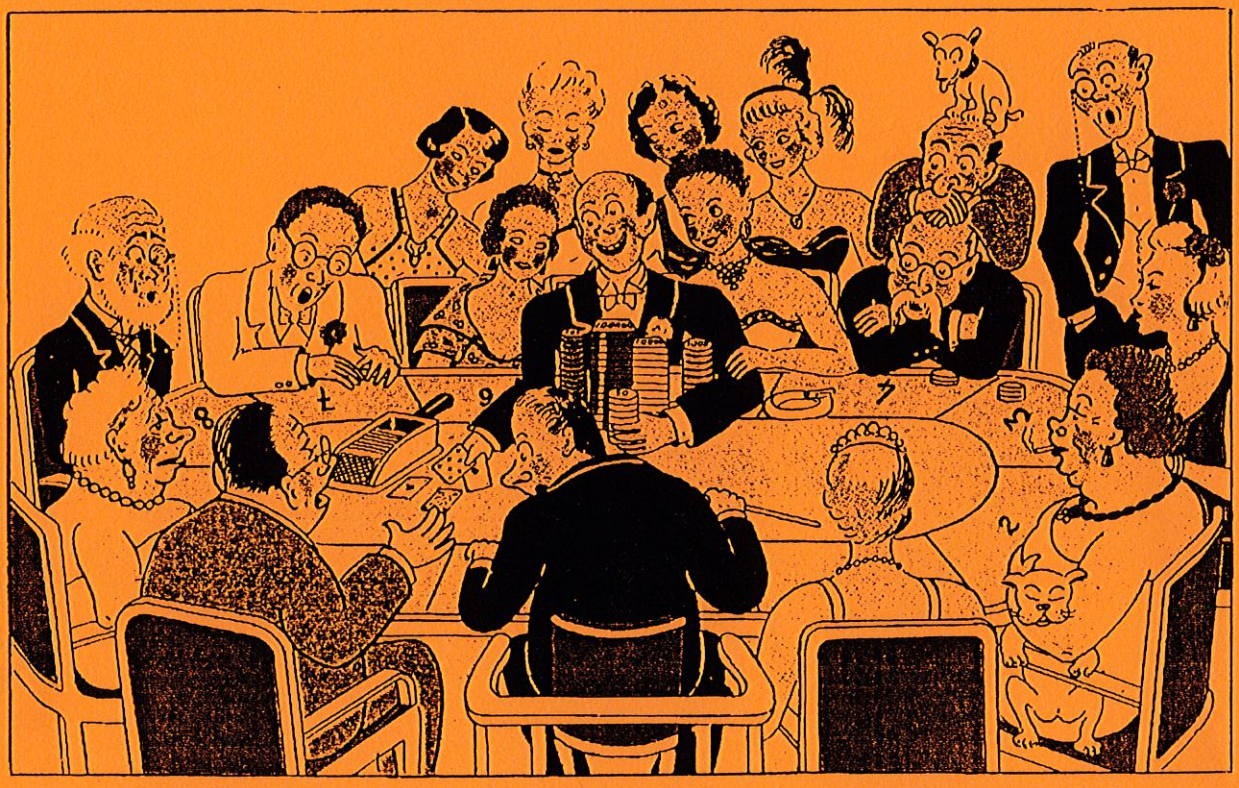
Publication n° 124 de la commission Inter-IREM Lycées techniques

LE NOUVEAU PROGRAMME DE STATISTIQUE ET PROBABILITES AU LYCEE



Carnet de stage

2000 : SECONDE
2001 : 1ère S – 1ère ES
2002 : T.S – T.ES

IREM PARIS-NORD

Institut Galilée
99, av. J.B. Clément
93430 VILLETANEUSE

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

LE NOUVEAU PROGRAMME DE STATISTIQUE ET PROBABILITES AU LYCEE

Carnet de stage

246 pages
Villetaneuse 2003

ISBN 2 86240 117 1

Dépôt légal : 4ème trimestre 2003

12 €

AVANT-PROPOS

La statistique inférentielle, qui est au centre des nouveautés introduites, en matière de statistique, dans les nouveaux programmes de lycée depuis la seconde, est au programme de la plupart des BTS tertiaires et industriels depuis une quinzaine d'années. Ce thème, absent de la formation initiale de la plupart des enseignants de mathématique, a nécessité de gros efforts de formation continue.

Les formateurs étant peu nombreux, notre commission inter-IREM a organisé de nombreuses journées d'information dans les académies. Les stages que notre commission a organisés sur ce thème à Paris pour les professeurs de mathématiques en BTS ont été ouverts aux professeurs des académies voisines (Amiens, Caen, Dijon, Orléans, Rouen). Bien que portant sur de "nouveaux programmes de 1988", ces stages n'ont pas désempli jusqu'en 2000... Les épreuves d'examen de BTS n'ont intégré que très progressivement ces notions, pour laisser aux enseignants le temps de les approfondir. Malgré cela il n'était pas rare, il y a peu, de trouver dans des sujets de BTS des questions où la distinction entre intervalle de confiance et test n'était pas évidente, ou entre paramètres à tester et observations... Quant à l'accueil réservé au premier exercice du bac ES Métropole 2003, il ne fait que confirmer la difficulté d'introduction et de maîtrise de ces notions.

Je m'occupe de la formation continue depuis 1977 et la statistique inférentielle est le seul thème de formation qui ait nécessité durant cette période un dispositif aussi étalé dans le temps.

Bien entendu l'outil informatique a permis de développer récemment les activités de simulation.

Notre réflexion sur l'enseignement de la statistique inférentielle en BTS depuis une quinzaine d'années nous a été très précieuse pour aborder les nouveaux programmes apparus en seconde à la rentrée 2000. Nous étions, par exemple, un des rares groupes IREM (pour ne pas dire le seul) a avoir travaillé et publié sur la simulation, à la demande notamment de Monsieur Jean-Louis Piednoir IGEN, dans le cadre de travaux commandés par la direction de l'enseignement scolaire du Ministère.

La pénurie de formateurs sur ce thème s'étendant aux séries générales, tout naturellement, nous avons mis en place des stages de statistique et probabilités sur les nouveaux programmes pour les classes de seconde, première et terminale S et ES, à Paris, pour les académies d'Ile de France. Philippe Dutarte, en s'appuyant notamment sur les travaux de la commission inter-IREM lycées techniques, anime ces formations, ainsi que des journées d'information sur l'usage des TICE en statistique et probabilités dans les académies, à la demande des IA-IPR.

La brochure suivante est le document papier remis aux participants à ces stages et séances d'information au cours desquels Philippe Dutarte présente différentes activités sur calculatrices ou ordinateurs.

Nous espérons que ce "carnet de stage" rendra service aux collègues et les remercions par avance de leurs remarques et suggestions.

Bernard VERLANT
Responsable de la Commission Inter-IREM
"Lycées technologiques"

Cette brochure a été réalisée par :

Philippe DUTARTE
Lycée Branly – CRETEIL
dutarte@club-internet.fr

Avec la participation de :

Christian KERN	Lycée Alain – ALENÇON
Marie-France NOUGUES	Faculté des Sciences de Montpellier
Christine DHERS	Lycée Newton ENREA – Clichy la Garenne
Dominique ARBRE	Lycée Paul Constans – Montluçon
Françoise DELZONGLE	Lycée Eiffel – Cachan
Isabelle BRUN	Lycée Pagnol – Athis-Mons
Loïc MAZO	Lycée Follereau - Belfort

de la
Commission Inter-IREM "Lycées techniques"

Réalisation :
Bernard VERLANT, responsable de la C.I.I. L.P.- L.T.

SOMMAIRE

Séance 1

Séance 2

Séance 3

Séance 4

Séance 5

Séance 1 : L'ARTICULATION STATISTIQUE / PROBABILITES

I – LES ENJEUX DE LA STATISTIQUE

"Pour comprendre les pensées de Dieu, il faut étudier les statistiques, car elles constituent la mesure de ses desseins."

Florence Nightingale[1] (1820-1910).

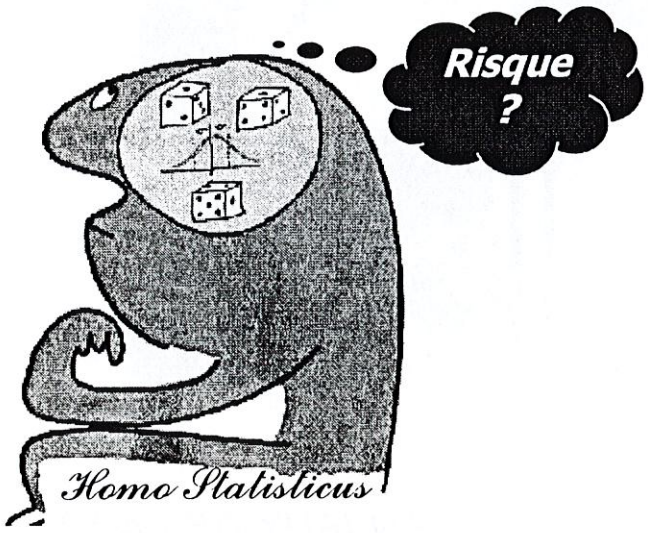
[1] Traduit du livre de D. Salsburg *"The lady tasting tea"* : *"To understand God's thoughts, we must study statistics, for these are the measure of his purpose."*

L'objectif de ce paragraphe d'introduction est de montrer l'importance de la méthode statistique.
Florence Nightingale prétendait au XIX^{e} siècle, que c'était là le plus sûr moyen de *"comprendre les pensées de Dieu"*. Dans un langage plus actuel (pour ne pas dire "à la mode"), nous dirions que la méthode statistique est un outil puissant de modélisation de la réalité. *Nightingale* ("l'ange de la statistique"), est très célèbre outre Manche. Par ses études statistiques, cette infirmière permit l'amélioration des conditions sanitaires dans les hôpitaux militaires britanniques. Elle utilisa les statistiques, développant les méthodes graphiques, tant pour étudier les causes de mortalité, l'effet des améliorations apportées, que pour convaincre du besoin de réforme sanitaire (à certaines périodes de la guerre de Crimée, on mourrait 7 fois plus dans les hôpitaux de campagne, que sur le champs de bataille, à cause d'épidémies).

Les différents textes officiels motivent à plusieurs reprise l'importance nouvelle donnée à l'enseignement de la statistique :

"L'usage de la statistique dans de nombreux domaines ***ne relève pas d'une mode passagère*** *mais de la diffusion d'une culture et d'un mode de pensée anciens, diffusion rendue possible par les progrès simultanés de la théorie mathématique et de la technologie informatique." (Programme 2001 de 1ère S.)*

"Former les élèves en statistique, c'est leur donner les moyens de développer une forme de pensée critique sans laquelle ils seront exclus du ***débat social et scientifique.***"
(Projet de programme de TS.)



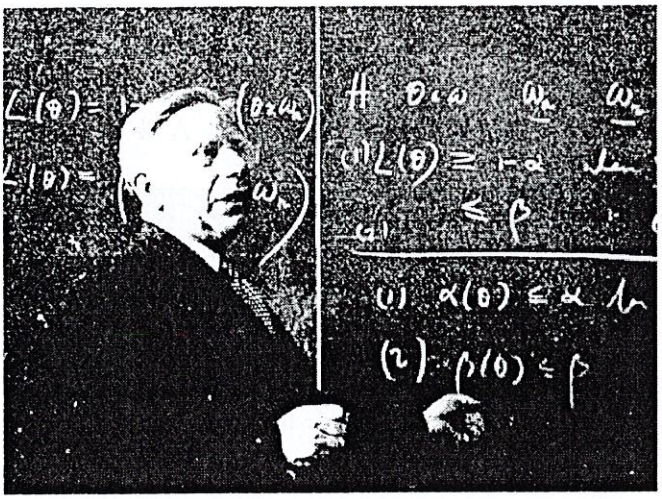
Lorsque l'on évoque la statistique, le plus souvent la méfiance s'installe. L'utilisateur de la statistique est souvent vu comme un manipulateur déguisant la réalité pour présenter ses préjugés comme une vérité objective d'où l'adage : "on fait dire ce que l'on veut à la statistique" ou cette citation d'un homme célèbre : "la statistique est la forme élaborée du mensonge", voir également cette couverture récente du magazine "Tangente".
Comme l'enseignement de la statistique, jusqu'à une date récente, était quasi confidentiel ou bien réduit à quelques recettes dans des

enseignements supérieurs spécialisés (économie, sciences humaines, biologie), on pensera souvent à quelques applications vulgarisées par la presse : sondages d'opinion, estimation des résultats électoraux à 20h00. 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**. Parmi les plus grands "consommateurs" de statistique, citons :

- les compagnies d'assurance,
- les services marketing,
- les sociétés de sondage,
- les laboratoires pharmaceutiques,
- les services de contrôle de qualité et fiabilité des produits industriels,
- les sociétés de placement, de finance.

...

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 déclassées par le gouvernement américain en 1947. *Wald* expose alors la technique des tests statistiques séquentiels.

Quelques exemples...

L'objectif visé dans les quelques exemples qui suivent est de montrer quelle est l'originalité de la démarche statistique (ce ne sont pas des mathématiques ordinaires bien

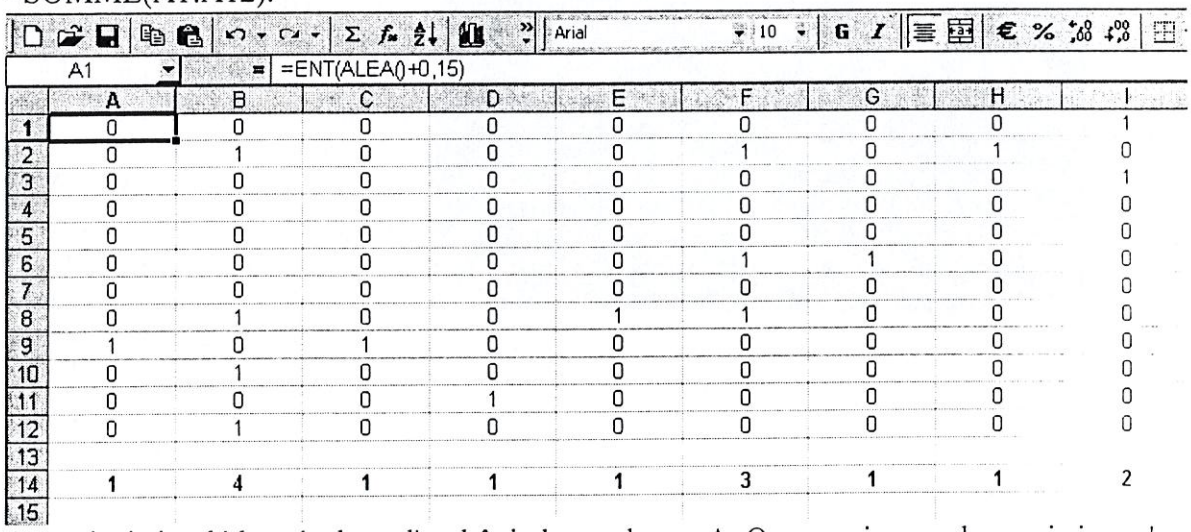
qu'il y ait beaucoup de mathématiques), quel est l'intérêt des techniques utilisées mais aussi leurs limites. Pour les décrire il nous faudra faire appel à des structures mathématiques diverses, certaines très simples comme la proportion, d'autres plus complexes comme la géométrie euclidienne et surtout le calcul des probabilités. 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.

Exemple 1 (Santé)

Dans le classement *"Science et avenir"* des hôpitaux de France, on lit, pour deux hôpitaux de province et pour une intervention analogue en chirurgie digestive :
Hôpital *A* (Briançon) : 1 décès sur 12 interventions au total.
Hôpital *B* (Méru, dans l'Oise) : 3 décès sur 12 interventions au total.
Faut-il suspecter l'hôpital *B* ? Le ministère de la santé devrait-il diligenter une enquête ? Est-il vraiment raisonnable d'établir un classement sur un "*échantillon*" de taille 12 ? La différence observée est-elle *significative* ?
Un raisonnement simple peut nous éclairer. Supposons que la probabilité p de décès dans ce type d'intervention soit $\boldsymbol{p = 0,15}$. Dans ces conditions, la variable aléatoire F qui, à tout échantillon de 12 interventions indépendantes, associe la fréquence des décès sur cet échantillon, est telle que $12F$ suit la loi binomiale $\mathscr{B}(12 ; 0,15)$.

La variabilité de F correspond alors à un écart type $\sigma = \frac{1}{12}\sqrt{12 \times 0,15 \times 0,85} \approx 0,10$.

Dans ces conditions les résultats des hôpitaux *A* ($f_A \approx 0,08$) et *B* ($f_B = 0,25$) sont dans les limites définies par cet "écart type". On ne peut donc pas considérer a priori que les conditions d'opération y soient différentes.
On peut simuler cette situation par l'instruction ***int(rand + 0,15)*** sur calculatrice, ou ***ENT(ALEA()+0,15)*** sur le tableur Excel. Cette instruction fournit la valeur 1 (décès) dans 15 % des cas et la valeur 0 sinon. On recopie ensuite l'instruction, avec la souris, jusqu'en A12 pour simuler 12 interventions. La somme est effectuée en A14 par l'instruction =SOMME(A1:A12).



	A	B	C	D	E	F	G	H	I
1	0	0	0	0	0	0	0	0	1
2	0	1	0	0	0	1	0	1	0
3	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	1	0	0
7	0	0	0	0	0	0	0	0	0
8	0	1	0	0	1	1	0	0	0
9	1	0	1	0	0	0	0	0	0
10	0	1	0	0	0	0	0	0	0
11	0	0	0	1	0	0	0	0	0
12	0	1	0	0	0	0	0	0	0
13									
14	1	4	1	1	1	3	1	1	2
15									

On a ainsi simulé les résultats d'un hôpital en colonne A. On recopie avec la souris jusqu'en colonne I pour avoir les résultats d'autres hôpitaux. Les écarts observés ne sont dus qu'au hasard.
En réalité la procédure statistique adaptée (qui ne suppose pas d'attribuer une valeur à p) est celle d'un test d'hypothèse sur la différence des fréquences.

Exemple 2 (Sondage d'opinion)

Lecture du "*baromètre présidentiel*" de la SOFRES (sondage effectué sur un échantillon de 1000 personnes).

	Mai 2001	Juin 2001
Jacques Chirac	0,50	0,49
Lionel Jospin	0,50	0,51

Dans un tout autre contexte, la situation est analogue à l'exemple 1.
Si on lance une pièce à pile ou face ($p = 0,5$), la variabilité de la fréquence des "pile" observée sur des échantillons aléatoires de taille 1000 correspond à un écart type $\sigma = \sqrt{\frac{0,5 \times 0,5}{1000}} \approx 0,016$. Il n'y a donc pas grand chose à déduire des sondages précédents, si ce n'est que les deux candidats sont, à l'époque du sondage (et dans l'hypothèse d'un second tour Chirac/Jospin), au coude à coude.
L'exemple précédent pose de façon plus générale le problème des sondages (électoraux) et de la confiance que l'on peut leur accorder. Lors du premier tour des élections présidentielles de 2002, le dernier sondage publié par l'institut B.V.A. , effectué sur 1000 électeurs le vendredi 19/04/02, prévoyait :

Jacques Chirac	19 %
Lionel Jospin	18 %
Jean-Marie Le Pen	14 %

La stupéfaction a été grande le dimanche 21/04/02 au vu des résultats :

Jacques Chirac	19,88 %
Lionel Jospin	16,18 %
Jean-Marie Le Pen	16,86 %

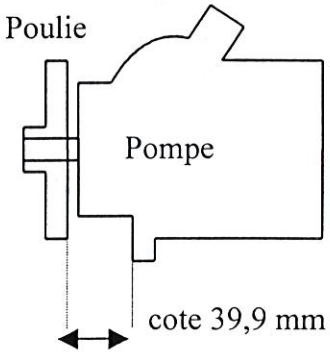
Doit-on considérer que le sondage était "faux" ? S'il s'agissait de prévoir l'ordre des trois candidats, la réponse est affirmative, mais c'est prêter à la méthode statistique plus de pouvoir qu'elle n'en possède. Il convient, en revanche, tenant compte de la variabilité inhérente à la méthode du sondage, de définir sous forme de "fourchettes" la précision que l'on est en "droit" d'attendre du sondage.

Exemple 3 (Relations commerciales)

Un commerçant d'articles de chasse reçoit de son fournisseur un lot important de cartouches. Le fournisseur affirme que celui-ci contient une proportion p de cartouches défectueuses inférieure à 0,10.
La seule façon de contrôler ces cartouches est de les tirer. On comprend que le commerçant ne peut contrôler tout le lot (quand le contrôle n'est pas, comme ici, destructif, il est souvent cependant onéreux, ce qui explique qu'un contrôle exhaustif de la qualité est généralement exclu).
Le commerçant prélève un échantillon aléatoire de 100 cartouches et les tire. Il en trouve 14 défectueuses. Doit-il renvoyer tout le lot à son fournisseur ? Ce serait compter sans l'étude de la variabilité inhérente au tirage aléatoire. Pour la loi binomiale $\mathscr{B}(100 ; 0,10)$ on a en effet $P(X \geq 14) \approx 0,22$ (ce qui n'est pas rare).
Dans cette situation, commerçant et fournisseur devront, avant la livraison, s'accorder sur la procédure statistique conduisant à l'acceptation ou au refus d'un lot.

Exemple 4 (Industrie)

La "révolution statistique" a joué un rôle déterminant dans la qualité de la production industrielle. Dans les années 1970, les voitures japonaises étaient de meilleure qualité que les automobiles américaines, pour un prix inférieur. En 1980 la chaîne de télévision NBC diffuse un documentaire intitulé "*If Japan Can, Why Can't We* ?" On y décrivait comment *W. Edwards Deming* (1900 – 1993) avait profondément influencé les industriels japonais. Après guerre, "made in Japan" était synonyme d'imitation bon marché, de piètre qualité. Lors d'une conférence effectuée là-bas à cette époque, *Deming* surprit son auditoire en affirmant qu'il était possible, en suivant des méthodes statistiques appropriées de contrôle de qualité, de produire des objets de haute qualité à bas coûts, qui leur permettrait de dominer le marché. Il prédit alors que ceci serait réalisable en cinq années environ. *Deming* affirma par la suite s'être trompé dans sa prédiction. Les japonais la réalisèrent en à peine deux ans[2].



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.

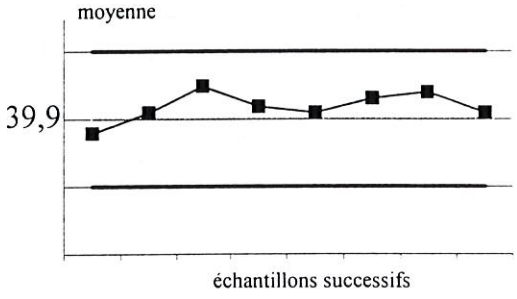
L'exemple étudié, issu de l'industrie automobile, est une presse d'emmanchement de poulie sur une pompe de direction assistée. Les performances de la presse sont variables, cette variabilité ayant de nombreuses causes possibles : main-d'œuvre, matériel, matière première, environnement de l'atelier, méthodes 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.

Pour la surveillance de la production à venir, on établit des cartes de contrôle "aux moyennes" (on suppose qu'une dérive sur la moyenne 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.

En supposant que ces moyennes se répartissent selon une loi normale (symétrique) de moyenne 39,9 , on a, à chaque échantillon, une chance sur deux d'être au dessus ou en dessous de 39,9.

Les industriels adoptent alors la règle selon laquelle, si apparaissent 7 échantillons successifs de moyenne supérieure à 39,9 (comme sur la figure suivante), l'alerte est donnée, car la probabilité d'une telle observation, à un instant donné, s'il n'y a pas dérive, est $0{,}5^7 \approx 0{,}008$.



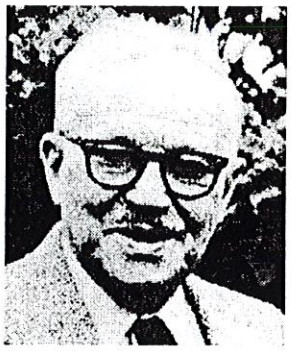
Cependant, sur 200 échantillons consécutifs par exemple, la probabilité d'observer au moins une fois, sans dérive, cette configuration est importante (voir page 8). Il y aura donc de fausses alertes.

Dans le cas de 9 échantillons successifs au dessus de la moyenne, la production est arrêtée pour réglage éventuel (car $0{,}5^9 \approx 0{,}002$).

---

[2] D'après *Salsburg – "The Lady tasting tea".*

Exemple 5 (Agronomie)



Dans un article de 1967 (*"Experimentation with weather control"*), *Jerzy Neyman* montre l'importance du choix de la méthode statistique employée dans les expérimentations.
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.

Tableau 1

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 %

Pour établir ces statistiques, une *estimation* en l'absence d'ensemencement des nuages est obtenue à partir de valeurs correspondant aux précipitations enregistrées, sur les différents sites, les années précédant l'expérience d'ensemencement.
Neyman souligne d'emblée l'absence de *randomisation* (choix aléatoire des jours où l'on pratiquera l'ensemencement ou non, analogue à l'utilisation d'un placebo en médecine) dont l'intérêt, pour écarter les biais, était pourtant depuis longtemps souligné par *Fischer* (le hasard est souvent le meilleur allié du statisticien).
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 ensemençait les nuages à l'iodure d'argent ou non, de façon à comparer ensuite les précipitations obtenues, avec ou sans ensemencement.

Tableau 2

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

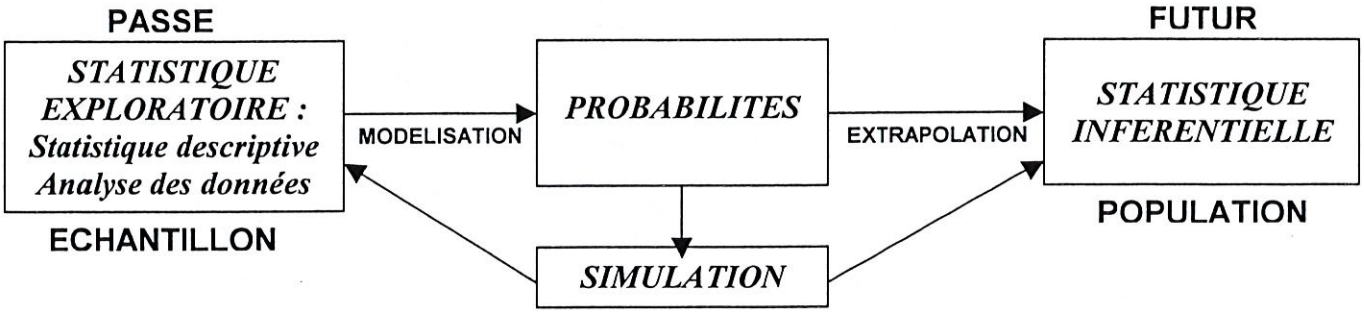
Australie 3	1957	6	+4
Australie 4	1958	6	+4
Australie 5	1959	4	-5
Australie 6	1960	1,5	-13
Mexique	1956	10	+20 ; -8
Suisse	1957	7	+32,9 ; +0 ; +78,8 ; -16,2
Québec	1959	4	-2
Japon	1960	0,25	+48
France	1961	1,5	-6,6
Israël	1961	4,5	+6,4 ; +24,8

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. Cette méthode conduit, à partir des données du second tableau, à penser que l'ensemencement des nuages n'est pas statistiquement plus efficace que la danse de la pluie des indiens d'Amérique.

L'un des biais de la méthode d'estimation des industriels (tableau 1) était la grande dépendance de leur méthode 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 décade en décade.

DES DEFINITIONS ...

Le schéma suivant présente les liens qu'entretiennent les notions de statistique, probabilités et simulation (et sur lesquels nous reviendrons), termes qui sont définis ensuite.



STATISTIQUE EXPLORATOIRE

Elle repose sur l'***observation*** de phénomènes (donc ***passés***). Son but est de *résumer*, *structurer* et *représenter* (***statistique descriptive***) l'information contenue dans les données. L'***analyse des données*** regroupe les techniques de visualisation de données multidimensionnelles (analyse en composantes principales...).

PROBABILITES

Théorie mathématique abstraite ***modélisant*** des phénomènes où le "hasard" intervient.

> *"Modéliser une expérience aléatoire, c'est lui associer une loi de probabilité (qui est un objet du monde mathématique)".*
> *Programme 2001 de 1ère S.*

STATISTIQUE INFERENTIELLE (OU INDUCTIVE)

Son but est d'***étendre*** (inférer = tirer les conséquences), ***à la population*** toute entière, les propriétés constatées sur un échantillon. Il s'agit d'*estimer* un paramètre, ou de *tester une hypothèse* statistique, concernant la population étudiée. La statistique inférentielle a un aspect ***décisionnel*** et le calcul des probabilités y joue un rôle fondamental.

SIMULATION

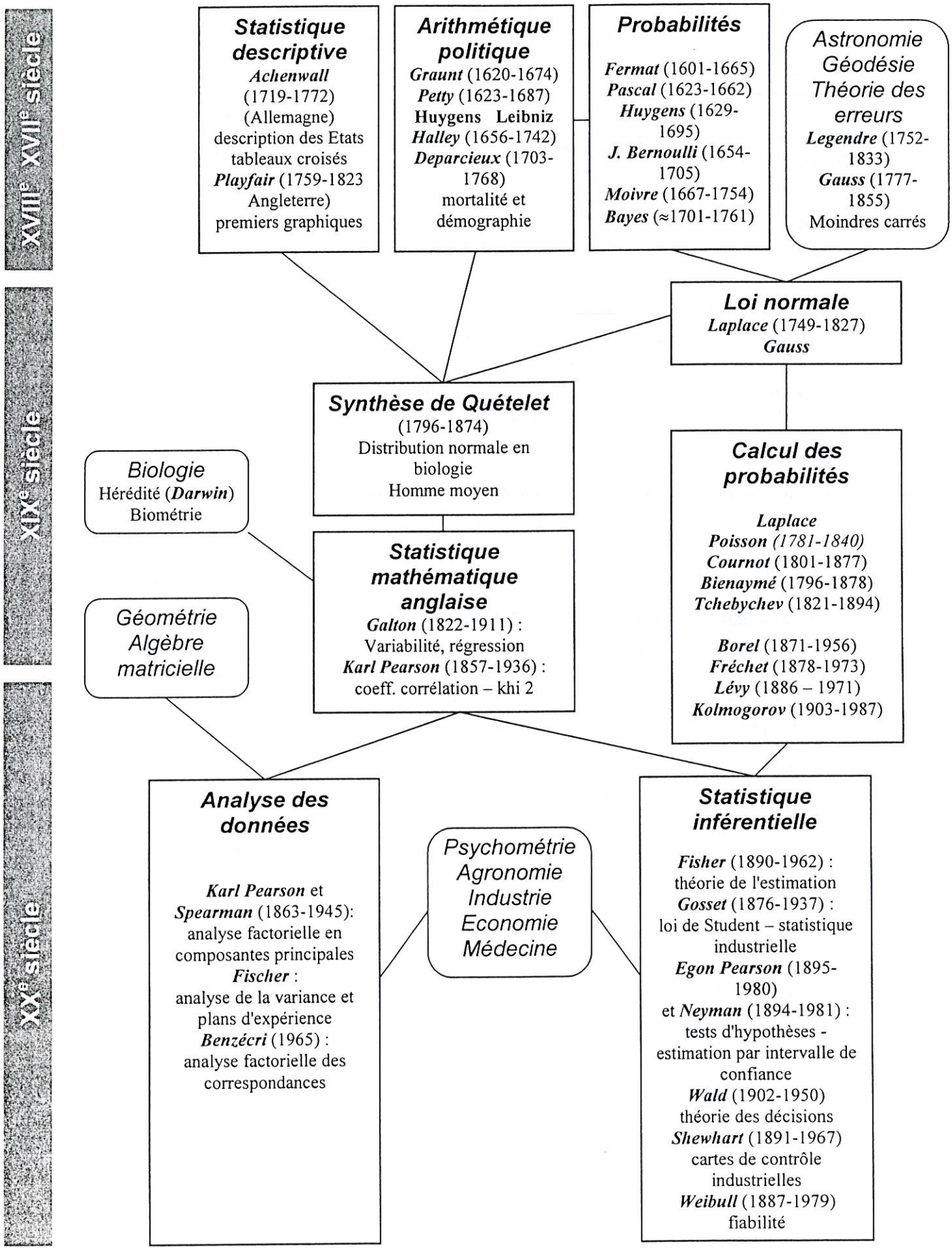
La simulation est la méthode statistique qui, en produisant des données sous un certain modèle (probabiliste), permet d'en examiner les conséquences, soit pour juger de l'adéquation du modèle à la réalité, soit pour obtenir (ou conjecturer) des résultats difficiles, ou impossibles, à calculer.

> *"La simulation permet d'une part d'avoir des estimations de résultats impossible à calculer explicitement et, d'autre part, par la comparaison de telles estimations avec des résultats expérimentaux, de valider le modèle choisi."*
> *Programme 2001 de 1ère S.*

QUELQUES REPERES HISTORIQUES

L'arbre de la page suivante schématise l'évolution et la filiation des différentes branches de la statistique, fournissant les grands noms de son histoire. Il y sera fait plusieurs fois allusion par la suite. L'histoire (et la petite histoire, car les anecdotes sont souvent frappantes et pleines d'enseignements) permettra d'éclairer certaines notions et de mieux comprendre leur rôle.

Une généalogie de la statistique du XVII^{e} au XX^{e} siècle



II – LA PLACE DE LA SIMULATION

*"La simulation joue un rôle important... et favorise l'émergence d'**un mode de pensée propre** à la statistique." (Projet de programme de TS.)*

1 – Définitions

• Une définition "académique" :
Rappelons la définition de *Yadolah Dodge* (*Statistique. Dictionnaire encyclopédique.* Dunod 1993) :
*"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."*

• **La simulation en seconde** :
L'aspect de modélisation par les probabilités n'étant abordé qu'en première, on adopte en seconde un point de vue pragmatique :

"Dans le cadre du programme de seconde, simuler une expérience consistera à produire une liste de résultats que l'on pourra assimiler à un échantillon de cette expérience". Document d'accompagnement du programme 2000 de seconde.

On peut dire que, pour nos élèves de seconde, 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 simulation en première** :
On peut replacer la simulation dans le cadre d'un ***modèle probabiliste*** que l'on explore ou dont on souhaite juger de la pertinence.

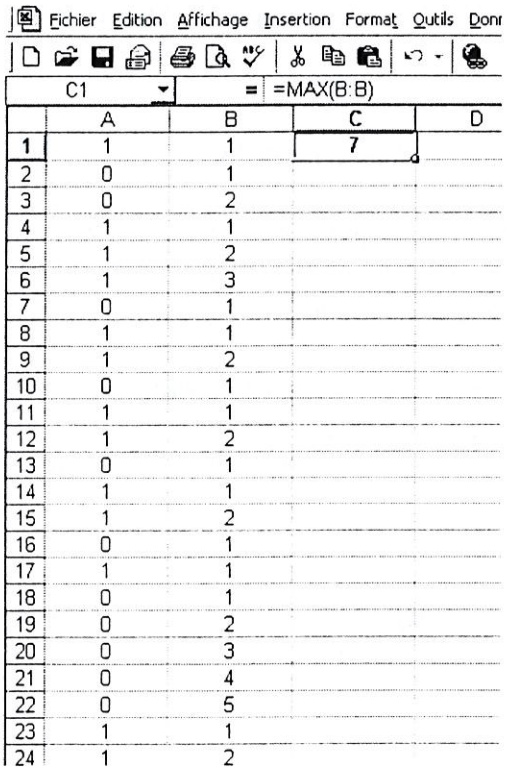
"Simuler consiste à produire des données à partir d'un modèle prédéfini. [...] Pour simuler une expérience, on associe d'abord un modèle à l'expérience en cours, puis on simule la loi du modèle." Document d'accompagnement du programme 2001 de 1ère S.

L'***intérêt pédagogique*** 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***).

2 – UN EXEMPLE EN SECONDE : Longueur maximale des suites de lancers consécutifs égaux à pile ou face

Pour illustrer les apports pédagogiques de la simulation d'expériences aléatoires en classe, examinons la question des séries de lancers consécutifs égaux au jeu de pile ou face.
La simulation est aisée sur Excel. On donne ensuite un exemple d'activités sur calculatrices, réalisables en seconde avec des petits groupes d'élèves.
La modélisation (admise) consiste à dire qu'à chaque lancer, on a "une chance sur deux" d'avoir pile ou d'avoir face. L'acceptation de ce modèle ne pose pas de difficulté. On va constater que, dans le cadre de ce modèle, la simulation permet de conjecturer des résultats non triviaux et, ici, non intuitifs.

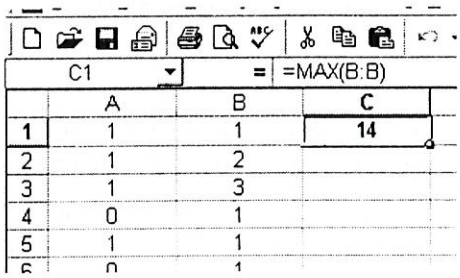
La question posée est la suivante :
au jeu de pile ou face, sur 200 lancers, quelle est la longueur maximale "habituelle" des séries de lancers consécutifs égaux ? Ou encore, est-il fréquent, sur 200 lancers, d'obtenir au moins une série de 5 lancers consécutifs égaux ?
Le tableur permet une simulation rapide de la situation.
Dans la colonne A, simulons les 200 lancers.
En ***A1***, on entre la formule :
=ENT(ALEA( ) + 0,5)
qui prend les valeurs 0 ou 1 avec une chance sur 2.
Recopions cette instruction vers le bas jusqu'en ***A200*** : on approche le pointeur de la souris du coin inférieur droit de la cellule A1, lorsque celui-ci prend la forme d'une croix noire, on enfonce le bouton gauche de la souris et l'on glisse jusqu'en A200 avec le bouton enfoncé.
Dans la colonne B, effectuons le décompte des lancers consécutifs égaux.
En ***B1*** on entre la valeur 1.
En ***B2*** on entre la formule :
SI(A2=A1 ; B1 + 1 ; 1)
qui affecte à la cellule B2 la valeur de B1 + 1 si la condition A2 = A1 est réalisée, et la valeur 1 sinon.
Recopier le contenu de B2 jusqu'en B200. Excel modifie automatiquement les références des cellules.
En ***C1***, il suffit de rechercher la plus longue série de lancers consécutifs égaux, en entrant la formule :
=MAX(B : B)


Fichier Edition Affichage Insertion Format Outils Donr

C1 = =MAX(B:B)

	A	B	C	D
1	1	1	7	
2	0	1		
3	0	2		
4	1	1		
5	1	2		
6	1	3		
7	0	1		
8	1	1		
9	1	2		
10	0	1		
11	1	1		
12	1	2		
13	0	1		
14	1	1		
15	1	2		
16	0	1		
17	1	1		
18	0	1		
19	0	2		
20	0	3		
21	0	4		
22	0	5		
23	1	1		
24	1	2		

Sur l'écran ci-contre, la plus grande série de lancers consécutifs égaux obtenus sur 200 lancers est de longueur 7.
En appuyant sur la touche F9, on réalise immédiatement une nouvelle simulation.
On constate que l'on ne descend guère en dessous d'au moins 5 lancers consécutifs égaux. On montre (voir page 26) que sur 200 lancers, on a 96% de chances d'obtenir au moins une série de 6 lancers consécutifs égaux. Ce que laisse bien soupçonner la simulation.
Ci-dessous par exemple, on a observé une série de 14 lancers consécutifs égaux.


C1 = =MAX(B:B)

	A	B	C
1	1	1	14
2	1	2	
3	1	3	
4	0	1	
5	1	1	
6	0	1	

Ces simulations mettent en évidence des propriétés non intuitives du hasard, qu'il est bon de connaître si l'on veut, par exemple, détecter une pièce truquée par des méthodes statistiques.
L'activité suivante utilise le support de la calculatrice, en petits groupes. Davantage peut-être que l'ordinateur, elle favorise le débat entre les élèves.

ACTIVITE EN SECONDE

T.P. AVEC CALCULATRICES

LANCERS CONSECUTIFS EGAUX A PILE OU FACE

TEMPS D'ATTENTE DE 3 LANCERS CONSECUTIFS EGAUX

Avec une pièce de monnaie

Lancer une pièce de monnaie, jusqu'à obtenir 3 piles ou 3 faces consécutifs. Effectuer cinq expériences, en notant dans le tableau ci-dessous, les résultats des lancers effectués jusqu'à obtenir trois coups consécutifs égaux, par exemple FFPFFPPFPPFFF , ainsi que le temps d'attente pour obtenir 3 consécutifs égaux, ici 13 lancers ont été nécessaires.

	Expérience	Temps d'attente
n° 1		
n° 2		
n° 3		
n° 4		
n° 5		

Simulation et temps d'attente moyen

Le programme suivant simule, pour N expériences, le temps d'attente moyen observé pour obtenir trois lancers consécutifs égaux.

CASIO Graph 25 → 100	T.I. 80	T.I. 82 - 83	T.I. 89 - 92
"N"↵ ? → N↵ 0 → S↵ For 1 → I To N↵ 3 → T↵ Int(Ran#+0.5) → A↵ Int(Ran#+0.5) → B↵ Int(Ran#+0.5) → C↵ While (A=B)+(B=C) ≠ 2↵ B → A↵ C → B↵ Int(Ran#+0.5) → C↵ T+1 → T↵ WhileEnd↵ S+T → S↵ Next↵ S÷N	:Input N :0 → S :For(I,1,N) :3 → T :int(rand + 0.5) → A :int(rand + 0.5) → B :int(rand + 0.5) → C :Lbl 1 :If (A=B) + (B=C) = 2 :Goto 2 :B → A :C → B :int(rand + 0.5) → C :T + 1 → T :Goto 1 :Lbl 2 :S + T → S :End :S/N	:Prompt N :0 → S :For(I,1,N) :3 → T :int(rand + 0.5) → A :int(rand + 0.5) → B :int(rand + 0.5) → C :While (A=B) + (B=C) ≠ 2 :B → A :C → B :int(rand + 0.5) → C :T + 1 → T :End :S + T → S :End :S/N	:Prompt n :0 → s :For i,1,n :3 → t :int(rand( ) + 0.5) → a :int(rand( ) + 0.5) → b :int(rand( ) + 0.5) → c :While a≠b or b≠c :b → a :c → b :int(rand( ) + 0.5) → c :t + 1 → t :EndWhile :s + t → s :EndFor :Disp s/n

⇨ **Pour obtenir certaines instructions :**

• CASIO Graph 25 → 100 : " par ALPHA ; **?** par PRGM ; **For To Next While WhileEnd EndFor** par PRGM puis COM ; **Int** par OPTN NUM ; **Ran#** par OPTN PROB ; = par PRGM REL.

• TI 80 → 92 :

Utilisation possible de la fonction **CATALOG** (sur TI 83 – 85 – 89 – 92).

Input Prompt par PRGM I/O ; → par STO ▸ ; **For While** par PRGM CTL ; **int** par MATH NUM ; **rand** par MATH PRB ; **If While Lbl Goto End** par PRGM CTL ; = ou ≠ par 2nd TEST (2nd MATH TEST sur TI 92).

Simulation d'une expérience

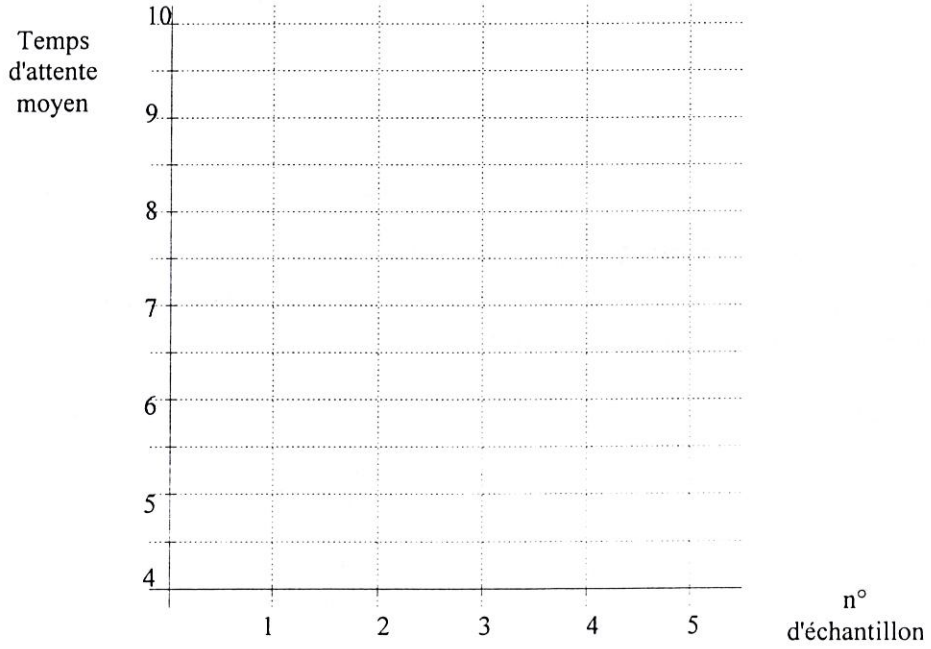
A l'invitation du programme précédent, entrer la valeur 1 pour N. Vous simulez l'expérience faite avec la pièce de monnaie au 1). Compléter le tableau suivant.

Temps d'attentes observés pour une expérience				

Temps d'attente moyen pour 10 expériences

A l'invitation du programme précédent, entrer la valeur 10 pour N.
Recommencer cinq fois et compléter le tableau, puis le graphique ci-dessous.

Temps d'attentes moyens observés pour 10 expériences				

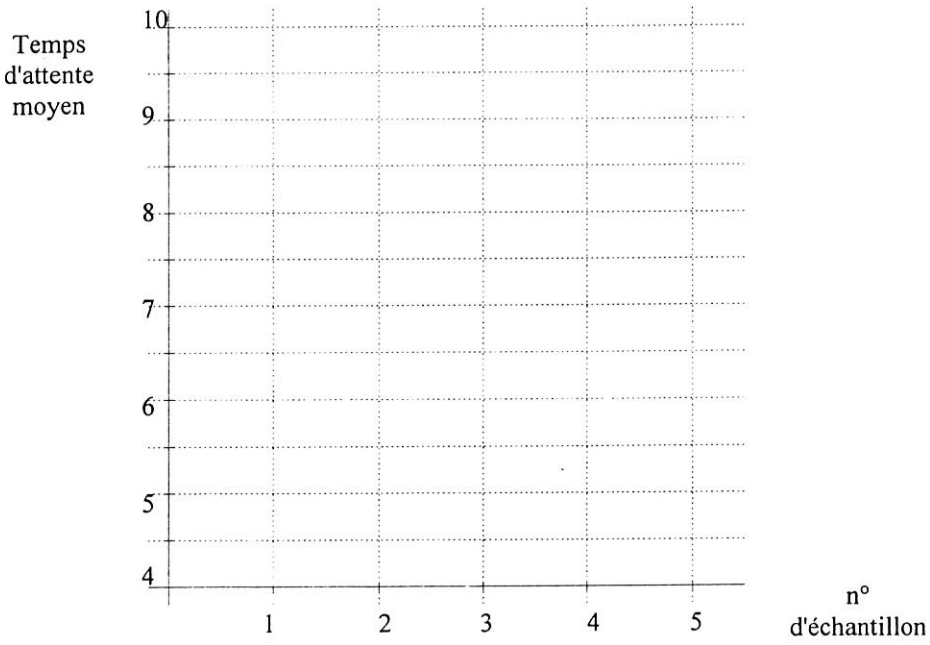


• Quelle est l'étendue de vos 5 valeurs ? .....

Temps d'attente moyen pour 100 expériences

A l'invitation du programme précédent, entrer la valeur 100 pour N (attention la durée de calcul est d'environ 1mn sur TI 83).
Recommencer cinq fois et compléter le tableau, puis le graphique.

Temps d'attentes moyens observés pour 100 expériences				



• Quelle est l'étendue de vos 5 valeurs ? .....

• Comparer aux observations obtenues au b- et donner une explication. .....

.....

• Calculer la moyenne de vos 5 temps d'attente moyens sur 100 expériences. A combien d'expériences correspond cette moyenne ? .....

.....

LONGUEUR MAXIMALE DE LANCERS CONSECUTIFS EGAUX SUR *n* LANCERS DE PILE OU FACE

Simulation

Le programme suivant calcule, pour *n* lancers de pile ou face simulés, la longueur maximale des lancers consécutifs égaux.

CASIO Graph 25 → 100	T.I. 80 - 82 - 83	T.I. 89 - 92
"N"↵ ? → N↵ Seq(1,I,1,2,1) → List 1↵ Int(Ran# + 0.5) → R↵ For 1 → I To N↵ Int(Ran# + 0.5) → S↵ If S = R↵ Then List 1[1] + 1 → List 1[1]↵ S → R↵ Max(List 1) → List 1[2]↵ Else 1 → List 1[1]↵ IfEnd↵ S → R↵ Next↵ List 1[2]	:Prompt N (Input N sur TI80) :1 → A :1 → L :int(rand + 0.5) → R :For(I,1,N) :int(rand + 0.5) → S :If S = R :Then :A + 1 → A :S → R :max(A,L) → L :Else :1 → A :End :S → R :End :L	:Prompt n :1 → a :1 → m :int(rand( ) + 0.5) → r :For i,1,n :int(rand( ) + 0.5) → s :If s = r :Then :a + 1 → a :s → r :max(a,m) → m :Else :1 → a :EndIf :s → r :EndFor :Disp m

⇨ **Pour obtenir certaines instructions :**
• CASIO Graph 25 → 100 : **Seq List** par OPTN LIST ; **If Then Else IfEnd** par PRGM COM ; **[ ]** par SHIFT ; **Max** par OPTN LIST.
• TI 80 → 92 : **max** par MATH NUM.

• A l'invitation du programme, entrer, pour N, la valeur 10. Quelle est la réponse fournie par le programme ? Que signifie-telle ? .....
.....

• A l'invitation du programme, entrer, pour N, la valeur 200 (attention à la durée de calcul). Recommencer 5 fois et compléter le tableau.

Longueur maximale des lancers consécutifs égaux, sur 200 lancers				

Commentez vos résultats. .....
.....
.....
.....

Le hasard n'est pas n'importe quoi ...

Il est 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 a été générée par un ordinateur (ou une réelle expérience de pile ou face).
Des deux tables suivantes (en lignes), quelle est celle imaginée au hasard par l'homme, et celle générée aléatoirement ?

Table 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

Table 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

Comment avez-vous fait ? Justifier votre démarche. .....
.....

Corrigé commenté de l'activité en 2nde : "LANCERS CONSECUTIFS EGAUX A PILE OU FACE"

CORRIGE COMMENTE

TEMPS D'ATTENTE DE 3 LANCERS CONSECUTIFS EGAUX

Simulation d'une expérience

On observe des résultats généralement très dispersés, par exemple :

```
prgmTPS3
N=?1
                    3
                 Done
N=?1
                    8
                 Done
```

Temps d'attentes observés pour une expérience				
3	8	3	15	6

Temps d'attente moyen pour 10 expériences

Temps d'attentes observés pour 10 expériences				
5,8	6,1	8,1	6,5	7,6

L'étendue de ces valeurs est 2,3.

Temps d'attente moyen pour 100 expériences

Temps d'attentes observés pour une expérience				
6,88	6,42	7,18	7,3	6,84

L'étendue est ici 0,76. Ce qui est bien inférieur au résultat précédent. Ceci est du au fait que cette fois la moyenne est faite sur 10 fois plus de valeur, ce qui atténue les effets du hasard. La moyenne de ces valeurs est 6,924 mn. Elle correspond à un temps d'attente moyen calculé sur 500 valeurs.

ENTRE NOUS ...

Nous allons étudier les variables aléatoires suivantes :

Soit L_n la variable aléatoire associant à n lancers de pile ou face, la longueur maximale des coups consécutifs égaux. Cherchons à déterminer $P(L_n < 3)$.

Soit T la variable aléatoire associant à des lancers successifs de pile ou face, le temps d'attente de 3 lancers consécutifs égaux. On montrera que $E(T) = 7$.

• **Expression de $P(L_n < 3)$ par récurrence :**

On note u_n le nombre de suites de n lancers de pile ou face (x_i) , i allant de 0 à n, ne contenant aucune séquence de 3 consécutifs égaux. Une telle suite peut être de 2 types disjoints :

⇨ $x_{n-1} \neq x_n$: 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.

On en déduit que $P(L_n < 3) = \dfrac{u_n}{2^n}$ où la suite (u_n) est définie par :

$u_1 = 2$; $u_2 = 4$ puis $\boxed{u_n = u_{n-1} + u_{n-2}}$ pour $n \geq 3$.

• **Expression directe de $P(L_n < 3)$:**

On l'a reconnu, la suite (u_n) est du type "*Fibonacci*". La recherche d'une suite géométrique (q^n) vérifiant la relation de récurrence conduit à l'équation caractéristique $q^2 - q - 1 = 0$ dont les solutions sont $\varphi = \dfrac{1+\sqrt{5}}{2}$ et $\overline{\varphi} = \dfrac{1-\sqrt{5}}{2}$.

Ainsi, $u_n = C_1\ \varphi + C_2\ \overline{\varphi}$ et les conditions initiales $u_0 = u_1 = 2$ donnent :

$P(L_n < 3) = (1+\frac{\sqrt{5}}{5})\left(\frac{\varphi}{2}\right)^n + (1-\frac{\sqrt{5}}{5})\left(\frac{\overline{\varphi}}{2}\right)^n$.

• **Expression du temps d'attente de trois lancers consécutifs égaux :**

pour tout $k \geq 3$, on a $P(T = k) = P(T > k-1) - P(T > k) = P(L_{k-1} < 3) - P(L_k < 3) = \frac{u_{k-1}}{2^{k-1}} - \frac{u_k}{2^k}$,

ou encore, $P(T = k) = \frac{u_{k-1} - u_{k-2}}{2^k} = \frac{1}{8}\left((1+\frac{\sqrt{5}}{5})\left(\frac{\varphi}{2}\right)^{k-3} - (1-\frac{\sqrt{5}}{5})\left(\frac{\overline{\varphi}}{2}\right)^{k-3}\right)$.

• **Espérance du temps d'attente de trois lancers consécutifs égaux :**

On a $E(T) = \sum_{k\geq 3} k\,P(T = k)$.

Pour $x < 1$, on a $\sum_{k\geq 3} kx^{k-3} = \frac{1}{x^2}\sum_{k\geq 3} kx^{k-1} = \frac{1}{x^2}\frac{d}{dx}\left(\frac{x^3}{1-x}\right) = \frac{3-2x}{(x-1)^2} = f(x)$.

On en déduit donc que $E(T) = \frac{1}{8}\left((1+\frac{\sqrt{5}}{5})\,f(\frac{\varphi}{2}) - (1-\frac{\sqrt{5}}{5})\,f(\frac{\overline{\varphi}}{2})\right) = 7$.

LONGUEUR MAXIMALE DE LANCERS CONSECUTIFS EGAUX SUR 200 LANCERS DE PILE OU FACE

Simulation

La simulation donne, par exemple,

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.

Le hasard n'est pas n'importe quoi

La table 1, où la longueur maximale de consécutifs égaux est de 4, a été imaginée par un être humain. Ce n'est pas le cas de la table 2, où on observe six 0 consécutifs, et qui a été obtenue par un générateur de nombres aléatoires.

ENTRE NOUS ...

La probabilité qu'une suite de 200 lancers de pile ou face contienne au moins une série de 6 lancers consécutifs égaux est environ 0,9653.

En effet, comme précédemment, notons u_n le nombre de suites de n lancers de pile ou face (x_i) , avec i de 0 à n, ne contenant aucune séquence de 6 consécutifs égaux. Une telle suite peut être de 5 types différents qui peuvent être 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.

On a donc $P(L_n < 6) = \frac{u_n}{2^n}$ où la suite (u_n) est définie par :

$u_1 = 2$; $u_2 = 4$; $u_3 = 8$; $u_4 = 16$; $u_5 = 32$ puis $\boxed{u_n = u_{n-1} + u_{n-2} + u_{n-3} + u_{n-4} + u_{n-5}}$ pour $n \geq 6$.

Ceci permet le calcul de proche en proche, de $P(L_{200} < 6)$. On a ainsi $P(L_{200} \geq 6) \approx 0{,}965313$.

3 – COMMENT 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.
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].

Il est important de noter qu'il n'y a pas de "garantie" qu'un générateur de nombres aléatoires fonctionne toujours correctement. En particulier, il y a toujours certaines limites à la simulation.

Un exemple de générateur de nombres (pseudo) aléatoires obtenu selon la méthode de Lucas-Lehmer[1]

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 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 :

$\boxed{r_{n+1} = a\, r_n \bmod(m)}$,

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 $x_n = \dfrac{r_n}{m}$ 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 :

$\boxed{a = 7^5 \ ; m = 2^{31} - 1 \ ; r_{n+1} = a\, r_n \bmod(m)}$

puis $\boxed{x_n = r_n / m}$.

Propriétés arithmétiques
de la suite $r_{n+1} = ar_n \bmod(m)$
$0 < r_0 < m$, 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 $a^k = 1 \bmod m$.
En effet $r_{n+t} = r_n \Leftrightarrow a^t r_n = r_n \bmod(m) \Leftrightarrow (a^t - 1)r_n = km$ avec k entier, c'est à dire, puisque r_n est premier avec m, $a^t = 1 \bmod m$.

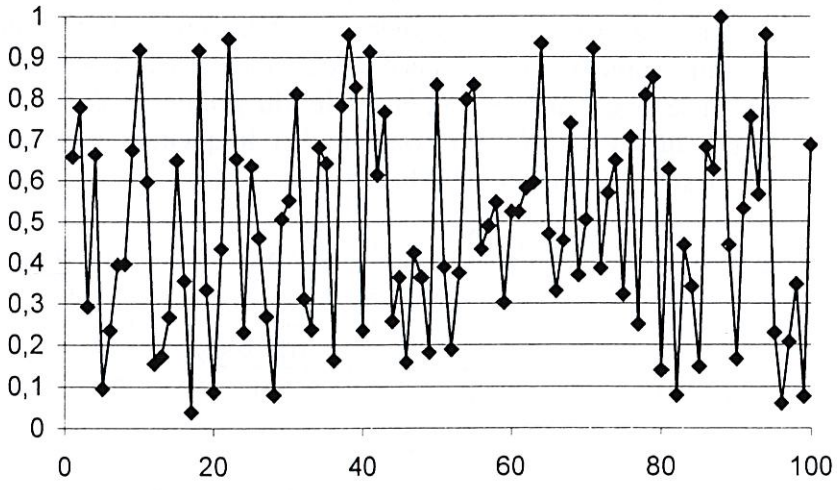
• Lorsque m est premier, le petit théorème de *Fermat* affirme que si m est premier et ne divise pas a, alors $a^{m-1} = 1 \bmod(m)$, donc, pour a premier avec m, l'ordre multiplicatif de a divise $m - 1$.

[1] Ce paragraphe peut être omis en première lecture.

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 $2^2 = 4$ mod(5) puis $2^3 = 3$ mod(5) et $2^4 = 1$ mod(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 $\{1 , 2 , ... , m - 2\}$.
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 $a^{\varphi(m)} = 1$ mod(m) où φ est l'indicateur d'*Euler* correspondant au nombre d'entiers premiers avec m dans $\{1 , 2 , ... , m - 1\}$. 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 ($5^6 = 1$ mod(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 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 Hanoï").

Tester un générateur de nombres aléatoires[4]

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 Ent(10 x_n) où 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 $\mathcal{N}\left(\frac{1}{10}, \sqrt{\frac{\frac{1}{10}\times\frac{9}{10}}{n}}\right)$, d'après le théorème limite central (ces notions sont rappelées plus loin).

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 2σ de 1/10, alors qu'un écart de 3σ 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.

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	$\frac{10\times9\times8\times7}{10^4}$ = 0,504	$C_4^2\frac{10\times9\times8}{10^4}$ = 0,432	$\frac{C_4^2}{2}\times\frac{10\times9}{10^4}$ = 0,027	$\frac{4\times10\times9}{10^4}$ = 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 1ère expérience	541	409	18	32	0
effectifs x_i observés à la 2ème 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

[4] Certains passages de ce paragraphe font appel à des techniques statistiques qui ne sont vues que par la suite, il peut être omis en première lecture.

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 $\chi^2_{obs} = \sum_{i=1}^{5} \frac{(x_i - t_i)^2}{t_i}$.

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 χ^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 χ^2_{obs} supérieure à 9,48.
Pour les échantillons obtenus précédemment avec le générateur aléatoire d'Excel, on a :
$\chi^2_{obs\,1} \approx 8{,}39$ et $\chi^2_{obs\,2} \approx 1{,}25$.

Un générateur de nombres aléatoires existe sur toutes les calculatrices sous la forme de la touche ***random***[5] : Ran# (CASIO) ou rand (T. I.).

Une activité élèves en seconde étudiant le générateur ALEA d'Excel

Une des difficultés de la simulation utilisant la calculatrice ou l'ordinateur est le côté "boîte noire" du générateur de nombres aléatoires. On peut avoir l'impression, en partie justifiée, de n'observer que les propriétés du générateur et non celles du hasard. La machine ne ressortirait, sous une autre forme, que ce que l'on a rentré dedans.
Est-ce bien le hasard que l'on étudie ? Pour le faire accepter par les élèves, il est conseillé de leur faire tout d'abord manipuler de vrais objets (pièces de monnaie, dés, ...), comme dans l'activité élèves décrite précédemment, afin de constater que les simulations fabriquées à partir du générateur sont de nature analogue.
Il nous semble même intéressant d'étudier en TP le générateur lui même, et les fluctuations des résultats. C'est ce que l'on propose dans l'activité suivante, à propos de celui d'Excel.
Les élèves, en demi-classe, sont au maximum deux par ordinateur. Ils doivent rendre un compte-rendu ("feuille réponse") à la fin de la séance, pour les obliger à analyser les résultats produits par l'ordinateur. Une "correction", ou synthèse est ensuite rapidement faite, en classe entière.

---

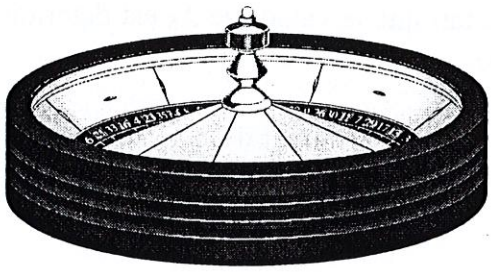
[5] Le mot *random* signifie "hasard" en anglais, il vient du vieux français *randon*, que l'on retrouve dans randonnée.

ACTIVITE ELEVES EN SECONDE

Travaux pratiques X STATISTIQUES

ETUDE DE LA FONCTION ALEA D'EXCEL

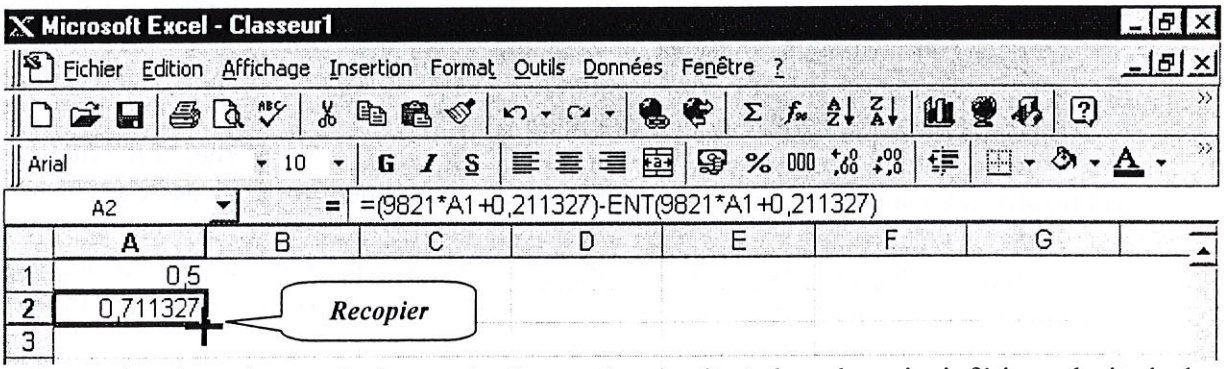
Simuler le hasard avec un ordinateur. Comment est-ce possible ?



Comment un ordinateur qui est fait pour calculer, selon des programmes précis qui ne doivent rien au hasard, peut-il fournir des nombres choisis "au hasard" ?
Bien sûr, ce n'est pas possible, le hasard n'habite pas la machine. En revanche, on peut calculer des nombres qui "ont l'air" d'arriver au hasard et ceci n'est pas nécessairement compliqué.

Une formule de simulation de nombres "au hasard"

Lancer Excel®.
Cliquer (avec le bouton gauche de la souris) dans la cellule A1, taper 0.5 puis ***ENTREE***.
Dans la cellule A2, taper la ***formule*** (*attention*, chaque formule doit commencer par le signe =) :
=(9821*A1+0.211327)-ENT(9821*A1+0.211327) puis ***ENTREE***.
Dans cette formule, la fonction ENT désigne la partie entière (ce qui est situé avant la virgule).



Approcher le pointeur de la souris du carré noir situé dans le coin inférieur droit de la cellule A2. Celui-ci se transforme en une croix noire, faire alors glisser, en maintenant le bouton gauche enfoncé pour ***recopier*** jusqu'en A15, puis relâcher le bouton de la souris.

Analyser les résultats sur la feuille réponse.

Les résultats semblent arriver au hasard. On effectue pourtant toujours le même calcul, assez simple, en boucle. Les nombres étranges, 9821 et 0,211327, ont été choisis (par tâtonnements) pour le caractère imprévisible des résultats.

La fonction ALEA()

Le générateur de nombres aléatoires d'Excel fonctionne selon un principe semblable.
En B1, entrer la ***formule*** : =ALEA() puis ***recopier*** comme précédemment jusqu'en B15.

✍ Comparer, sur la feuille réponse, aux résultats de la colonne A.

Obtenir un entier "au hasard" entre 0 et 9

En C1 entrer la ***formule*** : =ENT(10*ALEA()) puis ***recopier*** jusqu'en C15.

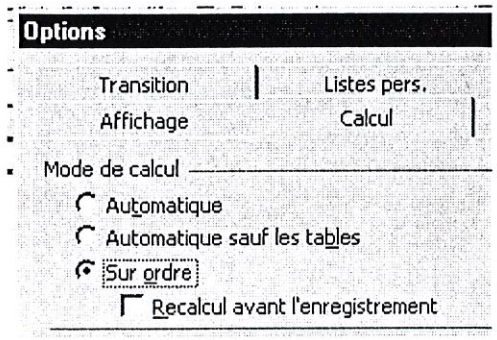
✍ Compléter la feuille réponse.

Il ne suffit cependant pas d'être imprévisible pour simuler le hasard, encore faut-il satisfaire à certains critères statistiques. Afin de tester la qualité d'un générateur de nombres aléatoires, on lui fait passer de nombreuses épreuves statistiques.

APPARITIONS DES DIFFERENTS CHIFFRES

Sur 100 chiffres fournis par ALEA

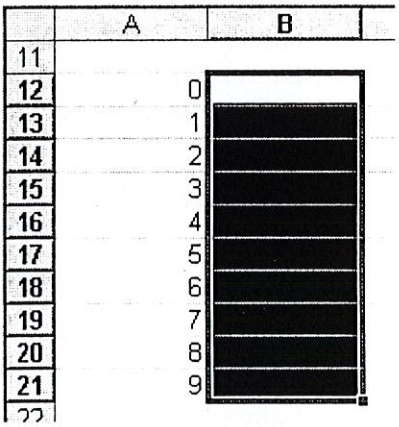
Examinons par exemple, dans le cas de 100 nombres entiers entre 0 et 9, fournis à l'aide du générateur ALEA, quelle est la fréquence d'apparition de chacun.



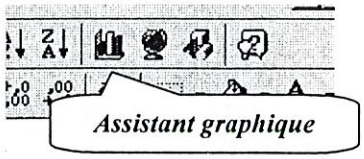
Cliquer sur l'onglet ***Feuil2*** en bas de l'écran.
Afin de ne lancer les calculs que lorsqu'on le désire, configurer Excel ainsi :
cliquer dans le menu ***Outils/Options...*** puis dans l'onglet ***Calcul*** , puis à la rubrique ***Mode de calcul***, choisir • ***Sur ordre*** puis ***OK***.
En A1 entrer la ***formule*** : =ENT(10*ALEA())
Puis ***recopier*** jusqu'en A10, relâcher le bouton gauche de la souris et, de nouveau, ***recopier*** jusqu'en J10.

Appuyer sur la touche ***F9*** pour lancer le calcul.
Vous avez simulé le tirage de 100 chiffres "au hasard".

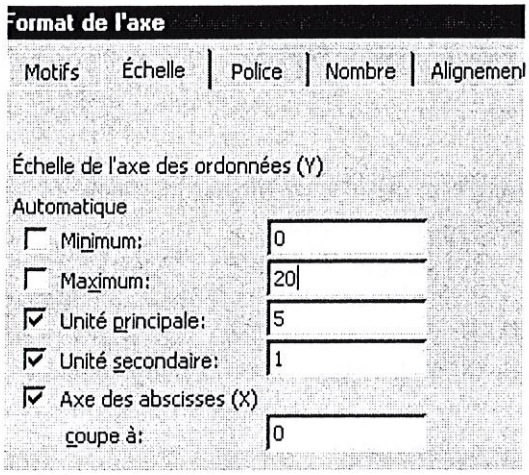
Pour étudier la répartition de ces 100 chiffres, on va compter le nombre de fois que chacun apparaît.
De A12 à A21, entrer les 10 chiffres (en A12 taper 0, en A13 taper 1, ... , en A21 taper 9).
Sélectionner les cellules de B12 à B21 (pour cela cliquer sur B12 et glisser, en gardant le bouton gauche de la souris enfoncé, jusqu'en B21, puis relâcher le bouton de la souris).
Alors que les cellules sélectionnées apparaissent en "vidéo inversée", écrire la ***formule*** :
=FREQUENCE(A1:J10;A12:A21) puis valider en appuyant *en même temps* sur les touches ***CTRL MAJUSCULE*** et ***ENTREE***.
Excel calcule alors les *effectifs* d'apparition de chacun des 10 chiffres (et non les *fréquences* comme semble l'indiquer le nom de la fonction utilisée).



Vous allez maintenant représenter ces effectifs sous forme d'un ***histogramme***.



Cliquer sur l'icône ***Assistant graphique***.
Etape 1 sur 4 : choisir Histogramme et le premier sous-type.
Cliquer sur Suivant.
Etape 2 sur 4 :
Dans l'onglet ***Plage de données***, taper dans la fenêtre ***Plage de données :*** B12:B21
puis cocher Série en : • Colonnes .
Dans l'onglet ***Série***, à la rubrique ***Etiquettes des abscisses***, taper =Feuil2!A12:A21
puis cliquer sur ***Suivant***.
Etape 3 sur 4 : cliquer sur Suivant.
Etape 4 sur 4 : cocher • en tant qu'objet dans Feuil2 puis ***Terminer***.
Déplacer l'histogramme, en gardant le bouton gauche de la souris enfoncé.
Cliquer avec le bouton *droit* de la souris sur la ***légende*** puis avec le bouton *gauche* sur ***Effacer***.



Cliquer avec le bouton *droit* de la souris, sur l'***Axe des ordonnées*** puis avec le bouton *gauche* sur ***Format de l'axe...***
Choisir l'onglet ***Echelle***.
Ne pas cocher ***Minimum*** ni ***Maximum*** mais indiquer comme valeurs respectives 0 et 20.
Cliquer sur ***OK***.
Pour obtenir un autre échantillon de 100 chiffres, appuyer sur ***F9*** et observer l'histogramme.

✍ Consigner les observations sur la feuille réponse.

Sur 1000 chiffres fournis par ALEA

Cliquer, en bas, sur l'onglet ***Feuil3***.
Entrer en A1 la ***formule*** : =ENT(10*ALEA()) puis ***recopier*** jusqu'en A100. Relâcher, puis ***recopier*** jusqu'en J100. Appuyer sur ***F9***.
Vous avez maintenant 1000 chiffres choisis "au hasard" à l'aide d'ALEA, dont on va étudier la répartition.
De A102 à A111, entrer les chiffres 0 ; 1 ; 2 ; ... ; 9.
Sélectionner les cellules de B102 à B111 et taper la ***formule*** :
=FREQUENCE(A1:J100;A102:A111) puis appuyer *en même temps* sur ***CTRL MAJUSCULE*** et ***ENTREE***.
Procéder comme au paragraphe précédent pour créer un ***Histogramme*** avec:
Plage de données : B102:B111 • en Colonnes puis
Série : Etiquettes des abscisses =Feuil3!A102:A111
Modifier le ***Format de l'axe*** des ordonnées avec une ***Echelle*** dont le ***Minimum*** soit 0 et le ***Maximum*** 200.
Faire plusieurs simulations en appuyant sur ***F9***.

✍ Consigner vos observations sur la feuille réponse.

TEST DU POKER

Bien sûr, étudier la proportion de chaque résultat possible n'est pas suffisant. On pourrait avoir des résultats du type 000111222333444555666777888999000111.... Les proportions sont respectées mais on n'imite pas le hasard !

Un autre test statistique consiste à faire des groupes de 4 chiffres consécutifs et à examiner la proportion de résultats où tous les nombres sont différents (5819), où on a une paire (1518), trois chiffres identiques (1151) etc.
C'est le test du poker.

Etudions le cas des groupes de 4 chiffres différents.

Etude théorique

On choisit au hasard, avec ordre et remise, 4 chiffres parmi les 10 (de 0 à 9). On peut obtenir, par exemple, 3893.

✍ Indiquer, sur la feuille réponse, quelles chances a-t-on, théoriquement, d'obtenir, à un tel tirage, un nombre de 4 chiffres différents ?

Simulation statistique

Dans le menu ***Insertion***, cliquer sur ***Feuille***.
En A1, entrer la ***formule*** : =ENT(10*ALEA()) puis ***recopier*** jusqu'en D1 et appuyer sur ***F9***. On a ainsi un premier tirage de quatre chiffres.
En E1, entrer la ***formule*** suivante, qui affichera 0 ou 1 selon qu'il existe ou non deux chiffres identiques : =SI(OU(A1=B1;A1=C1;A1=D1;B1=C1;B1=D1;C1=D1);0;1)
Sélectionner les cellules de A1 à E1 puis ***recopier*** vers le bas jusqu'en E100.
Appuyer sur ***F9***.
En E101, cliquer sur l'icône Σ de ***somme automatique*** et faire ***ENTREE***.

✍ Faire d'autres simulations et consigner vos observations sur la feuille réponse.

✍ FEUILLE REPONSE

NOMS : .....

Simuler le hasard avec un ordinateur. Comment est-ce possible ?

Une formule de simulation de nombres "au hasard"

Dans les cellules A1 à A15, est-il possible de prévoir le résultat d'une cellule à la cellule suivante ? Distingue-t-on un ordre ? .....
Cliquer dans la cellule A3. Examiner la formule inscrite. De même dans la cellule A4. Comment est effectué le calcul d'une cellule à la cellule suivante ? .....
.....

La fonction ALEA()

La fonction ALEA simule un nombre choisi "au hasard" entre 0 et 1. De même que dans les opérations précédentes, peut-on prévoir, dans la colonne B, le résultat d'une cellule à la cellule suivante ? .....

Obtenir un entier "au hasard" entre 0 et 9

La formule 10*ALEA() donne un nombre décimal compris entre ..... et .....
En prenant la partie entière, quels sont les résultats possibles de la formule ENT(10*ALEA()) ? .....
.....

APPARITIONS DES DIFFERENTS CHIFFRES

Sur 100 chiffres fournis par ALEA

Théoriquement, pour un hasard "parfait", quelles chances a-t-on d'obtenir le nombre 0 à un tirage ? .....
Quelle devrait être, théoriquement, la proportion d'apparition de chaque chiffre, *sur un grand nombre de tirages* ? .....
Sur 5 simulations de 100 tirages, indiquer la différence de hauteur entre le plus grand et le plus petit rectangle de l'histogramme :

Simulation n°	1	2	3	4	5
Différence maximale des effectifs					
Différence relative en poucentage (résultat du dessus divisé par 100)	%	%	%	%	%

Les fluctuations d'apparition des différents nombres entre les simulations sont assez importantes. Doit-on mettre en doute la qualité de la fonction ALEA ou considérer qu'un échantillon de taille 100 est insuffisant ? .....
.....

Sur 1000 chiffres fournis par ALEA

Sur 5 simulations de 1000 tirages, indiquer la différence de hauteur entre le plus grand et le plus petit rectangle de l'histogramme :

Simulation n°	1	2	3	4	5
Différence maximale des effectifs					
Différence relative en poucentage (résultat du dessus divisé par 1000)	%	%	%	%	%

Comparer aux fluctuations observées précédemment sur des échantillons de taille 100.

.....

.....

TEST DU POKER

Etude théorique

Pour compter le nombre de résultats possibles, on peut dire que l'on a 10 possibilités pour choisir le 1er chiffre, multiplié par 10 possibilités pour le choix du second, et ainsi de suite jusqu'au 4ème.

Soit au total $10\times10\times10\times10 = 10^4$ *résultats possibles*.

Combien peut-il y avoir de résultats où les 4 chiffres sont différents ?

.....

Montrer que l'on a, théoriquement, 50,4 % de chance d'obtenir, à un tirage, un nombre de 4 chiffres différents.

.....

.....

Simulation statistique

Sur 10 simulations, de 100 groupes de 4 chiffres, les pourcentages de tirages où les 4 chiffres sont différents ont été :

Simulation n°	1	2	3	4	5	6	7	8	9	10
%										

Faire la moyenne des pourcentages et comparer au résultat théorique.

.....

.....

.....

Corrigé de l'activité en 2nde : "ETUDE DE LA FONCTION ALEA D'EXCEL"

SIMULER LE HASARD AVEC UN ORDINATEUR

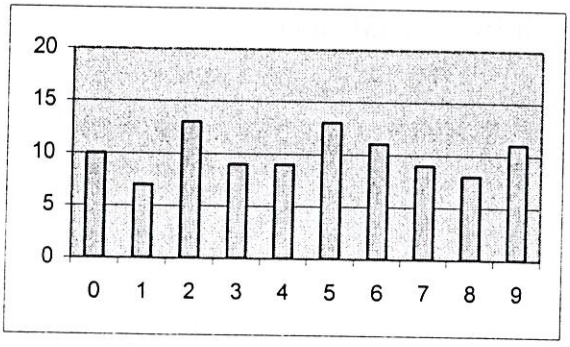
Une formule de simulation de nombres "au hasard"
Il s'agit de montrer que l'on peut générer de façon assez simple des nombres pseudo-aléatoires.
En recopiant vers le bas, le numéro de la cellule est automatiquement incrémenté. Le calcul d'une cellule est donc fait à partir du résultat de la cellule précédente.
Voici, ci-contre, les premiers résultats.

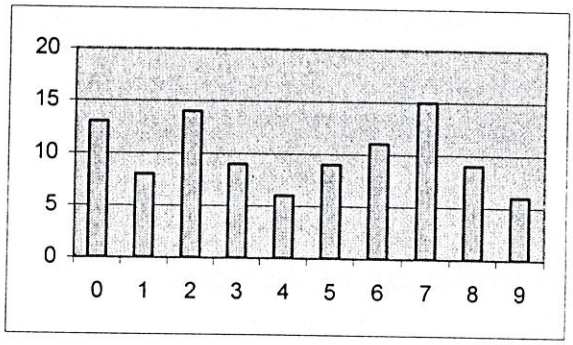
0,5
0,711327
0,153794
0,62219935
0,83119106
0,33872289
0,80881084
0,54262732
0,35427095
0,50631348
0,71599532
0,00132503
0,22447315
0,76211535
0,946207

La fonction ALEA()
Après avoir observé quelques résultats de la fonction ALEA() d'Excel, on se ramène, par ENT(10*ALEA()), à des entiers aléatoires entre 0 et 9, ce qu'il est plus aisé à manipuler que les réels entre 0 et 1.

APPARITION DES DIFFERENTS CHIFFRES

Sur 100 chiffres fournis par ALEA()
Sur 100 nombres, on observe des fluctuations assez importantes, de part et d'autre des 10% attendus.

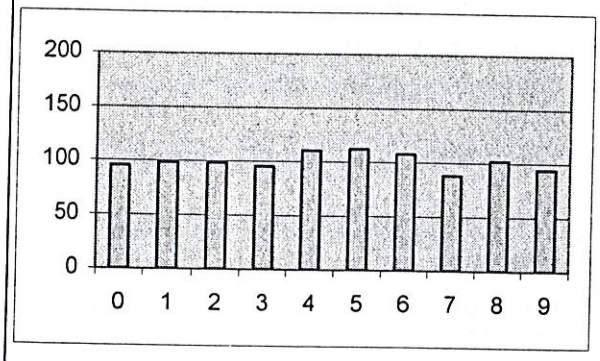


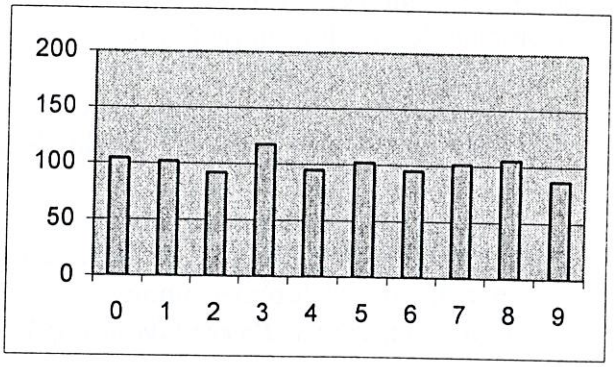


ENTRE NOUS ...

Considérons la variable aléatoire F, qui, à tout échantillon aléatoire de taille n, associe la fréquence d'apparition, dans cet échantillon, du zéro (par exemple).
On sait (voir séance 2) que, pour n assez grand, F suit approximativement la loi normale $\mathcal{N}\left(\frac{1}{10}, \sqrt{\frac{\frac{1}{10}\times\frac{9}{10}}{n}}\right)$.
Ainsi, l'écart type entre les fréquences d'apparition du zéro (ou de tout autre chiffre) sur différents échantillons de taille 100 est environ 0,03.

Sur 1000 chiffres fournis par ALEA()
En prenant une échelle relative analogue, on constate que les fluctuations sont très amorties, lorsque l'on passe à 1000 données.





ENTRE NOUS ...

L'écart type entre les fréquences d'apparition du zéro (ou de tout autre chiffre) sur différents échantillons de taille 1000 est environ 0,0095.

TEST DU POKER

Pourcentages, sur 100 groupes de 4 chiffres, des cas où les 4 chiffres sont différents :
On observe par exemple,
56% ; 47% ; 48% ; 52% ; 52% ; 49% ; 54% ; 46% ; 44% ; 55%.
Soit un pourcentage global de 50,3 %, proche de celui attendu.

ENTRE NOUS ...

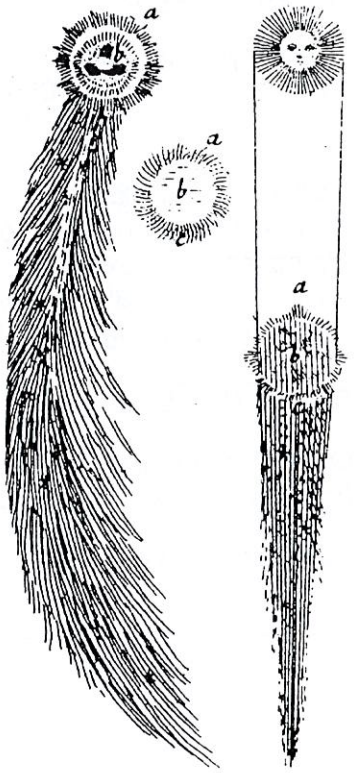
On peut juger ici de l'efficacité de la simulation, puisque la probabilité est connue : $p = 0{,}504$.
En simulant le prélèvement aléatoire de $n = 1000$ groupes de quatre chiffres, on doit s'attendre à un écart type (autour de $p = 0{,}504$) de l'ordre de $\sqrt{\frac{0{,}504 \times 0{,}496}{1000}} \approx 0{,}016$.

III – L'APPROCHE STATISTIQUE DES PROBABILITES : Loi normale et théorèmes limites

Est-ce que simuler permet de bonnes conjectures, dans le cadre d'une situation aléatoire ? Combien de fois doit-on répéter les expériences ? Quelle incertitude a-t-on ?
La simulation est fondée sur la loi des grands nombres, précisée par le théorème central-limite, où apparaît le rôle primordial de la loi normale.
"L'invention", au début du XIX^{e} siècle, de la loi "normale", dont l'usage est fondamental en statistique, s'est faite 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 Simon Laplace*** (1749-1827).

1 – La loi des erreurs



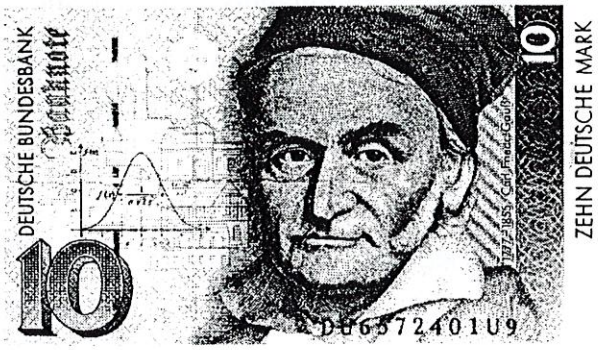
• La moyenne est la meilleure "estimation" du point de vue de la géométrie euclidienne :

Adrien-Marie Legendre (1752-1833) publie en 1805, dans ses *"Nouvelles méthodes pour la détermination des orbites des comètes"*, la méthode consistant à minimiser la somme des carrés des écarts. Cette méthode correspond à l'ajustement optimal pour la structure géométrique euclidienne. Soit une quantité θ inconnue, pour laquelle on possède plusieurs mesures différentes x_1 , x_2, ..., x_n (il y a toujours des erreurs aléatoires irréductibles, il ne s'agit pas des erreurs "systématiques" que l'on sait détecter et évaluer). On cherche à "résumer" au plus proche le vecteur $(x_1 , x_2, ..., x_n)$ par une valeur unique θ c'est à dire par le vecteur $(\theta, \ldots, \theta)$.
Pour la distance euclidienne, on a :
$d^2\big((\theta,\theta,...,\theta), (x_1, x_2,..., x_n)\big) = \sum(\theta - x_i)^2$.
L'estimation $\hat{\theta}$ de θ rendant minimale la somme des carrés des écarts $\sum(\theta - x_i)^2$ correspond à la moyenne.
En effet, ce minimum sera obtenu en annulant la dérivée par rapport à θ, soit $\frac{d}{d\theta}\sum(\theta - x_i)^2 = 2\sum(\theta - x_i) = 0$ qui donne $\hat{\theta} = \frac{x_1 + ...x_n}{n}$.

Ainsi, la moyenne est la caractéristique de position optimale, du point de vue de la géométrie euclidienne.

"Il est important de démontrer que la moyenne minimise la fonction $x \mapsto \sum_i (x_i - x)^2$ *qui mesure la dispersion, et que ce minimum est appelé la variance. Commentaires du programme 2001 de 1ère S.*

• La loi normale est celle qui rend l'estimation par la moyenne "optimale" d'un point de vue probabiliste :



Indépendamment de *Legendre*, ***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 ainsi à la "loi gaussienne".
Gauss pose ainsi le problème :
Quelle est la loi des erreurs pour laquelle $\frac{1}{n}\sum x_i$ est la valeur de θ rendant maximale la probabilité d'observer les valeurs $x_1, \ldots, x_n$?
En envisageant la question d'un point de vue probabiliste, on considérera donc que les erreurs $e_1 = x_1 - \theta$, ..., $e_n = x_n - \theta$ sont des réalisations de n variables aléatoires indépendantes E_1, ..., E_n de même loi continue de densité f , dépendant de la valeur inconnue θ.

La méthode du "maximum de vraisemblance" :

Pour θ donné, la probabilité d'effectuer des erreurs entre e_1 et $e_1 + de_1$, ..., e_n et $e_n + de_n$ est alors, en vertu de l'indépendance, le produit des probabilités, soit $f(e_1)\, de_1 \times \ldots \times f(e_n)\, de_n$.
On peut alors retourner le raisonnement (à la façon de *Bayes*) et se demander, les mesures x_1, ..., x_n étant connues, quelle est la valeur de θ la plus vraisemblable. C'est à dire, quelle est la valeur de θ qui rendra maximale la probabilité d'observation des mesures x_1, ..., x_n (réellement observées) donc des erreurs e_1, ...e_n.
Il s'agit donc de rechercher θ , donc f, de sorte que $f(x_1 - \theta) \times \ldots \times f(x_n - \theta)$ soit maximum (ce qu'on appellerait ultérieurement "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 θ, on obtient la condition $\sum \frac{d \ln f(x_i - \theta)}{d\theta} = 0$.
Sachant que la moyenne arithmétique $\hat{\theta} = \frac{x_1 + \ldots x_n}{n}$ correspond à la valeur recherchée de θ et que cette moyenne vérifie l'équation en θ : $\sum (x_i - \theta) = 0$, *Gauss* en déduit que, pour i allant de 1 à n, il suffit de poser : $\frac{d \ln f(x_i - \theta)}{d\theta} = k(x_i - \theta)$.
On a , 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 : le réel k est strictement positif (sans quoi il ne peut s'agir d'une densité de probabilité) et dépend de la dispersion (ce que l'on nommera plus tard l'écart type), le réel C est calculé de sorte que l'intégrale sur $\mathbb{R}$ fasse 1.

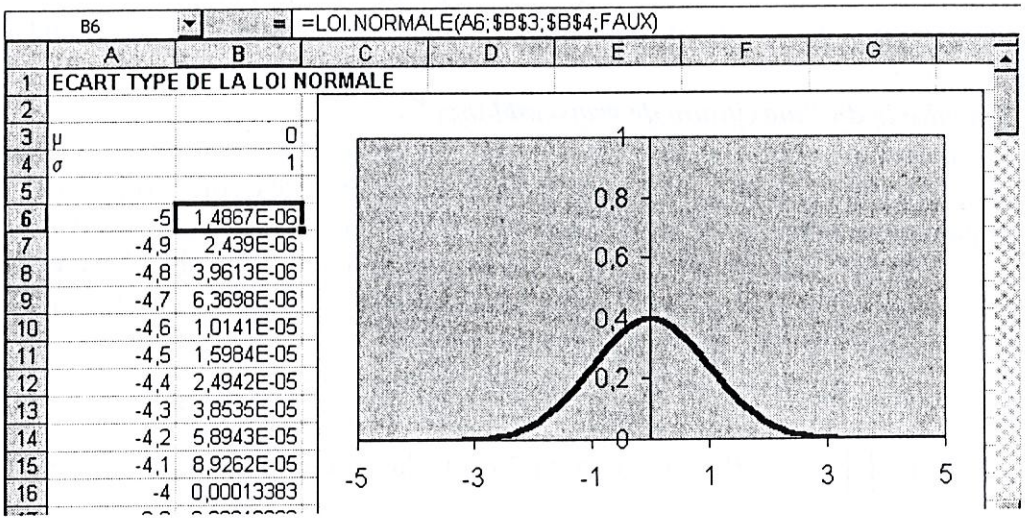
On remarque rétrospectivement que si $f(e_i) = C\, e^{-\frac{k}{2}e_i^2}$, alors $\prod f(e_i) = C\, e^{-\frac{k}{2}\sum e_i^2}$ est maximum lorsque la somme des carrés des écarts $\sum e_i^2$ est minimale.

Ainsi, la moyenne, paramètre déjà optimal du point de vue euclidien, l'est également du point de vue probabiliste (la moyenne est la valeur la plus probable) lorsque la "loi des erreurs" est normale.
C'est la justification réciproque moindres carrés + moyenne-écart type + loi normale, qui explique l'usage presque exclusif de la moyenne, jusqu'à une époque récente (voir la séance 2).

RAPPELS A PROPOS DE LA LOI NORMALE

> A propos de la loi normale : *"Tout le monde y croit cependant, me disait un jour Monsieur Lippmann, car les expérimentateurs s'imaginent qu'il s'agit d'un théorème de mathématiques et les mathématiciens que c'est un fait expérimental."*
> Boutade rapportée par *Henri Poincaré* – 1912.

Il s'agit d'une loi à densité *continue* pouvant prendre n'importe quelle valeur réelle.



Une variable aléatoire X dont l'ensemble des valeurs possibles est $\mathbb{R}$, suit la loi normale de paramètres μ et σ, notée $\mathcal{N}(\mu\ ,\ \sigma)$, lorsque :
pour tous les réels a et b (avec $a < b$ ou a valant $-\infty$ ou b valant $+\infty$), on a :

$$P(a \le X \le b) = \int_a^b f(x)dx \quad \text{avec} \quad \boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}.$$

La fonction f est la ***densité*** de la loi $\mathcal{N}(\mu\ ,\ \sigma)$. On a $E(X) = \mu$; $\sigma(X) = \sigma$.
La loi normale possède des propriétés fondamentales de linéarité. Une combinaison linéaire de variables aléatoires indépendantes de lois normales suit une loi normale. Si X suit la loi normale $\mathcal{N}(\mu\ ,\ \sigma)$, alors $T = \dfrac{X-\mu}{\sigma}$ suit la loi normale standard (centrée, réduite) de moyenne 0 et d'écart type 1. On peut donc facilement se ramener à cette dernière, dont la fonction de densité (courbe "en cloche") est représentée sur l'image ci-dessus et pour laquelle les résultats suivants sont particulièrement importants (et utilisés par la suite) :
$P(-1{,}96 \le T \le 1{,}96) \approx 0{,}95$ et $P(-2{,}58 \le T \le 2{,}58) \approx 0{,}99$.

2 – Les théorèmes limites

C'est ***Pierre Simon de Laplace*** qui va consacrer, avec le *théorème limite central*, la prépondérance en statistique de la loi normale.

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

a) Loi des grands nombres

"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.

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

Loi des grands nombres (énoncé vulgarisé pour la probabilité d'un événement) :

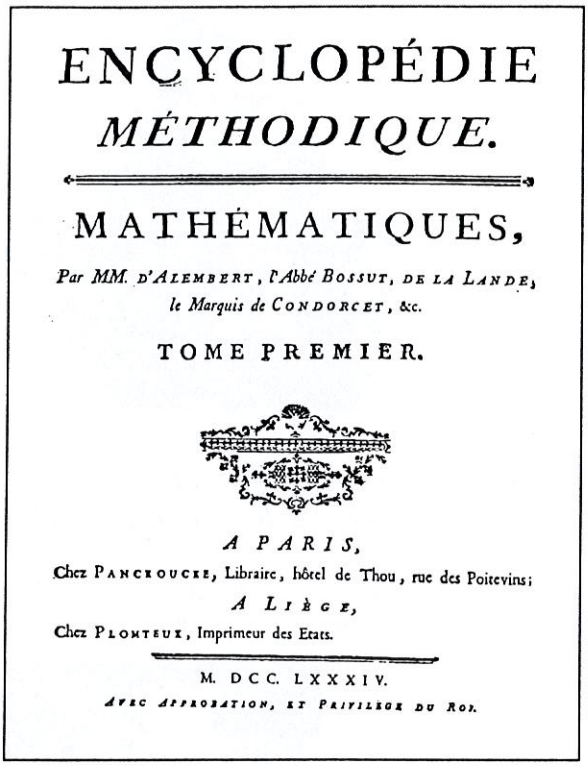
On répète de façon indépendante la même expérience aléatoire où l'on observe ou non l'événement A.
***Plus on fait d'expériences, plus la fréquence d'apparition de* A *se rapproche de la probabilité de* A.**

Le choix du programme de 1ère est, non pas d'introduire la notion de probabilité d'un événement, mais celle de loi de probabilité :

Le lien entre loi de probabilité et distributions des fréquences sera éclairé par un énoncé vulgarisé de la loi des grands nombres :
❑ *Pour une expérience donnée, dans le modèle défini par une loi de probabilité P, les distributions des fréquences calculées sur des séries de taille n se rapprochent de P quand n devient grand. (Programme 2001 de 1ère S).*
ou encore :
❑ *Si on choisit n éléments d'un ensemble fini E selon une loi de probabilité P, indépendamment les uns des autres, alors la distribution des fréquences est proche de la loi de probabilité P quand n est grand. (Document d'accompagnement 1ère S 2001).*

Voir la séance 3 pour cette approche statistique d'une loi de probabilité.

Voyons comment, dans l'édition "méthodique" de l'***Encyclopédie***, *Condorcet* montre l'intérêt de l'approche fréquentiste, en rapportant les travaux de *Jacques Bernoulli.*



ENCYCLOPÉDIE
MÉTHODIQUE.

MATHÉMATIQUES,

Par MM. d'Alembert, l'Abbé Bossut, de la Lande, le Marquis de Condorcet, &c.

TOME PREMIER.

A PARIS,
Chez Panckoucke, Libraire, hôtel de Thou, rue des Poitevins;
A Liège,
Chez Plomteux, Imprimeur des Etats.

M. DCC. LXXXIV.
Avec Approbation, et Privilege du Roi.

Article PROBABILITE de l'*Encyclopédie* par ***Condorcet*** (1751)

A propos des "*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 influer 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.

M. Bernoulli, le géomètre, qui entendoit le mieux ces ſortes de calculs, s'eſt propoſé l'objection & en donne la réponſe. On la trouvera dans ſon livre *de arte conjectandi*, *p.* 4, dans toute ſon étendue ; problême, ſuivant lui, auſſi difficile que la quadrature du cercle. Il y fait voir que la *probabilité*, qui naiſſoit de l'expérience répétée, alloit toujours en croiſſant, & croiſſoit tellement, qu'elle s'approchoit indéfiniment de la certitude. Son calcul nous apprend à déterminer (queſtion propoſée d'une manière fixe) combien de fois il faudroit réitérer l'expérience pour parvenir à un degré aſſigné de *probabilité*. Ainſi, dans le cas d'une urne pleine d'un grand nombre de boules blanches & noires, on veut s'aſſurer par l'expérience du rapport des blanches aux noires ; M. Bernoulli trouve que pour qu'il ſoit mille fois plus probable qu'il y en a deux noires ſur trois blanches, que non pas toute autre ſuppoſition, il faut avoir tiré de l'urne 25550 boules, & que, pour que cela fut deux mille fois plus probable, il falloit avoir fait 31258 épreuves ; enfin, pour que cela devînt ſept mille fois plus probable, il falloit 36960 tirages. La difficulté & la longueur du calcul ne permettent pas de le rapporter ici en entier, on peut le voir dans l'ouvrage cité.

Par-là il eſt démontré que l'expérience du paſſé eſt un principe de *probabilité* pour l'avenir ; que nous avons lieu d'attendre avec raiſon des évènemens conformes à ceux que nous avons vu arriver fréquemment, & plus nous avons lieu de les attendre de nouveau. Ce principe reçu, on ſent de quelle utilité ſeroient dans les queſtions de phyſique, de politique, & dans ce qui regarde la vie commune, des tables exactes qui fixeroient ſur une longue ſuite d'évènemens la proportion de ceux qui arrivent d'une certaine façon à ceux qui arrivent autrement. Les uſages qu'on a tiré des regiſtres baptiſtaires & mortuaires ſont ſi grands, que cela devroit engager non-ſeulement à les perfectionner, en marquant, par exemple, l'âge, la condition, le tempérament, le genre de mort, *&c.* mais auſſi à en faire de pluſieurs autres évènemens, que l'on dit très-mal-à-propos être l'effet du haſard ; c'eſt ainſi que l'on pourroit former des tables qui marqueroient combien d'incendies arrivent dans un certain tems, combien de maladies épidémiques ſe font ſentir en certains eſpaces de tems, combien de navires, *&c.* ce qui deviendroit très-commode pour réſoudre une infinité de queſtions utiles, & donneroit aux jeunes gens attentifs toute l'expérience des vieillards.

De façon plus précise, on a le théorème suivant :

Loi faible (convergence en probabilité) des grands nombres :

Soit un événement A avec P (A) = p .

Soit X_i , $1 \le i \le n$, 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 $\mathscr{B}(n\ ,\ p)$) et $F = \frac{1}{n} S_n$, la variable aléatoire correspondant à la fréquence d'observation de A sur les n expériences.

Alors, pour tout $t > 0$, $P\left(|F - p| > t\sqrt{\frac{p(1-p)}{n}}\right) \le \frac{1}{t^2}$.

Démonstration :

C'est une application de l'inégalité de ***Bienaymé-Tchebichev*** qui affirme que, pour une variable aléatoire X d'espérance $E(X) = \mu$ et d'écart type $\sigma(X) = \sigma \neq 0$, on a, pour tout $t > 0$,

$P(|X - \mu| > t\sigma) \le \frac{1}{t^2}$.

On peut, dans le cas d'une variable aléatoire prenant un nombre fini de valeurs x_1 , ..., x_n , justifier cette dernière inégalité ainsi :

On a $\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 P(X = x_i) \ge \sum_{j\ /\ |x_j - \mu| > t\sigma}^{n} (x_i - \mu)^2 P(X = x_i)$

et $\sum_{j\ /\ |x_j - \mu| > t\sigma}^{n} (x_i - \mu)^2 P(X = x_i) \ \ge\ t^2\sigma^2 \times \sum_{j\ /\ |x_j - \mu| > t\sigma}^{n} P(X = x_i)$.

D'où $\sigma^2 \ge t^2\ \sigma^2 \times P(|X - \mu| > t\sigma)$.

Remarques :

• La majoration est très grossière et conduit à des valeurs de n très grandes. Pour retrouver l'ordre de grandeur des nombres cités dans l'exemple de Bernoulli, on devra préciser la loi de la variable aléatoire F.

Prenons l'exemple de l'urne de *Bernoulli*, rapporté dans l'Encyclopédie.

On a p = 3/5 et *Bernoulli* recherche, dans "l'*Ars conjectandi*" une valeur de n de sorte que : $P\left(\frac{3}{5} - 0{,}02 \le \frac{S_n}{n} \le \frac{3}{5} + 0{,}02\right) \approx 0{,}999$ où S_n suit la loi $\mathscr{B}$ (n, 3/5). Problème énoncé ici avec des notations modernes (l'expression *"que non pas toute autre supposition"* n'est pas suffisamment claire dans l'article de l'Encyclopédie).

On veut donc $P\left(\left|\frac{S_n}{n} - \frac{3}{5}\right| > 0{,}02\right) \approx 0{,}001$. L'inégalité de *Bienaymé-Tchebichev* donne, en prenant $t = 32$: $P\left(\left|\frac{S_n}{n} - \frac{3}{5}\right| > 32\sqrt{\frac{\frac{3}{5} \times \frac{2}{5}}{n}}\right) \le 0{,}001$.

Ce qui conduit à choisir n tel que $32\sqrt{\frac{\frac{3}{5} \times \frac{2}{5}}{n}} = 0{,}02$ d'où $n = 614400$.

En étudiant soigneusement, dans le développement du binôme, les rapports d'un terme à son précédent, *Bernoulli* parvient à une majoration bien meilleure de n. C'est la valeur

25550 rapportée par *Condorcet*, mais, comme le dit ce dernier, *"la difficulté et la longueur du calcul ne permettent pas de le rapporter ici en entier."*
L'utilisation par *Moivre* de la formule de *Stirling* pour approcher la factorielle nous conduit sur la piste de la loi normale et permettra une meilleur évaluation de n (voir plus loin).

• La loi forte des grands nombres énonce un résultat analogue pour une convergence presque sûre (dite aussi convergence forte), c'est à dire sauf sur un ensemble de probabilité nulle.

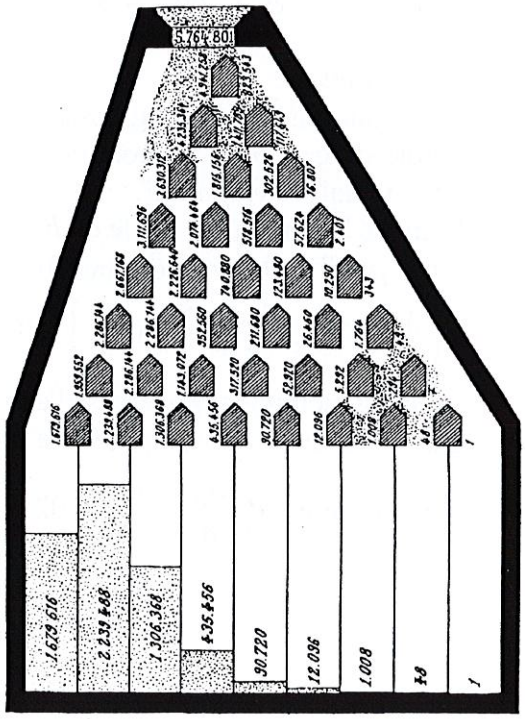
b) Théorème de Moivre-Laplace

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, ***Moivre***, dans la *doctrine des chances*, publiée en 1718, montre l'approximation de la distribution binomiale, pour n grand et dans le cas $p = 1/2$, par la loi "normale".
Pierre Simon Laplace généralise en 1812 ce résultat au cas p quelconque.

> Si les X_i sont des variables de *Bernoulli*, indépendantes et de même paramètre p, pas très voisin de 0 ou de 1, alors
> $S_n = \sum_{i=1}^{i=n} X_i$ suit approximativement, pour n assez grand, la loi normale $\mathcal{N}\left(np\ ,\ \sqrt{np(1-p)}\right)$.

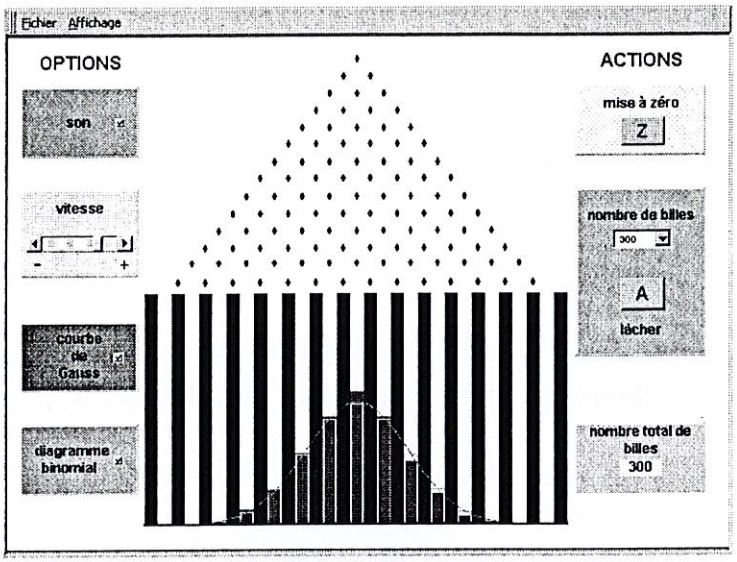
Ce théorème fournit ainsi une approximation d'une loi binomiale par la loi normale de même moyenne et même écart type.

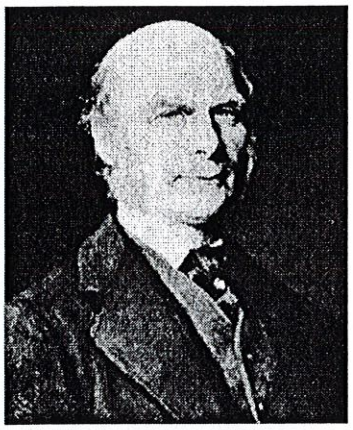
La planche de Galton



La *"planche de Galton"* est une illustration physique (classique... mais spectaculaire) de l'approximation d'une loi binomiale par une loi normale.
Galton, qui n'était pas mathématicien, éprouva le besoin d'imaginer et de faire réaliser des procédés physiques pour comprendre les propriétés de la loi normale.
Ecoutons *Lucien March*, introducteur en France de la statistique mathématique anglaise, nous décrire l'instrument dans son ouvrage *"Les principes de la méthode statistique"*, paru en 1930.
"La concentration des effets de l'association de séries primaires [il s'agit du théorème limite central] *peut être réalisé* ***mécaniquement*** *dans un appareil fort simple, analogue à cet ancien jouet dans lequel une bille descend à travers un quinconce de clous pour finir par se poser dans une case numérotée qui indique le gain du jeu. Cet appareil est figuré ci-après. Il comprend une trémie débouchant au-dessus d'un prisme dont l'arête partage l'orifice de la trémie dans une proportion donnée* [cette proportion est de 6/7 sur la figure], *de sorte que des grains, des grains de sable par*

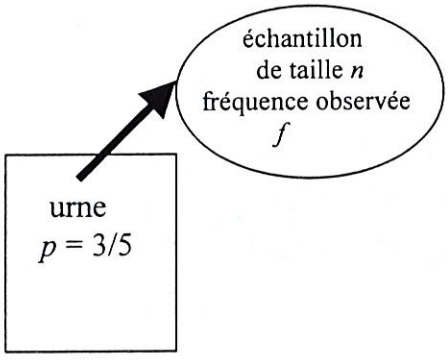
exemple, jetés dans la trémie se répartissent de chaque côté du prisme dans la proportion fixée sur l'autre face du prisme."
Puisqu'il y a 8 rangées de prismes et qu'un grain de sable a, pour chaque prisme rencontré, 6 chances sur 7 d'aller à gauche, la répartition s'effectue selon le modèle de la loi binomiale $\mathscr{B}(8\ ;\ 1/7)$.
En augmentant le nombre de rangées, on se trouvera dans les conditions d'approximation par la loi normale et l'on mettra ainsi en évidence le "profil" de la courbe de *Gauss*.




Francis Galton (1822–1911)

On a réalisé ci-dessus une simulation sous Excel d'une planche de Galton ayant 14 rangées avec $p = 0{,}5$. On peut comparer l'histogramme observé avec, d'une part, l'histogramme selon la loi binomiale et, d'autre part, la densité de la loi normale.

• *Fréquences observées dans le schéma de Bernoulli*



Si les X_i sont des variables de *Bernoulli* de même paramètre p alors, d'après le théorème de *Moivre-Laplace*, la variable aléatoire $F = \frac{1}{n} S_n = \frac{1}{n} \sum_{i=1}^{i=n} X_i$ (fréquence observée sur un échantillon de taille n) suit approximativement, pour n assez grand, la loi normale $\mathscr{N}\left(p\ ,\ \sqrt{\frac{p(1-p)}{n}}\right)$.

Appliquons ce résultat au problème posé par *Bernoulli* et rapporté dans l'*Encyclopédie*.
Si la proportion de boules blanches dans l'urne est $p = \frac{3}{5}$, la variable aléatoire F suit approximativement la loi $\mathscr{N}\left(\frac{3}{5}\ ,\ \frac{\sqrt{6}}{5\sqrt{n}}\right)$.

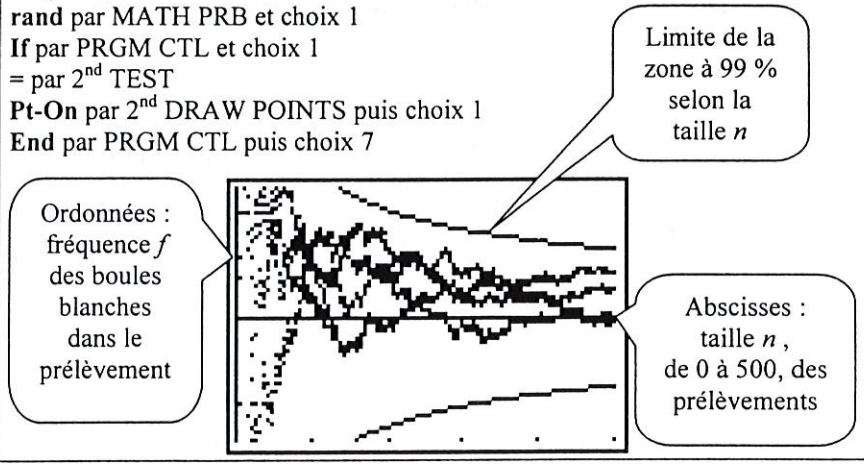
1) Calculons, en fonction de n, le réel positif h tel que $P(\frac{3}{5} - h \leq F \leq \frac{3}{5} + h) = 0{,}99$. On détermine ainsi un intervalle autour de 3/5 dans lequel la variable aléatoire F prend ses valeurs dans 99 % des cas.

On pose $T = \dfrac{F - \frac{3}{5}}{\frac{\sqrt{6}}{5\sqrt{n}}}$ de sorte que T suit la loi normale $\mathcal{N}(0, 1)$.

On a $P(\frac{3}{5} - h \leq F \leq \frac{3}{5} + h) = P\left(\dfrac{\frac{-h}{\sqrt{6}}}{5\sqrt{n}} \leq T \leq \dfrac{\frac{h}{\sqrt{6}}}{5\sqrt{n}}\right)$ or on sait (loi normale standard) que $P(-2{,}58 \leq T \leq 2{,}58) \approx 0{,}99$.

On en déduit que $h = \dfrac{2{,}58\sqrt{6}}{5\sqrt{n}}$.

2) On peut expérimenter, en fonction de la taille n des prélèvements, les fluctuations des fréquences de boules blanches observées en effectuant, sur calculatrice, le programme suivant.
L'écran illustre la "convergence" (presque sûre !) en œuvre dans la loi des grands nombres.

TI 83	Accès aux fonctions
:FnOff :ClrDraw :PlotsOff :0 → Xmin :500 → Xmax :100 → Xscl :0.5 → Ymin :0.7 → Ymax :0.1 → Yscl :DrawF 3/5 :DrawF 3/5+2.58√(6)/(5√(X)) :DrawF 3/5–2.58√(6)/(5√(X)) :For(J ,1 , 4) :0 → P :For (I , 1 , 500) :int(rand+3/5) → A :If A = 1 :P + 1 → P :Pt-On (I , P / I) :End :End	**FnOff** par VARS Y-VARS On/Off puis choix 2 **ClrDraw** par 2nd DRAW DRAW puis choix 1 **PlotsOff** par 2nd STATPLOT puis PLOTS et choix 4 **Xmin Xmax Xscl**... par VARS Window... **DrawF** par 2nd DRAW puis DRAW et choix 6 **For** par PRGM CTL et choix 4 **int** par MATH NUM et choix 5 **rand** par MATH PRB et choix 1 **If** par PRGM CTL et choix 1 = par 2nd TEST **Pt-On** par 2nd DRAW POINTS puis choix 1 **End** par PRGM CTL puis choix 7 

- Si p est inconnu et que la question est de savoir si $p = \frac{3}{5}$, on est dans la situation d'un ***test d'hypothèse*** (situation décrite dans l'article de l'*Encyclopédie*) et les limites sont alors celles de la *zone de rejet* de l'hypothèse $p = \frac{3}{5}$, selon la taille n de l'échantillon.
- Si p est "totalement inconnu" et que l'on cherche à l'estimer indépendamment de toute référence, on a la situation (plus délicate) de l'***estimation***. On parle alors d'***intervalles de confiance*** dont les bornes fluctuent en permanence, selon chaque échantillon et chaque valeur de n.
(Ces notions seront abordées dans les séances ultérieures de ce stage)

3) Déterminons n (pour comparer aux résultats de *Bernoulli*, cités par *Condorcet*) de sorte que : $P\left(\left|F-\frac{3}{5}\right|>0{,}02\right)=\frac{1}{1000}$.

On veut $P\left(|T| \le 0{,}02\frac{5\sqrt{n}}{\sqrt{6}}\right)=\frac{999}{1000}$. On recherche donc, pour la loi $\mathcal{N}(0, 1)$ la valeur de t telle que $2\,\Pi(t)-1=\frac{999}{1000}$ c'est à dire $t=\Pi^{-1}(\frac{1999}{2000})$ où Π est la fonction de répartition de la loi normale centrée réduite (tabulée). La table, ou la calculatrice, donne $t \approx 3{,}29$.

On a donc $0{,}02\frac{5\sqrt{n}}{\sqrt{6}} \approx 3{,}29$ d'où $n \approx 6495$ (à comparer avec 25550 obtenu par *Bernoulli*).

c) Théorème limite central

Laplace va plus loin que *Gauss* en montrant, en 1810, que sa *"seconde loi des erreurs"*[6] (la loi "normale") approche la distribution des moyennes arithmétiques de n erreurs indépendantes de même loi. Avec *Laplace*, la loi normale s'impose comme presque universelle, puisque, 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 statistique du XIX^{ème} siècle.
La dénomination de *"loi normale"* est 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"*.

Théorème limite central (TLC) :
Soit X_i des variables aléatoires indépendantes, de même loi, de moyenne μ et d'écart type σ. Pour n suffisamment grand, la variable aléatoire
$\overline{X}_n=\frac{1}{n}S_n=\frac{1}{n}\sum_{i=1}^{i=n}X_i$ *suit approximativement la loi normale* $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

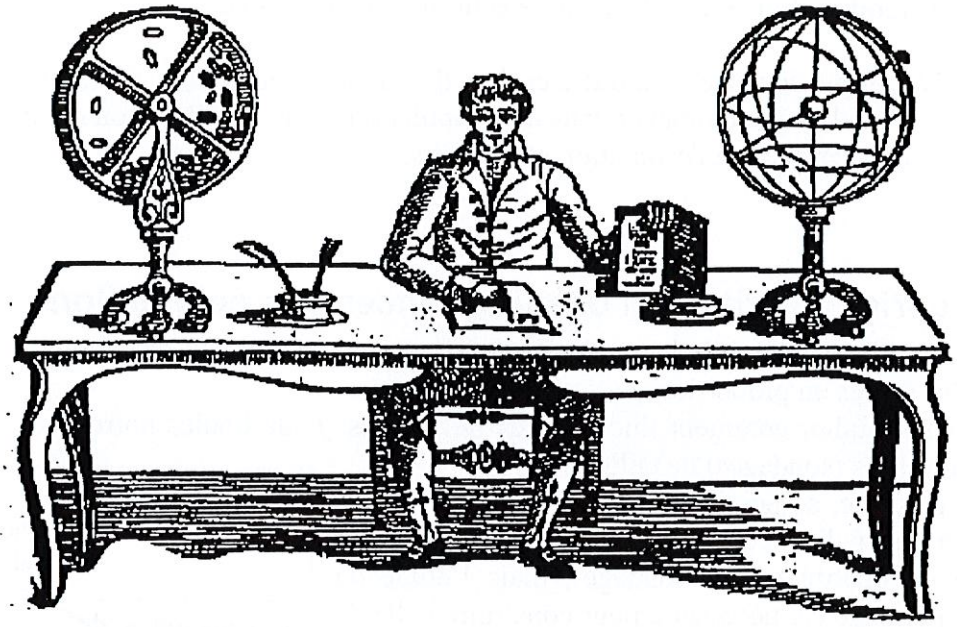
De façon plus précise, la suite de variables aléatoires $\left(\frac{\overline{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}}\right)_n$ ***converge en loi*** vers la loi $\mathcal{N}$ (0, 1) (c. à d. convergence simple des fonctions de répartitions vers celle de la loi normale centrée réduite). Ce théorème est, pour cette raison, aussi dénommé "*théorème de la limite centrée*".

L'étude statistique de la ***variabilité***, qui sera l'objet des séances suivantes, s'appuiera sur le théorème limite central.

---

[6] *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.

Séance 2 : FLUCTUATIONS - SONDAGES ET VARIABILITE



I – FLUCTUATIONS D'ECHANTILLONNAGE D'UNE FREQUENCE

"L'esprit statistique naît lorsque l'on prend conscience de l'existence de fluctuations d'échantillonnage." Document d'accompagnement du programme 2000 de seconde.

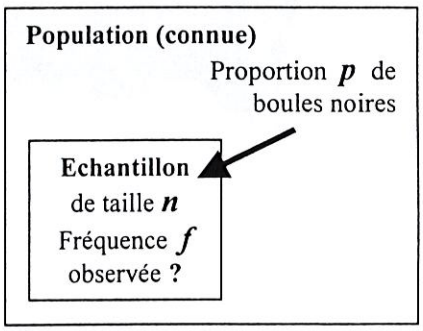
La citation précédente a pour but de montrer l'importance de l'étude, en classe de seconde, des fluctuations d'échantillonnage, pour la compréhension de l'approche statistique des situations variables. Elle est cependant un peu mystérieuse. Qu'est-ce que cet "esprit statistique" qui naîtrait de cette prise de conscience ?
Plutôt que d'esprit statistique, parlons de méthode statistique. Nombre de méthodes statistiques, comme l'estimation par intervalle de confiance ou la théorie des tests d'hypothèses, sont fondées sur l'étude de la variabilité d'un phénomène. Il s'agit souvent de distinguer entre variations "normales", habituelles, dues au hasard, et variations "significatives", signes d'autres causes. Prenons l'exemple du pile ou face. Sur 100 lancers,

il sera bien rare d'observer autant de pile que de face, même si la pièce est parfaitement équilibrée. La fréquence des piles fluctue. Cependant, on peut penser qu'un écart "significatif" (en un sens à préciser) entre la fréquence des piles et celle des faces pourra faire penser que la pièce est truquée.
Le premier objectif est de prendre conscience que les fluctuations d'une fréquence, par exemple la fréquence des piles sur n lancers, entres différents échantillons, dépend de la taille n de l'échantillon. Lorsque n est trop petit (par exemple $n = 10$ lancers) la fréquence fluctue trop pour que l'on puisse se faire une opinion au vu d'une expérience. C'est ainsi que dans l'exemple du classement des hôpitaux de la séance 1, le nombre $n = 12$ opérations chirurgicales est insuffisant pour distinguer sérieusement la qualité des différents établissements.
Ce n'est pas une notion évidente que lorsque la taille des échantillons augmente, la variabilité diminue. Il s'agit, en seconde, de l'expérimenter (cela ne sera théorisé, par la loi des grands nombres, ou le théorème limite central, que plus tard).

L'***échantillonnage*** consiste à modéliser les fluctuations observées entre les différents échantillons que l'on peut prélever dans une population. Sa compréhension est un préalable aux théories statistiques de l'estimation et des tests.

1 – Théorie : distribution des fréquences des échantillons

Un modèle utile est celui de l'urne de *Bernoulli* : une urne contient des boules blanches et des boules noires en proportion p.
On souhaite étudier comment fluctuent les fréquences f de boules noires, observées sur des échantillons (sondages) de taille n.
Dans la pratique, se pose bien sûr le problème inverse (inférer à partir d'un échantillon, par exemple pour un contrôle de qualité ou un sondage), mais l'étude de l'échantillonnage est nécessaire pour construire celle de l'estimation.



Rappelons le ***théorème de Moivre-Laplace*** (cas particulier du théorème limite central).

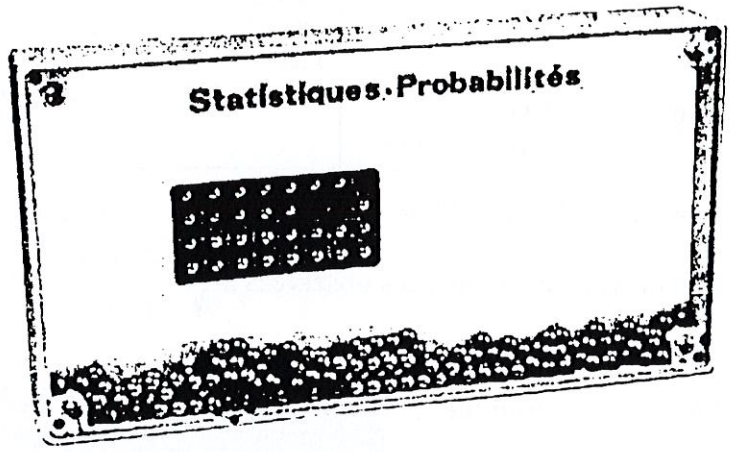
On effectue n tirages ***avec remise*** dans une urne contenant des boules blanches et noires.
Soit X_i la variable aléatoire prenant la valeur 1 si la $i^{ème}$ boule tirée est noire, et 0, si elle est blanche. Les X_i sont des variables indépendantes de *Bernoulli* de même paramètre p. Un calcul simple montre que l'espérance vaut $E(X_i) = p$ et l'écart type $\sigma(X_i) = \sqrt{p(1-p)}$.

La variable aléatoire $S_n = \sum_{i=1}^{n} X_i$ (nombre de boules noires sur un échantillon de taille n) suit alors la loi nommée binomiale, de paramètres n et p : $\mathcal{B}(n, p)$. On a $E(S_n) = np$ et $\sigma(S_n) = \sqrt{np(1-p)}$.
Le théorème de *Moivre-Laplace* affirme que cette loi binomiale est proche, pour n assez grand, de la loi normale de même espérance et écart type : $\mathcal{N}(np, \sqrt{np(1-p)})$ (que l'on se souvienne de la planche de *Galton* qui matérialise cette approximation).
En posant $F = \frac{1}{n} S_n$ (variable aléatoire correspondant à la fréquence observée des boules noires après n tirages), on obtient le résultat important suivant.

La variable aléatoire $F = \frac{1}{n}\sum_{i=1}^{i=n} X_i$ (fréquence observée sur un échantillon de taille n) suit approximativement, pour n assez grand, la loi $\mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

2 – Expérimentation

Expérience physique



L'appareil ci-contre permet de réaliser physiquement des échantillons et d'étudier la distribution d'échantillonnage des fréquences.
Il contient 200 billes dont 10 billes noires.
On a donc $p = 0{,}05$.
En secouant, on vient remplir les $n = 32$ alvéoles qui matérialisent un échantillon. On observe alors une fréquence f de billes noires.

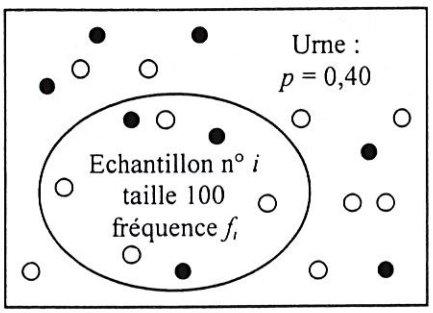
Il s'agit bien sûr d'un échantillonnage ***exhaustif*** (***sans remise***), ce qui complique un peu la théorie précédente.
La variable aléatoire X qui, à tout échantillon de taille $n = 32$ prélevé sans remise parmi les $N = 200$ billes, associe le nombre de billes noires contenues dans l'échantillon, suit la loi ***hypergéométrique*** $\mathcal{H}(N, n, p)$,

avec $P(X = k) = \frac{C_{Np}^{k} C_{N-Np}^{n-k}}{C_N^n}$, $E(X) = np = \mathbf{1{,}6}$ et $\sigma(X) = \sqrt{\frac{N-n}{N-1}\, np(1-p)} \approx \mathbf{1{,}13}$

(l'écart type avec la loi binomiale serait d'environ 1,23).

Avec une calculatrice ou l'ordinateur

La calculatrice (ou l'ordinateur) permet de simuler des tirages avec remise dans une urne, de façon simple, et répétable un grand nombre de fois.



On considère par exemple une urne qui contient 40% de boules noires ($p = \mathbf{0{,}40}$) et 60% de boules blanches.

Simulation d'un tirage sur calculatrice :
En mode "normal" :

Sur CASIO : Int(Ran# + 0.4)
⇨ Sur CASIO : ***Int*** s'obtient par OPTN puis NUM, et ***Ran#*** par OPTN puis PRB.

Sur T.I. : int(rand + 0.4) ou int(rand() + 0.4) sur TI89 et TI92.
⇨ Sur T.I. : ***int*** s'obtient par CATALOG ou MATH puis NUM, et ***rand*** par MATH puis PRB.

L'obtention du chiffre 1 signifiera "boule noire", celle du chiffre 0 signifiera "boule blanche".

Simulation d'un échantillon de taille n = 100, prélevé avec remise :

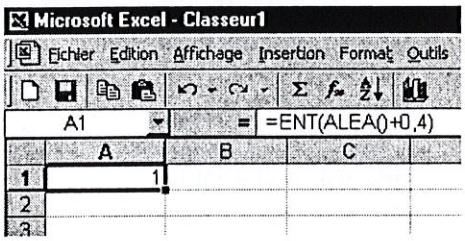
Commentaires	CASIO (anciens modèles sans l'instruction For)	CASIO Graph 25 à Graph 100	T.I. 81	T.I. 80 -82 - 83 - 85	T.I. 89 - 92
S compteur des boules noires. I compteur des 100 tirages 1 = boule noire ; 0 = boule blanche. Affichage de la fréquence f de l'échantillon.	0 → S ↵ 1 → I ↵ Lbl 1 ↵ Int (Ran# + 0.4) + S → S ↵ I + 1 → I ↵ I ≤ 100 ⇒ Goto 1 ↵ S ÷ 100	0 → S ↵ For 1 → I To 100 ↵ Int (Ran# + 0.4) + S → S ↵ Next ↵ S ÷ 100	:0 → S :1 → I :Lbl 1 :int (rand + 0.4) + S → S :I + 1 → I :If I ≤ 100 :Goto 1 :Disp S ÷ 100	:0 → S :For (I,1,100) :int (rand + 0.4) + S → S :End :Disp S ÷ 100	:0 → s :For i,1,100 :int (rand() + 0.4) + s → s :EndFor :Disp s ÷ 100

On peut alors simuler une vingtaine d'échantillons puis comparer la moyenne des fréquences observées à **0,40** et l'écart type des fréquences observées à $\sqrt{\frac{0,4\times 0,6}{100}} \approx$ **0,049**.

Simulation sur Excel :

Il suffit d'entrer dans la cellule A1 la formule : =ENT(ALEA() + 0,4)

En amenant le pointeur de la souris dans le coin inférieur droit, celui-ci se transforme en une croix noire. Il suffit alors de recopier vers la droite, jusqu'en CV1 pour constituer un échantillon de taille 100.



En CW1 on peut calculer la fréquence f observée sur l'échantillon, en entrant la formule : =SOMME(A1:CV1)/100

CW1 = =SOMME(A1:CV1)/100

	CS	CT	CU	CV	CW
1	0	1	0	0	0,38
2	0	1	0	0	0,35
3	0	1	0	1	0,42
4	0	0	0	0	0,43
5	1	1	0	0	0,42
6					0,28
7					0,46
8					0,35
9					0,45
10					0,54
11				0	0,52
12	1	0		0	0,33
13	1	0	1	1	0,4
14	0	0	1		0,35
15	0	0	0		0,32
16	1	1	0	1	0,31
17	1	1	0	1	0,36
18	0	1	0	0	0,49
19	1	0	0	1	0,43
20	0	0	1	0	0,34
21	1	0	0	0	0,34
22	0	1	1	0	0,41
23	0	1	0	1	0,51
24	0	0	0	1	0,41
25	1	1	0	0	0,37

Dans la colonne CW : fluctuation des fréquences sur des échantillons de taille 100

En sélectionnant la première ligne, puis en la recopiant vers le bas jusqu'à la ligne 1000 (par exemple), on visualise les fréquences observées sur 1000 échantillons de taille 100.

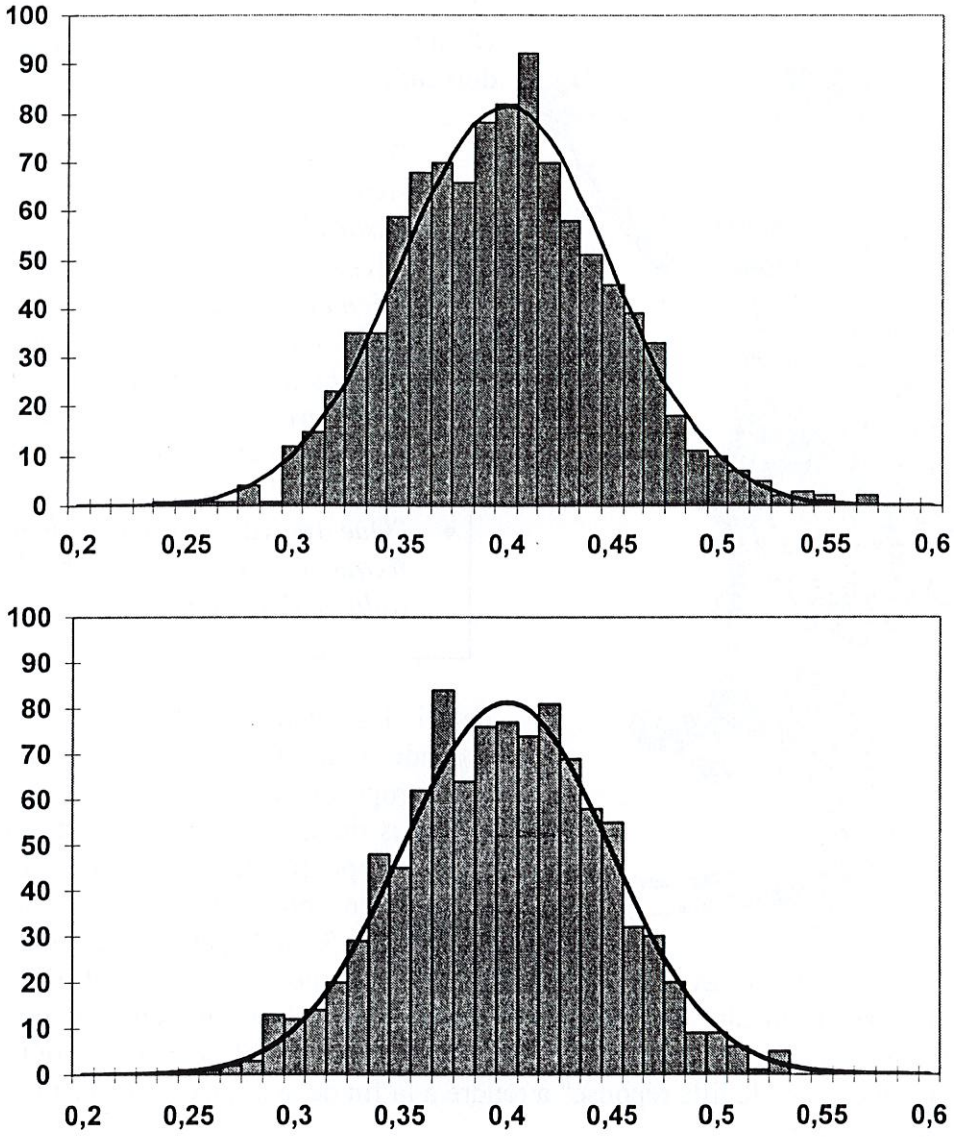
On peut visualiser cette distribution des fréquences en demandant un histogramme.

La théorie prévoit que les fréquences observées se distribuent approximativement selon la loi normale de moyenne 0,40 et d'écart type 0,049.

On peut superposer la courbe de *Gauss* à l'histogramme des fréquences observées sur les 1000 échantillons de taille 100, de façon à les comparer.

Il suffit d'appuyer sur la touche F9 pour simuler 1000 autres échantillons.

Voici deux exemples :



On pourra ainsi considérer que les variations "normales" de la fréquence observée sur les échantillons de taille 100 se situent ici entre 0,3 et 0,5 , alors qu'observer 60% de boules noires est pratiquement impossible sur un échantillon de taille 100 prélevé dans une urne contenant 40% de boules noires.
De façon générale, le calcul sur la loi normale montre que 95% des réalisations se situent entre la moyenne plus ou moins 1,96 fois l'écart type : ici 0,40 +/ – 1,96×0,049 c'est à dire entre 0,30 et 0,50.

L'indicateur des fluctuations est donc l'écart type $\frac{\sqrt{p(1-p)}}{\sqrt{n}}$ pour un échantillon de taille n. On peut dire que, lorsque la taille de l'échantillon augmente, les fluctuations diminuent en $\frac{1}{\sqrt{n}}$. C'est la prise de conscience de ce phénomène, par l'observation et la simulation, qui est l'objet du programme de seconde dans le domaine de l'aléatoire.

3 – ACTIVITES D'ECHANTILLONNAGE SUR LES THEMES DE SECONDE

D'après le programme de seconde, les élèves doivent être capable de :

- *"Concevoir et mettre en œuvre des simulations simples à partir de chiffres au hasard."*
- *"Avoir compris que le terme de fluctuation d'échantillonnage exprime le fait qu'entre plusieurs séries d'expériences, la distribution des fréquences varie, et que varie avec elle tout ce qui se calcule à partir de cette distribution, notamment la moyenne."*
- *"Que les variations des distributions des fréquences calculées sur des séries de taille n, diminuent lorsque n augmente."*

Parmi les thèmes d'étude du programme de seconde pour atteindre les objectifs ci-dessus, nous proposons ci-après trois activités.
On nous dit qu'il ne s'agit pas de faire un cours mais d'expérimenter. Pour les TP utilisant la calculatrice, nous avons réparti les élèves en groupes de 3 ou 4, pendant les séances en demi-classe, éventuellement avec un thème différent. Chaque groupe rendant un compte-rendu. Ce qui permet ensuite de faire une synthèse en classe entière, sur les principaux enseignement à tirer de l'expérience. Pour les TP sur Excel, les élèves sont à 1 ou 2 par machine, avec une "feuille réponse" à rendre à la fin de la séance (cela les oblige à analyser les résultats à l'écran).

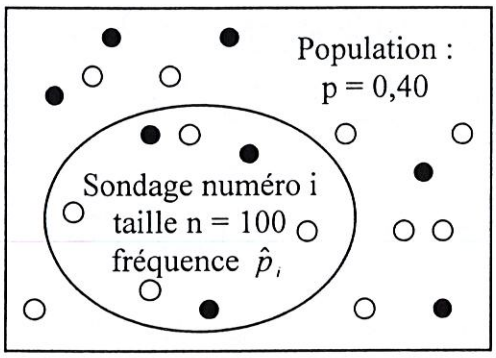
Les activités proposées sont les suivantes :

- T.P. sur Excel : "***Fluctuations des sondages***". Ce TP est nécessaire à l'élaboration d'une formule de fourchette de sondage. Il s'agit de montrer l'impact de la taille de l'échantillon sur les fluctuations des résultats du sondage. Ce TP pourra être suivi d'un autre sur l'estimation par "fourchette".
- T.P. Calculatrice : "***Une martingale***". Etude des fluctuations du temps d'attente du premier pile au jeu de pile ou face dans un jeu où on double la mise à chaque lancer.
- T.P. Calculatrice : "***Le cube et la fourmi***". Un classique des "marches aléatoires" ou l'on étudie les fluctuations de la moyenne, selon le nombre d'expériences effectuées.

T.P. SUR EXCEL EN SECONDE

FLUCTUATIONS DES SONDAGES

FLUCTUATION DES SONDAGES DE TAILLE 100



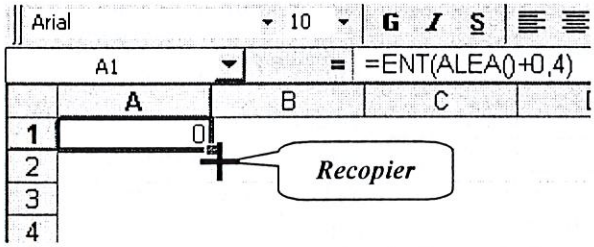
On considère une population importante parmi laquelle 40% (p = 0,40) sont satisfaits de l'action de leur président (ou encore, une urne qui contient 40% de boules noires et 60% de boules blanches).
On effectue des sondages en prélevant au hasard dans cette population des échantillons de taille n = 100 personnes.
Pour chaque sondage i, on calcule la fréquence $\hat{p}_i$ d'opinions favorables.
On désire étudier comment se répartissent les fréquences $\hat{p}_i$ des différents sondages.

Remarque : la population étant supposée importante, on considérera que la proportion p reste la même à chacun des 100 tirages.

Simulation d'un tirage

Lancer Excel®. Dans la cellule A1, taper la ***formule*** : =ENT(ALEA()+0.4) puis faire ***ENTREE***.
Le résultat est 1 si la personne interrogée est favorable au président (fréquence 0,4), et 0 si la personne n'est pas favorable (fréquence 0,6).

a) Simulation d'un premier sondage (100 tirages)



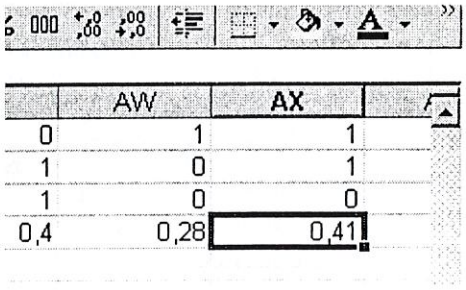
Approcher le pointeur de la souris du carré noir situé dans le coin inférieur droit de la cellule A1. Celui-ci se transforme en une croix noire, faire alors glisser, en maintenant le bouton gauche enfoncé pour ***recopier*** jusqu'en A100, puis relâcher le bouton de la souris.
Pour conserver cette simulation par la suite, cliquer dans le menu ***Outils/Options...*** puis dans l'onglet ***Calcul*** , puis à la rubrique ***Mode de calcul***, choisir • ***Sur ordre*** puis ***OK***.
La fréquence $\hat{p}_1$ d'opinions favorables de votre premier sondage, est la somme des 0 et des 1, divisée par 100. Calculer cette fréquence dans la cellule A 101 en tapant la ***formule*** : =SOMME(A1:A100)/100 puis ***ENTREE***.

✍ Indiquer le résultat sur la feuille réponse.

b) Simulation de 50 sondages de taille 100

	A	B
93	1	
94	0	
95	1	
96	0	
97	1	
98	1	
99	0	
100	1	
101	0,39	
102		

Revenir sur la cellule A1, à l'aide de l'ascenseur. ***Sélectionner*** les cellules de A1 à A101 : pour cela, glisser de A1 à A101 en gardant le bouton gauche de la souris enfoncé. Relâcher : les cellules sélectionnées sont en ***vidéo inversée*** (sur fond coloré).
Approcher le pointeur de la souris du carré noir situé dans le coin inférieur droit de la cellule A101. Celui-ci se transforme en une croix noire, faire alors glisser vers la droite, en maintenant le bouton gauche enfoncé pour ***recopier*** jusqu'en AX101, puis relâcher le bouton de la souris.



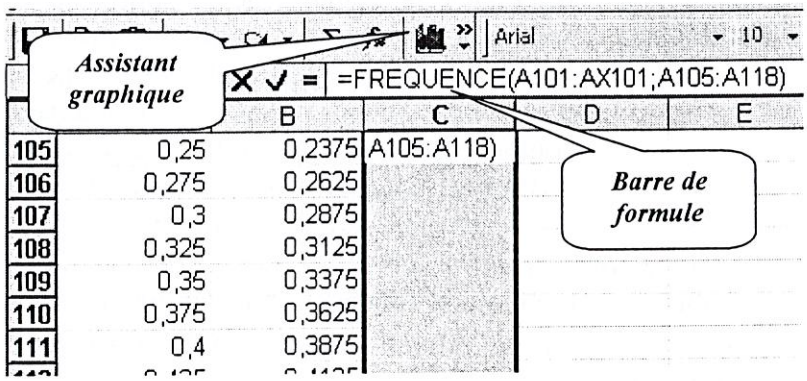
	AW	AX
0	1	1
1	0	1
1	0	0
0,4	0,28	0,41

Appuyer sur la touche ***F9*** pour lancer le calcul de la simulation des nouveaux sondages.

Calculons la fréquence moyenne d'opinions favorables pour les 50 sondages de taille 100.
En A103, taper : moyenne.
En B103, taper la ***formule*** : =SOMME(A101:AX101)/50. Puis faire ***ENTREE***.

Tri des données et histogramme

On va regrouper les données en 14 classes d'amplitude 0,025. Il faut, pour cela, entrer les bornes supérieures de ces classes.
En A105 entrer la valeur 0,25 . En A106, entrer la ***formule*** : =A105+0,025.
Recopier vers le bas jusqu'en A118, puis faire ***F9*** . La cellule A118 devrait ainsi contenir la valeur 0,575.
Indiquons en face les centres des classes : en B105 entrer la formule =A105–0,0125 puis ***recopier*** jusqu'en B118 et faire F9.
Sélectionner les cellules de C105 à C118, puis inscrire dans la ***barre de formules*** :
= FREQUENCE(A101:AX101;A105:A118) puis appuyer *simultanément* sur les touches ***CTRL MAJ ENTREE***.



Excel calcule et affiche les effectifs de chaque classe (et non les fréquences !).
Sélectionner les cellules C105 à C121.
Cliquer sur l'icône d'***assistant graphique*** 
Etape 1/4 :
Choisir ***Histogramme*** puis cliquer sur ***Suivant***.

Etape 2/4 : Cliquer sur l'onglet ***Série*** puis, à la rubrique ***Etiquettes des abscisses (X)***, cliquer sur l'icône de ***Sortie vers la feuille de calcul***  , sélectionner les cellules de puis revenir dans la boîte de dialogue de l'assistant graphique, par l'icône  .
Etape 3/4 : Dans l'onglet ***Légende*** désactiver ***Afficher la légende***. Cliquer sur ***Suivant***.
Etape 4/4 : Choisir ***En tant qu'objet dans Feuil 1*** puis cliquer sur ***Terminer***.
Déplacer et agrandir le graphique.

Dispersion des sondages de taille 100

Vous allez déterminer le pourcentage de sondages fournissant une fréquence comprise entre $0{,}4 - \frac{1}{\sqrt{n}}$ et $0{,}4 + \frac{1}{\sqrt{n}}$, c'est à dire, pour $n = 100$, entre 0,3 et 0,5.

En A120, entrer la formule :
=(NB.SI(A101:AX101;">=0,3")–NB.SI(A101:AX101;">0,5"))*2
puis en B120, taper "% entre 0,3 et 0,5".
Faire F9 pour de nouvelles simulations.

Compléter la feuille réponse.

FLUCTUATION DES SONDAGES DE TAILLE 1000

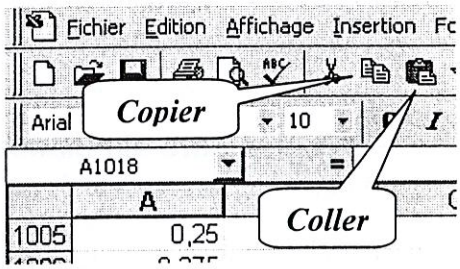
Simulation d'un premier sondage (1000 tirages)

Cliquer sur l'onglet ***Feuil2*** en bas de page.
Dans la cellule A1 de la feuille 2, taper la ***formule*** : =ENT(ALEA()+0.4) puis faire ***ENTREE***.
Recopier la cellule A1 vers le bas jusqu'en A1000 puis faire ***F9***.
En A1001, calculer la fréquence des opinions favorables de ce premier sondage de taille 1000.

Simulation de 50 sondages de taille 1000

Sélectionner les cellules de A1 à A1001, puis ***recopier*** jusqu'en AX1001. Faire ***F9***.

Histogramme et dispersion



Cliquer sur l'onglet ***Feuil1***. Sur la feuille 1, ***sélectionner*** les cellules de A105 à B118 puis cliquer sur l'icône ***Copier***.
Cliquer sur l'onglet ***Feuil2***. Sur la feuille 2, cliquer dans la cellule A1005 puis sur l'icône ***Coller***.
Sélectionner les cellules de C1005 à C1021 puis inscrire dans la ***barre de formule*** :
=FREQUENCE(A1001:AX1001;A1005:A1018)
puis faire simultanément ***CTRL MAJ ENTREE*** .

Cliquer sur l'icône de l'***assistant graphique*** et, en suivant la même procédure que précédemment, représenter l'histogramme des résultats des 50 sondages de taille 1000.

L'intervalle $[0{,}4 - \frac{1}{\sqrt{n}} \ ; \ 0{,}4 + \frac{1}{\sqrt{n}}]$ peut mesurer la précision du sondage en fonction du nombre n de personnes interrogées. Ainsi, pour $n = 1000$ on a l'intervalle [0,37 ; 0,43].
En A1020 taper la ***formule***
=(NB.SI(A1001:AX1001;">=0,37") – NB.SI(A1001:AX1001;">0,43"))*2
puis en B1020, taper : "% entre 0,37 et 0,43".
Faire F9 pour d'autres simulations (attendre le "recalcul").

Compléter la feuille réponse.

✍ FEUILLE REPONSE

NOMS : .....

FLUCTUATION DES SONDAGES DE TAILLE 100

a) Simulation d'un premier sondage

Fréquence des opinions favorables pour le premier sondage : $\hat{p}_1$ = ...

Comparer $\hat{p}_1$ avec la fréquence p de la population : .....

b) Simulation de 50 sondages de taille 100

• Sur une simulation de 50 sondages de taille 100 :

Quelle est la moyenne des 50 fréquences $\hat{p}_i$ des sondages : .....

Comparer avec la fréquence p de la population : .....

.....

Répartition par classe des fréquences $\hat{p}_i$:

Classes	$\hat{p}_i \leq 0,25$	]0,25;0,275]	]0,275;0,3]	]0,3;0,325]	]0,325;0,35]	]0,35;0,375]	]0,375;0,4]
Effectifs							

Classes	]0,4;0,425]	]0,425;0,45]	]0,45;0,475]	]0,475;0,5]	]0,5;0,525]	]0,525;0,55]	$\hat{p}_i > 0,55$
Effectifs							

Quel est le pourcentage des sondages donnant une fréquence $\hat{p}_i$ comprise entre 0,3 et 0,5 :

.....

• Sur plusieurs simulations de 50 sondages de taille 100 :

Indiquer le pourcentage de sondages donnant une fréquence comprise entre 0,3 et 0,5.

Simulation n°	1	2	3	4	5
%					

Conclure : .....

.....

FLUCTUATION DES SONDAGES DE TAILLE 1000

• Sur une simulation de 50 sondages de taille 1000 :

Répartition par classe des fréquences $\hat{p}_i$:

Classes	$\hat{p}_i \leq 0,25$	]0,25;0,275]	]0,275;0,3]	]0,3;0,325]	]0,325;0,35]	]0,35;0,375]	]0,375;0,4]
Effectifs							

Classes	]0,4;0,425]	]0,425;0,45]	]0,45;0,475]	]0,475;0,5]	]0,5;0,525]	]0,525;0,55]	$\hat{p}_i > 0,55$
Effectifs							

Quel est le pourcentage des sondages donnant une fréquence $\hat{p}_i$ comprise entre 0,375 et 0,425 ?

.....

Comparer les histogrammes de répartition des sondages de taille 100 et des sondages de taille 1000 : .....

• Sur plusieurs simulations de 50 sondages de taille 1000 :

Indiquer le pourcentage de sondages donnant une fréquence comprise entre 0,3 et 0,5.

Simulation n°	1	2	3	4	5
%					

Conclure : .....

.....

Quel est l'intervalle $[0{,}4 - \frac{1}{\sqrt{n}} \; ; \; 0{,}4 + \frac{1}{\sqrt{n}}]$ pour $n = 10000$?

.....

.....

A votre avis, pourquoi la plupart des sondages sont-ils effectués sur la base d'environ 1000 personnes interrogées et pas 10000 ?

.....

.....

Eléments de solution

FLUCTUATION DES SONDAGES DE TAILLE 100

a) Simulation d'un premier sondage

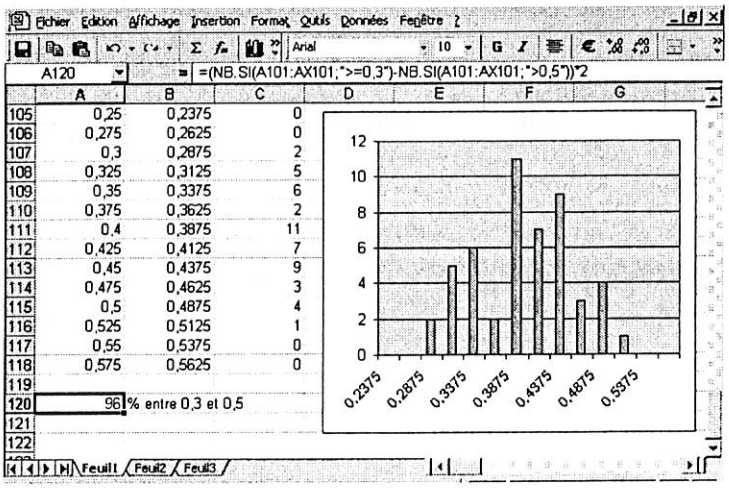
Fréquence des opinions favorables pour le premier sondage : $\hat{p}_1 = 0{,}47$ (par exemple).

Comparer $\hat{p}_1$ avec la fréquence p de la population : $\hat{p}_1 > p$.

b) Simulation de 50 sondages de taille 100

La moyenne des 50 fréquences $\hat{p}_i$ des sondages est proche de la fréquence p de la population.

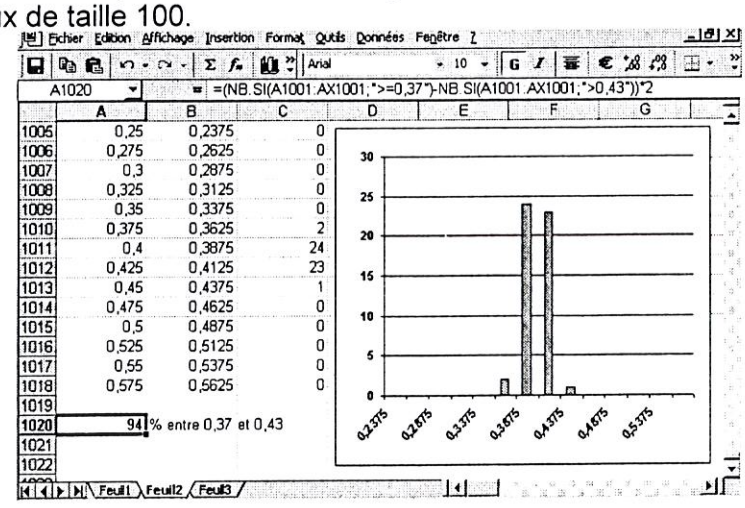
En faisant F9, on peut faire de nombreuses simulations. On constate, en moyenne, que 95 % des sondages donnent une fréquence comprise entre 0,3 et 0,5.



FLUCTUATION DES SONDAGES DE TAILLE 1000

Simulation de 50 sondages de taille 1000.

Les histogrammes de répartition des sondages de taille 1000 sont beaucoup moins dispersés que ceux de taille 100.



En faisant F9, on peut faire de nombreuses simulations. On constate, en moyenne, que 95 % des sondages donnent une fréquence comprise entre 0,37 et 0,43.

L'intervalle analogue pour $n = 10000$ est [0,39 ; 0,41].

Le gain en précision est assez minime par rapport au coût financier d'un tel sondage. C'est pourquoi la plupart des sondages sont effectués sur la base d'environ 1000 personnes interrogées.

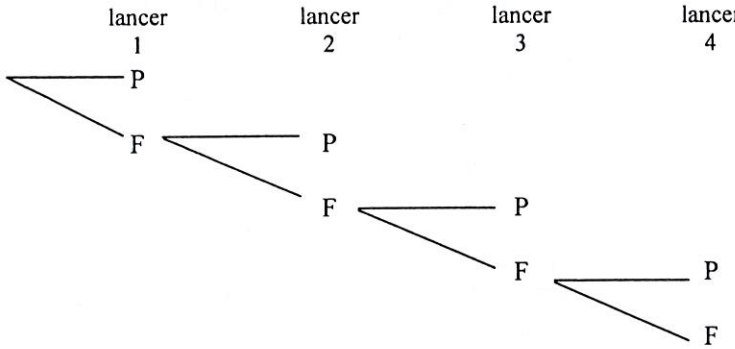
T. P. CALCULATRICES EN SECONDE

UNE MARTINGALE A PILE OU FACE

Règle du jeu : on joue à pile ou face, avec une mise de départ de 10 €. Lorsqu'on perd, la "banque" *conserve la mise*. Si l'on gagne, la "banque" *double la mise*.
Martingale : jouez pile (par exemple). Si vous perdez, rejouez pile et *doublez votre mise jusqu'à gagner*. Avec cette stratégie, on gagne à tous les coups !

Quelques essais, avec une pièce

Jouez quelques parties, en notant le nombre de lancers nécessaires avant de gagner, et combien vous rapporte chaque partie.
Vous est-il arrivé de devoir lancer la pièce plus de 10 fois ? .....
Combien rapportent les parties (gain final moins la mise totale) ? .....
Compléter le tableau suivant, pour chaque partie possible, d'au plus 5 lancers.



nb de lancers	mise sur le tapis	gain final	mise totale	bénéfice
1	10	20	10	10
2	20	40	30	10
3				
4				
5				

On constate que, pour un bénéfice assez modeste, il faut être capable de miser beaucoup, surtout si pile tarde à sortir.
Quelle est la mise totale si pile n'apparaît qu'au $10^{ème}$ lancer ? .....
.....
Quel est le bénéfice ? .....
On va étudier la question suivante : le risque d'être "ruiné" avant de pouvoir gagner est-il négligeable ? Pour cela, vous allez simuler un grand nombre de parties.

Simulation des temps d'attente de pile

Ce premier programme simule une partie et affiche le nombre de lancers nécessaires avant l'apparition de pile (*temps d'attente de pile*).

CASIO Graph 25 → 100	TI 80 (sans While)	TI 82 - 83	TI 89 - 92
1 → N↵ While Int(Ran# + 0.5) = 0 ↵ N + 1 → N↵ WhileEnd↵ N	:0 → N :Lbl 1 :N + 1 → N :If int(rand + 0.5) = 0 :Goto 1 :End :Disp N	:1 → N :While int(rand + 0.5) = 0 :N + 1 → N :End :Disp N	:1 → n :While int(rand() + 0.5) = 0 :n + 1 → n :EndWhile :Disp n

⇨ **Pour obtenir certaines instructions :**
• CASIO Graph 25 → 100 : **While et WhileEnd** par PRGM COM; **Int** par OPTN NUM ; **Ran#** par OPTN PROB ; = par PRGM REL.
• TI 80 82 83 : **While et End ou EndWhile** par PRGM CTL ; **int** par MATH NUM ; **rand** par MATH PRB ; **Disp** par PRGM I/O.

Observer, sur quelques parties, le temps d'attente de pile.
Quel a été le temps le plus long ?

.....

Ce second programme permet d'obtenir, sur 100 parties, la répartition des fréquences des différents temps d'attente de pile.

CASIO Graph 25(*) → 100	TI 80	TI 82 - 83	TI 89 - 92
Seq(I,I,1,15,1) → List 1↵ Seq(0,I,1,15,1) → List 2↵ For 1 → I To **100**↵ 1 → N↵ While Int(Ran# + 0.5) = 0↵ N + 1 → N↵ WhileEnd↵ List 2[N] + 1 → List 2 [N]↵ Next↵ List 2 ÷ **100** → List 2↵ S-WindMan↵ ViewWindow 1,15,1,0, 0.7,0.1↵ 0 → HStart↵ 1 → Hpitch↵ S-Gph1 DrawOn , Hist , List 1 , List 2 , Blue↵ DrawStat ⫽ List 2	:seq(I,I,1,15,1) → L_1 :seq(0,I,1,15,1) → L_2 :For(I,1,**100**) :0 → N :Lbl 1 :N + 1 → N :If int(rand + 0.5) = 0 :Goto 1 :End :L_2(N) + 1 → L_2(N) :End :L_2/**100** → L_2 :Plot 1 (Histogram, L_1,L_2) :PlotsOn 1 :1 → Xmin :15 → Xmax :1 → Xscl :0 → Ymin :0.7 → Ymax :0.1 → Yscl :Dispgraph :Pause :Disp L_1 , L_2	:seq(I,I,1,15,1) → L_1 :seq(0,I,1,15,1) → L_2 :For(I,1,**100**) :1 → N :While int(rand + 0.5) = 0 :N + 1 → N :End :L_2(N) + 1 → L_2(N) :End :L_2/**100** → L_2 :Plot 1 (Histogram, L_1,L_2) :PlotsOn 1 :1 → Xmin :15 → Xmax :1 → Xscl :0 → Ymin :0.7 → Ymax :0.1 → Yscl :Dispgraph :Pause :Disp L_1 , L_2	:seq(i,i,1,15,1) → L1 :seq(0,i,1,15,1) → L2 :For i,1,**100** :1 → n :While int(rand() + 0.5) = 0 :n + 1 → n :EndWhile :L2[n] + 1 → L2[n] :EndFor :L2/**100** → L2 :1 → xmin :15 → xmax :1 → xscl :0 → ymin :0.7 → ymax :0.1 → yscl :PlotsOn :NewPlot 1,4,L1,,L2,,,,1 :Disp L_1 , L_2

⇨ **Pour obtenir certaines instructions :**
• CASIO Graph 25 → 100 : **Seq** par OPTN LIST ; **List** par OPTN LIST ; **[et]** au clavier ; **S-Wind Man** par SET UP Man ou SHIFT SET UP S-WIN ; **Hstart et pitch** par VARS STAT GRPH ; **S-Graph1** par EXIT F4(MENU) STAT GRPH GPH1 ; **Draw On** par, éventuellement EXIT F4(MENU) STAT DRAW ON ; **Hist** par STAT GRPH ; **Blue** (si couleur !) par STAT COLR ; **DrawStat** par PRGM DISP Stat.
• TI 80 82 83 : **Seq** par 2nd LIST OPS ; **L_1** au clavier par 2nd ; **For End Pause** par PRGM CTL ; **Plot 1(Histogram,L1,L2)** par 2nd STAT PLOT PLOTS puis TYPE ; **Xmin** par VARS Window... ; **PlotsOn** par 2nd STAT PLOT PLOTS ; **Dispgraph** par PRGM I/O.
*** Particularités sur certains modèles :**
CASIO Graph 25 et 30 : Les instructions 0 → Hstart ↵ et 1 → pitch ↵ sont à supprimer.

Faire EXE ou ENTER après l'affichage de l'histogramme, puis pour faire une autre simulation.
Noter les fréquences des temps d'attentes, sur *100 parties*, pour quatre simulations :

temps d'attente	1	2	3	4	5	6	7	8	9	10	11	12
fréquences (1ère simulation)												
fréquences (2nde simulation)												
fréquences (3e simulation)												
fréquences (4e simulation)												

Quel est l'écart maximum entre vos quatre simulations de taille 100, pour les fréquences du temps d'attente égal à 1 ? .....

Modifier le programme en remplaçant, à deux endroits, **100** parties par **1000**.
Effectuer deux simulations (durée du programme sur CASIO Graph 60 : ≈ 1 mn).

temps d'attente	1	2	3	4	5	6	7	8	9	10	11	12
fréquences observées (1ère simulation)												
fréquences observées (2nde simulation)												

Quel est l'écart entre les deux simulations de taille 1000, pour les fréquences du temps d'attente égal à 1 ? .....
Comparer avec les simulations de taille 100. .....
Sur les données de votre première simulation de taille 1000, calculer le temps d'attente moyen du premier pile. .....

Eléments de solution

Quelques essais, avec une pièce

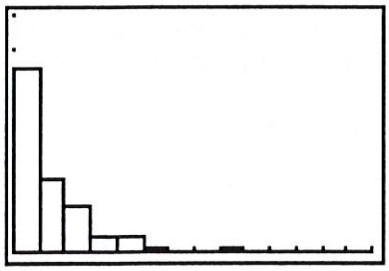
Il s'agit de s'approprier les règles du jeu et de donner du sens aux simulations qui suivent.
On observe très rarement un temps d'attente de pile supérieur à 10.
On constate qu'à chaque fois, certes on gagne, mais seulement 10 €.

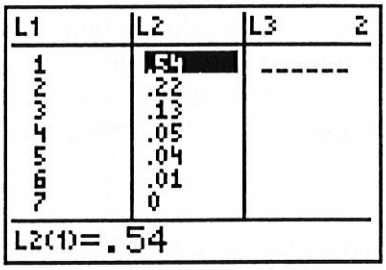
nb de lancers	mise sur le tapis	gain final	mise totale	bénéfice
1	10	20	10	10
2	20	40	30	10
3	40	80	70	10
4	80	160	150	10
5	160	*320*	310	10

Si pile n'apparaît qu'au 10ème lancer, il faudra, pour gagner 10 €, pouvoir miser :
10 +20 +40 + 80 + 160 + 320 + 640 + 1280 + 2560 + 5120 = 10 230 €.

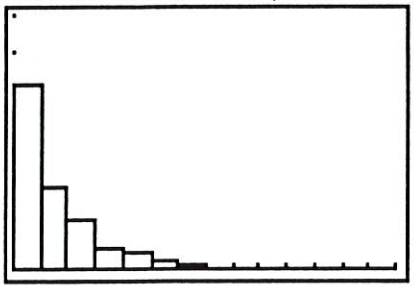
Simulation des temps d'attente de pile

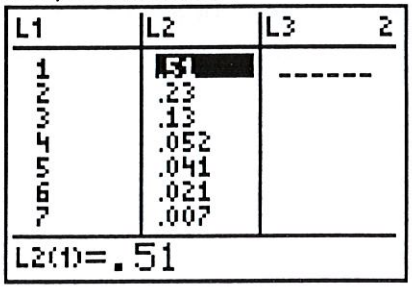
Exemple de distribution, sur 100 parties, des temps d'attentes :





Exemple de distribution, sur 1000 parties, des temps d'attentes :





ENTRE NOUS ...

Si l'on considère la variable aléatoire T qui associe, à chaque partie, le temps d'attente de pile, on a, pour tout $k \geq 1$, $P(T = k) = \left(\frac{1}{2}\right)^{k-1} \times \left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^{k}$.

C'est à dire que T suit la *loi géométrique* de paramètre $\frac{1}{2}$.

k	1	2	3	4	5	6	7	8	9	10	11	12
$P(T=k)$ $\approx$	0,5	0,25	0,125	0,063	0,031	0,016	0,008	0,004	0,002	0,001	0,0005	0,0002

(les hauteurs des rectangles de l'histogramme sont successivement divisées par 2).

La loi géométrique $\mathcal{G}(p)$ ayant pour espérance $\frac{1}{p}$ et pour écart type $\sqrt{\frac{1-p}{p^2}}$, on a :

$E(T) = 2$ et $\sigma(T) = \sqrt{2}$.

Fluctuations d'échantillonnage

On considère une urne, remplie de boules portant les temps d'attentes possibles, selon les proportions théoriques données par la loi géométrique.

• Pour les ***fluctuations de la fréquence*** F du temps d'attente égal à 1, sur des échantillons de taille n, on a $p = 0,5$ et l'on peut considérer que, pour n assez grand, F suit approximativement la loi normale de moyenne 0,5 et d'écart type $\sqrt{\frac{0,5 \times 0,5}{n}} = \frac{0,5}{\sqrt{n}}$.

• Pour les ***fluctuations du temps d'attente moyen*** $\overline{X}$, sur des échantillons de taille n, on a $\mu = E(T) = 2$ et $\sigma = \sigma(T) = \sqrt{2}$ et l'on peut considérer, d'après le T.L.C., que, pour n assez grand, $\overline{X}$ suit approximativement la loi normale de moyenne $\mu = 2$ et d'écart type $\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{2}{n}}$.

A propos du paradoxe de Saint-Petersbourg

Le thème de ce T.P. (au programme de seconde) est issu du ***"paradoxe de Saint Petersbourg"*** posé par *Nicolas Bernoulli* à *Montmort* en 1713 :

Pierre joue contre Paul à pile ou face. Paul s'engage à donner 1€ à Pierre si pile sort et à doubler cette somme tant que pile sort. Quelle doit être la somme que Pierre donne à Paul au début du jeu pour que celui-ci soit équitable ?

Le calcul de l'espérance de la somme que Paul donne à Pierre est :

$1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 2^2 \times \frac{1}{2^3} + \ldots = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \ldots$ soit une somme infinie !

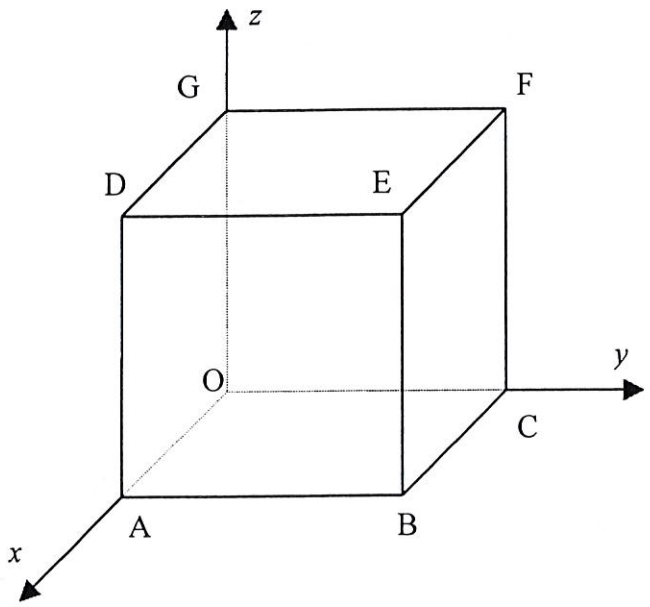
Le calcul de l'espérance mathématique est ici trompeur car il ne tient pas compte des conditions humaines du jeu : sommes d'argent supposées gigantesques mises en jeux, pour des probabilités extrêmement faibles, d'événements humainement non observables.

T. P. CALCULATRICES EN SECONDE

LE CUBE ET LA FOURMI

UNE FABLE CRUELLE

a) Description de l'expérience



Une fourmi se déplace sur les arêtes d'un cube. Elle part du point O. A chaque sommet, elle choisit au hasard l'une des trois arêtes qui s'offrent à elle. Au sommet E l'attend un oiseau, qui la mangera aussitôt.
On désire étudier la durée de vie de la fourmi, sachant qu'elle met une minute pour parcourir une arête.

On rapporte l'espace à un repère orthonormal, de sorte que O est repéré par les coordonnées (0 , 0 , 0) ;
A (1 , 0 , 0) ; C (0 , 1 , 0) et G (0 , 0 , 1).

Donner les coordonnées des points B, D, E et F.

.....

.....

b) Avec un dé

Pour simuler une promenade aléatoire de la fourmi, vous allez utiliser un dé. S'il tombe sur 1 ou 2, vous allez modifier l'abscisse x, s'il tombe sur 3 ou 4, vous modifierez l'ordonnée y et, enfin, pour 5 et 6, vous modifierez la cote z, et ce, jusqu'à atteindre le point E fatal.
Exemple :

dé :	3	1	4	5	6	5	4			
départ O	C	B	A	D	A	D	E			

La durée de vie de la fourmi a été de 7 mn.
A vous (ça peut être assez long) !

dé :										
départ O										

dé :										
point :										

dé :										
point :										

Quelle a été la durée de vie de votre fourmi ? .....

SIMULATION ET DUREE DE VIE MOYENNE

Le programme suivant simule, pour N expériences, la durée de vie moyenne des N fourmis.

CASIO Graph 25 → 100	T.I. 80	T.I. 82 - 83	T.I. 89 - 92
"N"↵ ? → N↵ 0 → S↵ For 1 → I To N↵ 0 → D↵ Seq(0,J,1,3,1) → List 1↵ 0 → T↵ While T ≠ 3↵ 1 + Int(3Ran#) → M↵ 1 − List 1[M] → List 1[M]↵ 1 + D → D↵ Sum(List 1) → T↵ WhileEnd↵ S + D → S↵ Next↵ S ÷ N	:Input N :0 → S :For(I,1,N) :0 → D :seq(0,J,1,3,1) → L_1 :0 → T :Lbl 1 :If T = 3 :Goto 2 :1 + int(3rand) → M :1 − L_1(M) → L_1(M) :1 + D → D :sum(L_1) → T :Goto 1 :Lbl 2 :S + D → S :End :S/N	:Prompt N :0 → S :For(I,1,N) :0 → D :seq(0,J,1,3) → L_1 :0 → T :While T ≠ 3 :1 + int(3rand) → M :1 − L_1(M) → L_1(M) :1 + D → D :sum(L_1) → T :End :S + D → S :End :S/N	:Prompt n :0 → s :For i,1,n :0 → d :seq(0,j,1,3) → L1 :0 → t :While t ≠ 3 :1 + int(3*rand( )) → m :1 − L1[m] → L1[m] :1 + d → d :sum(L1) → t :EndWhile :s + d → s :EndFor :Disp s/n

⇨ **Pour obtenir certaines instructions :**

• CASIO Graph 25 → 100 : " par ALPHA ; **?** par PRGM ; **Seq** List par OPTN LIST ; **For To Next** par PRGM COM ; **Int** par OPTN NUM ; **Ran#** par OPTN PROB ; ≠ par PRGM REL ; [ ] par SHIFT ; **Sum** par OPTN LIST.

• TI 80 → 92 :

Utilisation possible de la fonction **CATALOG** (sur TI 83 – 85 – 92).

Prompt Input par PRGM I/O ; → par STO ▸ ; L_1 au clavier par 2nd ; **seq** par 2nd LIST OPS ; **For** par PRGM CTL ; **int** par MATH NUM ; **rand** par MATH PRB ; **If While Lbl Goto End** par PRGM CTL ; = ou ≠ par 2nd TEST (2nd MATH TEST sur TI 92) ; **sum** par 2nd LIST MATH.

a) Simulation d'une expérience

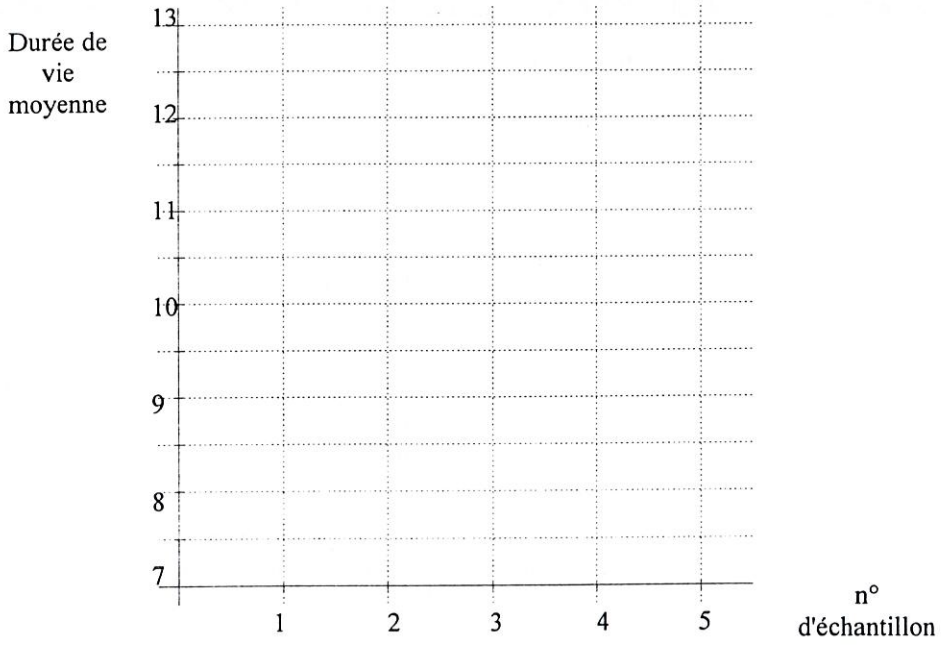
A l'invitation du programme précédent, entrer pour N, la valeur 1. Vous simulez la durée de vie d'une fourmi (expérience faite avec le dé au I). Compléter le tableau suivant.

Durées de vie observées pour une expérience				

b) Durée de vie moyenne pour 10 expériences

A l'invitation du programme précédent, entrer pour N, la valeur 10.
Recommencer cinq fois et compléter le tableau, puis le graphique ci-dessous.

Durées de vie moyennes observées pour 10 expériences				

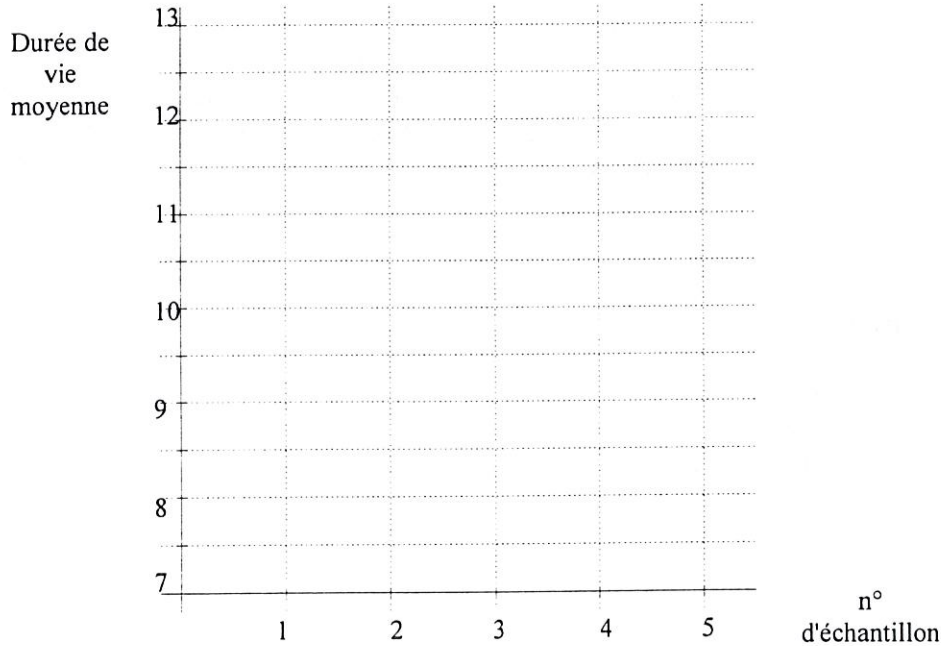


• Quelle est l'étendue de vos 5 valeurs ? .....

c) Durée de vie moyenne pour 50 expériences

A l'invitation du programme précédent, entrer pour N, la valeur 50 (attention la durée de calcul est d'environ 1mn sur TI 83). Recommencer cinq fois et compléter.

Durées de vie moyennes observées pour 50 expériences				



• Quelle est l'étendue de vos 5 valeurs ? .....
• Comparer aux observations obtenues au 2) et donner une explication. .....
.....
• Calculer la moyenne des 5 valeurs. A quoi correspond cette moyenne ? .....

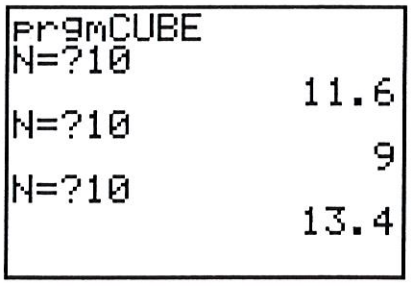
Eléments de solution

Simulation et durée de vie moyenne

a) Exemples de résultats obtenus pour une expérience :

Durées de vie observées pour une expérience				
3	15	5	13	25

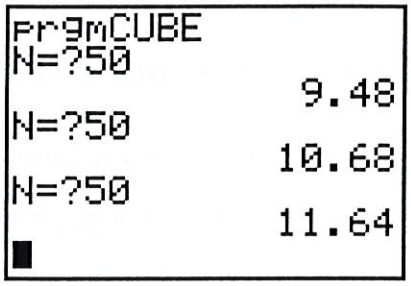
b) Exemples de durées de vie moyennes pour 10 expériences :



Durées de vie observées pour 10 expériences				
11,6	9	13,4	11	12,6

L'étendue est ici de 4,4.

c) Exemples de durées de vie moyennes pour 50 expériences :

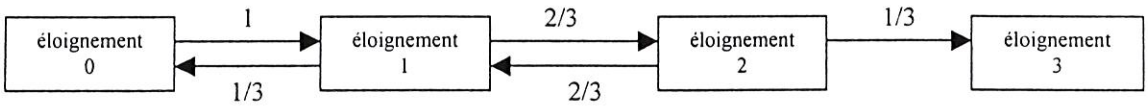


Durées de vie observées pour 50 expériences				
9,48	10,68	11,64	9,6	10,88

L'étendue est ici de 2,16. Elle est inférieure à celle observée sur des échantillons de taille 10, les effets du hasard sont atténués par la plus grande taille d'échantillon. Ce qui apparaît immédiatement sur les graphiques. La moyenne est de 10,456 et correspond à la durée de vie moyenne observée sur 250 fourmis.

ENTRE NOUS ...

Justifions que l'**espérance de vie** d'une fourmi est de **10** mn.
Pour cela, considérons l'éloignement possible d'une fourmi, par rapport à son point de départ. Celui-ci peut être de 0, 1 , 2 ou 3 (fin du voyage). Les probabilités avec lesquelles une fourmi passe d'un niveau d'éloignement à un autre sont résumées par le graphe ci-dessous :



Considérons la variable aléatoire *D* correspondant à la durée de vie d'une fourmi prise au hasard.
Le tableau suivant donne, en fonction du temps *t*, la probabilité d'éloignement de la fourmi par rapport à l'origine.

temps	variable aléatoire "durée de vie restante"	probabilité éloignement 0	probabilité éloignement 1	probabilité éloignement 2	probabilité éloignement 3
$t = 0$	D	1	0	0	0
$t = 1$	$D - 1$	0	1	0	0
$t = 2$	$D - 2$	1/3	0	2/3	0
$t = 3$	$D - 3$	0	7/9	0	2/9

On constate qu'à l'instant $t = 3$, la fourmi a 7 chances sur 9 d'être au même niveau (éloignement 1) qu'à l'instant $t = 1$, et 2 chances sur 9 d'être sur le point d'être mangée (au niveau 3).
En désignant par N la variable aléatoire correspondant au niveau atteint par une fourmi à l'instant $t = 3$, on peut faire le calcul d'espérance conditionnelle suivant :

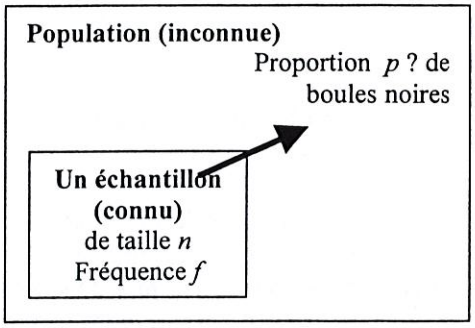
$$E(D-3) = E(D-3 \mid N=1) \times P(N=1) + E(D-3 \mid N=3) \times P(N=3)$$
$$= E(D-1) \times (7/9) + 0 \times (2/9).$$

D'où $E(D) - 3 = (E(D) - 1)(7/9)$ et donc $E(D) = 10$.

Référence : ❑ Engel - Varga - Walser - "Hasard ou stratégie" - Ed. OCDL - 1976.

II – INTERVALLE DE CONFIANCE POUR UNE FREQUENCE

1 – Le problème de l'estimation

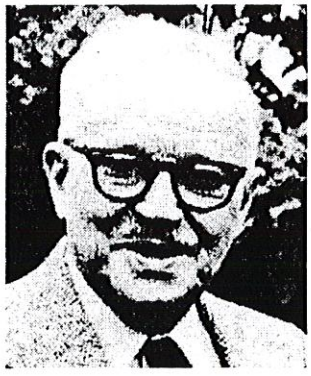


C'est le problème "inverse" de celui de l'échantillonnage, et celui qui se pose dans la pratique des sondages ou des contrôles de qualité : à partir de la fréquence f observée sur ***un*** échantillon, ***estimer*** la fréquence p correspondante, dans la population.
D'un point de vue logique, la situation est très différente de celle des fluctuations d'échantillons.
Alors que dans l'étude de l'échantillonnage la démarche est déductive (on connaît la fréquence p sur la population et on en déduit les variations des fréquences f sur les différents échantillons possibles), ***pour l'estimation, la démarche est inductive*** : de la valeur f obtenue sur un échantillon, on cherche à induire la valeur de la fréquence sur la population (à "inférer").
La démarche de la statistique inférentielle est ainsi inhabituelle en mathématiques. Son objectif est d'apporter une aide à la décision, qui n'a pas la "rigueur" habituelle en mathématiques.
On procède par sondages pour des raisons économiques. On ne peut pas toujours se payer le luxe d'un recensement et il se peut, en particulier, qu'un contrôle de qualité nécessite la destruction des pièces testées (mesures de résistance aux chocs...). C'est donc par nécessité que l'on procède par inférence, faute de mieux... et la résistance à ces pratiques, qui ne se sont imposées que par leur efficacité, fut forte.
Quelques repères de l'histoire de l'estimation :

- **1912 : *Ronald A. FISCHER*** crée les bases de la théorie moderne de l'estimation.
- **1934 – 1937 : *Jerzy NEYMAN*** expose la théorie de l'estimation "*par intervalle*".
- **3 novembre 1936 : *F.D. Roosevelt*** remporte l'élection présidentielle. Le *Literary Digest* avait prédit la victoire de *Landon* sur 2 000 000 de personnes interrogées (par téléphone... d'où un biais) alors que *George GALLUP* avait annoncé celle de Roosevelt sur un échantillon aléatoire réduit.

2 – Intervalle de confiance d'une fréquence

Une notion délicate...

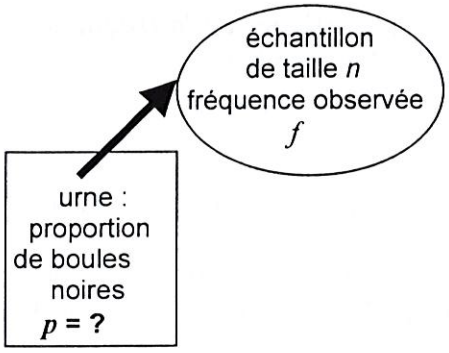


En 1934, ***Jerzy Neyman*** (1894-1981) fit un exposé devant la *Royal Statistical Society* intitulé *"Sur deux différents aspects de la méthode représentative"*. La partie la plus importante de cet exposé est contenue dans un appendice dans lequel *Neyman* propose une méthode nouvelle pour réaliser un intervalle d'estimation. Il intitule cette nouvelle procédure "intervalle de confiance". Comme nous allons le voir, ce n'est pas sans raison que *Neyman* évite d'utiliser le mot probabilité pour nommer sa procédure. Le professeur *Bowley*, présent dans l'assistance, se montre plutôt sceptique, lors de la discussion qui suit l'exposé : *"Je fais allusion aux limites de confiance du Dr. Neyman. Je ne suis pas du tout sûr que la "confiance" ("confidence")*

n'est pas une "escroquerie" ("confidence trick")."[7] Et de poursuivre : *"Les fondements de la théorie ne sont pas convaincants, et tant que je ne suis pas convaincu je douterai de sa validité."* Cette réaction, de la part d'un éminent statisticien en 1934, montre bien que la notion d'intervalle de confiance est délicate et ne va pas de soi. Une mauvaise compréhension conduit à des contre sens. L'expérimentation, par simulation, est sans doute la meilleure façon de comprendre le sens de l'expression "confiance" et d'être convaincu. Pour enseigner cette notion en sections de BTS depuis de nombreuses années, l'expérimentation nous a semblé absolument nécessaire. A plus forte raison sans doute en classe de seconde, si l'on souhaite aborder le thème d'étude (intéressant) sur les fourchettes de sondage.

Un peu de théorie...

On considère une urne où la proportion de boules noires est p.



Rappelons les résultats de l'étude de l'échantillonnage. On introduit la variable aléatoire d'échantillonnage F qui, à tout échantillon de taille n prélevé au hasard avec remise, associe la fréquence f des boules noires contenues dans l'échantillon.

On sait que nF, correspondant au nombre de boules noires dans l'échantillon, suit la loi binomiale $\mathcal{B}(n, p)$ laquelle est proche, lorsque $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$, de la loi normale $\mathcal{N}\left(np, \sqrt{np(1-p)}\right)$ de même espérance et de même écart type.

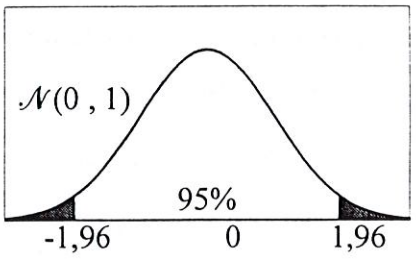
En divisant par n, on considérera donc que F suit approximativement, pour n assez grand, la loi normale $\mathcal{N}\left(p, \sqrt{\dfrac{p(1-p)}{n}}\right)$.

On peut, par exemple, déterminer le réel positif h tel que :
$P(p-h \leq F \leq p+h) = 0{,}95$.
Pour cela, on se ramène à la loi normale centrée réduite, en posant $T = \dfrac{F-p}{\sqrt{\dfrac{p(1-p)}{n}}}$ qui suit la loi normale $\mathcal{N}(0, 1)$.

On trouve alors $h = 1{,}96\sqrt{\dfrac{p(1-p)}{n}}$.

La théorie de l'***échantillonnage*** fournit ainsi, pour chaque valeur de p fixée, un ***intervalle de probabilité*** :

$\left[p - 1{,}96\sqrt{\dfrac{p(1-p)}{n}},\ p + 1{,}96\sqrt{\dfrac{p(1-p)}{n}}\right]$, où la variable aléatoire F prend ses valeurs avec une probabilité de 0,95.



Si T suit la loi $\mathcal{N}(0, 1)$, on a :

$P(\mathbf{-1{,}64} \leq T \leq \mathbf{1{,}64}) = \dfrac{1}{\sqrt{2\pi}} \int_{-1{,}64}^{1{,}64} e^{-\frac{x^2}{2}}\, dx \approx \mathbf{0{,}90}$

$P(\mathbf{-1{,}96} \leq T \leq \mathbf{1{,}96}) \approx \mathbf{0{,}95}$

$P(\mathbf{-2{,}58} \leq T \leq \mathbf{2{,}58}) \approx \mathbf{0{,}99}$.

[7] Rapporté dans le livre de D. Salsburg – *"The lady tasting tea"* : *"I am not at all sure that the "confidence" is not a "confidence trick"."*

Si l'on cherche à ***"retourner"*** l'égalité correspondant à l'intervalle de probabilité d'échantillonnage, on peut écrire l'équivalence :

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

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

La seconde écriture correspond à ce que l'on pourrait appeler un "intervalle dont les bornes sont aléatoires", centré sur les valeurs de F et contenant p avec une probabilité de 0,95 , sauf que l'expression exacte de cet intervalle contient la valeur p à estimer !! Et voici la pirouette, on remplacera, dans l'expression de l'intervalle de confiance, la quantité p inconnue par la valeur f observée. Mais on ne parlera pas de "probabilité" (si, comme le professeur *Bowley*, vous n'êtes pas convaincu, c'est normal, attendez la suite).

De façon générale, on proposera comme ***intervalle de confiance de la fréquence*** p sur la population totale, ***au coefficient de confiance de A %***, l'intervalle :

$$\left[f\ -t_A\sqrt{\frac{f(1-f)}{n}}\ ,\ f\ +t_A\sqrt{\frac{f(1-f)}{n}}\right]$$

où t_A est donné par la table de la loi normale $\mathcal{N}(0\ ,\ 1)$ tel que $2\times\Pi(t_A) - 1 = A\ \%$.

Par exemple, pour $f = 0{,}6$ obtenu sur un sondage de taille $n = 100$ et $A = 95\ \%$, la loi normale donne $t_A = 1{,}96$ et on prend comme intervalle de confiance pour p, l'intervalle $[0{,}504\ ;\ 0{,}696]$, centré sur la valeur observée $f = 0{,}6$.

Si l'on veut parler de probabilité, on peut dire que, sur un grand nombre d'intervalles de confiance (obtenus à partir d'un grand nombre d'échantillons), *environ* (à cause du remplacement de p par f) 95 % contiennent effectivement la valeur de p, ou encore, que l'on a environ 95% de chances d'exhiber un intervalle contenant p (avant le tirage de l'échantillon). ***La probabilité est dans la procédure.***

En observant que, ayant $0 \le p \le 1$, on a $1{,}96\sqrt{p(1-p)} \le 1$, on peut dire que :

$$P\left(F-\sqrt{\frac{1}{n}} \le p \le F+\sqrt{\frac{1}{n}}\right) \ge P\left(F-1{,}96\sqrt{\frac{p(1-p)}{n}} \le p \le F+1{,}96\sqrt{\frac{p(1-p)}{n}}\right) = 0{,}95$$

On alors une probabilité supérieure à 0,95 d'exhiber une ***fourchette de sondage*** contentant la valeur p à estimer.

Le programme de seconde suggère :

> *"On incitera les élèves à connaître l'approximation usuelle de la* ***fourchette*** *au niveau de confiance 0,95, issue d'un sondage sur n individus (n > 30), dans le cas où la proportion observée* $\hat{p}$ *est comprise entre 0,3 et 0,7 , à savoir :* $[\ \hat{p} - 1/\sqrt{n}\ ,\ \hat{p} + 1/\sqrt{n}\ ]$ *".*
> *(Programme 2000 de seconde)*

La condition $n > 30$ et $0{,}3 < \hat{p} < 0{,}7$ assure la validité de l'approximation normale.

De plus, dans ce cas, $1{,}96\times\sqrt{\hat{p}(1-\hat{p})}$ est proche de 1 et majoré par ce nombre : lorsque $0{,}3 < \hat{p} < 0{,}7$, l'étude des variations de $\hat{p}\mapsto 1{,}96\times\sqrt{\hat{p}(1-\hat{p})}$ montre qu'alors $0{,}898 < 1{,}96\times\sqrt{\hat{p}(1-\hat{p})} < 0{,}98$ (on peut toujours majorer $\sqrt{\hat{p}(1-\hat{p})}$ par 1/2).

Ainsi, l'intervalle (de "sécurité") [$\hat{p} - 1/\sqrt{n}$, $\hat{p} + 1/\sqrt{n}$] contient l'intervalle de confiance à 95%, tout en en étant très proche, on le nommera ***"fourchette de sondage de p au niveau 0,95"***.

Quelques remarques encore :

• ***On ne peut pas dire*** que p a 95 % de chances d'appartenir à un intervalle de confiance donné tel que [0,504 ; 0,696]. Cette expression ne contient rien d'aléatoire, p est, ou non, dans cet intervalle, sans que le hasard n'intervienne.

• Dans les cas $n < 30$ ou $np < 5$ ou $n(1-p) < 5$, on utilise la loi binomiale, ou un ***abaque***.

• A propos de la formule de la "fourchette" :

- Le ***gain en précision*** est en $\frac{1}{\sqrt{n}}$, c'est à dire qu'avec 100 fois plus de personnes interrogées, un sondage n'est que 10 fois plus précis.
- La formule ***ne dépend pas de la taille*** (supposée très grande devant n) ***de la population***. Pour une précision donnée, on doit interroger autant de personnes pour sonder toute la population française pour l'élection présidentielle, ou celle d'une ville pour les élections municipales.
- Ces formules ne sont valables que lorsque l'échantillon est tiré "au hasard". Le tirage au sort est encore la meilleure méthode pour obtenir un ***échantillon représentatif***. Cela ne va pas de soit et il y a eu longtemps débat au sein de l'Institut International de la Statistique dès 1895, entre le "choix raisonné" de l'échantillon et le choix aléatoire, au moins jusqu'à l'élection américaine de 1936. La raison humaine est mauvaise conseillère en la matière et introduit trop souvent des biais cachés.
Dans de nombreux sondages cependant, on utilise un échantillonnage "stratifié", selon le sexe, l'age, la catégorie socioprofessionnelle... suivant leurs proportions dans la population. Mais, à l'intérieur de chaque classe, le hasard est encore le meilleur allié pour garantir la représentativité (voir le paragraphe 4 à propos des sondages politiques).

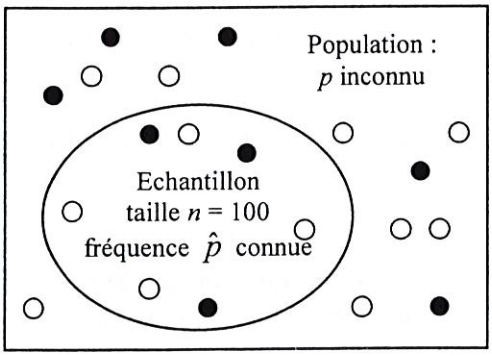
La compréhension de la notion (délicate) d'intervalle de confiance nécessite de se confronter à l'expérience.

3 – EXPERIMENTATION EN SECONDE : ESTIMATION APRES SONDAGE PAR UNE FOURCHETTE

Le T.P. sur Excel qui est présenté ci-après a pour objectif la compréhension, par les élèves, de la notion de coefficient de confiance. Ce fameux 95%, que l'on modifiera dans le TP, de façon à comprendre pourquoi on se prive souvent d'une confiance à 99%. On y expérimente également l'efficacité, toute relative (à 95%), de la procédure. De quoi être convaincu de son intérêt (à 100% ?).

T.P. SUR EXCEL EN SECONDE

ESTIMATION APRES SONDAGE PAR UNE FOURCHETTE

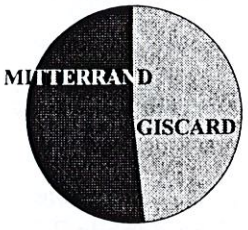


On considère une population importante où une élection oppose deux candidats X et Y. La proportion p des personnes qui voteront pour X est inconnue.
On prélève au hasard dans cette population un échantillon de taille $n = 100$ personnes et on calcule la fréquence $\hat{p}$ d'opinions favorables à X. Il s'agit, à partir de la valeur $\hat{p}$ donnée par le sondage, d'estimer la valeur inconnue de p.

L'observation des fluctuations d'échantillons a montré que, pour tenir compte de la précision plus ou moins grande apportée par la taille de l'échantillon, 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 :

$\left[\hat{p}-1{,}645\sqrt{\frac{\hat{p}(1-\hat{p})}{100}}\ ;\ \hat{p}+1{,}645\sqrt{\frac{\hat{p}(1-\hat{p})}{100}}\right]$, calculé avec la fréquence $\hat{p}$ sur un sondage de taille 100, on a environ 90 chances sur 100 d'obtenir une fourchette contenant la fréquence inconnue p.

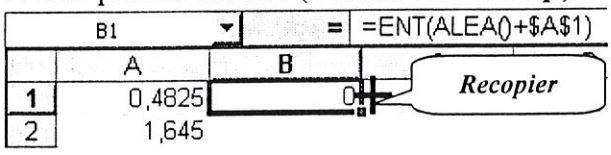
FOURCHETTES A 90% DE CONFIANCE



Travaillons sur un exemple où les scores étaient particulièrement serrés.
Le 10 mai 1981, François Mitterrand a été élu avec 51,75 % des voix, alors que Valery Giscard d'Estaing n'a recueilli que 48,25 % des suffrages.
On suppose que l'on effectue des sondages sur 100 électeurs, le jour de l'élection.

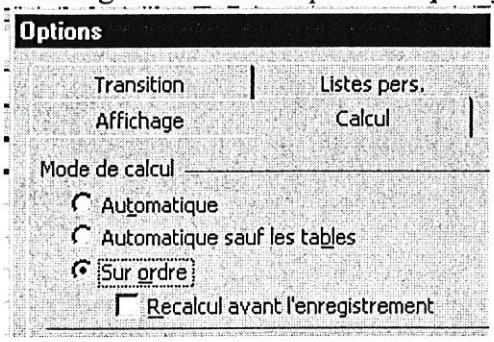
a) Premier sondage

Lancer Excel®. Cliquer (avec le bouton gauche de la souris) dans la cellule A1, taper 0.4825 puis ***ENTREE*** (**c'est la valeur de p**).



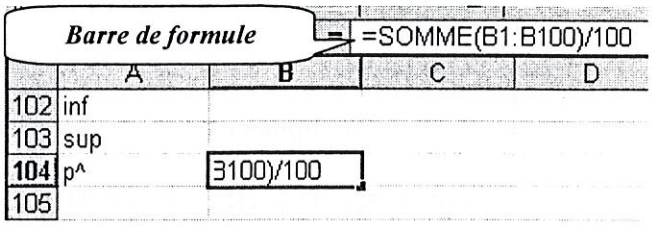
En A2, taper 1,645 (c'est la valeur qui apparaît dans l'expression de la fourchette à 90%).
Dans la cellule B1, entrer la ***formule*** :
=ENT(ALEA()+$A$1) (avec des parenthèses vides après ALEA).
(le symbole $ permet de fixer les références de la cellule A1 lors des opérations de recopie). **Le résultat est 1 si l'électeur est pour Giscard et 0 s'il est pour Mitterrand.**

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 B100, puis relâcher le bouton de la souris.



Vous avez simulé les résultats d'un sondage de 100 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***.

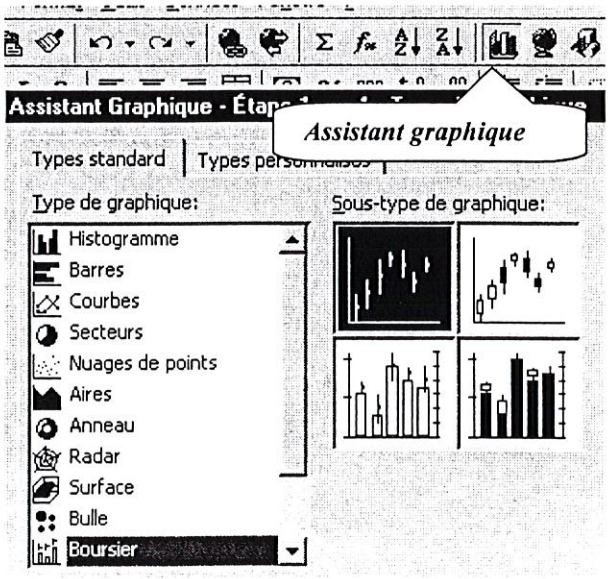
On va calculer la fréquence $\hat{p}$ sur le sondage et une **fourchette** pour p à 90% de confiance.



En A102 taper inf puis, en A103, taper sup et en A104 p^ (on obtient ^ en appuyant simultanément sur ALT GR et ç).
En B104, entrer la ***formule*** :
=SOMME(B1:B100)/100
En B103, entrer la ***formule*** : =B104+$A$2*RACINE(B104*(1-B104)/100)
pour la borne supérieure de la "fourchette".
En B102, entrer la ***formule*** : =B104-$A$2*RACINE(B104*(1-B104)/100)
pour la borne inférieure de la "fourchette".
En B105, entrer la ***formule*** :
=ET($A$1>=B102;$A$1<=B103)
(la formule ET(*condition 1* ; *condition 2*) renvoie la valeur VRAI si les *conditions* sont vérifiées et la valeur FAUX sinon).

b) Visualisation des fourchettes données par 10 sondages

Sélectionner les cellules de B1 à B105 (pour cela cliquer sur B1 et glisser, en gardant le bouton gauche de la souris enfoncé, jusqu'en B105, puis relâcher le bouton de la souris).
Recopier la sélection (en glissant après avoir obtenu la croix noire au coin de B105) jusqu'en K105. Appuyer sur ***F9*** pour lancer le calcul.



Vous avez maintenant 10 sondages de chacun 100 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 B102:K104
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***, décocher la rubrique ***Maximum*** et inscrire la valeur 0,8 (attention à taper une virgule et non un point). Cliquer sur ***OK***.
Appuyer sur la touche ***F9*** pour simuler 10 nouveaux sondages et observer le graphique.

c) Sondages donnant une fourchette ne contenant pas la valeur à estimer

Revenir sur la ***feuille 1***.
En A106 taper "FAUX". En B106, entrer la ***formule*** : =NB.SI(B105:K105;FAUX)
En C106 taper "sur 10" .

FOURCHETTES A 95% OU 99% DE CONFIANCE

a) Fourchettes à 95% de confiance

Sur la ***feuille 1***, remplacer en A2 la valeur 1,645 par 1,96.

b) Fourchettes à 99% de confiance

Sur la ***feuille 1***, remplacer en A2 la valeur 1,96 par 2,58.

ESSAIS D'ESTIMATION DE VALEURS INCONNUES

On considère cette fois une autre élection pour laquelle on cherche à estimer le score d'un candidat.
Sur la ***feuille 1***, inscrire en A2 la valeur 1.96 , puis entrer en A1 la ***formule*** suivante :
=(8*ALEA()+1)/10
Sur le ***graphique***, cliquer avec le bouton *droit* de la souris sur l'***Axe des ordonnées*** puis, dans l'onglet ***Echelle***, régler le ***Maximum*** à 1.
Sur la feuille du graphique, faire ***F9*** et essayer d'estimer la valeur de p grâce aux 10 fourchettes du graphique. Venir contrôler votre essai sur la ***feuille 1*** en considérant la valeur de la cellule A1.
Recommencer en faisant ***F9*** sur la feuille de graphique.

Eléments de solution

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 Mitterrand et 1 pour ceux favorables à Giscard.

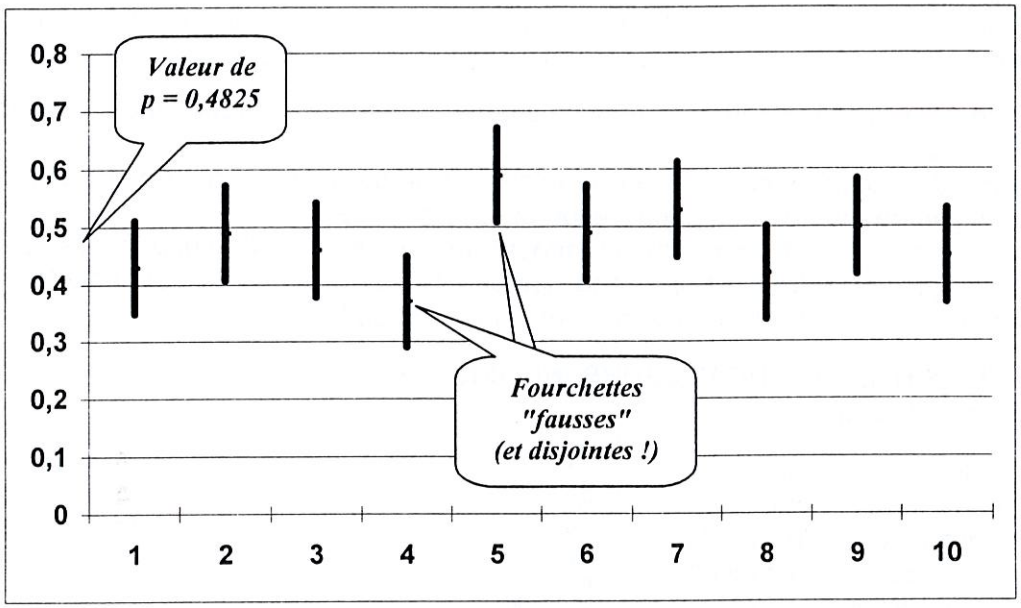
On obtient (par exemple !) $\hat{p} = 0,44$.

La fourchette correspondante pour *p* est alors [0,439 ; 0,522].

La valeur VRAI signifie que la fourchette contient effectivement la valeur p = 0,4825 (c'est le cas ici). La valeur FAUX signifie que le fourchette ne contient pas la valeur p = 0,4825.

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)



Sur cet exemple de 10 sondages, trois donnent une fréquence $\hat{p}$ (centre de la fourchette) en faveur (à tord) de Giscard ($\hat{p} > 0,5$) (on peut contrôler les valeurs sur la feuille de calcul).

Sur cet exemple de 10 sondages, une fourchette prévoit complètement la victoire de Mitterrand (fourchette entièrement située sous la valeur 0,5).

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 = 0,4825 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 0,4825	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 %).

FOURCHETTES A 95% OU 99% DE CONFIANCE

a) Fourchettes à 95% de confiance

Sur le graphique, on observe des "fourchettes" plus longues. Par exemple :

Simulations de 10 sondages	1	2	3	4	5	6	7	8	9	10
Nombres de fourchettes à 95% ne contenant pas 0,4825	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. Par exemple :

Simulations de 10 sondages	1	2	3	4	5	6	7	8	9	10
Nombres de fourchettes à 99% ne contenant pas 0,4825	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.

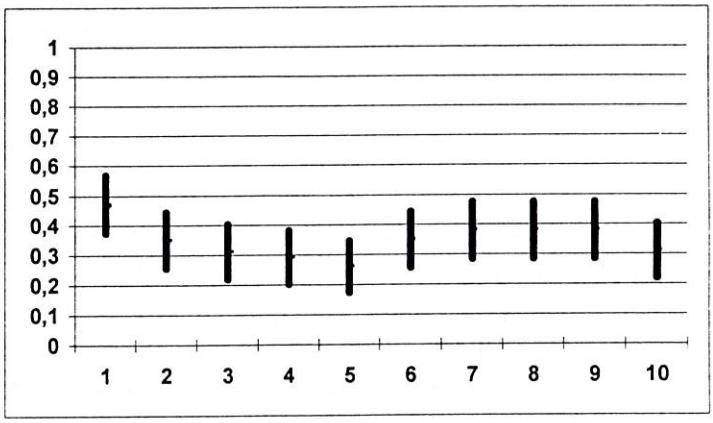
Le pourcentage de confiance le mieux adapté à cette situation de sondage est 95 %. C'est un bon compromis entre un trop grand nombre d'erreurs (90 % de confiance) et une trop grande amplitude (manque de précision au coefficient 99 %).

ESSAIS D'ESTIMATION DE VALEURS INCONNUES

Exemple de situation :

A partir du graphique ci-contre, on peut, en considérant une "moyenne" parmi les fluctuations observées, estimer que p est de l'ordre de 0,35.

Après vérification sur la feuille de calcul, on constate que p = 0,32.



4 – A propos des sondages politiques

En décembre 1965, à l'occasion de la première élection présidentielle au suffrage universel, les français découvrent les sondages pré-électoraux. Contre tous les observateurs, l'Ifop avait prévu le ballottage du général *De Gaulle*. *"C'est un triomphe pour l'Ifop, sinon pour le général"*, déclarait alors un ministre en privé. Cependant, lors des trois dernières grandes élections en France (présidentielles 1995, législatives 1997, présidentielles 2002), les instituts de sondage se sont en grande partie "trompés". Regardons de plus près.

a) Un exercice autour de l'élection présidentielle de 2002

Aucun sondage n'avait réellement prévu l'éviction de *Lionel Jospin* du second tour de l'élection. Et pourtant...si l'on avait fourni les fourchettes de sondage, on aurait vu que la prudence s'imposait ! Voici un exercice à proposer en seconde.

Exercice :

Lors du premier tour des élections présidentielles, le dernier sondage publié par l'institut B.V.A. , effectué sur 1000 électeurs le vendredi 19/04/02, prévoyait :

Jacques Chirac	19 %
Lionel Jospin	18 %
Jean-Marie Le Pen	14 %

La surprise a été grande le dimanche 21/04/02 au vu des résultats :

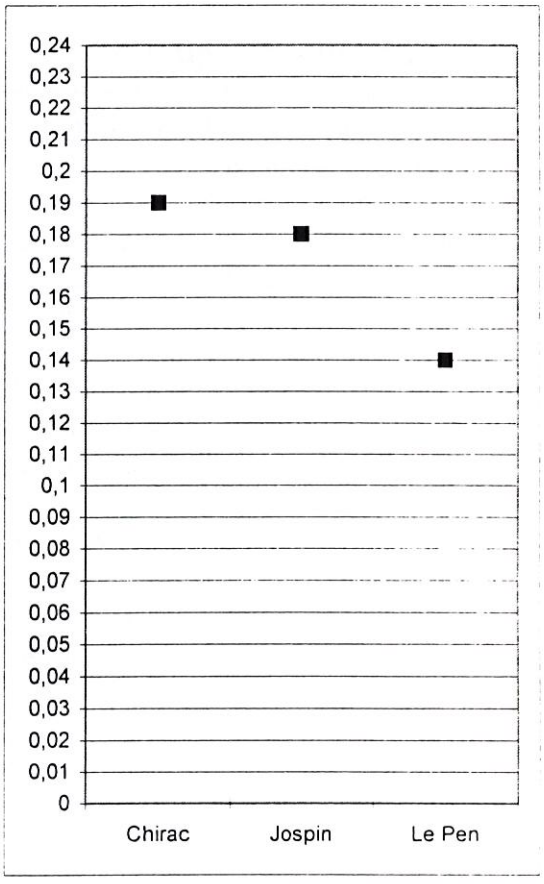
Jacques Chirac	19,88 %
Lionel Jospin	16,18 %
Jean-Marie Le Pen	16,86 %

1) On rappelle la formule des fourchettes de sondage à plus de 95 % de confiance, calculée à partir d'une fréquence f obtenue sur un échantillon aléatoire de taille 1000 :

$[f - \frac{1}{\sqrt{1000}}, f + \frac{1}{\sqrt{1000}}]$.

Calculer les trois fourchettes obtenues à partir du sondage B.V.A..

Représenter ces fourchettes sur le graphique ci-contre.

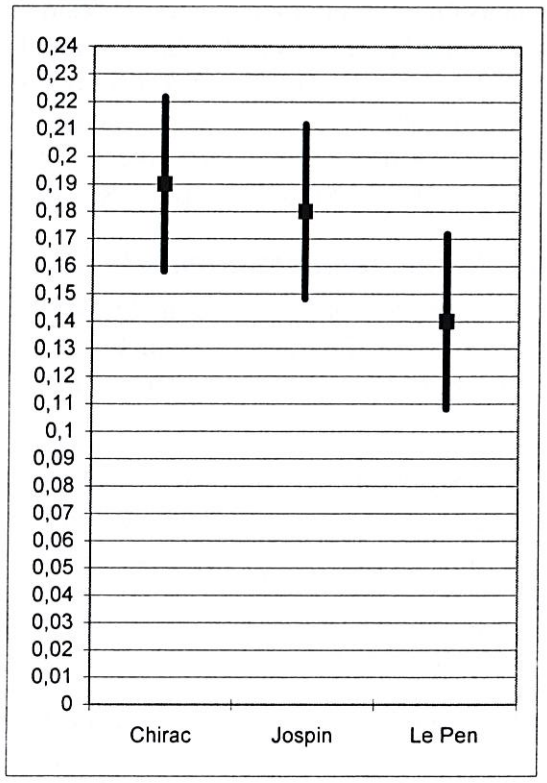


2) En se basant sur ces fourchettes, peut-on prévoir l'ordre des candidats à l'élection ?

3) Placer sur le graphique les résultats de l'élection.

Doit-on considérer que le dernier sondage B.V.A. était "faux" ? Expliquer votre réponse.

Eléments de réponse



b) Des sondages pas toujours aléatoires

La méthode des quotas

En France, à la différence semble-t-il des Etats-Unis, la méthode des sondages aléatoires est peu utilisée et les instituts ont souvent recours, pour leurs enquêtes d'opinion, à la méthode des quotas. Dans cette méthode, on exploite la structure connue de la population (par exemple grâce au recensement) pour reproduire la même structure dans l'échantillon. On choisit pour cela certains caractères de la population, que l'on pense devoir être en rapport avec l'enquête menée, comme le sexe, l'âge, la catégorie socioprofessionnelle, le type de commune... Ces caractères sont nommés variables de contrôle. Si l'on connaît la distribution de la population selon ces variables de contrôle, on obtient ainsi des quotas qui devront être respectés par les enquêteurs.

Le gros avantage est celui du coût, nettement moins élevé que celui des sondages aléatoires. Il y a moins de déplacements des enquêteurs et le rendement est environ deux fois plus élevé lorsque le choix des personnes à interroger est laissé libre, au lieu d'être imposé par une liste. Le gros inconvénient est d'être peu scientifique. Les personnes interrogées étant "choisies" par les enquêteurs, il est impossible de savoir quelle probabilité avait a priori chaque individu d'appartenir à l'échantillon. *"Avec la méthode des quotas, il n'existe pas de loi mathématique permettant de déterminer la marge d'erreur d'un sondage"*, explique *Jean-François Doridot*, directeur du département opinion d'Ipsos[8], *"en pratique toutefois, on considère que la marge d'erreur des sondages par quotas est égale, voire inférieure à celle des sondages aléatoires."* Des études ont cependant montré que

[8] Journal *"Le Monde"* du 17/03/02.

cette méthode avait tendance à sous représenter les travailleurs de l'industrie, les personnes les moins instruites ou ayant peu d'activités sociales...
On peut douter de l'affirmation ci-dessus selon laquelle la marge d'erreur par la méthode des quotas est égale "voire inférieure" à celle des sondages aléatoires. ***Le hasard est*** encore, pour éviter les biais, ***le meilleur allié du statisticien***.
Ecoutons, à propos du rôle du tirage au sort, *Daniel Schwartz*, pionnier de l'introduction de la statistique dans la médecine en France[9] :
"On ne dispose en général que d'estimations observées sur des échantillons qui s'écartent plus ou moins des vrais valeurs en raison des fluctuations d'échantillonnage. Ainsi *le hasard rend toute conclusion certaine impossible, il est notre maître, notre ennemi...*
Cependant l'intervention du hasard ne se limite pas là... L'établissement d'une fourchette dans la description d'une population suppose que l'échantillon considéré soit représentatif. On peut montrer que ceci n'est en principe réalisé que si l'échantillon résulte d'un tirage au sort. De même, dans la description comparée, par exemple dans la comparaison des taux de guéris avec deux traitements A et B, le test statistique permet de savoir si la différence est significative. Mais, dans ce cas, elle ne peut être attribuée au traitement que si les échantillons des deux groupes sont, à part le traitement, comparables à tous égards ; là encore on peut montrer que ceci nécessite que les deux groupes aient été constitués par tirage au sort. *Ainsi le hasard cette fois nous est utile, ce n'est plus notre ennemi, mais notre allié...*
On constitue souvent des échantillons par des procédés commodes, en s'imaginant qu'ils sont "représentatifs". Dans la population des étudiants suivant un cours, on choisira ceux du premier rang dans l'amphithéâtre. Dans un groupe de souris d'une race donnée, quand on souhaite faire une expérience sur 20 souris, on choisira les 20 premières attrapées dans la cage. Ces méthodes sont mauvaises. Dans un amphithéâtre, les élèves du premier rang (quand il y en a...) diffèrent des autres : souvent ce sont les plus consciencieux, les plus tôt arrivés ou ceux qui entendent ou voient moins bien. Les souris attrapées en premier... sont des nigaudes. L'expérience montre qu'elles sont plus vulnérables aux maladies."

La méthode de stratification

La stratification est, comme la méthode des quotas, fondée sur l'idée d'une exploitation des connaissances que l'on a de la population pour favoriser la représentativité de l'échantillon. Après le choix des critères de contrôle, on découpe la population en groupes homogènes, appelés strates. La différence essentielle avec la méthode précédente est, qu'à l'intérieur de chaque strate, on procède par tirage au sort. La stratification permet d'améliorer considérablement la précision de l'estimation et ce de façon mathématiquement quantifiable. On montre que pour obtenir la meilleur estimation possible, le taux de sondage dans chaque strate (rapport du nombre de sondés dans la state à l'effectif de celle-ci) doit être proportionnel à l'écart type dans la strate considérée. Ainsi, le taux de sondage est d'autant plus élevé que la dispersion de la variable étudiée à l'intérieur de la strate est plus grande. Un échantillon de ce type est nommé "échantillon de *Neyman*" d'après l'auteur de la méthode.
Le sondage *par grappes* consiste à constituer l'échantillon en interrogeant toutes les personnes appartenant à un même sous-ensemble, par exemple toutes les personnes d'un même foyer.

---

[9] Extrait de *"Chiffres en folie"* – Association Pénombre – Editions La Découverte 1999.

Des difficultés spécifiques

Des difficultés spécifiques aux sondages politiques (ou aux enquêtes d'opinion) tiennent non plus aux problèmes de biais affectant la constitution de l'échantillon, mais aux réponses des sondés : abstentionnistes répugnant à avouer qu'ils n'ont pas l'intention de voter, indécision jusqu'au dernier moment, sympathisants d'extrême droite hésitant à afficher leurs opinions... A la lumière des élections précédentes, des coefficients rectificatifs sont alors appliqués, faisant alors du sondage politique davantage un art alchimique qu'une science. Une difficulté également importante dans les sondages politiques est la faible détermination des intentions de vote. Selon IPSOS la proportion d'électeurs estimant pouvoir changer d'avis la dernière semaine est de 40%.

• Premier tour des élections présidentielles de 1995 :
Dernier sondage (non publié) du 21/04/95 :

Jacques Chirac	24 %
Edouard Balladur	20 %
Lionel Jospin	19 %

Résultats du premier tour (23/04/95) :

Jacques Chirac	20,5 %
Edouard Balladur	18,5 %
Lionel Jospin	23,2 %

Ainsi *Edouard Balladur* est éliminé du second tour et *Lionel Jospin* est en tête du premier tour, alors que les sondages le prévoyaient en troisième position, derrière *Balladur*. Réaction immédiate de *Nicolas Sarkozy* : *"C'est une formidable défaite pour les instituts de sondage."*

• Premier tour des élections législatives de 1997 (suite à la dissolution de l'assemblée par *Jacques Chirac*, sans doute mal conseillé par les sondages...) :
La gauche est en tête alors que les sondages prévoyaient une victoire en sièges de la droite. Le cas des législatives est en effet particulièrement difficile à modéliser. Dans le modèle utilisé par la SOFRES[10], on tient compte des votes antérieurs à une élection comparable pour accorder des primes aux "sortants" (déjà en place) pour le second tour, d'environ 3%, des "matrices" de reports de voix pour le second tour, on "redresse" le score du Front national apparaissant dans le sondage... C'est un tel modèle qui simule la projection en sièges après le second tour. Le mode de scrutin rendant l'exercice de simulation particulièrement délicat, très sensible aux modifications d'opinion. Ainsi, le modèle de la SOFRES estimait, avant la dissolution, l'avance de la droite à 95 sièges. Au vu du premier tour, la droite modérée a perdu 3% des intentions de votes. En tenant compte d'une nouvelle matrice de reports de voix, estimée après le premier tour, améliorés des Verts vers le PS et détériorés du Front national vers la droite modérée, le modèle prévoit alors la victoire de la gauche. Le modèle de la SOFRES, à défaut d'avoir prévu la victoire de la gauche avant le premier tour, a au moins permis de quantifier l'impact des différents transferts de voix.

[10] Source : *"Chiffres en folie"* – Editions La Découverte 1999.

• Election présidentielle de 2002 :
Paris, mardi 30 avril 2002, agence *Reuter* :
"Le président du tribunal de Paris a rejeté une demande présentée par neuf journalistes et particuliers qui souhaitaient contraindre les quatre principaux instituts de sondage (Ipsos, CSA, Ifop et BVA) à mentionner obligatoirement les méthodes de correction utilisées, les modes de calcul et les marges d'erreurs.
"Le matraquage de sondages faux au premier tour de l'élection présidentielle a eu pour résultat un abstentionnisme massif et les électeurs ont subi un préjudice", avait plaidé lors de l'audience l'avocat des demandeurs. Le juge a cependant estimé que "la responsabilité des manquements allégués ne sauraient incomber aux instituts de sondage, dès lors que seuls les organes d'information assument la mission, et la responsabilité qui y est afférente, de publier les résultats."
[...] Les demandeurs devront payer 3000 euros de frais de procédure."
Les seules données que les instituts de sondage sont actuellement contraints de mentionner sont les dates de leur réalisation, la taille de l'échantillon et la méthode générale. En revanche, les instituts (et à plus forte raison les médias) ne précisent jamais que les résultats bruts après enquête sont systématiquement "redressés" par des méthodes de pondération, faisant appel aux résultats des scrutins antérieurs, puis "lissés" lorsque les résultats redressés paraissent trop fantaisistes.
Ainsi, pour l'élection de 2002, *Lionel Jospin*, lorsqu'il recueille 26 à 27% en données brutes, est crédité de 22% après pondération, de même *Jacques Chirac* passerait de 30% en brut à 27% en pondéré, ou *Jean-Marie Le Pen* de 4% à 8% en pondéré (chiffres cités par *Philippe Méchet* de la Sofres). Ces pondérations sont établies à partir de plusieurs élections antérieures et de questions posées par le sondeur et permettant de mesurer le degré de certitude du choix de l'électeur. *"Toutes ces opérations obéissent à des règles fixées au début des campagnes électorales et ne changent plus en cours de route. La pondération d'un échantillon n'est pas du bricolage ou une manière d'anticiper ce qui pourrait se passer, mais bien un calcul statistique issu des mathématiques. Rien de plus, rien de moins"*, affirme *Philippe Méchet*, Directeur des études politiques de la Sofres.
Vous avez dit alchimie ?

III – MESURER LA VARIABILITÉ EN 1ère

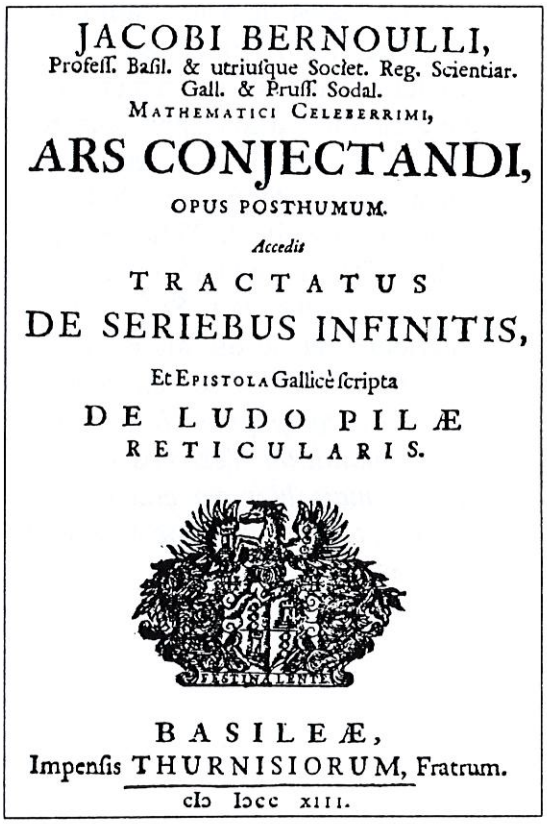
Comme on l'a vu à propos des fluctuations d'échantillonnage, l'étude de la variabilité est au centre des méthodes statistiques. Voyons par quels paramètres on peut la mesurer.
Il s'agit d'abord de déterminer une valeur "centrale" par rapport à laquelle sera mesurée la dispersion.

1 – Moyenne contre médiane : quel "milieu" ? Une approche historique

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.

Dans l'*Ars conjectandi* (1713) de *Jacques Bernoulli*, on trouve 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)."


JACOBI BERNOULLI,
Profeſſ. Baſil. & utriuſque Societ. Reg. Scientiar.
Gall. & Pruſſ. Sodal.
MATHEMATICI CELEBERRIMI,
ARS CONJECTANDI,
OPUS POSTHUMUM.
Accedit
TRACTATUS
DE SERIEBUS INFINITIS,
Et EPISTOLA Gallicè ſcripta
DE LUDO PILÆ
RETICULARIS.
BASILEÆ,
Impenſis THURNISIORUM, Fratrum.
cIↃ IↃcc xIII.

L'usage de la moyenne arithmétique sera développé dans le cadre de la ***théorie des erreurs***, dans un contexte probabiliste.
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^e siècle, une certaine ambiguïté existe dans la terminologie. Ainsi une expression telle que "erreur moyenne", désignera, dans les ouvrages français, la moyenne des écarts à la médiane $\frac{1}{n}\sum|x_i - Mé|$, 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 consacrera 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 comme meilleure estimation, est la loi normale.

Pierre Simon Laplace

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.

Le terme d'***écart type*** ("*standard deviation*"), ainsi que sa notation σ, sont introduits en 1893 par *Karl Pearson*.

Autres avantages de la moyenne arithmétique

Un avantage important de la moyenne est de pouvoir être calculée 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^e 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.

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 la moyenne. 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*, 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.*"

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 M est obtenue lorsque l'on minimise la somme $\sum|x_i - M|$ des écarts absolus.

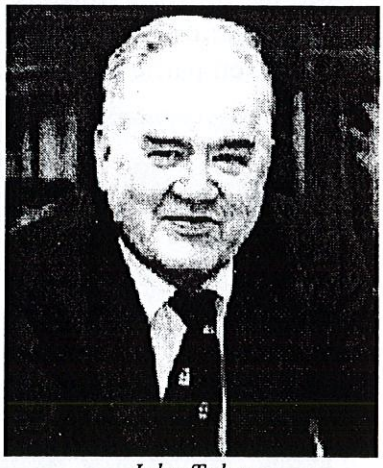
On en trouve l'usage chez *Gauss* (1816) et *Laplace* (1818), mais c'est surtout *Galton* qui lui accorde toute son attention en 1874, 1875, dans le cadre de ses études anthropologiques.
Mais l'usage de la médiane sera longtemps négligé, en raison, d'une part, de moins bonnes propriétés algébriques,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

La difficulté d'interprétation de la moyenne

Dès la fin du XIX^{e} 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 contestable ("l'homme moyen" de *Quételet*). 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 (critères non quantitatifs).

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, introduits par *John Tukey* (1915-2000) 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.



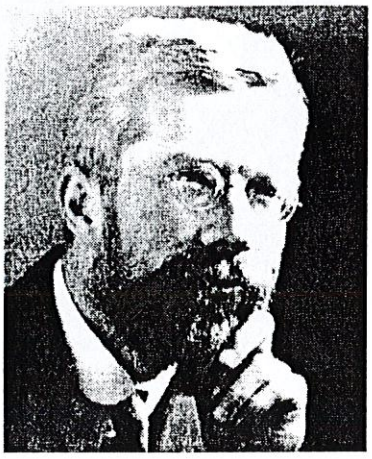
John Tukey

En résumé :

Moyenne		Médiane	
Avantage	Inconvénient	Avantage	Inconvénient
Minimise : $x \mapsto \sum (x_i - x)^2$. Indicateur de centralité euclidien.	Sensibilité aux valeurs extrêmes (recours possible à la moyenne tronquée).	Minimise : $x \mapsto \sum \lvert x_i - x \rvert$	Pas de calcul par paquets (recours possible à l'ordinateur).
Estimation optimale dans le cas d'une distribution normale.	Mal adapté dans le cas de distributions non normales.	Robustesse : peu sensible aux valeurs extrêmes.	Moins adapté que la moyenne pour les distributions normales.
Calcul possible par paquets (barycentre).	Interprétation parfois difficile (exemple : 2,5 enfants par famille).	Interprétation simple (en général une valeur de la série, milieu réel).	Pas de propriétés statistiques de type théorèmes limites.
Bonne connaissance de la loi de la moyenne de variables aléatoires indépendantes (théorèmes limites).	Réservé aux séries numériques.	Bonne illustration avec les boîtes à moustache.	

2 – Moyenne et écart type : aspect géométrique

L'intuition géométrique joue un rôle important en statistique. La partie de la statistique que l'on nomme "analyse des données" fait un usage important de la géométrie en dimension *n* (et de l'algèbre linéaire) pour mettre en évidence les "composantes principales" d'un ensemble de données. L'enseignement de la géométrie dans l'espace dans les sections ES trouve en partie sa justification dans ces utilisations.

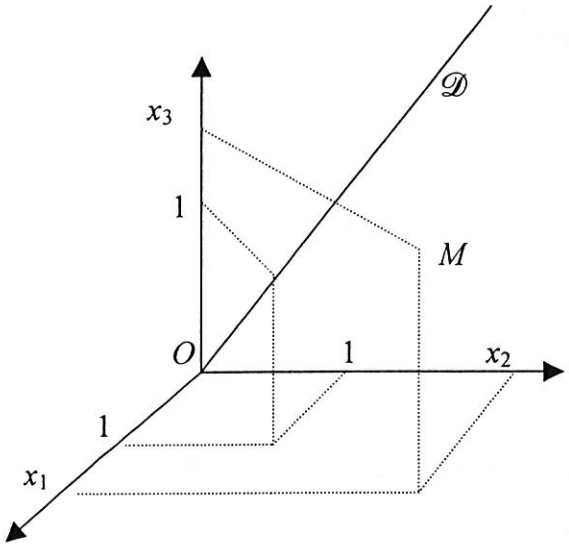


Pour l'anecdote, signalons que ***Ronald Aymler Fisher*** (1890-1962), le père de la statistique moderne, était doué d'une intuition géométrique hors du commun. *Fisher* a toujours souffert de problèmes de vue. Enfant, les médecins pour protéger ses yeux lui interdisaient de lire à la lumière artificielle. A l'école le soir, ses professeurs, pour éviter la lumière électrique, devaient lui enseigner les mathématiques sans faire usage de papier, crayon, ni aucune aide visuelle[11]. En conséquence, *Fisher* développa un profond sens géométrique, qui lui permit de résoudre bien des problèmes difficiles de statistique mathématique. A titre d'exemple, citons cet extrait d'une lettre de *William Gosset* (alias *Student*) à *Karl Pearson* : *"Je vous joins une lettre qui donne une preuve de mes formules pour la distribution des fréquences du [t de Student]... Pourriez-vous regarder cela pour moi ; je ne me sens pas à l'aise dans plus de trois dimensions, même si je peux le comprendre autrement..." Fisher* avait démontré les résultats de *Gosset* en utilisant la géométrie multidimensionnelle.

[11] Rapporté par D. Salsburg – *"The lady tasting tea"*.

TRAVAUX DIRIGES

UNE INTERPRETATION GEOMETRIQUE DE LA MOYENNE ET DE L'ECART TYPE



On suppose que l'on a effectué trois mesures x_1, x_2, x_3 (différentes à cause des erreurs de mesure) d'une même grandeur physique.

Pour représenter ces mesures, on se place dans l'espace, rapporté à un repère orthonormal $\left(O; \vec{i}, \vec{j}, \vec{k}\right)$, où l'on considère le point M de coordonnées (x_1, x_2, x_3).

Soit $\mathscr{D}$ la droite menée par O de vecteur directeur de coordonnées (1, 1, 1).

1) On souhaite résumer les trois mesures x_1, x_2, x_3 par une seule valeur t.

a) Soit $P(t, t, t)$ un point de la droite $\mathscr{D}$. Exprimer la distance PM.

b) Déterminer la valeur de t pour laquelle la fonction $f : t \mapsto (t - x_1)^2 + (t - x_2)^2 + (t - x_3)^2$ atteint son minimum (calculer la dérivée).

c) En déduire que le point P le plus proche de M sur $\mathscr{D}$ a pour coordonnées $(\bar{x}, \bar{x}, \bar{x})$ où $\bar{x}$ est la moyenne des trois mesures x_1, x_2, x_3.

d) En utilisant la condition d'orthogonalité, généralisée à trois coordonnées, vérifier que $\overrightarrow{OP} \perp \overrightarrow{PM}$.

2) Exprimer la distance PM en fonction de l'écart type $\sigma = \sqrt{\frac{1}{3}\sum_{i=1}^{3}(x_i - \bar{x})^2}$ des trois mesures x_1, x_2, x_3.

3) En utilisant le théorème de Pythagore, retrouver la formule : $\sigma^2 = \frac{1}{3}\sum_{i=1}^{3} x_i^2 - \bar{x}^2$

(la variance est égale à la moyenne des carrés moins le carré de la moyenne).

Eléments de solution

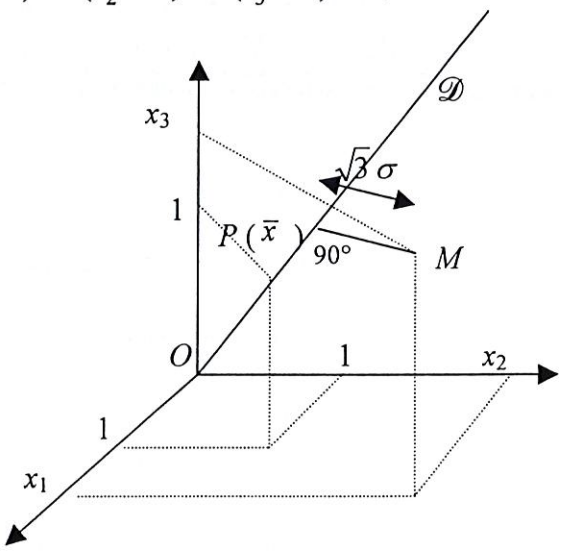
1) a) On a $PM = \sqrt{(t - x_1)^2 + (t - x_2)^2 + (t - x_3)^2}$.

b) On a $f'(t) = 2(t - x_1) + 2(t - x_2) + 2(t - x_3) = 0$ pour $3t = x_1 + x_2 + x_3$ soit $t = \bar{x}$.

On a bien un minimum puisque f un polynôme de degré 2 de coefficient dominant +3.
c) On a $PM = (f(t))^2$. Le minimum est donc atteint, comme pour f , pour $t = \bar{x}$.
d) On a $\bar{x}(x_1 - \bar{x}) + \bar{x}(x_2 - \bar{x}) + \bar{x}(x_3 - \bar{x}) = \bar{x}(x_1 + x_2 + x_3) - 3\bar{x}^2 = 0$.

Ainsi $\overrightarrow{OP} \perp \overrightarrow{PM}$.

2) On a $PM = \sqrt{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2} = \sqrt{3}\,\sigma$.

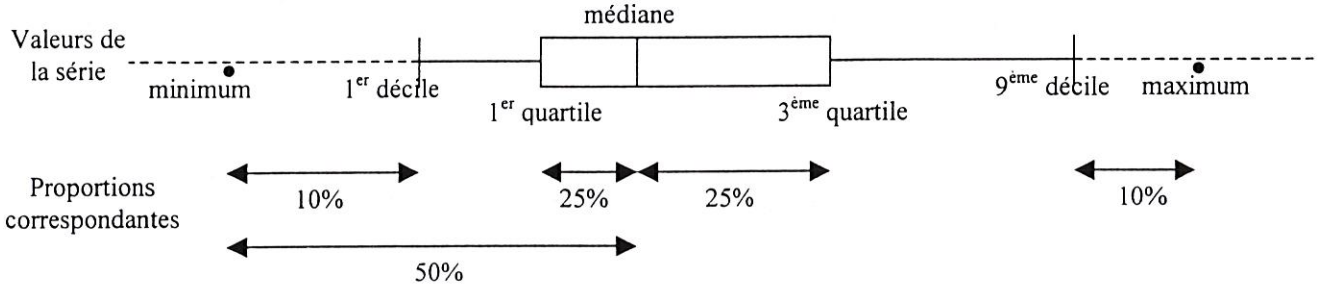


3) Puisque OMP est rectangle en P, le théorème de Pythagore donne $OM^2 = OP^2 + PM^2$ c'est à dire $\sum x_i^2 = 3\bar{x}^2 + 3\sigma^2$ d'où $\sigma^2 = \frac{1}{3}\sum_{i=1}^{3} x_i^2 - \bar{x}^2$.

3 – Quartiles et boîtes à moustache

Les diagrammes en boîtes permettent de visualiser la répartition des données : on partage l'ensemble des valeurs possibles en segments contenant une proportion prédéterminée des valeurs de la série statistique.

Le modèle ("par défaut") choisi par le programme officiel est le suivant (correspondant aux définitions qui suivent) :



Les définitions "officielles" sont les suivantes :
On ***ordonne*** la série des observations par ordre croissant.
Médiane : si la série est de taille $n = 2p + 1$, la médiane est la valeur du terme de rang $p + 1$, si la série est de taille $n = 2p$, la médiane est la demi-somme des valeurs des termes de rang p et $p + 1$.

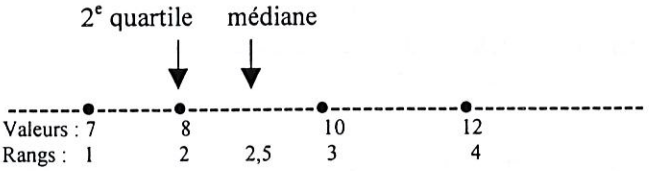
"La définition de la médiane n'est pas figée : certains définissent la médiane comme étant le deuxième quartile ou le cinquième décile ; ***dans la pratique de la statistique, les différences entre ces deux définitions sont sans importance. Au lycée, on évitera tout développement à ce sujet qui ne serait pas une réponse individuelle à une question d'un élève.***"
"La procédure qui consiste à tracer une courbe dite de fréquences cumulées croissantes, continue, obtenue par interpolation linéaire, et à définir la médiane comme l'intersection de cette courbe avec la droite d'équation y = 0,5 ou avec une courbe analogue dite des fréquences cumulées décroissante, ***n'est pas une pratique usuelle en statistique et ne sera pas proposée au*** *lycée."*
Document d'accompagnement du programme de 1ère.

Premier décile : la plus petite valeur de la série ordonnée telle qu'au moins 10% de ses valeurs en soient inférieures ou égales.
Premier quartile : même chose avec 25%.
Troisième quartile : même chose avec 75%.
Neuvième décile : même chose avec 90%.

En pratique : le 1er décile (resp. le 9ème décile) d'une série ordonnée de taille n est le terme dont le rang est le plus petit entier supérieur ou égal $\frac{1}{10}n$ (resp. $\frac{9}{10}n$).
Le 1er quartile (resp. le 3ème quartile) d'une série ordonnée de taille n est le terme dont le rang est le plus petit entier supérieur ou égal $\frac{1}{4}n$ (resp. $\frac{3}{4}n$).

SUR CALCULATRICES T.I. OU CASIO OU EXCEL:

La définition de la médiane est celle du programme, en revanche, celle des ***1er et 3ème quartiles*** diffère : sur les calculatrices, ils sont considérés comme les médianes (au sens du programme) des deux demi-séries obtenues en scindant la série initiale selon la médiane. Excel attribue, de plus, à la médiane un rang éventuellement fictif et procède par interpolation selon les rangs. Les différences ne sont en général sensibles que pour des séries de faible effectif, voici un exemple extrême (on ne fait pas de la statistique sur 4 valeurs !) :



Définitions : *Mé* = 9 *Q1* = 7 *Q3* = 10.
Calculatrices :
Mé = 9 *Q1* = 7,5 *Q3* = 11.
Excel : *Mé* = 9 *Q1* = 7,75 *Q3* = 10,5.

Pour les ***boîtes à moustache*** sur les calculatrices :
• T.I. : Deux types sont proposés. La boîte est toujours l'intervalle interquartile [Q1, Q3].
Dans le type 1 les moustaches correspondent à 1,5 fois m – Q1 et Q3 – m. Les points "aberrants" (en dehors des moustaches) sont représentés.
Dans le type 2 les moustaches sont limitées par les valeurs extrêmes min et max de la série.
‣ Faire ***STATPLOT***
Choisir ***Plot1*** , activer ***On*** , choisir la boîte puis les listes. Faire ***GRAPH*** pour l'affichage.

• CASIO : Dans la Med-Box, les moustaches sont limitées par les valeurs extrêmes min et max de la série.
‣ Faire Menu ***GRPH***, puis ***SEL*** pour sélectionner le numéro du graphique.
Faire ***SET*** pour choisir le type de graphique ***Box*** puis choisir les listes.

4 – Des exemples d'activités en 1ère S ou ES

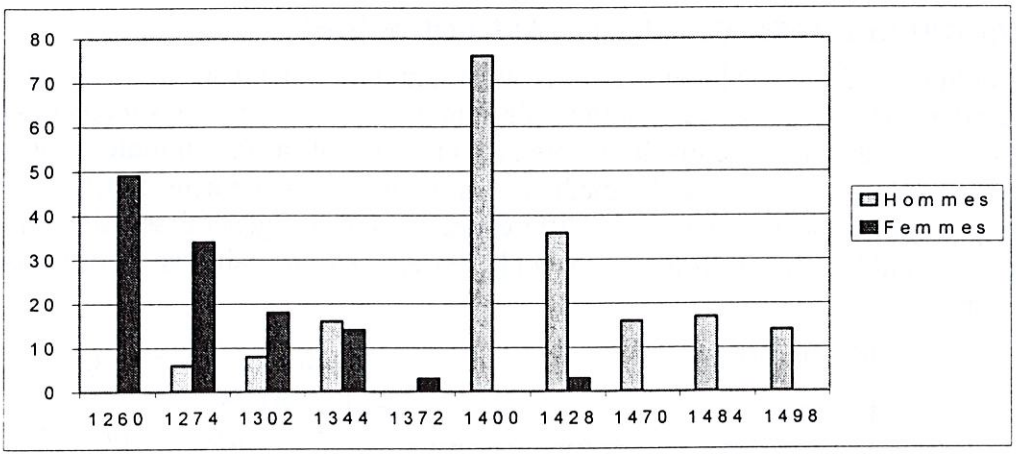
Exercice

REPARTITION DES SALAIRES DES OUVRIERS D'UNE ENTREPRISE

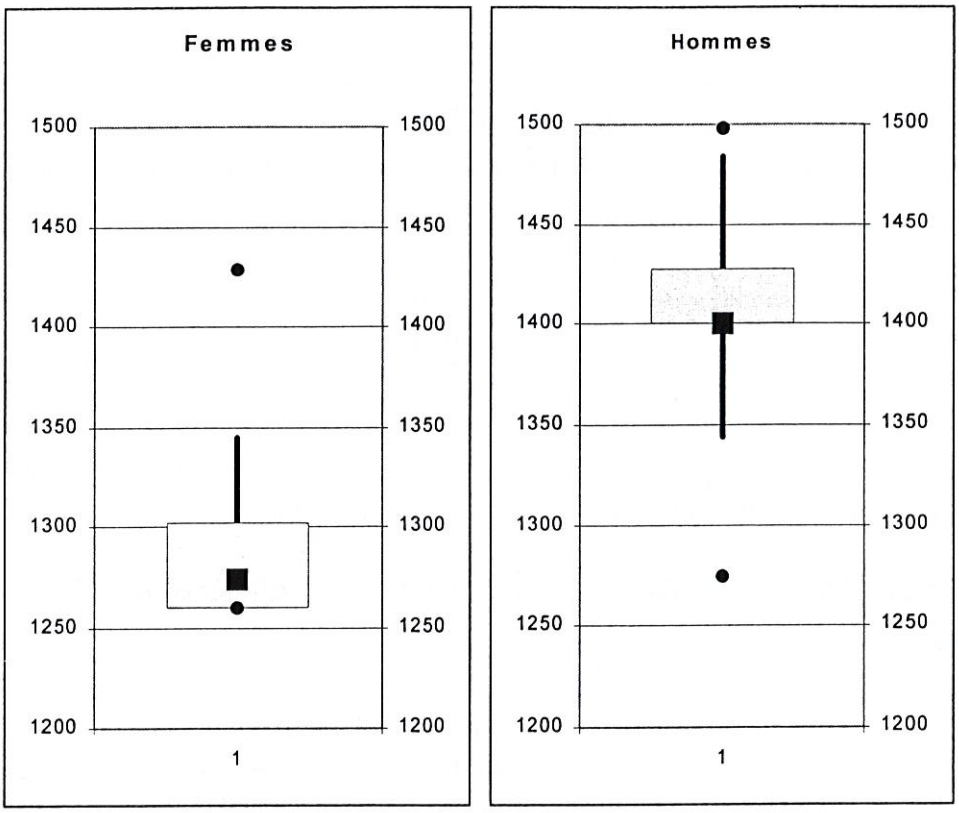
Une entreprise de fabrication métallique emploie 310 ouvriers, 189 hommes et 121 femmes. La répartition des salaires des ouvriers se fait selon le tableau suivant :

Salaire mensuel en euros	Hommes		Femmes	
	Effectifs	Effectifs cumulés	Effectifs	Effectifs cumulés
1260	0	0	49	49
1274	6	6	34	83
1302	8	14	18	101
1344	16	30	14	115
1372	0	30	3	118
1400	76	106	0	118
1428	36	142	3	121
1470	16	158	0	121
1484	17	175	0	121
1498	14	189	0	121

Un histogramme peut illustrer cette répartition.



1) Dans cette entreprise, les hommes sont en moyenne mieux rémunérés que les femmes (qui n'occupent pas le même type de poste) : $\bar{x}_H \approx 1413{,}19$ € et $\bar{x}_F \approx 1286{,}84$ €.
Comme paramètre de centralité, les moyennes apportent un renseignement comptable : sachant qu'il y a 189 ouvriers et 121 ouvrières, quelle est la somme approximative versée chaque mois aux employés de cette catégorie ?
2) Les écarts types correspondants sont $\sigma_H \approx 53{,}97$ € et $\sigma_F \approx 37{,}74$ €. Quelle information apportent ces valeurs ?
3) Déterminer le salaire médian m_H des ouvriers et celui m_F des ouvrières.
Expliquer les différences observées avec la moyenne.
Du point de vu social, le salaire médian est un indicateur de centralité plus fiable que la moyenne. Pour quelle raison ?



4) La comparaison des salaires entre hommes et femmes est facilitée par l'examen des boîtes à moustaches : le carré correspond à la médiane, le rectangle à l'intervalle interquartile, les moustaches au 1er au 9ème décile et les points au minimum et maximum.

a) Comment voit-on qu'au moins 50% des femmes ont un salaire inférieur au plus petit salaire masculin ?

b) Le 2ème salaire féminin est 1372 €. Comment constate-t-on qu'au moins 75% des hommes gagnent davantage ?

c) Le 1er décile des hommes est supérieur au 9ème décile des femmes. Qu'est-ce que cela signifie ?

Eléments de réponse

1) Les moyennes permettent d'évaluer la somme totale des salaires des ouvriers :
$1413{,}19 \times 189 + 1286{,}84 \times 121 \approx 422\,800$ €.

2) Les écarts types indiquent que les salaires masculins sont davantage dispersés.

3) On a (189 + 1)/2 = 95. Le salaire médian masculin est donc au 95ème rang, soit m_H = 1400 €. A comparer à la moyenne de 1413,19 € influencée par quelques "gros" salaires. D'un point de vue social, cela signifie que 50% des ouvriers ont un salaire inférieur ou égal à 1400 €.

On a (121 + 1)/2 = 61. Le salaire médian masculin est donc au 61ème rang, soit m_F = 1274 €. A comparer à la moyenne de 1286,84. La médiane rend compte ici de la masse des faibles salaires féminins à 1260 €.

4) Les boîtes à moustaches indiquent ici des structures non évidentes des données.

a) Le carré de la médiane féminine est en dessous du point minimum masculin.

b) Le rectangle masculin (et donc le 1er quartile) est supérieur à 1372.

c) Plus de 90% des hommes gagnent davantage que plus de 90% des femmes.

T.P. SUR EXCEL

STATISTIQUES A L'HOPITAL

Dans le classement des hôpitaux de France publié par un magazine, on distingue les deux hôpitaux de province suivants. Pour une intervention analogue (assez rare), on lit :
Hôpital *A* : 1 décès sur 12 interventions au total.
Hôpital *B* : 3 décès sur 12 interventions au total.
Faut-il suspecter l'hôpital *B* qui a connu 25% de décès sur ce type d'intervention ? Le ministère de la santé devrait-il diligenter une enquête ?

Avant de se faire une opinion, nous allons simuler la situation en supposant que le risque de décès, dans les conditions habituelles pour ce type d'opération, est $p = 15\ \%$.

1- Simulation d'une étude sur 12 opérations pour 999 hôpitaux

	A	B	C	D
1	p	0,15		
2	nb décès	effectif	cumul	
3	0			
4	1			
5	2			
6	3			
7	4			
8	5			
9	6			
10	7			
11	8			
12	9			
13	10			
14	11			
15	12			
16				

Lancer Excel®. Dans la cellule A1, taper *p* et dans la cellule B1 entrer la valeur 0,15.
Puis préparer, de A2 à C15, le tableau ci-contre.

En E1 entrer la ***formule*** : =ENT(ALEA()+$B$1)
Approcher le pointeur de la souris du coin inférieur droit de la cellule E1. Lorsque le pointeur s'est transformé en une croix noire, ***glisser*** vers la droite pour ***recopier vers la droite*** jusqu'en P1.
En D1 entrer la ***formule*** : =SOMME(E1:P1)

Vous avez simulé 12 opérations réalisées dans un premier hôpital. Combien y a-t-il eu de décès ? .....
Sélectionner les cellules de D1 à P1 puis recopier vers le bas jusqu'en P999.
Pour trier les résultats des 999 hôpitaux et recueillir les effectifs, ***sélectionner*** les cellules de B3 à B15, puis inscrire dans la ***barre de formules*** :
= FREQUENCE(D1:D999;A3:A15) puis appuyer *simultanément* sur les touches ***CTRL MAJ ENTREE***.
Pour compléter la colonne des effectifs cumulés, entrer en C3 la ***formule*** : =B3
Puis en C4 la ***formule*** : =C3+B4
Recopier la cellule C4 ***vers le bas*** jusqu'en C15.

Caractéristiques de position et dispersion

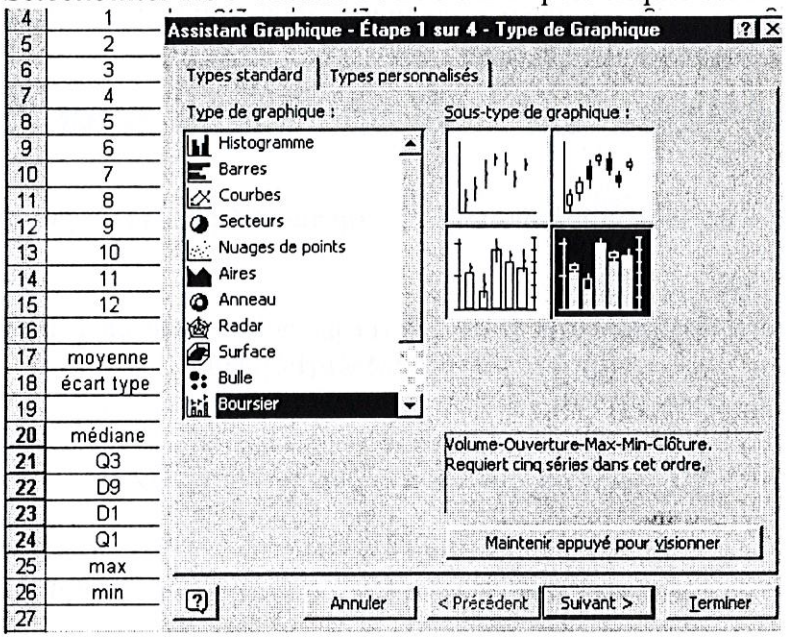
De A17 à B26, préparer le tableau ci-contre.
En B17 entrer la ***formule*** : =MOYENNE(D1:D999)
En B18 entrer la ***formule*** : =ECARTYPEP(D1:D999)
Pour obtenir la médiane, entrer en B20 la ***formule*** :
=NB.SI(C3:C15;"<500")
Pour obtenir le 3ème quartile, entrer en B21 la ***formule*** :
=NB.SI(C3:C15;"<750")

	A	B	C
16			
17	moyenne		
18	écart type		
19			
20	médiane		
21	Q3		
22	D9		
23	D1		
24	Q1		
25	max		
26	min		
27			

Pour obtenir le 9ème décile, entrer en B22 la ***formule*** :
=NB.SI(C3:C15;"<900")
En B23 entrer une ***formule*** fournissant le 1er décile et en B24 celle donnant le 1er quartile.
Pour obtenir le maximum, entrer en B25 la ***formule*** : =NB.SI(C3:C15;"<999")
Pour obtenir le minimum, entrer en B20 la ***formule*** : =NB.SI(C3:C15;"<1")
Où se situent les résultats des hôpitaux *A* et *B* (1 décès et 3 décès) par rapport aux quantiles précédents ? .....
.....

Boîtes à moustaches

Sélectionner les 5 cellules de B20 à B24 puis cliquer sur l'icône d'***assistant graphique*** 



4	1
5	2
6	3
7	4
8	5
9	6
10	7
11	8
12	9
13	10
14	11
15	12
16	
17	moyenne
18	écart type
19	
20	médiane
21	Q3
22	D9
23	D1
24	Q1
25	max
26	min
27	

Etape 1 sur 4 : choisir graphique ***Boursier*** et le 4ème ***Sous-type*** puis cliquer sur ***Suivant***.
Etape 2 sur 4 : choisir ***Série en • Lignes*** puis cliquer sur ***Terminer***.

Cliquer avec le bouton droit sur la ***Légende*** et choisir ***Effacer***.
Cliquer avec le bouton droit sur l'axe vertical gauche puis choisir ***Format de l'axe...*** Régler, dans l'onglet ***Echelle***, le ***minimum*** à 0, le ***maximum*** à 12 et l'***unité principale*** à 1. Procéder de même pour l'axe vertical droit.
Cliquer avec le bouton droit sur le rectangle noir, choisir ***Format des barres de baisse...*** et choisir une couleur claire.
Cliquer avec le droit sur une moustache, choisir ***Format des lignes Haut-Bas...*** et augmenter l'épaisseur.
Cliquer avec le droit sur le grand rectangle inférieur, choisir ***Type de graphique...*** puis ***Courbes***. La médiane est alors figurée par un point.
Pour ajouter la valeur maximale, copier la cellule B25 puis la coller sur le graphique. Procéder de même pour la valeur minimale.
On peut, en appuyant sur la touche ***F9***, refaire une simulation sur 999 hôpitaux.
Quel est l'intervalle interquartile le plus souvent observé, sur 999 hôpitaux ?
.....
Doit-on considérer comme suspect le résultat de l'hôpital *B* (3 décès sur 12 opérations) ?
.....
.....
La différence observée entre les hôpitaux *A* et *B* doit-elle être considérée comme suffisamment significative pour justifier une différence de qualité dans les soins ?
.....
.....

2- En changeant la valeur de *p*

Utiliser le fichier Excel pour répondre aux questions suivantes.

On lance 12 fois une pièce de monnaie et on observe 8 "pile". Doit-on suspecter la pièce d'être truquée ?

.....

.....

Même question si on observe 11 "pile".

.....

.....

Un Q.C.M. comporte 12 questions indépendantes, pour lesquelles 3 réponses sont proposées. On répond au hasard. Justifier que l'on a moins de 10% de chances d'avoir plus de la moyenne à ce Q.C.M.

.....

.....

3- Simulation d'une étude sur 120 opérations pour 999 hôpitaux

Nous allons cette fois simuler un échantillon de taille 120 opérations pour la proportion p de décès égale à 0,15.

En B1, entrer à nouveau la valeur 0,15.

Sélectionner la cellule P1 puis ***recopier vers la droite*** (croix noire) jusqu'en DT1 pour obtenir 120 opérations. En DU1 entrer la ***formule*** : =SOMME(E1:DT1)

Que représente le nombre calculé dans la cellule DU1 ? .....

.....

Sélectionner les cellules de E1 à DU1 puis ***recopier vers le bas*** jusqu'en DU999.

En A28 inscrire : nb décès.

En A29 entrer la valeur 0 puis en A30 entrer la ***formule*** =A29+1

Sélectionner la cellule A30 puis ***recopier vers le bas*** jusqu'en A149. Cette cellule devra alors contenir la valeur 120.

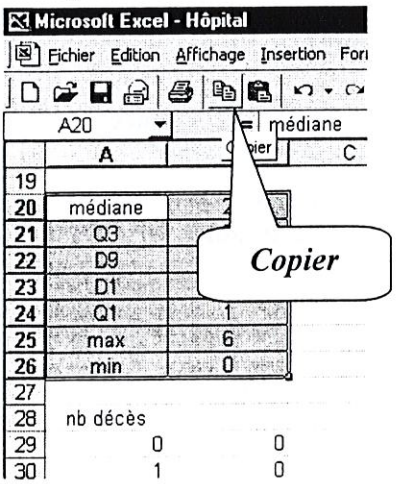
Sélectionner les cellules de B29 à B149 puis inscrire dans la ***barre de formules*** :

= FREQUENCE(DU1:DU999;A29:A149) puis appuyer *simultanément* sur les touches ***CTRL MAJ ENTREE***.

Pour compléter la colonne des effectifs cumulés, entrer en C29 la ***formule*** : =B29

Puis en C30 la ***formule*** : =C29+B30

Recopier la cellule C30 ***vers le bas*** jusqu'en C149.



Vous aller construire une boîte à moustaches pour illustrer l'étude sur des échantillons de taille 120.

Sélectionner les cellules de A20 à B26 puis cliquer sur l'icône ***Copier***.

Cliquer sur la cellule A152 puis ***Entrée***.

Dans les ***formules*** NB.SI situées dans les cellules B152 à B158 indiquer la plage C29:C149 pour obtenir les différents éléments de la boîte à moustaches.

Construire une boîte à moustaches en procédant comme au paragraphe 1 (avec ici ***Echelle***, ***minimum*** à 0, ***maximum*** à 120 et ***unité principale*** à 10).

Faire plusieurs fois ***F9***.

Que doit-on penser d'un hôpital qui connaîtrait 30 décès sur 120 (25%) ? .....

.....

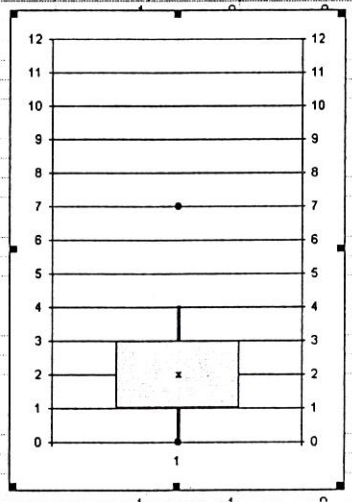
.....

Eléments de solution

1- Echantillons de taille 12

On obtient, pour p = 0,15 , des résultats du type suivant :

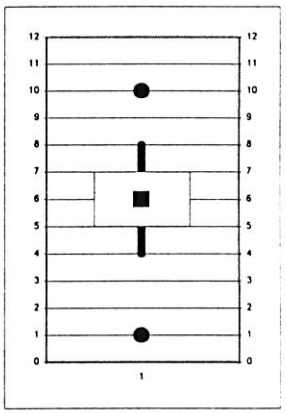
	A	B	C
2	nb décès	effectif	cumul
3	0	145	145
4	1	278	423
5	2	300	723
6	3	174	897
7	4	71	968
8	5	24	992
9	6	6	998
10	7	1	999
11	8	0	999
12	9	0	999
13	10	0	999
14	11	0	999
15	12	0	999
16			
17	moyenne	1,84884885	
18	écart type	1,2774224	
19			
20	médiane	2	
21	Q3	3	
22	D9	4	
23	D1	0	
24	Q1	1	
25	max	7	
26	min	0	



La plupart du temps, l'intervalle interquartile est [1, 3]. Les résultats des hôpitaux *A* et *B* sont donc dans cet intervalle. On peut alors considérer que l'écart observé est vraisemblablement du au hasard et n'est pas significatif d'une qualité différente des soins.

2- A pile ou face

En B1 on introduit pour p la valeur 0,5.
Sous l'hypothèse p = 0,5 l'observation d'au moins 8 "pile" se produit dans plus de 10% des cas. Ce nombre n'est pas significatif pour suspecter la pièce d'être truquée.
En revanche, l'observation de 11 "pile" est suffisamment rare pour considérer que si cela se produit en lançant 12 fois une pièce, il est possible que celle-ci soit truquée.

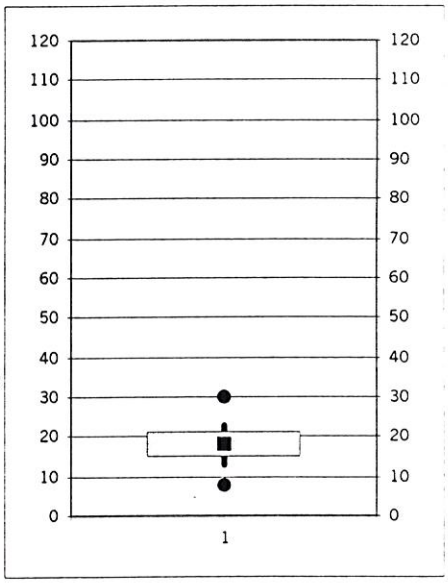


Le Q.C.M.

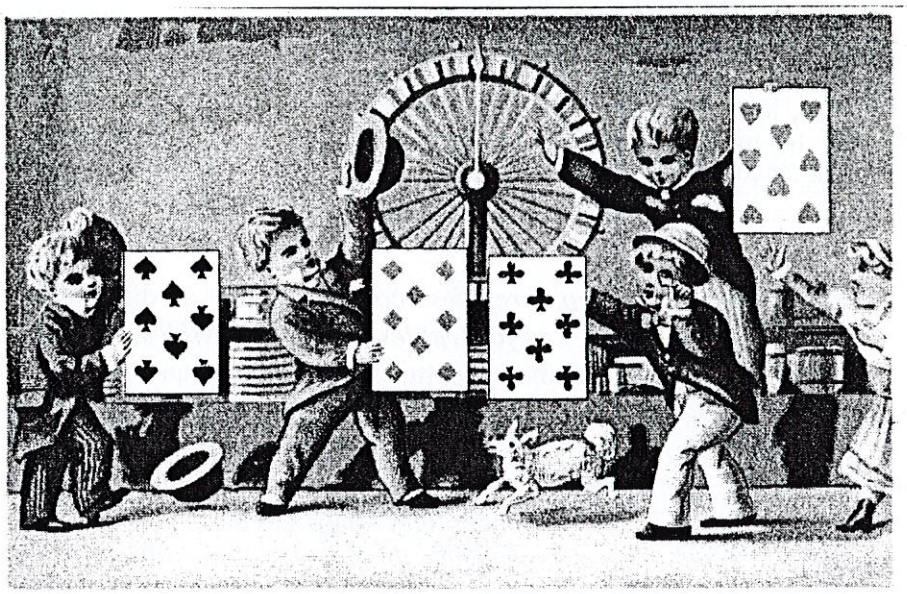
On introduit dans la cellule B1 la formule =1/3 et on constate que la valeur 7 est régulièrement supérieure au 9ème décile.

3- Echantillons de taille 120

On observe (à échelle équivalente) une boîte à moustaches beaucoup plus resserrée pour les échantillons de taille 120 que pour ceux de taille 12.
La valeur de 30 décès sur 120 opérations (25%) est très supérieure au 9ème décile et se produit donc très rarement dans un hôpital pour lequel le taux de décès théorique est de 15%.
On peut donc considérer comme douteux ce genre de résultat et diligenter, dans ce cas, une enquête quant au bon fonctionnement du service responsable de ce type d'opération.



Séance 3 :
LOI DE PROBABILITE ET MODELISATION



I – APPROCHE STATISTIQUE D'UNE LOI DE PROBABILITE

"On ne peut guère donner une définition satisfaisante de la Probabilité."
Henri Poincaré – *"Calcul des probabilités"* – 1856.

"[...] le terme de probabilité, seul, n'est pas défini à ce niveau d'étude."
Document d'accompagnement du programme de 1ère S – 2001.

1 – QU'EST-CE QU'UNE PROBABILITE ?

Nous avons connu plusieurs façons d'introduire la probabilité dans nos classes de lycée :

- Soit par la formule "nombre de cas favorables/nombre de cas possibles", dans le cadre de l'équiprobabilité (à définir ?). Ceci nous oriente sur la pratique du ***dénombrement*** et de la ***combinatoire*** mais offre peu d'intérêt au niveau des applications.
- Soit par une définition ***purement formelle*** d'espace probabilisé mais on perd le sens de ce dont on parle.
- Soit comme le nombre vers lequel se ***stabilise*** la ***fréquence observée***. Ce qui pose le problème de l'énoncé de la loi des grands nombres.

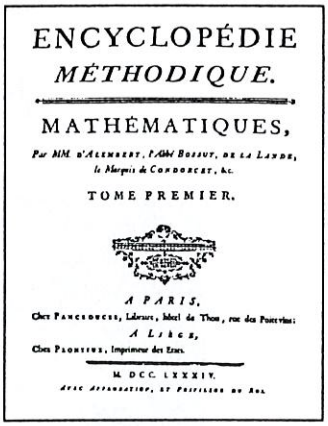
Cette dernière approche qui sera privilégiée parce qu'elle donne du sens au concept de probabilité et qu'elle se rapproche de la pratique dans les applications, faisant le ***lien entre statistique et probabilité par le biais de la modélisation***.

Si l'on se tourne vers l'histoire, on constate que la notion de probabilité a rapidement revêtu ces différentes facettes.

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 influer 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.


ENCYCLOPÉDIE
MÉTHODIQUE.

MATHÉMATIQUES,

TOME PREMIER.

A PARIS,

A LIEGE,

M. DCC. LXXXIV.

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 (évènements 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 évènement, 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"*.

Le point de vue subjectiviste

François Le Lionnais[1] 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 $i = 1, ..., n$, 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 donnée par : $P(A_j|E) = \dfrac{P(A_j \cap E)}{P(E)} = \dfrac{P(A_j) \times P(E|A_j)}{\sum_{i=1}^{i=n} P(E|A_i) \times P(A_i)}$.

Dans l'ignorance totale des $P(A_i)$, *Laplace*, selon le *"principe de la raison insuffisante"*, leur attribue une valeur uniforme $P(A_i) = \dfrac{1}{n}$. On a alors $P(A_j|E) = \dfrac{P(E|A_j)}{\sum_{i=1}^{i=n} P(E|A_i)}$.

De Finetti (1906-1985) 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"*[2].

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.

---

[1] dans *"Les grands courants de la pensée mathématique"* – Réédition "Rivages" 1986.

[2] cité par *Saporta* dans "*Théorie des probabilités*" 1970.

2 – L'APPROCHE FREQUENTISTE (C'EST A DIRE STATISTIQUE) DES PROGRAMMES DE 2001

Les nouveaux programmes de 1ère S et ES (2001) proposent une approche de type "fréquentiste" de la notion de "loi de probabilité", basée sur un énoncé "vulgarisé" de la loi des grands nombres, et faisant suite à l'expérimentation des fluctuations d'échantillonnage pratiquée en seconde. On y considère la loi de probabilité comme un objet mathématique permettant la "modélisation" de la réalité.

Loi de probabilité

> *"**Modéliser** une expérience aléatoire, c'est lui associer une **loi de probabilité**"*
> *"Une fréquence est empirique."*
> *"Les distributions des **fréquences** issues de la répétition d'expériences identiques et indépendantes varient **(fluctuent)** ; la loi de probabilité est un **invariant** associé à l'expérience."*
>
> *Document d'accompagnement du programme de 1ère S 2001.*

> *"On recensera les propriétés mathématiques élémentaires de l'objet "distributions de fréquences" et on définira une loi de probabilité comme un objet mathématique ayant les mêmes propriétés."*
>
> *Document d'accompagnement du programme de 1ère S 2001.*

La notion de modélisation était étrangère à l'ancien programme de première. Ici, on distingue bien les fréquences (et leurs fluctuations), qui sont du domaine de la réalité des observations (dont le traitement relève de la statistique) avec la loi de probabilité qui est du domaine mathématique (de la théorie des probabilités) et qui constitue un modèle de la réalité. Cette distinction est, selon Michel Henry[3] (IREM de Franche-Comté), essentielle : "Ce point de vue est en rupture avec celui de l'ancien programme de Première, qui faisait de la probabilité une sorte de limite de fréquence stabilisée. Du coup elle appartenait au même paradigme, celui de la description de la réalité et non de sa modélisation. [Ce nouveau point de vue] va permettre d'introduire en Terminale S les lois continues dans cette cohérence épistémologique."

> *"En classe de première, une **loi de probabilité P** sur un ensemble fini E est la **liste** des probabilités des éléments de E ; à partir de cette liste, on définit naturellement les probabilités des évènements (c'est à dire **implicitement** une application de $\mathcal{P}(E)$ dans [0, 1])".*
> *"Il est **inutilement complexe**, pour le cas des ensembles finis, de partir d'une **application** de $\mathcal{P}(E)$ dans [0, 1], vérifiant certains axiomes, puis de montrer ensuite que cette application est entièrement caractérisée par $(p_1, p_2, \ldots, p_n)$.*
>
> *Document d'accompagnement du programme de 1ère S 2001.*

> *"On parle ainsi d'une loi de probabilité, de la probabilité d'un événement, mais **le terme de probabilité, seul,** n'est pas défini à ce niveau d'étude".*
>
> *Document d'accompagnement du programme de 1ère S 2001.*

[3] Voir "Des lois de probabilité continues en terminale S, pourquoi et pour quoi faire ?" – revue "Repères – IREM" – Avril 2003.

On voit par ces instructions que l'on ne cherchera pas à définir formellement la notion de probabilité ou de loi de probabilité. On ne fera pas de la théorie de la mesure à un niveau élémentaire !
En revanche, l'élément essentiel de l'articulation statistique/probabilités, la loi des grands nombres, est incontournable, même si une présentation rigoureuse en est, au niveau du lycée, impossible.

Loi des grands nombres

> *"Le lien entre loi de probabilité et distributions de fréquences sera éclairé par un* ***énoncé vulgarisé de la loi des grands nombres.****"*
> *"... on illustrera ceci par des simulations dans des cas simples."*
> *Programme de 1ère S 2001.*

On nous propose plusieurs "énoncés vulgarisés" (plus ou moins simples !) de la loi des grands nombres. On comprend que la chose est un peu délicate.

> *Enoncé 1 :*
> *"Pour une expérience donnée, dans le modèle défini par une loi de probabilité P, les distributions des fréquences calculées sur des séries de taille n se rapprochent de P quand n devient grand."*
> *Programme de 1ère S 2001.*

> *Enoncé 2 :*
> *"Dans le monde théorique défini par une loi de probabilité P sur un ensemble fini, les fréquences des éléments de cet ensemble dans une suite de n expériences identiques et indépendantes "tendent" vers leur probabilité quand n augmente indéfiniment."*
>
> *Enoncé 3 :*
> *"Si on choisit n éléments d'un ensemble fini E selon une loi de probabilité P, indépendamment les uns des autres, alors la distribution des fréquences est proche de la loi de probabilité P lorsque n est grand."*
>
> *Document d'accompagnement du programme de 1ère S 2001.*

Pour les enseignants, on rappelle, dans le document d'accompagnement, l'énoncé de la ***loi forte des grands nombres***, en termes de convergence presque sûre (la loi faible s'énonce en termes de convergence en probabilité) :

> *"Dans l'ensemble des suites infinies d'éléments choisis selon P, le sous-ensemble des suites pour lesquelles la distribution des fréquences ne converge pas vers P est négligeable."*
>
> *Document d'accompagnement du programme de 1ère S 2001.*

3 – ACTIVITES ELEVES EN 1ère (introduction avant le cours) :

Il est peut-être difficile d'introduire directement la notion de loi de probabilité par le biais d'une loi des grands nombres définie sur la distribution des fréquences des r éléments de E. On peut commencer par la loi des grands nombres concernant un événement, c'est à dire pour $E = \{x_1, x_2\}$ (x_1 pour l'événement A).

On donne ici un exemple de T.P. avec calculatrice, permettant, avant le cours, une introduction de la notion de loi de probabilité, par observation de la loi des grands nombres.

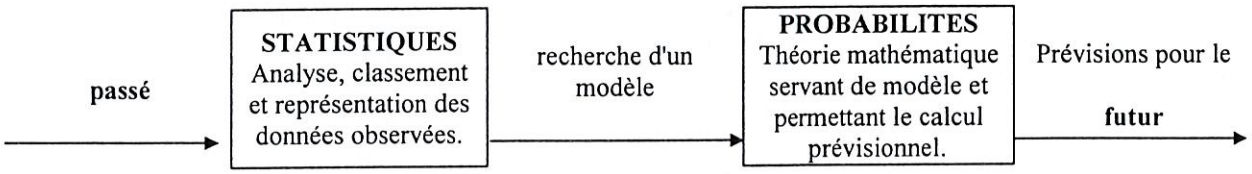
- Le premier exemple nous situe dans le cas, plus élémentaire, de l'observation de la loi des grands nombres dans le cas d'un seul événement (et de son contraire).
  La situation décrite est à la fois simple (facilement compréhensible, elle a un sens immédiat), avec une probabilité non intuitive, et est ensuite susceptible d'une "explication" géométrique, venant confirmer les observations par simulation.
- Le second exemple se place dans le cadre d'un loi de probabilité concernant 7 issues possibles. Il s'agit de la planche de *Galton*, dont le contexte, très parlant, plait aux élèves, et qui permet, de plus, un premier contact avec le triangle de *Pascal*.
- Le dernier exemple, à traiter éventuellement à la maison, met en lumière une probabilité inattendue à propos du jeu du loto.

T.P. CALCULATRICES

APPROCHE DE LA NOTION DE LOI DE PROBABILITE

"Comment oser parler des lois du hasard ?
Le hasard n'est-il pas l'antithèse de toute loi ?"
J. Bertrand – "Calcul des probabilités" – 1889.

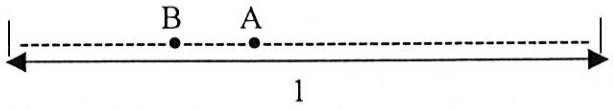
Prévoir et calculer des résultats dus au hasard, tel est le but des probabilités.



1 LE SEGMENT ALEATOIRE

L'expérience aléatoire

Sur un segment de longueur 1, on prend au hasard deux points A et B (différents des extrêmités).



On s'intéresse aux ***deux issues*** suivantes :

x_1 : la longueur AB est strictement supérieure à 0,5 ;

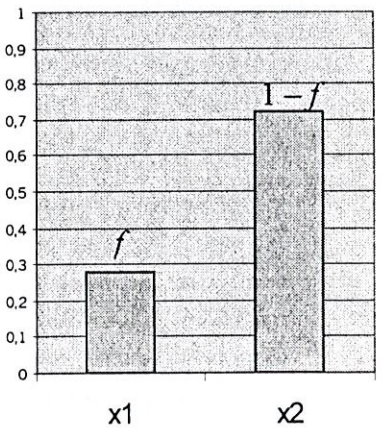
x_2 : la longueur AB est inférieure ou égale à 0,5.

On note E l'ensemble $\{x_1, x_2\}$.

Définir une ***loi de probabilité*** sur E consiste à rechercher un réel p entre 0 et 1 associé à x_1 (p est la probabilité de réalisation de l'événement x_1) et $1 - p$ associé à x_2.

Simulation

Pour déterminer p, on peut répéter un grand nombre de fois l'expérience et observer la fréquence f de réalisation de x_1.



La ***loi des grands nombres*** affirme :
La fréquence f de l'événement x_1 dans une suite de n expériences identiques et indépendantes, tend à se rapprocher de sa probabilité p quand n augmente indéfiniment.

Vous allez en faire l'expérience, par simulation.

1) Pourquoi l'instruction sur calculatrice abs(rand – rand) permet-elle de simuler l'expérience ? .....
.....
.....

Le programme suivant permet d'observer les variations de la fréquence de l'événement x_1 lors de la répétition $n = 500$ fois de l'expérience aléatoire.

CASIO Graph 25 → 100	T.I. 80 - 82 - 83	T.I. 89 - 92
ViewWindow 0,500 ,100,0,0.5,0.1↵ Graph Y= 0.25↵ For 1 → J To 4 ↵ 0 → N ↵ For 1 → I To 500 ↵ Abs(Ran# – Ran#) →A↵ A > 0,5 ⇒ N + 1 → N↵ Plot I , N ÷ I ↵ Next ↵ N÷500 ⁄⁄ Next	:FnOff :PlotsOff :0 → Xmin :500 → Xmax :100 → Xscl :0 → Ymin :0.5 → Ymax :0.1 → Yscl :DrawF 0.25 :For(J ,1 , 4) :0 → N :For (I , 1 , 500) :abs(rand – rand) → A : If A > 0.5 :N + 1 → N :Pt-On (I , N / I) :End :Disp N/500 :Pause :End	:FnOff :ClrDraw :PlotsOff :0 → xmin :500 → xmax :100 → xscl :0 → ymin :0.5 → ymax :0.1 → yscl :DrawFunc 0.25 :For j , 1 , 4 :0 → n :For i , 1 , 500 :abs(rand( ) – rand()) → a :If a > 0.5 :n + 1 → n :PtOn i , n/i :EndFor :Disp n/500 :Pause :EndFor

⇨ Pour obtenir certaines instructions :

• CASIO Graph 25 → 100 : **ViewWindow** par V-Window puis V.Win ; **Graph Y=** par Sketch GRPH Y= ; **For To Next** par PRGMpuis COM ; **Abs** par OPTN NUM ; **Ran#** par OPTN PROB ; > par PRGM REL ; ⇒ par PRGM JUMP ; **Plot** par Sketch PLOT puis Plot.

• TI 80 → 92 :

Utilisation possible de la fonction **CATALOG** (sur TI 83 – 85 – 92).

FnOff par Y-VARS On/Off... ou VARS Y-VARS puis Off ; **PlotsOff** par 2nd STATPLOT puis PLOTS ; **Xmin** par VARS Window ; **DrawF** par 2nd DRAW puis DRAW ; **For** par PRGM CTL ; **abs** par MATH NUM ; **rand** par MATH PRB ; **If** par PRGM CTL ; > par 2nd TEST (2nd MATH TEST sur TI 92) ; **Pt-On** par 2nd DRAW POINTS.

2) A quoi correspondent les axes des abscisses et des ordonnées sur le graphique de la calculatrice ? .....
.....
.....

3) Qu'observe-t-on sur les graphiques de la machine, lorsque le nombre de lancers augmente ? .....
.....
.....

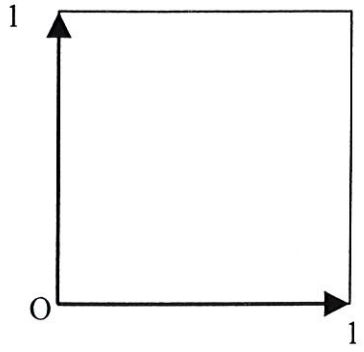
4) Noter, pour les quatre simulations les fréquences de l'événement x_1 après 500 réalisations indépendantes de l'expérience.

Simulation n° :	1	2	3	4
Fréquence f de l'événement x_1 pour $n = 500$ réalisations				

5) Quelle valeur peut-on, d'après vos expériences, attribuer à p ? .....

Un argument géométrique

On représente dans un repère orthonormal le résultat du tirage des points A et B par le point M(x, y) où x est l'abscisse de A sur le segment et y celle de B.

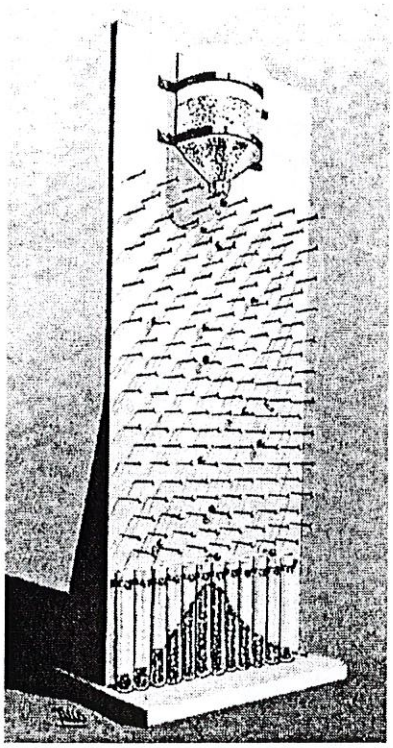


Hachurer la portion du carré correspondant aux points M pour lesquels $|x - y| > 0{,}5$.

6) Que vaut la surface hachurée ? Comparer avec les simulations.

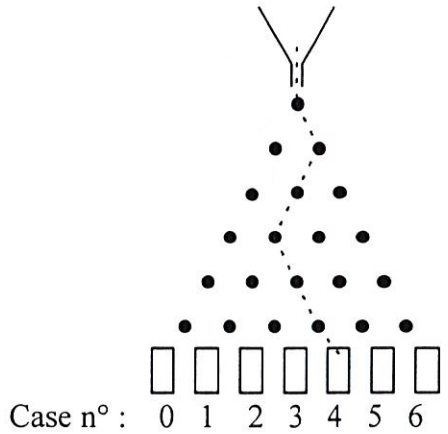
.....
.....
.....
.....
.....
.....
.....

2 LA PLANCHE DE GALTON



Considérons une planche de *Galton* où chaque bille rencontre 6 clous.
Chaque bille suit un trajet aléatoire, pour aboutir dans l'une des cases situées en bas. On suppose qu'à chaque clou rencontré, la bille a autant de chance d'aller à droite ou à gauche (la planche est bien verticale).

Francis GALTON (1822/1911) est un scientifique anglais, cousin de Charles DARWIN. S'intéressant essentiellement à la biologie et à l'anthropologie, il fut aussi un pionnier en statistique.



L'expérience aléatoire consiste à lâcher une bille.
L'ensemble E des issues possibles est $\{x_0, x_1, \ldots, x_6\}$ où x_i correspond à l'événement : "la bille arrive dans la case n°i".
Définir une ***loi de probabilité*** sur E consiste à rechercher des réels p_i entre 0 et 1 associés à chacun des x_i, avec $p_0 + \ldots + p_6 = 1$.

Pour déterminer les valeurs p_i, on peut répéter, par simulation, un grand nombre de fois l'expérience aléatoire (c'est à dire lancer un grand nombre de billes) et observer les fréquences f_i correspondant au nombre de billes arrivées dans chaque case.

La ***loi des grands nombres*** affirme :
Les fréquences f_i des événements x_i dans une suite de n expériences identiques et indépendantes, tendent à se rapprocher de leur probabilité p_i quand n augmente indéfiniment.

Simulation

Supposons qu'aller à droite correspond au tirage du chiffre 1, et qu'aller à gauche correspond au chiffre 0 (avec une chance sur deux dans chaque cas).
1) Pourquoi ceci peut-il être simulé par l'instruction : int(rand + 0.5) ?

.....
.....
.....
.....

Pour savoir dans quelle case arrive la bille, il suffit de compter les points :
Case n° 4 = ... points.
2) Entrer dans votre calculatrice le programme suivant (la partie sous la ligne pointillée concerne le tracé d'un histogramme et est facultative).

CASIO Graph 25 → 100	T.I. 80 - 82 - 83	T.I. 89 - 92
ClrList ↵ Seq(I,I,0,6,1) → List 1 ↵ Seq(0,I,1,7,1) → List 2 ↵ For 1 → I To 100 ↵ 0 → S ↵ For 1 → K To 6 ↵ Int (Ran# + 0.5) + S → S ↵ Next ↵ S + 1 → J ↵ List 2[J] + 1 → List 2[J] ↵ Next ↵ List 2	:ClrList L_1 , L_2 :seq(I,I,0,6,1) → L_1 :seq(0,I,1,7,1) → L_2 :For(I,1,100) :0 → S :For(K,1,6) :int(rand+0.5) + S → S :End :S + 1 → J :L_2(J) + 1 → L_2(J) :End :Disp L_2	:DelVar L1 , L2 :seq(i,i,0,6,1) → L1 :seq(0,i,1,7,1) → L2 :For i,1,100 :0 → s :For k,1,6 :int(rand( )+0.5) + s → s :EndFor :s + 1 → j :L2[j] + 1 → L2[j] :EndFor :Disp L2
S-WindMan ↵ ViewWindow 0,7,1, 0,50,10 ↵ 0 → Hstart ↵ 1 → Hpitch ↵ S-Gph1 DrawOn, Hist,List1,List2,Blue ↵ DrawStat	:Pause :Plot1(Histogram,L_1,L_2) :PlotsOn1 :0 → Xmin :7 → Xmax :1 → Xscl :0 → Ymin :50 → Ymax :10 → Yscl :DispGraph	:0 → xmin :7 → xmax :1 → xscl :0 → ymin :50 → ymax :10 → yscl :PlotsOn :NewPlot 1,4,L1,,L2 ,,,,1

⇨ Pour obtenir certaines instructions :
• CASIO Graph 25 → 100 : **ClrList** par PRGM CLR List ; **Seq** par OPTN LIST ; **List** par OPTN LIST ; **Int** par OPTN NUM ; **Ran#** par OPTN PROB ; **S-Wind Man** par SET UP Man ou SHIFT SET UP S-WIN ; **Hstart et pitch** par VARS STAT GRPH ; **S-Graph1** par F4(MENU) STAT GRPH GPH1 ; **Draw On** par F4(MENU) STAT DRAW ON ; **Hist** par F4(MENU) STAT GRPH ; **Blue** (si couleur) par STAT COLR ; **DrawStat** par PRGM DISP Stat.
• TI 80 82 83 : **ClrList** par STAT EDIT ; **Seq** par 2nd LIST OPS ; L_1 au clavier par 2nd ; **For End Pause** par PRGM CTL ; **int** par MATH NUM ; **rand** par MATH PRB ; **Disp** par PRGM I/O ; **Plot**

1(Histogram,L1,L2) par 2nd STAT PLOT PLOTS puis TYPE ; **Xmin** par VARS Window... ; **PlotsOn** par 2nd STAT PLOT PLOTS ; **Dispgraph** par PRGM I/O.
✳ Particularités sur certains modèles :
CASIO 6910 et 8930 : Les instructions 0 → Hstart ↵ et 1 → pitch ↵ sont à supprimer.

Pour n = **100** billes, indiquer les fréquences f_i de billes pour chaque case :

cases	0	1	2	3	4	5	6
Fréquences f_i							

Dans la seconde instruction For, remplacer 100 par 1000.
Pour n = **1000** billes (durée : 2 mn 30 sur CASIO 9990 ; 6 mn 30 sur TI 83):

cases	0	1	2	3	4	5	6
Fréquences f_i							

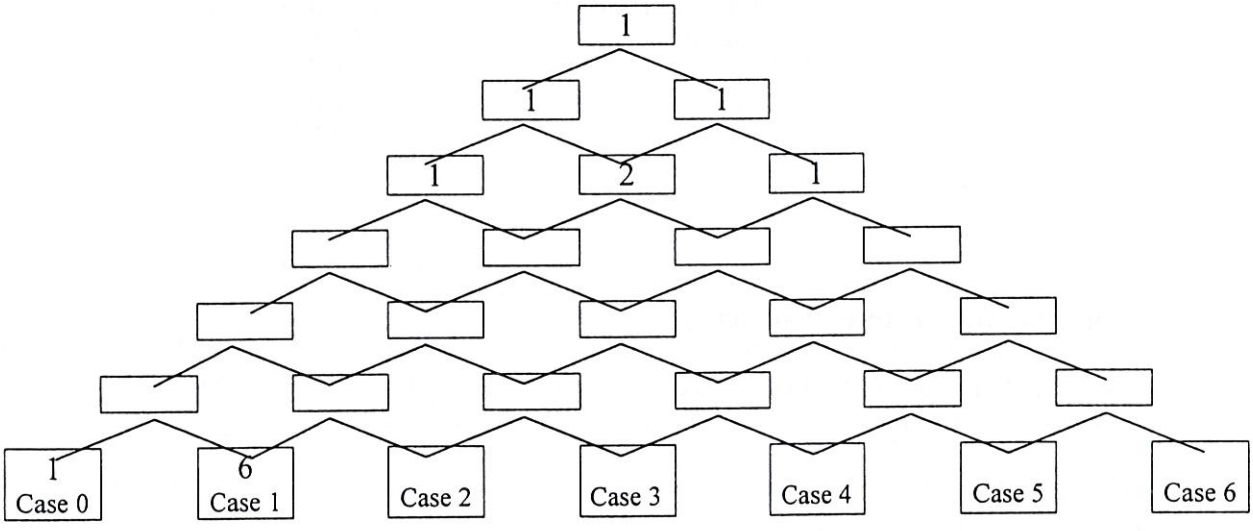
Moyenne des pourcentages obtenus dans la classe :

cases	0	1	2	3	4	5	6
Fréquences f_i							

Dénombrement des trajets possibles

Puisque la bille a, pour descendre, autant de chances de suivre un trajet plutôt qu'un autre (on dit que les trajets sont équiprobables), il s'agit de les compter. Pour faciliter ce dénombrement, vous allez exploiter un triangle, construit par *Pascal* pour résoudre une question de probabilité, et qui depuis porte son nom (quoique connu des Chinois et des Arabes bien avant).

3) Indiquer, dans chaque case, le nombre des chemins susceptibles d'y parvenir.



Déduire du triangle de Pascal le nombre de chemins possibles :

.....

En faisant le rapport des cas favorables aux cas possibles, indiquer dans le tableau ci-dessous les probabilités des événements x_i: "la bille arrive dans la case n°i."

Eléments de E	x_0	x_1	x_2	x_3	x_4	x_5	x_6
Probabilités p_i							

Comparer ces résultats théoriques à ceux obtenus par simulation.

.....

.....

.....

3 LE LOTO

Après avoir programmé votre calculatrice pour simuler le tirage du loto, vous déterminerez statistiquement, par l'étude des fréquences, la **probabilité d'avoir un tirage du loto comportant deux numéros consécutifs** au moins (numéro complémentaire excepté).

Entrer dans votre calculatrice le programme suivant, puis vérifier qu'il simule le tirage du loto (numéro complémentaire excepté) :

CASIO Graph 25 → 100	T.I. 80 - 82 - 83	T.I. 89 - 92
ClrList ↵ Seq(0,I,1,6,1) → List 1 ↵ Lbl 0 ↵ For 1 → I To 6 ↵ Int(1 + 49 Ran#) → List 1[I] ↵ Next ↵ For 1 → I To 5 ↵ I + 1 → J ↵ Lbl 1 ↵ List 1 [I] - List 1 [J] = 0 ⇒ Goto 0 ↵ J + 1 → J ↵ J ≤ 6 ⇒ Goto 1 ↵ Next ↵ **List 1** **Stop**	:ClrList L_1 :seq(0,I,1,6,1) → L_1 :Lbl 0 :For (I , 1 , 6) :int(1 + 49 rand) → L_1(I) :End :For (I , 1 , 5) :For (J , I + 1 , 6) :If L_1 (I) – L_1 (J)= 0 :Goto 0 :End :End :**Disp L_1**	:DelVar L1 , L2 :seq(0,I,1,6,1) → L_1 :Lbl a :For i , 1 , 6 :int(1 + 49 rand( )) → L1[i] :EndFor :For i , 1 , 5 :For j , i + 1 , 6 :If L1[i] – L1[j] = 0 :Goto a :EndFor :EndFor :**Disp L1**

⇨ **Pour obtenir certaines instructions :**
• CASIO 6910 → 9990 : **ClrList** par PRGM CLR List ; **Seq** par OPTN LIST ; **List** par OPTN LIST ; **Int** par OPTN NUM ; **Ran#** par OPTN PROB ; [au clavier ; = et ≤ par PRGM REL ; **Stop** par PRGM CTL.
• TI 80 → 83 : **Clrlist** par STAT ; L_1 au clavier par 2nd ; **Seq** par 2nd LIST OPS.

Modifier le programme précédent pour dénombrer, sur 100 tirages du loto, les tirages contenant au moins deux numéros consécutifs.
Ajouter en début de programme :

0 → S (initialisation du compteur du nombre de tirages)

0 → T (initialisation du compteur des tirages contenant au moins deux numéros consécutifs).

Remplacer les instructions d'affichage (en gras dans le programme précédent) par les suivantes.

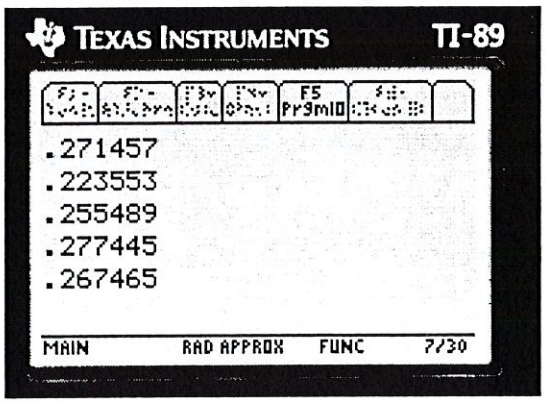
CASIO Graph 25 → 100	T.I. 80 - 82 - 83	T.I. 89 - 92
S + 1 → S ↵ For 1 → I To 5 ↵ I + 1 → J ↵ Lbl 2 ↵ If (List 1 [I] - List 1 [J])^2 = 1 ↵ Then T + 1 → T ↵ Goto 3 ↵ I-End ↵ J + 1 → J ↵ J ≤ 6 ⇒ Goto 2 ↵ Next ↵ Lbl 3 ↵ S ≤ 99 ⇒Goto 0 ↵ T ÷ S ↵ Stop	:S + 1 → S :For (I , 1 , 5) :For (J , I+1 , 6) :If L_1 (I) – L_1 (J))^2 = 1 :Then :T + 1 → T :Goto 1 :End :End :End :Lbl 1 :If S ≤ 99 :Goto 0 :Disp T / S	:s + 1 → s :For i , 1 , 5 :For j , i+1 , 6 :If (L1[i] – L1[j])^2 = 1 :Then :t + 1 → t :Goto b :EndIf :EndFor :EndFor :Lbl b :If s ≤ 99 :Goto a :Disp t / s

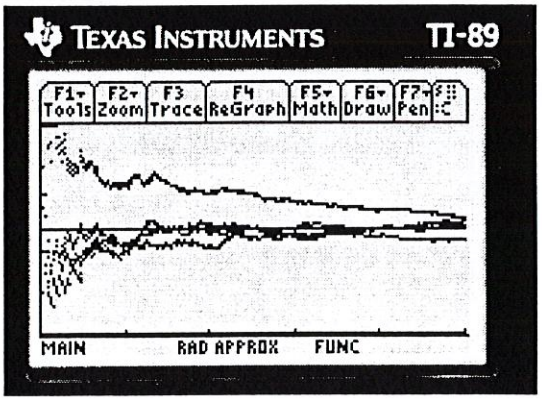
Quelle est la fréquence, sur 100 tirages, des tirages avec numéros consécutifs ? .....
Faire la moyenne des fréquences obtenues dans la classe : .....
Estimer la probabilité de l'événement : "le tirage des 6 numéros du loto comporte au moins deux numéros consécutifs".
.....
.....

Eléments de solution
"APPROCHE DE LA NOTION DE LOI DE PROBABILITE"

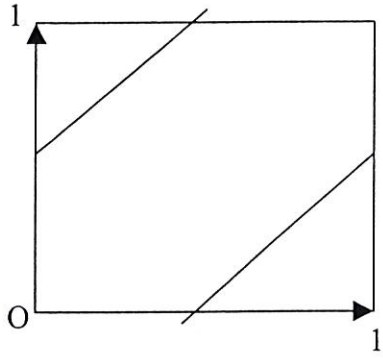
1- SEGMENT ALEATOIRE

1) La première instruction rand correspond au tirage aléatoire de l'abscisse de A sur le segment, dans l'intervalle]0, 1[. La seconde instruction rand correspond au tirage de l'abscisse de B.
Le calcul abs(rand – rand) est celui de la distance AB.
2) L'axe des abscisses correspond au nombre *n* d'expériences (de 0 à 500), celui des ordonnées à la fréquence de l'événement x_1 (de 0 à 0,5).
Cette question oblige à s'interroger sur la signification du graphique (bien que la réponse soit dans le texte, elle n'est souvent pas évidente pour l'axe des ordonnées).
3) On observe que, lorsque le nombre d'expériences augmente, les fluctuations de la fréquence de x_1 diminuent. Cette fréquence se rapproche de 0,25.
4) Quelques simulations...





5) Il semble que : $p \approx 0{,}25$.
6)



On trace les droites d'équation $y = x - ½$ et $y = x + ½$. La surface des deux triangles correspondant à $|x - y| > ½$ vaut ¼. Puisque M est choisi au hasard dans le carré unité, les probabilités sont proportionnelles aux surfaces d'où $p = ¼$.

2- PLANCHE DE GALTON ET TRIANGLE DE PASCAL

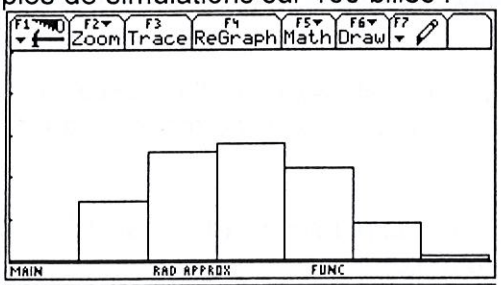
Avec la planche de Galton, le succès auprès des élèves est garanti. C'est un procédé concret et visuel.

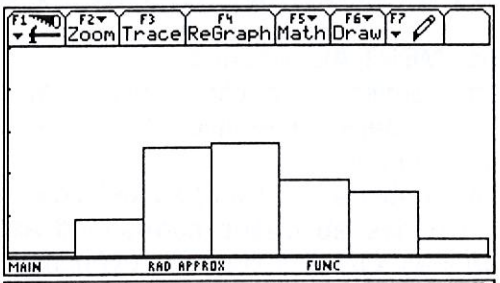
L'objectif de cette activité est de montrer que la répartition des billes est prévisible et qu'un simple dénombrement suffit à la justifier.

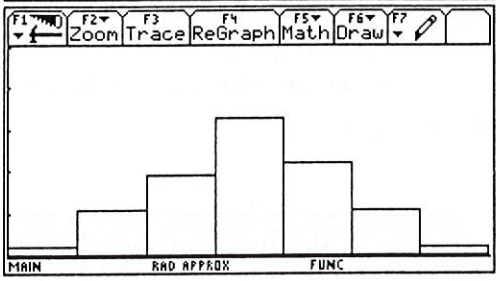
Les probabilités théoriques sont :

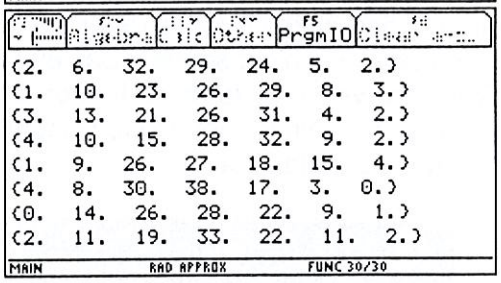
cases	0	1	2	3	4	5	6
probabilités ≈	0,016	0,094	0,234	0,313	0,234	0,094	0,016

Exemples de simulations sur 100 billes :

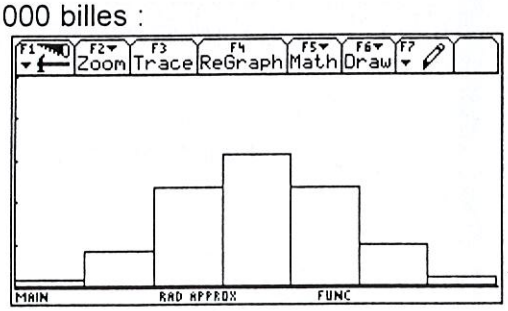


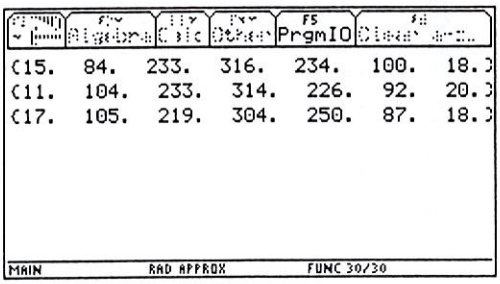






Sur 1000 billes :

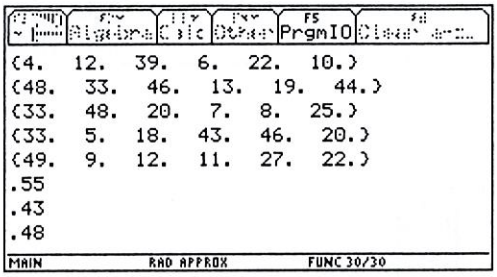




3- LE LOTO

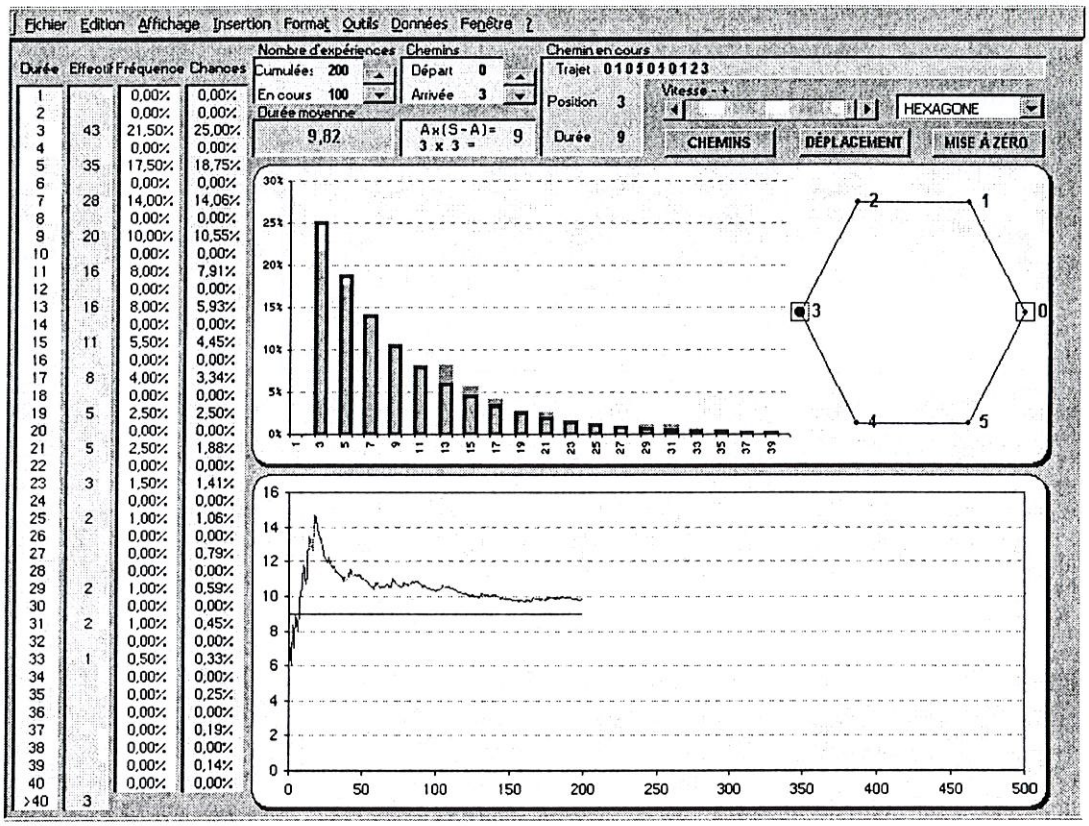
Le jeu du loto passionne évidemment nos élèves. Quant au résultat que l'on obtient ici (théoriquement : $1 - \frac{C_{44}^6}{C_{49}^6} \approx 0{,}495$), il est assez surprenant. L'astuce consiste à établir une bijection entre l'ensemble des tirages de 6 numéros non consécutifs de {1; ... ; 49} (sans remise, puis rangés en ordre croissant) et celui des tirages de 6 numéros quelconques de {1 ; ... ; 44} (sans remise, puis rangés en ordre croissant) :
$(n_1 , n_2 , \dots , n_6) \mapsto (n_1 , n_2 - 1 , \dots , n_6 - 5)$.
Cependant, les programmes étant plus longs, ils pourront être laissés aux élèves motivés.

Tirages du Loto puis fréquences, sur 100 tirages du Loto, des tirages avec au moins deux n^{os} consécutifs :



4 – OBSERVATION, SOUS EXCEL, DE LA CONVERGENCE DE LA DISTRIBUTION DES FREQUENCES OBSERVEES

On peut considérer l'exemple d'une ***marche aléatoire sur un hexagone***.
L'image d'écran ci-dessous est obtenue à partir d'un programme effectué sous Excel, et diffusé sur le CD-Rom accompagnant la brochure "Simulation en seconde" de l'IREM de Paris-Nord.



Le départ s'effectue au sommet noté 0 et l'arrivée à celui noté 3. A chaque sommet, le choix du sommet suivant se fait avec une chance sur deux.
On s'intéresse à la durée du trajet. On limite les résultats possibles au temps 40, soit 20 éléments dans l'ensemble E des résultats possibles (le temps de traversée est impair : 3, 5, …, 39 ou > 40).
On observe la "convergence" (non définie), ou stabilisation, de l'histogramme des fréquences observées sur n expériences répétées indépendamment, vers un histogramme "théorique" correspondant à la loi de probabilité sur E.
On peut également observer (graphique du bas) la convergence du temps moyen de traversée sur n expériences $\overline{x}_n$ vers la valeur 9 (on montre que la variable aléatoire T correspondant au temps de traversée a pour espérance 9).

II – MODELISATION D'UNE EXPERIENCE ALEATOIRE

"On modélise énormément. C'est maintenant la principale activité de l'ingénieur."

Nicolas Bouleau[15] - 1999.

"La modélisation, qui tente de s'habiller des habits de la science, demande à être critiquée [...] d'où l'importance de l'introduction de la modélisation dans l'enseignement."

Nicolas Bouleau- Commission *Kahane* – mars 2001.

1 – A PROPOS DE LA NOTION DE MODELE

Il s'agit ici de modèles probabilistes.

> *"**Modéliser une expérience aléatoire**, c'est lui associer une loi de probabilité."*
> *Document d'accompagnement du programme de 1ère S.*

La modélisation occupe une part importante des activités scientifiques (par exemple en sciences de l'ingénieur) pour comprendre ou prévoir ; les mathématiques fournissent le langage de la modélisation. C'est au travers de modèles mathématiques que nous affinons notre compréhension du monde. On connaît la célèbre phrase de *Galilée*, dans *L'essayeur* en 1623 :

"La philosophie est écrite dans ce très vaste livre qui est éternellement ouvert devant nos yeux – je veux dire l'Univers – mais on ne peut le lire avant d'avoir appris la langue et s'être familiarisé avec les caractères dans lesquels elle est écrite. Elle est écrite en langue mathématique et ses lettres sont des triangles, des cercles et d'autres figures géométriques, moyens sans lesquels il est humainement impossible de comprendre un seul mot, sans lesquels l'on erre en vain dans un obscur labyrinthe."

Aux triangles et aux cercles, s'ajoutent désormais les lois de probabilité. Si la nature n'est sans doute pas mathématique, comme le conçoit *Galilée*, les mathématiques constituent du moins, le moyen d'humainement la comprendre.

Nietzsche dénonce[16] *"cette foi dont se satisfont aujourd'hui tant de savants matérialistes qui croient que le monde doit avoir sa mesure dans nos petites échelles, et son équivalent dans notre petite pensée [...]. Que seule vaille une interprétation du monde qui ne permet que de compter, de peser, de voir et toucher, c'est balourdise et naïveté si ce n'est démence ou idiotie."*

> *"Les élèves devront bien distinguer ce qui est empirique (du domaine de l'expérience) de ce qui est théorique."* *Document d'accompagnement du programme de 1ère S.*

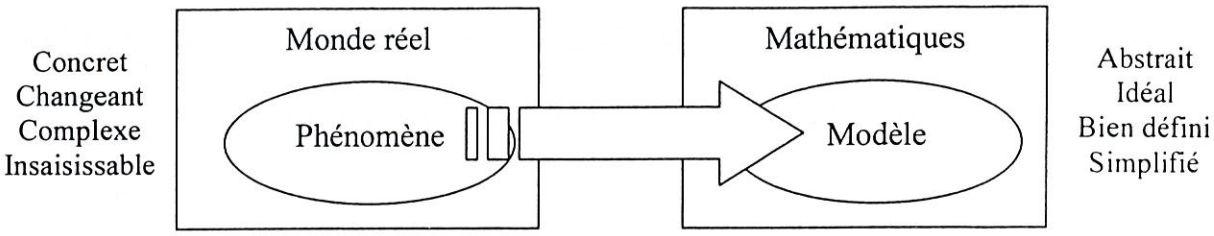
Un modèle est une représentation théorique d'une situation réelle dont il cherche à fournir la meilleure description possible. Il est fondé sur des observations et des hypothèses. Un modèle mathématique, constitué d'objets mathématiques, contient essentiellement des *variables*, concernant le phénomène et observables, dépendant de *paramètres* fixes, généralement inconnus.

Un exemple de modèle est le modèle linéaire fourni par la méthode des moindres carrés.

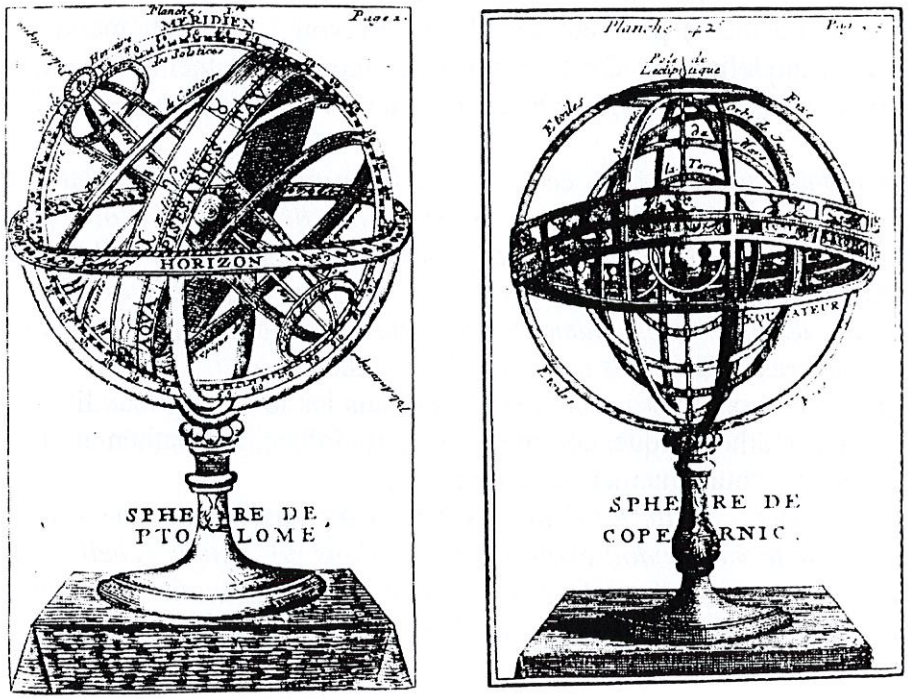
---

[15] *N. Bouleau - "Philosophie des mathématiques et de la modélisation"* – L'Harmattan 1999.

[16] Dans *"Le gai savoir"*, cité par *N. Bouleau.*



Cette notion d'adéquation aux phénomènes apparaît très tôt en astronomie, longtemps la plus développée des sciences. L'astronome *Ptolémée* (II[e] siècle ap. J;-C.) dit ainsi dans l'*Almageste* [17] : *"Il faut, autant qu'on le peut, adapter les hypothèses les plus simples aux mouvements célestes ; mais si elles ne suffisent pas, il faut en choisir d'autres qui les expliquent mieux."* Aux XVII[e] et XVIII[e] siècles, des débats sur le choix des "modèles" (on parle plutôt alors de "systèmes") apparaissent dans le cadre de problèmes d'astronomie et de géodésie (mouvements des planètes du système solaire, "figure" de la Terre). Ainsi, l'astronome *Jean-Sylvain Bailly* affirme, dans son "*Histoire de l'Astronomie moderne*" (1782) : *"N'espérons pas de jamais rien connaître dans les sciences de la nature sans les systèmes. [...] La seule loi qu'on leur impose est d'être vraisemblables, aussi bien que nécessaires à l'observation des faits observés."*



*Nicolas Bion – "L'usage des globes céleste et terrestre et des sphères suivant les **différents systèmes** du monde" – 3[e] édition 1710*

A la fin du XVI[e] siècle, *Tycho Brahé* parvient, dans ses observations du ciel, à une perfection inégalée, d'une précision jamais atteinte. Sur la base de ces observations, *Kepler* pourra établir ses lois des mouvements planétaires : un nouveau "modèle" du monde qui sera justifié par *Newton*.
Voici quelques lignes extraites de la "*Mécanique de l'Astronomie rénovée*[18]", publié en 1598 par *Tycho Brahé*.

[17] Cité dans *"Figures du ciel"* – Seuil/BnF – 1998.
[18] L'ouvrage, traduit du latin par *Jean Peyroux*, est paru en 1978 aux éditions Bergeret à Bordeaux.

TYCHONIS BRAHE

ASTRONOMIÆ INSTAURATÆ MECHANICA

"Car cela est le principal de tout, que de nombreuses et longues observations venant du Ciel soient prises dans des Instruments Astronomiques opportuns et non sujets à l'erreur, qui ensuite sont ***mises en ordre*** *au moyen de la Géométrie par des Hypothèses imaginées convenant aux quantités continues et au mouvement circulaire et uniforme (que Naturellement les choses Célestes et recherchent et poursuivent sans interruption) ; assurément au moyen de l'Arithmétique dans les quantités discrètes pour que les révolutions et les lieux des corps Célestes soient certains aux temps que tu veux. En vérité parmi tous ceux qui travaillèrent activement à cette chose, du moins les observations qui parviennent par là jusqu'à nous, furent notées par Timocharis, Hipparque, Ptolémée, Albategni, le Roi Alphonse, et au siècle précédent par Copernic bien que les enseignements des deux précédents sur ces choses dépendent de la relation de Ptolémée."*

Timocharis : astronome grec d'Alexandrie (vers – 230). *Albategni* ou *Al Battani* : astronome arabe mort en 929. *Alphonse X de Castille* (1232-1284) fit dresser les "tables Alphonsines".

Ce qui a profondément bouleversé la modélisation ces dernières années, c'est l'ordinateur, qui permet maintenant de faire de la modélisation un puissant outil de recherche. D'après *Nicolas Bouleau, "par le développement de l'informatique le rôle des mathématiques dans les applications s'est profondément modifié au point de changer de nature. Le principal changement dans les utilisations des mathématiques vient de l'usage de la notion de "modèle" qui s'est généralisé et intensifié dans tous les secteurs d'activité économique depuis les quelques décennies où l'informatique a pris son essor."*

Attention cependant, d'après les programmes, il ne s'agit, en aucun cas, de faire un cours sur la modélisation, mais seulement, par quelques activités, d'en montrer certaines caractéristiques.

> *"Il ne s'agit en aucun cas d'avoir des discours généraux sur les modèles et la modélisation."* *Document d'accompagnement du programme de 1ère S.*

Un modèle n'est ni juste, ni faux, il est plus ou moins bien adapté à la description d'un phénomène. Augmenter le nombre de paramètres, par exemple, permet souvent de l'améliorer. De plus, pour une situation donnée, il peut y avoir de nombreux modèles, dont le choix dépend de l'usage que l'on souhaite en faire. Citons encore Nicolas Bouleau : "Même avec le secours de l'expérience, plusieurs modèles subsistent pour représenter une même réalité. La meilleure critique devant une modélisation consiste en fait à créer un autre modèle : la solution n'est jamais unique."

> *"Le choix d'un modèle à partir de données expérimentales ne sera pas abordé dans l'enseignement secondaire."*
> *"La modélisation ne relève pas d'une logique du vrai ou du faux. Un modèle n'est ni vrai ni faux : il peut être validé ou rejeté au vu de données expérimentales. Une des premières fonctions de la statistique dite inférentielle est d'associer à une expérience aléatoire un modèle... et de définir des procédures de validation."*
> *Doc. d'accompagnement de 1ère S.*

Parmi ces procédures de validation, on peut citer la prise en compte du coefficient de corrélation linéaire pour la méthode des moindres carrés (assez rudimentaire) et, de façon plus générale, la procédure des ***tests d'hypothèses***, qui permet de quantifier les ***risques*** d'erreur dans l'acceptation ou le refus d'un modèle, et qui sera abordée lors de la 5ème séance (test du khi-2 entre autres).

Une dernière remarque, dans ces généralités, un bémol apporté par *Nicolas Bouleau*, à propos d'un usage parfois abusif des modèles probabilistes dans les sciences de l'ingénieur :

"Le calcul des probabilités est une des plus belles choses des mathématiques appliquées et il est de nombreux domaines où il rend des services immenses : trafic, gestion de stock, traitement du signal, thermodynamique statistique, etc.

Mais sous la pression de l'opinion publique, pour évaluer les ***risques****, les ingénieurs ont tendance à mettre des probabilités partout. A un récent colloque de génie civil, des modèles probabilistes ont été présentés à propos de : calcul des structures – bruit des infrastructures – mécanique des sols – fiabilité des ponts – séismes – usure des chaussées – bétons à fibres – comportement du bois – rupture des céramiques – endommagement des matériaux composites.*

Il y a là une tendance qui devient abusive pour deux raisons. D'abord les modèles probabilistes sont plus difficile à réfuter que les modèles déterministes, puisqu'une seule mesure ne suffit pas à disqualifier un résultat. Ensuite, ce qu'on redoute en situation de risque ce sont les évènements rares, car ils sont chargés de signification. Or les queues de lois de probabilité sont toujours mal connues, précisément parce qu'elles ne se nourrissent pas de faits fréquents."

Au lycée, on n'en est pas encore à abuser de la modélisation probabiliste et son importance, tant comme domaine d'application de plus en plus fréquent des mathématiques, que comme "fait de société", nous contraint à y initier nos élèves. Récemment dans le journal *Le Monde*, on lisait que les rapports "d'experts" à propos de la baisse des ressources de poissons, demandés par la Commission européenne, provoquaient la colère des marins – pêcheurs. Le journaliste[19] concluait : "*Pas la peine de "contester le thermomètre" et de s'en prendre à des chercheurs qui ne prétendent pas avec leurs modèles traduire toute la réalité des faits mais donner avec leurs simulations des tendances qui depuis des années paraissent se confirmer." Nous ne sommes pas là pour tuer la pêche et les marins, martèle l'un d'eux. Nous sommes là pour sonner l'alarme et aider les politiques à prendre des mesures de précaution"."*

2 – ACTIVITES ELEVES EN 1ère

On donne, dans les pages qui suivent :

- Un exemple de T.P. sur calculatrices, mêlant simulation et modélisation.
- Un exemple d'exercice, montrant que plusieurs modélisations sont possibles, lorsque la situation n'est pas clairement définie (ce qui est nécessairement le cas dans la pratique de l'ingénieur par exemple). Cet exercice est inspiré du document d'accompagnement des programmes.

[19] *J-F. Augereau – Le Monde* juillet 2002.

T.P. SUR CALCULATRICES

MODELISATION D'UNE EXPERIENCE ALEATOIRE

1 SUR LES BANCS PUBLICS
Du choix déterminant d'un modèle

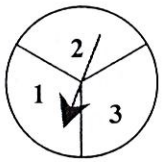
On souhaite répondre au petit problème suivant :

Dans un square se trouvent trois bancs à deux places. Deux personnes arrivent et s'assoient au hasard. Quelle est la probabilité qu'elles s'assoient côte à côte ?

On décide de *simuler* l'expérience un grand nombre de fois, "pour voir". Mais cela nous oblige à *modéliser* la situation.
L'expression "au hasard" indique par convention un modèle où le choix correspond à une probabilité équirépartie. Mais comment "choisit"-t-on sa place ?

1) PREMIER MODELE : Le choix des bancs

⇨ SIMULATION :
On numérote les trois bancs 1, 2, 3.
Chacune des personnes tire au hasard le numéro du banc sur lequel elle va s'asseoir, à l'aide de la roulette ci-contre.
Il suffit d'étudier le pourcentage des expériences où le même numéro a été tiré.



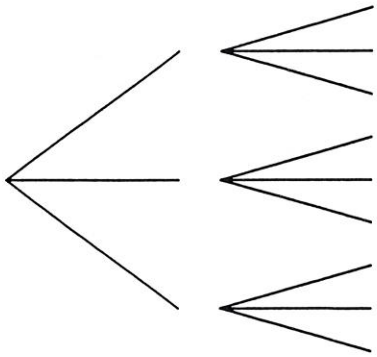
Justifier le fait que l'instruction Int(1 + 3Ran#) ou int(1 + 3rand) simule la roulette précédente. .....
.....

Le programme ci-contre simule 100 expériences décrites précédemment.
Il fournit le nombre d'expériences pour lesquelles le même numéro est sorti deux fois.

CASIO Graph 25 → 100	T.I. 80 - 82 - 83	T.I. 89 - 92
0 → S ↵ For 1 → I To 100 ↵ Int(1 + 3Ran#) → A ↵ Int(1 + 3Ran#) → B ↵ A = B ⇒ S + 1 → S ↵ Next ↵ S	:0 → S :For(I , 1 , 100) :int(1 + 3rand) → A :int(1 + 3rand) → B :If A = B :S + 1 → S :End :Disp S	:0 → s :For i , 1 , 100 :int(1 + 3rand()) → a :int(1 + 3rand()) → b :If a = b :s + 1 → s :EndFor :Disp s

Exécuter ce programme 6 fois, noter les résultats obtenus et calculer leur moyenne.
.....
.....
.....
.....
.....
.....

⇨ CALCUL DES PROBABILITES :



A l'aide de l'arbre ci-contre, où la première bifurcation correspond au choix équiprobable de la première personne et la seconde bifurcation à celui de la seconde personne, calculer la probabilité de l'événement "les deux personnes sont assises côte à côte" selon ce premier modèle.

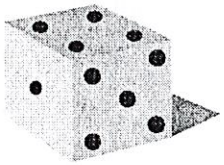
.....
.....
.....
.....
.....
.....

2) SECOND MODELE : Le choix des places

⇨ SIMULATION :

On numérote les places de 1 à 6. Le premier banc correspond aux places 1 et 2, le deuxième banc aux places 3 et 4, et le dernier aux places 5 et 6.
La première personne lance un dé pour choisir sa place.
La seconde personne fait de même, mais relance le dé si la place tirée est occupée.
Il suffit d'étudier le pourcentage d'expériences où les numéros des places tirées correspondent au même banc.

1	2		3	4		5	6
Banc 1			Banc 2			Banc 3	



Si l'on note A le numéro tiré par la première personne et B celui, différent, tiré par la seconde personne, pourquoi la connaissance de la somme A+B et du produit A×B caractérise-t-elle les nombres A et B ?

.....
.....
.....
.....

Le programme suivant simule 100 expériences décrites précédemment.
Il fournit le nombre d'expériences pour lesquelles les numéros sortis correspondent au même banc.
Exécuter ce programme 6 fois, noter les résultats obtenus et calculer leur moyenne.

CASIO Graph 25 → 100	T.I. 80 - 82 - 83	T.I. 89 - 92
0 → S↵ For 1 → I To 100 ↵ Int(1 + 6 Ran#) → A ↵ Lbl 1 ↵ Int(1 + 6 Ran#) → B ↵ A = B ⇒ Goto 1 ↵ A+B=3 ⇒ S+1→S ↵ A+B=7 And AB=12 ⇒ S+1→S ↵ A+B=11 And AB=30 ⇒ S+1→S ↵ Next ↵ S	:0 → S :For(I , 1 , 100) :int(1 + 6rand) → A :Lbl 1 :int(1 + 6rand) → B :If A = B :Goto 1 :If A+B=3 :S + 1 → S :If A+B=7 and AB = 12 :S + 1 → S :If A+B=11 and AB = 30 :S + 1 → S :End :Disp S	:0 → s :For i , 1 , 100 :int(1 + 6rand()) → a :Lbl c :int(1 + 6rand()) → b :If a = b :Goto c :If a+b=3 :s + 1 → s :If a+b=7 and a*b = 12 :s + 1 → s :If a+b=11 and a*b = 30 :s + 1 → s :EndFor :Disp s

⇨ **Pour obtenir certaines instructions** :

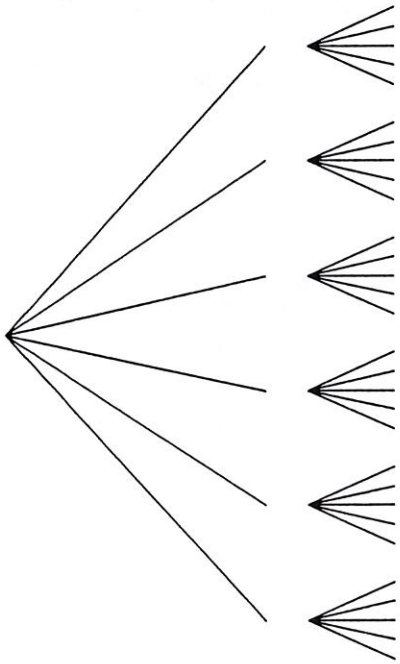
- CASIO Graph 30 et plus : **And** par OPTN LOGIC.
- TI 82 83 : **and** par 2nd TEST LOGIC.

✳ **Particularités sur certains modèles** :

- CASIO Graph 25 : En l'absence du connecteur And, remplacer A+B=7 And AB=12 ⇒ S+1→S ↵ par If A + B = 7 ↵ Then AB = 12 ⇒ S + 1 → S ↵ IfEnd.
- TI 80 : En l'absence du connecteur and, remplacer :If A+B=7 and AB = 12 ↵ :S + 1 → S par :If A + B = 7 ↵ :Then ↵ :If AB = 12 ↵ :S + 1 → S ↵ End.

.....
.....
.....
.....
.....

⇨ CALCUL DES PROBABILITES :

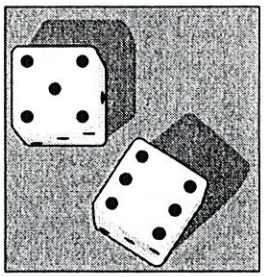


A l'aide de l'arbre ci-contre, calculer la probabilité de l'événement "les deux personnes sont assises côte à côte" selon ce second modèle.

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

Que conclure ?
L'énoncé ne dit pas comment le hasard intervient pour déterminer les positions des deux personnes. L'équiprobabilité porte-t-elle sur le choix du banc ou sur le choix d'une place ? Selon l'hypothèse suivie, la réponse est différente. L'énoncé de départ est incomplet et la question posée n'a pas de sens précis. Pour qu'une expérience aléatoire puisse être étudiée, il est nécessaire de définir un "protocole expérimental" permettant de la répéter.

2 SOMME DE DEUX DES
A la recherche de l'équiprobabilité



On lance deux dés et on s'intéresse à la somme des faces apparues.

Lancer dix fois les dés et compléter le tableau ci-dessous :

Sommes	2	3	4	5	6	7	8	9	10	11	12
effectifs											
fréquences											

Pensez-vous que les différentes sommes ont la même probabilité de survenir ?

.....

Pensez-vous que le fait de lancer les dés l'un après l'autre, ou en même temps, influe sur les probabilités des sommes ?

.....

MODELISATION(S)

Pour modéliser cette expérience aléatoire, on a les deux propositions suivantes.

• Modèle n°1 : on distingue les deux dés

On suppose que les dés sont de couleur différente : bleu et rouge. On note d'abord le résultat du dé bleu, puis celui du dé rouge.
L'ensemble E_1 des résultats possibles au lancer des deux dés est alors constitué des couples apparaissant dans ce tableau :

Dé rouge / Dé bleu	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

On muni E_1 de l'équiprobabilité et on désigne par A_k l'événement "la somme des dés vaut k".
Déterminer, dans ce modèle, les probabilités des événements A_k.

k	2	3	4	5	6	7	8	9	10	11	12
$P(A_k)$											

• Modèle n°2 : on ne distingue pas les deux dés

On considère que l'on lance les deux dés en même temps et qu'ils sont indiscernables. Il n'y a donc pas lieu de distinguer dans E des éléments tels que (3, 1) et (1, 3).

L'ensemble E_2 des résultats possibles au lancer des deux dés est alors constitué des couples apparaissant dans le tableau suivant :

Un dé / L'autre dé	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2		(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3			(3, 3)	(3, 4)	(3, 5)	(3, 6)
4				(4, 4)	(4, 5)	(4, 6)
5					(5, 5)	(5, 6)
6						(6, 6)

On muni E_2 de l'équiprobabilité et on désigne par A_k l'événement "la somme des dés vaut k".

Déterminer, dans ce modèle, les probabilités des événements A_k.

k	2	3	4	5	6	7	8	9	10	11	12
$P(A_k)$											

Quel modèle doit-on privilégier ? Les joueurs, qui ont l'expérience, ont peut-être leur opinion. Comme vous ne pouvez pas faire un grand nombre de lancers, vous allez les simuler.

SIMULATIONS

Sur 100 lancers

Le programme suivant simule 100 lancers de deux dés et affiche la distribution correspondante des sommes des faces.

CASIO Graph 25 → 100	TI 80 – 82 – 83	TI 89 - 92
Seq(O , I , 1 , 12 , 1) → List 1 ↵ For 1 → J To **100** ↵ Int(1+6Ran#) + Int(1+6Ran#) → N↵ List 1[N] + 1 → List 1[N] ↵ Next ↵ List 1 ÷ **100** → List 1 ↵ List 1 ↵ Stop	:seq(I , I , 1 , 12 , 1) → L_1 :seq(0 , I , 1 , 12 , 1) → L_2 :For(J , 1 , **100**) :int(1+6rand) + int(1+6rand) → N :L_2(N) + 1 → L_2(N) :End :L_2/**100** → L_2 :Disp L_1 , L_2	:seq(i , i , 1 , 12 , 1) → L1 :seq(0 , i , 1 , 12 , 1) → L2 :For j , 1 , **100** :int(1+6rand()) + int(1+6rand()) → n :L2[n] + 1 → L2[n] :EndFor :L2/**100** → L2 :Disp L1 , L2

⇨ **Pour obtenir certaines instructions :**

• CASIO : **Seq** par OPTN LIST ; **List** par OPTN LIST ; **For To Next** par PRGM COM ; **Int** par OPTN NUM ; **Ran#** par OPTN PROB ; = [au clavier.

• TI 80 82 83 : **Seq** par 2nd LIST OPS ; L_1 au clavier par 2nd ; **For End** par PRGM CTL ; **int** par MATH NUM ; **rand** par MATH PRB ; **Disp** par PRGM I/O.

Pour visualiser les listes 1 et 2 complètement, faire, après l'exécution du programme :

CASIO Graph 25 → 100	TI 80 82 83	TI 89 92
MENU STAT	STAT puis EDIT 1:Edit... puis ENTER	HOME L1 ENTER puis L2 ENTER

Exécuter trois fois le programme et inscrire les résultats dans le tableau suivant.

Sommes	2	3	4	5	6	7	8	9	10	11	12
fréquences simulation 1											
fréquences simulation 2											
fréquences simulation 3											
		Δ max			sur 3 simulations de 100 lancers						

Pour les sommes 4 et 5, indiquer dans le tableau l'écart maximum entre les 3 fréquences fournies par les 3 simulations.

Sur 1000 lancers

Dans le programme précédent, remplacer **100** (en gras) par **1000**.
Exécuter trois fois le programme (durée ≈ 1 mn) et compléter le tableau :

Sommes	2	3	4	5	6	7	8	9	10	11	12
fréquences simulation 1											
fréquences simulation 2											
fréquences simulation 3											
		Δ max			sur 3 simulations de 1000 lancers						

Comparer les écarts des fréquences de sortie, entre les différentes simulations, lorsqu'on fait 100 lancers et lorsqu'on fait 1000 lancers (considérer, par exemple, en particulier les sommes 4 et 5) : .....
.....
.....

Conclusion

En considérant en particulier le cas des sommes 4 et 5 dans les modèles 1 et 2, quel est des deux modèles, celui qui est le plus proche de l'expérience ?

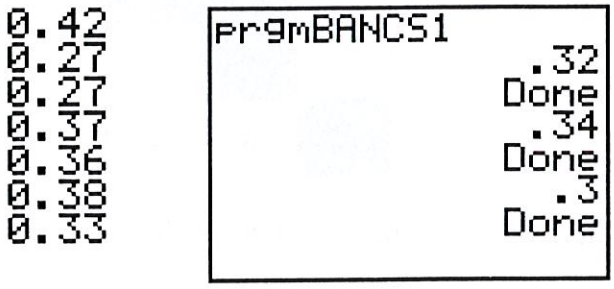
.....
.....
.....
.....
.....
.....
.....

Eléments de solution
"MODELISATION D'UNE EXPERIENCE ALEATOIRE"

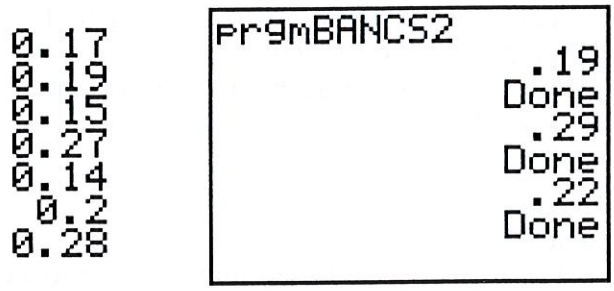
1 - SUR LES BANCS PUBLICS

L'activité de modélisation, rare dans nos exercices scolaires, est souvent déterminante dans les applications des probabilités. Chacun des modèles proposés est "raisonnable" et les simulations correspondantes sont calquées dessus. On observe des réactions parfois épidermiques de certains élèves qui refusent d'adhérer à l'un des modèles. Cette activité a pour objectif de montrer le caractère très relatif des résultats probabilistes et incite à une certaine tolérance scientifique.

Simulations du modèle 1 :



Simulations du modèle 2 :



2 – SOMME DE DEUX DES

modèle 1 :

Sommes	2	3	**4**	**5**	6	7	8	9	10	11	12
fréquences théoriques	1/36 ≈ 0,03	2/36 ≈ 0,06	**3/36 ≈ 0,08**	**4/36 ≈ 0,11**	5/36 ≈ 0,14	6/36 ≈ 0,17	5/36 ≈ 0,14	4/36 ≈ 0,11	3/36 ≈ 0,08	2/36 ≈ 0,06	1/36 ≈ 0,03

modèle 2 :

Sommes	2	3	**4**	**5**	6	7	8	9	10	11	12
fréquences théoriques	1/21 ≈ 0,05	1/21 ≈ 0,05	**2/21 ≈ 0,10**	**2/21 ≈ 0,10**	3/21 ≈ 0,14	3/21 ≈ 0,14	3/21 ≈ 0,14	2/21 ≈ 0,10	2/21 ≈ 0,10	1/21 ≈ 0,05	1/21 ≈ 0,05

On observe une "convergence" vers la distribution du modèle 1 (on est dans un cas où le choix du modèle n'est pas évident dans l'écriture du programme).

Le modèle n° 2 suppose une équiprobabilité qui n'est pas conforme aux observations.

Les fluctuations des distributions des fréquences des sommes se réduisent lorsque l'on passe de simulations de taille 100 à celles de taille 1000.

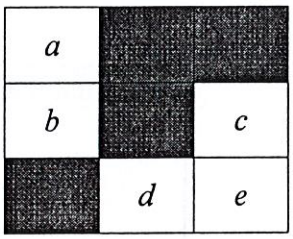
Les écarts de fréquence des sommes 4 et 5 (par exemple) observées sur 100 ou 1000 simulations correspondent aux fluctuations d'échantillonnage.

Pour la somme 4, on $p = \frac{3}{36}$ et les fréquences observées sur des échantillons de taille n se répartissent avec un écart type $\sigma \approx \sqrt{\frac{\frac{3}{36} \times \frac{33}{36}}{n}}$, soit 0,028 pour n = 100 et 0,009 pour n = 1000.

Exercice

TROIS MODELES

Dans la grille suivante, on doit choisir une case blanche.



Trois procédures, pour ce choix, sont proposées :

- (1) : On choisit une case blanche au hasard parmi les cinq cases blanches.
- (2) : On choisit au hasard une ligne, puis, dans la ligne, une case au hasard parmi les cases blanches.
- (3) : On choisit au hasard une colonne, puis, dans la colonne, une case au hasard parmi les cases blanches.

1) Trois modèles :
On note $E = \{a ; b ; c ; d ; e\}$. Déterminer les trois lois de probabilité, notées P_1 ; P_2 ; P_3 sur l'ensemble E modélisant chacune des trois procédures.

2) On a simulé n choix de cases blanches, selon l'une des trois procédures ci-dessus, pour $n = 100$, $n = 1000$ puis $n = 10000$.
On a obtenu les fréquences ci-dessous :

Case :	a	b	c	d	e
$n = 100$	0,19	0,16	0,22	0,31	0,12
$n = 1000$	0,15	0,159	0,174	0,368	0,149
$n = 10000$	0,167	0,168	0,164	0,336	0,165

Au vu de ce tableau, peut-on deviner le modèle utilisé ?

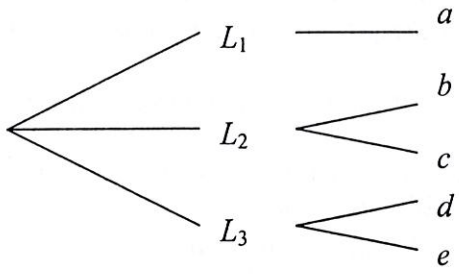
"On insistera sur le fait que l'observation des résultats simulés ne permet de remonter au modèle à coup sûr : une des fonctions de la statistique est de calculer la probabilité que l'on a de se tromper en "remontant d'une distribution de fréquences à une loi de probabilité."
Document d'accompagnement – 1ère S 2001.

Eléments de solution

1) Les lois de probabilités sont les suivantes (pour les modèles 2 et 3, on pourra utiliser un arbre) :

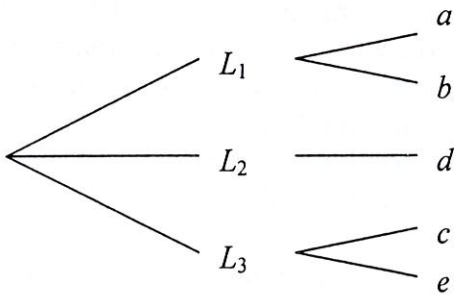
Loi P_1

x_i	a	b	c	d	e
p_i	1/5 = 0,2	0,2	0,2	0,2	0,2



Loi P_2

x_i	a	b	c	d	e
p_i	1/3	1/6	1/6	1/6	1/6



Loi P_3

x_i	a	b	c	d	e
p_i	1/6	1/6	1/6	1/3	1/6

2) D'après la loi des grands nombres, les fréquences observées se stabilisent, lorsque n augmente, vers les probabilités. Il semblerait donc que l'on ait simulé selon la loi P_3.

Remarque : un test statistique permettant de juger de l'adéquation des fréquences observées dans le cas n = 10000, à la loi P_3 est le test du khi-deux (voir séance 5).

III – NOTION DE VARIABLE ALEATOIRE (1ère S)

Cette notion n'est pas au programme de 1ère ES.

1 – LA PRESENTATION DES TEXTES OFFICIELS (2001)

*"Les **lois de probabilité non équiréparties** rencontrées en première et terminale sont le plus souvent l'image par une variable aléatoire d'une loi équirépartie, et on introduira donc **presque simultanément la notion de loi de probabilité et celle de variable aléatoire**.*

*"Une **variable aléatoire** T est une **application** définie sur un ensemble muni d'une loi de probabilité P ; son rôle est de **transporter** P sur un autre ensemble."*
Soit T une variable aléatoire de E (fini), muni d'une loi de probabilité P, à valeurs dans E' (fini).
*"La **loi de T** est une loi de probabilité définie sur E' par :*
$P'(x') = P(T = x')$ où $P(T = x')$ désigne la probabilité de l'ensemble des éléments de E dont l'image par T est x'."

Document d'accompagnement du programme de 1ère S.

La recherche de l'***équiprobabilité*** est fondamentale (par exemple en recherche physique). Quand on étudie une loi, le problème consiste à déterminer de quel espace avec équiprobabilité elle provient par une variable aléatoire. Dans toute épreuve d'univers fini, on peut se ramener à une situation d'équiprobabilité.

3 manuels, 3 définitions

• ***Bréal*** 1ère S p.232 :
*"Etant donnée une épreuve aléatoire, on appelle **variable aléatoire** associée à cette épreuve toute grandeur numérique dont la valeur dépend de l'issue de l'épreuve."*
Cette définition est conforme... aux programmes de terminale de 1992. Elle se justifie par le fait que les élèves de 1ère ne possèdent le concept de fonction que dans le cadre des fonctions d'une variable réelle et des transformations géométriques (et encore...).
Le problème est que, dans un souci de simplification, elle va entretenir la confusion entre l'opérateur variable aléatoire et ses résultats, ce qui peut poser de gros problèmes par la suite en statistique (pour l'estimation par exemple).

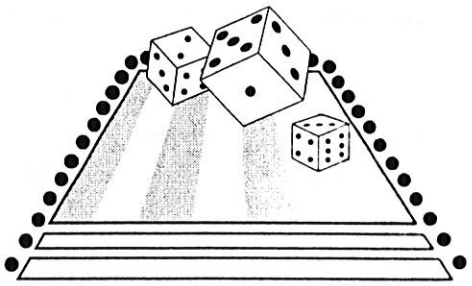
• ***Transmath*** 1ère S p.216 :
*"E est l'univers associé à une expérience aléatoire. Toute fonction définie sur E, à valeurs dans $\mathbb{R}$, est appelée une **variable aléatoire**."*
On se limite ici aux variables aléatoires réelles (elles seront ainsi la plupart du temps), mais cette limitation n'est pas dans le programme.
Par ailleurs le terme technique "d'univers" (ou la notation Ω) n'est pas non plus (volontairement) dans les textes officiels, il suffit de parler de l'ensemble des issues.

• ***Déclic*** 1ère S p.258 :
"Une variable aléatoire sur l'univers $E = \{x_1, ..., x_n\}$ est une fonction T définie sur E.
Si $\{t_1, ..., t_n\}$ est l'ensemble des images par T de toutes les éventualités de E (auquel cas, on a $k \leq n$), on note $(T = t_i)$ l'événement formé des éventualités qui ont pour image t_i par T, pour tout entier $1 \leq i \leq k$."
Cette définition, conforme au programme 2001, permet d'expliquer comment une variable aléatoire X "transporte" l'équiprobabilité de E = {(P, P) ; (P, F) ; (F, P) ; (F, F)} (lancers de deux pièces à pile ou face), en une loi de probabilité sur E' = {PP, PF, FF} en donnant à (P, F) et (F, P) la même image PF par X.

2 – ACTIVITE ELEVE EN 1ère S (après le cours) :

VARIABLES ALEATOIRES

1 ARNAQUE A LA FOIRE



Sur les foires anglo-saxonnes, on propose le jeu suivant :
Pour une mise de 1€, le joueur peut parier sur un nombre entier compris entre 1 et 6.
Il lance alors trois dés.
Si le nombre sur lequel il a parié sort 1 fois, 2 fois ou 3 fois, on lui rembourse sa mise plus 1€, 2€, ou 3€. Sinon, il perd sa mise.
Ce jeu peut sembler, à première vue, plutôt équitable. Nous allons examiner ce qu'il en est.

Simulation

Le programme suivant simule 100 parties, où, à chaque fois, un nombre N est choisi au hasard entre 1 et 6. Il fournit ensuite la répartition des gains ou pertes, puis la moyenne.

CASIO Graph 25 → 100	T.I. 80 - 82 - 83	T.I. 89 - 92
ClrList ↵ Seq(I,I,0,3,1) → List 1 ↵ -1 → List 1 [1] ↵ Seq(0,I,1,4,1) → List 2↵ For 1 → I To **100** ↵ 0 → S ↵ Int (1 + 6 Ran#) → N ↵ Int (1 + 6 Ran#) = N ⇒ S + 1 → S ↵ Int (1 + 6 Ran#) = N ⇒ S + 1 → S ↵ Int (1 + 6 Ran#) = N ⇒ S + 1 → S ↵ S + 1 → L ↵ List 2 [L] +1 → List 2 [L] ↵ Next ↵ List 2 ⫽ 1-Variable List 1 , List 2	:ClrList L_1 , L_2 :seq(I , I , 0 , 3 , 1) → L_1 : -1 → L_1(1) :seq(0 , I , 1 , 4 , 1) → L_2 :For(I, 1 , **100**) :0 → S :int(1 + 6rand) → N :If int(1 + 6rand) = N :S + 1 → N :If int(1 + 6rand) = N :S + 1 → S :If int(1 + 6rand) = N :S + 1 → S :S + 1 → L :L_2(L) + 1 → L_2(L) :End :Disp L_2 :Pause :1-Var Stats L_1 , L_2	:DelVar L1 , L2 :seq(i , i , 0 , 3 , 1) → L1 : -1 → L1[1] :seq(0 , i , 1 , 4 , 1) → L2 :For i, 1 , **100** :0 → s :int(1 + 6rand( )) → n :If int(1 + 6rand( )) = n :s + 1 → s :If int(1 + 6rand( )) = n :s + 1 → s :If int(1 + 6rand( )) = n :s + 1 → s :s + 1 → L :L2[L] + 1 → L2[L] :EndFor :Disp L2 :Pause :OneVar L1 , L2 :ShowStat

⇨ **Pour obtenir certaines instructions :**
• CASIO : **ClrList** par PRGM CLR List ; **Seq** par OPTN LIST ; **List** par OPTN LIST ; **For To Next** par PRGM COM ; **Int** par OPTN NUM ; **Ran#** par OPTN PROB ; **=** [au clavier ; **1-Variable** par MENU (F4) STAT CALC 1VAR
• TI 80 82 83 : **ClrList** par STAT EDIT ; **Seq** par 2nd LIST OPS ; L_1 au clavier par 2nd ; **For End** par PRGM CTL ; **int** par MATH NUM ; **rand** par MATH PRB ; **Disp** par PRGM I/O ; **1-Var Stats** par STAT CALC

On note X la variable aléatoire qui, à chaque jeu, associe le gain (ou la perte).

Compléter le tableau :

$X=$	-1	1	2	3	$\overline{x} \approx$
Nb d'observations					

Modifier le programme précédent pour faire une simulation de 1000 jeux.

$X=$	-1	1	2	3	$\overline{x} \approx$
Nb d'observations					

Recherche de l'espérance de gain

Pour confirmer vos observations, recherchons la ***loi de probabilité*** de X.
Soit $E = \{(n_1, n_2, n_3) / n_i = 1, 2, \ldots, 6\}$ l'ensemble de tous les triplets que l'on peut obtenir avec trois dés distincts.
Justifier que E possède 216 éléments.

.....
.....

On muni E de la loi de probabilité P correspondant à l'équiprobabilité.
La variable aléatoire X est l'application définie de E dans $E' = \{-1, 1, 2, 3\}$ en associant à chaque élément de E le gain qui lui correspond.
En considérant le rapport des cas favorables aux cas possibles, calculer la loi de probabilité de X.

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

x_i	-1	1	2	3
$p_i = P(X = x_i)$				

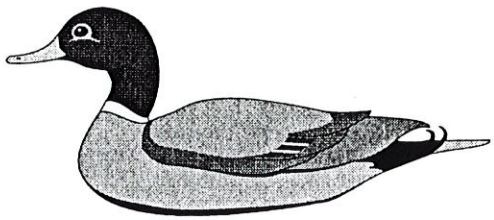
Le rôle de la variable aléatoire X a été de transporter la loi de probabilité P sur E'.

Calculer l'***espérance*** de X.

.....
.....
.....
.....

2 LA CHASSE AUX CANARDS

Trois chasseurs A, B, C, tirent simultanément sur trois canards numérotés 1, 2, 3.
Chaque chasseur choisit au hasard, et indépendamment des autres, de viser un canard et ne rate jamais son coup.
On note X la variable aléatoire qui, à toute chasse, associe le nombre de canards survivants.
On désire déterminer la loi de X.



Simulation sans programmer

La simulation va permettre de se faire une première idée de cette loi, sans pour autant commettre une hécatombe.
La simulation d'une chasse consiste à répéter, **sans programme**, 3 fois l'instruction :

1 + Int(3Ran#) ou 1 + int(3rand)

en appuyant sur EXE ou ENTER.

Effectuer ainsi 10 chasses et consigner ci-dessous vos résultats.

Simulation n°	1	2	3	4	5	6	7	8	9	10
Canards abattus										
Valeur de X										

Statistiques sur 10 chasses :

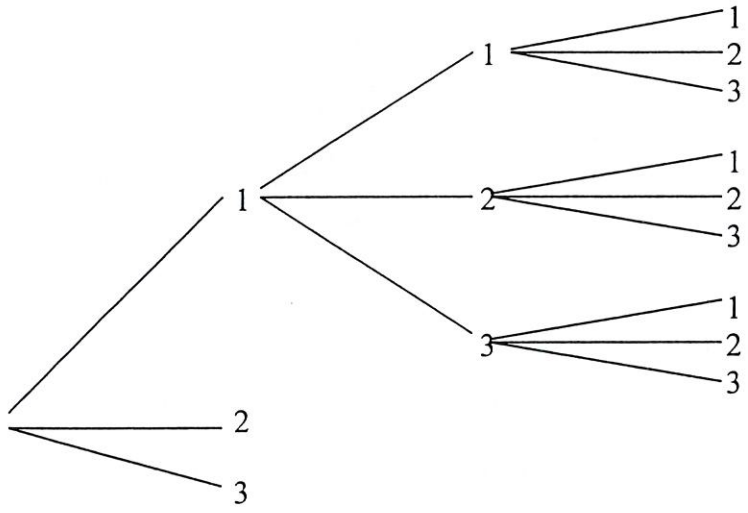
Valeur de X	0	1	2
Effectifs			

Faire des statistiques sur 10 chasses, ce n'est pas très significatif. Avant de concevoir un programme permettant d'aller beaucoup plus loin, réfléchissons à une solution théorique.

Résolution théorique

A l'aide de l'arbre (incomplet) suivant, déterminer la loi de X.

Valeur de X :



Choix du chasseur A | Choix du chasseur B | Choix du Chasseur C

Loi de X :

Valeur de X	0	1	2
Probabilité			

Calculer $E(X)$ et $\sigma(X)$:

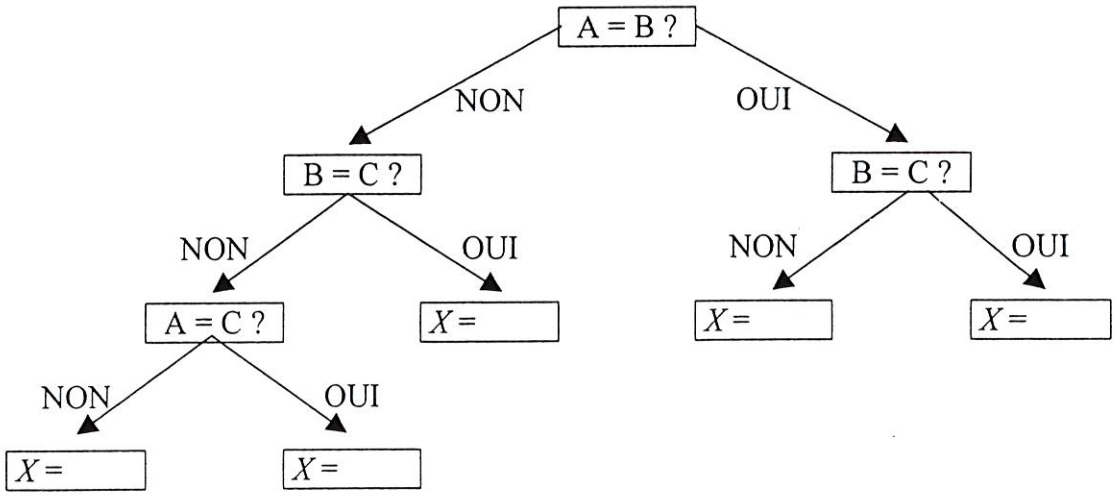
.....

.....

.....

Programmation

On désigne par A, B, C les variables où seront stockés les choix des chasseurs A, B, C :
1 + Int(3Ran#) → A puis 1 + Int(3Ran#) → B et 1 + Int(3Ran#) → C.
Il s'agit, selon les valeurs de A, B et C, de faire le tri et d'affecter la valeur correcte à X.
Pour ce faire, compléter l'organigramme suivant :



Effectuer le programme suivant pour 100, puis 1000 chasses.

CASIO Graph 25 → 100	T.I. 80 - 82 - 83	T.I. 89 - 92
ClrList ↵ Seq(I,I,0,2,1) → List 1 ↵ Seq(0,I,1,3,1) → List 2 ↵ For 1 → I To **100** ↵ 1 + Int(3Ran#) → A ↵ 1 + Int(3Ran#) → B ↵ 1 + Int(3Ran#) → C ↵ If A = B ↵ Then If B = C ↵ Then 2 → X ↵ Else 1 → X ↵ IfEnd ↵ Else If B = C ↵ Then 1 → X ↵ Else If A = C ↵ Then 1 → X ↵ Else 0 → X ↵ IfEnd ↵ IfEnd ↵ IfEnd ↵ X + 1 → X ↵ List 2 [X] + 1 → List 2 [X] ↵ Next ↵ List 2 ⫽ 1-Variable List 1 , List 2	:ClrList L_1 , L_2 :seq(I , I , 0 , 2 , 1) → L_1 :seq(0 , I , 1 , 3 , 1) → L_2 :For(I , 1 , **100**) :1 + int(3 rand) → A :1 + int(3 rand) → B :1 + int(3 rand) → C :If A = B :Then :If B = C :Then :2 → X :Else :1 → X :End :Else :If B = C :Then :1 → X :Else :If A = C :Then :1 → X :Else :0 → X :End :End :End :X + 1 → X :L_2(X) + 1 → L_2(X) :End :Disp L_2 :Pause :1-Var Stats L_1 , L_2	:DelVar L1 , L2 :seq(i , i , 0 , 2 , 1) → L1 :seq(0 , i , 1 , 3 , 1) → L2 :For i , 1 , **100** :1 + int(3 rand( )) → a :1 + int(3 rand( )) → b :1 + int(3 rand( )) → c :If a = b Then :If b = c Then :2 → x :Else :1 → x :EndIf :Else :If b = c Then :1 → x :Else :If a = c Then :1 → x :Else :0 → x :EndIf :EndIf :EndIf :x + 1 → x :L2[x] + 1 → L2[x] :EndFor :Disp L2 :Pause :OneVar L1 , L2 :ShowStat

Consigner vos résultats dans le tableau suivant :

	$X = 0$	$X = 1$	$X = 2$	$\bar{x}$	σ
100 chasses					
1000 chasses					

Comparer à vos résultats théoriques. .....

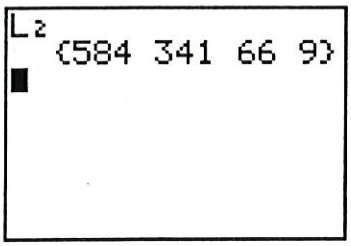
.....

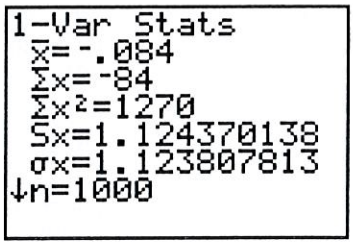
.....

Eléments de solution de l'activité "Variables aléatoires"

1 Arnaque à la foire :

1) On souhaite se faire d'abord une idée de l'espérance de gain par simulations.
Simulation de 1000 parties :





2) Loi de X et espérance :

$X=$	-1	1	2	3
Probabilités	125/216≈0,579	75/216≈0,347	15/216≈0,069	1/216≈0,005

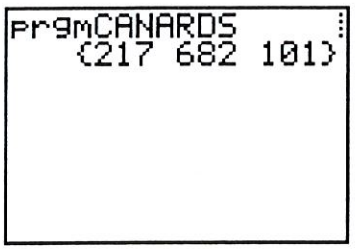
$E(X) = -17/216 \approx -0,079$. Ces résultats correspondent aux simulations et confirment que le jeu n'est pas équitable.

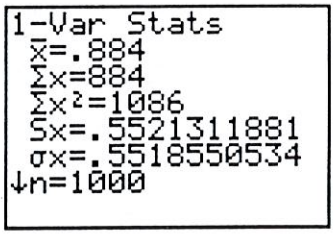
2 La chasse aux canards :

1) Il s'agit dans cette question de comprendre le principe de la simulation et de "vivre" les aléas du hasard.
2) Résolution théorique :

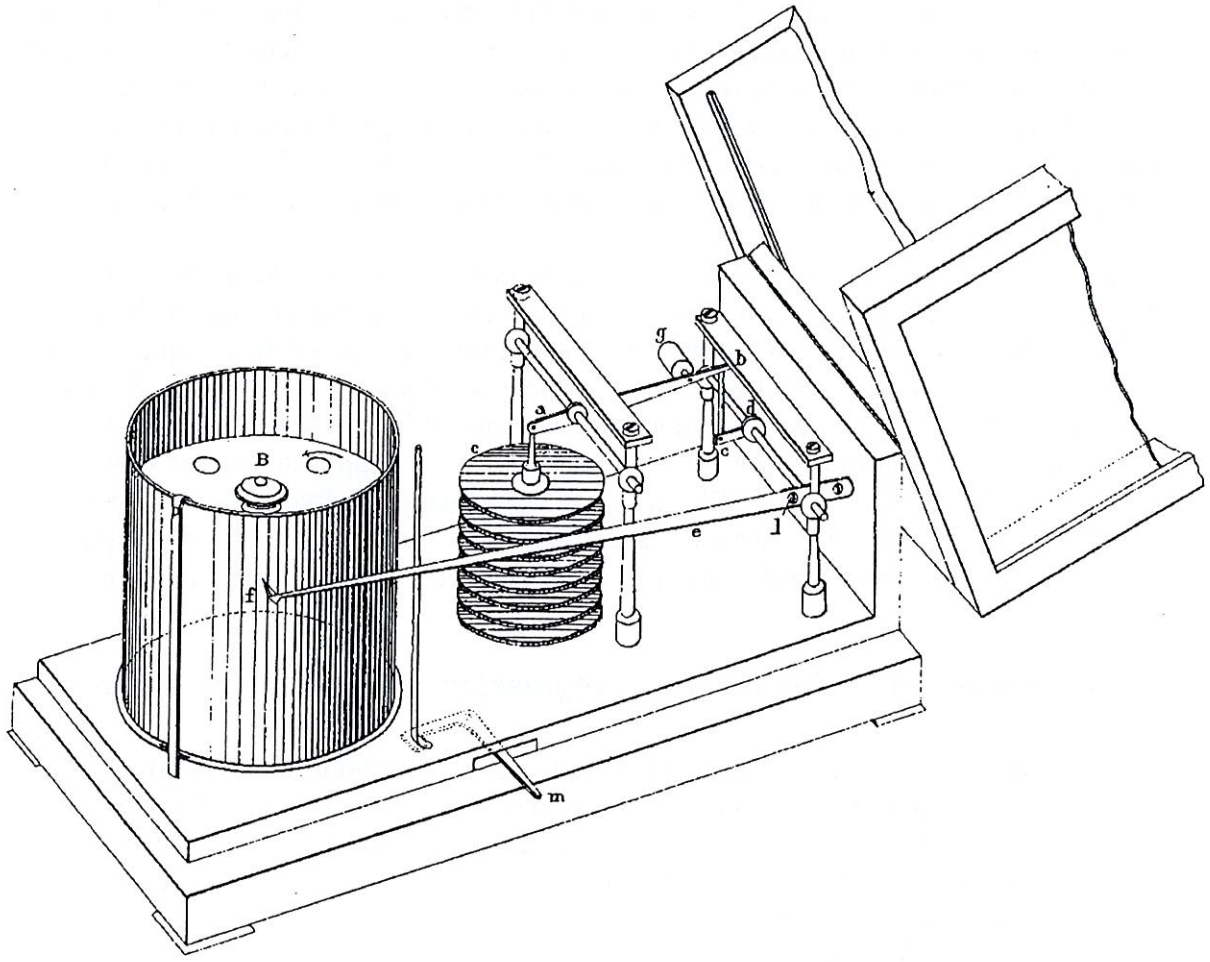
Valeur de X	0	1	2
Probabilité	$6/27 \approx 0,22$	$18/27 \approx 0,67$	$3/27 \approx 0,11$

$E(X) \approx 0,89$ et $\sigma(X) \approx 0,57$.
3) Simulation de 1000 chasses :





Séance 4 :
SERIES A DEUX VARIABLES - CONDITIONNEMENT ET INDEPENDANCE



I – REGRESSION LINEAIRE (Terminale ES)

1 – LE MODELE LINEAIRE

Nous avons décrit l'émergence de la technique d'ajustement linéaire selon les moindres carrés, au début du XIX^{ème} siècle, chez ***Legendre*** et ***Gauss***, dans le cadre de la théorie des erreurs de mesure.

De nombreux outils de la statistique mathématique apparaissent à la fin du XIX^{ème} siècle 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 khi-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 (eugénisme). 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 petite taille, 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.

Grand admirateur de *Galton* (il lui consacrera 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 σ.

a) Une vision géométrique de la régression linéaire selon les moindres carrés

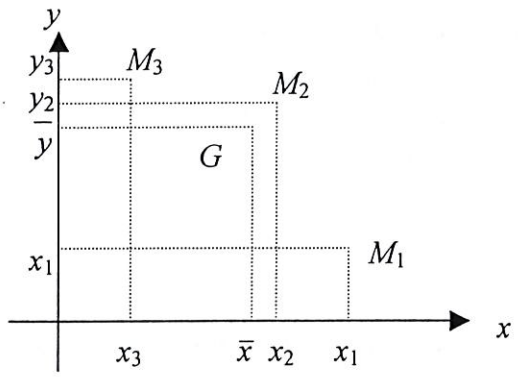
La géométrie euclidienne est d'une grande importance, tant pour la "fabrication" que pour l'interprétation d'indicateurs statistiques classiques (voir l'anecdote précédemment évoquée à propos de l'intuition géométrique sur-développée de *R. Fisher*).

Considérons ici deux variables x et y dont on étudie les valeurs sur n individus.

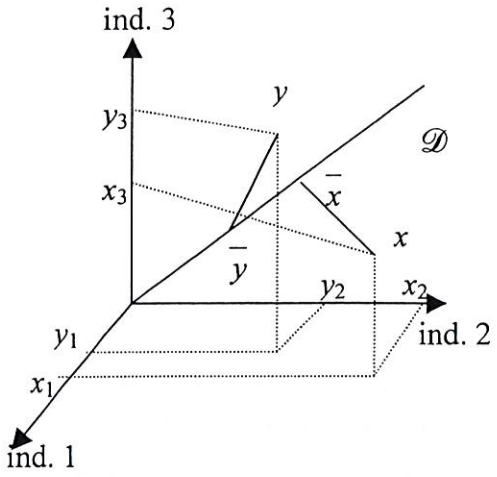
Deux représentations euclidiennes sont possibles :

- dans "l'espace des individus",
- ou dans "l'espace des variables".

Pour faire les figures suivantes nous avons supposé que $n = 3$ (c'est plus facile pour visualiser l'espace des variables !).



Espace des individus $\mathbb{R}^2$
Axes : les 2 variables x et y
Points : les $n = 3$ individus

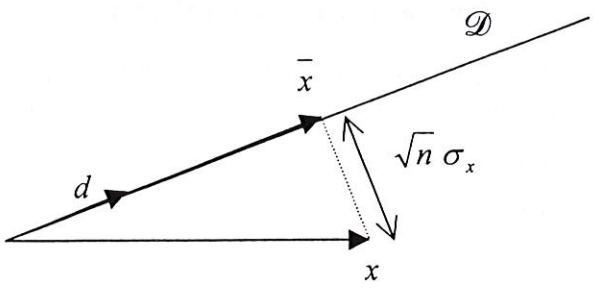


Espace des variables $\mathbb{R}^n$
Axes : les $n = 3$ individus
Points : les 2 variables x et y

(L'intérêt de ces représentations, outre l'interprétation géométrique des résumés classiques, est d'être généralisables au cas multivarié.)
On muni les espaces $\mathbb{R}^2$ et $\mathbb{R}^n$ de la structure euclidienne usuelle.
Soit, dans l'espace des variables, la droite $\mathcal{D}$ de vecteur directeur d de coordonnées (1,..., 1). La droite $\mathcal{D}$ représente les variables qui prennent la même valeur sur tous les individus.

La moyenne est la projection orthogonale de x sur $\mathcal{D}$

La droite $\mathcal{D}$ menée par l'origine et de vecteur directeur $d(1,\ldots, 1)$, 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 $\mathcal{D}$. Si la structure euclidienne usuelle est adéquate, le meilleur indicateur de centralité est le point de la droite $\mathcal{D}$ le plus proche du point x, c'est-à-dire la projection orthogonale de x sur la droite $\mathcal{D}$. Vérifions que l'on retrouve ainsi l'expression de la moyenne.



On a :

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

ou encore :

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

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

Ainsi, le meilleur indicateur de centralité est le point dont toutes les coordonnées sont égales à $\bar{x} = \dfrac{1}{n}\sum x_i$, la moyenne. On notera encore $\bar{x}$ le point dont toutes les coordonnées sont égales à $\bar{x}$.
De même y est projeté en $\bar{y}$ sur la droite $\mathcal{D}$.

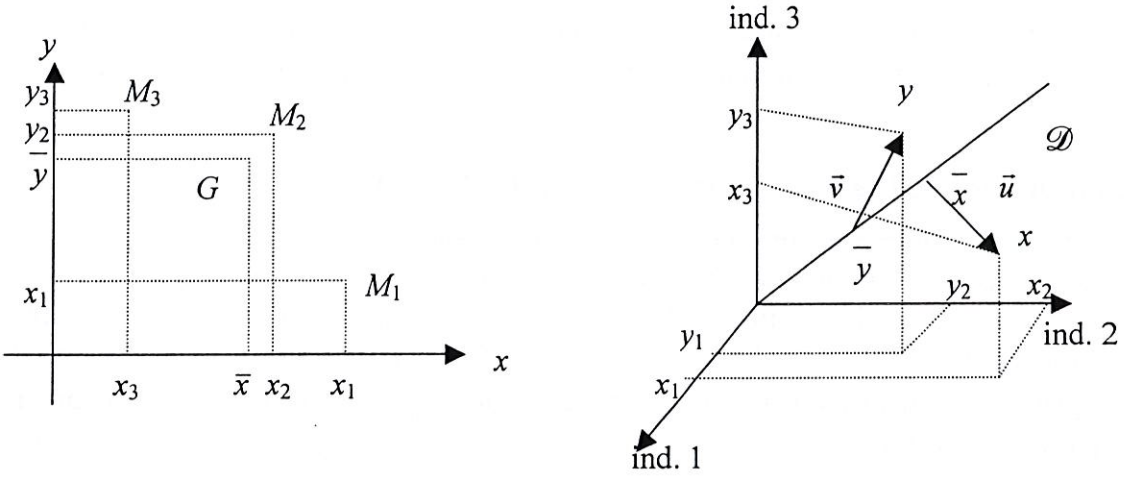
L'écart type est proportionnel à la distance entre x et $\bar{x}$

Le caractère x est d'autant plus dispersé que le point x, le représentant dans l'espace des variables, est éloigné de la droite $\mathcal{D}$. 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.
En notant $\vec{u} = x - \bar{x}$, l'écart type σ_x vérifie : $\sqrt{n}\,\sigma_x = \|\vec{u}\|$.

le coefficient de corrélation r est le cosinus de l'angle des deux vecteurs $\vec{u} = (x - \bar{x})$ et $\vec{v} = (y - \bar{y})$

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 "centrés" $\vec{u} = (x - \bar{x})$ et $\vec{v} = (y - \bar{y})$ 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.

Alignement des M_i ($y = k\,x$) ⟺ Colinéarité de $\vec{u}$ et $\vec{v}$ ($y - \bar{y} = k\,(x - \bar{x})$)



Nuage des individus — **Espace des variables**

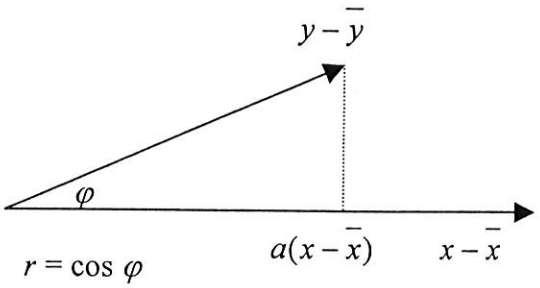
Une application du produit scalaire permet de retrouver l'expression du coefficient de corrélation linéaire r entre les variables x et y. On considère l'angle φ des deux vecteurs $\vec{u} = (x - \bar{x})$ et $\vec{v} = (y - \bar{y})$. Déterminons le cosinus de cet angle.

On a $\cos\varphi = \cos\left(x - \bar{x}, y - \bar{y}\right) = \dfrac{(x - \bar{x}).(y - \bar{y})}{\|x - \bar{x}\| \times \|y - \bar{y}\|}$,

c'est à dire $\cos\varphi = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{n}\sigma_x \times \sqrt{n}\sigma_y} = \dfrac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \times \sigma_y} = \dfrac{\sigma_{xy}}{\sigma_x \times \sigma_y}$, où σ_{xy} est la covariance.
donc $\cos\varphi = r$.

Le coefficient de régression linéaire a, de y en x, apparaît par projection de $\vec{v} = y - \bar{y}$ *sur* $\vec{u} = x - \bar{x}$

Pour des données bivariées, lorsque le nuage de points dans l'espace des individus est proche d'une droite, $|r|$ est voisin de 1 et l'angle entre $\vec{u} = x - \bar{x}$ et $\vec{v} = y - \bar{y}$ 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 $\vec{v} = y - \bar{y})$ sur $\vec{u} = x - \bar{x}$.
En effet :

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

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 $a = \frac{\sigma_{xy}}{\sigma_x^2}$.

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, 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 $(y - \bar{y})$.

b) Un T.P. sur Excel

Ce T.P. a comme objectifs :

- Analyser la signification du coefficient de corrélation.
- Rechercher, par tâtonnement numérique et graphique, un ajustement linéaire minimisant la somme des carrés des écarts verticaux.
- Apprendre à utiliser certaines fonctions du tableur Excel.
- Travailler sur le logarithme et l'exponentielle.

T.P. SUR EXCEL EN 1ère ES

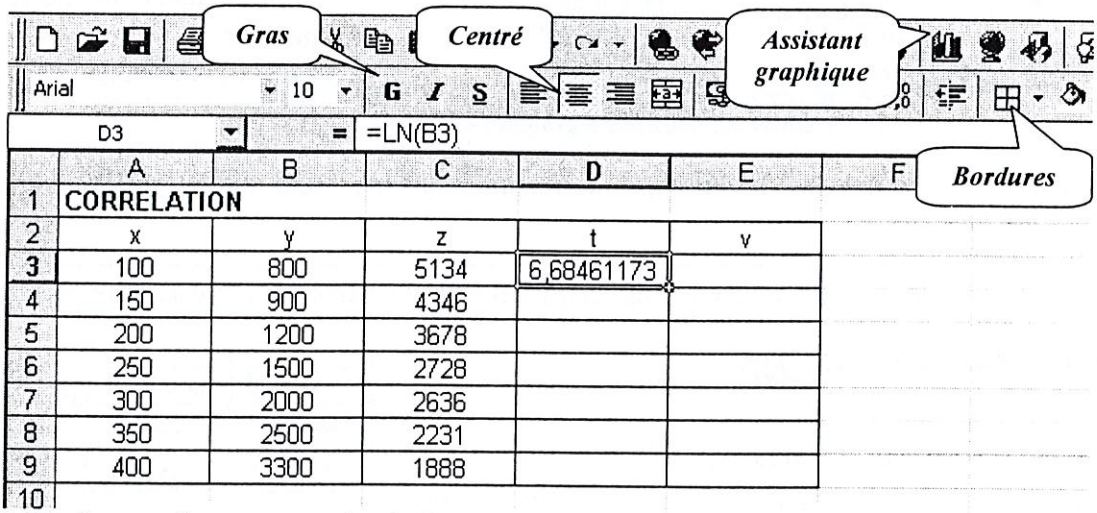
REGRESSION LINEAIRE SELON LES MOINDRES CARRES

Une société de voyages organisés dispose, chaque semaine, de 12 200 voyages de 7 jours à la vente, dont les prix varient entre 100 € et 400 €.
On désigne par x le prix d'un voyage et par y le nombre de voyages disponibles, en fonction du prix.
Pour la première semaine de juillet, cette société enregistre le nombre, noté z, de demandes qui lui sont parvenues. Le résultat figure dans le tableau ci-dessous.

prix d'un voyage : x_i	nombre de voyages disponibles : y_i	nombre de voyages demandées : z_i
100	800	5134
150	900	4346
200	1200	3678
250	1500	2728
300	2000	2636
350	2500	2231
400	3300	1888

CORRELATION

On considère les quatre séries : (x_i, y_i) , (x_i, t_i) où $t_i = \ln(y_i)$, (x_i, z_i) et (x_i, v_i) où $v_i = \ln(z_i)$.
Lancer Excel.
Cliquer dans la cellule A1, choisir ***Gras*** et taper le titre "CORRELATION".



	A	B	C	D	E	F
1	CORRELATION					
2	x	y	z	t	v	
3	100	800	5134	6,68461173		
4	150	900	4346			
5	200	1200	3678			
6	250	1500	2728			
7	300	2000	2636			
8	350	2500	2231			
9	400	3300	1888			
10						

Sélectionner (bouton gauche de la souris enfoncé) la plage rectangulaire des cellules A2 à E9. Cliquer sur les icônes ***Centré*** et ***Bordures*** (choisir le quadrillage en cliquant sur la petite flèche), puis entrer les données x_i , y_i , z_i comme ci-dessus.
En D3, entrer la ***formule*** : = LN(B3)
puis ***Recopier*** vers le bas jusqu'en D9 (en approchant le pointeur de la souris du carré noir situé dans le coin inférieur droit de la cellule, celui-ci se transforme en une croix noire, faire alors glisser en maintenant le bouton gauche enfoncé).

En E3, entrer la ***formule*** : = LN(C3) puis ***Recopier*** vers le bas jusqu'en E9.

• Création d'un graphique :

Sélectionner les cellules de A3 à B9, contenant les valeurs (x_i, y_i).

Cliquer sur l'icône ***Assistant graphique***.

Etape 1/4 : choisir ***Nuages de points*** (et le ***sous-type*** sans courbe) puis cliquer sur ***Suivant***.

Etape 2/4 : Cliquer sur ***Suivant***.

Etape 3/4 : Dans l'onglet ***Légende***, désactiver ***Afficher la légende*** (non coché). Dans l'onglet ***Quadrillage***, désactiver tout quadrillage. Dans l'onglet ***Titre***, à la rubrique ***Titre du graphique***, taper "Série (x,y)". Cliquer sur ***Suivant***.

Etape 4/4 : Cocher ***en tant qu'objet dans Feuil1*** puis cliquer sur ***Terminer***.

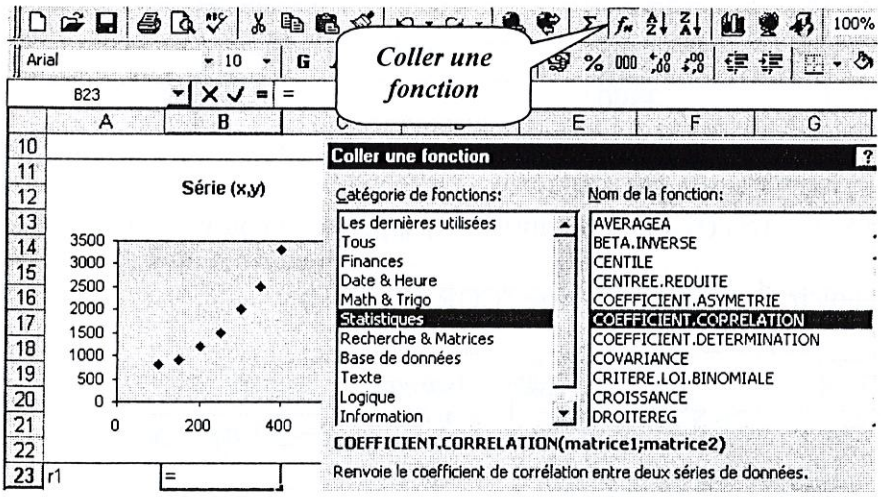
Déplacer le graphique obtenu et ajuster sa taille (poignées sur les côtés) de façon qu'il occupe trois colonnes de A11 à C22. Cliquer en dehors de la zone de graphique.

• Représenter de façon analogue la série (x_i, t_i) : pour sélectionner dans la feuille de calcul des colonnes non contiguës, maintenir enfoncée la touche ***CTRL*** et lâcher le bouton gauche de la souris.

Déplacer et ajuster le graphique de façon à ce qu'il recouvre les cellules D11 à F22.

• Représenter, sous les graphiques précédents mais en sautant deux lignes, les nuages de points (x_i, z_i) puis (x_i, v_i).

• On désigne par r_1 , r_2 , r_3 , r_4 les coefficients de corrélation linéaire respectifs des quatre séries représentées.



Sous le nuage de la série (x_i, y_i), taper en cellule A23 "r1", puis se placer en B23 et cliquer sur l'icône ***Coller une fonction***.

Dans la boîte de dialogue, choisir ***Statistiques*** puis ***COEFFICIENT.CORRELATION*** puis ***OK***.

Indiquer dans la boîte de dialogue :

Matrice 1 : A3:A9

Matrice 2 : B3:B9

Puis cliquer sur ***OK***.

Calculer de même, sous les nuages correspondant, les trois autres coefficients de corrélation linéaire.

Indiquer sur la feuille réponse les séries qui relèvent le mieux d'un ajustement affine.

AJUSTEMENT AFFINE

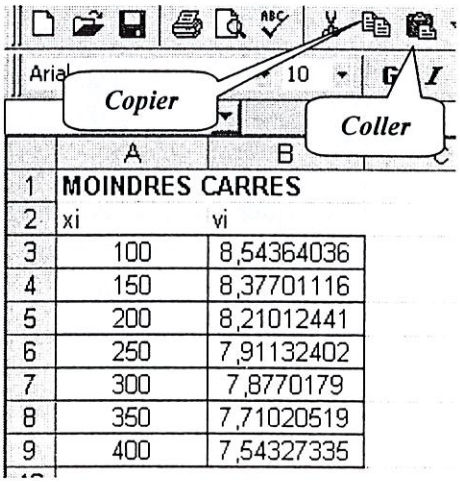
a) Ajustement de la Série (x_i , v_i)

Cliquer sur l'onglet ***Feuil2*** pour commencer une nouvelle feuille de calcul.

En A1 taper en ***Gras*** le titre : "MOINDRES CARRES".

En A2, inscrire xi et en B2, inscrire vi.

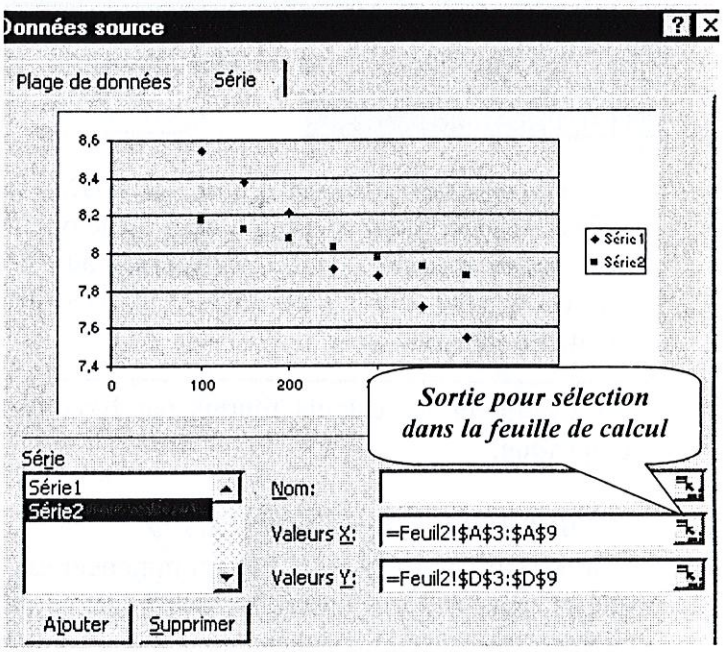
Revenir sur la feuille 1, sélectionner les données (x_i, v_i) (appuyer sur ***CTRL*** pour sélectionner des colonnes séparées) puis cliquer sur l'icône ***Copier***.



	A	B
1	MOINDRES CARRES	
2	xi	vi
3	100	8,54364036
4	150	8,37701116
5	200	8,21012441
6	250	7,91132402
7	300	7,8770179
8	350	7,71020519
9	400	7,54327335

Sur la feuille 2, cliquer en A3 puis sur l'icône ***Coller*** (éventuellement, faire ***Edition/Collage spécial/Valeurs***).
En A10 taper "Point moyen".
En A11, entrer la ***formule*** =MOYENNE(A3:A9)
En B11, entrer la ***formule*** =MOYENNE(B3:B9)
On va rechercher une droite de pente a, passant par le point moyen de coordonnées $(\bar{x}, \bar{v})$.
Cette droite aura pour équation $v = ax + b$ avec $\bar{v} = a\bar{x} + b$, d'où $b = \bar{v} - a\bar{x}$.
En C2 taper "pente a" et entrer en C3 la valeur $-0{,}001$ (pour un premier essai).
En C4 taper "ordonnée b" et entrer en C5 la ***formule*** =B11 – C3*A11
Vous allez ajouter une colonne contenant les valeurs v calculées par $ax + b$:
En D2 taper "axi+b" et entrer en D3 la ***formule*** =$C$3*A3+$C$5 puis ***recopier vers le bas*** jusqu'à la cellule D9 (le symbole $ empêche la modification des références de la cellule lors de la recopie).
En E2 taper "carrés écarts" et entrer en E3 la ***formule*** =(B3-D3)^2 (on obtient ^ par la touche ***ALT GR 9***) puis ***recopier vers le bas*** jusqu'à la cellule E9.

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 les données (x_i, v_i) des cellules A3 à B9. Revenir dans la boîte de dialogue de l'assistant graphique.
Onglet ***Série***, cliquer sur ***Ajouter*** puis, pour la Série 2 :
Valeurs X : sortir sélectionner les cellules de A3 à A9.
Valeurs Y : sortir sélectionner les cellules de D3 à D9.
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 • ***en tant qu'objet dans la Feuil2*** puis cliquer sur ***Terminer***.
Pour que la droite soit tracée proprement, cliquer sur le graphique *avec le bouton droit de la souris* sur un point de la ***Série 2*** puis cliquer (gauche) sur ***Format de la série de données...*** Dans l'onglet ***Motifs***, cocher ***Trait • automatique*** et ***Marque • automatique*** puis cliquer sur ***OK***.
Déplacer le graphique pour l'amener à côté des valeurs a de pente et b d'ordonnée à l'origine (cliquer sur le graphique et glisser en maintenant le bouton gauche de la souris enfoncé).

Vous allez déterminer la position de la droite qui minimise la somme des carrés des écarts verticaux.
Pour cela, taper en C6 "somme e^2" puis entrer en C7 la ***formule*** =SOMME(E3:E9)

Optimisation empirique de a

Modifier en C3 la 3ème décimale de a de façon à minimiser la somme des carrés des écarts verticaux (observer en même temps les positions sur le graphique).

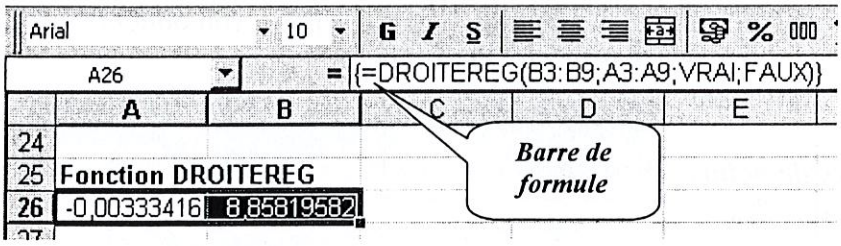
✍ Consigner les observations sur la feuille réponse.

Fixer de façon analogue, par essais successifs, la 4ème décimale de a.

✍ Consigner les observations sur la feuille réponse.

Déplacer le graphique en A12.

Utilisation de la fonction DROITEREG d'Excel



Sous le graphique, en A25 par exemple, taper en ***Gras*** "Fonction DROITEREG". ***Sélectionner*** les cellules A26 et B26 puis inscrire dans la ***Barre de formule*** : =DROITEREG(B3:B9;A3:A9;VRAI;FAUX) puis valider en appuyant simultanément sur ***CTRL MAJ ENTREE*** (car il s'agit d'une opération sur un tableau matriciel).
Dans cette formule, B3:B9 correspond aux ordonnées observées, A3:A9 aux abscisses. VRAI demande le calcul de b et FAUX demande de n'avoir que les coefficients de l'équation de droite.

✍ Comparer les calculs fournis par Excel avec ceux obtenus précédemment de façon empirique.

b) Ajustement de la Série (x_i , t_i)

Cliquer sur l'onglet ***Feuil3*** pour commencer une nouvelle feuille de calcul.
En A1 taper en ***Gras*** le titre "AJUSTEMENT (x,t)".
Revenir sur la feuille 1, sélectionner les colonnes des données de la série (x_i, t_i) (appuyer sur ***CTRL*** pendant la sélection) puis cliquer sur l'icône ***Copier***. Sur la feuille 3, se placer en A2 et cliquer sur l'icône ***Coller***.

Utilisation de la fonction DROITEREG d'Excel

En A10, taper en ***Gras***, "Fonction DROITEREG".
Sélectionner les cellules A11 et A12 puis mettre en œuvre, comme précédemment, dans la ***Barre de formule*** la fonction DROITEREG d'Excel pour obtenir les coefficients de la droite de régression de t en x selon la méthode des moindres carrés (valider en appuyant simultanément sur ***CTRL MAJ ENTREE***).

✍ Recopier les résultats sur la feuille réponse.

Utilisation de l'option graphique "Courbe de tendance"

On peut également demander à Excel d'ajouter sur le nuage de points (x_i , t_i) une courbe de tendance (ce sera une vérification supplémentaire).
Revenir sur la feuille 1. Cliquer sur le graphique de la série (x_i , t_i).

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*** puis cliquer sur ***OK***.

3- EXPLOITATION DU MODELE OBTENU

✍ Utiliser les résultats précédents pour donner, sur la feuille réponse, un modèle d'ajustement pour l'offre et la demande.

Vous pouvez contrôler vos calculs en demandant à Excel d'ajouter, sur les graphiques (x_i, y_i) et (x_i, z_i), une courbe de tendance.
Sur la feuille 1, cliquer sur le graphique de la série (x_i, y_i).
Aller dans le menu ***Graphique*** et cliquer sur ***Ajouter une courbe de tendance...***
Dans la boîte de dialogue, dans l'onglet ***Type*** choisir ***Exponentielle*** puis dans l'onglet ***Option*** cocher ***Afficher l'équation sur le graphique*** puis cliquer sur ***OK***.
Procéder de même avec le graphique de la série (x_i, z_i).

FEUILLE REPONSE

NOMS : .....

CORRELATION

Calcul des coefficients de corrélation linéaire (arrondis à 10^{-4} près) :

Série	(x_i, y_i)	(x_i, t_i)	(x_i, z_i)	(x_i, v_i)
Coefficient de corrélation linéaire				

A quoi correspond le signe du coefficient de corrélation ? .....

.....

Des deux séries (x_i, y_i) et (x_i, t_i), laquelle relève le mieux d'un ajustement affine ?
Pourquoi ?

.....

.....

.....

Même question pour les deux séries (x_i, z_i) et (x_i, v_i).

.....

.....

.....

AJUSTEMENT AFFINE

a) Ajustement de la Série (x_i, v_i)

Optimisation empirique de a

Recherche de la 3ème décimale :

valeur de la pente a	– 0,001	– 0,002	– 0,003	– 0,004
somme des carrés des écarts verticaux				

Conclusion : .....

Recherche de la 4ème décimale :

valeur de la pente a	valeur précédente	valeur optimale à 10^{-4} près	valeur suivante .....
somme des carrés des écarts verticaux			

Conclusion : indiquer, à 10^{-4} près, la pente a et l'ordonnée à l'origine b de la droite de régression de v en x, selon la méthode des moindres carrés.

.....

.....

Utilisation de la fonction DROITEREG d'Excel

Quels sont les résultats affichés ? .....
Donner, en arrondissant les coefficients à 10^{-4} près, une équation de la droite d'ajustement de v en x obtenue selon la fonction d'Excel :

.....

b) Ajustement de la Série (x_i, t_i)

Donner une équation de la droite d'ajustement de t en x, selon la méthode des moindres carrés (coefficients arrondis à 10^{-4} près).

.....
.....
.....

EXPLOITATION DU MODELE OBTENU

Déduire de ce qui précède une expression de y et z sous la forme :

$$y = a\,e^{bx} \text{ et } z = c\,e^{dx}$$

(dans les réponses, a et c seront remplacés par leurs arrondis à 1 près).

On a obtenu $t = \ln y =$ ..... $x +$ ..... d'où $y =$ .....
.....
.....
.....

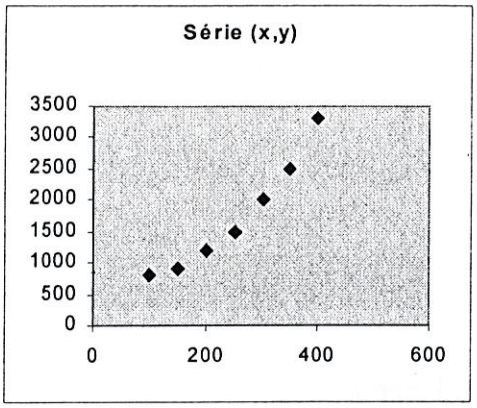
On a obtenu $v = \ln z =$ ..... $x +$ ..... d'où $z =$ .....
.....
.....
.....

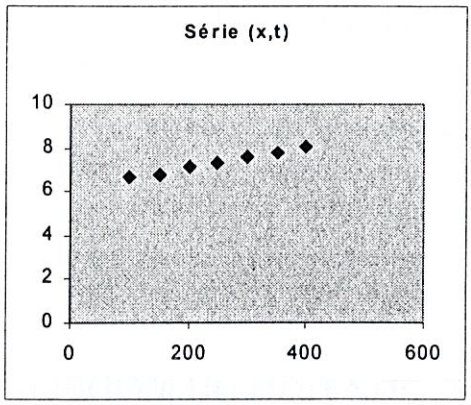
Le prix pour lequel l'offre est égal à la demande s'appelle le prix d'équilibre, noté x_0.
Calculer x_0 en utilisant les résultats de la question précédente.

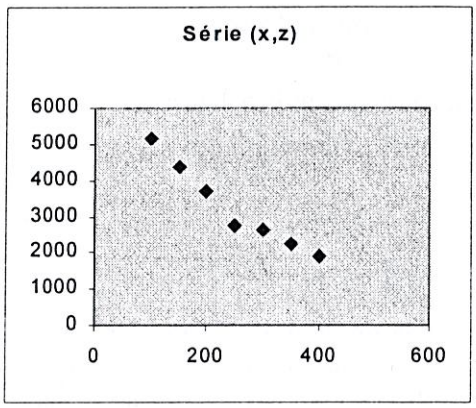
.....
.....
.....
.....
.....
.....

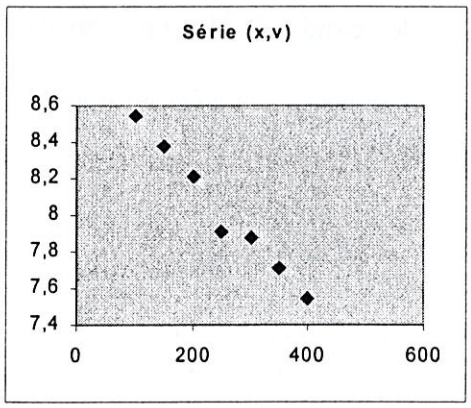
Eléments de réponse

CORRELATION









Calcul des coefficients de corrélation linéaire (arrondis à 10^{-4} près) :

Série	(x_i, y_i)	(x_i, t_i)	(x_i, z_i)	(x_i, v_i)
Coefficient de corrélation linéaire	**0,9703**	**0,9970**	**– 0,9744**	**– 0,9905**

Le signe du coefficient de corrélation correspond à la croissance (positif) ou à la décroissance (négatif) de la variable en ordonnée, en fonction de la variable en abscisse.
Des deux séries (x_i, y_i) et (x_i, t_i), celle qui relève le mieux d'un ajustement affine est la série (x_i, t_i) pour laquelle le coefficient de corrélation linéaire est, en valeur absolue, plus proche de 1.
Pour les deux séries (x_i, z_i) et (x_i, v_i), celle qui relève le mieux d'un ajustement affine est la série (x_i, v_i).

AJUSTEMENT AFFINE

a- Ajustement de la Série (x_i, v_i)

Optimisation empirique de a

Recherche de la 3ème décimale :

valeur de la pente a	-0,001	-0,002	-0,003	-0,004
somme des carrés des écarts verticaux	**0.3964**	**0.1396**	**0.0228**	**0.0460**

Conclusion : à 10^{-3} près, on a $a \approx -0,003$.

Recherche de la 4ème décimale :

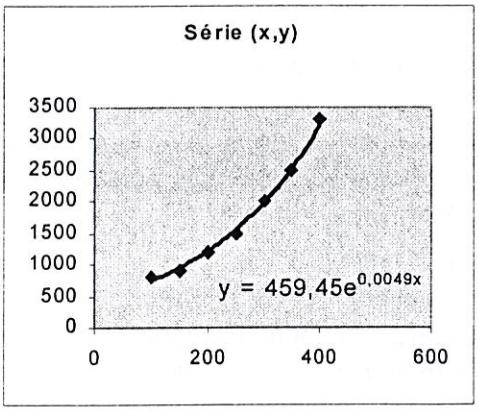
valeur de la pente a	Valeur précédente **– 0,0034**	Valeur optimale à 10^{-4} près **– 0,0033**	Valeur suivante **– 0,0032**
somme des carrés des écarts verticaux	**0.01529**	**0.01507**	**0.01624**

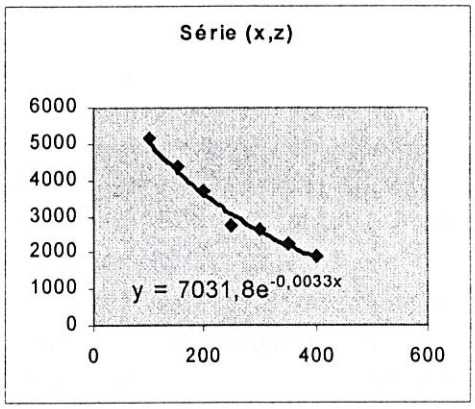
Conclusion :
A 10^{-4} près, selon la méthode des moindres carrés, $a \approx -0{,}0033$ et $b \approx 8{,}8497$.

Utilisation de la fonction DROITEREG d'Excel
Résultats affichés : a = – 0,00333416 et b = 8,85819582 .
En arrondissant les coefficients à 10^{-4} près, une équation de la droite d'ajustement de v en x est $v = -0{,}0033\,x + 8{,}8582$.
(Pour b, le résultat donné par la fonction d'Excel est préférable car celui obtenu par tâtonnement est calculé à partir de l'arrondi de a.)

b- Ajustement de la Série (x_i, t_i)
Une équation de la droite d'ajustement de v en x, selon la méthode des moindres carrés (coefficients arrondis à 10^{-4} près) est $v = 0{,}0049\,x + 6{,}1300$.

EXPLOITATION DU MODELE OBTENU
On obtient $y = 459\,e^{0{,}0049\,x}$ et $z = 7032\,e^{-0{,}0033\,x}$.





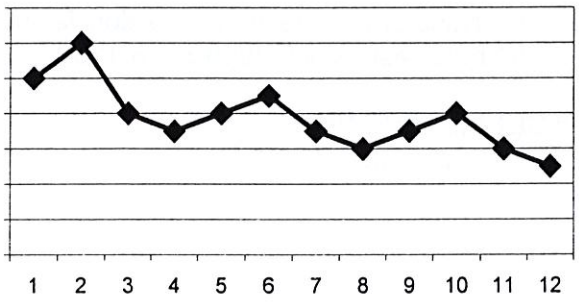
On en déduit que $y = z$ pour $x = (1/0{,}0082)\ln(7032/459) \approx 332{,}83$ €.

II – SERIES CHRONOLOGIQUES (1ère ES)

Ce thème n'est pas au programme de 1ères S. Seule la moyenne mobile est au programme de 1ère ES (comme il existe, selon les circonstances, plusieurs définitions, la définition utilisée devra être rappelée dans l'énoncé), les compléments apportés ci-dessous ne sont pas exigibles à l'examen.

1 – DECOMPOSITION D'UNE SERIE CHRONOLOGIQUE

Dès que l'on étudie les variations d'un phénomène au cours du temps, 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*** (utilisé pour la loi des grands nombres et les fluctuations d'échantillonnage) ***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***[1] suivant : $\boxed{X_t = f_t + s_t + \varepsilon_t}$,
où f_t représente la ***tendance***, s_t la ***saison*** et ε_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.

Lissage par 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***.
Bien qu'antérieure à l'avènement des ordinateurs, cette méthode "traditionnelle" est encore couramment utilisée, sous une forme ou sous une autre.
On compare la chronique $(X_t)_{t \in \{1, \ldots, 4n\}}$ avec la chronique $(Y_t)_{t \in \{3, \ldots, 4n-1\}}$ définie par :

$Y_t = \frac{1}{4}(X_{t-2} + X_{t-1} + X_t + X_{t+1})$ moyenne mobile d'ordre 4.

[1] On adopte ici un modèle additif, adapté lorsque les composantes agissent de façon "indépendante". Lorsque la composante saisonnière est proportionnelle à la tendance, le modèle multiplicatif $X_t = f_t \times s_t \times \varepsilon_t$ est plus adapté.

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 $Y_t = \frac{1}{2p+1}\sum_{j=-p}^{p} X_{t+j}$.

Exemple numérique[2]

On considère les ventes de voitures au Canada, en milliers, d'un constructeur automobile, sur cinq années consécutives.

Trimestre :	1	2	3	4
1997	710	1440	980	740
1998	770	1530	1020	760
1999	920	1300	940	660
2000	890	1220	820	840
2001	1000	1470	930	850

Cherchons l'effet du "***filtre***" moyenne mobile sur la chronique de départ $X_t = f_t + s_t + \varepsilon_t$.

On a $Y_t = \frac{1}{4}(f_{t-2} + f_{t-1} + f_t + f_{t+1}) + \frac{1}{4}(\varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t + \varepsilon_{t+1})$

puisque, par définition, $s_{t-2} + s_{t-1} + s_t + s_{t+1} = 0$.

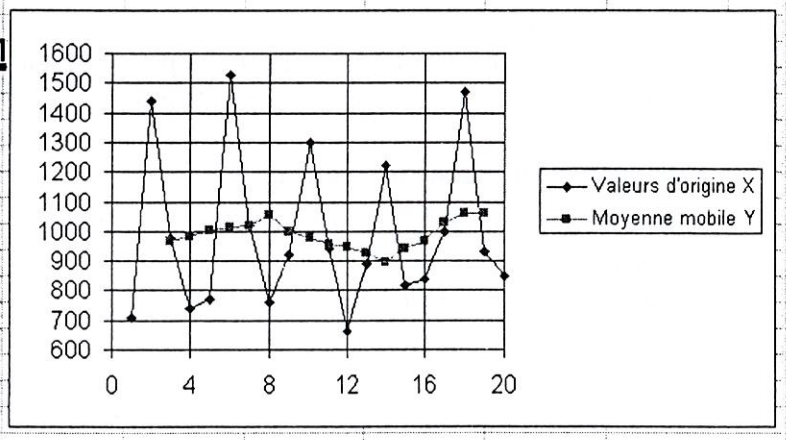
Comme f_t varie peu en fonction du temps, on a $\frac{1}{4}(f_{t-2} + f_{t-1} + f_t + f_{t+1}) \approx f_t$.

Si l'on suppose les aléas ε_t de même loi avec une moyenne nulle, un écart type σ 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 σ.

Les fluctuations aléatoires de Y_t sont donc plus faibles que celles de X_t, mais les aléas η_t ne sont plus indépendants.

C4 : =(B2+B3+B4+B5)/4

	A	B	C
1	t	X	Y
2	1	710	
3	2	1440	
4	3	980	967,5
5	4	740	982,5
6	5	770	1005
7	6	1530	1015
8	7	1020	1020
9	8	760	1057,5
10	9	920	1000
11	10	1300	980
12	11	940	955
13	12	660	947,5
14	13	890	927,5
15	14	1220	897,5
16	15	820	942,5
17	16	840	970
18	17	1000	1032,5
19	18	1470	1060
20	19	930	1062,5
21	20	850	
22			



[2] Source : adapté de Wonnacott – "Statistique" – Economica 1995.

Détermination de la composante saisonnière s_i

Comme $X_t - Y_t \approx s_t + (\varepsilon_t - \eta_t)$, on peut avoir un ***estimateur de la composante saisonnière*** (c'est *une* méthode) 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.

t	Série brute X_t	Moyenne mobile Y_t	$X_t - Y_t$	Composante saisonnière $\hat{s}$	$\hat{s}$ répété	Série corrigée des variations saisonnières $\hat{X}$
1	710				-96,25	806,25
2	1440				391,875	1048,125
3	980	967,5	12,5	-51,5	-51,5	1031,5
4	740	982,5	-242,5	-239,375	-239,375	979,375
5	770	1005	-235	-96,25	-96,25	866,25
6	1530	1015	515	391,875	391,875	1138,125
7	1020	1020	0		-51,5	1071,5
8	760	1057,5	-297,5		-239,375	999,375
9	920	1000	-80		-96,25	1016,25
10	1300	980	320		391,875	908,125
11	940	955	-15		-51,5	991,5
12	660	947,5	-287,5		-239,375	899,375
13	890	927,5	-37,5		-96,25	986,25
14	1220	897,5	322,5		391,875	828,125
15	820	942,5	-122,5		-51,5	871,5
16	840	970	-130		-239,375	1079,375
17	1000	1032,5	-32,5		-96,25	1096,25
18	1470	1060	410		391,875	1078,125
19	930	1062,5	-132,5		-51,5	981,5
20	850				-239,375	1089,375

On appelle ***série corrigée des variations saisonnières*** la chronique : $\hat{X}_t = X_t - \hat{s}_t$ avec $t \in \{1, 2, ..., 4n\}$.

La chronique (Y_t) des moyennes mobiles peut être interprétée comme une estimation de la tendance $(f_t)_{t \in \{3, ..., 4n-1\}}$, en l'absence d'autres informations.

Détermination de la tendance

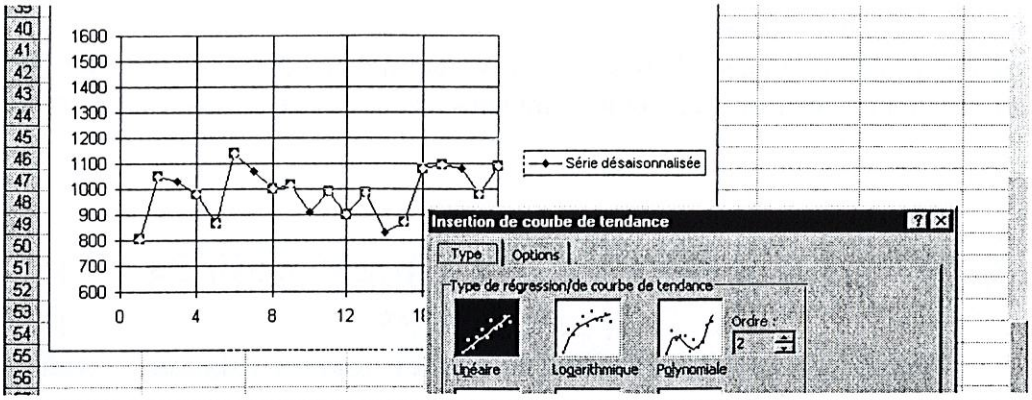
Il existe des situations où il est possible de proposer pour la tendance un certain type de fonction.

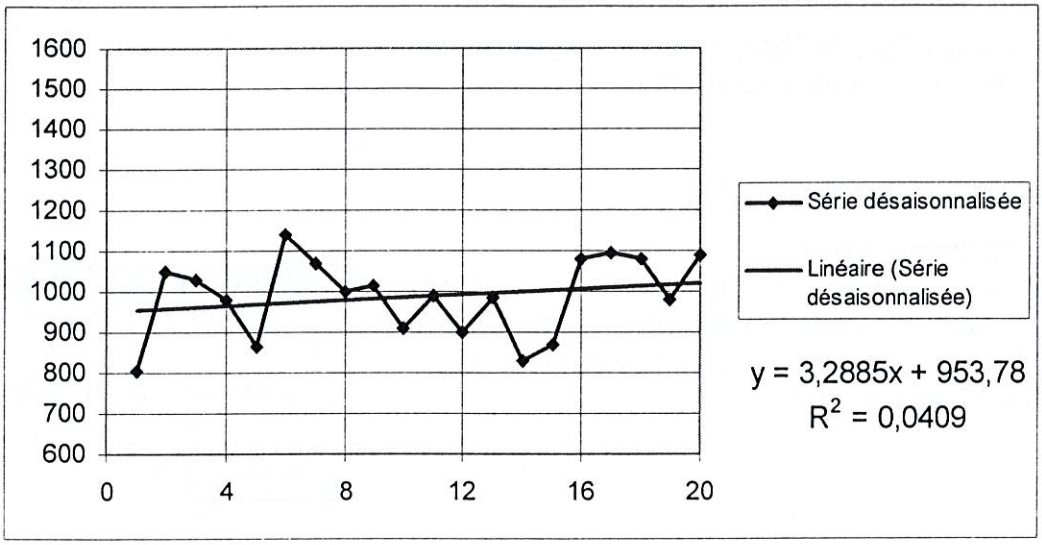
On peut avoir $f_t = a + bt$ ou $f_t = a + bt + ct^2$.

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.

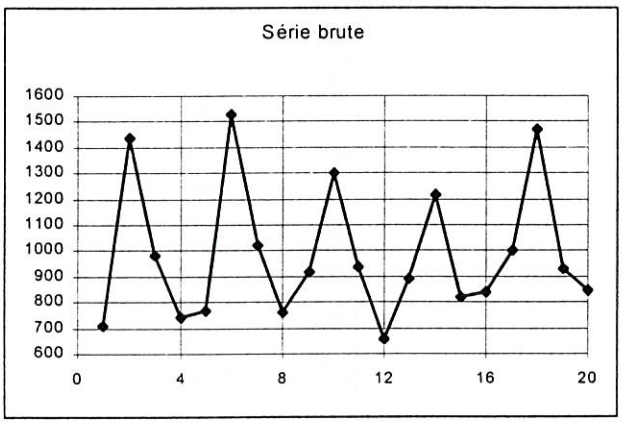
Dans le premier cas, on peut utiliser la méthode de ***régression linéaire selon les moindres carrés***.

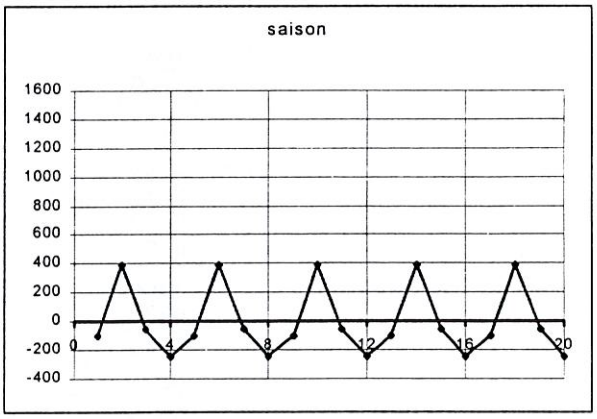
On obtient alors la tendance suivante :

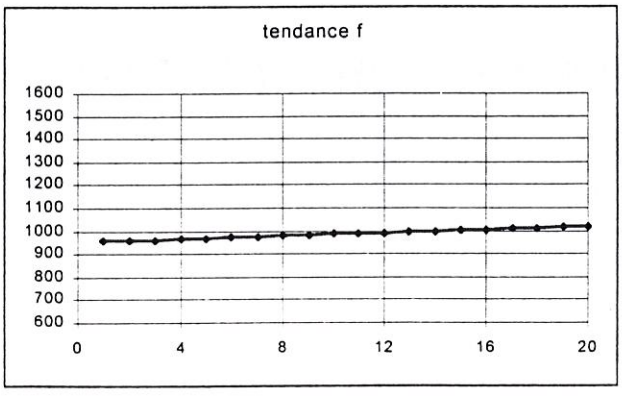


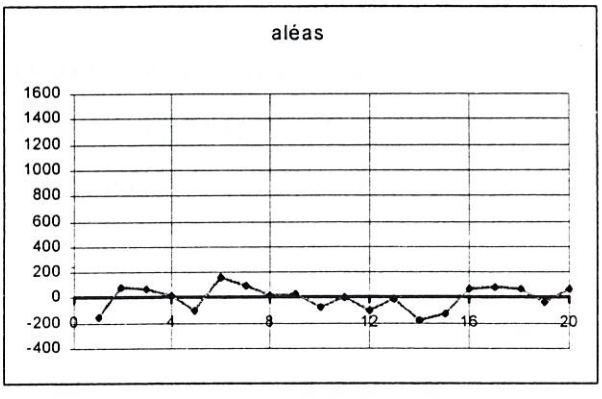


Décomposition de la série temporelle :









2 – UN T.P. EN PREMIERE ES SUR L'UTILISATION DE LA MOYENNE MOBILE

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.

UTILISATION DE MOYENNES MOBILES A LA BOURSE

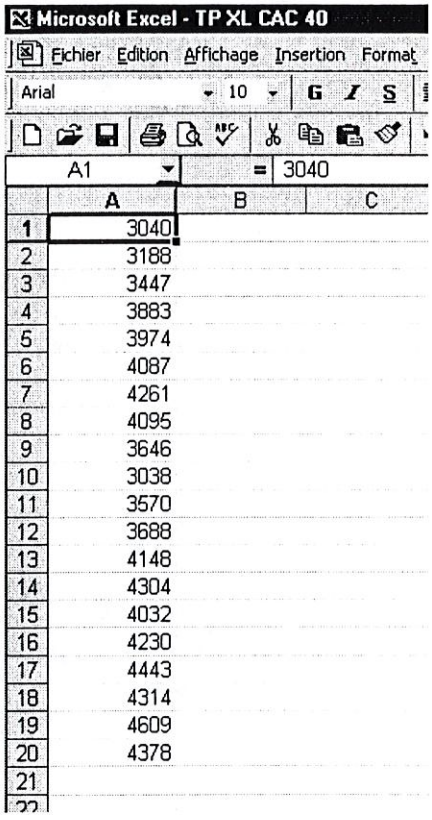
L'objectif est d'utiliser le lissage par les moyennes mobiles, pour obtenir, dans le contexte boursier, des signaux d'achat ou de vente.

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

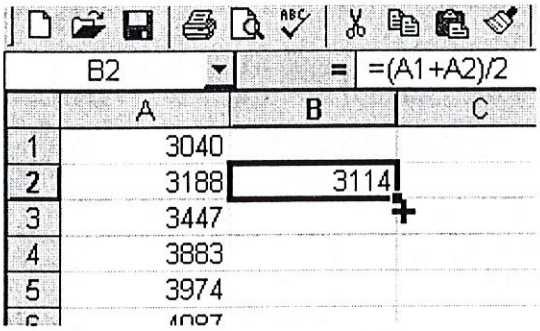


I – CALCUL ET REPRESENTATION DES MOYENNES MOBILES D'ORDRE 2 ET 5

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 2ème 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***).



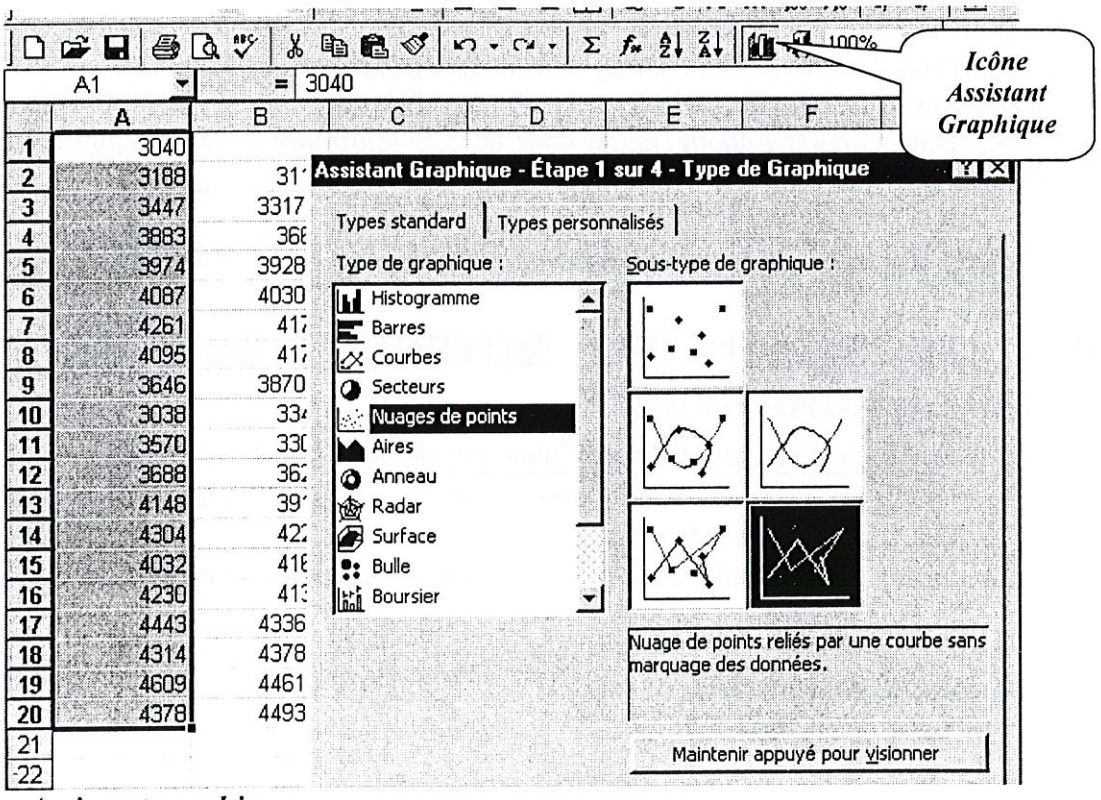
Approcher le pointeur de la souris du carré noir situé dans le coin inférieur droit de la cellule B2. Celui-ci se transforme en une croix noire, faire alors glisser, en maintenant le bouton gauche enfoncé pour ***recopier*** jusqu'en B20, puis relâcher le bouton de la souris.
La colonne B contient alors 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 : = SOMME(A1:A5)/5
puis recopier, comme précédemment, cette formule jusqu'en C20.

Représentations graphiques

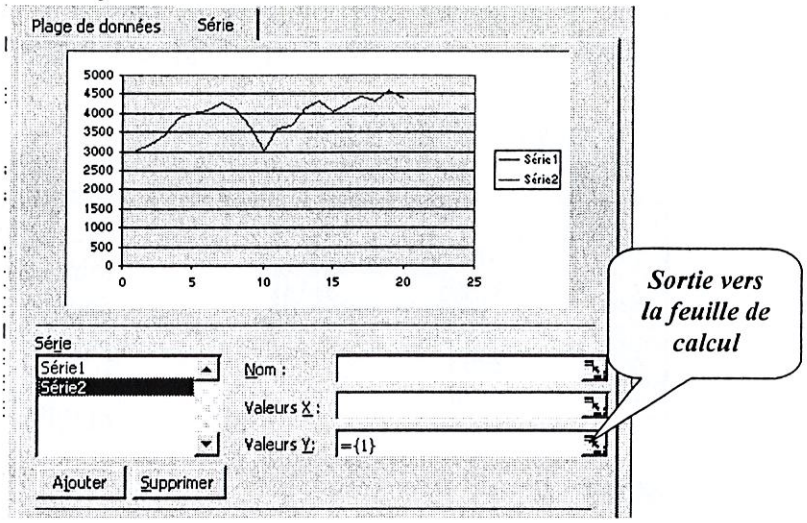
Sélectionner, avec le bouton gauche de la souris, les cellules de A1 à A20, puis appuyer sur



l'icône ***Assistant graphique***.

Etape 1/4 :
Dans ***Type de graphique***, choisir ***Nuages de points***, puis dans ***Sous-type de graphique***, ***Nuage de points reliées par une courbe sans marquage des données***.
Cliquer sur ***Suivant***.

Etape 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***.



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 (hausse supérieure à la tendance),
- **vendre** quand la moyenne mobile croise les cours à la baisse (baisse supérieure à la tendance).

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).

✍ FEUILLE REPONSE

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

.....

.....

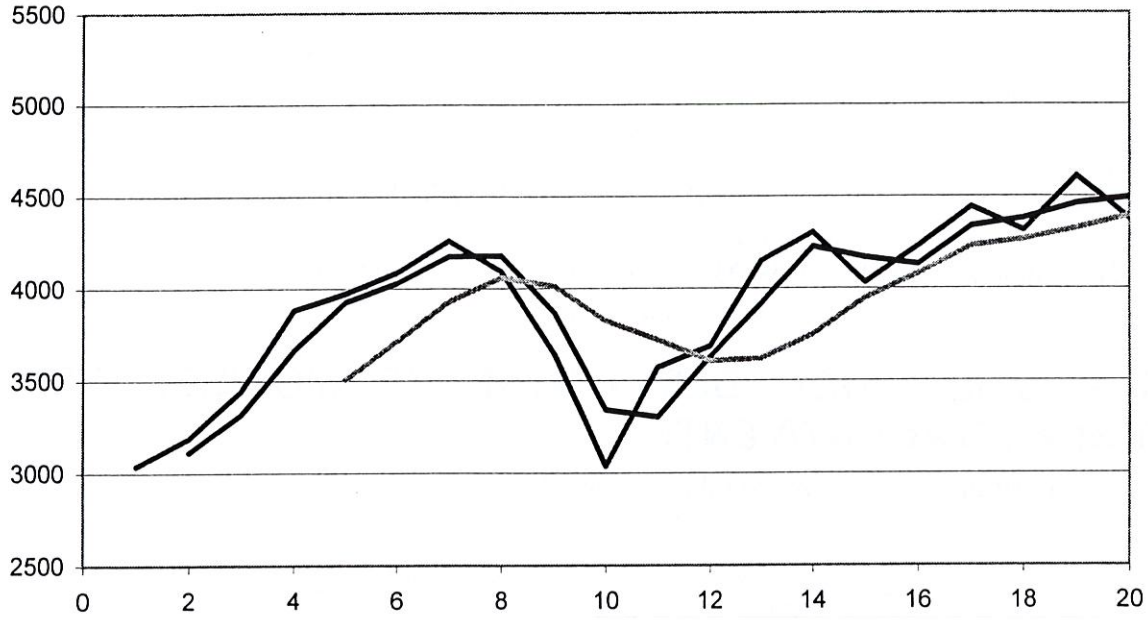
.....

.....

Corrigé du T.P.
"Utilisation des moyennes mobiles à la bourse"

On remarquera que la moyenne mobile utilisée dans ce contexte boursier ne correspond pas à la définition du document d'accompagnement du programme de 1ère ES (rentrée 2001) dans la mesure où, ignorant, bien entendu, les cours à venir, les moyennes mobiles ne sont calculées qu'à partir des valeurs qui précèdent.
Un énoncé d'examen devra préciser la définition de "moyenne mobile" utilisée, selon le contexte.

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 :	**+1362**

2) Utilisation de la moyenne mobile d'ordre 5 :

valeur d'achat	valeur de vente	gain (+ ou -)
3040	3646	+606
3688	4378	+690
	bilan :	**+1296**

3) Conclusion :
La stratégie à court terme, peut être plus risquée, s'est avérée, ici, plus avantageuse. Mais ce n'est pas un théorème...

III – CONDITIONNEMENT ET INDEPENDANCE

1 – NOTION DE PROBABILITE CONDITIONNELLE

Dans le paragraphe "Modalités de mise en œuvre" du programme de TS et TES de 2001, on lit, à propos de l'introduction de la notion de probabilité conditionnelle :

> *"On justifiera la définition de la probabilité de B sachant A, notée $P_A(B)$, par des calculs fréquentiels."*

Le document d'accompagnement des programmes donne quelques détails.

> *"Comme en classe de première, les calculs au niveau des fréquences sont transposés au niveau des probabilités d'événement et des lois de probabilité des variables aléatoires (conditionnement et indépendance). On garde constamment à l'esprit que les distributions des fréquences fluctuent, la loi de probabilité restant fixe..."*
> *"On peut [au niveau des fréquences], réfléchir au mode de calcul de la fréquence $f_A(B)$ et arriver à la formule :*
>
> $$f_A(B) = \frac{f(A \text{ et } B)}{f(A)}."$$
>
> *"Dans le paragraphe suivant, on donne un sens à l'égalité ci-dessus lorsque l'on remplace les fréquences d'évènements par des probabilités."*
> *"Par analogie avec les distributions des fréquences manipulées, on peut alors définir la probabilité de B sachant A par :*
>
> $$P_A(B) = \frac{P(A \text{ et } B)}{P(A)}$$
>
> *Document d'accompagnement du programme de TS.*

Le point de vue choisi, dans la continuité du programme de première, est celui de l'approche fréquentiste. Dans la pratique, les probabilités sont estimées à partir d'une étude statistique. Il s'agit d'une modélisation, justifiée par la loi des grands nombres, et dont l'objectif est l'inférence (passage d'un échantillon à la population ou passage d'un historique statistique à une prévision pour le futur). Les probabilités conditionnelles, outil théorique pour le calcul prévisionnel, n'échappent pas à ce schéma.
Cet aspect n'est pas toujours bien mis en valeur par les manuels scolaires de terminale. En particulier, certains exercices où l'on travaille sur un tableau d'effectifs, ne relèvent pas réellement d'un calcul probabiliste, mais d'un simple calcul statistique de fréquences conditionnelles. Dans un tel cas, où l'on considère que l'échantillon est la population totale où l'on prend au hasard un élément, le formalisme des probabilités conditionnelles paraît bien artificiel.

On propose, dans ce qui suit, une introduction au cours sous la forme d'un T.P. de simulation sur Excel, puis une manière d'écrire le début du cours, en guise de synthèse après ce T.P. .

2 – UN T.P. D'INTRODUCTION SUR EXCEL

Le contexte du T.P. est inspiré d'un article de *Ian Hacking* paru dans le magazine "Sciences et Avenir", numéro spécial "L'empire des probabilités" – Octobre/novembre 2001.

Pour estimer une probabilité conditionnelle (non intuitive), il s'agit d'étudier, par simulation, les fluctuations d'une fréquence conditionnelle et de la voir se stabiliser lorsque la taille de l'échantillon passe de $n = 100$ à $n = 1000$.

Le T.P. prend environ 1h40. C'est le prix à payer pour une introduction expérimentale de la notion de probabilité conditionnelle qui favorise l'approche du concept et permet de bien distinguer calcul statistique et calcul probabiliste.

Travaux pratiques STATISTIQUES

PROBABILITE CONDITIONNELLE : UN TAXI DANS LA BRUME

Fait divers :

Une ville a deux compagnies de taxi. Les taxis de l'une des compagnies sont bleus et 15% des taxis en circulation lui appartiennent. L'autre compagnie a des taxis verts et possède 85% des taxis en circulation. Par une nuit de brouillard, un taxi heurte une voiture garée et prend la fuite. Un témoin a vu la scène. On peut estimer que, dans de telles conditions de distance et de brouillard, son témoignage sur la couleur du taxi est fiable à 80%. Ce témoin dit que le taxi mis en cause est bleu.

Que pensez-vous ?

- ❑ Il y a une probabilité de 80% que le taxi soit bleu.
- ❑ Il est assez probable, mais pas à 80%, que le taxi soit bleu.
- ❑ Il est plutôt probable que le taxi soit vert.

✍ ***Compléter la feuille réponse.***

A – Simulation de l'expérience aléatoire

Préparer une feuille de calcul comme ci-dessous :

Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?

Arial 10 G I S € % ,00

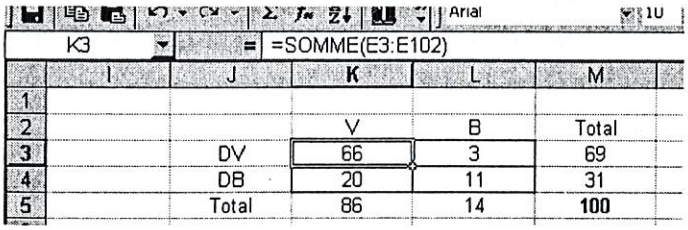
A3 = =SI(ENT(ALEA()+0,15)=1;"BLEU";"VERT")

	A	B	C	D	E	F	G	H
1	taux d'erreur :	0,2						
2	taxi	erreur ?	témoin		V et DV	V et DB	B et DV	B et DB
3	VERT	1	BLEU		0	1	0	0
4								
5								

La cellule ***A3*** contient la couleur du taxi, c'est à dire bleu, dans 15% des cas, ou vert dans 85% des cas. Entrer la ***formule*** : =SI(ENT(ALEA( ) + 0,15) = 1 ; "BLEU" ; "VERT")
La cellule ***B3*** contient la valeur 0 si le témoin ne fait pas d'erreur (dans 80% des cas) ou la valeur 1 lorsqu'il se trompe (20% des cas). Entrer la ***formule*** : =ENT(ALEA( ) + $B$1)
(le symbole $ bloque la référence de la cellule lorsqu'on recopie).
La cellule ***C3*** contient la couleur vue par le témoin : il dit vert ou il dit bleu. Entrer la ***formule*** : =SI(B3 = 0 ; A3 ; SI(A3 = "BLEU" ; "VERT" ; "BLEU"))
Dans les colonnes *E*, *F*, *G*, *H* on comptabilisera les quatre cas possible, en notant V ou B la couleur du taxi et DV ou DB la couleur dite par le témoin.
En ***E3*** entrer la ***formule*** : =SI(ET(A3="VERT";C3="VERT");1;0)
En ***F3*** entrer la ***formule*** : =SI(ET(A3="VERT";C3="BLEU");1;0)
En ***G3*** entrer la ***formule*** : =SI(ET(A3="BLEU";C3="VERT");1;0)
En ***H3*** entrer la ***formule*** : =SI(ET(A3="BLEU";C3="BLEU");1;0)

Sélectionner les cellules de ***A3*** à ***H3***, approcher le pointeur de la souris du coin inférieur droit de la sélection, puis, lorsque le pointeur s'est transformé en croix noire, ***recopier vers le bas*** jusqu'à la ligne 1002.
Vous avez ainsi simulé 1000 expériences aléatoires, selon les conditions rapportées dans le "fait divers".

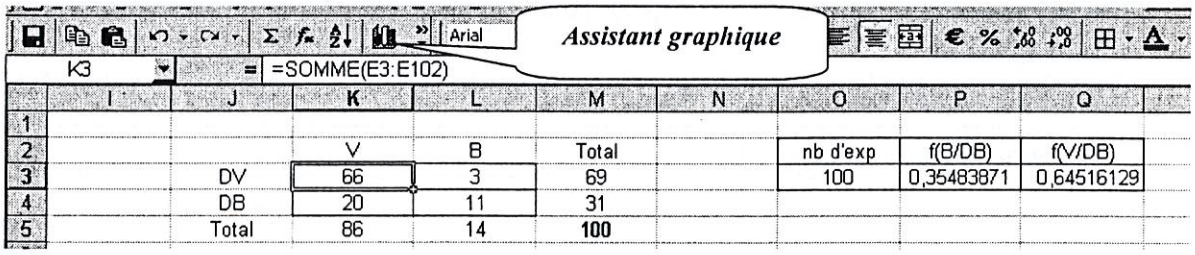
Fréquence conditionnelle sur 100 expériences



	I	J	K	L	M
1					
2			V	B	Total
3		DV	66	3	69
4		DB	20	11	31
5		Total	86	14	**100**

Dans les colonnes J à M, construire un tableau comptabilisant les quatre cas possibles.
En ***K3*** on entre la ***formule*** :
=SOMME(E3:E102)
En ***K4*** on entre la ***formule*** :
=SOMME(F3:F102)

En ***L3*** on entre la ***formule*** : =SOMME(G3:G102)
En ***L4*** on entre la ***formule*** : =SOMME(H3:H102)
Puis faire les sommes dans les marges du tableau.
Le témoin dit avoir vu un taxi bleu, événement noté DB. On examinera donc uniquement les cas DB. Dans ces cas là, la fréquence des situations où le taxi était réellement bleu est nommée **fréquence conditionnelle** de B sachant DB. On la note *f*(B/DB).



	I	J	K	L	M	N	O	P	Q
1									
2			V	B	Total		nb d'exp	f(B/DB)	f(V/DB)
3		DV	66	3	69		100	0,35483871	0,64516129
4		DB	20	11	31				
5		Total	86	14	**100**				

Dans la cellule ***P3*** calculer la fréquence conditionnelle *f*(B/DB) des expériences où, sachant que le témoin dit que le taxi est bleu, le taxi était réellement bleu, par la ***formule*** :
= L4 / M4
En ***Q3***, calculer la fréquence conditionnelle, dans le cas où le témoin voit un taxi bleu, des situations où il est vert *f*(V/DB) par la ***formule*** : = 1 – P3
Enfin, vous allez visualiser ces fréquences en construisant un histogramme. ***Sélectionner*** les cellules de ***P2*** à ***Q3*** puis cliquer sur l'icône de l'***Assistant graphique*** , choisir ***Histogramme*** et cliquer sur ***Terminer***. Effacez la légende en cliquant dessus avec le bouton droit de la souris.
Appuyer sur la touche ***F9*** pour visualiser d'autres séries de 100 simulations.

✍ ***Compléter la feuille réponse.***

Fréquence conditionnelle sur 1000 expériences

Comme dans au paragraphe précédent :

- Créer un tableau comptabilisant le nombre d'événements de chaque type, V et DV, V et DB, B et DV, B et DB pour 1000 expériences (lignes 3 à 1002).
- Calculer, pour ces 1000 simulations, les fréquences conditionnelles *f*(B/DB) et *f*(V/DB).
- Créer un histogramme représentant ces fréquences conditionnelles.

✍ ***Compléter la feuille réponse.***

B – Notion de probabilité conditionnelle

On a vu que la fréquence conditionnelle f(B/DB) de l'événement "le taxi est bleu sachant que le témoin a dit bleu" est obtenue par : $f(\text{B/DB}) = \frac{\text{nb de B et DB}}{\text{nb de DB}} = \frac{f(\text{B et DB})}{f(\text{DB})}$.

Lorsque le nombre n d'expériences indépendantes augmente, la fréquence f(DB) de l'événement DB, "le témoin a dit taxi bleu", se rapproche de sa probabilité P(DB). De même f(B et DB) se rapproche de la probabilité $P(\text{B} \cap \text{DB})$.

On a alors la fréquence conditionnelle f(B/DB) qui se rapproche de ce que l'on nomme **probabilité conditionnelle de B sachant DB**, notée P(B/DB) ou P_{DB}(B) , correspondant à $P(\text{B/DB}) = \frac{P(\text{B} \cap \text{DB})}{P(\text{DB})}$.

✍ ***Compléter la feuille réponse.***

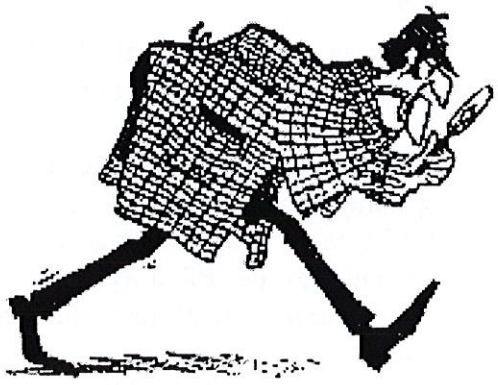
C – Taux de fiabilité du témoignage

Vous allez examiner l'impact du taux de fiabilité du témoin sur la fréquence conditionnelle f(B/DB).

Supposons que sont témoignage soit fiable à 85 %, remplacer le contenu de la cellule ***B1*** par la valeur 0,15 (c'est à dire 1 – 0,85). Faire plusieurs fois ***F9*** pour percevoir l'impact de ce changement.

Remplacer en suite le contenu de la cellule ***B1*** par la valeur 0, 10 puis par la valeur 0,50.

✍ ***Compléter la feuille réponse.***



✍ FEUILLE REPONSE

NOMS : .....

Le témoin affirme que le taxi est bleu. Que pensez-vous ? Cochez une case correspondant à votre opinion **avant tout calcul** (cette réponse n'est pas comptabilisée dans l'appréciation).

- ❑ Il y a une probabilité de 80% que le taxi soit bleu.
- ❑ Il est assez probable, mais pas à 80%, que le taxi soit bleu.
- ❑ Il est plutôt probable que le taxi soit vert.

A – Simulation de l'expérience aléatoire

Fréquence conditionnelle sur 100 expériences

Avec quels effectifs calcule-t-on la fréquence conditionnelle f(B/DB) en cellule P3 ?

$$f(\text{B} / \text{DB}) = \frac{\text{nombre de cas} \ldots\ldots}{\text{nombre de cas} \ldots\ldots} .$$

De même, comment calcule-t-on la fréquence conditionnelle f(V/DB) en cellule Q3 ?

$$f(\text{V} / \text{DB}) = \frac{\text{nombre de cas} \ldots\ldots}{\text{nombre de cas} \ldots\ldots} .$$

Comparer, sur vos simulations de taille 100, les deux fréquences conditionnelles, f(B/DB) et f(V/DB). Comment ses fréquences fluctuent-elles d'une simulation de 100 expériences à une autre ?

.....
.....
.....
.....

Fréquence conditionnelle sur 1000 expériences

Comparer, sur vos simulations de taille 1000, les deux fréquences conditionnelles, f(B/DB) et f(V/DB). Comment ses fréquences fluctuent-elles d'une simulation de 1000 expériences à une autre ?

.....
.....
.....
.....

Au vu des expériences précédentes, répondre à nouveau au questionnaire :
Le témoin affirme que le taxi est bleu. Que pensez-vous ?

- ❑ Il y a une probabilité de 80% que le taxi soit bleu.
- ❑ Il est assez probable, mais pas à 80%, que le taxi soit bleu.
- ❑ Il est plutôt probable que le taxi soit vert.

B – Notion de probabilité conditionnelle

L'énoncé du problème fournit les renseignements suivants :
$P(B) = \ldots\ldots$; $P(V) = \ldots\ldots$; $P(DB/B) = \ldots\ldots$ (probabilité que le témoin dise que le taxi est bleu, sachant qu'il est bleu) ; $P(DB/V) = \ldots\ldots$.

De la définition $P(DB/B) = \dfrac{P(B \cap DB)}{P(B)}$ on tire $P(B \cap DB) = P(B) \times P(DB/B)$.

Calculer : $P(B \cap DB) = \ldots\ldots$.
Calculer : $P(V \cap DB) = P(V) \times P(DB/V) = \ldots\ldots$.
En déduire $P(DB) = P(B \cap DB) + P(V \cap DB) = \ldots\ldots$.

Calculer la probabilité conditionnelle : $P(B/DB) = \dfrac{P(B \cap DB)}{P(DB)} = \dfrac{\ldots\ldots}{\ldots\ldots} \approx 0{,}41$.

Comparer la probabilité conditionnelle $P(B/DB)$ calculée précédemment avec la fréquence conditionnelle correspondante $f(B/DB)$ calculée sur 1000 simulations de l'expérience.
.....

C – Taux de fiabilité du témoignage

Lorsque B1 contient la valeur 0,15, que cela signifie-t-il ? .....
Qu'observez-vous pour les valeurs de $f(B/DB)$, pour des simulations de taille 1000, lorsque le contenu de B1 vaut 0,15 ?
.....
.....
Lorsque le contenu de B1 vaut 0,10 ?
.....
.....
On suppose que le témoin est fiable avec $P(DB/B) = P(DV/V) = x$.
Justifier que $P(DV/B) = P(DB/V) = 1 - x$. .....
Montrer, en reprenant les calculs du 2-, que : $P(B/DB) = \dfrac{0{,}15x}{0{,}15x + 0{,}85(1-x)}$.
.....
.....
.....
.....
.....
Rechercher x tel que $P(B/DB) = 0{,}5$.
.....
.....
Le résultat obtenu est-il compatible avec les fréquences $f(B/DB)$ observées lorsque le contenu de la cellule B1 était 0,15 ? Pourquoi ?
.....
.....
Lorsque le contenu de la cellule B1 vaut 0,50 , comparer les valeurs observées de $f(B/DB)$ avec $P(B) = 0{,}15$. .....
.....
Lorsque $x = 0{,}5$ montrer que $P(B/DB) = P(B)$.
.....
.....
Dans ce cas, le conditionnement par DB n'apporte rien à la réalisation de l'événement B, autrement dit le témoignage du témoin ne vaut rien. On dit que les évènements B ("le taxi est bleu") et DB ("le témoin dit que le taxi bleu") sont **indépendants**.

Eléments de solution pour l'activité EXCEL "PROBABILITE CONDITIONNELLE : UN TAXI DANS LA BRUME"

Durée : ≈ 1h 40.

A – Simulation de l'expérience aléatoire

Fréquence conditionnelle sur 100 expériences

On a $f(\text{B} / \text{DB}) = \dfrac{\text{nombre de cas B et DB}}{\text{nombre de cas DB}}$

et $f(\text{V} / \text{DB}) = \dfrac{\text{nombre de cas V et DB}}{\text{nombre de cas DB}}$.

De façon générale, on observe $f(\text{V} / \text{DB}) \geq f(\text{B} / \text{DB})$ mais les fluctuations entre les différents échantillons de taille 100 sont importantes, et on observe assez souvent $f(\text{V} / \text{DB}) < f(\text{B} / \text{DB})$.

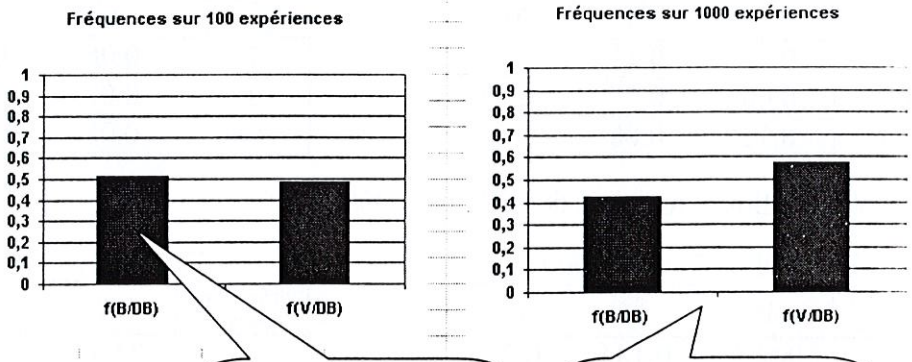
Fréquence conditionnelle sur 1000 expériences

Les fluctuations des fréquences conditionnelles sont beaucoup moins importantes et laissent à penser que lorsque le témoin voit le taxi bleu, le plus probable est qu'il soit vert.

	J	K	L	M
		V	B	Total
DV		68	3	71
DB		14	15	29
Total		82	18	**100**

	J	K	L	M
		V	B	Total
DV		656	33	689
DB		178	133	311
Total		834	166	**1000**

nb d'exp	f(B/DB)	f(V/DB)
100	0,51724138	0,48275862
1000	0,42765273	0,57234727



Sur 100 expériences, il se produit parfois que $f(\text{B/DB}) > f(\text{V/DB})$

Sur 1000 expériences, on observe toujours $f(\text{B/DB}) < f(\text{V/DB})$

B – Notion de probabilité conditionnelle

On a $P(\text{B} \cap \text{DB}) = 0{,}15 \times 0{,}80 = 0{,}12$.

On a $P(\text{DB}) = 0{,}15 \times 0{,}80 + 0{,}85 \times 0{,}20 = 0{,}29$.

On en déduit que $P(\text{B/DB}) = 0{,}12 / 0{,}29 \approx 0{,}41$.

Les fréquences $f(\text{B/DB})$ observées sur des simulations de taille 1000 sont assez proches de cette valeur.

C – Taux de fiabilité du témoignage
En posant $P(DB/B) = P(DV/V) = x$, on a $P(B/DB) = 0,5$ pour $x = 0,85$.
On constate que lorsque le contenu de la cellule B1 est $1 - x = 0,15$ les fréquences conditionnelles $f(B/DB)$ et $f(V/DB)$ sont relativement proches de 0,5.
En revanche, lorsque le témoin est fiable à 90% (B1 contient la valeur 0,10), on a régulièrement $f(B/DB) > f(V/DB)$.
On retrouve que lorsque le témoin répond "au hasard" (fiabilité 50%, $x = 0,5$), l'information apportée par son témoignage est sans intérêt : le fait que le taxi soit bleu est indépendant du témoignage "bleu" du témoin (ou de DV).

3 – UNE ECRITURE DU DEBUT DU COURS SUR LE CONDITIONNEMENT (PROBABILITE CONDITIONNELLE)

Situation du problème

Une population (très importante et supposée homogène) de bovins est susceptible d'avoir une maladie M. Un test de dépistage (positif $T+$, ou négatif $T-$) permet de détecter la maladie, sans toutefois être fiable à 100 %. On souhaite évaluer la fiabilité de ce test.
Pour ce faire, on prélève au hasard des échantillons dans la population, sur lesquels on pratique le test de dépistage, puis, par une analyse vétérinaire approfondie, on constate si l'animal est malade ou non.
On a obtenu les résultats suivants :

• Deux échantillons aléatoires de taille $n = 100$:

Echantillon 1 : $n = 100$

Fréquences observées	M	$\overline{M}$	total
$T+$	0,12	0,11	0,23
T–	0,00	0,77	0,77
total	0,12	0,88	1

Echantillon 2 : $n = 100$

Fréquences observées	M	$\overline{M}$	total
$T+$	0,06	0,23	0,29
T–	0,02	0,69	0,71
total	0,08	0,92	1

• Deux échantillons aléatoires de taille $n = 1000$:

Echantillon 3 : $n = 1000$

Fréquences observées	M	$\overline{M}$	total
$T+$	0,085	0,184	0,269
T–	0,011	0,720	0,731
total	0,096	0,904	1

Echantillon 4 : $n = 1000$

Fréquences observées	M	$\overline{M}$	total
$T+$	0,092	0,178	0,270
T–	0,009	0,721	0,730
total	0,101	0,899	1

On constate que ces fréquences fluctuent, mais cependant sensiblement moins lorsque $n = 1000$.
Pour chaque échantillon, on peut calculer, parmi les animaux malades, la fréquence de ceux qui ont été testés positifs. Cette fréquence est notée $f_M(T+)$. On l'obtient en faisant :

$$f_M(T+) = \frac{\text{nombre de } M \text{ et } T+}{\text{nombre de } M} = \frac{f(M \text{ et } T+)}{f(M)}.$$

Pour les échantillons prélevés, on trouve :

échantillon	1	2	3	4
$f_M(T+) \approx$	1	0,75	0,885	0,911

D'après la loi des grands nombres, lorsque n augmente, chacune des fréquences situées dans les quatre cases centrales des tableaux d'échantillon , $f(M \cap T+)$, $f(M \cap T-)$, $f(\overline{M} \cap T+)$, $f(\overline{M} \cap T-)$, tend à se rapprocher de la probabilité de l'événement correspondant dans la population totale : $P(M \cap T+)$, $P(M \cap T-)$, $P(\overline{M} \cap T+)$, $P(\overline{M} \cap T-)$.

On prendra ici, comme modèle, le tableau de probabilités suivant :

Probabilités	M	$\overline{M}$	total
$T+$	0,09	0,18	0,27
T–	0,01	0,72	0,73
total	0,10	0,90	1

Fiabilité du test dans le cas d'un animal malade :
La fréquence conditionnelle des animaux testés positifs parmi les malades d'un échantillon, tend, quand n augmente, à se rapprocher de ce que l'on nomme "probabilité conditionnelle de $T+$ sachant M " (lorsque l'on prend au hasard un animal dans la population) et que l'on peut calculer de façon analogue à celle de la fréquence conditionnelle :

$$P_M(T+) = \frac{P(M \cap T+)}{P(M)} = \frac{0,09}{0,10} = 0,90 .$$

On peut donc estimer que le test détecte un animal malade à 90%.

Fiabilité du test dans le cas non malade :
De même, la fréquence conditionnelle des animaux testés négatifs parmi les non malades d'un échantillon, tend, quand n augmente, à se rapprocher de la "probabilité conditionnelle de $T-$ sachant $\overline{M}$ ", que l'on peut calculer de façon analogue à celle de la fréquence conditionnelle :

$$P_{\overline{M}}(T-) = \frac{P(\overline{M} \cap T-)}{P(\overline{M})} = \frac{0,72}{0,90} = 0,80 .$$

On peut donc estimer que le test détecte un animal non malade à 80%.

Définition

Soit A et B deux évènements, avec $P(A) \neq 0$.

On appelle probabilité de B sachant A , notée $P_A(B)$, le nombre : $P_A(B) = \frac{P(A \cap B)}{P(A)}$.

Remarque : on a alors $P(A \cap B) = P(A) \times P_A(B)$.

Calculs probabilistes

On reprend l'exemple précédent.
Lorsque l'on pratique le test de dépistage sur un animal de la population, on connaîtra le résultat du test, mais on ignorera si l'animal est réellement malade, ou non malade. Le modèle probabiliste établi précédemment permet alors de se faire une opinion.
Un animal, pris au hasard dans la population, a été testé positif. On ignore s'il est réellement malade. On peut calculer, selon le modèle adopté ci-dessus, la probabilité qu'il soit malade, c'est à dire $P_{T+}(M)$.

On a $P_{T+}(M) = \dfrac{P(M \cap T+)}{P(T+)} = \dfrac{0{,}09}{0{,}27} \approx 0{,}33$.

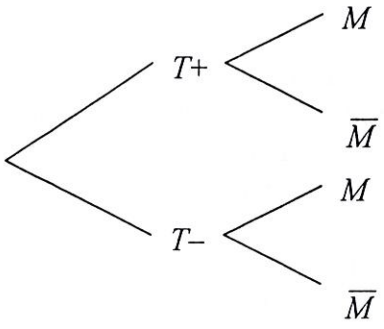
Ainsi, un animal testé positif, a plus de chances de ne pas être malade (2 chances sur 3). La question qui peut se poser est de savoir si, au nom du principe de précaution, il faut abattre tous les animaux testés positifs…

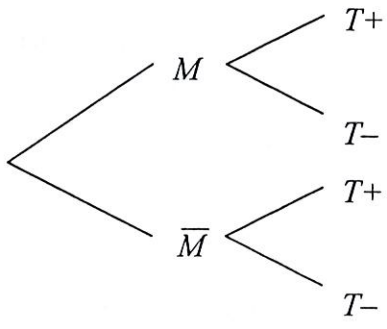
On peut aussi calculer $P_{T-}(M) = \dfrac{P(M \cap T-)}{P(T-)} = \dfrac{0{,}01}{0{,}73} \approx 0{,}01$.

Seulement 1% des animaux testés négatifs sont en fait malades.

Visualisation avec des arbres probabilistes

Compléter les branches des arbres ci-dessous, en indiquant les probabilités correspondantes.





4 – UN T.P. SUR CALCULATRICE

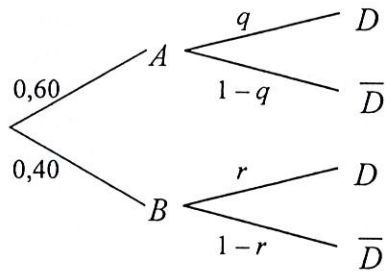
Le T.P. suivant confronte calculs probabilistes ("théoriques") et observation des fréquences obtenues par simulation, sur calculatrices.

T.P. SUR CALCULATRICES

CONDITIONNEMENT ET INDEPENDANCE : PIECES DEFECTUEUSES

NOMS : .....

Deux machines A et B fabriquent une pièce identique et fonctionnent indépendamment.
Dans la production de ces machines, on prend une pièce au hasard. La probabilité que cette pièce provienne de A est 0,6. La probabilité qu'elle provienne de B est donc 0,4.
On suppose que, parmi la production de A, la probabilité de tirer une pièce défectueuse est q, alors que parmi la production de B, la probabilité de tirer une pièce défectueuse est r.
On note D l'événement "la pièce tirée est défectueuse" et $\overline{D}$ l'événement contraire.
La situation est donc résumée par l'arbre suivant :



Arbre 1

On se pose la question suivante :

Lorsque l'on tire au hasard une pièce dans la production totale, et que l'on constate que la pièce est défectueuse, cette information donne-t-elle une idée de nature probabiliste sur sa fabrication éventuelle par A ?

Expérience statistique par simulation d'une production de 1000 pièces

Un programme permettra de simuler une production de 1000 pièces, en respectant les conditions précédentes. Il fournira le décompte, sur cette production, des pièces de chaque type, selon les deux critères d'origine, A ou B, et de qualité, D (défectueuse) ou non.
Dans ce programme, l'instruction 2 * int(rand + 0.6) → A attribue à la mémoire A la valeur 2 lorsque l'événement A est réalisé et la valeur 0 sinon, lorsque l'évènement B est réalisé.
Selon les cas, l'instruction int(rand + Q) → D ou int(rand + R) → D attribuera à la mémoire D la valeur 1 lorsque l'événement D est réalisé, ou 0 si $\overline{D}$ est réalisé.
Indiquer la valeur de A + D + 1 selon les évènements réalisés :

évènement	$A \cap D$	$A \cap \overline{D}$	$B \cap D$	$B \cap \overline{D}$
Valeur de A + D + 1				

Entrez dans votre calculatrice le programme suivant.

CASIO Graph 25→ 100	TI 80 82 83	TI 89 92
"Q" ↵ ? → Q ↵ "R" ↵ ? → R ↵ Seq(0,I,1,4,1) → List 1 ↵ For 1 → I To 1000 ↵ 2Int(Ran# + 0.6) → A ↵ If A = 2 ↵ Then Int(Ran#+Q) → D ↵ Else Int(Ran#+R) → D ↵ IfEnd ↵ List 1[A+D+1] + 1 → List 1[A+D+1] ↵ Next ↵ "AD" ↵ List 1[4] ↵ "A NON D" ↵ List 1[3] ↵ "BD" ↵ List 1[2] ↵ "B NON D" ↵ List 1[1]	:Disp "Q" :Input Q :Disp "R" :Input R :seq (0,I,1,4,1) → L_1 :For (I,1,1000) :2 int(rand + 0.6) → A :If A = 2 :Then :int(rand + Q) → D :Else :int(rand + R) → D :End :L_1(A+D+1) + 1 → L_1(A+D+1) :End :Disp "AD" :Disp L_1(4) :Pause :Disp "A NON D" :Disp L_1(3) :Pause :Disp "BD" :Disp L_1(2) :Pause :Disp "B NON D" :Disp L_1(1)	:DelVar s :Disp "q" :Input q :Disp "r" :Input r :seq (0,i,1,4,1) → s :For i,1,1000 :2*int(rand( ) + 0.6) → a :If a = 2 Then :int(rand( ) + q) → d :Else :int(rand( ) + r) → d :EndIf :s[a+d+1] + 1 → s[a+d+1] :EndFor :Disp "AD" :Disp s[4] :Pause :Disp "A NON D" :Disp s[3] :Pause :Disp "BD" :Disp s[2] :Pause :Disp "B NON D" :Disp s[1]

⇨ Pour obtenir certaines instructions :

• CASIO Graph 25 → 100 : **Seq** par OPTN LIST ; **List** par OPTN LIST ; **For To Next** par PRGM COM ; **If Then Else IfEnd** par PRGMpuis COM ; **Int** par OPTN NUM ; **Ran#** par OPTN PROB ; = par PRGM REL ; ⁄⁄ par PRGM ; [au clavier.

• TI 80 → 92 :

Utilisation possible de la fonction **CATALOG** (sur TI 83 – 89 – 92).

Seq par 2nd LIST OPS ; L_1 au clavier par 2nd ; **For End** par PRGM CTL ; → par STO ▸ ; **If Then Else End** par PRGM CTL ; **int** par MATH NUM ; **rand** par MATH PRB ; **If** par PRGM CTL ; = par 2nd TEST (2nd MATH TEST sur TI 92) ; **Disp** par PRGM I/O ; **Pause** par PRGM CTL.

Utiliser le programme précédent dans les deux cas suivants :

⇨ q = 0,10 et r = 0,03 :

Compléter le tableau ci-dessous, avec les effectifs donnés par votre simulation :

	A	B	Total
D			
$\overline{D}$			
Total			1000

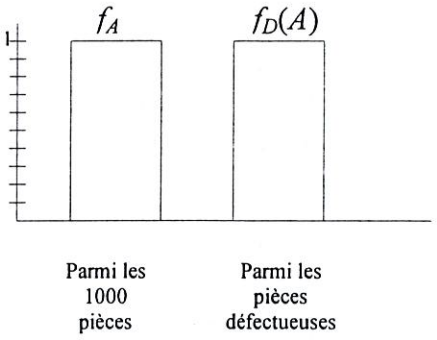
a) Quelle est, parmi la production simulée de 1000 pièces, la fréquence $f(A)$ des pièces provenant de A ? .....

b) Quelle est, parmi la production des 1000 pièces, la fréquence $f(A \cap D)$ des pièces provenant de A et défectueuses ? .....

c) Quelle est, *parmi les pièces défectueuses*, la fréquence conditionnelle notée $f_D(A)$ des pièces provenant de A ?

$f_D(A) = \dfrac{\text{nb pièces de type } A \text{ et } D}{\text{nb pièces } D} = \ldots\ldots\ldots\ldots$

Indiquer sur l'histogramme la proportion des pièces provenant de A, selon que l'on se place dans la production totale de 1000 pièces ou parmi les pièces défectueuses.



⇨ **$q = 0{,}10$ et $r = 0{,}10$:**

Compléter le tableau ci-dessous :

	A	B	Total
D			
$\overline{D}$			
Total			1000

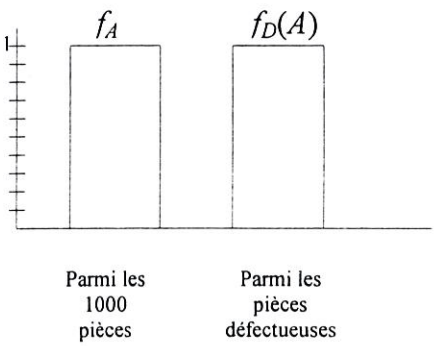
a) Quelle est, parmi la production simulée de 1000 pièces, la fréquence $f(A)$ des pièces provenant de A ? .....

b) Quelle est, parmi la production des 1000 pièces, la fréquence $f(A \cap D)$ des pièces provenant de A et défectueuses ? .....

c) Quelle est, *parmi les pièces défectueuses*, la fréquence conditionnelle notée $f_D(A)$ des pièces provenant de A ?

$f_D(A) = \dfrac{\text{nb pièces de type } A \text{ et } D}{\text{nb pièces } D} = \ldots\ldots\ldots\ldots$

Indiquer sur l'histogramme la proportion des pièces provenant de A, selon que l'on se place dans la production totale de 1000 pièces ou parmi les pièces défectueuses.



e) Dans chaque cas, d'après les graphiques, le fait de savoir que la pièce tirée est défectueuse modifie-t-il significativement la répartition statistique du critère "fabriquée par A" ?

$q = 0{,}10$ et $r = 0{,}03$: .....

$q = 0{,}10$ et $r = 0{,}10$: .....

Fréquence des pièces provenant de *A* parmi les défectueuses d'une production de 10 000 pièces

Si, au lieu de tirer une seule pièce au hasard dans la production, on recommence (avec remise) jusqu'à tirer une pièce défectueuse, il s'agit d'une nouvelle expérience aléatoire, que l'on peut résumer en disant que l'on tire au hasard une pièce défectueuse dans la production.
On s'intéresse à la probabilité que cette pièce défectueuse provienne de A.
En simulant plusieurs fois une production de 1000 pièces, on peut se faire une idée de cette probabilité. La moyenne des fréquences conditionnelles $f_D(A)$ devant s'en rapprocher.
En complétant vos simulations précédentes avec celles d'autres élèves, remplir le tableau suivant :

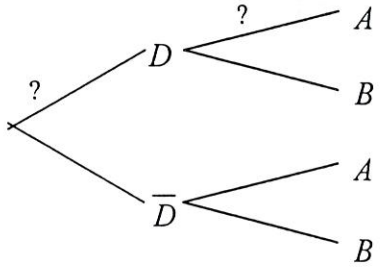
Simulations	1	2	3	4	5	6	7	8	9	10	Moyenne
$f_D(A)$ avec q=0,10 et r=0,03											
$f_D(A)$ avec q=0,10 et r=0,10											

A combien pouvez-vous estimer la probabilité qu'une pièce, tirée au hasard parmi les défectueuses, provienne de A ?
Lorsque $q = 0{,}10$ et $r = 0{,}03$: .....
Lorsque $q = 0{,}10$ et $r = 0{,}10$: .....

Etude probabiliste

On tire au hasard une pièce dans la production. On note A l'événement "la pièce tirée provient de la machine A", B l'événement "la pièce tirée provient de la machine B" et D l'événement "la pièce tirée est défectueuse".
a) On suppose que $q = 0{,}10$ et $r = 0{,}03$.
Calculer $P(A)$, $P(A \cap D)$ et $P(D) = P(A \cap D) + P(B \cap D)$ (on peut s'aider de l'arbre 1).
.....
.....
.....
.....
.....
.....

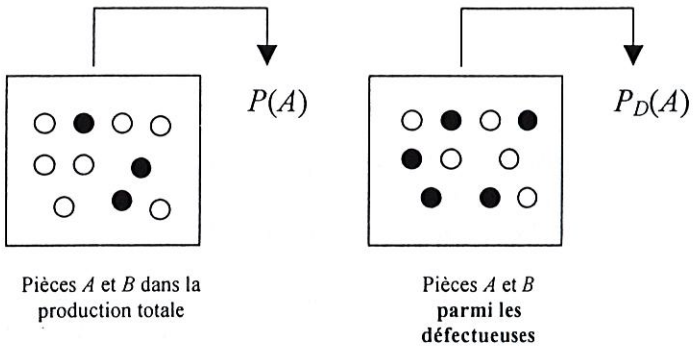
"Inverser" l'arbre 1, en indiquant sur l'arbre ci-dessous les probabilités correspondant aux points d'interrogation.



Arbre 2

On note $P_D(A)$ la ***probabilité conditionnelle*** de A sachant D correspondant à $P_D(A) = \dfrac{P(A \cap D)}{P(D)}$.

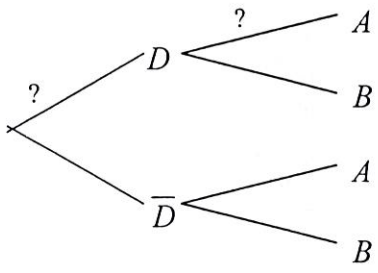
Que vaut $P_D(A)$? .....
Comparer avec l'estimation obtenue par simulation. .....



b) On suppose que $q = 0{,}10$ et $r = 0{,}10$.
Calculer $P(A)$, $P(A \cap D)$ et $P(D)$ (on peut s'aider de l'arbre 1).

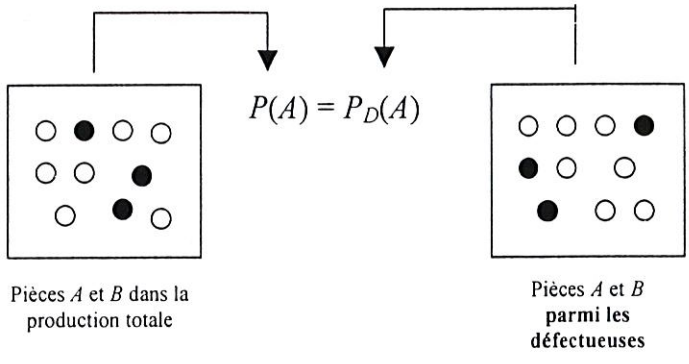
.....
.....
.....
.....
.....

"Inverser" l'arbre 1, en indiquant sur l'arbre ci-dessous les probabilités correspondant aux points d'interrogation .



Arbre 2

Lorsque $P(A \cap D) = P(A) \times P(D)$, on dit que les événements A et D sont **indépendants**.
On alors $P_D(A) = \dfrac{P(A \cap D)}{P(D)} = P(A)$, c'est à dire que l'information D n'influe pas sur la probabilité $P(A)$.
Que vaut $P_D(A)$? Que constate-t-on ? .....
.....



Les proportions sont les mêmes

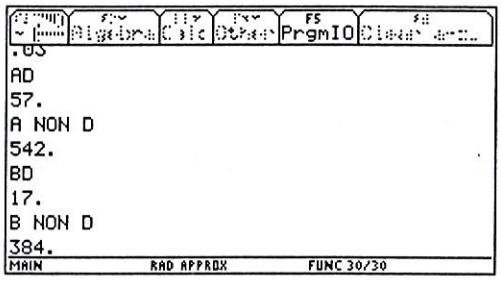
Eléments de réponse

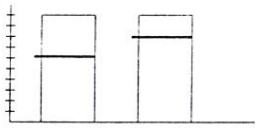
Durée : ≈ 1h30.

Production de 1000 pièces

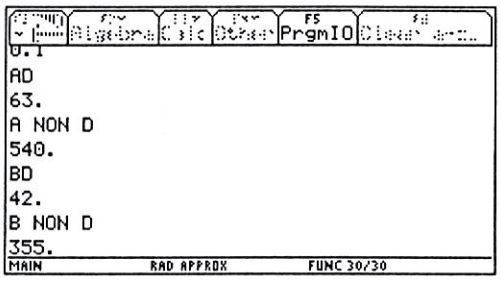
Il n'est pas question de probabilités dans cette question mais de simples statistiques sur une production de 1000 pièces, entièrement connue.
A titre d'exemple, voici un résultat possible du programme :

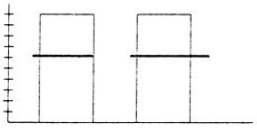
⇨ q = 0,10 et r = 0,03.
a) f_A = 0,599.
b) $f_{A\cap D}$ = 0,057.
c) $f_{A|D}$ = 57/(57+17) ≈ 0,77.





⇨ q = 0,10 et r = 0,10.
a) f_A = 0,603.
b) $f_{A\cap D}$ = 0,063.
c) $f_{A|D}$ = 63/(63+42) = 0,60.





e) On constate que c'est seulement dans le premier cas que la répartition du critère de provenance est significativement différente selon que l'on se place dans la production des 1000 pièces ou parmi les pièces défectueuses.

Production de 10000 pièces
En augmentant le nombre de simulations de la production, $f_{A|D}$ doit converger vers P(A | D).
Voici un exemple de 10 simulations de productions de 1000 pièces :

Simulations	1	2	3	4	5	6	7	8	9	10	moyenne ≈
$f_{A\|D}$ avec q=0,10 et r=0,03	0,77	0,83	0,93	0,84	0,81	0,85	0,83	0,78	0,79	0,84	0,83
$f_{A\|D}$ avec q=0,10 et r=0,10	0,60	0,58	0,56	0,64	0,56	0,62	0,58	0,51	0,54	0,69	0,59

On estime donc que la probabilité qu'une pièce défectueuse provienne de A est de l'ordre de 0,83 lorsque q=0,10 et r=0,03 et de l'ordre de 0,59 lorsque q=r=0,10.

Etude probabiliste

a) P(A) = 0,60 ; P(A∩D) = 0,60 × 0,10 = 0,06 d'après l'arbre 1 et P(D) = P(A∩D) + P(B∩D) (cas disjoints) d'où P(D) = 0,06 + 0,40×0,03 = 0,072.
Sur l'arbre 2, les données sont P(D) = 0,072 et P(A∩D) = 0,06, on en déduit que la probabilité figurant sur la branche menant de D à A est 0,06 ÷ 0,072 ≈ 0,83.
On a ainsi P(A/D) ≈ 0,83 , résultat approximativement égal à celui donné par les fréquences conditionnelles simulées.

b) P(A) = 0,60 ; P(A∩D) = 0,60 × 0,10 = 0,06 d'après l'arbre 1 et P(D) = P(A∩D) + P(B∩D) (cas disjoints) d'où P(D) = 0,06 + 0,40×0,10 = 0,10.
Sur l'arbre 2, les données sont P(D) = 0,10 et P(A∩D) = 0,06, on en déduit que la probabilité figurant sur la branche menant de D à A est 0,06 ÷ 0,10 = 0,60.
On a ainsi P(A/D) = 0,60 , résultat approximativement égal à celui donné par les fréquences conditionnelles simulées.
Dans ce cas, on constate que P(A/D) = 0,60 = P(A), il y a indépendance des événements A et D.

5 – DES EXERCICES DE CONDITIONNEMENT "A CONTEXTE"

Les deux exercices suivants complètent l'offre des manuels scolaires, en insistant sur l'importance des probabilités conditionnelles dans des questions de société (quand l'argumentation scientifique s'immisce dans le débat public).

Exercices

CONDITIONNEMENT

1 ALERTE A THREE MILE ISLAND

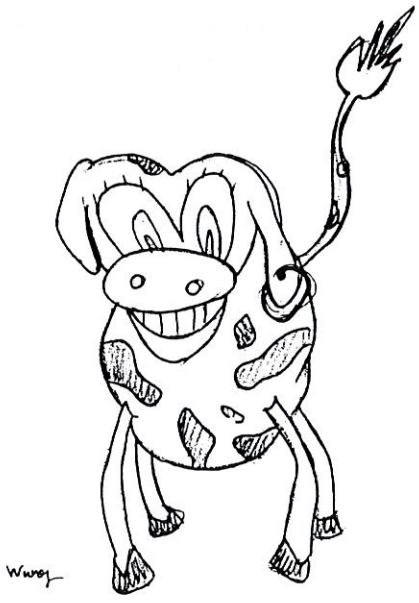
Dans les années 1980, une importante catastrophe nucléaire se produisit dans la centrale américaine de *Three Mile Island*. Différents voyants et indicateurs du tableau de bord de la salle de contrôle avaient pourtant alerté d'une baisse importante du niveau d'eau dans le réacteur. Mais dans les mois précédents, plusieurs fausses alarmes s'étaient produites, de sorte que l'opérateur ne pris pas au sérieux l'alerte qui était réelle, et tarda à réagir.

On suppose que s'il y a danger, l'alerte est donnée avec 99 % de certitude. S'il n'y a pas danger, l'alarme peut se déclencher (voyant défectueux par exemple, faux contact dans le tableau de commande…) et donner lieu à une fausse alerte avec la probabilité 0,005.
La probabilité pour qu'un jour tiré au hasard, un danger se présente est 0,001.
Pour un jour au hasard, on note D l'événement "un danger se présente" et A l'événement "l'alarme se déclenche".

Le système de contrôle déclenche une alerte.
Quelle est la probabilité que ce soit une fausse alerte ?

2 VACHE FOLLE [1]

Dans le journal *L'Express*, semaine du 1er au 7 mars 2001, l'ancien ministre *Claude Allègre* écrivait :
"Prenons pour exemple une maladie rare qui atteint, disons 1 individu sur 10000 (ce peut être la maladie de la vache folle). On met au point un test pour détecter si un individu est infecté par la maladie. Ce test donne des résultats fiables dans 99,9 % des cas. C'est donc a priori un excellent test. Supposons à présent qu'on pratique le test sur un individu pris au hasard et que ce test soit positif. Quelle est la probabilité que l'individu ainsi testé soit effectivement infecté ? Le calcul des probabilités nous répond sans hésiter 10 % (9 % exactement). Autrement dit, alors que le test est positif, l'individu a 90 % d'être sain ! "



1) L'affirmation ci-dessus est-elle exacte ? Pour un individu tiré au hasard, on notera :
M l'événement : "l'individu est malade" ; et $T+$ l'événement : "l'individu a été testé positif".
2) Dans son article, *C. Allègre* poursuit : *"Si on applique un tel test, excellent mais imparfait, pour dépister la maladie de la vache folle et qu'on abatte les vaches testées positives, on va déclencher un massacre bovin généralisé, ruineux et inutile."*
Sur un cheptel de 10000 vaches, combien de tests positifs peut-on s'attendre à obtenir ?

[1] D'après un article de *Gérard Kuntz* – Repères-IREM octobre 2001.

Commenter l'expression *massacre bovin généralisé*. Est-il "raisonnable" d'appliquer ici le "principe de précaution" ?

3) Lorsque l'on sait que le test est positif, par combien la probabilité d'être malade est-elle multipliée ?

4) On désigne maintenant par p la probabilité qu'un individu soit malade.

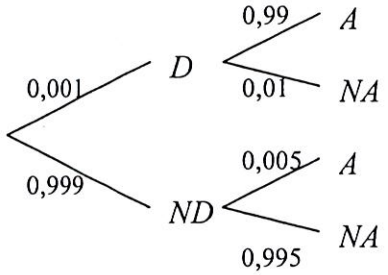
a) Exprimer, en fonction de p, la probabilité conditionnelle $P_{T+}(M)$, nommée "valeur prédictive" du test.

b) Pour quelle valeur de p la valeur prédictive du test dépasse-t-elle 50 % ?

Eléments de solution

1 - ALERTE A THREE MILE ISLAND

On note D l'évènement : "il y a danger".
On note A l'événement : "l'alerte se déclenche".
Les données sont résumées sur l'arbre suivant :



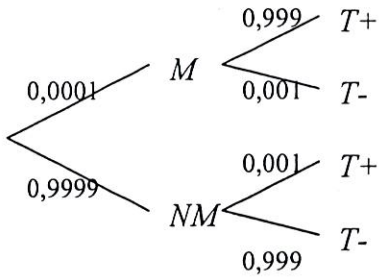
On cherche $P_A(ND)$.

On a $P_A(ND) = \dfrac{P(A \cap ND)}{P(A)} = \dfrac{0{,}999 \times 0{,}005}{0{,}999 \times 0{,}005 + 0{,}001 \times 0{,}99}$.

Soit $P_A(ND) \approx 0{,}83$. Ainsi, en cas d'alerte, on a 83 % de chances qu'il s'agisse d'une fausse alerte.

2 - VACHE FOLLE

1) Les données sont résumées sur l'arbre suivant :



On cherche $P_{T+}(M)$.

On a $P_{T+}(M) = \dfrac{P(T+ \cap M)}{P(T+)} = \dfrac{0{,}0001 \times 0{,}999}{0{,}9999 \times 0{,}001 + 0{,}0001 \times 0{,}999}$.

Soit $P_{T+}(M) \approx 0{,}0908$. L'ancien ministre a donc raison.

2) On peut estimer le nombre de vaches testées positives à :
$10000 \times P(T+) = 10{,}998$. On ne peut guère parler d'hécatombe pour 11 vaches, le principe de précaution est économiquement envisageable.

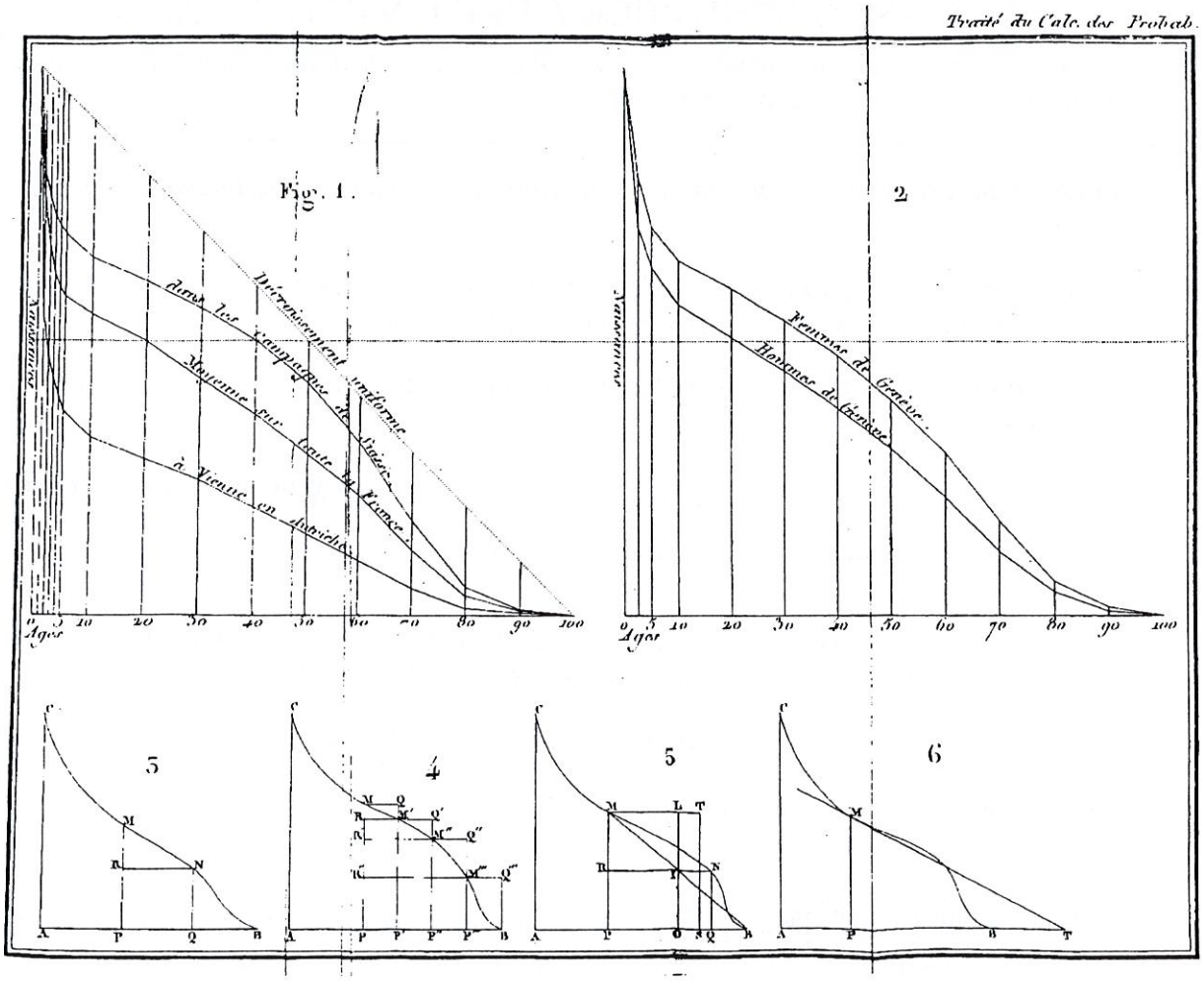
3) On cherche k tel que $P_{T+}(M) = k \times P(M)$ d'où $k \approx \dfrac{0{,}0908}{0{,}0001}$ soit 908 fois plus de "chances" d'être malade.

4) a) On a cette fois : $P_{T+}(M) = \dfrac{P(T+ \cap M)}{P(T+)} = \dfrac{p \times 0{,}999}{(1-p) \times 0{,}001 + p \times 0{,}999}$, c'est à dire

$P_{T+}(M) = \dfrac{0{,}999\, p}{0{,}001 + 0{,}998\, p}$.

b) La valeur prédictive du test dépasse 50 % lorsque le taux d'infection p dépasse 0,1 %.

Séance 5 : TEST D'ADEQUATION - LOIS CONTINUES



I – ADEQUATION A UN MODELE ET TEST STATISTIQUE

Il y a deux débouchés "naturels" à l'étude statistique des fluctuations d'échantillonnage.

• Le premier est celui de l'***estimation***, envisagé dans le thème "sondages" du programme de seconde. Dans ce cas, on n'a aucun a priori sur la fréquence p à estimer sur la population. On construira alors une "fourchette de sondage" (ou intervalle de confiance), généralement *centrée sur la fréquence f calculée sur l'échantillon.*

• La seconde situation où l'étude de la variabilité des fréquences d'échantillon trouve une application est celle des ***tests statistiques***.

Dans bien des situations en effet, on a une idée a priori de la valeur p_0 que devrait avoir la fréquence p sur la population. Si l'on se demande si un dé est truqué, on cherche à savoir si la fréquence p de sortie du 6 est égale à $p_0 = 1/6$. Dans le cas d'un contrôle de qualité, il existe sans doute une norme, ou un seuil de rentabilité, qui fait qu'on s'interroge si la proportion p de pièces défectueuses dans la production est inférieure à $p_0 = 0{,}10$ (par exemple). Dans ce cas, on construit, avant l'expérience, une "zone d'acceptation" fixée autour de p_0.
De façon générale il s'agit du problème de l'adéquation d'une distribution théorique (une loi de probabilité modélisant une situation) à une distribution empirique.

1 – EXEMPLES D'ADEQUATION A UNE LOI EQUIREPARTIE

Les programmes de terminale S et terminale ES comportent une initiation à la "problématique" des tests d'hypothèse.

> *"Etude d'un exemple traitant de l'adéquation de données expérimentales à une loi équirépartie."*
>
> *"L'élève devra être capable de poser le problème de l'adéquation à une loi équirépartie et de se reporter à des résultats de simulation qu'on lui fournit.*
>
> ***Le vocabulaire des tests** (test d'hypothèse, hypothèse nulle, risque de 1ère espèce) **est hors programme."***
>
> *Programme de TS et TES 2002.*

Le document d'accompagnement précise cependant :

> *"L'enjeu est de comprendre sur quoi porte le risque (refuser à tort le modèle) et que plus le risque est petit, plus on aura tendance à accepter le modèle de l'équiprobabilité."*
>
> *Document d'accompagnement du programme de TS et TES 2002.*

On verra plus loin le sens de ce rectificatif.

Le **TP ELEVE EN TERMINALE** suivant peut constituer une introduction expérimentale au thème de l'adéquation, que l'on lie à l'observation des fluctuations d'échantillonnage.

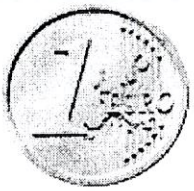
Travaux pratiques STATISTIQUES

DETECTEUR STATISTIQUE DE TRICHEURS

L'idée générale, pour détecter une pièce ou un dé truqué, est d'étudier d'abord les fluctuations d'échantillonnage d'une pièce ou d'un dé équilibré, de façon à connaître le comportement statistique "normal" lorsqu'il n'y a pas de truquage.

A – LA PIECE EST-ELLE TRUQUEE ?

ETUDE D'UNE PIECE NON TRUQUEE



Vous allez simuler 500 lancers d'une pièce parfaitement équilibrée, en considérant que *1* vaut *pile* et *0* vaut *face*.
En **A1**, entrer la ***formule*** =ENT(ALEA()+0,5) puis ***recopier vers le bas*** jusqu'en **A500** (cliquer en A1, approcher le pointeur de la souris du coin inférieur droit, lorsque celui-ci prend la forme d'une croix noire, enfoncer le bouton gauche et glisser vers le bas jusqu'en A500 où on relâche le bouton).
En **A502** entrer la ***formule*** =SOMME(A1:A500)/500
Vous allez maintenant effectuer 100 simulations de 500 lancers de la pièce.
Sélectionner la colonne A (pointeur sur le A de la tête de colonne), puis ***recopier vers la droite*** (glisser avec le pointeur en croix noire) jusqu'en **CV502**.

On considère l'intervalle $I = [0{,}5 - \frac{1}{\sqrt{500}}, 0{,}5 + \frac{1}{\sqrt{500}}]$ soit environ [0,455 ; 0,545].

En **A504** compter combien de simulations de 500 lancers donnent une fréquence de pile comprise dans l'intervalle *I* en entrant la ***formule*** :
=NB.SI(A502:CV502;">=0,455") – NB.SI(A502:CV502;"<=0,545")
En **A505** taper "% dans I".

✍ ***Compléter la feuille réponse.***

DETECTER UNE PIECE TRUQUEE PAR LA STATISTIQUE

On a observé que sur 500 lancers d'une pièce non truquée, on a plus de 95% de chances d'observer une fréquence de pile comprise entre 0,454 et 0,545.

On décide d'accepter comme non truquées les pièces qui, après 500 lancers, donnent une fréquence d'apparition de pile entre 0,454 et 0,545.

Cliquer sur l'onglet ***Feuil2***.
En **A1** entrer la ***formule*** =ENT(ALEA()+0,5)
En **A2** entrer la ***formule*** =ENT(ALEA()+0,5)
En **A3** entrer la ***formule*** =0,5+(–1)^A1*0,06*A2
Faire plusieurs fois **F9** et observer le contenu de la cellule A3.
En **B1** entrer la ***formule*** =ENT(ALEA()+A$3) puis ***recopier vers le bas*** jusqu'en **B500**.
En **B502** entrer la ***formule*** =SOMME(B1:B500)/500
Remonter en **A5** pour entrer la ***formule*** =ET(B502>=0,454;B502<=0,545)

✍ ***Compléter la feuille réponse.***

VOTRE DETECTEUR DE TRICHEURS EST-IL FIABLE ?

En **A6** entrer une ***formule*** indiquant si le test fournit une décision correcte :
=OU(ET(A3=0,5;A5=VRAI);ET(A3< >0,5;A5=FAUX))
Pour effectuer 100 tests, ***sélectionner*** les cellules de **A1** à **B102** puis ***recopier vers la droite*** jusqu'en **GR502**.
Remonter en **A6** pour entrer une ***formule*** comptant le nombre de tests fournissant une décision fausse =NB.SI(A6:GR6;FAUX)

Compléter la feuille réponse.

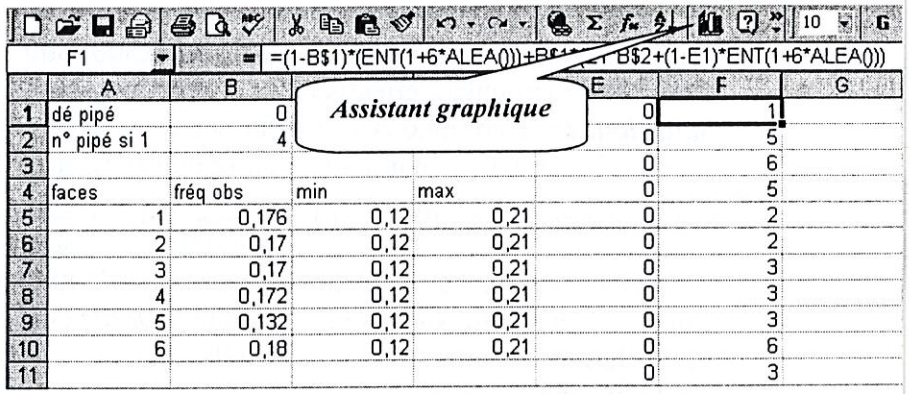
B – DETECTER UN DE PIPE

COMPARAISON DE CHAQUE FREQUENCE A L'INTERVALLE D'ECHANTILLONNAGE

On estime que pour un dé parfaitement équilibré, la fréquence de sortie d'une face donnée doit, sur 500 lancers, fluctuer dans plus de 95% des cas, dans l'intervalle :

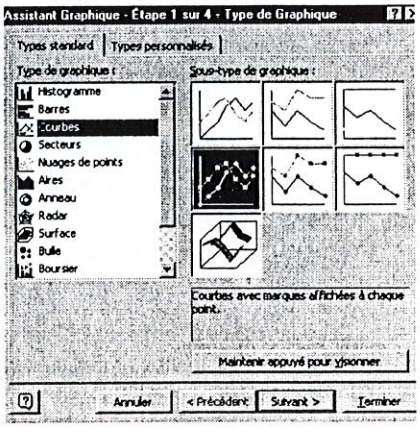
$[\frac{1}{6} - \frac{1}{\sqrt{500}}, \frac{1}{6} + \frac{1}{\sqrt{500}}]$, soit environ [0,12 ; 0,21].

Cliquer sur l'onglet ***Feuil3***.
En **A1** écrire : dé pipé.
En **B1** entrer la ***formule*** =ENT(ALEA()+0,5
Suivant que la cellule B1 vaut 0 ou 1, le dé ne sera pas, ou sera, pipé.
En **C1** entrer la ***formule*** =SI(B1=1;"PIPE";"NORMAL")
En **A2** écrire :
n° pipé si 1.



	A	B	C	D	E	F	G
1	dé pipé	0			0	1	
2	n° pipé si 1	4			0	5	
3					0	6	
4	faces	fréq obs	min	max	0	5	
5	1	0,176	0,12	0,21	0	2	
6	2	0,17	0,12	0,21	0	2	
7	3	0,17	0,12	0,21	0	3	
8	4	0,172	0,12	0,21	0	3	
9	5	0,132	0,12	0,21	0	3	
10	6	0,18	0,12	0,21	0	6	
11					0	3	

En **B2** entrer une ***formule*** choisissant aléatoirement la face pipée lorsque B1 vaut 1 : =ENT(1+6*ALEA())
En **E1** entrer la ***formule*** : =ENT(ALEA()+0,06)
En **F1** entrer une ***formule*** permettant de simuler le lancer d'un dé éventuellement pipé :
=(1–B$1)*(ENT(1+6*ALEA()))+B$1*(E1*B$2+(1–E1)*ENT(1+6*ALEA()))
Sélectionner les cellules E1 et F1 puis ***recopier vers le bas*** jusqu'en **F500**.
En **A4** écrire : faces. De **A5** à **A10** entrer les chiffres de 1 à 6.
En **B4** écrire : fréq obs.
Sélectionner les cellules de **B5** à **B10** puis taper, sans valider, la ***formule*** :
=FREQUENCE(F1:F500;A5:A10) et valider en ***appuyant simultanément*** sur les touches CTRL+SHIFT+ENTREE.
En **C4** écrire : min. ***Recopier*** de **C5** à **C10** la valeur 0,12.
En **D4** écrire : max. ***Recopier*** de **D5** à **D10** la valeur 0,21.
Sélectionner les cellules de **B5** à **D10** puis cliquer sur l'icône d'***Assistant graphique***.
A l'étape ¼ choisir ***Courbes*** puis cliquer sur ***Terminer***.
En cliquant dessus avec le bouton droit de la souris, effacer la légende.



Compléter la feuille réponse.

ETUDE DE L'ECART ENTRE LA DISTRIBUTION OBSERVEE ET LA DISTRIBUTION THEORIQUE

a) Etude d'un dé équilibré

On lance 500 fois un dé et on note les fréquences f_i de sortie de chaque face. Il s'agit de comparer la distribution observée de ces fréquences à la loi équirépartie théorique correspondant à un dé non truqué.

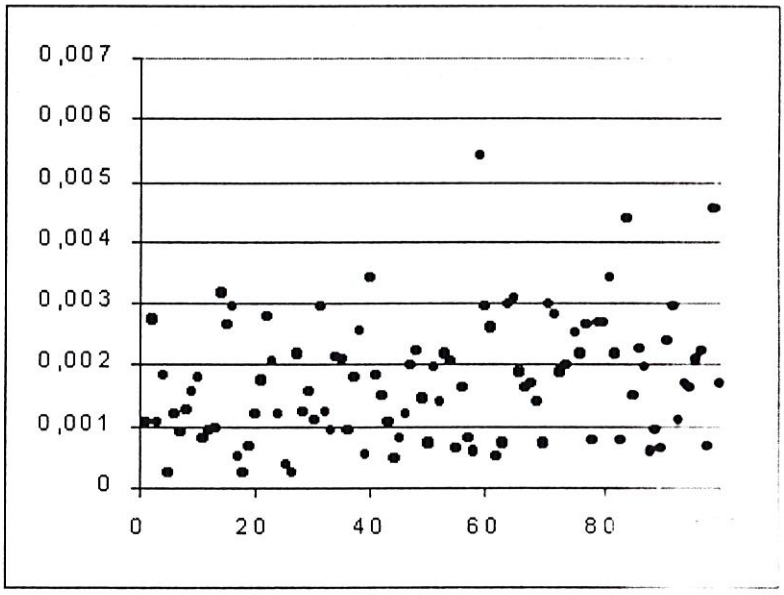
Face du dé	Fréquence observée sur 500 lancers	Fréquence théorique
1	f_1	1/6
2	f_2	1/6
3	f_3	1/6
4	f_4	1/6
5	f_5	1/6
6	f_6	1/6

Si le dé est parfaitement équilibré l'écart entre les deux colonnes des fréquences devrait être assez faible, aux fluctuations d'échantillonnage près, correspondant à 500 lancers.
Pour mesurer cet **écart**, on va considérer le carré de la "distance" habituelle en géométrie :

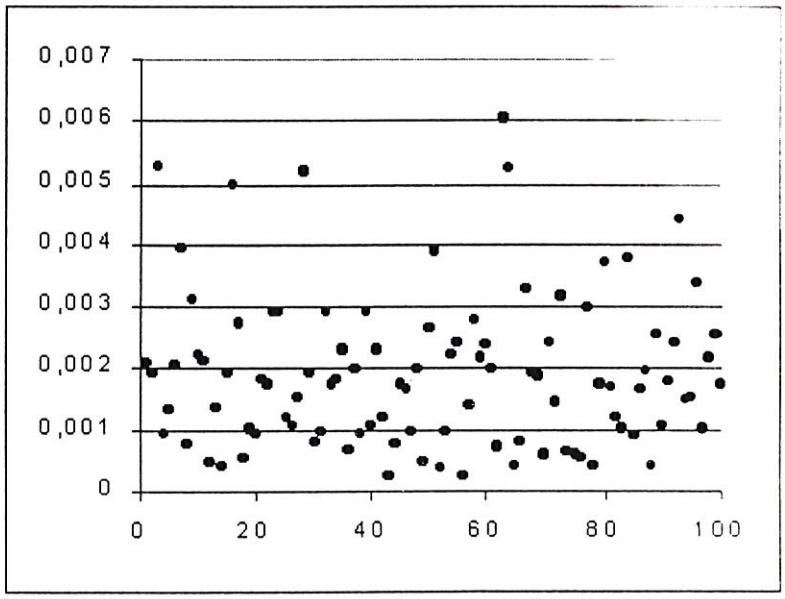
$$e = (f_1 - \frac{1}{6})^2 + (f_2 - \frac{1}{6})^2 + (f_3 - \frac{1}{6})^2 + (f_4 - \frac{1}{6})^2 + (f_5 - \frac{1}{6})^2 + (f_6 - \frac{1}{6})^2.$$

Pour étudier comment fluctue cet écart e lorsque le dé est normal, on a effectué deux séries de 100 simulations de 500 lancers d'un dé équilibré.

Les graphiques ci-contre indiquent (en ordonnée) les valeurs de e obtenues pour chacune des 100 simulations de 500 lancers.



✍ ***Compléter la feuille réponse.***



b) Détection d'un dé truqué

Au vu des observations précédentes, on prendra la règle suivante, pour décider qu'un dé est, ou non, pipé, selon l'écart e observé :

Si $e \leq 0{,}004$,	on déclare le dé équilibré.
Si $e > 0{,}004$,	on déclare le dé pipé.

Vous allez expérimenter cette nouvelle procédure.
Sur la feuille de calcul ***Feuil3***, en ***B12*** entrer la ***formule***
(B5 – 1/6)^2
Recopier vers le bas le contenu de la cellule ***B12***, jusqu'en ***B17***.
En ***A18*** écrire : e =
En **B18** entrer la ***formule***
=SOMME(B12:B17)
En ***C18*** écrire : considéré
En ***D18*** entrer la ***formule***
=SI(B18>0,004;"PIPE";"NORMAL")

D18 = =SI(B18>0,004;"PIPE";"NORMAL")

	A	B	C	D
1	dé pipé	**1**	PIPE	
2	n° pipé si 1	**6**		
3				
4	faces	fréq obs	min	max
5	1	0,142	0,12	0,21
6	2	0,188	0,12	0,21
7	3	0,12	0,12	0,21
8	4	0,162	0,12	0,21
9	5	0,182	0,12	0,21
10	6	0,206	0,12	0,21
11				
12		0,00060844		
13		0,00045511		
14		0,00217778		
15		2,1778E-05		
16		0,00023511		
17		0,00154711		
18	e =	**0,00504533**	considéré	PIPE
19				

✍ ***Compléter la feuille réponse.***

FEUILLE REPONSE

NOMS : .....

A – LA PIECE EST-ELLE TRUQUEE ?

ETUDE D'UNE PIECE NON TRUQUEE

Pourquoi la formule =ENT(ALEA()+0,5) permet-elle de simuler le lancer d'une pièce non truquée ? .....

.....

Que représente le contenu de la cellule A502 ? .....

.....

En appuyant sur la touche F9 , faites de nouvelles simulations de 500 lancers et compléter le tableau :

Simulation n°	1	2	3	4	5	6	7	8	9	10
Pourcentage des fréquences sur 500 lancers comprises entre 0,455 et 0,545										

DETECTER UNE PIECE TRUQUEE PAR LA STATISTIQUE

Montrer que la formule de la cellule A3 affiche les résultats 0,5 ; 0,56 et 0,44 avec les probabilités respectives 0,5 ; 0,25 et 0,25.

.....

.....

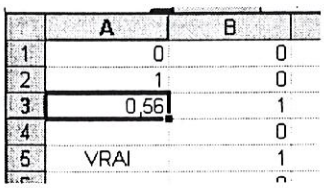
.....

.....

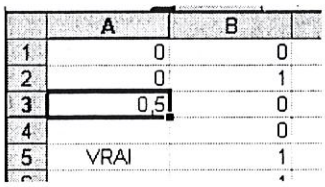
.....

Il y a quatre situations possibles :

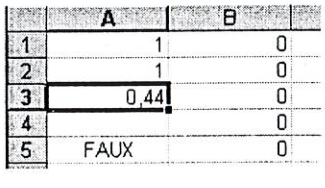
Décision / Réalité	Pièce non truquée	Pièce truquée
Pièce non truquée		ERREUR
Pièce truquée	ERREUR	



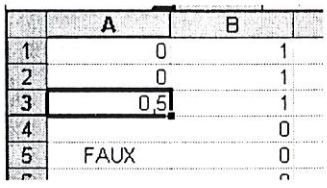
	A	B
1	0	0
2	1	0
3	0,56	1
4		0
5	VRAI	1



	A	B
1	0	0
2	0	1
3	0,5	0
4		0
5	VRAI	1



	A	B
1	1	0
2	1	0
3	0,44	0
4		0
5	FAUX	0



	A	B
1	0	1
2	0	1
3	0,5	1
4		0
5	FAUX	0

Relier par une flèche chacune des observations ci-dessus à la case correspondante.

VOTRE DETECTEUR DE TRICHEURS EST-IL FIABLE ?

Indiquer, pour 5 simulations, combien, sur 100 tests, ont conduit à prendre une mauvaise décision :

Simulation n°					
Nombre de tests (sur 100) conduisant à une décision fausse					

Commentez vos résultats : .....
.....
.....

B – DETECTER UN DE PIPE

COMPARAISON DE CHAQUE FREQUENCE A L'INTERVALLE D'ECHANTILLONNAGE

Que signifie le contenu de la cellule C 1 ? .....
.....

Selon quelle procédure peut-on détecter, au vu des fréquences observées des sorties de chaque face sur 500 lancers, qu'un dé est pipé ?
.....
.....
.....
.....
.....

Quel genre d'erreurs peut amener cette procédure de décision ?
.....
.....

Effectuer plusieurs fois F9.
Les erreurs de décision sont-elles fréquentes ?
.....
.....

Quel est le type d'erreur le plus fréquent ?
.....
.....
.....

ETUDE DE L'ECART ENTRE LA DISTRIBUTION OBSERVEE ET LA DISTRIBUTION THEORIQUE

a) Etude d'un dé équilibré

Pour chacun des deux graphiques, indiquer le pourcentage des simulations de 500 lancers pour lesquels l'écart e est strictement supérieur à 0,004 (dé équilibré) :

Graphique	n°1	n° 2
Pourcentage des cas où, sur 500 lancers d'un dé équilibré, on a $e > 0{,}004$		

On étudie un dé dont on ignore s'il est, ou non, truqué. On observe pour 500 lancers de ce dé un écart $e > 0{,}004$. Peut-on suspecter ce dé d'être truqué ? Pourquoi ?

.....
.....
.....

b) Détection d'un dé truqué

Quel genre d'erreurs peut amener cette procédure de décision ?

.....
.....

Effectuer plusieurs fois F9.
Les erreurs de décision sont-elles fréquentes ?

.....
.....

Quel est le type d'erreur le plus fréquent ?

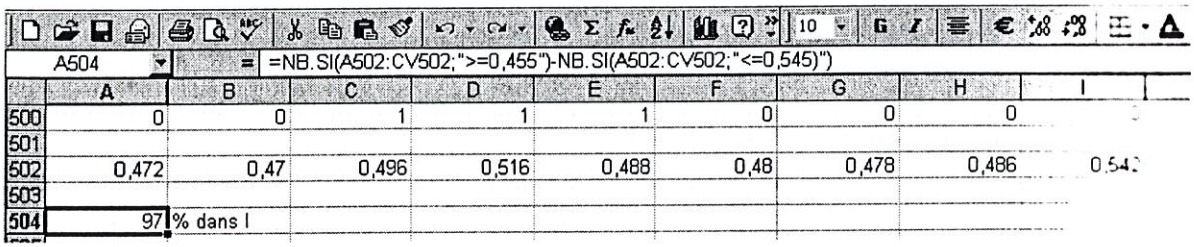
.....
.....
.....

Corrigé du TP "DETECTEUR STATISTIQUE DE TRICHEURS"

A – LA PIECE EST-ELLE TRUQUEE ?

ETUDE D'UNE PIECE NON TRUQUEE

L'instruction =ENT(ALEA()+0,5) permet de simuler le lancer d'une pièce non truquée puisque ALEA()+0,5 fournit un nombre aléatoire entre 0,5 et 1,5 de façon équidistribuée dont la partie entière a une chance sur 2 d'être 0 ou 1.
Le contenu de la cellule A502 représente la fréquence des "pile" sur 500 lancers simulés



A504 = =NB.SI(A502:CV502;">=0,455")-NB.SI(A502:CV502;"<=0,545")

	A	B	C	D	E	F	G	H	I
500	0	0	1	1	1	0	0	0	
501									
502	0,472	0,47	0,496	0,516	0,488	0,48	0,478	0,486	0,542
503									
504	97%	dans l							

On observe que plus de 95% des simulations fournissent une fréquence de "pile" sur 500 lancers comprise dans l'intervalle [0,455 ; 0,545].

DETECTER UNE PIECE TRUQUEE PAR LA STATISTIQUE

Montrons que l'instruction de la cellule A3 fournit les résultats 0,5 ; 0,56 et 0,44 avec les probabilités respectives 0,5 ; 0,25 et 0,25 :
La cellule A2 contient la valeur 0 dans 50% des cas, A3 vaut alors 0,5.
Dans les cas où A2 vaut 1, A1 vaut 0 dans 50% des cas et A3 vaut alors 0,5 + 0,06 = 0.56, et lorsque A1 vaut 1, A3 vaut 0,5 – 0,06 = 0,44.

VOTRE DETECTEUR DE TRICHEURS EST-IL FIABLE ?

A8 = =NB.SI(A6:GR6;FAUX)

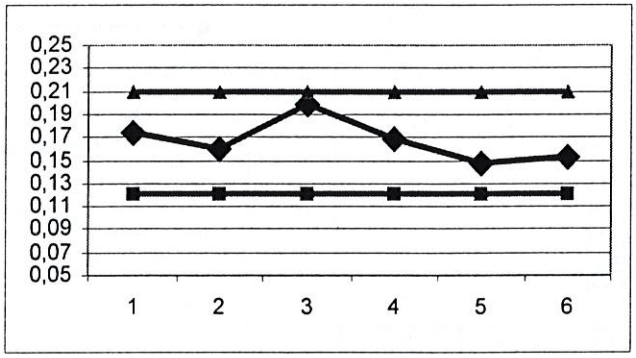
	A	B	C	D	E	F	G	H	I
1	0	1	1	1	0	0	1	0	
2	1	1	1	1	0	1	1	0	
3	0,56	1	0,44	0	0,5	1	0,44	0	
4		1		1		0		1	
5	FAUX	1	VRAI	0	VRAI	1	FAUX	1	VRAI
6	VRAI	0	FAUX	1	VRAI	0	VRAI	0	
7		1		0		0		1	
8	17%	1		1		0		1	
9		0		1		0		1	

Indiquons, pour 5 simulations, combien, sur 100 tests, ont conduit à prendre une mauvaise décision : 17% ; 10% ; 12% ; 18% ; 6%.
Le pourcentage d'erreurs observées de l'ordre de 15% reste raisonnable.

B – DETECTER UN DE PIPE

COMPARAISON DE CHAQUE FREQUENCE A L'INTERVALLE D'ECHANTILLONNAGE

La procédure de test consiste à déclarer pipé un dé pour lequel on observe, sur 500 lancers, une fréquence de sortie, pour l'une des faces, en dehors de l'intervalle [0,12 ; 0,21].
On observe très rarement l'erreur consistant à estimer à tort qu'un dé est pipé. En revanche l'erreur consistant à estimer normal un dé pipé est plus fréquente.



On peut imaginer déplacer les barres horizontales, mais on diminue alors un risque d'erreur pour augmenter l'autre...

ETUDE D'UNE "DISTANCE" ENTRE LES DISTRIBUTIONS OBSERVEE ET THEORIQUE

a) Etude d'un dé équilibré
Sur 100 simulations de 500 lancers d'un dé équilibré, on observe 3% des cas où $e > 0{,}004$ sur le premier graphique, et 6% sur le second.
Si, pour un dé inconnu, on observe un écart e tel que $e > 0{,}004$, il s'agit d'un cas rare pour un dé équilibré, il est donc raisonnable de suspecter ce dé d'être pipé.

b) Détection d'un dé truqué
Remarques analogues à celles du "test" précédent.

ENTRE NOUS...

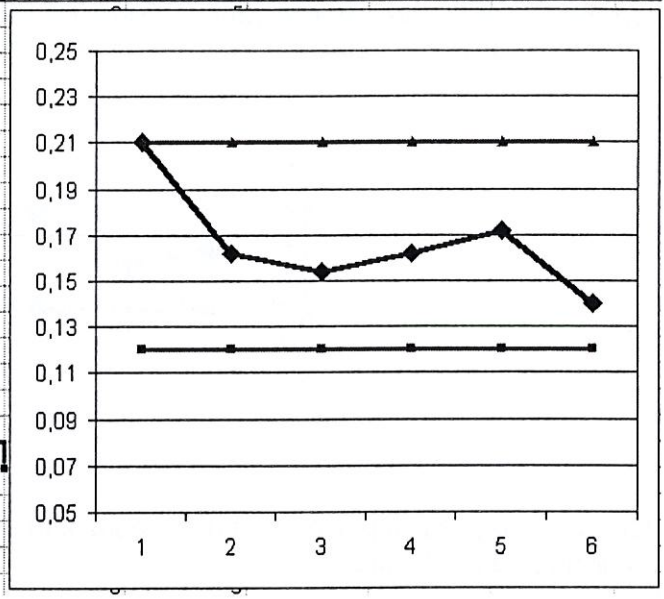
Le second test pratiqué est un embryon de **test du khi deux**. Celui-ci présente plusieurs avantages sur le premier test (comparaison à l'intervalle de fluctuation).
• Le premier est que la loi de l'écart e, sous l'hypothèse d'équidistribution, est connue. Ce qui permet d'évaluer facilement l'erreur consistant à déclarer pipé un dé qui, en fait, est équilibré (erreur dite de première espèce). En effet la distribution du khi 2 à 5 degrés de liberté montre que l'écart e est, dans le cas d'un dé équilibré, supérieur à 0,00369 dans seulement 5% des cas (pour pouvoir utiliser cette loi il faut que le nombre des observations de sortie de chaque face soit suffisamment important, ce qui est le cas sur 500 lancers). On montre que la variable aléatoire $T=\sum_{i=1}^{6}\frac{(nF_i-\frac{n}{6})^2}{\frac{n}{6}}=6n\sum_{i=1}^{6}(F_i-\frac{1}{6})^2$ suit approximativement, pour n assez grand, une loi indépendante de n dite loi du khi-deux à 5 degrés de liberté. La table de cette loi donne alors $P(T \leq 11{,}07) = 0{,}95$ d'où la valeur $0{,}00369=\frac{11{,}07}{6\times 500}$ que l'écart e ne dépasse que dans 5% des cas, lorsque le dé est équilibré.
• Le second avantage (assez décisif) est que l'on montre que, pour n, taille de l'échantillon, assez grand, la loi de l'écart entre effectifs observés (ici les nf_i) et théoriques (ici $n/6$) ne

dépend pas de *n* (ici, loi du khi 2 à 5 degrés de libertés), alors que la première procédure est étroitement liée à la taille de l'échantillon utilisé.

Voir, plus loin, davantage de détails sur le test du khi-deux.

D18 = =SI(B18>0,004;"PIPE";"NORMAL")

	A	B	C	D
1	dé pipé	1	PIPE	
2	n° pipé si 1	1		
3				
4	faces	fréq obs	min	max
5	1	0,21	0,12	0,21
6	2	0,162	0,12	0,21
7	3	0,154	0,12	0,21
8	4	0,162	0,12	0,21
9	5	0,172	0,12	0,21
10	6	0,14	0,12	0,21
11				
12		0,00187778		
13		2,1778E-05		
14		0,00016044		
15		2,1778E-05		
16		2,8444E-05		
17		0,00071111		
18	e =	**0,00282133**	considéré	NORMAL
19				
20				
21				
22				
23				



Sur cette image, on observe, pour la seconde procédure (on a $e \leq 0{,}004$), une erreur dite "de seconde espèce" : le dé est déclaré "normal" alors qu'en fait il est pipé (sur la face 1 dont la fréquence de sortie est un peu suspecte).

Documents Excel téléchargeables

Deux documents de travail du groupe Inter-IREM des Lycées Technologiques, sur le thème de l'adéquation à une loi équirépartie (format Excel), sont téléchargeables sur le site de l'IREM Paris-Nord à l'adresse :

http://www-irem.univ-paris13.fr/dossierLycee/Pour%20illustrer.htm

Auteur : **Christian Kern** ; Lycée Alain (Alençon) ; IREM de Caen ; cn.kern@wanadoo.fr

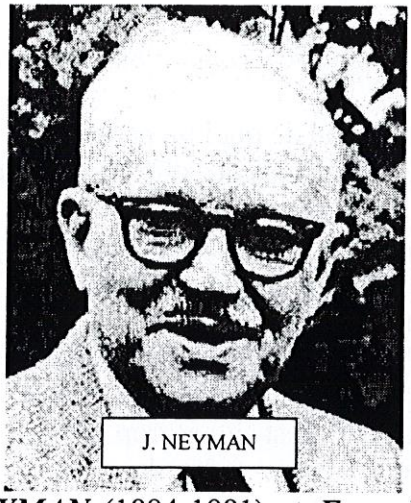
DESCRIPTION SOMMAIRE (contacter l'auteur, si nécessaire, pour plus de détails) :

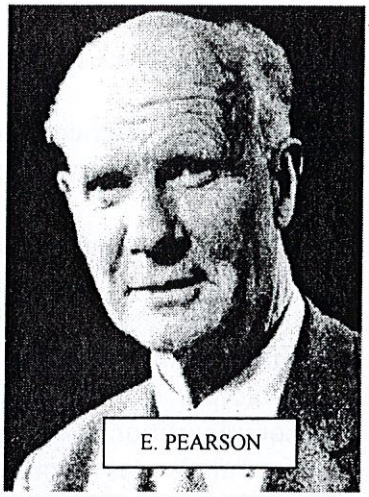
SUJET	FICHIER
ADÉQUATION À UNE LOI ÉQUIRÉPARTIE	
Adéquation à une loi	Adéquation d'un dé
Simulation de séries de lancers d'un dé équilibré (taille et nombre de séries définis par l'utilisateur) Calcul des $1000d^2 = 1000[(f_1-1/6)^2 + \ldots\ldots + (f_6-1/6)^2]$ où f_i est la fréquence observée de la face i. Résumés numériques des observations et représentations graphiques (histogramme et boîte à moustaches) Courbe de densité et boîte à moustaches théoriques optionnelles.	
Equilibre d'un dé	Équilibre d'un dé
Activité où l'utilisateur doit décider si un dé, choisi par l'ordinateur, est pipé ou équilibré en comparant expériences et théorie. L'utilisateur définit la taille et le nombre de séries de lancers ainsi que le seuil auquel il choisit de tester. Affichage du pourcentage d'inégalités « $1000d^2$ expérimental $\leq 1000d^2$ théoriques » vraies. Affichage optionnel de la nature du dé.	

2 – LES ENJEUX D'UN TEST STATISTIQUE : UN EXEMPLE DANS LE CADRE BINOMIAL

Repères historiques

Un test statistique s'inscrit dans une procédure décisionnelle particulière, et qui ne va pas de soi.


J. NEYMAN


E. PEARSON

C'est ***Jerzy NEYMAN*** (1894-1981) et ***Egon PEARSON*** (1895-1980, fils de Karl) qui proposent vers **1930** une démarche de décision universellement admise.
Ils ont remarqué que, dans la plupart des situations, **les deux hypothèses "alternatives"**, H_0 et H_1, **ne sont pas symétriques**. Il y en une dont le rejet, lorsqu'elle est vraie, peut avoir de graves conséquences, ou bien, une des deux hypothèses est privilégiée comme issue d'une théorie en vigueur, bien établie. La théorie des tests statistiques se présente donc comme un problème de choix entre deux décisions possibles : accepter ou refuser l'hypothèse privilégiée H_0.

De l'importance de cette procédure – étude des manuels scolaires

Certes, le programme officiel est clair :

> *"**Le vocabulaire des tests** (test d'hypothèse, hypothèse nulle, risque de 1ère espèce) **est hors programme**."*

Cependant, la notion de "risque" est au centre de cette problématique, et l'ignorer peut conduire à des contre-sens.

Les manuels scolaires de terminale S ou ES sont très inégaux sur ce thème.
Certains l'oublient complètement ou presque (collection *Math'x* TS éditions *Didier* 2002 par exemple), d'autres sont parfois maladroits dans la formulation.
Par exemple, dans la collection *Indice* (TS – *Bordas* 2002), on lit page 236, dans la partie "cours" (il n'est pas sûr qu'il faille faire un "cours" là dessus !) :
"Soit une épreuve conduisant aux issues a_1, a_2, ..., a_q.
Expérimentalement, si on répète n fois cette épreuve ($n \geq 100$) [on se demande pourquoi 100] , *on obtient les fréquences f_1, f_2, ..., f_q pour chacune des issues. Pour vérifier*

l'adéquation de ces données à la loi équirépartie sur $\{a_1, a_2, ..., a_q\}$, on calcule le nombre $d^2 = \sum_{i=1}^{q}\left(f_i - \frac{1}{q}\right)^2$.

La réalisation d'un grand nombre de simulations de cette épreuve [ou plutôt de ces n épreuves, réalisées sous l'hypothèse d'équirépartition – ce n'est pas dit] *conduit pour la variable d^2 à une série statistique de neuvième décile D_9.*

• Si $d^2 \leq D_9$, alors on dira que les données sont compatibles avec le modèle de la loi uniforme au seuil de risque 10%.

• Si $d^2 > D_9$, on dira que les données ne sont pas compatibles avec ce modèle au seuil de risque 10%."

Bravo à l'élève qui, du premier coup, comprend de quoi on parle.

Notons H_0 l'hypothèse d'équirépartition. Il faut bien sûr dire que les simulations sont faites sous l'hypothèse H_0.

Gagne-t-on en clarté en parlant de "neuvième décile" ? Disons plutôt que, sous l'hypothèse d'équirépartition, 90 % des simulations de ces n épreuves donnent une valeur d^2 inférieure à D_9 (pour reprendre les mêmes notations).

Quant aux affirmations concernant le "risque", seule la seconde est exacte. A savoir que l'on rejette H_0 avec un "risque" de 10 % (car dans ce cas, le "risque" est de rejeter à tort l'hypothèse d'équirépartition parce que d^2 est jugé trop grand, ce qui se produit, statistiquement, d'après les simulations, dans 10% des cas, sous l'hypothèse H_0 d'équirépartition).

En revanche, lorsque l'on accepte H_0 (c'est à dire $d^2 \leq D_9$), le "risque" est alors d'accepter l'équirépartition à tort, risque qui n'a aucune raison d'être de 10 % (et qui peut être difficilement évaluable) !

Conclusion : c'est plutôt risqué de parler du risque, alors qu'il y en a deux (première et seconde espèce) et qu'on ne les a pas définis.

Même genre de confusion dans le manuel de TS, collection *Fractale*, éditeur *Bordas* 2002. On lit, dans l'exercice proposé page 349 : *"L'affirmation du propriétaire peut-elle être acceptée avec un risque d'erreur de 10 % ?"*. Impossible de répondre. Une astuce consiste à rédiger la question sous la forme "au seuil de 10 %".

D'autres manuels sont plus rigoureux au niveau de l'expression, comme *Transmath* ou *Hyperbole* chez *Nathan*.

L'ouvrage de TS paru chez *Bréal* fait un lien intéressant entre fluctuations d'échantillonnage au pile ou face (plus de 95% des fréquences f observées se situent dans l'intervalle $\frac{1}{2} \pm \frac{1}{\sqrt{n}}$) et fluctuations de la distance au carré ($\sum_{i=1}^{2}(f_i - \frac{1}{2})^2 \leq \frac{2}{n}$ dans plus de 95% des cas). Dommage que dans le cadre de l'adéquation, on y parle régulièrement de "niveau de confiance", expression d'ordinaire réservée à l'estimation (il faudrait parler de seuil ou de seuil de risque).

Enfin, *Terracher* fait une très bonne présentation du test du khi-deux, page 314/315 de son manuel de TS. C'est très pédagogique et accessible aux élèves. Dommage que ce ne soit pas au programme.

Le TP qui suit a pour objectif, tout en travaillant sur la loi binomiale (au programme en TS), de faire prendre conscience des enjeux d'un test statistique. Les notions de risque sont en effet essentielles et constitutives de celle de test. On peut les faire comprendre sans pour autant entrer dans une étude systématique.
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 sont 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...).
- 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 et la taille n de l'échantillon (coût du contrôle). L'erreur de $1^{ère}$ espèce étant la plus facile à maîtriser, c'est sur elle, et l'hypothèse H_0, que sera construit le test. Cela conduit à ***privilégier H_0*** : les deux hypothèses ne jouent donc pas un rôle symétrique.

Il met en évidence les différentes ***étapes d'un test***, dont le plan est :

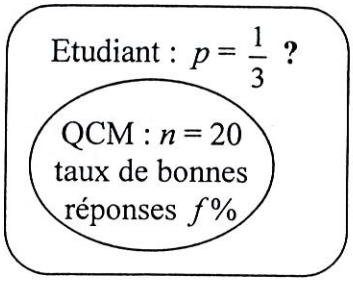
1. ***Construction du test*** :
   a - ***Choix des hypothèses*** H_0 et H_1.
   b - ***Calcul***, sous l'hypothèse H_0 , ***de la région critique au seuil*** α (ou de la zone d'acceptation).
   c - ***Enoncé de la règle de décision***.
2. ***Utilisation du test*** : prélèvement d'un échantillon et prise de décision.

TRAVAUX PRATIQUES

LOI BINOMIALE ET TEST

Dans un lycée syldave, les professeurs, exaspérés par le manque de travail d'une partie des élèves, décident d'établir un examen à la fin du premier trimestre pour renvoyer les élèves n'ayant fourni aucun travail (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 élève 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 recaler ce type d'élève, avec une probabilité d'environ 95 %.
Pour cela il faut définir la barre d'acceptation *avant* l'épreuve, de sorte que les élèves souscrivent au ***protocole*** ("règles du jeu") de l'examen.



A - CONSTRUCTION DU TEST

Choix des hypothèses

On teste l'hypothèse H_0 : " $p = \frac{1}{3}$ " (appelée "***hypothèse nulle***"), contre l'***hypothèse alternative*** H_1 : " $p > \frac{1}{3}$ ".

L'hypothèse nulle correspond à un élève répondant au hasard. L'hypothèse alternative doit "au contraire" correspondre à un élève qui a travaillé. Pourquoi ne prend-on pas " $p \neq \frac{1}{3}$ " ? .....

.....

Calcul de la zone d'acceptation de H_0

On suppose que H_0 est vraie : l'élève répond au hasard.
On désigne par X la variable aléatoire qui, à chaque élève 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 %.

.....

.....

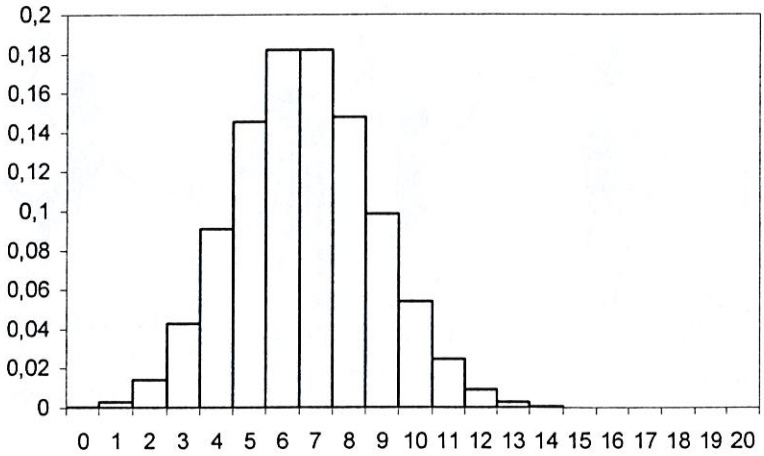
Quand le nombre de bonnes réponses est inférieur ou égal à k, on acceptera H_0 .

S'il est strictement supérieur à k, on supposera que l'élève 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 α ? .....

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



Règle de décision

Enoncer la ***règle de décision*** de l'examen.

.....

.....

B - UTILISATION DU TEST ET ERREURS

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 élève 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↵ 0 → X↵ For 1→I To **20**↵ Int(Ran# + P)+X → X↵ Next↵ X ⊿ P	:1/3+int(rand+.5)(.6-1/3) → P :0 → X :For(I,1,**20**) :int(rand+P)+X → X :End :Disp X , P	:1/3+int(rand()+.5)(.6-1/3) → p :0 → x :For i,1,**20** :int(rand()+p)+x → x :EndFor :Disp x , p

L'examen conduit-il toujours à une décision juste ?

.....

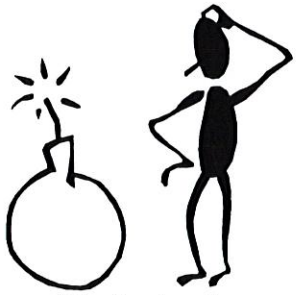
Les erreurs

Décision / Réalité	H_0 acceptée	H_1 acceptée
H_0 vraie	$1 - \alpha$	α ERREUR DE 1ère ESPECE
H_1 vraie	β ERREUR DE 2nde ESPECE	$1 - \beta$

Il y a quatre situations possibles.
Les ***erreurs de décision*** sont de deux types : "rejeter H_0 à tort" (erreur de première espèce correspondant au risque α) ou "accepter H_0 à tort".

Relier chaque dessin à la case qui lui correspond dans le tableau.

Etudiant qui n'a pas travaillé et est recalé



Etudiant qui a travaillé et est recalé

Etudiant qui a travaillé et est reçu à l'examen

Etudiant qui n'a pas travaillé et est reçu à l'examen

L'erreur de 2nde espèce

Un élève 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 !"

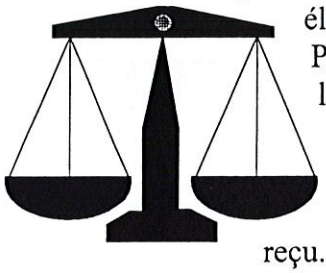
Vérifier l'affirmation de cet élève, qui craint d'être victime d'une erreur de 2nde espèce (utiliser la table ci-contre, des valeurs cumulées de la loi $\mathscr{B}(20 ; 0{,}60)$).

.....

.....

.....

Il faut avouer que ce n'est pas très moral vis à vis de cet élève.



Pour diminuer le risque de 2nde espèce, l'élève propose de baisser la barre d'admission à 8 : si le nombre x de bonnes réponses est tel que $x \leq 7$, l'élève est recalé, si $x \geq 8$, l'élève est reçu.

Quel est, dans ces conditions, le risque β de 2nde espèce, pour un élève tel que $p = 0{,}60$? .....

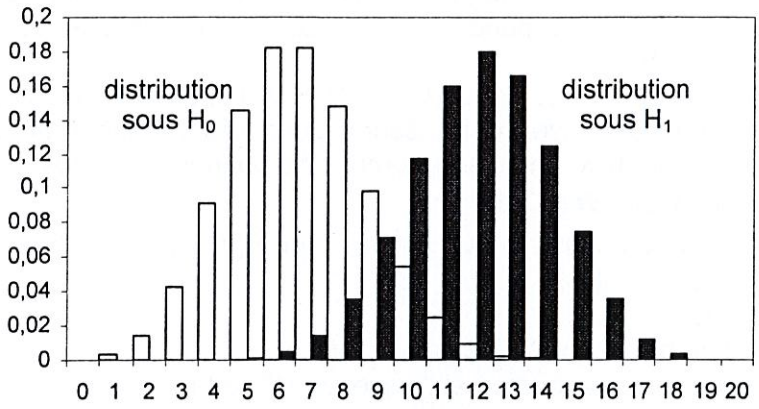
.....

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

Mais que devient le risque α d'admettre un élève n'ayant pas travaillé ? .....

.....

.....



C - TEST DE 100 QUESTIONS

Les professeurs jugeant ce risque de première espèce inacceptable, décident, pour diminuer β sans augmenter α, de proposer un QCM de 100 questions.

n = 100 et p = 1/3	
k	$P(X \leq k)$
30	0,276553852
31	0,352519827
32	0,434420644
33	0,518803305
34	0,601945044
35	0,680335826
36	0,751105282
37	0,812311298
38	0,863047864
39	0,90337693
40	0,934127842
41	0,95662851
42	0,97243255
43	0,983091089
44	0,989994915
45	0,994290629
46	0,996858719
47	0,998334005
48	0,999148486
49	0,999580659

Construction du test

• ***Choix des hypothèses :***

On teste H_0 : " $p = \frac{1}{3}$ ", contre H_1 : " $p > \frac{1}{3}$ ".

• ***Calcul de la zone d'acceptation de H_0, au seuil α de 5 % :***

On suppose que H_0 est vraie : $p = \frac{1}{3}$.

La variable aléatoire X qui, à chaque élève de ce type, associe le nombre de ses bonnes réponses au QCM, suit la loi $\mathscr{B}(100\ ,\ \frac{1}{3})$.

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 %.

.....

• ***Règle de décision :***

.....

.....

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).

.....

.....

Corrigé des travaux dirigés "INTRODUCTION AUX TESTS STATISTIQUES"

A – CONSTRUCTION DU TEST

❑ Le cas $p < 1/3$ n'est pas envisagé (que dire d'un élève cherchant à répondre faux ?). La forme de la région de rejet dépend de H_1 qui correspond uniquement à $p > 1/3$ (l'élève ne répond pas au hasard).

❑ 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 $\mathscr{B}(20, 1/3)$.

Avec la table fournie, on constate que $k = 10$, avec $P(X \leq 10) \approx 0{,}96$.

Le risque α est donc $\alpha \approx 4\%$.

❑ 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'élève est RECALE,
- si $x \geq 11$ alors H_0 est refusée et l'élève est ADMIS.

B – UTILISATION DU TEST ET ERREURS

❑ La simulation permet de "vivre" les aléas du hasard, et d'observer les deux types d'erreurs.

❑ On observe deux types d'erreurs de décision.

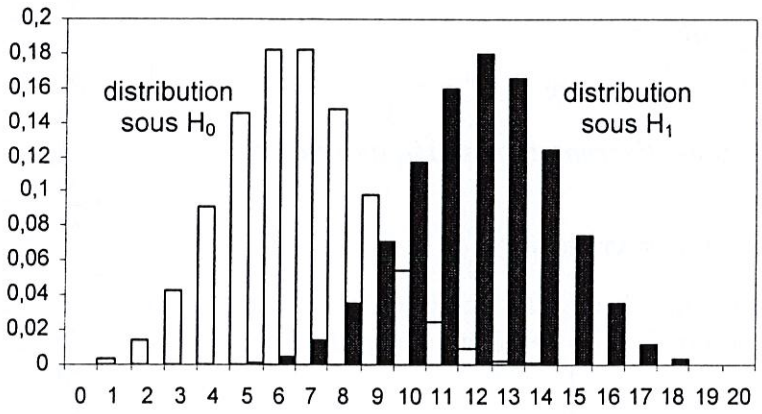
L'erreur de première espèce correspond à l'élève qui n'a pas travaillé et est reçu à l'examen.

L'erreur de seconde espèce correspond à l'élève qui a travaillé et est recalé.

❑ 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 $\mathscr{B}(20 ; 0{,}60)$.

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 1ère espèce (seuil) correspond aux rectangles clairs 11, 12, 13 ..., de la distribution sous H_0, et l'erreur de 2nde 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 $\mathscr{B}(20 ; 0{,}60)$, $\beta \approx 2\ \%$.

Mais alors, d'après la table de la loi $\mathscr{B}(20 ; 1/3)$, $\alpha \approx 100 - 66 = 34\ \%$!

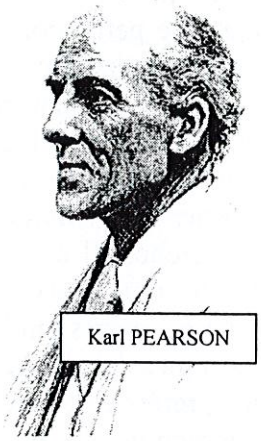
C – TEST DE 100 QUESTIONS

La règle de décision du test de 100 questions, au seuil de 5 % est :

Soit x le nombre de bonnes réponses. Si $x \leq 41$, le candidat est recalé. Si $x \geq 42$, le candidat est admis.

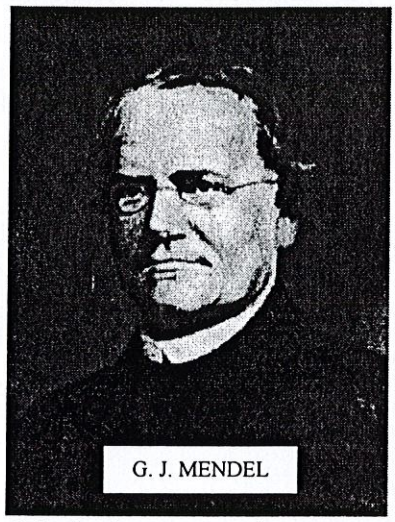
3 – ADEQUATION A UN MODELE : L'EXEMPLE DU TEST DU KHI 2

C'est à ***Karl Pearson*** (1857 – 1936) que l'on doit le critère du khi-deux, permettant de juger de la qualité d'ajustement d'une distribution théorique à une distribution observée. Pour cette étude, *Karl Pearson* eut recours à de nombreux lancers de pièces de monnaie ou de dés, effectués par lui-même, ses élèves ou ses proches. On ne disposait pas encore des techniques de simulation…


Karl PEARSON

LES PETITS POIS DE MENDEL

Paradoxalement, l'absence de variabilité peut être aussi suspecte que ses débordements et permet également de détecter statistiquement des fraudes.


G. J. MENDEL

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.

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.

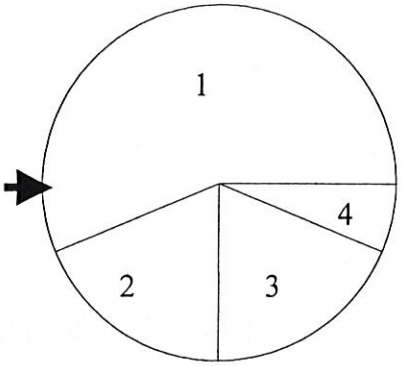
Types de petits pois	Proportions théoriques
1. Jaune rond	$p_1 = \frac{9}{16}$
2. Jaune anguleux	$p_2 = \frac{3}{16}$
3. Vert rond	$p_3 = \frac{3}{16}$
4. Vert anguleux	$p_4 = \frac{1}{16}$

La théorie de *Mendel* prévoit que le croisement de petits pois jaunes et ronds avec des petits poids 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 + int(16*rand) , où int correspond à la partie entière et 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.
Une simulation sur Excel fournit les résultats suivants :

Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?

F1 =SI(E1<10;1;SI(E1<13;2;SI(E1<16;3;4)))

	A	B	C	D	E	F	G	H
1	type	effectif théorique t	effectif observé x	(x-t)^2/t	8	1		
2	1	**312,75**	**321**	0,2176259	11	2		
3	2	**104,25**	**89**	2,23081535	16	4		
4	3	**104,25**	**108**	0,13489209	5	1		
5	4	**34,75**	**38**	0,30395683	15	3		
6	somme	556	556		15	3		
7					2	1		
8			chi-deux observé	**2,88729017**	3	1		
9			chi-deux Mendel	0,47	6	1		

La colonne E contient les instructions =ENT(1+16*ALEA()) puis la distinction des quatre types de petits pois est faite dans la colonne F grâce à la fonction SI(test;valeur si vrai;valeur si faux).

D'un point de vue probabiliste, on pourra introduire les quatre variables aléatoires X_1, ..., 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 $\mathscr{B}$ (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[1], noté χ^2_{obs} et valant $\chi^2_{obs} = \sum_{i=1}^{4} \frac{(x_i - t_i)^2}{t_i}$.

De grandes valeurs de χ^2_{obs} rendraient le modèle de *Mendel* suspect.

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

[1] Lorsqu'il n'y a pas équidistribution théorique, l'écart absolu entre effectif observé x_i et effectif théorique t_i serait trop systématiquement plus faible sur les classes "rares" et plus élevé sur les classes théoriquement "fréquentes".

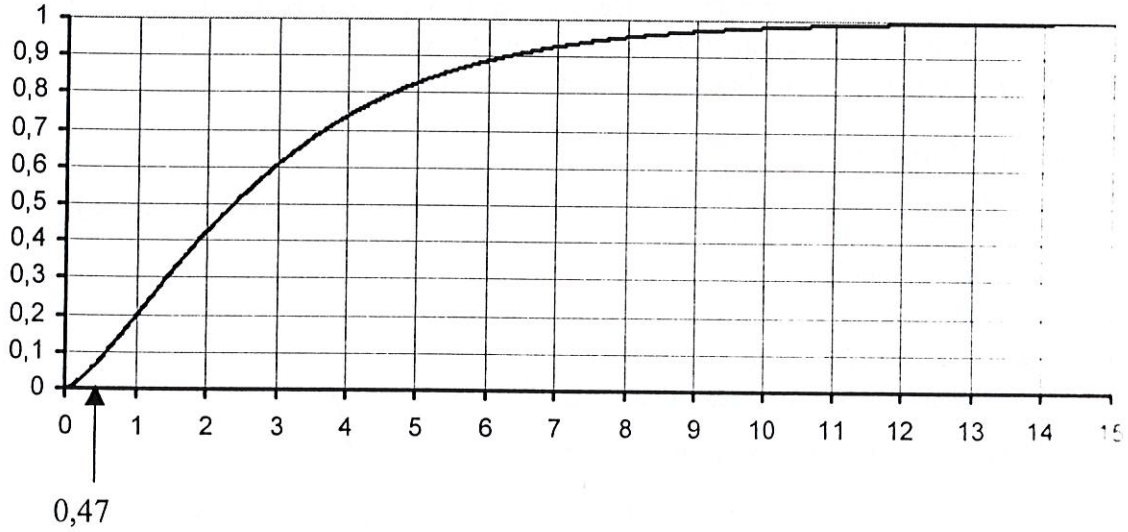
La simulation montre qu'un résultat aussi bon (sous l'hypothèse que le modèle est correct) est assez rare.

Pour étudier la variabilité de ce critère, introduisons la variable aléatoire $T = \sum_{i=1}^{4} \frac{(X_i - t_i)^2}{t_i}$

avec $\sum_{i=1}^{4} X_i = 556$ (n est le nombre total de données et $t_i = np_i$ est l'effectif théorique de la classe i).

Karl Pearson a démontré que, pour n assez grand, la variable aléatoire $T = \sum \frac{(X_i - np_i)^2}{np_i}$ suit approximativement une loi tabulée et connue sous le nom de ***loi du χ^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), qui ne dépend pas de n.
Sur Excel la fonction LOI.KHIDEUX(valeur t ; degrés de libertés) fournit la probabilité $P(T > t)$. Cette fonction (ou une table) permet alors d'obtenir : $P(T \geq 0{,}47) \approx 0{,}925$.
A titre d'illustration, la fonction de répartition $F(t) = P(T \leq t)$ pour T suivant la loi du khi-deux à 3 degrés de liberté est représentée ci-dessous.

Fonction de répartition du chi-deux à 3 degrés de liberté



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 séries d'expériences, *Fisher* a pu additionner les différents χ^2_{obs} et aboutir ainsi à $\chi^2_{obs} < 42$, avec 84 degrés de liberté.
C'est la probabilité que cet évènement se produise qui est de l'ordre de 0,00004.

B2 = =LOI.KHIDEUX(A2;84)

	A	B	C	D
1	t	P(T>=t)	F(t)=P(T<=t)	
2	42	0,99996465	3,5351E-05	
3	42,1	0,99996276	3,7237E-05	

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

RETOUR SUR LE DE PIPE

Dans le T.P. de terminale présenté au paragraphe II-1), on testait la conformité d'un dé en comparant les fréquences de sortie de chaque face à l'intervalle d'échantillonnage $[\frac{1}{6} - \frac{1}{\sqrt{500}}, \frac{1}{6} + \frac{1}{\sqrt{500}}]$ dans lequel se situe la fréquence d'une face sur 500 lancers, dans plus de 95% des cas.
S'il ne s'agissait que de tester la fréquence de sortie du 6, on pourrait évaluer le risque d'erreur de première espèce (considérer un dé comme pipé alors qu'il est normal) à moins de 5% (encore que, dans ce cas, on peut supposer que la possibilité de tricherie consiste seulement à favoriser le 6 et alors un test dit "unilatéral à droite", avec une zone de rejet seulement si la fréquence du 6 est significativement importante, est plus adapté).
Quand il s'agit de tester chacune des fréquences, c'est plus compliqué car il n'y a pas indépendance. Il y a en fait 5 degrés de liberté, la dernière fréquence étant connue dès que les cinq autres le sont. Le risque de première espèce est toutefois évaluable par simulation.
Un test adapté à cette situation est celui du khi-deux. Cependant ce test s'applique plutôt à des effectifs qu'à des fréquences. Il est en effet fondé sur un théorème limite, et on exige généralement un effectif au moins égal à 5 pour chaque classe.

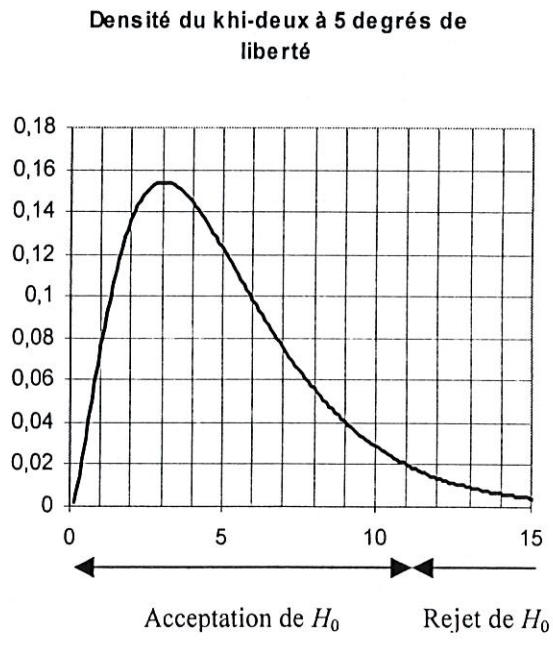
Nous allons effectuer le *test du khi-deux* sur un échantillon de taille 500, *au seuil de 5%*.
Soit X_1, X_2, ..., X_6 les variables aléatoires qui, à chaque réalisation de 500 lancers, associent l'effectif x_i de chaque face. Ces variables aléatoires suivent des lois binomiales $\mathscr{B}(500, p_i)$ où p_i représente la probabilité de sortie de la face i.

• *Choix des hypothèses* :

H_0 : pour tout $1 \leq i \leq 6$, $p_i = \frac{1}{6}$.

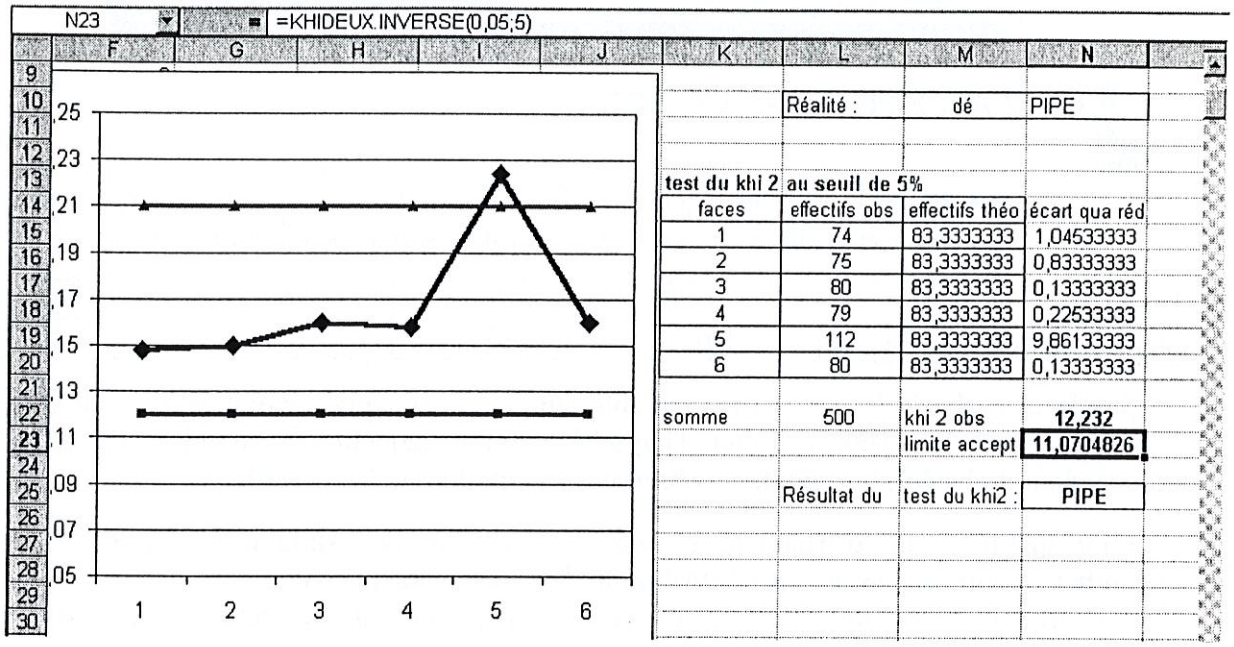
H_1 : il existe i tel que $p_i \neq \frac{1}{6}$.

• *Calcul de la zone d'acceptation de H_0* :
Si H_0 est vraie, les variables aléatoires X_i suivent la loi binomiale $\mathscr{B}$ (500, $\frac{1}{6}$) et la variable aléatoire $T = \sum_{i=1}^{6} \frac{(X_i - \frac{500}{6})^2}{\frac{500}{6}}$ suit approximativement la loi du khi-deux à 5 degrés de liberté.
On lit dans la table du khi-2 :
$P(T \leq 11{,}07) = 0{,}95$.
D'où la zone d'acceptation de H_0 : [0 ; 11,07].



• *Règle de décision* :
Soit χ^2_{obs} l'écart quadratique réduit obtenu entre les effectifs observés et les effectifs théoriques.
Si $\chi^2_{obs} \leq 11{,}07$ on accepte H_0 au seuil de 5%.
Si $\chi^2_{obs} > 11{,}07$ on rejette H_0 au risque de 5%.

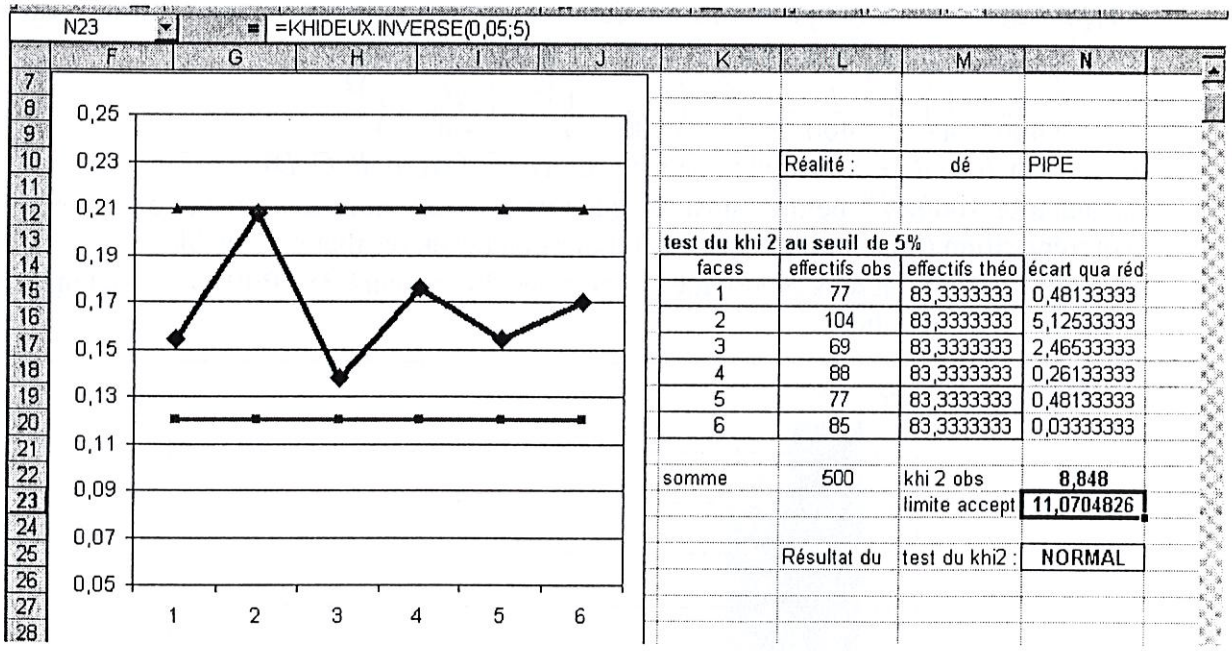
• *Utilisation du test du khi-deux au seuil de 5%* :
On effectue sur Excel 500 simulations de lancers.
Un exemple de décision correcte :



N23 =KHIDEUX.INVERSE(0,05;5)

	Réalité :	dé	PIPE
test du khi 2	**au seuil de 5%**		
faces	effectifs obs	effectifs théo	écart qua réd
1	74	83,3333333	1,04533333
2	75	83,3333333	0,83333333
3	80	83,3333333	0,13333333
4	79	83,3333333	0,22533333
5	112	83,3333333	9,86133333
6	80	83,3333333	0,13333333
somme	500	khi 2 obs	**12,232**
		limite accept	**11,0704826**
	Résultat du	test du khi2 :	PIPE

Un exemple d'erreur de seconde espèce :



N23 =KHIDEUX.INVERSE(0,05;5)

	Réalité :	dé	PIPE
test du khi 2	**au seuil de 5%**		
faces	effectifs obs	effectifs théo	écart qua réd
1	77	83,3333333	0,48133333
2	104	83,3333333	5,12533333
3	69	83,3333333	2,46533333
4	88	83,3333333	0,26133333
5	77	83,3333333	0,48133333
6	85	83,3333333	0,03333333
somme	500	khi 2 obs	**8,848**
		limite accept	**11,0704826**
	Résultat du	test du khi2 :	NORMAL

Exercices

ADEQUATION A UN MODELE EQUIREPARTI

1 LE CLASSEMENT DES LYCEES

L'année dernière, les quatre lycées d'une ville du centre de la France, notés A, B, C et D, ont présenté le même nombre d'élèves au bac S. Les résultats sont cependant différents :

Lycée	A	B	C	D
Nombre de reçus au bac S	101	82	117	100

Le proviseur du lycée B, voit ainsi son établissement classé en dernière position... Il soutient cependant que l'on peut accepter l'hypothèse selon laquelle "le nombre de reçu est indépendant du lycée", les variations observées n'étant dues qu'au hasard.

1. Pour justifier son affirmation, le proviseur produit des simulations effectuées sous l'hypothèse d'équiprobabilité dans l'attribution d'un reçu à chaque lycée.

a. Pourquoi la formule =ENT(4*ALEA( ) + 1) permet-elle de réaliser, sous Excel, une telle simulation ?

b. Le proviseur simule, à l'aide d'un ordinateur, l'attribution au hasard de 400 reçus suivant la loi équirépartie.

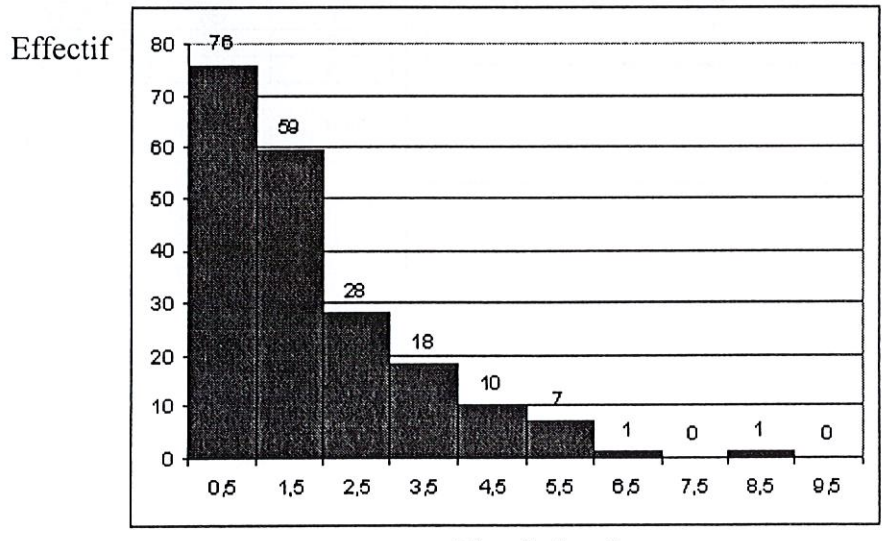
Il répète 200 fois cette opération et calcule à chaque fois la valeur de $1000\times d^2$ où d^2 correspond, pour chaque échantillon de taille 400 simulé, à :

$$d^2 = \left(\frac{a}{400} - \frac{1}{4}\right)^2 + \left(\frac{b}{400} - \frac{1}{4}\right)^2 + \left(\frac{c}{400} - \frac{1}{4}\right)^2 + \left(\frac{d}{400} - \frac{1}{4}\right)^2 ,$$

où a (resp. b, c, d) est le nombre de reçus du lycée A (resp. B, C, D).

La quantité $1000\times d^2$ permet ainsi de quantifier l'écart entre la distribution observée et l'équirépartition (on a multiplié par 1000 pour obtenir un résultat plus lisible).

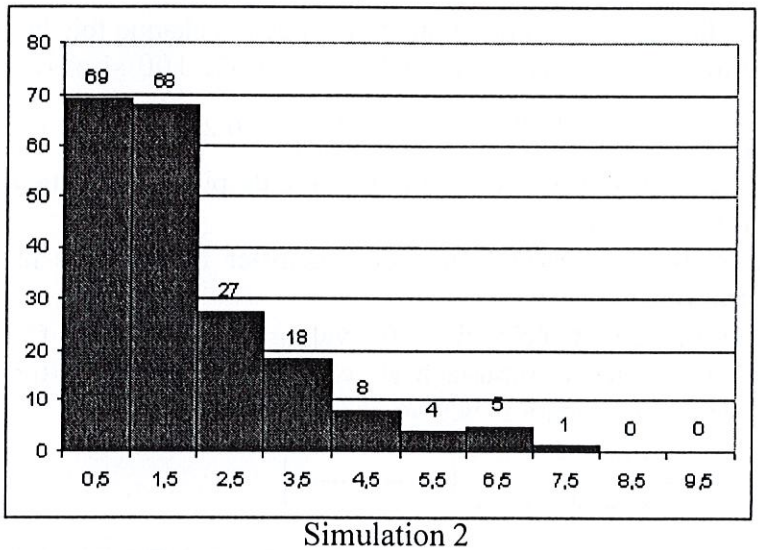
Le diagramme ci-dessous représente la série des 200 valeurs de $1000\times d^2$, obtenues sur une première simulation.



Simulation 1

Montrer que, d'après cette simulation, sous l'hypothèse que "le nombre de reçus est indépendant du lycée", on a, dans au moins 95% des cas, "$1000\times d^2 \leq 5$".

c. Une seconde simulation, où l'on a répété 200 fois l'attribution au hasard de 400 reçus selon la loi équirépartie, a fourni les valeurs de $1000\times d^2$ suivantes.



Simulation 2

Pourquoi les résultats sont-ils différents ?
Cette nouvelle simulation conforte-t-elle la conclusion de la question précédente ?

2. a. Sous l'hypothèse que "le nombre de reçus est indépendant du lycée", et en se fondant sur les simulations précédentes, donner un intervalle de la forme [0, h] dans lequel la valeur $1000\times d^2$ obtenue sur un échantillon aléatoire de 400 reçus, se situe dans 95% des cas.
b. Enoncer une règle de décision permettant d'utiliser ce test à partir d'un échantillon de 400 reçus, au seuil de risque de 5%.

3. a. L'ordinateur a calculé la valeur de $1000\times d^2$ correspondant aux résultats du bac donnés au début de l'exercice et a fourni $1000\times d^2 \approx 3{,}84$.
Poser le calcul qui a permis ce résultat.
b. Peut-on accepter, au seuil de 5%, l'hypothèse selon laquelle "le nombre de reçus est indépendant du lycée" ?

2 UNE APPLICATION INDUSTRIELLE

Dans une entreprise, quatre personnes travaillent en parallèle sur quatre postes différents et dans des conditions identiques, fabriquant en grande série des pièces de même type.
On souhaite savoir si l'on peut considérer la qualité de leur production comme analogue ou non.
Pour ce faire, on prendra au hasard, dans la fabrication d'une semaine, 100 pièces défectueuses dont on analysera l'origine.
On souhaite construire un test permettant de décider, au seuil de 10%, si au vu des résultats on peut accepter l'hypothèse selon laquelle "le nombre de pièces défectueuses est indépendant du poste de fabrication".

1. *Construction du test*

On simule, à l'aide d'un ordinateur, l'attribution au hasard de 100 pièces défectueuses, partagées sur quatre postes, suivant la loi équirépartie.

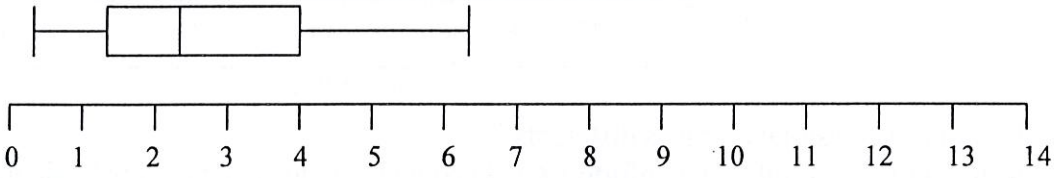
On répète 2000 fois cette opération et on calcule à chaque fois la valeur de $400\times d^2$ où d^2 correspond, pour chaque échantillon de taille 100 simulé, à :

$$d^2 = \left(\frac{a}{100} - 0{,}25\right)^2 + \left(\frac{b}{100} - 0{,}25\right)^2 + \left(\frac{c}{100} - 0{,}25\right)^2 + \left(\frac{d}{100} - 0{,}25\right)^2 ,$$

où a (respectivement b, c, d) est le nombre de pièces défectueuses provenant du poste A (respectivement B, C, D).

La quantité $400\times d^2$ permet ainsi de quantifier l'écart entre la distribution observée et l'équirépartition.

La simulation permet d'obtenir 2000 valeurs de $400\times d^2$. Ces valeurs ont permis de construire la "boîte à moustaches" ci-dessous, où les extrémités des "moustaches" correspondent aux premier et neuvième décile.



a. Lire sur le diagramme une valeur approchée du neuvième décile.

b. Sous l'hypothèse que "le nombre de pièces défectueuses est indépendant du poste de fabrication", et en se fondant sur la simulation précédente, donner un intervalle de la forme $[0, h]$ dans lequel la valeur $400\times d^2$ observée sur un échantillon aléatoire de 100 pièces défectueuses a 90% de chances de se situer (on prendra comme neuvième décile la valeur approchée obtenue à la question précédente).

2. *Règle de décision*

Enoncer une règle de décision permettant d'utiliser ce qui précède pour un test à partir d'un échantillon de 100 pièces défectueuses, au seuil de risque de 10%.

3. *Utilisation du test*

En fin de semaine, on a prélevé au hasard 100 pièces défectueuses dans la production. L'analyse de leur provenance a fourni le tableau ci-dessous :

Poste	*A*	*B*	*C*	*D*
Nombre de pièces défectueuses	33	22	13	32

a. Calculer la valeur $400\times d^2$ correspondant au tableau précédent.

b. Peut-on accepter, au seuil de 10%, l'hypothèse selon laquelle "le nombre de pièces défectueuses est indépendant du poste de fabrication" ?

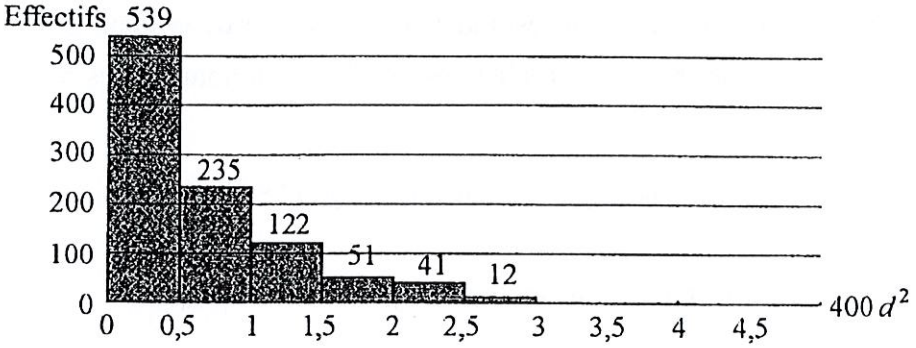
3 LES TRUITES DE PONDICHERY
(d'après BAC ES 2003 – Pondichéry)

Un pisciculteur possède un bassin qui contient 3 variétés de truites : communes, saumonées et arc-en-ciel. Il voudrait savoir s'il peut considérer que son bassin contient autant de truites de chaque variété. Pour cela il effectuera, au hasard, $n = 400$ prélèvements d'une truite dans le bassin, avec remise.

1) Pour juger des résultats que l'on obtiendrait à partir d'un bassin contenant réellement autant de truite de chaque variété, le pisciculteur simule, à l'aide d'un ordinateur, le prélèvement au hasard de 400 truites suivant la loi équirépartie.
Il répète 1000 fois cette opération et calcule à chaque fois la valeur de $400\times d^2$ où d^2 correspond, pour chaque prélèvement de taille $n = 400$ simulé, à :

$d^2 = \left(f_c - \frac{1}{3}\right)^2 + \left(f_s - \frac{1}{3}\right)^2 + \left(f_a - \frac{1}{3}\right)^2$, f_c étant la fréquence d'une truite commune, f_s celle d'une truite saumonée et f_a celle d'une truite arc-en-ciel.

Le diagramme à bandes ci-dessous représente la série des 1000 valeurs de $400\times d^2$, obtenues par simulation.



a. Déterminer une valeur approchée à 0,5 près par défaut, du neuvième décile D9 de cette série.
b. Sous l'hypothèse qu'il y a dans le bassin autant de truite de chaque variété, et en se fondant sur les simulations précédentes, donner un intervalle de la forme [0, D] dans lequel la valeur $400\times d^2$ observée sur un échantillon aléatoire de 400 prélèvements a 90% de chances de se situer (on prendra comme neuvième décile la valeur approchée obtenue à la question précédente).
c. Enoncer une règle de décision permettant d'utiliser ce test à partir d'un échantillon de 400 prélèvements, au seuil de risque de 10%.

2) Le pisciculteur effectue, au hasard, 400 prélèvements d'une truite avec remise dans son bassin, et obtient les résultats suivants :

Variétés	Commune	Saumonée	Arc-en-ciel
Effectifs	146	118	136

a. Calculer, pour cet échantillon, les fréquences de prélèvement f_c d'une truite commune, f_s d'une truite saumonée et f_a d'une truite arc-en-ciel. On donnera les valeurs décimales exactes.
b. Calculer, pour cet échantillon, la valeur $400\times d^2$ arrondie à 10^{-2} ; on note $400\times d^2_{obs}$ cette valeur.
c. Au vu de cet échantillon, peut-on accepter, au seuil de 10%, l'hypothèse selon laquelle il y a autant de truite de chaque variété dans le bassin ?

3) On considère désormais que le bassin contient autant de truites de chaque variété. Quand un client se présente, il prélève au hasard une truite du bassin.
Trois clients prélèvent chacun une truite. Le grand nombre de truites du bassin permet d'assimiler ces prélèvements à des tirages successifs avec remise.
Calculer la probabilité qu'un seul des trois clients prélève une truite commune.

4 AUX GUICHETS DE LA BANQUE
(d'après BAC ES – Métropole juin 2003)

Les guichets d'une agence bancaire d'une petite ville sont ouverts au public cinq jours par semaine : les mardi, mercredi, jeudi, vendredi et samedi.
Le tableau ci-dessous donne la répartition journalière des 250 retraits d'argent liquide effectués aux guichets une certaine semaine.

jour de la semaine	mardi	mercredi	jeudi	vendredi	samedi
rang i du jour	1	2	3	4	5
nombre de retraits	37	55	45	53	60

On veut tester l'hypothèse "le nombre de retraits est indépendant du jour de la semaine".
On suppose donc que le nombre de retraits journaliers est égal à $\frac{1}{5}$ du nombre des retraits de la semaine.

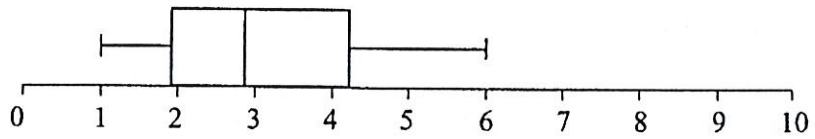
On pose $d^2_{obs} = \sum_{i=1}^{5} \left(f_i - \frac{1}{5} \right)^2$ où f_i est la fréquence des retraits du $i^{\text{ème}}$ jour.

1. Calculer les fréquences des retraits pour chacun des cinq jours de la semaine.

2. Calculer alors la valeur de $1000\, d^2_{obs}$ (la multiplication par 1000 permet d'obtenir un résultat plus lisible).

3. En supposant qu'il y a équiprobabilité des retraits journaliers, on simulé 2000 séries de 250 retraits hebdomadaires.

Pour chaque série, on a calculé la valeur du $1000\, d^2_{obs}$ correspondant. On a obtenu ainsi 2000 valeurs de $1000\, d^2_{obs}$.
Ces valeurs ont permis de construire le diagramme en boîte ci-dessous où les extrémités des "pattes" correspondent respectivement au premier décile et au neuvième décile.



Lire sur le diagramme une valeur approchée du neuvième décile.

4. a) Sous l'hypothèse qu'il y a équiprobabilité des retraits, et en se fondant sur les simulations précédentes, donner un intervalle de la forme [0, D] dans lequel la valeur $1000\, d^2_{obs}$ obtenue sur un échantillon aléatoire de 250 retraits, se situe dans 90% des cas.
   b) Enoncer une règle de décision permettant d'utiliser ce test à partir d'un échantillon de 250 retraits, au seuil de risque de 10%.
   c) D'après la série de 250 retraits observée au début de l'exercice, peut-on accepter, au seuil de 10%, l'hypothèse selon laquelle il y a équiprobabilité des retraits journaliers ?

Eléments de réponse

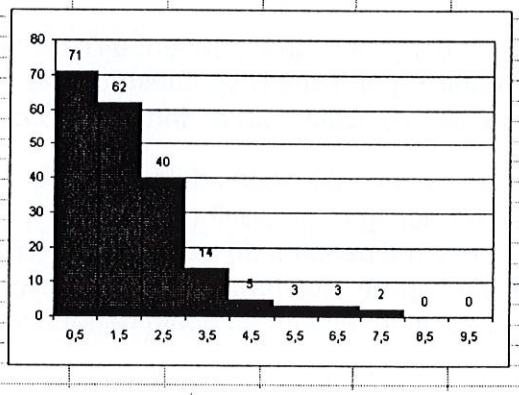
1 LE CLASSEMENT DES LYCEES

1.a. La fonction ALEA( ) produit un nombre au hasard dans l'intervalle [0, 1[.
Donc 4*ALEA( ) + 1 produit un nombre au hasard dans l'intervalle [1, 5[.
On en déduit que ENT(4*ALEA( ) + 1) vaut 1, 2, 3 ou 4, avec la même probabilité. Il suffit d'attribuer la valeur 1 au lycée A, 2 à B, 3 à C et 4 à D.

Remarque :
La simulation (ici non demandée) est aisée à mettre en place. On peut, par exemple, adopter la présentation ci-dessous :

B407 : =(B402-1/4)^2+(B403-1/4)^2+(B404-1/4)^2+(B405-1/4)^2

	A	B	C	D	E	F	G	H	I	J	K	L
395		2	2	2	3	4	1	3	1	3	1	2
396		1	4	1	2	2	2	3	2	1	2	2
397		3	1	3	3	2	3	2	1	4	2	1
398		2	4	2	1	2	4	1	1	4	1	2
399		3	4	1	4	1	4	1	1	3	4	2
400		3	4	3	1	4	4	2	1	1	4	2
401		fréq obs	fréq obs	fréq obs	fréq obs	fréq obs	fréq obs	fréq obs	fréq obs	fréq obs	fréq obs	fréq obs
402	1	0,23	0,2875	0,2675	0,2525	0,2775	0,245	0,2575	0,24	0,265	0,225	0,2325
403	2	0,25	0,24	0,2575	0,2475	0,245	0,225	0,2525	0,2475	0,25	0,27	0,2675
404	3	0,2625	0,22	0,25	0,23	0,265	0,28	0,2425	0,255	0,255	0,2525	0,26
405	4	0,2575	0,2525	0,225	0,27	0,2125	0,25	0,2475	0,2575	0,23	0,2525	0,24
406		D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11
407		0,0006125	0,0024125	0,0009875	0,0008125	0,0024125	0,00155	0,000125	0,0001875	0,00065	0,0010375	0,0008125
408		1000*D	1000*D	1000*D	1000*D	1000*D	1000*D	1000*D	1000*D	1000*D	1000*D	1000*D
409		0,6125	2,4125	0,9875	0,8125	2,4125	1,55	0,125	0,1875	0,65	1,0375	0,8125
410	répartition	1000*D^2										
411	1	71	0,5									
412	2	62	1,5									
413	3	40	2,5									
414	4	14	3,5									
415	5	5	4,5									
416	6	3	5,5									
417	7	3	6,5									
418	8	2	7,5									
419	9	0	8,5									
420	10	0	9,5									
421												
422	total	200										
423												
424												
425												
426												
427												
428												



On obtient 200 répartitions en recopiant la colonne B jusqu'en colonne GS.
1.b. Dans la simulation 1, on a 9 expériences sur 200 (soit moins de 5%) qui correspondent au cas où $1000 \times d^2 > 5$.
1.c. La simulation 2 est différente en raison des fluctuations d'échantillonnage (c'est à dire du hasard). Cependant, on constate que ces fluctuations sont limitées et ne modifient pas l'aspect global de l'histogramme. En particulier, on a, dans 95% des cas, $1000 \times d^2 \leq 5$.

2.a. D'après les simulations, on peut penser que, sous l'hypothèse que le nombre de reçus est indépendant du lycée, on a, dans 95% des cas, $1000 \times d^2 \in [0, 5]$.
2.b. Règle de décision, à partir d'un échantillon de taille 400, au seuil de 5% :
A partir d'une répartition de 400 reçus, on calcule la quantité $1000 \times d^2$.
• Si $1000 \times d^2 \leq 5$ on accepte, au seuil de 5%, l'hypothèse selon laquelle le nombre de reçus est indépendant du lycée.
• Si $1000 \times d^2 > 5$ on **rejette**, au **risque** de 5%, l'hypothèse selon laquelle le nombre de reçus est indépendant du lycée.

3.a. Le calcul effectué est :

$1000 \times d^2 = 1000 \times [(\frac{101}{400} - \frac{1}{4})^2 + (\frac{82}{400} - \frac{1}{4})^2 + (\frac{117}{400} - \frac{1}{4})^2 + (\frac{100}{400} - \frac{1}{4})^2]$.

3.b. On a 3,84 ≤ 5, on accepte donc, au seuil de 5%, l'hypothèse selon laquelle le nombre de reçus est indépendant du lycée.

Remarque (entre nous) :
Il est clair que le choix du seuil de risque n'est pas neutre. En prenant 5%, cela signifie que l'on ne rejettera l'hypothèse d'équidistribution que si les valeurs observées en sont très éloignées (correspondant à 5% des cas lorsque l'hypothèse est vraie). Le proviseur privilégie ainsi cette hypothèse (ce qui est son intérêt). Il faudrait supposer que ce seuil de 5% est celui habituellement utilisé dans ce genre de situation (avant que les résultats du bac ne soient connus).
Observons ce que donnerait ce test au seuil de 10%.
On montre que, sous l'hypothèse d'équirépartition, la variable aléatoire :

$4nD^2 = 1600\sum(F_i - \frac{1}{4})^2$, suit approximativement la loi du khi-2 à 3 degrés de liberté, dont les tables indiquent que le 9ème décile vaut 6,25.

Ainsi, sous l'hypothèse d'équirépartition, on a dans 90% des cas, $1000 \times d^2 \leq \frac{6,25 \times 1000}{1600}$

c'est à dire $1000 \times d^2 \leq 3,91$.
On a observé ici $1000 \times d^2 = 3,84$. L'hypothèse d'équirépartition est encore retenue au seuil de 10%, mais cela commence à être douteux.

2 UNE APPLICATION INDUSTRIELLE

1.a. On lit $D9 \approx 6,3$ (à 0,1 près) correspondant à l'extrémité de la "moustache" de droite
1.b. La simulation montre que, sous l'hypothèse que le nombre de pièces défectueuses est indépendant du poste de fabrication, on a, dans 90% des cas, $400 \times d^2 \in [0 ; 6,3]$.

2. A partir d'une répartition de 100 pièces défectueuses, on calcule la quantité $400 \times d^2$
• Si $400 \times d^2 \leq 6,3$ on accepte, au seuil de 10%, l'hypothèse selon laquelle le nombre de pièces défectueuses est indépendant du poste de fabrication.
• Si $400 \times d^2 > 6,3$ on **rejette**, au **risque** de 10%, l'hypothèse selon laquelle le nombre de pièces défectueuses est indépendant du poste de fabrication.

3.a. On observe $400 \times d^2 = 10,64$.
3.b. On rejette, au risque de 10%, l'hypothèse selon laquelle le nombre de pièces défectueuses est indépendant du poste de fabrication. Autrement dit, l'écart observé par rapport à l'équirépartition est jugé ici comme significatif d'une qualité de fabrication différente entre les différents postes. On en recherchera donc la cause...

3 LES TRUITES DE PONDICHERY

1.a. On constate que 89,6% des valeurs de $400 \times d^2$ sont dans l'intervalle [0 ; 1,5] et que 94,7% des valeurs de $400 \times d^2$ sont dans l'intervalle [0 ; 2]. On en déduit que D9, correspondant à 90%, se situe entre 1,5 et 2.
Donc D9 = 1,5 à 0,5 près par défaut.
1.b. Dans l'hypothèse où l'échantillon aléatoire de 400 truites est prélevé dans une population équirépartie, la valeur $400 \times d^2$ se situe dans 90% des cas dans l'intervalle [0, D9], soit, selon les simulations, et en arrondissant, dans l'intervalle [0 ; 1,5].

1.c. La règle de décision, au seuil de 10%, peut-être la suivante :
On prélève un échantillon aléatoire (avec remise puisque la simulation est ainsi faite) de 400 truites, à partir duquel on calcule la quantité $400\times d^{2}_{obs}$.
• Si $400\times d^{2}_{obs} \leq 1{,}5$ on accepte, au seuil de 10%, l'hypothèse selon laquelle les variétés de truites sont équiréparties dans le bassin.
• Si $400\times d^{2}_{obs} > 1{,}5$ on **rejette**, au **risque** de 10%, l'hypothèse selon laquelle les variétés de truites sont équiréparties dans le bassin.

2.a. On trouve $f_c = 0{,}365$; $f_s = 0{,}295$ et $f_a = 0{,}340$.
2.b. On trouve $400\times d^{2}_{obs} = 1{,}01$.
2.c. On a $400\times d^{2}_{obs} \leq 1{,}5$ on accepte donc, au seuil de 10%, l'hypothèse selon laquelle les variétés de truites sont équiréparties dans le bassin.

3. La variable aléatoire X correspondant au nombre de truites communes pêchées par trois clients suit la loi binomiale de paramètres 3 et 1/3 (répétition indépendante d'une épreuve de Bernoulli où la probabilité de succès a été supposée égale à 1/3).
On a donc $P(X = 1) = \binom{3}{1} \times \frac{1}{3} \times \left(\frac{2}{3}\right)^{2} = 4/9$.

Remarques (entre nous !) :
L'énoncé proposé à l'examen a été ici profondément modifié, pour les raisons suivantes (il en va de même pour l'exercice suivant).

◆ Il s'agit de distinguer soigneusement **trois étapes** :
• la construction d'une zone d'acceptation de l'hypothèse d'équirépartition fondée sur la simulation,
• l'énoncé d'une règle de décision,
• sa mise en œuvre à partir des données observées.
Ce n'est pas le cas dans l'énoncé de l'examen, où, de façon plutôt confuse, on demande d'un peu tout faire en même temps, en une seule question, rédigée ainsi :
"En argumentant soigneusement la réponse dire si on peut affirmer avec un risque d'erreur inférieur à 10% que « le bassin contient autant de truites de chaque variété »."
◆ A propos de l'ordre des questions :
Il est toujours un peu gênant de construire le test après l'étude d'un échantillon. En effet, on peut toujours modifier le choix du seuil, pour accepter, ou refuser, l'hypothèse testée. Ce choix résulte en fait habituellement d'un compromis entre des intérêts contradictoires et il est généralement essentiel, pour la déontologie statistique, que le test soit construit avant la prise d'échantillon.
◆ A propos de la notion de risque :
Du point de vue de la notion de risque, la formulation de l'exercice posé à l'examen est pour le moins confuse. Le risque porte sur le fait de **refuser à tort** le modèle. Il s'agit en fait du risque de 1ère espèce (ici de 10%), alors qu'un risque de 2ème espèce, accepter à tort le modèle, est toujours présent mais plus difficile à évaluer puisque dépendant de la nature de la distribution réelle de la population. Lorsque l'on dit, comme dans l'énoncé du bac, "Peut-on affirmer que la population est équirépartie ?", le risque d'erreur, dans cette affirmation est celui de seconde espèce, et n'est sans doute pas de 10%, ou inférieur à 10%. En fait, on ne sait pas ce qu'il est.
Une phrase du type "on peut affirmer, avec un risque d'erreur inférieur à 10%, que « le bassin contient autant de truites de chaque variété »" est donc un abus de langage, que, certes, les statisticiens peuvent parfois commettre, parce qu'ils savent bien de quoi ils parlent (du risque de 1ère espèce), mais qui ne peut que troubler des débutants et les conduire à de graves confusions. Comment alors comprendre qu'on prenne un risque de 10%, et pourquoi pas de 1% ?

◆ Pourquoi $400 \times d^2_{obs}$?
Les correcteurs de Pondichéry ont dû se demander s'ils n'avaient pas raté un épisode (ou plusieurs...) en lisant dans le corrigé fourni : "Le problème de d^2_{obs} est qu'il ne tient pas compte de n = 400. C'est la raison pour laquelle on considère 400 d^2_{obs}".
Il faut peut-être comprendre que la loi de la variable aléatoire dont 400 d^2_{obs} est une réalisation, est, à peu près, indépendante de n = 400 (la variable aléatoire $3nD^2$ suit approximativement, pour n grand, et sous l'hypothèse d'équirépartition, une loi du khi-deux à 2 degrés de liberté). Cette remarque, intéressante si l'on construit un test du khi-deux, en s'appuyant sur sa loi, est ici sans grand intérêt. L'énoncé de métropole (voir exercice suivant) dit, plus raisonnablement, que si l'on considère 1000 d^2_{obs} (et non d'ailleurs 250 d^2_{obs}) c'est pour "obtenir un résultat plus lisible."
◆ Est-ce bien raisonnable ?
Certains se sont interrogés sur l'aspect "réaliste" de cet exercice. On voit mal un pisciculteur opérer ainsi, d'autant qu'il est difficile de distinguer, extérieurement, une truite saumonée d'une truite arc-en-ciel...On aurait donc pu trouver une application plus pertinente de la statistique.

4 AUX GUICHETS DE LA BANQUE

2. On trouve $1000\,d^2_{obs}$ = 5,248 .

3. D'après le diagramme en boîte, D9 = 6 (extrémité de la "patte" ou "moustache" de droite).

4.a. Dans l'hypothèse où il y a équiprobabilité des retraits, la valeur $1000\,d^2_{obs}$ se situe dans 90% des cas dans l'intervalle [0, D9], soit, selon les simulations, dans l'intervalle [0 ; 6].
4.b. La règle de décision, au seuil de 10%, peut-être la suivante :
Sur un échantillon aléatoire de 250 retraits, on calcule la quantité $1000\,d^2_{obs}$.

• Si $1000\,d^2_{obs} \leq 6$ on accepte, au seuil de 10%, l'hypothèse selon laquelle "le nombre de retraits est indépendant du jour de la semaine".

• Si $1000\,d^2_{obs} > 6$ on rejette, au risque de 10%, l'hypothèse selon laquelle "le nombre de retraits est indépendant du jour de la semaine".

4.c. On a, avec l'échantillon du début de l'exercice, $1000\,d^2_{obs} \leq 6$, on accepte donc, au seuil de 10%, l'hypothèse selon laquelle "le nombre de retraits est indépendant du jour de la semaine".

II – LOIS CONTINUES

Voici, sur ce sujet, le contenu du programme de terminale S applicable à la rentrée 2002 :

> *"Contenus : Exemples de lois continues :*
> *- loi uniforme sur l'intervalle [0, 1] ;*
> *- loi de durée de vie sans vieillissement."*
> *"Modalités de mise en œuvre : application à la désintégration radioactive, loi exponentielle de désintégration des noyaux."*
> *"Commentaires : ce paragraphe est une application de ce qui aura été fait en début d'année sur l'exponentielle et le calcul intégral."*

Quelques précisions sont apportées par le document d'accompagnement :

> "[Il s'agit d'une] *introduction aux loi de probabilité à densité continue, qui fait naturellement suite au cours sur l'intégration et l'enrichit d'applications importantes, telles la modélisation de la durée de vie d'un noyau d'une substance radioactive."*
> *"Aucune difficulté technique ne sera soulevée ; en particulier on ne traitera que des cas menant à des calculs d'intégrales s'exprimant aisément à l'aide des fonctions étudiées en terminale. Pour une loi de $\mathbb{R}^+$, aucune notion d'intégrale généralisée n'est abordée formellement : l'outil limite à l'infini d'une fonction est suffisant."*
> *"Pour la classe terminale, on se limite à des lois de probabilité définies sur un intervalle I borné ou borné à gauche (i. e. [a, b] ou [a, + ∞[) et dites à densité continue."*
>
> *"Pour définir le choix au hasard d'un nombre réel dans [0, 1[, on ne peut plus passer par la probabilité p de chaque élément, puisqu'on devrait avoir p = 0 : cette difficulté a été un véritable défi pour les mathématiciens et a conduit à repenser la notion de loi de probabilité."*
> *"On pourra faire l'analogie avec les densités de masse (la masse d'un point est nulle, celle d'un segment est proportionnelle à sa longueur dans le cas d'une tige à densité constante)."*

1 – INTRODUIRE LES LOIS CONTINUES

Le passage des lois discrètes aux lois continues pose quelques difficultés, essentiellement celle de la notion de ***fonction de densité***, d'ailleurs trop souvent considérée comme une "fonction de probabilité" dans les modes d'emploi (mal traduits ?) de calculatrice ou l'aide du tableur. Cette fonction ne fournit pas une probabilité puisque la probabilité d'une valeur ponctuelle est nulle. C'est là le "mystère" de la continuité : la probabilité de chacune des valeurs ponctuelles observées est nulle (avant la réalisation de cette valeur).
Le modèle mathématique de la continuité est puissant et efficace. Il permet de rendre compte de nombreuses situations. Mais ce n'est qu'un modèle. La réalité physique n'est sans doute pas continue. Il n'empêche qu'à une certaine échelle, il est bien pratique de la décrire ainsi. La loi exponentielle, au programme de terminale S, modélise généralement un temps d'attente. Il est habituel de considérer un temps d'attente comme un intervalle de $\mathbb{R}^+$. Quelle est la nature physique du temps ? On sait, depuis Einstein, la complexité de la question.

Le passage, en terminale S, aux modèles probabilistes continus permettra donc d'étendre considérablement le champs d'application de la statistique et des probabilités. Le passage du discret au continu nécessite cependant de gros efforts conceptuels, qui historiquement, n'ont été possibles qu'après l'axiomatisation de Kolmogorov en 1933. Ce cadre rigoureux étant hors de propos en terminale, la simulation, sur calculatrice ou ordinateur, favorisera l'acceptation et la compréhension des outils mathématiques utilisés (intégration, passages à la limite...).
Selon Michel Henry[1], de l'IREM de Franche-Comté, "l'introduction de lois continues en Terminale, outre la résolution de problèmes plus intéressants que les problèmes traditionnels de combinatoire à propos de jeux de hasard, a pour importance essentielle de rendre incontournable la notion de modèle." Ainsi, certaines difficultés conceptuelles, comme celle de la probabilité ponctuelle nulle, pourraient être dues au fait qu'on oublie qu'il s'agit d'un modèle, où tout n'est pas susceptible d'une interprétation dans la réalité.

Dans les activités qui suivent, il s'agit d'introduire la notion fondamentale de fonction de densité de probabilité, en montrant que, dans le cadre continu, les calculs, liés à la notion d'aire, sont des calculs d'intégrales. On fera en particulier remarquer qu'ainsi, l'aire totale entre l'axe des abscisses et la courbe vaut 1, que la probabilité d'une valeur ponctuelle est nulle...
La manière la plus naturelle, semble-t-il, d'introduire la densité de probabilité est de relier sa représentation graphique à la notion d'histogramme pour les variables aléatoires discrètes. Dans chaque cas, la probabilité est proportionnelle à la surface.
On propose ici des travaux dirigés pour introduire les variables aléatoires continues en terminale S, mêlant simulation sur calculatrice et analyse.

[1] Revue "Repères" – Avril 2003.

TRAVAUX DIRIGES

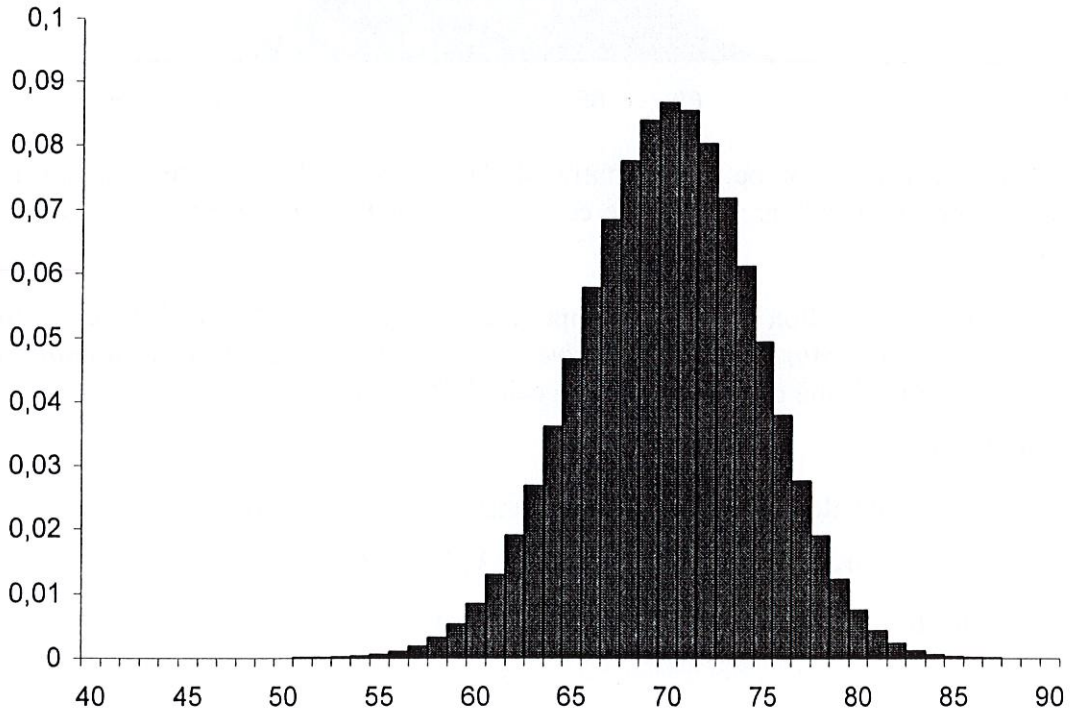
INTRODUCTION AUX VARIABLES ALEATOIRES CONTINUES

Généralités

➪ Une ***variable aléatoire discrète*** X prend des valeurs "isolées" :

Par exemple, une variable aléatoire X suivant la loi binomiale $\mathcal{B}(n, p)$ prend comme valeurs possibles les nombres entiers entre 0 et n avec des probabilités p_i telles que $p_i = \binom{n}{p} p^i (1-p)^{n-i}$.

La distribution d'une variable aléatoire discrète est souvent représentée sous forme d'un ***histogramme***.

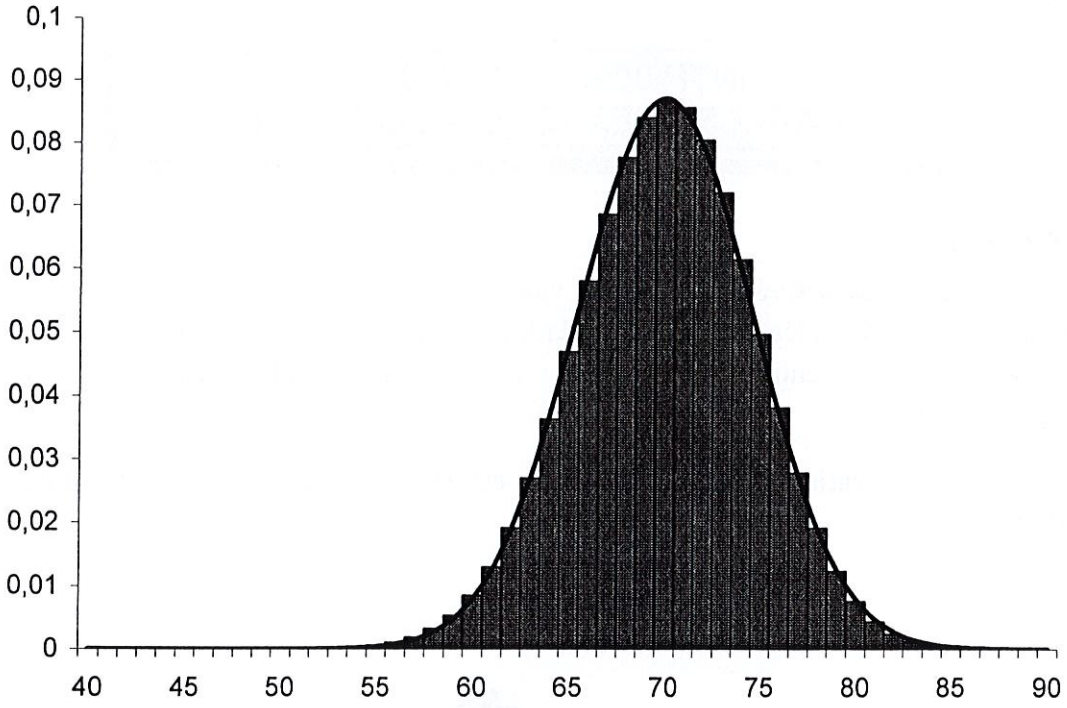


On a représenté ci-dessus la distribution de la loi binomiale $\mathcal{B}$ (100 ; 0,7). Les rectangles ayant pour largeur une unité, la probabilité $p_i = P(X = i)$ est égale à ***l'aire du rectangle*** correspondant (en unités d'aire).

L'***espérance*** d'une variable aléatoire discrète X prenant n valeurs possibles $x_1, \ldots, x_n$ est :

$E(X) = \sum_{i=1}^{n} x_i \, p_i$, sa ***variance*** est $V(X) = \sum_{i=1}^{n} (x_i - E(X))^2 \, p_i$.

➪ Une ***variable aléatoire continue*** Y peut prendre comme valeurs possibles tous les nombres réels d'un certain intervalle.
Sa distribution est donnée par sa ***fonction de densité*** f.



On a représenté ci-dessus la courbe représentative de la densité f d'une variable aléatoire Y continue fournissant des résultats proches de ceux d'une variable aléatoire de loi binomiale $\mathscr{B}(100\ ;\ 0{,}7)$.

La ***probabilité*** qu'une réalisation de Y soit comprise entre a et b est alors ***l'aire*** (en unités d'aire) ***située sous la courbe représentative de f, entre les droites d'équation x = a et x = b.*** Cette probabilité est obtenue par un calcul d'intégrale :

$P(a \le Y \le b) = \int_a^b f(x)\,dx$.

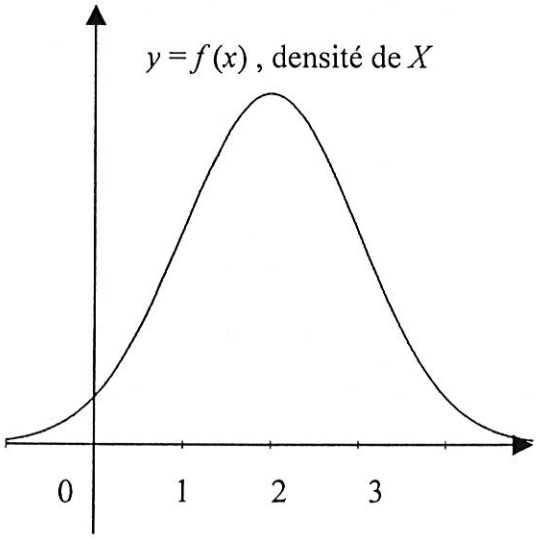
L'***espérance*** d'une variable aléatoire continue Y de densité f définie sur $[\alpha, \beta]$ est :

$E(Y) = \int_\alpha^\beta x\,f(x)\,dx$ sa ***variance*** est $V(Y) = \int_\alpha^\beta (x - E(X))^2\,f(x)\,dx$

(à comparer avec les formules analogues dans le cas discret).

1 A partir d'une courbe de densité

La variable aléatoire X est à pour fonction de densité la fonction f représentée ci-après.



1) Dans le cas ci-contre, on a :

$P\,(1 \leq X \leq 3) = \int_1^3 f(x)\,dx$.

Hachurer l'aire correspondante.

2) Déterminer :

$P\,(X = 2) = \int_2^2 f(x)\,dx$ = ...

Commenter votre résultat : .....
.....
.....
.....
.....

2 LOI UNIFORME SUR [0;1]

1) Soit Y la variable aléatoire qui correspond au tirage au hasard d'un nombre d'au plus 10 décimales dans l'intervalle [0, 1[.

a) Justifier qu'il y a 10^{10} résultats possibles. .....

b) Quelle est la probabilité de l'événement "$Y = 0{,}4536694833$" ? .....

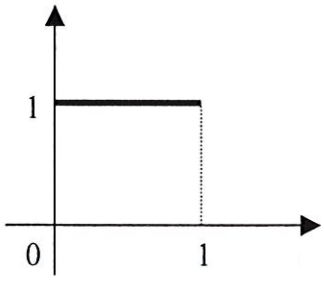
et de l'événement "$Y = 0{,}5$" ? .....

2) Soit X la variable aléatoire qui correspond au tirage au hasard d'un nombre réel de l'intervalle [0, 1] (il y a une infinité de réels dans cet intervalle).

Quelle est la probabilité de l'événement "$X = 0{,}5$ " ? .....

Quelle est, intuitivement, la probabilité de l'événement " $0 \leq X \leq \frac{1}{2}$ " ? .....

3) La variable aléatoire X admet comme ***densité*** la fonction f définie par :



$f(x) = 1$ si $x \in [0, 1]$ et $f(x) = 0$ si $x \notin [0, 1]$.

Utiliser la fonction f pour calculer :

a) $P\,(0 \leq X \leq 0{,}5) = \int_0^{0,5} f(x)\,dx$.

.....

b) $P\,(0{,}2 \leq X \leq 0{,}3)$.

.....

c) $E\,(X) = \int_0^1 x\,f(x)\,dx$.

.....

d) $V\,(X) = \int_0^1 [x - E(X)]^2\,f(x)\,dx$, puis $\sigma\,(X) = \sqrt{V(X)}$

.....

e) Que vaut $\int_0^1 f(x)\,dx$? Justifier le résultat par un argument probabiliste.

.....

4) ***Simulation*** de X avec la calculatrice :
Votre calculatrice contient une fonction "***Random***", symbolisée par Ran# ou rand, qui simule une réalisation d'une variable aléatoire de loi uniforme sur [0, 1] (du moins, le choix au hasard d'un nombre décimal).
Effectuer Ran# ou rand sur votre calculatrice puis plusieurs fois EXE ou ENTER. Observer.
Le programme ci-dessous effectue 100 fois la fonction "Random" puis détermine la moyenne et l'écart type des résultats.

CASIO GRAPH 25 30 65 80 100	TI 82 83	TI 89 92
ClrList ↵ Seq(0,I,1,100,1) → List 1 ↵ For 1 → I To 100 ↵ Ran# → List 1[I] ↵ Next ↵ 1-Variable List 1 , 1	:ClrList L_1 :seq(0,I,1,100,1) → L_1 :For (I,1,100) :rand → L_1(I) :End :Disp mean(L_1) :stdDev(L_1)	:DelVar L1 :seq(0,i,1,100,1) → L1 :For i , 1 , 100 :rand( ) → L1[i] :EndFor :Disp mean(L1) :Disp stdDev(L1)

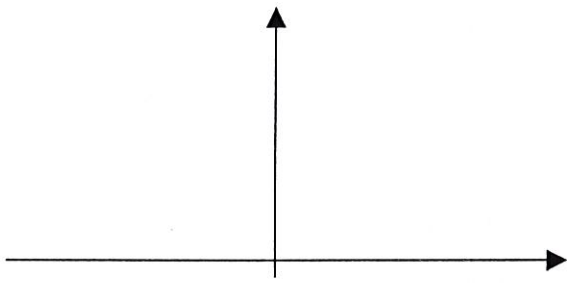
Sur CASIO : On obtient **CLRList** par PRGM CLR ; **Seq** et **List** par OPTN LIST ; **Ran#** par OPTN PROB ; **1-Variable** par PRGM EXIT MENU (F4) STAT CALC.
Sur TI : On obtient **mean** et **stdDev** par 2nd List MATH.

Comparer avec les valeurs théoriques $E(X)$ et $\sigma(X)$:
.....

3 RECHERCHE DE FONCTIONS DE DENSITES

1) Fonction "chapeau" :
Rechercher une fonction f telle que :

- $f(x) = 0$ si $x \notin [-1, 1]$,
- $f(-1) = f(1) = 0$,
- $f(0) > 0$,
- f est paire, représentée par deux segments ,
- $\int_{-1}^{1} f(x)\,dx = 1$.



.....
.....
.....
.....
.....
.....

2) Remplacer les segments de droites précédents par des arcs de courbes d'équation $y = a\cos(\omega x)$ où a et ω sont à déterminer en conservant les autres conditions.
.....
.....
.....
.....
.....
.....
.....
.....

3) Soit f définie sur $\mathbb{R}$ par $f(x) = \begin{cases} 0 \text{ si } x < 0 \\ ae^{-2x} \text{ si } x \geq 0 \end{cases}$, où a est une constante strictement positive.

a) Calculer, pour $x \geq 0$, $f'(x)$.

.....
.....
.....
.....
.....
.....

b) En déduire les variations de f.

.....
.....

c) Déterminer $\lim\limits_{x \to +\infty} f(x)$ puis dresser le tableau de variation de f.

.....
.....
.....
.....
.....

d) Calculer , pour $t > 0$, $I(t) = \int_0^t f(x)\, dx$.

.....
.....
.....
.....
.....

e) Déterminer $\lim\limits_{t \to +\infty} I(t)$.

.....
.....
.....
.....

Quelle valeur donner au réel a pour que f soit la fonction de densité d'une variable aléatoire X ?

.....
.....
.....

f) Calculer, dans ces conditions, pour $t > 0$, $J(t) = \int_0^t x f(x)\, dx$,

puis l'espérance de X : $E(X) = \lim\limits_{t \to +\infty} J(t)$.

.....
.....
.....
.....
.....
.....

2 – LA LOI EXPONENTIELLE

"Une circonstance bien digne d'attention, et sur laquelle nous aurons l'occasion de revenir, c'est que les faits qui paraissent les plus accidentels quand ils sont considérés un à un, manifestent un ordre lorsqu'on peut en observer un grand nombre de simultanés et consécutifs ; et le calcul fait voir comment, sans connaître la nature de leurs causes ni le nombre des combinaisons qui les produisent ou les contrarient, on peut assigner des limites à leurs possibilités respectives, et par conséquent spéculer sur l'avenir conformément aux règles de la prudence."

S. F. Lacroix – "Traité élémentaire du calcul des probabilités" – 3ème édition, 1833.

La loi exponentielle est l'une des plus impressionnantes de la théorie des probabilités, elle instaure de l'ordre là il semble qu'il y en ait le moins et permet la prévision (via la loi des grands nombres) dans des situations qui paraissent extrêmes. L'expérimentation, par simulation, permettra d'en prendre conscience. En outre, son étude est particulièrement riche, d'un point de vue mathématique, parce qu'elle mêle probabilités et analyse (limites, intégrales, fonction exponentielle), comme d'un point de vue interdisciplinaire et sociologique, par ses applications en physique ou dans l'étude des risques de catastrophes au centre de certains débats de société.
Voici des extraits du document d'accompagnement du programme de TS (rentrée 2002) :

"Les évènements rares ont-ils une loi ? [...] Il semble que cette question n'ait pas beaucoup de sens [...] et pourtant, il est possible d'émettre quelques hypothèses en utilisant des propositions classiques sur les limites ; [ceci] s'appliquera, plus généralement, aux évènements rares qui surviennent lors d'une activité fréquente :
par exemple, pannes de matériel, accidents d'avion, randonneurs frappés par un éclair, désintégration des noyaux d'une substance radioactive, etc."

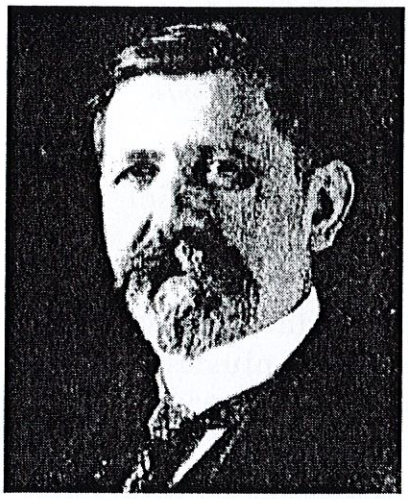
"Une convergence thématique forte apparaît avec le chapitre "Radioactivité" du programme de physique."
"C'est l'occasion de traduire dans le champs des mathématiques la notion d'absence d'usure ; ce travail de modélisation illustre une pratique que les élèves n'ont en général pas eu l'occasion de rencontrer."

a) UN TP D'INTRODUCTION DE LA LOI EXPONENTIELLE

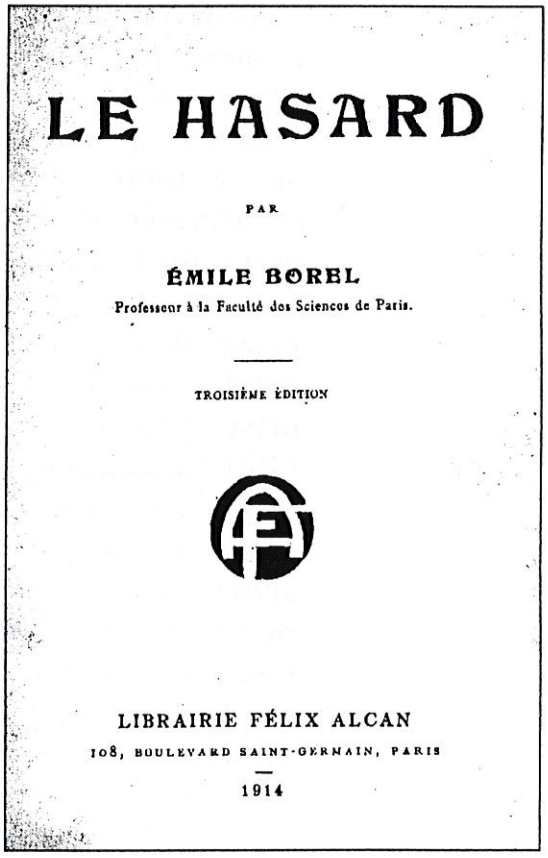
Le TP suivant, sur le thème de la désintégration nucléaire, permet l'introduction de la partie du cours sur la loi exponentielle.

TRAVAUX DIRIGES

LOI EXPONENTIELLE DE DESINTEGRATION DES NOYAUX



Emile Borel (1871-1956) est un grand mathématicien français, qui contribua à la théorie des fonctions et à celle des probabilités. Les travaux de *Marie Curie* sont à l'origine de la théorie de la radioactivité. Victime, en 1911, d'une campagne de presse calomnieuse (on met en cause sa vie privée, ses origines étrangères...), elle sera farouchement défendue par *Borel*. De cette affaire date une solide amitié entre *Borel* et *Marie Curie*.


LE HASARD

PAR

ÉMILE BOREL

Professeur à la Faculté des Sciences de Paris.

TROISIÈME ÉDITION

LIBRAIRIE FÉLIX ALCAN

108, BOULEVARD SAINT-GERMAIN, PARIS

1914

Le texte suivant, extrait du livre d'*Emile Borel*, "*Le hasard*" (1914), est contemporain des découvertes de *Pierre* et *Marie Curie* dans le domaine atomique. Il décrit, sous forme vulgarisée, la loi statistique de désintégration des atomes radioactifs.

La radioactivité est une suite de désintégrations par lesquels les noyaux des éléments radioactifs évoluent spontanément vers un état plus stable. Ce faisant, ils peuvent émettre un noyau d'hélium (particule α), un électron (particule β^-) ou un positron (particule β^+). Les noyaux obtenus sont en général dans des états excités et se désexcitent en émettant un photon énergétique (rayon γ).

Lire le texte suivant.

— On a beaucoup étudié dans ces dernières années des phénomènes importants et nouveaux, dans lesquels la théorie des probabilités intervient pour ainsi dire à chaque instant : ce sont les phénomènes de radioactivité. Je ne puis m'étendre ici sur l'historique de la découverte ni sur le détail de ces phénomènes, qui ont pris rapidement une si grande place dans la physique[1]. Il est cependant nécessaire d'en préciser brièvement la nature. Le caractère essentiel de la radioactivité paraît être la décomposition *spontanée* de certains atomes. Cette décomposition se distingue très nettement de la dissociation chimique d'une molécule en plusieurs atomes, par plusieurs caractères dont le principal peut-être est son *invariance* à l'égard de tous les agents physiques . En d'autres termes, en un temps donné, une substance radioactive déterminée se trouve perdre, en vertu du phénomène de la radioactivité, une proportion rigoureusement déterminée de son poids. Tout se passe donc comme si chaque atome radioactif avait à chaque instant la même probabilité de se briser pendant la seconde suivante, cette probabilité ne pouvant être modifiée ni par les agents physiques (température, pression, champ électrique ou magnétique), ni par le vieillissement spontané de l'atome lui-même. Si l'on admet ce point de vue comme une interprétation exactement adéquate des faits expérimentaux — et il semble bien qu'on ne puisse point ne pas l'admettre — l'étude mathématique des phénomènes radioactifs est manifestement du domaine de la théorie des probabilités.

1. Voir, pour l'historique, dans la *Revue du Mois* du 10 janvier 1913, la *Conférence Nobel* 1903 par Pierre CURIE et la *Conférence Nobel* 1912 par Mme Pierre CURIE.

[...]

[...]

On conçoit donc sans peine qu'il ait été possible, même en opérant sur des corps plus radioactifs que le radium, d'arriver à déceler expérimentalement les émissions de particules α et à mesurer les intervalles de temps qui les séparent. La répartition de ces intervalles de temps autour de leur valeur moyenne est un problème de probabilités continues. On peut en effet, la durée de l'expérience étant faible par rapport à la durée nécessaire pour une diminution appréciable de la masse radioactive utilisée, considérer que la probabilité de l'émission est constante ; l'espérance mathématique du joueur qui recevrait une somme fixe par particule émise, est donc proportionnelle au temps. Le problème de probabilités continues peut être posé sous la forme géométrique suivante : *Sur une droite indéfinie sont marqués au hasard un certain nombre de points, de telle manière qu'il y ait en moyenne* hx *points sur une longueur* x *; quelle est la probabilité pour que la distance d'un point marqué à celui qui est situé immédiatement à sa droite soit supérieure à une longueur donnée* y. Un calcul facile[1] montre que cette probabilité est

1. Voir E. Borel, *Introduction géométrique à quelques théories physiques*. Note V (Gauthier-Villars.)

e^{-hy}. Si l'on mesure un grand nombre de distances (c'est-à-dire d'intervalles de temps écoulés entre deux émissions successives), on peut vérifier l'accord entre ce résultat et l'expérience. Cette vérification a été faite d'une manière satisfaisante[1], rendant par suite très vraisemblable le point de départ du calcul.

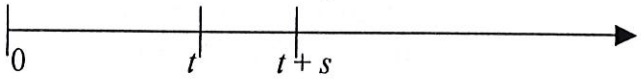
. . . . . .

1. Je citerai notamment une expérience très complète faite par Mme Curie, avec l'aide de ses préparateurs, et non encore publiée au moment où j'écris ces lignes. Cette expérience a porté sur 10.000 émissions et l'étude numérique, faite avec le plus grand soin par Mme Curie, concorde admirablement avec les prévisions théoriques. Cette concordance est la preuve expérimentale la plus complète de l'invariance de la radioactivité.

On considère une matière radioactive et on note T la variable aléatoire qui à tout atome radioactif pris au hasard associe le temps d'attente avant sa désintégration.
On suppose que T est une variable aléatoire continue de densité f, définie sur $[0, +\infty[$.

On désigne par F la primitive de f définie sur $[0, +\infty[$ par $F(t) = \int_0^t f(x)dx = P(T \leq t)$.

Soit $t > 0$ quelconque et considérons l'intervalle de temps $[t, t + s]$.
1) *Emile Borel* affirme : *"en un temps donné, une substance radioactive déterminée se trouve perdre, en vertu du phénomène de la radioactivité, une proportion rigoureusement déterminée de son poids."*
Justifier que la proportion de poids perdu par la substance pendant l'intervalle de temps $[t, t + s]$ est $\frac{P(t \leq T \leq t + s)}{P(T > t)}$.



.....
.....
.....
.....

Montrer que ce rapport peut s'écrire $\frac{F(t + s) - F(t)}{1 - F(t)}$.

.....
.....
.....

2) On désigne par "taux moyen de désintégration par unité de temps entre t et $t + s$ " la quantité $\frac{F(t + s) - F(t)}{1 - F(t)} \times \frac{1}{s}$ et par "taux instantané de désintégration au temps t " la limite

$h(t) = \lim_{s \to 0} \frac{F(t + s) - F(t)}{1 - F(t)} \times \frac{1}{s}$.

Montrer que $h(t) = \frac{F'(t)}{1 - F(t)}$.

.....
.....
.....

3) D'après le texte, la décomposition radioactive se *"distingue"* par son invariance et, pour tout $t \in \mathbb{R}$, $h(t) = h$ est constant.

On a donc, pour tout $t \in \mathbb{R}$, $\frac{F'(t)}{1 - F(t)} = h$. En déduire que $\ln(1 - F(t)) = -ht + k$ où k est une constante réelle.

.....
.....
.....

4) Déterminer $F(0)$ et en déduire que, pour tout $t \geq 0$, $F(t) = 1 - e^{-ht}$.
Donner l'expression de f.

.....
.....
.....
.....

5) A la fin du texte, *E. Borel* dit qu'un "*calcul facile*" donne $P(T > y) = e^{-hy}$. Vérifier cette affirmation.

.....

6) Soit a un nombre réel strictement positif fixé. On note $I(a) = \int_0^a t\,f(t)dt$.

Démontrer, à l'aide d'une intégration par parties, que $I(a) = \frac{1}{h} + e^{-ha}(-a - \frac{1}{h})$.

.....
.....
.....
.....

7) Déterminer l'espérance de T (temps "moyen" d'attente d'une désintégration) : $E(T) = \lim_{a \to +\infty} I(a)$.

.....
.....
.....

8) Simulation et comparaison aux résultats théoriques :

a) On suppose que $h = 0{,}07$.

Montrer que l'instruction, sur calculatrice, int(rand + 0.07) simule pour une unité de temps, la désintégration éventuelle de l'atome.

.....
.....

Effectuer plusieurs fois le programme suivant. Que simule-t-il ?

.....

CASIO sans instruction While	CASIO avec instruction While	T.I. 80 - 81 (sans instruction While)	T.I. 82 83	T.I. 89 92
-1 → I ↵ Lbl 1 ↵ I + 1 → I ↵ Int (Ran# + 0.07) = 0 ⇒ Goto 1 ↵ I	0 → I ↵ While Int(Ran# + 0.07) = 0 ↵ I + 1 → I ↵ WhileEnd ↵ I	:-1 → I :Lbl 1 :I + 1 → I : If int (rand+0.07) = 0 :Goto 1 :Disp I	:0 → I :While int(rand + 0.07) = 0 :I + 1 → I :End :Disp I	:0 → i :While int(rand( ) + 0.07) = 0 :i + 1 → i :EndWhile :Disp i

b) Compléter le programme précédent afin de simuler 200 temps de désintégration et de calculer le temps moyen.

CASIO sans instruction While	CASIO avec instruction While	T.I. 80 - 81 (sans instruction While)	T.I. 82 83	T.I. 89 92
Seq(0 , I , 1 , 200 , 1) → List 1↵ For 1 → N To 200↵ -1 → I ↵ Lbl 1 ↵ I + 1 → I ↵ Int (Ran# + 0.07) = 0 ⇒ Goto 1 ↵ I → List 1[N]↵ Next↵ Mean (List 1)	Seq(0 , I , 1 , 200 , 1) → List 1↵ For 1 → N To 200↵ 0 → I ↵ While Int(Ran# + 0.07) = 0 ↵ I + 1 → I ↵ WhileEnd ↵ I → List 1[N]↵ Next↵ Mean (List 1)	:seq(0 , I , 1 , 200 , 1) → L_1 :For(N , 1 , 200) :-1 → I :Lbl 1 :I + 1 → I : If int (rand+0.07) = 0 :Goto 1 :End :I → L_1(N) :End :Disp mean(L_1)	:seq(0 , I , 1 , 200 , 1) → L_1 :For(N , 1 , 200) :0 → I :While int(rand + 0.07) = 0 :I + 1 → I :End :I → L_1(N) :End :Disp mean(L_1)	:seq(0 , i , 1 , 200 , 1) → L1 :For n , 1 , 200 :0 → i :While int(rand( ) + 0.07) = 0 :i + 1 → i :EndWhile :i → L1[n] :EndFor :Disp mean(L1)

Comparer le temps moyen $\bar{x}$ obtenu sur 200 temps simulés avec $E(T) = \frac{1}{h} = \frac{1}{0{,}07}$.

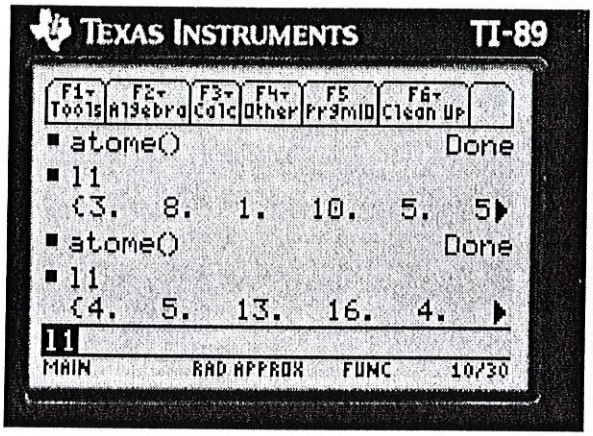
.....
.....
.....

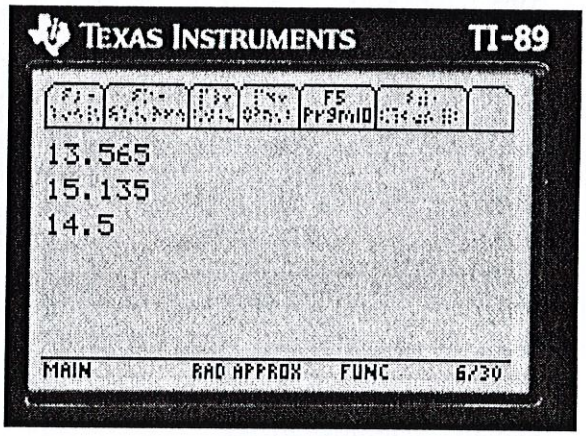
Corrigé de l'activité "Loi exponentielle de désintégration des noyaux"

1) $P(T > t)$ correspond au pourcentage (théorique) d'atomes non désintégrés au temps t.
$P(t \leq T \leq t + s)$ correspond au pourcentage (théorique) d'atomes se désintégrant durant l'intervalle de temps $[t, t + s]$.
On a ensuite $P(T > t) = 1 - P(T \leq t) = 1 - F(t)$ et

$$P(t \leq T \leq t + s) = \int_t^{t+s} f(x)dx = \int_0^{t+s} f(x)dx - \int_0^t f(x)dx = F(t + s) - F(t).$$

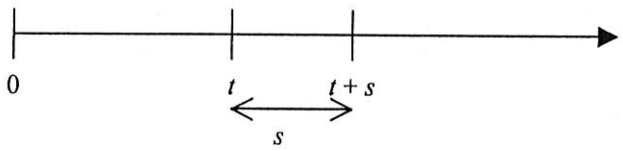
2) C'est la définition de la dérivée (on retrouve la notion de vitesse instantanée.
3) On intègre.
4) On a la condition initiale $F(0) = P(T \leq 0) = P(T = 0) = 0$.
5) On a $P(T > y) = 1 - P(T \leq y) = 1 - F(y) = e^{-hy}$.
6) On intègre par parties.
7) On trouve $E(T) = \frac{1}{h}$.
8) Les temps d'attentes sont extrêmement imprévisibles (voir premier écran ci-dessous).
Les moyennes observées sur 200 temps sont très proches de l'espérance théorique (obtenue au 6) : $1/h \approx 14{,}3$.





ENTRE NOUS ...

❑ Approche de la loi exponentielle par les probabilités conditionnelles et l'équation fonctionnelle de l'exponentielle :
C'est, par exemple, la démarche suivie dans le document d'accompagnement du nouveau programme de terminale S.



Puisque le phénomène est "sans mémoire", on a, pour tout $t \geq 0$ et tout $s \geq 0$:

$$P(T > t + s \mid T > t) = P(T > s).$$

En utilisant la définition des probabilités conditionnelles, on obtient :

$$\frac{P((T > t + s) \cap (T > t))}{P(T > t)} = P(T > s).$$

C'est à dire : $$P(T > t + s) = P(T > t) \times P(T > s).$$

On a ainsi, pour tout $t \geq 0$ et tout $s \geq 0$, $R(t + s) = R(t) \times R(s)$.
On retrouve l'équation fonctionnelle de la fonction exponentielle et on peut en déduire que $R(t) = e^{kt}$.

Cette démarche peut sembler plus élégante mais la manipulation des probabilités conditionnelles est conceptuellement plus difficile. Par ailleurs, elle occulte le rôle central joué par le taux d'avarie ("fonction de hasard" ou taux de désintégration).
La présentation précédente, par l'analyse, est davantage celle utilisée en physique (voir les manuels de physique).

❑ On pourra remarquer que la loi simulée dans le TP est une discrétisation de la loi exponentielle. Il s'agit de la ***loi géométrique*** (temps d'attente du premier succès dans un schéma de Bernoulli).
Pour une simulation directe de la loi exponentielle, voir l'exercice n° 3 ci-après (par la méthode dite de l'anamorphose, ou déformation, de la loi uniforme sur]0, 1[ ).

b) DES EXERCICES A PROPOS DE LA LOI EXPONENTIELLE (Terminale S)

Les exercices suivants peuvent compléter l'offre des manuels scolaires.
Le premier, à propos des risques d'éruption volcanique, cherche à modéliser et revient sur l'adéquation à une loi équidistribuée.
Le deuxième, inspiré de sujets de BTS industriels, mêle calculs d'analyse et application à la gestion des pannes.
Le troisième envisage la simulation (spectaculaire) de loi exponentielle de l'exercice précédent.

Exercices

LOI EXPONENTIELLE

1 ETUDE STATISTIQUE DES ERUPTIONS DU VOLCAN ASO

Le volcan *Aso*, situé sur l'île de *Kyushu* au Japon, est l'un des plus actifs au monde. On possède les statistiques de ses éruptions, régulièrement tenues depuis le XIII^{ème} siècle[2] .
Le tableau suivant fournit les années d'éruption du volcan Aso, pour chaque siècle, du XIII^{ème} au XIX^{ème}.

1229	*1286*	1377	1533	1587	*1675*	*1814*	*1884*
1239	1305	1387	1542	1598	*1683*	*1815*	*1894*
1240	1324	1388	1558	*1611*	*1691*	*1826*	*1897*
1265	1331	*1434*	1562	*1612*	1708	*1827*	
1269	1335	*1438*	1563	*1613*	1709	*1828*	
1270	1340	*1473*	1564	*1620*	1765	*1829*	
1272	1346	*1485*	1576	*1631*	1772	*1830*	
1273	1369	1505	1582	*1637*	1780	*1854*	
1274	1375	1506	1583	*1649*	*1804*	*1872*	
1281	1376	1522	1584	*1668*	*1806*	*1874*	

1) Répartition des éruptions par siècle

On souhaite savoir si l'on peut supposer que, pour chacun de ces 7 siècles, les éruptions se répartissent selon un modèle uniforme, les fluctuations entre les siècles étant dues au hasard.

a) Indiquer à quoi correspondent les notations x_i et f_i dans le tableau ci-dessous.

siècle	xi	fi	(fi - 1/7)^2
13	11	0,15068493	6,12743E-05
14	12	0,16438356	0,000463387
15	4	0,05479452	0,007755025
16	15	0,20547945	0,003921554
17	11	0,15068493	6,12743E-05
18	5	0,06849315	0,005530003
19	15	0,20547945	0,003921554

total	73	1	0,021714071

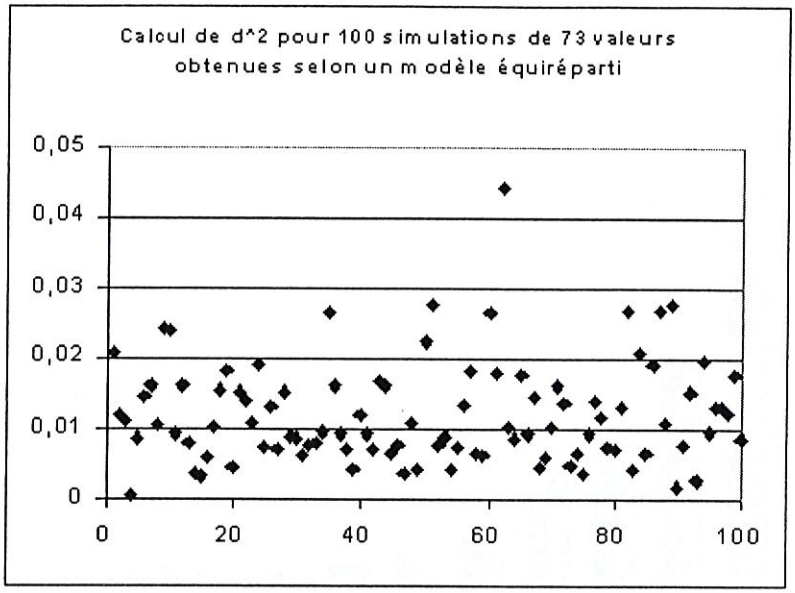
A quel calcul correspond le résultat (environ 0,02) obtenu dans la case en bas à droite du tableau ?

b) Donner un moyen de simuler, à l'aide de la calculatrice ou d'un tableur, la répartition de 73 éruptions entre 7 siècles, selon un modèle uniforme.

c) On a simulé 100 fois la répartition de 73 éruptions entre 7 siècles, selon un modèle uniforme. Pour chaque simulation, on a calculé la quantité $d^2 = \sum_{i=1}^{7} (f_i - \frac{1}{7})^2$.

Le graphique suivant fournit les résultats de ces simulations.

[2] On peut consulter www.volcanolive.com .



D'après ces simulations, peut-on rejeter, au risque de 5%, le modèle d'une loi équirépartie entre les siècles ?

2) Etude des temps d'attente entre deux éruptions

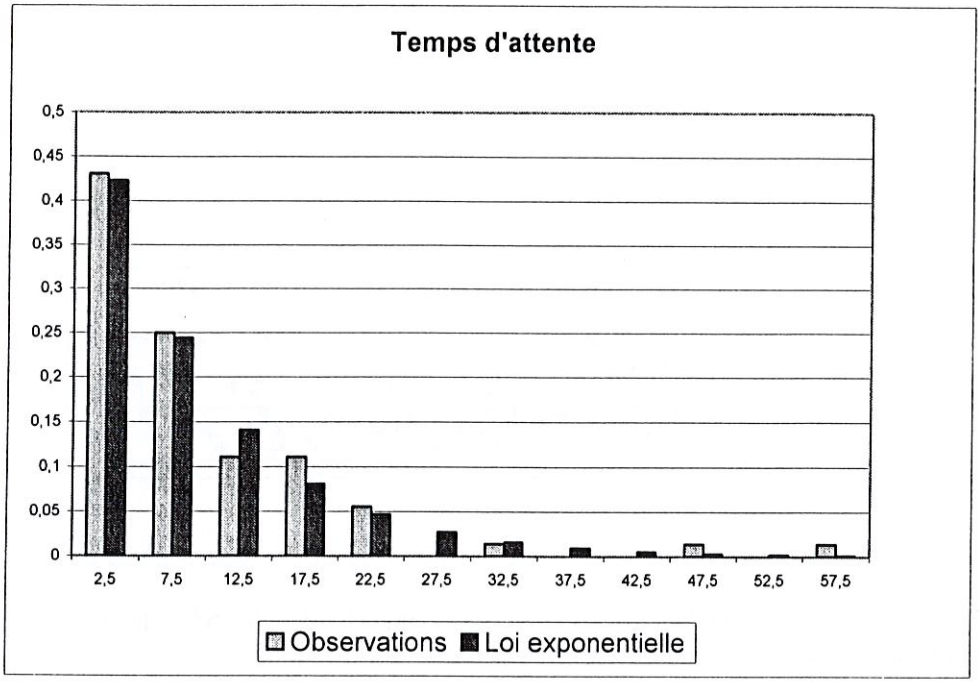
On considère la variable aléatoire T qui, à chaque éruption prise au hasard, associe le temps d'attente de la prochaine éruption.

a) On souhaite appliquer au temps d'attente le modèle d'une loi exponentielle. Sachant que la moyenne des temps d'attente observés est d'environ 9,28 années (à 10^{-2} près) et que l'on veut choisir la loi exponentielle de sorte que $E(T)$ soit environ égal à cette moyenne, quelle valeur du paramètre λ de cette loi doit-on prendre (à 10^{-2} près) ?

b) Les temps d'attentes (en années) entre deux éruptions du tableau donné au début de l'exercice, ont été regroupés en classes d'amplitude 5.

classes tps d'attente	effectifs	fréquences	probabilités
[0, 5]	31	0,43055556	0,42305019
]5, 10]	18	0,25	0,244078727
]10, 15]	8	0,11111111	0,140821175
]15, 20]	8	0,11111111	0,08124675
]20, 25]	4	0,05555556	0,046875297
]25, 30]	0	0	0,027044694
]30, 35]	1	0,01388889	0,015603431
]35, 40]	0	0	0,009002397
]40, 45]	0	0	0,005193931
]45, 50]	1	0,01388889	0,002996637
]50, 55]	0	0	0,001728909
]55, 60]	1	0,01388889	0,000997494

Justifier les calculs de $P(T \leq 5)$ et $P(5 < T \leq 10)$ figurant dans le tableau ci-dessus.



c) Calculer $P(T \geq 56)$.
Un temps de repos tel que celui qu'a connu le volcan entre 1709 et 1765 doit-il, selon le modèle exponentiel, être considéré comme exceptionnel ?

2 EN PANNE...

1) **Calcul d'intégrales**
Calculer en fonction du nombre réel positif t, les intégrales suivantes :

a) $F(t) = \frac{1}{200}\int_0^t \mathrm{e}^{-0,005x}\,\mathrm{d}x$;

b) $J(t) = \frac{1}{200}\int_0^t x\,\mathrm{e}^{-0,005x}\,\mathrm{d}x$;

c) $K(t) = \frac{1}{200}\int_0^t x^2\,\mathrm{e}^{-0,005x}\,\mathrm{d}x$.

2) **Interprétation en probabilités**

Soit la fonction f définie par : $\begin{cases} f(x) = 0 & \text{pour} \quad x < 0, \\ f(x) = \frac{1}{200}\mathrm{e}^{-0,005x} & \text{pour} \quad x = 0. \end{cases}$

a) Calculer $I = \lim\limits_{t \to +\infty} F(t)$ où F est définie au 1°.

On admet que f est la densité de probabilité d'une variable aléatoire T.

b) Calculer l'espérance mathématique $E(T) = \lim\limits_{t \to +\infty} J(t)$ de la variable aléatoire T.

c) Calculer l'espérance mathématique $E(T^2) = \lim\limits_{t \to +\infty} K(t)$ de la variable T^2.

En déduire la variance $V(T) = E(T^2) - [E(T)]^2$ et l'écart type de la variable aléatoire T.

3) **Loi exponentielle**

On s'intéresse à un type de constituant intervenant dans des robots de peinture, utilisés dans l'industrie automobile. On note T la variable aléatoire qui, à tout constituant de ce type, associe sa durée de vie en jours. Une étude statistique à montré qu'on peut considérer que T suit la loi exponentielle de paramètre $\lambda = 0{,}005$ dont la densité est la fonction f précédente.

a) Quelle est la durée de vie "moyenne" d'un tel composant ?

b) Montrer que pour tout $t \in [0, +\infty[$, $P(T > t) = e^{-0{,}005\,t}$.
Quelle est la probabilité, à 10^{-3} près, qu'un composant fonctionne correctement plus de 100 jours ?

c) Soit t et h deux nombres réels strictement positifs. On sait qu'un composant a bien fonctionné pendant h heures, calculer la probabilité conditionnelle $P_{(T \geq h)}(T \geq t + h)$ qu'il fonctionne encore correctement pendant t heures.
Pourquoi dit-on que la variable aléatoire T est "sans mémoire" ?

d) Calculer la valeur entière approchée au mieux de t_0 telle que $P(T \leq t_0) = 0{,}5$.

e) Etant donné qu'après un temps t_0 suivant sa mise en marche, la probabilité qu'un constituant soit encore en état de fonctionnement est de 50%, déterminer, pour ce temps t_0, la probabilité qu'un système de deux constituants montés en "parallèle" soit encore en état de marche, si l'on admet que les deux constituants fonctionnent de façon indépendante.

3 SIMULATION D'UNE LOI EXPONENTIELLE

1) Soit X une variable aléatoire de loi uniforme sur l'intervalle]0, 1].

a) Comment peut-on simuler, à l'aide d'une calculatrice ou d'un tableur, une série de réalisations de la variable aléatoire X ?

b) Soit a un réel de l'intervalle]0, 1], calculer $P(X \geq a)$.

2) Soit T la variable aléatoire définie par $T = -\frac{1}{\lambda} \ln X$ où λ est un réel strictement positif.
Montrer que les valeurs prises par la variable aléatoire T appartiennent à l'intervalle $[0, +\infty[$.

3) Soit t un réel de l'intervalle $[0, +\infty[$, montrer que $P(T \leq t) = 1 - e^{-\lambda t}$.

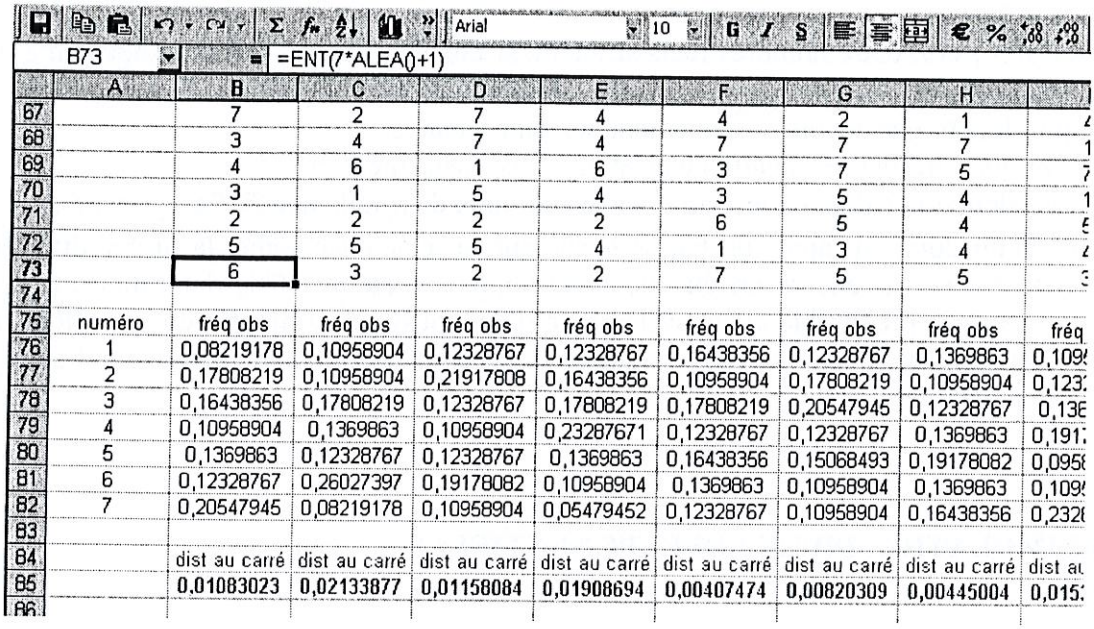
4) Déduire de la question précédente la fonction de densité f de la variable aléatoire T (on pourra montrer que, pour $t \geq 0$, $f(t) = \frac{d}{dt} P(T \leq t)$).
Quelle est la loi de T ?

5) Comment peut-on simuler, à l'aide d'une calculatrice ou d'un tableur, une série de réalisations d'une variable aléatoire de loi exponentielle de paramètre $\lambda = 0{,}005$?
Effectuer une simulation d'une telle série de 10 valeurs.

Eléments de réponse

1 LE VOLCAN ASO

1) La simulation du modèle équiréparti s'effectue en répétant 73 fois une instruction telle que ENT(7*ALEA()+1) sur Excel, ou int(7*rand+1) sur calculatrice.
Un exemple sur Excel (non demandé dans l'exercice) :



B73 : =ENT(7*ALEA()+1)

	A	B	C	D	E	F	G	H
67		7	2	7	4	4	2	1
68		3	4	7	4	7	7	7
69		4	6	1	6	3	7	5
70		3	1	5	4	3	5	4
71		2	2	2	2	6	5	4
72		5	5	5	4	1	3	4
73		6	3	2	2	7	5	5
74								
75	numéro	fréq obs	fréq obs	fréq obs	fréq obs	fréq obs	fréq obs	fréq obs
76	1	0,08219178	0,10958904	0,12328767	0,12328767	0,16438356	0,12328767	0,1369863
77	2	0,17808219	0,10958904	0,21917808	0,16438356	0,10958904	0,17808219	0,10958904
78	3	0,16438356	0,17808219	0,12328767	0,17808219	0,17808219	0,20547945	0,12328767
79	4	0,10958904	0,1369863	0,10958904	0,23287671	0,12328767	0,12328767	0,1369863
80	5	0,1369863	0,12328767	0,12328767	0,1369863	0,16438356	0,15068493	0,19178082
81	6	0,12328767	0,26027397	0,19178082	0,10958904	0,1369863	0,10958904	0,1369863
82	7	0,20547945	0,08219178	0,10958904	0,05479452	0,12328767	0,10958904	0,16438356
83								
84		dist au carré	dist au carré	dist au carré	dist au carré	dist au carré	dist au carré	dist au carré
85		0,01083023	0,02133877	0,01158084	0,01908694	0,00407474	0,00820309	0,00445004

D'après la simulation, où au moins 10% des expériences ont fourni un écart d^2 supérieur à 0,02 , l'hypothèse d'une répartition uniforme ne doit pas être rejetée (au seuil de 5%).
Remarque :

La table de la loi du khi-2 à 6 degrés de liberté donne $P\left(\sum\frac{(X_i-t_i)^2}{t_i}\leq 10{,}645\right)=0{,}90$.

On a ici $\sum\frac{(x_i-t_i)^2}{t_i}=\sum\frac{(73f_i-\frac{73}{7})^2}{\frac{73}{7}}=7\times 73d^2$ donc $d^2=\frac{10{,}645}{7\times 73}\approx 0{,}021$. On a donc $P(D^2\geq 0{,}02)>10\%$. De ce point de vue, la simulation fournie dans l'exercice est assez significative.

2) a) On prend $\lambda \approx 1 / 9{,}28 \approx 0{,}11$ à 10^{-2} près.

b) On a $P(T \leq 5) = \int_0^5 0{,}11\,e^{-0{,}11t}\,dt \approx 0{,}423$.

De même, $P(5 < T \leq 10) = \int_5^{10} 0{,}11\,e^{-0{,}11t}\,dt \approx 0{,}244$.

c) On a $P(T \geq 56) = 1 - P(T < 56) = 1 - \int_0^{56} 0{,}11\,e^{-0{,}11t}\,dt \approx 0{,}002$.

Une telle période de repos est donc, dans ce modèle, assez exceptionnelle.

Remarque : Les données des éruptions du volcan Aso au XX^{ème} siècle (que l'on peut trouver sur Internet) n'ont pas été retenues car elles s'éloignent trop du modèle exponentiel ! Il y a

une éruption presque chaque année. Le volcan a peut-être connu un regain significatif d'activité au cours de ce siècle, à moins que ce ne soit les vulcanologues. Cela pose le problème du recueil des données et de sa méthodologie (qui est absolument pas au programme des lycées). Qu'entend-t-on exactement par "éruption" ? Par ailleurs, cet exercice n'a aucune prétention scientifique. Il s'agit d'un exemple de modélisation, avec de "vraies" valeurs (mais dont on ne maîtrise pas la provenance), restant dans le cadre du programme et non d'une étude statistique dans sa globalité. Il illustre cependant bien ce que peut être la pratique de la modélisation statistique, la loi exponentielle étant, dans ce genre de situation (comme dans le cas des phénomènes de crues des rivières), le modèle le plus simple. D'autres lois, plus sophistiquées, permettront une meilleure adéquation (loi de *Rayleigh*, loi de *Gumbel*, loi de *Weibull* en fiabilité...).

2 EN PANNE...

1° a) $F(t) = \int_0^t 0{,}005\, e^{-0{,}005x}\, dx = [e^{-0{,}005\,x}]_0^t = 1 - e^{-0{,}005\,t}$.

b) $J(t) = \int_0^t 0{,}005\, x\, e^{-0{,}005x}\, dx$.

On intègre par parties :
$J(t) = [-x\, e^{-0{,}005\,x}]_0^t + 200\, F(t) = -t\, e^{-0{,}005\,t} - 200\, e^{-0{,}005\,t} + 200$.

c) $K(t) = \int_0^t 0{,}005\, x^2\, e^{-0{,}005x}\, dx = [-x^2\, e^{-0{,}005\,x}]_0^t + 400\, J(t)$,

$K(t) = -t^2\, e^{-0{,}005\,t} - 400\, t\, e^{-0{,}005\,t}\ 80000\, e^{-0{,}005\,t} + 80000$.

2° a) $l = \lim_{t \to +\infty} 1 - e^{-0{,}005\,t} = 1$.

b) $E(T) = \lim_{t \to +\infty} (-t\, e^{-0{,}005\,t} - 2\, e^{-0{,}005\,t} + 200) = 200$.

c) $E(T^2) = \lim_{t \to +\infty} (-t^2\, e^{-0{,}005\,t} - 400\, t\, e^{-0{,}005\,t} + 80000\, e^{-0{,}005t} + 80000) = 80000$.

$V(T) = E(T^2) - [E(T)]^2 = 80000 - 200^2 = 40000$, $\sigma(T) = \sqrt{V(T)} = 200$.

3° a) La durée de vie moyenne d'un composant est fournie par $E(T) = 200$ jours.
b) On a $P(T > t) = 1 - P(T \leq t) = 1 - F(t)$.
D'après ce qui précède $F(t) = 1 - e^{-0{,}005\,t}$ donc $P(T > t) = e^{-0{,}005\,t}$.
En particulier, $P(T > 100) = e^{-0{,}005 \times 100} \approx 0{,}607$.

c) On a $P_{(T \geq h)}(T \geq t + h) = \dfrac{P(T \geq t + h \text{ et } T \geq h)}{P(T \geq h)} = \dfrac{P(T \geq t + h)}{P(T \geq h)}$.

D'où $P_{(T \geq h)}(T \geq t + h) = \dfrac{e^{-0{,}005(t+h)}}{e^{-0{,}005h}} = e^{-0{,}005\,t}$.

On constate que ce résultat est indépendant de h : $P_{(T \geq h)}(T \geq t + h) = P(T \geq t)$. Autrement dit, la probabilité du temps de bon fonctionnement après h heures de fonctionnement est la même que si l'on venait de mettre en service le composant. On sait que cette propriété (pas d'effet de "vieillissement") est caractéristique de la loi exponentielle.

d) $P(T \leq t_0) = 0{,}5$ équivaut à $F(t_0) = 0{,}5$ et à $e^{-0{,}005\,t_0} = 0{,}5$ d'où :
$-0{,}005\, t_0 = \ln 0{,}5$, $t_0 = -200 \ln 0{,}5$, $t_0 \approx 139$ jours.
Remarque : cette valeur est l'analogue de la notion de médiane, à distinguer de la durée moyenne de fonctionnement (200 jours) donnée par l'espérance. On peut interpréter ce résultat en disant que, sur un très grand nombre de composants mis en marche en même temps, au bout de 139 jours, environ 50% est en panne.

e) Un système de deux éléments en parallèle est en état de marche lorsque l'un ou l'autre des éléments est en état de marche. Cette union n'étant pas facilement exploitable, considérons le cas contraire. Le système en parallèle est en panne lorsque l'un ET l'autre des composants est en panne, ce qui, en vertu de l'indépendance de ces pannes, se produit à l'instant t_0 avec la probabilité $0,5 \times 0,5 = 0,25$.
La probabilité que le système fonctionne encore après t_0 heures de marche est donc :
$1 - 0,25 = 0,75$.
Comme on l'imagine aisément, un système de deux éléments en parallèle, est plus fiable qu'un seul élément (la "fiabilité" au bout de 139 jours passe ici de 50% à 75%).

3 SIMULATION D'UNE LOI EXPONENTIELLE

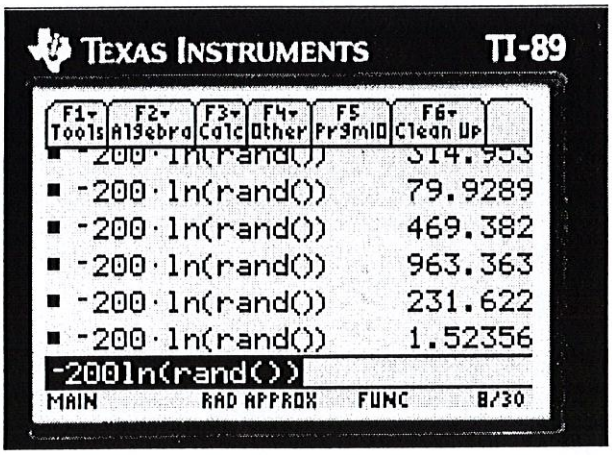
1a) On simule X en faisant rand ou Ran# sur la calculatrice, ou ALEA() sur Excel et en répétant autant de fois que désiré.

b) On a, pour tout $a \in]0, 1]$, $P(X \geq a) = \int_a^1 1 \, dx = 1 - a$.

2) Si $x \in]0, 1]$, $-(1/\lambda) \ln x \in [0, +\infty[$ donc T est à valeurs dans l'intervalle $[0, +\infty[$.
3) On a, pour tout $t \in [0, +\infty[$, $P(T \leq t) = P(-(1/\lambda) \ln X \leq t) = P(X \geq e^{-\lambda t}) = 1 - e^{-\lambda t}$ car $e^{-\lambda t} \in]0, 1]$.
4) Si $t < 0$ alors $f(t) = 0$ car T est à valeurs dans $[0, +\infty[$.

Si $t \geq 0$, $P(T \leq t) = \int_0^t f(x) \, dx = 1 - e^{-\lambda t}$ d'où $f(t) = (1 - e^{-\lambda t})' = \lambda e^{-\lambda t}$.

On reconnaît la fonction de densité de loi exponentielle de paramètre λ. Il s'agit donc de la loi de la variable aléatoire T.
5) D'après ce qui précède, on peut simuler une réalisation d'une variable aléatoire de loi exponentielle de paramètre 0,005 par l'instruction :
– ln(rand) / 0,005 c'est à dire – 200 ln(rand) ou – 200 ln(Ran#) sur calculatrice ;
ou – 200*LN(ALEA( )) sur Excel.
On a beau le savoir, les réalisations successives d'une variable aléatoire de loi exponentielle sont toujours impressionnantes (ça c'est du hasard !) :



c) DES OUVERTURES A PROPOS DE LA LOI EXPONENTIELLE

La loi exponentielle est celle du temps d'attente d'un "succès" dans un processus de Poisson

Le cadre général d'intervention de la loi exponentielle est celui des "processus de Poisson". Un processus de Poisson possède trois caractéristiques :

- il est sans mémoire,
- s'effectue à rythme contant,
- et correspond à des événements "rares".

Il en est ainsi des appels à un standard téléphonique, des arrivées à un guichet de banque, ou arrivées de véhicules, des pannes d'un matériel électronique, de certaines "catastrophes"...

Dans ces circonstances, on montre que la variable aléatoire X correspondant au nombre d'évènements ("succès") dans un intervalle de temps de longueur donnée, pris au hasard, suit une loi de *Poisson* de paramètre a. C'est à dire que la probabilité d'avoir k succès ($k \in \mathbb{N}$) pendant cet intervalle de temps est :

$P(X = k) = e^{-a} \frac{a^k}{k!}$ où a correspond à l'espérance de X, c'est à dire au nombre moyen de succès, sur la répétition d'un grand nombre de telles expériences aléatoires.

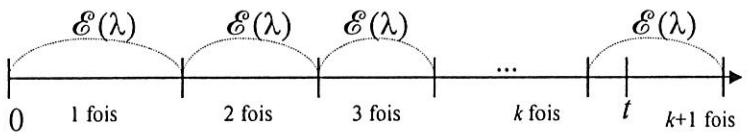
Soit t un réel strictement positif. La variable aléatoire X_t correspondant au nombre de "succès" pendant un intervalle de temps de longueur t suit une loi de Poisson dont le paramètre a est proportionnel à t. Notons $a = \lambda t$.

Désignons par T la variable aléatoire correspondant au temps d'attente du premier succès, à partir d'une origine prise au hasard.

On a $P(T > t) = P(X_t = 0) = e^{-\lambda t} \frac{(\lambda t)^0}{0!} = e^{-\lambda t}$ car on ne doit avoir aucun "succès" pendant l'intervalle $[0, t]$. Il s'en suit que la loi de T est exponentielle de paramètre λ.

En vertu du caractère "sans mémoire" du processus de Poisson, la situation est analogue à chaque nouveau succès et l'on peut dire que la loi exponentielle correspond au temps d'attente entre deux succès d'un processus de Poisson.

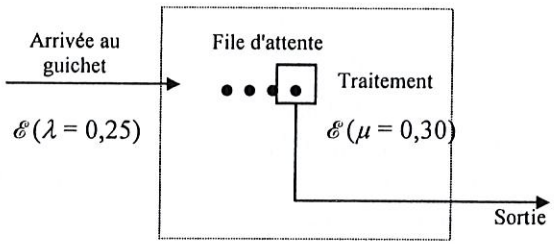
Sur le schéma ci-dessous, sur l'intervalle $[0, t]$, on a observé k succès. On a $X_t = k$ où X_t suit une loi de Poisson de paramètre $\lambda \times t$. Les temps d'attente entre chaque succès correspondent à des réalisations d'une variable aléatoire de loi exponentielle de paramètre λ.



Loi exponentielle et files d'attente

On considère ici la question de l'attente à un guichet, d'une administration par exemple (on peut remplacer cette situation par celle d'appels à un standard téléphonique, ou d'attente d'impression pour une imprimante desservant plusieurs ordinateurs en réseau…).

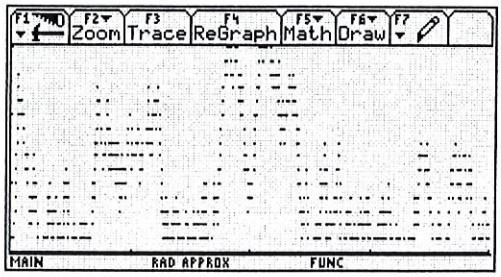
On a vu que l'on peut souvent modéliser le rythme des arrivées au guichet par une loi exponentielle.
On considère ici que le temps d'attente, en minutes, entre deux arrivées suit une loi exponentielle de paramètre $\lambda = 0{,}25$. On a donc, en moyenne une arrivée toutes les 4 minutes (l'espérance est $1/\lambda$).



De même, le temps de traitement de chaque dossier au guichet est aléatoire. On suppose qu'il suit une loi exponentielle de paramètre $\mu = 0{,}30$. Le traitement d'un dossier prend donc, en moyenne, 3,33 minutes.

On peut naïvement penser que la différence entre le temps moyen de traitement et le temps moyen entre deux arrivées permet à ce service de fonctionner dans des conditions convenables (encore qu'on a suffisamment vécu de phénomènes d'accumulation dans les files d'attente pour nourrir quelques doutes…).

Le programme suivant, après introduction des valeurs de λ et μ, simule sur un graphique, l'évolution de la longueur de la file d'attente en fonction du temps, puis affiche la longueur moyenne de la file d'attente (y compris le client en traitement) sur l'intervalle de temps [0 ; ≈2000 mn].
L'échelle en ordonnée, de 0 à 15 personnes dans la file, montre que sur cette simulation, pour une durée d'environ 33 heures (2000 minutes), on a plusieurs fois dépassé 15 personnes dans la file…



CASIO	T.I.
ViewWindow 0,2000,500,0,15,5 ↵ ? → L ↵ ? → M ↵ 0 → B ↵ 1 → Q ↵ (-ln Ran#) ÷ L → C ↵ Plot C,Q ↵ (-ln Ran#) ÷ M → R ↵ (-ln Ran#) ÷ L → A ↵ Lbl 1 ↵ A > R ⇒ Goto 3 ↵ R - A → R ↵ Lbl 2 ↵ C + A → C ↵ B + QA → B ↵ Q + 1 → Q ↵ (-ln Ran#) ÷ L → A ↵ Goto 4 ↵ Lbl 3 ↵ A - R → A ↵ C + R → C ↵ B + QR → B ↵ Q - 1 → Q ↵ (-ln Ran#) ÷ M → R ↵ Lbl 4 ↵ Plot C , Q ↵ C ≥ 2000 ⇒ Goto 5 ↵ Q = 0 ⇒ Goto 2 ↵ Goto 1 ↵ Lbl 5 ↵ B ÷ C	:0 → Xmin :2000 → Xmax :500 → Xscl :0 → Ymin :15 → Ymax :5 → Yscl :PlotsOff :ClrDraw :Input L :Input M :0 → B :1 → Q :(-ln (rand)) / L → C :Pt-On (C,Q) :(-ln (rand)) / M → R :(-ln (rand)) / L → A :Lbl 1 :If A > R :Goto 3 :R - A → R :Lbl 2 :C + A → C :B + Q×A → B :Q + 1 → Q :(-ln (rand)) / L → A :Goto 4 :Lbl 3 :A - R → A :C + R → C :B + Q×R → B :Q - 1 → Q :(-ln (rand)) / M → R :Lbl 4 :Pt-On(C , Q) :If C ≥ 2000 :Goto 5 :If Q = 0 :Goto 2 :Goto 1 :Lbl 5 :Disp B / C

BIBLIOGRAPHIE

• Eléments théoriques au niveau B.T.S. :

✍ "***Statistique et probabilités - BTS industriels***" - B. Verlant et G. Saint-Pierre - Ed. Foucher.

✍ "***Statistique et probabilités - BTS tertiaires***" - B. Verlant et G. Saint-Pierre - Ed. Foucher.

• Eléments théoriques au niveau supérieur :

✍ "***Probabilités, analyse des données et statistiques***" - G. Saporta - Ed. Technip.

✍ "***Statistique***" - Wonnacott - Ed. Economica.

✍ "***Contes et décomptes de la statistique. Une initiation par l'exemple***" – C. Robert - Ed. Vuibert 2003.

• Brochures diffusées par l'IREM Paris-Nord :

✍ "***Simulation d'expériences aléatoires de la 1ère au BTS***".

✍ "***Simulation et statistique en seconde***".

✍ "***Enseigner la statistique au lycée : des enjeux aux méthodes***".

• Histoire :

✍ "***Histoire de la statistique***" - J.-J. Droesbeke et P. Tassi - "Que sais-je ?" n° 2527 - PUF.

✍ "***La politique des grands nombres***" – A. Desrosières – La Découverte/Poche

✍ "***The lady tasting tea – How statistics revolutionized science in the twentieth century***" – David Salsburg – First Owl Books New York 2002.

TABLE DES MATIERES