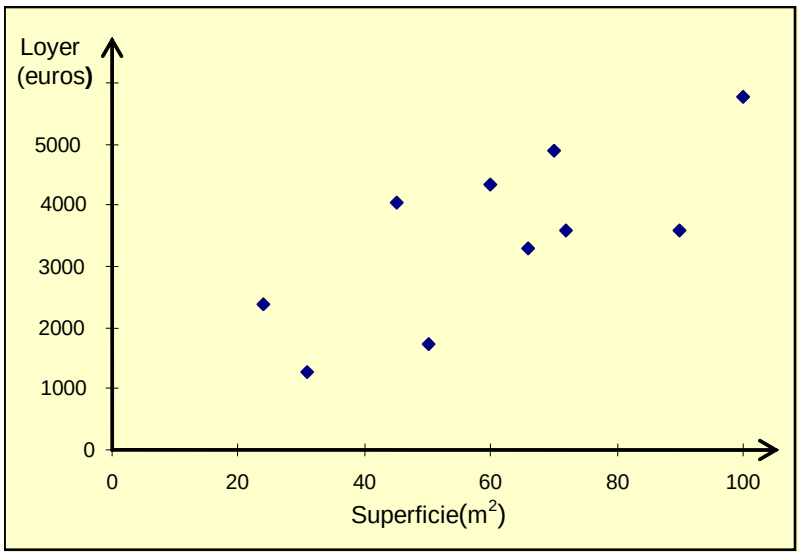


c) Liaison entre deux variables, covariance et corrélation, Khi2, rapport de corrélation

L'objectif d'une étude statistique est bien souvent d'étudier une liaison éventuelle entre deux variables.

c1) Dans le cas de deux variables réelles, on pourra représenter les données par un *nuage de points* dans un repère orthogonal. Voici, par exemple, la superficie en m^2 (variable *X*) et le loyer mensuel en euros (variable *Y*) de 10 locaux commerciaux.

Loyer *Y* (en Euros)	**Superficie *X* (en m^2)**
2395	24
1265	31
4050	45
1730	50
4350	60
3300	66
3600	72
4900	70
3600	90
5760	100



Le centre de gravité G du nuage de points a pour coordonnées $(\bar{x}, \bar{y})$, moyennes de *X* et de *Y*. Deux indices permettent de mesurer la liaison entre *X* et *Y* : ***la covariance***, égale à la moyenne des produits des écarts à la moyenne (mesurée ici en m^2 x €) et ***le coefficient de corrélation linéaire***, égal au rapport de la covariance par le produit des écart-types (indice

sans unité de mesure). On montre que ce coefficient est compris entre – 1 et +1 et qu'il indique une forte liaison linéaire négative lorsqu'il est proche de – 1, une forte liaison linéaire positive lorsqu'il est proche de +1 et une absence de liaison linéaire lorsqu'il est proche de 0.

Dans le cas d'une liaison linéaire, on propose un modèle de la forme $Y = a\,X + b$ qui traduirait, à une erreur additive près, la liaison entre les variables. Dans ce modèle, *X* est la ***variable explicative*** (dite aussi indépendante) alors que *Y* est la ***variable à expliquer*** (ou dépendante). Elles ne jouent pas des rôles symétriques. La recherche de *a* et *b* minimisant la somme des carrés des erreurs $\sum_{i=1}^{n}(y_i - ax_i - b)^2$ (***critère de moindres carrés***) conduit à la solution : $\hat{a} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)}$ et $\hat{b} = \bar{y} - a\bar{x}$ et la droite d'équation $y = \hat{a}\,x + \hat{b}$ solution du problème est appelée ***droite de régression linéaire*** de *Y* en *X* ; elle passe par le centre de gravité du nuage de points et elle est croissante lorsque la covariance est positive, décroissante lorsqu'elle est négative. Ce modèle permettra de « prévoir » le loyer d'un local commercial de même type en fonction de sa superficie. Un modèle a deux objectifs, qui peuvent être contradictoires, l'un est ***descriptif***, l'autre ***prédictif***.

On peut appliquer ce traitement statistique à l'étude d'une ***série chronologique,*** c'est-à-dire, d'une variable dépendant du temps *Y*(*t*), le chiffre d'affaire mensuel d'une entreprise par exemple. On fait alors jouer au temps le rôle de la variable explicative.

c2) Dans le cas de deux variables catégorielles *X* et *Y*, on désigne par $n_{i,j}$ l'effectif ***conjoint*** de la $i^{\text{ème}}$ modalité de *X* et $j^{\text{ème}}$ modalité de *Y*, n_{i+} l'effectif ***marginal*** de la $i^{\text{ème}}$ modalité de *X*, n_{+j} l'effectif ***marginal*** de la $j^{\text{ème}}$ modalité de *Y* et *n* l'effectif total. Le vocabulaire se réfère à la présentation de la distribution d'effectifs du couple de variables dans un tableau, appelé ***tableau de contingence***, avec, au centre, la distribution conjointe des effectifs du couple de variables et, dans les marges, les distributions marginales des effectifs des deux variables. On obtient les fréquences correspondantes $f_{i,j}$, f_{i+}, f_{+j} en divisant les effectifs par l'effectif total *n*. L'absence de liaison (ou ***indépendance***) entre les deux variables se traduit par :

$$n_{i,j} = n_{i+} \times n_{+j}\,/n \text{ ou } f_{i,j} = f_{i+} \times f_{+j}.$$

L'indice $n\sum_{i,j}\frac{(f_{i,j}-f_{i+}f_{+j})^2}{f_{i+}f_{+j}}$, appelé ***Khi2 de contingence***, est le carré d'une distance entre le tableau observé et le tableau d'indépendance. Il est proche de 0 lorsqu'il y a absence de liaison et il est d'autant plus élevé que la liaison est plus forte. C'est dans un cadre inférentiel que cet indice permet de tester l'indépendance des variables *X* et *Y*.

c3) Dans le cas d'une variable réelle et d'une variable catégorielle *X* et *Y*, on note $\bar{x}_k$ et σ_k^2 la moyenne et la variance de *X* sur les individus de la $k^{\text{ème}}$ catégorie de *Y*, d'effectif n_k, et on note toujours *n* l'effectif total, $\bar{x}$ et σ^2 la moyenne et la variance de *X*.
On vérifie la propriété suivante : $\bar{x} = \sum\frac{n_k}{n}\bar{x}_k$, la moyenne de *X* est égale à la moyenne pondérée des moyennes des catégories.
On pose : $\sigma^2_{\text{inter}} = \sum\frac{n_k}{n}(\bar{x}_k - \bar{x})^2$, variance des moyennes, et $\sigma^2_{\text{intra}} = \sum\frac{n_k}{n}\sigma_k^2$, moyenne des variances. On vérifie alors aussi : $\sigma^2 = \sigma^2_{\text{inter}} + \sigma^2_{\text{intra}}$.
Une liaison entre les deux variables est traduite par une forte homogénéité des individus à l'intérieur des catégories et une hétérogénéité entre les catégories (variance intra faible et variance inter forte). On introduit alors le ***rapport de corrélation*** $\eta = \frac{\sigma_{inter}}{\sigma}$, compris entre 0 et 1, proche de 0 en l'absence de liaison entre les variables, proche de 1 lorsque les variables sont liées.

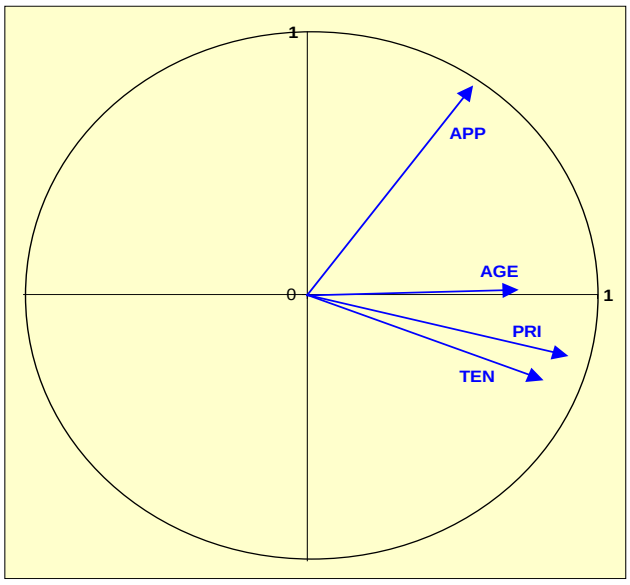
d) Analyses de données multidimensionnelles

Les analyses factorielles de données multidimensionnelles se sont en grande partie développées en France dans les années 60-80. Ces méthodes descriptives reposent sur l'algèbre linéaire et la géométrie euclidienne : l'espace de représentation des individus et celui des variables sont mis en dualité ; les résultats, présentés sous forme graphique, permettent de visualiser les liaisons entre variables et les proximités entre individus. Ces méthodes reposent sur les liaisons deux à deux des variables et utilisent donc les indices introduits dans le paragraphe précédent ; elles apportent une vision géométrique à des méthodes développées dans un cadre probabiliste dans la première moitié du XXème siècle.

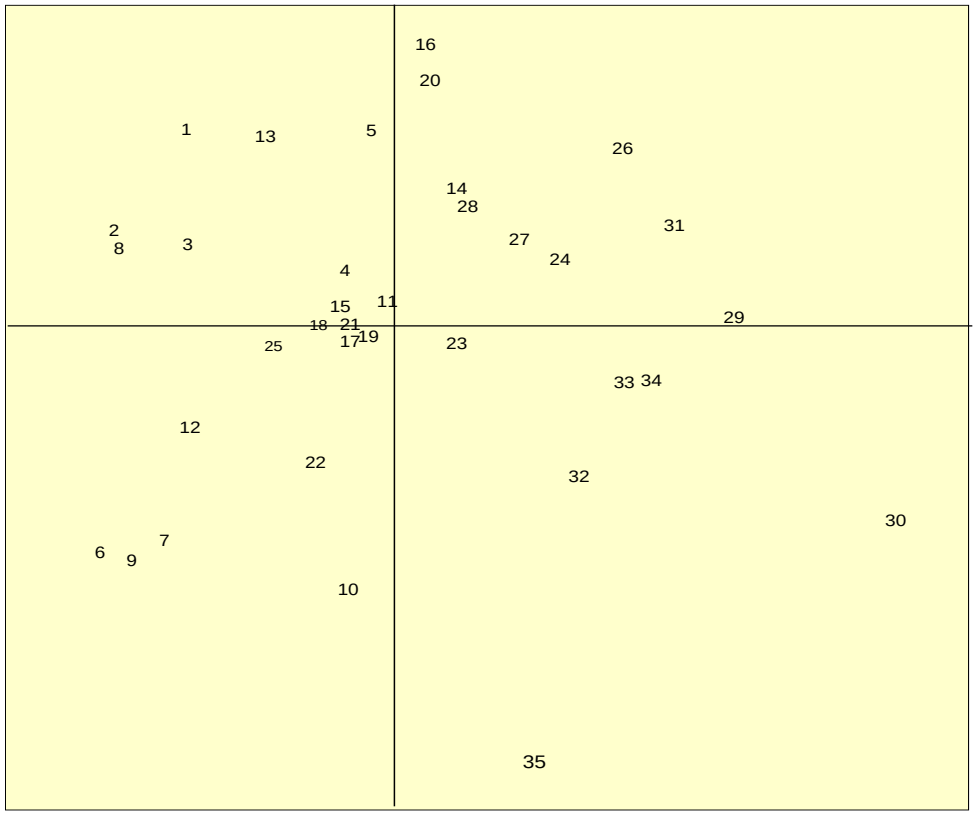
L'Analyse en Composantes Principales traite d'un ensemble de variables réelles, ***l'Analyse Factorielle des Correspondances Multiples*** de variables catégorielles, ***l'Analyse Factorielle Discriminante*** permet d'expliquer une variable catégorielle à partir de plusieurs variables réelles, ***l'Analyse Canonique*** permet d'étudier les corrélations entre deux groupes de variables réelles, … Contrairement aux trois autres analyses, qui sont descriptives et font jouer des rôles symétriques aux variables, l'analyse discriminante est explicative et sera utilisée dans un cadre de modélisation. On pourrait ajouter la ***Régression linéaire multiple***, dans sa version descriptive, qui consiste à expliquer une variable réelle en fonction d'un ensemble d'autres variables réelles.

Voici un exemple de sorties graphiques d'une **Analyse en Composantes Principales**. Il s'agit d'une étude sur le rapport qualité prix de 35 marques de whisky. Les quatre variables (prix au litre, âge, teneur en malt et appréciation d'un jury) sont centrées sur la moyenne et réduites (c'est-à-dire divisées par leur écart-type) et sont représentées sur ce graphique par des vecteurs. Le cosinus de l'angle entre deux vecteurs est égal au coefficient de corrélation des deux variables correspondantes. On observe ici que le prix est lié à l'âge et la teneur en malt des whiskies mais que l'appréciation donnée par le jury est non corrélée au prix. Sur le graphique représentant les 35 whiskies, graphique sur lequel on peut reporter les directions des vecteurs représentant les variables, on notera que le whisky 35 est cher mais pas très apprécié alors que les whiskies 26 et 31 sont appréciés et de prix moyen.

Représentation des variables



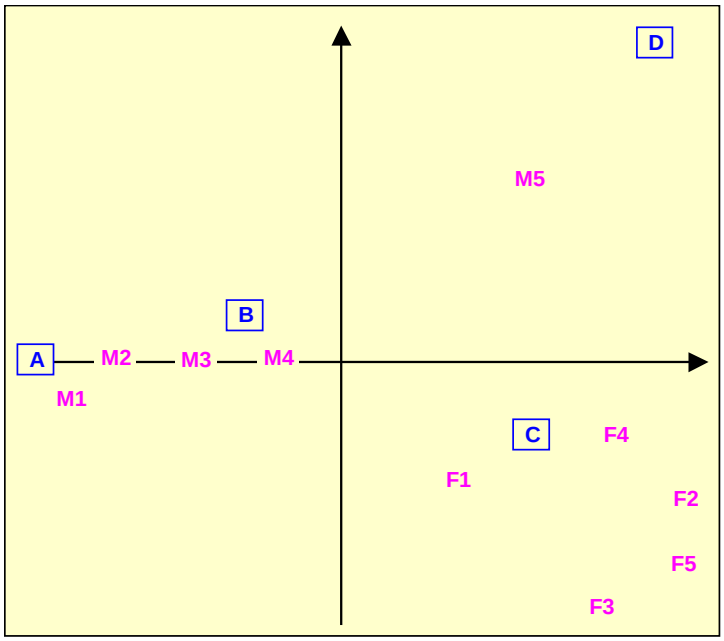
Représentation des 35 marques de whisky



Voici, à présent, un exemple **d'Analyse Factorielle des Correspondances**. Il s'agit de l'opinion sur la guerre du Vietnam de 3147 étudiants de l'Université de Chapel Hill (Caroline du Nord) en 1967. Ils ont à choisir une stratégie parmi quatre stratégies A, B, C, D qui vont de la plus guerrière (bombardements intensifs et invasion terrestre) à la plus pacifique (retrait immédiat des forces militaires du Vietnam).

La représentation des effectifs selon, d'une part, la stratégie choisie (4 modalités), d'autre part, le sexe et l'année d'étude (1ère à 5ème année) (10 modalités) est la suivante. On observe une opinion très concentrée autour de la stratégie D (arrêt des bombardements et ouverture de négociations) des filles alors que les garçons de 1ère année sont pour la stratégie la plus guerrière, que les opinions des garçons des années 2, 3 et 4 se décalent vers la stratégie B (statu quo) et que les garçons de 5ème année sont plus enclins à choisir la stratégie D.

Opinion sur la guerre du Vietnam selon l'âge et l'année d'étude des étudiants de l'Université de Chapel Hill en 1967



Les exemples présentés sont simples afin d'obtenir une lecture rapide et facile, mais il est évident que ces méthodes prennent tout leur intérêt lorsque l'on dispose d'un grand nombre d'individus et de variables, par exemple pour le traitement d'enquêtes.

Il existe également tout un ensemble de ***méthodes de classification*** d'individus ou de variables qu'il est difficile de présenter ici. Toutes ces méthodes sont dites descriptives ou exploratoires ; elles permettent de repérer des regroupements d'individus et des corrélations entre variables qu'il sera peut-être possible de confirmer, dans un deuxième temps, par des méthodes dites inférentielles ou confirmatoires.

2.3. Les limites et les extensions

Corrélation n'implique pas causalité. C'est le chercheur qui émet l'hypothèse que telle variable peut être la cause de telle autre. Une étude de corrélation pourra confirmer ou infirmer l'hypothèse. De plus, une corrélation positive entre deux variables peut être due à une troisième variable ; dégagée de l'influence de la troisième, les deux premières variables peuvent être non corrélées, voire corrélées négativement.

Ce paradoxe est connu sous le nom de *paradoxe de Simpson* et illustré par l'exemple suivant. Il s'agit de la sentence (condamnation à mort ou non) de 4764 meurtres jugés en Floride de 1973 à 1979. (Cf. Kripendorf, "Information Theory and Statistics", Wiley, 1986).
Selon la couleur de peau du meurtrier, on observe les résultats suivants, qui laissent penser que la sentence est plus sévère pour un meurtrier blanc que pour un meurtrier noir:

Meurtrier \ Condam à mort	Oui	Non	Taux
Blanc	72	2185	3.2%
Noir	59	2448	2.4%

Pourtant, lorsqu'on considère aussi la couleur de peau de la victime, on observe les résultats suivants. Quelle que soit la couleur de peau de la victime, la sentence est plus sévère pour un meurtrier noir que pour un meurtrier blanc.

		Condamnation à mort		
Victime	Meurtrier	Oui	Non	Taux
Blanche	Blanc	72	2074	3.4%
	Noir	48	239	16.7%
Noire	Blanc	0	111	0.0%
	Noir	11	2209	0.5%

Il faut donc se méfier d'une interprétation hâtive d'une liaison observée en terme de causalité. Des variables non observées peuvent être la cause de la liaison observée.

Les analyses de correspondances peuvent être utilisées sur des corpus de textes ou pour le traitement de réponses à des questions ouvertes (***analyse de données textuelles***). La Statistique n'est pas réservée au traitement de données quantitatives. Nombre de méthodes concernent l'étude des relations entre variables catégorielles c'est-à-dire l'étude des positions relatives de catégories d'individus. L'indicatrice d'une catégorie (variable prenant la valeur 1 si l'individu appartient à la catégorie, 0 sinon) est une variable réelle dont la moyenne est égale à la proportion p de la catégorie et la variance égale à $p\,(1-p)$. Ces indicatrices jouent un rôle considérable en statistique et probabilité pour le traitement de données qualitatives. Il est possible de construire une variable catégorielle à partir d'une variable réelle ; par exemple, la classe d'âge d'une population à trois modalités : moins de 25 ans, de 25 à 65 ans, plus de 65 ans. Lorsque l'on ne s'intéresse qu'aux effectifs et fréquences d'une variable et non aux valeurs de la variable, on la considère comme catégorielle.

3. Les Probabilités

3.1. Historique

On fait en général remonter les premiers calculs de probabilités au XVIème siècle en Italie, Tartaglia (1499-1577), Cardan (1501-1576), Galilée (1564-1642), essentiellement pour répondre à des questions de jeux de hasard.

En France, Pascal (1623-1662) et Fermat (1601-1665) ont échangé une importante correspondance sur ces sujets. Pascal écrit en 1654 un traité qu'il présente ainsi : « … traité tout à fait nouveau, d'une matière absolument inexplorée jusqu'ici, à savoir la répartition du hasard dans les jeux … Ainsi, joignant la rigueur des démonstrations de la science à l'incertitude du hasard, et conciliant ces choses en apparence contraires, elle peut s'arroger à bon droit ce titre stupéfiant : La Géométrie du Hasard ». Il faut savoir qu'à l'époque les mathématiciens étaient appelés géomètres.

La famille Bernoulli a joué ensuite un rôle étonnant. Ils sont cinq à avoir travaillé sur les probabilités avec des résultats plus ou moins importants : Jacques (1654-1705) et Jean (1667-1748) pour la première génération, Nicolas (1687-1759) et Daniel (1700-1782) pour la deuxième, enfin Jean (1744-1807) pour la troisième.

Bayes (1702-1761) a laissé son nom à une formule fondamentale, désignée à l'origine par « formule de probabilités des causes ». En effet, elle permet de calculer la probabilité a

posteriori d'une cause en fonction d'une probabilité a priori de cette cause et des probabilités des conséquences sous cette cause ou sous une cause alternative. Le fait de partir d'une probabilité ***subjective*** a priori et de la modifier par l'expérience conduit à une approche de la Statistique dite ***bayésienne*** ; cette approche diffère de l'approche dite ***objective*** ou ***fréquentiste*** présentée plus loin. Les probabilités subjectives sont particulièrement utiles pour la construction des arbres de défaillance dans le cadre de l'évaluation et de la gestion des risques, pour laquelle on manque de catastrophes pour utiliser les méthodes fréquentistes !

Laplace (1749-1827) est l'auteur d'une impressionnante œuvre sur le calcul des probabilités. Son « Essai philosophique sur les probabilités » est édité cinq fois de 1814 à 1825 et tous les savants du XIXème siècle ont lu cet essai. Il y énonce et commente, sans formules, les deux théorèmes fondamentaux de la théorie des probabilités : ***la loi des grands nombres*** et le ***théorème central limite***.

Il est pourtant déterministe : selon lui, les lois de la nature sont absolument déterministes, le hasard ne joue aucun rôle, le calcul des probabilités n'est utile que pour maîtriser les erreurs de mesure, pour corriger les faiblesses de nos instruments et de nos sens, en l'attente de progrès ultérieurs. Cette vision du hasard, dite ***épistémique***, s'oppose à la vision ***ontique*** pour laquelle le hasard est intrinsèque aux phénomènes physiques, hypothèse à la base de la théorie quantique.

Il est reproché à Laplace d'avoir, à cause de l'immense autorité qu'il avait à son époque, entraîné les scientifiques français sur la voie de garage déterministe dont ils n'ont pu se dégager qu'au début du XXème siècle.

Laplace et Gauss (1777-1855) ont laissé leurs noms à la célèbre loi de probabilité, appelée aussi ***loi normale*** que nous présentons plus loin.

En 1933, Kolmogorov fonde le calcul des probabilités sur la théorie de la mesure et de l'intégration, qui date du début du siècle pour l'ensemble des nombres réels mais qui est tout juste achevée sur un ensemble abstrait avec les théorèmes de décomposition de Lebesgue-Nikodym et l'existence de densités (1930).

Ainsi, les théories de la mesure et de l'intégration ont jeté les bases d'une formulation homogène et cohérente des probabilités en y introduisant la rigueur et le raisonnement mathématique qui lui manquaient ; rappelons en effet que Keynes, en 1921, écrivait à propos des probabilités que « les savants y décèlent un relent d'astrologie ou d'alchimie » et que von Mises en 1919 affirmait que « le calcul des probabilités n'est pas une discipline mathématique ».

3.2. Quelques concepts et les deux théorèmes fondamentaux

Nous avons vu en introduction que les ***Probabilités*** ont pour objet d'étude les phénomènes ***aléatoires***.

Lors d'une expérience aléatoire, on commencera par affecter aux différents résultats possibles des réels positifs de somme 1 avant de déduire d'autres résultats à l'aide du calcul des probabilités. La première étape est une étape de modélisation qui repose sur des hypothèses ou sur l'expérience. Si on lance un dé équilibré, on affectera la probabilité 1/6 à chacune des six faces du dé. L'hypothèse de dé équilibré est traduite par ***l'équiprobabilité*** des différentes faces. C'est ***l'approche classique*** ou laplacienne des probabilités car c'est Laplace qui a énoncé le premier la règle selon laquelle la probabilité d'un événement est égale au rapport du nombre de cas favorables sur le nombre de cas possibles, à condition que tous les cas possibles aient la même chance de se produire. On peut aussi lancer un grand nombre de fois le dé et constater que la fréquence de sortie de chacune des six faces s'approche de 1/6

lorsque le nombre d'épreuves augmente. Il s'agit alors de ***l'approche fréquentiste*** des probabilités ; c'est par l'expérience que l'on peut se faire une idée de la notion de probabilité.

Si on note *X* le chiffre qui apparaît lors du lancer d'un dé, *X* est une ***variable aléatoire*** prenant ses valeurs dans l'ensemble {1, …, 6} avec même probabilité 1/6 pour chacune des 6 valeurs. On peut à partir de là représenter l'histogramme de cette distribution de probabilité (appelée aussi ***loi de probabilité***) comme on a représenté, en statistique descriptive, la distribution de fréquence d'une variable réelle. On peut également résumer la distribution par la moyenne (appelée aussi ***espérance mathématique***) et par l'écart-type.

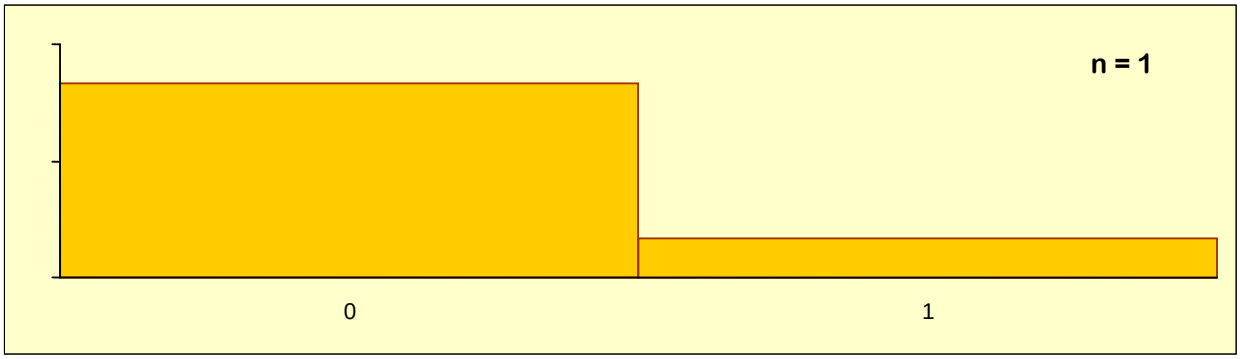
Dès que l'on répète une expérience aléatoire, se pose le problème de ***l'indépendance en probabilité*** des événements concernant la première expérience avec ceux concernant la deuxième. Intuitivement, si d'une urne contenant deux boules, une rouge et une blanche, on extrait au hasard et successivement les deux boules (***tirage sans remise***), la couleur de la seconde boule tirée dépend de la couleur de la première. En revanche, si l'on remet la première boule tirée dans l'urne avant de tirer la seconde (***tirage avec remise***), tirer une boule rouge ou une boule blanche lors du premier tirage n'a aucune influence sur le résultat du deuxième tirage (indépendance en probabilité des résultats des deux tirages). Dans le cadre de la statistique descriptive de deux variables catégorielles, on a vu que l'indépendance des deux variables est traduite par la propriété suivante : « la fréquence conjointe de la $i^{\text{ème}}$ modalité de la première variable avec la $j^{\text{ème}}$ modalité de la deuxième variable est égale au produit des fréquences marginales correspondantes ». L'indépendance en probabilité de deux variables aléatoires est traduite par la même propriété, en remplaçant les fréquences par les probabilités.

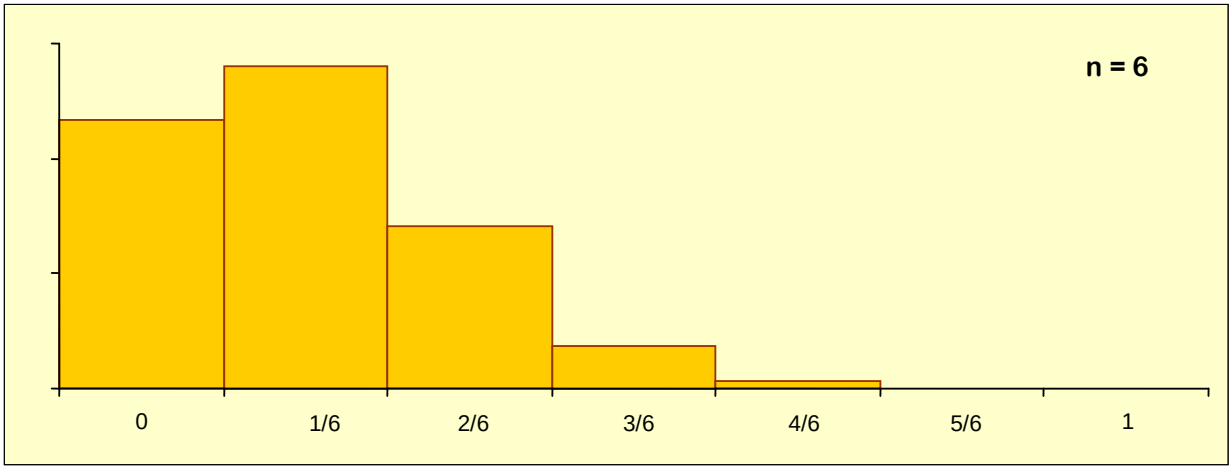
L'expérience consistant à répéter de façon indépendante une même épreuve aléatoire dont on note *p* la probabilité de succès (par exemple, le tirage avec remise dans l'urne ci-dessus, en appelant succès le fait d'obtenir la boule rouge) est appelée ***schéma de Bernoulli*** et le nombre *X* de succès sur *n* épreuves est une variable aléatoire dont la loi de probabilité est la ***loi binomiale*** (de paramètres *n* et *p*). On montre que cette variable aléatoire a pour moyenne *np* et pour écart-type $\sqrt{np(1-p)}$. On en déduit que la ***fréquence d'échantillonnage***, c'est-à-dire, la fréquence de succès sur *n* épreuves $F = X / n$ a pour moyenne *p* et pour écart-type $\sqrt{p(1-p)/n}$. Aussi, lorsque *n* tend vers l'infini, l'écart-type tend vers zéro et la fréquence de succès a pour limite une constante égale à la probabilité *p* de succès lors d'une épreuve. Ce résultat est la première version, obtenue par Jacques Bernoulli, de ***la loi des grands nombres*** et justifie l'approche fréquentiste des probabilités.

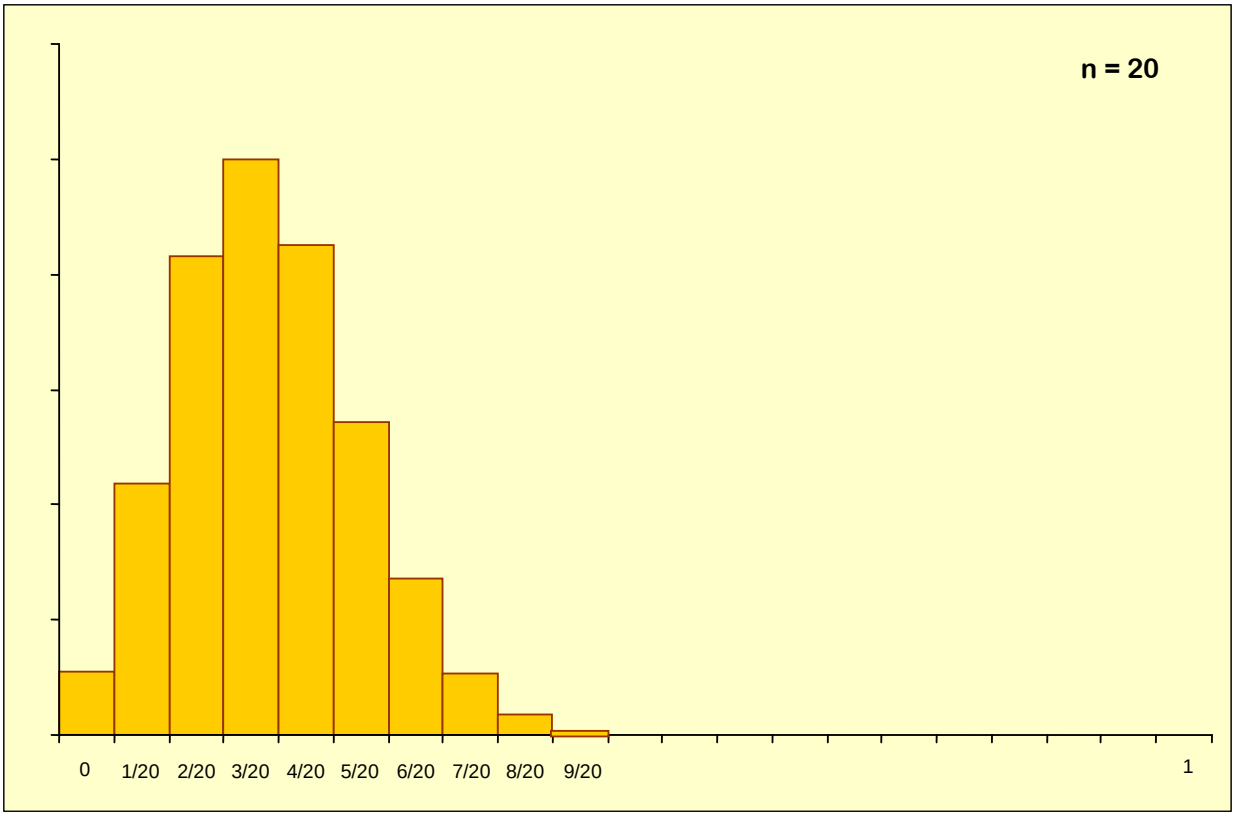
Dans sa version plus générale, obtenue par Laplace, la loi des grands nombres exprime que la ***moyenne d'échantillonnage***, c'est-à-dire, la moyenne de *n* variables aléatoires indépendantes et de même loi, d'espérance mathématique μ et d'écart-type σ, converge (en probabilité) vers μ lorsque *n* tend vers l'infini. Cette loi est annoncée comme universelle : Poisson (1781-1840) annonce la mécanique statistique en affirmant que la loi des grands nombres peut s'appliquer aux molécules des gaz. Elle s'applique même à la société (cf. la physique sociale de Quételet).

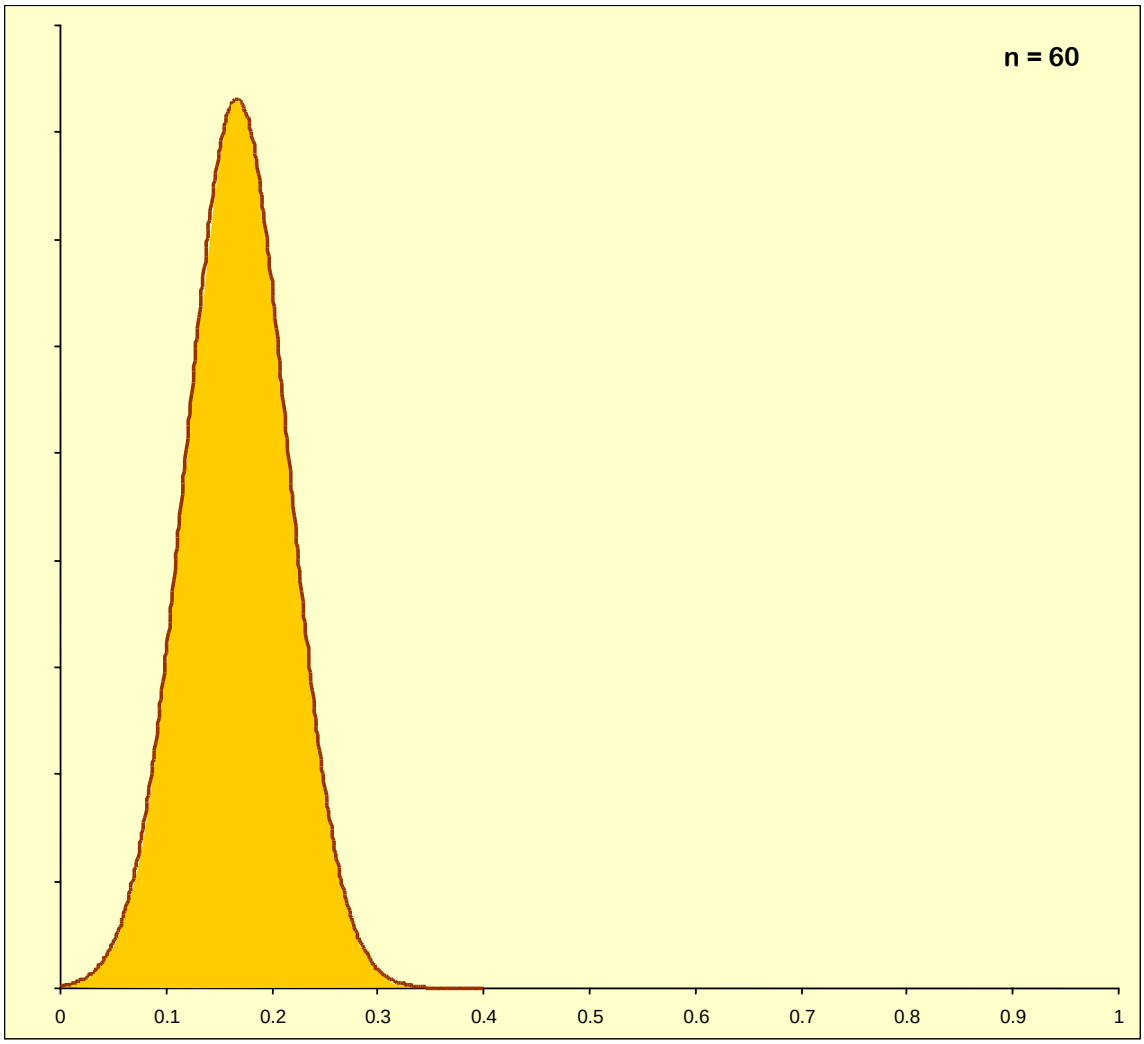
Le théorème central limite précise la vitesse de convergence de la moyenne d'échantillonnage vers la moyenne μ. (et donc aussi de la fréquence d'échantillonnage vers la probabilité *p)*. On en déduit que la loi de probabilité de la moyenne (ou fréquence) d'échantillonnage, pour un nombre d'épreuves *n* assez grand, peut être approchée par une loi normale ce que nous pouvons visualiser par l'exemple suivant.

On lance un dé *n* fois de suite et on s'intéresse à la fréquence d'obtention du chiffre 1. Les distributions de probabilité de la fréquence pour : *n* = 1, *n* = 6, *n* = 20 et *n* = 60 sont représentées ci-après. Les résultats sont représentés par des intervalles pour mieux visualiser la convergence. Pour *n* "infini", la probabilité est concentrée sur la probabilité 1/6 d'obtenir le chiffre 1.







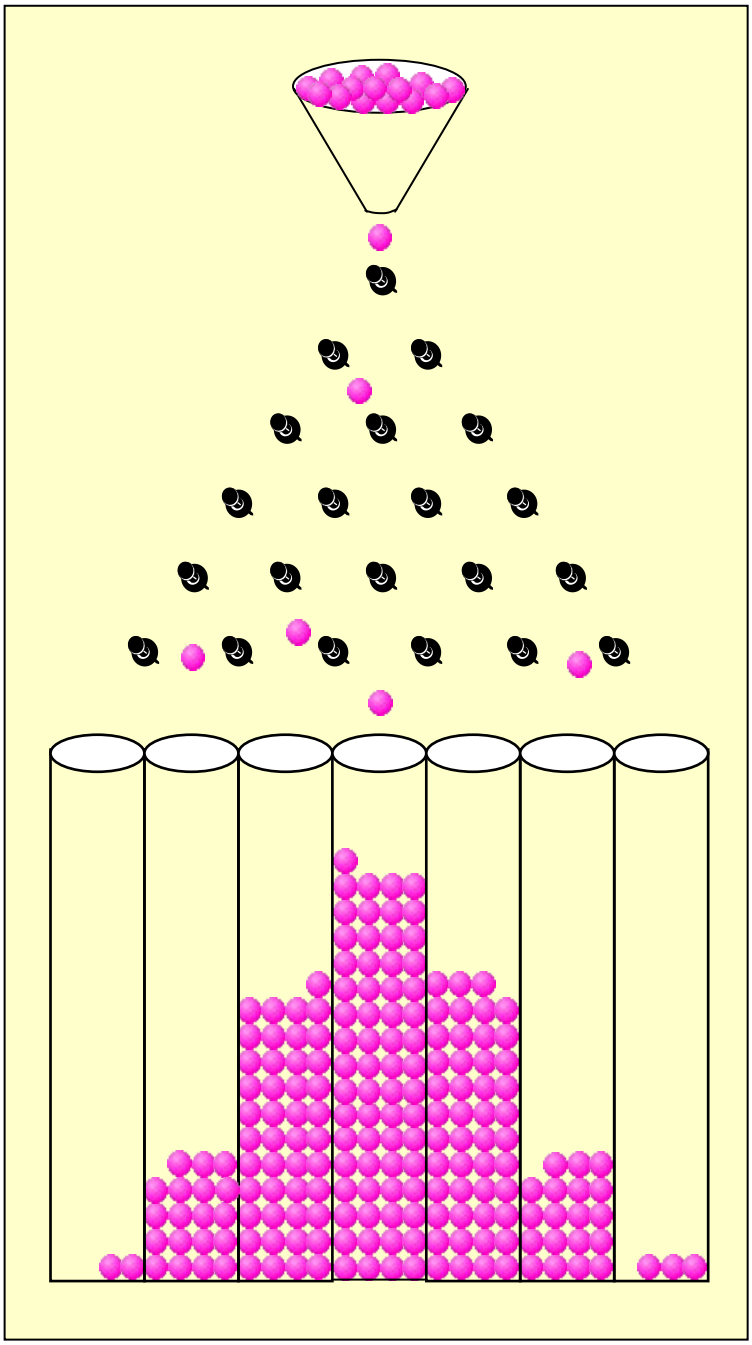


La ***planche de Galton*** est une illustration physique de la loi binomiale et de la loi normale. On lâche une à une un grand nombre de billes sur une planche inclinée sur laquelle sont plantés des clous, d'abord un clou, puis deux clous, trois clous, …, jusqu'à six clous, placés en quinconce et sept godets récupèrent les billes. Chaque bille bute sur le 1er clou et dévie sur la gauche ou sur la droite avec probabilité ½, elle bute ensuite sur un des deux clous de la deuxième rangée et dévie de même sur la gauche ou sur la droite avec probabilité ½, etc.

A la fin de l'expérience, on observe que les hauteurs de billes dans les godets sont symétriques et décroissantes par rapport au godet central. Si on prend comme unité de hauteur celle des deux godets extérieurs, on obtient, en remontant vers le godet central, les hauteurs 1, 6, 15 et 20.

Pour une loi binomiale de paramètres $n = 6$ et $p = ½$ les probabilités d'obtenir 0, 1, 2, 3, 4, 5, 6 succès (le succès correspondant à une bifurcation vers la droite par exemple) sont égales à 1/64, 6/64, 15/64, 20/64, 15/64, 6/64, 1/64 respectivement ; l'histogramme obtenu pour cette distribution de probabilité est celui fourni par les billes dans les godets de la planche de Galton.

Schématisation de la planche de Galton



Si, à présent on augmente le nombre n de rangées de clous (et donc le nombre de godets, égal à une unité de plus que le nombre de rangées de clous) alors l'histogramme obtenu se rapproche de celui obtenu sous la courbe en cloche.

On ne compte plus les propriétés d'optimalité de la loi normale. Elle intervient dans des domaines parfois inattendus. Citons son lien avec la théorie de l'information. ***La théorie de l'information*** s'est développée dans les années 1940 à partir des travaux de Shannon et Wiener et a pour objet la mesure du degré de complexité d'un système.

La notion d'***entropie*** en est un des outils principaux. Si un système est décrit par une variable aléatoire admettant un nombre fini de valeurs x_i , $i = 1, \ldots, n$ avec les probabilités p_i

($\sum p_i = 1$), alors l'entropie statistique mesure le degré d'indétermination du système ou la quantité d'information à laquelle on ne peut pas accéder.

Elle doit vérifier trois propriétés :

1) elle est nulle quand un événement est certain,

2) elle est maximale dans le cas d'équiprobabilité,

3) l'entropie d'un couple de variables aléatoires indépendantes est égale à la somme des entropies des variables individuelles.

Si on pose : $H = -\sum p_i \, ln(p_i)$, alors cette fonction vérifie les trois propriétés. L'analogue pour une variable aléatoire admettant une densité *f* sera : $H = -\int f(x) \ln f(x) dx$. On montre que parmi toutes les densités de probabilité paires, de moyenne 0 et d'écart-type 1, la densité de la loi normale centrée réduite maximise l'entropie. Elle est ainsi bien adaptée pour la modélisation de système sur lequel on a peu d'information.

3.3. Les limites et les extensions

Le paradigme statistique qui vient d'être décrit ne convient que pour le hasard que Mandelbrot (le père de la géométrie fractale) appelle ***le hasard bénin*** et qu'il oppose au ***hasard sauvage.***

La loi des grands nombres et le théorème central limite, même s'ils peuvent être encore généralisés, supposent que les variables aléatoires dont on étudie la moyenne d'échantillonnage admettent une moyenne et un écart-type.

Toutes les variables aléatoires n'ont pas une telle propriété. Par exemple, l'histogramme de la loi de Cauchy ressemble à une courbe en cloche mais les queues de la distribution sont plus épaisses et sa moyenne est infinie.

Les théorèmes fondamentaux ne s'appliquent pas à de telles variables. Pourtant, force est de constater que l'on rencontre des phénomènes aléatoires qui relèvent de ce hasard sauvage : historiquement, la série des crues du Nil, quotidiennement les cours de la bourse, mais aussi l'intensité des courants électriques à travers les lames métalliques fines et de nombreux autres phénomènes physiques.

Entre le hasard bénin et le hasard sauvage, Mandelbrot définit aussi le ***hasard lent*** pour lesquels les théorèmes limites s'appliquent mais la convergence est tellement lente que la loi normale ne peut pas être utilisée comme approximation dans les problèmes réels. La variable lognormale est un exemple de ce hasard lent.

4. La Statistique Inférentielle

4.1. Historique

C'est entre 1885 et 1935 qu'émerge la statistique mathématique. Le contexte est favorable : d'une part, il existe déjà des outils statistiques puissants, tels que les moindres carrés, la régression, la corrélation, la théorie des erreurs, d'autre part, les théories mathématiques sur lesquelles la statistique va s'appuyer sont en plein essor, enfin, la prise en compte de l'aléatoire devient de plus en plus nécessaire, non seulement en astronomie mais aussi en agronomie, biométrie, psychométrie, économétrie, …

Les résultats les plus importants (théorie de l'estimation, théorie des tests) sont obtenus par l'école anglaise (Pearson, père et fils, Fisher, Neyman, Spearman, Student).

Ce n'est qu'à la fin de cette période que Neyman publie les premiers résultats relevant de la ***théorie des sondages*** (sondage stratifié proportionnel, sondage stratifié optimal, sondage en grappes, ...) qui seront utilisés par les instituts de statistique.

En 1895, lors du congrès international de statistique, un statisticien norvégien, Kiaer, propose une communication faisant état de deux expériences norvégiennes d'enquêtes « partielles » (c'est-à-dire, sur échantillons) ayant fourni de bons résultats.

Il introduit la notion ***d'échantillon représentatif*** bien qu'il ne la définisse pas explicitement ; évidemment la qualité est mesurée par des enquêtes exhaustives (recensements), les seules autorisées à l'époque avec les ***monographies***, qui pouvaient être considérées comme des recensements d'une infime partie d'une population.

Les réactions de l'assemblée ont été extrêmement violentes et les débats se sont poursuivis pendant quarante ans, tout d'abord sur la controverse recensement versus sondage, puis, parmi les convaincus de l'intérêt des sondages, sur la controverse échantillon par ***choix d'expert*** versus échantillon par ***choix au hasard*** (c'est-à-dire selon une procédure aléatoire contrôlée). Il a bien fallu admettre que le hasard était préférable aux experts … que seul le hasard était capable de tenir compte de l'influence possible sur les résultats des variables cachées.

A condition de se placer dans le cadre de l'échantillonnage aléatoire, il sera alors possible d'écrire une théorie des sondages (Neyman, 1934) fondée sur la statistique inférentielle développée peu avant.

Le premier à appliquer cette théorie de façon éclatante est Gallup aux États-Unis, en 1936, pour prévoir qui serait le futur président. Les grands journaux avaient l'habitude de lancer à cette occasion de vastes consultations, appelées "votes de paille" ; l'idée qui prévalait à l'époque était que « plus grand est l'échantillon, plus près d'un recensement on se situe et meilleurs sont les résultats ». Le Litterary Digest annonce ainsi la victoire du conservateur Landon à partir d'un échantillon de plus de 2 millions de votes, alors que Gallup annonce la réélection du démocrate Roosevelt à partir d'un échantillon de 4000 personnes, choisi en tenant compte des principes mis en évidence par Neyman ! La victoire de Gallup est éclatante.

La même année, il ouvre l'institut qui porte son nom et, l'année suivante, l'IFOP (Institut Français de l'Opinion Publique) est créé en France. On sait depuis le succès que connaissent les sondages d'opinion !

Depuis 1989, on découvre les travaux réalisés en URSS. Il apparaît ainsi que les résultats, dits de Neyman, ont été publiés par Kovalevsky en 1924 et immédiatement utilisés au niveau de l'Etat.

4.2. Échantillonnage, estimation, tests

On a vu en introduction que la Statistique Inférentielle a pour objectif d'inférer à la population tout entière les résultats obtenus sur un échantillon. Il s'agit en fait d'estimer des paramètres concernant la population (proportion, moyenne, écart-type, …) à partir des indices analogues calculés sur l'échantillon. Dans le cas de données expérimentales, il s'agit d'estimer les paramètres de la loi de probabilité dont on suppose que les données peuvent être issues.

On distingue alors trois parties :

- la théorie de l'échantillonnage qui est la partie déductive : elle consiste à établir les lois de probabilités de toutes les variables d'échantillonnage susceptibles d'être utilisées (moyenne d'échantillonnage, variance d'échantillonnage, …),

- la théorie de l'estimation qui consiste à établir des critères (méthodes des moments, moindres carrés, maximum de vraisemblance, …) permettant d'estimer les paramètres

inconnus concernant la population ou la loi de probabilité du modèle à partir de variables d'échantillonnage et à étudier les propriétés de tels estimateurs (estimateurs sans biais, convergents, efficaces …), enfin

- la théorie des tests qui consiste à infirmer ou non une hypothèse, posée sur la population ou sur la loi de probabilité du modèle, à partir des observations obtenues sur un échantillon.

Il est impossible de présenter ces théories, seul un exemple d'estimation par intervalle de confiance et un exemple de test statistique seront présentés.

a) Estimation par intervalle de confiance d'une proportion

On pose à 1000 français adultes la question : « Avez-vous confiance en votre premier ministre ? » ; 64% des personnes interrogées répondent « oui ». Peut-on inférer ce résultat à la population adulte française ?

S'il s'agit d'un échantillon aléatoire simple (c'est-à-dire si tous les échantillons de 1000 adultes français ont la même probabilité d'être tirés), et si on note p la proportion inconnue d'adultes français ayant confiance en leur premier ministre, alors la fréquence observée f = 0.64 (qui n'est autre que la proportion calculée sur l'échantillon) est l'observation d'une variable aléatoire F de moyenne p et d'écart-type $\sqrt{p(1-p)/1000}$. On sait de plus (théorème central limite) que la loi de F peut être approchée par une loi normale de mêmes paramètres. Enfin, on a la propriété suivante sur les lois normale : 95% des valeurs d'une variable normale sont comprises entre la moyenne moins deux écart-types et la moyenne plus deux écart-types. On en déduit que 95% des observations f sont comprises dans l'intervalle $\left[p-2\sqrt{p\left(1-p\right)/1000}\ ;\ p+2\sqrt{p\left(1-p\right)/1000} \right]$ appelé ***intervalle de fluctuation*** de la fréquence d'échantillonnage à 95%.

Approximation : pour une proportion p, qui est un nombre réel compris entre 0 et 1, on montre que $p(1-p)$ est compris entre 0 et 1/4, et égal à 1/4 pour p = 1/2 ; aussi l'intervalle de fluctuation de la fréquence d'échantillonnage à 95% est inclus dans l'intervalle suivant, de forme très simple et utilisé comme approximation lorsque p est compris entre 0.2 et 0.8 : $\left[p-1/\sqrt{1000}\ ;\ p+1/\sqrt{1000} \right]$ c'est-à-dire p plus ou moins 3%.

Comme 95% des échantillons de taille 1000 fournissent une fréquence f appartenant à cet intervalle, **étant donné un échantillon** sur lequel on observe f = 64%, une ***estimation ponctuelle*** de p est f = 64%, et une estimation de p par ***intervalle de confiance*** à 95% est : f plus ou moins 3%, c'est-à-dire ici [61% ; 67%]. C'est la fameuse fourchette de sondage.

En résumé, dans le cas d'un échantillon aléatoire de taille n, sur lequel est observée la fréquence f, pour $n \geq 30$ et f compris entre 0.2 et 0.8, l'estimation par intervalle de confiance à 95% de la proportion inconnue p sur la population est donnée par :

$$\left[f-1/\sqrt{n}\,,\ f+1/\sqrt{n} \right].$$

On notera que la ***précision*** de l'estimation de p, d'autant plus grande que l'intervalle est petit à un niveau de confiance donné, ne dépend pas de la taille N de la population ni du taux de sondage n/N. Ces notions sont au programme de 2nde en France depuis la rentrée 2009.

b) Test d'hypothèse d'égalité de deux proportions

Si, dans les mêmes conditions d'échantillonnage, on avait obtenu le mois précédent 62% de « oui », peut-on dire que la côte de popularité du premier ministre a augmenté ?

On note p_1 et p_2 les proportions inconnues de « oui » le mois précédent et le mois actuel sur la population, f_1 et f_2 les fréquences de « oui » de l'échantillon du mois précédent et de l'échantillon du mois actuel ; alors on observe : $f_2 - f_1 = 64\%–62\%$ soit 2%.

Faisons l'hypothèse que la côte de popularité n'ait pas évolué, c'est-à-dire $p_1 = p_2$; est-il possible d'obtenir, sur deux échantillons aléatoires simples et indépendants chacun de taille 1000, une différence de fréquences de oui égale à 2% ? Autrement dit, la différence de 2% observée entre les deux échantillons est-elle due à la fluctuation naturelle de l'échantillonnage ou révèle-t-elle une différence entre les proportions sur la population ?

Encore une fois, nous nous fixons une probabilité d'erreur, par exemple 5%. Il s'agit alors de calculer la probabilité que la différence soit supérieure ou égale à 2% sous l'hypothèse $p_1 = p_2$; si cette probabilité est inférieure à 5%, on conclut que les résultats ne sont pas compatibles avec l'hypothèse et on rejette l'hypothèse ; dans le cas contraire, les résultats sont jugés compatibles avec l'hypothèse et on ne rejette pas l'hypothèse. Finalement, le seuil de 5% fixé à l'avance correspond à la probabilité de rejeter l'hypothèse $p_1 = p_2$ alors qu'elle est "vraie". Sans poursuivre le développement du test, dans notre exemple, les résultats obtenus pour l'intervalle à 95% de confiance nous laissent entrevoir que la différence de 2% peut être due à l'échantillonnage ; cette différence est dite non significative au seuil de 5%.

Un exemple historique de l'application de la statistique inférentielle est l'étude réalisée par Fisher sur les données expérimentales à partir desquelles Mendel découvrit les lois fondamentales de l'hérédité. Fisher a montré que les données d'observations de Mendel sont tellement proches de sa théorie que, si l'on calculait la probabilité d'obtenir les résultats qu'il a trouvés expérimentalement, on arriverait à une probabilité de 7 sur 100 000. Mendel aurait falsifié des données afin de faire avancer sa théorie qui lui semblait irréfutable. Cet argument serait renforcé par la présence d'altérations sur les feuilles de mesures de Mendel. Si c'est le cas, Mendel pensait que les fluctuations observées étaient dues à des erreurs de mesure alors qu'il s'agissait de fluctuations naturelles de l'échantillonnage.

4.3. Les limites et les extensions

a) Tests non paramétriques pour les échantillons de petites tailles

Évidemment, dans le cas d'échantillons de petites tailles, il n'est plus possible d'utiliser l'approximation normale.

On veut, par exemple, tester l'efficacité d'une crème pour la peau ; 12 personnes ont utilisé la crème, 12 ont utilisé un placebo. L'hypothèse d'absence d'efficacité se traduit par « les 24 observations proviennent d'une même distribution ». On range alors les 24 valeurs observées de la plus petite à la plus grande en retenant de quel échantillon elles proviennent et on étudie comment les rangs des deux échantillons se situent dans la liste ordonnée de 1 à 24 des 24 rangs.

Des calculs de probabilités permettent de conclure que la disposition observée peut provenir des fluctuations naturelles d'échantillonnage ou, au contraire, que la disposition observée est très peu probable, ce qui nous conduit à rejeter l'hypothèse.

Ces tests sur les rangs sont des ***tests non paramétriques*** dans le sens où l'on ne raisonne plus sur des paramètres résumant les distributions mais sur les distributions elles-mêmes. Ce champ se développe beaucoup grâce à la puissance des ordinateurs.

b) Retour sur les sondages aléatoires stratifiés

Supposons que l'on désire connaître le loyer annuel moyen d'un logement de l'agglomération toulousaine en l'an 2010, que l'on dispose de la liste de tous les logements en location, appelée ***base de sondage***, mais qu'il semble impossible de réaliser une enquête exhaustive.

Bien souvent dans la base de sondage, on dispose d'informations auxiliaires qui permettent de constituer des partitions de la population. On dispose ainsi, dans notre exemple, de la zone géographique (6 zones), du type de logement (5 types) et d'autres renseignements dont nous ne tiendrons pas compte. Ces deux variables croisées forment une partition de la population de logements en 30 sous-populations que nous appellerons ***strates***, et l'on connaît, pour chacune d'elles, la taille de la strate c'est-à-dire le nombre de logements en location.

Plutôt que de tirer un échantillon aléatoire simple de taille 1000, on se propose de tirer 30 échantillons aléatoires simples, un dans chaque strate, dont les tailles sont proportionnelles aux tailles des strates (avec une somme des tailles égale à 1000). On parle alors ***d'échantillon stratifié proportionnel***. On a constitué ainsi un modèle réduit de la population.

L'estimateur de stratification du loyer moyen de l'agglomération toulousaine est beaucoup plus précis que l'estimateur construit à partir d'un échantillon aléatoire simple. Il n'est pas rare d'obtenir pour une même taille globale de l'échantillon, un intervalle de confiance dix fois plus petit avec un échantillon stratifié proportionnel qu'avec un échantillon aléatoire simple.

Le second résultat important est que, parmi les échantillons stratifiés de même taille globale, la répartition proportionnelle n'est pas forcément la meilleure. Si l'on a une strate de 5000 logements de même et bas standing, sensiblement de même loyer et donc avec un écart-type faible, et une autre strate de 250 logements de standing plus élevé et plus varié, donc un écart-type important, un échantillon stratifié proportionnel conduit à ce que la taille de l'échantillon de la première strate soit 20 fois celle de l'échantillon de la deuxième, par exemple 200 et 10. Intuitivement, on a envie de réduire la taille de l'échantillon de la première strate où l'on sait avoir peu de variabilité pour augmenter la taille de l'échantillon de la deuxième strate où il y a une plus grande variabilité et on aura raison. On montre en effet que, à taille globale fixée, ***l'échantillon stratifié optimal*** (celui qui donne la fourchette la plus petite) conduit à une taille proportionnelle au produit de la taille de la strate et de l'écart-type de la strate. Il faut charger l'échantillon aux endroits de plus grande variabilité.

c) Méthode empirique des quotas

Pour terminer sur les sondages disons deux mots de ***la méthode des quotas***. C'est une méthode empirique. Elle ressemble à la stratification proportionnelle mais le choix de chaque échantillon ne se fait pas selon une procédure aléatoire à partir d'une base de sondage ; c'est l'enquêteur qui choisit les individus de l'échantillon en respectant certains quotas, c'est-à-dire un nombre fixé à l'avance d'individus à interroger pour chaque catégorie de quelques

variables de contrôle (par exemple, région d'habitation, sexe, classe d'âge, catégories sociales et professionnelles), variables que l'on pense liées au sujet de l'enquête.

Dans les sondages à plusieurs degrés, par exemple, « régions », « classes de taille de communes », « communes », « quartiers », « individus », il est possible de faire un sondage aléatoire stratifié au premier degré et au second degré (toutes les régions et toutes les classes de taille de communes sont enquêtées), un échantillonnage aléatoire au troisième degré à partir d'une base de sondage de communes déjà existante, puis de découper en quartiers les communes sélectionnées pour constituer les bases de sondage aléatoire du quatrième degré, enfin de faire une enquête par quotas au niveau des individus dans les quartiers sélectionnés. L'INSEE a testé avec succès cette méthodologie pour des enquêtes répétées, après l'avoir contrôlée par une méthode purement aléatoire pour vérifier que les variables de quotas étaient correctement choisies.

Le problème de la non-réponse dans les sondages aléatoires est difficile à résoudre. Lorsqu'un individu sélectionné ne répond pas, il n'est pas possible de le remplacer au pied-levé par un autre qui lui « ressemblerait ». Pour les enquêtes par quota, il suffit de respecter les quotas et il faut parfois interroger dix personnes pour en trouver une qui accepte de répondre. Si l'ensemble des répondants et l'ensemble des non-répondants n'ont pas la même structure d'opinion par rapport au sujet de l'enquête, le risque d'avoir une estimation biaisée est important.

5. La Modélisation Aléatoire

5.1. Historique

C'est par souci de simplification qu'ont été présentées séparément les méthodes descriptives, la statistique inférentielle et la modélisation aléatoire (on dit indifféremment modèle aléatoire, modèle stochastique, modèle probabiliste ou modèle statistique). C'est a posteriori qu'il peut être intéressant de faire la part, dans un modèle aléatoire, entre les propriétés algébriques et géométriques du modèle, l'aléatoire dû au modèle et l'aléatoire dû à l'échantillonnage.

5.2. Modèle linéaire, analyse de variance, plans d'expériences

La modélisation non aléatoire consiste à relier des variables statistiques entre elles. Nous avons déjà décrit dans la partie « statistique descriptive », le modèle le plus simple entre deux variables réelles X et Y : on suppose qu'il existe deux paramètres réels a et b tels que l'on ait : $Y = aX + b$. A partir d'un échantillon de n observations (x_i, y_i), i = 1, …, n, un ***critère des moindres carrés*** donne pour solution la droite de ***régression linéaire*** de Y en X.

Dans sa version aléatoire la plus simple, ***le modèle linéaire*** s'écrira $Y = a\,X + b + \varepsilon$ où ε est une variable aléatoire normale centrée d'écart-type σ et où la variable X est supposée non aléatoire ; on aura alors pour les n observations l'égalité : $y_i = a\,x_i + b + \varepsilon_i$ et les variables aléatoires (ε_i) seront supposées indépendantes.

On montre alors que les estimateurs des moindres carrés possèdent de bonnes propriétés et, connaissant de nouvelles valeurs de X, on pourra prévoir les valeurs de Y à l'aide du modèle.

On peut disposer de plusieurs variables explicatives, le modèle linéaire s'écrira alors $Y = a_1 X_1 + \ldots + a_p X_p + b + \varepsilon$ avec les mêmes hypothèses sur l'erreur aléatoire. Les variables Y, X_1, …, X_p peuvent être le résultat de transformations sur d'autres variables (logarithme, carré, quotient, …) ; la linéarité s'entend en fait par rapport aux coefficients (cadre de l'algèbre linéaire et la géométrie euclidienne). Utilisés dans le champ économique, ces modèles sont souvent appelés ***modèles économétriques***.

Les variables explicatives peuvent être catégorielles, on affectera alors un coefficient à chaque catégorie, il s'agit alors de ***l'analyse de variance*** (connue sous le nom d'ANOVA).

Reprenons l'exemple du loyer annuel moyen des logements de l'agglomération toulousaine. On note Z le loyer annuel, $X_1,\ldots,X_6$ les indicatrices des zones et $Y_1,\ldots,Y_5$ les indicatrices des types de logement. On rappelle que l'indicatrice d'une catégorie affecte la valeur 1 à un logement de cette catégorie, 0 sinon.

Le modèle s'écrit : $Z = a_1 X_1 + \ldots + a_6 X_6 + b_1 Y_1 + \ldots + b_5 Y_5 + c + \varepsilon$. Les indicatrices des catégories d'une même variable catégorielle sont liées linéairement puisque leur somme est égale à 1, aussi on posera $a_6 = b_5 = 0$ et le coefficient c sera estimé par le loyer moyen d'un logement de la zone 6 et du type 5.

Au lieu d'estimer, à partir de l'échantillonnage, les loyers des 30 strates, ce qui ne manquera pas de donner de très mauvais résultats pour les strates ayant un faible échantillon, on estime seulement 10 paramètres et l'on prédit par le modèle les loyers de chacune des 30 strates.

Dans le cas d'un modèle plus complexe avec 10 variables catégorielles (28 zones, 5 types, 5 standings, …) on peut arriver, en croisant toutes les variables, à plus d'un million de « cellules » différentes, qu'il est impossible d'utiliser comme strates d'un sondage stratifié. Un modèle est alors préférable car l'estimation d'une cinquantaine de paramètres seulement permettra de prévoir le loyer moyen de n'importe quelle cellule et donc le loyer d'un logement toulousain en fonction de ses caractéristiques.

L'***analyse de covariance*** permet d'expliquer une variable réelle par un ensemble de variables réelles ou catégorielles.

Des généralisations de ces modèles linéaires (***logit***, ***probit***) permettent de traiter une variable à expliquer dichotomique, c'est-à-dire, une variable catégorielle ayant deux modalités seulement.

L'analyse de variance est utilisée dans ***les plans d'expériences***. Comme son nom l'indique, il s'agit ici de planifier des expériences.

Supposons que l'on dispose d'une machine de levage et transport de caisses et qu'il y ait cinq boutons de réglage afin qu'elle fonctionne correctement. On ne sait pas a priori quelles sont les bonnes positions de chacun de ces boutons. Une pratique non planifiée consiste à les placer en position intermédiaire et à faire une première expérience, puis à changer la position d'un bouton pour voir l'effet de ce bouton sur le réglage, puis à changer la position d'un autre bouton, etc. Petit à petit, après de nombreuses expériences, après avoir éventuellement noté les résultats et avec beaucoup de chance, il est possible d'obtenir une machine dont le réglage est acceptable. Lorsque le problème est plus complexe, la combinatoire est telle qu'il est illusoire d'arriver au but par ce moyen.

Planifier les expériences consiste à décider à l'avance d'une liste d'expériences que l'on sait ne pas être optimales mais qui permettront de mesurer les effets de chacune des conditions expérimentales afin, par des calculs, d'en déduire les conditions optimales.

C'est Fisher qui développe ces méthodes à la Rothamsted Experimental Station, station d'expérimentation agricole près de Londres où il travaille de 1919 à 1933. Il est aisé de comprendre que les agronomes réfléchissent longuement avant de faire une expérience puisqu'ils n'obtiendront les résultats qu'à la saison prochaine.

C'est donc dans ce cadre qu'a débuté la planification d'expériences mais les méthodes se sont généralisées dans l'industrie grâce à un japonais, excellent vulgarisateur, Taguchi. A la fin de la deuxième guerre mondiale, les américains ont aidé les japonais à la reconstruction de leur industrie et ils ont formé de nombreux experts dont fait partie Taguchi. Très rapidement, ces experts révolutionnent la ***gestion de la qualité*** et deviennent, depuis les années 80, des exemples pour les industriels américains et européens. En 2010, ces outils semblent avoir atteint leurs limites. Les procédures qualité mises au point pour un certain volume de production ne fonctionnent pas correctement à une plus grande échelle ; Toyota qui était considéré le maître es qualité en a fait l'expérience.

Pendant longtemps les techniques statistiques de qualité et de ***gestion de production*** en entreprise se résumaient à des contrôles de réception de lots. Le principe en est le suivant : on compte le nombre d'objets défectueux sur un échantillon d'objets tiré au hasard ; le lot est rejeté lorsque ce nombre dépasse un certain seuil fixé à l'avance entre le fournisseur et le client à partir de calculs de probabilités.

Ensuite, se mettent en place les cartes de contrôle en cours de fabrication. On tire régulièrement des petits échantillons d'objets fabriqués et on vérifie que leurs caractéristiques restent bien entre des limites fixées à l'avance. Une sortie des limites entraîne une vérification immédiate du réglage des machines concernées sans attendre la fin de la production pour rejeter éventuellement le lot.

Les plans d'expériences se situent en amont de la production puisqu'ils ont pour objectif de définir les conditions optimales de production.

La première étape d'un plan d'expériences consiste, après discussion avec les cadres et les techniciens du secteur où est fabriqué le produit, à faire la liste de toutes les variables (appelées facteurs dans ce contexte) susceptibles d'avoir une influence sur la qualité du produit.

Il faut ensuite les hiérarchiser, définir les différents niveaux des facteurs qui vont être expérimentés. S'il y a 10 facteurs à 2 niveaux, il y a 2 puissance 10, c'est-à-dire, 1024 conditions expérimentales différentes. Il est bien souvent impossible de réaliser un aussi grand nombre d'expériences. Si le nombre d'expériences qu'il est possible de réaliser est de 128 (= 2 puissance 7), il faut alors choisir les interactions que l'on souhaite conserver et planifier les expériences selon une combinatoire particulière.

Après la réalisation de cette série d'expériences, l'analyse de la variance permet d'estimer les effets des différents niveaux des facteurs ainsi que des interactions retenues. On en déduit les conditions expérimentales optimales qui bien souvent correspondent à une expérience non encore réalisée.

C'est à peu près ainsi que s'est déroulé, en 1993, le stage dans une importante entreprise, dont le secret professionnel oblige à taire le nom, de deux étudiants stagiaires de sciences et technologies. Avant d'investir un important budget dans l'achat d'une nouvelle presse permettant d'améliorer le collage de pièces métalliques, la direction de l'entreprise

décide de recevoir les deux stagiaires pour étudier les conditions de collage par un plan d'expériences.

Après consultation, c'est effectivement l'insuffisance de la puissance de la presse qui est le plus souvent mentionnée comme cause des défauts mais de nombreux autres facteurs sont cités. Les défauts sont des bulles dont les normes ISO décrivent en long et en large le nombre maximal au cm^2 et le diamètre maximal de chacune d'elles qui restent acceptables dans le cadre des normes de qualité. Les stagiaires retiennent une dizaine de facteurs à deux niveaux et, en suivant la méthodologie Taguchi, proposent un plan d'expériences avec 128 expériences. C'est la première fois que les techniciens réalisent une série de 128 expériences de collage programmées à l'avance sans analyse de résultats intermédiaires. Une fois toute la série réalisée, les défauts sont dénombrés et mesurés, les calculs d'analyse de variance sont effectués et les conditions optimales sont déduites et proposées aux techniciens. Le résultat est remarquable puisqu'aucune bulle n'apparaît et ceci avec la presse incriminée. Finalement, c'est un facteur considéré comme très secondaire qui jouait en fait un rôle considérable sur la qualité du collage.

5.3. Extensions et limites

a) Processus stochastiques

Pour les séries chronologiques, ont été développés des modèles non aléatoires décomposant la série de façon additive ou multiplicative en une composante tendancielle linéaire ou exponentielle et une composante saisonnière périodique. Pour s'adapter aux évolutions dans le temps, des modèles dynamiques, avec des coefficients dépendant du temps, ont été proposés.

Mais l'analyse statistique des séries chronologiques (appelées aussi ***processus stochastiques***) a connu un essor considérable suite aux travaux de Box et Jenkins publiés en 1970 sur les modèles linéaires stochastiques autorégressifs (AR), moyennes mobile (MA) ou mixtes (ARMA). Au lieu de décomposer la série chronologique en somme ou produit d'autres séries dépendant du temps, l'idée tout à fait nouvelle est d'exprimer la valeur présente d'une variable comme fonction linéaire de ses valeurs passées et des valeurs passées et présentes d'un processus aléatoire, appelé bruit, modélisant les erreurs. Le bruit est supposé avoir une variance constante ce qui ne convient pas à certains processus, notamment en finance. Les modèles ARCH sont alors proposés avec une variance dépendant du temps.

Ces processus stochastiques, qui peuvent être unidimensionnels (une seule variable dépendant du temps) ou multidimensionnels (plusieurs variables dépendant du temps et qui peuvent interagir) sont à ***temps discret***, c'est-à-dire, indexés par l'ensemble des nombres entiers.

Parallèlement a été développée en probabilité la ***théorie des processus stochastiques*** à ***temps continu***, c'est-à-dire, indexés par l'ensemble des nombres réels. Les modèles ARCH apparaissent alors comme de bonnes approximations des modèles en temps continu lorsque l'intervalle de temps entre deux observations tend vers zéro. Le passage au continu permet de développer des résultats théoriques nouveaux et des applications encore plus efficaces. Ces modèles se sont développés en finance avec le succès que l'on sait !

b) Modèles à équations structurelles

Contrairement aux modèles linéaires présentés précédemment pour lesquels les nouvelles variables sont combinaisons linéaires des variables initiales, l'objectif des ***modèles à équations structurelles*** est de rechercher des ***variables latentes*** expliquant au mieux les corrélations entre variables initiales, dites aussi ***manifestes***. D'autres types de modèles sont proposés pour la statistique en sciences humaines et sociales : les ***modèles hiérarchiques*** ou ***multiniveaux***, les modèles issus de la ***théorie de la réponse aux items*** en éducation, …

6. Perspectives et conclusion

Le développement des capacités de calcul des ordinateurs a complètement transformé la recherche en mathématiques, et plus particulièrement en probabilités et statistique. Nous présentons ci-après deux exemples d'utilisation importante des ordinateurs : ***la simulation*** et ***la fouille dans les données*** (plus connue sous le nom de Data Mining).

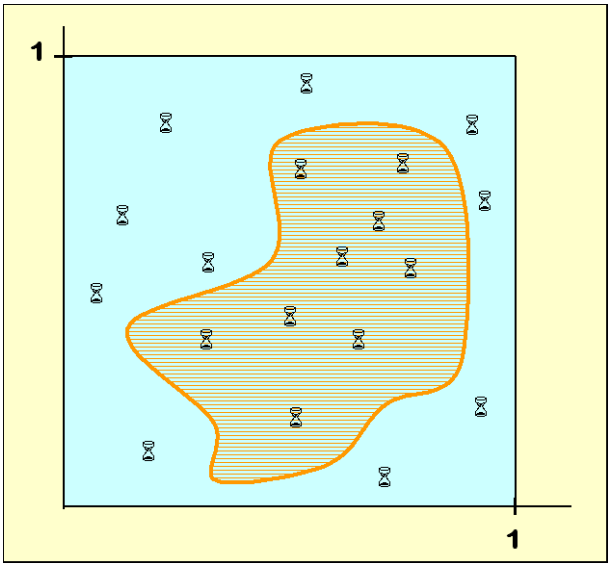
a) La simulation et les générateurs de nombres au hasard

La simulation aléatoire repose sur la loi des grands nombres : en répétant un grand nombre de fois de façon indépendante une expérience aléatoire à laquelle est liée une variable réelle X, on obtient une approximation de plus en plus fiable de l'espérance mathématique de X.

Plutôt que de réaliser des calculs formels trop complexes, par exemple l'évolution de processus aléatoire ou le calcul de variance d'estimateur d'un plan de sondage aléatoire à plusieurs degrés, la simulation permet d'obtenir des résultats numériques sur un très grand nombre d'observations.

Toute simulation nécessite l'utilisation de suites de nombres au hasard, c'est-à-dire de nombres issus de la loi uniforme.

Prenons un exemple très simple. Comment mesurer l'aire de la surface hachurée ?



On inscrit la surface dans un carré dont on prend le côté comme unité de longueur. On simule un échantillon de couples (x_i, y_i), $i = 1, \ldots, 100$, de variables aléatoires, indépendantes, et de même loi uniforme sur [0,1]. La proportion de points qui tombent à l'intérieur de la surface hachurée est une estimation de la mesure d'aire cherchée.

La simulation, très utilisée aujourd'hui dans un grand nombre de disciplines grâce aux possibilités des ordinateurs, a besoin de suites de nombres aléatoires toujours plus longues. Le fait d'obtenir des résultats à partir d'échantillons tirés au hasard, comme la roulette au Casino, a donné le nom de méthodes de Monte Carlo à ces méthodes de simulation. Mais, ces générateurs n'offrent qu'une approche ***pseudo-aléatoire***. Il s'agit en fait d'un procédé de construction déterministe de suites périodiques dont la période est tellement longue que la suite très longue des chiffres d'une même période paraît aléatoire.

Ainsi, on résout un problème non aléatoire, l'aire d'une surface, par un procédé aléatoire et on construit des suites de nombres aléatoires par un procédé déterministe !

Comment peut-on juger du ***caractère aléatoire d'une suite de nombres*** ? Il n'est pas suffisant de savoir que la série a été produite par une procédure aléatoire. Par exemple, dans une série de pile ou face (représentés par 1 et 0), les 3 suites de 12 chiffres 110100101110, 111111111111 et 101010101010 ont toutes les trois la même probabilité de sortir, probabilité égale à $(1/2)^{12}$; la première semble pourtant plus « aléatoire » que les deux autres. On souhaite en effet retrouver dans la suite le même nombre de « 0 » et de « 1 » mais on souhaite aussi avoir absence de périodicité : les observations doivent sembler provenir de variables aléatoires ***indépendantes en probabilité***. En regroupant les observations deux à deux, les quatre configurations « 00 », « 01 », « 10 », « 11 » doivent être également présentes, en regroupant les observations trois à trois, les huit configurations « 000 », « 001 », …, « 111 » doivent être également présentes, etc.

La définition d'une suite aléatoire est issue de la théorie de l'information. Une suite de chiffres est aléatoire lorsque le plus petit algorithme nécessaire pour l'introduire dans un ordinateur contient à peu près le même nombre de bits que la suite. On sait définir ce qu'est une suite aléatoire et mesurer son caractère aléatoire mais on montre aussi qu'il est impossible de démontrer qu'une suite donnée est aléatoire. C'est à partir de tests statistiques que l'on décidera de rejeter ou non la propriété d'***aléarité*** d'une suite donnée.

b) Fouille dans les données (data mining)

Le développement de l'informatique et l'augmentation de la puissance des ordinateurs fournissent des moyens exceptionnels pour la gestion des données. Faut-il savoir transformer ces données en information utile.

De fait, les administrations publiques, les collectivités territoriales et les entreprises sont à présent envahies de données, mais ces données sont souvent de mauvaise qualité et mal exploitées.

Comment repérer, trier, analyser les résultats déversés par les logiciels ? Comment obtenir les données pertinentes qui permettront de répondre aux questions posées ? Comment rechercher de façon automatique et intelligente les connaissances utiles sur le Web ?

Les méthodes d'extraction automatique de connaissances dans les bases de données (internes ou externes à l'entreprise), en particulier sur le Web, tentent de répondre à ces questions. Ces méthodes, au carrefour de plusieurs disciplines, sont en plein développement.

c) La statistique et ses utilisations

La présentation faite ici est partielle et subjective ; partielle car seules quelques méthodes statistiques de base ont été présentées, subjective car la présentation est faite autour de la typologie des méthodes en fonction du type de "données", c'est le point de vue de la statistique mathématique. Dès que l'on passe à l'utilisation de la statistique, les difficultés méthodologiques et les problèmes d'interprétation ne manquent pas. Les difficultés sont présentes avant même le premier traitement statistique dès l'étape du recueil de "données" … car les "données" portent mal leur nom : elles reposent sur des choix et des compromis et n'ont rien de "naturels".

Le paragraphe précédent peut laisser penser que la statistique mathématique aurait un développement autonome et pourrait, à l'occasion, trouver des applications. En réalité, la statistique se développe en grande partie autour de problèmes posés dans les champs disciplinaires auxquels elles s'appliquent. On invite le lecteur à consulter, sur le site de la Société Française de Statistique (http://www.sfds.asso.fr/), la liste des groupes de travail. Outre le groupe « statistique mathématique » dont l'objectif est de fédérer la communauté académique en statistique mathématique et de favoriser les interactions entre théorie, méthodologie et applications, ainsi qu'entre statistique et probabilités, on trouve les groupes suivants : « agro-industrie », « analyse d'images », « banque, finance, assurance », « biopharmacie et santé », « chimiométrie », « data mining et apprentissage », « enquêtes et modèles », « environnement », « fiabilité et incertitudes », enfin trois groupes aux thèmes plus transversaux : « enseignement de la statistique », « histoire de la statistique » et « statistique et société ».

Concernant la statistique d'État, Alain Desrosières montre bien, dans les deux ouvrages cités en références, comment elle est historiquement et socialement construite : les différentes formes de l'État (ingénieur, libéral, providence, keynésien, néo-libéral) sont associées à différentes façons de penser la société et l'économie et différents modes d'action mais aussi à différentes formes de statistiques. Les nouvelles formes de statistiques associées à l'État néo-libéral qui émerge dans les années 1990 sont la construction et l'usage d'indicateurs de performance pour évaluer, classer, comparer, établir des palmarès (cf. tableau page 56 du tome I et page 12 du tome II).

En France, la LOLF (loi organique relative aux lois de finances), adoptée en août 2001 et mise en application le 1er janvier 2006, va bien au-delà d'une simple réforme de la comptabilité publique. Elle institue de nouvelles règles d'élaboration et d'exécution du budget de l'État et introduit une démarche de performance pour améliorer l'efficacité des politiques publiques. Le budget général de l'État est segmenté en 48 missions, 171 programmes, 499 objectifs et 1030 indicateurs de performance (Cf. sur le site : http://www.performance-publique.gouv.fr/ le Projet de Loi de Finances 2010). Depuis 2008, ***l'évaluation des politiques publiques*** est inscrite dans la constitution comme mission officielle du Parlement. Il s'agit de la Nouvelle Gestion Publique prônée par l'OCDE et l'Union Européenne.

Le ***traité de Lisbonne***, signé le 13 décembre 2007, est entré en vigueur le 1er décembre 2009 après avoir été ratifié par chacun des 27 États membres de l'UE. Les chefs d'État ou de gouvernement se sont mis d'accord sur de nouvelles règles qui régissent l'étendue et les modalités de l'action future de l'Union. Dans des domaines qui relèvent de la compétence des États membres tels que l'emploi, la protection sociale, l'inclusion sociale, l'éducation, la

jeunesse et la formation, la convergence est assurée par la ***MOC (méthode ouverte de coordination)***. Cette méthode repose sur :

- l'identification et la définition en commun d'objectifs à remplir (adoptés par le Conseil),
- des instruments de mesure définis en commun (statistiques, indicateurs, lignes directrices),
- le "benchmarking", c'est-à-dire la comparaison des performances des États membres et l'échange des meilleures pratiques (surveillance effectuée par la Commission).

(cf. le portail de l'Union Européenne : http://europa.eu/geninfo/atoz/fr/index_1_fr.htm).

L'argument statistique est plus que jamais invoqué pour la justification des politiques publiques. Il est donc aussi un élément incontournable de communication ce qui n'est pas sans poser problèmes aux Services Statistiques Nationaux dont l'indépendance professionnelle est inscrite comme premier principe du code de bonnes pratiques de la statistique européenne : (http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/code_practicefr.pdf)

d) Conclusion

C'est en se référent à la statistique et aux probabilités que certaines pratiques sociales se développent (sondages d'opinion, évaluation de la qualité et de la performance, marchés du risque financier et des produits dérivés). Inversement, ces pratiques modifient profondément les "réalités" que la statistique et les probabilités sont sensées décrire et modéliser.

Après les physiciens et les biologistes, les mathématiciens probabilistes et statisticiens s'interrogent sur leurs responsabilités vis-à-vis de la société. Une bonne formation scientifique des futurs citoyens est nécessaire mais une formation à l'histoire et à l'épistémologie des sciences des futurs scientifiques est également nécessaire. C'est par un réel travail interdisciplinaire dès l'enseignement secondaire que cela peut commencer.

Références

1. Le hasard, Dossier Hors Série, Pour la Science, Avril 1996.

2. Histoire de la Statistique, Jean-Jacques Droesbeke et Philippe Tassi,
Que sais-je ? n° 2527, PUF, 2ème édition, Paris 1997.

3. Pour une sociologie historique de la quantification. L'argument statistique I,
Alain Desrosières, Presses de l'École des Mines, coll. Sciences sociales, Paris 2008

4. Gouverner par les nombres, L'argument statistique II,
Alain Desrosières, Presses de l'École des Mines, coll. Sciences sociales, Paris 2008