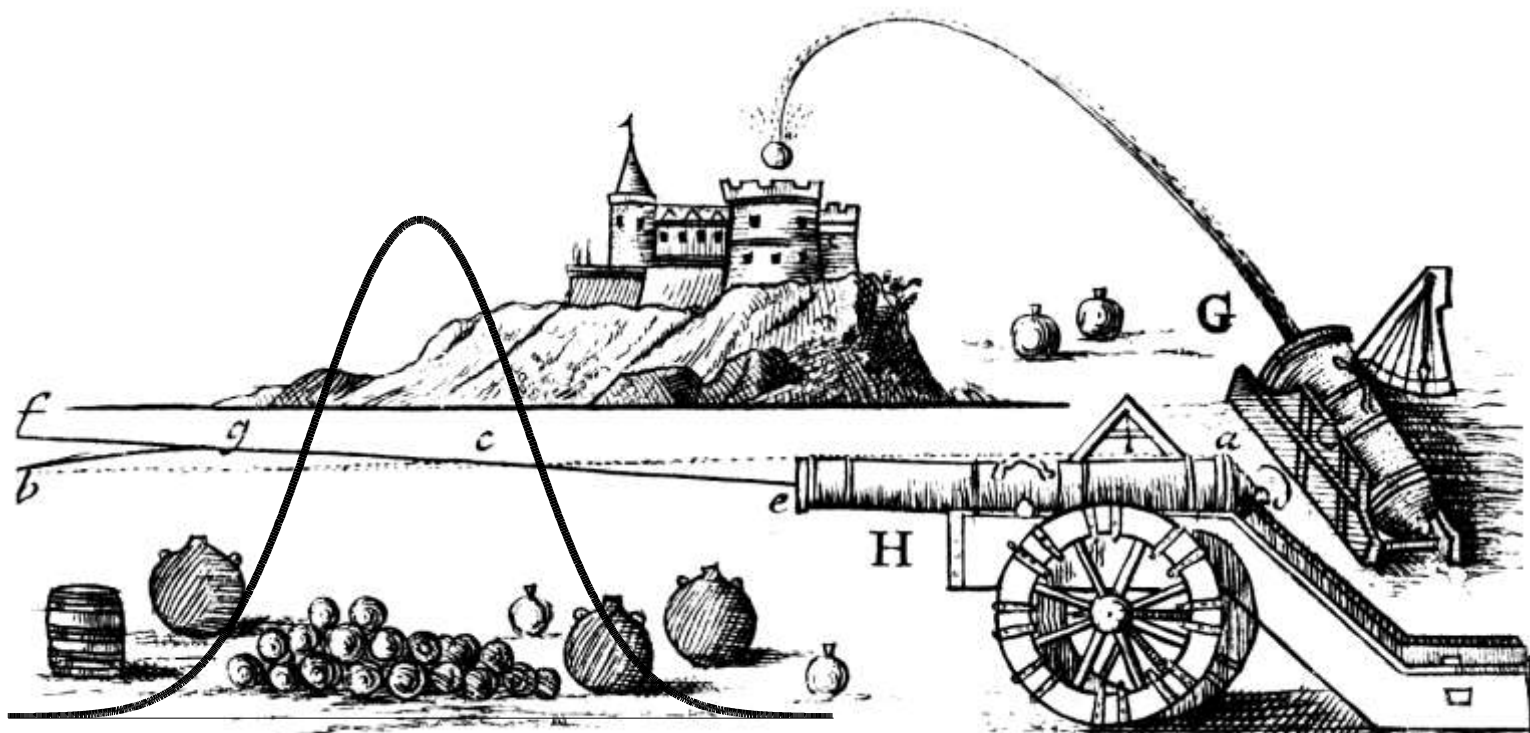


Brochure n°112  
de la Commission Inter-I.R.E.M.  
Lycées Technologiques



# ENSEIGNER LA STATISTIQUE AU LYCEE : DES ENJEUX AUX METHODES



Direction de l'enseignement scolaire  
Bureau A 11  
Valorisation des innovations pédagogiques

Commission Inter-I.R.E.M.  
Lycées Technologiques  
I.R.E.M.  
Institut Galilée  
av. J.B. Clément  
93430 VILLETANEUSE

**UNIVERSITE Paris-Nord - I.R.E.M.**

**ENSEIGNER LA STATISTIQUE AU LYCEE :  
DES ENJEUX AUX METHODES**

173 pages  
Villetaneuse 2001

Dépôt légal : 4<sup>ème</sup> trimestre 2001

Cette brochure a été réalisée dans le cadre de travaux de recherche lancés conjointement par l'ADIREM (Assemblée des Directeurs d'IREM), la Direction de l'enseignement scolaire et l'Inspection générale de mathématiques.

Elle constitue une information à destination des enseignants du second degré, au moment où l'enseignement de la statistique se développe dans les lycées. Pour mieux comprendre les nouvelles inflexions des programmes dans ce domaine, il s'agit d'en montrer les enjeux (utilisations de la méthode statistique dans les secteurs industriels et tertiaires), les techniques, les origines historiques, et d'illustrer le propos de quelques activités pédagogiques.

**Bernard VERLANT**

Responsable de la Commission Inter-IREM  
"Lycées technologiques"

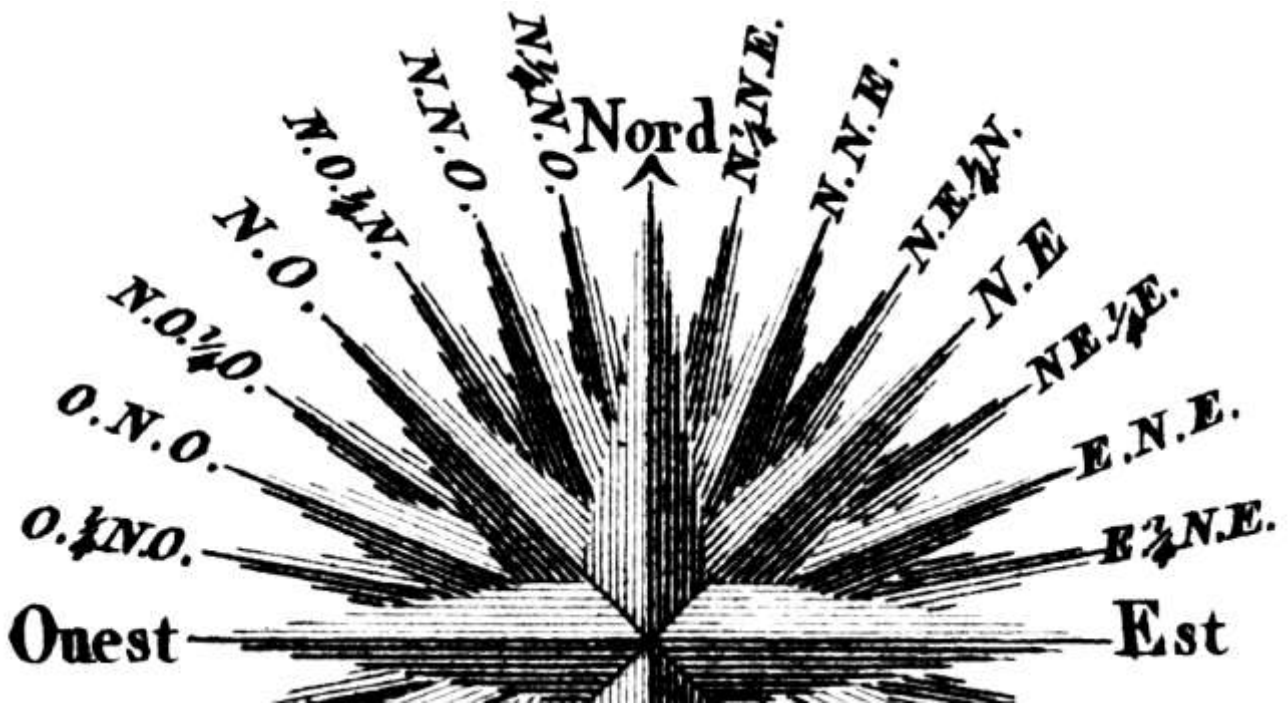
Cette brochure a été réalisée par :

**Jean-Louis PIEDNOIR**  
**Inspecteur général de mathématiques**

et

**Philippe DUTARTE**  
Professeur au lycée E. Branly de Créteil

avec la collaboration des enseignants de la  
**Commission Inter IREM "Lycées technologiques"**



# S O M M A I R E

AVANT PROPOS .....	6
VARIABILITE, INCERTITUDE ET HASARD .....	7
QU'EST-CE QUE LA STATISTIQUE ?.....	22
GRANDE ET PETITE HISTOIRE DE LA STATISTIQUE .....	37
LA STATISTIQUE EUCLIDIENNE .....	67
LA STATISTIQUE INFERENTIELLE .....	79
LA SIMULATION EN STATISTIQUE.....	93
POUR ALLER PLUS LOIN .....	110
ANNEXE 1 – Simulation de fourchettes de sondages.....	124
ANNEXE 2 – Utilisation de moyennes mobiles à la bourse.....	135
ANNEXE 3 – Expérimentation du théorème limite central.....	141
ANNEXE 4 – Un exemple d'introduction à la notion de test .....	153
ANNEXE 5 – La maîtrise statistique des procédés de production.....	161
ANNEXE 6 – Loi de Poisson et temps d'attente .....	168
BIBLIOGRAPHIE.....	172

# AVANT PROPOS

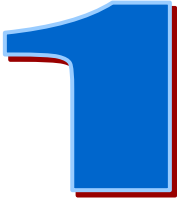
On a assisté, depuis quelques décennies, à un développement considérable de l'utilisation des méthodes statistiques, dans des domaines très variés de l'activité humaine. Elles sont maintenant d'utilisation courante en biologie, dans les sciences humaines, dans l'industrie, partout où il existe de la variabilité. La généralisation de l'usage des ordinateurs a rendu possible ce développement, du fait des possibilités nouvelles de recueil, de stockage et de traitement des données.

Il n'était plus possible de laisser les élèves de nos lycées ignorer un mode de raisonnement et des outils, quand leur usage est aussi répandu dans la société. La statistique fait donc une entrée remarquable dans les programmes. Le professeur de mathématiques est chargé, à la fois, d'initier au mode de raisonnement, et de décrire quelques outils. Malheureusement, sa formation initiale lui donne très peu de recul sur ce sujet. Sauf exception, il n'a rien vu de tout cela lors de ses études, sauf, peut-être, un aperçu en DEUG. Il revient donc à la formation continue de combler cette lacune. Celle-ci est certes affaire de stages, de rencontres, mais aussi de travail personnel de formation autodidacte. On demande bien à nos élèves de faire des exercices, de résoudre des problèmes.

La présente brochure veut aider les professeurs de mathématiques dans ce travail personnel, afin qu'ils puissent mieux comprendre ce qu'ils enseignent en seconde, première, terminale, mais aussi en classes de techniciens supérieurs ou en DECF. L'accent est mis sur la démarche statistique, assez différente des démarches scientifiques classiques. Le substrat mathématique justifiant, à travers des modèles, les procédures statistiques est évoqué, mais n'est pas traité complètement. Il ne s'agit pas d'écrire un manuel de statistique, il en existe d'excellents (cf. bibliographie), mais de montrer, à partir d'exemples, la pertinence de méthodes utilisant des mathématiques parfois très sophistiquées.

Nous aurons atteint notre but si nous avons donné le goût au lecteur d'enseigner la statistique et aussi de parfaire sa culture à l'aide d'autres lectures.

**Jean-Louis Piednoir**  
**Inspecteur général de mathématiques**



# VARIABILITE, INCERTITUDE ET HASARD

Dans des situations à forte variabilité, génératrices d'incertitude, le hasard est un modèle possible, et la statistique propose, à partir des observations, des méthodes originales, aidant à la prise de décision, sur la base d'un risque d'erreur consenti.

## Variabilité

Nous vivons dans le *variable* et l'*incertain*. Considérons la taille des individus d'une certaine population, la durée de vie d'une ampoule électrique, la qualité d'une production, les commandes des clients d'une entreprise, le nombre d'accidents ou d'incendies, la sensibilité aux maladies ou aux traitements, l'opinion des électeurs, les cours de la bourse, la météo... On peut multiplier les exemples, pour lesquels d'ailleurs l'aléatoire n'est pas nécessairement la seule réponse possible. La géométrie fractale, ou la théorie du chaos, par exemple, proposent également des modèles, surtout qualitatifs. Reste que l'on se tournera souvent vers le statisticien qui, à partir des données observées, saura, face à la variabilité et à l'incertitude qui en découle, aider scientifiquement à la prise de décision. C'est plus rigoureux et satisfaisant que de recourir à un astrologue, même si l'art de la statistique peut paraître ésotérique du fait de méthodes inhabituelles en mathématiques. Ainsi *Keynes* écrit encore en 1921 (à propos des méthodes probabilistes) "*les savants y décèlent un relent d'astrologie ou d'alchimie*"<sup>1</sup>. Par rapport à l'astrologue, le statisticien met surtout en place des procédures normalisées sur lesquelles un consensus pourra être trouvé.

L'idée de variabilité n'est pas récente. Ainsi, au II<sup>e</sup> siècle après J.-C., l'astronome (et astrologue ! ) alexandrin *Ptolémée*, établissant un gigantesque catalogue d'étoiles qui fit longtemps autorité, fut confronté à des mesures différentes d'une même valeur. Il propose alors une mesure unique assortie d'une variation basée, semble-t-il, sur l'étendue des mesures.

### **Les observations astronomiques de Bradley**

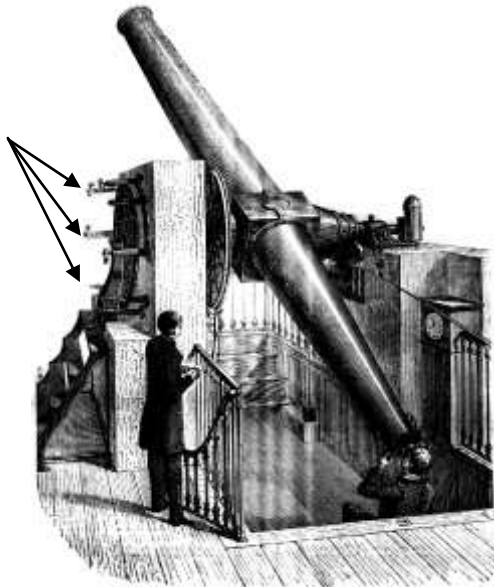
Dans un traité de topographie du début du XX<sup>e</sup> siècle<sup>2</sup>, on lit que l'astronome *James Bradley* (1693 – 1762) exécuta 470 observations (aux résultats variables) pour déterminer la déclinaison (fixe) d'une étoile, par mesure d'un angle zénithal. *Friedrich Bessel* (1784 – 1846) examina les écarts à la moyenne de ces mesures et les ajusta à une loi normale :

Écarts à la moyenne	Fréquences en % des observations de <i>Bradley</i>	Valeurs théoriques selon un ajustement normal
0",0 à 0",4	22	19.5
0",4 à 0",8	19.3	18.3
0",8 à 1",2	18.3	16.2
1",2 à 1",6	9.3	13.6
1",6 à 2",0	9	10.6
2",0 à 2",4	7.7	7.9

<sup>1</sup> Cité par *Droesbeke* et *Tassi* dans "*Histoire de la statistique*" – "*Que-sais-je ?*" – 1997.

<sup>2</sup> *A. Pelletan* – "*Traité de topographie*" – 1911.

2",4 à 2",8	3.3	5.5
2",8 à 3",2	5	3.6
3",2 à 3",6	2.7	2.2
3",6 à 4",0	1.3	1.3



Cercle méridien de l'Observatoire de Paris. Sur le côté gauche, plusieurs viseurs permettent la répétition de la lecture de la mesure, de façon à en faire ensuite la moyenne.

plus de 4"	2	1.4
------------	---	-----

Le statisticien d'aujourd'hui considérerait, dans ces conditions, que les 470 mesures sont un échantillon aléatoire extrait de la population (virtuelle) de toutes les mesures possibles, population répartie selon une loi normale, centrée sur la valeur réelle (inconnue) de la déclinaison de l'étoile. Il montrerait que la moyenne des 470 mesures est, sous l'hypothèse gaussienne<sup>3</sup>, la meilleure estimation de la déclinaison réelle (si l'hypothèse gaussienne n'est pas réalisée, alors la moyenne peut être un mauvais estimateur). Il estimerait l'écart type de la population et fournirait un intervalle de confiance dans lequel il situerait la déclinaison exacte, avec un pourcentage d'erreur déterminé.

La variabilité crée du probabilisable et la statistique fournit, sans certitude, une réponse basée sur ce qui est le plus vraisemblable, assortie d'un pourcentage d'erreur. Il faut admettre que, dans certains cas, les calculs mathématiques effectués conduisent à une réponse fautive au problème posé, mais ce risque est maîtrisé.

La moyenne joue un rôle essentiel en statistique, mais ce n'est pas l'arbre qui cache la variabilité. *Daniel Schwartz*<sup>4</sup> signale que "les mauvaises langues prétendent qu'un statisticien se noya dans un cours d'eau dont la profondeur moyenne était de 20 cm. C'est qu'à l'endroit où il souhaitait patauger elle atteignait 2 m !" mais d'ajouter, "comment oublierait-il la variabilité, raison d'être de la statistique ?" S'il n'y avait pas de variabilité, toutes les valeurs seraient identiques, et la statistique inutile. C'est en s'appuyant sur la variabilité, sur les *fluctuations* entre les échantillons, que le statisticien calculera la "fourchette" du sondage, ou établira que la différence de durée de vie des ampoules électriques fabriquées par A est *significativement* supérieure à celles fabriquées par B, c'est à dire vraisemblablement au delà de la variabilité "naturelle" observée d'une ampoule à l'autre.

<sup>3</sup> On peut effectuer un test statistique, pour juger de la normalité des valeurs observées.

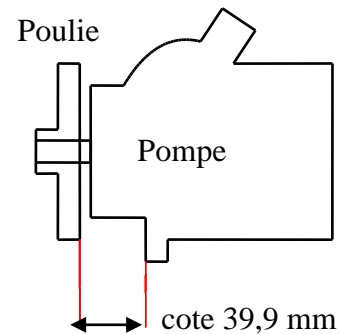
<sup>4</sup> "Le jeu de la science et du hasard" – Flammarion 1997.



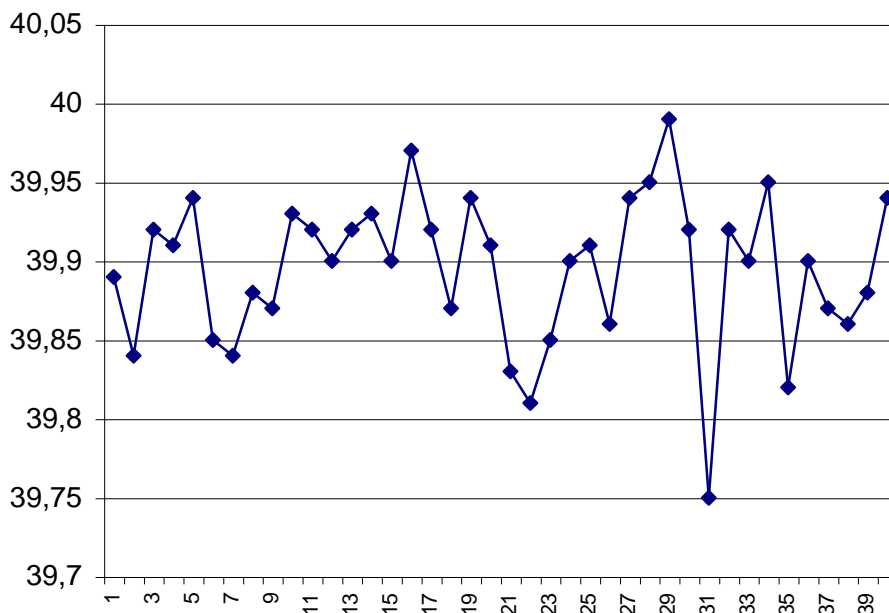
### Carte de contrôle industrielle

Voici un exemple précis, dans le contexte industriel, où la "Maîtrise Statistique des Procédés" s'appuie sur l'étude de la variabilité de la production, pour détecter les anomalies, en temps réel. Le processus étudié, issu de l'industrie automobile<sup>5</sup>, est une presse d'emmanchement de poulie sur une pompe de direction assistée. Les performances de la presse sont variables, cette variabilité ayant de nombreuses causes possibles : main-d'œuvre, matériel, matière première, environnement de l'atelier, méthodes d'organisation...

L'emmanchement de la poulie sur l'axe de la pompe est mesuré par la cote de 39,9 mm indiquée sur le schéma ci-contre. On a mesuré cette cote sur 40 ensembles pompe-poulie, produits de façon successive dans la production en série.



Les variations sont représentées sur le schéma suivant.



L'objectif du statisticien est ici d'utiliser la variabilité observée pour faire du "film" de son observation un outil de décision. Cet outil sera la *carte de contrôle*, qui permettra de détecter en temps réel les anomalies des caractéristiques statistiques étudiées, pour envisager les actions à suivre (contrôle supplémentaire, arrêt de la production, réglages selon le type de dérive statistique observée).

1<sup>ère</sup> étape : Etude de la normalité.

Un premier test sera mis en oeuvre pour voir si l'hypothèse selon laquelle "les observations sont le résultat d'un phénomène suivant une loi normale" est raisonnable (test de "normalité"). C'est le cas ici, on pourra donc supposer que la variable aléatoire  $X$  qui, à chaque pièce mesurée, associe sa cote en mm, suit la loi normale de moyenne  $\mu = 39,9$  et d'écart type  $\sigma$ .

<sup>5</sup> Source : constructeur automobile français.

2<sup>ème</sup> étape : Etude de la "capabilité".

On vérifiera ensuite si, selon cette hypothèse de normalité, le procédé est apte à fabriquer des pièces conformes aux normes de tolérance. L'intervalle de tolérance pour cette cote est 39,9 +/- 0,2 mm (normes du constructeur) dont l'amplitude est  $a = 0,4$ .

Le quotient  $C = \frac{a}{6\sigma}$  est nommé *capabilité* du procédé (à fabriquer selon la norme). Le

dénominateur  $6\sigma$  correspond au fait que, pour une loi normale de moyenne  $\mu$  et d'écart type  $\sigma$ , 99,7% des valeurs observées sont comprises entre  $\mu - 3\sigma$  et  $\mu + 3\sigma$ .

On demande à ce que cette capabilité soit au moins égale à 1, sinon, le procédé fabriquera, à coup sûr, une quantité non négligeable de pièces hors normes, dans ce cas, avant de le mettre sous contrôle, on essaiera d'abord de l'améliorer. Lorsque  $C = 1$ , on a  $\sigma = \frac{a}{6} = \frac{0,2}{3}$

et donc :

$P(39,9 - 0,2 \leq X \leq 39,9 + 0,2) = P(-3 \leq T \leq 3) \approx 0,997$  où  $T = \frac{X - 39,9}{\sigma}$  suit la loi

normale standard  $N(0, 1)$ . Pour une capabilité égale à 1, on a donc 99,7 % des pièces fabriquées qui sont conformes.

L'écart type observé sur les 40 pièces qui ont été prélevées est  $\sigma_e \approx 0,047$  d'où une estimation de  $\sigma$  à  $\sqrt{\frac{40}{39}} \sigma_e \approx 0,048$  donc une capabilité  $C \approx \frac{0,4}{6 \times 0,048} \approx 1,39$  ce qui est

convenable.

3<sup>ème</sup> étape : Construction d'une carte de contrôle<sup>6</sup>.

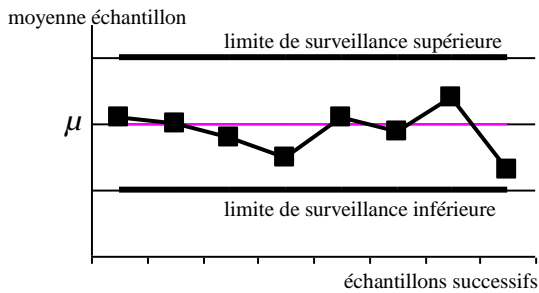
On peut alors envisager, pour la surveillance de la production à venir, d'établir une "carte de contrôle aux moyennes". On supposera ici que la valeur de  $\sigma$  reste stable mais qu'une dérive sur la valeur de  $\mu$  est à craindre. La caractéristique mesurée (cote de l'emmanchement pompe-poulie) étant un point "critique" (c'est un "point Sécurité-Réglementation"), la norme prévoit de prélever régulièrement des échantillons de  $n = 5$  ensembles pompe-poulie, sur lesquels on calculera la moyenne des 5 cotes. Des **limites de surveillance**, de part et d'autre de 39,9, seront calculées<sup>7</sup> de sorte que si  $X$  suit effectivement la loi  $N(39,9 ; 0,048)$ , 95% des moyennes d'échantillons de taille 5 doivent se situer entre ces limites. Si ce n'est pas le cas, pour l'un des échantillons, l'alerte sera donnée. Dans 5% des cas, il s'agira d'une fausse alerte, mais on peut également raisonnablement penser que si la moyenne d'un échantillon est située en dehors des limites de surveillance, c'est que l'hypothèse  $\mu = 39,9$  est fautive et qu'une dérive de fabrication est vraisemblable.

Des **limites d'acceptations** sont de même calculées, entre lesquelles se situent 99,8% des échantillons sous l'hypothèse d'un processus "sous contrôle" (où  $X$  suit la loi normale  $N(39,9 ; 0,048)$ ), et dont le dépassement provoque l'arrêt immédiat de la production. On admet donc dans ce cas un risque de 2<sup>0</sup>/<sub>00</sub> de fausses alertes conduisant à l'arrêt de la production (d'une part cet arrêt est très coûteux, d'autre part la multiplication de fausses alertes conduit à ne plus prendre suffisamment au sérieux les alertes à venir).

<sup>6</sup> On pourra, sur ce sujet, consulter la brochure de l'IREM de Clermont-Ferrand : "Une application industrielle des statistiques : la carte de contrôle." – Mars 2001.

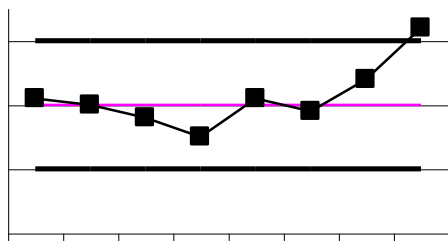
<sup>7</sup> Les calculs sont développés dans le TP sur Excel donné en annexe 5.

L'étude statistique de la variabilité de la production permettra ainsi différents "diagnostics", en voici quelques exemples (interprétation des moyennes de huit échantillons de taille 5 prélevés successivement) :



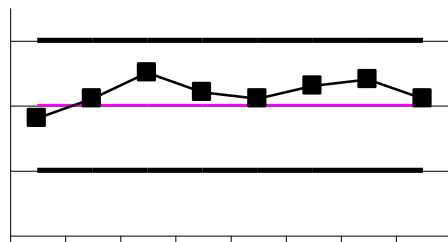
Carte 1 :

La moyenne reste entre les limites de surveillance, sans variation remarquable, le processus est considéré comme stable et aucune action n'est à envisager.



Carte 2 :

La dernière moyenne sort des limites de surveillance. Le processus dérive probablement. Il faut en trouver la cause et corriger.



Carte 3 :

On constate une série de sept points consécutifs du même côté de la moyenne. Le processus dérive probablement, ce qui peut être dû à un mauvais réglage initial.

### **Les petits pois de Mendel**

Paradoxalement, l'absence de variabilité peut être aussi suspecte que ses débordements (dans le cas d'un dé, par exemple, l'observation d'un "débordement" amenant de façon significative davantage de 6 que la normale conduira à suspecter que le dé est truqué) et permet également de détecter statistiquement des fraudes.

C'est ainsi que *Ronald Fisher* examina les données expérimentales qui permirent à *Gregor Johann Mendel* (1822-1884) d'étayer sa théorie de l'hérédité. Dans la plupart des cas, les résultats expérimentaux de *Mendel* étaient étonnamment proches de ceux prévus par sa théorie. *Fisher* montra que *Mendel*, sous l'hypothèse que sa théorie était exacte, et compte tenu de la variabilité naturelle des expériences, ne pouvait observer des résultats si proches des valeurs théoriques, qu'avec une probabilité inférieure à 0,00004<sup>8</sup>. On peut ainsi raisonnablement penser que *Mendel* avait truqué ses chiffres, ou du moins n'avait retenu que les expériences les plus favorables, pour mieux imposer sa théorie dans un environnement plutôt hostile.

Voyons l'exemple des petits pois.

<sup>8</sup> Exemple cité par A. Engel – "Les certitudes du hasard" – ALEA 1990.

Types de petits pois	Proportions théoriques
1. Jaune rond	$p_1 = \frac{9}{16}$
2. Jaune anguleux	$p_2 = \frac{3}{16}$
3. Vert rond	$p_3 = \frac{3}{16}$
4. Vert anguleux	$p_4 = \frac{1}{16}$

La théorie de *Mendel* prévoit que le croisement de petits pois jaunes et ronds avec des petits pois verts et anguleux donnera naissance à quatre nouvelles variétés, dans les proportions ci-contre.

*Mendel* a réalisé 556 observations sur les résultats de ces croisements, et comparé ses observations aux valeurs théoriques attendues.

Types de petits pois	Effectifs observés	Valeurs théoriques
1. Jaune rond	$x_1 = 315$	$t_1 = 312,75$
2. Jaune anguleux	$x_2 = 101$	$t_2 = 104,25$
3. Vert rond	$x_3 = 108$	$t_3 = 104,25$
4. Vert anguleux	$x_4 = 32$	$t_4 = 34,75$

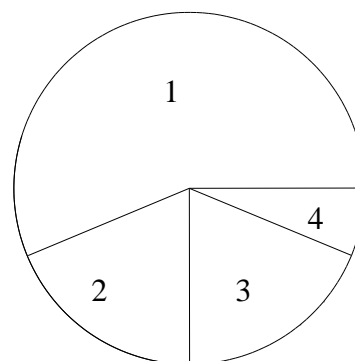
*Mendel* a-t-il eu de la chance ?

Bien sûr, sa théorie est un modèle. D'abord parce qu'on n'espère pas observer des quarts de petits pois. Ensuite parce que le hasard intervient et qu'une certaine variabilité "naturelle" entre les observations possibles est à prévoir.

Pour évaluer cette variabilité, on peut faire "tourner" le modèle de *Mendel*. En l'occurrence, il s'agit de faire tourner 556 fois la roue ci-contre, où les différents secteurs correspondent aux proportions théoriques des quatre espèces de petit pois.

On préférera sans doute recourir à une simulation, par exemple sur calculatrice, l'instruction :

$1 + \text{int}(16 * \text{rand})$  , où  $\text{int}$  correspond à la partie entière et  $\text{rand}$  au générateur de nombres aléatoires dans  $]0, 1[$ , fournit un entier aléatoire entre 1 et 16, de manière équirépartie. Il suffit ensuite de décider qu'entre 1 et 9, il s'agit de l'espèce 1, qu'entre 10 et 12, il s'agit de l'espèce 2, etc.



D'un point de vue probabiliste, on pourra introduire les quatre variables aléatoires  $X_1, \dots, X_4$  qui, à chaque réalisation de 556 expériences indépendantes, font respectivement correspondre l'effectif  $x_i$  de chaque espèce de petit pois. Les variables aléatoires  $X_i$ , sous l'hypothèse que le modèle de *Mendel* est le bon, suivent des lois binomiales  $B(556, p_i)$ , pour  $i$  allant de 1 à 4.

Il faut décider d'un critère de qualité (ou d'adéquation) des observations par rapport au modèle théorique. De façon classique, on choisira l'écart quadratique réduit, noté  $\chi_{\text{obs}}^2$  et

valant  $\chi_{\text{obs}}^2 = \sum_{i=1}^4 \frac{(x_i - t_i)^2}{t_i}$ . De grandes valeurs de  $\chi_{\text{obs}}^2$  rendraient le modèle de *Mendel*

suspect. Les observations de *Mendel* conduisent à  $\chi_{\text{obs}}^2 \approx 0,47$ . Est-ce bon, jusqu'à quel point ?

Pour étudier la variabilité de ce critère, introduisons la variable aléatoire :

$$T = \sum_{i=1}^4 \frac{(X_i - t_i)^2}{t_i} \quad \text{avec} \quad \sum_{i=1}^4 X_i = 556.$$

La loi de  $T$  suit approximativement une loi tabulée et connue sous le nom de loi du  $\chi^2$  à 3 degrés de liberté (en effet la relation ci-dessus fait que la valeur de  $X_4$  est déterminée dès que les valeurs de  $X_1$ ,  $X_2$  et  $X_3$  sont connues).

La table permet alors d'obtenir :  $P(T \geq 0,47) \approx 0,925$ .

*Mendel* avait donc environ 7,5% de chances d'obtenir de si bons résultats. On veut bien croire en cette chance, le problème est qu'en raison de l'indépendance des différentes sortes d'expériences, *Fisher* a pu additionner les différents  $\chi_{\text{obs}}^2$  et aboutit ainsi à  $\chi_{\text{obs}}^2 < 42$ , avec 84 degrés de liberté. C'est la probabilité que cet événement se produise qui est de l'ordre de 0,00004.

Une enquête sur les manuscrits de *Mendel* a, par la suite, montré que certains résultats d'expérience avaient été grattés et "corrigés"...

La variabilité est l'essence même de la statistique, et son omission un écueil qu'il faut éviter. *Daniel Schwartz* rapporte, à ce propos, l'anecdote suivante. Alors qu'il fit venir un agronome parce que ses rhododendrons fleurissaient mal, celui-ci diagnostiqua un manque de terre de bruyère. Il fit alors remarquer qu'un *roseum elegans*, qui fleurissait très bien, n'avait pas reçu du tout de terre de bruyère.

L'agronome rétorqua : - Voyons Monsieur Schwartz, vous qui êtes statisticien, vous raisonnez sur *un cas* ?

Mais quelques minutes après l'agronome griffonnait sur un carnet.

- Qu'écrivez-vous là ?
- Mais que le *roseum elegans* se passe de terre de bruyère.
- Eh là ! Vous raisonnez sur *un cas* !

Un autre exemple, où le particulier est pris pour le général, est celui de la réussite au bac scientifique (bac C à l'époque de l'enquête). Combien de fois, citant des élèves ayant réussi (ou échoué) "à cause" des matières littéraires, n'a-t-on fait des résultats dans ces matières un critère de réussite déterminant. Une analyse statistique plus poussée, fondée sur l'analyse des données multidimensionnelle, montre en fait que l'opposition littéraire – scientifique ou la réussite dans les disciplines littéraires n'entrent pratiquement pas dans les facteurs explicatifs de la réussite au baccalauréat C. En revanche, l'opposition mathématiques – sciences expérimentales ou sciences physiques – sciences naturelles sont des facteurs importants<sup>9</sup>. Cela ne veut pas dire qu'individuellement, la réussite dans les matières littéraires ne peut pas être déterminante pour la réussite au bac S, mais ce n'est pas, statistiquement (du point de vue de la gestion globale des élèves), un critère primordial.

L'idée d'appliquer la statistique dans certains domaines du variable n'a pas toujours été naturelle. Le traitement statistique d'une population conduit en effet à un certain **effacement de l'individu**. Le cas isolé n'intéresse pas le statisticien. Le médecin, par exemple, s'intéresse au contraire à un patient particulier. C'est ainsi que dans le domaine médical, la méthode statistique a rencontré de vives résistances (même de la part du grand *Claude Bernard*, pour qui le médecin ne peut soigner "en moyenne").

### **La variabilité "sauvage" de la bourse**

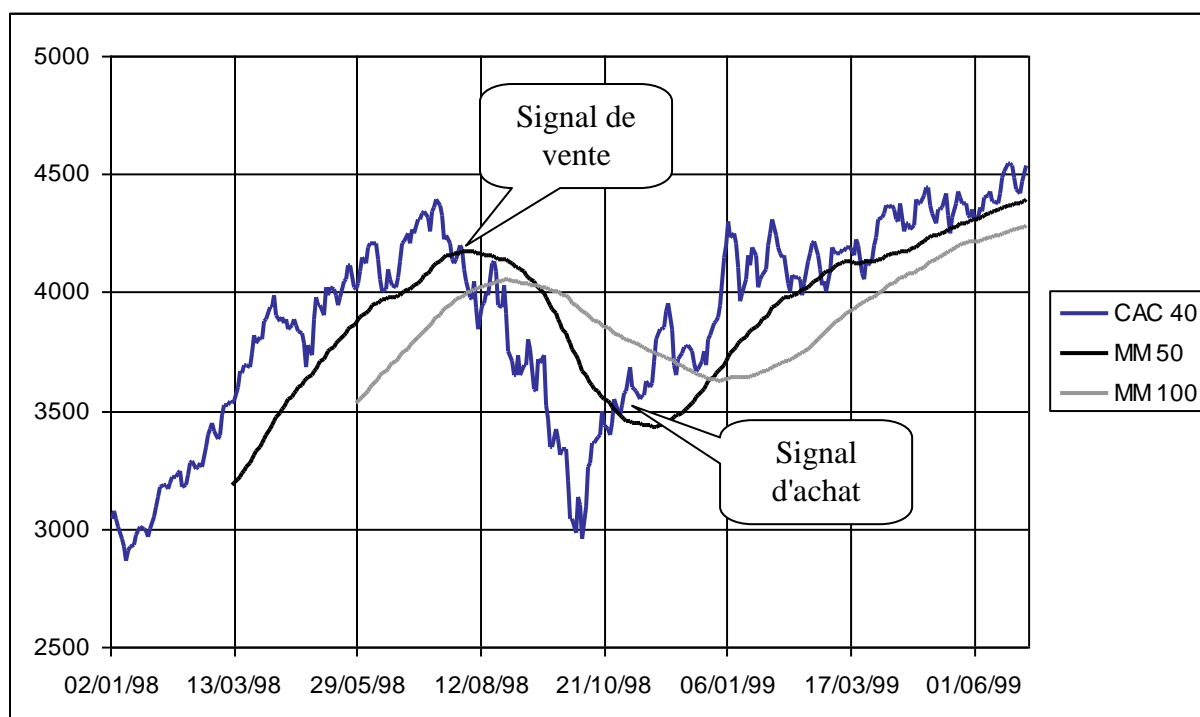
S'il est un domaine où la variabilité est particulièrement insaisissable, c'est bien celui de la bourse. Pour les variations des marchés financiers, *Benoît Mandelbrot* parle de "hasard

<sup>9</sup> J.L. Piednoir – Etude pour le ministère de l'Education Nationale. Voir chapitre 4.

sauvage", par opposition au "hasard bénin", celui qui, grosso modo, relève de la loi normale<sup>10</sup>. La finance moderne fait une forte consommation de statistique. Certains<sup>11</sup> y sentent parfois une odeur de "recettes alchimiques" : "la découverte de subtiles régularités dans les variations boursières permettraient à ceux qui en ont connaissance de transformer le plomb en or". Toutefois, il vaut mieux connaître ces différentes méthodes d'analyse statistique, afin de combattre à armes égales.

Un des outils statistiques les plus anciens, et les plus pratiqués dans le domaine financier, est celui des *moyennes mobiles*. C'est un moyen de lissage qui permet de gommer les mouvements erratiques des cours, pour n'en conserver que la tendance de fond, et obtenir ainsi un indicateur pour l'achat ou la vente.

Prenons l'exemple ci-dessous où la courbe montrant la plus forte variabilité est celle du niveau de clôture de l'indice CAC 40, pour la période de janvier 1998 à juin 1999.



La moyenne mobile d'ordre 50 (MM50) est tout simplement ici la moyenne arithmétique de 50 valeurs : celle du cours du jour et des 49 cours des séances précédentes (en général, on ne connaît pas le cours du lendemain). Son graphique ne débute donc qu'avec la 50<sup>ème</sup> donnée. Cette moyenne est dite "mobile" du fait que le calcul de la moyenne mobile consécutive ne diffère que par glissement d'une valeur (la plus ancienne disparaît au profit de la nouvelle).

On a également tracé la courbe des moyennes mobiles sur 100 jours. La moyenne mobile d'ordre 50 donne ici la tendance à court terme, alors que celle d'ordre 100, dont l'inertie est bien sûr plus grande, correspond à une tendance à plus moyen terme.

Les analystes financiers adoptent alors la règle suivante :

- acheter quand la moyenne mobile croise les cours à la hausse,
- vendre quand la moyenne mobile croise les cours à la baisse.

<sup>10</sup> "Du hasard bénin au hasard sauvage" – Benoît Mandelbrot – Dossier "Pour la Science" – "Le hasard" – Avril 1996.

Pour une utilisation du modèle fractal, voir "Randonnées multifractales à travers Wall Street" – Benoît Mandelbrot – Dossier "Pour la Science" – "Les mathématiques sociales" – Juillet 1999.

<sup>11</sup> "Les marchés aléatoires" – Jean-Philippe Bouchaud et Christian Walter – Dossier "Pour la Science" – "Le hasard" – Avril 1996.

Cela fonctionne plutôt bien avec la courbe MM50, début août 98 et fin octobre 98, mais bien sûr, ça ne marche pas à tous les coups (ça se saurait...). Il arrive fréquemment que, quelques jours après, les cours croisent de nouveau la moyenne mobile. Dans ce cas, ce sont de mauvais signaux. On peut attendre un peu la confirmation d'un signal... mais pas trop. On peut aussi utiliser plusieurs moyennes mobiles, ou croiser avec d'autres indicateurs. Vous avez dit alchimie ?

L'analyse précédente de la série temporelle qu'était le niveau du CAC 40 n'était que descriptive (déterministe). On peut chercher à raffiner le modèle en introduisant l'aléatoire. On pourra, par exemple, dans un modèle additif, décomposer une série temporelle  $y(t)$  sous la forme :  $y(t) = T(t) + C(t) + S(t) + \varepsilon(t)$  où  $T$  est la tendance linéaire,  $C$  la composante cyclique,  $S$  la composante saisonnière et le résidu  $\varepsilon$  une variable aléatoire. La tendance linéaire  $T$  pourra par exemple s'obtenir à l'aide de la méthode des moindres carrés. La méthode des moyennes mobiles correspondra à  $T + C$  (elle gomme les variations saisonnières et les aléas). On pourra en déduire  $S$  et chercher à modéliser la loi de  $\varepsilon$ .

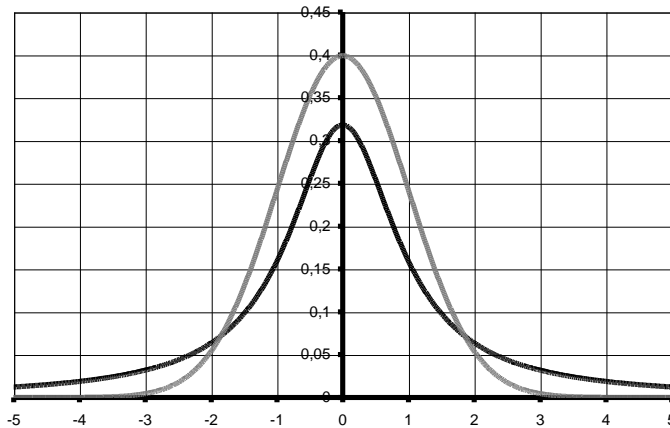
L'aléa  $\varepsilon$  pourra lui-même être décomposé en une partie auto-corrélée, c'est à dire dépendant du passé, et une partie de "bruit pur", indépendante.

Dans le cas des marchés financiers, la modélisation du "**bruit**" par une loi normale est déficiente car celle-ci s'écarte trop rarement de façon importante de sa moyenne (la queue de la courbe de *Gauss* est très aplatie), alors que les soubresauts du marchés correspondent à de violentes déviations (dues au comportement mimétique des opérateurs ou à l'impact d'un événement d'actualité). *Christian Walter*<sup>12</sup> cite l'exemple de l'étude sur 70 ans de la rentabilité de l'indice boursier américain SP500, pour lequel on observe une moyenne de 0,81 % avec un écart type de 5,82 %. Cependant, la baisse mensuelle maximale observée durant ces 70 ans a été de - 35,28 %. Dans un modèle de variabilité gaussien, une telle observation est quasi impossible : si une variable aléatoire  $X$  suit la loi normale  $N(0,81 ; 5,82)$ , la probabilité  $P(X \leq - 35,28)$  est de l'ordre de  $3.10^{-10}$ , autant dire que l'évènement est humainement inobservable. En effet, si pendant  $n$  mois on réalise l'expérience aléatoire consistant à voir si l'évènement " $X \leq - 35,28$ " se réalise (avec la probabilité  $p = 3.10^{-10}$ ), ou non, l'espérance du nombre de ces réalisations est  $n \times p$  (loi binomiale). Pour espérer voir une réalisation, il faut donc que  $n \times 3.10^{-10} = 1$  d'où  $n = (1/3) \times 10^{10}$  mois c'est à dire de l'ordre de 300 millions d'années.

---

<sup>12</sup> "*Les impossibles de la finance*" – Pour la science n° 225 – Juillet 1996.

Si la loi normale n'est pas capable de prendre en compte les krachs boursiers, on pourra songer à une distribution de *Cauchy* ou à celles de *Paul Lévy*, découvertes dans les années 1930. Mais la difficulté de ces distribution est qu'elles possèdent une espérance et un écart type infinis et donc que les théorèmes limites classiques ne s'appliquent pas.



Ci-contre, est représentée la densité de la loi de Cauchy, définie par  $f(x) = \frac{1}{\pi(1+t^2)}$ , "à queue épaisse", comparée à la densité de la loi normale centrée réduite.

## Incertitude, hasard et déterminisme

Une façon de mesurer l'incertitude issue de la variabilité est d'introduire le "*hasard*", même dans des situations a priori déterministes. Un hasard qui résulte de notre ignorance.

*Henri Poincaré*<sup>13</sup> affirmait que "*Tout phénomène, si minime qu'il soit, a une cause, et un esprit infiniment puissant, infiniment bien informé des lois de la nature, aurait pu le prévoir depuis le commencement des siècles. Si pareil esprit existait, on ne pourrait jouer avec lui à un jeu de hasard, on perdrait toujours. Pour lui, en effet, ce mot de hasard n'aurait pas de sens, ou plutôt, il n'y aurait pas de hasard. C'est à cause de notre faiblesse et de notre ignorance qu'il y en a un pour nous.*"

C'est cette sorte de hasard que l'on rencontre dans les phénomènes de *chaos déterministe*, comme le lancer d'un dé, dont on connaît les lois mécaniques déterministes, mais dont la sensibilité aux conditions initiales fait que l'issue est tout à fait incertaine. Que l'on songe également aux suites récurrentes, parfaitement déterministes, puisque calculées, que l'on utilise comme générateurs de nombres pseudo aléatoires : lorsque l'on ignore leur mode de calcul, les résultats paraissent parfaitement aléatoires et passent même tous les tests statistiques. C'est à dire qu'ils se comportent "comme si".

Remarquons qu'inversement, le hasard peut créer du déterminisme (heureusement pour la statistique), par la loi des grands nombres. En particulier, les lois déterministes de la physique à l'échelle macroscopique résultent de l'accumulation d'un très grands nombre de phénomènes aléatoires à l'échelle microscopique.

Depuis la mécanique quantique, il semblerait en effet qu'existe bien un "*hasard primordial*" en physique, existant sans qu'il soit possible de trouver une explication plus profonde, bien que notre esprit répugne à cette idée, en témoigne le "*Dieu ne joue pas aux dés*" d'*Einstein*.

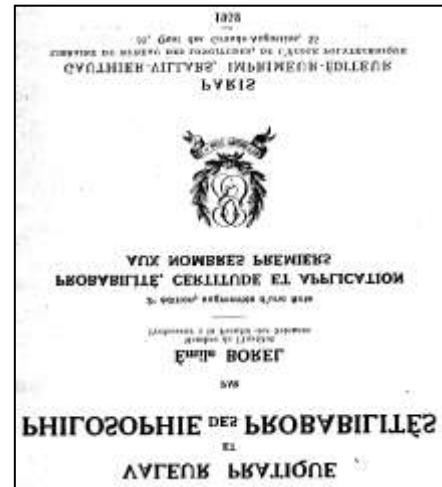
<sup>13</sup> "Science et méthode" – 1908.



## Le hasard n'est pas raisonnable

La psychologie humaine nous prépare mal aux phénomènes aléatoires, dont les résultats nous paraissent souvent paradoxaux. Selon *Emile Borel*<sup>14</sup>, "Il n'est pas possible à l'esprit humain d'imiter parfaitement le hasard, c'est à dire de substituer un mécanisme rationnel quelconque à la méthode empirique qui consiste à effectuer une suite indéfinie d'épreuves répétées".

Ainsi, il sera généralement facile, entre deux listes de 200 chiffres 0 et 1 "au hasard" de détecter celle qui a été imaginée par un être humain et celle qui résulte d'une réelle expérience de pile ou face.



Des deux séries suivantes (en lignes), quelle est celle imaginée "au hasard" par l'homme ?

Série 1 :

0	1	0	0	1	0	1	0	0	0	1	1	0	1	1	0	0	1	1	0
0	0	1	1	1	0	1	0	1	0	0	1	1	0	0	1	1	1	0	0
1	0	1	1	0	1	1	0	0	0	0	1	1	1	0	1	0	1	0	0
1	1	0	0	0	1	1	1	0	1	0	1	0	0	1	1	0	0	0	1
0	0	0	1	1	0	1	1	0	1	1	0	1	1	0	0	1	1	0	0
0	0	1	0	0	1	0	0	0	1	1	0	1	1	0	0	1	1	1	1
0	0	0	1	0	1	0	1	1	0	0	1	1	1	0	1	0	0	1	1
1	0	0	1	0	1	1	0	0	1	1	0	0	1	1	0	1	0	1	0
1	0	0	1	1	0	1	1	1	0	1	0	1	0	1	0	1	1	0	1
0	1	0	0	1	0	1	0	0	1	1	0	0	0	0	1	1	0	0	1

Série 2 :

1	1	1	1	0	1	0	1	0	0	0	1	0	0	0	0	1	1	0	1
0	0	0	0	0	1	1	0	1	0	0	1	0	0	1	1	1	1	0	0
1	1	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	1	1	1
1	0	0	1	0	1	0	0	0	0	1	1	0	0	1	1	0	0	1	1
1	1	0	0	0	1	1	0	1	1	1	0	0	0	1	0	0	0	0	1
0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	1	1	0	1	0
0	1	1	0	0	1	1	1	1	0	0	1	1	0	0	1	1	0	0	1
1	0	1	1	0	1	1	1	0	1	1	0	1	1	1	0	1	0	0	1
0	0	0	0	1	0	1	0	0	0	1	1	0	0	1	1	0	1	0	0
0	1	1	1	0	0	0	0	0	0	1	1	1	0	1	1	0	1	0	0

Et bien, d'une part on montre que la probabilité d'avoir, à pile ou face, au moins 6 résultats consécutifs égaux, sur 200 lancers, est d'environ 0,96 , d'autre part les psychologues ont constaté que l'esprit humain n' imagine pas qu'une telle série puisse être aussi fréquemment l'effet du hasard. Ce qui permet de penser que la première série ci-dessus a été imaginée par l'homme. Encore une fois, il n'y a pas de certitude dans cette affirmation, mais le risque d'erreur est de l'ordre de 4%, et le fait est que la première série a bien été imaginée par un être humain.

<sup>14</sup> "Valeur pratique et philosophie des probabilités" – 1938.

Paradoxalement, le caractère "inhumain" du hasard est parfois un allié du statisticien. Lorsqu'il s'agit, pour un sondage, de prélever un *échantillon représentatif* de la population, il y a eu longtemps débat au sein de l'*Institut International de la Statistique*, dès 1895, d'abord sur le bien fondé de l'échantillonnage, ensuite sur la méthode, entre celle du "choix raisonné" de l'échantillon et celle du choix aléatoire. Ce débat eut lieu au moins jusqu'à l'élection américaine de 1936<sup>15</sup>, où la méthode aléatoire a spectaculairement démontré sa supériorité sur un choix "raisonné" qui, d'une part induit trop souvent des biais cachés, et, d'autre part, empêche l'application de la théorie des probabilités lorsque l'aléatoire n'y est pas quantifiable.

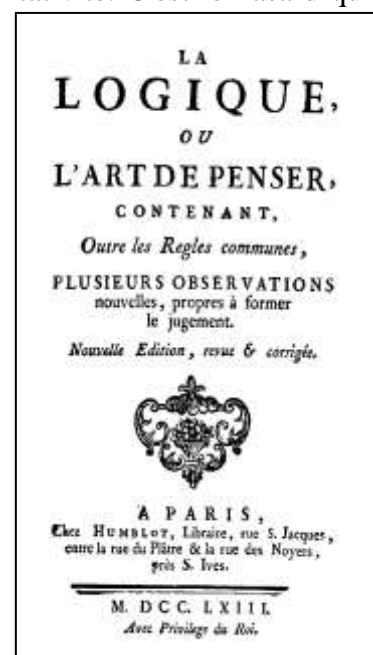
Dans de nombreux sondages cependant (dont celui de 1936), on utilise des quotas, ou un "échantillonnage stratifié", selon le sexe, l'âge, la catégorie socioprofessionnelle... suivant leurs proportions dans la population. On profite ainsi d'une information supplémentaire, fournie par le recensement, et en rapport avec l'étude. Mais, à l'intérieur de chaque classe, le hasard est encore le meilleur allié pour garantir la représentativité. C'est le hasard qui fonde la théorie mathématique des sondages.

## Mesurer le hasard

Paradoxe encore, ce hasard, par définition non raisonnable, on le mesure. Ce qui étonnait tant le pionnier que fut *Pascal*, présentant ce "titre stupéfiant : *La géométrie du Hasard*".

La statistique, dont la fonction est de mesurer l'incertain à partir d'observations, n'est pas restée à l'écart du débat sur le concept de probabilité, qui quantifie cette incertitude.

Le sens du mot *probabilité* lui-même a évolué. En effet, *probabilité* signifiait au Moyen Age "avis certifié par l'autorité" (un avis "probant"). Au XVII<sup>e</sup> siècle, on le trouve employé, dans "*La logique de Port-Royal*" d'*Antoine Arnauld* et *Pierre Nicole* (première édition en 1662 et dont se réclame *Jacques Bernoulli*) avec le sens de "raison de croire". Ainsi, on y lit que "*la conduite de la vie, [...] ne demande pas de plus grande certitude que cette certitude morale, et qui doit même se contenter en plusieurs rencontres, de la plus grande probabilité, [...] le mieux que nous puissions faire, quand nous sommes engagés à prendre parti, est d'embrasser le plus probable, puisque ce serait un renversement de la raison d'embrasser le moins probable.*"



<sup>15</sup> Voir chapitre 3.

### Le point de vue objectiviste

A l'article "Probabilité" de l'*Encyclopédie méthodique* (1784), Condorcet attribue deux "sources de probabilité" :

Nous les réduisons à deux espèces ; l'une renferme les *probabilités* tirées de la considération de la nature même, & du nombre des causes ou des raisons qui peuvent influencer sur la vérité de la proposition dont il s'agit ; l'autre n'est fondée que sur l'expérience du passé, qui peut nous faire tirer avec confiance des conjectures pour l'avenir, lors du moins que nous sommes assurés que les mêmes causes qui ont produit le passé existent encore, & sont prêtes à produire l'avenir.



Les probabilités tirées "du nombre des causes", c'est à dire, dans le cadre de l'équiprobabilité, du *rapport des cas favorables aux cas possibles*, ne sont calculables que dans un cadre très limité, grosso modo, celui des jeux de hasard et des modèles d'urnes, dont les règles sont déterminées et la modélisation assez simple. Cette approche, reposant sur l'équiprobabilité des issues, qui n'est pas assurée dans nombre de problèmes, est inopérante dans l'étude de la mortalité, des risques à assurer, des contrôles de qualité dans l'industrie..., en fait des probabilités rencontrées "dans la vie".

L'autre approche, "fondée sur l'expérience du passé", c'est-à-dire sur l'observation des fréquences lorsque l'on répète un grand nombre de fois l'expérience, ouvre davantage de perspectives (au moins dans les cas d'expériences répétables).

Cette *approche fréquentiste* des probabilités est fondée sur la *loi des grands nombres* dont Jacques Bernoulli est à l'origine ("L'Art de conjecturer" 1713).

La principale critique de l'approche fréquentiste est que la définition de probabilité repose alors sur la loi des grands nombres qui, elle-même, suppose définie la probabilité.

Il faut attendre Kolmogorov (1933) pour que soit fondée une *théorie axiomatique des probabilités*, en ignorant volontairement l'interprétation de la probabilité et l'utilisation qui peut en être faite. Cette théorie des probabilités, alors purement mathématique, est basée sur la théorie de la mesure de Borel (1897), l'intégrale de Lebesgue (1901) et la mesure abstraite de Radon (1913). Le calcul des probabilités fournit des *modèles*, mais il ne dit pas quelles lois choisir. L'évaluation des probabilités "physiques" pose le problème de la modélisation et de l'adéquation du modèle à la réalité, éventuellement avec essais et rectifications.

Dans les applications, et pour le statisticien, le point de vue "classique" reste l'approche fréquentiste, privilégiée lorsque l'on possède des données en nombre suffisant (événements répétables) pour utiliser les théorèmes limites. C'est une conception *objective* de la probabilité, en ce sens que l'on admet l'existence d'une valeur déterminée à la probabilité d'un événement, valeur que l'on cherche à estimer. Dans cette optique, Emile Borel affirmait que "les probabilités doivent être regardées comme des grandeurs physiques, [...] avec une certaine approximation".

Cette conception est en particulier celle de Jerzy Neyman pour la mise en place de la théorie des tests d'hypothèses : le paramètre du modèle est une valeur certaine, fixée, mais inconnue et l'on teste une hypothèse la concernant (hypothèse nulle). Cette démarche fut combattue par Ronald Fisher, dont la philosophie de la statistique était différente.

### Le point de vue subjectiviste

François Le Lionnais<sup>16</sup> affirmait en 1948 que "[quant au] problème crucial de l'origine, subjective ou objective, de la notion de Probabilité. Il s'agit là d'une confrontation dont la connaissance devrait faire partie, de nos jours, du bagage de tout homme cultivé."

En effet, une autre approche de la probabilité la considère davantage comme la mesure des raisons que nous avons de croire en la réalisation d'un évènement, dans l'ignorance où nous sommes et *relativement* aux informations dont nous disposons. La probabilité est une opinion sur les choses. Un assureur sur la vie ne donnera pas à la probabilité qu'a un nouveau client de décéder dans les 10 années qui viennent, la même valeur selon les informations dont ils dispose (âge, sexe, antécédents médicaux ou familiaux...). La probabilité est donc révisable en fonction d'informations nouvelles, elle peut varier selon les circonstances ou l'observateur, elle est *subjective*. C'est le point de vue *bayésien* (Thomas Bayes 1702-1761). Le *théorème de Bayes*, énoncé et démontré indépendamment par Laplace en 1774, dans son "*Mémoire sur la probabilité des causes*" permet une méthode d'estimation basée sur les probabilités conditionnelles. Si différentes "causes", les évènements  $A_i$  avec  $i = 1, \dots, n$ , peuvent "provoquer" l'évènement  $E$ , la probabilité *a priori* de  $E$  est donnée par :

$$P(E) = \sum_{i=1}^{i=n} P(E|A_i) \times P(A_i).$$

Inversement, si l'évènement  $E$  est observé, la probabilité *a posteriori* de la "cause"  $A_j$  est

$$\text{donnée par : } P(A_j|E) = \frac{P(A_j \cap E)}{P(E)} = \frac{P(A_j) \times P(E|A_j)}{\sum_{i=1}^{i=n} P(E|A_i) \times P(A_i)}.$$

Dans l'ignorance totale des  $P(A_i)$ , Laplace, selon le "*principe de la raison insuffisante*",

$$\text{leur attribue une valeur uniforme } P(A_i) = \frac{1}{n}. \text{ On a alors } P(A_j|E) = \frac{P(E|A_j)}{\sum_{i=1}^{i=n} P(E|A_i)}.$$

De Finetti alla jusqu'à affirmer : "*La probabilité n'existe pas*". Et de la ranger au rayon des antiques croyances comme celles de "*l'éther, de l'espace et du temps absolu... ou des fées. La probabilité, considérée comme quelque chose ayant une existence objective est également une conception erronée et dangereuse*"<sup>17</sup>.

De ce point de vue, la répétition n'est plus nécessaire pour probabiliser, et on peut probabiliser l'incertain même s'il n'est pas aléatoire. Ainsi, pour l'estimation d'un paramètre (non aléatoire), on considère ce paramètre comme une variable aléatoire, dont les différentes valeurs sont les "causes" possibles des observations. Ronald Fischer, le père de la théorie de l'estimation, était relativement proche de cette conception philosophique de la statistique<sup>18</sup>.

## En conclusion

Dans de nombreux domaines, scientifiques, industriels, économiques, la prise en compte de la variabilité passe par la méthode statistique, pour estimer, prévoir ou aider à la prise de décision. Les deux philosophies de la statistique ont, chacune à leur manière, enrichi les

<sup>16</sup> dans "*Les grands courants de la pensée mathématique*" – Réédition "Rivages" 1986.

<sup>17</sup> cité par Saporta dans "*Probabilités, analyse des données et statistique*" – Technip 1990.

<sup>18</sup> à propos des oppositions entre estimation subjectiviste et test d'hypothèses fréquentiste, on pourra se rapporter au chapitre 3 et, en particulier, à l'anecdote des "faiseurs de pluie".

techniques. L'une ou l'autre approche étant plus performante, selon la nature du problème (répétable ou non répétable, valeurs du paramètre à privilégier ou non).

Mais qu'est-ce que la statistique ? Quelles sont précisément ses méthodes ? Beaucoup se fondent sur le calcul des probabilités, d'autres sur une vision géométrique. C'est le propos du chapitre suivant.

# 2

## QU'EST-CE QUE LA STATISTIQUE ?

### I – VOUS AVEZ DIT STATISTIQUE ?

Prononcez le mot statistique devant un public même cultivé quelles représentations mentales suscitez-vous ? Comme son enseignement, jusqu'à une date récente, était quasi confidentiel ou bien réduit à quelques recettes dans des enseignements supérieurs spécialisés (économie, sciences humaines, biologie), vos interlocuteurs penseront à quelques applications vulgarisées par la presse : sondages d'opinion, estimation des résultats électoraux à 20h01, ou bien à des procédures fort complexes utilisées par des ingénieurs spécialisés dans les entreprises. Le plus souvent la méfiance s'installe. L'utilisateur de la statistique est souvent vu comme un manipulateur déguisant la réalité pour présenter ses préjugés comme une vérité objective d'où l'adage : "on fait dire ce que l'on veut à la statistique" ou cette citation d'un homme célèbre : "la statistique est la forme élaborée du mensonge".

Pourtant la pratique de la statistique a beaucoup progressé ces dernières années dans la recherche, dans les services, dans la production. Le système éducatif en a pris acte. Depuis les années quatre-vingt, un chapitre statistique est apparu dans les programmes de collège, de seconde générale et technologique. Les applications industrielles, commerciales ont amené l'introduction de ces techniques dans les classes de techniciens supérieurs, dans celles préparant au diplôme d'études comptables et financières. Récemment plusieurs secteurs économiques ont demandé d'en intensifier l'enseignement.

On est loin de la méfiance exprimée plus haut. Les techniques statistiques se sont imposées du fait même de leur efficacité. L'objectif visé dans les quelques pages qui suivent est de montrer, à partir de quelques exemples, quelle est l'originalité de la démarche statistique, quel est l'intérêt des techniques utilisées mais aussi leur limite. Pour les décrire il nous faudra faire appel à des structures mathématiques diverses, certaines très simples comme la proportion, d'autres plus complexes comme l'algèbre linéaire, la géométrie euclidienne et surtout le calcul des probabilités. On distinguera la description d'une population et le jugement sur échantillon. A chaque fois, on montrera que le choix d'une technique statistique est profondément lié à l'usage que l'on veut en faire, et que sa pertinence est liée à un protocole rigoureux. La statistique a aussi à voir avec l'action.

### II – QUELQUES EXEMPLES

Comme on fait de la statistique pour émettre des hypothèses de recherche, pour prendre des décisions, les exemples abondent et beaucoup de conclusions, tirées à partir d'une approche statistique d'un problème, concernent la vie courante de nombreux citoyens. Ceux qui suivent sont choisis pour leur intérêt historique ou didactique, d'autres seront évoqués, par la suite, à propos de telle ou telle technique.

### **Exemple 1 : le commerçant**

Le commerce est une très vieille activité humaine. Chaque jour, chaque semaine, le commerçant note ses ventes. Evidemment elles varient d'une période à une autre. Pourtant il lui faut se réapprovisionner et passer commande à son fournisseur. S'il commande trop peu il y a manque à gagner, s'il commande trop il reste des invendus. C'est à partir des informations sur le volume des ventes antérieures qu'il prend sa décision. Voilà quelqu'un qui fait de la statistique sans le savoir !

### **Exemple 2 : Florence Nightingale, l'ange soigneur (1820-1910)**

*Florence Nightingale*, fille de bonne famille ayant eu connaissance des travaux statistiques de *Quételet*, statisticien précurseur belge, est frappée par la mortalité sévissant dans l'armée anglaise pendant la guerre de Crimée (1854-1855). Elle note le nombre de morts selon les causes de décès : morts violentes, morts évitables, morts autres. Elle poursuit son étude en Angleterre, sur l'armée qui y est stationnée. Comme il n'y a pas de décès pour fait de guerre, elle compare la mortalité des soldats anglais et celle des Anglais mâles du même âge. La différence est telle, qu'après publication des résultats dans la presse, un mouvement d'opinion se fait jour. Après cela, malgré de fortes résistances, l'Etat-Major britannique est obligé de réformer de fond en comble l'organisation de son système de santé. Bel exemple d'une application statistique élémentaire (on ne calcule que des proportions), efficace.

### **Exemple 3 : la réussite des élèves de terminale**

Pour 993 élèves de terminale C (1989) répartis dans 21 lycées, on dispose des notes obtenues à chacun des trois trimestres et au baccalauréat dans 5 disciplines, soit 20 notes pour chaque élève.

On se pose un certain nombre de questions :

Y a-t-il relation entre les notes trimestrielles et celles du baccalauréat ?

Y a-t-il relation entre les notes des différentes disciplines ?

Y a-t-il des différences entre lycées dans les modes de notation ?

### **Exemple 4 : le lot de cartouches**

A la veille de l'ouverture de la chasse, un commerçant vend à ses clients des cartouches. Il commande à un fabricant une caisse afin de pouvoir la détailler. Il voudrait connaître la qualité de ce lot, c'est à dire la proportion  $\theta$  de mauvaises cartouches. Il y a évidemment une méthode pour connaître  $\theta$ , tirer toutes les cartouches et compter celles qui font long feu. Mais après, il n'y en a évidemment plus aucune à vendre ! Comment avoir des informations sur  $\theta$ , sans détruire tout le lot ?

### **Exemple 5 : les calculs astronomiques**

L'observation des étoiles et des planètes est une très vieille histoire, de près de 5 000 ans. *Ptolémée*, au premier siècle de notre ère, à partir des mesures faites précédemment, a mis au point le premier modèle capable de prévoir le mouvement apparent des planètes du système solaire. Avec l'apparition des lunettes astronomiques, au XVI<sup>ème</sup> siècle, les mesures se multiplient et gagnent en précision. Hélas, d'une observation à l'autre, la mesure fluctue. Il y a des erreurs. Comment réduire leur importance en utilisant toutes les observations pour obtenir la meilleure précision possible sur les positions futures des planètes dans le ciel ? *Simpson*, mathématicien anglais, s'attaqua au problème en 1756. Vingt ans plus tard les français *Lagrange* et *Laplace* apportèrent leur contribution. C'est

*Carl Friedrich Gauss* (1777-1855) qui devait fournir un cadre théorique adéquat pour trouver la solution encore mise en œuvre.

### **Exemple 6 : le médicament est-il efficace ?**

Un laboratoire pharmaceutique met au point un médicament susceptible de faire baisser la tension artérielle des individus. Immédiatement se pose la question : est-il efficace ? Si oui, de combien ? Mais la tension artérielle n'est pas un concept simple, elle varie d'un individu à l'autre et, pour un même individu, d'un moment à l'autre. Comment caractériser cette donnée, comment évaluer l'action du médicament ?

### **Exemple 7 : les clients du supermarché**

Une innovation majeure est apparue récemment : le code barre. Quand vous passez à la caisse, l'employée enregistre tous vos achats : les produits que vous avez achetés, avec pour chacun la quantité prise. Toutes ces données sont enregistrées. Que faire de cette masse énorme d'informations pour améliorer les performances de la grande surface ?

## **III – PREMIERE FORMALISATION**

Dans les exemples précédents que peut-on observer ? A chaque fois il y a des unités appelées aussi individus sur lesquelles on fait des observations : de quoi est mort tel soldat anglais, quelles sont les notes obtenues par un élève, la cartouche choisie est bonne ou mauvaise, le patient a, à un instant donné, avec ou sans prise de médicament, une tension artérielle etc. Mais on voit bien que **l'individu lui-même ne nous intéresse pas**. L'objet de l'étude est plus vaste, l'individu n'est observé que parce qu'il participe au phénomène étudié : causes de mortalité dans l'armée anglaise, notation des élèves de terminale C, qualité du lot de cartouches, efficacité du médicament. Mais d'un individu à l'autre, on note des **variations** relativement importantes. Si toutes les observations étaient identiques, il n'y aurait plus de problème. L'objet de la statistique est de prendre en compte ces variations.

Pour cela on va définir un cadre mathématique adéquat. Il nous faut rattacher les individus observés à un cadre plus vaste. L'objet d'étude sera donc un collectif appelé souvent **population** et noté  $P$ . Il est composé d'éléments appelés aussi individus. Dans certains exemples précédents,  $P$  a une existence physique évidente et il faut le définir précisément. Ce sont les soldats de l'armée anglaise, les élèves de terminale C des 21 lycées étudiés, le lot de cartouches. Dans d'autres, il faut l'imaginer, le conceptualiser. Il en est ainsi de l'ensemble des états de tensions possibles dans l'étude de l'efficacité du médicament.

On veut donc définir et étudier diverses caractéristiques de  $P$ , mais  $P$  n'est pas accessible directement. On ne peut accéder à ces caractéristiques qu'en effectuant des mesures sur les individus qui le composent. Dans certains cas, on considère tous les éléments de  $P$ , c'est le cas par exemple des élèves de terminale C des 21 lycées étudiés. Dans d'autres, on n'en prend qu'une partie, soit par commodité (pour les enquêtes sur les intentions de vote, il serait dispendieux d'interroger tous les électeurs), soit par structure (il est impossible de mesurer toutes les tensions artérielles de toute la population concernée à tout moment), soit pour des raisons économiques (quand il s'agit par exemple d'essais destructifs comme dans le cas des cartouches). On appelle **échantillon** la partie de la population dont les individus seront étudiés. Elle sera notée  $E$  avec  $E \subset P$ .

Sur chaque individu de  $E$ , on va procéder à des **observations**. Celles-ci peuvent être décrites par des éléments d'un ensemble  $X$ . On établit donc une application notée  $x$  de  $E$  dans  $X$ . On notera  $i$  l'individu de  $E$  et  $x_i$  la valeur prise par  $x$  en  $i$ ,  $x$  est appelé un



**caractère.** Dans le cas des notes des élèves, on a  $X = \mathbb{R}^{21}$ , dans le cas des soldats anglais  $X = \{\text{mort violente ; mort évitable ; mort autre ; vivant}\}$ . Pour les cartouches  $X = \{0, 1\}$ , en codant 1 quand la cartouche est mauvaise, 0 quand elle est bonne. Pour la tension artérielle on a  $X$  intervalle de  $\mathbb{R}$ . A noter que l'ensemble  $E$  est toujours fini, car un observateur ne peut faire qu'un nombre fini de mesures et donc  $E$  pourra être noté  $\{1, 2, \dots, n\}$ .

L'ensemble  $X$  peut être sans structure, comme dans le cas des soldats anglais, le **caractère**  $x$  est dit alors **nominal** (ou **qualitatif**). Il est dit **numérique**(ou **quantitatif**) si  $X = \mathbb{R}$ , **multidimensionnel** si  $X = \mathbb{R}^k$  (cf. les notes des élèves). Il existe des cas où  $X$  est un ensemble muni d'une structure d'ordre total ; quand on définit par exemple des échelles d'attitudes en psychologie, elles sont souvent notées par un nombre réel, mais il est bien évident que seule la structure d'ordre de  $\mathbb{R}$  intervient. Il n'est pas possible d'additionner des attitudes, on peut seulement les comparer. A noter que l'ensemble  $X$  est souvent un produit cartésien. Sur un même individu, on peut procéder à l'observation de plusieurs caractères de natures différentes. Tel est le cas si, pour les élèves des 21 lycées, outre leurs notes, on prend en compte la catégorie socioprofessionnelle du père, le diplôme le plus élevé de la mère etc... On a alors  $X = X_1 \times X_2 \times \dots \times X_k$ , lorsqu'il y a  $k$  caractères élémentaires étudiés.

**On a donc fait sur les  $n$  individus de l'échantillon des observations qui sont des valeurs prises dans un ensemble  $X$ . Mais les individus ne nous intéressent pas, c'est la population  $P$  que l'on veut étudier.**

Peu importe que telle ou telle cartouche soit bonne ou mauvaise, c'est la proportion de mauvaises cartouches dans le lot acheté que l'on cherche à connaître. Que l'élève *Dupont* ait obtenu au baccalauréat des notes supérieures à celles de ses devoirs surveillés ne nous intéresse pas. Par contre, pour un lycée donné, comment sont notés les élèves, quelles sont leurs performances au baccalauréat, quelle valeur prédictive ont les notes obtenues au cours de l'année, voilà notre sujet d'étude.

**Il va donc falloir combiner les observations faites sur les individus pour aboutir à une caractérisation de la population.** Cette opération peut également être énoncée dans le langage de l'information au sens commun du terme. On a recueilli sur les individus certaines informations, parfois très nombreuses, pensons aux notes des élèves de terminale des 21 lycées où l'enquêteur a récolté  $993 \times 20$  soit 19 860 notes. Pour les tickets de supermarché, le nombre d'informations est encore plus considérable, nous avons des millions de données stockées dans l'ordinateur. Il est bien évident que l'examen de données brutes aussi nombreuses est impossible, et, sauf cas exceptionnel, ne donne aucun renseignement sur la population. Il faut trouver des **résumés** pertinents de cette masse de données, résumés caractérisant la population dont font partie les individus. A noter, bien évidemment, que si tous les individus étaient identiques, la question ne se poserait pas. La caractéristique de la population serait celle de l'individu. **Il n'y a démarche statistique que parce qu'il y a variabilité des observations** faites sur les individus.

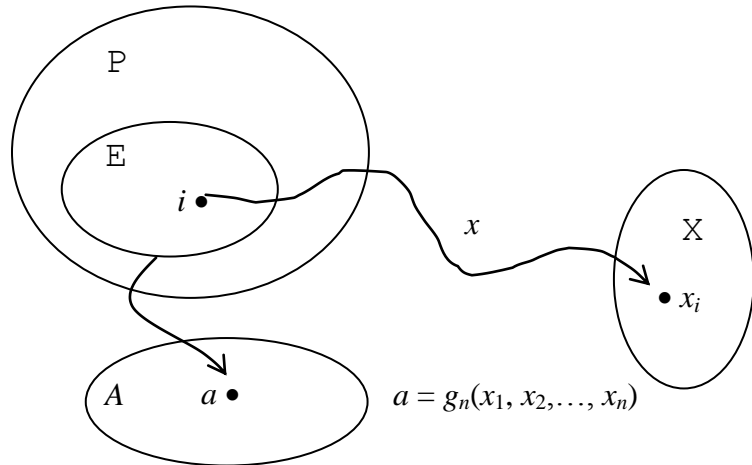
On peut traduire mathématiquement l'opération de combinaison des observations faites sur les individus pour aboutir à une caractérisation de la population. On a noté  $E$  l'échantillon observé, avec  $E = \{1, 2, \dots, n\}$ . Soit  $x$  le caractère étudié prenant ses valeurs dans l'ensemble  $X$ . On a  $x : E \rightarrow X$ . Si  $i \in E$ ,  $x_i$  sera l'observation faite sur l'individu  $i$ .

On dispose donc de  $n$  données  $(x_1, x_2, \dots, x_n)$ . Appelons  $A$  l'ensemble dans lequel la caractéristique de la population prend sa valeur. Le choix d'un élément  $a$  dans  $A$  va dépendre évidemment de  $(x_1, x_2, \dots, x_n)$ . Il faut donc trouver une application que l'on notera

$g_n$ , avec  $g_n : X^n \rightarrow A$ . L'application  $g_n$  est un **résumé statistique**. Le travail du statisticien consiste à trouver les cadres théoriques permettant de choisir des résumés  $g_n$  pertinents, en fonction des situations étudiées et des objectifs recherchés.

Le schéma ci-dessous résume la démarche statistique :

P est la population.  
 E est l'échantillon.  
 $i$  est un individu.  
 $x$  est le caractère étudié, prenant ses valeurs dans  $X$ .  
 $x_i$  est la valeur du caractère, observée sur l'individu  $i$ .  
 $g_n : X^n \rightarrow A$  est le résumé statistique.  
 A est l'ensemble dans lequel  $g_n$  prend ses valeurs.  
 $a$  est la valeur de A qui caractérise la population.



## IV – LA RECHERCHE DU RÉSUMÉ

La formulation précédente est évidemment trop générale pour être opérationnelle. A supposer que  $A$  soit bien défini, ce qui n'est pas toujours le cas, il existe une infinité ou au moins un très grand nombre d'applications  $g_n$  possibles. Il est des cas où le choix de  $g_n$  s'impose de lui-même. Il en est d'autres où il faut déterminer des critères de choix, d'autres encore où un modèle plus complet est indispensable. On illustrera cela par divers exemples.

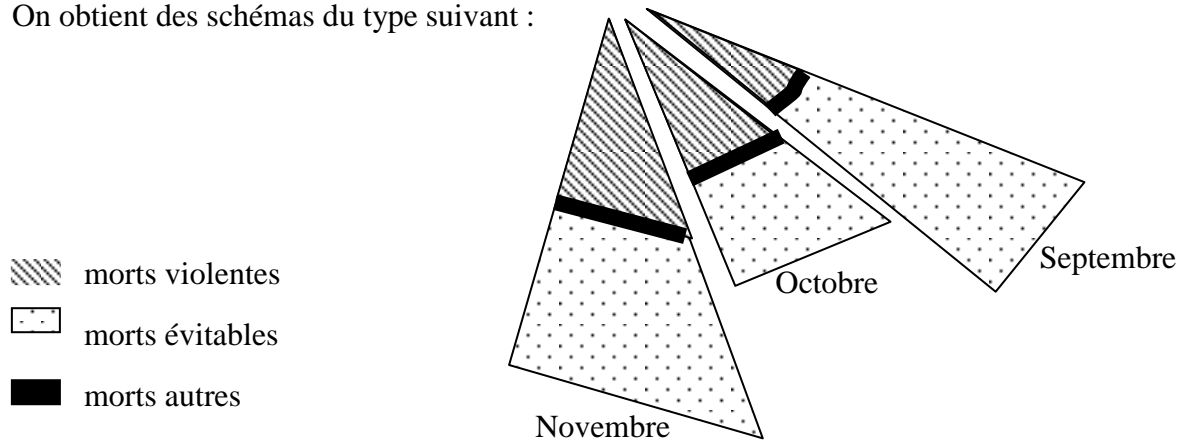
### 1 – La proportion

Reprenons l'étude de l'infirmière anglaise *Florence Nightingale*. Elle est une des premières à présenter sous forme de diagrammes ses résultats statistiques. Ces dessins sont évidemment des résumés, ils sont lisibles par tous, ils font partie maintenant de notre univers familier. Voici deux diagrammes publiés et qui ont fait scandale à l'époque.

#### **Les crêtes de Coq**

Pendant la guerre de Crimée, on dénombre mois par mois les décès dans l'armée anglaise en les classifiant en trois catégories : morts violentes, morts évitables, morts autres. Chaque mois est représenté par un secteur angulaire de  $30^\circ$  ( $360/12$ ). Dans un secteur donné une aire est attribuée à un type de décès, elle est proportionnelle au nombre de décès de ce type.

On obtient des schémas du type suivant :



### Les diagrammes en bâton

Après avoir découvert que l'état déplorable des hôpitaux de campagne faisait que les soldats avaient plus à craindre une maladie, bénigne au départ, que le combat, elle compare la mortalité des soldats et celle de la population civile en Angleterre et obtient le tableau suivant

Relative mortality of the Army at Home and of the English male population at corresponding ages		
Age	Deaths Annually to 1000 Living	Death
20-25	8.4	————— English men
	17.0	- - - - - English soldiers
25-30	9.2	—————
	18.3	- - - - -
30-35	10.2	—————
	18.4	- - - - -
35-40	11.6	—————
	19.3	- - - - -

L'énormité de ces résultats statistiques déclencha un vaste mouvement d'opinion. En effet les résumés statistiques s'imposaient d'eux-mêmes : il suffisait de compter et de calculer des proportions, puis de les représenter. De plus leur interprétation en terme de causalité est évidente : les normes d'hygiène et l'organisation du système de santé de l'armée anglaise sont déplorable. Une action est alors possible.

Même quand les résumés s'imposent à l'évidence et qu'il ne s'agit que de proportions, les interprétations ne sont pas toujours aussi simples et de nombreux pièges guettent l'utilisateur de statistiques. Le statisticien américain *Simpson* fournit un paradoxe simple.

Une université sélectionne ses étudiants et donne les proportions d'hommes et de femmes admis :

hommes 534 admis sur 1 198 candidats ..... soit 44,6% d'admis

femmes 113 admis sur 449 candidats ..... soit 25,2% d'admis

La sélection semble favoriser les hommes.

Cette université étant formée de deux départements *A* et *B*, faisons l'analyse par département.

département *A* :

hommes 512 admis sur 825 candidats..... soit 62,1% d'admis

femmes 89 admis sur 108 candidats ..... soit 82,4% d'admis

département *B* :

hommes 22 admis sur 373 candidats..... soit 5,9% d'admis

femmes 24 admis sur 341 candidats ..... soit 7,0% d'admis

Au niveau de chaque département, les femmes réussissent mieux que les hommes les examens d'entrée. Que s'est il passé ? On voit que 31,1% des hommes et 76,9% des femmes choisissent le département le plus sélectif. On a à faire à **deux populations distinctes** : celle du département *A* et celle du département *B*. Les réunir en une seule ne permet pas des interprétations pertinentes.

Un exemple analogue est fourni par une étude sur la réussite scolaire des élèves de nationalité étrangère. On note en 1983 la situation des élèves entrés au cours préparatoire en 1978. Ils peuvent être en 6<sup>ème</sup>, scolarité normale, en CM2 donc avoir une année de retard ou bien dans d'autres classes : CM1 deux ans de retard ou diverses classes pour élèves en grande difficulté. Le tableau suivant donne les résultats en % de la catégorie. Comme on sait depuis les années 50 que la variable catégorie socioprofessionnelle joue un rôle important, on a distingué les enfants d'ouvriers non qualifiés (OS), d'ouvriers qualifiés (OP), des catégories supérieures (autres CSP), car bien évidemment il y a une liaison entre la catégorie socioprofessionnelle (CSP) et la nationalité.

Situation en 1983 des élèves entrés au CP en 1978

	Elèves français			Elèves étrangers		
	6 <sup>ème</sup>	CM2	Autres	6 <sup>ème</sup>	CM2	Autres
Total	64	26	10	43	36	21
Enfants d'OS	47	35	18	39	39	22
Enfants d'OP	59	30	11	47	34	19
Autres CSP	70	22	8	51	30	19

A première vue, on voit que même à C.S.P. semblable, les élèves français réussissent mieux que les élèves étrangers. La première idée qui vient à l'esprit est d'expliquer cela par un déficit linguistique donc de proposer, en terme de remédiation, des cours de Français. Mais l'enquête s'est développée et on a eu l'idée de **faire intervenir une autre variable** : la taille de la fratrie. Si on regarde la réussite des enfants d'ouvriers non qualifiés, membres d'une famille de trois enfants ou plus, on obtient :

Elèves français			Elèves étrangers		
6 <sup>ème</sup>	CM2	Autres	6 <sup>ème</sup>	CM2	Autres
33	37	30	35	41	24

Les résultats s'inversent, les élèves étrangers réussissent mieux que les élèves français à CSP et à taille de fratrie égale (pour les autres C.S.P. que celle des ouvriers non qualifiés les résultats sont identiques). L'introduction de cette nouvelle variable, la taille de la fratrie, bouleverse les conclusions. En terme de remédiation, les études surveillées sont certainement préférables au cours de français, la taille de la fratrie intervenant sur la capacité à faire ses devoirs et à apprendre ses leçons.

Toujours dans le cas où les résumés sont des proportions, il faut se méfier de comparer ce qui n'est finalement pas comparable. Chargé par le ministre *Claude Allègre* d'un rapport sur l'organisation des études supérieures, *Jacques Attali*, par ailleurs polytechnicien, citant

une étude *Claude Thélot*, concluait à une dé-démocratisation du recrutement des élites. En effet depuis 1950 les jeunes sortis de trois grandes écoles, Polytechnique, Ecole Nationale d'Administration, Ecole Nationale Supérieure (rue d'Ulm), forment 1,2 ‰ d'une classe d'âge. On compare leur origine sociale en 1950 et en 1993. Deux milieux sont pris en compte : le milieu intellectuel (diplôme du père supérieur au baccalauréat), le milieu populaire. On obtient, en pourcentage de recrutement, les chiffres suivants

Elites	1950	1993
milieu populaire	25 %	9 %
milieu intellectuel	60 %	80 %

Le verdict paraît sans appel : la proportion d'enfants du peuple passe de 25 % à 9 %. Mais il faut comparer ce qui est comparable. Poursuivant son étude, *Claude Thélot* introduit la composition sociale française, qui s'est modifiée entre ces deux dates. Si on a l'évolution de celle-ci, on s'aperçoit que l'on comparait ce qui n'était pas comparable...

Voici l'évolution de la répartition, dans la population française, des deux milieux définis précédemment :

Population française	1950	1993
milieu populaire	80%	60%
milieu intellectuel	5%	20%

Si on appelle respectivement  $q_1$  et  $q'_1$  la proportion de jeunes du milieu populaire élèves de ces écoles en 1950 et 1993 et, de même,  $q_2$  et  $q'_2$  celle des jeunes du milieu intellectuel, un calcul simple montre qu'en 1950  $\frac{q_2}{q_1} = 40$  et qu'en 1993 les mêmes proportions donnent

$\frac{q'_2}{q'_1} = 23$ . On en conclut alors que l'écart entre les catégories sociales s'est amenuisé.

**Ces exemples montrent que, même quand les résumés statistiques sont évidents à choisir, leur interprétation en terme de caractéristique de la population peut être délicate.** Pour tirer des conclusions valides, il est nécessaire d'avoir une certaine prudence, marque distinctive du statisticien, et une connaissance autre de la population étudiée. On verra ultérieurement comment le modèle probabiliste peut la formuler.

## 2 – Premières qualités d'un résumé

Dans les exemples précédents, les résumés s'imposaient d'eux-mêmes, il s'agissait, pour des caractères qualitatifs, de calculer et de comparer des proportions. Il est bien évident que cela n'est pas toujours aussi simple.

En attendant des modèles plus précis, comme les modèles probabilistes qui guideront la recherche de résumés pertinents (comme dans le cas des cartouches ou celui de l'efficacité du médicament), on peut énoncer quelques principes qualitatifs valides y compris quand il s'agit d'un recensement (où  $E = P$  ).

**1<sup>er</sup> principe :** Si on dispose de plusieurs populations identiques ou analogues pour le phénomène étudié, on s'attend à ce que la **fluctuation des résumés**, caractérisant les diverses populations, soit inférieure à celle des mesures sur les individus d'une population donnée. Il s'agit d'un indicateur de bon sens. On caractérise des populations à partir d'individus différents les uns des autres. Il est illusoire d'obtenir des caractéristiques identiques pour des populations semblables mais on espère une certaine stabilité. Cela est évident si on se réfère à l'exemple du commerçant. Pour tenir boutique, il lui faut des

résumés relativement stables pour passer ses commandes, dès lors que le comportement des consommateurs ne change pas globalement, au delà des fluctuations journalières de ses ventes.

**2<sup>ème</sup> principe :** Le résumé doit être **peu sensible à la présence ou à l'absence d'un individu** dans la population. Cela semble évident. Si un individu donné joue un rôle trop important dans le résumé celui-ci ne peut caractériser la population. Dans la pratique, il n'est pas toujours facile de s'en rendre compte. Un exemple concret peut le montrer. A la fin du XVIII<sup>ème</sup> siècle, on voulait caractériser la capacité productive en "bled" (orthographe de blé à l'époque) d'un terrain, voire d'une région. La seule façon d'aborder la question est de mesurer la production faite année par année. Si on dispose de  $n$  années de production  $(x_1, x_2, \dots, x_n)$ , comment faire pour obtenir un nombre caractérisant la production au delà des fluctuations annuelles, liées à la météorologie en particulier ? On trouve dans le journal de Liouville, "*Annales de mathématiques pures et appliquées*", un article anonyme publié en 1821 et intitulé "*Dissertation sur la recherche du milieu le plus probable*", où il est dit : "*...il existe certaines provinces françaises où, pour déterminer le rendement moyen d'un terrain, on observe ce rendement durant vingt années consécutives, on enlève la plus petite et la plus grande valeur et on prend la dix-huitième partie de la somme des valeurs restantes*".

Pourquoi cette règle (à l'époque purement empirique) ? Pourquoi ne pas caractériser le terrain par la moyenne arithmétique des vingt observations ? C'est qu'en matière agricole, il arrive des années exceptionnellement bonnes ou exceptionnellement mauvaises. Si on représente sur un axe les différentes possibilités on peut avoir :

—xxxxx—x—                      ou                      —x—xxxxx—

or la moyenne arithmétique  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  est attirée par les valeurs extrêmes :  $\bar{x}$  tend vers

l'infini quand le plus grand des  $x_i$  tend vers l'infini, et, du coup, le résumé est très sensible aux valeurs extrêmes pour peu qu'elles s'écartent notablement du gros du peloton. Ceci justifie l'introduction de la **moyenne tronquée** dans les programmes de mathématiques de seconde en vigueur en 2000.

**3<sup>ème</sup> principe :** Le résumé doit être **opérationnel** pour le phénomène étudié.

Très souvent on mène une étude statistique pour prendre une **décision** : quelle commande faire ? Doit-on mettre le médicament sur le marché ? Quel type de notation recommander ? etc. C'est pour cela que l'ensemble dans lequel la caractéristique de la population prend ses valeurs a été appelé  $A$ ,  $A$  comme **action**. Il n'y a pas de bons résumés en soi, un résumé est **pertinent** suivant l'usage que l'on veut en faire.

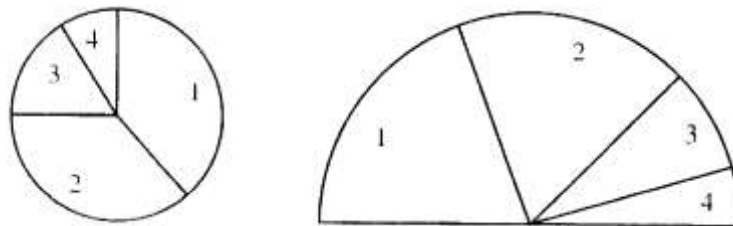
Prenons l'exemple de l'étude du nombre d'enfants par fratrie pour les élèves d'un lycée :

$P$  : les élèves du lycée ;  $x$  : la taille de la fratrie de chaque élève.

valeurs de $x$	1	2	3	4 et plus
effectifs	$n_1$	$n_2$	$n_3$	$n_4$

Présenter les résultats sous la forme d'un diagramme circulaire est un résumé statistique : c'est bien une caractéristique de la population des élèves du lycée. Mais est-ce une bonne représentation ?

Comparons la représentation par un diagramme circulaire et celle par un diagramme semi-circulaire.



La première ne permet pas de bien voir ce qui se passe car elle ne respecte pas la structure d'ordre sous-jacente, le secteur représentant les fratries, de quatre enfants et plus étant à côté de celui représentant les fratries d'un enfant. Par contre, le diagramme semi-circulaire respecte cette structure d'ordre. La remarque ne serait pas valide s'il s'agissait d'un caractère nominal comme la couleur des cheveux.

Soit, pour un second exemple, une étude sur les revenus des foyers français. On cherche un résumé donnant une valeur centrale pour ce caractère. Depuis le collège, on enseigne aux élèves que l'on peut prendre la **moyenne**, ou bien la **médiane**, ou bien encore le **mode**. Comme il y a beaucoup plus de petits revenus que de très gros, on sait que ces trois indicateurs sont différents. Lequel choisir ? Si vous faites de l'économie, votre indicateur doit être lié à la richesse globale du pays et la moyenne s'impose. Si vous faites de la sociologie, votre revenu moyen vous indiquera quelqu'un de relativement aisé, il vaut mieux prendre la médiane qui définit le foyer tel qu'il y en a autant de plus pauvres que de plus riches que lui. Si vous voulez mener une campagne de publicité pour un produit de consommation courante, alors mieux vaut prendre le mode.

**4<sup>ème</sup> principe** : Le résumé doit être **interprétable** dans l'étude faite.

Dans l'exemple de la capacité productive d'un terrain ou du revenu des foyers, les résumés exhibés sont interprétables aisément. Ils représentent une tendance centrale opérationnelle. Cela n'est pas toujours le cas. Prenons l'exemple des notes obtenus par les lycéens. On cherche des profils caractéristiques, des proximités entre résultats, toutes choses complexes, et il est difficile de définir a priori un résumé. On peut alors utiliser des techniques standards fournissant des résumés sophistiqués (cf. chapitre 4) et ensuite essayer d'interpréter ces derniers. Si une interprétation solide apparaît, alors le résumé est validé après coup.

### 3 – Description ou induction ?

Pour aller plus loin que les généralités précédentes et entrer dans quelques grandes techniques statistiques, il faut préciser les conditions d'obtention des données, faire des hypothèses sur la population et les formaliser en un modèle mathématique. En première analyse, on peut distinguer deux cas. Dans le premier, on observe tous les individus de la population :  $E = P$ . Les résumés statistiques que l'on met en œuvre servent à décrire  $P$ . On fait alors de la **statistique descriptive**. Les méthodes utilisées vont des plus simples, enseignées dès le collège, à de plus complexes, enseignées dans les cours spécialisés de l'enseignement supérieur sous le nom d'**analyse des données**.

#### 3.1 – De la description

La description commence quand on calcule des proportions représentées par des **diagrammes** divers : **circulaires**, quand il n'y a pas d'ordre dans les catégories nominales

étudiées (comme quand on étudie la couleur des cheveux par exemple), **semi-circulaire**, **en bâtons**, quand un ordre total existe (comme quand on étudie la taille des familles). Pour les variables quantitatives, la description commence quand on regroupe les données en **classes**, comme lors d'une étude sur la taille des conscrits d'une année donnée. Il s'agit d'un premier résumé. On avait  $N$  individus donc  $N$  tailles, on a  $K$  classes  $(1, 2, \dots, K)$  avec  $n_j$  individus dans la classe  $j$ . On résume l'information par les  $K$  entiers  $n_j$  soit  $K + (K+1)$  informations (les  $K+1$  bornes des intervalles) au lieu des  $N$  mesures, avec  $K$  très petit par rapport à  $N$ .

La description se poursuit quand on introduit **moyenne**, **médiane**, **mode** ainsi que les caractères de dispersion : **écart inter-quartile**, **inter-décile**, **écart type**, **étendue**, quand on dessine les **boîtes à moustaches**, tel que cela est prévu dans les programmes de lycée. Elle devient plus compliquée quand on étudie deux variables quantitatives à la fois et que l'on parle de **corrélations**, que l'on fait de la **régression**.

Les méthodes de description statistique qui viennent d'être évoquées permettent de calculer quelques résumés élémentaires, indices de centralité, de dispersion, dans le cas de variables quantitatives par exemple. Elles sont évidemment insuffisantes dès lors que l'on a affaire à des réalités plus complexes. Des méthodes plus sophistiquées ont été mises au point pour analyser ces réalités. Elles portent souvent le nom d'**analyse des données**. A titre d'exemple on évoquera deux ou trois grandes méthodes.

Supposons que, sur les  $n$  individus de la population, on mesure  $k$  caractères numériques. Faisons l'hypothèse que la **géométrie euclidienne** est adéquate pour la description de la population. On peut alors considérer les deux représentations suivantes. Dans l'espace euclidien  $E_k$ , le point  $M_i$  de coordonnées  $(x_{i1}, x_{i2}, \dots, x_{ik})$  représentera l'**individu**  $i$  ( $x_{ij}$  est la valeur de la variable  $j$  pour l'individu  $i$ ). De même, dans l'espace euclidien  $E_n$  le point  $P_j$  de coordonnées  $(x_{1j}, x_{2j}, \dots, x_{nj})$  représentera la **variable**  $j$ , par ses valeurs sur les  $n$  individus. On pourra dès lors utiliser les structures euclidiennes pour bâtir des résumés en faisant par exemple des projections, en calculant des distances, des angles. Cette statistique, que l'on peut appeler euclidienne, permet une interprétation des notions de moyenne, d'écart type, de coefficient de corrélation. Elle comprend des méthodes modernes d'analyse des données : l'**analyse en composantes principales**<sup>19</sup>, l'**analyse canonique** et son adaptation à des données qualitatives : l'**analyse factorielle des correspondances** où il s'agit de comparer deux caractères qualitatifs. Un exemple célèbre est fourni par l'étude des liens entre l'appartenance politique et le vocabulaire utilisé chez les parlementaires de la 3<sup>ème</sup> république.

Il est des cas où l'on soupçonne une population d'être composée de sous populations plus homogènes. Considérons par exemple une population de terroirs agricoles. On souhaite regrouper dans une même classe les terroirs qui se ressemblent. Pour cela on définit une **distance** entre les individus et on regroupe, selon une procédure rigoureuse, dans une même classe les individus les plus proches. De nombreux algorithmes de **classification automatique** existent. Le choix de l'un d'entre eux, de la finesse de la classification, dépend des données traitées et des objectifs poursuivis : détermination d'une réglementation agricole, typologie des clients d'une banque etc... Pour toute information sur ces méthodes on se reportera à la bibliographie.

### 3.2 – De l'induction

Le plus souvent il n'est pas possible d'examiner tous les éléments de la population. Celle-ci peut être infinie, comme pour l'étude de l'efficacité d'un médicament, où la population est l'ensemble des personnes susceptibles d'être traitées. Il peut être absurde d'examiner

<sup>19</sup> Voir le chapitre 4.



tous les individus quand il s'agit d'essais destructifs, pensons aux cartouches ! On observe alors  $n$  individus d'un échantillon et, à partir de ces observations, on veut caractériser la population entière. On fait alors une **inférence**. D'où le nom de **statistique inférentielle**, ou **statistique inductive**, donné aux procédures. On dira que les conclusions, que l'on induira de l'échantillon, seront valides si celui-ci est représentatif, au sens courant du terme. Mais comment s'en assurer ? Le seul moyen est de faire appel au **hasard**. Celui-ci est aveugle. Avec lui, pas de risque de choisir inconsciemment les individus de l'échantillon selon une variable liée à celle que l'on étudie. On élimine alors les **biais**, pour parler comme un statisticien.

Dans le cas des cartouches, le hasard doit être provoqué. Il ne faut pas choisir les cartouches à tirer selon votre fantaisie, vous risquez de les prendre là où les mauvaises sont plus nombreuses. Seule façon d'éviter cet inconvénient : numéroter vos cartouches et tirer comme à une loterie les numéros qui seront essayés.

Dans la pratique il existe évidemment des procédures plus opérationnelles que celle qui vient d'être imaginée, mais qui assurent que chaque cartouche a la même probabilité qu'une autre de figurer dans l'échantillon. Si de plus on conduit l'expérience de telle façon que le résultat d'un tirage n'a pas d'influence sur les suivantes, il est possible alors d'exprimer la loi de probabilité du nombre de mauvaises cartouches en fonction du paramètre  $\theta$ ,  $\theta$  étant la proportion de mauvaises cartouches dans la population totale. L'objectif de l'essai est de recueillir des informations sur ce  $\theta$  inconnu. **On a donc transformé la caractéristique de la population en un paramètre inconnu d'un loi de probabilité.**

Dans la plupart des cas traités par la statistique inductive, le hasard n'est pas seulement provoqué, il fait partie de la réalité ou tout au moins de la façon dont on se représente la réalité. Prenons l'exemple du médicament. L'expérience commune montre que la tension artérielle est un phénomène éminemment variable dans le temps, d'un individu à l'autre, et peu de régularités sont observables. Le **modèle** adopté, validé par de nombreuses études, nous fait représenter la tension artérielle comme une variable aléatoire suivant une loi partiellement ou totalement inconnue (cf. chapitre 5), selon les modèles choisis. Les observations faites sur les individus sont considérées comme autant de réalisations indépendantes de cette variable aléatoire. On caractérise la population étudiée par cette loi de probabilité. Le résumé statistique recherché consiste, à partir des observations faites, à dire des choses sur cette loi de probabilité, inconnue totalement ou partiellement, ou, plus fréquemment, à caractériser ce qui différencie les lois de probabilité de deux populations : celle qui est traitée par le médicament testé et celle qui n'a pas bénéficié du dit médicament.

Comme dans le cas précédent, la validité du modèle dépend des conditions dans lesquelles se déroulent les expériences. On observe des individus provenant de deux populations : celle qui est traitée et la population témoin. Il faut s'assurer que les tirages faits le sont dans les populations globales et non dans des sous-populations qui se distingueraient par un autre critère que le traitement. Si par exemple les expériences sont faites sur des cobayes de laboratoire, attribuer le traitement aux animaux les premiers attrapés dans la cage et considérer les suivants comme représentatifs de la population témoin introduit un biais. Les premiers attrapés sont probablement les moins vifs et donc représentatifs d'une sous-population plus malingre que l'autre. Là encore une seule solution pour garantir que l'hypothèse faite selon laquelle "on a des réalisations de variables aléatoires de même loi dans chaque population (traitée ou témoin), les deux populations ne différant que par le traitement", soit à peu près réalisée : tirer au hasard les animaux soumis au médicament et les autres. On verra aussi qu'il est important de garantir l'indépendance des variables aléatoires : une observation particulière ne doit pas influencer les autres. Quand on fait des

analyses de sang par exemple, il faut alors veiller à la propreté absolue du matériel, qu'il n'y ait pas de traces de sang de l'individu  $i$  quand on pratique l'analyse de celui de l'individu  $i+1$ .

En résumé, faire de la statistique inductive c'est caractériser la population étudiée par une **loi de probabilité**. Mais celle-ci n'étant pas connue complètement, on cherche sur elle des informations. Pour cela on tire au hasard des individus et on considère les observations faites comme des réalisations de **variables aléatoires indépendantes** et de même loi, au moins dans les **modèles** les plus simples. Conformément à la pratique statistique générale, c'est à partir de ces **observations** que l'on caractérisera la loi de probabilité inconnue. Au vu de cette caractérisation, on pourra prendre les **décisions** attendues.

Mais qu'est-ce qui justifie une telle pratique ? Raisonnons sur un cas simple. On produit en grande série une certaine sorte de pièces. Dans le mécanisme de production on fabrique une proportion  $\theta$  de pièces hors norme. On veut connaître  $\theta$ . On tire des échantillons de taille  $n$ . Soit  $k_j(n)$  le nombre de mauvaises pièces de l'échantillon  $j$ . L'observation

courante montre que la fréquence  $\frac{k_j(n)}{n}$  varie autour de  $\theta$ , quand on connaît  $\theta$ . On appelle

ce phénomène **fluctuation d'échantillonnage**. On le fait observer aux élèves des classes de seconde.

Une série d'expériences faites avec  $\theta = 0,1$  et  $n = 100$  donnerait : 0,12 ; 0,10 ; 0,12 ; 0,3 ; 0,10 ; 0,09 ; 0,09. Avec  $\theta = 0,1$  et  $n = 1000$  on aurait : 0,087 ; 0,092 ; 0,103 ; 0,114 ; 0,095 ; 0,096 ; 0,079 ; 0,093. Quand  $n$  vaut 1000 les fluctuations sont moins importantes que quand  $n$  vaut 100. La **loi des grands nombres**, démontrée par *Jacques Bernoulli* au

XVIII<sup>ème</sup> siècle, nous garantit qu'en probabilité  $\frac{k(n)}{n}$  tend vers  $\theta$  quand  $n$  tend vers l'infini.

Si la probabilité est une qualité objective du phénomène étudié, un nombre infini d'expériences permettrait de la connaître complètement. Malheureusement tout être humain ne peut faire qu'un nombre fini d'expériences. La connaissance sera donc approximative et de plus l'approximation ne pourra s'exprimer qu'en termes issus du calcul des probabilités. Avec en plus une difficulté supplémentaire le résultat expérimental "la pièce est bonne ou est mauvaise" n'a rien d'aléatoire, il est. Ce qui est aléatoire c'est l'acte qui a amené le résultat. **La probabilité n'est pas dans le constat, elle est dans la procédure**. Quand un dé honnête roule sur la table il y a une probabilité  $\frac{1}{6}$  que la face marquée 6 sorte, quand il est arrêté il y a un résultat observable mais plus de probabilité. Les remarques précédentes illustrent la difficulté qu'il y a à appréhender les techniques statistiques indispensables à la connaissance de la loi de probabilité sous-jacente à un phénomène, à juger de leur validité, à interpréter les conclusions qu'elles suggèrent. Pourtant elles sont incontournables, pensons à l'exemple des cartouches. On aboutit donc à une connaissance de nature différente de la connaissance déterministe qui nous est familière. La refuser revient à se priver de toute connaissance, ce qui est absurde. Une des façons de se familiariser avec ce mode d'approche de la réalité consiste à se donner un phénomène aléatoire que l'on connaît, à le **simuler** par des moyens informatiques, à exécuter les procédures statistiques, à comparer ce que l'on obtient aux caractéristiques, connues, du phénomène aléatoire de départ. La simulation se fait par l'utilisation des fonctions pseudo-aléatoires des logiciels classiques comme la touche RAND des calculatrices. C'est pourquoi la simulation est recommandée pour présenter les procédures statistiques des programmes des sections de techniciens supérieurs et fait partie, pour étudier les fluctuations d'échantillonnage, du programme de seconde.

## POUR ALLER PLUS LOIN

Dans une première définition, la statistique se caractérise comme l'**art de passer des observations sur des individus à des caractéristiques de la population**. Ces dernières sont choisies en fonction des objectifs poursuivis. Ainsi on résume un nombre parfois considérable de données en quelques indications synthétiques. Cette opération nécessite l'existence d'un **modèle**. En **statistique descriptive** on suppose, par exemple, qu'une **structure euclidienne** est adéquate pour décrire la réalité. En **statistique inductive** on considère les données comme des réalisations de **variables aléatoires**. La caractéristique de la population est alors une loi de probabilité incomplètement connue. **La statistique est alors l'art de mesurer les probabilités inconnues**.

On s'efforce donc d'extraire des données tout ce qui peut contribuer à la connaissance de la population, en surmontant la variabilité des observations sur les individus. Quand on augmente le nombre des expériences, on améliore mécaniquement la qualité de l'information. Mais il y a une limite à cette augmentation. L'expérimentation peut coûter cher ou prendre trop de temps. Comment faire les expériences pour optimiser la qualité de l'information recueillie ? Quand on a un modèle probabiliste précis, on peut répondre à ce type de questions. Les **plans d'expériences** formalisent la réponse. C'est une branche de la statistique qui intervient très en amont de l'étude à faire. On programme à l'avance les lieux où l'on doit observer. Voilà une technique qui se répand dans l'industrie. Elle vient de faire son apparition dans les programmes des sections de techniciens supérieurs de la branche chimie.

Il existe des cas où, sur la loi de probabilité inconnue, on dispose d'informations de type non déterministe. On les formalise sous forme d'une loi de probabilité sur les paramètres inconnus de la loi du phénomène. Celle-là est appelé connaissance a priori. On combine cette connaissance a priori avec le résultat des observations pour aboutir à une connaissance a posteriori. L'outil de cette transformation est le théorème de *Bayes* sur les probabilités conditionnelles, d'où l'adjectif de **bayésienne** associé à cette statistique. On peut aussi l'interpréter autrement : au lieu de mesurer un paramètre d'une loi de probabilité, vue comme une réalité **objective**, on modifie une opinion a priori pour tenir compte des faits et aboutir à une opinion a posteriori. C'est l'interprétation **subjective** de la probabilité.

La coupure entre statistique descriptive et statistique inductive n'est pas toujours aussi profonde que cela pourrait être déduit de l'exposé précédent. Soit par exemple à étudier, pour une série de villes de population  $y$ , le nombre des naissances  $x_1$  et le nombre des décès  $x_2$ . On peut écrire le modèle (ici supposé linéaire) :  $y = b + a_1 x_1 + a_2 x_2 + \varepsilon$ . On fait des observations dans  $n$  villes et par une technique de statistique euclidienne, on cherche  $b$ ,  $a_1$  et  $a_2$  qui sont tels que la somme des  $\varepsilon^2$  pour les  $n$  villes soit minimum (méthode des moindres carrés). Pour affiner les conclusions, on fait l'hypothèse que ces  $n$  villes forment un échantillon d'une population mythique de villes ayant ou pouvant exister. Cela permet de considérer  $\varepsilon$  comme une variable aléatoire et donc de quantifier des précisions sur les paramètres  $b$ ,  $a_1$  et  $a_2$  (par exemple par des intervalles de confiance).

Tout ce qui précède montre que **la statistique est un mode incontournable de connaissance** et ses applications sont de plus en plus nombreuses. Il faut dire en sus que le développement de l'**informatique** a, d'une part, permis un stockage commode d'un nombre important de données, voir les tickets de caisse du supermarché, et, d'autre part, de réaliser des calculs infaisables à la main ou très pénibles à exécuter. Pensons par exemple à la diagonalisation de matrices carrées symétriques de grande taille ou à la réalisation des algorithmes de classification. Le développement de logiciels conviviaux a beaucoup fait

pour la diffusion de la statistique. **Le risque est d'aboutir à des procédures presse-bouton** fournissant des résumés dont la justification risque d'échapper à l'opérateur si ce dernier n'a pas fait l'effort de comprendre la logique interne des procédures présentes dans le logiciel.

Un dernier mot, le développement des capacités de traitement d'une masse considérable de données a engendré de nouvelles procédures. On peut associer maintenant techniques statistiques et intelligence artificielle pour explorer ces données. Ainsi est né le "data-mining". Il a été appliqué à la masse des tickets de caisse du supermarché. Cela a permis de réorganiser ce dernier et d'avoir une augmentation importante du volume des ventes. Vous avez dit efficacité.

# 3

## GRANDE ET PETITE HISTOIRE DE LA STATISTIQUE

Ce chapitre présente quelques grandes lignes de l'histoire de la statistique, intimement liée à celle des probabilités et de la société. L'exposé linéaire qui en est fait pourra parfois sembler schématique, les lecteurs souhaitant approfondir se rapporteront à la bibliographie. Il est par ailleurs agrémenté de quelques anecdotes historiques, mettant en lumière certaines spécificités.

### I – LA « PREHISTOIRE »

Si le terme *statistique*, lié à la notion d'Etat, est relativement récent (issu du latin *statisticum*, il apparaît à la fin du XVII<sup>e</sup> siècle), l'activité de recueil de données est très ancienne et répond aux besoins d'organisation et de gouvernement des grands empires : dénombrements liés en particulier, à l'armée, aux impôts et à l'estimation des richesses. C'est un élément de contrôle comme le sont les levés topographiques ou cadastraux.

Les premiers recensements connus apparaissent sur les tablettes d'argile sumériennes, pour des listes d'hommes et de biens, 3000 ans avant notre ère. On rapporte qu'en 2238 av. J.-C., l'empereur chinois Yao organise le recensement des productions agricoles. Vers 2500 av. J.-C., en Egypte, une lourde administration gère l'impôt et le cadastre. Désignant une année par l'un de ses évènements remarquables, on peut lire dans certains textes égyptiens des expressions du type "*au commencement du temps du deuxième recensement du bétail*". Les Incas tenaient, eux, leurs statistiques agricoles à l'aide de cordes de couleur.

Les recensements romains sont connus, en particulier pour les circonstances de la naissance du Christ. Ciceron<sup>20</sup> insistait sur l'importance des statistiques (avant le mot) : "*Il est nécessaire au sénateur d'avoir une notion complète de l'Etat ; et cela s'étend loin : savoir l'effectif de l'armée, la puissance financière, les alliés, amis et tributaires que possède l'Etat ; [...] connaître les précédents traditionnels des décisions à prendre, l'exemple des ancêtres... Vous voyez enfin tout ce que cela comporte en général de savoir, d'application, de mémoire, et sur quoi un sénateur ne saurait en aucune manière se trouver pris au dépourvu.*" Le recensement romain permettait à la fois, de connaître les ressources en hommes mobilisables et en biens, et de classer les citoyens afin de répartir charges et avantages. Le recensement était également une démonstration de puissance, permettant de proclamer publiquement l'ampleur de la domination romaine. Selon Tacite, l'empereur Auguste aurait été le premier à faire un bilan des richesses de l'empire romain (soldats, navires, ressources privées et publiques). Au III<sup>e</sup> siècle apparaissent à Rome des tables d'estimation des rentes viagères.

A partir du XIII<sup>e</sup> siècle, les données deviennent plus nombreuses. Les commerçants de Venise amassent des données sur le commerce extérieur, évaluent les risques maritimes.

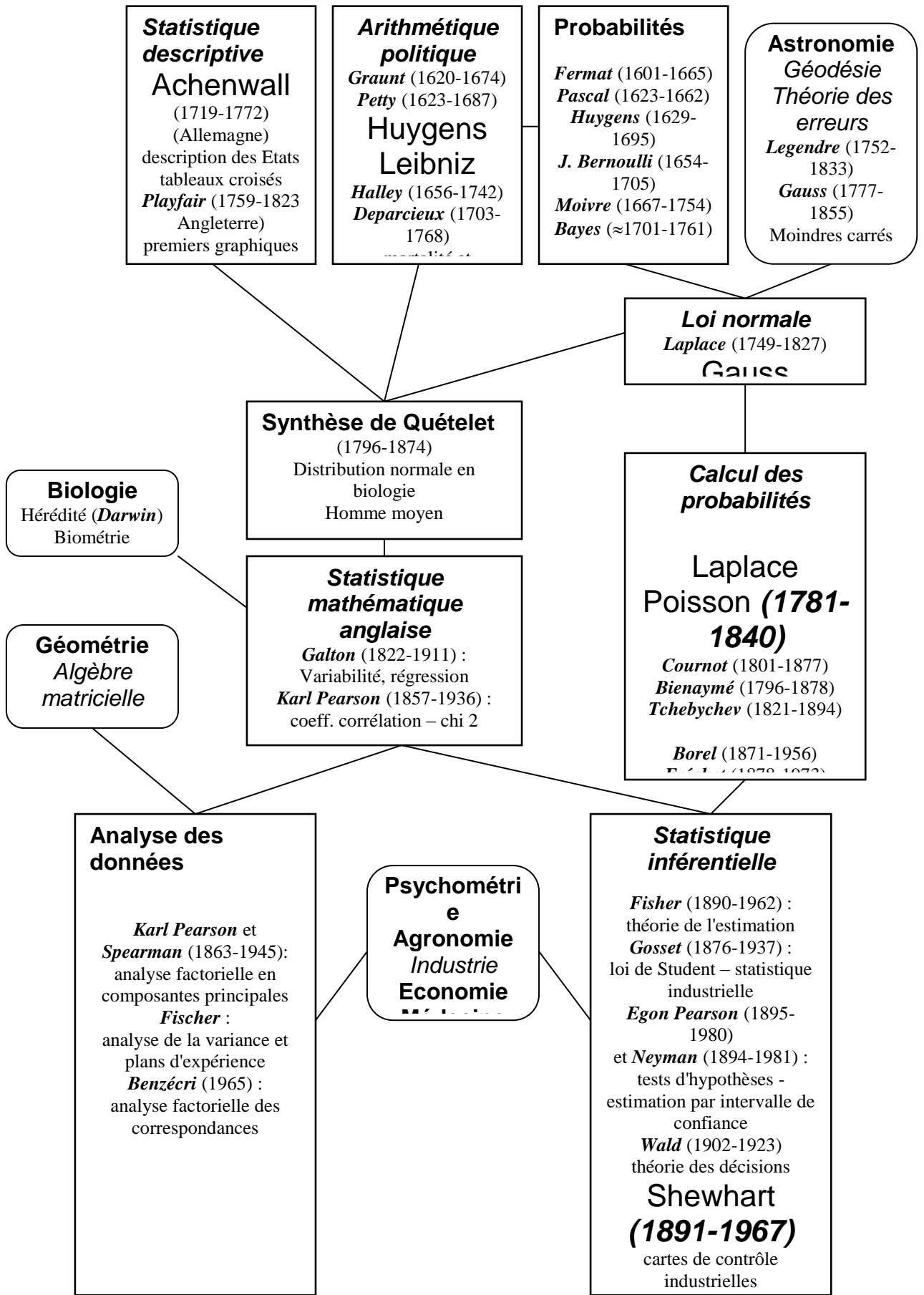
<sup>20</sup> Cité par Claude Nicolet dans "L'inventaire du Monde" – Fayard 1988.

En Hollande, on étudie les rentes viagères. Au XVI<sup>e</sup> siècle la tenue des registres des naissances est rendue obligatoire en France, par François I<sup>er</sup>, puis, sous Henri III, ceux des mariages et naissances.

Des progrès fondamentaux en statistique vont apparaître à la fin du XVII<sup>e</sup> siècle.

Une généalogie de la statistique du XVII<sup>e</sup> au XX<sup>e</sup> siècle

XVIII<sup>e</sup> XVII<sup>e</sup> siècle  
 XIX<sup>e</sup> siècle  
 XX<sup>e</sup> siècle



## II – LA STATISTIQUE AUX XVII<sup>e</sup> ET XVIII<sup>e</sup> SIECLES

### 1 – La statistique descriptive

Les progrès de la statistique descriptive sont en grande partie liés à ceux de la représentation graphique et de l'organisation des données en tableaux. Si les mathématiciens arabes du XII<sup>e</sup> siècle, en particulier les astronomes, utilisent des tables à double entrée et même des graphiques de type polygonaux, on doit attendre l'introduction des coordonnées par *Descartes* dans *La géométrie* en 1637.

La statistique descriptive du XVIII<sup>e</sup> siècle est dominée par *l'école allemande*, qui aurait légué le mot statistique (*Statistik* par *Achenwall* en 1746, encore que le terme, dérivé du latin, apparaisse plus tôt) ainsi qu'une tradition de **description** globale des Etats, plus qualitative et littéraire que quantitative. A la suite des travaux de *Conring* (1606-1681), vers 1660, visant à la classification de savoirs hétéroclites sur l'Etat et à la nomenclature (basée sur la logique d'*Aristote*), se développe à l'université de Göttingen une "école de statistique" menée par *Achenwall* (1719-1772) puis son successeur *Schlözer* (1735-1809). C'est dans le cadre d'une Allemagne divisée en une multitude de micro-Etats que se développe cette activité statistique de recueil et de classement de données, à laquelle on recourt pour tout type de litige ou conflit. Au début du XIX<sup>e</sup> siècle, on construira, à partir des tables de la statistique allemande, des **tableaux croisés** avec, en lignes, les pays et, en colonnes, les différents éléments littéraires de la description.

Vers 1760, *Jean-Henri Lambert* (1728-1777), mathématicien de langue allemande de Mulhouse (on lui doit la projection conique *Lambert* actuellement utilisée pour cartographier la France) développe des **représentations graphiques** remarquables.

En 1786, *William Playfair* (1759-1823) publie à Londres *"The Commercial and Political Atlas"* contenant le premier **diagramme en barres** connu, puis en 1801 *"The Statistical Breviary"* illustré de **diagrammes en secteurs**.

En France, la particularité est d'avoir, depuis 1660 environ, un pouvoir (royal) central fort, soutenu par une puissante administration. Pour *Colbert*, les intendants font parvenir des descriptions des provinces, de plus en plus codifiées (enquêtes sur les manufactures, le commerce, la population). Pour remplacer la taille par la dîme royale, *Vauban* rédige en 1686 une *"Méthode générale et facile pour faire le dénombrement des peuples"*.

De 1795 à 1806 sont organisées des enquêtes globales sur les nouveaux départements, dans le même esprit que la statistique allemande.

En 1800 *Bonaparte* institue les préfets et un "bureau de la statistique de la République".

#### Barème et Bottin

Avant de tomber dans l'anonymat des noms communs, Barème et Bottin sont les auteurs de tables de standardisation. *François –Bertrand Barrême* (≈1640-1703), arithméticien et poète, publia des tables mathématiques et de conversion. Il dédia plusieurs de ses ouvrages à son protecteur *Colbert*. *Sébastien Bottin* publie en 1799 un *"Annuaire politique et économique du Bas-Rhin"* que le ministre de l'intérieur salue comme *"le premier ouvrage vraiment statistique de cette nature que nous ayons en France"* (cité par Alain Desrosères). Il lance ensuite une entreprise éditant ses "bottins" (almanach).

### 2 – Le calcul des probabilités et l'arithmétique politique

C'est dans le cadre de l'Angleterre de la fin du XVII<sup>e</sup> siècle, où la société civile gagne en indépendance par rapport à l'Etat (ce qui permet de penser cette société civile en tant que



telle), que naissent des techniques liées à la collecte de données et surtout à leur traitement et à leur extrapolation, rassemblées sous la dénomination d'*arithmétique politique*, ancêtre de la statistique mathématique.

### Etude de la première table de mortalité

En **1662**, paraît dans les *Observations naturelles et politiques sur les bulletins de mortalité de la ville de Londres*, la première table de mortalité jamais construite (sans doute due à Petty). La voici reproduite ici :

"Puisque nous avons trouvé que sur 100 conceptions prises au départ, à peu près 36 n'atteignent pas l'âge de 6 ans, et que peut-être une seule survit à 76 ans, ayant sept décennies entre 6 et 76 ans, nous avons recherché six moyennes proportionnelles entre 64, ceux qui sont encore vivants à 6 ans, et l'unique survivant à 76 ans, et nous trouvons que les nombres suivants sont pratiquement assez près de la vérité ; car les hommes ne meurent pas selon des proportions exactes, ni selon des fractions : de là procède la table suivante, à savoir que sur 100 il en meurt

dans les six premières années	36
dans la décennie suivante	24
dans la seconde décennie	15
dans la troisième décennie	9
dans la quatrième	6
dans la suivante	4
dans la suivante	3
dans la suivante	2
dans la suivante	1

De là, il s'ensuit que sur 100 personnes conçues, il en reste,

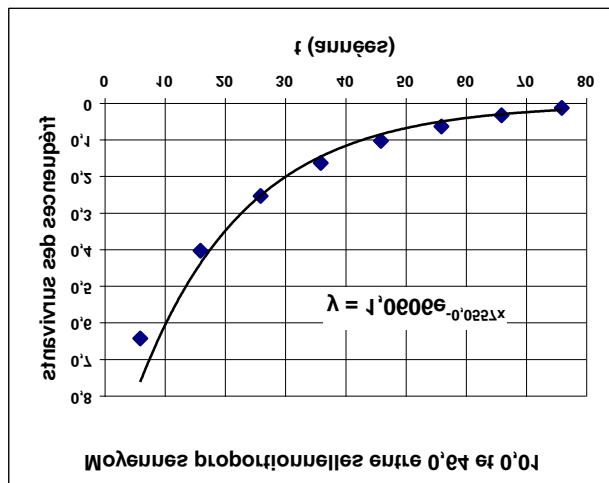
au bout de six années pleines	64
au bout de 16 ans	40
au bout de 26	25
au bout de 36	16
au bout de 46	10
au bout de 56	6
au bout de 66	3
au bout de 76	1
à 80	0

Considérons le second tableau, correspondant aux survivants. Il s'agit d'une "modélisation". A partir de l'observation qu'au bout de 6 ans, il y a 64 % de survivants et 1 % au bout de 76 ans, on a recherché une *série géométrique* pour les pourcentages de survivants à chaque décennie intermédiaire.

Les calculs ont été faits de manière approximative. Les logarithmes sont connus à l'époque mais leur usage se limite à l'astronomie.

Rechercher une raison constante  $q$ , avec  $0 < q < 1$ , comme coefficient multiplicateur du nombre de survivants, revient à considérer que le *taux de mortalité* entre 6 et 76 ans est le même (ici  $q \approx \frac{5}{8}$  correspond à 3 décès pour 8 personnes sur 10 ans).

(ajustement exponentiel sur *Excel*.)



Les calculs de cette "arithmétique" s'appuient sur la théorie naissante des probabilités et en alimentent également la réflexion (*Huygens* et l'espérance de vie, *Jacques Bernoulli* et sa loi des grands nombres).

A la suite des travaux fondateurs de *Graunt* (1620-1674) sur les bulletins de décès et les naissances (il découvre ainsi la proportion plus grande de naissances masculines : 107 pour 100 naissances féminines), l'économiste *William Petty* (1623-1687) systématise et théorise les études démographiques sur les naissances, décès (voir encadré<sup>21</sup>), nombres de personnes par famille...

Article "*Arithmétique politique*" de l'*Encyclopédie méthodique* (1784)

\* **ARITHMÉTIQUE politique** ; c'est celle dont les opérations ont pour but des recherches utiles à l'art de gouverner les peuples, telles que celles du nombre des hommes qui habitent un pays ; de la quantité de nourriture qu'ils doivent consommer ; du travail qu'ils peuvent faire ; du tems qu'ils ont à vivre ; de la fertilité des terres ; de la fréquence des naufrages, &c. On conçoit aisément que ces découvertes & beaucoup d'autres de la même nature, étant acquises par des calculs fondés sur quelques expériences bien constatées, un ministre habile en tireroit une foule de conséquences pour la perfection de l'agriculture, pour le commerce tant intérieur qu'extérieur, pour les colonies, pour le cours & l'emploi de l'argent, &c. Mais souvent les ministres (je n'ai garde de parler sans exception) croient n'avoir pas besoin de passer par des combinaisons & des suites d'opérations arithmétiques : plusieurs s'imaginent être doués d'un grand génie naturel qui les dispense d'une marche si lente & si pénible, sans compter que la nature des affaires ne permet ni ne demande presque jamais la précision géométrique. Cependant si la nature des affaires la demandoit & la permettoit, je ne doute point qu'on ne parvint à se convaincre que le monde politique, aussi bien que le monde physique, peut se régler à beaucoup d'égards par poids, nombre & mesure.

Il s'agit de travaux "d'experts", liés à des préoccupations pratiques. *Graunt* a été commerçant, *Petty*, fut successivement, médecin, mathématicien, parlementaire, fonctionnaire et homme d'affaire.

En 1696, l'astronome anglais *Edmond Halley* (1662 – 1742), en se basant sur cinq ans d'état civil de la ville de Breslau (Pologne), établit une **table de mortalité**, préfigurant les travaux d'actuariat.

En Hollande, le calcul des probabilités est appliqué à l'espérance de vie humaine (*Christian et Louis Huygens* en 1669) et à l'estimation du prix d'achat d'une rente, à l'aide de tables de mortalité (*Jan De Witt* en 1671).

Un argument théorique de poids, en faveur des techniques de l'arithmétique politique, est apporté par la **loi des grands nombres**

dont *Jacques Bernoulli* (1654-1705) est à l'origine ("*L'Art de conjecturer*" 1713).

*Condorcet* rapporte, dans l'*Encyclopédie* (édition "*méthodique*" de 1784), que ce théorème "aussi difficile que la quadrature du cercle [...] fait voir que la probabilité (au sens "raison de croire") qui naissait de l'expérience répétée, allait toujours en croissant, et croissait tellement, qu'elle s'approchait indéfiniment de la certitude.[...] Par là il est démontré que l'expérience du passé est un principe de probabilité pour l'avenir. [...] Ce principe reçu, on sent de quelle utilité seraient dans les questions de physique, de politique, et dans ce qui regarde la vie commune, des tables exactes qui fixeraient sur une longue suite d'évènements la proportion de ceux qui arrivent d'une

*Jacob Bernoulli*

<sup>21</sup> Source : "*Les mathématiques sociales*" - Dossier *Pour la Science* de juillet 1999 - article de Hervé Le Bras.

certaine façon à ceux qui arrivent autrement. Les usages qu'on a tiré des registres baptistaires et mortuaires sont si grands, que cela devrait engager non seulement à les perfectionner, en marquant, par exemple, l'âge, la condition, le tempérament, le genre de mort, etc. mais aussi à en faire de plusieurs autres évènements, que l'on dit très mal à propos être l'effet du hasard, c'est ainsi que l'on pourrait former des tables qui marqueraient combien d'incendies arrivent dans un certain temps, combien de maladies épidémiques se sont fait sentir en certains espaces de temps, combien de navires, etc. ce qui deviendrait très commode pour résoudre une infinité de questions utiles, et donnerait aux jeunes gens attentifs toute l'expérience des vieillards."

En France, **Antoine Deparcieux**, de la Société Royale des Sciences de Montpellier, à la demande du gouvernement, s'occupe en 1746 à rectifier la table établie par *Halley*. *Deparcieux* se base, pour ce faire, sur l'examen des registres d'une tontine et les registres de décès des couvents parisiens sur une longue période (1607 – 1745) et fait paraître "*l'essai sur les probabilités de la durée de la vie humaine*". Sa table fut longtemps en usage auprès des assureurs français, pour le calcul de leurs rentes viagères. En 1806, **Duvillard**, alors directeur des services de la statistique publie également une table de mortalité utilisée longtemps, elle, pour le calcul des primes d'assurance en cas de décès.

Parmi les techniques de l'arithmétique politique, apparaît celle de "**l'estimation**" par **coefficient multiplicateur**, ancêtre des sondages (voir le texte extrait des "*Récréations mathématiques*" d'*Ozanam* - édition de 1778) :

Ayant observé sur quelques paroisses que le nombre  $x$  d'habitants était proportionnel au nombre  $n$  de naissances annuelles, soit  $x = kn$ , avec un facteur  $k$  à peu près constant, on estimait la population d'une région en multipliant par ce facteur  $k$  le nombre des naissances dans l'année, recueilli sur les registres.

Une des raisons obligeant les anglais à recourir à ces méthodes détournées d'estimation est leur conception libérale de l'Etat, empêchant celui-ci d'organiser des enquêtes systématiques, comme en France par exemple.

En 1783, **Laplace**, en formulant des hypothèses quant à la distribution de probabilité du coefficient multiplicateur, calcule "*l'erreur à craindre*" pour une population ainsi estimée, à partir d'enquêtes sur des régions tirées au hasard.

#### 254 RÉCRÉATIONS MATHÉMATIQUES.

##### §. V.

*Sur le rapport des naissances & des morts au nombre total des habitants d'un pays : Conséquences de ces observations.*

Comme il seroit bien difficile de faire l'énumération des habitants d'un pays, sur-tout s'il falloit la réitérer autant de fois que des intérêts politiques peuvent exiger qu'on connoisse sa population, on a tâché d'y suppléer, en déterminant le rapport des naissances ou des morts avec le nombre total des habitants de ce pays : car, comme dans tous les pays de l'Europe civilisés on tient des registres des naissances & des morts, on peut, en les compulsant, juger de la population, voir si elle augmente ou diminue, & examiner, dans le dernier cas, les causes qui produisent cette diminution.

On déduit, par exemple, des tables de *M. Halley*, qui présentent l'état de la population de *Breslaw* vers l'année 1690, que sur 34000 habitants il y arrivoit annuellement, calcul moyen, 1238 naissances ; ce qui donne le rapport des premiers aux secondes, de  $27 \frac{1}{2}$  à 1. Pour des villes telles que *Breslaw*, où il n'y a pas un grand abord d'étrangers, on peut donc prendre pour règle, de multiplier les naissances par  $27 \frac{1}{2}$ , & l'on aura le nombre des habitants.

### Le débat sur l'inoculation de la variole

Ce débat, sur une question éminemment sensible et pratique, mettant en jeu la notion d'espérance mathématique, illustre l'écart qui peut exister entre les résultats statistiques, fondés sur une approche globale, et le sentiment individuel, face à une décision personnelle. Alors qu'au milieu du XVIII<sup>e</sup> siècle la petite vérole (ou variole) est responsable de 50 à 80000 morts par an, les médecins observent une certaine résistance à la réinfection. On tente alors l'inoculation de la variole. Suite à la mort de Louis XV causée par la variole le 10 mai 1774, la famille royale sera inoculée un mois après. Les futurs Louis XVIII et Charles X se feront également inoculer. Mais le risque est grand, puisque statistiquement, on constate qu'une personne sur 300 meurt dans l'année, de ses suites. La question se pose sur la pertinence de sa généralisation et suscite un grand débat d'opinion.

L'élite "éclairée", *Voltaire* en tête, milite pour l'inoculation. Du côté des "géomètres", *Daniel Bernoulli* calcule que l'espérance de vie d'une personne inoculée augmente de plus de trois années. Ce résultat, de type fréquentiste (loi des grands nombres), s'il peut justifier une politique sanitaire, est cependant insuffisamment important pour une décision individuelle sans arrières pensées.

Vers 1760, l'opinion de *d'Alembert* est plus nuancée. Il note<sup>22</sup> que les mathématiques s'appliquent mal dans cette situation, relevant du "cas de conscience" : *"Si les avantages de l'inoculation ne sont pas de nature à être appréciés mathématiquement, il est néanmoins vraisemblable que ces avantages sont réels pour ceux qui la subiront avec les précautions convenables [...]. [Ces] objections n'attaquent que les mathématiciens qui pourraient trop se presser de réduire cette matière en équations et en formules. [...] On a trop souvent confondu l'intérêt de l'Etat en général pour avoir l'inoculation, avec celui que les particuliers peuvent y trouver car ces deux intérêts peuvent être fort différent."*

Condorcet exprime une opinion analogue vers 1772 : *"L'on voit que ces déterminations de la vie moyenne peuvent servir avantageusement pour les Etats mais sont presque inutiles pour chaque homme."*

Après les observations du médecin britannique *Edward Jenner* en 1796, on immunise par injection à l'homme d'une maladie voisine (la variole de la vache), procédé moins risqué qui préfigure la vaccination.

## III – L'EMERGENCE DE LA STATISTIQUE MATHÉMATIQUE AU XIX<sup>e</sup> SIÈCLE

### 1 – La loi « normale »

L'établissement, au début du XIX<sup>e</sup> siècle, de la loi "normale", dont l'usage est fondamental en statistique, s'est fait par deux voies : celle, dans le cadre de la "théorie des erreurs", de la méthode des moindres carrés, qui aboutit avec *Carl Friedrich Gauss* (1777-1855), et celle des théorèmes limites, avec l'énoncé d'une première version du théorème limite central par *Pierre*



<sup>22</sup> Cité par Eric Brian dans *"La mesure de l'Etat"*.

**Simon de Laplace** (1749-1827).

L'astronomie et la géodésie sont à l'origine des questions théoriques sur la répartition des erreurs de mesure. Il ne s'agit pas des erreurs "systématiques", dues par exemple à un défaut de l'instrument, que l'on peut évaluer et corriger facilement, mais des erreurs "accidentelles" (que l'on peut qualifier d'aléatoires), dues à l'addition de nombreux facteurs indépendants (conditions de la mesure, erreurs de lecture, de visée...), qui peuvent induire une erreur dans un sens ou dans l'autre. L'objectif est de pouvoir aller au delà de la précision de l'instrument, en "combinant" plusieurs mesures de la même quantité, de façon à calculer la "meilleure estimation" de cette dernière. Ces questions se posent en astronomie lors des calculs des trajectoires planétaires, en particulier pour détecter d'éventuelles planètes "perturbatrices" inconnues, et, en géodésie, dans les opérations de triangulation visant à la mesure d'un méridien, d'abord pour la détermination de la forme de la Terre (aplatissement aux pôles), puis pour la détermination de la valeur du mètre.

L'astronome allemand **Tobias Mayer** en 1750, pour des calculs concernant l'observation d'un cratère de Lune, donne une méthode d'ajustement par regroupement des données (**méthode de Mayer**).

**Adrien-Marie Legendre** (1752-1833) publie en 1805 la méthode consistant à minimiser la somme des carrés des écarts, correspondant à l'ajustement optimal pour la structure géométrique euclidienne. Indépendamment, **Gauss**, alors directeur de l'observatoire de Göttingen, parvient, dans le cadre de l'étude des orbites planétaires, à cette même **méthode des moindres carrés**, dit-il dès 1794 (il en conteste la paternité à *Legendre*, mais ne publiera qu'en 1809). L'originalité de *Gauss* est d'établir les liens qui existent entre cette méthode et les lois de probabilité, aboutissant à la "loi gaussienne" :

Soit une quantité  $\theta$  inconnue, pour laquelle on possède plusieurs mesures  $x_1, x_2, \dots, x_n$ . On constate tout d'abord que l'estimation  $\hat{\theta}$  de  $\theta$  rendant minimale la somme des carrés des erreurs correspond à la moyenne :  $\hat{\theta} = \bar{x} = \frac{x_1 + \dots + x_n}{n}$ .

En effet, ce minimum sera obtenu en annulant la dérivée par rapport à  $\hat{\theta}$  de la somme des carrés des écarts, soit

$$\frac{d}{d\hat{\theta}} \sum (\hat{\theta} - x_i)^2 = 2 \sum (\hat{\theta} - x_i) = 0 \text{ qui donne } \hat{\theta} = \frac{x_1 + \dots + x_n}{n}.$$

En envisageant la question d'un point de vue probabiliste, on considérera que les erreurs  $e_1 = x_1 - \theta, \dots, e_n = x_n - \theta$  sont des réalisations de  $n$  variables aléatoires indépendantes  $E_1, \dots, E_n$  de même loi continue de densité  $f$ , dépendant de la valeur inconnue  $\theta$ .

Pour  $\theta$  donné, la probabilité d'effectuer des erreurs entre  $e_1 + de_1, \dots, e_n + de_n$  est alors, en vertu de l'indépendance,  $f(e_1) \times \dots \times f(e_n) de_1 \dots de_n$ .

On peut alors retourner le raisonnement (à la façon de *Bayes*) et se demander, les mesures  $x_1, \dots, x_n$  étant connues, quelle est la valeur de  $\theta$  la plus vraisemblable. C'est à dire, quelle est la valeur de  $\theta$  qui rendra maximale la probabilité d'observation des mesures  $x_1, \dots, x_n$  (réellement observées) donc des erreurs  $e_1, \dots, e_n$ . Il s'agit de rechercher  $\theta$ , donc  $f$ , de sorte que  $f(x_1 - \theta) \times \dots \times f(x_n - \theta)$  soit maximum ("**maximum de vraisemblance**").

Le produit  $\prod f(x_i - \theta)$  est maximum lorsque la somme  $\sum \ln(f(x_i - \theta))$  est maximale.

En dérivant par rapport à  $\theta$ , on obtient la condition  $\sum \frac{d \ln f(x_i - \theta)}{d\theta} = 0$ .

Sachant que la moyenne arithmétique  $\hat{\theta} = \frac{x_1 + \dots + x_n}{n}$  correspond à la valeur recherchée de  $\theta$  et que cette moyenne vérifie l'équation en  $\theta$ :  $\sum (x_i - \theta) = 0$ , Gauss en déduit que, pour  $i$  allant de 1 à  $n$ , on a  $\frac{d \ln f(x_i - \theta)}{d\theta} = k(x_i - \theta)$ .

On a enfin, en intégrant,  $\ln f(x_i - \theta) = -k \frac{(x_i - \theta)^2}{2} + \text{cte}$  soit  $f(x_i - \theta) = C e^{-\frac{k}{2}(x_i - \theta)^2}$ , où

l'on retrouvera l'expression de la densité de la **loi normale**.

Rétrospectivement, on constate que si  $f(e_i) =$

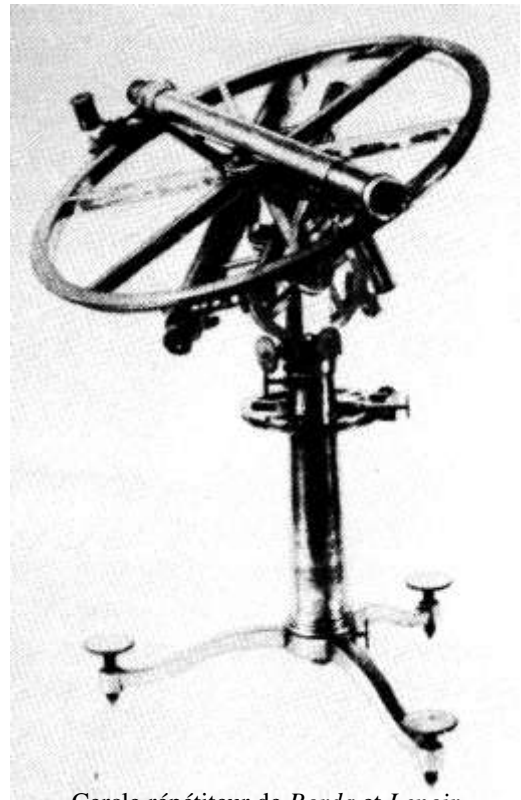
$C e^{-\frac{k}{2}e_i^2}$ , alors  $\prod f(e_i) = C e^{-\frac{k}{2}\sum e_i^2}$  est maximum lorsque la somme des carrés des écarts  $\sum e_i^2$  est minimale.

Ainsi lorsque, lors de mesures, l'addition de plusieurs facteurs aléatoires indépendants (et sensiblement équivalents) induit des erreurs, celles-ci se répartissent selon la loi de Gauss et la moyenne arithmétique des mesures fournit l'estimation qui minimise la somme des carrés des erreurs.

C'est ainsi par exemple que le "cercle répétiteur" conçu par le "mathématicien marin" Borda et réalisé par "l'artiste mécanicien" Lenoir en 1787 permit un gain important de précision dans la mesure des angles et put être utilisé par Delambre et Méchain dans la mesure du méridien pour déterminer la valeur du mètre. L'appareil, muni de deux lunettes de visée, permettait, en débrayant une des lunettes du cercle, de cumuler  $n$  mesures de l'angle. En divisant la somme par  $n$ , on obtenait ainsi une mesure moyenne, pour laquelle la dispersion (l'écart type) est divisée par  $\sqrt{n}$  (voir théorème limite central et échantillonnage).

L'approche de **Laplace** se situe dans la voie des **lois limites**, ouverte par Jacques Bernoulli et la loi des grands nombres.

On sait que la somme de  $n$  variables aléatoires de Bernoulli, valant 1 avec la probabilité  $p$  et 0 sinon, suit la loi binomiale de paramètres  $n$  et  $p$ . En utilisant la formule de Stirling pour approcher la factorielle, Abraham de Moivre (1667 – 1754), dans la "doctrine des chances", publiée en 1718, montre l'approximation de la distribution binomiale, pour  $n$  grand, dans le cas  $p = 1/2$ , par la loi normale. Laplace généralise en 1812 ce résultat au cas  $p$  quelconque et montre, en 1810, que sa "seconde loi des erreurs"<sup>23</sup> approche la distribution des moyennes arithmétiques de  $n$  erreurs indépendantes de même loi.



Cercle répétiteur de Borda et Lenoir

<sup>23</sup> Laplace avait introduit en 1774 une "première loi des erreurs", de densité  $f(x) = (k/2)e^{-k|x|}$  avec  $x \in \mathbf{R}$ , en considérant les écarts absolus des mesures par rapport à la médiane.



Pierre Simon de Laplace

statistique du XIX<sup>ème</sup> siècle, à partir de *Quételet* et de son "homme moyen".

La dénomination de "*loi normale*" est quant à elle utilisée par *Pearson* en 1893. Quant au nom de "*théorème limite central*", il a été proposé par *Polya* en 1920 qui parle de "*central limit theorem of probability theory*".

*Laplace* et *Gauss* réalisent ainsi, au début du XIX<sup>ème</sup> siècle, une synthèse entre l'approche empirique des moindres carrés et celle, probabiliste, des lois limites. Avec *Laplace*, la loi normale s'impose comme presque universelle. En effet, même si la distribution individuelle des erreurs ne suit pas une loi normale, celle des moyennes des erreurs suit approximativement, sous certaines conditions (indépendance, lois identiques), une loi normale. C'est sur ce résultat que va s'appuyer toute la

## 2 – La synthèse de Quételet

La figure phare de la statistique du XIX<sup>e</sup> siècle est celle de l'astronome belge *Adolphe Quételet* (1796 – 1874). C'est lui qui fit la synthèse entre la tradition de la statistique descriptive "à l'allemande", littéraire et sociale, et celle, fondée sur le calcul des probabilités, de "l'arithmétique politique", "à l'anglaise". Cette synthèse se base sur la loi normale, dont le théorème limite central explique l'omniprésence, y compris dans le domaine humain. Les **régularités statistiques** qui émergent, en moyenne, des grands nombres sont spectaculaires, parfois effrayantes lorsqu'il s'agit par exemple du nombre des crimes et que celui-ci apparaît comme implacablement déterminé. *Quételet* affirme<sup>24</sup> : "*En se plaçant dans des conditions favorables pour bien observer, on trouve que, chez les êtres organisés, tous les éléments sont sujet à varier autour d'un état moyen, et que les variations qui naissent sous l'influence des causes accidentelles, sont réglées avec tant d'harmonie et de précision, qu'on peut les classer d'avance numériquement et par ordre de grandeur, dans les limites entre lesquelles elle s'accomplissent. Tout est prévu, tout est réglé : notre ignorance seule nous porte à croire que tout est abandonné au caprice du hasard*". Il défendra une statistique scientifique, une "*Physique sociale*" (c'est le titre d'un de ses ouvrages, publié en 1835).

Par son passage en 1823 à l'Observatoire de Paris, il entre en contact avec les mathématiciens français *Fourier*, *Laplace* et *Poisson*. Mais, en ce début du XIX<sup>e</sup> siècle apparaît un être nouveau : la "société", étudiée en tant que telle, de l'extérieur. Des probabilistes français, *Quételet* conservera les résultats concernant la loi des grands nombres et la convergence de la loi binomiale vers la loi normale. Appliqués au domaine social, ces résultats le conduiront à la notion "**d'homme moyen**" dont toutes les caractéristiques seraient les moyennes de celles des hommes réels, ayant des mensurations moyennes, mais aussi des désirs moyens, une propension au crime moyen... Les humains réels se distribueraient autour de cet "homme moyen" abstrait comme les erreurs dans la

<sup>24</sup> Cité par Pierre Crépel dans "*Les mathématiques sociales*" – Dossier "*Pour la Science*" Juillet 1999.

mesure d'une grandeur physique. La méthode, très critiquée à l'époque (cet homme moyen est-t-il un homme ? Il est, selon *Cournot*, une monstruosité), montre cependant l'intérêt d'une démarche statistique faisant émerger une tendance globale des aléas individuels (voir le débat à propos du choléra). *Quételet* fut très célèbre à son époque, sachant constituer de vastes réseaux, organisant les premiers congrès internationaux de statistique (le premier se tient à Bruxelles en 1853). Dans le domaine des enquêtes, il préférera la rigueur des **recensements** (organisant plusieurs recensements de population, rendus possibles par les structures administratives) aux incertitudes des sondages. La voie ouverte, à la fin du XVIII<sup>e</sup> par les techniques du coefficient multiplicateur, ou de la probabilité des causes est alors, jusqu'au XX<sup>e</sup> siècle, négligée.

### La controverse lors de l'épidémie de choléra de 1832

Cet épisode est caractéristique de l'effacement, en statistique, du cas individuel, derrière les propriétés globales de la population, ce qui, parfois, peut conduire à controverses. C'est en particulier le cas en matière médicale, où les méthodes statistiques ont rencontré une certaine résistance. Même *Claude Bernard* estimait qu'on ne peut pas soigner "en moyenne".

En 1832, une épidémie de choléra, en provenance d'Asie, se répand en France. Les médecins sont alors partagés en deux camps : les *contagionnistes*, qui pensent que le choléra est transmis par contact avec un malade, et les *infectionnistes*, qui estiment que sa propagation est due à l'insalubrité et au manque d'hygiène et sont les tenants de la théorie des "miasmes". Ces derniers s'appuient sur l'étude statistique de la mortalité à Paris, rue par rue, et selon les caractéristiques du "milieu social".

L'histoire donnera, d'une certaine façon, raison aux deux camps, montrant la complémentarité des deux approches. L'étude individuelle de la maladie conduira en 1883 à la découverte du vibron du choléra, dont la diffusion est favorisée par l'absence d'égoûts, ce que l'étude statistique a pu laisser supposer.

### 3 – La statistique mathématique anglaise

De nombreux outils de la statistique mathématique apparaissent en Angleterre, dans un contexte biologique, hérité de *Darwin*. Il s'agit, d'une part de la politique eugéniste, visant à l'amélioration de l'espèce humaine, dont on connaît les abominables dérives, et, d'autre part, de son aspect scientifique, la biométrie. De ces préoccupations naîtront, entre autres, la régression linéaire, la corrélation et le test du chi-deux.

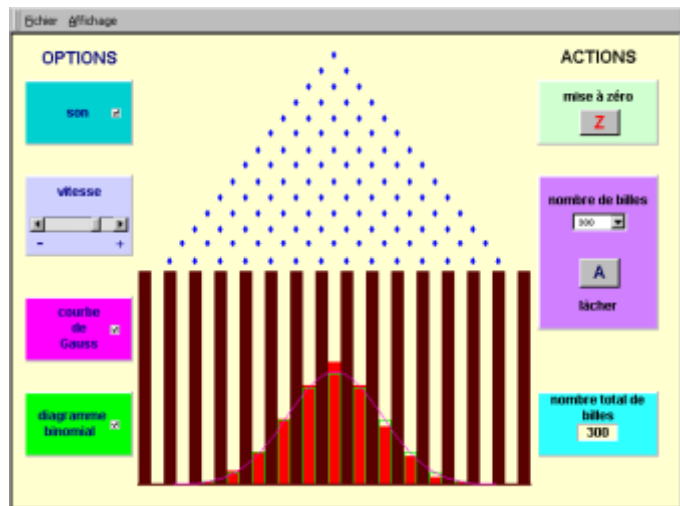
La première figure à évoquer est celle de **Francis Galton** (1822 – 1911), cousin de *Charles Darwin*. C'est un savant polyvalent, géographe et biologiste, qui s'intéresse à des questions statistiques dans le cadre de la génétique, de l'hérédité biologique et du comportement humain. Plutôt que sur les qualités moyennes, comme *Quételet*, *Galton* fera davantage porter son attention sur la **variabilité**, les différences entre les individus, dans l'espoir de conserver, ou favoriser, les meilleurs. Son apport essentiel aux techniques statistiques est celui de la **corrélation** (1880) et de sa mesure par un "indice de corrélation" (1888).

En 1877, en étudiant la taille des enfants par rapport à celle de leurs parents, il constate, qu'en moyenne, il y a une "**régression**". C'est à dire que, dans les familles de grande taille, la taille des enfants est, en moyenne, inférieure à celle des parents, alors que dans les familles de petites tailles, celle des enfants est, en moyenne, supérieure. La taille des enfants est bien "corrélée" à celle des parents, mais il y a une "régression" vers une taille plus "moyenne". La régression vers la moyenne étant inversement proportionnelle à la corrélation.



Dans le cadre de l'eugénisme, *Galton* examina la dispersion des résultats, en développant les notions de **médiane** et de **quartiles**. Ces outils statistiques visent à quantifier les effets de l'hérédité. Il est difficile, aux vues des exactions commises par la suite dans l'Allemagne nazie, d'imaginer comment, à l'époque, les idées eugénistes étaient perçues. Les ouvrages de *Galton* furent vus comme progressistes et participant au combat de la science contre l'obscurantisme de la religion traditionnelle.

Peu versé dans les mathématiques, *Galton* eut le souci de représenter physiquement les phénomènes aléatoires, en faisant construire des machines, comme la fameuse "planche de Galton" (la "quinconce"), mettant en évidence la convergence de la loi binomiale vers la loi normale : des billes sont lâchées du haut d'une planche verticale où sont régulièrement plantés des clous, en quinconce. Des tubes transparents recueillent les billes en bas de la planche. Il s'y dessine une courbe de *Gauss*. Puis il compliqua sa machine. Plaçant des tubes à un rang intermédiaire, puis



Répartition des billes, après leur trajectoire aléatoire, rencontrant 14 clous d'une planche de *Galton* virtuelle (Excel)

### Bertillon, Galton et les empreintes digitales

Expert auprès de la préfecture de police de Paris, *Alphonse Bertillon* (1853 – 1914), effectua des travaux de mesures anthropométriques, de façon à identifier de façon unique les délinquants et faciliter leur découverte (1879). Une fois adopté, le système de *Bertillon* se diffusa rapidement (dès 1888 aux Etats-Unis). *Bertillon*, exploitant les progrès de son temps, est également à l'origine de la photographie anthropométrique figurant dans les fichiers de police. Il contribua à l'identification de l'anarchiste *Ravachol*. En revanche, expert au procès *Dreyfus*, il attribua à tort à ce dernier le bordereau annonçant l'envoi de documents militaires secrets.

*Galton*, reprenant les travaux de *Bertillon*, étudia les relations entre les différentes mesures du corps humain, effectuant des calculs de corrélation. Au problème d'identification des personnes, *Galton* apportera la solution des empreintes digitales, complémentaire aux techniques de *Bertillon*.

libérant un de ces tubes. Les billes libérées viendront se répartir selon une densité normale. La distribution globale étant la somme de ces sous distributions.

Grand admirateur de *Galton* (il lui consacra une biographie), *Karl Pearson* (1857 – 1936) est, contrairement à ce dernier, un mathématicien professionnel, mais également attiré par la physique, l'histoire et la philosophie. Professeur de mathématiques appliquées à l'University College de Londres, et lié d'amitié à son collègue de zoologie, *Weldon*, il se tourne, âgé de 33 ans, vers la statistique, dans le cadre de la théorie de la sélection naturelle de *Darwin* et dans la mouvance des travaux de *Galton*. S'appuyant sur les travaux de ce dernier, *Karl Pearson* poursuit l'étude de la **corrélation** et donne de son coefficient  $r$  l'expression que nous lui connaissons actuellement. Dans l'étude de la dispersion, il introduit le terme "**standard deviation**" (**écart type**), en 1893, ainsi que sa notation  $\sigma$ . On lui doit également le critère du **chi-deux**, permettant de caractériser la qualité d'ajustement

d'une distribution théorique à une distribution observée, ayant recours, pour ce faire, à de nombreux lancers de pièces de monnaies ou de dés, effectués par lui-même, ses élèves et ses proches (on ne dispose pas encore de techniques de simulation). Enfin, *Karl Pearson* crée un réseau scientifique et des institutions autour des questions statistiques liées à la biométrie (qu'il définit comme "*l'étude de l'application des méthodes mathématiques à l'examen des formes multiples de la vie.*"). Il est co-fondateur, en 1901, de la revue "*Biometrika*", "*in consultation with F. Galton*", qui publiera de nombreux articles de statistique et dont l'influence sera très grande. En 1906, il crée deux laboratoires, très proches, l'un de biométrie, l'autre d'eugénique. A sa mort, en 1936, cet ensemble aura donné trois laboratoires : statistique appliquée (dirigé par son fils *Egon Pearson*), eugénique (dirigé par *Ronald Fischer*) et génétique.

En cette fin de XIX<sup>e</sup> siècle, sont ainsi nés en Angleterre, dans le contexte particulier de la théorie de l'évolution, puis de l'eugénisme et de la biométrie, des outils statistiques puissants, une statistique fortement mathématisée et liée aux probabilités, dont les applications dépasseront rapidement le cadre de leur conception, pour répondre aux besoins de nombreux domaines (psychométrie, agronomie, industrie, économie...) et trouver, dans ces applications, de nombreuses causes de développements futurs. C'est ainsi que le sens du mot "statistique" lui-même connut un glissement, de "statistique administrative" ou "morale et sociale", au début du XIX<sup>e</sup> siècle, on en arrive au sens de "statistique mathématique" au début du XX<sup>e</sup>.

### La « saga » de la moyenne

Longtemps considérée comme paramètre idéal, la moyenne arithmétique atteint son apogée au XIX<sup>e</sup> siècle, avant d'être parfois remise en cause, les développements récents de l'informatique changeant un peu la donne. C'est l'histoire de ce paramètre, à travers le temps, que nous évoquons dans cet encadré.

#### Les débuts de la moyenne arithmétique

Il est difficile de dater avec précision la première utilisation statistique de la moyenne arithmétique en tant que telle, avec utilisation explicite de son expression mathématique. L'astronome danois *Tycho Brahé* (1546–1601), qui fut un observateur hors pair, accumula de nombreuses observations pour une même donnée astronomique, remplaçant celles-ci par une valeur "centrale" sans que l'on ait de preuve explicite de l'utilisation de la moyenne arithmétique pour cela.

Dans l'*Ars conjectandi* (1713) de *Jacques Bernoulli*, on trouve<sup>25</sup> ce conseil, pour obtenir la hauteur "moyenne" sur plusieurs mesures à l'aide d'un baromètre : "*Rassemble toutes les hauteurs que tu as observées, qu'elles soient différentes ou identiques, en une somme que tu divises par le nombre d'observations, ou, ce qui est plus avantageux, si les mêmes hauteurs ont été observées plusieurs fois, les différentes hauteurs sont multipliées par le nombre d'observations qui ont été faites de chacune d'entre elles, la somme de tous ces produits divisée par le nombre de ces observations donne la hauteur "moyenne" (mediam dans le texte original en latin) [...] il est évident qu'ils se trompent ceux qui, pour rechercher la quantité moyenne de mercure, font la moyenne arithmétique des extrêmes (aujourd'hui nommée étendue moyenne).*"

En 1725, l'astronome anglais *Flamsteed*, utilise la moyenne arithmétique pour obtenir le "milieu" et tenir compte des erreurs produites par son quart de cercle mural lors de la mesures d'ascensions droites (équivalent de la longitude) d'étoiles<sup>26</sup> :

<sup>25</sup> Cité par Jean Claude Girard dans "A bas la moyenne" – Repères IREM n°33.

<sup>26</sup> Exemple cité par Jean-Jacques Droesbeke dans la revue "Culture et Sociétés" – 1994.

"*Rectarum Solis Adscensionum Differentia inter 14um Martii ac 15um Septembris* [1690]  
*ex Observationibus circa Solem pro istis Diebus reperitur, viz.*  
*per Calcem Castoris* ..... 178° 36' 0"  
*per Procyonem*..... 178° 36' 5"  
*per Pollucem*..... 178° 36' 20"  
**Media** *inter has Differentia* ..... 178° 36' 8"."

En 1722, Roger Cotes utilise dans "*Harmonia Mensurarum*" une moyenne pondérée lorsqu'il dispose de quatre observations pour la position d'un même point.

L'usage de la moyenne arithmétique sera glorifié dans le cadre de la **théorie des erreurs**, dans un contexte probabiliste. En 1755 Thomas Simpson publie "*A letter to the Right Honourable George Earl of Macclesfield, President of the Royal Society, on the advantage of taking the mean of a number of observations in practical astronomy.*"

Suivant des idées analogues, Lagrange publie en 1774 un "*Mémoire sur l'utilité de la méthode de prendre le milieu entre les résultats de plusieurs observations, dans lequel on examine les avantages de cette méthode par le calcul des probabilités, et où l'on résout différents problèmes relatifs à cette matière.*"

L'article "*Milieu*" de l'*Encyclopédie méthodique* (1784), dont le début est reproduit ici, précise quels étaient les enjeux (l'article est de Jean Bernoulli).

**MILIEU à prendre entre les observations,**  
*(Arith.)* Ce sujet me paroît être devenu un de ceux qui sont le plus d'un ressort d'un ouvrage tel que celui-ci. Le *Dictionnaire raisonné des Sciences*, &c. semble promettre au mot ARITHMÉTIQUE de le traiter au mot MOYEN, mais on n'y trouve pas son attente remplie; je tâcherai de suppléer du moins en partie à cette omission.

Quand on a fait plusieurs observations d'un même phénomène, & que les résultats ne sont pas tout-à-fait d'accord entr'eux, on est sûr que ces observations sont toutes, ou au moins en partie peu exactes, de quelque source que l'erreur puisse provenir; on a coutume alors de prendre le milieu entre tous les résultats, parce que de cette manière les différentes erreurs se répartissant également dans toutes les observations, l'erreur qui peut se trouver dans le résultat moyen devient aussi moyenne entre toutes les erreurs. Il n'est pas douteux que cette pratique ne soit très-utile pour diminuer l'incertitude qui naît de l'imperfection des instrumens & des erreurs inévitables des observations; mais il est aisé de s'appercevoir qu'elle ne la diminue pas autant qu'on le desireroit, & qu'elle est susceptible à plus d'un égard d'être perfectionnée, parce qu'en prenant simplement le milieu arithmétique, on ne tient pas compte du plus ou moins de probabilité de l'exactitude des observations, des différens degrés d'habileté des observateurs, &c. Différens grands géomètres ont entrepris cette utile recherche, ils l'ont considérée sous différens points de vue, & l'ont traitée plus ou moins en détail; il est fort à souhaiter que les astronomes, les physiciens & généralement tous les observateurs, profitent des résultats de ces recherches dans la discussion de leurs observations.

En 1774, *Laplace* obtient cependant sa première loi des erreurs, de densité  $f(x) = \frac{k}{2} e^{-k|x|}$ ,

avec  $x \in \mathbb{R}$ , en considérant les écarts des observations à la médiane.

Dans les ouvrages d'astronomie ou de topographie du début du XIX<sup>e</sup> siècle, une certaine ambiguïté existe dans la terminologie<sup>27</sup>. Ainsi une expression telle que "erreur moyenne",

désignera, dans les ouvrages français, la moyenne des écarts à la médiane  $\frac{1}{n} \sum |x_i - Me|$ ,

et, dans les ouvrages allemands, l'écart type  $\sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2}$ .

### **Le triptyque "moindres carrés – moyenne – loi normale"**

C'est l'ensemble moindres carrés – moyenne – loi normale, les uns justifiant les autres, qui consacra l'usage de la moyenne (et de l'écart type).

En 1805, *Legendre*, dans ses "*Nouvelles méthodes pour la détermination de l'orbite des comètes*" explique que "*La règle par laquelle on prend le milieu (il s'agit de la moyenne arithmétique) entre les résultats de diverses observations (pour un seul élément), n'est que la conséquence très simple de notre méthode générale, que nous appelons méthode des moindres carrés.*" Et de poursuivre en indiquant que ce minimum des carrés des écarts est

obtenu en annulant  $\frac{d}{dx} \sum (x - x_i)^2 = 2 \sum (x - x_i)$  d'où  $x = \frac{1}{n} \sum x_i = \bar{x}$ .

En 1809, *Gauss* fait le lien avec la loi "normale", montrant que si l'on considère que les erreurs sont aléatoires, alors la loi de probabilité validant la moyenne et la méthode des moindres carrés comme meilleure estimation, est la loi normale (voir au paragraphe précédent sur la loi "normale"). On en conclut que, lorsque les erreurs se répartissent selon une loi normale, alors la moyenne fournit la meilleure estimation du paramètre mesuré.

En 1810, *Laplace*, en établissant ce qu'on appellera le théorème limite central, ira plus loin, en montrant que même si la distribution des erreurs n'est pas normale, celle de leur moyenne tend, en général, vers une loi normale.

Ces résultats feront donc de la moyenne (et donc de l'écart type) un paramètre incontournable, d'autant que *Quételet* en étendra l'usage à d'autres domaines, comme celui des sciences humaines.

### **Autres avantages de la moyenne arithmétique**

La moyenne est alors au faîte de sa gloire, d'autant qu'elle possède d'autres atouts mathématiques.

Le premier est sa linéarité : si les valeurs observées subissent une transformation linéaire, la moyenne  $\bar{x}$  est soumise à la même transformation.

Le second est qu'elle peut se calculer par regroupement des données (comme un

barycentre), par exemple en deux groupes de tailles  $n_1$  et  $n_2$  :  $\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$ .

Un autre avantage sera développé au XX<sup>e</sup> siècle dans le cadre de l'estimation et des sondages aléatoires. C'est la bonne connaissance que l'on a de la loi de probabilité de la variable aléatoire correspondant à la moyenne arithmétique d'un échantillon de taille  $n$ . C'est la continuation des théorèmes limites, notamment avec, en 1908, la loi de *Student*, dans le cas de petits échantillons extraits d'une population normale d'écart type inconnu.

<sup>27</sup> Signalé par Marie Françoise Jozeau de l'IREM Paris 7.

### Les autres "moyennes"

Bien sûr, ce succès de la moyenne arithmétique n'allait pas de soi et l'on avait bien d'autres façons de déterminer le "milieu" de plusieurs valeurs.

**La moyenne géométrique**  $G = \sqrt[n]{x_1 \times \dots \times x_n}$  :

La comparaison de deux valeurs n'est pas seulement basée sur leur différence, mais aussi sur leur rapport. Ainsi, lorsque que la variable étudiée correspond à une loi d'agrégation multiplicative, c'est la moyenne géométrique qui convient (c'est par exemple le cas bien connu des pourcentages d'augmentation, où le taux moyen  $i$  d'augmentation sur deux mois consécutifs sera donné par  $1 + i = \sqrt{(1 + i_1)(1 + i_2)}$ ).

En 1879 *Galton* écrit ainsi un article intitulé "*The Geometric Mean in Vital and Social Statistics*."

**La moyenne harmonique**  $H = \frac{n}{\sum \frac{1}{x_i}}$  :

La moyenne harmonique donne une valeur "centrale" chaque fois qu'une variable intervient par son inverse, par exemple dans le cas de la recherche de la vitesse moyenne, connaissant la vitesse aller et la vitesse retour.

**La médiane :**

C'est bien entendu la médiane qui a constitué (et constitue toujours) l'alternative la plus sérieuse à la moyenne. La médiane n'est pas de même nature que les autres moyennes. Elle est définie comme étant l'observation centrale, dans le cas d'un nombre impair d'observations rangées en ordre croissant, ou la moyenne des deux valeurs centrales, dans le cas d'un nombre pair d'observations.

Dès 1669, *Christian Huygens*, dans sa correspondance avec son frère *Louis*<sup>28</sup>, distingue l'espérance de vie (correspondant à la moyenne arithmétique) et ce que l'on appellera la "durée de vie probable" (correspondant à la médiane) qui est la valeur pour laquelle la probabilité d'être inférieur à cette valeur égale celle d'y être supérieur. Il écrit ainsi : "[Ce sont] deux choses différentes que l'espérance ou la valeur de l'âge futur d'une personne, et l'âge auquel il y a égal apparence qu'il parviendra ou ne parviendra pas. Le premier est pour régler les rentes à vie, et l'autre pour les gageures."

On a vu que si la moyenne  $\bar{x}$  correspond aux moindres carrés (c'est à dire minimise la somme des distances au sens de la distance euclidienne usuelle), la médiane  $Mé$  est obtenue lorsque l'on minimise la somme  $\sum |x_i - M|$  des écarts absolus.

On en trouve l'usage chez *Gauss* (1816) et *Laplace* (1818), mais c'est surtout *Galton* qui lui accorde toute son attention en 1874, 1875, dans le cadre de ses études anthropologiques. Mais l'usage de la médiane sera longtemps négligé, en raison, d'une part, de moins bonnes propriétés algébriques, et, d'autre part, de qualités statistiques moindres dans le cadre d'une population normale.

En 1882 cependant, *Simon Newcomb* examine 684 résidus basés sur l'observation de passages de Mercure devant le Soleil. Il constate que leur distribution avait une queue plus épaisse que celle de la loi normale. Afin de remédier à la moyenne arithmétique, ici défailante, il repensera à la médiane.

### La difficulté d'interprétation de la moyenne

Dès la fin du XIX<sup>e</sup> siècle, l'usage de la moyenne est critiqué dans la mesure où son interprétation est abstraite (que penser d'une famille possédant 2,51 enfants ?) et parfois

<sup>28</sup> Voir Bernard Parzysch – Bulletin A.P.M.E.P. n°416.

contestable (voir "l'homme moyen"). Ce n'est pas le cas de la médiane, qui est en général une des valeurs observées. De plus, la médiane existe dans des cas où la moyenne  $\bar{x}$  n'existe pas.

### La "robustesse" de la médiane

Ce qui a surtout redoré le blason de la médiane c'est sa "robustesse", c'est à dire sa faible dépendance aux valeurs aberrantes.

On avait rapidement constaté la mauvaise influence des valeurs extrêmes sur  $\bar{x}$ . L'une des solutions, préconisée par exemple par l'astronome *Roger Boscovich*, était la **moyenne tronquée** (c'est à dire calculée après suppression des valeurs aberrantes). Les fermiers généraux de l'Ancien régime utilisaient cette technique pour calculer l'impôt : on se basait sur la récolte moyenne des cinq dernières années, après suppression de la meilleure et de la plus mauvaise récolte.

La formulation actuelle de la robustesse a été introduite par *George Box* en 1953, puis par *Peter Huber* en 1964. L'usage accru de la médiane doit ensuite beaucoup aux diagrammes en boîtes (Box Plot), particulièrement parlants, de *John Turkey* dans les années 1970.

L'usage accru de l'informatique en statistique permit un recours plus fréquent, quand il se justifie à la médiane, dont l'obtention ne pose plus de problèmes.

## IV – LES METHODES STATISTIQUES DU XX<sup>e</sup> SIECLE

### 1 – La statistique inférentielle

La statistique connut, au début du XX<sup>e</sup> siècle, une véritable révolution, celle de l'inférence. Il s'agit, à partir de résultats statistiques limités, observés sur un échantillon, d'inférer à toute une population pour laquelle on induira des estimations, à partir des observations. Cette démarche trouve sa motivation dans des domaines appliqués spécifiques (laboratoires agronomiques, industriels...), essentiellement en Grande Bretagne et aux Etats-Unis, mais les outils ainsi créés trouveront, de part leur efficacité, des applications quasi universelles. La France restera à l'écart de ce mouvement. A l'époque du *bourbakisme* montant, si de grands noms comme *Emile Borel*, *Maurice Fréchet*, *Paul Levy*, s'illustrent dans la théorie (formelle) des probabilités, les méthodes statistiques ont mauvaise presse dans les milieux universitaires. Il faudra attendre les années cinquante (et encore plus tard dans l'enseignement) pour que, face à la pression économique des applications et à l'efficacité des nouvelles méthodes statistiques, la France commence à refaire son retard dans ce domaine.

Depuis le XVIII<sup>e</sup> siècle, deux approches des probabilités coexistent : le point de vue *subjectif* (représenté par *Bayes* et *Laplace*), où l'on considère la probabilité en termes de "raisons de croire", d'estimation du degré de confiance dans la réalisation d'un événement aléatoire et le point de vue *fréquentiste*, s'appuyant sur la loi des grands nombres de *Bernoulli* et les théorèmes limites (*Laplace*), où l'on conçoit la probabilité comme stabilisation limite de la fréquence lors de la répétition, un grand nombre de fois, de l'évènement. Au XIX<sup>e</sup> siècle, *Quételet* et ses successeurs ne retiendront de la synthèse de *Laplace* que l'aspect fréquentiste, laissant en sommeil les techniques d'estimations amorcées au XVIII<sup>e</sup>.

Les formulations probabilistes de l'estimation réapparaissent, dans la mouvance de l'école de *Karl Pearson*, avec *Ronald Aylmer Fischer* (1890 – 1962) et *William Sealy Gosset* – alias *Student* – (1876 – 1937), tous deux confrontés à un nombre insuffisant de données, rendant plus difficile la perspective fréquentiste. *Fischer*, travaillant à l'étude des engrais

dans un centre agronomique, ne peut recourir qu'à un nombre limité d'essais contrôlés, *Gosset*, à la brasserie Guinness de Dublin, ne dispose, pour ses études de qualité, que d'échantillons de petite taille, en raison de la grande variabilité de la production qui n'en permet pas l'homogénéité. *Fischer* et *Gosset* sont alors amenés à utiliser des notations différentes pour la valeur théorique  $\theta$  du paramètre d'une distribution de probabilité, et pour l'estimation  $\hat{\theta}$  de ce paramètre, aux vues des observations. Cette innovation dans les notations rend possible le développement de la statistique inférentielle dans deux directions, celle de l'estimation (dominée par un esprit subjectif) et celle de la théorie des tests d'hypothèses (s'inscrivant davantage dans la tradition fréquentiste).

**Ronald Fischer** est généralement présenté comme le père de la théorie de l'**estimation** selon le principe du "*maximum de vraisemblance*" (introduit en 1912 et développé jusqu'en 1922-1925). Il présente cette notion de vraisemblance comme descendante de celle de "probabilité inverse" de *Laplace*. Gauss, on l'a vu, s'en servit déjà pour établir la loi normale, mais *Fischer* justifia et développa la méthode.

Il s'agit de retenir comme estimation, parmi un ensemble de valeurs possibles d'un paramètre  $\theta$ , celle qui rend la plus vraisemblable les observations effectuées, c'est à dire celle pour laquelle, rétrospectivement, la probabilité des observations (réellement effectuées) est la plus grande. Si la fonction  $f$  est la densité théorique d'une variable aléatoire  $X$  dépendant d'un paramètre  $\theta$  inconnu, il faut, pour les observations indépendantes  $x_1, \dots, x_n$  de la variable aléatoire  $X$ , prendre comme estimation de  $\theta$  la valeur

de  $\theta$  qui rend maximum la vraisemblance  $L$  définie par :  $\ln L(x_1, \dots, x_n; \theta) = \sum_{i=1}^n \ln f(x_i; \theta)$ .

En effet, en vertu de l'indépendance, la probabilité de l'intersection est obtenue à partir du produit des densités (d'où la somme en prenant le logarithme).

Les apports de *Fischer* à la statistique ne se limitent pas à l'estimation. En liaison avec *Gosset*, il introduit la notion de "**degrés de liberté**". A la station agronomique expérimentale de Rothamsted en Angleterre, il conçoit et utilise les méthodes de "**plans d'expérience**" (voir l'encadré suivant<sup>29</sup>) et de "**randomisation**" (néologisme fabriqué sur le mot anglais random, signifiant hasard). La randomisation des expériences consiste à introduire le hasard dans la procédure expérimentale, de façon à éviter les biais. Si, par exemple, on veut comparer quatre types d'engrais et que l'on dispose, pour ce faire de 20 parcelles, chaque type d'engrais sera affecté à 5 parcelles. Mais comment les choisir ? Ces parcelles sont peut-être de fertilité inégale et l'on commettra une erreur en attribuant à un type donné d'engrais les parcelles les plus fertiles. La seule façon de procéder est de tirer au hasard les 5 parcelles attribuées à un engrais spécifié.

*Fisher* est, selon Dreesbeke et Tassi<sup>30</sup>, "*l'homme qui a fait de la statistique une science moderne*". Son ouvrage "*Statistical Methods for Research Workers*", publié en 1925 sera le "best-seller" de la statistique, avec 14 éditions et des traductions en 6 langues.

**William Gosset** est le précurseur des statisticiens industriels. Il fit toute sa carrière dans les brasseries Guinness, délaissant les possibilités qui lui furent offertes d'une carrière universitaire. Considéré par les brasseurs comme l'un des leurs, occupant ses loisirs à la statistique en vue de l'amélioration de la production, ses échanges avec les statisticiens universitaires étaient parfois vus d'un mauvais oeil par ses employeurs. Ceci explique le surnom de *Student* utilisé pour dénommer la loi dont il est à l'origine. En effet, la société Guinness l'autorisa à publier ses articles à condition qu'il use au choix, du pseudonyme "Pupil" ou "Student". *Gosset* choisit le second.

<sup>29</sup> On pourra, à propos des plans d'expérience, lire l'exemple, simple et spectaculaire, donné à la fin du chapitre 7.

<sup>30</sup> "Histoire de la Statistique" – "Que sais-je ?" – P.U.F. 1997.

### Où une récréation mathématique devient une procédure statistique

Lorsque *Fisher* travaille à la station agricole expérimentale de Rothamsted, il est amené à faire de nombreuses expériences. Compte tenu de la précision qu'il veut obtenir et vu le temps que cela prend (il faut attendre que ça pousse !), il veut réduire le nombre d'expériences. Il faut donc qu'il les organise de façon à tirer de celles qui sont faites le maximum d'informations utiles, ce sont les "plans d'expérience". Cela l'amène, en particulier, à utiliser des structures mathématiques de nature arithmétique, géométrique, combinatoire, étudiées depuis l'Antiquité (voir l'œuvre de *Diophante*) et qui, jusque là figuraient dans la rubrique des récréations mathématiques. La plus célèbre est celle des carrés latins, également appelés carrés Eulériens (*Euler* 1782).

Supposons, par exemple, que le rendement  $y$  d'une culture dépende de 3 facteurs  $x_1, x_2, x_3$ , correspondant respectivement au nitrate (N), au potassium (K) et au phosphore (P) et ayant chacun 5 niveaux possibles 1, ..., 5. Une première méthode consisterait, pour chaque niveau de l'un des facteurs, à faire varier le niveau des autres, mais cela conduirait à  $5^3 = 125$  expériences, ce qui est beaucoup trop.

Supposant un modèle linéaire, on posera donc a priori  $y = \lambda_0 + \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \varepsilon$  où  $\varepsilon$  correspond aux erreurs de mesure, à l'influence variable de la météo etc. On supposera que  $\varepsilon$  peut être considérée comme une variable aléatoire de loi normale  $N(0, \sigma)$ . Les expériences serviront alors à estimer les valeurs de  $\lambda_0, \lambda_1, \lambda_2, \lambda_3$  et  $\sigma$ .

Il faut alors déterminer quelles sont les combinaisons des différents niveaux des facteurs  $x_1, x_2$  et  $x_3$  qui feront l'objet des expériences, afin d'obtenir la meilleure information possible. Dans cette situation, une bonne solution consiste à organiser 25 expériences selon le premier carré suivant, où chaque ligne correspond au niveau (1, 2, ..., 5) de nitrate, chaque colonne au niveau de potassium et la lettre latine inscrite dans chaque case au niveau de phosphore.

A	B	C	D	E
E	A	B	C	D
D	E	A	B	C
C	D	E	A	B
B	C	D	A	A

A $\alpha$	B $\beta$	C $\gamma$	D $\delta$	E $\varepsilon$
E $\beta$	A $\gamma$	B $\delta$	C $\varepsilon$	D $\alpha$
D $\gamma$	E $\delta$	A $\varepsilon$	B $\alpha$	C $\beta$
C $\delta$	D $\varepsilon$	E $\alpha$	A $\beta$	B $\gamma$
B $\varepsilon$	C $\alpha$	D $\beta$	E $\gamma$	A $\delta$

A $\alpha$	B $\beta$	C $\gamma$	D $\delta$	E $\varepsilon$	F $\varphi$
F $\beta$	A $\gamma$	B $\delta$	C $\varepsilon$	D $\varphi$	E $\alpha$
E $\gamma$	F $\delta$	A $\varepsilon$	B $\varphi$	C $\alpha$	D $\beta$
D $\delta$	E $\varepsilon$	F $\varphi$	A $\alpha$	B $\beta$	C $\delta$
C $\varepsilon$	D $\varphi$	E $\alpha$	F $\beta$	A $\delta$	B $\varepsilon$
B $\varphi$	C $\alpha$	D $\beta$	E $\delta$	F $\varepsilon$	A $\varphi$

Ainsi, chaque niveau de phosphore (lettre latine) est testé une fois avec chaque niveau de potassium et de nitrate. Cette façon de "brouiller les cartes" permet de réduire les erreurs systématiques provenant d'un niveau particulier de l'un des éléments.

Si l'on doit faire intervenir un quatrième facteur, soit le niveau de magnésium (Mg), correspondant à un nouveau terme  $\lambda_4 x_4$ , on construira un carré graeco-latin, où chaque lettre grecque (niveau de Mg) coïncide exactement avec chaque lettre latine (deuxième tableau).

Il existe des plans carrés graeco-latin pour toutes les dimensions  $n$ , sauf  $n = 1 ; 2$  et  $6$  (troisième tableau).

La loi de *Student* est utilisée dans les questions d'estimation basées sur de petits échantillons, lorsque l'écart type est inconnu et que l'usage du théorème limite central et de la loi normale n'est pas possible. En effet, *Gosset* s'aperçoit rapidement, lors de ses études de qualité de fabrication, que les conditions varient tellement (température, provenance du houblon, du malt, conditions de fabrication...) que les données homogènes sont peu nombreuses. La "loi des erreurs" classique ne peut pas s'appliquer dans ces conditions.



Considérons, qu'au sein d'une production répartie selon une loi normale de moyenne  $\mu$  et d'écart type  $\sigma$ , on prélève au hasard un échantillon de taille  $n$  (petit). On note  $X_i$  la variable aléatoire qui, au  $i^{\text{ème}}$  tirage, associe son résultat. On suppose que les  $X_i$  sont indépendantes, de même loi normale  $N(\mu, \sigma)$ . On note encore  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  la variable aléatoire qui, à chaque échantillon prélevé, associe sa moyenne. *Gosset* étudie la variable aléatoire  $S_n^2 = \frac{1}{n} \sum X_i^2 - \bar{X}_n^2$  correspondant à la variance des échantillons, puis étudie la loi de la variable aléatoire  $T_{n-1} = \frac{\bar{X}_n - \mu}{S_n} \sqrt{n-1}$  (**loi de Student** à  $n - 1$  degrés de liberté).

**Egon Pearson** (1895 – 1980), fils de *Karl*, et **Jerzy Neyman** (1894 – 1981), ont laissé leurs noms attachés à la notion de **tests d'hypothèses**.

**Egon Pearson** n'est pas que le fils de son père, duquel il sera relativement indépendant. Il collaborera étroitement avec *Gosset*, sur la loi de *Student* et sur des questions de **robustesse** (sensibilité aux valeurs "aberrantes") en utilisant des tables de nombres au hasard, et avec *Neyman*, pour les tests d'hypothèses. Il jouera également un rôle important dans le développement des techniques statistiques dans l'industrie, devenant l'un des fondateurs de la "*Société de Recherche opérationnelle britannique*" en 1948 (domaine développé en Angleterre durant la seconde guerre mondiale).

Mathématicien d'origine russo-polonaise, **Neyman** fait un séjour d'étude en 1924 à l'University College qui le met en contact avec *Fischer*, *Gosset*, *Karl* et *Egon Pearson* (avec lequel il correspondra beaucoup). En 1925, il assiste à Paris aux cours de *Lebesgue*, *Hadamard* et *Borel*. De retour en Pologne, il assure la direction du département statistique d'un institut de biologie. En 1934, il est en poste à l'University College aux côtés d'*Egon Pearson* mais en 1938, à l'approche de la guerre, il part pour *Berkeley*, aux Etats-Unis.

A la différence de la théorie de l'estimation de *Fischer*, celle des tests d'hypothèse d'*Egon Pearson* et *Neyman* est davantage élaborée, dans les années 1925, à partir de la vision fréquentiste de probabilité. Cette différence de point de vue sera à l'origine d'une vive controverse avec *Fischer*. Dans les tests de *Pearson* et *Neyman*, au lieu d'estimer un paramètre selon la vraisemblance avec les données, on affecte une valeur a priori (en quelque sorte "objective") à ce paramètre ("hypothèse nulle"), et on évalue les risques de l'acceptation d'une hypothèse fautive ou du rejet d'une hypothèse vraie. Cette valeur objective est, par exemple, la norme d'une production industrielle. Selon les valeurs observées, on considèrera que l'hypothèse selon laquelle cette norme est respectée, est, ou non, acceptée. Le terme d'hypothèse "nulle" est dû à *Fisher* (1951), il correspond, à l'origine, au cas d'un traitement sans effet, où lorsqu'il n'y a pas d'effet significatif entre deux traitements (voir l'encadré sur "la danse de la pluie"). De façon générale, *E. Pearson* et *Neyman* mirent en évidence en 1928 le rôle dissymétrique joué par l'hypothèse  $H_0$ , que l'on privilégie, et l'hypothèse  $H_1$  qu'on lui oppose. En 1933, ils introduisent la distinction entre erreur de première espèce (rejet à tort de  $H_0$ ) et de seconde espèce (acceptation à tort de  $H_0$ ).

### Jerzy Neyman réhabilite la danse de la pluie

Dans un article de 1967 ("*Experimentation with weather control*"), *Jerzy Neyman* montre l'importance du choix de la méthodologie statistique employée dans les expérimentations, privilégiant les tests d'hypothèses (lorsqu'ils sont praticables), jugés plus objectifs, plutôt que les techniques d'estimation, davantage subjectives.

Les nuages sont presque toujours constitués d'eau surfondue (c'est à dire à l'état liquide à une température inférieure à 0°) coexistant en état instable avec des cristaux de glace. En 1946 *Schaefer* découvre que l'apport d'iodure d'argent provoque alors artificiellement des précipitations vers le sol. Cette découverte suscite beaucoup d'espoirs pour favoriser les précipitations dans les régions arides et des industriels (les "faiseurs de pluie") proposent à partir des années 1950 l'ensemencement des nuages par iodure d'argent. A l'appui de leurs services, ces industriels fournissaient les statistiques suivantes.

Expérience	Année	Durée	Pourcentage d'augmentation des précipitations
Pennsylvanie	1954	1 mois	+ 17 %
Pennsylvanie	1955	2 mois	+ 33 %
Caroline du Sud	1957	2 mois	+ 19 %
New Hampshire	1957	2 mois	+ 21 %
Massachusset	1957	1 mois	+ 30 %
Pennsylvanie	1957-58	5 mois	+ 6 %
New York	1962	1 mois	+ 57 %
Pennsylvanie	1963	3 mois	+ 5 %
Connecticut	1964	1 mois	+ 29 %
New York	1964	1 mois	+ 37 %
Maryland	1964	3 mois	+ 14 %
Massachusset	1964	1 mois	+ 8 %
New Hampshire	1964	19 jours	+ 14 %
New Jersey	1964	3 mois	+ 0 %

D'abord surpris par la très courte durée des expériences, *Neyman* examine la façon dont ces résultats sont obtenus.

Les précipitations sur la cible sont mesurées par différents instruments, en différents lieux. Si l'on désigne par  $y$  une moyenne des mesures simultanées, une *estimation* de  $y$  en l'absence d'ensemencement des nuages est obtenue à partir des valeurs  $x_1, \dots, x_n$  correspondant aux précipitations enregistrées sur les différents sites les années précédant l'expérience d'ensemencement, à partir d'une équation de régression de  $y$  en  $x_1, \dots, x_n$ . La comparaison de la valeur estimée de  $y$  sans ensemencement, avec la valeur de  $y$  obtenue à l'aide des observations durant l'expérience d'ensemencement, est alors considérée comme une mesure de l'effet de celle-ci.

*Neyman* admet qu'à première vue cette méthode peut sembler attrayante, mais souligne d'emblée l'absence de randomisation (choix aléatoire des jours où l'on pratiquera l'ensemencement ou non) dont l'intérêt était pourtant depuis longtemps souligné par *Fischer*. Selon *Neyman*, "ordinairement, une attitude raisonnable à l'égard des résultats d'une expérience non randomisée, est qu'elle peut être biaisée. A l'encontre de cela, le second tableau [ qui suit] peut être vu comme une preuve que les résultats du premier tableau sont réellement biaisés."

Il faut dire que, selon *Neyman*, les pressions furent fortes, tant de la part des industriels, que de certains lobbies agricoles.

Parallèlement à ces résultats, obtenus par les industriels, 19 autres expériences, cette fois randomisées, et beaucoup plus longues, avaient été menées dans différents pays par des laboratoires indépendants. Chaque jour où les conditions météorologiques semblaient favorables à l'ensemencement (présence de nuages), on tirait au sort pour savoir si on ensemencait les nuages à l'iodure d'argent ou non, de façon à comparer ensuite les précipitations obtenues, avec ou sans ensemencement.

Expérience	Année (début)	Durée en années	Pourcentage de changement dans les précipitations attribuable à l'ensemencement
USA 1	1953	1,5	-3,3 ; +12,3 ; -0,4 ; +0,6 ; +23,9 ; +0,4
USA 2	1953	1,5	-5,6 ; -33,9 ; -16
USA Santa Barbara	1957	3	-8 ; +125 ; -39 ; +124 ; -16 ; +40 ; -27 ; +58
USA Arizona 1	1957	4	-30 ; -7
USA Arizona 2	1961	3	-30
USA Whitetop	1960	5	-54,8 ; -39,4 ; -23,5
USA Lake Almanor	1962	1	+41,6 ; -12,1
Australie 1	1955	5	+19
Australie 2	1957	3	-5
Australie 3	1957	6	+4
Australie 4	1958	6	+4
Australie 5	1959	4	-5
Australie 6	1960	1,5	-13
Mexique	1956	10	+20 ; -8
Suisse	1957	7	+32,9 ; +0 ; +78,8 ; -16,2
Québec	1959	4	-2
Japon	1960	0,25	+48
France	1961	1,5	-6,6
Israël	1961	4,5	+6,4 ; +24,8

L'un des biais de la méthode d'estimation des industriels (tableau 1) était la grande dépendance de leur formule de régression à la durée des statistiques météorologiques utilisées : pour une zone donnée, on s'aperçoit que la fréquence d'un certain type d'orage peut varier non seulement d'année en année mais aussi de décennie en décennie.

*Neyman* conclut à la nécessité de développer une méthodologie appropriée pour juger si les différences (positives ou négatives) entre les quantités de précipitations ensemencées ou non sont significatives ou l'effet du hasard. Cette méthodologie est celle d'un test de comparaison sur des échantillons aléatoires, l'un avec nuages ensemencés, l'autre non.

Entre le 1<sup>er</sup> juin et le 23 août 1975 (soit 83 jours), une expérience s'est par exemple tenue en Floride. Les 24 jours où les conditions météorologiques furent favorable à l'ensemencement des nuages à l'iodure d'argent, les expérimentateurs tirèrent au sort pour savoir si, oui (codé 1) ou non (codé 0), ils interviendraient sur les nuages. Ils relevèrent les volumes de pluie (unité 10 millions de m<sup>3</sup>) tombées sur le bassin versant, sur une durée de 6 heures, le jour de l'essai.

Expérience	1	2	3	4	5	6	7	8	9	10	11	12
Date	0	1	3	4	6	9	18	25	27	28	29	32
Ensemencement 0 / 1 pour non / oui	0	1	1	0	1	0	0	0	0	1	1	1
Précipitations (10 <sup>7</sup> m <sup>3</sup> )	12,85	<b>5,52</b>	<b>6,29</b>	6,11	<b>2,45</b>	3,61	0,47	4,56	6,35	<b>5,06</b>	<b>2,76</b>	<b>4,05</b>

Expérience	13	14	15	16	17	18	19	20	21	22	23	24
Date	33	35	38	39	53	55	56	59	65	68	82	83
Ensemencement 0 / 1 pour non/oui	0	1	1	0	0	1	0	1	1	0	1	0
Précipitations (10 <sup>7</sup> m <sup>3</sup> )	5,74	<b>4,84</b>	<b>11,86</b>	4,45	3,66	<b>4,22</b>	1,16	<b>5,45</b>	<b>2,02</b>	0,82	<b>1,09</b>	0,28

Désignons par  $X_0$  la variable aléatoire qui, à une période de 6 heures choisie au hasard, associe le volume des précipitations sans ensemencement (en 10<sup>7</sup> m<sup>3</sup>). Selon un historique statistique important, on peut considérer que  $X_0$  suit la loi normale  $N(\mu_0 = 4, \sigma = 3)$ . La

variable aléatoire  $X_1$  correspondant au volume des précipitations avec ensemencement suivra la loi normale  $N(\mu_1; \sigma = 3)$ . On construit un test permettant de décider si l'hypothèse  $H_0: \mu_0 = \mu_1$  (ensemencement inefficace) ou l'hypothèse  $H_1: \mu_0 < \mu_1$  (ensemencement efficace) doit être retenue.

Compte tenu du coût important de l'ensemencement des nuages, les agriculteurs n'accepteront l'investissement que si l'augmentation observée est significative, c'est à dire très au-delà de la variabilité naturelle.

Soit  $\bar{X}_0$  la variable aléatoire associant à 12 expériences sans ensemencement, pratiquées aléatoirement et de façon indépendante, la moyenne  $\bar{x}_0$  des précipitations observées. Cette variable aléatoire suit la loi  $N(4, \frac{3}{\sqrt{12}})$ .

Sous l'hypothèse  $H_0$ , la variable aléatoire  $\bar{X}_1$ , correspondant aux moyennes observées sur 12 expériences d'ensemencement, suivra la même loi.

Sous l'hypothèse  $H_0$ , où l'ensemencement est inefficace,

la variable aléatoire différence  $\bar{X}_1 - \bar{X}_0$  suit donc la loi

normale  $N(0, \sqrt{\frac{9}{12} + \frac{9}{12}})$  (les variances s'ajoutent) dont

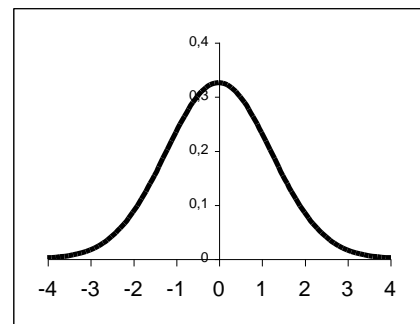
la densité est représentée ci-contre.

Ainsi, sous l'hypothèse que l'ensemencement est inefficace, la différence des moyennes obtenues sur 12 expériences avec et sans traitement s'observe selon cette densité. On pourra alors considérer que

l'ensemencement des nuages est significativement efficace lorsque la différence observée  $\bar{x}_1 - \bar{x}_0$  est supérieure à 2,01 car, d'après la courbe ci-dessus, une telle différence ne peut être due au hasard (variabilité sous l'hypothèse  $H_0$ ) que dans 5% des cas (seulement 5% de la surface sous la courbe est située à droite de 2,01).

Dans l'étude menée, la moyenne des 12 expériences sans ensemencement est  $\bar{x}_0 \approx 4,17$  alors que celle des 12 expériences avec ensemencement est  $\bar{x}_1 \approx 4,63$ . La différence  $\bar{x}_1 - \bar{x}_0$  vaut donc environ 0,46 qui est en plein dans la bosse de la variabilité sous  $H_0$ .

Ainsi l'ensemencement des nuages n'est pas statistiquement plus efficace que la danse de la pluie des indiens d'amérique.



*Jerzy Neyman* est également, par sa théorie de l'**estimation par intervalle de confiance**, à l'origine des techniques modernes de **sondage**. Si les premiers essais d'application des probabilités aux sondages aléatoires datent de la fin XVIII<sup>e</sup> siècle, la notion de probabilité étant dans les esprits liée à l'incertitude, au défaut de connaissance, les enquêtes les plus exhaustives possibles seront privilégiées au cours du XIX<sup>e</sup> siècle. De 1895 jusque dans les années 1930, le débat sur la représentativité des échantillons agite les congrès de l'Institut International de la Statistique. Dans un premier temps, jusque vers 1925, on se demande si l'on peut raisonnablement procéder par échantillonnage, ou s'il faut privilégier systématiquement les recensements. Puis à partir de 1925, avec le développement de l'utilisation des sondages, en particulier aux Etats-Unis avec les enquêtes d'opinion, la discussion porte sur la façon de procéder à l'échantillonnage, entre les tenants du "choix raisonné" et ceux de l'aléatoire.

Le premier, *Neyman* prend nettement parti pour la méthode aléatoire. Le hasard seul permettant d'appliquer la théorie des probabilités et d'encadrer le risque, alors que la raison humaine est vecteur d'introduction de biais. En 1934, *Neyman* montre que les hypothèses garantissant la convergence des estimateurs obtenus par "choix raisonné" ne peuvent être obtenues dans la pratique. De 1934 à 1937, il expose la théorie de l'estimation par intervalles de confiance : selon la répartition possible des différents échantillons que l'on est susceptible de prélever (répartition donnée par le théorème limite central), on construit réciproquement, à partir de l'échantillon effectivement prélevé, un intervalle avec par exemple 95% de chances que cette construction aboutisse à un intervalle contenant effectivement le pourcentage à estimer dans la population. Au delà de l'échantillonnage aléatoire simple, *Neyman*, en étudiant

l'**échantillonnage stratifié** établit le lien entre l'aléatoire et ce que l'on sait déjà par ailleurs (par exemple par un recensement donnant la répartition par critères socioprofessionnels, types de familles, âges...), par tirage au hasard à l'intérieur de strates établies a priori dans la population.

Les techniques d'estimation et de tests d'hypothèses seront formellement unifiées dans le cadre de la **théorie des décisions statistiques**, élaborée dans les années 1940 par **Abraham Wald** (1902 – 1950). Dans le même ordre d'idées que la théorie des jeux, développée à la même époque par *Von Neumann* et *Morgenstern*, estimation et tests se réduisent alors à la recherche d'une "fonction de décision". Dans le cas d'un test, cette fonction de décision est binaire (acceptation ou refus d'une hypothèse, 1 ou 0), dans le cas de l'estimation, l'ensemble des décisions est isomorphe à celui des paramètres (voir cette présentation au chapitre 5). Ces recherches sont publiées en 1939, aux Etats-Unis, où *Abraham Wald*, juif d'origine hongroise, vient de s'établir, fuyant l'annexion de l'Autriche par les nazis en 1938. Outre le domaine des tests séquentiels, *Wald* travailla également à différentes applications dans le domaine économique, par exemple, à la correction saisonnière des séries temporelles.

Les méthodes statistiques de **contrôle de qualité** dans l'**industrie** furent particulièrement développées aux Etats-Unis. **Walter Shewhart** (1891-1967) fit la plus grande partie de sa carrière à la *Bell Telephone Company* où il met au point, en 1924, les **cartes de contrôle** à

### Sondages : quand le hasard l'emporte sur la quantité

Le berceau des sondages se trouve aux Etats-Unis, lors des couvertures, par la presse, des campagnes présidentielles.

Dès 1824, le *Harrisbourg Pennsylvanian* et le *Raleigh Star* organisent des "votes de paille".

L'habitude, vu l'intérêt du public, sera vite prise mais le choix des échantillons ne repose sur aucun critère.

Seule la quantité, de plus en plus impressionnante (30000 personnes en 1905, plus de 2 millions en 1936) compte.

Une date décisive pour l'affirmation de la méthode par échantillon aléatoire représentatif est celle du 3 novembre 1936, où *F. D. Roosevelt* remporte l'élection présidentielle :

Le *Literary Digest* avait prédit la victoire du républicain *Alf Landon* à partir d'un "vote de paille" effectué sur plus de deux millions de personnes, alors que **George Gallup** avait annoncé celle de *Roosevelt* selon un échantillon représentatif réduit.

Il faut dire que le sondage du *Literary Digest* avait été effectué à partir de listes du type annuaire du téléphone (en 1936 !), membres de clubs... ce qui introduisait un biais certain. *Alf Landon*, dont on a aujourd'hui oublié le nom, a pu ainsi se croire président quelques heures, alors qu'il ne reçut que 40% des suffrages.

la base de la "maîtrise statistique des procédés". Il s'agit de définir statistiquement des limites de contrôle de certains paramètres de la production (fréquence, moyenne, étendue ou écart type) telles que si, au cours d'un échantillonnage périodique, ces limites sont dépassées, des actions de correction puissent être menées.

En 1935, *Egon Pearson* publiait à Londres un ouvrage fondamental dans ce domaine : "*The application of statistical methods to industrial standardisation and quality control*". Les applications statistiques dans le domaine du contrôle industriel se développèrent alors de façon importante dans les pays anglo-saxons, la seconde guerre mondiale donnant dans ce domaine, et celui de la recherche opérationnelle (organisation et optimisation), une impulsion décisive, dans le cadre de la

production de guerre. Un plan d'ensemble de développement des méthodes statistiques de contrôle de qualité, de cours de statistique industrielle dans les universités et les écoles techniques, fut organisé aux Etats-Unis dans le cadre de l'"*Engineering Science and Management War Training Program*".

En France, ce n'est qu'en 1947 que l'INSEE consacre une brochure à l'exposé des méthodes décrites dans les "standards" de guerre anglo-saxons.

En 1950, *William Edwards Deming* (1900 – 1993), un ancien élève de *Walter Shewhart* ayant développé toute une "philosophie" de la qualité, est invité par les chefs d'entreprise japonais à enseigner ses méthodes au Japon. Son influence y fut très importante et le boom économique du Japon, dont la production était à l'origine d'une qualité médiocre, y doit beaucoup. C'est ainsi qu'au Japon est encore décerné chaque année le "prix *Deming*", récompensant une entreprise ayant excellé dans le domaine de la qualité. *Deming* insista sur la nécessité de l'enseignement des techniques statistiques et sur un usage fréquent de contrôles statistiques de qualité pour faire que chacun, dans l'entreprise, soit sensible aux normes de qualité. Il préconisait également l'embauche dans chaque entreprise d'un expert statisticien capable d'identifier les problèmes, de collecter des données pour aider à trouver une solution, et d'analyser ces données pour inférer des estimations.

La **fiabilité** est le domaine de la statistique inférentielle traitant des durées de vie (ou de bon fonctionnement) des matériels et donc de l'étude statistique de leurs pannes. Le nom de *Wallodi Weibull* (1887 – 1979) y est attaché. D'origine suédoise, *Weibull* travailla comme inventeur (roulements à billes, marteau électrique... ) et ingénieur conseil dans de nombreuses sociétés suédoises ou allemandes, par exemple chez *SAAB*. Il s'intéressa aux

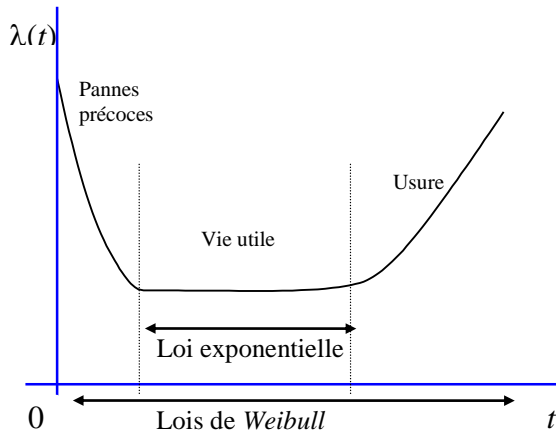
### Quand une méthode statistique est classée "secret défense"

S'il fallait une preuve de l'importance que peut avoir une méthode statistique, son classement "secret défense" en est sans doute une.

Avec l'entrée en guerre en 1941 des Etats-Unis, le statisticien *Abraham Wald* travailla sur des projets militaires, au sein du groupe de recherches statistiques de l'Université Columbia de New York. Ses compétences en statistique lui permirent de développer une méthode d'estimation de la

vulnérabilité des avions. Il y inventa également le concept d'**analyse séquentielle** en réponse à la demande de méthodes plus efficaces de contrôle de qualité dans la production industrielle de guerre. Une procédure d'analyse des données intégrant la dimension temporelle est préférable à celle qui consiste à d'abord collecter toutes les données, puis à les analyser. Dans cette approche, on ne fixe pas a priori la taille de l'échantillon, mais on analyse en temps réel et l'on stoppe l'échantillonnage lorsque les résultats le justifient. Ces procédures séquentielles furent classées "secret défense" et déclassifiées par le gouvernement américain en 1947. *Wald* expose alors la technique des tests statistiques séquentiels.

problèmes de résistance des matériaux, en particulier à ceux de fatigue et de rupture des tubes à vide. C'est dans ce cadre qu'apparaît en 1939 pour la première fois la distribution de *Weibull*. Mais l'article qui eut le plus d'influence fut publié en 1951 dans le "*Journal of Applied Mechanics*" sous le titre "*A Statistical Distribution Function of Wide Applicability*" où sont décrit sept cas d'utilisation de la distribution de *Weibull*. En effet, l'intérêt de cette distribution, outre ses propriétés analytiques satisfaisantes, est de permettre un bon ajustement d'une grande variété de problèmes de durée de vie. On constate expérimentalement, que pour la plupart des matériels, la courbe représentative du taux  $\lambda$  d'avarie (taux de variation du nombre de pannes) en fonction du temps, a la forme d'une "courbe en baignoire".



En période de jeunesse, la fréquence des pannes a tendance à diminuer, alors que pour un matériel vieillissant, elle a tendance à augmenter.

Si l'on désigne par  $T$  la variable aléatoire qui, à tout matériel choisi au hasard, associe son temps de bon fonctionnement avant défaillance, lorsque  $\lambda$  est constant, on montre que  $T$  suit une loi exponentielle, mais lorsque  $\lambda$  varie avec le temps, on cherchera un modèle parmi les lois de *Weibull*.

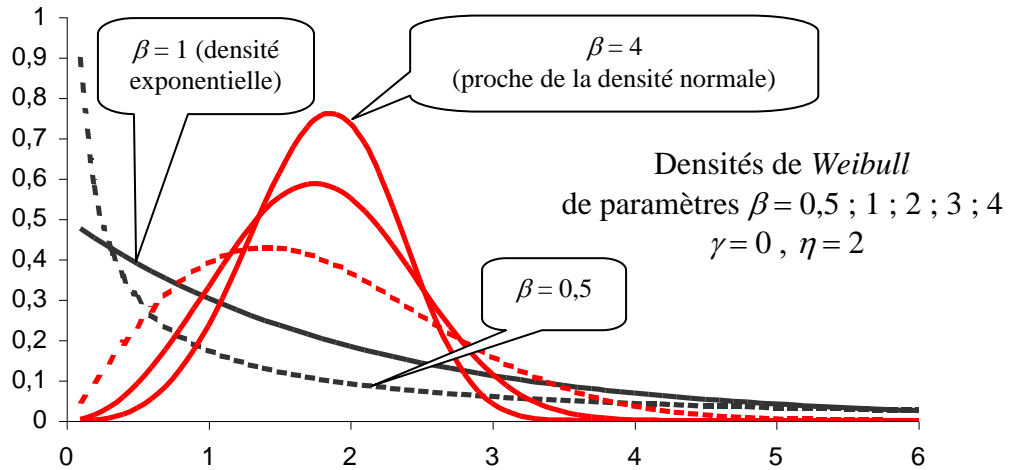
Pour couvrir tous les cas, *Weibull* a

choisi pour  $\lambda$  une fonction dépendant de trois paramètres :  $\gamma$ ;  $\beta$  et  $\eta$  : 
$$\lambda(t) = \frac{\beta}{\eta} \left( \frac{t-\gamma}{\eta} \right)^{\beta-1}$$

avec  $t > \gamma$ ,  $\beta > 0$ ,  $\eta > 0$  (le paramètre important étant  $\beta$ , paramètre "de forme", les autres terminant l'ajustement). Ainsi, lorsque la variable aléatoire  $T$ , correspondant au temps de bon fonctionnement, suit la loi de *Weibull* de paramètres  $\gamma$ ,  $\beta$ ,  $\eta$ , on montre que sa densité

est donnée par 
$$f(t) = \frac{\beta}{\eta} \left( \frac{t-\gamma}{\eta} \right)^{\beta-1} e^{-\left(\frac{t-\gamma}{\eta}\right)^\beta}.$$

Dans le cas de pannes précoces (rodage), avec  $\beta = 0,5$  par exemple, la densité est rapidement décroissante avec le temps. Dans le cas de matériel vieillissant (c'est également valable pour la mortalité humaine...), avec  $\beta = 4$  par exemple, les pannes surviennent selon une densité approximativement normale. Depuis sa création, la distribution de *Weibull* a prouvé sa valeur pour l'analyse des durées de vie dans des domaines aussi variés que l'industrie automobile, aérospatiale, électrique, nucléaire, électronique et médicale. Pour ses travaux, *Weibull* reçut en 1978 du roi de Suède la médaille d'or de l'Académie Suédoise des Sciences de l'Ingénierie.



## 2 – L'analyse des données

Parallèlement à la statistique inférentielle, s'est développée l'analyse des données, branche de la statistique qui s'appuie essentiellement, non pas sur les modèles probabilistes, mais sur les représentations géométriques et dont le développement a été favorisé par celui de l'algèbre linéaire et des moyens de calcul et de représentation informatiques. Bien que constituant une approche complémentaire à celle de la statistique inférentielle, les rapports entre ces deux branches n'ont pas toujours été simples, comme en témoigne l'introduction<sup>31</sup> donnée, en 1973 par Benzécri au tome 2 de son ouvrage "L'analyse des données" :

*"Statistique n'est pas probabilité. Sous le nom de statistique, des auteurs [...] ont édifié une pompeuse discipline, riche en hypothèses qui ne sont jamais satisfaites dans la pratique. Ce n'est pas de ces auteurs qu'il faut attendre la solution de nos problèmes typologiques."*

Les méthodes de l'analyse des données permettent de dégager l'information essentielle d'une grande quantité d'observations portant sur l'étude de plusieurs caractéristiques.

### Une douzaine de définitions de la statistique

Dans un article publié en 1935 dans la "Revue de l'Institut International de la Statistique", Willcox dénombreait 115 définitions de la statistique. En étant plus raisonnable, voici quelques exemples.

1. Dans le style dénigrement (par ignorance ?), celle attribuée à Bismarck : *"Il y a trois formes de mensonge, le mensonge ordinaire, le damné mensonge et la statistique."*
2. Achenwall (1749) : *"La statistique est la connaissance approfondie de la situation respective et comparative des Etats."*  
Définition correspondant à l'origine de la statistique, liée à la notion d'Etat.
3. Sinclair (1785) : *"La statistique a pour but de constater la somme de bonheur dont jouit une population et les moyens de l'augmenter."*  
Dans cette définition, l'activité statistique, au delà du constat, est aussi une aide à l'action.
4. Aftalon (1929) : *"Etude numérique des faits et de leurs rapports."*  
Au delà de la simple collecte de chiffres, cette définition introduit une idée de comparaison.
5. Willcox (1936) : *"Etude numérique des groupes ou des masses, par l'étude des unités qui les composent, que ces unités soient des hommes ou non, des êtres animés ou inanimés."*  
Cette définition insiste sur la notion de collectivité, de population, sur laquelle s'applique la méthode statistique. La statistique traite davantage des populations que des individus.

<sup>31</sup> Cité dans "Histoire de la Statistique" – "Que sais-je ?" – P.U.F. 1997.



Les fondements mathématiques de l'**analyse en composantes principales** sont développés à partir de 1901 par *Karl Pearson*. A partir d'un nuage de points, correspondant aux individus observés, dans l'espace où les axes sont associés aux caractères étudiés, il recherche l'axe du plus grand allongement du nuage.

La première utilisation véritable de cette "analyse factorielle" apparaît en 1904 en *psychométrie* (mesure des capacités intellectuelles) avec **Charles Spearman** (1863 – 1945), disciple de *Karl Pearson*, dans son article, publié dans *l'American Journal of Psychology*, "*General Intelligence Objectively Determined and Measured*" (excusez du peu). Considérant les résultats des tests d'aptitude passés par des enfants, il constate que les réussites aux différents tests sont très corrélées. Considérant, dans l'espace vectoriel des différents tests, le nuage des points constitués par les résultats des enfants, la corrélation se traduit par un allongement sur l'axe principal d'inertie. Cette sorte de moyenne des différents tests permet à *Spearman* d'isoler un "facteur" général unique mesurant l'intelligence, qu'il nomme "*intelligence générale*" ou "*facteur g*". Ce système de tests, selon l'échelle *g*, a fonctionné en Angleterre de 1944 à 1965.

Dans les années 1930, le psychologue américain **Louis Léon Thurstone** (1887 – 1955) décompose *g* en sept "aptitudes mentales primaires" en s'appuyant sur un regroupement des tests en sous ensembles particulièrement bien corrélés.

**L'analyse (factorielle) des correspondances**, dont les principes remontent aux "tables de contingences" de *Fisher* (1940), a été développée vers 1965 par le français **Jean Paul Benzécri**. Le premier exposé sous le nom d'analyse

### Définitions de la statistique

6. *Julin* (1921) : "*Méthode qui, par le relevé en masse et l'expression numérique de ses résultats, arrive à la description des phénomènes collectifs et permet de reconnaître ce qu'ils présentent de permanent et de régulier dans leur variété, comme de variable dans leur apparente uniformité.*"

Tendance moyenne et variabilité sont les deux mamelles de la statistique...

7. *Vessereau* : "*Méthodes de recherches dans lesquelles le grand nombre et l'enchevêtrement des facteurs exigent une technique d'interprétation basée sur la connaissance des lois du hasard.*"

On souligne ici l'appui que prend la statistique mathématique sur les modèles probabilistes.

9. *H. Laurent* (1908) : "*La statistique mathématique a pour but d'indiquer et de rechercher les méthodes pour faire de bonnes observations, lorsqu'il s'agit de faire des évaluations numériques.*"

Cette définition a le mérite de distinguer les méthodes de recueil des observations et celles de l'évaluation, à l'aube de statistique inférentielle.

10. Dictionnaire *Le Robert* (1996) : "*Science et techniques d'interprétation mathématique de données complexes et nombreuses.*"

Définition un peu courte. Par "interprétation de données complexes" il faut sans doute comprendre que la statistique met en évidence l'ordre enfoui dans des données chaotiques.

11. *Encyclopedia Universalis* : "*Le mot statistique désigne à la fois un ensemble de données d'observation et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation.*"

Il semble plutôt que pour les données recueillies, on parle de statistiques (au pluriel). Le singulier correspondant, soit à l'activité scientifique, soit à la variable aléatoire permettant l'estimation.

12. *G. Saporta* (1990) : "*La statistique consiste, après le recueil des données, à traiter et interpréter les informations. Elle comporte deux grands aspects, l'aspect descriptif ou exploratoire (analyse des données) et l'aspect inférentiel ou décisionnel.*"

factorielle des correspondances fut donné par *Benzécri* en hiver 1963. Il s'agit d'une méthode d'exploration des dépendances ("correspondances") entre des caractères qualitatifs, par exemple la dépendance du lieu d'habitation des parisiens (parmi les 80 quartiers) et de leur catégorie socioprofessionnelle.

La détermination des axes factoriels s'effectuant par le calcul des valeurs propres et vecteurs propres d'une matrice (voir chapitre 4), pour 40, 100 ou 1000 variables, ces calculs étaient impensables avant les ordinateurs. L'utilisation de l'informatique a donné une impulsion décisive à la statistique dans ce domaine et l'école française d'analyse des données a très bien saisi la nécessité de refondre la statistique en fonction de l'informatique.

## Bibliographie à propos de l'histoire de la statistique

### ***On pourra essentiellement consulter :***

**DESROSIERES Alain** – *"La politique des grands nombres – Histoire de la raison statistique"* – La découverte / Poche 2000.

**DROESBEKE Jean-Jacques** et **TASSI Philippe** – *"Histoire de la statistique"* – Que sais-je ? n° 2527 – P.U.F. 1997.

### ***Des compléments seront apportés par :***

**BRIAN Eric** – *"La mesure de l'Etat – Administrateurs et géomètres au XVIII<sup>e</sup> siècle"* – Albin Michel 1994.

**CREPEL Pierre** – *"La naissance des mathématiques sociales"* – et **LE BRAS Hervé** – *"L'invention des concepts en démographie"* – **Dossier "Pour la science" : "Les mathématiques sociales" – Hors série juillet 1999.**

**DODGE Yadolah** – *"Statistique – Dictionnaire encyclopédique"* – Dunod 1993.

### **Encyclopédie Universalis**

**LE BRAS Hervé** – *"Naissance de la mortalité – L'origine politique de la statistique et de la démographie"* – Seuil/Gallimard 2000.

### **Sur INTERNET :**

[www-groups.dcs.st-and.ac.uk/~history/BiogIndex.html](http://www-groups.dcs.st-and.ac.uk/~history/BiogIndex.html)

[www.math.uah.edu/stat/biographies/](http://www.math.uah.edu/stat/biographies/)

[www.mrs.umm.edu/~sungurea/introstat/history/](http://www.mrs.umm.edu/~sungurea/introstat/history/)

[www.weibullnews.com/ybullbio.htm](http://www.weibullnews.com/ybullbio.htm)

# 4

## LA STATISTIQUE EUCLIDIENNE

On montre dans ce chapitre l'importance de la géométrie euclidienne pour la "fabrication" et l'interprétation d'indicateurs statistiques classiques, tout d'abord dans le cas de deux variables. On aborde ensuite le domaine plus général de "l'analyse des données" (analyse en composantes principales ou analyse factorielle) pour plus de deux variables.

### I – DES INTERPRETATIONS GEOMETRIQUES

Soit  $X$  l'ensemble dans lequel sont prises les mesures du caractère (ou variable)  $x$  étudié. On note  $x_i$  la valeur prise par le caractère  $x$  pour l'individu  $i$ . Si  $X = \mathbb{R}^k$ ,  $x$  est formé de  $k$  caractères numériques. On dispose ainsi d'un ensemble de  $n$  mesures  $\{x_1, x_2, \dots, x_n\}$  où  $n$  est le nombre d'individus et  $x_i \in \mathbb{R}^k$ .

Dans un premier temps, on s'intéresse à des situations où  $X = \mathbb{R}$  ou  $X = \mathbb{R}^2$  (c'est à dire, un ou deux caractères étudiés) et où l'on cherche à caractériser :

- la ou les valeurs centrales ;
- la ou les valeurs de dispersion ;
- une liaison à peu près linéaire (dans le cas  $X = \mathbb{R}^2$ ).

On fait une "pétition de principe" : on suppose que **l'espace euclidien est adapté à la représentation des données**<sup>32</sup>. Cette pétition de principe ne peut être justifiée qu'*a posteriori*, c'est impossible *a priori*.

On suppose dans la suite que l'on étudie deux caractères  $x$  et  $y$ , c'est à dire que  $k = 2$ .

Géométriquement, dans le cas de données bivariées ( $k = 2$ ), on possède les observations :

$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$ , avec ici  $x_i \in \mathbb{R}$  et  $y_i \in \mathbb{R}$ , sur  $n$  individus.

On peut représenter ces données de deux façons distinctes, **duales** l'une de l'autre :

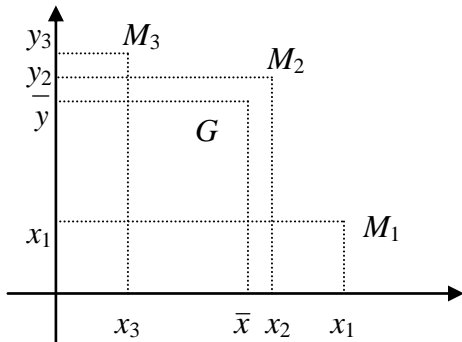
- soit  $n$  points dans un espace à  $k = 2$  dimensions (**espace des individus**) ;
- soit  $k = 2$  points dans un espace à  $n$  dimensions (**espace des variables**).

---

<sup>32</sup> On adoptera ici la distance euclidienne usuelle, définie par  $\delta(M_1, M_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$  dans le cas de deux variables  $x$  et  $y$ , ce qui donne beaucoup de poids aux valeurs aberrantes. Pour la distance non euclidienne définie par  $\delta(M_1, M_2) = |x_1 - x_2| + |y_1 - y_2|$ , on obtiendrait la médiane  $Mé$  comme tendance centrale associée, et  $\frac{1}{n} \sum |x_i - Mé|$  comme mesure de dispersion.

Par ailleurs, avec la distance euclidienne usuelle, chaque dimension est de même nature. Dans l'espace des individus, où chaque dimension correspond à une variable (qui peuvent être l'âge, la taille, le poids...), s'exprimant avec une unité particulière, on peut être amené à choisir une autre métrique, par exemple pondérée :  $\delta^2(M_1, M_2) = a(x_1 - x_2)^2 + b(y_1 - y_2)^2$ , avec  $a \geq 0$  et  $b \geq 0$  ou toute autre métrique liée à une forme quadratique.

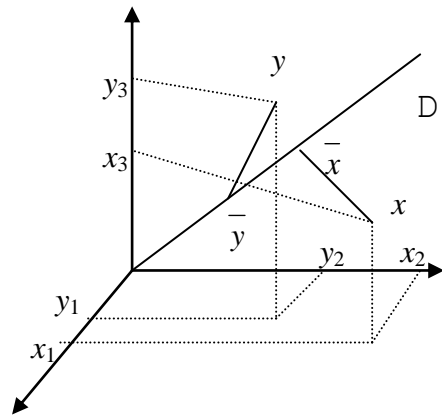
Pour faire les figures nous avons supposé que  $n = 3$  :



### Espace des individus

Axes : les  $k = 2$  variables  $x$  et  $y$

Points : les  $n = 3$  individus



### Espace des variables

Axes : les  $n = 3$  individus

Points : les  $k = 2$  variables  $x$  et  $y$

L'intérêt de ces représentations, outre l'interprétation géométrique des résumés classiques, est d'être généralisables au cas multivarié où, sur un même individu, on effectue  $k$  mesures (avec  $k > 2$ ), cas que l'on envisagera par la suite. On aura alors, dans l'espace des individus,  $n$  points dans un espace à  $k$  dimensions et, dans l'espace des variables,  $k$  points dans un espace à  $n$  dimensions (ceci est étudié dans les classes des lycées préparant au DECF).

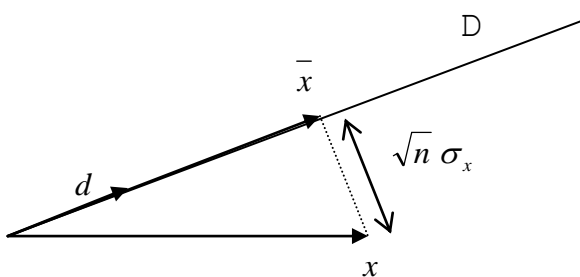
Revenons au cas bivarié, où  $k = 2$  (les variables sont notées  $x$  et  $y$ ).

L'individu  $i$  est représenté, dans l'espace des individus  $\mathbb{R}^2$ , par le point de coordonnées  $(x_i, y_i)$ , la variable  $x$  est représentée, dans l'espace des variables  $\mathbb{R}^n$ , par le point de coordonnées  $(x_1, x_2, \dots, x_n)$ . Munissons  $\mathbb{R}^2$  et  $\mathbb{R}^n$  de la structure euclidienne usuelle.

Soit, dans l'espace des variables, la droite  $D$  de vecteur directeur  $d$  de coordonnées  $(1, \dots, 1)$ .

Dans  $\mathbb{R}^n$ , la droite  $D$  représente les variables qui prennent la même valeur sur tous les individus.

Comme on recherche un indicateur de centralité, notons que si tous les individus sont identiques, l'indicateur de centralité sera la valeur commune. La droite  $D$  menée par l'origine et de vecteur directeur  $d(1, \dots, 1)$ , est le lieu des points dans l'espace des variables dont toutes les coordonnées sont égales. La variabilité inhérente à la situation statistique fait que le point  $x$ , correspondant aux valeurs de la première variable sur les différents individus, n'est pas sur  $D$ . Si la structure euclidienne usuelle est adéquate, le meilleur **indicateur de centralité** est le point de la droite  $D$  le plus proche du point  $x$ , c'est-à-dire la **projection orthogonale** de  $x$  sur la droite  $D$ .



On a :

$$\text{proj}_d(x) = \|x\| \cos(x, d) \frac{d}{\|d\|},$$

ou encore :

$$\text{proj}_d(x) = \frac{x \cdot d}{\|d\|^2} d$$

$$\text{soit } \text{proj}_d(x) = \frac{\sum x_i}{n} d.$$

Ainsi, le meilleur indicateur de centralité est le point dont toutes les coordonnées sont égales à  $\bar{x} = \frac{1}{n} \sum x_i$ , la moyenne. On notera encore  $\bar{x}$  le point dont toutes les coordonnées sont égales à  $\bar{x}$ .

**On voit que, comme indicateur de centralité, la moyenne est fortement liée à la structure euclidienne.**

De même  $y$  est projeté en  $\bar{y}$  sur la droite  $D$ .

Le caractère  $x$  est d'autant plus dispersé que le point  $x$ , le représentant dans l'espace des variables, est éloigné de la droite  $D$ . La **distance** du point  $x$  au point  $\bar{x}$  est un **indicateur de la dispersion**. Cette distance est  $\sqrt{\sum (x_i - \bar{x})^2}$ ; c'est l'**écart type**, au facteur  $\sqrt{n}$  près, fait pour normaliser, afin de pouvoir comparer des populations d'effectifs différents. L'écart type  $\sigma_x$  vérifie alors :  $n(\sigma_x)^2 = d^2(x, \bar{x})$ .

**L'indicateur de dispersion, écart type, est lui aussi très lié à la structure euclidienne.**

Pour des données bivariées, si le nuage de points, dans l'espace à deux dimensions des individus, est aplati autour d'une droite, alors, dans l'espace des variables, les vecteurs  $(x - \bar{x})$  et  $(y - \bar{y})$  sont pratiquement colinéaires et doivent former un angle très faible ou à peu près plat. En revanche, un nuage "rond", dans l'espace des individus, ferait que ces vecteurs sont orthogonaux.

Une application du **produit scalaire** permet de retrouver l'expression du **coefficient de corrélation linéaire**  $\rho$  entre les variables  $x$  et  $y$ .

On considère l'angle  $\varphi$  des deux vecteurs centrés  $(x - \bar{x})$  et  $(y - \bar{y})$ . Déterminons le cosinus de cet angle.

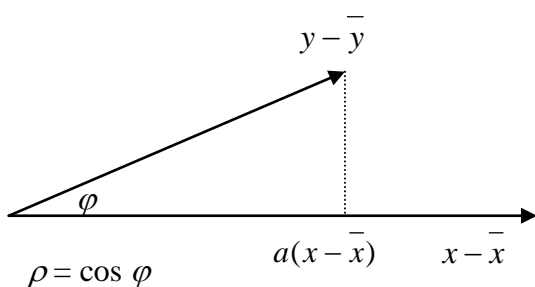
$$\text{On a } \cos \varphi = \cos(x - \bar{x}, y - \bar{y}) = \frac{(x - \bar{x}) \cdot (y - \bar{y})}{\|x - \bar{x}\| \times \|y - \bar{y}\|},$$

$$\text{c'est à dire } \cos \varphi = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{n}\sigma_x \times \sqrt{n}\sigma_y} = \frac{\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \times \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y},$$

donc  $\cos \varphi = \rho$ .

**Ainsi, le coefficient de corrélation  $\rho$  est égal au cosinus de l'angle des deux vecteurs  $(x - \bar{x})$  et  $(y - \bar{y})$ , ce qui indique la liaison entre ce coefficient et la structure euclidienne de l'espace.**

Donc, pour des données bivariées, si le nuage de points dans l'espace des individus est proche d'une droite,  $|\rho|$  est voisin de 1 et l'angle de  $(x - \bar{x})$  et  $(y - \bar{y})$  est petit ou à peu près plat.



La **droite de régression** de  $y$  en  $x$  selon les **moindres carrés** apparaît alors comme le meilleur ajustement euclidien de  $(y - \bar{y})$  selon  $(x - \bar{x})$ , c'est à dire fourni par la **projection orthogonale** du vecteur  $(y - \bar{y})$  sur  $(x - \bar{x})$ .

En effet :

$$\text{proj}_{(x - \bar{x})}(y - \bar{y}) = \|y - \bar{y}\| \cos \varphi \frac{x - \bar{x}}{\|x - \bar{x}\|},$$

c'est à dire  $\text{proj}_{(x-\bar{x})}(y-\bar{y}) = \sqrt{n}\sigma_y \times \rho \times \frac{1}{\sqrt{n}\sigma_x} (x-\bar{x})$ ,

soit  $\text{proj}_{(x-\bar{x})}(y-\bar{y}) = \sigma_y \times \frac{\sigma_{xy}}{\sigma_x \sigma_y} \times \frac{1}{\sigma_x} (x-\bar{x}) = \frac{\sigma_{xy}}{\sigma_x^2} (x-\bar{x})$ .

On reconnaît là l'expression du coefficient  $a$  de la droite de régression de  $y$  en  $x$  selon la méthode des moindres carrés, c'est à dire  $a = \frac{\sigma_{xy}}{\sigma_x^2}$ . Si la droite de régression des moindres

carrés a pour équation  $y = ax + b$  alors on a donc  $\text{proj}_{x-\bar{x}}(y-\bar{y}) = a(x-\bar{x})$ .

**Ainsi, le coefficient de régression linéaire  $a$ , de  $y$  en  $x$ , apparaît comme un rapport entre la projection de  $(y-\bar{y})$  sur  $(x-\bar{x})$  et  $(x-\bar{x})$ .**

Ces remarques montrent l'intérêt de la vision géométrique pour la compréhension statistique de ces indicateurs. Avoir ces schémas de géométrie à l'esprit est susceptible d'aider à l'interprétation : ainsi, un coefficient de corrélation de 0,8 ne peut que conduire à penser (par le cosinus) à un angle assez faible. De même, un contresens trop classique est révélé : si on a calculé la régression de  $y$  en  $x$  par la méthode des moindres carrés en minimisant la somme des carrés des résidus<sup>33</sup>, il ne peut pas être question de pouvoir "prévoir" une valeur de  $x$  à partir d'une valeur de  $y$ . En effet, la droite de régression de  $y$  en  $x$  est distincte de la droite de régression  $x$  en  $y$ , puisque dans ce cas on considère la projection de  $(x-\bar{x})$  sur  $(y-\bar{y})$ .

Cette image géométrique est destinée à favoriser des interprétations statistiques plus correctes.

## II – APRES LE LYCEE... L'ANALYSE DE DONNEES MULTIDIMENSIONNELLE

Pour une même population, sur  $n$  individus de cette population, on mesure désormais  $k$  caractères ( $k$  éventuellement grand et non seulement 2 comme précédemment).

Par **exemple**, lorsqu'on s'intéresse aux notes obtenues à l'écrit par des candidats au bac C<sup>34</sup>, chaque individu a un vecteur de notes : la note de philo, la note de math, la note de physique, celle de sciences naturelles... Il s'agit de comprendre la structure globale de cette collection de notes.

On dispose d'une matrice  $(n, k)$ . Soit  $X$  cette matrice où  $x_{ij}$  est la note de l'élève  $i$  dans la discipline  $j$ .

$$\begin{array}{c} \text{individus} \uparrow \\ \begin{array}{c} 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} \left[ \begin{array}{ccc} 1 & j & k \\ \vdots & \vdots & \vdots \\ \dots & \dots & x_{ij} \\ \vdots & \vdots & \vdots \end{array} \right] \longleftarrow \text{variables} \end{array}$$

Avec un grand nombre d'individus, c'est parfaitement illisible. Il faut essayer de transformer, de résumer.

<sup>33</sup> On cherche une droite d'équation  $y = ax + b$  ; en fait  $y_i = ax_i + b + c_i$  où  $c_i$  est un résidu de sorte que la somme des carrés des résidus de 1 à  $n$  soit minimale.

<sup>34</sup> Etude statistique effectuée dans les années 1980.

On veut chercher à préciser ce qui est le plus important dans la réussite au baccalauréat. Est-ce la réussite de certains élèves dans les matières littéraires comme on l'a prétendu tant de fois ?

On commence, en général, par regarder ce que donne chaque discipline : moyenne, écart type pour chaque variable (chaque discipline). C'est ainsi qu'on peut constater que la plus forte dispersion des notes de mathématiques leur fait jouer un rôle plus fort (la philosophie voit ses notes beaucoup plus centrées). L'importance d'ensemble du rôle joué par une discipline ne dépend pas seulement de son coefficient, mais aussi de l'écart type des notes attribuées. Le fait que la discipline n'a pas de rôle est manifeste si tous les candidats ont la même note (la moyenne).

On peut en outre calculer des coefficients de corrélation entre deux variables ... Mais on voudrait aller plus loin. L'étude séparée des variables laisse de côté les liaisons qui peuvent exister entre elles. Il faut donc analyser les données en tenant compte de leur caractère multidimensionnel. L'**analyse en composantes principales** (ou analyse factorielle) est une méthode puissante pour en explorer la structure.

Reprenons le modèle euclidien usuel<sup>35</sup>, en généralisant ce qui a été fait précédemment.

Les données correspondent à la matrice  $X = (x_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq k}}$ .

On peut représenter ces données comme  $n$  points (lignes de  $X$ )  $M_i$  dans un espace euclidien à  $k$  dimensions (**espace des individus**), puis comme  $k$  points (colonnes de  $X$ ) dans un espace euclidien à  $n$  dimensions (**espace des variables**).

En  $k$  dimensions, c'est bien peu lisible, mais une dimension, deux dimensions seraient plus commodes à lire. De là une question : **quel est le sous-espace à une, deux, trois dimensions le plus proche du nuage de points dans l'espace des individus ?**

En termes euclidiens, ces sous-espaces passent nécessairement par le **centre de gravité** (ou point moyen) du nuage. Le centre de gravité  $G$  a pour coordonnées, dans l'espace des

individus, les moyennes de chacune des  $k$  variables sur les  $n$  individus :  $G \begin{pmatrix} \overline{x_1} \\ \vdots \\ \overline{x_k} \end{pmatrix}$  avec

$\overline{x_j} = \frac{1}{n} \sum_{i=1}^n x_{ij}$ . On peut également écrire  $G = \frac{1}{n} {}^t X d$ , où  ${}^t$  désigne la transposée et  $d \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ .

Puisque les sous espaces recherchés passent par  $G$ , on peut centrer les données et poser dorénavant  $X = (x_{ij} - \overline{x_j})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq k}}$

On considérera la **matrice des variances-covariances** définie par  $V = {}^t X X$  où  $X$  est la matrice des données centrées.

<sup>35</sup> On suppose donc ici que toutes les variables de même nature (par exemple les notes dans les différentes matières du baccalauréat).

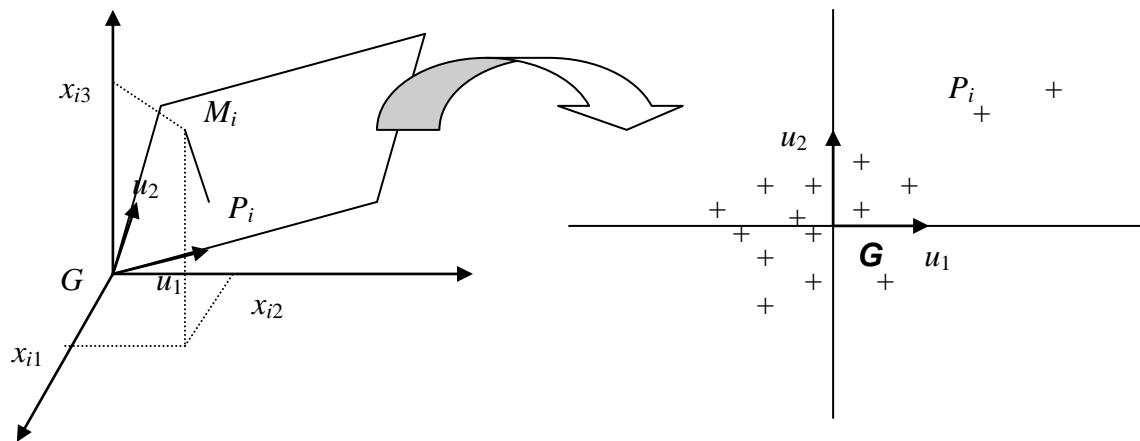
La matrice  $V$  est donc de dimensions  $k \times k$  avec  $V = \begin{bmatrix} \sigma_1^2 & & & \\ & \ddots & & \\ & & \sigma_{ij} & \\ & & & \ddots \\ & & & & \sigma_k^2 \end{bmatrix}$  et

$\sigma_{ij} = \frac{1}{n} \sum_{l=1}^n (x_{li} - \bar{x}_i)(x_{lj} - \bar{x}_j)$  pour  $1 \leq i \leq k$  et  $1 \leq j \leq k$ , correspondant à la covariance des variables  $x_i$  et  $x_j$ .

Soit à rechercher le plan, passant par  $G$  et dirigé par les vecteurs  $u_1$  et  $u_2$ , le plus proche du nuage de points, dans l'espace des individus.

Chaque individu  $M_i$  est projeté en  $P_i$  sur ce plan :  $P_i = \text{proj}_{(u_1, u_2)}(M_i)$ .

De la sorte, en deux dimensions, on va pouvoir regarder, obtenir de la lisibilité, en balance avec un peu de perte d'information... naturellement<sup>36</sup>.



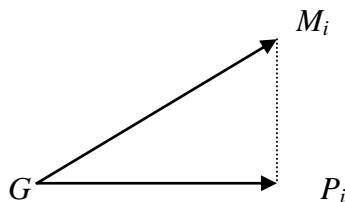
Espace des individus (avec  $k = 3$ )

Projection du nuage de point dans le sous-espace de dimension 2, mené par  $G$  et dirigé par  $u_1$  et  $u_2$

Dans l'espace des variables, une étude duale est possible : rechercher le sous-espace de dimension 1, 2 ou 3 approchant au mieux le nuage des variables.

Revenons à l'espace des individus. Perdre le moins d'information possible s'interprète géométriquement par une déformation minimale du nuage par projection. Ceci peut se traduire à l'aide de la notion d'**inertie** totale du nuage. Cette inertie  $I$  est définie comme la

moyenne des carrés des distances au centre de gravité :  $I = \frac{1}{n} \sum_{i=1}^n GM_i^2$ .



En effet, d'après le théorème de *Pythagore*,

$$GM_i^2 = GP_i^2 + P_iM_i^2$$

et minimiser les carrés des écarts  $P_iM_i^2$  (moindres carrés) équivaut à maximiser les  $GP_i^2$  (maximum d'inertie).

<sup>36</sup> En permanence, le statisticien est devant un dilemme : réduire les données, les ordonner, les classer en vue de mettre en évidence la structure du phénomène ou bien perdre le moins d'information possible et, donc, classifier et réduire le moins possible car toute transformation induit une perte d'information. La tâche du statisticien est de représenter les données avec un minimum de perte d'information... et un maximum d'explication.



L'inertie peut s'exprimer à l'aide de la matrice  $V$ .

$$\text{On a en effet } I = \frac{1}{n} \sum_{i=1}^n GM_i^2 = \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^k (x_{ij} - \bar{x}_i)^2 \right),$$

$$\text{ou encore } I = \sum_{j=1}^k \left( \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_i)^2 \right) = \sum_{j=1}^k \sigma_j^2 \quad \text{c'est à dire } I = \text{Tr } V.$$

**L'inertie est ici égale à la somme des variances des  $k$  variables**, c'est à dire à la trace de la matrice  $V = {}^tXX$  où  $X$  est la matrice des données centrées.

Algébriquement, il s'agira de **diagonaliser** la matrice symétrique  $V = {}^tXX$ , donc de trouver ses valeurs propres  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$ . La trace étant invariante par diagonalisation, l'inertie globale  $\sum GM_i^2$  est la somme des valeurs propre  $\lambda_1 + \lambda_2 + \dots + \lambda_k$ . On montre (voir l'encadré suivant, que l'on peut omettre en première lecture) que les vecteurs  $u_1$  et  $u_2$  cherchés sont les **vecteurs propres associés** respectivement **aux deux plus grandes valeurs propres** de  $V$ <sup>37</sup>.

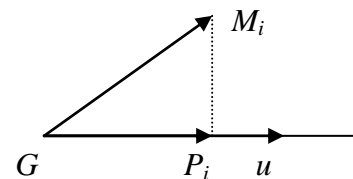
### Détermination des axes principaux d'inertie

On cherche d'abord, dans l'espace des individus, le sous-espace de dimension 1 d'inertie maximale. Il s'agit de la droite passant par  $G$  et telle que l'inertie du nuage projeté sur cette droite est maximale.

Examinons l'expression de l'opérateur de projection sur la droite passant par  $G$  et dirigée par  $u$

$$\text{On a } \overrightarrow{GP_i} = \frac{\overrightarrow{GM_i} \cdot u}{\|u\|} \frac{u}{\|u\|},$$

$$\text{ou encore } \overrightarrow{GP_i} = \frac{{}^t u \overrightarrow{GM_i}}{{}^t u u} u = \frac{u}{{}^t u u} {}^t u \overrightarrow{GM_i}.$$



On notera donc  $P = \frac{1}{{}^t u u} u {}^t u$  l'opérateur de projection ( ${}^t u u$  est un scalaire mais  $u {}^t u$  est une matrice). Les données du nuage projeté correspondent alors à la matrice  $X {}^t P$  dont la ligne

$$i \text{ contient les coordonnées du point projeté } P_i, \text{ en effet } (x_{i1}, \dots, x_{ik}) {}^t P = \begin{pmatrix} P \\ \vdots \\ P \end{pmatrix} \begin{pmatrix} x_{i1} \\ \vdots \\ x_{in} \end{pmatrix}.$$

L'inertie du nuage projeté est donc  $\text{Tr}({}^t(X {}^t P) X {}^t P) = \text{Tr}(P {}^t X X {}^t P)$

ou encore  $\text{Tr}(P {}^t X X P)$  car  ${}^t P = P$ ,

ou encore  $\text{Tr}({}^t X X P P)$  car  $\text{Tr}(AB) = \text{Tr}(BA)$ ,

ou enfin  $\text{Tr}({}^t X X P)$  car  $P^2 = P$ .

<sup>37</sup> Dans les études statistiques, les valeurs propres sont distinctes du fait même de leur provenance de situations réelles ; il faut bien voir que le cas de matrices à valeurs propres, multiples, est fortement improbable dans la réalité statistique.

On déduit de ce qui précède que l'inertie du nuage projeté est donnée par :

$$\text{Tr}\left({}^tXX \frac{1}{{}^tuu} u^t u\right) = \frac{1}{{}^tuu} \text{Tr}({}^tXXu^t u) \quad \text{car } {}^tuu \text{ est un scalaire,}$$

ou encore  $\frac{1}{{}^tuu} \text{Tr}({}^t u^t XXu) = \frac{{}^t(Xu)Xu}{{}^tuu}$  car  $\text{Tr}(AB) = \text{Tr}(BA)$  et que  ${}^t(Xu)Xu$  est un scalaire.

On a donc obtenu que l'inertie du nuage projeté est  $\frac{{}^t uVu}{{}^tuu}$  avec  $V = {}^tXX$ .

Une condition nécessaire sur  $u$  pour que cette inertie soit maximale sera obtenue en annulant la différentielle de la fonction définie de  $\mathbb{R}^k$  dans  $\mathbb{R}$  par  $u \mapsto \frac{{}^t uVu}{{}^tuu}$ .

Examinons d'abord la différentielle de la fonction définie de  $\mathbb{R}^k$  dans  $\mathbb{R}$  par  $u \mapsto {}^tuu = \|u\|^2$ .

Par définition, la différentielle de cette fonction en  $u$  est l'application linéaire  $L$  de  $\mathbb{R}^k$  dans  $\mathbb{R}$  telle que  $\lim_{\|v\| \rightarrow 0} \frac{1}{\|v\|} (\|u+v\|^2 - \|u\|^2 - L(v)) = 0$ .

On constate que  $L(v) = 2u.v$ , qui est linéaire en  $v$ , convient, car

$$\lim_{\|v\| \rightarrow 0} \frac{1}{\|v\|} (\|u+v\|^2 - \|u\|^2 - 2u.v) = \lim_{\|v\| \rightarrow 0} \frac{1}{\|v\|} \|v\|^2 = 0.$$

De même, la différentielle en  $u$  de la fonction définie de  $\mathbb{R}^k$  dans  $\mathbb{R}$  par  $u \mapsto {}^t uVu = u.Vu$  est donnée par  $L(v) = 2Vu.v$  car :

$$\lim_{\|v\| \rightarrow 0} \frac{1}{\|v\|} ((u+v).V(u+v) - u.Vu - 2Vu.v) = \lim_{\|v\| \rightarrow 0} \frac{1}{\|v\|} (u.Vv - Vu.v + v.Vv) = \lim_{\|v\| \rightarrow 0} \frac{1}{\|v\|} v.Vv = 0,$$

en effet,  $V = {}^tXX$  étant symétrique, on a  $u.Vv = {}^t v^t Vu = {}^t vVu = v.Vu$ .

On déduit de ce qui précède que  $\frac{d}{du} \left( \frac{{}^t uVu}{{}^tuu} \right) = \frac{({}^t uu)2Vu - ({}^t uVu)2u}{{}^t uu)^2}$ ,

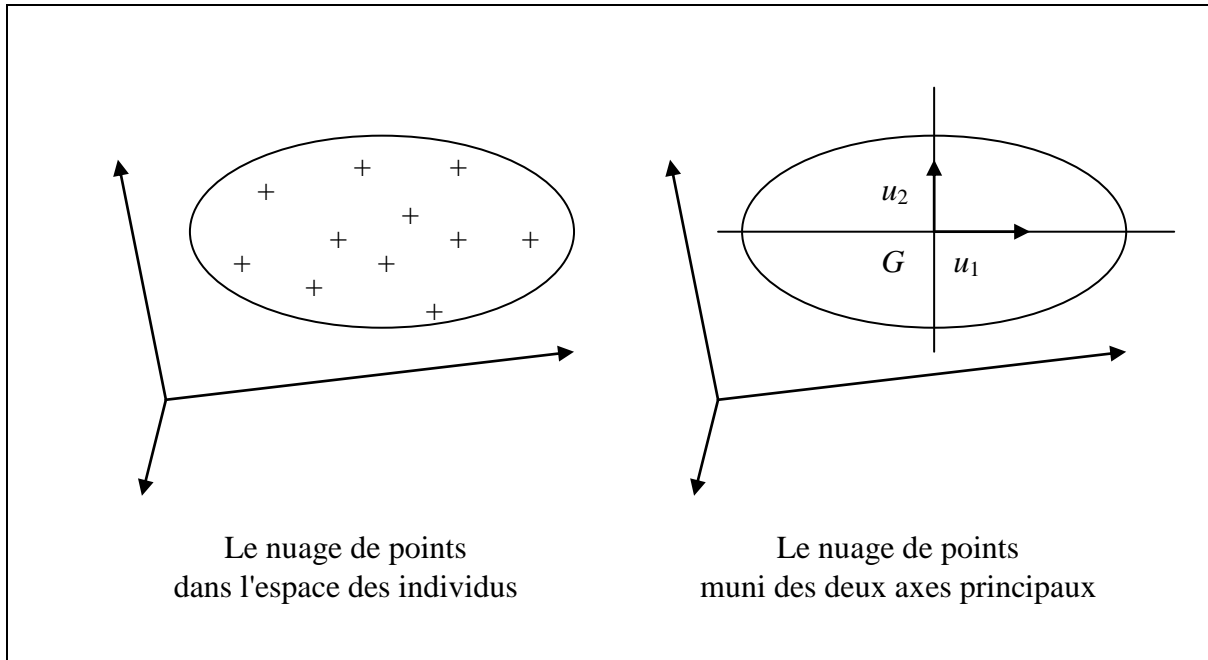
qui s'annule lorsque  $({}^t uu)Vu = ({}^t uVu)u$  c'est à dire  $Vu = \frac{{}^t uVu}{{}^tuu} u$ .

Or  $\frac{{}^t uVu}{{}^tuu}$  est un scalaire (et c'est l'inertie du nuage projeté). Une condition nécessaire sur  $u$

est donc qu'il soit vecteur propre de  $V$  pour la valeur propre  $\lambda = \frac{{}^t uVu}{{}^tuu}$ .

Cette valeur étant l'inertie du nuage projeté, il suffit de prendre pour  $u$  le vecteur de propre  $u_1$  de  $V$  correspondant à sa plus grande valeur propre  $\lambda_1$ .

La matrice  $V$  étant symétrique, elle possède des vecteurs propres deux à deux orthogonaux. On peut alors réitérer le même raisonnement, en considérant l'espace supplémentaire orthogonal à la droite  $(G, u_1)$ , recherchant dans ce supplémentaire, la droite  $(G, u_2)$  sur laquelle l'inertie du nuage projeté est maximale. On trouvera que  $u_2$  est le vecteur propre de  $V$  associé à la plus grande valeur propre  $\lambda_2$  restante.



Les logiciels actuellement disponibles font directement les calculs de valeurs propres et vecteurs propres, et produisent la représentation graphique du plan  $(u_1, u_2)$  et des points projetés : il s'agit alors de rendre compte qualitativement de ce résultat : comment peut-on interpréter ?

La capacité à **lire une analyse factorielle** est très importante pour débrouiller des données multivariées.

Cette présentation est-elle **pertinente** ?

Un des premiers critères de pertinence est de regarder quel est le pourcentage d'inertie expliquée.

La variabilité globale, l'inertie globale  $\sum GM_i^2$ , est la somme des valeurs propres  $\lambda_1 + \lambda_2 + \dots + \lambda_k$ , elle est à comparer à l'inertie du nuage projeté  $\sum GP_i^2$  qui vaut  $\lambda_1 + \lambda_2$ . Le rapport  $\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \dots + \lambda_k}$  rend compte de l'inertie expliquée par la

représentation. Proche de 0,9 ce rapport signale plutôt un nuage peu dispersé autour du plan  $(u_1, u_2)$ . En trois dimensions il serait en forme de soucoupe volante. S'il était en forme de fuseau,  $\lambda_1$  serait nettement plus grand que  $\lambda_2$ .

Quels sont, par ailleurs, les points qui sont bien représentés ? Les points bien représentés sont tels que l'angle entre le vecteur  $\overline{GM_i}$  et le plan de projection soit petit. Il est vraisemblable que les points les plus éloignés de  $G$  sont les mieux représentés ; les points les plus proches de  $G$  ont toutes chances de ne pas l'être. Encore une fois, l'apprentissage de la lecture d'une représentation graphique dépend de la capacité à se situer vis-à-vis de la représentation géométrique dans l'espace.

Dans l'**exemple des résultats au baccalauréat**, le premier axe a une interprétation évidente : il est expliqué par le niveau général des élèves (pas besoin d'un gros outil statistique pour mettre en évidence cela). En revanche, pour le deuxième axe, c'est l'opposition entre réussite en mathématiques et réussite en sciences expérimentales. Un troisième axe met en évidence l'influence de l'opposition entre réussite en sciences physiques et réussite en sciences naturelles. En revanche, l'opposition littéraire-scientifique, ou la réussite dans les disciplines littéraires, n'entre pratiquement pas dans les

facteurs explicatifs de la réussite au baccalauréat C. Statistiquement, il était erroné de donner crédit à l'idée selon laquelle on pouvait réussir le bac C avec les matières littéraires.

Il est possible que la représentation euclidienne ne soit pas pertinente :

- soit qu'on soit dans l'impossibilité d'interpréter les axes,
- soit que la forme du nuage ne se prête pas au traitement, par exemple un nuage en forme de "banane"...

### III – DES GENERALISATIONS

Dans l'exemple des notes du bac, les mesures sont homogènes en ce sens qu'il s'agit de la même grandeur pour chaque caractère (chaque note). Il n'en irait pas de même si on traitait, par exemple, de la longueur du pied, du poids, de la taille : ces **mesures** sont **hétérogènes**. Pour pouvoir faire une analyse factorielle, on centre et on réduit chacune des mesures : on retranche  $\bar{x}_j$  à  $x_{ij}$  et on divise par l'écart type  $\sigma_j$ . Ceci fait que tous les individus étudiés

appartiennent à une hypersphère de rayon  $\sqrt{n}$  : 
$$\sum_i \frac{(x_{ij} - \bar{x})^2}{\sigma_j^2} = n.$$

C'est à dire que toutes les variables interviennent de la même façon, sans tenir compte de l'échelle dans lesquelles elles sont mesurées.

Une autre généralisation plus fréquente a lieu dans le cas où on souhaite **comparer deux classements**, par exemple, les performances dans une discipline donnée (avec  $k$  classes) et les résultats à des tests psycho-génétiques selon  $l$  classes que nous nommerons modalités, pour plus de clarté.

On a la matrice de données suivante :

$$\begin{array}{c}
 \begin{array}{cccc}
 & 1 & j & k \\
 \begin{array}{c} 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} & \left[ \begin{array}{ccc} & & \\ & \vdots & \\ & \vdots & \\ \cdots & \cdots & n_{ij} \\ & & \end{array} \right] & \longleftarrow \text{classes (selon les} \\
 \begin{array}{c} \uparrow \\ \text{modalités} \\ \text{(selon les notes} \\ \text{aux tests)} \end{array} & & & \text{notes dans une discipline)}
 \end{array}
 \end{array}$$

où  $n_{ij}$  est le nombre d'individus qui appartiennent à la classe  $i$  pour le premier caractère (performance dans la discipline) et à la modalité  $j$  pour le deuxième caractère (résultats aux tests). Il s'agit d'étudier la structure interne de cette observation.

On introduit les fréquences  $f_{ij} = \frac{n_{ij}}{N}$  où  $N$  est l'effectif total, proportion d'apparition du couple  $(i, j)$  dans l'ensemble de tous les couples apparus, puis les fréquences marginales  $f_{i\cdot} = \sum_j f_{ij}$  et  $f_{\cdot j} = \sum_i f_{ij}$  qui permettent d'obtenir le tableau des fréquences et ses marges suivant.

$$\begin{array}{c}
 \begin{array}{ccc}
 & 1 & j & k \\
 \begin{array}{c} 1 \\ \vdots \\ i \\ n \end{array} & \left[ \begin{array}{ccc}
 & & \\
 & \vdots & \\
 \cdots & \cdots & f_{ij} \\
 & & 
 \end{array} \right] & \begin{array}{c} \\ \\ f_{i\bullet} \\ \\ \end{array} \\
 & & & f_{\bullet j}
 \end{array}
 \end{array}$$

Ainsi on peut comparer des profils : par exemple, deux colonnes ou deux lignes en calculant la contribution relative  $\frac{f_{ij}}{f_{i\bullet}}$  du couple  $(i, j)$  à l'élément  $i$  ou celle  $\frac{f_{ij}}{f_{\bullet j}}$  du couple  $(i, j)$  à l'élément  $j$ .

Pour reprendre le traitement géométrique précédent, il faut **définir une distance** entre deux modalités .

On adopte en général :  $\delta(i, i') = \sqrt{\sum_{j=1}^n \frac{1}{f_{\bullet j}} \left( \frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2}$ , dite distance du khi-deux<sup>38</sup>.

Pour cette distance, deux classes  $j$  ayant le même profil étant fusionnées, la distance  $\delta(i, i')$  ne change pas. Cette propriété, dite de l'équivalence distributionnelle, est utile car elle permet de remédier à l'arbitraire toujours possible des classifications. De là une possibilité de réexploiter la géométrie euclidienne, afin de trouver de même des axes factoriels permettant d'interpréter des proximités entre modalités des deux classements. C'est ce qu'on appelle l'**analyse des correspondances**.

## CONCLUSION

Pourquoi fait-on de la **géométrie dans l'espace** dans les sections de ES ? Ce n'est pas la capacité technique à effectuer les calculs qui est recherchée : ils sont très bien effectués par les logiciels. On fait de la géométrie dans l'espace pour donner aux élèves une **vision** les rendant capables d'interpréter correctement des résultats alors que l'exécution des calculs est désormais l'enfance de l'art. Ces lycéens vont rencontrer de telles démarches dans diverses voies d'études : économie, sociologie, histoire... Il faut qu'ils soient capables de comprendre de quoi il s'agit. C'est donc comme instrument d'illustration des concepts statistiques que l'apprentissage de la géométrie importe.

Il est important aussi comme constitution d'images mentales qui aideront à apprendre l'algèbre linéaire dont on sait qu'elle est fortement présente dans les études économiques. Ce qu'il est important que les élèves maîtrisent est finalement relativement élémentaire : les notions de points, de droite, de plan, d'intersection, de parallélisme, d'orthogonalité, de projection. Cela ne peut pas être appris d'un coup en Première ou Terminale : un apprentissage continu et conséquent de la géométrie dans l'espace depuis l'école et particulièrement en classe de Seconde constitue aussi la base de cette formation.

<sup>38</sup> distance qui est euclidienne, sinon la géométrie de représentation est beaucoup plus confuse.

## Bibliographie à propos de la statistique euclidienne

**AVENEL Michèle** – *"DECF – Mathématiques appliquées"* – FOUCHER.

**BRY Xavier** – *"Analyses factorielles simples"* – ECONOMICA 1995.

Présentation intuitive de la technique statistique avec, en encarté, le traitement mathématique.

**FOUCART Thierry** – *"L'analyse des données, mode d'emploi"* – PRESSES UNIVERSITAIRES DE RENNES 1997.

A partir de nombreux exemples, avec leurs sorties informatiques, on procède à des interprétations. Les parties mathématiques sont réduites.

**SAPORTA Gilbert** – *"Probabilités, analyse des données et statistique"* – TECHNIP 1990.

Ouvrage de référence pour la statistique inductive comme pour l'analyse des données. Les modèles mathématiques sont explicités, de nombreux exemples sont donnés.



# LA STATISTIQUE INFÉRENTIELLE

## I – LE MODELE DE LA DECISION STATISTIQUE

La **statistique inductive**, ou **statistique inférentielle**, a comme objectif de préciser une loi de probabilité inconnue sur une population, à partir de l'observation de tirages aléatoires. Cette idée est présente dès la fin du XVIII<sup>ème</sup> siècle chez des mathématiciens comme **Simon Laplace** ; **Gauss** l'utilise dans la première moitié du XIX<sup>ème</sup> siècle dans le cadre de la théorie des erreurs construite pour les astronomes qui font beaucoup de mesures entachées d'erreurs pour calculer quelques paramètres. Au tournant du XIX<sup>ème</sup> et du XX<sup>ème</sup> siècle des statisticiens anglais comme **K. Pearson** inventent des procédures. Il reviendra à **Ronald Fischer** et à **Jerzy Neyman** entre 1920 et 1940 de fournir les outils théoriques nécessaires au modèle. Après la 2<sup>ème</sup> guerre mondiale **Abraham Wald** proposera le modèle décisionnel dont s'inspire l'exposé ci-dessous<sup>39</sup>.

### A propos du modèle probabiliste

On sait que quand on peut interpréter la variabilité des observations faites comme le résultat d'épreuves aléatoires (que l'on pourra nommer "tirages"), alors on caractérise la population étudiée par une loi de probabilité (qui n'est pas complètement connue). Pour déterminer les aspects intéressants, pour l'expérimentateur, de cette loi, on fait plusieurs tirages, que nous supposons indépendants en probabilité. Il faut maintenant représenter mathématiquement ces tirages.

#### Le cas discret

Prenons l'exemple des cartouches : un lot important de cartouches contient une proportion  $\theta$  de mauvaises cartouches. Le modèle est celui du schéma de *Bernoulli*.

Un tirage peut être représenté par un ensemble  $X$  à deux éléments : la cartouche est bonne ou mauvaise. On codera par 0 la bonne cartouche et par 1 la mauvaise (pour la qualité, on s'intéresse en priorité aux mauvaises). On a donc  $X = \{0, 1\}$ . Comme on effectue  $n$  tirages, l'ensemble des tirages possibles sera représenté par l'ensemble produit  $X^n$  qui sera noté  $\Omega$ . Un élément  $\omega$  de  $\Omega$  sera donc une suite de  $n$  termes 0 ou 1. Le premier terme est le résultat du premier tirage, le  $j^{\text{ème}}$  terme celui du  $j^{\text{ème}}$  tirage.

L'ensemble  $\Omega$  ayant  $2^n$  éléments, donc un nombre fini, une loi de probabilité sur  $\Omega$  est donc une loi discrète représentable par la probabilité de chaque élément de  $\Omega$ . Posons  $k(\omega)$  le nombre de 1 de la suite  $\omega$ ,  $k$  est une application de  $\Omega$  dans  $\{0, 1, \dots, n\}$ , ensemble des  $n$  premiers entiers naturels. Si  $\theta$  est la probabilité de tirer une mauvaise cartouche, comme les tirages sont supposés indépendants (ce qui est admis si  $n$  est petit face à la taille du lot), on peut écrire :  $P(\{\omega\}) = \theta^{k(\omega)} (1 - \theta)^{n - k(\omega)}$ .

<sup>39</sup> Cet historique est développé au chapitre 3.

### Le cas continu

La plupart des phénomènes où la statistique est appliquée ne sont pas des phénomènes discrets mais des phénomènes continus. Prenons l'exemple d'un essai thérapeutique, où la tension artérielle prise à un moment donné sur un individu peut être vue comme la réalisation d'une variable aléatoire  $X$  prenant ses valeurs dans  $\mathbb{R}$  et dont la loi de probabilité sera décrite par une fonction  $f$ , appelée **densité de probabilité**, définie de  $\mathbb{R}$  dans  $\mathbb{R}$ , supposée intégrable et telle que  $P(a \leq X \leq b) = \int_a^b f(x) dx$ .

Si on raisonne comme un physicien du XVII<sup>ème</sup> siècle, avec les quantités évanescences (dont l'analyse non standard donne un statut rigoureux), on peut dire que  $f(x) dx$  est la probabilité pour que  $X$  soit dans un intervalle infiniment petit autour de  $x$  et de longueur  $dx$ . Si on fait  $m$  mesures indépendantes de la tension artérielle, l'ensemble de tous les tirages possibles sera l'ensemble  $\mathbb{R}^m$ , encore noté  $\Omega$ . On appellera  $X_i$  la variable aléatoire modélisant la  $i^{\text{ème}}$  mesure. La probabilité de l'évènement " $X_i$  est dans un intervalle infiniment petit autour de  $x_i$  et de longueur  $dx_i$ , pour  $i$  variant de 1 à  $n$ " sera la quantité infinitésimale  $l(x_1, \dots, x_m) dx_1 \dots dx_m$  où  $l$  est la densité multidimensionnelle du vecteur aléatoire  $(X_1, \dots, X_m)$ . Puis, les probabilités sur des intervalles seront obtenues par

$$\text{intégration : } P\left[\bigcap_{i=1}^m (a_i < X_i < b_i)\right] = \iint_{\cap_{i=1}^m [a_i, b_i[} \dots \int l(x_1, \dots, x_m) dx_1 \dots dx_m .$$

Si on suppose que les variables  $X_i$  sont indépendantes et de même loi (on répète  $m$  fois le même phénomène sans interférence entre les répétitions), on peut écrire :

$$l(x_1, \dots, x_m) dx_1 \dots dx_m = f(x_1)dx_1 \times f(x_2)dx_2 \times \dots \times f(x_m)dx_m \text{ et donc } l(x_1, \dots, x_m) = \prod_{i=1}^m f(x_i) .$$

Dans le cas de l'essai thérapeutique, on mesure  $m$  tensions artérielles pour l'échantillon témoin et  $n$  pour l'échantillon traité. Si  $f$  est le densité de probabilité de la variable tension artérielle en l'absence de médicament et  $g$  la densité de probabilité de la variable tension artérielle en présence de médicament, si les expériences sont indépendantes, alors les  $m + n$  observations peuvent être considérées comme la réalisation d'un vecteur aléatoire  $(X_1, X_2, \dots, X_m, X_{m+1}, \dots, X_{m+n})$  admettant une densité de probabilité :

$$l(x_1, \dots, x_m, x_{m+1}, \dots, x_{m+n}) = \prod_{i=1}^m f(x_i) \prod_{i=m+1}^{m+n} g(x_i) .$$

En termes plus savants, on dit que l'ensemble des observations  $\Omega = \mathbb{R}^{m+n}$  est muni d'une loi de probabilité  $P$  admettant une densité qui est la fonction  $l : \mathbb{R}^{m+n} \rightarrow [0, 1]$ . La densité est dite prise par rapport à la mesure de *Lebesgue*  $m$  qui n'est autre que la mesure du volume dans l'espace à  $m + n$  dimensions : soit un hyper parallélépipède construit sur les intervalles  $I_1, I_2, \dots, I_{m+n}$ , on a  $m(I_1 \times \dots \times I_{m+n}) = \lambda(I_1) \cdot \lambda(I_2) \dots \lambda(I_{m+n})$  où  $\lambda$  est la mesure de la longueur d'un intervalle.

Synthétiquement, on note  $\frac{dP}{dm}(x_1, \dots, x_m, x_{m+1}, \dots, x_{m+n}) = l(x_1, \dots, x_m, x_{m+1}, \dots, x_{m+n})$ .

## 1 – Le modèle probabiliste de recueil des données

Le constat expérimental observé est considéré comme le résultat d'épreuves aléatoires, il est une issue possible parmi d'autres. On introduit alors l'**ensemble  $\Omega$  de toutes les issues a priori possibles**.

Dans le cas où on étudie la qualité des cartouches d'un lot en contenant un très grand nombre, on procède à  $n$  essais,  $n$  très petit devant le nombre de cartouches du lot, ce qui garantit l'indépendance des tirages (les tirages ne modifient pas sensiblement les



proportions parmi les cartouches restantes). On a alors  $\Omega = \{0,1\}^n$ , avec les notations précédentes.

Dans le cas de l'essai thérapeutique, on mesure  $m + n$  tensions artérielles, chacune pouvant être interprétée comme un nombre réel. On peut donc prendre  $\Omega = \mathbb{R}^{m+n}$ . On objectera que des tensions négatives, ou prenant de très grandes valeurs, sont impossibles et que  $\Omega$  est plus gros que l'ensemble des issues a priori possibles. Cela n'a pas d'importance. On verra que l'on muni  $\Omega$  d'une loi de probabilité. Il suffira que la probabilité d'une issue impossible soit quasi nulle pour que le modèle proposé reste acceptable.

A ce stade, pour munir  $\Omega$  d'une loi de probabilité, il faudrait définir sur  $\Omega$  l'ensemble des sous-ensembles de  $\Omega$  probabilisables. Cet ensemble a une structure portant le nom de tribu. Comme cette notion n'intervient pas, pratiquement, dans les problèmes statistiques abordés dans cette première présentation, on fera l'impasse sur les questions de mesurabilité.

L'ensemble  $\Omega$  sera donc muni d'une **loi de probabilité**  $P$  qui régit le phénomène. L'expression mathématique de  $P$  dépend bien entendu des conditions dans lesquelles ont été faites les expériences.

Dans le cas des cartouches, on a vu que  $\Omega = \{0, 1\}^n$  et que si les conditions expérimentales sont telles qu'il est possible que tous les tirages soient considérés comme des répétitions indépendantes du même phénomène, alors on peut écrire :

$$P(\{(x_1, x_2, \dots, x_n)\}) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}, \text{ où } x_i \in \{0, 1\} \text{ et } \theta \text{ est la}$$

proportion de mauvaises cartouches dans le lot.

Encore faut-il que l'hypothèse "répétitions indépendantes" soit valide. Où sont les mauvaises cartouches dans la caisse ? Nul ne le sait. Peut-être sont-elles plus nombreuses dans le haut car à un moment la caisse a été exposée aux intempéries, ou au contraire sur un côté du fait d'un remplissage avec une petite série défectueuses ? Il faut donc neutraliser cette géométrie de la place des mauvaises cartouches dans la caisse. Comme elle est inconnue, il n'y a qu'une seule solution possible : mettre ce facteur inconnu dans le hasard. On va donc, par la pensée, affecter un numéro à chaque cartouche et procéder au tirage de ces numéros comme à une loterie : chaque numéro a la même probabilité d'être tiré. Notre processus expérimental doit simuler ce tirage. On pourra par exemple chercher les coordonnées dans la caisse des cartouches tirées grâce à la touche random de la calculatrice.

Dans le cas de l'essai thérapeutique, le même type de précaution est à prendre si on veut que la densité de probabilité des tensions artérielles que l'on mesurera puisse s'écrire comme on l'a vu précédemment sous la forme d'une densité de probabilité  $l$  telle que :

$$l(x_1, \dots, x_m, x_{m+1}, \dots, x_{m+n}) = \prod_{i=1}^m f(x_i) \prod_{i=m+1}^{m+n} g(x_i).$$

Il faudra aussi tirer au hasard, dans la population des malades souffrant d'hypertension, ceux qui feront partie de l'échantillon témoin et ceux qui recevront le médicament actif. On délivre aux individus de l'échantillon témoin un placebo identique au médicament par son aspect extérieur. Ni le malade, ni son médecin traitant, ne sait si le patient fait partie de l'échantillon témoin ou de l'échantillon traité. L'expérience est dite "en double aveugle" afin d'éviter des effets psychosomatiques : une personne qui croit être traitée développe des réactions qui la distinguent des autres. Laisser choisir le hasard pour déterminer qui est dans un échantillon ou qui est dans un autre permet également d'éviter les biais comme on dit en statistique. C'est la seule façon de se garantir contre le risque de choisir l'un des échantillons dans une sous-population qui diffère de la population globale par une autre caractéristique : la taille, le poids (on prendrait les traités parmi les personnes les plus grosses par exemple), le groupe sanguin ( le groupe O est très répandu en pays basque...)

etc. Cette procédure, dite de **randomisation**, rend le modèle probabiliste précédent plus crédible.

Prendre comme modèle probabiliste que les tensions artérielles mesurées sont des réalisations de  $m + n$  variables aléatoires admettant comme densité dans  $\mathbb{R}^{m+n}$  la fonction :

$$l(x_1, \dots, x_m, x_{m+1}, \dots, x_{m+n}) = \prod_{i=1}^m f(x_i) \prod_{i=m+1}^{m+n} g(x_i) \quad (\text{modèle 1})$$

suppose que l'on n'ait pas de renseignement complémentaire sur le phénomène.

Souvent des études antérieures ont permis de préciser certains aspects de celui-ci. Il convient alors de les traduire en termes probabilistes en spécifiant davantage la fonction  $l$ . Par exemple, on sait que si le médicament agit, il fait baisser la tension artérielle mais qu'il ne modifie pas la forme de la densité de probabilité en rendant compte. Cela se traduit par :

$\forall x \in \mathbb{R}, g(x) = f(x - \Delta)$  avec  $\Delta \leq 0$   
(modèle 2)

ce qu'illustre le schéma ci-contre.

Notons que si  $\Delta > 0$ , cela voudrait dire que le médicament, au lieu de diminuer la tension artérielle, l'augmenterait ! Ce cas de figure est supposé impossible.

Des études antérieures ont pu montrer qu'en sus, le phénomène "tension artérielle" pouvait se modéliser par une loi gaussienne. On peut alors écrire :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (\text{modèle 3})$$

Ces précisions supplémentaires seront évidemment utilisées pour mettre au point des procédures mieux adaptées donc plus performantes.

Sur l'ensemble  $\Omega$  des issues a priori possibles, on a une loi de probabilité  $P$  qui n'est pas complètement connue :

- dans le cas des cartouches, elle dépend de la proportion  $\theta$  de mauvaises cartouches ;
- dans le cas de l'essai thérapeutique, elle dépend de deux fonctions  $f$  et  $g$  que l'on peut éventuellement préciser.

Il est alors possible de dire que  $P$  appartient à une famille  $\Pi$  de lois de probabilités sur  $\Omega$  a priori possibles. Faire de la statistique sera d'une certaine façon choisir  $P$  dans  $\Pi$ . La famille  $\Pi$  étant peu maniable, on la met en correspondance biunivoque avec un ensemble  $\Theta$  appelé **espace des paramètres** :

- dans le cas des cartouches on a évidemment  $\Theta = [0,1]$ , ensemble des proportions de mauvaises cartouches a priori possibles ;
- dans le cas de l'essai thérapeutique, si on appelle  $F$  l'ensemble des densités de probabilités sur  $\mathbb{R}$ , dans les trois modèles énoncés on a respectivement :

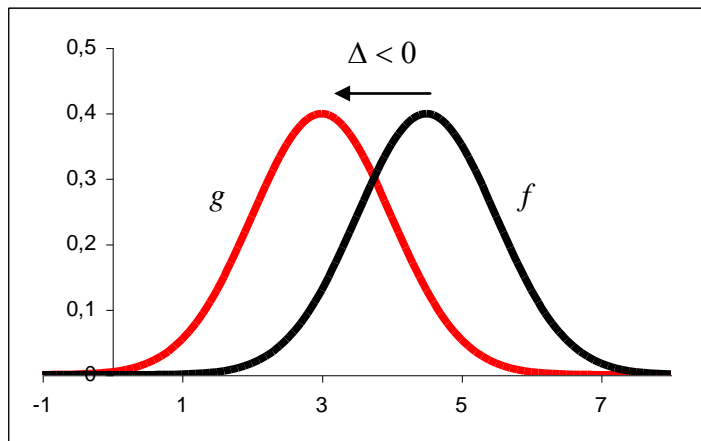
$$\Theta = F \times F \quad (\text{modèle 1, où il faut estimer les densités } f \text{ et } g) ;$$

$$\Theta = \mathbb{R}^- \times F \quad (\text{modèle 2, où il faut estimer } \Delta \text{ et } f) ;$$

$$\Theta = \mathbb{R}^- \times \mathbb{R}^2 \quad (\text{modèle 3, où il faut estimer } \Delta, \mu \text{ et } \sigma).$$

## 2 – La formalisation de l'action à mener

La formalisation suivante, développée par *Abraham Wald* dans les années 1940, permet d'unifier les points de vue, exposés plus loin, de l'estimation et des tests d'hypothèses.



On fait de la statistique pour prendre des décisions, pour avoir les connaissances recherchées sur le phénomène. La connaissance complète de la loi  $P$  qui régit le phénomène n'est pas toujours nécessaire. Ainsi, pour les cartouches, le commerçant qui reçoit le lot veut parfois connaître  $\theta$ , donc  $P$  (cadre de l'estimation), mais le plus souvent, si son seuil de rentabilité est à 10% de mauvaises cartouches, il lui suffit de savoir si  $\theta \leq 0,1$  ou si  $\theta > 0,1$  (contexte du test d'hypothèse). Pour l'essai thérapeutique, s'il s'agit de savoir si le médicament est efficace pour le mettre sur le marché on se pose la question, dans le cas du modèle 2 :  $\Delta = 0 ? \Delta < 0 ?$  On peut aussi vouloir connaître l'intensité de son action, donc estimer  $\Delta$ , mais  $f$  ne nous intéresse pas.

Les remarques précédentes sont susceptibles de la formalisation suivante. Soit  $A$  l'**ensemble des actions à mener**. Une action dépend de l'état de la nature symbolisé par un élément  $\theta$  de  $\Theta$ . Il existe donc une application  $h : \Theta \rightarrow A$  exprimant ce lien.

Dans le cas de **décisions dichotomiques**  $A = \{0,1\}$  (c'est le cadre des tests).

Pour les cartouches on a

$$h(\theta) = 0 \text{ si } \theta \leq 0,1$$

$$h(\theta) = 1 \text{ si } \theta > 0,1 .$$

Pour l'essai thérapeutique, avec le modèle 2,

$$h(0, f) = 0$$

$$h(\Delta, f) = 1 \text{ si } \Delta < 0$$

Dans le cas où il faut **donner une valeur à un paramètre** réel, on a  $A = \mathbf{R}$  (c'est le cadre de l'estimation) et pour l'essai thérapeutique  $h(\Delta, f) = \Delta$

### 3 – Stratégies et résumés

**Faire de la statistique c'est donc prendre une décision au vu d'un constat expérimental.** L'ensemble  $\Omega$  étant celui des constats et  $A$  celui des actions, le travail du statisticien consiste donc à trouver une **règle de décision**  $S$ , appelée aussi **stratégie**, qui est une application  $S : \Omega \rightarrow A$ , et à en déterminer les qualités qui ne pourront être exprimées qu'en termes probabilistes,  $S$  est une variable aléatoire. Du coup les qualités ne concernent que la règle et non le résultat obtenu par application de la règle au constat expérimental effectué.

Il n'est pas souvent facile de trouver directement une stratégie  $S$ . En règle générale, on cherche  $S$  à partir de résumés condensant l'information provenant des données. Soit  $Y$  l'ensemble dans lequel le **résumé**  $T$  (appelé aussi **statistique**) prend ses valeurs.

On a donc l'application  $T : \Omega \rightarrow Y$ , qui est une variable aléatoire.

Pour les cartouches, un résumé intuitif s'impose : la proportion de mauvaises cartouches

dans l'échantillon. On a alors  $Y = [0, 1]$  et si  $\omega = (x_1, x_2, \dots, x_n)$ , on a  $T(\omega) = \frac{1}{n} \sum_{i=1}^n x_i$ , où  $x_i$

vaut 0 ou 1 selon que la  $i^{\text{ème}}$  cartouche est bonne ou mauvaise. La variable aléatoire  $T$  associe ainsi à tout échantillon de taille  $n$ , prélevé dans la caisse de cartouches, la proportion de mauvaises cartouches dans l'échantillon.

Pour l'essai thérapeutique, on peut être tenté de résumer les données en considérant les variables aléatoires  $\bar{X}_1$  et  $\bar{X}_2$ , associant aux deux sous-échantillons aléatoires, sans ou avec traitement, les moyennes des tensions artérielles. On peut aussi considérer les variables aléatoires correspondant aux variances  $S_1^2$  et  $S_2^2$ . On a alors  $Y = \mathbf{R}^4$  et  $T = (\bar{X}_1, \bar{X}_2, S_1^2, S_2^2)$  avec des notations évidentes. Au lieu des  $m + n$  variables aléatoires représentant les données, on en considère quatre :  $\bar{X}_1, \bar{X}_2, S_1^2, S_2^2$ .

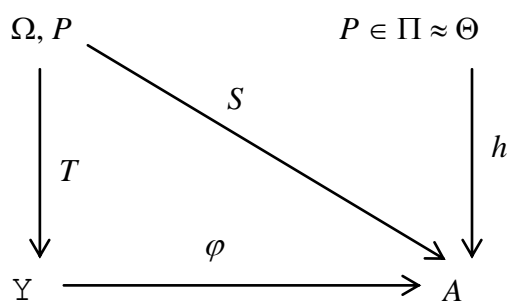
Dans certains cas, il suffit de prendre une statistique plus simple. Par exemple,  $W$  correspondant au nombre de couples d'observations, pour lesquels la valeur du 1<sup>er</sup> échantillon est supérieure à la valeur du 2<sup>e</sup> échantillon : si  $u$  est la fonction échelon,

$$u(x) = 0 \text{ si } x < 0 \text{ et } u(x) = 1 \text{ si } x \geq 0, \text{ on posera } Y = \mathbb{N} \text{ et } W(\omega) = \sum_{i=1}^m \sum_{j=1}^{n+1} u(x_i - x_{m+j}).$$

Dans un modèle donné, certains résumés peuvent avoir une importance particulière. Prenons l'exemple de l'essai thérapeutique avec le modèle 3, où les observations sont gaussiennes. Alors le résumé  $T$  a la propriété suivante : soit  $X_i$  la variable aléatoire représentant la mesure de la tension de l'individu  $i$ , si on cherche la loi de probabilité conditionnelle de  $(X_1, X_2, \dots, X_n)$  sachant  $T$ , celle-ci est indépendante des paramètres  $(\Delta, \mu, \sigma)$ . En d'autres termes la connaissance de  $(X_1, X_2, \dots, X_n)$  sachant  $T$  résume toute l'information que les données peuvent donner sur ces trois paramètres. Il est inutile de s'encombrer d'autres résumés ou informations. On dit que le **résumé**  $T$  est **exhaustif**. Quand elle existe, toute stratégie optimale sera évidemment une fonction de  $T$ . Bien entendu l'exhaustivité est relative au modèle. Si le modèle 3 n'est pas valide, car les lois sous-jacentes ne sont pas gaussiennes, alors  $T$  n'est plus exhaustif.

#### 4 – Schéma statistique

La formalisation précédente peut se synthétiser dans le schéma simple suivant :



$\Omega$  est l'ensemble de toutes les données a priori possibles.

$P$  est la loi de probabilité sur  $\Omega$ , partiellement inconnue.

$\Pi$  est l'ensemble des lois de probabilité a priori possibles représentant le phénomène.

$\Pi$  est en correspondance biunivoque avec un ensemble  $\Theta$  de paramètres, appelé ensemble des "états de la nature".

$A$  est l'ensemble des actions à mener.

L'action pertinente dépend de l'état de la nature, la liaison se fait par l'application  $h : \Theta \rightarrow A$ .

A chaque constat expérimental, il nous faut associer une action. Le travail du statisticien est donc de trouver une application  $S : \Omega \mapsto A$ . Pour ce faire, il bâtit sa stratégie  $S$  à partir de résumés des observations. Ces résumés peuvent être représentés par une application  $T : \Omega \rightarrow Y$  où  $Y$  est un ensemble à préciser,  $T$  est appelé une statistique.

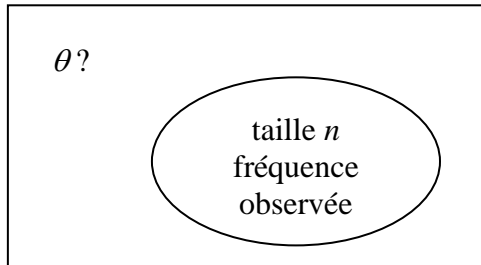
La recherche de  $S$  se fait alors par l'intermédiaire de  $T$  : il faut trouver une fonction  $\varphi : Y \rightarrow A$  et alors  $S = \varphi \circ T$ .

Pour aller plus loin, il est nécessaire de préciser le type de problème à résoudre. On abordera dans la suite les deux questions statistiques les plus courantes : d'une part l'estimation, ponctuelle et par intervalle, d'un paramètre réel (dans ce cas  $A = \mathbb{R}$ ), d'autre part les tests d'hypothèses (dans ce cas  $A = \{0, 1\}$ ).

## II.- L'ESTIMATION D'UN PARAMETRE REEL

### 1 – Estimateur et estimation ponctuelle

Pour illustrer la procédure d'estimation ponctuelle d'un paramètre réel, reprenons l'exemple des cartouches.



Nous voulons, au vu du constat expérimental, attribuer une valeur à la proportion  $\theta$  de mauvaises cartouches dans un stock.

Si, par exemple, sur un échantillon de  $n = 1000$  cartouches prélevé dans un stock important de cartouches, on en observe 62 défectueuses, on pourra prendre 0,062 comme estimation ponctuelle de  $\theta$ .

En termes de "stratégie", on a ici  $\Theta = [0, 1]$  et  $A = [0, 1]$ , l'application  $h$  est l'identité. On appellera **estimateur** de  $\theta$  (proportion de mauvaises cartouches dans la "population"), la stratégie qui à toute issue expérimentale possible, fait correspondre une valeur de  $\theta$ . Traditionnellement, un estimateur de  $\theta$  sera noté  $\hat{\theta}$ ,  $\hat{\theta}$  est une variable aléatoire.

Si  $\omega$  est le constat expérimental fait, l'application de la stratégie  $\hat{\theta}$  à  $\omega$  donne le résultat  $\hat{\theta}(\omega)$  qui est l'**estimation** de  $\theta$  pour la donnée observée  $\omega$ . Il faut, pour comprendre la méthode statistique, distinguer l'estimateur de l'estimation :  $\hat{\theta}$  est une variable aléatoire, alors que  $\hat{\theta}(\omega)$  est un nombre. Dans ce cas, il est intuitif de considérer comme estimation de la proportion de mauvaises cartouches dans la caisse, la proportion de mauvaises cartouches dans l'échantillon.

On a donc, pour tout  $\omega \in \Omega$ ,  $\hat{\theta}(\omega) = \frac{1}{n} \sum_{i=1}^n x_i$  avec  $\omega = (x_1, x_2, \dots, x_n)$ .

Plus généralement, si  $\Theta$  est un ensemble à préciser et  $A = \mathbb{R}$ , nous devons estimer  $h(\theta)$ .

L'essai thérapeutique en fournit un exemple. Plaçons nous dans le cas du modèle 2. Sans traitement, la loi de probabilité de la tension artérielle admet une densité  $f$ . Avec traitement, cette densité devient, pour tout  $x \in \mathbb{R}$ ,  $f(x - \Delta)$ . On a vu que si l'état de la nature est caractérisé par  $\theta = (\Delta, f)$ , on a  $h(\theta) = \Delta$ , que nous devons estimer. On verra plus loin plusieurs estimateurs possibles de  $\Delta$ . Comme précédemment, l'estimateur de  $h(\theta)$  sera la variable aléatoire  $\hat{h}(\theta) : \Omega \rightarrow \mathbb{R}$  et l'estimation, pour le constat expérimental  $\omega$ , sera  $\hat{h}(\theta)(\omega)$  qui est un nombre réel.

On s'intéresse maintenant à la qualité de l'estimation. Pour le constat expérimental  $\omega$ , l'estimation de  $h(\theta)$  est  $\hat{h}(\theta)(\omega)$ , l'erreur faite est donc  $h(\theta) - \hat{h}(\theta)(\omega)$ . Pour mesurer l'importance de cette erreur, il est d'usage de prendre son carré :  $[h(\theta) - \hat{h}(\theta)(\omega)]^2$ , ce qui pénalise fortement les grandes erreurs. L'erreur dépend donc du constat expérimental. Si on veut la caractériser, il nous faut considérer la variable aléatoire  $[h(\theta) - \hat{h}(\theta)]^2$ . On appelle **risque** son espérance mathématique, quand l'état de la nature est  $\theta$ .

On pose alors :  $R[h(\theta), \theta] = E_{\theta}[(\hat{h}(\theta) - h(\theta))^2]$  où l'indice  $\theta$  de  $E_{\theta}$  indique que l'espérance est calculée quand l'état de la nature est  $\theta$ , donc pour la loi de probabilité caractéristique de cet état.

Illustrons cette démarche par l'exemple des cartouches. Si la variable aléatoire  $X$  correspond au nombre de mauvaises cartouches dans l'échantillon, on sait que  $X$  suit la loi binomiale de paramètres  $n$  et  $\theta$ .

On a alors  $P_{\theta}(X = k) = C_n^k \theta^k (1 - \theta)^{n-k}$  puis  $E_{\theta}(X) = n\theta$  et  $\text{Var}_{\theta}(X) = n\theta(1 - \theta)$ .

L'estimateur de  $\theta$  est  $\hat{\theta} = \frac{X}{n}$ . On a donc  $E_{\theta}(\hat{\theta}) = \frac{n\theta}{n} = \theta$

et  $E_{\theta}[(\hat{\theta} - \theta)^2] = E_{\theta}[(\frac{X}{n} - \theta)^2] = \frac{1}{n^2} E_{\theta}[(X - n\theta)^2] = \frac{1}{n^2} \text{Var}_{\theta}(X) = \frac{\theta(1 - \theta)}{n}$ .

Le risque de l'estimateur  $\hat{\theta}$  est donc la fonction  $\theta \mapsto \frac{\theta(1 - \theta)}{n}$ .

L'erreur d'un estimateur peut donc être assimilée au résultat d'une loterie. Comparer deux estimateurs revient à comparer deux loteries, c'est à dire deux situations aléatoires.

Depuis *Pascal*, on compare des situations aléatoires réelles en comparant leurs espérances mathématiques. Ainsi, en changeant de notations, soit  $S_1$  et  $S_2$  deux estimateurs de  $h(\theta)$ , on préférera  $S_1$  à  $S_2$  si pour tout  $\theta \in \Theta$ ,  $R(\theta, S_1) \leq R(\theta, S_2)$ .

Mais ce critère est très décevant, il n'est pas possible de trouver une stratégie uniformément meilleure que toutes les autres. Il va falloir réduire le champ des stratégies possibles a priori. Par exemple la stratégie constante qui consiste à donner a priori une valeur à  $h(\theta)$  sans regarder les observations n'est dominée par aucune autre. Si cette valeur est la bonne alors le risque est nul !

On va donc définir des qualités pour les estimateurs. On exigera en premier lieu qu'un **estimateur** soit **convergent**. Notre problème est plongé dans une suite de problèmes identiques qui ne diffèrent que par le nombre  $n$  des individus de l'échantillon. On pourra alors  $S_n$  est dit **convergent en probabilité** si pour tout  $\theta \in \Theta$  et pour tout  $\varepsilon > 0$ ,  $P_{\theta} \left[ |S_n - h(\theta)| \geq \varepsilon \right] \xrightarrow{n \rightarrow \infty} 0$ .

Quand on fait un très grand nombre d'expériences, il est très peu probable d'obtenir une estimation éloignée de ce qu'il faut estimer. A l'évidence l'estimateur constant n'est pas convergent.

Une autre qualité recherchée, quand cela est possible, est qu'en moyenne l'estimateur soit proche de la chose estimée. Mathématiquement cela s'exprime par :  $E_{\theta}(S_n) = h(\theta)$ . On dit que  $S_n$  est **sans biais**. Quand un estimateur est sans biais, son risque est donné par sa variance :  $R(\theta, S_n) = E_{\theta}[(S_n - h(\theta))^2] = E_{\theta}[(S_n - E_{\theta}(S_n))^2] = \text{Var}_{\theta}(S_n)$ . On va donc, quand cela est possible, chercher l'**estimateur de variance minimum** parmi les estimateurs sans biais.

Il existe des méthodes de recherche d'estimateurs, on se reportera à un manuel de statistique pour l'exposé de la théorie. On va illustrer les propos précédents par les exemples déjà choisis.

Pour les cartouches, le bon sens nous pousse à choisir comme estimation de la proportion de mauvaises cartouches dans le lot, celle observée dans l'échantillon. On a, avec les

notations antérieures,  $\hat{\theta}_n(\omega) = \frac{1}{n} \sum_{i=1}^n x_i$ .

On sait que  $n\hat{\theta}_n$  est une variable aléatoire qui suit une loi binomiale (en supposant les tirages indépendants) donc on peut écrire  $E_\theta(\hat{\theta}_n) = \theta$ , donc  $\hat{\theta}_n$  est un estimateur sans biais, et  $\text{Var}_\theta \hat{\theta}_n = \frac{\theta(1-\theta)}{n}$ , donc  $\hat{\theta}_n$  est un estimateur convergent, puisque  $\text{Var}_\theta \hat{\theta}_n \xrightarrow{n \rightarrow \infty} 0$ .

On montre de plus qu'il est de variance minimum parmi les estimateurs sans biais. Pour l'essai thérapeutique, avec le modèle (2), on peut définir plusieurs estimateurs possibles :

$$\hat{\Delta}_2 = \bar{X}_2 - \bar{X}_1,$$

$$\hat{\Delta}_2 = \text{med}_{j \geq m+1}(X_j) - \text{med}_{i \leq m}(X_i) \quad (\text{med signifie médiane}),$$

$$\hat{\Delta}_3 = \text{med}_{j \geq m+1}(X_j - X_i)_{i \leq m}.$$

Ils sont tous convergents donc asymptotiquement sans biais (le biais de  $\hat{\Delta} - \Delta$  tend vers zéro quand  $n$  et  $m$  tendent vers l'infini).

On montre que  $\hat{\Delta}_1$  est optimal pour le modèle 3 avec lois gaussiennes, mais que  $\hat{\Delta}_3$  devient meilleur que  $\hat{\Delta}_1$  dès lors que l'on s'écarte quelque peu du modèle gaussien.

## 2 – Estimation par intervalle de confiance

Avoir une valeur approximative pour un paramètre  $h(\theta)$ , c'est bien, mais on aimerait connaître sa précision.

Commençons par donner une estimation par intervalle de confiance dans l'exemple des cartouches. La précision est liée à la taille  $n$  de l'échantillon prélevé. Pour définir ce qu'est cette précision, nous allons raisonner en termes de variables aléatoires.

La variable aléatoire  $X$  qui, à tout échantillon de taille  $n$  prélevé de façon indépendante, associe le nombre de cartouches défectueuses suit la loi binomiale de paramètres  $n$  et  $\theta$ , laquelle est proche, sous certaines conditions, de la loi normale de même moyenne  $n\theta$  et de même écart type  $\sqrt{n\theta(1-\theta)}$ . Ainsi, la variable aléatoire  $F = \frac{X}{n}$  suit approximativement la

loi normale de moyenne  $\theta$  et d'écart type  $\sqrt{\frac{\theta(1-\theta)}{n}}$ . Une variable aléatoire normale étant

telle que 95 % de ses observations apparaissent dans l'intervalle de rayon  $1,96 \times \sigma$  autour de la moyenne ( $\sigma$  désignant l'écart type), on peut donc écrire :

$$P\left(\theta - 1,96\sqrt{\frac{\theta(1-\theta)}{n}} \leq F \leq \theta + 1,96\sqrt{\frac{\theta(1-\theta)}{n}}\right) = 0,95.$$

Ce qui signifie que 95 % des échantillons de taille  $n$  que l'on est susceptible de prélever, dans le stock où la proportion de défectueuses est  $\theta$ , feront apparaître une fréquence de cartouches défectueuses comprise dans l'encadrement ci-dessus.

En remarquant que la fonction  $\theta \mapsto \theta(1 - \theta)$  est majorée par  $\frac{1}{4}$ , on peut agrandir l'encadrement en majorant  $1,96\sqrt{\theta(1-\theta)}$  par 1. On a alors :

$$P\left(\theta - \frac{1}{\sqrt{n}} \leq F \leq \theta + \frac{1}{\sqrt{n}}\right) \geq 0,95.$$

Retournant l'encadrement, on a également :  $P\left(F - \frac{1}{\sqrt{n}} \leq \theta \leq F + \frac{1}{\sqrt{n}}\right) \geq 0,95.$

Ce qui précède peut alors s'interpréter en disant que l'**intervalle aléatoire**  $\left[F - \frac{1}{\sqrt{n}}, F + \frac{1}{\sqrt{n}}\right]$  a une probabilité supérieure à 0,95 de recouvrir effectivement la valeur  $\theta$  à estimer.

Ainsi, supposons qu'un échantillon de taille  $n = 1000$  laisse apparaître 62 cartouches défectueuses, on pourra donner comme intervalle de confiance (à plus de 95 % de confiance) pour  $\theta$ , l'intervalle  $\left[0,062 - \frac{1}{\sqrt{1000}} ; 0,062 + \frac{1}{\sqrt{1000}}\right]$  soit  $[0,030 ; 0,094]$ . La procédure suivie ayant plus de 95 % de chances d'aboutir à un intervalle contenant effectivement  $\theta$ .

Généralisons l'exposé précédent. Pour simplifier les notations, par abus de langage,  $h(\theta)$  sera noté  $\theta$ . Comme on est dans l'aléatoire, vouloir qu'un intervalle recouvre à coup sûr la valeur inconnue du paramètre revient à prendre  $\mathbb{R}$  tout entier comme intervalle, ou  $[0,1]$  dans le cas d'une proportion, ce qui est évidemment sans intérêt. On va donc se fixer un **risque**  $\alpha$ , petit, et on va rechercher deux variables aléatoires  $\hat{\theta}^*$  et  $\hat{\theta}^{**}$  qui sont donc des applications de  $\Omega$  dans  $\mathbb{R}$  qui soient telles que :  $P_{\theta}(\hat{\theta}^* \leq \theta \leq \hat{\theta}^{**}) = 1 - \alpha$ .

L'intervalle aléatoire  $[\hat{\theta}^*, \hat{\theta}^{**}]$  a donc une probabilité  $(1 - \alpha)$  de recouvrir la valeur inconnue  $\theta$  (inconnue mais non aléatoire), si  $\theta$  correspond à la loi de probabilité régissant le phénomène.

Classiquement, on prend pour  $\alpha$  les valeurs suivantes : 0,10 ; 0,05 ; 0,01.

Evidemment plus  $\alpha$  est petit, plus le risque est faible mais plus grande est la largeur de l'intervalle.

Une fois la procédure mise au point, on l'applique au constat expérimental fait et on dit alors que :  $\hat{\theta}^*(\omega) < \theta < \hat{\theta}^{**}(\omega)$  au **niveau de confiance**  $(1 - \alpha)$ .

Attention : le niveau de confiance n'est pas la probabilité. Dans la double inégalité précédente rien n'est aléatoire,  $\theta$  est un paramètre inconnu,  $\hat{\theta}^*(\omega)$  et  $\hat{\theta}^{**}(\omega)$  sont des nombres bien déterminés. Le mot confiance est là pour rappeler que la procédure utilisée pour l'encadrement de  $\theta$  avait a priori une probabilité  $1 - \alpha$  de recouvrir la vraie valeur  $\theta$ . Illustrons ceci par l'exemple de l'essai thérapeutique dans le cas du modèle 3 (gaussien). On démontre que la quantité

$$T = \frac{\bar{X}_2 - \bar{X}_1 - \Delta}{S} \sqrt{\frac{mn}{m+n}} \quad \text{avec} \quad S^2 = \frac{1}{n+m-2} ((m-1)S_1^2 + (n-1)S_2^2)$$



suit une loi de probabilité dite de *Student* à  $n + m - 2$  degrés de liberté et indépendante des paramètres sans intérêt  $\mu$  et  $\sigma$ , loi dont il existe des tables et qui est symétrique. On trouve dans la table la valeur  $t_{\alpha, \nu}$  avec  $\nu = n + m - 2$  telle que  $P[-t_{\alpha, \nu} \leq T \leq t_{\alpha, \nu}] = 1 - \alpha$ .

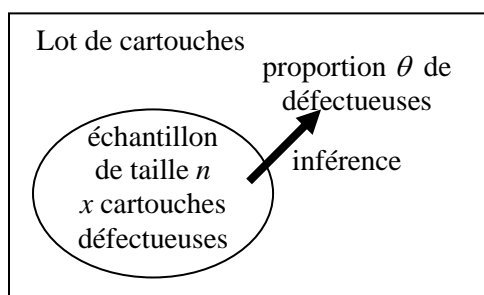
Un calcul simple montre alors que  $\hat{\Delta}^* = \bar{X}_{(2)} - \bar{X}_{(1)} - t_{\alpha, \nu} S \sqrt{\frac{m+n}{mn}}$  et

$$\hat{\Delta}^{**} = \bar{X}_2 - \bar{X}_1 + t_{\alpha, \nu} S \sqrt{\frac{m+n}{mn}}.$$

### III – LES TESTS STATISTIQUES

#### 1 – Un exemple de test statistique

Commençons, pour situer le problème, par un exemple, dans le cas des cartouches. Le fournisseur d'un lot important de cartouches affirme que celui-ci contient une proportion  $\theta$  de cartouches défectueuses, inférieure à 0,1.



Pour contrôler cette affirmation, on prélèvera, dans le lot, un *échantillon* aléatoire de  $n = 100$  cartouches dont on vérifiera la qualité.

En supposant que le prélèvement est effectué avec remise, la variable aléatoire  $X$  qui, à tout échantillon prélevé, associe le nombre  $x$  de cartouches défectueuses, suit la *loi binomiale* de paramètres  $n$  et  $\theta$ .

Deux hypothèses sont possibles. Soit  $\theta$  est inférieur à 0,1, soit  $\theta$  est supérieur à 0,1. On va voir que, dans la construction du test statistique, **ces deux hypothèses ne sont pas symétriques**.

• Supposons qu'au vu des expériences antérieures avec ce fournisseur, on lui fasse plutôt confiance.

On ne rejettera l'hypothèse "le lot est acceptable" que si on a de sérieuses raisons de le faire. Ce qui conduit à privilégier, a priori, l'hypothèse  $\theta \leq 0,1$ .

On appelle *hypothèse nulle* l'hypothèse  $H_0 : \theta \leq 0,1$ .

On désigne par  $H_1$  l'hypothèse alternative " $\theta > 0,1$ ".

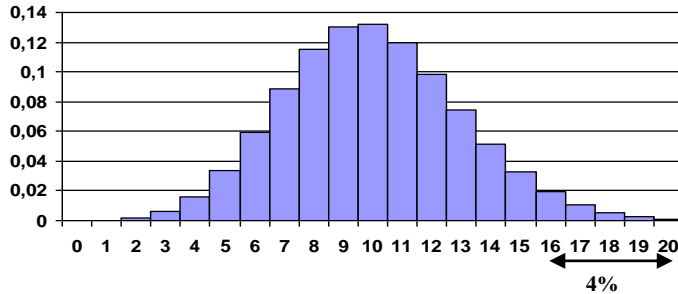
Nature \ Décision	$H_0$ Vraie	$H_1$ Vraie
$H_0$ acceptée	Exact	<b>Erreur 2</b>
$H_1$ acceptée	<b>Erreur 1</b>	Exact

Deux erreurs antagonistes sont possibles : décider  $H_1$  quand  $H_0$  est vraie (**erreur 1**) ou décider  $H_0$  quand  $H_1$  est vraie (**erreur 2**).

On limitera l'erreur de première espèce, en calculant la région de rejet de l'hypothèse  $H_0$  de sorte que, la probabilité de rejeter  $H_0$  alors que  $H_0$  est vraie, soit inférieure à 0,05.

Supposons que  $H_0$  soit vraie ( $\theta \leq 0,1$ ), on recherche le premier entier  $k$  tel que  $P(X \geq k) \leq 0,05$  où  $X$  suit la loi binomiale  $B(100, \theta)$ . C'est à dire que,  $H_0$  étant vraie, on aura moins de 5% de chances d'observer sur un échantillon, plus de  $k$  cartouches défectueuses, et, dans ce cas, on estimera plutôt que  $H_0$  est fautive, avec une erreur 1 de probabilité inférieure à 0,05.

Examinons les résultats fournis par la loi binomiale  $B(100; 0,1)$  (on prend  $\theta = 0,1$  mais lorsque  $\theta < 0,1$  les probabilités sont encore inférieures à celles-ci).



$k$	$P(X = k)$	$P(X \leq k)$
0	2,65614E-05	2,65614E-05
1	0,000295127	0,000321688
2	0,001623197	0,001944885
3	0,005891602	0,007836487
4	0,015874596	<b>0,023711083</b>
5	0,033865804	0,057576886
6	0,059578729	0,117155615
7	0,088895246	0,206050862
8	0,114823027	0,320873888
9	0,130416277	0,451290165
10	0,131865347	0,583155512
11	0,119877588	0,7030331
12	0,098788012	0,801821113
13	0,074302095	0,876123207
14	0,051303827	0,927427035
15	0,032682438	<b>0,960109473</b>
16	0,019291717	0,97940119
17	0,010591531	0,989992721
18	0,005426525	0,995419246
19	0,002602193	0,998021439
20	0,001170987	0,999192426

On constate que  $k = 16$ .

La **règle de décision** du test sera donc la suivante :

Soit  $x$  le nombre de cartouches défectueuses de l'échantillon.

Si  $x \leq 15$ , on accepte l'hypothèse  $H_0$  selon laquelle  $\theta \leq 0,1$ .

Si  $x \geq 16$ , on rejette  $H_0$  et l'on suppose que  $\theta > 0,1$ .

Ainsi, on ne rejettera le lot de cartouches que si l'on obtient au moins 16 cartouches défectueuses dans l'échantillon.

Ceci correspond à ce qu'on appelle un **test unilatéral à droite**, au **seuil** de 5% (ou 4% ici).

• Supposons maintenant que le fournisseur soit soupçonné d'être un margoulin. On prendra comme hypothèse nulle que le lot est mauvais donc que  $\theta \geq 0,1$ .

On a donc  $H_0$  : " $\theta \geq 0,1$ " et l'hypothèse alternative  $H_1$  : " $\theta < 0,1$ ".

On limitera l'erreur de première espèce, en calculant la région de rejet de l'hypothèse  $H_0$  de sorte que, la probabilité de rejeter  $H_0$  alors que  $H_0$  est vraie, soit inférieure à 0,05.

Supposons que  $H_0$  soit vraie ( $\theta \geq 0,1$ ), on recherche le plus grand entier  $k$  tel que  $P(X \leq k) \leq 0,05$  où  $X$  suit la loi binomiale  $B(100, \theta)$ . C'est à dire que,  $H_0$  étant vraie, on aura moins de 5% de chances d'observer sur un échantillon, moins de  $k$  cartouches défectueuses.

D'après les résultats de la loi binomiale  $B(100; 0,1)$  (lorsque  $\theta \geq 0,1$  les probabilités sont encore inférieures), on a  $k = 4$  (voir le tableau précédent).

La **règle de décision** du test sera donc la suivante :

Soit  $x$  le nombre de cartouches défectueuses de l'échantillon.

Si  $x \leq 4$ , on accepte l'hypothèse  $H_0$  selon laquelle  $\theta \geq 0,1$ .

Si  $x \geq 5$ , on rejette  $H_0$  et l'on suppose que  $\theta < 0,1$ .

Ainsi, on rejettera le lot de cartouches dès que l'on obtient 5 cartouches défectueuses dans l'échantillon.

Ceci correspond à ce qu'on appelle un **test unilatéral à gauche**, au **seuil** de 5%.

• On voit comment la décision dépend des a priori du décideur...

## 2 – Formalisation de la notion de test statistique

Si l'on reprend la formalisation due à *Abraham Wald*, on dira que l'on a affaire à un problème de test quand la **décision** à prendre est **dichotomique** :  $A = \{0,1\}$ . Les états de la nature sont alors partitionnés en deux sous-ensembles  $\Theta_0 = h^{-1}(\{0\})$  et  $\Theta_1 = h^{-1}(\{1\})$ . Pour les états de la nature deux possibilités :  $\theta \in \Theta_0$ , c'est l'hypothèse  $H_0$ , ou  $\theta \in \Theta_1$ , c'est l'hypothèse  $H_1$ . De même l'ensemble des constats expérimentaux est, par une stratégie  $S$ , partitionné en deux sous-ensembles  $C = S^{-1}(\{1\})$  ensemble des constats où on décidera  $H_1$  et  $\bar{C} = S^{-1}(\{0\})$  ensemble des constats où on décidera  $H_0$ . On peut alors faire le schéma :

Nature Décision	$H_0$ Vraie	$H_1$ Vraie
$H_0$ acceptée	Exact	<b>Erreur 2</b>
$H_1$ acceptée	<b>Erreur 1</b>	Exact

Deux erreurs antagonistes sont possibles : décider  $H_1$  quand  $H_0$  est vraie (erreur 1), décider  $H_0$  quand  $H_1$  est vraie (erreur 2). On peut essayer de quantifier ces erreurs.

$P_\theta(C)$  est la probabilité de décider  $H_1$  quand  $\theta$  est l'état de la nature, donc l'erreur 1 pourra être "mesurée" par la fonction  $P_\theta(C)$  avec  $\theta \in \Theta_0$  et l'erreur 2 par la fonction  $(1 - P_\theta(C))$  avec  $\theta \in \Theta_1$ .

On veut minimiser les erreurs. Pour simplifier supposons que  $\Theta_0$  et  $\Theta_1$  soient des ensembles à un élément. On peut choisir  $C$  (dans  $S$ ) tel, qu'à partir d'un certain moment, il ne sera plus possible de diminuer les deux erreurs à la fois, diminuer une erreur revient à augmenter l'autre. A la limite, si  $C = \emptyset$  (on accepte  $H_0$  "les yeux fermés"), l'erreur 1 est nulle. Il faut donc proposer une problématique de choix.

Il reviendra à *Neyman* et *E. Pearson* de proposer en 1930 une problématique de décision universellement admise. Ils ont remarqué que, dans la plupart des problèmes, les deux hypothèses n'étaient pas symétriques. Il y en a une dont le rejet, quand elle est vraie, peut avoir de graves conséquences, ou bien, qui est issue d'une théorie en vigueur. Pour le médicament il est grave de mettre sur le marché un médicament inefficace. Appelons  $H_0$  cette hypothèse, elle sera dite "hypothèse nulle". On va essayer de se garantir contre un rejet trop fréquent de  $H_0$  quand elle est vraie. On va donc limiter l'erreur 1, dite erreur de première espèce, à un niveau  $\alpha$  fixé d'avance. On choisira alors la région  $C$  dite **région critique** (si  $\omega \in C$  on rejette  $H_0$ ) telle que  $\sup_{\theta \in \Theta_0} P_\theta(C) \leq \alpha$ ,  $\alpha$  est appelé aussi **seuil**.

Parmi ces régions critiques, on choisira celle (si elle existe) telle que les  $(1 - P_\theta(C))_{\theta \in \Theta_1}$  soient les plus petits possibles, ils mesurent la probabilité de déclarer à tort que l'alternative  $H_0$  est vraie si l'état de la nature est  $\theta \in \Theta_1$ .

On a vu, dans l'exemple des cartouches, comment la décision dépend des a priori du décideur : choix de l'hypothèse nulle et aussi choix du seuil. Plus on a confiance dans l'hypothèse nulle, plus on choisira un  $\alpha$  faible, pour avoir une probabilité très faible de rejeter  $H_0$  quand elle est vraie. En contre-partie, on aura une probabilité relativement forte de rejeter  $H_1$  quand  $H_1$  est vraie. En statistique il n'est pas possible "d'avoir le beurre et l'argent du beurre" !

Le choix du seuil dépend donc du problème auquel est appliqué la statistique et n'est pas de la compétence du seul statisticien. Reste au spécialiste du problème à comprendre la démarche statistique pour faire des interprétations valides mais cela est une autre histoire.

# 6

## LA SIMULATION EN STATISTIQUE

### Qu'est-ce que la simulation ?

Le développement des *moyens de calcul informatiques* a modifié bien des pratiques scientifiques. La statistique ne fait pas exception : l'ordinateur ou la calculatrice permettent assez simplement d'expérimenter des situations aléatoires, par simulation à partir d'une loi donnée, en les répétant un grand nombre de fois. On verra que la loi des grands nombres en permet une justification. Il pourra s'agir, soit d'étudier les conséquences d'un modèle, en le faisant "tourner", soit de conjecturer certains résultats, là où le calcul est trop compliqué, ou impossible.

Selon *Emile Borel*<sup>40</sup>, "*le but principal du calcul des probabilités [...] est de calculer les probabilités de phénomènes complexes en fonction des probabilités, supposées connues, de phénomènes plus simples*". **En statistique, c'est à partir d'observations que l'on évalue des probabilités.** La simulation permet de fabriquer de telles observations, à partir de probabilités simples, supposées connues, et ainsi de se faire une idée de la qualité des procédures employées.

#### Des exemples

- Dans le *jeu de pile ou face*, intéressons nous, par exemple, aux séries de lancers consécutifs égaux. La modélisation (admise) consiste à dire qu'à chaque lancer, on a "une chance sur deux" d'avoir pile ou d'avoir face (cette probabilité de 1/2 est admise). Dans le cadre de ce modèle, la simulation permet de conjecturer des résultats non triviaux et non intuitifs, comme par exemple d'évaluer la probabilité d'observer au moins six lancers consécutifs égaux sur 200 lancers.

Le programme sur calculatrice suivant calcule, pour 200 lancers de pile ou face simulés (on verra comment plus loin), la longueur maximale des lancers consécutifs égaux.

CASIO Graph 25 → 100	T.I. 80 - 82 - 83	T.I. 89 - 92
Seq(1,I,1,2,1) → List 1↵	:1 → A	:1 → a
Int(Ran# + 0.5) → R↵	:1 → M	:1 → m
For 1 → I To 200↵	:int(rand + 0.5) → R	:int(rand( ) + 0.5) → r
Int(Ran# + 0.5) → S↵	:For(I,1,200)	:For i,1,200
If S = R↵	:int(rand + 0.5) → S	:int(rand( ) + 0.5) → s
Then List 1[1] + 1 → List 1[1]↵	:If S = R	:If s = r
S → R↵	:Then	:Then
Max(List 1) → List 1[2]↵	:A + 1 → A	:a + 1 → a
Else 1 → List 1[1]↵	:S → R	:s → r
IfEnd↵	:max(A,M) → M	:max(a,m) → m
S → R↵	:Else	:Else
Next↵	:1 → A	:1 → a
List 1[2]	:End	:EndIf

<sup>40</sup> *Emile Borel* – "Les probabilités et la vie".

	:S → R :End :M	:s → r :EndFor :Disp m
--	----------------------	------------------------------

Cinq simulations ont donné , par exemple, les résultats suivants :

Longueur maximale des lancers consécutifs égaux, sur 200 lancers				
8	7	6	11	7

Sur 200 lancers consécutifs, on a, à chaque fois, observé au moins 6 lancers consécutifs égaux (et parfois beaucoup plus). Ce qui est assez *contraire à l'intuition*. Bien sûr, cinq simulations, ce n'est pas assez pour faire des statistiques. On verra plus loin comment utiliser le théorème limite central pour déterminer combien effectuer de simulations pour évaluer correctement la probabilité d'avoir au moins 6 lancers consécutifs égaux sur 200 lancers de pile ou face.

Dans cette situation, le calcul est en fait possible (mais pas immédiat). La probabilité qu'une suite de 200 lancers de pile ou face contienne au moins une série de 6 lancers consécutifs égaux est environ 0,965 (voir l'encadré).

• Prenons un second exemple dans un cas pratique.

On cherche à étudier *le temps de rotation moyen d'un bus*, selon la configuration d'une ligne urbaine. On a modélisé la situation. Le temps d'arrêt à chaque station est aléatoire, en fonction du nombre de descentes et de montées (selon un processus de Poisson). De même, le temps entre deux stations dépend des aléas de la circulation.

On peut alors, dans un programme, simuler les différentes variables aléatoires intervenant ici (dont le choix aura été déterminé selon un historique statistique), puis les combiner de façon à simuler une rotation du bus. Ayant simulé ce modèle,

on peut alors le faire "tourner" un grand nombre de fois et évaluer, entre autres, le temps moyen de rotation d'un bus (mais aussi observer les phénomènes d'attente).

### Une définition

Comme on vient de le voir, simuler une expérience aléatoire consiste à produire "virtuellement" des résultats analogues à ceux que l'on aurait obtenus en réalisant "physiquement" l'expérience aléatoire.

### **La probabilité qu'une suite de 200 lancers de pile ou face contienne au moins une série de 6 lancers consécutifs égaux est environ 0,965**

Notons  $u_n$  le nombre de suites de  $n$  lancers de pile ou face ( $x_i$ ), avec  $i$  de 1 à  $n$ , ne contenant *aucune* séquence de 6 consécutifs égaux.

Une telle suite peut être de 5 types différents, en considérant les 6 derniers termes, dénombrés ainsi :

⇒  $x_{n-1} \neq x_n$  : il y a  $u_{n-1}$  telles suites.

⇒  $x_{n-1} = x_n$  et  $x_{n-2} \neq x_{n-1}$  : il y a  $u_{n-2}$  telles suites.

⇒  $x_{n-1} = x_n$  ;  $x_{n-2} = x_{n-1}$  et  $x_{n-3} \neq x_{n-2}$  : il y a  $u_{n-3}$  telles suites.

⇒  $x_{n-1} = x_n$  ;  $x_{n-2} = x_{n-1}$  ;  $x_{n-3} = x_{n-2}$  et  $x_{n-4} \neq x_{n-3}$  : il y a  $u_{n-4}$  telles suites.

⇒  $x_{n-1} = x_n$  ;  $x_{n-2} = x_{n-1}$  ;  $x_{n-3} = x_{n-2}$  et  $x_{n-4} = x_{n-3}$  et  $x_{n-5} \neq x_{n-4}$  : il y a  $u_{n-5}$  telles suites.

La probabilité cherchée est donc  $\frac{u_{200}}{2^{200}}$  où la suite  $(u_n)$  est définie par :

$u_1 = 2$  ;  $u_2 = 4$  ;  $u_3 = 8$  ;  $u_4 = 16$  ;  $u_5 = 32$

puis  $u_n = u_{n-1} + u_{n-2} + u_{n-3} + u_{n-4} + u_{n-5}$  pour  $n \geq 6$ .

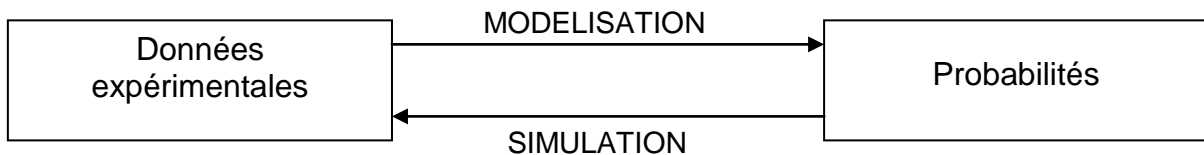
Ceci permet le calcul de  $u_{200}$  de proche en

proche. On a ainsi  $\frac{u_{200}}{2^{200}} \approx 0,965313$ .

Une définition plus précise de la simulation est donnée par *Yadolah Dodge*<sup>41</sup>.

"La simulation est la méthode statistique permettant la reconstitution  **fictive**  de l'évolution d'un phénomène. C'est une  **expérimentation**  qui suppose la constitution d'un modèle  **théorique**  présentant une similitude de propriétés ou de relations avec le phénomène faisant l'objet de l'étude."

Ceci peut être schématisé ainsi :



En produisant des données, sous un certain modèle, la simulation permettra d'examiner les conséquences, souvent non évidentes, de ce modèle, et, éventuellement, son adéquation aux données réelles. De façon générale, la simulation permet d'obtenir (ou de conjecturer) des résultats difficiles, ou impossibles, à calculer (c'est dans ce cadre qu'elle est utile aux statisticiens)<sup>42</sup>.

Pour imiter le hasard, les simulations sont basées sur le calcul de nombres pseudo-aléatoires qui, non seulement sont imprévisibles, mais encore ont le "goût" (statistiquement) du hasard.

## Comment peut-on simuler le hasard ?

### **Hasard ou pseudo-hasard ?**

Les premières tables de nombres au hasard ont été construites à partir des  **numéros gagnants de la loterie** . Cette pratique a conduit à désigner par "*méthode de Monte-Carlo*" les procédés de calcul d'aire utilisant ces nombres au hasard. Ainsi, alors que le statisticien *Karl Pearson* (1857-1936) eut beaucoup recours à des lancements de pièces ou de dés, embauchant pour ce faire amis et élèves, son fils *Egon Pearson* (1895-1980), à l'origine de la théorie des tests, utilisa ce qu'on appela plus tard la simulation, grâce à des tables de nombres au hasard produites dans les années 1925. En 1955, la *Rand Corporation* édita une table "*A Million Random Digits*" obtenue à partir de  **bruits de fond électroniques**  (fluctuations du débit de tubes électroniques). Il s'agit alors d'un générateur aléatoire physique.

Avec l'apparition des ordinateurs, on chercha à générer des nombres aléatoires, à l'aide d' **algorithmes** . Il ne s'agit plus de hasard physique mais d'un hasard calculé. On comprend bien l'antagonisme entre les deux termes. On ne peut pas calculer des nombres au hasard, puisqu'il sont alors le résultat d'un algorithme déterministe.

Cela nous conduit à nous poser la question : "quand peut-on dire qu'une suite de nombres est une suite au hasard ?" On peut se limiter à une suite de 0 et de 1, et la question devient : "quand peut-on considérer qu'une suite de 0 et de 1 est une suite au hasard ?" C'est à dire résultant d'un tirage à pile ou face, ou encore, de façon plus mathématique, comme étant les résultats successifs d'une suite de variables aléatoires  $X_i$  indépendantes et valant 0 ou 1 avec une probabilité 0,5.

<sup>41</sup> *Statistique. Dictionnaire encyclopédique*. Dunod 1993

<sup>42</sup> Il existe également des simulations non aléatoires (dans l'étude de systèmes dynamiques, par exemple).

Cette question est mathématiquement très difficile. Une réponse théorique à été apportée en 1966 par *Martin-Löf* : "Une suite de chiffres est aléatoire quand le plus petit algorithme nécessaire pour l'introduire dans l'ordinateur contient à peu près le même nombre de bits que la suite". Cette définition, exclut donc toute possibilité d'une règle effective.

Un objectif plus raisonnable est de trouver un algorithme produisant une suite de nombres, telle qu'un statisticien en l'analysant, ne soit pas capable de détecter si elle a été produite par un procédé mathématique ou un réel phénomène aléatoire physique : qu'il lui soit impossible, par exemple, sur une suite assez grande de 0 et de 1 (disons 200) de savoir s'ils ont été générés par un ordinateur, ou en lançant une pièce de monnaie bien équilibrée. Une telle suite est *pseudo-aléatoire*. Ces suites, construites sur des procédés récurrents, sont nécessairement périodiques, puisque l'on travaille avec un nombre fini de décimales. On cherche donc à ce que la période soit très grande et il faut être sûr de son générateur lorsque l'on a besoin d'une très grande quantité de nombres au hasard.

Pour simuler une variable aléatoire de loi donnée, le principe consiste à "déformer" un générateur de nombres pseudo-aléatoires correspondant à une distribution uniforme sur l'intervalle  $[0, 1]$ .

### **Simuler une distribution uniforme sur $[0, 1]$**

La plupart des générateurs de nombres (pseudo) aléatoires, simulent le tirage au hasard d'un nombre réel (ou plutôt décimal) entre 0 et 1. De façon plus précise, on simule les réalisations d'une suite de variables aléatoires indépendantes  $X_i$  de même loi  $U$  ( $[0, 1]$ ).

Les procédés les plus courants consistent en des suites récurrentes, de grande période, et dont le comportement chaotique *satisfait à divers tests statistiques*, permettant de valider l'hypothèse qu'il s'agit de réalisations de  $X$  (un premier test peut se construire sur l'observation des fréquences d'apparition des différents chiffres).

### **Un premier exemple de générateur de nombres aléatoires dans $[0, 1]$**

Si les constructeurs de calculatrices actuels ne donnent pas de renseignement quant à leur générateur de nombres aléatoires, ce n'était pas le cas dans les années soixante dix, où le générateur suivant correspond à un ancien modèle de calculatrice *Hewlett-Packard*.

En mode habituel de calcul, effectuer :

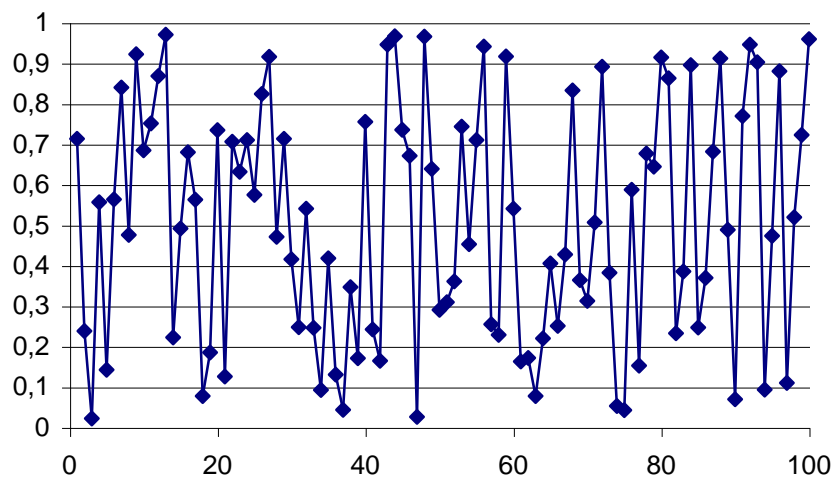
CASIO	TEXAS INSTRUMENTS
0.5 → X EXE Frac (9821X + 0.211327) → X EXE Puis appuyer plusieurs fois sur EXE.	0.5 → X ENTER fPart (9821X + 0.211327) → X ENTER Puis appuyer plusieurs fois sur ENTER.

Il s'agit donc d'une suite  $(x_n)$  définie par récurrence par  $x_0 = 0,5$  et  $x_{n+1}$  est la partie fractionnaire de  $9821x_n + 0,211327$ .



On indique ci-dessous les premiers termes de cette suite. Le graphique montre son caractère chaotique. Reste à vérifier qu'elle possède les propriétés statistiques de la loi uniforme sur  $[0, 1]$ .

0,71327
0,237940002
0,022028641
0,55655523
0,142180072
0,563753843
0,839757939
0,475991181
0,92266285
0,685117385
0,751110381
0,868318686
0,971083347
0,222816259
0,49174896
0,67980342
0,562658
0,077484166
0,185268453
0,734743145



### Un second générateur obtenu selon la méthode de *Lucas-Lehmer*

De nombreux générateurs sont obtenus à partir de propriétés arithmétiques, en particulier suite aux travaux de *Lehmer*, dans les années soixante dix. Certaines suites congruentes possèdent, en effet, des propriétés structurelles démontrées, comme la grande longueur de leur période (pour les propriétés arithmétiques, voir l'encadré suivant), qui en font, a priori, de bons candidats pour servir de générateur aléatoire. On leur fait subir ensuite toutes sortes de tests statistiques (on en verra des exemples plus loin) pour sélectionner le plus satisfaisant. Mais, cette fois, il n'y aura pas de certitude. La méthode statistique ne démontre pas qu'un générateur donné est toujours satisfaisant pour une simulation donnée. On choisit des entiers  $a$  et  $m$  premiers entre eux ( $m$  grand, souvent un nombre premier), puis on construit la suite  $(r_n)$  d'entiers positifs de  $[0, m - 1]$ , définie à partir d'une valeur  $r_0$ , non nulle et première avec  $m$ , et de la relation de récurrence :

$$r_{n+1} = a r_n \pmod{m},$$

c'est à dire que  $r_{n+1}$  est le reste de la division euclidienne de  $a r_n$  par  $m$ .

La suite  $(x_n)$  définie par  $x_n = \frac{r_n}{m}$  fournit, pour certains choix de  $a$  et  $m$ , un générateur de nombres aléatoires dans  $[0, 1]$ .

Le choix de  $a$  et  $m$  est effectué selon des critères statistiques et dépend de la configuration de l'ordinateur.

Pour un modèle **IBM** des années 80, on avait par exemple choisi :

$$a = 7^5 ; m = 2^{31} - 1 ; r_{n+1} = a r_n \pmod{m}$$

$$\text{puis } x_n = r_n / m.$$

**Propriétés arithmétiques  
de la suite  $r_{n+1} = ar_n \pmod{m}$   
 $0 < r_0 < m$ ,  $r_0$  et  $a$  premiers avec  $m$**

- Tout d'abord, pour tout  $n$  dans  $\mathbb{N}$  on a  $r_n$  non nul et premier avec  $m$ .

En effet,  $r_n = 0$  impliquerait l'existence d'un entier  $k$  tel que  $ar_{n-1} = km$  mais, puisque  $m$  est premier avec  $a$ , le lemme de *Gauss* donnerait que  $m$  divise  $r_{n-1}$ , c'est à dire  $r_{n-1} = 0$  ( $r_{n-1}$  est un reste modulo  $m$ ). Par récurrence, on remonterait à  $r_0 = 0$ , ce qui est exclu.

De même, s'il existait  $d$  diviseur commun à  $r_n$  et  $m$ , alors  $d$  diviserait  $ar_{n-1}$  et, puisque  $m$  est premier avec  $a$ , le lemme de *Gauss* donnerait que  $d$  divise  $r_{n-1}$ . On remonterait à  $d$  divise  $r_0$  qui est exclu.

- Dans ces conditions, la suite  $(r_n)$  a pour période l'ordre multiplicatif de  $a$  modulo  $m$ , c'est à dire le plus petit entier  $k$  tel que  $a^k = 1 \pmod{m}$ .

En effet  $r_{n+t} = r_n \Leftrightarrow a^t r_n = r_n \pmod{m} \Leftrightarrow (a^t - 1)r_n = km$  avec  $k$  entier, c'est à dire, puisque  $r_n$  est premier avec  $m$ ,  $a^t = 1 \pmod{m}$ .

- Lorsque  $m$  est premier, le petit théorème de *Fermat* affirme que si  $m$  est premier et ne divise pas  $a$ , alors  $a^{m-1} = 1 \pmod{m}$ , donc, pour  $a$  premier avec  $m$ , l'ordre multiplicatif de  $a$  divise  $m-1$ .

Un théorème de *Legendre* assure alors, pour  $m$  premier, l'existence de nombres d'ordre multiplicatif maximum  $m-1$  modulo  $m$ .

Par exemple, modulo 5, le nombre 2 est d'ordre multiplicatif maximum 4 car  $2^2 = 4 \pmod{5}$  puis  $2^3 = 3 \pmod{5}$  et  $2^4 = 1 \pmod{5}$ .

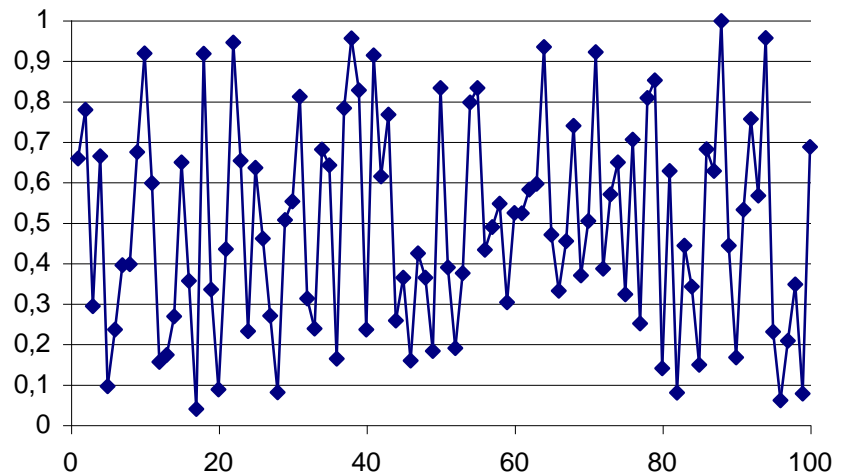
Il n'existe malheureusement pas d'algorithme permettant de trouver ces nombres d'ordre maximum. Le théorème de *Legendre* précise cependant qu'entre 1 et  $m-1$ , il en existe  $\varphi(m-1)$  correspondant au nombre d'entiers premiers avec  $m-1$  dans  $\{1, 2, \dots, m-2\}$ . Les nombres d'ordre maximum ne sont donc pas rares, et, avec l'aide de l'ordinateur, on peut trouver un tel  $a$ , qui nous assurera une suite dont la plus petite période est  $m-1$ ,  $m$  étant un très grand nombre premier.

- Lorsque  $m$  n'est pas premier, un théorème d'*Euler* donne que, si  $a$  et  $m$  sont premiers entre eux, alors  $a^{\varphi(m)} = 1 \pmod{m}$  où  $\varphi$  est l'indicateur d'*Euler* correspondant au nombre d'entiers premiers avec  $m$  dans  $\{1, 2, \dots, m-1\}$ . Ainsi, l'ordre multiplicatif de  $a$  est alors un diviseur de  $\varphi(m)$ . Mais on n'est pas assuré qu'il existe un tel nombre d'ordre multiplicatif maximum  $\varphi(m)$ , modulo  $m$ .

Par exemple,  $\varphi(21) = 12$  et, modulo 21, l'ordre multiplicatif de 5, par exemple, est 6 ( $5^6 = 1 \pmod{21}$ ) qui est un diviseur de 12 et l'ordre multiplicatif maximum, modulo 21.

On donne ci-dessous les premières valeurs de  $(x_n)$ , obtenue sur Excel, avec  $a = 7^5$  ;  $m = 2^{31} - 1$  et en débutant avec la valeur  $r_0 = 5$ .

(n=2) 0,657688941
0,778026611
0,29325066
0,663836187
0,094795932
0,235223081
0,394323584
0,396482029
0,67346448
0,917510387
0,59708186
0,154826731
0,172860553
0,267308175
0,648500967
0,35574692
0,038490931
0,917078254
0,334211188
0,087429872



Remarque :

Le nombre  $2^{31} - 1$  est un nombre de *Mersenne* premier. Les nombres de *Mersenne* ne sont pas tous premiers (la calculatrice TI 89 donne par exemple, en faisant  $\text{factor}(2^{30} - 1)$  :

$2^{30} - 1 = 3^2 \times 7 \times 11 \times 31 \times 151 \times 331$ ) mais leur intérêt réside dans le fait qu'il existe un test (découvert par *Lucas* en 1878) permettant de savoir s'ils sont premiers, et que leur manipulation est très commode dans le système binaire des ordinateurs, puisque  $2^{31} - 1$  s'écrit avec 31 chiffres 1 consécutifs.

*Edouard Lucas* (1842-1891), professeur au lycée St-Louis, puis Charlemagne, à Paris, est célèbre pour ses résultats en théorie des nombres, et ses "Récréations mathématiques" (il est l'inventeur du problème des "Tours de Hanoi").

### Tester un générateur de nombres aléatoires

On construit, à partir de la suite  $(x_n)$  fournie par le générateur, une suite de chiffres aléatoires parmi 0, 1, 2, ..., 9, en faisant  $\text{Ent}(10 x_n)$  où  $\text{Ent}$  désigne la partie entière, instruction qui ne retient que la première décimale.

Le premier test à effectuer est celui des **fréquences d'apparition des différents chiffres**. Chaque chiffre doit avoir une probabilité 1/10 de sortir, et sur  $n$  chiffres consécutifs fournis par le générateur, la fréquence observée d'un chiffre doit se répartir, suivant les

échantillons, approximativement selon la loi normale  $N\left(\frac{1}{10}, \sqrt{\frac{\frac{1}{10} \times \frac{9}{10}}{n}}\right)$  (d'après le théorème

limite central).

La dispersion "normale" des fréquences observées, sur des échantillons de taille  $n$ , doit donc se faire, si le générateur est bon, avec un écart type  $\sigma = \frac{0,3}{\sqrt{n}}$ , c'est à dire 0,03 si

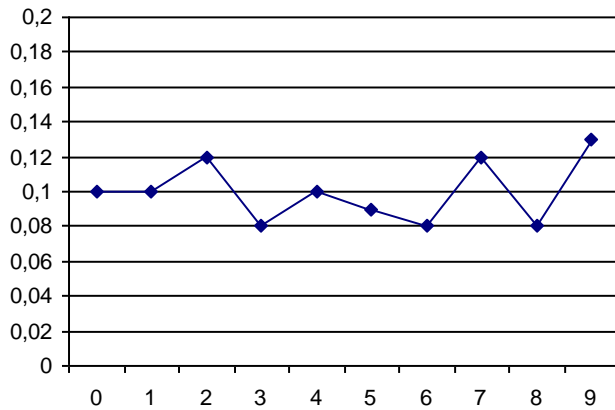
$n = 100$  et 0,0095 si  $n = 1000$  (il est à noter qu'une dispersion trop faible est aussi suspecte que le contraire !). D'après les propriétés de la loi normale, on devrait donc avoir 95 chances sur 100 d'observer les fréquences d'apparition d'un chiffre à moins de  $2\sigma$  de 1/10, alors qu'un écart de  $3\sigma$  est peu probable.

Ainsi, sur des échantillons de taille  $n = 100$ , on a 95% de chances d'observer la fréquence de sortie d'un chiffre dans la bande  $[0,04 ; 0,16]$ , alors que pour des échantillons de taille  $n = 1000$ , les fréquences doivent, à 95%, se situer dans la bande  $[0,08 ; 0,12]$ . On peut

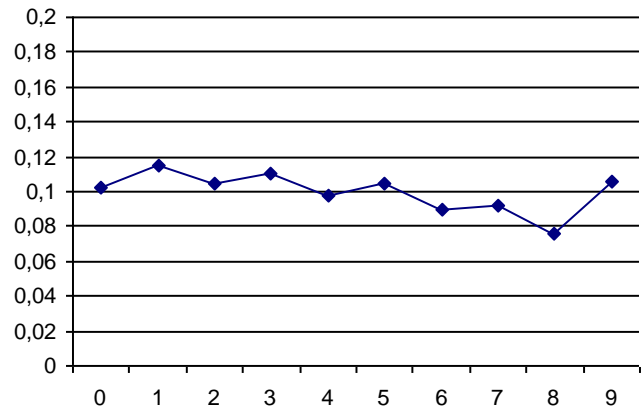
donc construire un premier test statistique, en écartant un générateur fournissant trop fréquemment une fréquence en dehors de ces intervalles.

En conservant la première décimale des résultats des deux générateurs précédents, on obtient les graphiques suivants.

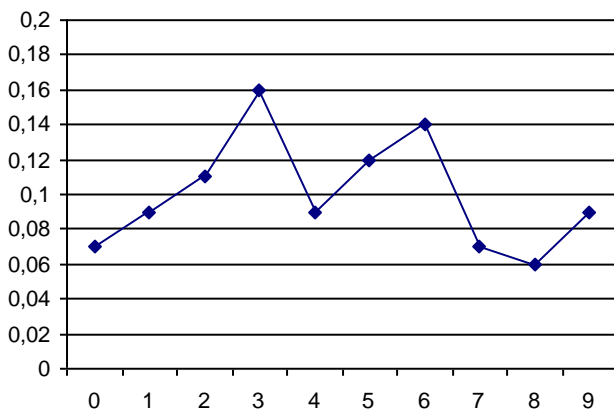
Générateur 1 : fréquence des chiffres de 0 à 9  
pour  $n = 100$  valeurs



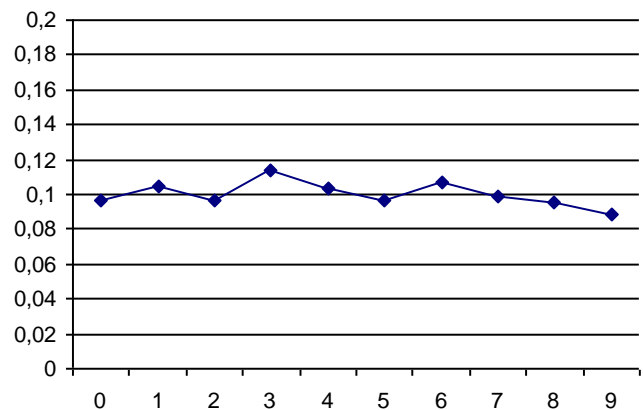
Générateur 1 : fréquence des chiffres de 0 à 9  
pour  $n = 1000$  valeurs



Générateur 2 : fréquence des chiffres de 0 à 9  
pour  $n = 100$  valeurs



Générateur 2 : fréquence des chiffres de 0 à 9  
pour  $n = 1000$  valeurs



On observe que le premier générateur a fourni, pour  $n = 1000$ , un nombre exceptionnellement bas de 8, ce qui, sans le discréditer totalement, le rend un peu suspect.

Un second contrôle peut être celui du *poker*, où l'on regroupe consécutivement les chiffres par quatre et où l'on compare les fréquences observées des différentes configurations possibles à leur probabilité :

Configuration	Chiffres différents : 5872	Une paire : 4849	Deux paires : 7337	Trois chiffres identiques : 5515	Quatre chiffres identiques : 6666
Probabilité	$\frac{10 \times 9 \times 8 \times 7}{10^4} = 0,504$	$C_4^2 \frac{10 \times 9 \times 8}{10^4} = 0,432$	$\frac{C_4^2}{2} \times \frac{10 \times 9}{10^4} = 0,027$	$\frac{4 \times 10 \times 9}{10^4} = 0,036$	$\frac{10}{10^4} = 0,001$

Pour le générateur ALEA() d'Excel, on obtient par exemple, sur deux expériences, et pour 1000 groupes de quatre chiffres, les effectifs  $x_i$  suivants, à comparer aux valeurs théoriques  $t_i$  :

Configuration	4 chiffres différents	Une paire	Deux paires	3 chiffres identiques	4 chiffres identiques
effectifs $x_i$ observés à la 1 <sup>ère</sup> expérience	541	409	18	32	0
effectifs $x_i$ observés à la 2 <sup>ème</sup> expérience	497	439	31	32	1
valeurs théoriques $t_i$	$t_1 = 504$	$t_2 = 432$	$t_3 = 27$	$t_4 = 36$	$t_5 = 1$

Une certaine fluctuation des observations est attendue, mais dans quelles limites ? On peut mesurer l'adéquation des observations  $x_i$  aux valeurs théoriques correspondantes  $t_i$  en

introduisant l'écart quadratique réduit  $\chi_{\text{obs}}^2 = \sum_{i=1}^5 \frac{(x_i - t_i)^2}{t_i}$ .

Pour étudier la variabilité de ce critère, on introduit les variables aléatoires  $X_i$  qui, à chaque échantillon de 1000 groupes de quatre chiffres consécutifs, associent le nombre de configurations de type  $i$ , ainsi que la variable aléatoire  $T = \sum_{i=1}^5 \frac{(X_i - t_i)^2}{t_i}$ , avec

$$\sum_{i=1}^5 X_i = 1000.$$

La loi de  $T$  suit approximativement une loi tabulée et connue sous le nom de loi du  $\chi^2$  à 4 degrés de liberté (en effet la relation ci-dessus fait que la valeur de  $X_5$  est déterminée dès que les valeurs de  $X_1, X_2, X_3$  et  $X_4$  sont connues).

La table permet alors d'obtenir :  $P(T \leq 9,48) \approx 0,95$ .

On pourra alors considérer comme suspect d'observer une valeur de  $\chi_{\text{obs}}^2$  supérieure à 9,48.

Pour les échantillons obtenus précédemment avec le générateur aléatoire d'Excel, on a :

$$\chi_{\text{obs}1}^2 \approx 8,39 \text{ et } \chi_{\text{obs}2}^2 \approx 1,25.$$

Un générateur de nombres aléatoires existe sur toutes les calculatrices sous la forme de la touche *random*<sup>43</sup> : **Ran#** (CASIO) ou **rand** (T. I.).

### Simuler d'autres distributions

A partir de la distribution uniforme sur  $[0, 1]$ , obtenue par la fonction *random*, on a recours à différentes techniques pour simuler toute autre distribution.

A l'aide de la partie entière (notée ici *int*), si l'instruction *rand* simule la loi  $U([0, 1])$ , l'instruction **int(10rand)** simule le tirage d'un chiffre entre 0 et 9, **1 + int(6rand)** le lancer d'un dé et **int(rand + 0.5)** le lancer d'une pièce, codé par 0 ou 1.

### Distribution de Bernoulli

Les résultats d'une variable aléatoire  $X$  suivant la loi de Bernoulli de paramètre  $p$ , notée  $B(1, p)$  (c'est à dire telle que  $P(X = 1) = p$  et  $P(X = 0) = 1 - p$ , avec  $p \in [0, 1]$ ) sont simulés par l'instruction : **int(rand + p)**.

En effet, l'instruction *rand + p* correspond à une distribution uniforme sur  $[p; 1 + p]$  et la partie entière d'un nombre choisi dans cet intervalle est 0 s'il appartient à  $[p; 1[$  et 1 s'il

<sup>43</sup> Le mot *random* signifie "hasard" en anglais, il vient du vieux français *random*, que l'on retrouve dans *randonnée*.



Ce résultat repose sur le *théorème de la limite centrée*. La somme de  $n$  variables aléatoires  $X_i$  uniformes sur  $[0 ; 1]$  et indépendantes suit approximativement, pour  $n$  assez grand, une loi normale.

⇒ Simulation "exacte" :

Une simulation "exacte" de la loi  $N(m; \sigma)$  est possible sous la forme de l'instruction suivante<sup>44</sup> (en mode Degrés) :  $m + \sigma \cos(360 \text{ rand}) \sqrt{-2 \ln \text{ rand}}$ .

L'idée repose sur un changement de variable des coordonnées polaires aux coordonnées cartésiennes. On montre que, si  $U$  et  $V$  sont deux variables aléatoires indépendantes de loi uniforme sur  $[0, 1]$ , alors  $\rho = \sqrt{-2 \ln V}$  et  $\theta = 2\pi U$  sont indépendantes, respectivement de loi exponentielle de paramètre 1/2 et de loi uniforme sur  $[0, 2\pi]$ , puis on obtient que la variable aléatoire  $X = \rho \cos \theta$  suit la loi normale  $N(0, 1)$ .

## Comment justifier la simulation

Est-ce que simuler permet toujours de bonnes conjectures, dans le cadre d'une situation aléatoire ? Combien de fois doit-on répéter les expériences ? Quelle incertitude a-t-on ?

La réponse à ces questions, et la justification de la simulation, est fondée sur la loi des grands nombres, elle-même précisée par le théorème limite central.

La *loi des grands nombres* permet d'affirmer qu'en simulant une expérience aléatoire un grand nombre de fois, de façons indépendantes, les fréquences observées se rapprochent des probabilités à évaluer.

De façon plus précise, on a le théorème suivant :

### **Loi faible des grands nombres :**

Soit un événement  $A$  avec  $P(A) = p$ .

Soit  $X_i$ ,  $1 \leq i \leq n$ , des variables aléatoires de Bernoulli, indépendantes, de paramètre  $p$  ( $X_i$  vaut 1 si  $A$  est réalisé à l'expérience  $i$  et 0 sinon).

On note  $S_n = \sum_{i=1}^{i=n} X_i$  (qui suit la loi binomiale  $B(n, p)$ ) et  $F_n = \frac{1}{n} S_n$ , la variable aléatoire correspondant à la fréquence d'observation de  $A$  sur les  $n$  expériences.

Alors, pour tout  $t > 0$ , 
$$P\left(|F_n - p| > t \sqrt{\frac{p(1-p)}{n}}\right) \leq \frac{1}{t^2}.$$

Prenons l'exemple des 200 lancers de pile ou face, où l'on cherche à évaluer, par simulation, la probabilité de l'évènement  $A$  : "Sur 200 lancers, on a eu au moins une série de 6 lancers consécutifs égaux". On a vu que  $P(A) = p \approx 0,965$ , mais bien sûr, lorsque l'on simule, on ignore cette valeur.

Déterminons, à l'aide de la loi des grands nombres, le nombre  $n$  de simulations de 200 lancers qu'il suffit d'effectuer, pour être pratiquement assuré que la fréquence  $f$  de l'évènement  $A$  sur les  $n$  simulations approchera  $p$  à  $10^{-2}$  près. C'est de la statistique. On n'aura pas de certitude, mais un risque mesuré de se tromper, disons moins d'une chance sur 100.

On cherche donc, avec les notations du théorème, un nombre suffisant  $n$  de simulations tel

que  $P(|F_n - p| > \frac{1}{100}) \leq \frac{1}{100}$ .

Comme la valeur de  $p$ , entre 0 et 1, est inconnue, on peut majorer  $\sqrt{p(1-p)}$  par 1/2.

<sup>44</sup> Certaines calculatrices possèdent une instruction randNorm.

On a alors  $P\left(|F_n - p| > t\sqrt{\frac{p(1-p)}{n}}\right) \leq P\left(|F_n - p| > \frac{t}{2\sqrt{n}}\right) \leq \frac{1}{t^2}$ .

On prend  $t = 10$ , d'où  $\frac{10}{2\sqrt{n}} = \frac{1}{100}$  et  $n = 250\,000$ .

En simulant 250 000 fois 200 lancers, on aura donc au moins 99% de chances d'obtenir une valeur de  $p$  à  $10^{-2}$  près.

En fait, cette valeur de  $n$  est très exagérée, la majoration donnée par la loi des grands nombres, et résultant de l'inégalité de *Bienaymé-Tchebitchev*, étant très grossière.

Le *théorème limite central* permet, lorsque cela est possible, d'utiliser la répartition de la loi normale.

### **Théorème limite central :**

Soit  $X_i$  des variables aléatoires indépendantes, de même loi, de moyenne  $\mu$  et d'écart type  $\sigma$ . Pour  $n$  suffisamment grand, la variable aléatoire

$$\bar{X}_n = \frac{1}{n} S_n = \frac{1}{n} \sum_{i=1}^{i=n} X_i \text{ suit approximativement la loi normale } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Dans l'exemple qui nous concerne, les  $X_i$  sont des variables de *Bernoulli* de même paramètre  $p$  (d'espérance  $p$  et d'écart type  $\sqrt{p(1-p)}$ ) alors, le théorème affirme que la

variable aléatoire  $F_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{i=n} X_i$  (fréquence observée sur un échantillon de taille  $n$ ) suit

approximativement, pour  $n$  assez grand, la loi normale  $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ .

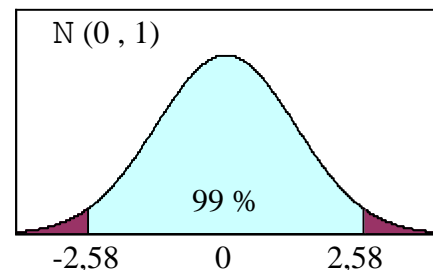
On cherche  $n$  tel que  $P(|F_n - p| \leq 0,01) \approx 0,99$ . On se ramène à la loi normale centrée

réduite (tabulée) en posant  $T = \frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}$ .

On a alors  $P(|F_n - p| \leq 0,01) = P(|T| \leq \frac{0,01}{\sqrt{\frac{p(1-p)}{n}}})$  et

l'on sait que, pour la loi normale centrée réduite,  $P(|T| \leq 2,58) \approx 0,99$ .

On en déduit que  $\sqrt{n} \approx 258\sqrt{p(1-p)} \leq \frac{258}{2}$ , soit  $n$  de l'ordre de 16 600. Ceci est possible, avec la puissance de calcul de l'ordinateur.



## **Ce que peut apporter la simulation à l'enseignement de la statistique**

### **La simulation permet de représenter la probabilité dans son aspect fréquentiste**

Les probabilités définies, dans le cadre de l'équiprobabilité, par le rapport des cas favorables aux cas possibles, ne sont calculables que dans un cadre très limité, grosso modo, celui des jeux de hasard, dont les règles sont déterminées et la modélisation assez



simple. Cette approche est inopérante en statistique. Ce n'est pas par un dénombrement exact que l'assureur évaluera la probabilité de naufrage d'un navire, ou le technicien la probabilité de panne d'une machine.

L'approche fréquentiste, fondée sur la loi des grands nombres de *Bernoulli*, consiste à lier la notion de probabilité à celle de fréquence observée après la répétition un grand nombre de fois d'une expérience<sup>45</sup>. La simulation permet, à peu de frais, d'y parvenir.

*Emile Borel*<sup>46</sup> affirme que "les probabilités doivent être regardées comme analogues à la mesure des grandeurs physiques, c'est à dire qu'elles ne peuvent jamais être connues exactement, mais seulement avec une certaine approximation". Cette démarche statistique, pour évaluer les probabilités "dans la vie", peut être assistée par la simulation. Quand cela est possible, les élèves effectuent quelques expériences réelles (avec une pièce de monnaie, des dés ...), pour se frotter à la réalité et accepter que le générateur de nombres aléatoires la prolonge.

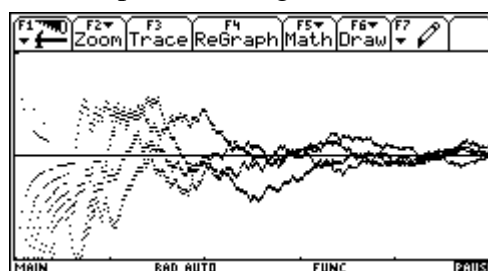
La simulation permet également aux élèves de visualiser cette "convergence" vers la probabilité.

Le programme suivant montre, par exemple, comment, sur 500 lancers de pile ou face (en abscisses) évolue la proportion de "piles" (ici entre 0,4 et 0,6 en ordonnées). Sur l'écran apparaît la trajectoire des fréquences cumulées des piles sur quatre simulations de 500 lancers. On constate l'aspect chaotique des résultats du hasard sur les 100 premiers lancers,

puis une convergence (en  $\frac{1}{\sqrt{n}}$  selon le théorème limite central) vers la probabilité  $p = 1/2$ .

CASIO Graph 25 ou +	T.I. 89 - 92
ViewWindow 0,500,100,0.4, 0.6,0.1	:FnOff
Graph Y=0.5	:ClrDraw
For 1 → J To 4	:PlotsOff
0 → P	:0 → xmin
For 1 → I To 500	:500 → xmax
Int(Ran#+0.5) → A	:100 → xscl
A ≠ 0 ⇒ Goto 1	:0.4 → ymin
P + 1 → P	:0.6 → ymax
Lbl 1	:0.1 → yscl
Plot I, P ÷ I	:DrawFunc 0.5
Next	:For j, 1, 4
Next	:0 → p
	:For i, 1, 500
	:int(rand( )+0.5) → a
	:If a = 0
	:p + 1 → p
	:PtOn i, p/i
	:EndFor
	:EndFor

Exemple d'affichage sur TI 92 :



### **L'épreuve de l'expérience et l'attrait pour les nouvelles technologies**

L'un des principaux intérêts pédagogiques de la simulation réside dans la *nature expérimentale* qu'elle donne à l'enseignement de la statistique et des probabilités, donnant davantage de sens aux concepts et motivant les élèves par l'aspect novateur de cette approche (utilisation des calculatrices programmables et de l'ordinateur). On voit comment la statistique fonctionne et cela rend les formules moins austères.

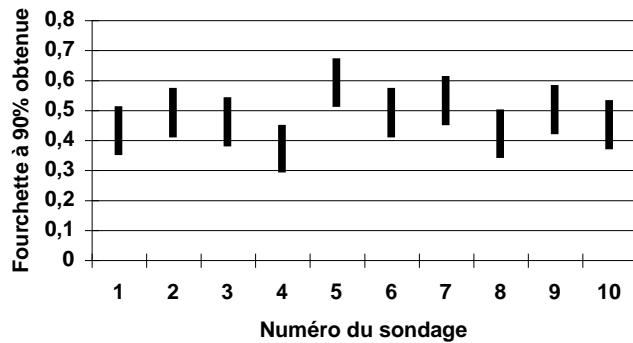
La simulation permet de mettre à l'épreuve de l'expérience certains résultats (parfois admis) du cours. Elle favorise le débat scientifique, obligeant les élèves à confronter leurs

<sup>45</sup> La loi des grands nombres ne permet pas cependant de définir la probabilité, puisque faisant déjà appel à cette notion.

<sup>46</sup> dans "Les probabilités et la vie".

observations et à les analyser. Dans le cadre de travaux pratiques (un peu au sens de la physique), on constate l'efficacité de la théorie, on donne une réalité aux formules. On peut également, en sens inverse, expérimenter d'abord, pour émettre des conjectures ou introduire une notion. Voyons deux exemples, au niveau seconde, puis B.T.S.

### Expérimentation des "fourchettes" de sondage



L'observation des "fourchettes" de sondages permet de comprendre immédiatement leur dépendance par rapport à l'échantillon ("fourchettes" éventuellement disjointes ou très différentes, pourcentage à estimer non nécessairement contenu dans la "fourchette"...). Ces questions qui, posées dans le cadre du cours, poseraient problème, trouvent, grâce à l'expérience, une réponse toute naturelle<sup>47</sup>.

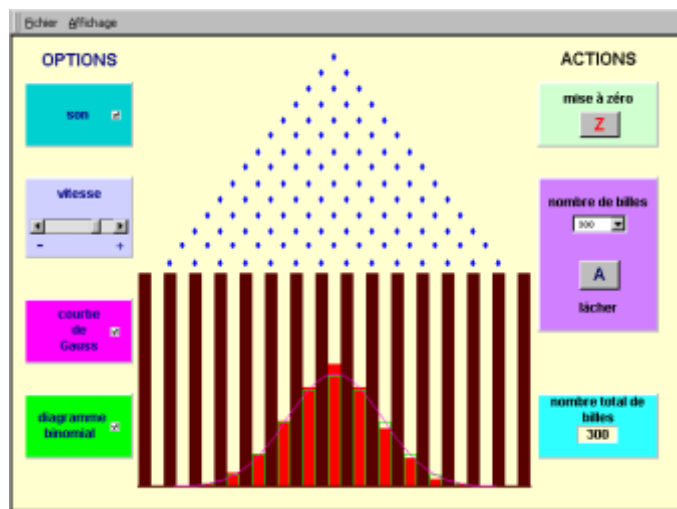
On peut ensuite, profitant par exemple, de l'interactivité du tableur, faire varier le seuil de confiance ou la taille de l'échantillon et en évaluer l'impact sur la qualité des "fourchettes" (taux d'erreur, précision de l'information...).

### Expérimentation du théorème limite central

Dans les sections où la distribution de *Laplace-Gauss* est enseignée, l'élève peut se demander pourquoi, avec une expression analytique paradoxalement compliquée, elle est si répandue, au point d'être qualifiée de "normale" ?

La réponse à cette question est donnée par le théorème limite central. La somme de  $n$  variables aléatoires indépendantes de même loi suit approximativement, pour  $n$  assez grand, une loi normale. Mais ce résultat, alors qu'il est essentiel, est généralement admis. Il est donc instructif de l'expérimenter, par simulation.

Une simulation physique est fournie par la planche de *Galton* (illustration ci-contre).



On peut également utiliser directement la fonction random de la calculatrice.

En admettant que  $n = 12$  est assez grand, l'instruction

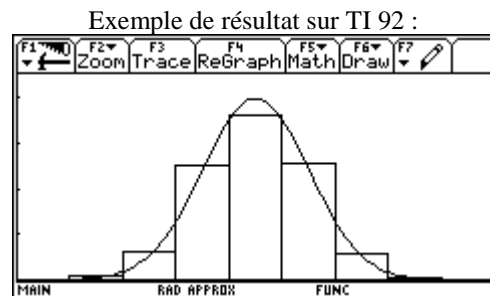
$$\underbrace{\text{Ran\#} + \text{Ran\#} + \dots + \text{Ran\#}}_{12 \text{ fois}} - 6 + 4,5$$

doit approximativement simuler la loi normale  $N(4,5 ; 1)$ .

Le programme suivant regroupe 100 résultats consécutifs de cette instruction en 9 classes :  $[0;1[ ; [1;2[ \dots [8;9[$ , puis compare l'histogramme avec la courbe de densité de la loi normale  $N(4,5 ; 1)$ .

<sup>47</sup> Voir le TP donné en annexe 1.

CASIO Graph 40 ou +	TI 89 - 92
ClrList ↵	:DelVar L1 , L2
Seq(I,I,0,8,1) → List 1 ↵	:seq (i,i,0,8,1) → L1
Seq(0,J,1,9,1) → List 2 ↵	:seq (0,j,1,9,1) → L2
For 1 → K To 100 ↵	:For k , 1 , 100
<b>Ran#+Ran#+Ran#+Ran#+Ran#</b>	<b>:rand()+rand()+rand</b>
<b>n#+Ran#+Ran#+Ran#+Ran#</b>	<b>( )+rand()+rand( )+</b>
<b>+Ran#+Ran#+Ran# - 1.5</b> →	<b>rand()+rand()+rand</b>
N ↵	<b>( )+rand()+rand( )+</b>
1 + Int N → N ↵	<b>rand()+rand( ) - 1.5</b> → n
N ≥ 1 And N ≤ 9 ⇒	:int ( n )+1 → n
List 2[N]+1 → List 2[N]↵	:If n ≥ 1 and n ≤ 9
Next ↵	:L2[n] + 1 → L2[n]
List 2 ↵	:EndFor
S-WindMan ↵	:Disp L2
ViewWindow 0,9,1,0,45,10 ↵	:Pause
0 → Hstart ↵	:0 → xmin
1 → Hpitch ↵	:9 → xmax
S-Gph1 DrawOn,Hist,List	:1 → xscl
1,List 2,Blue ↵	:0 → ymin
DrawStat ↵	:45 → ymax
Graph Y= (100÷	:10 → yscl
$\sqrt{(2\pi)}e^{-.5(X-4.5)^2}$ )	:PlotsOn
	:Newplot 1,4,L1,,
	L2,,,1
	:DrawFunc 100/( $\sqrt{(2\pi)}$
	) $\times e^{-.5(x-4.5)^2}$ )



La conclusion à en tirer est que, lorsqu'un phénomène quelconque subit des variations dues à l'addition d'un grand nombre de perturbations aléatoires indépendantes (sans que l'une d'elle soit dominante), celles-ci suivront approximativement une loi normale.

### Retrouver le hasard et l'ordre

*"De façon apparemment paradoxale, l'accumulation d'événements au hasard aboutit à une répartition parfaitement prévisible des résultats possibles. Le hasard n'est capricieux qu'au coup par coup."*

*"Le Trésor" - M. SERRES et N. FAROUKI,*  
article loi des grands nombres.

Un des principaux effets des activités de simulation est de réintroduire le "hasard" au cœur de notre enseignement de statistique et probabilités, lequel devient trop souvent du "dressage" aux techniques de résolution de problèmes (simple application de formules). Les probabilités ne consistent-elles pas à mettre un peu d'ordre là où le néophyte ne voit que l'intervention du "hasard" ? Dès que l'on a décelé un certain ordre, on peut prévoir. Donnons l'exemple de l'étude de pannes à taux d'avarie constant.

A partir d'un historique statistique de pannes (ici simulé), on recherche la nature de leur loi. Une pièce est supposée avoir un taux d'avarie par heure de 0,007 , c'est à dire que, pour toute durée d'une heure choisie au hasard, la probabilité de panne est 0,007.

CASIO Graph 25 ou +	T.I. 82 83 89 92
0 → I ↓	:0 → I
While Int( Ran# + 0.007 )	:While int( rand + 0.007 )
= 0 ↓	= 0
I + 1 → I ↓	:I + 1 → I
WhileEnd ↓	:End (ou EndWhile)
I	:Disp I

Un simple programme sur calculatrice, permet de simuler le temps de bon fonctionnement de la pièce (le temps n'est pas continu mais "saute" d'heure en heure, et, chaque heure l'instruction :

`int( rand + 0.007 )`

simule la probabilité de panne).

Lorsque les élèves expérimentent, avec ce programme, les temps de bon fonctionnement successifs, ils constatent la très forte présence du hasard. On passe ainsi, par exemple, de 192 heures sans panne, à 37 heures, puis 407...

En regroupant un nombre important d'observations (200 par exemple), en classes, et en reportant les résultats sur papier semi-logarithmique, on verra surgir un ordre... celui du modèle de la loi exponentielle.

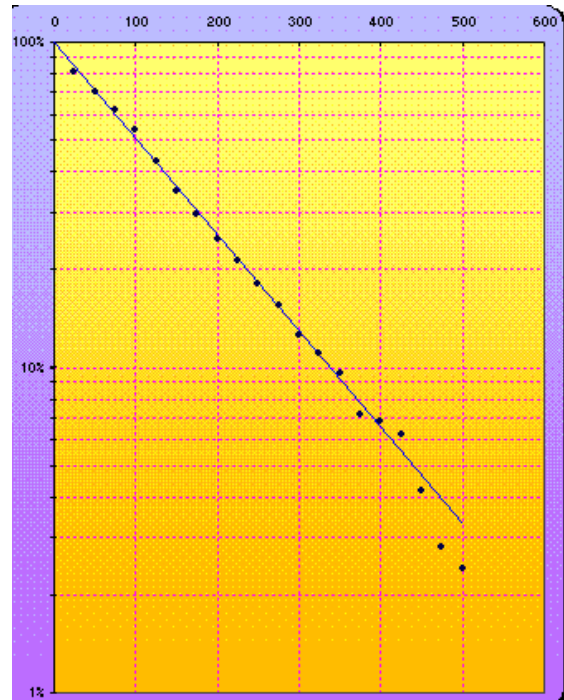
On pourra dès lors déterminer l'espérance des temps de bon fonctionnement de cette pièce.

### **Explorer, quand les outils mathématiques font défaut**

De la même manière que le statisticien, l'économiste ou l'ingénieur exploite la puissance des ordinateurs pour étudier des situations aléatoires pour lesquelles le calcul est impossible ou trop compliqué, on peut, par simulation, explorer avec les élèves des situations riches, pour lesquelles ils ne possèdent pas les outils mathématiques d'un traitement complet.

En seconde, l'étude mathématique des séries de lancers consécutifs à pile ou face n'est pas possible. En revanche leur simulation, outre l'étude des fluctuations d'échantillonnage, mettra en évidence une propriété non triviale du hasard.

En B.T.S. par exemple, l'étude des files d'attente ou de la gestion de stocks, n'est pas au programme. Leur simulation permettra l'étude, dans un contexte intéressant et pratique, de lois figurant au programme (loi de Poisson, exponentielle, normale...).



## **Bibliographie à propos de la simulation et des générateurs aléatoires**

**BOULEAU Nicolas** – *"Probabilités de l'ingénieur : variables aléatoires et simulation"* – Hermann 1986.

**CHAITIN Gregory** – *"Les suites aléatoires"* – Dossier *"Pour la science"* : *"Le hasard"* – Hors série avril 96.

### **Commission Inter-IREM Lycées technologiques**

– *"Simulations d'expériences aléatoires – Une expérimentation du hasard de la première au BTS"* – IREM Paris-Nord - 1998.

– *"Simulation et statistique en seconde"* – IREM Paris-Nord - 2000.

### **Commission Inter-IREM Statistique et probabilités**

– *"Enseigner les probabilités au lycée"* – 1997.

**DELAHAYE Jean-Paul** – *"Aléas du hasard informatique"* – *Pour la science* mars 98.

**DEWDNEY Alexander** – *"Les hasards simulés"* – Dossier *"Pour la science"* : *"Le hasard"* – Hors série avril 96.

**SAPORTA Gilbert** – *"Probabilités, analyse des données et statistique"* – TECHNIP 1990.



## POUR ALLER PLUS LOIN

Dans ce qui précède, la statistique a été définie comme l'art de caractériser une population à partir de mesures effectuées sur tout ou partie des individus composant ladite population, étant bien entendu que les mesures diffèrent d'un individu à un autre. Quand cette variabilité peut être interprétée comme étant due au hasard, hasard modélisable par une loi de probabilité partiellement inconnue, la statistique devient un mode de détermination de celle-ci. On a présenté des modèles supposés rendre compte d'expériences déjà faites, et les procédures mises en oeuvre avaient un caractère d'évidence, qu'il fallait justifier scientifiquement. Quoi de plus intuitif en effet que d'estimer une proportion de mauvaises cartouches dans un lot important par la proportion observée dans un échantillon tiré au hasard ?

Mais les situations réelles sont rarement aussi simples. L'objectif de ce chapitre est de montrer, à partir de quelques exemples, que la statistique inductive est susceptible de répondre à de nombreuses questions. Voici quelques unes d'entre elles, que nous allons développer.

- ❑ Etant donné un modèle probabiliste plus ou moins complexe, chercher les procédures les plus efficaces.
- ❑ Juger de la validité d'un modèle.
- ❑ Se garantir, autant que possible, contre une erreur de modèle en adoptant des procédures peu sensibles ou insensibles à ces erreurs.
- ❑ Faire des prévisions quand le futur dépend du passé selon un certain modèle (c'est le traitement des séries chronologiques).
- ❑ Organiser les expériences à faire pour tirer de ces dernières le plus d'informations possibles.

Ces questions ont été choisies, parmi d'autres, parce que les élèves du second cycle ou les étudiants des classes de techniciens supérieurs, peuvent les rencontrer lors de leurs études. Certaines font partie des programmes des brevets de techniciens supérieurs (BTS) (étude de la fiabilité avec les procédures liées à la loi de *Weibull*, initiation aux plans d'expériences dans les BTS de la branche chimie). D'autres peuvent être rencontrées lors des travaux personnels encadrés en 2<sup>nd</sup> cycle ou lors des stages en entreprise dans les sections de techniciens supérieurs.

Dans le chapitre 4 (la statistique euclidienne), on a vu un exemple de mise en oeuvre de techniques statistiques ne nécessitant pas l'introduction d'un modèle probabiliste. On a parlé à ce propos de statistique descriptive, d'analyse des données. Les méthodes exposées alors partaient du postulat que la géométrie euclidienne était apte à représenter ces données. En dehors de l'exemple étudié, de nombreuses méthodes ont été mises au point pour résumer des données quand la variabilité de celles-ci est difficilement interprétable à l'aide d'un modèle probabiliste. Par exemple, on peut citer les méthodes où l'on partitionne la population étudiée en sous-populations regroupant des individus se ressemblant, méthodes de *classification*

*automatique*, ou bien celles où l'on compare plusieurs caractères, chacun d'entre eux ayant un nombre fini de modalités, *analyse factorielle multiple*.

La plupart des procédures citées ci-dessus sont extrêmement faciles à mettre en oeuvre. Il existe sur le marché de nombreux logiciels intégrant des algorithmes de calculs parfois complexes. Les faire tourner sur des données recueillies auparavant donne toujours une réponse. Reste à connaître le domaine de validité de la méthode dont les procédures de calcul ont été automatisées, puis, si celle-ci est valable, à interpréter cette dernière et à savoir si elle est pertinente. C'est ce qui sera fait pour les exemples ci-dessous.

## I – RECHERCHE DE PROCEDURES EFFICACES

### 1 – Les tests de Student

Considérons les deux problèmes suivants :

(I) Test d'une moyenne :

Les observations faites sont les réalisations de  $n$  variables aléatoires  $X_1, X_2, \dots, X_n$  indépendantes, de même loi : loi normale de moyenne  $\mu$  et d'écart type  $\sigma$ , notée  $N(\mu, \sigma)$ ,  $\mu$  et  $\sigma$  sont inconnus et on veut tester l'hypothèse suivante  $\mu = \mu_0$ ,  $\mu_0$  donnée d'avance.

(II) Test de comparaison de deux moyennes :

On est en présence de deux populations sur lesquelles on réalise des observations qui sont des réalisations de variables aléatoires indépendantes. Celle du premier échantillon seront notées  $X_1, X_2, \dots, X_m$ , elles suivent une loi normale  $N(\mu_1, \sigma)$ , celles du deuxième échantillon seront notées  $X_{m+1}, X_{m+2}, \dots, X_{m+n}$ , elles suivent une loi normale  $N(\mu_2, \sigma)$ ;  $\mu_1, \mu_2, \sigma$  sont inconnus. On veut tester si ces deux populations sont identiques, c'est à dire si  $\mu_1 = \mu_2$ . Il s'agit d'une situation classique quand, par exemple, on veut tester l'efficacité d'un médicament, ou d'un engrais sur le rendement d'une culture donnée.

On sait que si  $X_1, X_2, \dots, X_n$  sont des variables aléatoires indépendantes, identiquement distribuées selon une loi normale  $N(\mu, \sigma)$ , alors  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  suit une loi normale

$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ . Si on veut tester l'hypothèse nulle  $\mu = \mu_0$  contre, par exemple, l'alternative

$\mu > \mu_0$ , il est tentant de comparer  $\bar{X}$  à  $\mu_0$ . Mais  $\bar{X} - \mu_0$  suit une loi normale  $N\left(0, \frac{\sigma}{\sqrt{n}}\right)$

qui n'est pas connue car  $\sigma$  est inconnu. On ne peut pas déterminer le seuil  $x_\alpha$  tel que la probabilité de l'évènement  $\{\bar{X} - \mu \geq x_\alpha\}$  soit inférieure ou égale à  $\alpha$  quand l'hypothèse  $\mu = \mu_0$  est vraie. En statistique il est tentant, quand on a un paramètre  $\sigma$  inconnu, on dit aussi nuisible ou importun, de le remplacer par son estimation.

On sait que la variable aléatoire  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  est un estimateur sans biais de  $\sigma^2$ ,  $E(S^2) = \sigma^2$ , et convergent, pour tout  $\varepsilon > 0$ ,  $P(|S^2 - \sigma^2| > \varepsilon) \xrightarrow{n \rightarrow \infty} 0$ .

L'idée de *Student*, au début du XX<sup>e</sup> siècle, a été d'introduire la variable aléatoire

$$T = \sqrt{n(n-1)} \frac{\bar{X} - \mu_0}{\sqrt{\sum (X_i - \bar{X})^2}}.$$

Il a montré que si  $\mu = \mu_0$  alors  $T$  suit une loi de probabilité, indépendante de  $\sigma$ , que l'on peut calculer et appelée loi de *Student* à  $n - 1$  degrés de liberté. On lit alors dans une table, ou on trouve par la calculatrice, le nombre  $t_{\alpha, n-1}$  tel que :  $P_{\mu_0}(T \geq t_{\alpha, n-1}) = \alpha$ . Avec un risque de  $\alpha$ , on rejette l'hypothèse nulle  $\mu = \mu_0$  si l'observation faite est dans la région critique  $C = \{\omega; T(\omega) \geq t_{\alpha, n-1}\}$ .

La même variable aléatoire  $T$  permet également de trouver un intervalle de confiance pour l'estimation du paramètre  $\mu$  :

on montre que l'intervalle  $\left[ \bar{X} - t_{\frac{\alpha}{2}, n-1} \frac{\sqrt{\sum (X_i - \bar{X})^2}}{\sqrt{n(n-1)}}, \bar{X} + t_{\frac{\alpha}{2}, n-1} \frac{\sqrt{\sum (X_i - \bar{X})^2}}{\sqrt{n(n-1)}} \right]$  est un

intervalle de confiance pour  $\mu$  de niveau  $\alpha$ , c'est à dire que

$$P_{\mu} \left( \bar{X} - t_{\frac{\alpha}{2}, n-1} \frac{\sqrt{\sum (X_i - \bar{X})^2}}{\sqrt{n(n-1)}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2}, n-1} \frac{\sqrt{\sum (X_i - \bar{X})^2}}{\sqrt{n(n-1)}} \right) = 1 - \alpha \quad (\text{on remarquera que}$$

la distribution de  $T$  est symétrique).

On peut montrer que les **procédures** précédentes sont **optimales**, en un sens que nous allons préciser. Pour tester l'hypothèse  $\mu = \mu_0$  contre  $\mu > \mu_0$  au seuil  $\alpha$  (la probabilité de rejeter l'hypothèse nulle quand elle est vraie) utilisons le test de *Student*, basé sur  $T$ , et dont  $C$  est la région critique. Il est de **puissance maximum**. Cela veut dire que l'erreur de deuxième espèce (probabilité d'accepter l'hypothèse  $\mu = \mu_0$  quand elle est fautive, c'est à dire ici quand  $\mu > \mu_0$ ) est la plus petite possible.

Le test de *Student* peut être adapté au problème (II), dit problème à deux échantillons. On définit alors la quantité  $T_2$  par :

$$T_2 = \sqrt{\frac{nm}{n+m}} \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sum_{i=1}^m (X_i - \bar{X}_1)^2 + \sum_{i=1}^n (X_{m+i} - \bar{X}_2)^2}{n+m-2}}} \quad \text{où } \bar{X}_1 \text{ et } \bar{X}_2 \text{ sont les variables}$$

aléatoires correspondant aux moyennes du premier et du deuxième échantillon.

On montre encore que, si  $\mu_1 = \mu_2$ ,  $T_2$  suit, quel que soit  $\mu_1 = \mu_2$  et  $\sigma$ , une loi de *Student* à  $n + m - 2$  degrés de liberté. Si on teste l'hypothèse  $\mu_1 = \mu_2$  contre l'hypothèse  $\mu_1 > \mu_2$ , on rejettera l'hypothèse nulle si l'observation faite  $\omega$  est telle que  $T_2(\omega) \geq t_{\alpha, n+m-2}$  où  $t_{\alpha, n+m-2}$  sera lu dans la table. Comme pour le problème (I), on peut montrer que ce test est le meilleur possible.

## 2 – Le problème de *Behrens-Fisher*

Il n'est pas toujours possible de trouver une procédure statistique qui soit optimale, comme dans le cas précédent, quand on est en présence de paramètres inconnus. Reprenons le problème (II) en élargissant les hypothèses. Le premier échantillon est constitué de la réalisation de  $m$  variables aléatoires  $X_1, X_2, \dots, X_m$  suivant une loi normale  $N(\mu_1, \sigma_1)$ . De même pour le second échantillon  $X_{m+1}, X_{m+2}, \dots, X_{m+n}$  mais avec une loi normale  $N(\mu_2, \sigma_2)$  avec, ici,  $\sigma_1 \neq \sigma_2$  et inconnus. On veut tester l'hypothèse  $\mu_1 = \mu_2$  contre l'alternative  $\mu_1 > \mu_2$  au seuil de  $\alpha$ .



On peut évidemment trouver des procédures raisonnables, qui, dans la pratique, donnent satisfaction, mais *Linnick*, qui était professeur à Leningrad, a montré vers 1970 qu'il n'existait pas de procédure optimale valide quel que soit  $\sigma_1$  et  $\sigma_2$ ,  $\sigma_1 \neq \sigma_2$ .

La démonstration, fort longue, fait appel à des théorèmes d'*Henri Cartan* sur les fonctions de variable complexe.

### 3 – Le papier *Weibull*

La procédure du papier *Weibull* est au programme du B.T.S. maintenance industrielle<sup>48</sup>.

Quand on cherche à modéliser la **durée de vie** d'un équipement industriel, on utilise une loi de probabilité, dite loi de *Weibull*. Soit  $S$  la variable aléatoire correspondant à la durée de vie de l'équipement. On peut très souvent écrire la probabilité que cette durée de vie

dépasse le temps  $x$  sous la forme :  $P(S > x) = e^{-\left(\frac{x-\gamma}{\eta}\right)^\beta}$  où  $\gamma$ ,  $\beta$  et  $\eta$  sont des paramètres à estimer à partir de  $n$  durées de vie  $x_1, x_2, \dots, x_n$  observées, considérées comme réalisations de  $n$  variables aléatoires  $X_1, X_2, \dots, X_n$  indépendantes de même loi que  $S$ .

Pour simplifier, supposons que  $\gamma = 0$ .

Notons  $F_{\beta, \eta}(x) = P(S \leq x)$  et  $\bar{F}_{\beta, \eta}(x) = 1 - F_{\beta, \eta}(x) = P(S > x)$ .

La densité de probabilité  $f$  de la variable aléatoire  $S$  peut s'écrire :

$$f(x, \beta, \eta) = \frac{d}{dx} F_{\beta, \eta}(x) = \frac{d}{dx} \left( 1 - e^{-\left(\frac{x}{\eta}\right)^\beta} \right) = \frac{\beta}{\eta} \left(\frac{x}{\eta}\right)^{\beta-1} e^{-\left(\frac{x}{\eta}\right)^\beta} \quad (\text{loi de Weibull}).$$

Il existe une méthode pour trouver des estimateurs de paramètres inconnus qui, en général, donne des estimateurs qui ont les qualités requises, c'est la méthode du **maximum de vraisemblance**. On cherche les valeurs de  $\beta$  et de  $\eta$  qui maximisent la probabilité d'observation des valeurs  $x_1, \dots, x_n$  (réellement observées), c'est à dire qui maximisent le produit (en raison de l'indépendance des  $X_i$ )  $f(x_1) \times f(x_2) \times \dots \times f(x_n)$ . En passant au logarithme, il s'agit de déterminer  $\beta$  et  $\eta$  de sorte à maximiser la quantité :

$$L(\beta, \eta, x_1, \dots, x_n) = \sum_{i=1}^n \ln f(x_i, \beta, \eta).$$

La solution peut être obtenue en résolvant en  $\beta$  et  $\eta$  le système de deux équations à deux inconnues  $\frac{\partial L}{\partial \beta} = 0$  ;  $\frac{\partial L}{\partial \eta} = 0$ . Mais il y a une difficulté : il est impossible de trouver des

expressions explicites des solutions du type  $\beta = \hat{\beta}(x_1, \dots, x_n)$ ,  $\eta = \hat{\eta}(x_1, \dots, x_n)$ .

En revanche,  $x_1, \dots, x_n$  étant donnés, il est possible de trouver, par des méthodes numériques, des solutions approchées qui donnent des estimations de  $\beta$  et  $\eta$  pour le constat expérimental fait.

Il existe maintenant des logiciels statistiques fournissant ces solutions. Il n'en était pas de même il y a quelques années. Il fallait pourtant trouver une solution et donc mettre au point une méthode pratique. L'inventeur du papier *Weibull* est parti d'une toute autre idée.

Appelons  $\bar{F}_n(x)$  la fréquence observée des temps de survie supérieurs ou égaux à  $x$  :

$$\bar{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{[x, +\infty[}(x_i).$$

<sup>48</sup> Pour davantage d'information, on peut consulter la brochure "A propos de fiabilité" de la C.I.I. Lycées technologiques.

La loi des grands nombres garantit que la variable aléatoire dont  $\bar{F}_n(x)$  est une réalisation converge vers  $\bar{F}_{\beta,\eta}(x)$  en probabilité. Donc si  $n$  est grand  $\bar{F}_n$  aura de grandes chances d'être proche de  $\bar{F}_{\beta,\eta}$ , correspondant à la loi régissant le phénomène. D'où l'idée : cherchons  $\beta$  et  $\eta$  tels que la courbe de  $\bar{F}_{\beta,\eta}$  relative à ces valeurs soit le plus proche possible de  $\bar{F}_n$  et cela "à vue d'œil". Mais cet organe est surtout sensible aux alignements. On va donc faire une anamorphose pour transformer les courbes  $\bar{F}_{\beta,\eta}$  en droites.

On peut écrire :  $\ln(-\ln \bar{F}_{\beta,\eta}) = \beta \ln x - \beta \ln \eta$ . En posant  $y = \ln(-\ln \bar{F}_{\beta,\eta})$  et  $t = \ln x$ , on a une liaison entre  $y$  et  $t$ , représentable par une droite.

Réordonnons les observations  $x_1, \dots, x_n$  selon les valeurs croissantes, elles seront alors notées  $x_{(1)}, \dots, x_{(n)}$  avec  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . La fonction  $\bar{F}_n$ , pour les fréquences observées, est alors telle que  $\bar{F}_n(x_{(i)}) = \frac{n-i}{n}$  et constante entre  $x_{(i)}$  et  $x_{(i+1)}$ .

Reportons sur le graphique les points  $M_i$  de coordonnées  $(t_i = \ln x_{(i)} ; y_i = \ln(-\ln(1 - \frac{i}{n})))$ . Si ces points ne sont pas trop éloignés d'une droite, alors on admet que le modèle de *Weibull* avec  $\gamma = 0$  est adapté à la situation. Traçons au jugé cette droite, la pente fournit une estimation de  $\beta$  et l'abscisse du point de cette droite d'ordonnée  $y = 0$  fournit une estimation du paramètre  $\eta$ .

Les qualités des estimateurs précédents ne peuvent être calculées, mais, par simulation, on peut en avoir une idée. De même, par des méthodes dites de rééchantillonnage, on peut obtenir des intervalles de confiance des paramètres  $\beta$  et  $\eta$ .

Malgré l'existence de logiciels performants, pour des raisons pratiques, la méthode graphique est toujours utilisée, ce qui justifie sa présence dans les programmes.

## II – TESTER LA VALIDITE DU MODELE

Dans les problèmes qui ont servi à illustrer la méthode statistique, on a très souvent supposé que les observations  $x_1, \dots, x_n$  étaient des réalisations de variables aléatoires  $X_1, \dots, X_n$  indépendantes et de même loi, par exemple une loi normale  $N(\mu, \sigma)$ . Il s'agit d'un modèle, mais que vaut ce modèle ? On peut poser la question autrement : est-il vraisemblable que les données recueillies soient des réalisations de variables aléatoires indépendantes, de même loi, loi normale  $N(\mu, \sigma)$  ?

Il est intuitif que si l'on observe  $x_1 < x_2 < \dots < x_{n-1} < x_n$ , le modèle est peu vraisemblable dès que  $n \geq 5$ , car la probabilité d'une telle observation est  $\frac{1}{n!}$  (il y a  $n!$  rangements

possibles). Dans ce cas, c'est l'indépendance ou la notion de même loi qui est à mettre en cause, mais on ne peut alors a priori rien dire de l'hypothèse de normalité. Cela veut dire qu'il n'est pas possible, à partir d'un échantillon, de répondre globalement à la question posée, mais que l'on peut apporter des réponses partielles.

Supposons que les conditions de recueil des données soient telles que nous sommes assurés qu'il s'agit de réalisations de variables aléatoires indépendantes et de même loi et que l'interrogation porte sur le caractère gaussien de la loi. Il s'agit d'une question fréquemment posée. On a vu précédemment, quand a été explicitée la technique du papier *Weibull*, que l'on pouvait déjà juger de la validité de l'hypothèse : "la loi de survie est une loi de

*Weibull*". Une technique analogue, d'ajustement linéaire, est encore très usitée pour la loi normale, elle porte le nom de **droite de Henry**<sup>49</sup>.

On utilise pour cela un papier millimétré particulier dit papier gaussio-arithmétique. On

appellera  $\phi$  la fonction de répartition de la loi normale  $N(0, 1)$  :  $\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$  et

$\phi^{-1} : ]0, 1[ \rightarrow \mathbb{R}$  la fonction réciproque.

Sur le graphique, le point noté  $(x, y)$  sera de coordonnées  $(x, z)$  avec  $z = \phi^{-1}(y)$ . Avec une telle anamorphose de l'axe des ordonnées, on peut représenter par une droite la fonction de répartition  $F$  d'une loi normale  $N(\mu, \sigma)$ . En effet, celle-ci s'écrit  $y = F(x) = \phi\left(\frac{x-\mu}{\sigma}\right)$  et on

a donc  $z = \phi^{-1}(y) = \phi^{-1} \circ \phi\left(\frac{x-\mu}{\sigma}\right) = \frac{x-\mu}{\sigma}$ .

La pente de la droite est donc  $\frac{1}{\sigma}$  et elle coupe l'axe des abscisses au point d'abscisse  $\mu$ .

Considérons maintenant la fonction de répartition de l'échantillon, correspondant aux fréquences cumulées observées :  $F_n(x)$  est la proportion d'observations inférieures ou

égales à  $x$ ,  $F_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{]-\infty, x]}(x_i)$ .

En notant encore  $F_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{]-\infty, x]}(X_i)$  la variable aléatoire correspondant à la réalisation précédente, on sait que si  $F$  est la fonction de répartition de la loi commune des variables aléatoires  $X_i$  dont  $x_i$  est la réalisation, on a :  $P[\sup_x |F_n(x) - F(x)| > \varepsilon] \xrightarrow{n \rightarrow \infty} 0$ .

Donc, si  $n$  est suffisamment grand,  $F_n$  est probablement proche de  $F$ .

On va représenter  $F_n$  sur le papier gaussio-arithmétique. Comme précédemment, réordonnons les observations selon les valeurs croissantes. Elles seront notées  $x_{(1)}, \dots, x_{(n)}$

avec  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . On a alors  $F_n(x_{(i)}) = \frac{i}{n}$ .

Représentons sur le graphique le point  $M_i$  de coordonnées notées  $(x_{(i)}, \frac{i}{n})$  et donc de

coordonnées réelles  $(x_{(i)}, \phi^{-1}\left(\frac{i}{n}\right))$ , pour  $i$  variant de 1 à  $n-1$ . Si ces points sont à peu près

alignés, alors le modèle gaussien est vraisemblablement adapté et de plus la droite la plus proche du nuage des points, tracée au jugé, permet une première estimation de  $\mu$  et de  $\sigma$ , en exploitant une remarque précédente.

La technique graphique qui vient d'être exposée est ancienne, commode, mais très empirique et non quantifiée. L'appréciation de la qualité de l'ajustement est subjective. Les statisticiens ont mis au point des procédures pour tester l'hypothèse : "les  $n$  variables aléatoires supposées indépendantes et de même loi sont gaussiennes", contre l'alternative qu'elles ne le sont pas. Il s'agit alors, au sens précis du terme, d'un test statistique, avec un niveau (erreur de première espèce) précisé à l'avance. Ces procédures (test du  $\chi^2$ , test de *Kolmogorov-Smirnov*, ou encore, plus spécialisés, test d'*Anderson*, test de *Shapiro* et *Wilks*) remplacent peu à peu la droite de *Henry* et sont intégrées dans certains logiciels<sup>50</sup>.

<sup>49</sup> Voir également, à ce propos, les activités pédagogiques proposées en annexe 3 et 5.

<sup>50</sup> Un exposé des test du  $\chi^2$  et de *Kolmogorov-Smirnov* se trouve dans le livre de *Saporta* – "Probabilités, analyse des données et statistique".

### III – ET SI LE MODELE N'EST PAS ADEQUAT ?

Dans ce qui suit, on supposera toujours que les observations faites sont des réalisations de variables aléatoires indépendantes et de même loi, mais que celle-ci n'est pas toujours connue.

On discutera, à titre d'exemple, des deux cas :

(I)  $X_1, \dots, X_n$  sont des variables aléatoires indépendantes de même loi supposée être une loi normale  $N(\mu, \sigma)$ . En réalité, le phénomène est légèrement perturbé et la vraie loi n'est pas normale mais elle en est proche. On veut estimer  $\mu$ . Que deviennent les procédures mises au point pour la loi normale ?

(II) Soit le problème à deux échantillons pour, par exemple, tester l'efficacité d'un traitement agricole. Les  $m$  variables aléatoires  $X_1, \dots, X_m$  sont indépendantes de fonction de répartition  $F$  et  $X_{m+1}, \dots, X_{m+n}$  sont  $n$  variables aléatoires indépendantes de fonction de répartition  $G$  définie, pour tout réel  $x$ , par  $G(x) = F(x - \Delta)$ .

Les  $X_i$ , pour  $i \leq m$ , représentent le rendement de parcelles non traitées, et les  $X_i$ , pour  $i \geq m + 1$ , représentent le rendement de parcelles traitées.

Si le traitement est inefficace, alors  $\Delta = 0$ , s'il est efficace, le rendement est supérieur et  $\Delta > 0$ . On veut donc tester l'hypothèse  $\Delta = 0$  contre l'alternative  $\Delta > 0$ .

#### 1 – Robustesse des procédures

Prenons le premier exemple et supposons qu'au phénomène étudié se superpose une perturbation qui se produit rarement (avec une probabilité  $\varepsilon$ ,  $\varepsilon \ll 1$ ), mais qui est très

dispersée : sa variance est  $\tau$  avec  $\frac{\tau}{\sigma} \gg 1$ . La loi du phénomène admet donc une densité  $f$

$$\text{avec } f(x) = \frac{1}{\sqrt{2\pi}} \left( \frac{1-\varepsilon}{\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} + \frac{\varepsilon}{\tau} e^{-\frac{1}{2}\left(\frac{x-\mu}{\tau}\right)^2} \right).$$

On veut estimer  $\mu$ . Si  $\varepsilon = 0$ , on sait que le meilleur estimateur est la **moyenne**  $\bar{X}$ . Quand il y a perturbation,  $\bar{X}$  est encore un estimateur convergeant sans biais :  $E(\bar{X}) = \mu$ , mais sa variance est augmentée de la quantité  $\frac{\varepsilon(\tau^2 - \sigma^2)}{n}$  :  $Var \bar{X} = \frac{1}{n}[\sigma^2 + \varepsilon(\tau^2 - \sigma^2)]$ .

Le théorème limite central nous garantit que si  $n$  est grand, la loi de  $\bar{X}$  est à peu près gaussienne. D'autres estimateurs convergents de  $\mu$  peuvent être fournis. Comme ils sont, eux aussi, asymptotiquement gaussiens, on peut, quand  $n$  est grand, les comparer en comparant leurs variances. Ainsi, on peut montrer que la variance de la **médiane** des  $X_i$

$$\text{notée } Med_i X_i, \text{ est à peu près } Var(Med_i X_i) \approx \frac{1}{n} \times \frac{\pi}{2} \frac{\sigma^2}{\left[1 - \varepsilon \left(1 - \frac{\sigma}{\tau}\right)\right]^2}.$$

Ainsi, si  $\varepsilon = 0,05$  et  $\tau = 5\sigma$ , on a :

$$Var \bar{X} = \frac{1}{n} \times 2,2\sigma^2 \text{ et } Var(Med_i X_i) \approx \frac{1}{n} \times 1,6\sigma^2.$$

Même avec une perturbation très faible (5 %), la moyenne n'est plus l'estimateur optimal et de plus il se dégrade très vite. On peut, pour s'en apercevoir, refaire le calcul avec  $\varepsilon = 0,10$ . On peut donner de ce phénomène une explication intuitive. La perturbation est très dispersée, elle peut donc donner une valeur très grande en valeur absolue. Certes elle est de probabilité faible, mais comme on répète l'expérience, la probabilité d'obtenir une grande

valeur (en valeur absolue), appelée **valeur aberrante**, est grande. On sait que la moyenne est sensible à ces valeurs, d'où la dégradation des performances de l'estimateur moyenne. Intuitivement, les agronomes français du XVIII<sup>e</sup> siècle, cités au chapitre 2 paragraphe IV, s'en étaient aperçus. C'est pour cela qu'ils estimaient la moyenne  $\mu$ , non avec la moyenne de toutes les observations, mais avec la **moyenne tronquée**.

Réordonnons les variables  $X_1, \dots, X_n$  suivant les valeurs croissantes. On obtient la statistique d'ordre  $X^{(1)}, \dots, X^{(n)}$  avec  $X^{(1)} \leq X^{(2)} \leq \dots \leq X^{(n)}$ . Ainsi  $X^{(1)} = \min\{X_1, \dots, X_n\}$  ;  $X^{(2)} = \min\{X_1, \dots, X_n\} - \{X^{(1)}\}$  etc...

La moyenne tronquée d'ordre  $\alpha$ , notée  $\bar{X}_\alpha$ , s'écrit : 
$$\bar{X}_\alpha = \frac{1}{n(1-\alpha)} \sum_{i=1+\lfloor n\frac{\alpha}{2} \rfloor}^{n-\lfloor n\frac{\alpha}{2} \rfloor} X^{(i)}$$

(pour  $\alpha = 0,05$  par exemple, on supprime 5 % des valeurs).

La variable aléatoire  $\bar{X}_\alpha$  est un estimateur de  $\mu$  qui reste de bonne qualité en l'absence de perturbation, et qui ne se dégrade pas comme  $\bar{X}$  en présence de celle-ci. On dit que  $\bar{X}_\alpha$  est un **estimateur robuste**.

## 2 – Une procédure non paramétrique

Regardons maintenant le problème (II), celui du rendement agricole, avec une fonction de répartition  $F$  en l'absence de traitement et  $G$  telle que  $G(x) = F(x - \Delta)$  avec traitement.

A la suite de **Franck Wilcoxon** (1945), introduisons la statistique  $W$  suivante, correspondant au nombre de couples d'observations dont la valeur sans traitement est

supérieure à celle avec traitement : 
$$W = \sum_{i=1}^m \sum_{j=1}^n u(X_i - X_{m+j})$$

où  $u$  est la fonction échelon définie par  $u(x) = 1$  si  $x \geq 0$  et  $u(x) = 0$  si  $x < 0$ .

Si l'hypothèse nulle  $\Delta = 0$  est vraie, alors  $G = F$ ,  $X_i$  et  $X_{m+j}$  suivent la même loi et

$$P_0[u(X_i - X_{m+j}) = 1] = P_0[u(X_i - X_{m+j}) = 0] = \frac{1}{2}.$$

En extrapolant, on devine, et cela se démontre, que si  $\Delta = 0$ , la loi de probabilité de  $W$  est indépendante de  $F$ . Si  $\Delta$  est strictement positif (traitement efficace), les faibles valeurs de

$W$  sont les plus probables car alors on a  $P_\Delta[u(X_i - X_{m+j}) = 1] > \frac{1}{2}$ .

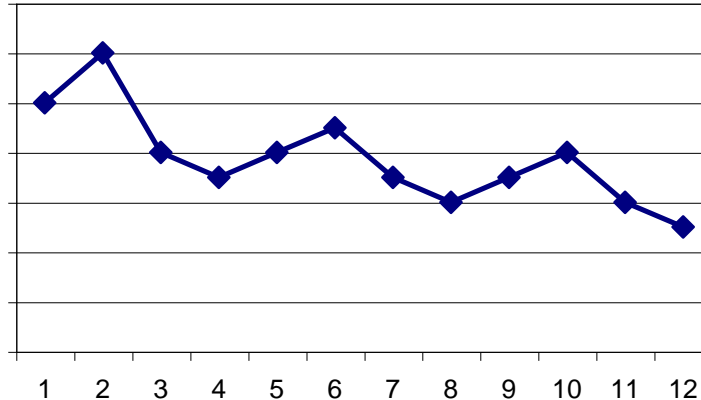
Sur ces bases, on peut construire un test de niveau  $\alpha$  (la probabilité de rejeter l'hypothèse  $\Delta = 0$  quand elle est vraie est égale à  $\alpha$ ) quelle que soit la loi de probabilité  $F$  régissant le phénomène. Une telle procédure est dite **non paramétrique** car  $F$  n'est pas décrite à l'aide de paramètres réels tels que  $\mu$  et  $\sigma$ , comme lorsqu'on supposait que  $F$  était une loi normale  $N(\mu, \sigma)$ .

Si  $\Delta = 0$ , la loi de  $W$  est tabulée. On cherche donc le nombre  $w(\alpha, m, n)$  tel que  $P_{\Delta=0}(W \leq w(\alpha, m, n)) = \alpha$ . On rejettera alors l'hypothèse  $\Delta = 0$  si le constat expérimental  $\omega = (x_1, \dots, x_m, x_{m+1}, \dots, x_{m+n})$  est tel que  $W(\omega) \leq w(\alpha, m, n)$ , en effet, si  $W(\omega)$  est petit, cela signifie que très peu d'observations sans traitement sont supérieures aux observations avec traitement.

On peut s'interroger sur la performance d'une telle procédure. Evidemment, si  $F$  correspond à une loi normale  $N(\mu, \sigma)$ , ce test est moins bon que le test de *Student* vu précédemment, qui est optimal, mais on montre que la perte d'efficacité est faible. En revanche, si  $F$  ne correspond pas à une loi gaussienne, alors la procédure de *Wilcoxon* est plus performante que celle de *Student* appliquée à tort, et parfois de beaucoup. C'est pour cela que, récemment, ces techniques non paramétriques se sont développées.

## IV – LES SERIES CHRONOLOGIQUES

Dès que l'on étudie les variations d'un phénomène au cours du temps<sup>51</sup>, comme par exemple le taux de chômage avec le trimestre comme unité de temps, ou les ventes annuelles d'un produit donné, le modèle des répétitions indépendantes d'un même phénomène ne s'applique plus. D'autres représentations du réel sont à mettre en oeuvre. Prenons l'exemple du taux de chômage, mesuré à la fin de chaque trimestre. Si on représente graphiquement les données, on obtient un graphique du type suivant.



Clairement, on repère une tendance à la baisse sur le long terme, mais aussi une composante saisonnière : les trimestres 2, 6, 10 correspondent à des pointes de chômage, tandis que les trimestres 4, 8, 12 correspondent à des creux.

D'où l'idée de poser le modèle<sup>52</sup> suivant :  $X_t = f_t + s_t + \varepsilon_t$ , où  $f_t$  représente la **tendance**,  $s_t$  la **saison** et  $\varepsilon_t$  un **aléa** de moyenne nulle.

La saison correspond ici au trimestre dans l'année, ce peut être toute autre unité. Par définition, on a  $s_i = s_{i+4k}$  et, pour assurer l'unicité du modèle, on pose  $s_1 + s_2 + s_3 + s_4 = 0$ . Très souvent, on suppose les aléas indépendants et de même loi. De plus,  $f_t$  représentant la tendance à long terme, on suppose donc que  $f_t$  varie lentement avec  $t$ .

### La moyenne mobile

On a souvent besoin de connaître l'importance de la composante saisonnière, afin de pouvoir donner des données "corrigées des variations saisonnières" comme cela se lit dans les pages économiques de diverses publications. Pour le modèle que nous avons posé, on utilise très souvent la méthode de la **moyenne mobile**.

On compare la chronique  $(X_t)_{t \in \{1, \dots, 4n\}}$  avec la chronique  $(Y_t)_{t \in \{3, \dots, 4n-1\}}$  définie par :

$$Y_t = \frac{1}{4}(X_{t-2} + X_{t-1} + X_t + X_{t+1}).$$

Plus généralement, s'il y a, non pas 4, mais  $2p$  saisons, on pose  $Y_t = \frac{1}{2p} \sum_{j=-p}^{p-1} X_{t+j}$ .

S'il y a  $2p + 1$  saisons, on pose  $Y_t = \frac{1}{2p+1} \sum_{j=-p}^p X_{t+j}$ .

Cherchons l'effet du "filtre" moyenne mobile sur la chronique de départ  $X_t = f_t + s_t + \varepsilon_t$ .

$$\text{On a } Y_t = \frac{1}{4}(f_{t-2} + f_{t-1} + f_t + f_{t+1}) + \frac{1}{4}(\varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t + \varepsilon_{t+1})$$

<sup>51</sup> A propos des séries chronologiques, on peut trouver des compléments dans *Wonnacott – "Statistique"*.

<sup>52</sup> On adopte ici un modèle additif.

puisque, par définition,  $s_{t-2} + s_{t-1} + s_t + s_{t+1} = 0$ .

Comme  $f_t$  varie peu en fonction du temps, on a  $\frac{1}{4}(f_{t-2} + f_{t-1} + f_t + f_{t+1}) \approx f_t$ .

Si l'on suppose les aléas  $\varepsilon_t$  de même loi avec une moyenne nulle, un écart type  $\sigma$  et indépendants, alors les aléas  $\eta_t = \frac{1}{4}(\varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t + \varepsilon_{t+1})$  ont une moyenne nulle et un écart type  $\frac{\sigma}{2}$ , plus petit que  $\sigma$ .

Les fluctuations aléatoires de  $Y_t$  sont donc plus faibles que celles de  $X_t$ , mais les aléas  $\eta_t$  ne sont plus indépendants.

Il résulte de cela que :  $X_t - Y_t \approx s_t + (\varepsilon_t - \eta_t)$ .

On peut donc avoir un estimateur de la composante saisonnière en faisant la moyenne des  $X_t - Y_t$  pour les  $t$  appartenant à une saison donnée  $j$ , où  $j \in \{1, 2, 3, 4\}$  :

$$\hat{s}_j = \frac{1}{n} \sum_{k=0}^{n-1} (X_{j+4k} - Y_{j+4k}).$$

L'espérance mathématique et la variance de  $\hat{s}_j$  sont calculables.

On appelle série corrigée des variations saisonnières la chronique :  $\hat{X}_t = X_t - \hat{s}_t$  avec  $t \in \{1, 2, \dots, 4n\}$ .

La chronique ( $Y_t$ ) des moyennes mobiles peut être interprétée comme une estimation de la tendance  $(f_t)_{t \in \{3, \dots, 4n-1\}}$ , en l'absence d'autres informations.

## La tendance

Il existe des situations où il est possible de proposer pour la tendance un certain type de fonction.

On peut avoir  $f_t = a + bt$  ou  $f_t = a + bt + ct^2$ .

La chronique désaisonnalisée  $\hat{X}_t = X_t - \hat{s}_t$ , est alors utilisée pour estimer les coefficients  $a$  et  $b$  ou  $a$ ,  $b$  et  $c$ . Dans le premier cas, on peut utiliser la méthode de régression linéaire, telle qu'elle figure dans les programmes de ES.

## V – ORGANISER LES EXPERIENCES

Tous les expérimentateurs savent qu'une mesure expérimentale est toujours entachée d'une erreur. Si  $x$  est la mesure faite,  $\mu$  la mesure exacte et  $e$  l'erreur, on écrit :  $x = \mu + e$ .

L'erreur<sup>53</sup> est par nature aléatoire et, le plus souvent, on suppose que  $e$  est une variable aléatoire suivant une loi normale  $N(0, \sigma)$  où  $\sigma$  caractérise la précision de la mesure.

Quand un phénomène dépend de  $k$  variables et de  $p$  paramètres inconnus, selon une formule connue, et qu'il donne une réponse numérique  $y$ , en l'absence d'erreur, il suffirait de faire  $p$  expériences différentes en faisant varier les valeurs données aux  $k$  variables, et on aurait un système de  $p$  équations à  $p$  inconnues, qu'il suffirait de résoudre pour connaître les valeurs des paramètres. Mais la présence de l'erreur interdit de telles pratiques. Pour augmenter la précision de l'estimation des paramètres, on peut multiplier les expériences. Mais il y a une limite. Les expériences peuvent coûter cher en argent, songeons aux essais destructifs, ou en temps, comme en agronomie.

<sup>53</sup> On exclut l'erreur systématique, aisément décelable.

D'où l'idée directrice de ce paragraphe : comment organiser les expériences pour tirer de celles qui sont faites un maximum d'informations ? Comme l'erreur est supposée aléatoire, c'est la statistique inductive qui sera le cadre de référence.

## La pesée de trois objets

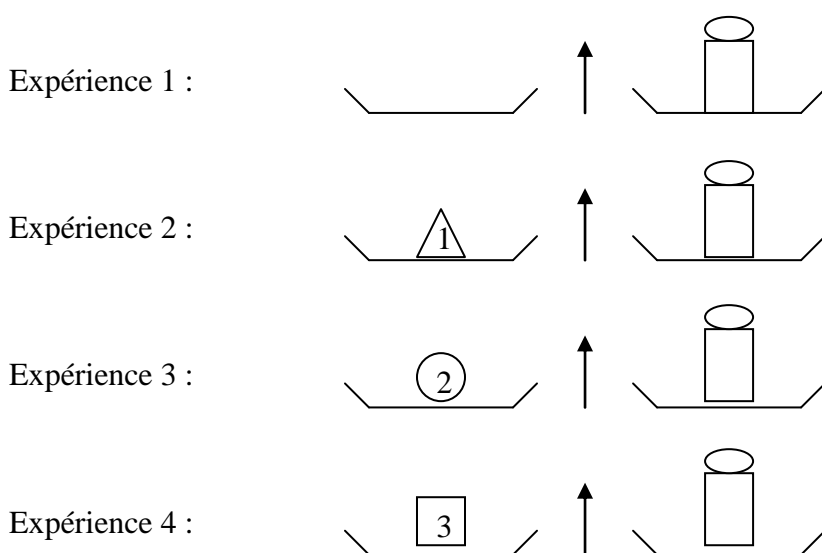
Pour illustrer la problématique du *plan d'expérience*<sup>54</sup>, nous prendrons un exemple simple. On dispose de trois objets notés 1, 2, 3, d'une balance *Roberval* fautive (les plateaux ne sont pas exactement bien équilibrés) et nous pouvons faire quatre pesées. Chaque pesée est faite avec une erreur modélisable par une loi normale  $N(0, \sigma)$ . Comment faire les pesées pour obtenir des estimations des trois masses avec le maximum de précision ?

Appelons  $y_1, y_2, y_3, y_4$  les résultats des pesées 1, 2, 3, 4. Notons  $m_0$  la masse (inconnue) qu'il faut mettre pour équilibrer les deux plateaux, et  $m_1, m_2, m_3$  les masses (inconnues) des objets 1, 2, 3. On désignera par  $\hat{m}_i$  l'estimateur de la masse  $m_i$  pour  $i \in \{1, 2, 3\}$ .

### Première stratégie (ou 1<sup>er</sup> plan d'expérience)

L'idée la plus simple consiste à faire une expérience à vide, pour connaître le déséquilibre des plateaux, puis de peser successivement chaque objet.

La figure ci-dessous illustre le procédé :



On peut aussi représenter cette stratégie par une matrice. La ligne  $i$  représentera l'expérience, avec  $i \in \{1, 2, 3, 4\}$ . La colonne  $j$  représentera l'objet  $j$ , avec  $j \in \{0, 1, 2, 3\}$ .

On notera 1 quand l'objet  $j$  est dans le plateau de gauche dans l'expérience  $i$ . On notera  $-1$  quand il est dans le plateau de droite, et 0 quand il est absent.

Si, pour équilibrer une pesée, les poids sont à mettre dans le plateau de gauche, alors on les comptera négativement.

Avec ces conventions, la matrice de notre premier plan d'expérience est :

$$A_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

<sup>54</sup> Des compléments à propos des plans d'expérience peuvent être apportés par la brochure "Les plans d'expérience pour le B.T.S. chimiste" de la C.I.I. Lycées technologiques.



On a évidemment : 
$$\begin{cases} y_1 = m_0 + e_1 \\ y_2 = m_0 + m_1 + e_2 \\ y_3 = m_0 + m_2 + e_3 \\ y_4 = m_0 + m_3 + e_4 \end{cases}, \text{ où les } e_i \text{ sont des variables aléatoires}$$

indépendantes suivant la loi normale  $N(0, \sigma)$ . On trouve facilement les estimateurs des masses, en résolvant ce système en supposant les erreurs nulles.

On a : 
$$\begin{cases} \hat{m}_1 = y_2 - y_1 \\ \hat{m}_2 = y_3 - y_1 \\ \hat{m}_3 = y_4 - y_1 \end{cases}.$$

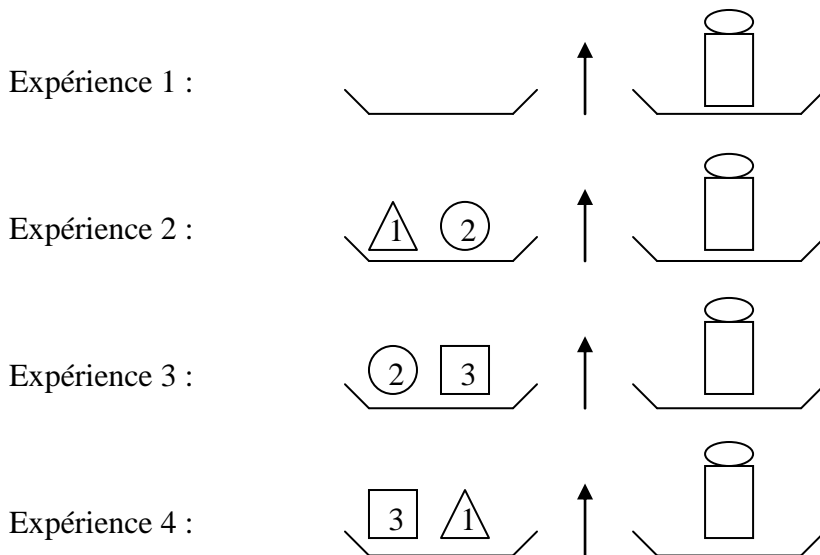
Les expériences étant indépendantes, la variance d'une somme est la somme des variances, et on voit facilement que :  $E(\hat{m}_i) = m_i$  et  $Var(\hat{m}_i) = 2\sigma^2$ .

La valeur  $\sigma$  mesurant l'imprécision de la balance, l'imprécision des estimations est donc  $\sqrt{2} \times \sigma$ . L'imprécision est donc, dans ce premier plan, multipliée par un facteur  $\sqrt{2}$ . Si l'on répétait  $n$  fois cette série de 4 expériences, alors la variance de l'estimateur obtenu en faisant la moyenne des  $n$  expériences serait :  $\frac{2}{n}\sigma^2$ .

A noter qu'en posant  $y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$ ,  $m = \begin{bmatrix} m_0 \\ m_1 \\ m_2 \\ m_3 \end{bmatrix}$  et  $e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$ , on a  $y = A_1 m + e$  et  $\hat{m} = A_1^{-1} y$ .

### Deuxième stratégie

Au lieu de peser les objets un par un, on peut penser qu'il vaudrait mieux les peser deux par deux.



La matrice de ce plan d'expérience est alors  $A_2$  :

$$A_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

à l'expérience 2, l'objet 2 (colonne 3) est dans le plateau de gauche.

On a :

$$\begin{cases} y_1 = m_0 + e_1 \\ y_2 = m_0 + m_1 + m_2 + e_2 \\ y_3 = m_0 + m_2 + m_3 + e_3 \\ y_4 = m_0 + m_1 + m_3 + e_4 \end{cases}$$

Avec le même procédé que précédemment, on parvient à :

$$\begin{cases} \hat{m}_1 = \frac{1}{2}(-y_1 + y_2 - y_3 + y_4) \\ \hat{m}_2 = \frac{1}{2}(-y_1 + y_2 + y_3 - y_4) \\ \hat{m}_3 = \frac{1}{2}(-y_1 - y_2 + y_3 + y_4) \end{cases}$$

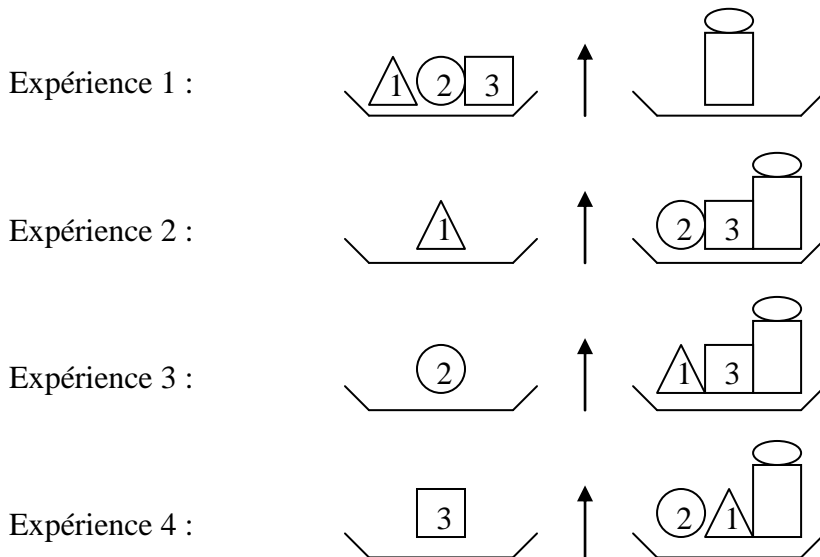
avec  $E(\hat{m}_i) = m_i$  et  $Var(\hat{m}_i) = \left(\frac{1}{2}\right)^2 \times 4Var(y_i) = \sigma^2$ .

Cette deuxième stratégie est meilleure que la précédente, car l'imprécision est  $\sigma$ , celle de la balance.

Là encore  $\hat{m} = A_2^{-1} y$ . Peut-on mieux faire ?

**Troisième stratégie**

On conjecture qu'en posant les trois objets à la fois, on améliore encore la stratégie. On réalise le plan d'expérience suivant :



La matrice de ce plan d'expérience est  $A_3$  :

$$A_3 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

à l'expérience 2, l'objet 2 est dans le plateau de droite.

$$\text{On a : } \begin{cases} y_1 = m_0 + m_1 + m_2 + m_3 + e_1 \\ y_2 = m_0 + m_1 - m_2 - m_3 + e_2 \\ y_3 = m_0 - m_1 + m_2 - m_3 + e_3 \\ y_4 = m_0 - m_1 - m_2 + m_3 + e_4 \end{cases} . \quad \text{Il vient : } \begin{cases} \hat{m}_1 = \frac{1}{4}(y_1 + y_2 - y_3 - y_4) \\ \hat{m}_2 = \frac{1}{4}(y_1 - y_2 + y_3 - y_4) \\ \hat{m}_3 = \frac{1}{4}(y_1 - y_2 - y_3 + y_4) \end{cases}$$

$$\text{avec } E(\hat{m}_i) = m_i \text{ et } \text{Var}(\hat{m}_i) = \left(\frac{1}{4}\right)^2 \times 4\text{Var}(y_i) = \frac{\sigma^2}{4}.$$

Cette troisième stratégie est encore meilleure ! L'imprécision est, par rapport à la deuxième stratégie, divisée par 2. Si l'on voulait obtenir, avec la première stratégie, la même précision qu'avec la troisième, il faudrait répéter 8 fois cette stratégie ( $\frac{\sqrt{2}}{\sqrt{n}}\sigma = \frac{\sigma}{2}$  pour  $n =$

8). En d'autres termes, quand les expériences sont bien faites, avec 4 expériences bien combinées, on fait aussi bien qu'avec 32 faites selon le simple "bon sens". Le gain n'est pas mince.

### Analyse de la troisième stratégie

Peut-on faire mieux ? La réponse est non. Le mathématicien français *Hadamard* (1862-1961) en a fait la démonstration (dans un cas plus général !). La qualité du troisième plan d'expérience peut être analysée en étudiant sa matrice.

$$\text{Celle-ci s'écrit } A_3 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

Cette matrice, où ne figurent que des 1 et des -1, est dite orthogonale car  ${}^tA_3 A_3 = 4I$  où  $I$  est la matrice identité. On peut aussi caractériser ces matrices en disant que, quand on prend un couple d'objets quelconque, il est dans l'état  $(k, l)$ ,  $k \in \{-1, 1\}$ ,  $l \in \{-1, 1\}$ , un nombre  $p$  de fois dans les  $n$  expériences faites,  $p$  étant indépendant de  $k$ , de  $l$  et du couple choisi. Dans notre exemple, on a  $p = 1$ . Une telle matrice, qui peut être rectangulaire, est dite *matrice d'Hadamard*.

## Bibliographie... pour aller plus loin

### Commission Inter-IREM Lycées technologiques

- "A propos de fiabilité" – IREM Paris-Nord – Brochure n°48.
- "Les plans d'expérience pour le BTS chimiste" – IREM Paris-Nord – Brochure n°88.

**SAPORTA Gilbert** – "Probabilités, analyse des données et statistique" – TECHNIP 1990.

**WONNACOTT T.H.** et **WONNACOTT R.J.** – "Statistique" – Economica 1995.

## ANNEXE 1 – Simulation de fourchettes de sondages

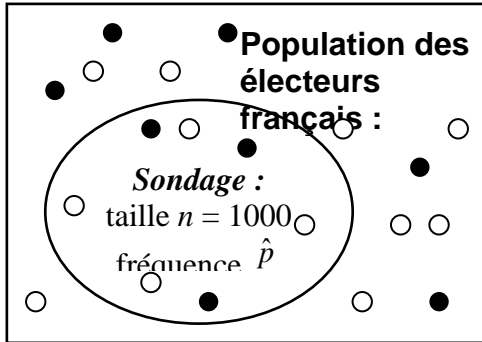
Forme	Niveau	Prérequis	Durée approximative
Travaux pratiques sur <b>EXCEL</b>	2 <sup>nde</sup>	Formule d'une fourchette à 95% sur un échantillon de taille $n$ .	1h (1h30 avec la partie III)

Ce T.P., proposé à des élèves de seconde, en salle informatique (un ou deux élèves par ordinateur), permet, par simulation, de donner du sens à la notion de "fourchette" de sondage : comment les fourchettes fluctuent-elles d'un sondages à un autre, deux fourchettes peuvent-elles être disjointes, une fourchette contient-elle le pourcentage à estimer ... ? On travaillera en particulier sur la signification du pourcentage de confiance.

On trouvera dans les pages suivantes :

- Le document élève comprenant une "feuille réponse".
- Un corrigé.

## SIMULATION DE FOURCHETTES DE SONDAGES



On considère une élection présidentielle en France. Le second tour oppose deux candidats  $X$  et  $Y$ .

On effectue un sondage sur  $n = 1000$  personnes, pour donner une estimation de la proportion  $p$  d'électeurs favorables à  $X$ , dans la population française.

Sur les 1000 personnes, tirées au sort, on obtient la proportion  $\hat{p}$  d'électeurs favorables à  $X$ .

Prenons un exemple.

Peu avant l'élection présidentielle de mai 1995, la chaîne de télévision *France 3* fournit les résultats d'un sondage qu'elle a fait effectuer sur un échantillon de 1000 personnes inscrites sur les listes électorales. D'après ce sondage, *Jacques Chirac* obtient 55% des intentions de vote, et *Lionel Jospin* 45%.

Le résultat de l'élection sera en fait de 52,6% pour *Jacques Chirac* et 47,4% pour *Lionel Jospin*.

– Compléter la feuille réponse.

## I – FOURCHETTES A 90% DE CONFIANCE

L'observation des fluctuations d'échantillons a montré qu'il est préférable de donner, à partir de  $\hat{p}$ , une *fourchette*, dans laquelle se situerait, avec plus ou moins de *confiance*, la valeur inconnue  $p$ .

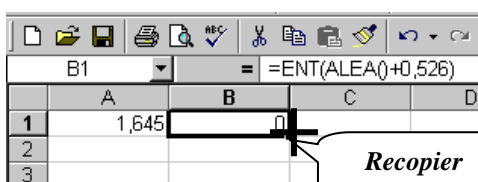
C'est ainsi que les statisticiens ont établi qu'avec l'intervalle :

$$\left[ \hat{p} - 1,645 \sqrt{\frac{\hat{p}(1-\hat{p})}{1000}} ; \hat{p} + 1,645 \sqrt{\frac{\hat{p}(1-\hat{p})}{1000}} \right], \text{ calculé avec la fréquence } \hat{p} \text{ sur un}$$

sondage de taille 1000, on a environ 90 chances sur 100 d'obtenir une fourchette contenant la fréquence inconnue  $p$ .

### a) Premier sondage

Travaillons avec les données de mai 1995. On considère que l'on effectue un sondage sur 1000 personnes dans une population où la proportion  $p$  d'électeurs en faveur de *Jacques Chirac* est  $p = 0,526$ .



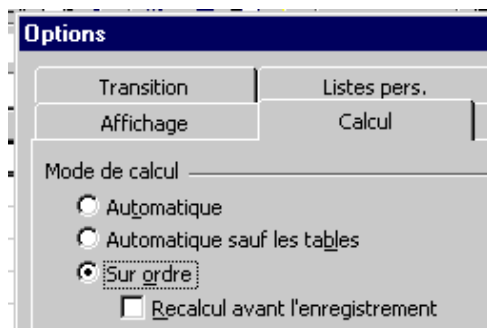
Lancer **Excel**<sup>®</sup>. Cliquer (avec le bouton gauche de la souris) dans la cellule A1, taper 1,645 (c'est la valeur qui apparaît dans l'expression de la fourchette à 90%).

Dans la cellule B1, entrer la *formule* :

=ENT(ALEA()+0,526) (avec des parenthèses vides après ALEA).

**Le résultat est 1 si l'électeur est pour Jacques Chirac et 0 s'il est pour Lionel Jospin.**

Approcher le pointeur de la souris du carré noir situé dans le coin inférieur droit de la cellule B1. Celui-ci se transforme en une croix noire, faire alors glisser, en maintenant le bouton gauche enfoncé pour **recopier** jusqu'en B1000, puis relâcher le bouton de la souris. Vous avez simulé les résultats d'un sondage de 1000 personnes.



Afin de ne lancer les calculs que lorsqu'on le désire, configurer Excel ainsi :

Cliquer dans le menu **Outils/Options...** puis dans l'onglet **Calcul**, puis à la rubrique **Mode de calcul**, choisir **• Sur ordre** puis **OK**.

On va calculer la fréquence  $\hat{p}$  sur le sondage et une **fourchette** pour  $p$ , à 90% de confiance.

En A1002 taper `inf`

puis, en A1003, taper `sup` et en

A1004 `p^` (on obtient `^` en appuyant simultanément sur ALT GR et `ç`).

En B1004, entrer la **formule** :

`=SOMME(B1:B1000)/1000`

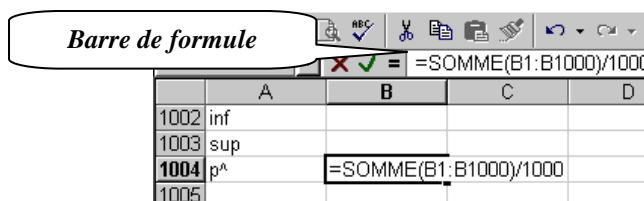
En B1003, entrer la **formule** : `=B1004+$A$1*RACINE(B1004*(1-B1004)/1000)` pour la borne supérieure de la "fourchette".

En B1002, entrer la **formule** : `=B1004-$A$1*RACINE(B1004*(1-B1004)/1000)` pour la borne inférieure de la "fourchette".

En B1005, entrer la **formule** :

`=ET(0,526>=B1002;0,526<=B1003)`

(la formule `ET(condition 1 ; condition 2)` renvoie la valeur VRAI si les *conditions* sont vérifiées et la valeur FAUX sinon).



— Conserver les résultats de ce sondage sur la feuille réponse.

## b) Visualisation des fourchettes données par 10 sondages

**Sélectionner** les cellules de B1 à B1005 (pour cela cliquer sur B1 et glisser, en gardant le bouton gauche de la souris enfoncé, jusqu'en B1005, puis relâcher le bouton de la souris).

**Recopier** la sélection (en glissant après avoir obtenu la croix noire au coin de B1005) jusqu'en K1005.

Appuyer sur **F9** pour lancer le calcul.

Vous avez maintenant 10 sondages de chacun 1000 personnes.

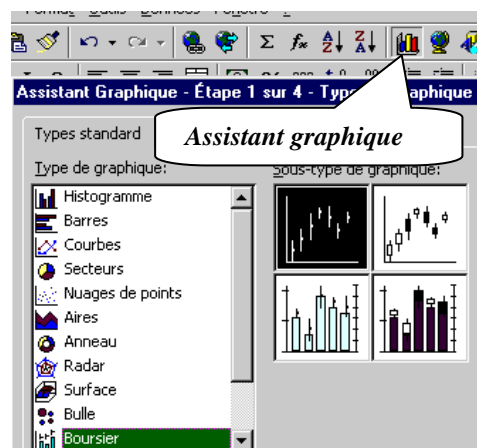
Cliquer sur l'icône **Assistant graphique**.

**Etape 1/4** : choisir **Boursier** et le premier **Sous-type** puis cliquer sur **Suivant**.

**Etape 2/4** : dans **Plage de données**, inscrire B1002:K1004

puis cocher **Série en • Lignes**

et cliquer sur **Suivant**.



**Etape 3/4** : cliquer sur *Suivant*.

**Etape 4/4** : cocher *Insérer le graphique* • *Sur une nouvelle feuille* puis cliquer sur *Terminer*.

Sur le graphique, cliquer avec le bouton *droit* de la souris sur la légende et choisir *Effacer*, puis cliquer avec le bouton *droit* de la souris sur l'*Axe des ordonnées* et choisir *Format de l'axe...* . Dans l'onglet *Echelle*, à la rubrique *Minimum*, inscrire la valeur 0,45 (attention à taper une virgule et non un point), à la rubrique *Maximum*, inscrire la valeur 0,61. Cliquer sur *OK*.

Appuyer sur la touche **F9** pour simuler 10 nouveaux sondages et observer le graphique.

– Consigner vos observations sur la feuille réponse.

### c) Sondages donnant une fourchette ne contenant pas la valeur à estimer

Revenir sur la *feuille 1*.

En A1006 taper "FAUX". En B1006, entrer la *formule* : =NB.SI(B1005:K1005;FAUX)

En C1006 taper "sur 10" .

– Faire **F9** pour de nouvelles simulations et compléter la feuille réponse.

## 2- FOURCHETTES A 95% OU 99% DE CONFIANCE

### a) Fourchettes à 95% de confiance

Sur la *feuille 1*, remplacer en A1 la valeur 1,645 par 1,96.

– Faire **F9** pour relancer les simulations et compléter la feuille réponse.

### b) Fourchettes à 99% de confiance

Sur la *feuille 1*, remplacer en A1 la valeur 1,96 par 2,58.

– Faire **F9** pour relancer les simulations et compléter la feuille réponse.

## III – PERSPECTIVES DU BAROMETRE "SOFRES"

### 1 – Résultats SOFRES

Le "baromètre présidentiel" de la SOFRES<sup>55</sup> est un sondage effectué sur un échantillon de  $n = 1000$  personnes.

Voici les résultats de quatre de ces sondages, pour le second tour de l'élection :

	mars 2001	mai 2001	juin 2001	septembre 2001
<i>Jacques Chirac</i> $\hat{p}$	0,48	0,50	0,49	0,49
<i>Lionel Jospin</i> $1 - \hat{p}$	0,52	0,50	0,51	0,51

Doit-on, d'après ces sondages, affirmer que la popularité de *Jacques Chirac* dans la population française a augmenté au mois de mai, avant de rebaisser les mois suivants ?

<sup>55</sup> Source SOFRES : <http://www.sofres.com>

Avant de répondre, comme on sait qu'un sondage peut ne pas représenter exactement l'opinion de la population entière, on va calculer les fourchettes de confiance à 95%, en

utilisant la formule simplifiée :  $\left[ \hat{p} - \frac{1}{\sqrt{1000}} ; \hat{p} + \frac{1}{\sqrt{1000}} \right]$

Cliquer sur l'onglet **Feuil2**, pour ouvrir une nouvelle feuille de calcul.

	A	B	C	D
1	0,48	0,5	0,49	
2	0,44837722			
3				
4				

Dans les cellules A1, B1 et C1, entrer respectivement les valeurs 0,48 ; 0,5 et 0,49.

En A2, entrer la formule :

=A1 - 1/RACINE(1000)

En A3, entrer la formule : =A1 + 1/RACINE(1000)

Sélectionner les cellules A2 et A3, et les recopier vers la droite jusqu'en C3. Appuyer sur **F9** pour obtenir les trois fourchettes.

– Compléter la feuille réponse.

## 2 – Comparaison avec le pile ou face

Pour confirmer l'impression donnée par le calcul des fourchettes, nous allons effectuer des sondages virtuels sur la base  $p = 0,5$ , où l'opinion est également partagée entre les deux candidats.

Dans la cellule D4, entrer la formule : =ENT(ALEA() + 0,5)

Recopier cette cellule vers le bas jusqu'en D1003. Faire F9 pour lancer le calcul.

En D1, entrer la formule : =SOMME(D4:D1003)/1000

– Faire **F9** pour relancer les simulations et compléter la feuille réponse.



<b>– FEUILLE REPONSE</b>
--------------------------

NOMS : .....

Pour l'élection de mai 1995 :

Quelle est la fréquence  $\hat{p}$  des électeurs favorables à *Jacques Chirac*, obtenue sur le sondage de *France 3* (donner la réponse sous forme d'un nombre décimal entre 0 et 1) ?

On a  $\hat{p} = \dots\dots\dots$

Quelle est la fréquence  $p$  des voix reçues par *Jacques Chirac* à l'élection ?

On a  $p = \dots\dots\dots$

Considérez-vous que le sondage était faux ?

.....  
 .....  
 .....

## I – FOURCHETTES A 90% DE CONFIANCE

### a) Premier sondage

Quelle est la fréquence  $\hat{p}$  des électeurs de *Jacques Chirac* sur les 1000 personnes de votre sondage ? .....

Quelle fourchette, à 90%, lui correspond pour  $p$  ? .....

Que signifie la valeur VRAI ou la valeur FAUX, calculée dans la cellule B1005 ?

.....  
 .....

### b) Visualisation des fourchettes données par 10 sondages

Deux fourchettes sont-elles obligatoirement les mêmes ? .....

Deux fourchettes ont-elles obligatoirement le même centre ? .....

Deux fourchettes peuvent-elles n'avoir aucun élément commun ? .....

Est-ce que  $p = 0,526$  appartient nécessairement à la fourchette donnée par un sondage ?

.....

### c) Sondages donnant une fourchette ne contenant pas la valeur à estimer

Indiquer dans le tableau suivant, pour chaque groupe de 10 sondages, le nombre de ceux qui fournissent une fourchette ne contenant pas la valeur à estimer.

Simulations de 10 sondages	1	2	3	4	5	6	7	8	9	10
Nombres de fourchettes à 90% ne contenant pas 0,526										

Quel pourcentage de fourchettes à 90% de confiance "fausses" avez vous globalement obtenu ? .....

## 2- FOURCHETTES A 95% OU 99% DE CONFIANCE

### a) Fourchettes à 95% de confiance

Sur le graphique, observer l'aspect des fourchettes. Quelle différence a-t-on ? .....

.....

Indiquer dans le tableau suivant, pour chaque groupe de 10 sondages, le nombre de ceux qui fournissent une fourchette ne contenant pas la valeur à estimer.

Simulations de 10 sondages	1	2	3	4	5	6	7	8	9	10
Nombres de fourchettes à 95% ne contenant pas 0,526										

Quel pourcentage de fourchettes à 95% de confiance "fausses" avez vous globalement obtenu ? .....

### b) Fourchettes à 99% de confiance

Sur le graphique, quelle différence observe-t-on ? .....

.....

Indiquer dans le tableau suivant, pour chaque groupe de 10 sondages, le nombre de ceux qui fournissent une fourchette ne contenant pas la valeur à estimer.

Simulations de 10 sondages	1	2	3	4	5	6	7	8	9	10
Nombres de fourchettes à 99% ne contenant pas 0,526										

Quel pourcentage de fourchettes à 99% de confiance "fausses" avez vous globalement obtenu ? .....

Quel avantage a-t-on à donner une fourchette à 99 % de confiance ? .....

.....

Mais quel est l'inconvénient ? .....

.....

Quel vous semble le pourcentage de confiance le mieux adapté à cette situation de sondage et pourquoi ? .....

.....

.....

## III – PERSPECTIVES DU BAROMETRE "SOFRES"

### 1 – Résultats SOFRES

Pour  $\hat{p} = 0,48$ , la fourchette de confiance pour  $p$  à 95% est environ : [..... ; .....].

Pour  $\hat{p} = 0,50$ , la fourchette de confiance pour  $p$  à 95% est environ : [..... ; .....].

Pour  $\hat{p} = 0,49$ , la fourchette de confiance pour  $p$  à 95% est environ : [..... ; .....].

En obtenant  $\hat{p} = 0,49$  après avoir obtenu  $\hat{p} = 0,50$  le mois précédent, doit-on en déduire que la popularité  $p$  de *Jacques Chirac* a diminuée ? .....

.....

.....

## 2 – Comparaison avec le pile ou face

L'instruction ALEA() fournit un nombre au hasard entre 0 et 1.

Quel est le résultat de l'instruction ALEA() + 0,5 ? .....

Pourquoi l'instruction ENT(ALEA() + 0,5) , où ENT donne la partie entière d'un nombre, peut-elle être utilisée pour simuler le lancer d'une pièce de monnaie dans le jeu de pile ou face ? .....

Noter, pour 5 simulations, les résultats obtenus dans la cellule D1 :

--	--	--	--	--

On peut interpréter ces résultats comme ceux de sondages, sur 1000 personnes, fournissant la proportion  $\hat{p}$  en faveur de *Jacques Chirac*, dans une population où cette proportion est  $p$ . Quelle est la valeur de  $p$  utilisée dans ces simulations ? .....

Pourquoi ces simulations montrent-elles qu'il est possible que l'opinion française soit restée la même de mars à septembre 2001, malgré les variations du baromètre SOFRES ?

.....  
 .....  
 .....  
 .....

## Description et compte-rendu du TP Excel "SIMULATION DE FOURCHETTES DE SONDAGE"

### DUREE

1h30 en salle informatique (demi-classe), à un ou deux élèves par poste.

### CORRIGE COMMENTE

#### 1- FOURCHETTES A 90% DE CONFIANCE

##### a) Premier sondage

La colonne de 0 et de 1 correspond aux résultats du sondage (0 pour les sondés favorables à Chirac et 1 pour ceux favorables à Jospin).

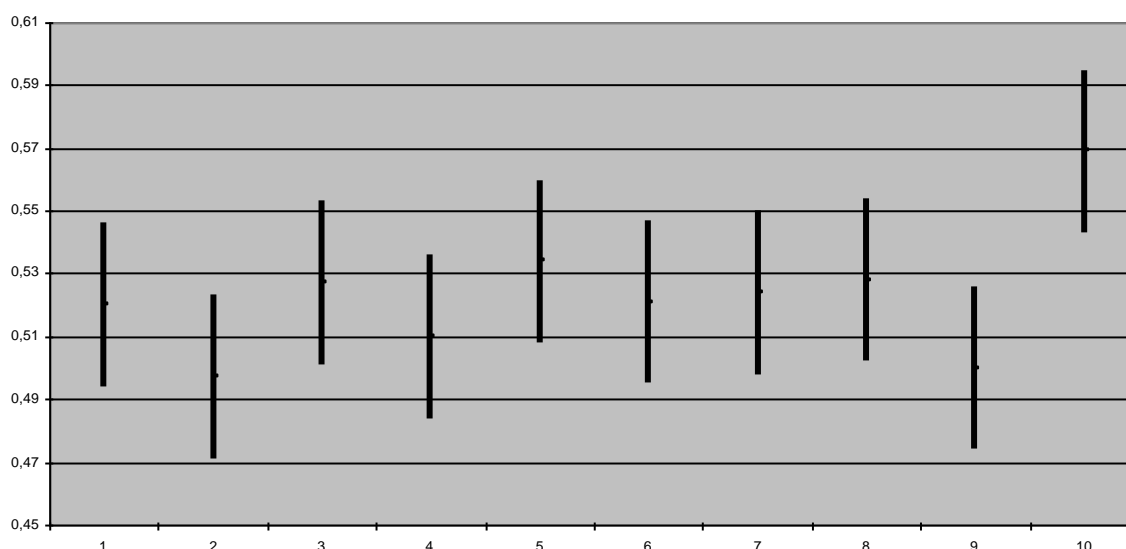
On obtient (par exemple !)  $\hat{p} = 0,511$ .

La fourchette correspondante pour  $p$  est alors  $[0,484 ; 0,537]$ .

La valeur VRAI signifie que la fourchette contient effectivement la valeur  $p = 0,526$  (c'est le cas ici). La valeur FAUX signifie que la fourchette ne contient pas la valeur  $p = 0,526$ .

##### b) Visualisation des fourchettes données par 10 sondages

Chaque simulation de 10 sondages (touche de "recalcul" F9) donne lieu à un graphique de ce type (10 fourchettes)



Deux fourchettes sont-elles obligatoirement les mêmes ? Non.

Deux fourchettes ont-elles obligatoirement le même centre ? Non.

Deux fourchettes peuvent-elles n'avoir aucun élément commun ? Oui (mais il faut parfois faire un certain nombre de simulations pour l'observer).

Est-ce que  $p = 0,526$  appartient nécessairement à la fourchette donnée par un sondage ? Non (fourchettes fausses).

##### c) Sondages donnant une fourchette ne contenant pas la valeur à estimer

Le tableau suivant donne (par exemple), pour chaque groupe de 10 sondages simulés, le nombre de ceux qui fournissent une fourchette ne contenant pas la valeur à estimer.

Simulations de 10 sondages	1	2	3	4	5	6	7	8	9	10
Nombres de fourchettes à 90% ne contenant pas 0,526	2	1	0	0	3	1	0	1	1	0

Le pourcentage de fourchettes à 90% de confiance "fausses", globalement obtenu sur ces exemples, est 9 % (sur un grand nombre de sondages, on obtiendrait 10 %).

## 2- FOURCHETTES A 95% OU 99% DE CONFIANCE

### a) Fourchettes à 95% de confiance

Sur le graphique, on observe des "fourchettes" plus longues.

On a par exemple :

Simulations de 10 sondages	1	2	3	4	5	6	7	8	9	10
Nombres de fourchettes à 95% ne contenant pas 0,526	0	0	0	0	2	1	0	1	1	1

Le pourcentage de fourchettes à 95% de confiance "fausses", globalement obtenu sur ces exemples, est 6 %.

### b) Fourchettes à 99% de confiance

Sur le graphique, on observe des "fourchettes" plus longues.

On a par exemple :

Simulations de 10 sondages	1	2	3	4	5	6	7	8	9	10
Nombres de fourchettes à 99% ne contenant pas 0,526	0	0	0	0	0	1	0	0	1	0

Le pourcentage de fourchettes à 99% de confiance "fausses", globalement obtenu sur ces exemples, est 2 %.

L'avantage des fourchettes à 99 % est qu'on se trompe moins, mais l'inconvénient est leur grande amplitude, qui fait que l'information est peu précise.

Le pourcentage de confiance le mieux adapté à cette situation de sondage est 95 %. C'est un bon compromis entre un trop grand nombre d'erreurs (90 % de confiance) et une trop grande amplitude (manque de précision au coefficient 99 %).

## III – PERSPECTIVES DU BAROMETRE "SOFRES"

### 1 – Résultats SOFRES

Sur l'image ci-contre apparaissent les trois fourchettes de confiance à 95%.

C'est à dire que l'on a plus de 95% de chances (formule simplifiée), en donnant une telle fourchette, de recouvrir la valeur réelle de  $p$ .

	A	B	C	D
1	0,48	0,5	0,49	0,475
2	0,44837722	0,46837722	0,45837722	
3	0,51162278	0,53162278	0,52162278	
4				1
5				1
6				0
7				n

Compte-tenu de l'amplitude de ces fourchettes, ce n'est pas parce que l'on observe la valeur 0,49 après avoir obtenu la valeur 0,5, que l'on doit en déduire que la valeur de  $p$  a certainement diminuée. Cette impression sera confirmée par la simulation suivante.

### 2 – Comparaison avec le pile ou face

Les simulations montrent les fluctuations des sondages de taille 1000 effectués sur une même population où  $p = 0,5$ .

Les observations confirment que les différentes valeurs du "baromètre SOFRES" peuvent très bien correspondre au même état de l'opinion (en particulier avec  $p = 0,5$ ).

N.B.

- On pourrait analyser différemment une suite continue de hausses (ou de baisses), même faibles.
- Les fourchettes à 95% que l'on pourrait construire sur les résultats SOFRES sont, en fait, meilleures que celles calculées ici, en ce sens que la méthode d'échantillonnage utilisée n'est qu'en partie aléatoire (méthode des quotas – sexe, âge, activité – et stratification par

région et catégorie d'agglomération). C'est à l'intérieur de chaque strate, que l'on procède à un tirage au sort.

## ANNEXE 2 – Utilisation de moyennes mobiles à la bourse

Forme	Niveau	Prérequis	Durée approximative
Travaux pratiques sur <b>EXCEL</b>	<b>1<sup>ère</sup> ES</b>	Moyenne	45 mn

On étudie, à l'aide du tableur, le lissage de l'évolution de l'indice boursier CAC 40 selon les moyennes mobiles d'ordre 2 et 5 et les stratégies d'achat/vente qui leurs sont associées.

On remarquera que la moyenne mobile utilisée dans ce contexte boursier ne correspond pas à la définition du document d'accompagnement du programme de 1<sup>ère</sup> ES (rentrée 2001), ou donnée dans cette brochure page 119, dans la mesure où, ignorant, bien entendu, les cours à venir, les moyennes mobiles ne sont calculées qu'à partir des valeurs qui précèdent.

On trouvera dans les pages suivantes :

- Le document élève comprenant une "feuille réponse".
- Un corrigé.

Remarque : les élèves plus performants pourront ajouter des légendes sur le graphique.



## UTILISATION DE MOYENNES MOBILES A LA BOURSE

L'objectif est d'utiliser le lissage par les moyennes mobiles, pour obtenir, dans le contexte boursier, des signaux d'achat ou de vente.

### SAISIE DES DONNEES

Ouvrir un dossier Excel.

Entrer en colonne A les données ci-dessous, correspondant à l'indice **CAC 40** au premier jour ouvrable de chaque mois, sur 20 mois consécutifs, à partir du 02 janvier 1998.

3040
3188
3447
3883
3974
4087
4261
4095
3646
3038
3570
3688
4148
4304
4032
4230
4443
4314
4609
4378

	A	B	C
1	3040		
2	3188		
3	3447		
4	3883		
5	3974		
6	4087		
7	4261		
8	4095		
9	3646		
10	3038		
11	3570		
12	3688		
13	4148		
14	4304		
15	4032		
16	4230		
17	4443		
18	4314		
19	4609		
20	4378		
21			
22			

## I – CALCUL ET REPRESENTATION DES MOYENNES MOBILES D'ORDRE 2 ET 5

Un des outils statistiques les plus anciens, et les plus pratiqués dans le domaine financier, est celui des *moyennes mobiles*. C'est un moyen de lissage qui permet de gommer les mouvements erratiques des cours, pour n'en conserver que la tendance de fond, et obtenir ainsi un indicateur pour l'achat ou la vente.

### MM2

La moyenne mobile d'ordre 2 (MM2) est tout simplement, dans le cadre présent, la moyenne arithmétique de 2 valeurs : celle du mois présent et celle du mois précédent (en général, on ne connaît pas l'avenir). Son graphique ne débute donc qu'avec la 2<sup>ème</sup> donnée. Cette moyenne est dite "mobile" du fait que le calcul de la moyenne mobile consécutive ne diffère que par glissement d'une valeur (la plus ancienne disparaît au profit de la nouvelle).



Dans la cellule B2, entrer la **formule** (attention, les formules doivent commencer par le symbole =) :  $= (A1+A2)/2$  (puis appuyer sur **Entrée**).

	A	B	C
1	3040		
2	3188	3114	
3	3447		
4	3883		
5	3974		
6	4087		

Approcher le pointeur de la souris du carré noir situé dans le coin inférieur droit de la cellule B2. Celui-ci se transforme en une croix noire, faire alors glisser, en maintenant le bouton gauche enfoncé pour **recopier** jusqu'en B20, puis relâcher le bouton de la souris.

La colonne B contient alors les moyennes mobiles d'ordre 2.

## MM5

Vous allez calculer en colonne C les moyennes mobiles d'ordre 5.

Entrer en cellule C5 la formule :  $= \text{SOMME}(A1:A5)/5$

puis recopier, comme précédemment, cette formule jusqu'en C20.

## Représentations graphiques

Sélectionner, avec le bouton gauche de la souris, les cellules de A1 à A20, puis appuyer sur l'icône **Assistant graphique**.

**Assistant Graphique - Étape 1 sur 4 - Type de Graphique**

Types standard | Types personnalisés

Type de graphique :  
 Histogramme  
 Barres  
 Courbes  
 Secteurs  
 Nuages de points  
 Aires  
 Anneau  
 Radar  
 Surface  
 Bulle  
 Boursier

Sous-type de graphique :  
 Nuage de points reliés par une courbe sans marquage des données.

Maintenir appuyé pour visionner


**Étape 1/4 :**

Dans **Type de graphique**, choisir **Nuages de points**, puis dans **Sous-type de graphique**, **Nuage de points reliés par une courbe sans marquage des données**.

Cliquer sur **Suivant**.

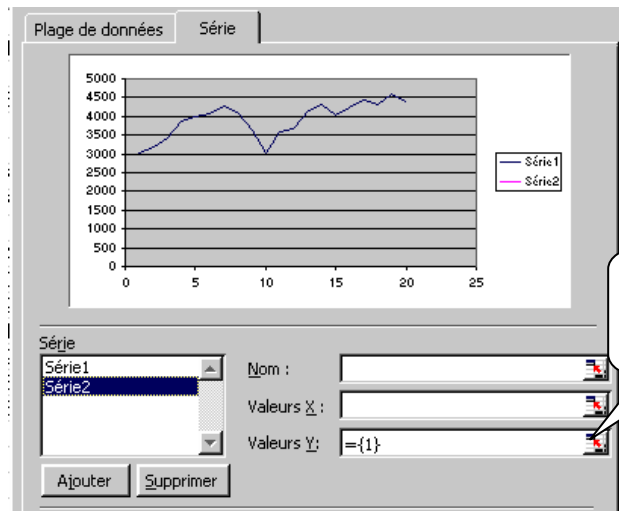
**Étape 2/4 :**

Dans l'onglet **Série**, cliquer sur le bouton **Ajouter**.

Pour la *Série 2*, à la rubrique *Valeurs Y*, cliquer sur l'icône  de sortie sur la feuille de calcul.

Sur la feuille de calcul, sélectionner les cellules de B1 à B20, puis revenir, par l'icône analogue, dans la boîte de dialogue de l'assistant graphique. Procéder de même pour *Ajouter* en *Série 3* les valeurs des cellules de C1 à C20.

Cliquer sur *Suivant*.



Sortie vers la feuille de calcul

**Etape 3/4 :**

Dans l'onglet *Légende*, désactiver la case *Afficher la légende*. Cliquer sur *Suivant*.

**Etape 4/4 :**

Sélectionner *Placer le graphique en tant qu'objet dans Feuil1* puis cliquer sur *Terminer*.

On peut déplacer le graphique et l'agrandir à l'aide des poignées.

Cliquer, avec le *bouton droit* de la souris, sur l'axe des abscisses et choisir *Format de l'axe...* Dans l'onglet *Echelle*, entrer dans la case *Maximum* la valeur 20.

Cliquer, avec le *bouton droit* de la souris, sur l'axe des ordonnées et choisir *Format de l'axe...* Dans l'onglet *Echelle*, entrer dans la case *Minimum* la valeur 2500 et dans la case *Maximum* la valeur 5500.

– Compléter, sur la feuille réponse, les questions d'analyse du graphique.

## II – UTILISATION DES MOYENNES MOBILES COMME SIGNAL D'ACHAT/VENTE

Pour l'achat et la vente, les analystes financiers adoptent la règle suivante :

- **acheter** quand la moyenne mobile croise les cours à la hausse,
- **vendre** quand la moyenne mobile croise les cours à la baisse.

On suppose que l'on achète, le 02/01/98, un groupement de titres indexé sur l'indice CAC 40, qui vaut, à cette date, 3040 points.

– Comparer, sur la feuille réponse, les décisions prises suivant que l'on suive la moyenne mobile MM2 (à court terme) ou la moyenne mobile MM5 (à moyen terme).

<b>– FEUILLE REPONSE</b>
--------------------------

**NOMS :** .....

## I – CALCUL ET REPRESENTATION DES MOYENNES MOBILES D'ORDRE 2 ET 5

1) Comparer l'aspect des trois courbes : CAC 40, MM2 et MM5.

.....  
 .....  
 .....

2) Combien de fois la courbe MM2 traverse-t-elle celle du CAC 40 ?

.....

3) Combien de fois la courbe MM5 traverse-t-elle celle du CAC 40 ?

.....

## II – UTILISATION DES MOYENNES MOBILES COMME SIGNAL D'ACHAT/VENTE

1) Utilisation de la moyenne mobile d'ordre 2 :

valeur d'achat	valeur de vente	gain (+ ou -)
3040		
bilan		

2) Utilisation de la moyenne mobile d'ordre 5 :

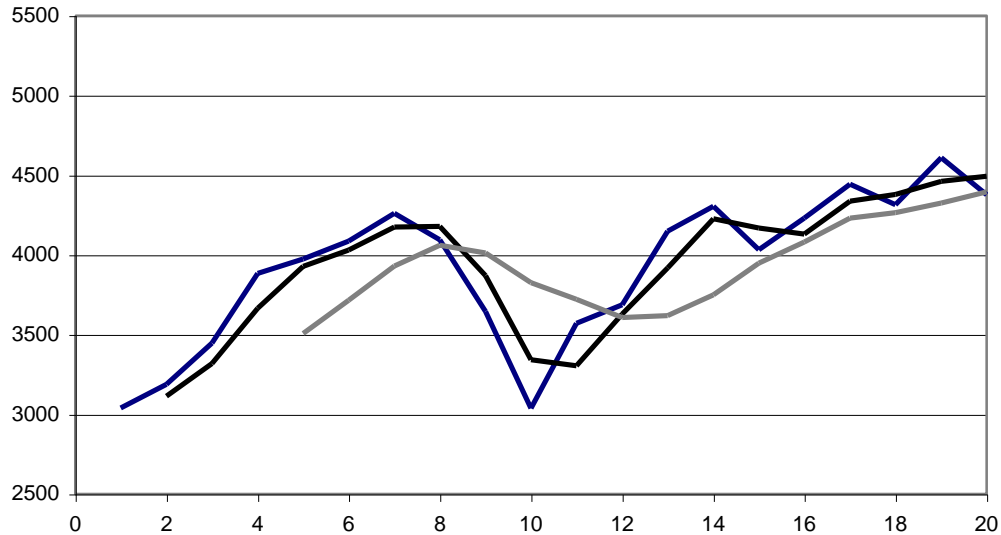
valeur d'achat	valeur de vente	gain (+ ou -)
3040		
bilan		

3) Conclusion

.....  
 .....  
 .....

## Corrigé du T.P. "Utilisation des moyennes mobiles à la bourse"

### I – CALCUL ET REPRESENTATION DES MOYENNES MOBILES D'ORDRE 2 et 5



1) Les variations des moyennes mobiles sont inférieures à celles du CAC 40, l'effet de lissage augmentant avec l'ordre de la moyenne, avec un décalage de plus en plus important (effet d'inertie).

2) La courbe de la moyenne mobile d'ordre 2 croise celle du CAC 40 à 7 reprises.

3) La courbe MM5 croise 3 fois celle du CAC 40.

### II – UTILISATION DES MOYENNES MOBILE COMME SIGNAL D'ACHAT/VENTE

1) Utilisation de la moyenne mobile d'ordre 2 :

valeur d'achat	valeur de vente	gain ( + ou - )
3040	4095	+1055
3570	4032	+462
4230	4314	+84
4609	4370	-239
	bilan :	<b>+1362</b>

2) Utilisation de la moyenne mobile d'ordre 5 :

valeur d'achat	valeur de vente	gain ( + ou - )
3040	3646	+606
3688	4378	+690
	bilan :	<b>+1296</b>

3) Conclusion :

La stratégie à court terme, peut être plus risquée, s'est avérée, ici, plus avantageuse. Mais ce n'est pas un théorème...

## ANNEXE 3 – Expérimentation du théorème limite central

Forme	Niveau	Prérequis	Durée approximative
Travaux pratiques sur <b>EXCEL</b>	B.T.S.	Loi normale Régression linéaire	2h

### **L'importance, en statistique, de la loi normale tient en grande partie au théorème limite central.**

Ce T.P., proposé à des étudiants de deuxième année en B.T.S. Industriel, permet, par simulation, d'expérimenter ce théorème, d'apprécier l'approximation normale à l'aide d'une régression linéaire, puis d'appliquer ce procédé dans le cadre d'une situation industrielle.

On trouvera dans les pages suivantes :

- Le document élève comprenant une "feuille réponse".
- Un corrigé.

## EXPERIMENTATION DU THEOREME LIMITE CENTRAL

### Objectifs

- Expérimenter le théorème limite central sur la simulation de 1000 données.
- Trier des données pour comparer leur distribution à une densité normale.
- Contrôler la normalité d'une distribution empirique par régression linéaire (droite de *Henry*).
- Appliquer ce procédé pour un test de normalité d'une production industrielle.

## 1 - GENERATION D'UNE DISTRIBUTION DE 1000 VALEURS

### Loi uniforme sur [0 , 1]

	A1	=	=ALEA()
	A	B	
1	0,28289904		
2			
3			

Lancer Excel®.

Dans la cellule A1, entrer la formule : =ALEA()

Faire **ENTREE** puis, en approchant le pointeur de la souris du carré noir situé dans le coin inférieur droit de la cellule A1, celui-ci se transforme en une croix noire,

faire alors glisser en maintenant le bouton gauche enfoncé pour **recopier** jusqu'en J1.

La fonction ALEA( ) génère un nombre aléatoire compris entre 0 et 1, c'est à dire qu'elle simule un résultat d'une variable aléatoire  $X$  suivant la loi uniforme sur  $[0 ; 1]$  :  $U([0 , 1])$ .

Un peu de théorie avant de poursuivre... Cette loi continue admet comme fonction de

densité la fonction  $f$  définie par :  $f(x) = \begin{cases} 1 & \text{si } x \in [0 , 1] \\ 0 & \text{si } x \notin [0 , 1] \end{cases}$

– Calculer, pour cette loi, sur la feuille réponse, l'espérance et la variance.

### Somme de $n = 12$ variables aléatoires indépendantes de même loi $U[0 , 1]$

On désigne par  $Y$  une variable aléatoire somme de  $n = 12$  variables aléatoires de même loi  $U[0 , 1]$ . Pour simuler 1000 réalisations de  $Y$ , procéder ainsi :

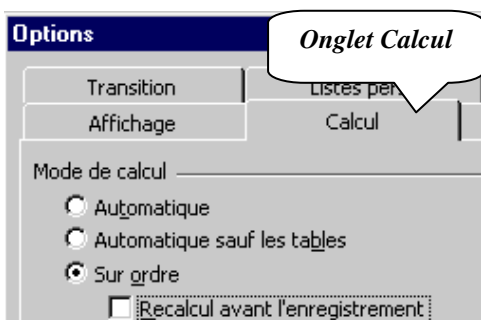
Cliquer sur la cellule A1. Dans la **Barre de formules**, sélectionner ALEA() (mettre en vidéo inversée en balayant, bouton gauche de la souris enfoncé) puis cliquer sur l'icône **Copier**. Cliquer dans la barre de formules, ajouter à la fin + puis cliquer sur l'icône **Coller**.

	I	J
1	4,4138467	9,03957893
2		

Répéter l'opération de façon à avoir la somme de 12 fois la fonction ALEA( ) (rien à voir avec  $12 * ALEA( )$  !). Faire **ENTREE**.

En approchant le pointeur de la souris du carré noir situé dans le coin inférieur droit de la cellule A1, celui-ci se transforme en une croix noire, **Recopier** alors jusqu'en J1. Relâcher

le bouton gauche de la souris puis **Recopier vers le bas** jusqu'en J100 (il faut que la première ligne soit sélectionnée avant de recopier vers le bas).



Vous avez maintenant une simulation de 1000 réalisations aléatoires de la variable  $Y$ .

Pour conserver ces 1000 valeurs, ou refaire une nouvelle simulation quand on le désire, cliquer dans le **menu Outils / Options...** puis dans l'**onglet Calcul** de la boîte de dialogue, à la rubrique **Mode de calcul**, choisir **• Sur ordre** puis **OK**.

### **Moyenne $\bar{x}$ et écart type $s_e$ d'un échantillon de 1000 valeurs**

En A102 taper "moyenne" et en B102 entrer la formule : =MOYENNE(A1:J100) puis **ENTREE**.

En A103 taper "écart type" et en B103 entrer la formule : =ECARTYPEP(A1:J100) (Attention à bien taper ECARTYPEP et non ECARTYPE qui correspond à l'estimation de l'écart type de la population dont est extrait l'échantillon) puis **ENTREE**.

Faire **F9** pour une nouvelle simulation de 1000 valeurs.

– Consigner vos résultats sur la feuille réponse.

### **Valeurs théoriques de $\mu$ et $\sigma$**

En répétant 12 fois la fonction ALEA() et en faisant la somme, on simule la somme de 12 variables aléatoires indépendantes de loi  $U$  ( $[0, 1]$ ). On note  $\mu$  et  $\sigma$  l'espérance et l'écart type de cette somme.

– Déterminer les paramètres théoriques  $\mu$  et  $\sigma$  sur la feuille réponse.

## **2 - THEOREME LIMITE CENTRAL**

### **Un énoncé du théorème**

Pourquoi, avec une expression analytique paradoxalement compliquée, la loi de *Laplace-Gauss* est-elle si répandue au point d'être qualifiée de "*normale*" ?

La réponse des mathématiciens à cette question est le **Théorème limite central** :

*La somme de  $n$  variables aléatoires indépendantes de même loi suit approximativement, pour  $n$  assez grand, une loi normale.*

– Expliquer, sur la feuille réponse, pourquoi, selon ce théorème, la variable aléatoire  $Y$  suit approximativement une loi normale.

### **Comparaison graphique de l'histogramme des 1000 données à la densité normale**

#### **Tri des données**

On va regrouper les 1000 données en 17 classes de part et d'autre de la valeur 6.

Il faut entrer les bornes supérieures de ces classes.

Cliquer sur l'onglet **Feuil2** (en bas).

En A1 taper "sup classes".

En A2 entrer la valeur 2,25 . En A3, entrer la formule : =A2+0,5

FREQUENCE		X	✓	=	=FREQUENCE(Feuil1!A1:...
	A	B	C	D	
1	sup classes	effectifs ni			
2	2,25	10;A2:A18)			
3	2,75				
4	3,25				
5	3,75				
6	4,25				
7	4,75				
8	5,25				
9	5,75				
10	6,25				
11	6,75				
12	7,25				
13	7,75				
14	8,25				
15	8,75				
16	9,25				
17	9,75				
18	10,25				

Recopier vers le bas jusqu'en A18 puis faire **F9**. Le calcul s'effectue et la dernière cellule contient alors la valeur 10,25.

En B1 taper "effectifs ni".

**Sélectionner** les cellules de B2 à B18 (pour cela, cliquer sur B2 et glisser, en gardant le bouton gauche de la souris enfoncé, jusqu'en B18, puis relâcher le bouton de la souris).

Alors que les cellules sélectionnées apparaissent en "vidéo inversée", cliquer dans la **barre de formules** et taper :  
=FREQUENCE(Feuil1!A1:J100;A2:A18)  
puis valider en appuyant *en même temps* sur les touches **CTRL MAJUSCULE** et **ENTREE**.

Excel calcule alors les effectifs de chacune des classes (et non les fréquences comme semble l'indiquer le nom de la fonction utilisée).

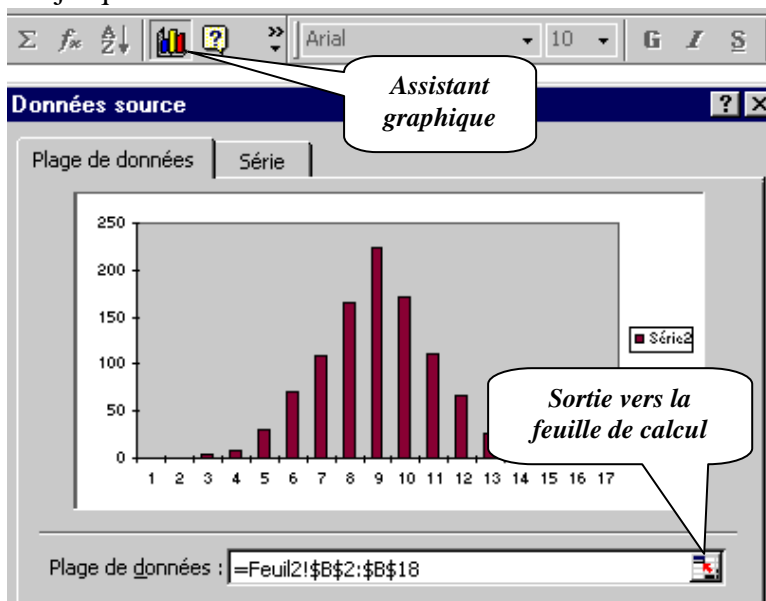
En B19, cliquer sur l'icône  $\Sigma$  de **somme automatique** puis sur **ENTREE**. Vous devriez obtenir l'effectif total de 1000.

### Comparaison avec les résultats que fournirait la loi $N(6 ; 1)$

La fonction de densité de la normale  $N(6 ; 1)$  est définie sur  $\mathbb{R}$  par :  $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-6)^2}$ .

L'histogramme étant construit à partir de classes d'amplitude 0,5, chaque rectangle devrait, dans le cadre de la loi normale  $N(6 ; 1)$ , avoir comme hauteur  $500 \times f(x)$  données.

En C1, taper "valeurs xi". En C2, entrer la valeur 2. En C3, entrer la formule : =C2+0,5 puis recopier cette formule vers le bas jusqu'en C18. Après **F9**, cette cellule contiendra la valeur 10. En D1, taper "loi normale". En D2, entrer la formule :  
=(500/RACINE(2\*PI()))\*EXP(-0,5\*(C2-6)^2) puis recopier cette formule vers le bas jusqu'en D18. Faire **F9**.



Cliquer sur l'icône **Assistant graphique**.

**Etape 1 sur 4** : dans l'onglet **Types personnalisés** choisir **Courbes-Histogramme** puis cliquer sur **Suivant**.

**Etape 2 sur 4** : dans l'onglet **Plage de données**, sortir, par l'icône indiqué ici, vers la feuille de calcul. Y sélectionner les valeurs  $n_i$  puis revenir, par l'icône analogue, dans la boîte de dialogue.

Dans l'onglet **Série**, pour la **Série 2**, **Etiquettes des abscisses X**,



sortir, sur la feuille de calcul, sélectionner les valeurs  $x_i$ . Cliquer sur **Ajouter** puis, pour la **Série 1**, sortir, sur la feuille de calcul, sélectionner les **Valeurs** (colonne des valeurs "loi normale"). Cliquer sur **Suivant**.

**Etape 3 sur 4** : dans l'onglet **Légende**, désélectionner **Afficher la légende**. Cliquer sur **Suivant**.

**Etape 4 sur 4** : cocher • **Sur une nouvelle feuille** puis **Terminer**.

Cliquer, avec le *bouton droit* de la souris, sur un point de la courbe (Série 1) puis choisir **Format de la série de données...** Dans l'onglet **Motifs**, pour **Traits**, augmenter un peu l'épaisseur et cocher la case **Lissage**, pour **Marque**, cocher **Aucune** puis faire **OK**.

Faire **F9** pour une nouvelle simulation de 1000 valeurs.

– Consigner vos commentaires sur la feuille réponse.

### 3- DROITE DE HENRY

On souhaite, dans ce paragraphe, utiliser la technique de la régression linéaire selon les moindres carrés pour quantifier la qualité de l'ajustement de la distribution observée avec une loi normale.

• On désigne maintenant par  $X$  une variable aléatoire suivant la loi  $N(\mu, \sigma)$ .

Sa fonction de répartition  $F$  est donnée, pour tout  $x \in \mathbb{R}$ , par :

$$y = F(x) = P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P\left(\frac{X - \mu}{\sigma} \leq t\right) = \Pi(t) \quad , \text{ avec } t = \frac{x - \mu}{\sigma} \text{ et où } \Pi$$

est la fonction de répartition de la loi  $N(0, 1)$ , tabulée dans le formulaire officiel.

• Statistiquement, pour une valeur  $x_i$  de la distribution,  $y_i$  est la *fréquence cumulée croissante* (analogue statistique de la fonction de répartition). Notons  $t_i$  la valeur, donnée par *lecture inverse* de la table de la loi  $N(0, 1)$ , telle que  $y_i = \Pi(t_i) \Leftrightarrow t_i = \Pi^{-1}(y_i)$ .

S'il s'agit d'une distribution normale, le nuage de points  $(x_i, t_i)$  devrait donc être ajusté par la droite d'équation  $t = \frac{x - \mu}{\sigma}$ , que l'on nomme **droite de Henry**.

• Sur la **feuille 2**, taper en E1 "ni cumulés" et E2, entrer la formule : =B2 puis en E3, la formule : =E2+B3 puis **Recopier vers le bas** jusqu'en E18 puis faire **F9**. La valeur calculée devrait être 1000.

En F1, taper "fréq cumul yi", puis en F2, entrer la formule : =E2/\$B\$19 puis recopier vers le bas jusqu'en F18 (le symbole \$ empêche la modification de la référence de la cellule lors de la recopie vers le bas). Faire **F9**.

En G1, taper "ti invnorm(yi)".

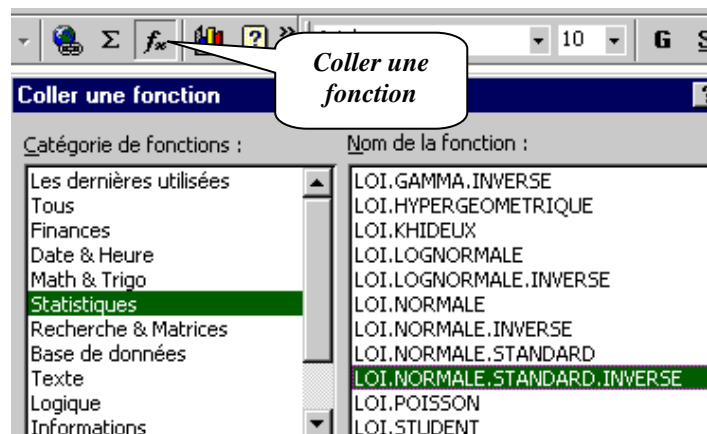
Cliquer en G2, puis sur l'icône **Coller une fonction**.

Choisir **Statistiques** puis **LOI.NORMALE.STANDARD.INVERSE** et cliquer sur **OK**.

Dans la boîte de dialogue, entrer pour **Probabilité** : F2.

Cliquer sur **OK**.

**Recopier vers le bas** jusqu'en G18 puis faire **F9**.



– Pour les valeurs 0 et 1 de  $y$ , Excel répond pour  $t$  : **#NOMBRE!**

Pouvez-vous expliquer pourquoi ?

On va maintenant représenter le nuage de points  $(x_i ; t_i)$ .

Cliquer sur l'icône *Assistant graphique*.

**Etape 1/4** : choisir *Nuages de points* (sous-type n°1 sans courbe).

Cliquer sur *Suivant*.

**Etape 2/4** : Onglet *Plage de données*, sortir vers la feuille de calcul pour y sélectionner, parmi les valeurs  $t_i$ , celles ne contenant pas #NOMBRE!

Revenir dans la boîte de dialogue de l'assistant graphique.

Onglet *Série*, Valeurs X : sortir sélectionner les cellules contenant les valeurs  $x_i$  correspondantes.

Cliquer sur *Suivant*.

**Etape 3/4** : Onglet *Légende*, désactiver l'option *Afficher la légende*. Onglet *Quadrillage*, désactiver les options. Cliquer sur *Suivant*.

**Etape 4/4** : cocher  *Sur une nouvelle feuille* puis cliquer sur *Terminer*.

On peut demander à Excel d'ajouter sur le nuage de points  $(x_i , t_i)$  une courbe de tendance.

Cliquer sur le graphique. Aller dans le menu *Graphique* (en haut de l'écran) et cliquer sur *Ajouter une courbe de tendance...* Dans la boîte de dialogue, dans l'onglet *Type* choisir *Linéaire* puis dans l'onglet *Option* cocher

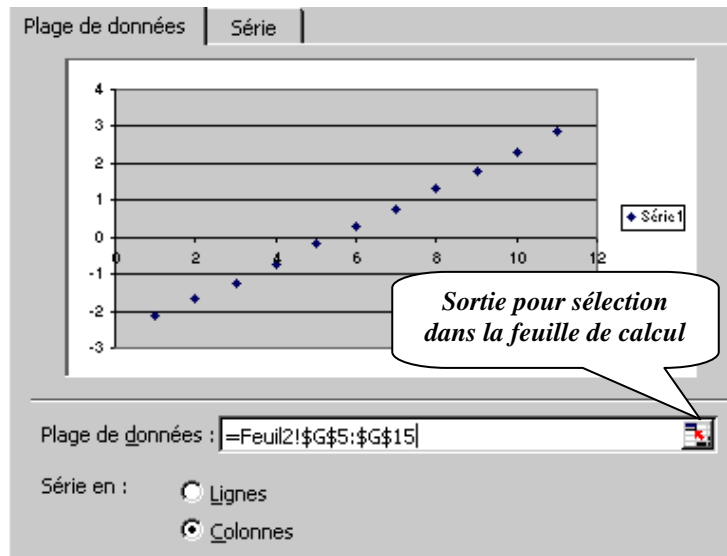
• *Afficher l'équation sur le graphique* et

• *Afficher le coefficient de détermination ( $R^2$ ) sur le graphique*

puis cliquer sur *OK*.

La droite affichée est la droite de Henry, obtenue par ajustement linéaire de  $t$  en  $x$  selon la méthode des moindres carrés. Faire **F9** pour une autre simulation de 1000 valeurs.

– Exploiter sur la feuille réponse les résultats affichés.



## 4- TEST DE NORMALITE D'UNE PRODUCTION

La durée de vie, exprimée en heures, des joints à lèvres PLAUSTRA - type IE - définit une variable aléatoire continue  $X$ .

L'étude de la durée de vie de 500 de ces joints a permis d'obtenir l'historique suivant :

Temps de bon fonctionnement $x_i$	500	700	900	1100	1300	1500	1700
Effectifs $n_i$	24	67	108	126	109	51	15

Reprendre, sur la feuille 3 du classeur Excel, les techniques du §3 pour :

1) Créer un tableau contenant les valeurs  $x_i$ ,  $n_i$ , les effectifs cumulés croissants  $y_i$  et les valeurs  $t_i = \Pi^{-1}(y_i)$  correspondantes.

2) Représenter graphiquement le nuage de points  $(x_i ; t_i)$ .

3) Y indiquer la droite de Henry et le coefficient de détermination.

— Exploiter les calculs d'Excel pour montrer, sur la feuille réponse, que l'on peut ajuster la distribution de la durée de vie des joints PLAUSTRA à une loi normale dont on précisera les paramètres.

**– FEUILLE REPONSE**

**NOMS :** .....

**1- GENERATION D'UNE DISTRIBUTION DE 1000 VALEURS**

**Loi uniforme sur [0 , 1]**

Soit  $X$  suivant la loi  $U ([0 , 1])$ .

$E(X) = \int_0^1 x dx = \dots\dots\dots$

$V(X) = E(X^2) - [E(X)]^2 = \int_0^1 x^2 dx - [E(X)]^2 = \dots\dots\dots$

**Somme de  $n = 12$  variables aléatoires indépendantes de même loi  $U [0 , 1]$**

**Moyenne  $\bar{x}$  et écart type  $s_e$  d'un échantillon de 1000 valeurs**

Compléter, pour quatre simulations, le tableau ci-dessous :

Simulation n°	1	2	3	4
$\bar{x}$				
$s_e$				

**Valeurs théoriques de  $\mu$  et  $\sigma$**

Soit  $Y = X_1 + X_2 + \dots + X_{12}$  où les  $X_i$  sont indépendantes de loi  $U ([0 ; 1])$ .

$\mu = E(Y) = 12 \times E(X) \dots\dots\dots$

$\sigma = \sqrt{V(Y)} = \sqrt{12V(X)} = \dots\dots\dots$

Comparer  $\bar{x}$  à  $\mu$  et  $s_e$  à  $\sigma$  : .....

**2 - THEOREME LIMITE CENTRAL**

**Un énoncé du théorème**

Pourquoi peut-on considérer que  $Y$  suit approximativement une loi normale ? .....

.....  
 .....  
 .....

Quels sont, théoriquement, les paramètres de cette loi normale ? .....

.....

**Comparaison graphique de l'histogramme des 1000 données à la densité normale**

Comparer, sur plusieurs simulations, l'histogramme avec le profil de la densité normale : ..

.....  
 .....

L'utilisation du théorème limite central était-elle justifiée ? .....

.....

.....

***Imprimer le graphique (si possible) ou enregistrer le fichier***

### 3- DROITE DE HENRY

On pose  $y = F(x) = \Pi(t)$  et  $t = \Pi^{-1}(y)$ .

Pour les valeurs 0 et 1 de  $y$ , Excel répond pour  $t$  : #NOMBRE! car  $\Pi^{-1}(0)$  correspondrait à .....

.....et  $\Pi^{-1}(1)$  correspondrait à .....

Déterminer, pour plusieurs simulations, la valeur du coefficient de corrélation linéaire  $R$  de  $t$  en  $x$  :

Simulation	1	2	3	4
$R \approx$				

Peut-on estimer que la série des 1000 données est issue d'une distribution normale ? .....

Pour une simulation où  $R$  est au moins de 0,9, donner l'équation  $t = ax + b$  de la droite de Henry fournie par Excel :

.....

S'il s'agit de loi  $N(\mu ; \sigma)$ , la droite de Henry a comme équation :  $t = \frac{1}{\sigma}x - \frac{\mu}{\sigma}$ .

En déduire une estimation de  $\mu$  et de  $\sigma$  :

$\sigma$  est estimé à .....

.....

$\mu$  est estimé à .....

.....

Comparer aux valeurs de  $\mu$  et de  $\sigma$  attendues (§1) : .....

### 4- TEST DE NORMALITE D'UNE PRODUCTION

Est-il raisonnable d'ajuster la distribution des durées de vie des joints PLAUSTRA à une loi normale ? .....

.....

Quelle est l'équation  $t = ax + b$  de la droite de Henry calculée par Excel ?

.....

En déduire une estimation de  $\mu$  et  $\sigma$  : .....

.....

.....

.....

***Imprimer le graphique (si possible) ou enregistrer le fichier***

## Corrigé et compte-rendu de l'activité EXCEL "AJUSTEMENT A UNE LOI NORMALE"

### 1- GENERATION D'UNE DISTRIBUTION DE 1000 VALEURS

#### Loi uniforme sur [0 , 1]

Si  $X$  suit la loi  $\mathcal{U}$  ([0 , 1], on a  $E(X) = \frac{1}{2}$  et  $V(X) = \frac{1}{12}$ .

#### Somme de $n = 12$ variables aléatoires indépendantes de même loi $\mathcal{U}$ [0 , 1]

Des valeurs expérimentales sur 1000 réalisations de  $Y$  :

Simulation n°	1	2	3	4
$\bar{x}$	6.001	5.974	5.974	6.026
$s_e$	1.008	1.016	1.004	1.008

Les valeurs théoriques sont  $\mu = E(Y) = 6$  et  $\sigma = \sigma(Y) = \sqrt{12 \times \frac{1}{12}} = 1$ .

Les valeurs observées sur les 1000 données sont très proches

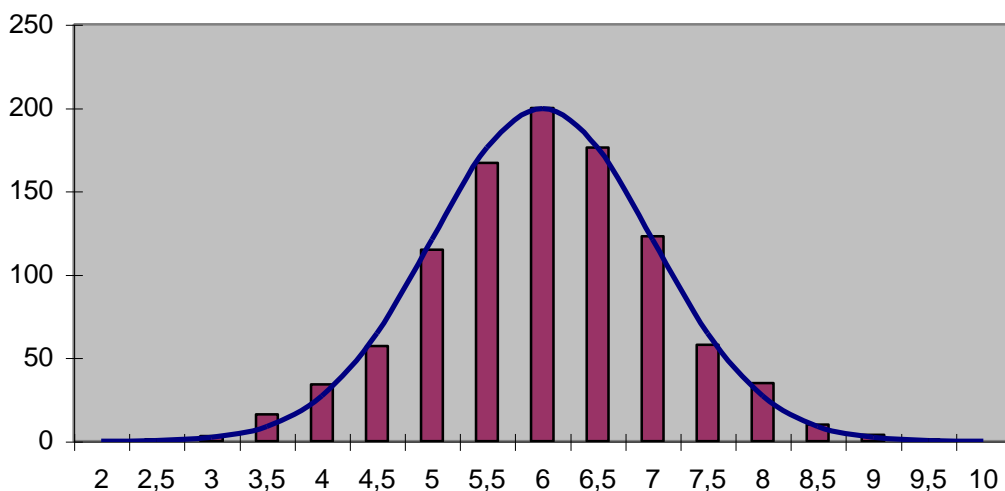
### 2 - THEOREME LIMITE CENTRAL

#### Un énoncé du théorème

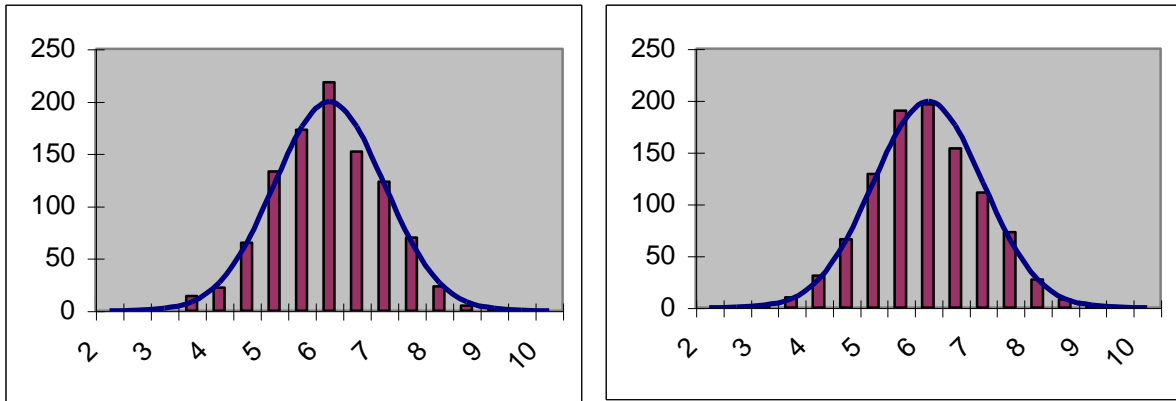
La variable aléatoire  $Y$  étant la somme de  $n = 12$  variables aléatoires indépendantes de même loi, elle suit approximativement (en supposant que  $n = 12$  est assez grand) une loi normale.

Il est clair que la moyenne et l'écart type de cette loi normale doivent être ceux de  $Y$  c'est à dire  $\mu = 6$  et  $\sigma = 1$ .

#### Comparaison de l'histogramme des 1000 réalisations de $Y$ avec la densité de la loi $\mathcal{N}(6, 1)$



Il suffit de faire F9, pour avoir aussitôt une autre simulation. On expérimente ainsi l'approximation donnée par le théorème limite central. Il apparaît ici que  $n = 12$  est suffisant pour une bonne approximation normale.



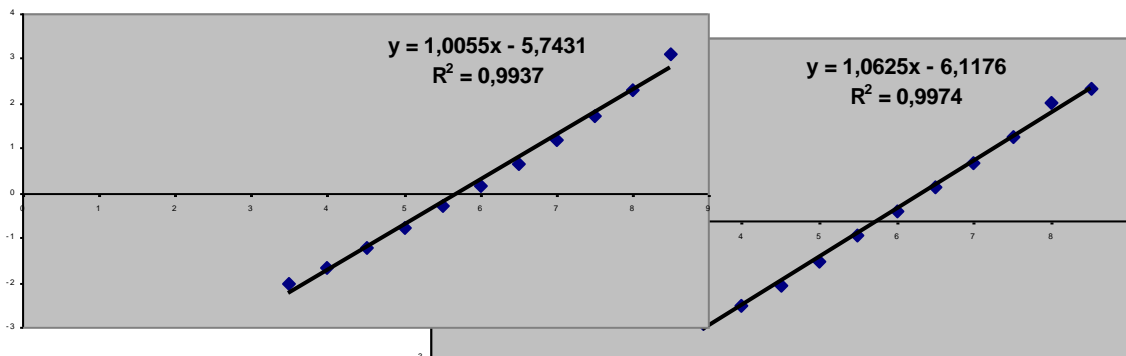
### 3- DROITE DE HENRY

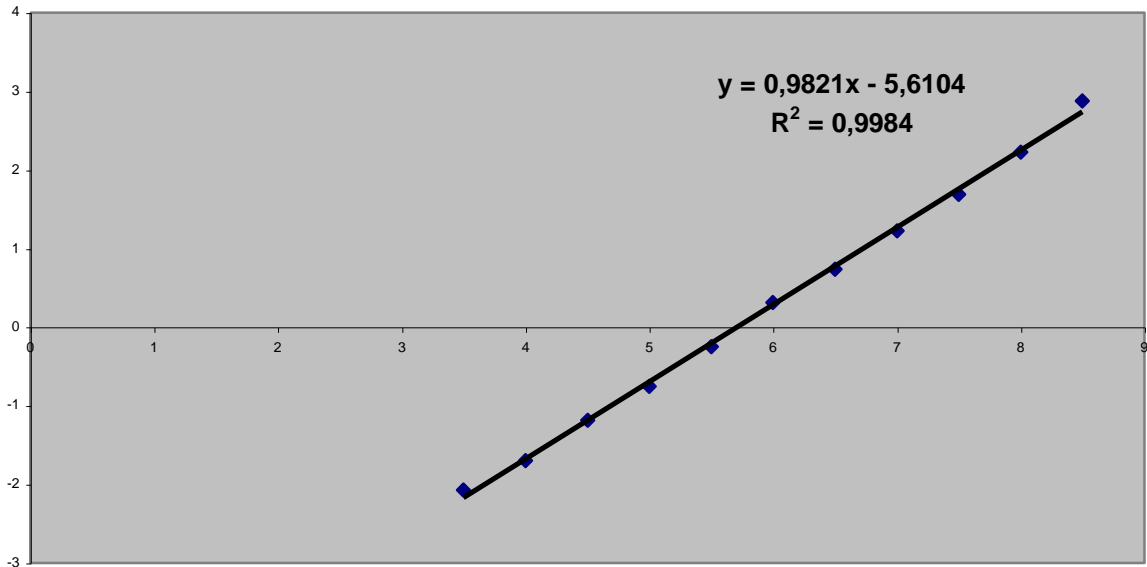
	A	B	C	D	E	F	G
1	sup classes	effectifs ni	valeurs xi	loi normale	ni cumulés	fréq cumul yi	ti invnorm(yi)
2	2,25	0	2	0,06691511	0	0	#NOMBRE!
3	2,75	0	2,5	0,43634135	0	0	#NOMBRE!
4	3,25	0	3	2,21592421	0	0	#NOMBRE!
5	3,75	7	3,5	8,76415025	7	0,007	-2,45727279
6	4,25	27	4	26,9954833	34	0,034	-1,82500571
7	4,75	71	4,5	64,7587978	105	0,105	-1,25356564
8	5,25	144	5	120,985362	249	0,249	-0,67763949
9	5,75	180	5,5	176,032663	429	0,429	-0,17892035
10	6,25	182	6	199,47114	611	0,611	0,2819263
11	6,75	156	6,5	176,032663	767	0,767	0,7290032
12	7,25	122	7	120,985362	889	0,889	1,22122856
13	7,75	75	7,5	64,7587978	964	0,964	1,79911694
14	8,25	22	8	26,9954833	986	0,986	2,19728463
15	8,75	9	8,5	8,76415025	995	0,995	2,57583451
16	9,25	5	9	2,21592421	1000	1	#NOMBRE!
17	9,75	0	9,5	0,43634135	1000	1	#NOMBRE!
18	10,25	0	10	0,06691511	1000	1	#NOMBRE!

On aurait  $\Pi^{-1}(0)$  qui vaudrait  $-\infty$  et  $\Pi^{-1}(1)$  qui vaudrait  $+\infty$ , c'est la raison de la réponse #NOMBRE!

Simulation	1	2	3	4
R ≈	0,999	0,999	0,998	0,862

De façon générale, R est très proche de 1 et on peut donc affirmer que les 1000 données sont issues d'une distribution approximativement normale.





Pour le graphique ci-dessus (par exemple), on a  $t = 0,9821 x - 5,6104$ .

D'où  $\frac{1}{\sigma} = 0,9821$  et  $\sigma$  est estimé à 1,02. Puis  $\frac{\mu}{\sigma} = 5,6104$  qui permet d'estimer  $\mu$  à 5,71.

Ces valeurs sont assez proches des valeurs théoriques :  $\mu = 6$  et  $\sigma = 1$ .

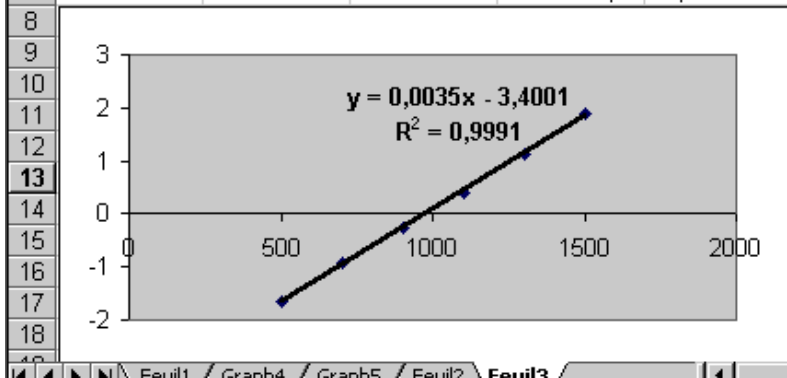
#### 4- TEST DE NORMALITE D'UNE PRODUCTION

	A	B	C	D	E
1	xi	ni	ni cumulés	yi	ti
2	500	24	24	0,048	-1,66456175
3	700	67	91	0,182	-0,90776894
4	900	108	199	0,398	-0,25852728
5	1100	126	325	0,65	0,38532107
6	1300	109	434	0,868	1,11698682
7	1500	51	485	0,97	1,88078957

On obtient un coefficient de corrélation  $R \approx 0,9995$  qui justifie l'ajustement des durées de vie à une distribution normale.

On a  $\frac{1}{\sigma} = 0,0035$  d'où  $\sigma$  estimé à 285,7 heures.

Puis  $\frac{\mu}{\sigma} = 3,4001$  qui conduit à estimer  $\mu$  à 971,5 heures.



Feuil1 / Graph4 / Graph5 / Feuil2 / Feuil3



## ANNEXE 4 – Un exemple d'introduction à la notion de test

Forme	Niveau	Prérequis	Durée approximative
Travaux dirigés utilisant la <b>calculatrice programmable</b>	B.T.S.	Loi binomiale Loi normale	1h30

Les travaux dirigés qui suivent ont été pratiqués en sections de techniciens supérieurs, comme introduction à la notion de test d'hypothèses. Bien que les notions d'erreur de seconde espèce et de puissance d'un test ne soient pas au programme de BTS, ni les choix de construction du test demandés à l'examen, ce sont des enjeux essentiels du test. On se situe alors dans un souci de formation, sans exiger de connaissance sur ces sujets.

On se place dans un cas simple (test d'une fréquence dans un cadre binomial), sur un sujet sensible (la réussite à un examen) où les différents enjeux (risques) ont une signification claire. Les élèves ont été intéressés, et motivés par le contexte.

Il s'agit de faire comprendre les éléments essentiels suivants :

- La **construction du test** doit se faire **avant** la prise d'échantillon. Elle doit faire l'objet d'un **protocole** sur lequel se mettent d'accord les deux partis en présence, ici professeurs et élèves, dans les relations commerciales, vendeur et acheteur. D'où la nécessité d'une **normalisation** des tests.
- Les **erreurs** sont inévitables. Elles sont de deux types et les **choix** effectués pour la construction du test correspondent à un **compromis** entre la maîtrise des risques  $\alpha$  et  $\beta$  et la taille  $n$  de l'échantillon (coût du contrôle).

Bien qu'hors programme des BTS, ce sont des enjeux importants du test, et l'on peut les faire comprendre, sans entrer dans une étude systématique.

L'erreur de 1<sup>ère</sup> espèce  $\alpha$  étant la plus facile à maîtriser ( $\beta$  ne peut être déterminée que si l'on connaît les lois de probabilité sous  $H_1$ ), c'est sur elle, et l'hypothèse  $H_0$ , que sera construit le test. Cela conduit à **privilegier  $H_0$**  : si la forme de la région d'acceptation dépend de la nature de  $H_1$  (bilatéral ou unilatéral), ses limites ne dépendent que de  $H_0$  (à partir de laquelle, on les calcule). Les deux hypothèses ne jouent donc pas un rôle symétrique.

Il faut mettre en évidence les différentes **étapes d'un test**, dont le plan est :

### 1. Construction du test :

a - **Choix des hypothèses**  $H_0$  et  $H_1$  (test bilatéral ou unilatéral).

Ce choix est dirigé par l'énoncé. Celui de  $\alpha$  et  $n$  est imposé à l'examen.

b - **Calcul**, sous l'hypothèse  $H_0$ , **de la région critique au seuil  $\alpha$**  (ou de la zone d'acceptation).

c - **Enoncé de la règle de décision**.

### 2. Utilisation du test :

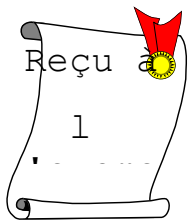
Prélèvement d'un échantillon et prise de décision, selon le résultat observé.

Un corrigé est donné à la suite des "documents élèves".

# INTRODUCTION AUX TESTS STATISTIQUES

Dans un lycée syldave, les professeurs, exaspérés par le manque de travail d'une partie des étudiants de BTS, décident d'établir un examen de passage à la fin du premier semestre (les mœurs syldaves sont assez rudes ...).

L'examen se présentera sous la forme d'un QCM de 20 questions indépendantes. A chaque question, trois réponses sont proposées, dont une seule est exacte. Un étudiant n'ayant fourni aucun travail, répondra au hasard et donc, correctement, avec une probabilité  $p = \frac{1}{3}$  à chaque question.



L'objectif des professeurs est de recalser ce type d'étudiant, avec une probabilité d'environ 95 %.

Pour cela il faut définir la barre d'acceptation *avant* l'épreuve, de sorte que les étudiants souscrivent au **protocole** ("règles du jeu") de l'examen.

On peut considérer ce QCM comme un **test statistique** devant permettre de détecter si l'étudiant qui le passe répond au hasard ( $p = \frac{1}{3}$ ). Le QCM est un **échantillon aléatoire** non exhaustif de ses réponses.

Etudiant :  $p = \frac{1}{3}$  ?

QCM :  $n = 20$   
taux de bonnes  
réponses  $f\%$

## I - CONSTRUCTION DU TEST

### 1) Choix des hypothèses

On teste l'hypothèse  $H_0$  : " $p = \frac{1}{3}$ " (appelée "**hypothèse nulle**"), contre l'**hypothèse alternative**  $H_1$  : " $p > \frac{1}{3}$ " (c'est un test "**unilatéral**").

L'hypothèse nulle correspond à un étudiant répondant au hasard. L'hypothèse alternative doit "au contraire" correspondre à un étudiant qui a travaillé. Pourquoi ne prend-on pas " $p \neq \frac{1}{3}$ " ? .....

.....

### 2) Calcul de la zone d'acceptation de $H_0$

On suppose que  $H_0$  est vraie : l'étudiant répond au hasard. On désigne par  $X$  la variable aléatoire qui, à chaque étudiant de ce type, associe le nombre de ses bonnes réponses au QCM.

Quelle est le loi de  $X$  (justifier) ? .....

.....

.....

A l'aide de la table ci-contre, déterminer le nombre  $k$  de bonnes réponses tel que  $P(X \leq k)$  soit le plus proche possible de 95 %.

.....  
 .....

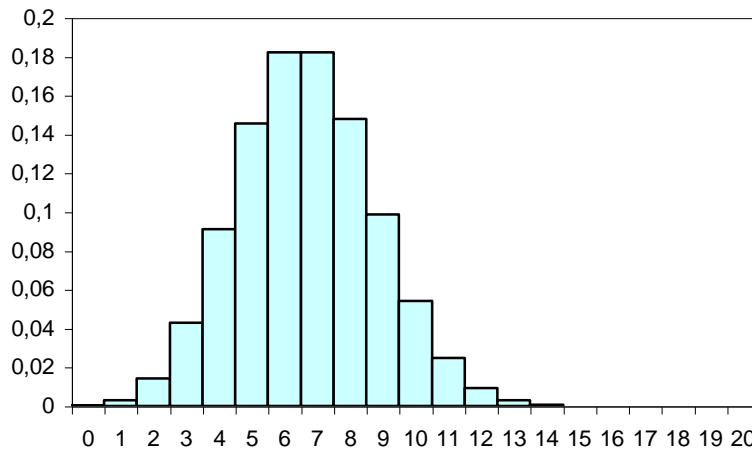
Quand le nombre de bonnes réponses est inférieur ou égal à  $k$ , on acceptera  $H_0$ .

S'il est strictement supérieur à  $k$ , on supposera que l'étudiant a travaillé et l'on rejettera  $H_0$ , avec un *risque* de rejet à tort de :  $\alpha = P(X > k)$ .

Quel est, ici, le risque  $\alpha$  ? .....

Sur le graphique suivant, indiquer la zone d'acceptation et la zone de rejet de  $H_0$ .

n = 20 et p = 1/3	
k	P(X ≤ k)
0	0,00030073
1	0,00330802
2	0,01759263
3	0,06044646
4	0,15151086
5	0,29721389
6	0,47934269
7	0,66147148
8	0,80945113
9	0,90810423
10	0,96236343
11	0,9870267
12	0,99627543
13	0,99912119
14	0,99983263
15	0,99997492
16	0,99999715
17	0,99999977
18	0,99999999
19	1
20	1



### 3) Règle de décision

Enoncer la *règle de décision* de l'examen.

.....  
 .....

## II - UTILISATION DU TEST ET ERREURS

### 1) Expérimentation du test

Le programme suivant choisit aléatoirement une valeur de  $p$  : avec une chance sur deux,  $p = \frac{1}{3}$  ou  $p = 0,60$  (cas d'un étudiant ayant moyennement travaillé). Puis, il simule le passage de l'examen et affiche le nombre  $x$  de réponses correctes ainsi que la valeur de  $p$ .

CASIO	TI 82 - 83	TI 89 - 92
1÷3+Int(Ran#+.5).(6-1÷3) → P.␣	:1/3+int(rand+.5).(6-1/3) → P	:1/3+int(rand()+.5).(6-1/3) → p
0 → X.␣	:0 → X	:0 → x
For 1→I To 20.␣	:For(I,1,20)	:For i,1,20
Int(Ran# + P)+X → X.␣	:int(rand+P)+X → X	:int(rand()+p)+x → x
Next.␣	:End	:EndFor
X //	:Disp X , P	:Disp x , p
P		

L'examen conduit-il toujours à une décision juste ?

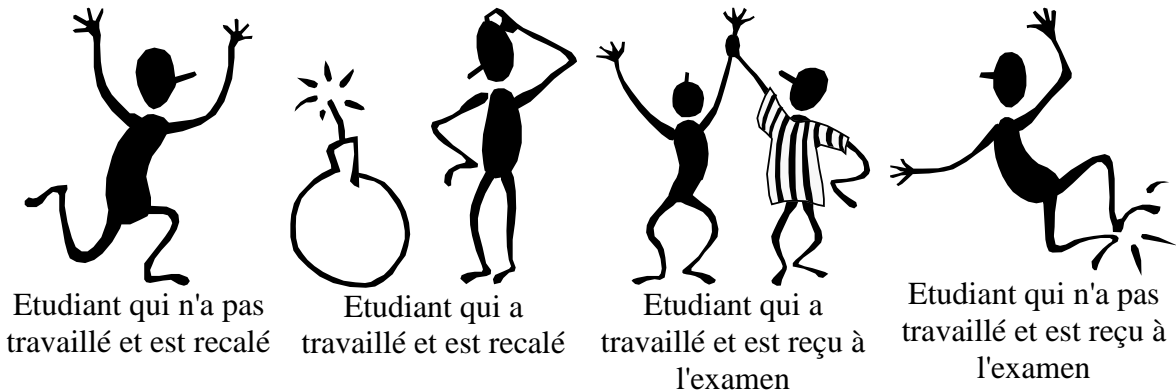
.....

## 2) Les erreurs

Décision \ Réalité	H <sub>0</sub> acceptée	H <sub>1</sub> acceptée
H <sub>0</sub> vraie	1 - α	α ERREUR DE 1 <sup>ère</sup> ESPECE
H <sub>1</sub> vraie	β ERREUR DE 2 <sup>nde</sup> ESPECE	1 - β

Il y a quatre situations possibles. Les **erreurs de décision** sont de deux types : "rejeter H<sub>0</sub> à tort" (erreur de première espèce correspondant au risque α) ou "accepter H<sub>0</sub> à tort".

Relier chaque dessin à la case qui lui correspond dans le tableau.



## 3) L'erreur de 2<sup>nde</sup> espèce

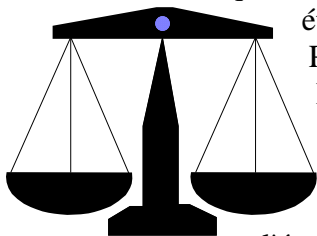
Un étudiant se manifeste alors. Il est sérieux et travailleur, mais, mal assuré, il perd souvent une partie de ses moyens à l'examen. Il estime cependant sa probabilité de bien répondre à une question à  $p = 0,6$ .

*"C'est pas juste ! Bien qu'ayant une probabilité de bonne réponse de 60 %, j'ai une chance sur quatre d'être recalé !"*

Vérifier l'affirmation de cet étudiant, qui craint d'être victime d'une erreur de 2<sup>nde</sup> espèce (utiliser la table ci-contre, des valeurs cumulées de la loi B (20 ; 0,60)).

n = 20	p = 0,60
k	P(X ≤ k)
0	1,09951E-08
1	3,40849E-07
2	5,04126E-06
3	4,7345E-05
4	0,000317031
5	0,001611525
6	0,006465875
7	0,021028927
8	0,056526367
9	0,127521246
10	0,244662797
11	0,404401275
12	0,584107062
13	0,749989328
14	0,874401027
15	0,949048047
16	0,984038837
17	0,996388528
18	0,999475951
19	0,999963438
20	1

Il faut avouer que ce n'est pas très moral vis à vis de cet étudiant.



Pour diminuer le risque de 2<sup>nde</sup> espèce, l'étudiant propose de baisser la barre d'admission à 8 : si le nombre  $x$  de bonnes réponses est tel que  $x \leq 7$ , l'étudiant est recalé, si  $x \geq 8$ , l'étudiant est reçu.

Quel est, dans ces conditions, le risque β de 2<sup>nde</sup> espèce, pour un étudiant tel que  $p = 0,60$  ? .....

Mais que devient le risque α d'admettre un étudiant n'ayant pas travaillé ? .....

### III - TEST DE 100 QUESTIONS

Les professeurs jugeant ce risque de première espèce inacceptable, décident, pour diminuer  $\beta$  sans augmenter  $\alpha$ , de proposer un QCM de 100 questions.

#### 1) Construction du test

• *Choix des hypothèses :*

On teste l'hypothèse  $H_0 : " p = \frac{1}{3} "$  , contre  $H_1 : " p > \frac{1}{3} "$  .

• *Calcul de la zone d'acceptation de  $H_0$ , au seuil  $\alpha$  de 5 % :*

On suppose que  $H_0$  est vraie :  $p = \frac{1}{3}$  . On désigne par  $X$  la variable aléatoire qui, à chaque étudiant de ce type, associe le nombre de ses bonnes réponses au QCM. On sait que  $X$  suit la loi  $B(100, \frac{1}{3})$  . Pour simplifier les calculs, on approche la loi de  $X$  par une loi normale.

Quels en sont les paramètres ? .....

On note  $F = \frac{1}{100}X$  , la variable aléatoire correspondant aux fréquences des bonnes réponses à un QCM. En supposant que  $F$  suive une loi normale, quels en sont les paramètres, sous l'hypothèse  $H_0$  ? .....

Déterminer le réel  $h$  tel que, sous l'hypothèse  $H_0$ ,  $P(F \leq h) = 0,95$ .

.....  
 .....  
 .....

• *Règle de décision :*

.....  
 .....

#### 2) Simulation

Reprendre le programme précédent, en remplaçant **20** par **100**.

Comparer l'efficacité de ce Q.C.M. au précédent (observer la fréquence des erreurs de première et seconde espèce).

.....  
 .....  
 .....  
 .....

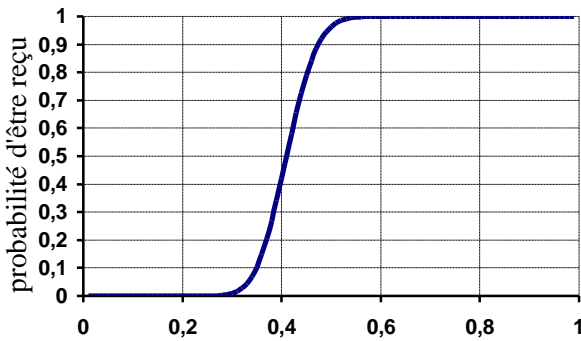
**POUR ALLER PLUS LOIN ...**

**Puissance du test**

Plaçons nous maintenant du point de vue de l'étudiant, pour lequel  $p = p_0$  et qui se préoccupe de sa probabilité d'être reçu :  $1 - \beta(p_0) = P(F > 0,41 \mid p = p_0)$ .

Montrer que, si  $p = p_0$ , on a  $1 - \beta(p_0) = P(F > 0,41) = 1 - \Pi \left( 10 \times \frac{0,41 - p_0}{\sqrt{p_0(1 - p_0)}} \right)$ , où  $\Pi$  est

la fonction de répartition de la loi  $N(0, 1)$ . .....



La fonction  $p \mapsto 1 - \beta(p)$  se nomme "**puissance**" du test et est représentée ci-contre.

Lire sur le graphique, la probabilité d'être reçu au nouvel examen, d'un étudiant tel que  $p = 0,40$  , puis tel que  $p = 0,60$ .

Confirmer par le calcul.

Est-ce raisonnable ?

.....  
 .....  
 .....

.....  
 .....

## Corrigé des travaux dirigés "INTRODUCTION AUX TESTS STATISTIQUES"

### I – CONSTRUCTION DU TEST

1) Le cas  $p < 1/3$  n'est pas envisagé (que dire d'un étudiant cherchant à répondre faux ?). La forme de la région de rejet dépend de  $H_1$  qui correspond uniquement à  $p > 1/3$  (l'étudiant ne répond pas au hasard).

2) Si  $H_0$  est vraie, on a  $p = 1/3$ . Répondre au Q.C.M. est alors la répétition de 20 expériences aléatoires indépendantes, avec deux issues possibles (bonne réponse avec  $p = 1/3$ , ou mauvaise réponse) et où  $X$  associe le nombre de bonnes réponses. La variable aléatoire  $X$  suit donc la loi binomiale  $B(20, 1/3)$ .

Avec la table fournie, on constate que  $k = 10$ , avec  $P(X \leq 10) \approx 0,96$ .

Le risque  $\alpha$  est donc  $\alpha \approx 4\%$ .

3) Règle de décision

Soit  $x$  le nombre de bonnes réponses au Q.C.M.,

- si  $x \leq 10$  alors  $H_0$  est acceptée et l'étudiant est RECALE,
- si  $x \geq 11$  alors  $H_0$  est refusée et l'étudiant est ADMIS.

### II – UTILISATION DU TEST ET ERREURS

1) La simulation permet de "vivre" les aléas du hasard, et d'observer les deux types d'erreurs.

2) On observe deux types d'erreurs de décision.

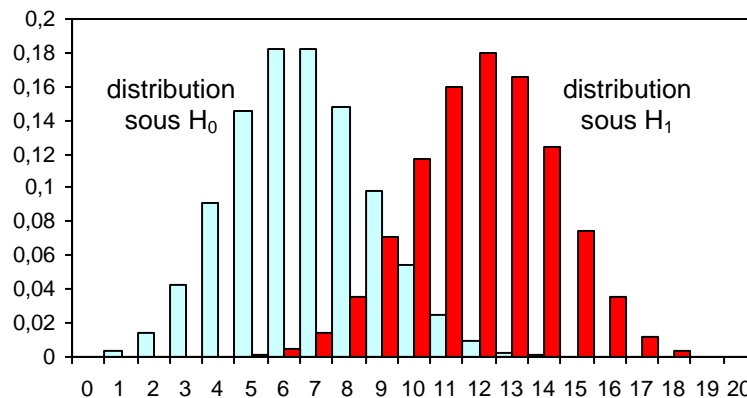
L'erreur de première espèce correspond à l'étudiant qui n'a pas travaillé et est reçu à l'examen.

L'erreur de seconde espèce correspond à l'étudiant qui a travaillé et est recalé.

3) L'erreur de seconde espèce :

Si  $H_1$  est vraie, on a  $p = 0,6$  et la variable aléatoire  $X$  suit alors la loi binomiale  $B(20 ; 0,60)$ .

On a alors  $P(X \leq 10) \approx 0,24$  et donc  $\beta \approx 24\%$ .



L'acceptation de  $H_0$  étant fixée à  $x \leq 10$ , l'erreur de 1<sup>ère</sup> espèce (seuil) correspond aux rectangles clairs 11, 12, 13 ..., de la distribution sous  $H_0$ , et l'erreur de 2<sup>nd</sup>e espèce (pour  $p = 0,6$ ) aux rectangles foncés 10, 9, 8, 7, 6, 5 ...

En abaissant la barre d'admission à 8 ( $H_0$  acceptée lorsque  $x \leq 7$ ), on alors, d'après la table de la loi  $B(20 ; 0,60)$ ,  $\beta \approx 2\%$ .

Mais alors, d'après la table de la loi  $B(20 ; 1/3)$ ,  $\alpha \approx 100 - 66 = 34\%$  !

### III – TEST DE 100 QUESTIONS

#### 1) Construction du test

Sous l'hypothèse  $H_0$ , on approche la loi  $B(100, 1/3)$  de la variable aléatoire  $X$  par la loi normale  $N\left(\frac{100}{3}, \sqrt{100} \times \frac{1}{3} \times \frac{2}{3}\right)$ .

La variable aléatoire  $F = \frac{1}{100}X$  suit alors approximativement la loi  $N\left(\frac{1}{3}, \frac{\sqrt{2}}{30}\right)$ .

On pose  $T = \frac{F - \frac{1}{3}}{\frac{\sqrt{2}}{30}}$ , qui suit la loi  $N(0, 1)$ . On a  $P(F \leq h) = P\left(T \leq \left(h - \frac{1}{3}\right) \times \frac{30}{\sqrt{2}}\right) = 0,95$  d'où,

d'après la table de la loi normale centrée réduite,  $h = \frac{1}{3} + 1,645 \frac{\sqrt{2}}{30} \approx 0,41$ .

La règle de décision du test de 100 questions, au seuil de 5 % est donc :

Soit  $x$  le nombre de bonnes réponses. Si  $x \leq 41$ , le candidat est recalé. Si  $x \geq 42$ , le candidat est admis.

#### 2) Puissance du test

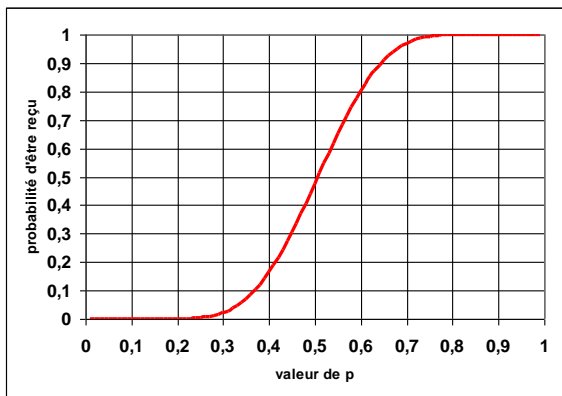
Sous l'hypothèse  $p = p_0$ , la variable aléatoire  $F$  suit approximativement la loi normale  $N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{100}}\right)$ . On a donc  $P(F > 0,41) = P\left(T > \frac{0,41 - p_0}{\sqrt{p_0(1-p_0)}} \times 10\right)$  et la probabilité

d'être reçu est donc bien  $1 - \beta(p_0) = P(F > 0,41) = 1 - \Pi\left(10 \times \frac{0,41 - p_0}{\sqrt{p_0(1-p_0)}}\right)$ .

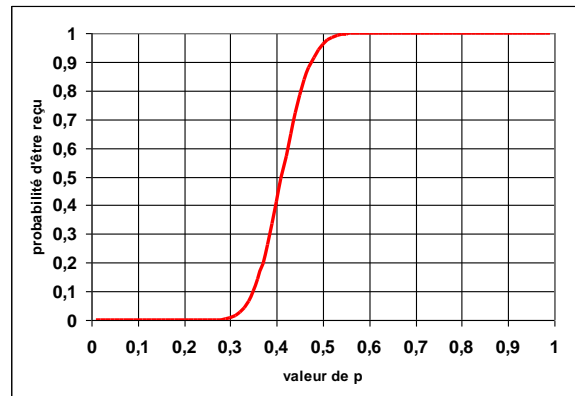
Pour un étudiant tel que la probabilité de bonne réponse est  $p = 0,4$ , la probabilité d'être reçu à l'examen est  $1 - \beta(0,4) = 1 - \Pi(0,204) \approx 0,42$ .

Pour un étudiant tel que la probabilité de bonne réponse est  $p = 0,6$ , la probabilité d'être reçu à l'examen est  $1 - \beta(0,6) = 1 - \Pi(-3,88) \approx 0,99995$ .

Ces résultats sont "raisonnables", c'est à dire conformes à ce que l'on attend d'un examen. Ce test est donc "puissant" en ce sens que son pouvoir de discrimination est important.



$n = 20$



$n = 100$

Comparaison des courbes de puissance des tests au seuil  $\alpha$  de 5 %, pour  $n = 20$  questions et pour  $n = 100$  questions.



## ANNEXE 5 – La maîtrise statistique des procédés de production

Forme	Niveau	Prérequis	Durée approximative
Travaux pratiques	B.T.S. (deuxième année)	Régression linéaire Loi binomiale Loi normale Echantillonnage	2h

Il s'agit d'étudier une application industrielle importante de la statistique : le contrôle de production.

Dans les activités proposées (ici sous forme de travaux dirigés pendant le cours de maths), on suit le "film" du contrôle de qualité, dans l'ordre où il est pratiqué en entreprise, selon les étapes suivantes :

- 1- Etude de la normalité de la production.
- 2- Etude de la "capabilité" des moyens de production (selon les tolérances).
- 3- Etablissement de "cartes de contrôle".

On a choisi, dans ce TP, d'étudier la normalité par régression linéaire selon les moindres carrés, pour illustrer cet outil puissant, au programme des sections de techniciens supérieurs. Cependant, dans la pratique en entreprise, on effectue, soit un ajustement linéaire "à l'œil" par utilisation du papier gaussien-arithmétique, soit, usant de moyens informatiques, un test statistique du type  $\chi^2$  ou *Kolmogorov-Smirnov*.

On trouvera dans les pages suivantes :

- Le document fourni aux élèves.
- Un corrigé.

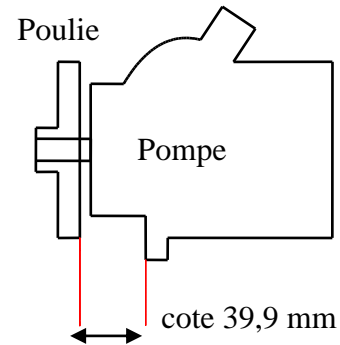
# TRAVAUX PRATIQUES

## LA MAITRISE STATISTIQUE DES PROCÉDES DE PRODUCTION

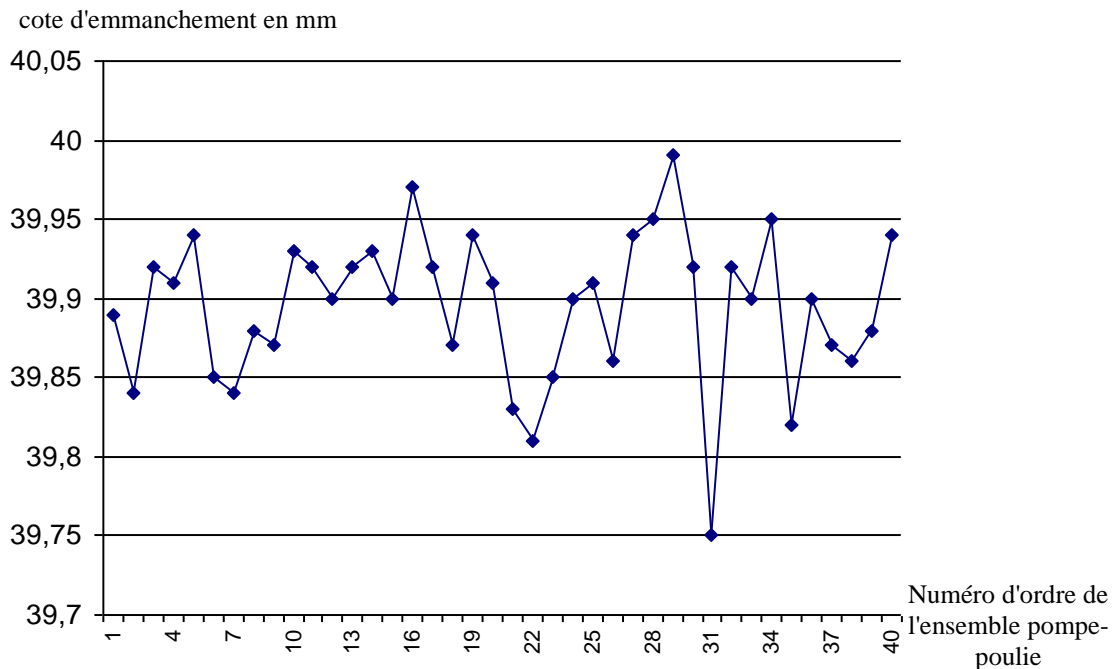
Dans le cadre de la "Maîtrise Statistique des Procédés", on étudie la variabilité de la production. Un des objectifs est de détecter les anomalies, en temps réel.

L'exemple étudié, issu de l'industrie automobile, est une presse d'emmanchement de poulie sur une pompe de direction assistée. Les performances de la presse sont variables, cette variabilité ayant de nombreuses causes possibles : main-d'œuvre, matériel, matière première, environnement de l'atelier, méthodes d'organisation...

L'emmanchement de la poulie sur l'axe de la pompe est mesuré par la cote de 39,9 mm indiquée sur le schéma ci-contre.



On a mesuré cette cote, à  $10^{-2}$  mm près, sur 40 ensembles pompe-poulie, produits de façon successive dans la production en série. Les observations sont représentées, dans l'ordre chronologique, sur le schéma suivant.

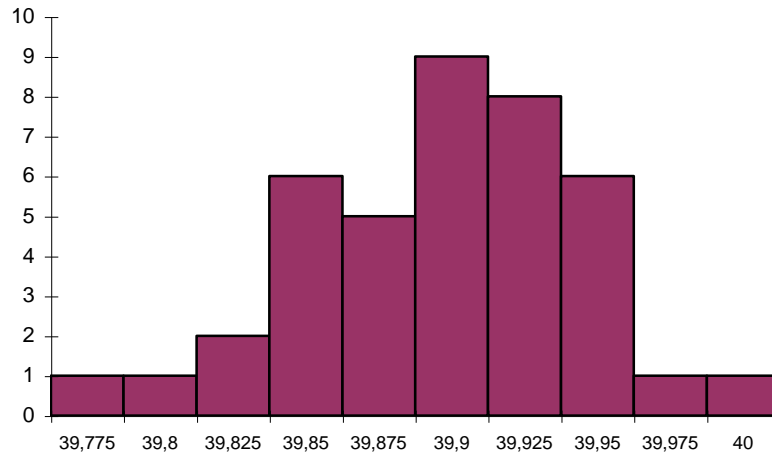


## I – ETUDE DE NORMALITE

Une première étape consiste à voir si les variations observées peuvent raisonnablement résulter d'un phénomène suivant une loi normale.

Nous allons d'abord regrouper les 40 observations en 10 classes d'égale amplitude 0,25 et construire un histogramme. Les résultats sont les suivants :

centres des classes	39,775	39,8	39,825	39,85	39,875	39,9	39,925	39,95	39,975	40
effectifs	1	1	2	6	5	9	8	6	1	1



Dans un deuxième temps, nous allons comparer ces résultats observés à ceux, théoriques, correspondant à une variable aléatoire  $X$  de loi normale  $N(\mu, \sigma)$ . Cette comparaison se fera à l'aide, d'une part des fréquences cumulées observées, et, d'autre part, de la fonction de répartition de  $X$ .

1) Fréquences cumulées observées :

On note  $x_i$  les bornes supérieures des classes et  $y_i$  les fréquences cumulées correspondantes. C'est à dire que  $y_i$  est la fréquence des observations inférieures ou égales à  $x_i$ .

Calculer la fréquence cumulée de la case "oubliée".

bornes sup des classes $x_i$	39,7875	39,8125	39,8375	39,8625	39,8875	39,9125	39,9375	39,9625	39,9875	40,0125
fréquences cumulées $y_i$	0,025	0,05		0,25	0,375	0,6	0,8	0,95	0,975	1

2) Fréquences théoriques selon la loi normale :

Si  $X$  suit la loi  $N(\mu, \sigma)$  alors  $T = \frac{X - \mu}{\sigma}$  suit la loi normale centrée réduite  $N(0, 1)$  et

$$y_i = F(x_i) = P(X \leq x_i) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x_i - \mu}{\sigma}\right) = P(T \leq t_i) = \Pi(t_i) \text{ avec } t_i = \frac{x_i - \mu}{\sigma} \text{ et où } \Pi$$

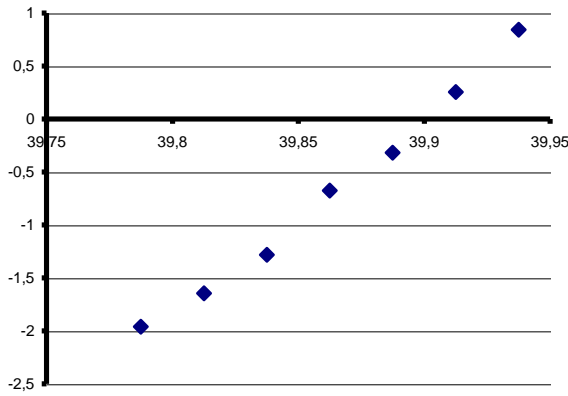
est la fonction de répartition de la loi  $N(0, 1)$ , tabulée dans le formulaire officiel de BTS.

On peut, à partir des fréquences cumulées observées  $y_i$ , calculer, par lecture inverse de la table de la loi  $N(0, 1)$ , les valeurs  $t_i$  telles que  $y_i = \Pi(t_i) \Leftrightarrow t_i = \Pi^{-1}(y_i)$ .

Calculer, à  $10^{-2}$  près, la valeur  $t_7$  oubliée dans le tableau ci-dessous, c'est à dire telle que  $P(T \leq t_7) = 0,8$  où  $T$  suit la loi normale  $N(0, 1)$ .

$y_i$	0,025	0,05	0,1	0,25	0,375	0,6	0,8	0,95	0,975	1
$t_i \approx$	-1,96	-1,645	-1,281	-0,674	-0,319	0,253		1,645	1,96	

3) Régression linéaire (droite de *Henry*) :



Si la distribution observée est extraite d'une population normale, on devrait avoir  $t_i \approx \frac{1}{\sigma} x_i - \frac{\mu}{\sigma}$ .

a) Sur le graphique ci-contre, on a représenté les points de coordonnées  $(x_i, t_i)$ . Les points semblent-ils pratiquement alignés ?

b) Calculer une équation  $t = ax + b$  de la droite d'ajustement de  $t$  en  $x$  selon la méthode des moindres carrés pour le

tableau ci-dessous (on arrondira à  $10^{-3}$  près) :

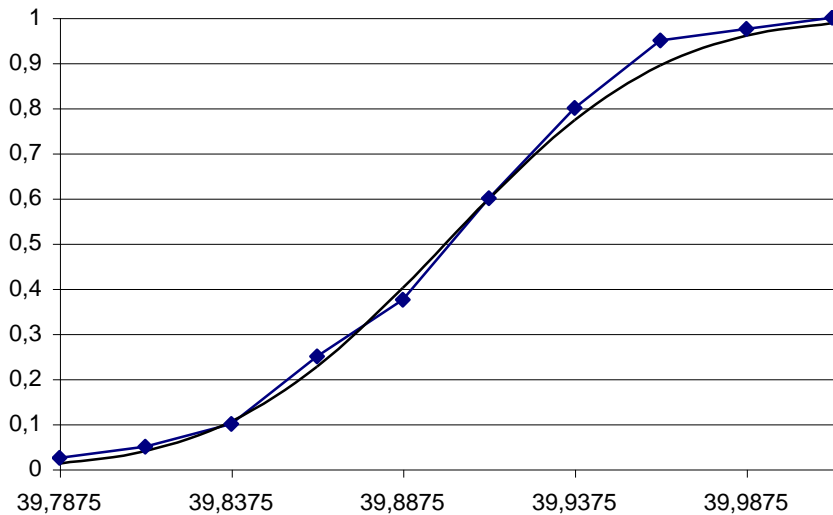
$x_i$	39,7875	39,8125	39,8375	39,8625	39,8875	39,9125	39,9375	39,9625	39,9875
$t_i$	-1,96	-1,645	-1,281	-0,674	-0,319	0,253	0,842	1,645	1,96

c) Que vaut le coefficient de corrélation  $r$  ? Comment peut-on l'interpréter ?

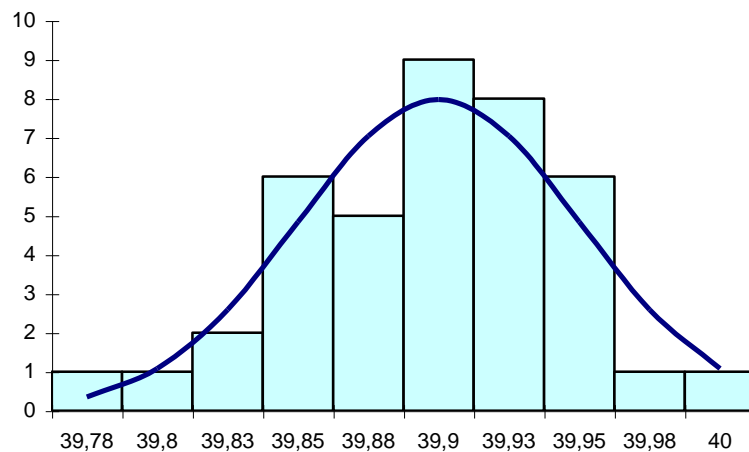
d) Vérifier que, d'après les résultats précédents, on peut approximativement prendre

$$\sigma = \frac{1}{a} \approx 0,05 \text{ et } \mu = -\frac{b}{a} \approx 39,9.$$

Voici, à titre indicatif, d'autres graphiques permettant de visualiser la normalité des données :



Fréquences cumulées croissantes observées  $y_i$   
et fonction de répartition  $F$  de la loi  $N(39,9 ; 0,05)$



Histogramme des observations  
et densité de la loi normale  $N(39,9 ; 0,05)$

## II – CAPABILITE

Le bureau d'étude a défini, pour cette cote, l'**intervalle de tolérance** suivant :  
 $[39,9 - 0,2 \text{ mm} ; 39,9 + 0,2 \text{ mm}]$ . C'est à dire, qu'en dehors de cet intervalle, l'emmanchement sera considéré comme non conforme.

Le procédé de fabrication est considéré comme "**capable**" lorsque la probabilité de fabrication d'une pièce hors tolérance est inférieure à 2 ‰.

Si le procédé n'est pas capable, il fabriquera, en quantité inacceptable, des pièces hors normes. Dans ce cas, avant de le mettre sous contrôle, on essaiera au préalable de l'améliorer, pour le rendre capable.

1) On suppose que la variable aléatoire  $X$  qui, à chaque ensemble poulie-pompe choisi au hasard, associe sa cote d'emmanchement en mm, suit la loi normale  $N(39,9 ; 0,05)$ .

Calculer la probabilité que cette cote soit acceptable, c'est à dire :  $P(39,7 \leq X \leq 40,1)$ .

2) Peut-on considérer le procédé de fabrication comme capable ?

## III – CARTE DE CONTROLE POUR LES MOYENNES

Pour la surveillance de la production à venir, on envisage d'établir une "carte de contrôle aux moyennes". On supposera que la valeur de  $\sigma$  reste stable mais qu'une dérive sur la valeur de  $\mu$  est à craindre. La cote de l'emmanchement pompe-poulie étant un "point Sécurité-Réglementation", la norme prévoit de prélever régulièrement des échantillons de  $n = 5$  ensembles pompe-poulie, sur lesquels on calculera la moyenne des 5 cotes d'emmanchement.

### 1 – Dérive systématique d'un côté de la moyenne

On suppose ici que  $X$  suit la loi  $N(39,9 ; 0,05)$ .

On considère l'expérience aléatoire consistant à prélever avec remise un échantillon de 5 ensembles pompe-poulie dans la production et à calculer la moyenne  $\bar{x}$  des cotes d'emmanchement de cet échantillon.

Soit  $A$  l'évènement "la moyenne  $\bar{x}$  est supérieure ou égale à 39,9". Compte-tenu de la symétrie de la loi normale, on a  $P(A) = 0,5$ .

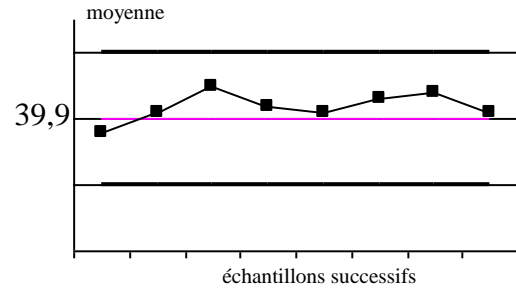
On prélève, de façon indépendante,  $N$  échantillons.

a) Montrer que la variable aléatoire  $Y$  qui, à ces  $N$  échantillons, associe le nombre de fois que l'évènement  $A$  s'est produit, suit la loi binomiale de paramètres  $N$  et  $0,5$ .

b) Calculer, en fonction de  $N$ , la probabilité que les  $N$  échantillons amènent une moyenne supérieure à  $39,9$ .

c) Déterminer le plus petit entier  $N$  tel que  $0,5^N \leq 0,01$ , puis le plus petit entier  $N$  tel que  $0,5^N \leq 0,002$ .

d) Justifier la règle selon laquelle, si apparaissent 7 échantillons successifs de moyenne supérieure à  $39,9$  (comme sur la figure), l'alerte est donnée et, dans le cas de 9 échantillons successifs au dessus de la moyenne, la production est arrêtée pour réglage.



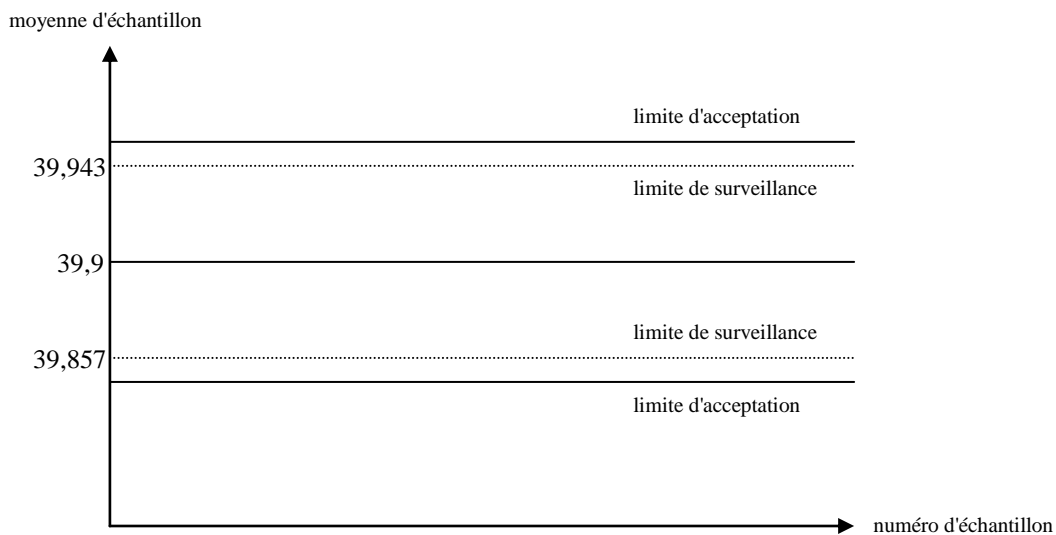
## 2 – Limites de surveillance et d'acceptation

On suppose ici que la variable aléatoire  $X$  qui, à chaque ensemble pompe-poulie prélevé au hasard dans la production, associe la cote d'emmanchement exprimée en mm, suit la loi normale  $N(39,9 ; 0,05)$ .

a) Soit  $\bar{X}$  la variable aléatoire qui, à chaque échantillon de 5 ensembles pompe-poulie prélevé au hasard et avec remise, associe la moyenne des cotes d'emmanchement de cet échantillon.

Justifier que  $\bar{X}$  suit la loi normale de moyenne  $39,9$  et d'écart type  $0,022$  à  $10^{-3}$  près.

b) Des limites de surveillance sont fixées, sur la carte de contrôle, à  $39,857$  mm et  $39,943$  mm. Calculer, à  $10^{-2}$  près,  $P(39,857 \leq \bar{X} \leq 39,943)$ .



c) Pour calculer les limites d'acceptation, dont le dépassement provoque l'arrêt de la production, calculer, à  $10^{-3}$  près, le réel  $h$  tel que  $P(39,9 - h \leq \bar{X} \leq 39,9 + h) = 0,998$ .

**Corrigé des travaux pratiques**  
**"Maîtrise statistique des procédés de production"**

**I – Normalité**

1) La fréquence cumulée manquante est  $y_3 = \frac{4}{40} = 0,1$ .

2) On a  $P(T \leq t) = 0,8 \Leftrightarrow \Pi(t) = 0,8$  ce qui donne, par lecture inverse de la table des valeurs de la fonction de répartition  $\Pi$ ,  $t \approx 0,84$ .

3) a) Les points de coordonnées  $(x_i, t_i)$  sont pratiquement alignés.

3 b) On obtient, à l'aide de la calculatrice, une équation de la droite de régression de  $t$  en  $x$  selon la méthode des moindres carrés :  $t = ax + b$  avec  $a \approx 17,521$  et  $b \approx -699,164$ .

3 c) Le coefficient de corrélation linéaire est  $r \approx 0,96$  qui est très proche de 1. Ce qui va dans le sens d'un ajustement par une loi normale.

3 d) On alors  $\sigma = 1/a \approx 0,057$  puis  $\mu = -b/a \approx 39,904$ .

**II – Capabilité**

1) On trouve que  $P(39,7 \leq X \leq 40,1) \approx 1$  (la calculatrice donne 0,99994).

Remarque : la condition d'une probabilité de 2‰ hors tolérance correspond à un écart à la moyenne de l'ordre de  $3\sigma$  pour une loi normale.

2) La probabilité précédente étant supérieure à 0,9973, on peut donc considérer le processus comme capable.

**III – Carte de contrôle pour la moyenne**

**1) Dérive systématique d'un côté de la moyenne**

a) On a la répétition de  $N$  expériences aléatoires indépendantes, avec deux issues possibles ( $A$  se réalise, avec la probabilité 0,5, ou non), où  $Y$  associe à ces  $N$  expériences, le nombre de fois que  $A$  s'est réalisé. Donc  $Y$  suit la loi  $B(N; 0,5)$ .

b) On a  $P(Y = N) = 0,5^N$ .

c) On a  $0,5^N \leq 0,01 \Leftrightarrow N \geq \ln(0,01) / \ln(0,5)$  soit  $N \geq 6,64$ . Le plus petit  $N$  est donc 7.

On a  $0,5^N \leq 0,002 \Leftrightarrow N \geq \ln(0,002) / \ln(0,5)$  soit  $N \geq 8,96$ . Le plus petit  $N$  est donc 9.

d) Il y a moins de 1% de chances que le processus soit sous contrôle (avec  $\mu = 39,9$ ) lorsqu'on observe 7 points consécutifs au-dessus de 39,9. La probabilité tombe à moins de 2‰ dans le cas de 9 points consécutifs au-dessus de 39,9. Il y a donc lieu, dans ce dernier cas, d'arrêter la production, pour contrôle.

On a, bien sûr, des règles analogues, pour l'observation de points consécutifs en-dessous de 39,9.

**2) Limites de surveillance et d'acceptation**

a) On sait, d'après le cours sur l'échantillonnage, que si  $X$  suit la loi  $N(39,9; 0,05)$  alors  $\bar{X}$  suit la loi normale  $N(39,9; \frac{0,05}{\sqrt{5}})$ , soit un écart type d'environ 0,022 à  $10^{-3}$  près.

b) On a  $P(39,857 \leq \bar{X} \leq 39,943) \approx 0,95$  à  $10^{-2}$  près.

c) On cherche  $h$  tel que  $2\Pi(h/0,022) - 1 \approx 0,998$  d'où  $h \approx 0,063$  à  $10^{-3}$  près.

## ANNEXE 6 – Loi de Poisson et temps d'attente

Forme	Niveau	Prérequis	Durée
<b>Devoir</b>	<b>B.T.S.</b> (deuxième année)	Loi normale Loi de Poisson Loi exponentielle Test d'hypothèse	1h

Ce devoir d'une heure a été proposé en B.T.S. industriel deuxième année. Il s'agit d'un bilan sur les lois de probabilité et les tests. Il s'agissait de proposer une situation concrète issue de données fournies par le secteur industriel (ici, bâtiment et travaux publics).

On trouvera dans les pages suivantes :

- L'énoncé du devoir.
- Des éléments de correction.



**DEVOIR**  
**Loi de Poisson et temps d'attente**

Lors de la construction d'un centre commercial, le terrassement nécessite l'extraction et l'évacuation d'une grande quantité de matériaux. L'entreprise responsable du chantier dispose, pour ce faire, d'une pelle sur chenilles et de camions bennes. L'extraction ne peut avoir lieu que si un camion est présent pour recevoir les matériaux. L'étude suivante est menée pour juger du rendement de l'extraction.

**Les parties A, B et C peuvent être traitées de façon indépendante.**  
**Dans ce qui suit, tous les résultats approchés seront donnés à  $10^{-2}$  près.**

*A . Rendement de la pelle*

La variable aléatoire  $X$  qui, à une heure choisie au hasard pendant la durée des travaux de terrassement, associe le nombre de  $m^3$  extraits par la pelle suit une loi normale de moyenne  $\mu$  et d'écart type  $\sigma = 10$ .

1° Compte tenu de la nature du matériau à extraire, on a calculé, à l'aide la fiche technique de la pelle, un rendement "théorique", pour celle-ci, de  $120 m^3/h$ .

On suppose, dans cette question, que  $\mu = 120$ .

Déterminer alors la probabilité d'extraire, pendant une heure, moins de  $100 m^3$ .

2° Dans cette question, la valeur de  $\mu$  est inconnue.

Pour contrôler le rendement "théorique" calculé, on se propose de construire un test d'hypothèse bilatéral. Pour cela, on fera fonctionner  $n = 16$  fois la pelle pendant une heure.

On admet que ces 16 épreuves sont indépendantes.

On désigne par  $\bar{X}$  la variable aléatoire qui, à chaque échantillon de 16 heures prises au hasard pour des fonctionnements indépendants d'une heure, associe le rendement moyen observé, en  $m^3/h$ . On admet que  $\bar{X}$  suit une loi normale.

L'hypothèse nulle est  $H_0 : \mu = 120$ . L'hypothèse alternative est  $H_1 : \mu \neq 120$ .

Le seuil de signification du test est fixé à 0,05.

a) Justifier que, sous l'hypothèse nulle  $H_0$ ,  $\bar{X}$  a pour moyenne 120 et comme écart type 2,5.

b) Sous l'hypothèse nulle  $H_0$ , déterminer le nombre réel positif  $h$  tel que :

$$P(|\bar{X} - 120| > h) = 0,05 \text{ c'est à dire } P(120 - h \leq \bar{X} \leq 120 + h) = 0,95.$$

c) Enoncer la règle de décision permettant d'utiliser ce test.

d) Après 16 essais indépendants, d'une heure chacun, la pelle a extrait, au total,  $1896 m^3$ . Peut-on, au seuil de 5 %, accepter l'hypothèse  $\mu = 120$  ?

*B . Temps d'attente des arrivées des camions*

1° Compte tenu du nombre de camions dont on dispose, et de la durée aléatoire (liée au trafic) du trajet qu'ils doivent faire pour déposer en décharge les matériaux extraits, on peut considérer que la variable aléatoire  $Y$  qui, à toute durée d'une heure travaillée, associe le nombre de camions entrant sur le chantier suit la loi de Poisson de paramètre 5.

Quelle est la probabilité que se présentent, pendant une heure, au plus 4 camions sur le chantier ?

2° Pour  $t$  réel positif fixé, on note  $Y_t$  la variable aléatoire qui, à tout intervalle de temps de durée  $t$  (en heures), associe le nombre de camions entrant sur le chantier.

On admet que  $Y_t$  suit la loi de Poisson de paramètre  $\lambda = 5t$ .

a) Exprimer, en fonction de  $t$ ,  $P(Y_t = 0)$ .

b) Quelle est la probabilité que, pendant 10 minutes, n'arrive aucun camion ?

3° On désigne par  $T$  la variable aléatoire correspondant au temps d'attente (en heures) entre les arrivées consécutives de deux camions sur le chantier.

a) Soit  $t > 0$ . La probabilité d'un temps d'attente supérieur à  $t$  est égale à la probabilité qu'aucun camion ne se présente pendant un intervalle de temps de durée  $t$ .

Montrer, en utilisant la question B.2.a), que, pour tout  $t > 0$ ,  $P(T > t) = e^{-5t}$ .

b) On déduit de la question précédente que  $T$  suit la loi exponentielle de paramètre 5.

Quelle est l'espérance du temps d'attente d'un nouveau camion (donner le résultat en minutes) ?

### *C. Rendement prévisible de l'extraction*

1° Le matériau à extraire a une masse volumique de 1,6 tonne par  $m^3$ , et la pelle un rendement moyen de  $120 m^3/h$ . Chaque camion benne a une charge utile de 26 t.

Calculer la durée moyenne (à une minute près) de chargement d'un camion.

2° S'il n'y a pas de camion sous la pelle, l'extraction cesse, et le rendement baisse. La simulation des arrivées et chargements des camions, selon les lois de probabilité décrites dans les parties A et B, permet de constater que pendant environ 32 % du temps, aucun camion n'est présent sur le chantier.

Quel est le rendement effectif moyen (en  $m^3/h$ ) prévisible de l'extraction ?

**ELEMENTS DE SOLUTION****EXERCICE 1 :**

- A.1)  $P(X < 100) \approx \mathbf{0,02}$  à  $10^{-2}$  près.
- A.2) a)  $\bar{X}$  suit la loi  $N(120, \frac{10}{\sqrt{16}})$ .
- b)  $2\pi(\frac{h}{2,5}) - 1 = 0,95 \Rightarrow \frac{h}{2,5} = 1,96 \Rightarrow h = \mathbf{4,9}$ .
- c) Soit  $\bar{x}$  le rendement moyen observé sur un échantillon de taille 16.  
 $H_0$  est acceptée au seuil de 5% lorsque  $\bar{x} \in [\mathbf{115,1 ; 124,9}]$ .
- d)  $\frac{1896}{16} = 118,5$   $H_0$  est donc **acceptée** au seuil de 5%.
- B.1)  $P(Y \leq 4) \approx \mathbf{0,44}$  à  $10^{-2}$  près.
- B.2) a)  $P(Y_t = 0) = e^{-5t}$ .
- b)  $P(Y_{1/6} = 0) = e^{-5/6} \approx \mathbf{0,43}$  à  $10^{-2}$  près.
- B.3) a)  $P(T > t) = P(Y_t = 0)$ .
- b)  $E(T) = \frac{1}{5}$  h = **12 mn**.
- C.1) 26 t correspondent à  $\frac{26}{1,6} = 16,25 \text{ m}^3$  de matériaux,  
ce qui nécessite, en moyenne,  $\frac{16,25}{120} \times 60 \approx \mathbf{8 \text{ mn}}$ .
- C.2) Rendement moyen prévisible  $120 \times 0,68 = \mathbf{81,6 \text{ m}^3/\text{h}}$ .

# BIBLIOGRAPHIE

**AVENEL Michèle** – *"DECF – Mathématiques appliquées"* – FOUCHER.

**BOULEAU Nicolas** – *"Probabilités de l'ingénieur : variables aléatoires et simulation"* – Hermann 1986.

**BRIAN Eric** – *"La mesure de l'Etat – Administrateurs et géomètres au XVIII<sup>e</sup> siècle"* – Albin Michel 1994.

**BRY Xavier** – *"Analyses factorielles simples"* – ECONOMICA 1995.  
Présentation intuitive de la technique statistique avec, en encarté, le traitement mathématique.

**CHAITIN Gregory** – *"Les suites aléatoires"* – Dossier *"Pour la science"* : *"Le hasard"* – Hors série avril 96.

## **Commission Inter-IREM Lycées technologiques**

- *"A propos de fiabilité"* – IREM Paris-Nord – Brochure n°48.
- *"Les plans d'expérience pour le BTS chimiste"* – IREM Paris-Nord – Brochure n°88.
- *"Simulations d'expériences aléatoires – Une expérimentation du hasard de la première au BTS"* – IREM Paris-Nord – Brochure n°93 – 1998.
- *"Simulation et statistique en seconde"* – IREM Paris-Nord – Brochure n°102 – 2000.

## **Commission Inter-IREM Statistique et probabilités**

- *"Enseigner les probabilités au lycée"* – 1997.

**CREPEL Pierre** – *"La naissance des mathématiques sociales"* – et **LE BRAS Hervé** – *"L'invention des concepts en démographie"* – Dossier *"Pour la science"* : *"Les mathématiques sociales"* – Hors série juillet 1999.

**DELAHAYE Jean-Paul** – *"Aléas du hasard informatique"* – *Pour la science* mars 98.

**DESROSIERES Alain** – *"La politique des grands nombres – Histoire de la raison statistique"* – La découverte / Poche 2000.

**DEWDNEY Alexander** – *"Les hasards simulés"* – Dossier *"Pour la science"* : *"Le hasard"* – Hors série avril 96.

**DODGE Yadolah** – *"Statistique – Dictionnaire encyclopédique"* – Dunod 1993.

**DROESBEKE Jean-Jacques** et **TASSI Philippe** – *"Histoire de la statistique"* – Que sais-je ? n° 2527 – P.U.F. 1997.

**Encyclopédie Universalis**

... / ...

**FOUCART Thierry** – *"L'analyse des données, mode d'emploi"* – PRESSES UNIVERSITAIRES DE RENNES 1997.

**IREM de Clermont-Ferrand** - *"Une application industrielle des statistiques : la carte de contrôle"* – Mars 2001.

**LE BRAS Hervé** – *"Naissance de la mortalité – L'origine politique de la statistique et de la démographie"* – Seuil/Gallimard 2000.

**SAPORTA Gilbert** – *"Probabilités, analyse des données et statistique"* – TECHNIP 1990.

**WONNACOTT T.H.** et **WONNACOTT R.J.** – *"Statistique"* – Economica 1995.

**Sur INTERNET :**

De nombreuses informations sont accessibles sur Internet, cours de statistique, histoire, et surtout données statistiques récentes, souvent directement importables dans Excel (et utilisables en citant sa source). N'hésitez pas à utiliser les moteurs de recherche...

Pour l'histoire, nous avons, en particulier, consulté :  
[www-groups.dcs.st-and.ac.uk/~history/BiogIndex.html](http://www-groups.dcs.st-and.ac.uk/~history/BiogIndex.html)  
[www.math.uah.edu/stat/biographies/](http://www.math.uah.edu/stat/biographies/)  
[www.mrs.umm.edu/~sungurea/introstat/history/](http://www.mrs.umm.edu/~sungurea/introstat/history/)  
[www.weibullnews.com/ybullbio.htm](http://www.weibullnews.com/ybullbio.htm)

Nous avons utilisé des données provenant de :  
[www.insee.fr](http://www.insee.fr)  
[www.sofres.com](http://www.sofres.com)  
[www.travail.gouv.fr](http://www.travail.gouv.fr)