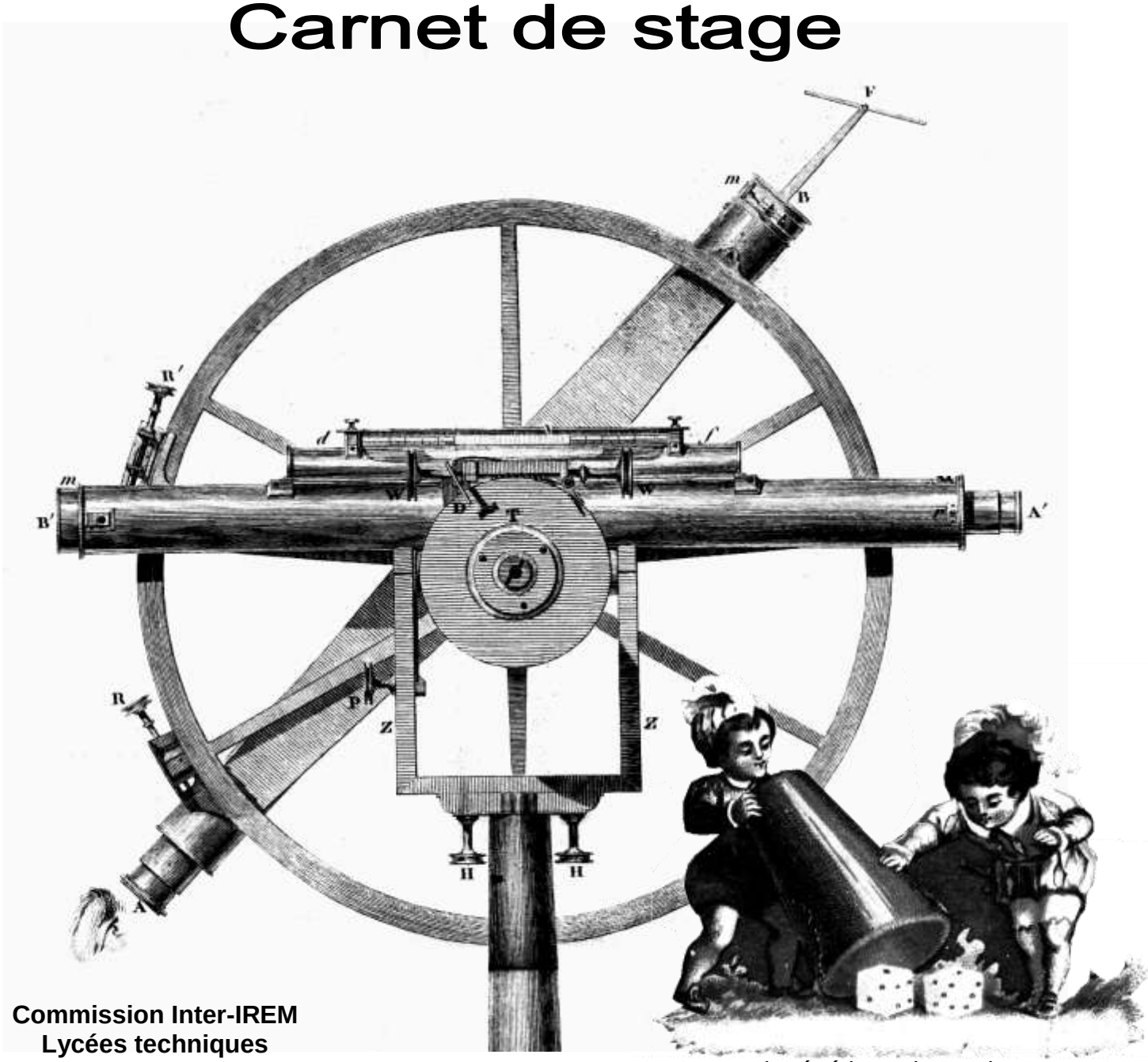
Publication n°118 de la commission Inter-IREM Lycées techniques

LA STATISTIQUE INFERENTIELLE EN QUATRE SEANCES

Carnet de stage



Le cercle répétiteur de Borda :
une solution technologique fondée sur
l'étude statistique des erreurs aléatoires

Commission Inter-IREM
Lycées techniques
Institut Galilée
99, av. J.B. Clément
93430 VILLETANEUSE

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

LA STATISTIQUE INFERENTIELLE EN QUATRE SEANCES
Carnet de stage

272 pages

Dépôt légal : 4ème trimestre 2002

AVANT-PROPOS

La statistique inférentielle 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 organisé 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 désemplis 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…

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, même si, comme plus personne ne l'ignore maintenant, les "fourchettes" ne sont pas des "intervalles de confiance"… 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.

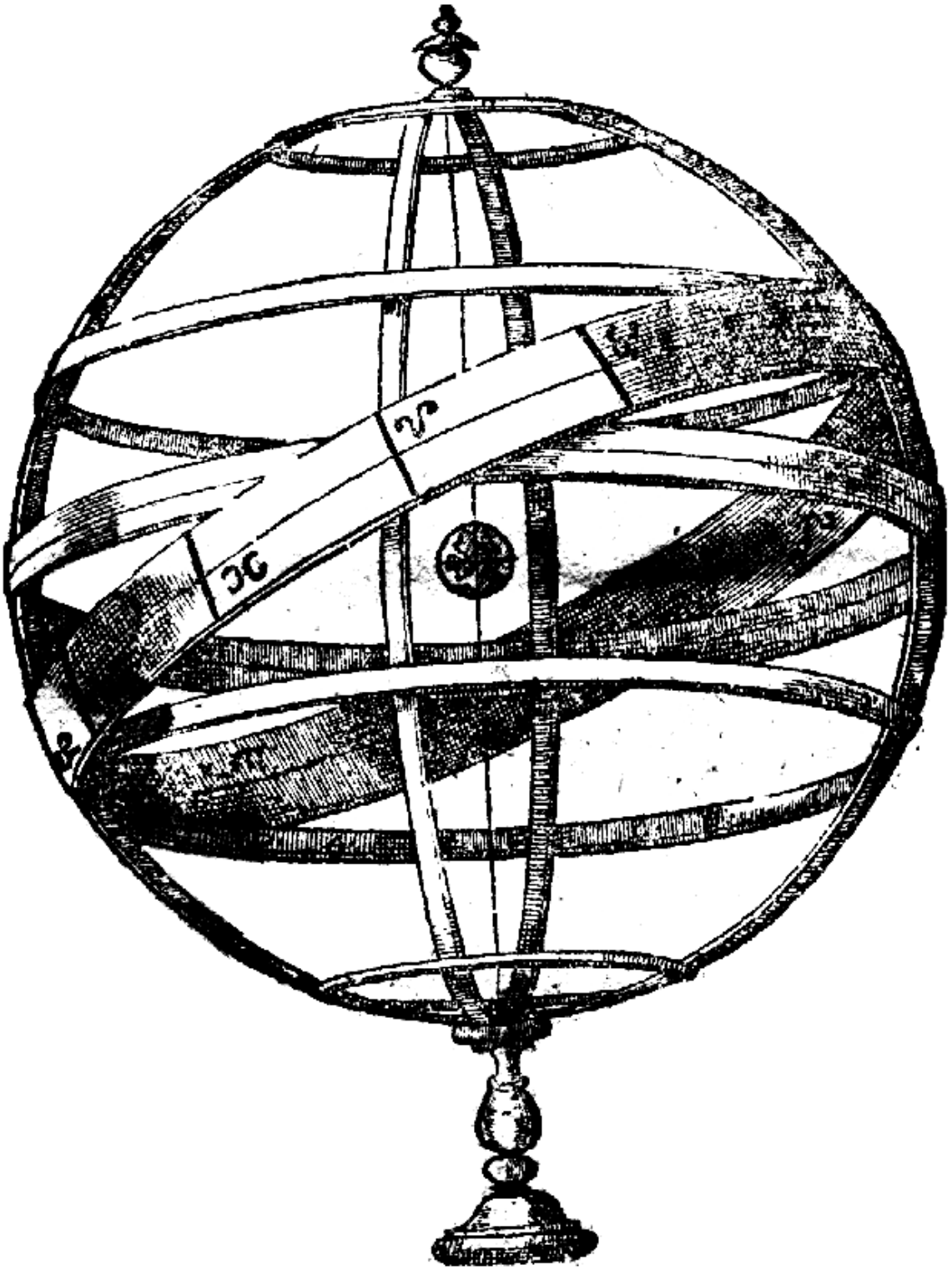
La pénurie de formateurs sur ce thème s'étendant aux séries générales, tout naturellement, nous avons mis en place des stages de statistique et probabilités sur les nouveaux programmes pour les classes de seconde, première et terminale, à Paris, pour les académies d'Ile de France.

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 techniques"



Cette brochure a été réalisée par :

Philippe DUTARTE
dutarte@club-internet.fr

Christian KERN
Lycée Alain – ALENÇON

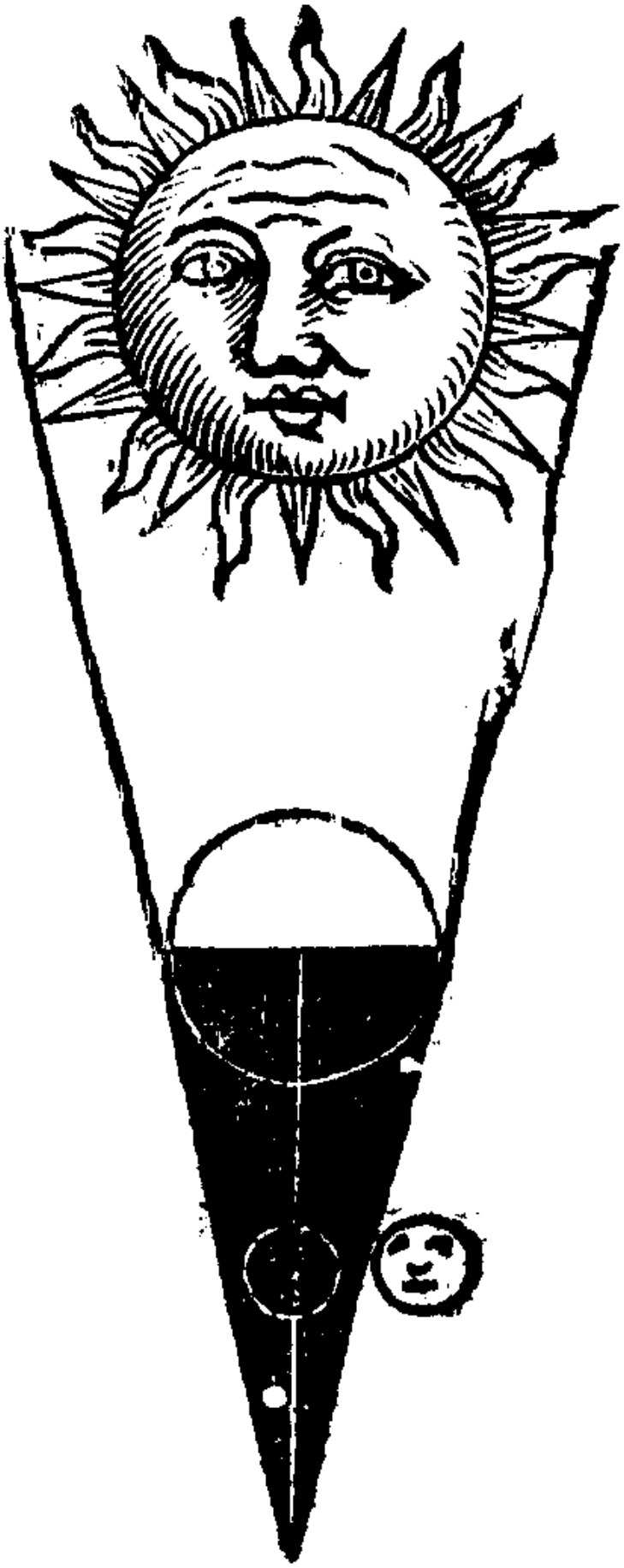
Avec la participation de :

Marie-France NOUGUES	Faculté des Sciences de Montpellier
Christine DHERS	Lycée Newton ENREA – Clichy la Garenne
Geneviève SAINT-PIERRE	IREM de Paris-Nord
Dominique ARBRE	Lycée Paul Constans – Montluçon
Françoise DELZONGLE	Lycée Eiffel – Cachan
Isabelle BRUN	Lycée Pagnol – Athis-Mons
Loïc MAZO	Lycée Follereau - Belfort

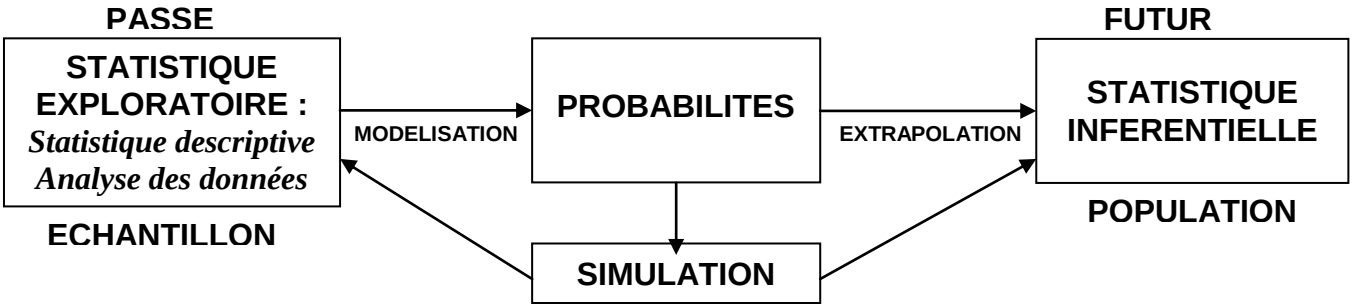
de la
Commission Inter-IREM "Lycées techniques"

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

Sommaire



QUELQUES DEFINITIONS ...



STATISTIQUE EXPLORATOIRE

Elle repose sur l'***observation*** de phénomènes concrets (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.

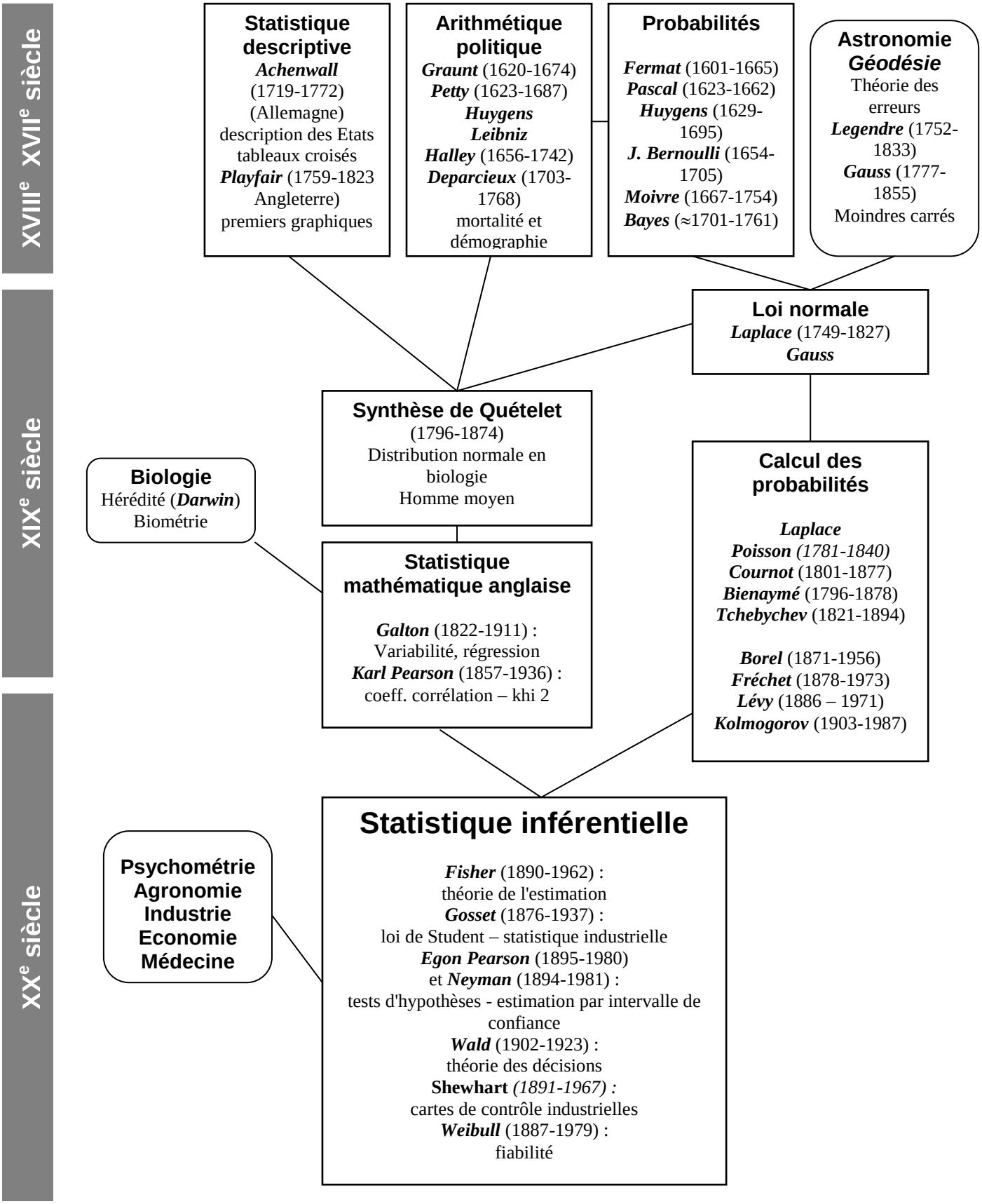
STATISTIQUE INFERENTIELLE

D'abord désignée comme "***statistique mathématique***" (parce que la théorie des probabilités y a une large place) ou "***statistique inductive***" (parce que la démarche y est souvent inductive, plutôt que déductive, avec toute l'incertitude que cela sous-tend), elle prend son essor en Angleterre au début du XX^{e} siècle, avec *Ronald A. Fisher* et *Karl Pearson*, pour répondre à des problèmes pratiques, en agronomie et en biologie.
Son but est d'***étendre*** (inférer = tirer les conséquences), ***à la population*** toute entière, les propriétés constatées sur un échantillon. Il s'agit, ayant conscience des fluctuations d'échantillonnage, 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, en particulier pour calculer les risques d'erreur.

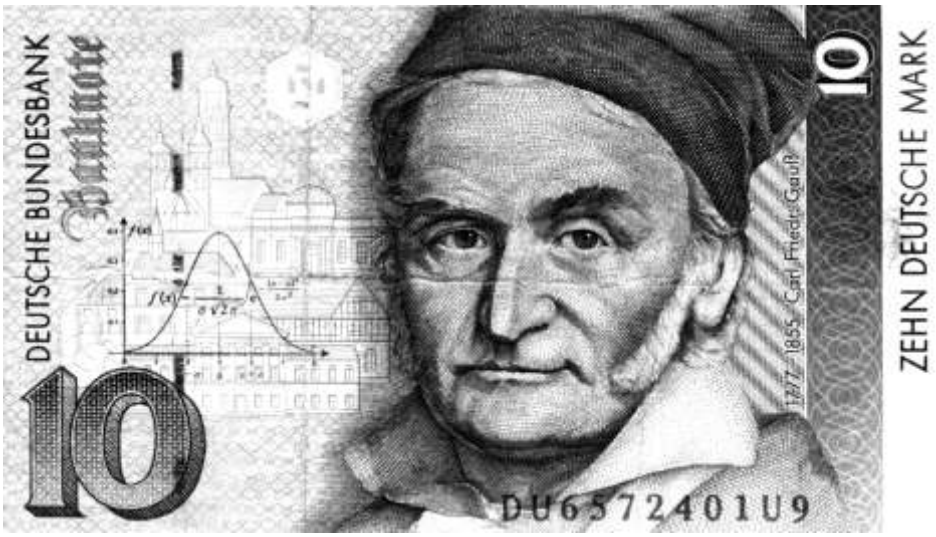
SIMULATION

La simulation est la méthode statistique permettant la reconstitution fictive de l'évolution d'un phénomène. En produisant des données, sous un certain modèle (probabiliste), la simulation 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.

ARBRE GENEALOGIQUE DE LA STATISTIQUE INFERENTIELLE



Séance 1 :
POURQUOI LA LOI DE LAPLACE-GAUSS EST-ELLE 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.

La loi de *Laplace–Gauss* joue un rôle central en statistique (on la désigne souvent comme la "*reine des statistiques*"). Il s'agit ici de voir pourquoi, et comment on peut le faire comprendre aux étudiants de B.T.S.

La statistique inférentielle dans les programmes de B.T.S. est axée sur cette loi (sauf pour ce qui concerne la fiabilité). Bien sûr ce n'est pas la seule loi utile en statistique, mais on ne peut pas, dans un horaire limité, trop embrasser, et les méthodes statistiques d'estimation ou de tests basées sur d'autres lois reposent sur des principes analogues à celles fondées sur la loi normale et faisant partie du programme. Ce sont ces méthodes qu'il s'agit de bien faire comprendre à propos du modèle "normal", en se laissant la possibilité d'envisager d'autres modèles, pour répondre aux sollicitations des matières techniques ou des études rencontrées par les étudiants en stage.

Extraits du programme de BTS 2001 :
A propos des variables aléatoires continues :
*"**L'exemple de la loi normale est suffisant**. On pourra, en vue de l'étude de la fiabilité, présenter la loi exponentielle."*
A propos de la statistique inférentielle :
"Sous l'impulsion notamment du mouvement de la qualité, les méthodes statistiques sont aujourd'hui largement utilisées dans les milieux économique, social ou professionnel."
"... l'objectif essentiel est d'initier les étudiants, sur quelques cas simples, au raisonnement et méthodes statistiques et à l'interprétation des résultats obtenus."
"Il s'agit de faire percevoir, à partir d'exemples figurant au programme, ce que sont les procédures de décision en milieu aléatoire, ainsi que leur pertinence."
"... dans le cadre de [la liaison avec les enseignements d'autres disciplines] on pourra donner quelques exemples d'autres procédures que celles figurant au programme de mathématiques (par exemple l'utilisation du test du khi 2, de la loi de Student) […] mais aucune connaissance à leur sujet n'est exigible dans le cadre de ce programme."

Avant d'étudier le rôle de la loi normale, se pose le problème pédagogique (délicat) de l'introduction des variables aléatoires continues : passage du discret au continu, notion de densité de probabilité, approximations et correction de continuité.

A - INTRODUIRE LES VARIABLES ALEATOIRES CONTINUES

Il s'agit d'introduire la notion de fonction de densité, en montrant que, dans ce cadre, les calculs, liés à la notion d'aire, sont des calculs d'intégrales. On fera en particulier remarquer que l'aire totale entre l'axe des abscisses et la courbe vaut 1, que la probabilité d'une valeur ponctuelle est nulle...

• On pourra commencer par une étude de la loi uniforme sur [0, 1], où les calculs sont simples et peuvent être reliés à une étude statistique du "random" de la calculatrice.
⇨ ***Voir le T.D. "élève" en annexe.***

Soit *X* une variable aléatoire de loi U [0 , 1]. Si son espérance $\int_0^1 x dx = \frac{1}{2}$ tombe sous le sens, il n'en est pas de même pour son écart type $\sigma(X) = \sqrt{\int_0^1 (x-\frac{1}{2})^2 dx} = \sqrt{\frac{1}{12}} \approx 0{,}29$.
On peut valider ce calcul barbare, par une expérimentation statistique de la fonction *random* de la calculatrice ou la fonction *ALEA* d'Excel.

Le programme incite au recours à la simulation :
"La réalisation de simulations dans le cadre du modèle probabiliste de référence peut fournir un éclairage intéressant."

```
prgmRANDOM
      .4740628068
      .3079930102
prgmRANDOM
      .5508297068
      .2676641947
```

La moyenne affichée $\bar{x}$ et l'écart type calculé s_e sont obtenus sur un échantillon de taille $n = 100$ extrait aléatoirement d'une population répartie selon la loi U [0, 1] de moyenne $\mu = \frac{1}{2}$ et d'écart type $\sigma \approx 0{,}29$.

On observe des fluctuations entre les différents échantillons choisis.

Savoir si $n = 100$ est suffisant pour des observations correctes, demande de connaître les lois de ces fluctuations (données par la "loi des grands nombres" et le "théorème limite central"), ce que l'on abordera en fin de séance.

U [0, 1]
$\mu = 0{,}5$
$\sigma \approx 0{,}29$

$n = 100$
$\bar{x}$; s_{100}

• On peut ensuite se placer sur le terrain de l'analyse, pour rechercher des fonctions de densités de variables aléatoires.

⇨ ***Voir le T.D. "élève" en annexe .***

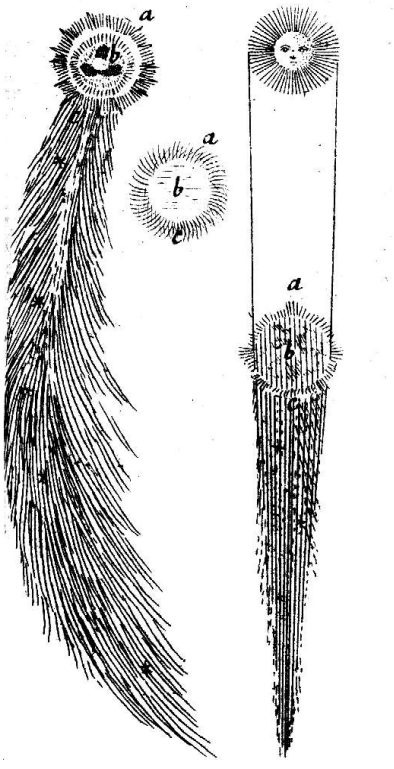
⇨ ***Voir les exercices corrigés d'épreuves de B.T.S. (exercices d'analyse).***

B – LE RÔLE 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).

I – 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 (θ, …, θ).

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 $\hat{\theta} = \frac{x_1 + ...x_n}{n}$.

• 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 $f(e_1)\, de_1 \times ... \times f(e_n)\, de_n$.
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 $f(x_1 - \theta) \times ... \times f(x_n - \theta)$ 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 $\hat{\theta} = \frac{x_1 + ...x_n}{n}$ 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 $f(x_i - \theta) = Ce^{-\frac{k}{2}(x_i - \theta)^2}$, 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 $f(e_i) = Ce^{-\frac{k}{2}e_i^2}$, alors $\prod f(e_i) = Ce^{-\frac{k}{2}\sum e_i^2}$ est maximum lorsque la somme des carrés des écarts $\sum e_i^2$ est minimale.

II – LES THEOREMES LIMITES

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

1 – 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).

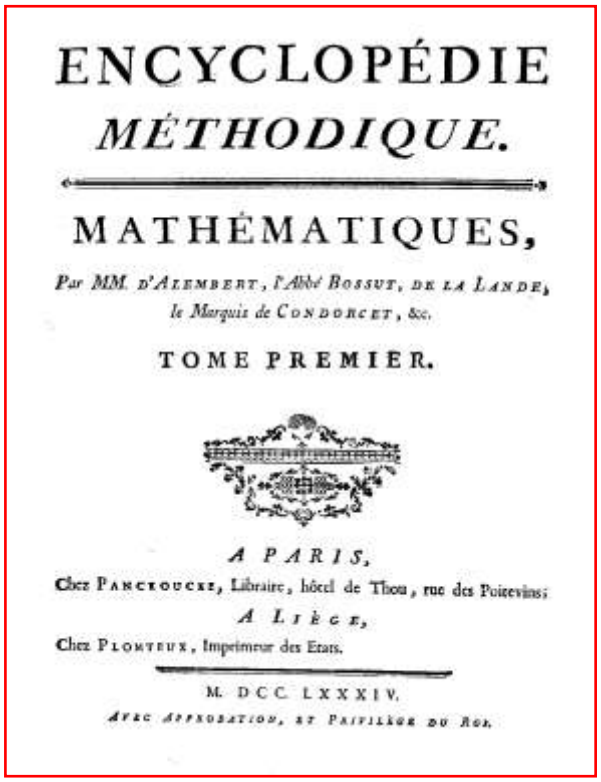
Iacob Bernoulli

Jacques Bernoulli (1654-1705)

Loi des grands nombres (énoncé "intuitif") ***:***

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.

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 Privilège du Roi.

Article PROBABILITE de l'*Encyclopédie* par ***Condorcet***.

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 des grands nombres, c'est à dire pour une convergence en probabilité) :

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

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 $t > 0$,
$P(\left|X - \mu\right| > t\sigma) \le \frac{1}{t^2}$.
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 $\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 P(X = x_i) \ge \sum_{j\ /\left|x_i - \mu\right| > t\sigma}^{n} (x_i - \mu)^2 P(X = x_i)$
et $\sum_{j\ /\left|x_i - \mu\right| > t\sigma}^{n} (x_i - \mu)^2 P(X = x_i) \ge t^2\sigma^2 \times \sum_{j\ /\left|x_i - \mu\right| > t\sigma}^{n} P(X = x_i)$.
D'où $\sigma^2 \ge t^2\ \sigma^2 \times P(\left|X - \mu\right| > t\sigma)$.

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 $\mathcal{B}$ (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[1]).
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.

[1] L'éclaircissement nous a été apporté par *Michel Henry* de l'IREM de Besançon.

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.

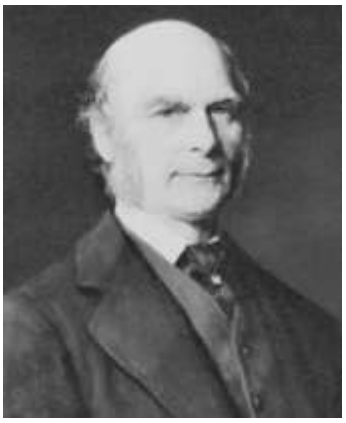
2 – Théorème de Moivre-Laplace

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 $\mathbb{N}\left(np\ ,\ \sqrt{np(1-p)}\right)$.

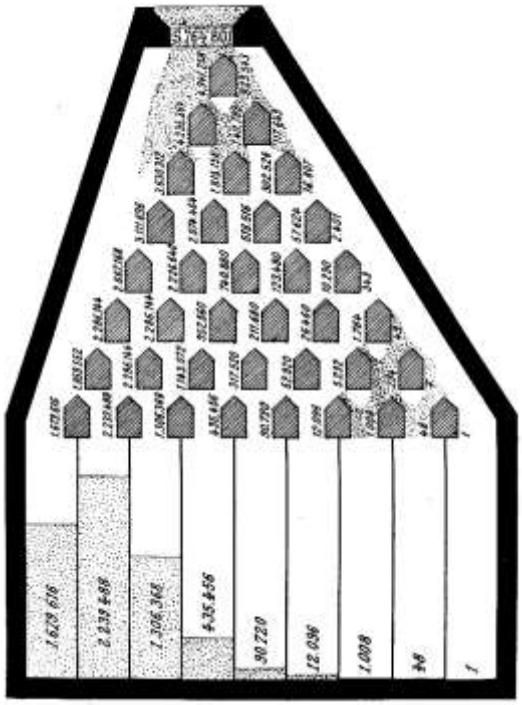
Ce théorème fournit ainsi une approximation d'une loi binomiale par la loi normale de même moyenne et même écart type.

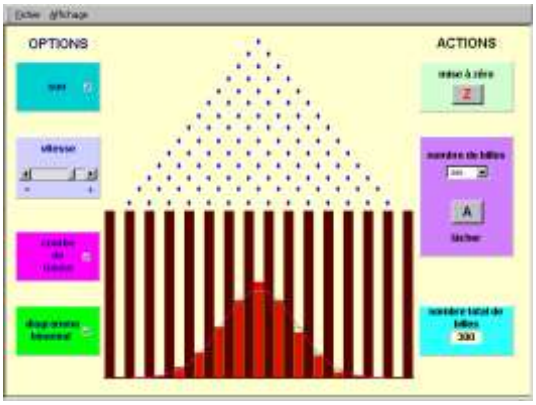
La planche de Galton


Francis Galton (1822–1911)

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 $\mathcal{B}(8 ; 1/7)$.
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*.





On a réalisé ci-contre 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.

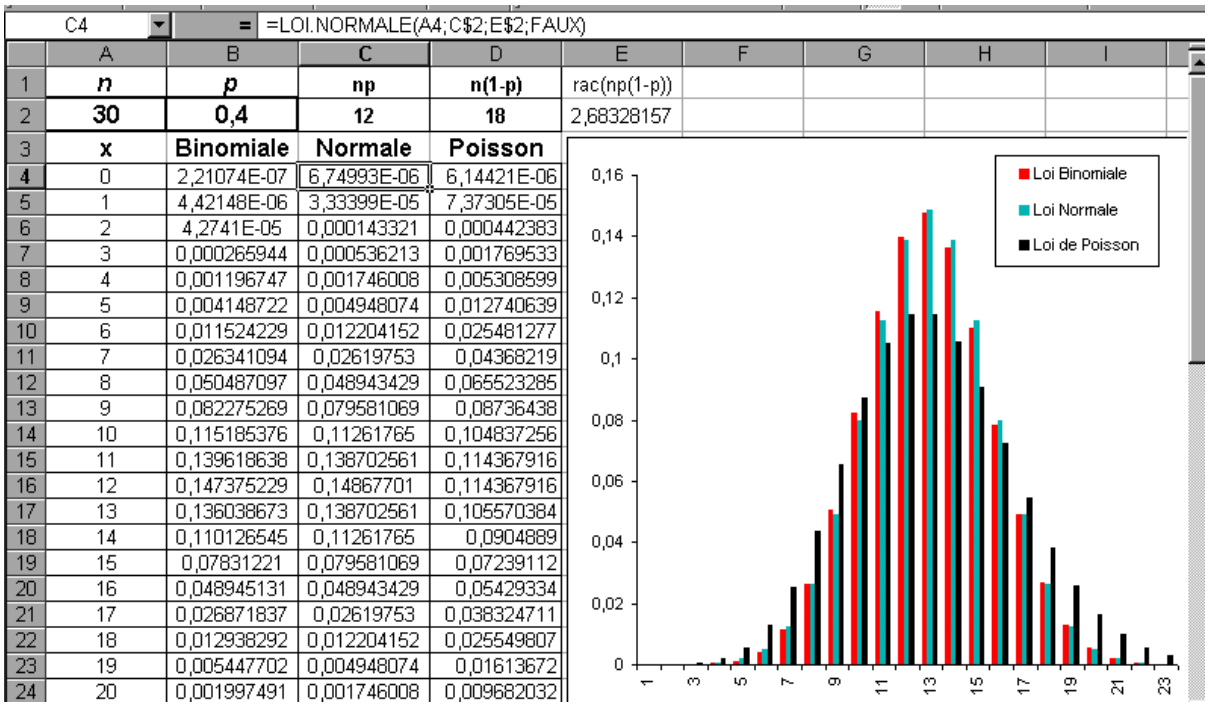
• Condition empirique d'approximation de $\mathcal{B}(n, p)$ par $\mathcal{N}(np, \sqrt{np(1-p)})$:

On donne généralement $n > 30$, np et $n(1 – p)$ supérieurs à 5 (ou parfois à 15, ou encore à 20). La convergence est d'autant plus rapide que p est plus voisin de 0,5 , on ajoute donc parfois que p ne doit être "ni trop petit, ni trop grand".

Les programmes donnent, à propos de l'approximation d'une loi binomiale par une loi de *Poisson* ou une loi normale, les indications suivantes :
"Les résultats sont admis, mais l'outil informatique peut permettre des approches expérimentales."
"Aucune connaissance sur les critères d'approximation n'est exigible dans le cadre des programmes de mathématique."
"Les étudiants doivent savoir déterminer les paramètres."
"Il conviendra de mettre en évidence la raison d'être de la correction de continuité lors de l'approximation d'une loi binomiale par une loi normale ; toutes les indications seront fournies."

⇨ ***Voir les exercices corrigés d'épreuves de B.T.S.***

Une expérimentation sur tableur est très facile à mettre en place.
L'instruction LOI.BINOMIALE(x ; n ; p ; FAUX) calcule $P(X = x)$ où X suit la loi $\mathbb{B}$ (n, p).
L'instruction LOI.NORMALE(x ; μ ; σ ; FAUX) calcule $f(x)$ où f est la densité d'une variable aléatoire Y de loi $\mathbb{N}$ (μ, σ).
L'instruction LOI.POISSON(x ; λ ; FAUX) calcule $P(Z = x)$ où Z suit la loi $\mathbb{P}$ (λ).
On peut alors présenter les calculs comme l'indique l'image d'écran suivante (on recopiera les instructions vers le bas, jusqu'à x = 50) :



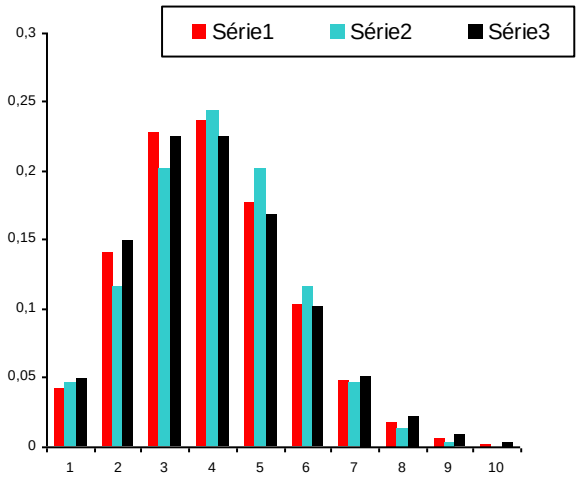
	A	B	C	D	E
1	*n*	*p*	np	n(1-p)	rac(np(1-p))
2	30	0,4	12	18	2,68328157
3	x	Binomiale	Normale	Poisson	
4	0	2,21074E-07	6,74993E-06	6,14421E-06	
5	1	4,42148E-06	3,33399E-05	7,37305E-05	
6	2	4,2741E-05	0,000143321	0,000442383	
7	3	0,000265944	0,000536213	0,001769533	
8	4	0,001196747	0,001746008	0,005308599	
9	5	0,004148722	0,004948074	0,012740639	
10	6	0,011524229	0,012204152	0,025481277	
11	7	0,026341094	0,02619753	0,04368219	
12	8	0,050487097	0,048943429	0,065523285	
13	9	0,082275269	0,079581069	0,08736438	
14	10	0,115185376	0,11261765	0,104837256	
15	11	0,139618638	0,138702561	0,114367916	
16	12	0,147375229	0,14867701	0,114367916	
17	13	0,136038673	0,138702561	0,105570384	
18	14	0,110126545	0,11261765	0,0904889	
19	15	0,07831221	0,079581069	0,07239112	
20	16	0,048945131	0,048943429	0,05429334	
21	17	0,026871837	0,02619753	0,038324711	
22	18	0,012938292	0,012204152	0,025549807	
23	19	0,005447702	0,004948074	0,01613672	
24	20	0,001997491	0,001746008	0,009682032	

On représente sous forme d'histogramme les trois colonnes B, C, D. Le pas des x étant de 1, les valeurs données par la densité de la loi normale correspondent approximativement à la probabilité $P(x - 0{,}5 \leq Y \leq x + 0{,}5)$.
Sur l'image précédente, on a n = 30 et p = 0,4. L'approximation normale est assez satisfaisante, alors que les résultats donnés par la loi de *Poisson* de paramètre np = 12 (bâtons en 3ème position) sont trop éloignés.

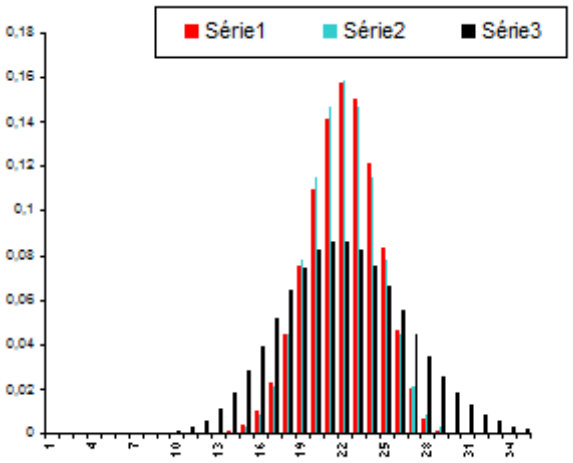
Il suffit de modifier les valeurs de n et p dans les cellules A2 et B2 pour voir évoluer les graphiques et expérimenter "à vu d'œil" les conditions d'approximation (on peut ajuster l'échelle en cliquant dans la zone de graphique puis en ne sélectionnant qu'une partie des trois séries représentées).

Voir les quelques exemples qui suivent.

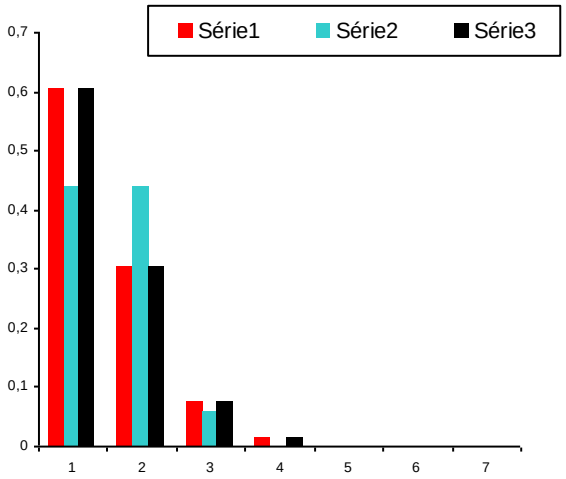
Les conditions d'approximation sont cependant très empiriques et dépendent de la précision nécessaire à chaque domaine d'utilisation. Comme on va le voir plus loin, l'approximation peut-être assez convenable dans un certain secteur de l'histogramme et beaucoup plus défectueuse ailleurs.



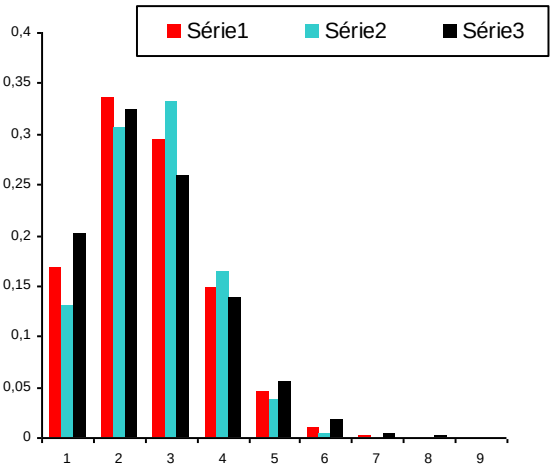
$n = 30$; $p = 0,1$
La loi de *Poisson* (série 3) est globalement préférable pour approcher la loi binomiale (série 1)



$n = 30$; $p = 0,7$
La loi Normale (série 2) peut approcher la loi binomiale (série 1). La loi de *Poisson* est totalement inadaptée



$n = 50$; $p = 0,01$
La loi de Poisson (série 3) peut approcher la loi binomiale (série 1). La loi Normale est totalement inadaptée



$n = 8$; $p = 0,20$
Ni la loi normale, ni la loi de *Poisson* ne sont réellement adaptées (pour n petit il n'est guère utile d'approcher la loi binomiale).

• *Correction de continuité :*

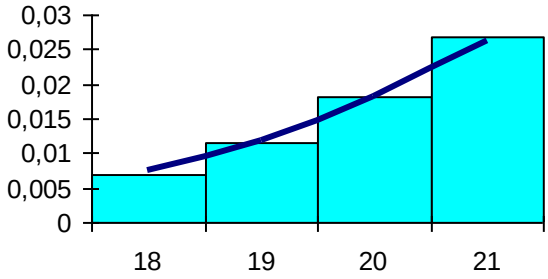
L'approximation de X suivant la loi binomiale $\mathcal{B}(n\ ,\ p)$ par Y de loi normale $\mathcal{N}(np\ ,\ \sqrt{np(1-p)})$ pourra être améliorée par une ***"correction de continuité"*** correspondant au passage du discret au continu :

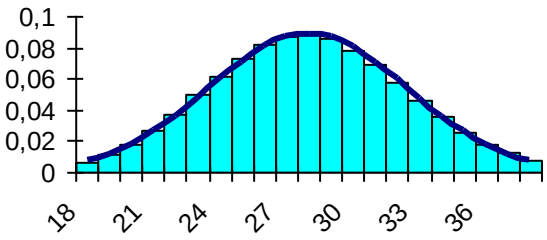
$\boxed{P(X = k) \approx P(k - 0,5 \leq Y \leq k + 0,5)}$.

Envisageons deux exemples posés au B.T.S.

Exemple 1 : (d'après BTS informatique de gestion 1999)

Une usine fabrique des objets d'un certain type.
La probabilité de l'événement E : "l'objet est de deuxième choix" est $P(E) = 0{,}28$.
On prélève un échantillon de 100 objets, pris au hasard et avec remise dans la production.
On note Y la variable aléatoire qui, à chaque échantillon de ce type, associe le nombre d'objets de deuxième choix dans cet échantillon.
1) Quelle est la loi suivie par Y ?
2) On approche la loi de Y par une loi normale. Quels sont les paramètres de cette loi normale ?
3) On veut calculer le nombre $\alpha = P(Y \geq 20)$. Expliquer pourquoi le calcul de α, par utilisation directe de la loi de Y , est long.
On note Z une variable aléatoire de loi $\mathcal{N}$ (28 ; 4,49) et l'on admet que $P(Z \geq 19{,}5)$ est une approximation satisfaisante de α. Calculer $P(Z \geq 19{,}5)$.





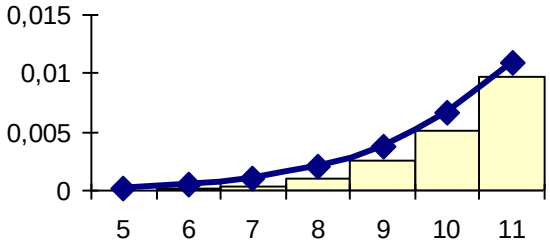
Comparaison de
Y de loi $\mathcal{B}$ (100 ; 0,28)
et Z de loi $\mathcal{N}$ (28 ; 4,49)

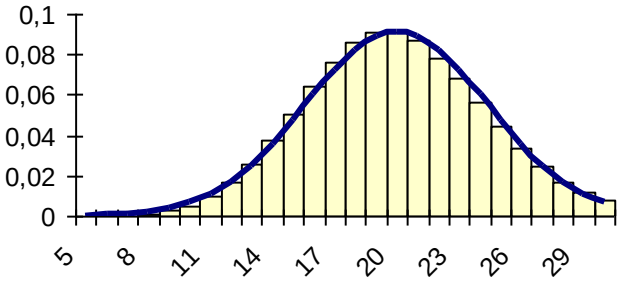
	AVEC	SANS
P(Y >= 20)	P(Z >= 19,5)	P(Z >= 20)
0,97410863	0,9708275	0,962604
	CORRECTION	CORRECTION

Exemple 2 : (BTS informatique de gestion 1995)

Dans cet énoncé d'examen on approche la variable aléatoire Y de loi $\mathcal{B}$ (400 ; 0,05) par la variable aléatoire Z de loi $\mathcal{N}$ (20 , $\sqrt{19}$), ce qui est raisonnable.
L'énoncé demande donc d'approcher $P(Y \leq 10)$ par $P(Z \leq 10{,}5)$, ce qui correspond à la correction de continuité.

Dans la zone du calcul, la dissymétrie de la loi binomiale va à l'encontre de la correction de continuité : pas de chance, $P(Z \leq 10)$ aurait donné un meilleur résultat.



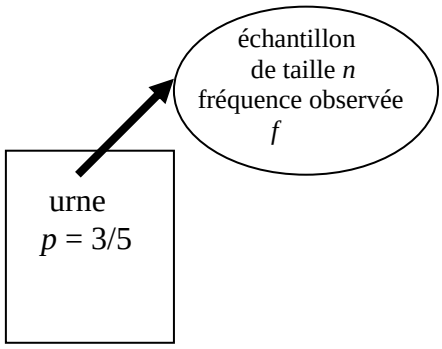


Comparaison de
Y de loi $\mathcal{B}$ (400 ; 0,05)
et de Z de loi $\mathcal{N}$ (20 ; $\sqrt{19}$)

	AVEC	SANS
P(Y <= 10)	P(Z <= 10,5)	P(Z <= 10)
0,0093986	0,01466953	0,01090735
	CORRECTION	CORRECTION

⇨ ***Voir autres énoncés de B.T.S.***

• ***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 $\mathrm{N}\left(p , \sqrt{\frac{p(1-p)}{n}} \right)$.

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 $\mathrm{N}\left(\frac{3}{5} , \frac{\sqrt{6}}{5\sqrt{n}} \right)$.

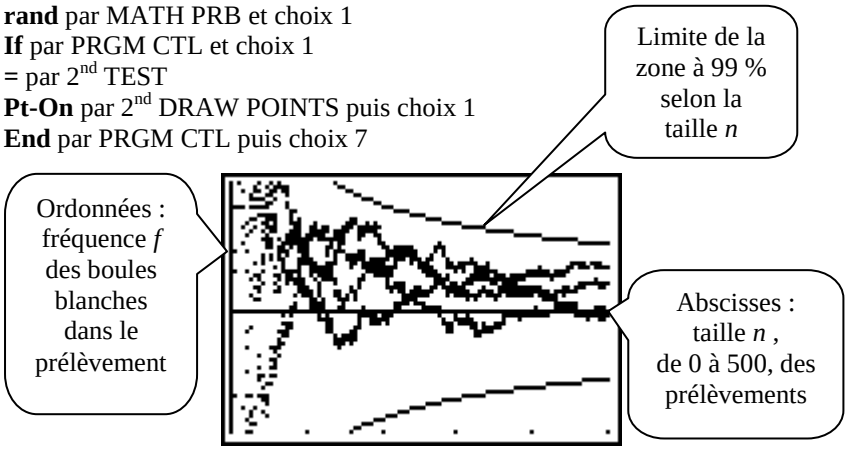
1) Calculons, en fonction de n, le réel positif h tel que $P(\frac{3}{5} - h \le F \le \frac{3}{5} + h) = 0{,}99$. On détermine ainsi un intervalle autour de 3/5 dans lequel la variable aléatoire F prend ses valeurs dans 99 % des cas.

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

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

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 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 : $P\left(\left|F - \frac{3}{5}\right| > 0{,}02\right) = \frac{1}{1000}$.

Ces calculs ne sont pas à poser aux élèves !!

On veut $P\left(|T| \leq 0{,}02\frac{5\sqrt{n}}{\sqrt{6}}\right) = \frac{999}{1000}$. 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 $t = \Pi^{-1}(\frac{1999}{2000})$. 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*).

3 – Théorème limite central

Laplace va plus loin que *Gauss* en montrant, en 1810, que sa *"seconde loi des erreurs"* [2] (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

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

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* $N\left(\mu\ ,\ \frac{\sigma}{\sqrt{n}}\right)$.

• 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 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*".

• ***Pour n petit*** ($n < 30$), l'approximation par la loi de ***Student*** est préférable. On y reviendra plus loin.

• En revanche, ***si les X_i , indépendantes, suivent la même loi normale N (μ , σ)***, alors $\overline{X}_n$ suit *exactement* la loi $N\left(\mu\ ,\ \frac{\sigma}{\sqrt{n}}\right)$ même pour *n* petit.

• ***L'écart type*** $\boxed{\frac{\sigma}{\sqrt{n}}}$ s'explique facilement et n'est pas spécifique au cadre gaussien. En raison de l'indépendance des X_i , on a :
$V\left(\overline{X}_n\right) = V\left(\frac{1}{n}\sum_{i=1}^{i=n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{i=n} V(X_i) = \frac{1}{n^2} n \times \sigma^2 = \frac{\sigma^2}{n}$ d'où $\sigma\left(\overline{X}_n\right) = \frac{\sigma}{\sqrt{n}}$.

• *Cas des fluctuations des moyennes d'échantillon*

D'après le T.L.C., les *moyennes des échantillons* se répartissent approximativement selon la loi normale $N\left(\mu\ ,\ \frac{\sigma}{\sqrt{n}}\right)$ (**voir la séance n° 2** de ce stage, à propos "d'échantillonnage" où l'on reviendra sur l'interprétation du T.L.C. en termes d'échantillonnage).

Sur l'exemple d'une population répartie selon la loi uniforme U [0, 1] (voir au début de cette séance), on peut comparer les observations sur des simulations avec les résultats théoriques. On complètera les colonnes "observée" à l'aide de plusieurs échantillons de taille $n = 100$ simulés par la calculatrice.

Les résultats des cases en ***gras*** *sont* ***hors-programme B.T.S.***

Distributions d'échantillonnage	Variables aléatoires	Moyenne (sur plusieurs échantillons de taille n)		Ecart type (sur plusieurs échantillons de taille n)	
		théorique (espérance)	observée	théorique	observé
Moyennes des échantillons de taille n	$\overline{X} = \frac{1}{n}\sum X_i$	$\mu = 0{,}5$ (T.L.C.)		$\frac{\sigma}{\sqrt{n}} \approx 0{,}029$ (T.L.C.)	
Variances des échantillons de taille n	$S^2 = \frac{1}{n}\sum (X_i - \overline{X})^2$	$\frac{n-1}{n}\sigma^2 = \frac{99}{100\times 12} = 0{,}0825$ (voir estimation ponctuelle de l'écart type - séance n° 2)		$\sqrt{\frac{\mu_4 - \sigma^4}{n}} \approx 0{,}0075$	

Moyenne des variances obtenues sur les différents échantillons

Indicateurs des fluctuations de la moyenne et de la variance, entre plusieurs échantillons de taille n

Le moment μ_4 d'ordre 4 vaut, dans le cas de l'exemple : $\mu_4 = E\left[\left(X - E(X)\right)^4\right] = \int_0^1 (x - \frac{1}{2})^4 \, dx = \frac{1}{80}$.

Lorsque les variables aléatoires X_i sont normales, S^2 suit la loi du $\boxed{\chi^2_{n-1}}$, de moyenne $n - 1$ et d'écart type $2(n - 1)$. Pour ces questions (hors-programme), consulter "*Saporta*" p.269 à p.275.

On reviendra sur l'échantillonnage dans la 2ème séance.

• *Sommes de variables aléatoires indépendantes*

Dans le cas des sommes de variables aléatoires indépendantes, il est important de noter que $\boxed{X + X \neq 2 \times X}$.

En particulier, si $E(X + X) = 2\times E(X) = E(2\times X)$, en revanche, $V(X + X) = 2\times V(X)$ alors que $V(2\times X) = E(4X^2) - 4\,E^2(X) = 4\,V(X)$.

⇨ ***voir exercices corrigé de B.T.S.***

• *Pourquoi la loi "normale" ?*

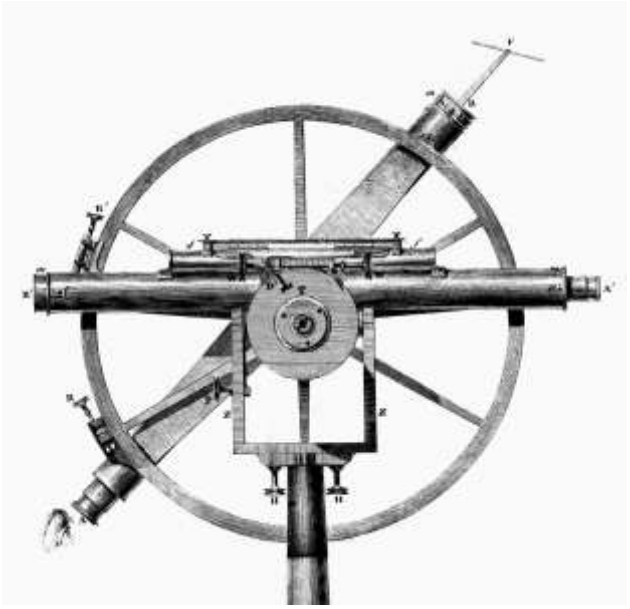
Le théorème limite central explique ainsi l'apparition de la loi normale dans l'étude de fluctuations dues à l'*addition* de nombreux facteurs aléatoires indépendants.

De façon générale, la loi normale modélisera les situations aléatoires possédant de nombreuses causes indépendantes dont les effets s'*ajoutent*, sans que l'un d'eux soit prépondérant (qualité d'une production industrielle, erreurs de mesure, gestion de comptes bancaires...).

C'est ainsi qu'au XIX^e siècle le statisticien belge *A. Quételet*, puis l'anglais *Galton* retrouvèrent expérimentalement cette distribution normale dans quantité de domaines. Ce qui contribua à attribuer un rôle prépondérant à la moyenne (sous l'hypothèse d'une distribution gaussienne, il s'agit du meilleur estimateur).

Le cercle répétiteur de *Borda*, construit par *Lenoir* (visible au musée du CNAM à Paris) permit ainsi, à la fin du XVIII^e et au tout début du XIX^e siècle, de faire un gain important en précision dans les mesures géodésiques. Il favorisa la mesure de la méridienne de

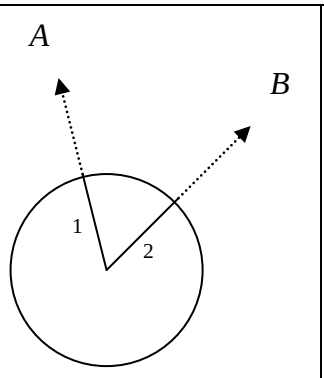
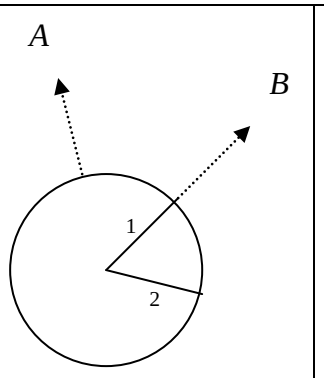
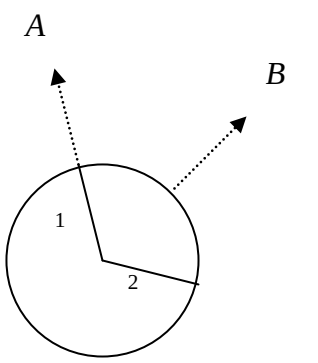
France en vue de la définition du mètre. Il s'agit d'un instrument mettant concrètement en œuvre le théorème de *Laplace*.



Après avoir déterminé les erreurs systématiques produites notamment par l'instrument, demeurent, dans la mesure des angles, des erreurs fortuites que l'on considérera comme aléatoires. Ces erreurs résultent de l'addition de nombreux facteurs (habileté du géomètre, conditions atmosphériques, hydrométrie, chaleur, jeu ou mouvement de l'instrument…), sans qu'aucun soit prépondérant. On vérifie ainsi expérimentalement qu'elles se distribuent selon un loi normale.

Dans ces conditions, une estimation optimale de l'angle consistera à effectuer la moyenne de *n* mesures. C'est ce que permet mécaniquement l'instrument. On sait que le gain en précision, mesuré par l'écart type entre les moyennes, est multiplié par un facteur $\frac{1}{\sqrt{n}}$, du moins en supposant indépendantes les mesures successives, ce qui n'est pas acquis (il faudrait en particulier que l'origine de chaque mesure soit rigoureusement la division finale de la mesure précédente, ce qui est rarement réalisé en raison du jeu des axes[3]).

La mesure se déroule ainsi :

		
Etape 1	Etape 2	Etape 3
Les lunettes 1 et 2 étant indépendantes (plateaux débrayés), on vise les points *A* et *B*.	Les lunettes 1 et 2 sont solidaires (plateaux embrayés), on vise alors le point *B* avec la lunette 1.	Les lunettes 1 et 2 sont débrayées, on vise alors le point *A* avec la lunette 1. La première mesure a été "mémorisée" et la mesure actuelle est la somme de deux mesures de l'angle. Il suffit de réitérer le procédé.

[3] Nous remercions M.F. Jozeau, de l'IREM de Paris 7, de nous avoir signalé ce problème.

Malgré le rôle capital qu'elle joue en statistique, la loi normale est loin de décrire tous les phénomènes et il ne faut pas considérer comme "*anormale*" une variable aléatoire qui ne suit pas une loi de *Laplace-Gauss*.
Le *"modèle normal"* sera validé par des tests comme celui de la ***droite de Henry*** présenté ci-dessous (il existe des tests plus sophistiqués, comme celui de *Kolmogorov*, présenté à la fin de la séance 4, où l'on quantifie les risques d'erreur).

III - AJUSTEMENT A UNE DISTRIBUTION NORMALE : DROITE DE HENRY

Programme 2001 de BTS :
"En liaison avec les enseignements d'autres disciplines, on pourra donner quelques exemples d'autres procédures que celles figurant au programme de mathématiques (par exemple utilisation de la droite de Henry)".

1 - Le principe

On souhaite utiliser la technique de la régression linéaire selon les moindres carrés pour quantifier la qualité de l'ajustement d'une distribution statistique observée avec celle d'une loi normale.

• Du point de vue des *probabilités* :

On désigne par X une variable aléatoire suivant la loi $\mathcal{N}(\mu ; \sigma)$.

Sa fonction de répartition F est donnée, pour tout $x \in \mathbb{R}$, par :

$y = F(x) = P(X \leq x) = P(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}) = P(\frac{X-\mu}{\sigma} \leq t) = \Pi(t)$, avec $t = \frac{x-\mu}{\sigma}$ et où Π est la fonction de répartition de la loi $\mathcal{N}(0, 1)$, tabulée dans le formulaire officiel.

• *Statistiquement*, l'analogue de la fonction de répartition est la fréquence cumulée : pour une valeur x_i de la distribution statistique, y_i est la *fréquence cumulée croissante*.

Notons t_i la valeur, donnée par *lecture inverse* de la table de la loi $\mathcal{N}(0 ; 1)$, telle que $y_i = \Pi(t_i) \Leftrightarrow t_i = \Pi^{-1}(y_i)$.

Si la distribution statistique est extraite d'une population normale, le nuage de points $(x_i ; t_i)$ devrait donc être ajusté par la droite d'équation $t = \frac{x-\mu}{\sigma}$, que l'on nomme **droite de *Henry*** (lequel a appliqué ce procédé lors de l'étude de la précision des tirs d'artillerie).

Historique statistique	**Modèle probabiliste**
Série d'observations x_i	Variable aléatoire X
Fréquences cumulées y_i	Fonction de répartition F
Série des valeurs $t_i = \Pi^{-1}(y_i)$ $t_i \approx \frac{x_i - \mu}{\sigma}$ (ajustement linéaire)	$T = \frac{X-\mu}{\sigma}$, de fonction de répartition Π, suit la loi $\mathcal{N}(0, 1)$

2 - Un exemple

La durée de vie, exprimée en heures, des joints à lèvres PLAUSTRA - type IE - définit une variable aléatoire continue X.

L'étude de la durée de vie de 500 de ces joints a permis d'obtenir l'historique suivant.

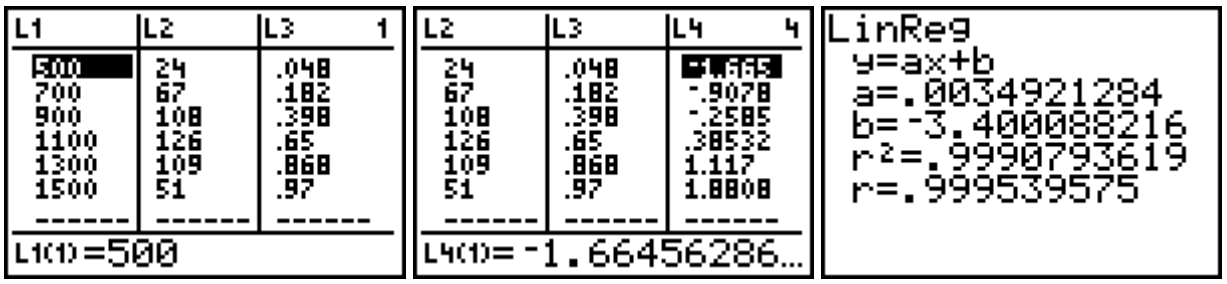
Temps de bon fonctionnement x_i	500	700	900	1100	1300	1500	1700
Effectifs n_i	24	67	108	126	109	51	15

a - Sur calculatrice

Sur TI 83, on peut procéder ainsi :

Opérations	Procédure	Accès aux fonctions
Effacer les listes.	ClrList L_1 , L_2 , L_3 , L_4	**ClrList** par STAT EDIT choix 4 **L_1** au clavier par 2^{nd}
Entrer en liste L_1 les temps de bon fonctionnement x_i (**sauf le dernier qui donnerait une fréquence 1**). Entrer en liste L_2 les effectifs n_i (sauf le dernier).	STAT EDIT choix 1 Edit... puis ENTER Entrer les 6 nombres dans chacune des colonnes L_1 et L_2.	
Quitter l'éditeur de listes	2^{nd} QUIT	
Calculer les fréquences cumulées y_i (en liste L_3). Calculer les valeurs $t_i = \Pi^{-1}(y_i)$ correspondantes (en liste L_4).	cumSum(L_2)/500 → L_3 ENTER seq(invNorm(L_3(X)),X,1,6) → L_4 ENTER	**cumSum** par 2^{nd} LIST OPS puis choix 6 → par la touche STO ▸ **seq** par 2^{nd} LIST OPS puis choix 5 **invNorm** par 2^{nd} DISTR choix 3
Faire la régression linéaire sur les listes L_1 et L_4.	LinReg(ax+b) L_1 , L_4 ENTER	**LinReg(ax+b)** par STAT CALC choix 4

On obtient les résultats suivants :



La corrélation $r \approx 0{,}9995$ justifie l'ajustement des durées de vie à une distribution normale.

On a $\frac{1}{\sigma} \approx 0{,}0035$ d'où σ estimé à 285,7 heures.

Puis $\frac{\mu}{\sigma} \approx 3{,}4001$ qui conduit à estimer μ à 971,5 heures.

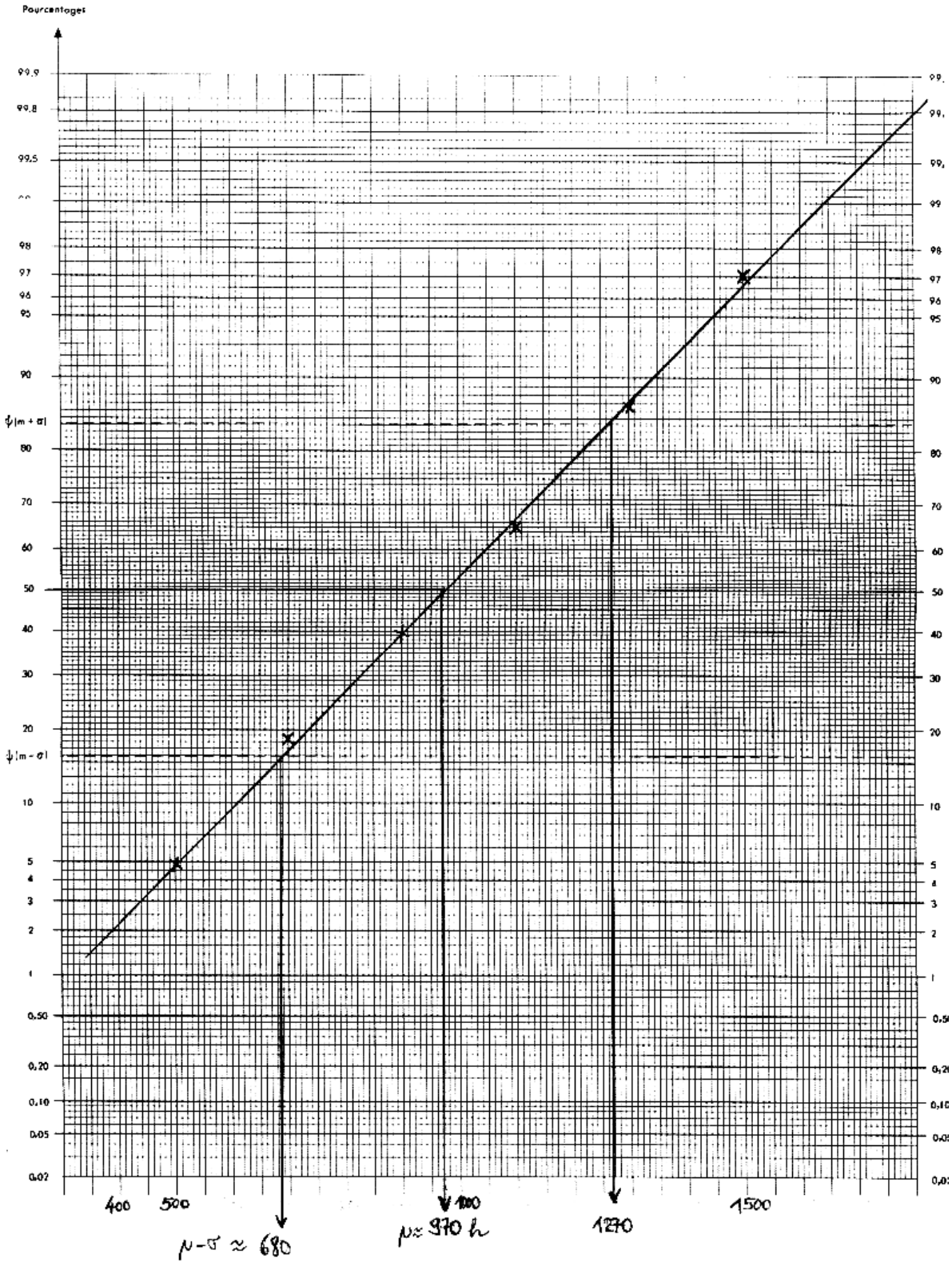
b - Sur ordinateur

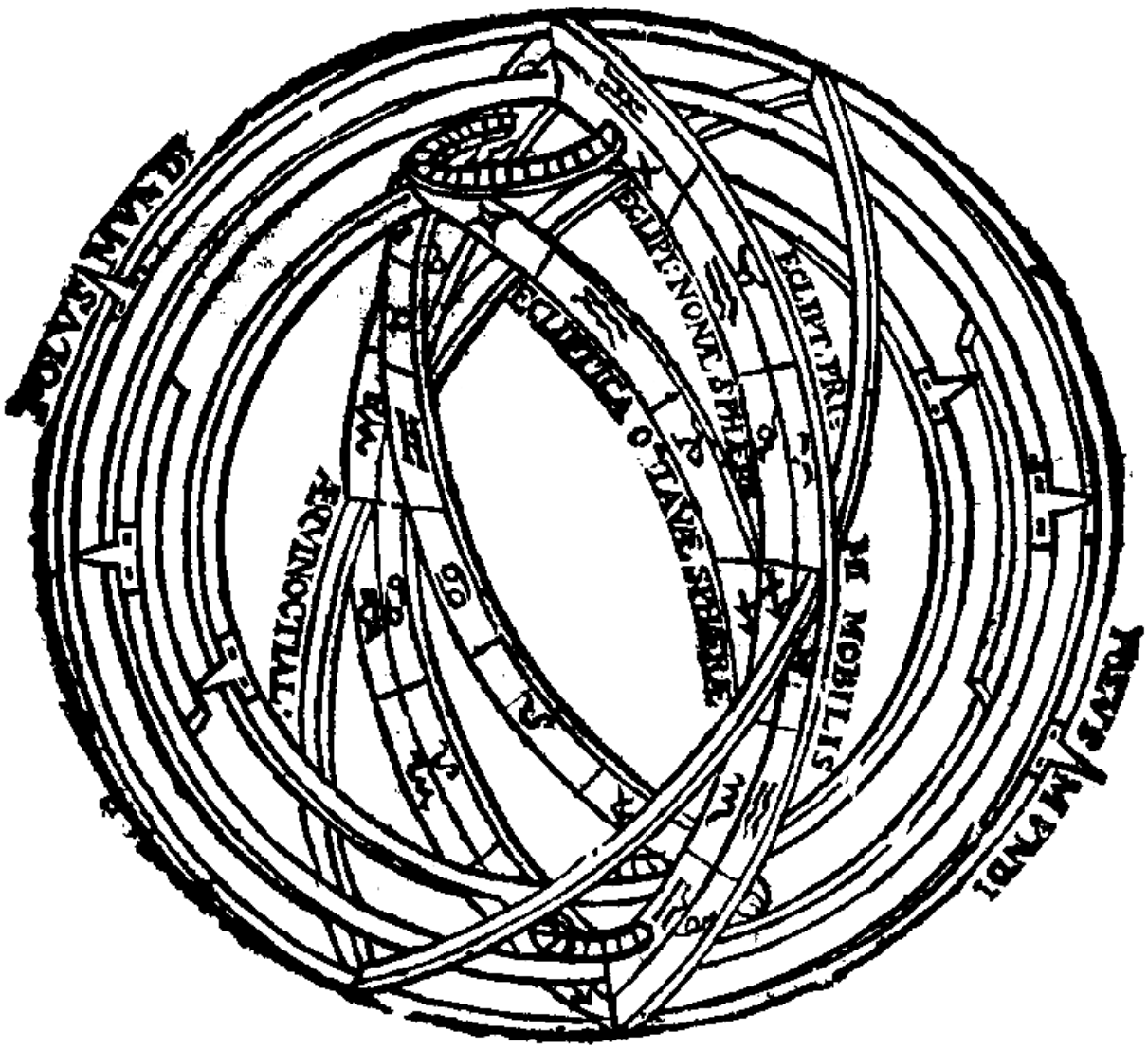
⇨ ***Voir TP sur Excel en annexe.***

c - A l'aide du papier Gausso-arithmétique

C'est une méthode graphique ("au jugé"). Son intérêt est sa grande simplicité, raison pour laquelle elle est encore utilisée dans le domaine technologique.

On a reporté en abscisses (échelle arithmétique), les temps de bon fonctionnement x_i et en ordonnées (échelle gaussienne proportionnelle à Π^{-1}), les fréquences cumulées y_i .


Pourcentages
99.9
99.8
99.5
98
97
96
95
90
ψ|m + σ|
80
70
60
50
40
30
20
ψ|m - σ|
10
5
4
3
2
1
0.50
0.20
0.10
0.05
0.02
400
500
1500
μ-σ ≈ 680
μ≈ 970 h
1270

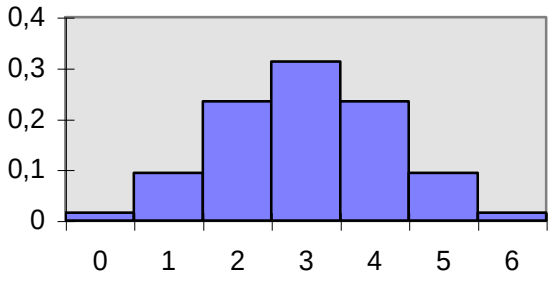

POLVS MVNDI
ECLIPT. NOVA
ECLIPTICA
AEQVINOCTIALIS
MOBILIS
POLVS MVNDI

TRAVAUX DIRIGES

T. D. : INTRODUCTION AUX VARIABLES ALEATOIRES CONTINUES

Une ***variable aléatoire discrète*** prend des valeurs "isolées" :

- Une variable aléatoire X suivant la loi binomiale $\mathcal{B}(n, p)$ prend comme valeurs possibles les nombres entiers entre 0 et n.
- Une variable aléatoire X suivant la loi de Poisson $\mathcal{P}(\lambda)$ prend comme valeurs possibles les nombres entiers positifs.

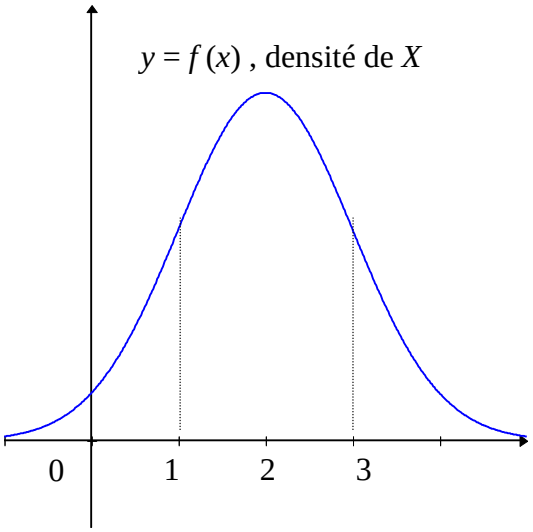


On peut représenter la distribution d'une variable aléatoire discrète sous forme d'histogramme avec des rectangles de base 1, où la probabilité $P(X = k)$ est proportionnelle (question d'échelle) à ***l'aire du rectangle*** correspondant.

Par exemple, ci-contre, la distribution de la loi binomiale $\mathcal{B}(6, \frac{1}{2})$.

Une ***variable aléatoire continue*** peut prendre comme valeurs tous les nombres réels d'un certain intervalle.
Sa distribution est donnée par sa ***fonction de densité*** f. La probabilité qu'une réalisation de X soit comprise entre a et b est alors (à un facteur d'échelle près) ***l'aire située sous la courbe représentative de f, entre les droites d'équation x = a et x = b.***

1 A partir d'une courbe de densité



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 :

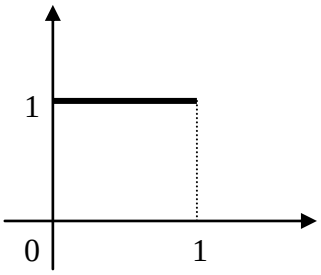
$P(X = 2) = \int_2^2 f(x)\,dx$ = ……..

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.

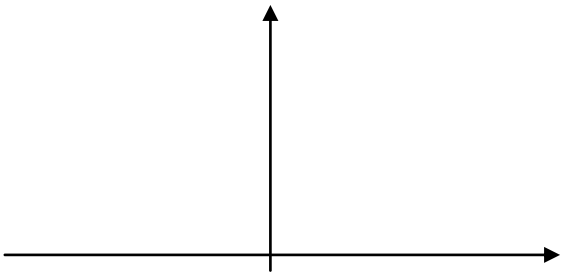
Comparer avec les valeurs théoriques $E(X)$ et $\sigma(X)$:
.....

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 $\lim_{x \to +\infty} f(x)$ puis dresser le tableau de variation de f.

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

d) Calculer , pour $t > 0$, $I(t) = \int_{0}^{t} f(x)\,dx$.

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

e) Déterminer $\lim_{t \to +\infty} I(t)$.

.....
.....

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 : $E(X) = \lim\limits_{t \to +\infty} J(t)$.

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

4) Soit la fonction f définie sur $\mathbb{R}$ par $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.

a) Montrer que f est paire. Qu'en déduit-on quant à sa courbe représentative ?

……
……

b) Déterminer $\lim\limits_{x \to +\infty} f(x)$. En déduire $\lim\limits_{x \to -\infty} f(x)$.

……
……
……
……
……
……
……

c) Calculer, pour tout $x \in \mathbb{R}$, $f'(x)$. Dresser le tableau de variations de f.

……
……
……
……
……
……
……
……
……

d) Tracer, dans un repère orthogonal $\left(O; \vec{i}; \vec{j}\right)$, la courbe $\mathcal{C}$, représentant la fonction f.

e) Donner, à l'aide de votre calculatrice, des valeurs approchées des intégrales suivantes :

$I(1) = \int_{-1}^{1} f(x)dx \approx$ …… ; $I(2) = \int_{-2}^{2} f(x)\, dx \approx$

……...

$I(10) = \int_{-10}^{10} f(x)\, dx \approx$ …… .

Travaux pratiques STATISTIQUES

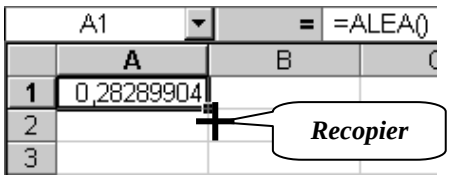
T.P. Excel : EXPERIMENTATION DU THEOREME LIMITE CENTRAL

Objectifs

- Expérimenter le théorème limite central sur la simulation de 1000 données.
- Trier des données pour comparer leur distribution à une densité normale.
- Contrôler la normalité d'une distribution empirique par régression linéaire (droite de *Henry*).
- Appliquer ce procédé pour un test de normalité d'une production industrielle.

1 - GENERATION D'UNE DISTRIBUTION DE 1000 VALEURS

Loi uniforme sur [0 , 1]

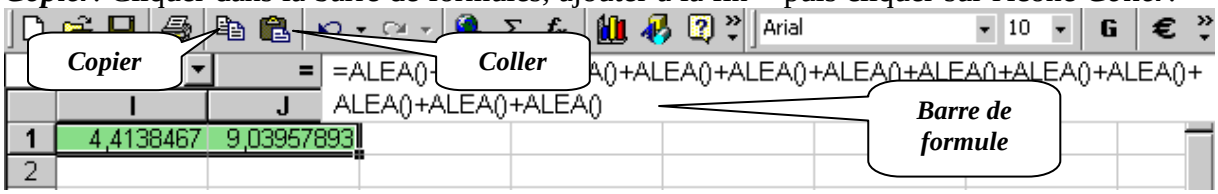


Lancer Excel®.
Dans la cellule A1, entrer la formule : =ALEA( )
Faire ***ENTREE*** puis, en approchant le pointeur de la souris du carré noir situé dans le coin inférieur droit de la cellule A1, celui-ci se transforme en une croix noire, faire alors glisser en maintenant le bouton gauche enfoncé pour ***recopier*** jusqu'en J1.
La fonction ALEA( ) génère un nombre aléatoire compris entre 0 et 1, c'est à dire qu'elle simule un résultat d'une variable aléatoire X suivant la loi uniforme sur [0 ; 1] : U ([0 , 1]).
Un peu de théorie avant de poursuivre... Cette loi continue admet comme fonction de densité la fonction f définie par : $f(x) = \begin{cases} 1 \text{ si } x \in [0,1] \\ 0 \text{ si } x \notin [0,1] \end{cases}$.

– Calculer, pour cette loi, sur la feuille réponse, l'espérance et la variance.

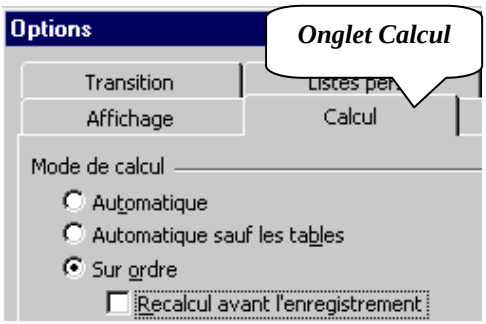
Somme de n = 12 variables aléatoires indépendantes de même loi U [0 , 1]

On désigne par Y une variable aléatoire somme de n = 12 variables aléatoires de même loi U [0 , 1]. Pour simuler 1000 réalisations de Y, procéder ainsi :
Cliquer sur la cellule A1. Dans la ***Barre de formules***, sélectionner ALEA() (mettre en vidéo inversée en balayant, bouton gauche de la souris enfoncé) puis cliquer sur l'icône ***Copier***. Cliquer dans la barre de formules, ajouter à la fin + puis cliquer sur l'icône ***Coller***.



Répéter l'opération de façon à avoir la somme de 12 fois la fonction ALEA( ) (rien à voir avec 12*ALEA( ) !). Faire ***ENTREE***.
En approchant le pointeur de la souris du carré noir situé dans le coin inférieur droit de la cellule A1, celui-ci se transforme en une croix noire, ***Recopier*** alors jusqu'en J1. Relâcher

le bouton gauche de la souris puis ***Recopier vers le bas*** jusqu'en J100 (il faut que la première ligne soit sélectionnée avant de recopier vers le bas).



Vous avez maintenant une simulation de 1000 réalisations aléatoires de la variable *Y*.
Pour conserver ces 1000 valeurs, ou refaire une nouvelle simulation quand on le désire, cliquer dans le ***menu Outils / Options...*** puis dans l'***onglet Calcul*** de la boîte de dialogue, à la rubrique ***Mode de calcul***, choisir • ***Sur ordre*** puis ***OK***.

Moyenne $\bar{x}$ et écart type s_e d'un échantillon de 1000 valeurs

En A102 taper "moyenne" et en B102 entrer la formule : =MOYENNE(A1:J100) puis ***ENTREE***.
En A103 taper "écart type" et en B103 entrer la formule : =ECARTYPEP(A1:J100)
(Attention à bien taper ECARTYPEP et non ECARTYPE qui correspond à l'estimation de l'écart type de la population dont est extrait l'échantillon) puis ***ENTREE***.
Faire ***F9*** pour une nouvelle simulation de 1000 valeurs.

– Consigner vos résultats sur la feuille réponse.

Valeurs théoriques de μ et σ

En répétant 12 fois la fonction ALEA() et en faisant la somme, on simule la somme de 12 variables aléatoires indépendantes de loi $U([0 , 1])$. On note μ et σ l'espérance et l'écart type de cette somme.

– Déterminer les paramètres théoriques μ et σ sur la feuille réponse.

2 - THEOREME LIMITE CENTRAL

Un énoncé du théorème

Pourquoi, avec une expression analytique paradoxalement compliquée, la loi de *Laplace-Gauss* est-elle si répandue au point d'être qualifiée de "*normale*" ?
La réponse des mathématiciens à cette question est le ***Théorème limite central*** :

> ***La somme de n variables aléatoires indépendantes de même loi suit approximativement, pour n assez grand, une loi normale.***

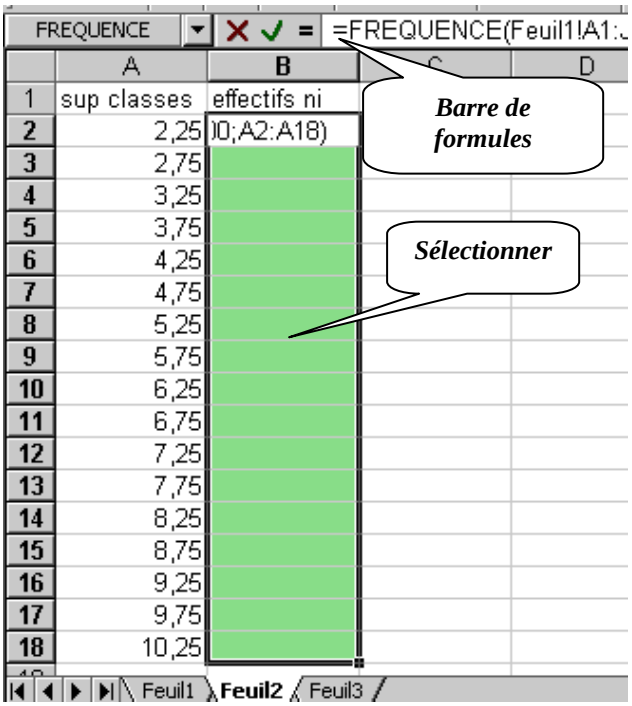
– Expliquer, sur la feuille réponse, pourquoi, selon ce théorème, la variable aléatoire *Y* suit approximativement une loi normale.

Comparaison graphique de l'histogramme des 1000 données à la densité normale

Tri des données

On va regrouper les 1000 données en 17 classes de part et d'autre de la valeur 6.
Il faut entrer les bornes supérieures de ces classes.
Cliquer sur l'onglet ***Feuil2*** (en bas).
En A1 taper "sup classes".

En A2 entrer la valeur 2,25 . En A3, entrer la formule : =A2+0,5



	A	B
1	sup classes	effectifs ni
2	2,25	)0;A2:A18)
3	2,75	
4	3,25	
5	3,75	
6	4,25	
7	4,75	
8	5,25	
9	5,75	
10	6,25	
11	6,75	
12	7,25	
13	7,75	
14	8,25	
15	8,75	
16	9,25	
17	9,75	
18	10,25	

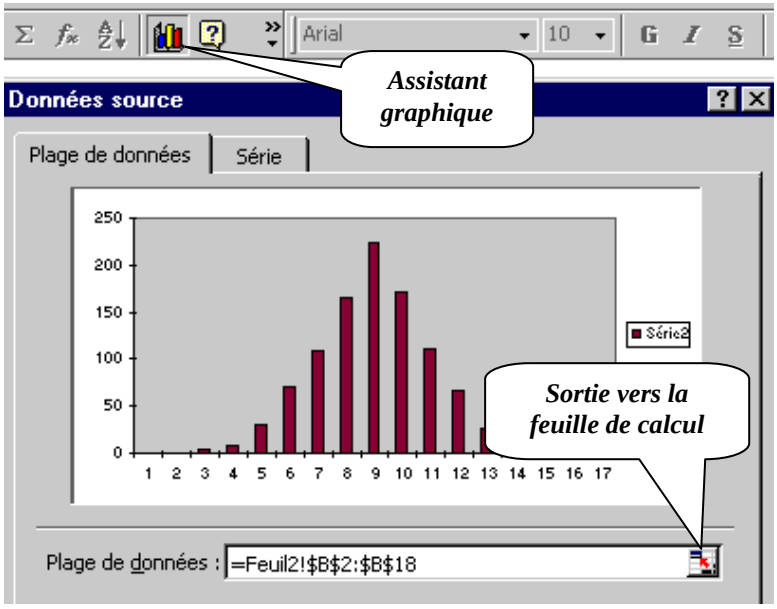
Recopier vers le bas jusqu'en A18 puis faire ***F9***. Le calcul s'effectue et la dernière cellule contient alors la valeur 10,25.
En B1 taper "effectifs ni".
Sélectionner les cellules de B2 à B18 (pour cela, cliquer sur B2 et glisser, en gardant le bouton gauche de la souris enfoncé, jusqu'en B18, puis relâcher le bouton de la souris).
Alors que les cellules sélectionnées apparaissent en "vidéo inversée", cliquer dans la ***barre de formules*** et taper :
=FREQUENCE(Feuil1!A1:J100;A2:A18)
puis valider en appuyant *en même temps* sur les touches ***CTRL MAJUSCULE*** et ***ENTREE***.
Excel calcule alors les effectifs de chacune des classes (et non les fréquences comme semble l'indiquer le nom de la fonction utilisée).

En B19, cliquer sur l'icône **Σ** de ***somme automatique*** puis sur ***ENTREE***. Vous devriez obtenir l'effectif total de 1000.

Comparaison avec les résultats que fournirait la loi N (6 ; 1)

La fonction de densité de la normale N (6 ; 1) est définie sur ℝ par : $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-6)^2}$.

L'histogramme étant construit à partir de classes d'amplitude 0,5 , chaque rectangle devrait, dans le cadre de la loi normale N (6 ; 1), avoir comme hauteur $500 \times f(x)$.
En C1, taper "valeurs xi". En C2, entrer la valeur 2. En C3, entrer la formule : =C2+0,5 puis recopier cette formule vers le bas jusqu'en C18 . Après ***F9***, cette cellule contiendra la valeur 10. En D1, taper "loi normale". En D2, entrer la formule :
=(500/RACINE(2*PI()))*EXP(-0,5*(C2-6)^2) puis recopier cette formule vers le bas jusqu'en D18. Faire ***F9***.



Cliquer sur l'icône ***Assistant graphique***.
Etape 1 sur 4 : dans l'onglet ***Types personnalisés*** choisir ***Courbes-Histogramme*** *puis* cliquer sur ***Suivant***.
Etape 2 sur 4 : dans l'onglet ***Plage de données***, sortir, par l'icône indiqué ici, vers la feuille de calcul. Y sélectionner les valeurs n_i puis revenir, par l'icône analogue, dans la boîte de dialogue.
Dans l'onglet ***Série***, pour la ***Série 2*** , ***Etiquettes des abscisses X***,

sortir, sur la feuille de calcul, sélectionner les valeurs x_i . Cliquer sur ***Ajouter*** puis, pour la ***Série 1***, sortir, sur la feuille de calcul, sélectionner les ***Valeurs*** (colonne des valeurs "loi normale"). Cliquer sur ***Suivant***.
Etape 3 sur 4 : dans l'onglet ***Légende***, désélectionner ***Afficher la légende***. Cliquer sur ***Suivant***.
Etape 4 sur 4 : cocher • ***Sur une nouvelle feuille*** puis ***Terminer***.
Cliquer, avec le *bouton droit* de la souris, sur un point de la courbe (Série 1) puis choisir ***Format de la série de données...*** Dans l'onglet ***Motifs***, pour ***Traits***, augmenter un peu l'épaisseur et cocher la case ***Lissage***, pour ***Marque***, cocher ***Aucune*** puis faire ***OK***.
Faire ***F9*** pour une nouvelle simulation de 1000 valeurs.

– Consigner vos commentaires sur la feuille réponse.

3- DROITE DE HENRY

On souhaite, dans ce paragraphe, utiliser la technique de la régression linéaire selon les moindres carrés pour quantifier la qualité de l'ajustement de la distribution observée avec une loi normale.

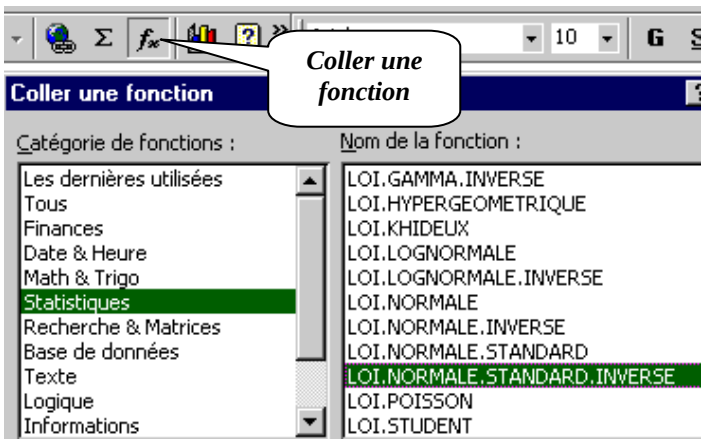
• On désigne maintenant par X une variable aléatoire suivant la loi $\mathcal{N}(\mu , \sigma)$.

Sa fonction de répartition F est donnée, pour tout $x \in \mathbb{R}$, par :

$y = F(x) = P(X \leq x) = P(\dfrac{X-\mu}{\sigma} \leq \dfrac{x-\mu}{\sigma}) = P(\dfrac{X-\mu}{\sigma} \leq t) = \Pi(t)$, avec $t = \dfrac{x-\mu}{\sigma}$ et où Π est la fonction de répartition de la loi $\mathcal{N}(0 , 1)$, tabulée dans le formulaire officiel.

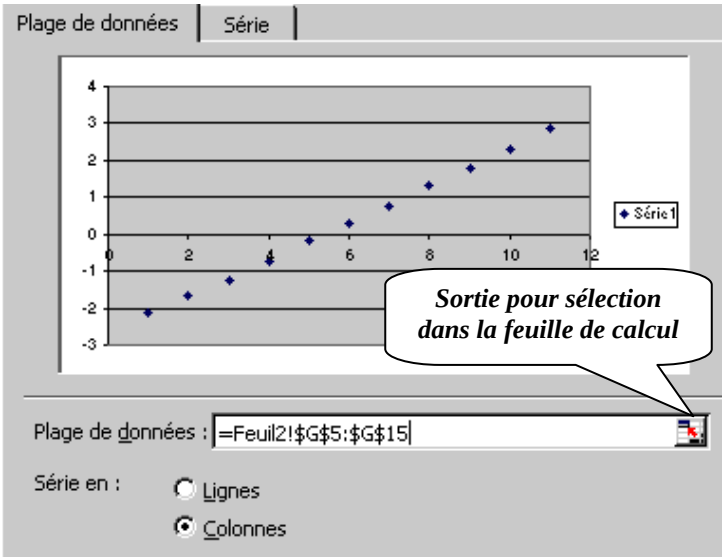
• Statistiquement, pour une valeur x_i de la distribution, y_i est la *fréquence cumulée croissante* (analogue statistique de la fonction de répartition). Notons t_i la valeur, donnée par *lecture inverse* de la table de la loi $\mathcal{N}(0 , 1)$, telle que $y_i = \Pi(t_i) \Leftrightarrow t_i = \Pi^{-1}(y_i)$.
S'il s'agit d'une distribution normale, le nuage de points (x_i , t_i) devrait donc être ajusté par la droite d'équation $t = \dfrac{x-\mu}{\sigma}$, que l'on nomme **droite de Henry**.

• Sur la ***feuille 2***, taper en E1 "ni cumulés" et E2, entrer la formule : =B2 puis en E3, la formule : =E2+B3 puis ***Recopier vers le bas*** jusqu'en E18 puis faire ***F9***. La valeur calculée devrait être 1000.
En F1, taper "fréq cumul yi", puis en F2, entrer la formule : =E2/$B$19 puis recopier vers le bas jusqu'en F18 (le symbole $ empêche la modification de la référence de la cellule lors de la recopie vers le bas). Faire ***F9***.
En G1, taper "ti invnorm(yi)".
Cliquer en G2, puis sur l'icône ***Coller une fonction***.
Choisir ***Statistiques*** puis ***LOI.NORMALE.STANDARD.INVERSE*** et cliquer sur ***OK***.
Dans la boîte de dialogue, entrer pour ***Probabilité*** : F2.
Cliquer sur ***OK***.
Recopier vers le bas jusqu'en G18 puis faire ***F9***.



– Pour les valeurs 0 et 1 de y, Excel répond pour t : **#NOMBRE!**
Pouvez-vous expliquer pourquoi ?

On va maintenant représenter le nuage de points (x_i ; t_i).
Cliquer sur l'icône ***Assistant graphique***.



Etape 1/4 : choisir ***Nuages de points*** (sous-type n°1 sans courbe).
Cliquer sur ***Suivant***.
Etape 2/4 : Onglet ***Plage de données***, sortir vers la feuille de calcul pour y sélectionner, parmi les valeurs t_i , celles ne contenant pas #NOMBRE!
Revenir dans la boîte de dialogue de l'assistant graphique.
Onglet ***Série***, Valeurs X : sortir sélectionner les cellules contenant les valeurs *x*i correspondantes.
Cliquer sur ***Suivant***.
Etape 3/4 : Onglet ***Légende***, désactiver l'option ***Afficher la légende***. Onglet ***Quadrillage***, désactiver les options. Cliquer sur ***Suivant***.
Etape 4/4 : cocher • ***Sur une nouvelle feuille*** puis cliquer sur ***Terminer***.
On peut demander à Excel d'ajouter sur le nuage de points (x_i , t_i) une courbe de tendance.
Cliquer sur le graphique. Aller dans le menu ***Graphique*** (en haut de l'écran) et cliquer sur ***Ajouter une courbe de tendance...*** Dans la boîte de dialogue, dans l'onglet ***Type*** choisir ***Linéaire*** puis dans l'onglet ***Option*** cocher
• ***Afficher l'équation sur le graphique*** et
• ***Afficher le coefficient de détermination (R^2) sur le graphique***
puis cliquer sur ***OK***.
La droite affichée est la droite de Henry, obtenue par ajustement linéaire de t en x selon la méthode des moindres carrés. Faire ***F9*** pour une autre simulation de 1000 valeurs.

– Exploiter sur la feuille réponse les résultats affichés.

4- TEST DE NORMALITE D'UNE PRODUCTION

La durée de vie, exprimée en heures, des joints à lèvres PLAUSTRA - type IE - définit une variable aléatoire continue *X*.
L'étude de la durée de vie de 500 de ces joints a permis d'obtenir l'historique suivant :

Temps de bon fonctionnement x_i	500	700	900	1100	1300	1500	1700
Effectifs n_i	24	67	108	126	109	51	15

Reprendre, sur la feuille 3 du classeur Excel, les techniques du §3 pour :
1) Créer un tableau contenant les valeurs x_i , n_i , les effectifs cumulés croissants y_i et les valeurs $t_i = \Pi^{-1}(y_i)$ correspondantes.
2) Représenter graphiquement le nuage de points (x_i ; t_i).
3) Y indiquer la droite de Henry et le coefficient de détermination.

– Exploiter les calculs d'Excel pour montrer, sur la feuille réponse, que l'on peut ajuster la distribution de la durée de vie des joints PLAUSTRA à une loi normale dont on précisera les paramètres.

– FEUILLE REPONSE

NOMS : *.....*

1- GENERATION D'UNE DISTRIBUTION DE 1000 VALEURS

Loi uniforme sur [0 , 1]

Soit X suivant la loi U ([0 , 1]).

$E(X) = \int_0^1 x\mathrm{d}x$ = ...

$V(X) = E(X^2) - [E(X)]^2 = \int_0^1 x^2\mathrm{d}x - [E(X)]^2$ = ...

Somme de *n* = 12 variables aléatoires indépendantes de même loi U [0 , 1]

Moyenne $\bar{x}$ et écart type s_e d'un échantillon de 1000 valeurs

Compléter, pour quatre simulations, le tableau ci-dessous :

Simulation n°	1	2	3	4
$\bar{x}$				
s_e				

Valeurs théoriques de μ et σ

Soit $Y = X_1 + X_2 + ... + X_{12}$ où les X_i sont indépendantes de loi U ([0 ; 1]).

μ = $E(Y) = 12 \times E(X)$ .....

σ = $\sqrt{V(Y)} = \sqrt{12V(X)}$ = ...

Comparer $\bar{x}$ à μ et s_e à σ : .....

.....

2 - THEOREME LIMITE CENTRAL

Un énoncé du théorème

Pourquoi peut-on considérer que Y suit approximativement une loi normale ? .....

.....

.....

.....

Quels sont, théoriquement, les paramètres de cette loi normale ? .....

.....

Comparaison graphique de l'histogramme des 1000 données à la densité normale

Comparer, sur plusieurs simulations, l'histogramme avec le profil de la densité normale : .....

.....

.....

L'utilisation du théorème limite central était-elle justifiée ? .....
.....
.....

Imprimer le graphique (si possible) ou enregistrer le fichier

3- DROITE DE HENRY

On pose $y = F(x) = \Pi(t)$ et $t = \Pi^{-1}(y)$.
Pour les valeurs 0 et 1 de y, Excel répond pour t : #NOMBRE! car $\Pi^{-1}(0)$ correspondrait à .....
.....et $\Pi^{-1}(1)$ correspondrait à .....
Déterminer, pour plusieurs simulations, la valeur du coefficient de corrélation linéaire R de t en x :

Simulation	1	2	3	4
R ≈				

Peut-on estimer que la série des 1000 données est issue d'une distribution normale ? .....
Pour une simulation où R est au moins de 0,9, donner l'équation $t = ax + b$ de la droite de Henry fournie par Excel :
.....

S'il s'agit de loi $\mathbb{N}(\mu ; \sigma)$, la droite de Henry a comme équation : $t = \frac{1}{\sigma} x - \frac{\mu}{\sigma}$.

En déduire une estimation de μ et de σ :
σ est estimé à .....
.....
μ est estimé à .....
.....

Comparer aux valeurs de μ et de σ attendues (§1) : .....

4- TEST DE NORMALITE D'UNE PRODUCTION

Est-il raisonnable d'ajuster la distribution des durées de vie des joints PLAUSTRA à une loi normale ? .....
.....
Quelle est l'équation $t = ax + b$ de la droite de Henry calculée par Excel ?
.....
En déduire une estimation de μ et σ : .....
.....
.....
.....

Imprimer le graphique (si possible) ou enregistrer le fichier

Corrigé et compte-rendu de l'activité EXCEL
"AJUSTEMENT A UNE LOI NORMALE"

1- GENERATION D'UNE DISTRIBUTION DE 1000 VALEURS

Loi uniforme sur [0 , 1]

Si X suit la loi $\mathcal{U}$ ([0 , 1], on a $E(X) = \frac{1}{2}$ et $V(X) = \frac{1}{12}$.

Somme de n = 12 variables aléatoires indépendantes de même loi $\mathcal{U}$ [0 , 1]

Des valeurs expérimentales sur 1000 réalisations de Y :

Simulation n°	1	2	3	4
$\bar{x}$	6.001	5.974	5.974	6.026
s_e	1.008	1.016	1.004	1.008

Les valeurs théoriques sont $\mu = E(Y) = 6$ et $\sigma = \sigma(Y) = \sqrt{12 \times \frac{1}{12}} = 1$.

Les valeurs observées sur les 1000 données sont très proches

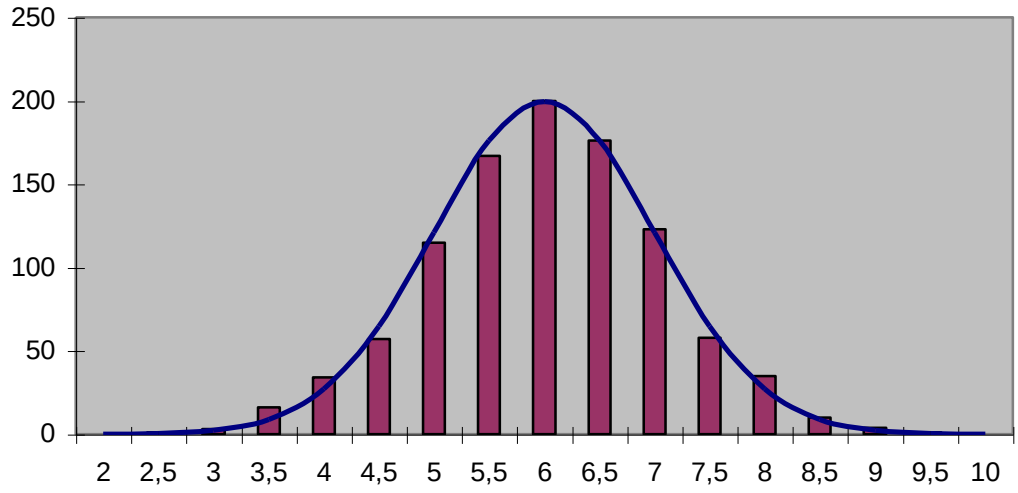
2 - THEOREME LIMITE CENTRAL

Un énoncé du théorème

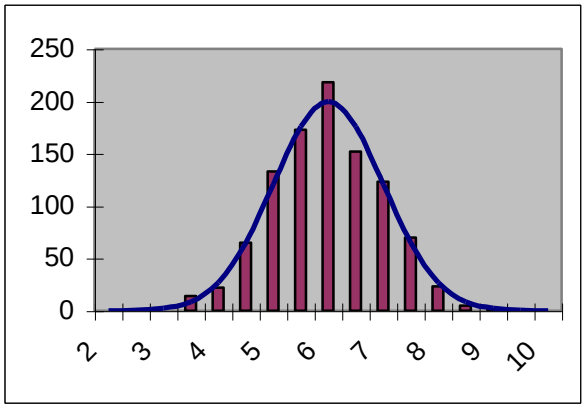
La variable aléatoire Y étant la somme de n = 12 variables aléatoires indépendantes de même loi, elle suit approximativement (en supposant que n = 12 est assez grand) une loi normale.

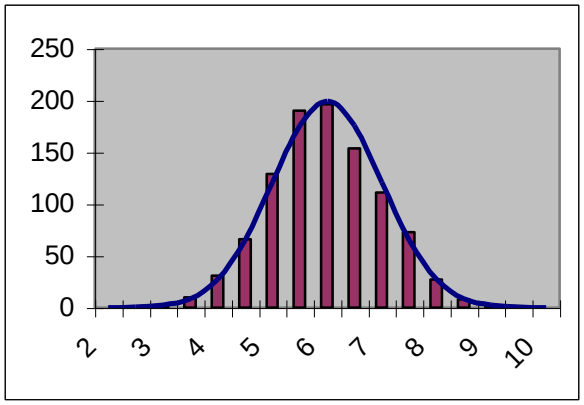
Il est clair que la moyenne et l'écart type de cette loi normale doivent être ceux de Y cet à dire $\mu = 6$ et $\sigma = 1$.

Comparaison de l'histogramme des 1000 réalisations de Y avec la densité de la loi $\mathcal{N}$ (6 , 1)



Il suffit de faire F9, pour avoir aussitôt une autre simulation. On expérimente ainsi l'approximation donnée par le théorème limite central. Il apparaît ici que n = 12 est suffisant pour une bonne approximation normale.





3- DROITE DE HENRY

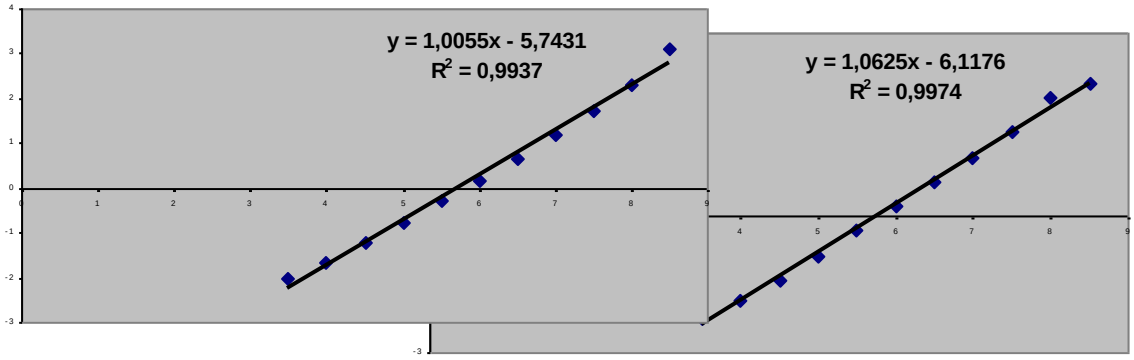
	A	B	C	D	E	F	G
1	sup classes	effectifs ni	valeurs xi	loi normale	ni cumulés	fréq cumul yi	ti invnorm(yi)
2	2,25	0	2	0,06691511	0	0	#NOMBRE!
3	2,75	0	2,5	0,43634135	0	0	#NOMBRE!
4	3,25	0	3	2,21592421	0	0	#NOMBRE!
5	3,75	7	3,5	8,76415025	7	0,007	-2,45727279
6	4,25	27	4	26,9954833	34	0,034	-1,82500571
7	4,75	71	4,5	64,7587978	105	0,105	-1,25356564
8	5,25	144	5	120,985362	249	0,249	-0,67763949
9	5,75	180	5,5	176,032663	429	0,429	-0,17892035
10	6,25	182	6	199,47114	611	0,611	0,2819263
11	6,75	156	6,5	176,032663	767	0,767	0,7290032
12	7,25	122	7	120,985362	889	0,889	1,22122856
13	7,75	75	7,5	64,7587978	964	0,964	1,79911694
14	8,25	22	8	26,9954833	986	0,986	2,19728463
15	8,75	9	8,5	8,76415025	995	0,995	2,57583451
16	9,25	5	9	2,21592421	1000	1	#NOMBRE!
17	9,75	0	9,5	0,43634135	1000	1	#NOMBRE!
18	10,25	0	10	0,06691511	1000	1	#NOMBRE!

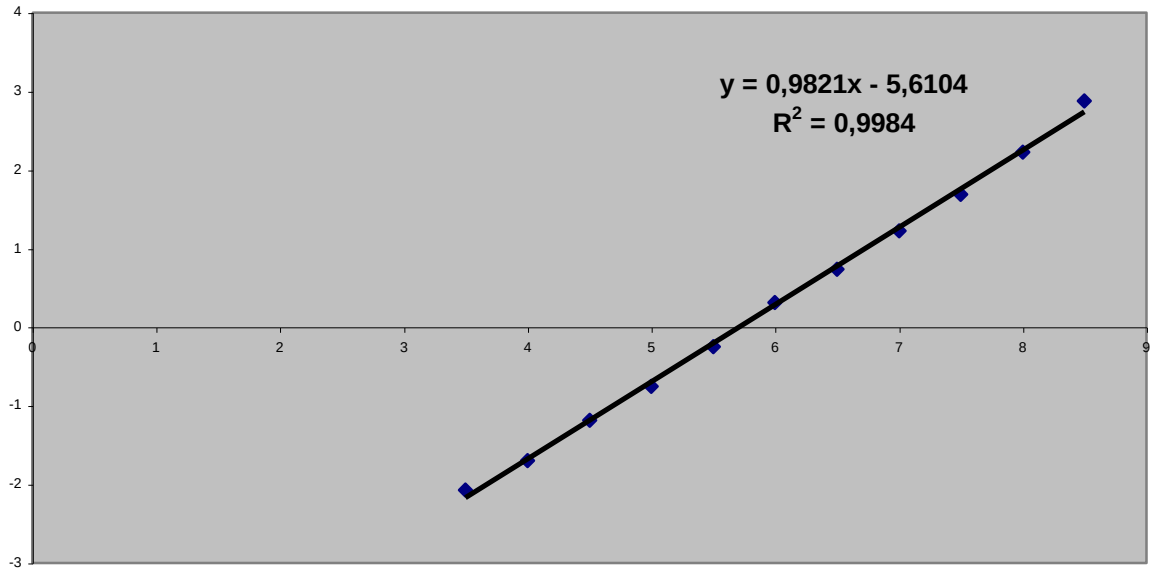
Feuil1 / Graph4 / Graph5 / **Feuil2** / Feuil3 /

On aurait $\Pi^{-1}(0)$ qui vaudrait $-\infty$ et $\Pi^{-1}(1)$ qui vaudrait $+\infty$, c'est la raison de la réponse #NOMBRE!

Simulation	1	2	3	4
R ≈	0,999	0,999	0,998	0,862

De façon générale, R est très proche de 1 et on peut donc affirmer que les 1000 données sont issues d'une distribution approximativement normale.





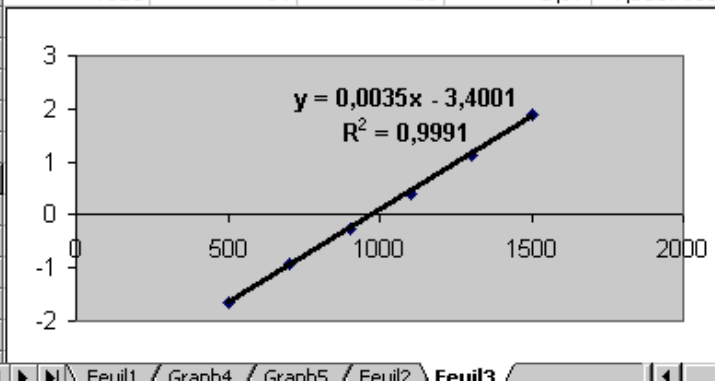
Pour le graphique ci-dessus (par exemple), on a $t = 0{,}9821\,x - 5{,}6104$.

D'où $\frac{1}{\sigma} = 0{,}9821$ et σ est estimé à 1,02. Puis $\frac{\mu}{\sigma} = 5{,}6104$ qui permet d'estimer μ à 5,71.

Ces valeurs sont assez proches des valeurs théoriques : $\mu = 6$ et $\sigma = 1$.

4- TEST DE NORMALITE D'UNE PRODUCTION

	A	B	C	D	E
1	xi	ni	ni cumulés	yi	ti
2	500	24	24	0,048	-1,66456175
3	700	67	91	0,182	-0,90776894
4	900	108	199	0,398	-0,25852728
5	1100	126	325	0,65	0,38532107
6	1300	109	434	0,868	1,11698682
7	1500	51	485	0,97	1,88078957



On obtient un coefficient de corrélation R ≈ 0,9995 qui justifie l'ajustement des durées de vie à une distribution normale.

On a $\frac{1}{\sigma} = 0{,}0035$ d'où σ estimé à 285,7 heures.

Puis $\frac{\mu}{\sigma} = 3{,}4001$ qui conduit à estimer μ à 971,5 heures.

Epreuves corrigées de B.T.S.

Annales du B.T.S. : VARIABLES ALEATOIRES CONTINUES ET LOI NORMALE

EXERCICES D'ANALYSE A PROPOS D'UNE LOI CONTINUE

1 – Conception des produits industriels 1997

Soient le nombre réel strictement positif a et f la fonction, de la variable réelle t, définie sur $\mathbb{R}$ par $\begin{cases} \text{si } t \geq 0 \text{ alors } f(t) = a\,e^{-at} \\ \text{si } t < 0 \text{ alors } f(t) = 0 \end{cases}$.

1°a) Montrer que, pour tout t de $\mathbb{R}$, $f(t) \geq 0$.

b) Pour tout réel strictement positif, calculer, en fonction de x et de a, l'intégrale $F(x) = \int_0^x f(t)\mathrm{d}t$, en déduire $\lim\limits_{x \to +\infty} F(x)$.

c) Calculer à l'aide d'une intégration par parties, en fonction de x et de a, l'intégrale $J(x) = \int_0^x t\ f(t)\mathrm{d}t$, en déduire $\lim\limits_{x \to +\infty} J(x)$

2° Soit Y la variable aléatoire qui à toute ampoule prélevée au hasard dans la production de B associe la mesure, en heure, de sa durée de vie. On admet que la mesure, en heure, de la durée de vie moyenne des ampoules de la production B est l'espérance mathématique de la variable aléatoire Y notée E(Y) et vérifiant E(Y) = $\lim\limits_{x \to +\infty} J(x)$.

a) Exprimer E(Y) en fonction de a.

b) Quelle doit être la valeur de a pour que la durée de vie moyenne des ampoules de la production B soit égale à 1000 heures. Dans ce cas, pour tout réel positif, exprimer $f(t)$ en fonction de t.

2 – D'après Groupement C 1999

1° On appelle f la fonction définie sur $\mathbb{R}$ par : $f(t) = e^{-\frac{t^3}{10^9}}$.
a) Démontrer que f est une fonction décroissante.
Déterminer sa limite en $+\infty$ et interpréter géométriquement ce résultat.
b) Déterminer la limite de f en $-\infty$.
c) Tracer soigneusement la courbe représentative de f dans un repère orthogonal pour t variant de 0 à 1500 (échelle : 1 cm pour 100 unités sur l'axe des abscisses et 10 cm pour une unité sur l'axe des ordonnées).

2° a) Résoudre algébriquement dans $\mathbb{R}$ l'équation $f(t) = 0{,}5$; donner la valeur exacte de la solution, puis sa valeur approchée arrondie à l'unité.
b) En déduire l'ensemble des solutions de l'inéquation $f(t) < 0{,}5$.

4° On appelle T la variable aléatoire associant à toute machine d'un certain type sa durée, en heures, de fonctionnement sans panne.
On admet que, pour t réel positif ou nul, $f(t)$ représente la probabilité que T soit supérieur à t ainsi $P(T > t) = f(t)$.
a) Calculer la probabilité qu'une telle machine fonctionne plus de 1000 heures sans panne.
b) Pourquoi peut-on affirmer qu'il y a plus de neuf chances sur dix qu'une telle machine fonctionne sans panne plus de 400 heures ?

3 – Maintenance Nouvelle Calédonie 1996

1) *Calcul d'intégrales*
Calculer en fonction du nombre réel positif t, les intégrales suivantes :

a) $F(t) = \dfrac{1}{200}\displaystyle\int_0^t \mathrm{e}^{-0{,}005x}\,\mathrm{d}x$;

b) $J(t) = \dfrac{1}{200}\displaystyle\int_0^t x\,\mathrm{e}^{-0{,}005x}\,\mathrm{d}x$

c) $K(t) = \dfrac{1}{200}\displaystyle\int_0^t x^2\,\mathrm{e}^{-0{,}005x}\,\mathrm{d}x$

(On pourra utiliser des intégrations par parties pour calculer $J(t)$ et $K(t)$).

2) *Interprétation en probabilités*
Soit la fonction f définie par : $\begin{cases} f(x) = 0 & \text{pour} \quad x < 0, \\ f(x) = \dfrac{1}{200}\mathrm{e}^{-0{,}005x} & \text{pour} \quad x = 0. \end{cases}$

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) *Utilisation*
On considère que la variable aléatoire T correspond au temps de bon fonctionnement d'un certain type de matériel.
a) La fonction de fiabilité associée à variable aléatoire T est définie sur $[0, +\infty[$ par $R(t) = P(T > t)$. Montrer que $R(t) = \mathrm{e}^{-0{,}005\,t}$.
b) Calculer à 10^{-3} près : $P(T > 300)$, $P(T > 100)$ et $P(100 < T \leq 300)$.
c) Calculer la valeurs entière approchée de t_0 telle que $P(T \leq t_0) = 0{,}1$.

APPROXIMATION D'UNE LOI BINOMIALE PAR UNE LOI NORMALE – CORRECTION DE CONTINUITE

4 – Groupement D 2001

Un magicien prétend qu'il peut souvent deviner à distance la couleur d'une carte tirée au hasard d'un jeu de cartes bien battu et comportant des cartes de deux couleurs différentes en nombre égal.
On appelle p la probabilité que le magicien donne une réponse juste (succès) lors d'un tirage.

Si le magicien est un imposteur on a $p = \frac{1}{2}$, sinon $p > \frac{1}{2}$.

On appellera échantillon de taille n toute réalisation de n tirages successifs d'une carte dans le jeu, avec remise.

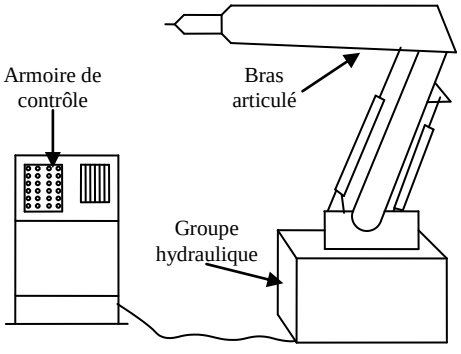
On suppose $p = \frac{1}{2}$ et on note Y la variable aléatoire qui, à tout échantillon de taille n, associe le nombre de succès du magicien.
(On arrondira les probabilités au dix millième le plus proche.)

1. Dans cette question on prend $n = 20$.
Quelle est la loi suivie par Y ? Donner ses paramètres.
Calculer la probabilité $P(Y = 15)$.

2. Dans cette question on prend $n = 100$. On admet que la variable aléatoire Y peut-être approchée par une variable aléatoire Z suivant une loi normale.
Préciser les paramètres de cette loi normale.
Utiliser cette approximation pour calculer $P(Y > 60)$.

5 – Groupement B 2000

Les ateliers de peinture d'un grand constructeur automobile fonctionnent à l'aide de robots permettant de positionner les pistolets autour de la carrosserie.



Ces robots sont constitués de trois parties : un bras articulé actionné par des vérins hydrauliques, un groupe hydraulique et une armoire de contrôle

Pannes mécaniques sur le bras articulé
Les pannes mécaniques sur le bras articulé, assez fréquentes, sont souvent sans gravité. Elles sont généralement dues à l'encrassage par la peinture, à du jeu ou à un blocage dans les articulations mécaniques.

Les ateliers de peinture comptent 300 robots équipés de bras articulés identiques dont les pannes mécaniques surviennent de façon indépendante. Pour chaque bras articulé, la probabilité, qu'une semaine choisie au hasard, ce type de panne se produise est 0,05.
Une semaine étant choisie au hasard, on réalise, pour chacun des 300 robots, la même expérience aléatoire consistant à observer si son bras connaît une défaillance mécanique.

On désigne par X la variable aléatoire qui, à toute semaine, associe le nombre de robots dont le bras a connu une panne mécanique.

1) Expliquer pourquoi X suit une loi binomiale ; déterminer les paramètres de cette loi.
2) On approche X par la variable aléatoire Y de loi normale de moyenne $\mu = 15$ et d'écart type $\sigma = 3{,}77$. Justifier le choix des paramètres μ et σ.
Soit E l'événement "en une semaine, strictement plus de 20 robots ont connu une panne mécanique sur leur bras articulé".
3) Calculer, à 10^{-2} près, $P(Y \geq 20{,}5)$ (c'est, en utilisant l'approximation de X par Y, la valeur de P(E)).

6 – Hygiène – Propreté – Environnement 1998 (énoncé modifié)

Un contrôle de la qualité d'une eau de baignade consiste à mesurer le nombre de coliformes contenus dans 100 ml (cent millilitres) de cette eau.
On désigne par X la variable aléatoire qui, à tout prélèvement au hasard de 100 ml d'eau associe le nombre de coliformes, exprimés en nombre entier de milliers, contenus dans cette eau.
On admet que cette variable aléatoire discrète X suit approximativement la loi normale de moyenne $\mu = 4{,}5$ et d'écart type $\sigma = 2{,}4$.
On désigne par Z la variable aléatoire qui suit cette loi normale $\mathcal{N}(4{,}5 ; 2{,}4)$

a) Donner une approximation de la probabilité de l'événement "$X > 4$" en calculant $P(Z \geq 4{,}5)$.

b) Donner une approximation de la probabilité de l'événement "$X > 8$" en calculant $P(Z \geq 8{,}5)$.

c) Calculer approximativement la probabilité d'avoir "$X > 8$" , sachant que l'événement "$X > 4$" est réalisé en calculant $P(Z \geq 8{,}5 / Z \geq 4{,}5)$.

SOMME DE VARIABLES ALEATOIRES DE LOI NORMALE

7 – Industries céréalières 1998

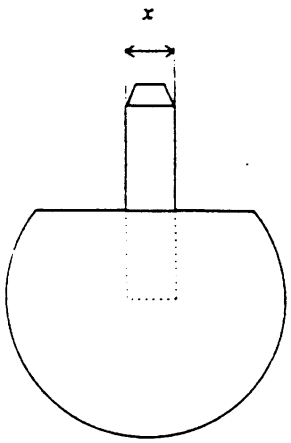
On suppose que le délai entre une commande et sa livraison est de deux semaines.
Soit X_1 (respectivement X_2) la variable aléatoire qui, à la première (respectivement la deuxième) semaine du délai. associe la consommation de blé cette semaine là.
On suppose que X_1 et X_2, sont indépendantes et suivent la loi normale de moyenne 5 et d'écart type 0,2.
a) Quelle est la loi suivie par la somme $Y = X_1 + X_2$? Préciser ses paramètres.

b) Pour quelle valeur du nombre réel y, la probabilité que Y prenne une valeur inférieure à y est-elle égale à 0,99 ?

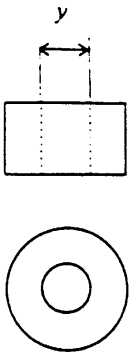
Remarque : dans cet exercice la loi de Y n'est pas donnée.

8 – Productique bois 1998

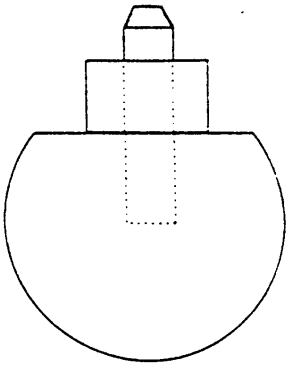
Une entreprise commercialise des pieds de lit de type boule.



PIECE A



PIECE B



ASSEMBLAGE C

Ces pieds sont constitués d'une portion de boule en bois dans laquelle est emboîté un axe métallique de diamètre x (pièce A) et d'une bague en matière plastique (pièce B) de diamètre intérieur y. La pièce A est fabriquée par l'entreprise et la pièce B par un fournisseur.

1° A toute pièce A tirée au hasard dans la production, on associe le diamètre x de son axe métallique, mesuré en millimètres. On définit ainsi une variable aléatoire X.
On admet que X suit la loi normale de moyenne 12 et d'écart type 0,03.

A toute pièce B tirée au hasard dans la production, on associe son diamètre intérieur y mesuré en millimètres. On définit ainsi une variable aléatoire Y.
On admet que la variable aléatoire Y suit la loi normale $\mathcal{N}$ (12,1 ; 0,04).

On note Z la variable aléatoire $Y - X$ qui, à deux pièces A et B tirées au hasard, associe le "jeu" $y - x$.
Pour que le montage des pièces soit possible, il faut que ce jeu soit au moins égal à 0,01 mm. Les pièces A et B étant produites dans des usines différentes, les variables X et Y sont indépendantes.

1° La variable aléatoire Z suit une loi normale. Quels sont ses paramètres ?

2° Calculer la probabilité de l'événement « le montage est possible ».

9 – Construction navale 1997

Les chantiers de l'Atlantique utilisent des remorques pour transporter des pièces des ateliers vers l'aire de prémontage des paquebots.
Une remorque ne peut transporter plus de 490 tonnes.

Les charges sont constituées de n plaques, prises au hasard, et numérotées de 1 à n (n entier naturel non nul). On note Z_i la variable aléatoire qui, à toute pièce numérotée i, choisie au hasard, associe la masse, en tonnes, de cette pièce.
On suppose que les variables aléatoires Z_i ($1 \leq i \leq n$) sont indépendantes et qu'elles suivent la même loi normale de paramètres $m = 0{,}9$ et $\sigma = 0{,}05$.
On note $T_n = Z_1 + Z_2 + ...+ Z_n$ la variable aléatoire qui à tout lot de n pièces, associe la masse totale des n pièces et on suppose que T_n suit une loi normale .

1° On sait que :
L'espérance mathématique de T_n est $E[T_n] = E(Z_1) + E(Z_2) + ...+ E(Z_n)$.
La variance de T_n est $V[T_n] = V(Z_1) + V(Z_2) + ...+ V(Z_n)$.
Préciser, en fonction de n, les paramètres de la loi normale suivie par T_n.

2° Quelle est la probabilité d'être en surcharge si on met 544 pièces sur la remorque ?

1 – Conception des produits industriels 1997

1° a) Pour tout réel t, $e^{-at} > 0$, d'après l'énoncé a > 0 donc $a\,e^{-at} > 0$ et $f(t) \geq 0$.

b) $F(x) = [e^{-at}]_0^x$, $F(x) = 1 - e^{-ax}$,

$\lim\limits_{x\to+\infty} e^{-ax} = 0$, $\lim\limits_{x\to+\infty} F(x) = 1$.

En posant :

$$\begin{cases} u(t) = t & u'(t) = 1 \\ v'(t) = ae^{-at} & v(t) = -e^{-at} \end{cases},$$

$$J(x) = [-t\,e^{-at}]_0^x - \int_0^x -e^{-at}\,dt,$$

$$J(x) = [-t\,e^{-at} - \frac{e^{-at}}{a}]_0^x,$$

$$J(x) = -x\,e^{-ax} - \frac{1}{a}e^{-ax} + \frac{1}{a},$$

$\lim\limits_{x\to+\infty} e^{-ax} = 0$, $\lim\limits_{x\to+\infty} x\,e^{-ax} = 0$, donc

$\lim\limits_{x\to+\infty} J(x) = \frac{1}{a}$.

2° a) $E(Y) = \lim\limits_{x\to+\infty} J(x) = \frac{1}{a}$.

b) $E(Y) = 1000$ équivaut à $\frac{1}{a} = 1000$ et à $a = 0{,}001$; alors $f(t) = 0{,}001\,e^{-0{,}001t}$.

2 – Groupement C 1999

1 a) $f'(t) = -\frac{3t^2}{10^9}e^{-\frac{t^3}{10^9}}$, pour tout réel $t \neq 0$, $e^{-\frac{t^3}{10^9}} > 0$ et $\frac{3t^2}{10^9} > 0$ donc pour tout $t \neq 0$ $f'(t) < 0$ et $f'(0) = 0$,

f est strictement décroissante sur $\mathbb{R}$.

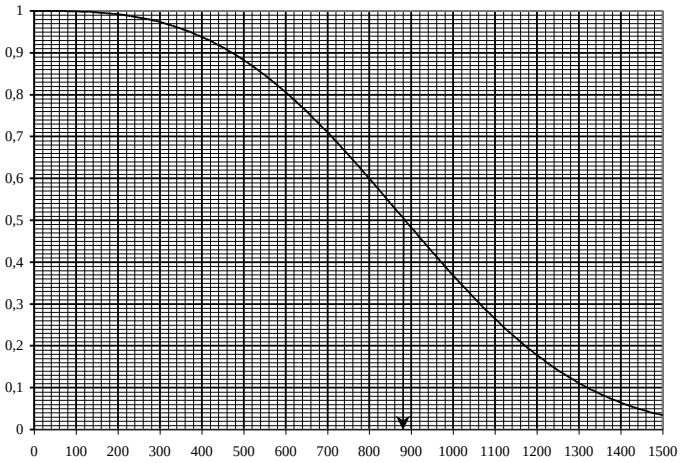
$\lim\limits_{t\to+\infty} -\frac{t^3}{10^9} = -\infty$, $\lim\limits_{t\to+\infty} e^{-\frac{t^3}{10^9}} = 0$ puisque $\lim\limits_{u\to-\infty} e^u = 0$ donc $\lim\limits_{t\to+\infty} f(t) = 0$.

L'axe des abscisses est asymptote à la courbe représentative de f.

b) $\lim\limits_{t\to-\infty} -\frac{t^3}{10^9} = +\infty$, $\lim\limits_{t\to-\infty} e^{-\frac{t^3}{10^9}} = +\infty$ puisque $\lim\limits_{u\to+\infty} e^u = +\infty$ donc $\lim\limits_{t\to-\infty} f(t) = +\infty$.

c)

Représentation graphique.



2° a) Les équations suivantes sont équivalentes dans $\mathbb{R}$:

$f(t) = 0{,}5$, $e^{-\frac{t^3}{10^9}} = 0{,}5$, $-\frac{1}{10^9}t^3 = \ln 0{,}5$,

$t^3 = 10^9 \ln 2$, $t = 10^3 \times \sqrt[3]{\ln 2}$, $t \approx 885$.

b) f est strictement décroissante sur $\mathbb{R}$,

$f(10^3 \times \sqrt[3]{\ln 2}) = 0{,}5$ donc $f(t) < 0{,}5$ équivaut à $t > 10^3 \times \sqrt[3]{\ln 2}$, l'ensemble des solutions de l'inéquation est donc l'intervalle $]\,10^3 \times \sqrt[3]{\ln 2}\,, +\infty[$.

3° a) La probabilité qu'une telle machine fonctionne plus de 1000 heures est $P(T > 1000) = f(1000)$, or

$f(1000) = e^{-\frac{10^9}{10^9}}$,

$P(T > 1000) = e^{-1}$, $P(T > 1000) \approx 0{,}368$.

b) La probabilité qu'une telle machine fonctionne plus de 400 heures est $P(T > 400) = f(400)$.

Il s'agit de vérifier que $f(400) > 0{,}9$ or

$f(400) = e^{-\frac{400^3}{10^9}}$, $f(400) \approx 0{,}93$ donc on a bien une probabilité supérieure à 0,9.

3 – Maintenance Nouvelle Calédonie 1997

1° a) $F(t) = \int_0^t 0{,}005\,e^{0{,}005x}\,dx$,

$F(t) = [e^{-0{,}005x}]_0^t$, $F(t) = 1 - e^{-0{,}005t}$.

b) $J(t) = \int_0^t 0{,}005x\,e^{0{,}005x}\,dx$.

On intègre par parties en posant :

$\begin{cases} u(x) = x & u'(x) = 1 \\ v'(x) = 0{,}005\mathrm{e}^{0{,}005x} & v(x) = \mathrm{e}^{0{,}005x} \end{cases}$.

$J(t) = [-x\,\mathrm{e}^{-0{,}005\,x}]_0^t + 200\,F(t)$,

$J(t) = -\,t\,\mathrm{e}^{-0{,}005\,t} - 200\,\mathrm{e}^{-0{,}005\,t} + 200$.

c) $K(t) = \int_0^t 0{,}005x^2\,\mathrm{e}^{0{,}005x}\,\mathrm{d}x$.

On intègre par parties en posant :

$\begin{cases} u(x) = x^2 & u'(x) = 2x \\ v'(x) = 0{,}005\mathrm{e}^{0{,}005x} & v(x) = \mathrm{e}^{0{,}005x} \end{cases}$,

$K(t) = [-x^2\,\mathrm{e}^{-0{,}005\,x}]_0^t + 400\,J(t)$,

$K(\mathrm{t}) = -\,t^2\,\mathrm{e}^{-0{,}005\,t} - 400\,t\,\mathrm{e}^{-0{,}005\,t} + 80000\,\mathrm{e}^{-0{,}005\,t} + 80000$.

2° $I = \lim\limits_{t\to+\infty} -\mathrm{e}^{-0{,}005\,t}$, $\lim\limits_{t\to+\infty} \mathrm{e}^{-0{,}005\,t} = 0$, $I = 1$.

$E(T) = \lim\limits_{t\to+\infty} (-t\,\mathrm{e}^{-0{,}005\,t} - 2\,\mathrm{e}^{-0{,}005\,t} + 200)$,

$E(T) = 200$.

$E(T^2) = \lim\limits_{t\to+\infty} (-t^2\,\mathrm{e}^{-0{,}005\,t} - 400\,t\,\mathrm{e}^{-0{,}005\,t} + 80000\,\mathrm{e}^{-0{,}005t} + 80000)$ $K = 80000$.

$V(T) = E(T^2) - [\mathrm{E}(T)]^2 = K - 200^2$,

$V(T) = 40000$, $\sigma(T) = \sqrt{\mathrm{V}(T)} = 200$.

3° a) En fiabilité F est la fonction de défaillance, R définie sur $[0, +\infty[$ par $R(t) = 1 - F(t)$ est la fonction de fiabilité.

D'après ce qui précède $F(t) = 1 - \mathrm{e}^{-\,0{,}005\,t}$ donc $R(t) = \mathrm{e}^{-\,0{,}005\,t}$, la variable aléatoire T suit donc une loi exponentielle de paramètre $\lambda = 0{,}005$.

b) $P(T > 300) = R(300) = \mathrm{e}^{-\,0{,}005\,\times\,300}$,

$P(T > 300) = \mathrm{e}^{-\,1{,}5}$,

$P(T > 300) = 0{,}223$ à 10^{-3} près.

$P(T \leq 100) = 1 - \mathrm{e}^{-\,0{,}005\,\times 100}$, $P(T \leq 100) = 1 - \mathrm{e}^{-\,0{,}5}$,

$P(T \leq 100) = 0{,}393$ à 10^{-3} près.

$P(100 < T \leq 300) = \mathrm{F}(300) - \mathrm{F}(100)$,

$P(100 < T \leq 300) = 1 - 0{,}616 = 0{,}384$ à 10^{-3} près.

c) $P(\mathrm{T} \leq t_0) = 0{,}1$ équivaut à $F(t_0) = 0{,}1$ et à $R(t_0) = 0{,}9$ et à $\mathrm{e}^{-\,0{,}005\,t_0} = 0{,}9$ d'où :

$-\,0{,}005\,t_0 = \ln 0{,}9$, $t_0 = -\,200 \ln 0{,}9$, $t_0 \leq 21$.

4 – Groupement D 2001

1° a) On est en présence d'une succession de 20 épreuves indépendantes (tirage avec remise), chacune ayant deux issues : la réponse est juste avec la probabilité 0,5 ; la réponse est fausse avec la probabilité 0,5.

La variable aléatoire Y qui, à tout tirage de 20 cartes, associe le nombre de réponses exactes suit donc la loi binomiale B (20 ; 0,5).

$P(Y = 15) = C_{20}^{15}(0{,}5)^{15}(0{,}5)^5$, $P(Y = 15) = C_{20}^{15}(0{,}5)^{20}$, $P(Y = 15) \approx 0{,}0179$.

b) Si l'on approche Y par une variable aléatoire de loi normale, on doit choisir μ et σ correspondant à la moyenne et à l'écart type de la loi binomiale.

L'espérance mathématique de Y est $\mathrm{E}(Y) = n \times p$, donc $\mu = \mathrm{E}(Y) = 100 \times 0{,}5 = 50$.

L'écart type de Y est $\sigma(Y) = \sqrt{n \times p \times q}$,

donc $\sigma = \sigma(Y) = \sqrt{100 \times 0{,}5 \times 0{,}5} = 5$.

Z suit la loi normale N (50 ; 0,5), $T = \dfrac{Z - 50}{5}$ suit la loi normale N (0, 1).

On a ainsi :

$P(Y > 60) \approx P(Z \geq 60{,}5) = P\left(T \geq \dfrac{60{,}5 - 50}{5}\right)$,

$P(Z \geq 60{,}5) = P\left(T \geq \dfrac{10{,}5}{5}\right)$, $P(Z \geq 60{,}5) = P(T \geq 2{,}1)$,

$P(\mathrm{Z} \geq 60{,}5) = 1 - \Pi(2{,}1)$, $P(Z \geq 60{,}5) = 1 - 0{,}9821$

$P(Z \geq 60{,}5) = 0{,}0179$ à 10^{-4} près, $P(Y > 60) \approx 0{,}0179$ à 10^{-4} près.

5 – Groupement B 2000

1. Chaque semaine, on est en présence de n épreuves aléatoires *indépendantes* pouvant, chacune, déboucher sur *deux issues possibles* : le bras du robot n° i connaît une panne (avec la probabilité $p = 0{,}05$) ou pas.

La variable aléatoire X qui, à chaque semaine, associe *le nombre de robots dont le bras a connu une panne* suit donc la loi binomiale de paramètres $n = 300$ et $p = 0{,}05$.

2. Si l'on approche X par une variable aléatoire de loi normale, on doit choisir μ et σ correspondant à la moyenne et à l'écart type de la loi binomiale.

Donc $\mu = \mathrm{E}(X) = 15$

et $\sigma = \sigma(X) = \sqrt{14{,}25} \approx 3{,}77$ à 10^{-2} près.

3. On pose $Z = \dfrac{Y - 15}{3{,}77}$ qui suit la loi normale N (0 ; 1).

On a ainsi : $P(Y \geq 20{,}5) = \mathrm{P}\left(Z \geq \dfrac{20{,}5 - 15}{3{,}77}\right)$

$P(Y \geq 20{,}5) \approx 1 - \pi(1{,}46)$

$P(Y \geq 20{,}5) \approx 1 - 0{,}9279$

$P(Y \geq 20{,}5) \approx 0{,}07$ à 10^{-2} près.

6 – Hygiène – Propreté – Environnement 1998

La variable aléatoire Z suit la loi normale $\mathcal{N}(4,5 ; 2,4)$, la variable aléatoire $T = \dfrac{Z - 4,5}{2,4}$ suit la loi normale $\mathcal{N}(0, 1)$.

a) $P(Z \geq 4,5) = P(T > 0)$, $P(Z \geq 4,5) = \mathbf{0,5}$ alors que

$P(X > 4) = P(T > -\dfrac{0,5}{2,4})$, $P(X > 4) = \Pi(\dfrac{0,5}{2,4})$,

$P(X > 4) \approx 0,583$ soit environ 58 %.

b) $P(Z \geq 8,5) = P(T > \dfrac{4}{2,4})$,

$P(Z \geq 8,5) = 1 - \Pi(1,67)$,

$P(Z \geq 8,5) \approx 0,0475$ soit moins de 5 % alors que :

$P(X > 8) = P(T > \dfrac{3,5}{2,4})$, $P(X > 8) = 1 - \Pi(\dfrac{3,5}{2,4})$,

$P(X > 8) \approx 0,072$ soit plus de 7 %.

c) $P(Z \geq 8,5 / Z \geq 4,5) = \dfrac{(P(Z \geq 8,5 \text{ et } P(Z \geq 4,5))}{P(Z \geq 4,5)}$,

$P(Z \geq 8,5 / Z \geq 4,5) = \dfrac{(P(Z \geq 8,5)}{P(Z \geq 4,5)}$,

$P(Z \geq 8,5 / Z \geq 4,5) = \dfrac{0,0475}{0,5}$,

$P(Z \geq 8,5 / Z \geq 4,5) = 0,095$ à 10^{-3} près, alors que :

$P(X > 8 / Z > 4) = \dfrac{0,072}{0,5832}$,

$P(X > 8 / Z > 4) = 0,123$ à 10^{-3} près.

7 – Industries céréalières 1998

a) On sait que pour deux variables aléatoires indépendantes X_1 et X_2, suivant les lois normales $\mathcal{N}(m_1, \sigma_1)$ et $\mathcal{N}(m_2, \sigma_2)$ alors la variable aléatoire $X_1 + X_2$ suit la loi normale de moyenne $m = m_1 + m_2$ et d'écart type $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$

X_1 et X_2 suivent la loi normale $\mathcal{N}(5 ; 0,2)$ donc

$m = 5 + 5$, $m = 10$.

$\sigma = \sqrt{0,2^2 + 0,2^2}$, $\sigma(Y) = 0,2 \times \sqrt{2}$.

b) La variable aléatoire Y suit la loi normale $\mathcal{N}(10 ; 0,2\sqrt{2})$ la variable aléatoire $T = \dfrac{Y - 10}{0,2\sqrt{2}}$ suit la loi normale centrée, réduite $\mathcal{N}(0, 1)$.

$P(Y \leq y) = 0,99$ équivaut à $P(T \leq \dfrac{y - 10}{0,2\sqrt{2}}) = 0,99$,

et à $\pi(\dfrac{y - 10}{0,2\sqrt{2}}) = 0,99$ et à $\dfrac{y - 10}{0,2\sqrt{2}} = 2,33$

$y = 10 + 0,2 \times \sqrt{2} \times 2,33$,

$y = 10,66$ à 10^{-2} près.

C'est pour y ≈ 10,66 que l'on a $P(Y \leq y) = 0,99$.

8 – Productique Bois 1998

1° Z est la différence de deux variables aléatoires indépendantes.

X suit la loi normale $\mathcal{N}(12 ; 0,03)$ et Y suit la loi normale $\mathcal{N}(12,1 ; 0,04)$ alors Z suit la loi normale $\mathcal{N}(m, \sigma)$ avec $m = 12,1 - 12$ et

$\sigma = \sqrt{0,4^2 + 0,3^2}$.

Z suit la loi normale de moyenne $m = 0,1$ et d'écart type $\sigma = \sqrt{0,0025}$, $\sigma = 0,05$.

2° Z suit la loi normale $\mathcal{N}(0,1 ; 0,05)$.

La variable aléatoire $T = \dfrac{Z - 0,1}{0,05}$ suit la loi normale centrée réduite $\mathcal{N}(0, 1)$.

$P(Z \geq 0,01) = P(T \geq -\dfrac{0,09}{0,05})$,

$P(Z \geq 0,01) = P(T \geq -1,8)$

$P(Z \geq 0,01) = \pi(1,8)$,

$P(Z \geq 0,01) = 0,964$ à 10^{-3} près.

9 – Construction navale 1997

1° les variables aléatoires Z_i $(1 \leq i \leq n)$ sont indépendantes et suivent la même loi normale de paramètres $m = 0,9$ et $\sigma = 0,05$.

L'espérance mathématique de T_n est

$E[T_n] = E(Z_1) + E(Z_2) + ...+ E(Z_n)$,

$E[T_n] = 0,9\ n$.

La variance de T_n est

$V[T_n] = V(Z_1) + V(Z_2) + ...+ V(Z_n)$,

$V[T_n] = n\ (0,05)^2$.

$\sigma(T_n) = 0,05\sqrt{n}$, la variable aléatoire T_n suit la loi normale $\mathcal{N}(0,9\ n ; 0,05\sqrt{n})$.

2° La variable aléatoire notée T_{544} qui, à tout lot de 544 pièces associe la masse de ces 544 pièces, suit la loi normale $\mathcal{N}(0,9\ n ; 0,05\sqrt{n})$ avec $n = 544$.

T_{544} suit la loi normale $\mathcal{N}(489,6 ; 0,05\sqrt{544})$, la variable aléatoire $U = \dfrac{T_{544} 489,6}{0,05\sqrt{544}}$ suit la loi normale centrée, réduite $\mathcal{N}(0, 1)$.

$P(T_{544} \leq 490) = P(U \leq \dfrac{0,4}{0,05\sqrt{544}})$,

$P(T_{544} \leq 490) = P(U \leq 0,343)$,

$P(T_{544} \leq 490) = \pi(0,343)$,

$P(T_{544} \leq 490) = 0,004$ à 10^{-3} près.

Supplément à la séance n°1
De l'intérêt de l'urne de Bernoulli

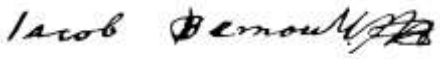
« Si la notion de *vérité statistique* devenait familière à tous ceux qui parlent ou écrivent au sujet de questions où la vérité statistique est la seule vérité, bien des sophismes et bien des paradoxes seraient évités. »

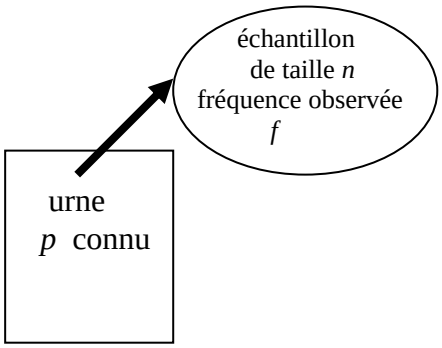
Emile Borel – *Le Hasard* – Alcan, 3e édition, 1914.

1– L'urne bicolore

On considère la situation dite de « l'**urne de Bernoulli** » comprenant deux sortes de boules, noires et blanches, et où la proportion des boules noires est p.
On effectue n tirages au hasard et avec remise dans cette urne. Le résultat est nommé échantillon aléatoire de taille n.

Jacques Bernoulli (1654-1705)





On note X la variable aléatoire correspondant au nombre de boules noires dans un échantillon aléatoire de taille n. Cette variable suit la **loi binomiale** de paramètres n et p dont l'espérance est $E(X) = np$ et l'écart type $\sigma(X) = \sqrt{n\,p\,(1-p)}$.

De manière à pouvoir comparer des tirages de tailles différentes, il est préférable, plutôt que de considérer le nombre de boules noires, d'en considérer la **fréquence**. On introduit donc la variable aléatoire $F = \frac{1}{n} X$.

Pour cette variable aléatoire F, on a comme espérance $E(F) = \frac{1}{n} \times E(X) = \frac{1}{n} \times np = p$ et comme écart type $\sigma(F) = \frac{1}{n}\ \sigma(X) = \frac{1}{n} \times \sqrt{n\,p\,(1-p)} = \sqrt{\frac{p(1-p)}{n}}$.

L'interprétation de ces résultats est que si l'on prélève un grand nombre d'échantillons aléatoires de tailles n et que l'on considère la distribution des fréquences observées, ces fréquences ont pour moyenne p (la fréquence dans l'urne) et pour écart type $\sqrt{\frac{p(1-p)}{n}}$.

C'est bien sûr cet indicateur de dispersion qui rend compte de la qualité d'un échantillon de taille n pour témoigner de la fréquence p dans l'urne. Plus n est grand, plus l'information est précise, mais, ce qui est important c'est que le gain en qualité d'information est en $\frac{1}{\sqrt{n}}$.

2– Inquiétudes à Woburn : le cadre binomial

Woburn est une petite ville industrielle du Massachusetts, au Nord-Est des Etats-Unis. Du milieu à la fin des années 1970, la communauté locale s'émeut d'un grand nombre de leucémies infantiles survenant dans certains quartiers de la ville. Les familles se lancent alors dans l'exploration des causes et constatent la présence de décharges et de friches industrielles ainsi que l'existence de polluants. Dans un premier temps, les experts gouvernementaux concluent qu'il n'y a rien d'étrange. Mais les familles s'obstinent et saisissent leurs propres experts. Une étude statistique montre qu'il se passe sans doute quelque chose « d'étrange ».

Le tableau suivant résume les données statistiques concernant les enfants de Woburn de moins de 15 ans, pour la période 1969-1979 (Sources : *Massachusetts Department of Public Health* et *Harvard University*).

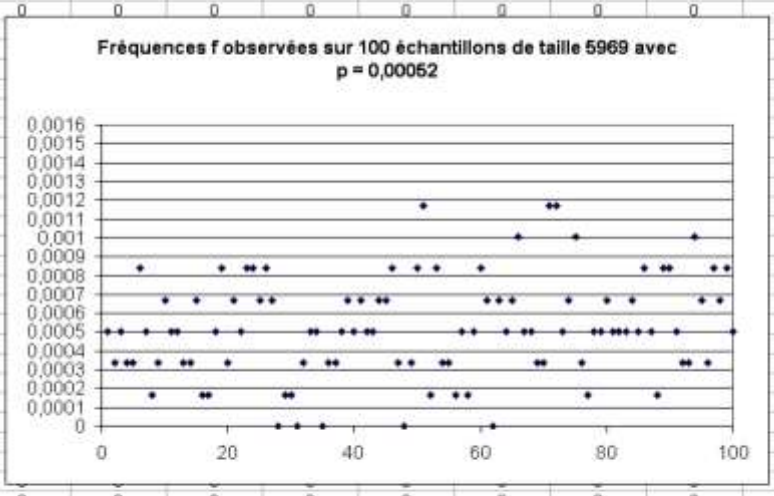
Enfants entre 0 et 14 ans	Population de Woburn selon le recencement de 1970 ***n***	Nombre de cas de leucémie infantile observés à Woburn entre 1969 et 1979	Fréquence des leucémies à Woburn ***f***	Fréquence des leucémies aux Etats-Unis ***p***
Garçons	**5969**	9	**0,00151**	**0,00052**
Filles	**5779**	3	**0,00052**	**0,00038**
Total	11748	12	0,00102	0,00045

La question statistique qui se pose est de savoir si le hasard seul peut raisonnablement expliquer les fréquences observées à Woburn, considérées comme résultant d'un échantillon prélevé dans la population américaine.
La population des Etats-Unis étant très grande par rapport à celle de Woburn, on peut considérer que l'échantillon résulte d'un tirage avec remise et simuler des tirages de taille n avec le tableur.

Dans le cas des garçons, simulons sur le tableur 100 échantillons de taille $n = 5969$ prélevés dans une population où $p = 0{,}00052$. On représente sur un graphique les 100 fréquences f des cas de leucémie observés sur les échantillons simulés. La taille de cette simulation peut demander quelques secondes de calcul, selon la puissance de l'ordinateur. Voici l'affichage que nous avons obtenu.

B2 = =ENT(ALEA()+0,00052)

	A	B	C	D	E	F	G	H	I	J	K	L
1	simulations avec p = 0,00052 (garçons)	échantillon 1	échantillon 2	échantillon 3	échantillon 4	échantillon 5	échantillon 6	échantillon 7	échantillon 8	échantillon 9	échantillon 10	échantillon 11
2		0	0	0	0	0	0	0	0	0	0	0
3		0	0	0	0	0	0	0	0	0	0	0
4		0	0	0	0	0	0	0	0	0	0	0
5		0	0	0	0	0	0	0	0	0	0	0
6		0	0	0	0	0	0	0	0	0	0	0
5942		0	0	0	0	0	0	0	0	0	0	0
5943		0	0	0	0	0	0	0	0	0	0	0
5944		0	0	0	0	0	0	0	0	0	0	0
5945		0	0									0
5946		0	0									0
5947		0	0									0
5948		0	0									0
5949		0	0									0
5950		0	0									0
5951		0	0									0
5952		0	0									0
5953		0	0									0
5954		0	0									0
5955		0	0									0
5956		0	0									0
5957		0	0									0
5958		0	0									0
5959		0	0									0
5960		0	0									0
5961		0	0									0
5962		0	0									0
5963		0	0									0
5964		0	0									0
5965		0	0									0
5966		0	0									0
5967		0	0									0
5968		0	0									0
5969		0	0	0	0	0	0	0	0	0	0	0
5970		0	0	0	0	0	0	0	0	0	0	0
5971	nombre de cas	**3**	**2**	**3**	**2**	**2**	**5**	**3**	**1**	**2**	**4**	**3**
5972	fréquence f	**0,0005026**	**0,00033506**	**0,0005026**	**0,00033506**	**0,00033506**	**0,00083766**	**0,0005026**	**0,00016753**	**0,00033506**	**0,00067013**	**0,0005026**
5973												
5974	nombre d'échantillons où f >= 0,00151			**0**	%							



Cette simulation montre que plus de 95 % des fluctuations aléatoires des valeurs de *f* s'effectuent dans l'intervalle [0 ; 0,001].
L'instruction =NB.SI(B5972:CW5972;">=0,00151") entrée en cellule D5974 confirme qu'ici aucun échantillon simulé n'a montré une fréquence de leucémie infantile chez les garçons atteignant le niveau de 0,00151 observé à Woburn.
On ne peut donc pas raisonnablement attribuer au seul hasard le **niveau très « significativement » élevé des leucémies infantiles observées chez les garçons à Woburn**.

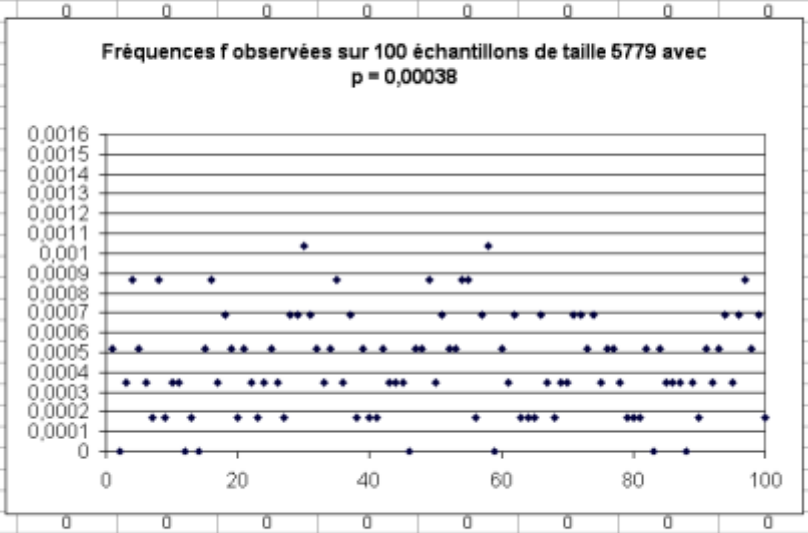
Pour ce qui est des filles, nous simulons de manière analogue sur le tableur 100 échantillons de taille n = 5779 prélevés dans une population où p = 0,00038. L'ordinateur affiche les résultats suivants :

B2 = =ENT(ALEA()+0,00038)

	A	B	C	D	E	F	G	H	I	J	K	L
1	simulations avec p = 0,00038 (filles)	échantillon 1	échantillon 2	échantillon 3	échantillon 4	échantillon 5	échantillon 6	échantillon 7	échantillon 8	échantillon 9	échantillon 10	échantillon 11
2		0	0	0	0	0	0	0	0	0	0	0
3		0	0	0	0	0	0	0	0	0	0	0
4		0	0	0	0	0	0	0	0	0	0	0
5		0	0	0	0	0	0	0	0	0	0	0

D5784 = =NB.SI(B5782:CW5782;">=0,00052")

	A	B	C	D	E	F	G	H	I	J	K	L
5752		0	0	0	0	0	0	0	0	0	0	0
5753		0	0	0	0	0	0	0	0	0	0	0
5754		0	0	0	0	0	0	0	0	0	0	0
5755		0	0	0	0	0	0	0	0	0	0	0
5756		0	0	0	0	0	0	0	0	0	0	0
5757		0	0	0								
5758		0	0	0								
5759		0	0	0								
5760		0	0	0								
5761		0	0	0								
5762		0	0	0								
5763		0	0	0								
5764		0	0	0								
5765		0	0	0								
5766		0	0	0								
5767		0	0	0								
5768		0	0	0								
5769		0	0	0								
5770		0	0	0								
5771		0	0	0								
5772		0	0	0								
5773		0	0	0								
5774		0	0	0								
5775		0	0	0								
5776		0	0	0								
5777		0	0	0								
5778		0	0	0								
5779		0	0	0								
5780		0	0	0	0	0	0	0	0	0	0	0
5781	nombre de cas	**3**	**0**	**2**	**5**	**3**	**2**	**1**	**5**	**1**	**2**	**2**
5782	fréquence f	**0,00051912**	**0**	**0,00034608**	**0,0008652**	**0,00051912**	**0,00034608**	**0,00017304**	**0,0008652**	**0,00017304**	**0,00034608**	**0,00034608**
5783												
5784	nombre d'échantillons où f >= 0,00052			**25**	%							



Cette fois, l'instruction =NB.SI(B5782:CW5782;">=0,00052") entrée en D5784 montre que 25 % des échantillons simulés avec p = 0,00038 font apparaître une fréquence f supérieure ou égale à celle observée avec les données de Woburn. On peut donc penser que le **taux de leucémies infantiles observé chez les filles à Woburn n'est pas « significativement » élevé**. Le hasard pourrait l'expliquer. La taille de l'échantillon est en tout cas trop faible pour mettre en évidence ici un phénomène « anormal ».

Le taux anormalement élevé de leucémies infantiles chez les garçons à Woburn est officiellement confirmé par le Département de Santé Publique du Massachusetts en avril 1980. Les soupçons se portent alors sur la qualité de l'eau de la nappe phréatique qui, par des forages, alimente la ville. On découvre alors le syndrome du trichloréthylène. Les industriels responsables de cette pollution sont traduits en justice, les familles obtiendront des « réparations » financières et la dépollution des sites sera engagée. Suite à cette affaire, le discours du nouveau maire montre bien le changement d'attitude des autorités : « notre première priorité, dira-t-il, est de nous assurer d'avoir un approvisionnement en eau propre et saine ».

Pour approfondir un peu le traitement probabiliste de cet exemple, on peut songer à approcher la loi binomiale.

Les faibles valeurs de p empêchent, dans cet exemple, d'utiliser la formule de fluctuation de plus de 95 % des fréquences au programme de seconde : $[\ p - \frac{1}{\sqrt{n}}\ ,\ \ p + \frac{1}{\sqrt{n}}\ ]$ (voir paragraphe suivant).

En revanche, avec les données concernant les enfants des deux sexes, on peut appliquer l'approximation par la **loi normale**, puisque $n \geq 30$ et $np = 5{,}3 \geq 5$. On sait alors qu'environ 95 % des fréquences obtenues sur les échantillons de taille $n = 11748$ dans une population où $p = 0{,}00045$ fluctuent à l'intérieur de l'intervalle :

$[\ p - 1{,}96 \times \sqrt{\frac{p\,(1-p)}{n}}\ ,\ \ p + \sqrt{\frac{p\,(1-p)}{n}}\ ]$ c'est à dire [0,00032 ; 0,00057]. La fréquence $f = 0{,}00102$ observée avec les enfants de Woburn est loin d'appartenir à cet intervalle.

Lorsque p est trop petit, on peut considérer la variable aléatoire X correspondant au nombre d'observations et qui suit la **loi binomiale** de paramètres n et p. Vu la taille de n, il n'est pas très pratique d'utiliser ici cette loi, que l'on peut en revanche approcher par la **loi de Poisson** de paramètre $n \times p$ qui est bien adaptée aux cas rares. Dans le cas des garçons, on peut ainsi calculer la probabilité d'observer un nombre de leucémies infantiles supérieur où égal à celui de Woburn en utilisant la loi de Poisson de paramètre $n \times p = 5969 \times 0{,}00052 \approx 3{,}1$. On obtient $P\,(\,X \geq 9\,) \approx 0{,}0047$. Ce que l'on peut interpréter en disant que l'explication « il ne se passe rien d'étrange » a moins de 0,5 % de chances d'être exacte.

Qui peut encore, après un tel exemple, affirmer que « la statistique est la forme la plus élaborée du mensonge » ?

3 – Le cas Aamjiwnaag : émergence de l'approximation normale

Pour préciser le comportement de la variable aléatoire d'échantillonnage F, on peut, dans certains cas, recourir à la **loi normale**. En effet, pour n « assez grand » (dans la pratique, on prend $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$), la loi binomiale donne des résultats proches de ceux d'une loi normale. On peut donc considérer que la variable aléatoire F suit approximativement une loi normale de moyenne p et d'écart type $\sqrt{\frac{p(1-p)}{n}}$.

On sait que la loi normale a comme propriété qu'environ **95 % des observations** se situent dans un intervalle de rayon deux écarts types autour de la moyenne. On pourra donc vérifier qu'environ 95 % des échantillons aléatoires de taille n fournissent une fréquence comprise dans l'intervalle $[\ p - 2 \times \sqrt{\frac{p(1-p)}{n}}\ \ ,\ \ p + 2 \times \sqrt{\frac{p(1-p)}{n}}\ ]$. Ce résultat est très important car il mesure la **variabilité « naturelle » des phénomènes aléatoires**.

On peut donner une version simplifiée de cet intervalle, en le majorant.

La fonction $p \mapsto p(1-p)$ atteint son maximum pour $p = \frac{1}{2}$ donc, pour tout p, on a :

$p(1-p) \leq \frac{1}{2} \times (1 - \frac{1}{2})$ c'est-à-dire $p(1-p) \leq \frac{1}{4}$.

On en déduit que $2 \times \sqrt{\frac{p(1-p)}{n}} \leq 2 \times \sqrt{\frac{1}{4\,n}}$ c'est-à-dire $2 \times \sqrt{\frac{p(1-p)}{n}} \leq \frac{1}{\sqrt{n}}$.

Ainsi, l'intervalle $[\ p - 2 \times \sqrt{\frac{p(1-p)}{n}}\ ,\ p + 2 \times \sqrt{\frac{p(1-p)}{n}}\ ]$ est inclus dans l'intervalle $[\ p - \frac{1}{\sqrt{n}}\ ,\ p + \frac{1}{\sqrt{n}}\ ]$.

Les élèves de seconde peuvent expérimenter qu'environ plus de 95 % des échantillons de taille n fournissent une fréquence comprise dans l'intervalle $[\ p - \frac{1}{\sqrt{n}}\ ,\ p + \frac{1}{\sqrt{n}}\ ]$.

Aamjiwnaag

Sources : Science et Vie février 2006 – Environnemental Health Perspectives octobre 2005 (article en ligne).

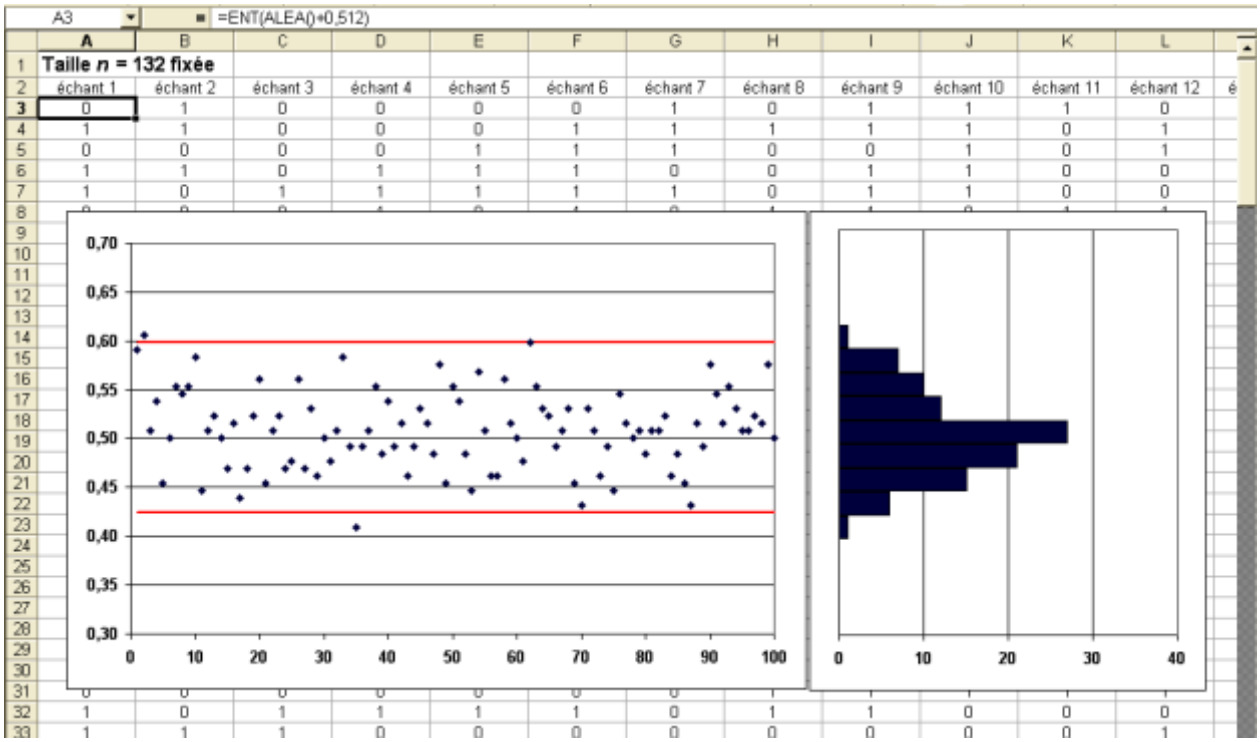
Les données proviennent d'une étude effectuée au Canada et montrant une différence (très) significative du sex-ratio à la naissance (déficit de garçons) sur une population exposée à une pollution chimique. Dans ce cas particulier, l'inquiétude provient du fait que bien que ces industries canadiennes respectent les normes, une exposition prolongée à de faibles doses de polluants puisse avoir un impact sanitaire mesurable.

Le « sex-ratio » est le rapport du nombre de garçons à celui des filles à la naissance. Il est habituellement de 105 garçons pour 100 filles.

Dans la réserve indienne d'Aamjiwnaag, située au Canada, il est né entre 1999 et 2003, n =132 enfants dont 46 garçons.

Question statistique : la fréquence des garçons observée à Aamjiwnaag pour la période 1999-2003 présente-t-elle une « différence significative » au niveau 0,95 avec p = 0,512 ?

Etude des fluctuation des échantillons de taille n = 132



L'image ci-dessus correspond à 100 simulations d'un échantillon de taille 132 extraits au hasard avec remise d'une urne où la proportion du caractère considéré est p = 0,512.

L'histogramme correspond à l'observation des 100 fréquences d'échantillons. Il s'agit de la répartition de 100 valeurs observées de la variable aléatoire F. Cette répartition empirique est

à comparer avec le modèle de loi normale de moyenne $p = 0{,}512$ et d'écart type $\sqrt{\frac{0{,}512\times(1-0{,}512)}{132}} \approx 0{,}044$ donné par le théorème limite central.

La fréquence observée à Aamjiwnaag est $f = \frac{46}{132} \approx 0{,}348$.

Cette observation est très éloignée de la bande (apparaissant sur le graphique) des fluctuations dans 95 % des cas sous l'hypothèse $p = 0{,}512$.
La différence observée est significative au seuil de 5 %.
Sous l'hypothèse que F suit la loi normale de moyenne $p = 0{,}512$ et d'écart type 0,044 on a : $P(F \geq 0{,}410) \approx 0{,}99$. La valeur observée à Aamjiwnaag est donc inférieure à ce que l'on observe dans 99 % des cas sur des échantillons de taille 132 lorsque $p = 0{,}512$.

L'étude statistique n'établit pas de causalité entre la pollution chimique et le « défaut » observé du sex-ratio. A la question « Que peut-on tirer comme conclusion ? », on peut seulement ici répondre que « cette étude pose question ». La statistique donne l'alerte, ce qui est déjà beaucoup. Le fait que la réserve soit située au cœur d'industries chimiques devient un élément troublant sur lequel on doit enquêter.
De façon générale, c'est la notion de « preuve statistique » qui est ici en jeu. Il ne s'agit pas d'une « preuve » au sens habituel mais d'un élément probant, d'une présomption. Plutôt que de parler de « preuve statistique », les anglo-saxons disent plus justement *piece of evidence*.
D'autres explications possibles du déséquilibre du sex-ratio pourraient être liées au mode de vie de ces indiens ou à leur patrimoine génétique. Une étude statistique comparative a été menée sur des indiens de la même tribu vivant dans un autre environnement et a (dé)montré que ce n'était (sans doute) pas le cas. En revanche l'influence de certains produits chimiques sur le sex-ratio a été établie « statistiquement » par d'autres études.
Une recherche sur Internet permettra d'avoir d'autres éléments sur ce dossier (qui a fait polémique au Canada).

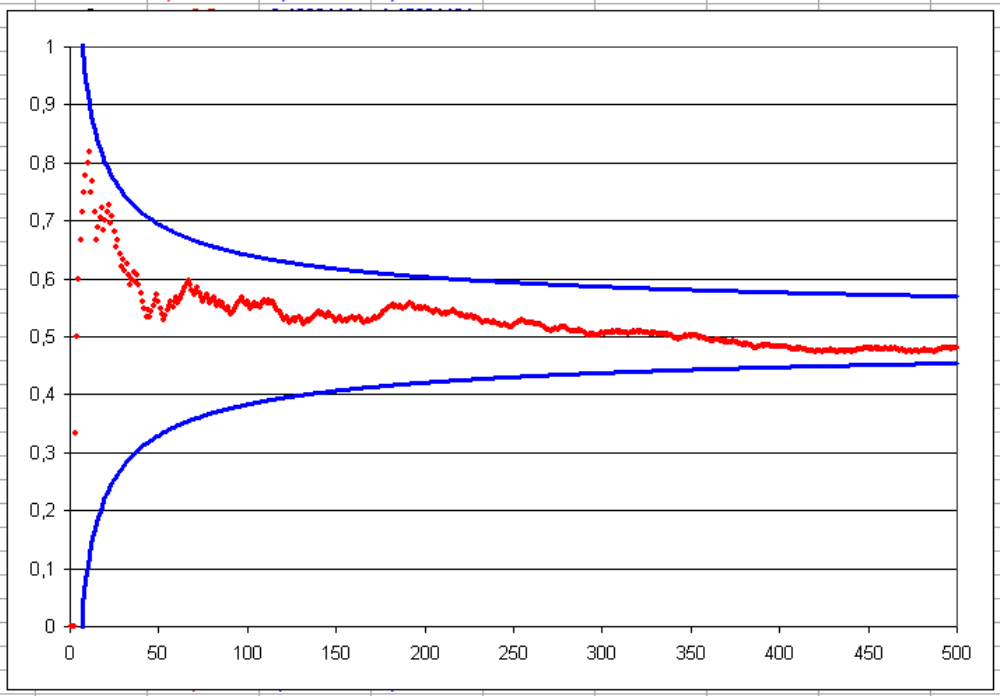
Influence de la taille de l'échantillon : convergence quand n augmente

On peut expérimenter, en fonction de la taille n des prélèvements, la « convergence » (presque sûre !) en œuvre dans la loi des grands nombres.
On entre en B2 la formule
`=ENT(ALEA()+0,512)`
que l'on recopie vers le bas jusqu'à la ligne 501.

Le cumul des résultats est obtenu en colonne C en entrant en C2 la formule : `=B2`
puis en C3 la formule `=C2+B3` que l'on recopie vers le bas jusqu'à la ligne 501.
Les fréquences sont obtenues en colonne D, en entrant en D2 la formule : `=C2/A2`
que l'on recopie vers le bas.
Pour le rang n, les bornes de l'intervalle de fluctuation de 99 % des fréquences sont calculées en colonne E et F.
En E2 on entre la formule
`=0,512-2,58*RACINE(0,512*(1-0,512)/A2)`
que l'on recopie vers le bas.
En E2 on entre la formule
`=0,512+2,58*RACINE(0,512*(1-0,512)/A2)`
que l'on recopie vers le bas.

B2 = =ENT(ALEA()+0,512)

	A	B	C	D	E	F	G	H	I	J	K
1	rang	alea	cumul	fréquence (cumulée)	borne inf à 99%	borne sup à 99 %					
2	1	0	0	0	-0,77762843	1,80162843					
3	2	0	0	0	-0,39990501	1,42390501					
4	3	1	1	0,33333333	-0,23256732	1,25656732					
5	4	1									
6	5	1									
7	6	1									
8	7	1									
9	8	1									
10	9	1									
11	10	1									
12	11	1									
13	12	0									
14	13	1									
15	14	0									
16	15	0									
17	16	1									
18	17	1									
19	18	1									
20	19	0									
21	20	1									
22	21	1									
23	22	1									
24	23	0									
25	24	1									
26	25	0									
27	26	0									
28	27	1									
29	28	0									
30	29	0									
31	30	1									
32	31	0									
33	32	1									



D’autres simulations sont obtenues en faisant F9.

Séance 2 :
ECHANTILLONNAGE
ET ESTIMATION

"Mieux vaut prévoir sans certitude que de ne pas prévoir du tout."
Henri Poincaré – *"Science et hypothèse"*.

"Il est toujours dangereux d'appliquer des formules dont on n'a bien saisi ni le sens profond ni l'esprit."
Lucien March,
"Les principes de la méthode statistique" – 1930.

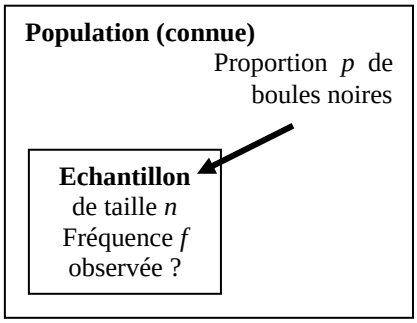
A – ECHANTILLONNAGE

"L'esprit statistique naît lorsque l'on prend conscience des fluctuations d'échantillonnage."
Programme de seconde 2000.

C'est la connaissance de la variabilité naturelle d'un phénomène, due au seul hasard, qui permet la méthode statistique. Cette connaissance de la variabilité "normale" conduit, dans le cadre de l'estimation, à définir la "*confiance*" à accorder aux estimations, et, dans le cadre des tests, à calculer des limites au delà desquelles les variations seront considérées comme suffisamment *"significatives*" pour ne pas être dues au seul hasard.
C'est donc en se fondant sur l'étude des fluctuations d'échantillonnage, que nous pourrons conduire nos prises de décision statistiques.
Nous commencerons par la situation des fréquences, plus simple parce que ne mettant en jeu qu'un seul paramètre, bien que ce cas puisse être vu comme cas particulier de la moyenne.

I – FLUCTUATIONS D'ECHANTILLONNAGE D'UNE FREQUENCE

1 – Le problème de l'échantillonnage

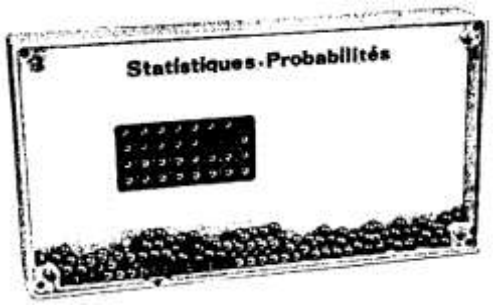


Une urne (connue) contient des boules blanches et des boules noires en proportion p.
On souhaite étudier comment fluctuent les fréquences f de boules noires, observées sur des échantillons (sondages) de taille n.
Dans la pratique, se pose le problème inverse (inférer à partir d'un échantillon, par exemple pour un contrôle de qualité), mais l'étude de l'échantillonnage est nécessaire d'un point de vue théorique. C'est sur cette théorie que sera construite l'estimation.

2 – Expérimentation et théorie des fluctuations d'échantillonnage d'une fréquence

Les activités de se paragraphes peuvent se pratiquer avant le cours sur l'échantillonnage. Seule la connaissance de la loi binomiale et de son approximation par la loi normale sont nécessaires.

a – Une expérimentation "physique" dans le cadre d'un échantillonnage sans remise



On peut fabriquer un appareil, semblable à l'appareil ci-contre, contenant 200 billes dont 10 billes noires.

La "population" est donc connue, avec $p = 0{,}05$.
En secouant, on vient remplir les $n = 32$ alvéoles qui matérialisent un échantillon. On observe alors une fréquence f de billes noires parmi les 32 billes.

Sur l'image ci-dessus, par exemple, $f = \frac{3}{32} \approx 0{,}09$.

Il s'agit bien sûr d'un échantillonnage ***exhaustif*** (sans remise).

Soit X la variable aléatoire 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.
La variable aléatoire X suit la loi ***hypergéométrique*** H(N , n , p), avec $P(X = k) = \frac{C_{Np}^{k} C_{N-Np}^{n-k}}{C_N^n}$.

NORME AFNOR

5.4 PLAN D'ECHANTILLONNAGE

Plan selon lequel on prélève un ou plusieurs échantillons en vue d'une information à recueillir et éventuellement d'une décision à prendre.

5.5 ECHANTILLONNAGE AU HASARD

Prélèvement de n *individus dans une population de* N *individus de telle manière que toutes les combinaisons possibles de* n *individus aient la même probabilité d'être prélevées.*

5.6 ECHANTILLONNAGE EXHAUSTIF (sans remise)

La loi hypergéométrique n'est pas au programme des sections de B.T.S., cependant l'expérimentation physique en classe laisse aux étudiants des images mentales très fortes concernant la distribution d'échantillonnage.
On pourra comparer le nombre moyen de billes noires dans un échantillon avec l'espérance $E(X) = np = 1{,}6$ et l'écart type avec $\sigma(X) = \sqrt{\frac{N-n}{N-1}\, np(1-p)} \approx 1{,}13$ (l'écart type avec la loi binomiale serait d'environ 1,23).

La variable aléatoire $F = \frac{1}{32} X$ qui, à tout échantillon associe la fréquence observée des boules noires, a pour espérance $E(F) = \frac{E(X)}{32} = 0{,}05$ et pour écart type $\frac{\sigma(X)}{32} \approx 0{,}035$.

b – Simulation avec la touche random (ou ALEA) dans le cadre d'un échantillonnage avec remise

Avec une calculatrice

En mode "normal" de calcul, taper l'instruction :
Int (Ran# + 0.05) ou int (rand + 0.05) ou int(rand() + 0.05) , selon les modèles, puis faire plusieurs fois EXE ou ENTER.

⇨ Sur T.I. : ***int*** s'obtient par (MATH puis NUM) et ***rand*** par MATH puis PRB.
⇨ Sur CASIO : ***Int*** s'obtient par (OPTN puis NUM) et ***Ran#*** par OPTN puis PRB.

Le résultat affiché est 1 (= boule noire) avec une fréquence de 0,05 ou 0 (= boule blanche) avec une fréquence de 0,95.
Le programme suivant simule le calcul de f sur un échantillon de taille $n = 32$, ***prélevé avec remise***.

Commentaires	CASIO (anciens modèles de la fx 7000G à la CFX 9900GC)	CASIO 6910G - 9930 - 9940 - 9960 – 9990 et Graph xx	T.I. 81	T.I. 80 -82 - 83 - 85	T.I. 89 - 92
S compteur des boules noires. I compteur des 32 tirages 1 = boule noire ; 0 = boule blanche. Affichage de la fréquence f de l'échantillon.	0 → S ↵ 1 → I ↵ Lbl 1 ↵ Int (Ran# + 0.05) + S → S ↵ I + 1 → I ↵ I ≤ 32 ⇒ Goto 1 ↵ S ÷ 32	0 → S ↵ For 1 → I To 32 ↵ Int (Ran# + 0.05) + S → S ↵ Next ↵ S ÷ 32	:0 → S :1 → I :Lbl 1 :int (rand + 0.05) + S → S :I + 1 → I :If I ≤ 32 :Goto 1 :Disp S ÷ 32	:0 → S :For (I,1,32) :int (rand + 0.05) + S → S :End :Disp S ÷ 32	:0 → s :For i,1,32 :int (rand() + 0.05) + s → s :End :Disp s ÷ 32

Avec Excel

C'est extrêmement simple, puisqu'il n'y a pas de programmation à faire.
Dans la cellule A1, on entre la formule : =ENT(ALEA()+0,05)
que l'on recopie jusqu'à la cellule A32, pour simuler un échantillon de taille 32, ***prélevé avec remise***.
En A33, on peut calculer la fréquence de l'échantillon en faisant =SOMME(A1:A32)/32 .
Il suffit alors de recopier la première colonne vers la droite pour simuler d'autres échantillons.

Introduction d'une variable aléatoire

On supposera dorénavant que l'échantillonnage se fait avec remise (exhaustif) pour rester dans le cadre des programmes. La norme AFNOR, indiquée ci-après précise que l'on peut raisonnablement faire cette hypothèse si la proportion entre la taille *n* de l'échantillon et celle de la population est inférieure à 10 % (ce n'était pas le cas de la machine à billes).

La variable aléatoire *X* qui, à tout échantillon de taille *n* prélevé au hasard avec remise, associe le nombre de boules noires observé, suit la ***loi binomiale*** $\mathcal{B}(n, p)$.

On appelle ***distribution d'échantillonnage*** *de la fréquence* des boules noires dans les échantillons de taille *n*, la loi de probabilité de la variable aléatoire $F = \frac{1}{n} X$.

On a $E(F) = \frac{1}{n} E(X) = p$,

et $\sigma(F) = \frac{1}{n} \sigma(X) = \sqrt{\frac{p(1-p)}{n}}$.

On peut, avec les élèves, expérimenter la distribution d'échantillonnage avec les calculatrices, dans l'exemple précédent (en parallèle avec l'instrument à bille), en restant dans le cadre de la loi binomiale (on est d'ailleurs dans un cas où l'approximation normale n'est pas valable puisque $np = 32 \times 0{,}05 = 1{,}6$ est inférieur à 5).

On inscrit au tableau les fréquences f qu'ils ont obtenues sur leurs échantillons simulés.

On comparera la moyenne des fréquences des billes noires sur au moins une vingtaine d'échantillons avec l'espérance de F $E(F) = \frac{E(X)}{32} = \frac{32 \times 0{,}05}{32} = 0{,}05$ et l'écart type entre les différents échantillons avec l'écart type de F $\sigma(F) = \frac{\sigma(X)}{32} = \frac{\sqrt{32 \times 0{,}05 \times 0{,}95}}{32} \approx 0{,}038$.

NORME AFNOR

NORME STATISTIQUE 5.6 (suite)
Lorsque chaque individu prélevé est remis dans la population avant prélèvement de l'individu suivant, l'échantillonnage est dit NON EXHAUSTIF (avec remise).
NOTE : Dans un échantillonnage exhaustif et au hasard, l'indépendance des individus tirés peut être considérée comme pratiquement

Remarque (non destinée aux élèves) :

Pour que l'expérience fonctionne à peu près bien (il y a toujours un peu de suspense, mais ça rend la statistique vivante), un minimum de 20 échantillons est nécessaire. Examinons cela sur le cas des moyennes de 20 fréquences, que l'on espère très proche de $p = 0{,}05$.

Soit la variable aléatoire $\overline{F} = \frac{1}{20} \sum_{i=1}^{20} F_i$ où F_i est la variable aléatoire correspondant à la fréquence obtenue sur le i ème échantillon. On suppose les variables aléatoires F_i indépendantes (il ne faut pas faire cette expérience avec un lot de calculatrices neuves de même type où le random sera à peu près analogue !).

On a donc $Var(\overline{F}) = \frac{1}{20^2} \sum Var(F_i) = \frac{\sigma^2}{20}$ d'où $\sigma(\overline{F}) = \frac{\sigma}{\sqrt{20}} \approx \frac{0{,}038}{\sqrt{20}} \approx 0{,}008$.

Un écart type de 0,008 sur une valeur à mettre en évidence qui est 0,05 c'est encore beaucoup.

On peut dire ensuite que pour simplifier les calculs (manipulation des coefficients du binôme...encore que les calculatrices soient très performantes) lorsque *n* est grand (ou que

l'on cherche à le déterminer car alors on a besoin d'une expression pratiquable ce qui n'est pas le cas de la formule du binôme), on utilisera l'approximation normale de la loi binomiale :

D'après le ***théorème limite central*** (cas du théorème de *Moivre – Laplace*), pour n assez grand (et p et $1 - p$ pas trop petits : on donne parfois np et $n(1 - p)$ supérieurs à 5),
la variable aléatoire X suit approximativement la loi normale $\mathrm{N}\left(np, \sqrt{np(1-p)}\right)$ et donc

F suit approximativement la loi normale $\mathrm{N}\left(p, \sqrt{\dfrac{p(1-p)}{n}}\right)$.

Ceci sera mis en pratique dans l'exemple suivant.

3 – Un exemple utile au citoyen

Voici un exemple déjà historique : 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 (avec l'élection présidentielle de 2002 ça marche moins bien).
On effectue un sondage sur 1000 électeurs, le jour de l'élection.
1) Soit F la variable aléatoire qui, à tout échantillon aléatoire non exhaustif de taille 1000, associe la fréquence des électeurs de Giscard sur l'échantillon. On peut supposer que F suit une loi normale, en déterminer les paramètres.

On a $\sqrt{\dfrac{0{,}4825 \times 0{,}5175}{1000}} \approx 0{,}0158$.

La variable aléatoire F suit donc (approximativement) la loi normale N (0,4825 ; 0,0158).

Ce qui veut dire qu'entre deux sondages aléatoires de taille 1000, l'écart type est de 1,6 % ce qui, dans la situation présente, est énorme.

2) Quelle est la probabilité que ce sondage se trompe en donnant plus de 50 % de votants pour Giscard (on calculera $P\,(F > 0{,}5)$).

On pose $T = \dfrac{F - 0{,}4825}{0{,}0158}$ qui suit la loi N (0, 1).

On a $P(F > 0{,}5) = P(T > 1{,}1076) = 1 - \Pi(1{,}1076)$ où Π est la fonction de répartition de la loi normale centrée réduite.

On trouve $P\,(F > 0{,}5) \approx 0{,}1340$.

On peut remarquer que la calculatrice TI83 ne peut ici faire le calcul avec la loi binomiale B (1000 ; 0,4825).

3) Quelle devrait-être la taille n du sondage, pour que cette probabilité soit inférieure à 0,01 ?

On pose $T = \dfrac{F - 0{,}4825}{\sqrt{\dfrac{0{,}4825 \times 0{,}5175}{n}}}$ qui suit la loi N (0, 1).

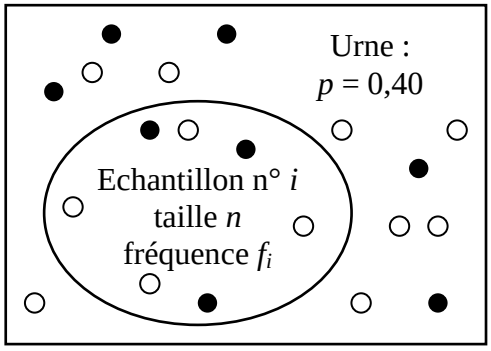
On a $P(F > 0{,}5) = P\left(T > \dfrac{0{,}0175}{\sqrt{\dfrac{0{,}4825 \times 0{,}5175}{n}}}\right) = 1 - \Pi\left(\dfrac{0{,}0175\sqrt{n}}{\sqrt{0{,}4825 \times 0{,}5175}}\right)$.

On recherche t tel que $P(T > t) = 0{,}01 \Leftrightarrow \Pi(t) = 0{,}99$.

La lecture inverse de la table donne $t \approx 2{,}33$ (l'instruction invNorm(0.99) de la calculatrice donne 2,3263…). D'où $n > \left(\dfrac{2{,}327}{0{,}0175}\right)^2 \times 0{,}4825 \times 0{,}5175$.

Soit $n \approx 4415$.
Pour ce type de calcul, l'approximation normale est nécessaire, sauf à utiliser un tableur.

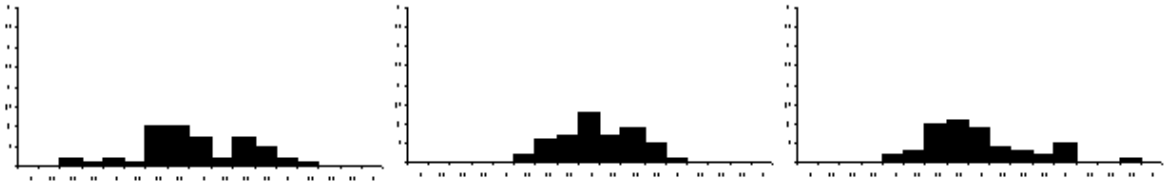
4 – Expérimentation sur Excel de la normalité de la distribution d'échantillonnage



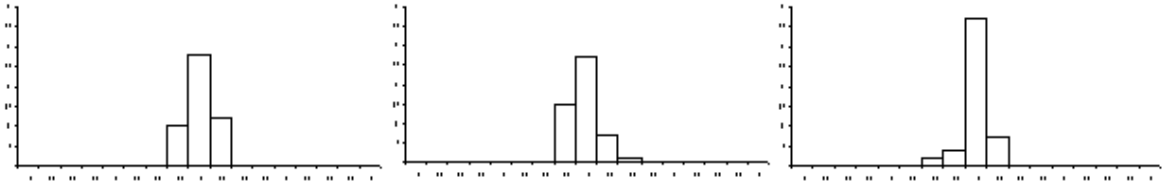
On considère une urne qui contient 40% de boules noires (p = 0,40) et 60% de boules blanches.
On prélève dans cette urne, au hasard et avec remise, des échantillons de taille n .
Pour chaque échantillon i, on calcule la fréquence f_i des boules noires.
On désire étudier comment se répartissent les fréquences f_i des différents échantillons.

Influence de la taille n de l'échantillon, sur les fluctuations d'échantillonnage

On peut comparer la moyenne et surtout l'écart type d'échantillonnage, observés sur 50 échantillons, avec $E(F) = p = 0{,}4$ et $\sigma(F) = \sqrt{\frac{p(1-p)}{n}}$, qui vaut environ 0,049 lorsque n = 100 et 0,015 lorsque n = 1000.
On observera donc qu'en multipliant par 10 la taille n de l'échantillon, on divise par $\sqrt{10} \approx 3$ l'écart type de la distribution d'échantillonnage.
Sur les graphiques ci-dessous, on a regroupé, à la *même échelle*, les fréquences observées f_i en classes d'amplitude 0,025.



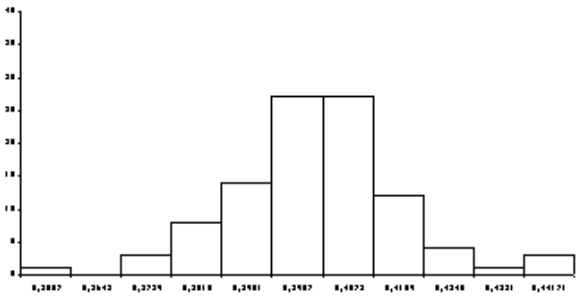
50 échantillons de taille n = 100

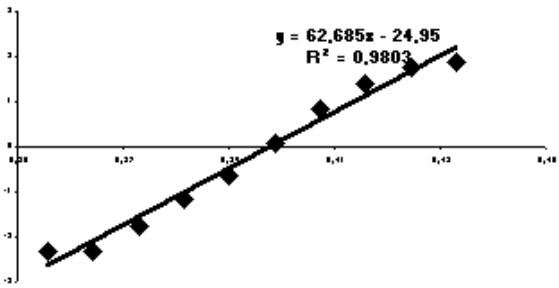


50 échantillons de taille n = 1000

Etude de la normalité de la distribution d'échantillonnage

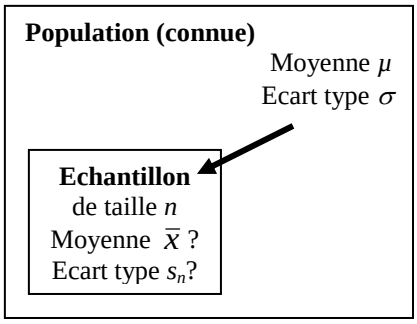
On a regroupé ci-après les fréquences f_i observées sur 100 échantillons de taille n = 1000, en 11 classes. Puis on a tracé la *droite de Henry* correspondante.
Le coefficient de corrélation linéaire est un témoin de la normalité de la distribution d'échantillonnage.





II - FLUCTUATIONS D'ECHANTILLONNAGE D'UNE MOYENNE

1 – Distribution d'échantillonnage d'une moyenne



On considère une population de moyenne μ et d'écart type σ.

On note X_i la variable aléatoire associant, à tout échantillon de taille n prélevé au hasard avec remise, la valeur obtenue au i^{ème} tirage, pour $1 \leq i \leq n$.

On désigne par $\overline{X}$ la variable aléatoire $\overline{X} = \frac{1}{n}\sum_{i=1}^{i=n} X_i$ qui, à tout échantillon, associe la moyenne observée.
Les variables aléatoires X_i sont indépendantes et de même loi (tirages avec remise).

D'après le ***théorème limite central***, pour n assez grand,

$\overline{X}$ suit approximativement la loi normale $\mathbb{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Remarques :

- Dans la pratique, pour l'utilisation de ce théorème limite, on demande généralement $n \geq 30$.
- Si la population est normale, c'est à dire si les X_i suivent la loi normale $\mathbb{N}(\mu, \sigma)$, alors $\overline{X}$ suit (exactement) la loi normale $\mathbb{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, même si n est petit.
- Si n est petit ($n < 30$) et que la population n'est pas normale, l'approximation de ***Student*** est préférable (voir plus loin).

2 – Exemple d'application aux cartes de contrôle

Un exemple intéressant d'utilisation des fluctuations d'échantillonnage est celui de l'établissement de "cartes de contrôle" dans le cadre du contrôle de qualité.

⇨ ***Voir l'exercice de B.T.S. corrigé "Construction navale 1998".***

⇨ ***Voir le T.P. en annexe sur "La maîtrise statistique des procédés de production."***

B – ESTIMATION

"Rien de plus aisé que d'inventer des méthodes d'estimation. [Reste à] distinguer les méthodes satisfaisantes de celles qui ne le sont pas."

R.A. Fisher,
"Les méthodes statistiques adaptées à la recherche scientifique" – Traduction 1947.

La statistique connut, au début du XX^{e} siècle, une véritable révolution, celle de l'**inférence**. Il s'agit, à partir de résultats statistiques limités, observés sur un échantillon, d'inférer à toute une population pour laquelle on induira des estimations, à partir des observations. Cette démarche trouve sa motivation dans des domaines appliqués spécifiques (laboratoires agronomiques, industriels...), essentiellement en Grande Bretagne et aux Etats-Unis, mais les outils ainsi créés trouveront, de part leur efficacité, des applications quasi universelles.
Depuis le XVIII^{e} siècle, deux approches des probabilités coexistent : le point de vue *subjectif* (représenté par *Bayes* et *Laplace*), où l'on considère la probabilité en termes de "raisons de croire", d'estimation du degré de confiance dans la réalisation d'un événement aléatoire et le point de vue *fréquentiste*, s'appuyant sur la loi des grands nombres de *Bernoulli* et les théorèmes limites (*Laplace*), où l'on conçoit la probabilité comme stabilisation limite de la fréquence lors de la répétition, un grand nombre de fois, de l'évènement. Au XIX^{e} siècle, *Quételet* et ses successeurs ne retiendront de la synthèse de *Laplace* que l'aspect fréquentiste, laissant en sommeil les techniques d'estimations amorcées au XVIII^{e}.
Les formulations probabilistes de l'estimation réapparaissent, dans la mouvance de l'école de *Karl Pearson*, avec *Ronald Aylmer Fischer* (1890 – 1962) et *William Sealy Gosset* – alias *Student* – (1876 – 1937), tous deux confrontés à un nombre insuffisant de données, rendant plus difficile la perspective fréquentiste. *Fischer*, travaillant à l'étude des engrais dans un centre agronomique, ne peut recourir qu'à un nombre limité d'essais contrôlés, *Gosset*, à la brasserie Guinness de Dublin, ne dispose, pour ses études de qualité, que d'échantillons de petite taille, en raison de la grande variabilité de la production qui n'en permet pas l'homogénéité. *Fischer* et *Gosset* sont alors amenés à utiliser des notations différentes pour la valeur théorique θ du paramètre d'une distribution de probabilité, et pour l'estimation $\hat{\theta}$ de ce paramètre, aux vues des observations. Cette innovation dans les notations rend possible le développement de la statistique inférentielle dans deux directions, celle de l'estimation (dominée par un esprit subjectif) et celle de la théorie des tests d'hypothèses (s'inscrivant davantage dans la tradition fréquentiste) que l'on abordera à la 3ème séance.


R.A. Fisher

Ronald Fischer est généralement présenté comme le père de la théorie de l'**estimation selon le principe du *"maximum de vraisemblance"*** (introduit en 1912 et développé jusqu'en 1922-1925). Il présente cette notion de vraisemblance comme descendante de celle de "probabilité inverse" de *Laplace*.

Fisher est, selon Droesbeke et Tassi[4], *"l'homme qui a fait de la statistique une science moderne"*. Son ouvrage *"Statistical Methods for Research Workers"*, publié en 1925 sera le "best-seller" de la statistique, avec 14 éditions et des traductions en 6 langues.

Jerzy Neyman (1894 – 1981), est quant à lui à l'origine de la théorie de l'**estimation par intervalle de confiance**. Mathématicien d'origine russo-polonaise, *Neyman* fait un séjour d'étude en 1924 à l'University College qui le met en contact avec *Fischer, Gosset, Karl* et *Egon Pearson* (avec lequel il correspondra beaucoup). En 1925, il assiste à Paris aux cours de *Lebesgue*, *Hadamard* et *Borel*. De retour en Pologne, il assure la direction du département statistique d'un institut de biologie. En 1934, il est en poste à l'University College aux côtés d'*Egon Pearson* mais en 1938, à l'approche de la guerre, il part pour *Berkeley*, aux Etats-Unis.

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

Le premier, *Neyman* prend nettement parti pour la méthode aléatoire. Le hasard seul permettant d'appliquer la théorie des probabilités et d'encadrer le risque, alors que la raison humaine est vecteur d'introduction de biais. En 1934, *Neyman* montre que les hypothèses garantissant la convergence des estimateurs obtenus par "choix raisonné" ne peuvent être obtenues dans la pratique. De 1934 à 1937, il expose la théorie de l'estimation par intervalles de confiance.

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

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

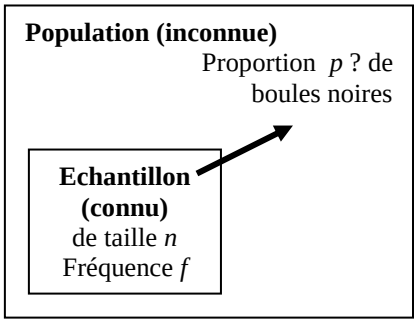
Il faut dire que le sondage du *Literary Digest* avait été effectué à partir de listes du type annuaire du téléphone (en 1936 !), membres de clubs... ce qui introduisait un biais certain. *Alf Landon*, dont on a aujourd'hui oublié le nom, a pu ainsi se croire président quelques heures, alors qu'il ne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

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

[4] "Histoire de la Statistique" – "Que sais-je ?" – P.U.F. 1997.

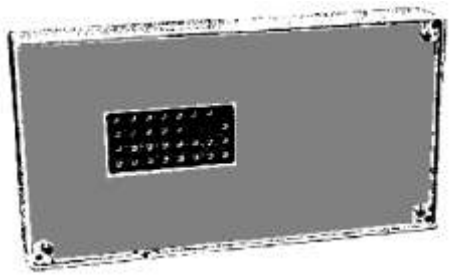
I – ESTIMATION D'UNE FREQUENCE

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.
Sauf à prendre comme échantillon la population toute entière (c'est un recensement et non un sondage) cette estimation est aléatoire (elle se fait à l'aide d'une variable aléatoire nommée "estimateur", notion qui n'est pas au programme de B.T.S.) et dépend du hasard qui a amené l'échantillon. Il n'y aura aucune certitude. C'est le type même du raisonnement inductif, qui n'a pas, on le verra, le caractère de la rigueur habituelle en mathématiques. Mais c'est ça ou rien.
On procède par sondages pour des raisons économiques. Il se peut, en particulier, qu'un contrôle de qualité nécessite la destruction des pièces testées (mesures de résistance aux chocs...).

NORME AFNOR
5.12 ESTIMATION
a) Opération ayant pour but, à partir des observations obtenues sur un ou plusieurs échantillons, d'attribuer des valeurs numériques aux paramètres de la population dont ceou ces échantillons sont issus, ou de la loi de probabilité considérée comme représentant cette population.
b) Résultat de cette opération



On peut matérialiser la situation en apposant un cache sur l'appareil à billes.
La proportion p de billes noires dans l'appareil est inconnue. On cherche à l'estimer à partir de la proportion f observée dans la lucarne correspondant à un échantillon.

Cependant, dans la suite, on supposera que les échantillons sont prélevés avec remise.

2 – Estimation ponctuelle

On prend comme ***estimation ponctuelle*** de p , la fréquence f observée sur l'échantillon.
Cette estimation a le défaut de dépendre fortement de l'échantillon (et donc du hasard qui l'a amené). Elle ne contient aucune indication de qualité de l'estimation, en particulier de la taille n de l'échantillon utilisé. On comprend que plus n est grand, meilleure est l'information, mais de quelle façon ? Pour cela, on va utiliser nos connaissances en matière d'échantillonnage, pour donner une "fourchette".
L'estimation est alors considérée comme le résultat d'une variable aléatoire nommée estimateur. Lorsque l'on connaît la loi de l'estimateur, on peut construire des intervalles de confiance.
La notion d'estimateur n'est pas au programme des sections de techniciens supérieurs.

Estimation :	Nombre ou intervalle	f est une *estimation* de p
Estimateur :	Variable aléatoire	F est un *estimateur sans biais* de p car $E(F) = p$

Le programme officiel indique cependant :

"Une illustration qualitative succincte des notions de biais et de convergence d'un estimateur peut être proposée, mais toute étude mathématique de ces qualités est hors programme."

Nous procéderons ainsi pour la question délicate (quant à sa présentation aux élèves) de l'estimation de l'écart type, raison supplémentaire pour commencer avec les fréquences plutôt qu'avec le couple moyenne et écart type.

3 – Intervalle de confiance d'une fréquence

a) CAS DES GRANDS ECHANTILLONS ($n \geq 30$; $np \geq 5$; $n(1-p) \geq 5$)

On considère une urne où la proportion de boules noires est p (inconnu).
Introduisons 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 suit la loi binomiale $\mathcal{B}(n, p)$ laquelle est proche (théorème de *Moivre-Laplace*), sous les conditions précisées ci-dessus, de la loi normale $\mathcal{N}\left(np, \sqrt{np(1-p)}\right)$.

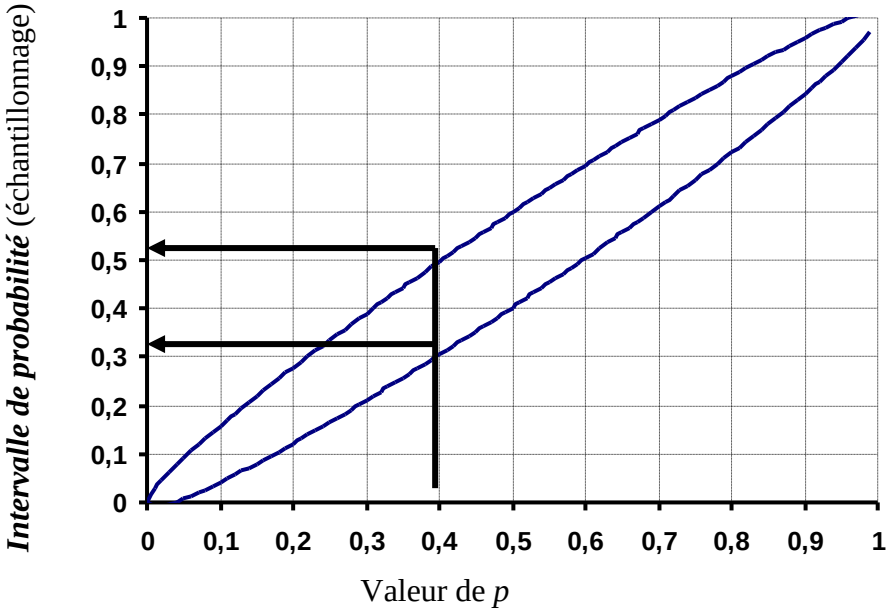
On considérera donc que F suit approximativement la loi $\mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

On peut, par exemple, déterminer le réel positif h tel que $P(p - h \leq F \leq p + h) = 0{,}95$.

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

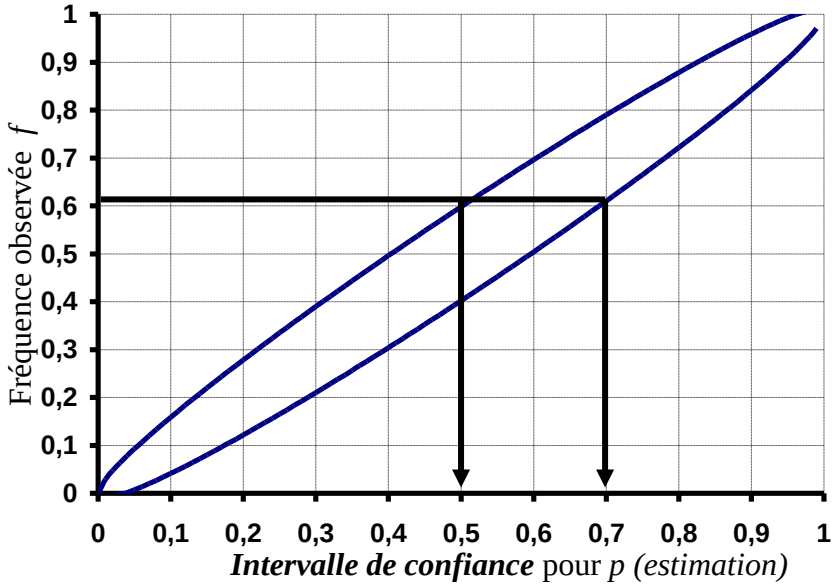
La théorie de l'***échantillonnage*** fournit ainsi, pour chaque valeur de p fixée, un ***intervalle de probabilité*** $\left[p - 1{,}96\sqrt{\frac{p(1-p)}{n}}, p + 1{,}96\sqrt{\frac{p(1-p)}{n}}\right]$ où la variable aléatoire F prend ses valeurs avec une probabilité de 0,95.

Supposons que $n = 100$ (la taille de l'échantillon est connue) et représentons, en fonction de p, les courbes d'équation :
$y = p - 0{,}196\sqrt{p(1-p)}$ et $y = p + 0{,}196\sqrt{p(1-p)}$.
On obtient le graphique suivant.



Le graphique se lit ainsi : supposons, on n'en sait rien, que l'urne contienne 40 % de boules noires, soit p = 0,4. L'intervalle de probabilité fourni par la théorie de l'échantillonnage est [0,304 , 0,496]. C'est à dire, qu'*avant de prélever un échantillon*, on a, sous cette hypothèse, 95 % de chances d'obtenir sur l'échantillon une fréquence f comprise dans cet intervalle.

L'estimation consiste en un renversement de point de vue : supposons que l'on observe une fréquence f = 0,6 de boules noires sur un échantillon (en ordonnées sur le graphique), on prendra comme ***intervalle de confiance à 95 %*** de p, l'intervalle obtenu, sur l'axe des abscisses, à partir des courbes précédentes, soit, environ (par lecture graphique) : [0,5 ; 0,7].



Si l'on recherche l'expression algébrique des bornes de cet intervalle de confiance, il faut rechercher p tel que : $0{,}6 = p \pm 0{,}196\sqrt{p(1-p)}$.

D'où l'équation du second degré : $(0{,}6-p)^2 = 0{,}196^2\, p(1-p)$,

c'est à dire, $p^2(1+0{,}196^2) - p(0{,}196^2+1{,}2) + 0{,}6^2 = 0$,

dont les solutions sont, à 10^{-3} près, 0,502 et 0,691.

On remarque que l'intervalle de confiance n'est pas, a priori, exactement centré sur la valeur $f = 0{,}6$.

On constate que les solutions obtenues sont très proches de celles données par la formule de l'intervalle de probabilité d'échantillonnage, où l'on remplacerait p par f !!

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 $\mathbb{N}(0\ ,\ 1)$ tel que $2\times\Pi(t_A) - 1 = A\ \%$.

Dans le cas précédent, pour $f = 0{,}6$; $n = 100$ et $A = 95\ \%$, on obtient $t_A = 1{,}96$ et comme intervalle de confiance pour p : [0,504 ; 0,696] , centré sur la valeur observée $f = 0{,}6$.

Remarques (assez nombreuses !) :

• ***A propos de la formule de l'intervalle de confiance d'une fréquence*** :

Exposer la résolution de l'équation du second degré aux étudiants n'est guère utile. On peut se contenter de dire que cette formule peut "s'expliquer" par le fait que f est une estimation ponctuelle de p et présenter les choses comme suit.

On a l'équivalence : $P\left(p - 1{,}96\sqrt{\frac{p(1-p)}{n}} \le F \le p + 1{,}96\sqrt{\frac{p(1-p)}{n}} \right) = 0{,}95$

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

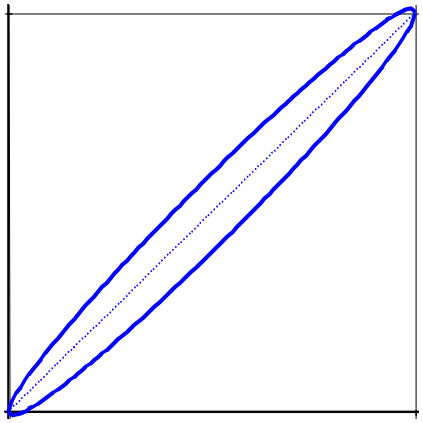
Le problème est que la valeur inconnue p se situe dans les bornes. On y substitue, faute de mieux, son estimation ponctuelle f .

Certains utilisent l'estimation ponctuelle de l'écart type (dont on parlera plus loin) pour donner la formule :

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

Pourquoi pas ? De toutes façons ça ne change pas grand chose et il faut bien comprendre qu'en estimation les décimales ont rapidement peu de signification (il faut être modeste en matière d'induction).

On peut également fournir un argument géométrique, ou, pour les sportifs qui voudront tenter ça en classe, un développement limité (voir les encadrés suivants).



Un argument géométrique

L'ellipse d'équation :

$$y^2 + p^2\left(1+\frac{1{,}96^2}{100}\right) - 2py - \frac{1{,}96^2}{100}p = 0$$

a son grand axe (de symétrie) proche de la droite d'équation $y = x$.

L'expression de l'intervalle de confiance justifiée par un développement limité

Pour une valeur f observée, l'intervalle de confiance de p à 95% s'obtient en résolvant en p l'équation :

$$p^2\left(1+\frac{1{,}96^2}{n}\right) - p\left(\frac{1{,}96^2}{n} + 2f\right) + f^2 = 0.$$

On pose $x = \dfrac{1{,}96}{\sqrt{n}}$ (proche de 0 pour n grand).

On a :

$$p = \frac{2f + x^2 \pm \sqrt{x^4 + 4f(1-f)x^2}}{2(1+x^2)}$$

$$= \left(f + \frac{1}{2}x^2 \pm x\sqrt{f(1-f)}\left(1+\frac{1}{4f(1-f)}x^2\right)^{\frac{1}{2}}\right) \times \left(1+x^2\right)^{-1} .$$

On effectue un développement limité de cette expression à l'ordre 1, au voisinage de $x = 0$:

$$p = \left(f \pm x\sqrt{f(1-f)}(1+x\varepsilon_1(x))\right)(1+x\varepsilon_2(x)) ,$$

puis $p = f \pm x\sqrt{f(1-f)} + x\varepsilon(x)$

avec $\lim\limits_{x\to 0} \varepsilon(x) = 0$.

On retrouve ainsi l'expression des bornes de l'intervalle de confiance à 95 % :

$$f \pm \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)} .$$

(d'après Saporta p. 307)

• ***A propos de "probabilité" et "confiance"*** :
Le programme officiel insiste sur le point suivant :

"On distinguera confiance et probabilité :
- ***avant*** *le tirage d'un échantillon, la procédure d'obtention de l'intervalle de confiance a une* ***probabilité*** $1 - \alpha$ *que cet intervalle contienne le paramètre inconnu,*
- ***après*** *le tirage, le paramètre est dans l'intervalle calculé avec une* ***confiance*** $1 - \alpha$ *."*

En effet, 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.
La probabilité porte sur la procédure d'estimation :

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

La seconde écriture correspond à un ***intervalle dont les bornes sont aléatoires***, centré sur les valeurs de F et contenant p avec une probabilité de 0,95 (le problème est que cet intervalle contient la valeur inconnue p dans son expression).
On peut dire que, sur un grand nombre d'intervalles de confiances (obtenus à partir d'un grand nombre d'échantillons), environ 95 % contiennent effectivement la valeur de p, ou

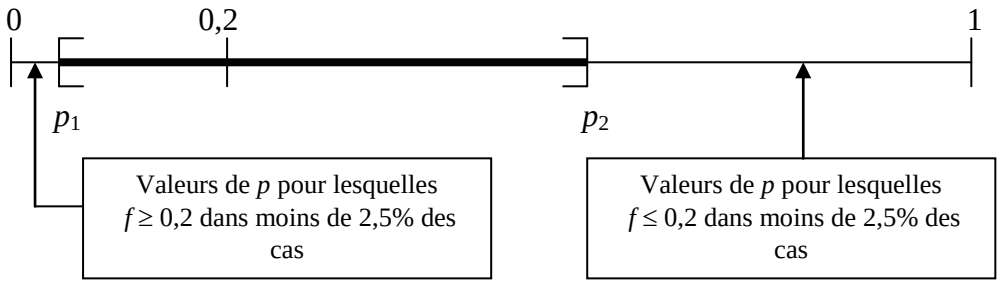
encore, que l'on a 95% de chances d'exhiber un intervalle contenant p (avant le tirage de l'échantillon).

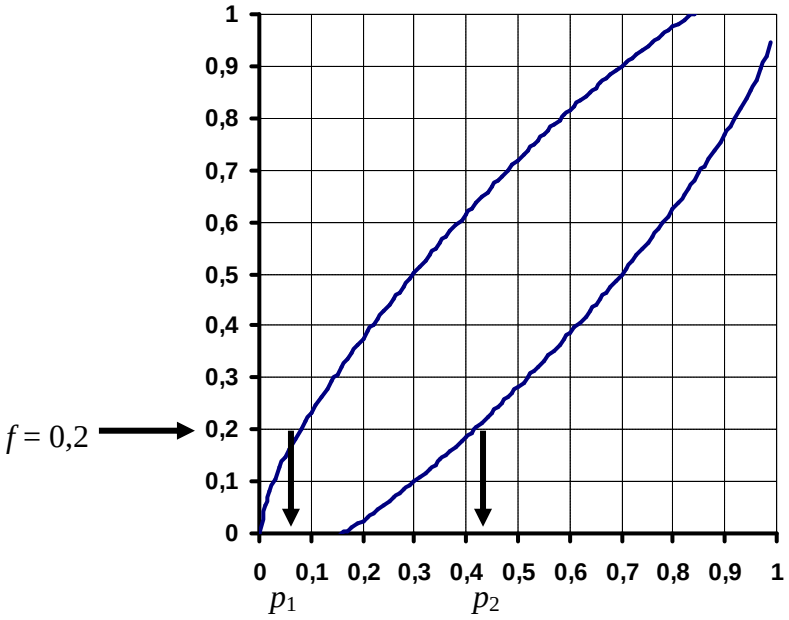
b) CAS DES PETITS ECHANTILLONS ($n < 30$ ou $np < 5$ ou $n(1 – p) < 5$)

On ne peut plus utiliser l'approximation par la loi normale, on utilise alors soit la loi binomiale, soit un abaque (voir plus loin) ou des tables.
Ce paragraphe répond au T.P. 1 du module "Statistique inférentielle" du programme.

"Estimation ponctuelle et par intervalle de confiance de la fréquence, dans le cas d'une loi binomiale connue, à partir d'échantillons simulés.
La connaissance a priori de la loi sous-jacente permet de comparer le paramètre réel et les estimations obtenues à partir d'échantillons.
Aucune connaissance sur ce TP n'est exigible dans le cadre du programme de mathématiques."

On a p inconnu et supposons que la fréquence du seul échantillon prélevé est $f = 0,2$.
A partir de la fréquence de cet échantillon, on détermine les extrémités p_1 et p_2 de l'intervalle de confiance de p au coefficient de confiance 95% de la façon suivante :





N.B. : les courbes représentées sur le schéma ci-dessus, l'on été sous l'approximation normale et ne sont pas exactes (voir l'abaque).

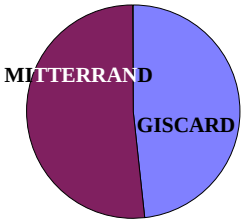
⇨ ***Voir le T.P. Excel "Loi binomiale et intervalle de confiance" donné en annexe.***

4 – Expérimentation de la notion d'intervalle de confiance sur Excel

La compréhension de la notion (délicate) d'intervalle de confiance nécessite de se confronter à l'expérience. Ceci est possible grâce à la simulation sur tableur.

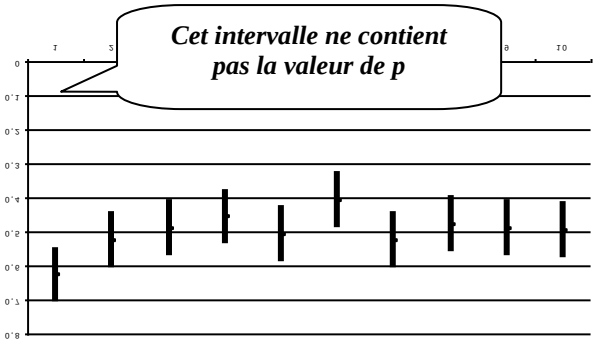
⇨ ***Voir le T.P. Excel "Intervalles de confiance" donné en annexe.***

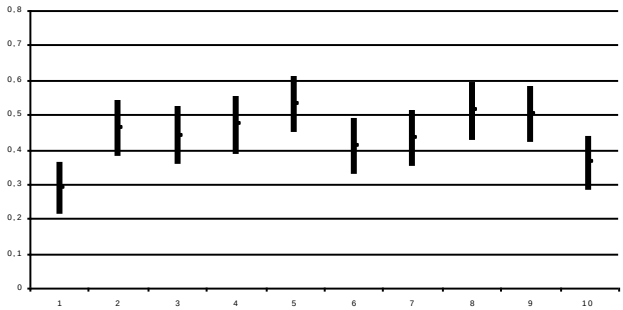
Travaillons sur un exemple où les scores étaient particulièrement serrés :
Le 10 mai 1981, François Mitterrand a été élu avec 51,75 % des voix, alors que Valery Giscard d'Estaing n'a recueilli que **48,25 %** des suffrages.
On suppose que l'on effectue des sondages le jour de l'élection, pour estimer la proportion p des partisans de Giscard dans l'électorat (en réalité, $p = 0{,}4825$).



a – Taille n = 100, coefficient de confiance 90 %

Pour simuler un sondage (échantillon prélevé au hasard avec remise) de n = 100 personnes, il suffit de recopier 100 fois la formule =ENT(ALEA()+0,4825), qui renvoie la valeur 1 dans le cas d'un électeur de Giscard et 0 sinon.
On peut faire calculer les bornes d'un intervalle de confiance à 90 % pour p, puis représenter les "fourchettes" ainsi obtenues, pour 10 sondages, à l'aide d'un graphique de type "boursier".
En choisissant l'option "calcul sur ordre", il suffit ensuite de faire F9, pour obtenir 10 nouveaux sondages et visualiser les fluctuations des différents intervalles de confiance calculés :





On peut, par l'observation, répondre aux questions suivantes :

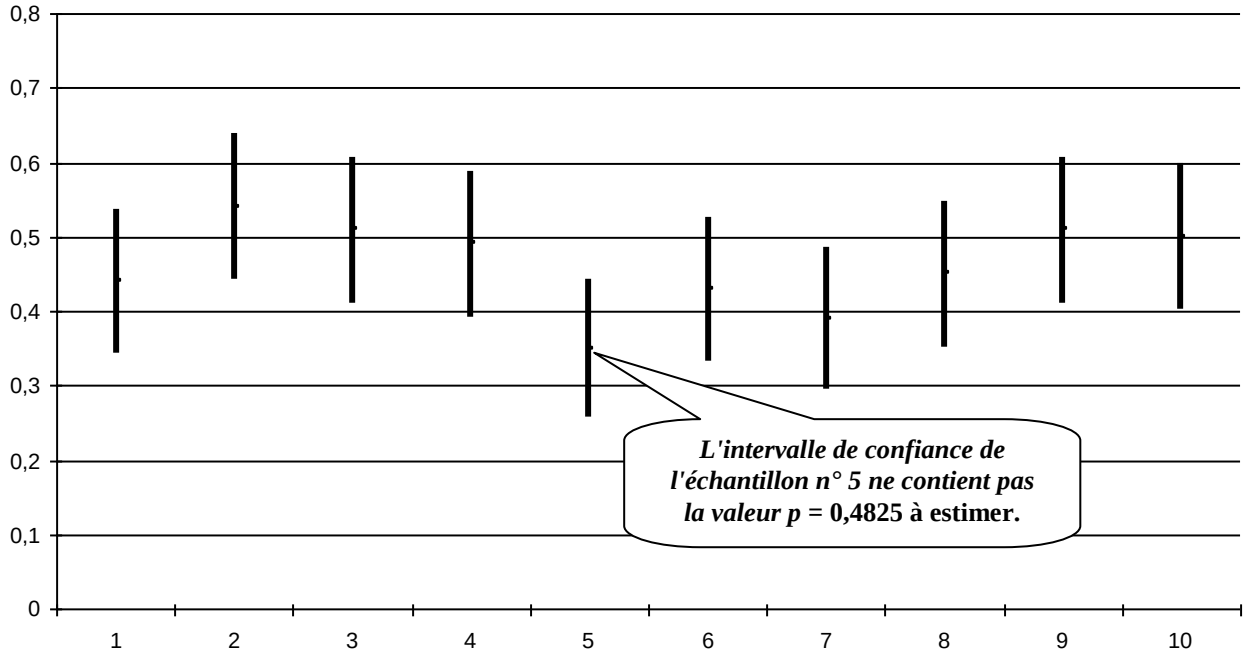
- Sur un exemple de 10 sondages,

- combien donnent une fréquence f en faveur (à tort) de Giscard ($f > 0{,}5$) ?
- combien d'intervalles prévoient complètement la victoire de Mitterrand (fourchette entièrement située sous la valeur 0,5) ?

- Deux intervalles de confiances ont-ils obligatoirement le même centre ?
- Deux intervalles de confiance peuvent-ils n'avoir aucun élément commun ?
- Est-ce que $p = 0{,}4825$ appartient nécessairement à l'intervalle de confiance donnée par un sondage ?
- Quel est, sur 100 sondages observés, le pourcentage d'intervalles à 90% de confiance ne contenant pas la valeur p à estimer ?

b – Taille n = 100, coefficient de confiance 95 %

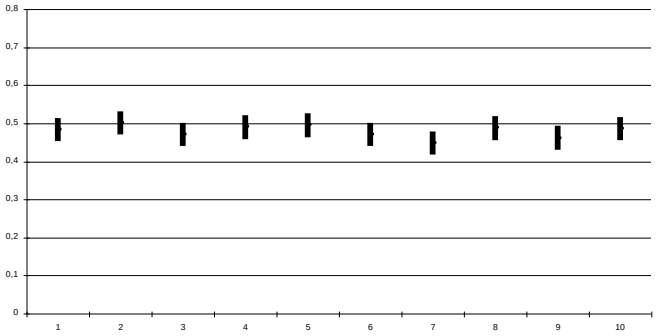
Il suffit de changer la valeur 1,645 , utilisée pour les intervalles de confiance à 90 %, par 1,96 , pour faire de nouvelles observations.



Les intervalles de confiance à 95 % se "trompent" moins souvent (statistiquement, on peut observer qu'environ 95 % contiennent la valeur *p*) parce qu'ils sont plus "longs".
On peut observer de légers écarts d'amplitude entre les différents intervalles, selon que *f* est petit (échantillon n° 5) ou grand (échantillon n° 2).

c – Taille n = 1000, coefficient de confiance 95 %

La plus grande qualité de l'information se traduit par une amplitude notablement réduite des intervalles.
Il y a toujours, statistiquement, 95 % qui ne contiennent pas *p* (sur l'image, c'est le cas de l'échantillon n° 7).



5 – Des questions pour tester la compréhension des élèves

Certains énoncés à l'examen posent parfois, après la détermination d'un intervalle de confiance, certaines questions permettant de juger de la compréhension des élèves (qui souvent ne font qu'appliquer une formule sans comprendre).
A la suite d'un TP de simulation tel que le précédent, le nombre des réponses correctes des étudiants à ce genre de questions est en nette augmentation.

⇨ ***Voir les annales de B.T.S. données en annexe.***

Examinons ici deux exemples du BTS.

BTS Groupement B 2002 :

On considère un échantillon de 100 véhicules prélevés au hasard dans le parc de véhicules neufs mis en circulation par l'entreprise en septembre 2001. Ce parc contient suffisamment de véhicules pour qu'on puisse assimiler ce tirage à un tirage avec remise.

On constate, qu'au bout de 6 mois de mise en circulation, 91 véhicules de cet échantillon n'ont pas eu de sinistre.

a) Donner une estimation ponctuelle du pourcentage p de véhicules de ce parc qui n'ont pas eu de sinistre 6 mois après leur mise en circulation.

b) Soit F la variable aléatoire qui à tout échantillon de 100 véhicules prélevés au hasard et avec remise dans ce parc, associe le pourcentage de véhicules qui n'ont pas encore eu de sinistre 6 mois après leur mise en circulation.

On suppose que F suit la loi normale

$\mathrm{N}\left(p\ ,\ \sqrt{\frac{p(1-p)}{100}}\right)$, où p est le pourcentage inconnu de véhicules du parc qui n'ont pas eu de sinistre 6 mois après leur mise en circulation.

Déterminer un intervalle de confiance du pourcentage p avec le coefficient de confiance 95 %.

c) On considère l'affirmation suivante :

« le pourcentage p est obligatoirement dans l'intervalle de confiance obtenu à la question b)»

Est-elle vraie ? (On ne demande pas de justification.)

a) On estime p à 0,63.

b) On trouve [0,535 ; 0,725].

c) Non, p n'est pas obligatoirement dans l'intervalle de confiance (la procédure amenant à la détermination de cet intervalle a 95 % de chances d'aboutir à un intervalle contenant p).

BTS Comptabilité gestion 1992 :

Dans cet énoncé, on détermine une estimation de la moyenne μ d'une population (le problème est analogue à celui d'une fréquence), par intervalle de confiance centré sur la moyenne $\overline{x}$ obtenue sur un échantillon, avec le coefficient de confiance 95%.

On pose ensuite la question suivante :

Répondre **par oui ou par non** aux cinq questions suivantes :

a) Le nombre μ appartient-il obligatoirement à cet intervalle de confiance ?

b) Le nombre μ a-t-il plus de chances de se trouver près du centre que d'une extrémité de **cet** intervalle de confiance ?

c) Avec un nouvel échantillon de 100 jours ouvrables prélevé comme précédemment, on obtient de la même façon un second intervalle de confiance de μ avec le coefficient de confiance 95%.

- Les deux intervalles de confiance sont-ils obligatoirement les mêmes ?
- Ces deux intervalles ont-ils obligatoirement le même centre ?
- Ces deux intervalles de confiance peuvent-ils n'avoir aucun élément commun ?

a) Non. Il est clair qu'il ne s'agit pas d'un intervalle à 100 % de confiance. La procédure suivie ne conduit à un intervalle contenant μ que dans 95 % des cas.

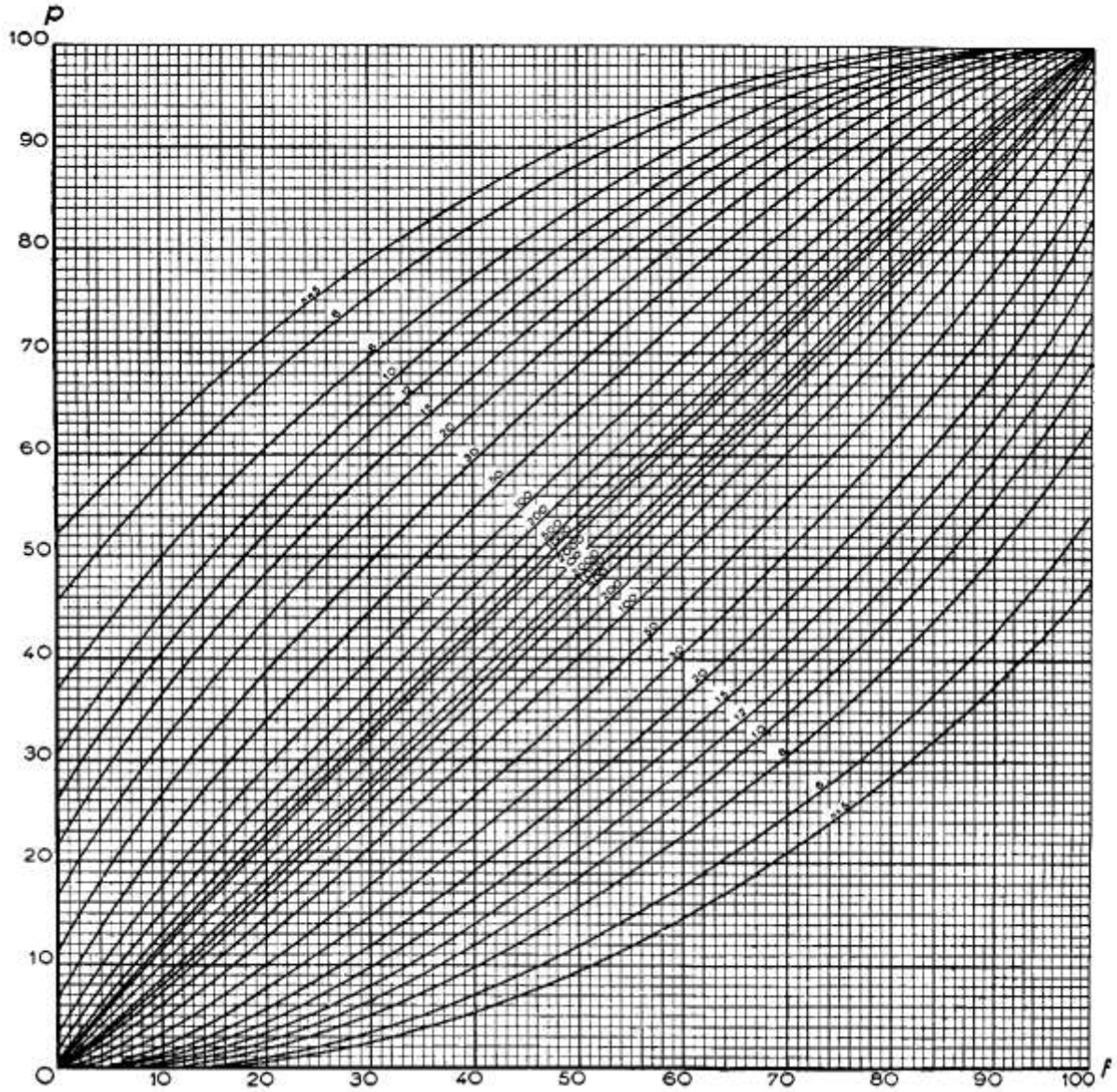
b) Il est impossible de répondre par oui ou par non à cette question, qui est mal posée. Pour un intervalle de confiance déjà calculé et pour μ déterminé, quoique inconnu, il n'y a rien d'aléatoire. Il ne peut donc être question de probabilités ou de "chances".

Maintenant, d'après le profil de la loi normale, on peut dire que la procédure de calcul des intervalles de confiance a plus de chances d'amener un intervalle recouvrant μ vers son centre que vers ses extrémités. Mais ne compliquons pas et il vaut mieux éviter de poser ce genre de question.

c) La pratique des simulations montre clairement que les réponses aux trois questions posées sont : non, non (les centres f_i sont rarement les mêmes) et oui (on peut avoir des intervalles disjoints, auquel cas un au moins se trompe, même si c'est relativement rare).

6 – Utilisation d'abaques ou des fonctions intégrées de certaines calculatrices

a – Abaque des intervalles de confiance à 95 % d'une fréquence



ABAQUE donnant, en fonction de la fréquence ***f*** *(**en % en abscisses**) observée sur un échantillon de taille n , l'intervalle de confiance à 95% de la PROPORTION* ***p*** *(en %) dans la population*

• En dehors des conditions de l'approximation normale ($n < 30$ ou $np < 5$ ou $n(1 - p) < 5$), l'abaque est calculé à partir de la loi binomiale et les intervalles de confiance peuvent ne pas être symétriques autour de la valeur observée f .

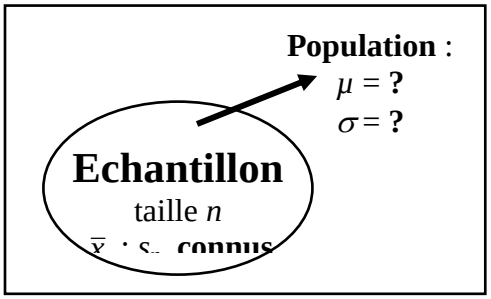
• Pour n "moyen" (de l'ordre de 30 à 100), on lit sur l'abaque (ou dans les tables) des intervalles de confiance encore sensiblement différents de ceux donnés par la formule $\left[f - t_A\sqrt{\frac{f(1-f)}{n}} \ , \ f + t_A\sqrt{\frac{f(1-f)}{n}} \right]$, formule qui ne tient pas compte du fait qu'on peut faire une correction de continuité dans l'approximation par la loi normale.

b – Calculatrices

Modèles de calculatrices :	**CASIO Graph 80**	**TI 83**	**SHARP EL 9600**
Intervalle de confiance d'une fréquence ou proportion *p* d'une population, à partir de la fréquence *f* calculée sur un échantillon de taille *n*.	STAT INTR Z 1-P C-Level : entrer 0.95 ou 0.99 ou ... x: entrer le nombre de succès (x=f×n) n: entrer taille échantillon affichage de l'intervalle de confiance	STAT TESTS A:1-PropZInt x: entrer le nombre de succès (x=f×n) n: entrer taille échantillon C-Level : entrer 0.95 ou 0.99 ou ... affichage de l'intervalle de confiance	STAT E TEST 17 InputStats ENTER 14 Zint1prop

II – ESTIMATION D'UNE MOYENNE

1 – Estimation ponctuelle



On prélève, au hasard et avec remise, un échantillon de taille *n* dans une population.
On calcule la moyenne $\bar{x}$ et l'écart type s_n de cet échantillon.
Il s'agit d'estimer la moyenne μ et l'écart type σ, inconnus, de la population.

On prend comme ***estimation ponctuelle*** de μ, la moyenne $\bar{x}$ observée sur l'échantillon.

On prend comme ***estimation ponctuelle*** de σ, le nombre $s = \sqrt{\frac{n}{n-1}}\, s_n$.

Estimateurs sans biais (démonstration hors programme en B.T.S.)

Soit X_i la variable aléatoire qui, au i^{ème} tirage, associe la mesure obtenue.
Les X_i sont indépendantes, de même loi de moyenne μ et d'écart type σ (tirages avec remise).

Alors $\overline{X} = \frac{1}{n}\sum X_i$ est un *estimateur sans biais* de μ car $E(\overline{X}) = \frac{1}{n} \times n\mu = \mu$ et l'estimation ponctuelle $\bar{x}$ est une réalisation de la variable aléatoire $\overline{X}$.

L'estimation s^2 est une réalisation de la variable aléatoire $S_{n-1}{}^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}\left(X_i - \overline{X}\right)^2$, qui est un *estimateur sans biais* de la variance σ^2 .
Pour le justifier, on observe d'abord que :

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum\left(X_i - \overline{X} + \overline{X} - \mu\right)^2 = \sum\left(X_i - \overline{X}\right)^2 + 2(\overline{X} - \mu)\sum(X_i - \overline{X}) + \sum\left(\overline{X} - \mu\right)^2 .$$

or $\sum(X_i - \overline{X}) = \sum X_i - n\overline{X} = 0$, d'où : $\boxed{\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + n\left(\overline{X} - \mu\right)^2}$.

On a ainsi $E\left(\sum\left(X_i - \overline{X}\right)^2\right) = \sum V(X_i) - nV(\overline{X})$,

c'est à dire $E\left(\sum\left(X_i - \overline{X}\right)^2\right) = n\sigma^2 - n\frac{\sigma^2}{n}$ et donc $E\left(S_{n-1}{}^2\right) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$

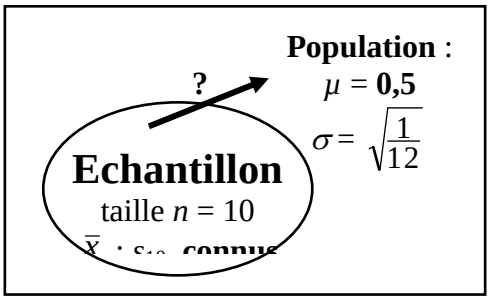
Remarques :

• L'estimation $s = \sqrt{\frac{n}{n-1}}\, s_n$ est donnée par la touche $x\sigma_{n-1}$ ou s de la calculatrice.

• L'estimation de σ peut étonner. Elle est ainsi faite pour que , sur un grand nombre d'échantillons, la moyenne des estimations soit égale à σ. En effet, s_n a tendance à être généralement inférieur à σ (*biais*).

La démonstration (voir encadré) est tout à fait hors programme. En revanche, celui-ci nous suggère une expérimentation par simulation (une "*illustration qualitative succincte"* peut être proposée mais "*toute étude mathématique est hors programme*").

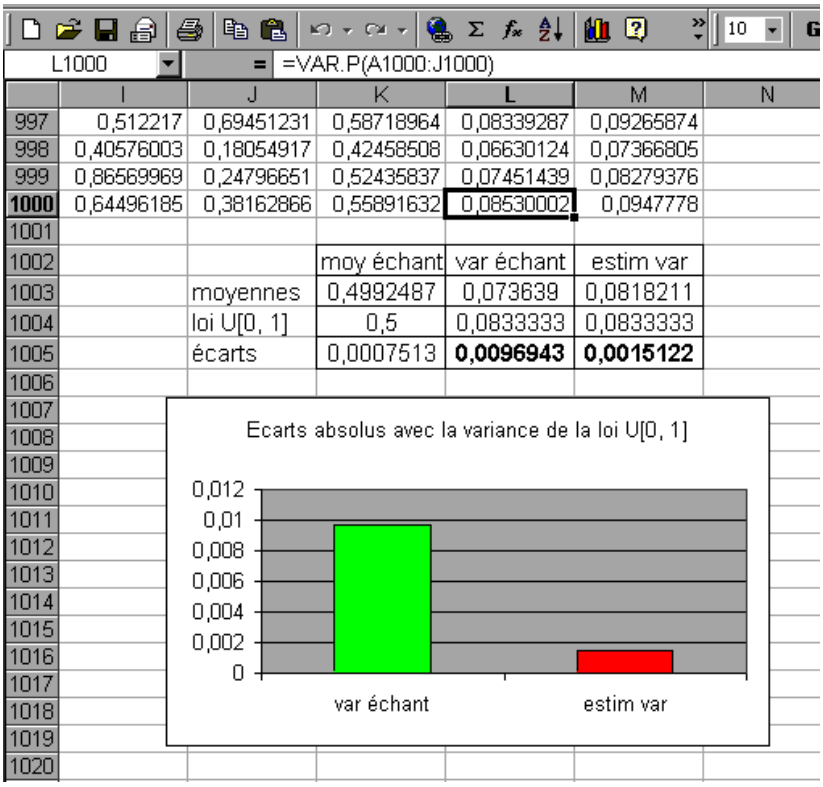
Une telle illustration est rapidement possible, par simulation sur Excel.



On a choisi l'étude de l'estimateur de la variance, dans le cadre d'une population distribuée selon la loi uniforme U [0, 1].

On prend comme ***estimation ponctuelle*** de σ , le nombre $s = \sqrt{\frac{10}{9}}\, s_{10}$.

On vérifie expérimentalement que s^2 est "en moyenne" plus proche de σ^2 que s_{10}^2 .



	I	J	K	L	M	N
997	0,512217	0,69451231	0,58718964	0,08339287	0,09265874	
998	0,40576003	0,18054917	0,42458508	0,06630124	0,07366805	
999	0,86569969	0,24796651	0,52435837	0,07451439	0,08279376	
1000	0,64496185	0,38162866	0,55891632	0,08530002	0,0947778	
1001						
1002			moy échant	var échant	estim var	
1003		moyennes	0,4992487	0,073639	0,0818211	
1004		loi U[0, 1]	0,5	0,0833333	0,0833333	
1005		écarts	0,0007513	**0,0096943**	**0,0015122**	
1006						
1007						
1008						
1009						
1010						
1011						
1012						
1013						
1014						
1015						
1016						
1017						
1018						
1019						
1020						

Pour obtenir les résultats de l'écran ci-dessus, on a simulé 1000 échantillons de taille 10 selon la formule =ALEA() .

La variance s_{10}^2 sur un échantillon correspond à la formule =VAR.P(A*x*:J*x*) (le P semble signifier que l'on considère la plage de cellules suivante, ici l'échantillon, comme la population de l'étude), alors que l'estimation s^2 de la variance de la population à partir de cet échantillon correspond à la formule =VAR(A*x*:J*x*) .

En faisant F9 on simule une autre série de 1000 échantillon.

2 – Estimation par intervalle de confiance

On considère la variable aléatoire $\overline{X}$ qui, à tout échantillon de taille *n*, associe sa moyenne.

On sait que, au moins pour *n* assez grand (théorème limite central), $\overline{X}$ suit (approximativement) la loi normale $\mathbb{N}\left(\mu\,,\,\frac{\sigma}{\sqrt{n}}\right)$.

On sait alors que $P\,(\mu - h \leq \overline{X} \leq \mu + h) = 0{,}95$ pour $h = 1{,}96\frac{\sigma}{\sqrt{n}}$.

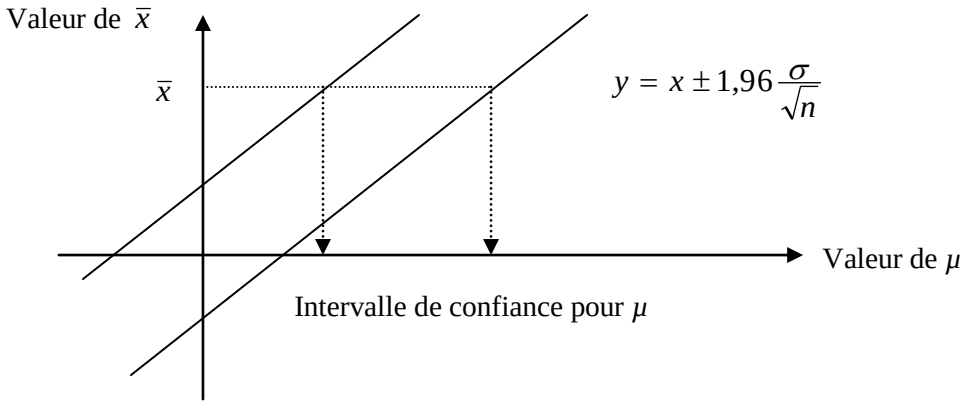
Ce qui équivaut à $P\,(\overline{X} - h \leq \mu \leq \overline{X} + h) = 0{,}95$ (intervalle dont les bornes sont aléatoires et qui contient μ avec une probabilité de 0,95).

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

$\left[\bar{x} - t_A\frac{\sigma}{\sqrt{n}}\,;\,\bar{x} + t_A\frac{\sigma}{\sqrt{n}}\right]$, calculé à partir de la moyenne $\bar{x}$ observée sur un échantillon de taille *n*, où t_A est donné par la table de la loi normale $\mathbb{N}\,(0\,,\,1)$ tel que $2\times\Pi(t_A) - 1 = A\,\%$.

Remarques :

• Lorsque σ est connu et que *n* est grand, la situation est plus simple que pour les fréquences.



• Ce qui précède est valable pour les grands échantillons ($n \geq 30$) ou les petits échantillons issus d'une population normale, lorsque σ est connu.

• Lorsque σ est inconnu et $n \geq 30$, on utilise la formule précédente, en remplaçant σ par son estimation $s = \sqrt{\frac{n}{n-1}}\,s_n$.

• Pour les ***petits échantillons*** issus d'une population normale, on utilise, lorsque σ est inconnu, une loi de ***Student*** .

Dans certains domaines d'application, les petits échantillons sont la règle. C'était de le cas pour *William Gosset* (1876-1937) alias *Student*, qui, pour ses contrôles de qualité aux brasseries Guinness, ne pouvait disposer de grands échantillons extraits d'une population homogène à cause de la grande variabilité des conditions de fabrication (température, provenance du houblon, du malt, …).

Le programme prévoit que, pour répondre à la demande des matières technologiques, on peut donner des exemples de procédures fondées sur la loi de *Student* mais *"aucune connaissance à leur sujet n'est exigible"*.

On considère une population répartie selon une loi normale de moyenne μ et d'écart type σ, tous deux inconnus.
On prélève des échantillons aléatoires non exhaustifs de taille n.
Soit X_i la variable aléatoire qui associe au $i^{\text{ème}}$ tirage, son résultat. Les X_i sont indépendantes de loi normale $\mathcal{N}(\mu\ ,\ \sigma)$.
On note :

$$\overline{X} = \frac{1}{n}\sum X_i \quad \text{et} \quad S^2 = \frac{1}{n}\sum X_i^2 - \overline{X}^2 .$$

On sait que $\dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ suit la loi normale centrée réduite et on montre que $\dfrac{nS^2}{\sigma^2}$ est indépendante de la précédente et suit la loi du khi 2 à $n-1$ degrés de liberté.
Dans ces conditions (voir l'encadré) la variable aléatoire $T_{n-1} = \dfrac{\dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}\sqrt{n-1}}{\sqrt{\dfrac{nS^2}{\sigma^2}}}$ suit la loi de *Student* à $n-1$ degrés de liberté.
Le calcul des intervalles de confiance sera donc basé sur cette variable aléatoire $\boxed{T_{n-1} = \dfrac{\overline{X} - \mu}{S}\sqrt{n-1}}$ suivant la loi de *Student* à $n-1$ degrés de liberté.

LOI DE STUDENT

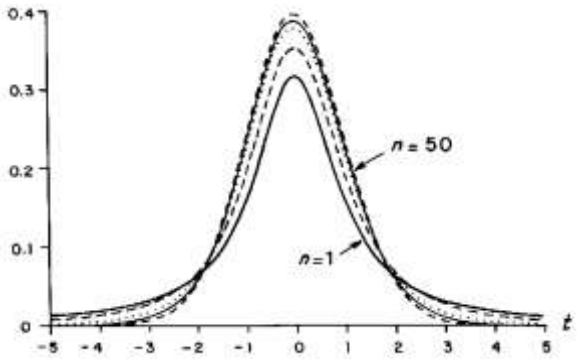
Si X est une variable aléatoire de loi normale centrée réduite et Y une variable aléatoire indépendante de X de loi du khi 2 à n degrés de liberté, alors, par définition, la variable aléatoire $\dfrac{X\sqrt{n}}{\sqrt{Y}}$ suit la loi de *Student* à n degrés de liberté.
La loi de *Student* à n degrés de liberté admet comme fonction de densité la fonction f définie sur $\mathbb{R}$ par :

$$f(t) = \frac{1}{\sqrt{n}\ B(\frac{1}{2}, \frac{n}{2}) \times \left(1 + \frac{t^2}{n}\right)^{(n+1)/2}},$$

où la fonction eulérienne B est définie par

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

avec $\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}\ dt$.



Densité de *Student* pour 1, 2, 5, 10, 50 degrés de liberté.

Pour un niveau de confiance $1-\alpha$, on lira dans la table le nombre t tel que :

$$P(-t \le T \le t) = 1-\alpha \ \Leftrightarrow\ P(-t \le \frac{\overline{X} - \mu}{S}\sqrt{n-1} \le t) = 1-\alpha \ \Leftrightarrow$$

$$P(\overline{X} - t\frac{S}{\sqrt{n-1}} \le \mu \le \overline{X} + t\frac{S}{\sqrt{n-1}}) = 1-\alpha .$$

Et on prendra ainsi comme intervalle de confiance de μ au niveau $1-\alpha$:

$\boxed{\left[\overline{x} - t\dfrac{s_n}{\sqrt{n-1}}\ ,\ \overline{x} + t\dfrac{s_n}{\sqrt{n-1}}\right]}$ où $\overline{x}$ et s_n sont la moyenne et l'écart type de l'échantillon,

ou encore, $\boxed{\left[\overline{x} - t\dfrac{s}{\sqrt{n}}\ ,\ \overline{x} + t\dfrac{s}{\sqrt{n}}\right]}$ où s est l'estimation de σ, calculée sur l'échantillon.

Exemple :
Sur un échantillon de taille $n = 20$, extrait avec remise d'une population normale, on obtient une moyenne $\bar{x} = 24{,}75$ et un écart type $s_n = 1{,}12$.
a) Donner un intervalle de confiance de la moyenne μ de la population, avec un niveau de confiance de 95 %.

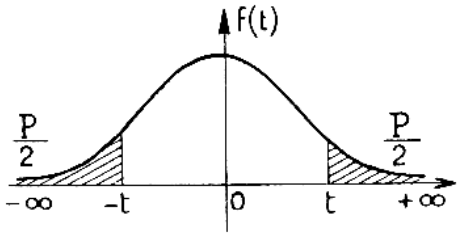
Le nombre de degrés de liberté est $\nu = n - 1 = 19$. Pour un coefficient de confiance de 95 % la table donne (P = 0,05) $t = 2{,}093$. Les bornes de l'intervalle de confiance sont donc données par $24{,}75 \pm 2{,}093\dfrac{1{,}12}{\sqrt{19}}$, d'où l'intervalle : [24,21 ; 25,29].

b) Reprendre le calcul avec $n = 30$ et comparer avec l'intervalle obtenu avec la loi normale.

Pour 29 degrés de libertés, on lit $t = 2{,}042$ alors que la loi normale donne $t = 1{,}96$.
Les intervalles de confiance sont très voisins :
[24,32 ; 25,18] avec la loi de *Student* et [24,34 ; 25,16] avec la loi normale.

Table de distribution de T (loi de Student)

ν : degrés de liberté. P : probabilité que T prenne une valeur extérieur à $[-t\ ,\ t]$.



ν \ P	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,929
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
80	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

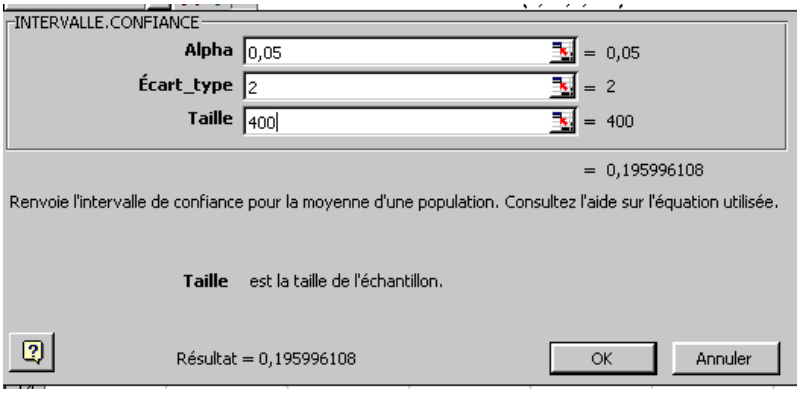
Pour les grands échantillons, on retrouve les valeurs fournies par la loi $N(0\ ,\ 1)$

3 – Utilisation des fonctions intégrées de certaines calculatrices ou de l'ordinateur

a - Calculatrices

Modèles de calculatrices :	CASIO Graph 80	TI 83	SHARP EL 9600
Intervalle de confiance d'une moyenne *m* d'une population, à partir de la moyenne $\bar{x}$ calculée sur un échantillon de taille *n*.	STAT INTR Z (si σ pop; connu) ou t (si σ pop. Inconnu) 1-S Data : Var C-Level : 0.95 ou 0.99 ou ... σ (ou xσn-1) : entrer σ pop. ou estimé $\bar{x}$: entrer $\bar{x}$ échantillon n: entrer taille échantillon affichage de l'intervalle de confiance	STAT TESTS 7:ZInterval (Si σ pop connu) ou 8:TInterval (Si σ inconnu) Inpt: Stats $\bar{x}$: entrer $\bar{x}$ échantillon σ (ou Sx) : entrer σ pop. ou son estimé n : entrer taille échantillon C-Level : entrer 0.95 ou 0.99 ou ... affichage de l'intervalle de confiance	STAT E TEST 17 InputStats ENTER (mode valeurs numériques) 12 Zint1samp (Si σ pop connu) ou 06 Tint1samp (Si σ inconnu)

b – Sur Excel

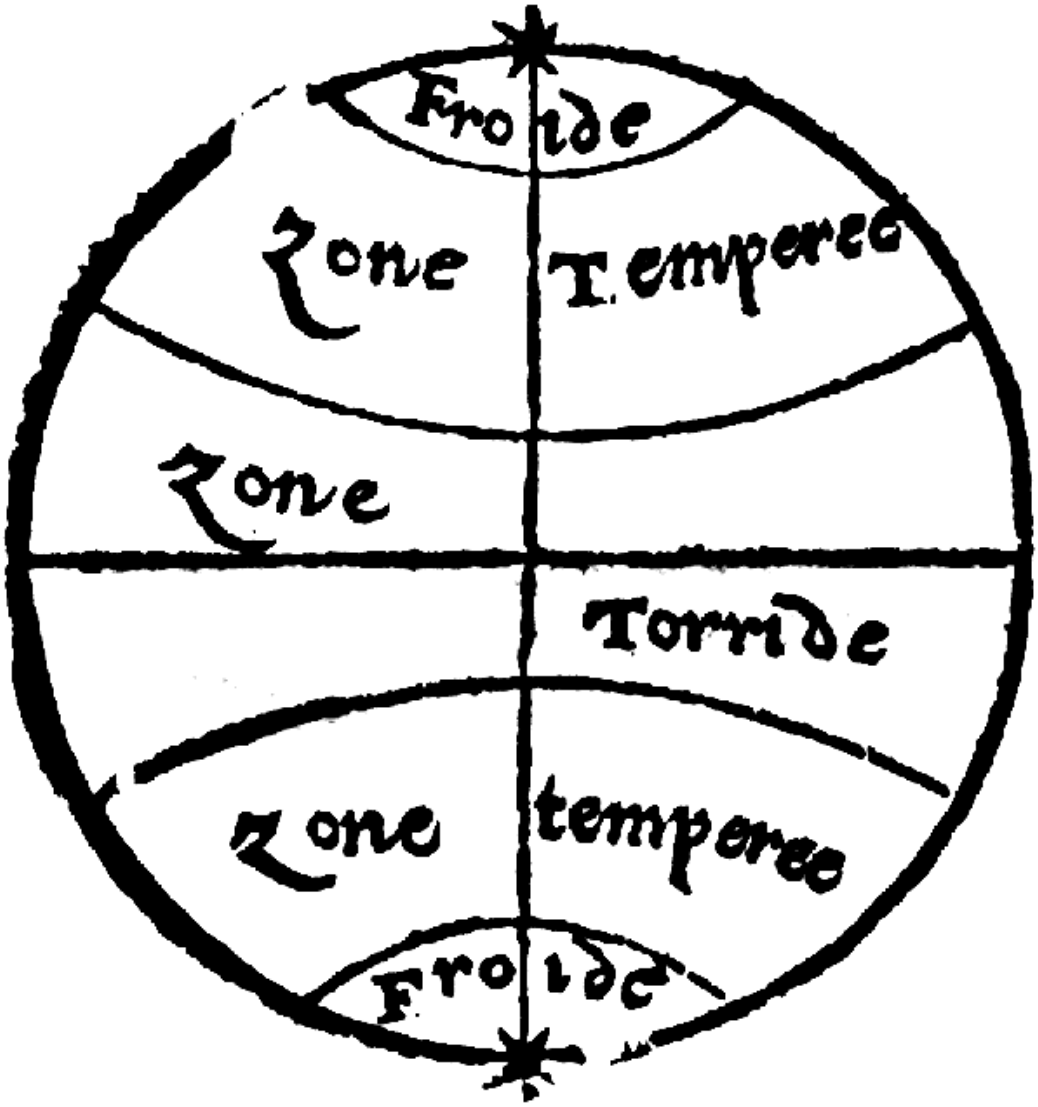


Pour un intervalle de confiance d'une moyenne μ, on peut, lorsque l'écart type σ de la population est connu, utiliser la fonction INTERVALLE.CONFIANCE .
Par exemple, si l'on entre :
$\alpha = 0{,}05$ (confiance à 95%)
$\sigma = 2$ et $n = 400$,
on obtient la valeur ≈ 0,196
qui est le rayon de l'intervalle de confiance. Ainsi, si l'on a obtenu $\bar{x} = 239$ sur un échantillon, l'intervalle de confiance de μ à 95 % sera [239 – 0,196 ; 239 + 0,196].

III – Tableau résumé des notations

	Population	**Echantillon**	Estimation	Variable aléatoire (estimateur)	Loi de probabilité
fréquence	p	f	f	$F = \frac{1}{n}\sum X_i$ où X_i suit la loi $\mathcal{B}(1, p)$	$\mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ Approximativement, pour $n \geq 30$
moyenne	μ	$\bar{x}$	$\bar{x}$	$\bar{X} = \frac{1}{n}\sum X_i$ avec $E(X_i) = \mu$ et $\sigma(X_i) = \sigma$	$\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, si X_i normales, ou approximativement pour $n \geq 30$
écart type	σ	s_n	$s = \sqrt{\frac{n}{n-1}}\, s_n$	$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}(X_i - \bar{X})^2$	$\frac{(n-1)S_{n-1}^2}{\sigma^2}$ suit la loi χ^2_{n-1}, si X_i normales

Les résultats encadrés en gras sont hors programme


froide
zone temperee
zone
torride
zone temperee
froide

T.P. Excel : INTERVALLES DE CONFIANCE

Objectifs

Sur la base d'un énoncé d'examen :

- Etudier un échantillon issu d'une distribution normale (moyenne pondérée , écart type, histogramme).
- Déterminer un intervalle de confiance pour la moyenne de la population et étudier l'impact du coefficient de confiance et de la taille de l'échantillon.
- Expérimenter la dépendance des intervalles de confiance à l'échantillon choisi.

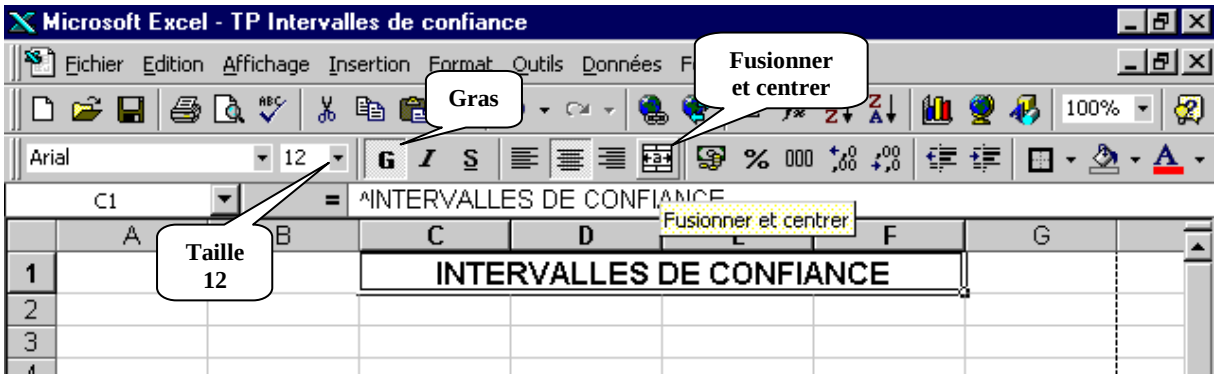
OUVERTURE D'UN FICHIER ET MISE EN PLACE DE L'ENVIRONNEMENT

Lancer Excel.

Pour marquer le titre :

Cliquer dans la cellule C1, choisir ***Taille 12 Gras*** et taper le titre : "INTERVALLES DE CONFIANCE" puis sur la touche ***ENTREE***.

Sélectionner (en maintenant enfoncé le bouton gauche de la souris), les cellules de C1 à F1 puis cliquer sur l'icône ***Fusionner et centrer*** puis sur l'icône d'encadrement.



1 – ETUDE D'UN ECHANTILLON

Un sujet de BTS

(*D'après BTS Bâtiment 1998*)

Une usine produit des fils de 6 mètres de long utilisés pour la fabrication de panneaux de treillis soudé.

On suppose que la variable aléatoire X qui, à chaque fil choisi dans la production de la journée, associe son diamètre en millimètres, suit une loi normale de moyenne μ et d'écart type $\sigma = 0{,}1$.

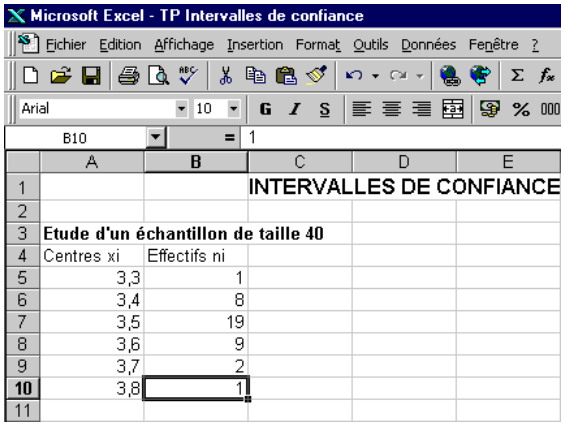
On cherche à estimer μ à partir des résultats d'un échantillon de la production journalière.

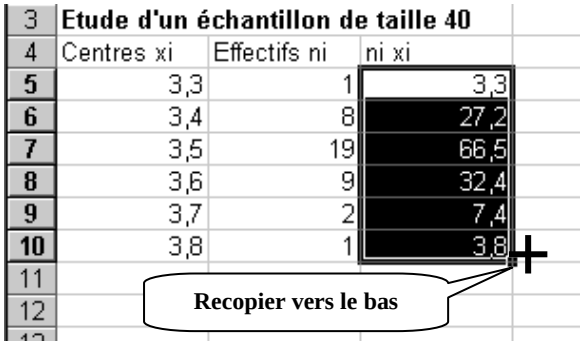
On mesure le diamètre exprimé en millimètres de chacun des n = 40 fils d'un échantillon choisi au hasard et avec remise dans la production d'une journée. On regroupe les résultats en classes dont les effectifs sont les suivants :

classes	]3,25 ; 3,35]	]3,35 ; 3,45]	]3,45 ; 3,55]	]3,55 ; 3,65]	]3,65 ; 3,75]	]3,75 ; 3,85]
effectifs	1	8	19	9	2	1

Etude de l'échantillon

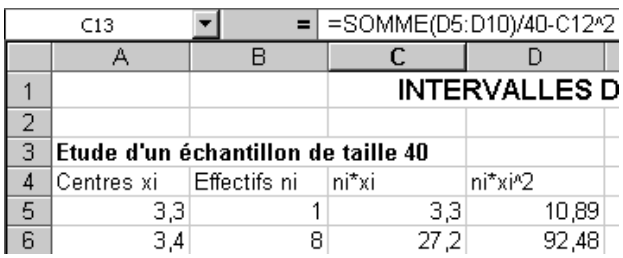
Cliquer sur A3 et taper, en gras, "Etude d'un échantillon de taille 40".
En A4 taper "Centres xi" et entrer de A5 à A10 les centres des classes.
En B4 taper "Effectifs ni" et entrer de B5 à B10 les effectifs correspondants.
En supposant que les effectifs sont regroupés aux centres des classes, vous allez calculer avec le tableur, la moyenne $\bar{x}$ et l'écart type s de l'échantillon.





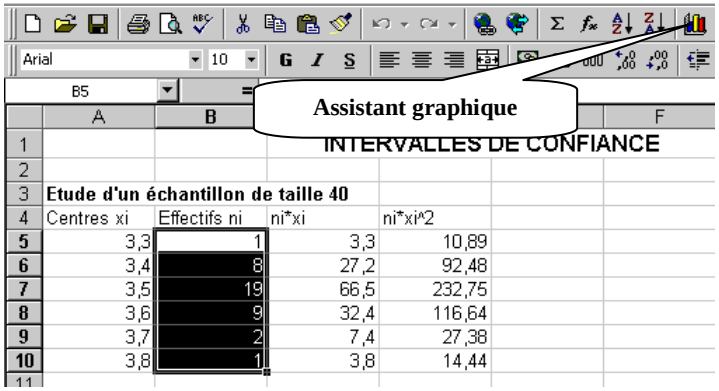
Comme Excel ne possède pas de fonction intégrée permettant le calcul direct d'une moyenne (et d'un écart type) pondérée, vous allez en détailler le calcul.
En C4, taper "ni xi".
En C5, entrer la formule : =A5*B5
puis approcher le pointeur de la souris du carré noir situé dans le coin inférieur droit de la cellule C5. Lorsque le pointeur se transforme en une croix noire, faire glisser en maintenant le bouton gauche enfoncé, jusqu'à la cellule C10 (on a ***recopié la formule vers le bas***).
En A12, inscrire "Moyenne échantillon" et en C12 entrer la formule :
=SOMME(C5:C10)/40
(la formule s'inscrit en même temps dans la ***barre de formule*** en haut).

Pour déterminer l'écart type, procéder ainsi :



En D4 taper "ni*xi^2" (le symbole ^ de puissance s'obtient en appuyant simultanément sur CTRL ALT et la touche ç).
En D5 entrer la formule :
=B5*A5^2 que vous recopiez vers le bas jusqu'en D10.
En A13, taper "Variance échantillon" et en C13, entrer la formule :
=SOMME(D5:D10)/40-C12^2
Enfin, en A14 taper "Ecart type échantillon" et en C14 entrer la formule :
=RACINE(C13)

Représentation sous forme d'histogramme

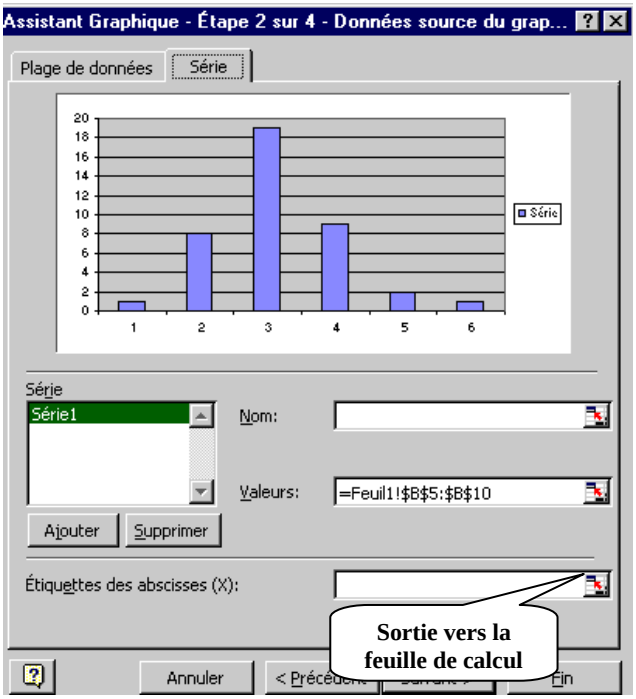


	A	B	C	D
4	Centres xi	Effectifs ni	ni*xi	ni*xi^2
5	3,3	1	3,3	10,89
6	3,4	8	27,2	92,48
7	3,5	19	66,5	232,75
8	3,6	9	32,4	116,64
9	3,7	2	7,4	27,38
10	3,8	1	3,8	14,44

Sélectionner (bouton gauche de la souris enfoncé) la plage des valeurs d'effectifs (de B5 à B10).

Cliquer sur l'icône de l'***assistant graphique***.

Etape 1/4 : Choisir Histogramme et cliquer sur ***Suivant***.
Etape 2/4 : Aller dans l'onglet ***Série*** et à la rubrique ***Etiquette des abscisses*** cliquer sur l'icône de sortie vers la feuille de calcul pour venir y sélectionner les valeurs de A5 à A10 (retour dans l'assistant graphique par le même icône située dans la boîte Assistant graphique).
Cliquer sur ***Suivant***.



Etape 3/4 : Dans l'onglet ***Légende*** désactiver ***Afficher la légende*** puis cliquer sur ***Suivant***.
Etape 4/4 : Cocher l'option ***En tant qu'objet dans la Feuil1*** puis cliquer sur ***Terminer***.

L'histogramme peut être en dehors de la zone d'impression. Pour le déplacer, cliquer (gauche) dans la ***Zone de graphique*** de l'histogramme puis, en maintenant le bouton gauche enfoncé, déplacer le graphique.

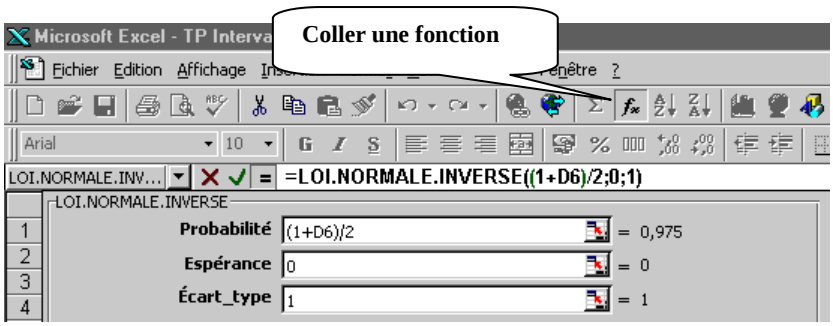
– Compléter la feuille réponse.

2 – DETERMINATION D'INTERVALLES DE CONFIANCE POUR LA MOYENNE μ

Intervalle de confiance à 95%

On cherche, à partir de l'échantillon étudié sur la feuille de calcul 1, à donner un intervalle de confiance pour la moyenne μ des diamètres en millimètres pour la production du jour.
Cliquer sur l'onglet ***Feuil 2*** pour accéder à une autre feuille de calcul.
Dans la cellule A1, taper, en gras, "Intervalles de confiance".
En A3, taper "Taille échantillon" puis en D3 taper 40.
En A4, taper "Moyenne échantillon" puis en D4 taper 3,515.
En A5, inscrire "Ecart type population" et en D5 la valeur 0,1.
En A6, inscrire "Coefficient de confiance" et en D6 la valeur 0,95.
En A7 inscrire "t pour 2π(t)-1=coef confiance" (la lettre π s'obtient en choisissant la ***police Symbol*** puis avec la touche p).

Cliquer sur D7 puis sur l'icône ***Coller une fonction***, choisir ***Statistiques*** puis ***LOI.NORMALE.INVERSE*** et cliquer sur ***OK***.



Dans la boîte de dialogue, inscrire à la rubrique ***Probabilité*** (1 + D6) / 2
(la cellule D6 contient le coefficient de confiance).
Pour ***Espérance*** inscrire la valeur 0 et pour ***Ecart_type*** la valeur 1
(il s'agit de la loi N (0;1)).

Puis cliquer sur ***OK***.

En A8 inscrire "Borne inférieure IC" puis en D8, taper la formule :
=D4-D7*0,1/RACINE(D3)

En A9 inscrire "Borne supérieure IC" puis en D9, taper la formule :
=D4+D7*0,1/RACINE(D3)
En A10, inscrire "Amplitude IC" et en D10, taper la formule : =D9-D8 .

– Inscrire les résultats sur la feuille réponse.

Impact du coefficient de confiance et de la taille de l'échantillon

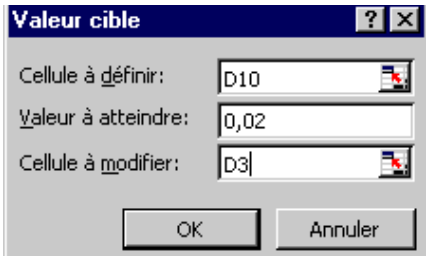
Double cliquer (gauche) sur la cellule D6 de façon à modifier le coefficient de confiance. Remplacer 0,95 par 0,99. Faire ***ENTREE***.

– Inscrire les résultats sur la feuille réponse.

Revenir en D6 à 0,95. Modifier cette fois en D3 la taille de l'échantillon : remplacer 40 par 100 (*on suppose que l'on conserve la même moyenne d'échantillon*).

– Inscrire les résultats sur la feuille réponse.

On désire connaître la taille de l'échantillon de sorte à avoir un intervalle de confiance à 95% pour μ d'amplitude 0,02. Procéder ainsi :
Dérouler le menu ***Outils/Valeur cible...***
Dans la boîte de dialogue, entrer :
Cellule à définir : D10
Valeur à atteindre : 0,02
(attention Excel n'accepte pas 0.02)
Cellule à modifier : D3
Cliquer sur ***OK***.



La réponse étant approximative, modifier le contenu de la cellule D3 de façon à déterminer la taille minimale.

– Inscrire les résultats sur la feuille réponse.

3 – DEPENDANCE D'UN INTERVALLE DE CONFIANCE A L'ECHANTILLON CHOISI

Un intervalle de confiance est calculé à partir d'un échantillon. Pour prendre conscience de cette dépendance, vous allez supposer que X suit la loi normale N (μ ; 0,1) avec μ = 3,504 et simuler la prise de plusieurs échantillons de taille n = 100.

Cliquer sur l'onglet ***Feuil3***.

En A1, taper "échantillons", puis, en B1, entrer la ***formule*** suivante, qui simule une réalisation de la variable aléatoire X de loi N (3,504 ; 0,1) :

=3,504+0,1*COS(2*PI()*ALEA())*RACINE(-2*LN(ALEA()))

Recopier vers le bas la cellule B1 jusqu'en B40. Vous avez simulé un échantillon aléatoire de taille 40.

B44 = =B45+B42

	A	B	C
40		3,49126655	
41			
42	rayon IC95%	0,02600741	
43	inf IC	3,4901088	
44	sup IC	3,54212361	
45	moyenne xi	3,5161162	

En A45, taper "moyenne xi" et, en B45, entrer la ***formule*** : =MOYENNE(B1:B40)

En A42, taper "rayon IC90%". Vous allez calculer le rayon d'un intervalle de confiance à 90 % de μ, grâce à la ***fonction*** INTERVALLE.CONFIANCE d'Excel, connaissant σ=0,1 et n = 40.

En B42, entrer la ***formule*** : =INTERVALLE.CONFIANCE(0,10;0,1;40)

En A43 taper "inf IC", puis, en A44, "sup IC".

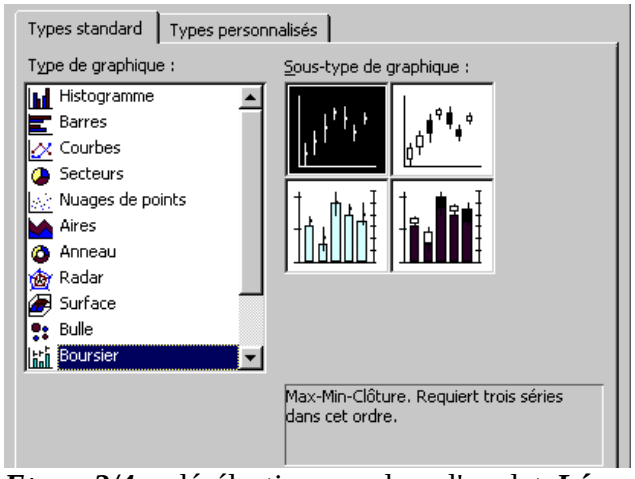
Entrer les ***formules***, en B43 : =B45-B42 puis, en B44 : =B45+B42 .

Ce sont les bornes de l'intervalle de confiance. Celui-ci contient-il la valeur réelle de μ, ici supposée être égale à 0,504 ? Pour le faire apparaître, inscrire en A47 "valeur IC", puis entrer en B47 la ***formule*** : =ET(3,504>=B43;3,504<=B44)

Vous avez sans doute remarqué qu'à chaque nouveau calcul, un autre échantillon aléatoire est généré. On va exploiter ceci pour observer de nombreuses simulations.

Sélectionner les cellules de B1 à B47, puis ***recopier***-les jusqu'à la colonne K.

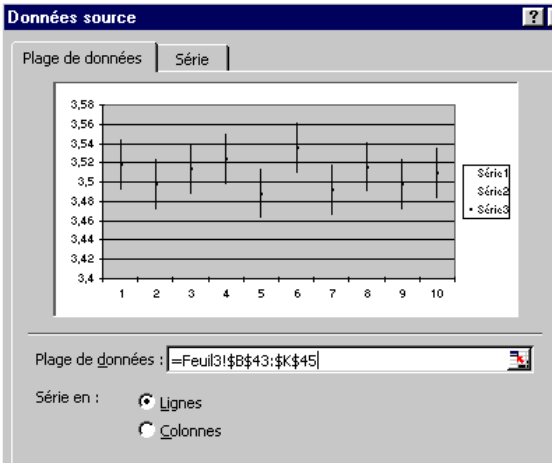
Vous disposez alors des résultats de 10 échantillons. Pour totaliser le nombre d'intervalles ne contenant pas μ, taper en A48, "total FAUX/10" et entrer en B48 la ***formule*** : =NB.SI(B47:K47;FAUX)



Cliquer sur l'icône de l'***assistant graphique***.

Etape 1/4 : choisir le ***type Boursier***, premier sous-type.

Etape 2/4 : sortir sur la feuille de calcul, sélectionner les trois lignes de B43 à K45.



Etape 3/4 : désélectionner, dans l'onglet ***Légende***, l'option ***Afficher la légende***.

Etape 4/4 : cocher • ***sur une nouvelle feuille***.

Pour faire de nouvelles simulations, appuyer sur F9.

– ***Répondre aux questions de la feuille réponse.***

– FEUILLE REPONSE

NOMS : *.....*

1 – ETUDE D'UN ECHANTILLON

Moyenne de l'échantillon de taille 40 : .....
Ecart type de l'échantillon de taille 40 : .....

Si possible, joindre l'impression de la feuille 1 ou enregistrer sur disquette.

2 – DETERMINATION D'INTERVALLES DE CONFIANCE

Impact du coefficient de confiance et de la taille de l'échantillon

Intervalle de confiance à A % pour μ, centré sur $\bar{x}$, obtenu à partir d'un échantillon de taille n (arrondir les résultats à 10^{-3} près) :

Coefficient de confiance	Taille n de l'échantillon	Intervalle de confiance obtenu pour μ	Amplitude de l'intervalle de confiance
95 %	40		
99 %	40		
95 %	100		

• Expliquer les différences d'amplitude des intervalles de confiance,
- Selon la valeur du coefficient de confiance :

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

- Selon la taille de l'échantillon :

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

• Taille n de l'échantillon de sorte que l'intervalle de confiance à 95% soit d'amplitude inférieure à 0,02 : .....
.....
.....

Vérification de l'amplitude pour cette valeur de n : .....
.....

Si possible, joindre l'impression de la feuille 2 ou enregistrer sur disquette.

... / ...

3 – DEPENDANCE D'UN INTERVALLE DE CONFIANCE A L'ECHANTILLON CHOISI

Soit un intervalle de confiance à 90 %, calculé sur un échantillon de 40 valeurs.

- Le nombre μ appartient-il *toujours* à l'intervalle de confiance ? .....

Avec un nouvel échantillon de 40 mesures, on obtient de la même façon un second intervalle de confiance de μ avec le coefficient de confiance 90 %.

- Les deux intervalles de confiance sont-ils obligatoirement les mêmes ? .....
- Ces deux intervalles ont-ils obligatoirement le même centre ? .....
- Ces deux intervalles de confiance peuvent-ils n'avoir aucun élément commun ? .....

Faire 10 simulations de 10 échantillons de taille 40, et noter, à chaque fois, le nombre d'intervalles de confiance à 90 % ne contenant pas μ :

Simulation	1	2	3	4	5	6	7	8	9	10
Nombre de FAUX										

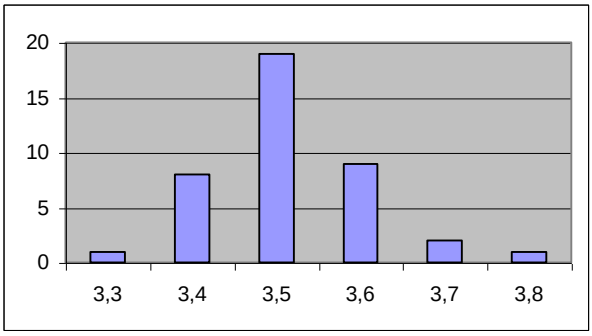
- Sur ces 100 intervalles de confiances, combien, au total, ne contenaient pas μ ? .....
- Sur un grand nombre d'intervalles de confiance à 90 % de μ, obtenus sur un grand nombre d'échantillons, combien ne contiennent pas la valeur à estimer ? .....

Si possible, joindre l'impression du graphique ou enregistrer sur disquette.

Corrigé du TP sur Excel
"INTERVALLES DE CONFIANCE" Durée : 1h30

1 – ETUDE D'UN ECHANTILLON

Moyenne de l'échantillon de taille 40 : $\bar{x} = 3{,}515$.
Ecart type de l'échantillon de taille 40 : $s \approx 0{,}096$ à 10^{-3} près.



2 – DETERMINATION D'INTERVALLES DE CONFIANCE

Impact du coefficient de confiance et de la taille de l'échantillon

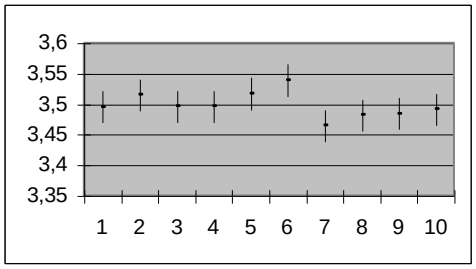
Coefficient de confiance	Taille n de l'échantillon	Intervalle de confiance obtenu pour μ	Amplitude de l'intervalle de confiance
95 %	40	[3,484 ; 3,546]	0,062
99 %	40	[3,474 ; 3,556]	0,081
95 %	100	[3,495 ; 3,535]	0,039

• Expliquer les différences d'amplitude des intervalles de confiance,
- Selon la valeur du coefficient de confiance :
Si l'on augmente le coefficient de confiance (de façon à avoir plus de chances de contenir μ), l'amplitude de l'intervalle augmente (fourchette moins précise).
- Selon la taille de l'échantillon :
Si l'on augmente la taille de l'échantillon, l'amplitude de l'intervalle diminue, car on a une information plus précise avec un échantillon plus grand.
• Taille n de l'échantillon de sorte que l'intervalle de confiance à 95% soit d'amplitude inférieure à 0,02 :
L'outil "valeur cible", donne $n \approx 367$ qui correspond à une amplitude légèrement supérieure à 0,02.
En tatônnant, on trouve que $n = 385$ est la plus petite taille donnant une amplitude, de l'intervalle de confiance à 95 %, inférieure à 0,02.

3 – DEPENDANCE D'UN INTERVALLE DE CONFIANCE A L'ECHANTILLON CHOISI

On obtient des graphiques du type ci-contre :

Sur cette image, les intervalles de confiance obtenus à partir des échantillons 6 et 7, ne contiennent pas la moyenne $\mu = 3{,}504$.



Soit un intervalle de confiance à 90 %, calculé sur un échantillon de 40 valeurs.

- Le nombre μ appartient-il à l'intervalle de confiance ? Pas forcément.

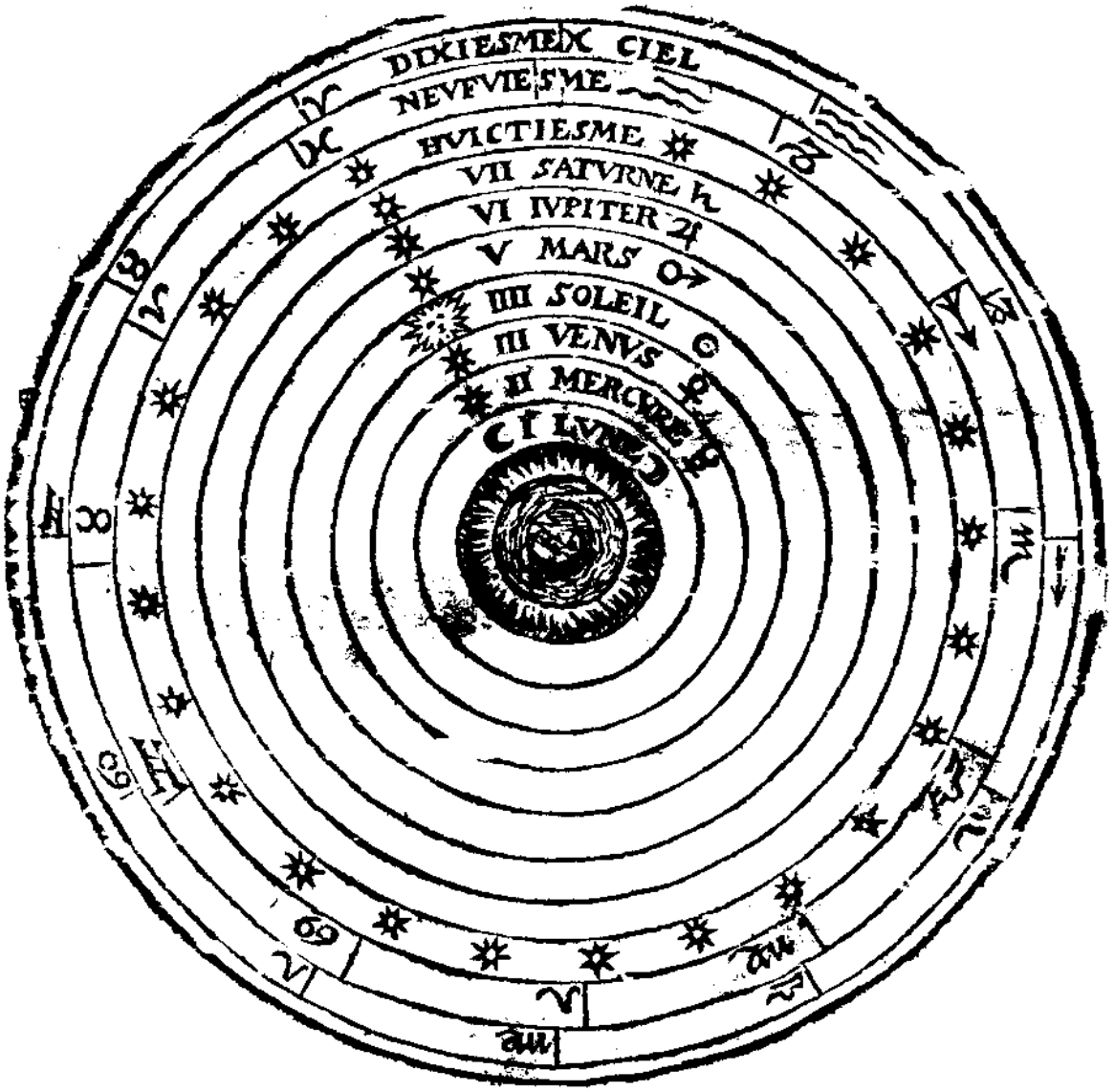
Avec un nouvel échantillon de 40 mesures, on obtient de la même façon un second intervalle de confiance de μ avec le coefficient de confiance 90 %.

- Les deux intervalles de confiance sont-ils obligatoirement les mêmes ? Non.
- Ces deux intervalles ont-ils obligatoirement le même centre ? Non.
- Ces deux intervalles de confiance peuvent-ils n'avoir aucun élément commun ? Oui (dans un cas extrême).

Exemple de 10 simulations de 10 échantillons de taille 40 (soit 100 échantillons), avec le nombre d'intervalles de confiance à 90 % ne contenant pas μ :

Simulation	1	2	3	4	5	6	7	8	9	10
Nombre de FAUX	0	1	0	1	0	3	0	1	2	1

- Sur ces 100 intervalles de confiances, 7 ne contenaient pas μ .
- Sur un grand nombre d'intervalles de confiance à 90 % de μ, obtenus sur un grand nombre d'échantillons, combien ne contiennent pas la valeur à estimer ? 10 % .

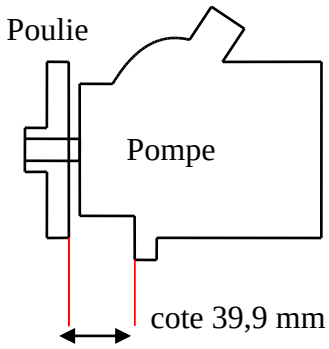

DIXIESME CIEL
NEVFVIESME
HVICTIESME
VII SATVRNE
VI IVPITER
V MARS
IIII SOLEIL
III VENVS
II MERCVRE
I LVNE

TRAVAUX PRATIQUES

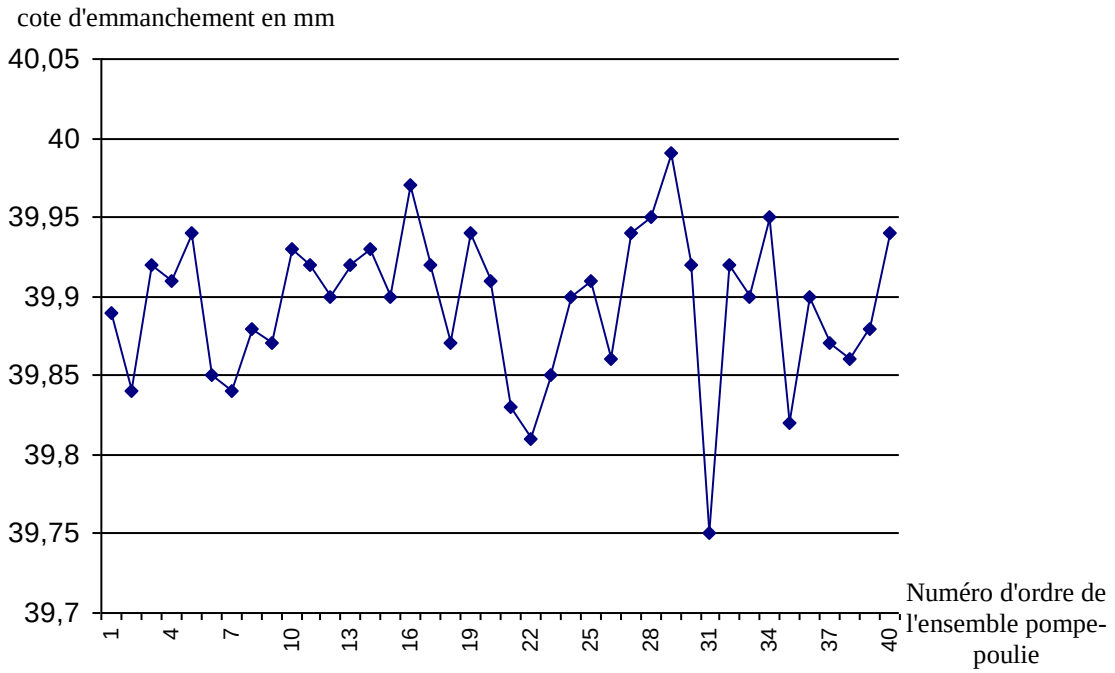
T.P. : LA MAITRISE STATISTIQUE DES PROCEDES DE PRODUCTION

Dans le cadre de la "Maîtrise Statistique des Procédés", on étudie la variabilité de la production. Un des objectifs est de détecter les anomalies, en temps réel.

L'exemple étudié, issu de l'industrie automobile, est une presse d'emmanchement de poulie sur une pompe de direction assistée. Les performances de la presse sont variables, cette variabilité ayant de nombreuses causes possibles : main-d'œuvre, matériel, matière première, environnement de l'atelier, méthodes d'organisation...
L'emmanchement de la poulie sur l'axe de la pompe est mesuré par la cote de 39,9 mm indiquée sur le schéma ci-contre.



On a mesuré cette cote, à 10^{-2} mm près, sur 40 ensembles pompe-poulie, produits de façon successive dans la production en série. Les observations sont représentées, dans l'ordre chronologique, sur le schéma suivant.

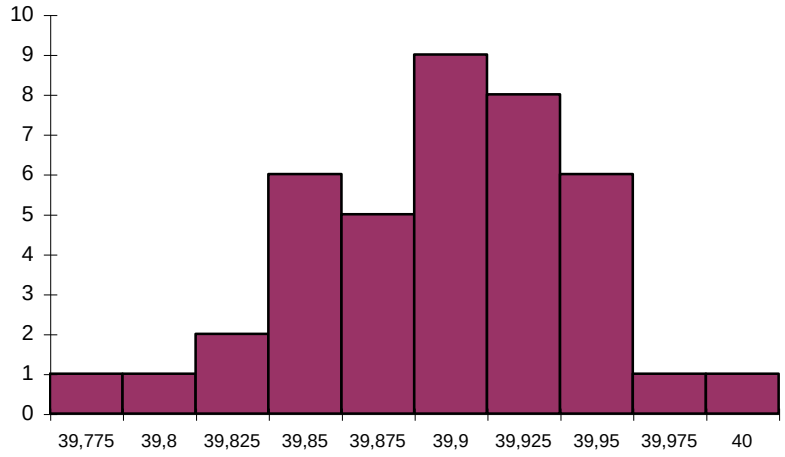


I – ETUDE DE NORMALITE

Une première étape consiste à voir si les variations observées peuvent raisonnablement résulter d'un phénomène suivant une loi normale.

Nous allons d'abord regrouper les 40 observations en 10 classes d'égale amplitude 0,025 et construire un histogramme. Les résultats sont les suivants :

centres des classes	39,775	39,8	39,825	39,85	39,875	39,9	39,925	39,95	39,975	40
effectifs	1	1	2	6	5	9	8	6	1	1



Dans un deuxième temps, nous allons comparer ces résultats observés à ceux, théoriques, correspondant à une variable aléatoire X de loi normale $\mathcal{N}(\mu, \sigma)$. Cette comparaison se fera à l'aide, d'une part des fréquences cumulées observées, et, d'autre part, de la fonction de répartition de X.

1) Fréquences cumulées observées :
On note x_i les bornes supérieures des classes et y_i les fréquences cumulées correspondantes. C'est à dire que y_i est la fréquence des observations inférieures ou égales à x_i.
Calculer la fréquence cumulée de la case "oubliée".

bornes sup des classes x_i	39,7875	39,8125	39,8375	39,8625	39,8875	39,9125	39,9375	39,9625	39,9875	40,0125
fréquences cumulées y_i	0,025	0,05		0,25	0,375	0,6	0,8	0,95	0,975	1

2) Fréquences théoriques selon la loi normale :

Si X suit la loi $\mathcal{N}(\mu, \sigma)$ alors $T = \dfrac{X-\mu}{\sigma}$ suit la loi normale centrée réduite $\mathcal{N}(0, 1)$ et

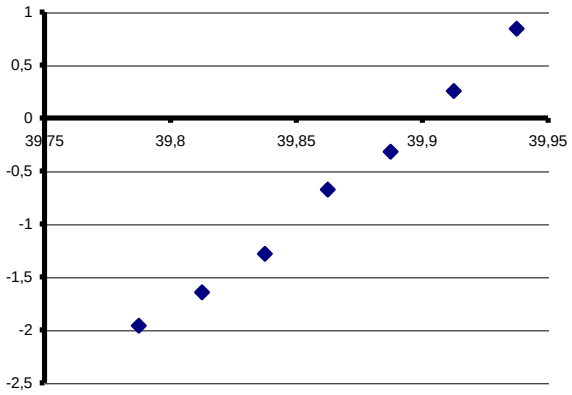
$y_i = F(x_i) = P(X \le x_i) = P\left(\dfrac{X-\mu}{\sigma} \le \dfrac{x_i-\mu}{\sigma}\right) = P(T \le t_i) = \Pi(t_i)$ avec $t_i = \dfrac{x_i-\mu}{\sigma}$ et où Π est la fonction de répartition de la loi $\mathcal{N}(0, 1)$, tabulée dans le formulaire officiel de BTS.

On peut, à partir des fréquences cumulées observées y_i, calculer, par lecture inverse de la table de la loi $\mathcal{N}(0, 1)$, les valeurs t_i telles que $y_i = \Pi(t_i) \Leftrightarrow t_i = \Pi^{-1}(y_i)$.

Calculer, à 10^{-2} près, la valeur t_7 oubliée dans le tableau ci-dessous, c'est à dire telle que $P(T \le t_7) = 0{,}8$ où T suit la loi normale $\mathcal{N}(0, 1)$.

y_i	0,025	0,05	0,1	0,25	0,375	0,6	0,8	0,95	0,975	1
$t_i \approx$	– 1,96	– 1,645	– 1,281	– 0,674	– 0,319	0,253		1,645	1,96	

3) Régression linéaire (droite de *Henry*) :



Si la distribution observée est extraite d'une population normale, on devrait avoir $t_i \approx \frac{1}{\sigma} x_i - \frac{\mu}{\sigma}$.

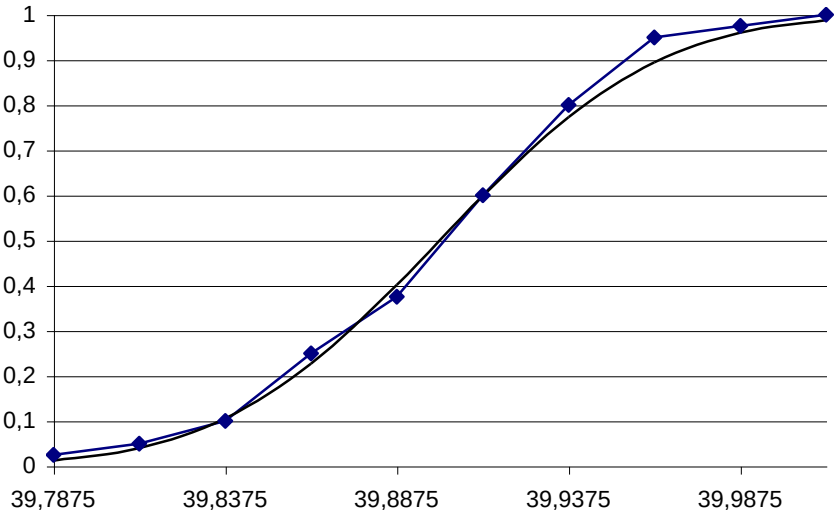
a) Sur le graphique ci-contre, on a représenté les points de coordonnées (x_i, t_i).
Les points semblent-ils pratiquement alignés ?

b) Calculer une équation $t = ax + b$ de la droite d'ajustement de t en x selon la méthode des moindres carrés pour le tableau ci-dessous (on arrondira à 10^{-3} près) :

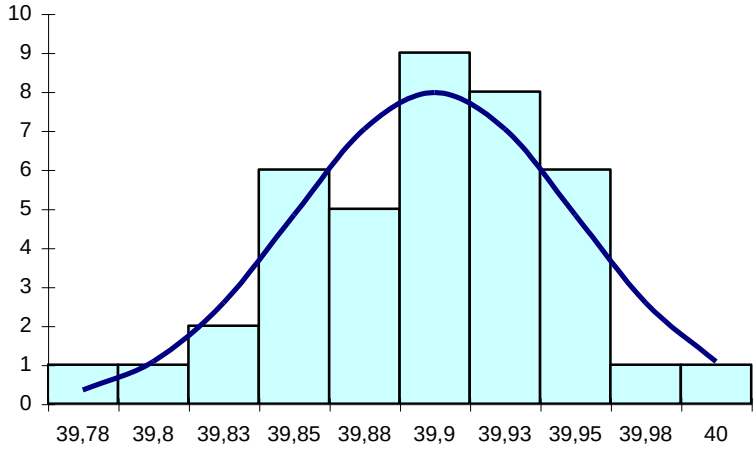
x_i	39,7875	39,8125	39,8375	39,8625	39,8875	39,9125	39,9375	39,9625	39,9875
t_i	– 1,96	– 1,645	– 1,281	– 0,674	– 0,319	0,253	0,842	1,645	1,96

c) Que vaut le coefficient de corrélation r ? Comment peut-on l'interpréter ?
d) Vérifier que, d'après les résultats précédents, on peut approximativement prendre $\sigma = \frac{1}{a} \approx 0{,}05$ et $\mu = -\frac{b}{a} \approx 39{,}9$.

Voici, à titre indicatif, d'autres graphiques permettant de visualiser la normalité des données :



Fréquences cumulées croissantes observées y_i
et fonction de répartition F de la loi $\mathcal{N}$ (39,9 ; 0,05)



Histogramme des observations
et densité de la loi normale $\mathcal{N}$ (39,9 ; 0,05)

II – CAPABILITE

Le bureau d'étude a défini, pour cette cote, l'**intervalle de tolérance** suivant :
[39,9 – 0,2 mm ; 39,9 + 0,2 mm] . C'est à dire, qu'en dehors de cet intervalle, l'emmanchement sera considéré comme non conforme.
Le procédé de fabrication est considéré comme "**capable**" lorsque la probabilité de fabrication d'une pièce hors tolérance est inférieure à 2 °/$_{oo}$.
Si le procédé n'est pas capable, il fabriquera, en quantité inacceptable, des pièces hors normes. Dans ce cas, avant de le mettre sous contrôle, on essaiera au préalable de l'améliorer, pour le rendre capable.
1) On suppose que la variable aléatoire X qui, à chaque ensemble poulie-pompe choisi au hasard, associe sa cote d'emmanchement en mm, suit la loi normale $\mathcal{N}$ (39,9 ; 0,05).
Calculer la probabilité que cette cote soit acceptable, c'est à dire : $P(39{,}7 \leq X \leq 40{,}1)$.
2) Peut-on considérer le procédé de fabrication comme capable ?

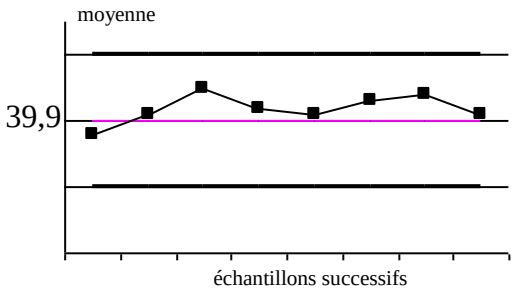
III – CARTE DE CONTROLE POUR LES MOYENNES

Pour la surveillance de la production à venir, on envisage d'établir une "carte de contrôle aux moyennes". On supposera que la valeur de σ reste stable mais qu'une dérive sur la valeur de μ est à craindre. La cote de l'emmanchement pompe-poulie étant un "point Sécurité-Réglementation", la norme prévoit de prélever régulièrement des échantillons de n = 5 ensembles pompe-poulie, sur lesquels on calculera la moyenne des 5 cotes d'emmanchement.

1 – Dérive systématique d'un côté de la moyenne

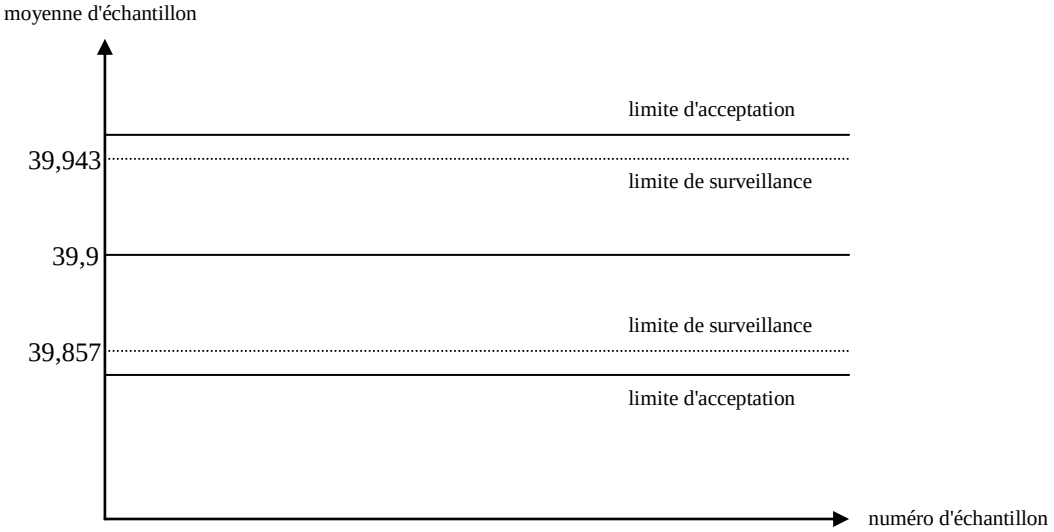
On suppose ici que X suit la loi $\mathcal{N}$ (39,9 ; 0,05).
On considère l'expérience aléatoire consistant à prélever avec remise un échantillon de 5 ensembles pompe-poulie dans la production et à calculer la moyenne $\bar{x}$ des cotes d'emmanchement de cet échantillon.
Soit A l'évènement "la moyenne $\bar{x}$ est supérieure ou égale à 39,9". Compte-tenu de la symétrie de la loi normale, on a $P(A) = 0{,}5$.
On prélève, de façon indépendante, N échantillons.

a) Montrer que la variable aléatoire *Y* qui, à ces *N* échantillons, associe le nombre de fois que l'évènement *A* s'est produit, suit la loi binomiale de paramètres *N* et 0,5.
b) Calculer, en fonction de *N*, la probabilité que chacun des *N* échantillons amène une moyenne supérieure à 39,9.
c) Déterminer le plus petit entier *N* tel que $0{,}5^{N} \leq 0{,}01$, puis le plus petit entier *N* tel que $0{,}5^{N} \leq 0{,}002$.
d) Justifier la règle selon laquelle, si apparaissent 7 échantillons successifs de moyenne supérieure à 39,9 (comme sur la figure), l'alerte est donnée et, dans le cas de 9 échantillons successifs au dessus de la moyenne, la production est arrêtée pour réglage.



2 – Limites de surveillance et d'acceptation

On suppose ici que la variable aléatoire *X* qui, à chaque ensemble pompe-poulie prélevé au hasard dans la production, associe la cote d'emmanchement exprimée en mm, suit la loi normale $\mathcal{N}$ (39,9 ; 0,05).
a) Soit $\overline{X}$ la variable aléatoire qui, à chaque échantillon de 5 ensembles pompe-poulie prélevé au hasard et avec remise, associe la moyenne des cotes d'emmanchement de cet échantillon.
Justifier que $\overline{X}$ suit la loi normale de moyenne 39,9 et d'écart type 0,022 à 10^{-3} près.
b) Des limites de surveillance sont fixées, sur la carte de contrôle, à 39,857 mm et 39,943 mm. Calculer, à 10^{-2} près, $P(39{,}857 \leq \overline{X} \leq 39{,}943)$.



c) Pour calculer les limites d'acceptation, dont le dépassement provoque l'arrêt de la production, calculer, à 10^{-3} près, le réel *h* tel que $P(39{,}9 - h \leq \overline{X} \leq 39{,}9 + h) = 0{,}998$.

Corrigé des travaux pratiques
"Maîtrise statistique des procédés de production"

I – Normalité

1) La fréquence cumulée manquante est $y_3 = \frac{4}{40} = 0{,}1$.

2) On a $P(T \leq t) = 0{,}8 \Leftrightarrow \Pi(t) = 0{,}8$ ce qui donne, par lecture inverse de la table des valeurs de la fonction de répartition Π, $t \approx 0{,}84$.

3) a) Les points de coordonnées (x_i , t_i) sont pratiquement alignés.

3 b) On obtient, à l'aide de la calculatrice, une équation de la droite de régression de t en x selon la méthode des moindres carrés : $t = ax + b$ avec $a \approx 20{,}482$ et $b \approx -817{,}107$.

3 c) Le coefficient de corrélation linéaire est $r \approx 0{,}994$ qui est très proche de 1. Ce qui va dans le sens d'un ajustement par une loi normale.

3 d) On a alors $\sigma = 1/a \approx 0{,}049$ puis $\mu = -b/a \approx 39{,}894$.

II – Capabilité

1) On trouve que $P(39{,}7 \leq X \leq 40{,}1) \approx 1$ (la calculatrice donne 0,99994).

Remarque : la condition d'une probabilité de 2°/$_{oo}$ hors tolérance correspond à un écart à la moyenne de l'ordre de 3σ pour une loi normale.

2) La probabilité précédente étant supérieure à 0,9973 , on peut donc considérer le processus comme capable.

III – Carte de contrôle pour la moyenne

1) Dérive systématique d'un côté de la moyenne

a) On a la répétition de N expériences aléatoires indépendantes, avec deux issues possibles (A se réalise, avec la probabilité 0,5 , ou non), où Y associe à ces N expériences, le nombre de fois que A s'est réalisé. Donc Y suit la loi $\mathbb{B}$ (N ; 0,5).

b) On a $P(Y = N) = 0{,}5^N$.

c) On a $0{,}5^N \leq 0{,}01 \Leftrightarrow N \geq \ln(0{,}01) / \ln(0{,}5)$ soit $N \geq 6{,}64$. Le plus petit N est donc 7.

On a $0{,}5^N \leq 0{,}002 \Leftrightarrow N \geq \ln(0{,}002) / \ln(0{,}5)$ soit $N \geq 8{,}96$. Le plus petit N est donc 9.

d) Il y a moins de 1% de chances que le processus soit sous contrôle (avec μ = 39,9) lorsqu'on observe 7 points consécutifs au-dessus de 39,9. La probabilité tombe à moins de 2 $^0/_{00}$ dans le cas de 9 points consécutifs au-dessus de 39,9. Il y a donc lieu, dans ce dernier cas, d'arrêter la production, pour contrôle.

On a, bien sûr, des règles analogues, pour l'observation de points consécutifs en-dessous de39,9.

2) Limites de surveillance et d'acceptation

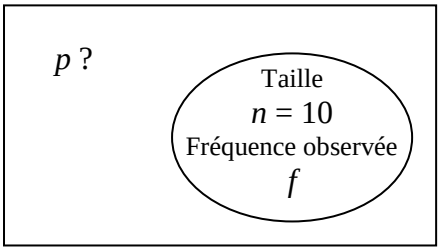
a) On sait, d'après le cours sur l'échantillonnage, que si X suit la loi $\mathbb{N}$ (39,9 ; 0,05) alors $\overline{X}$ suit la loi normale $\mathbb{N}$ (39,9 ; $\frac{0{,}05}{\sqrt{5}}$), soit un écart type d'environ 0,022 à 10^{-3} près.

b) On a $P(39{,}857 \leq \overline{X} \leq 39{,}943) \approx 0{,}95$ à 10^{-2} près.

c) On cherche h tel que $2\Pi(h/0{,}022) - 1 \approx 0{,}998$ d'où $h \approx 0{,}068$ à 10^{-3} près.

Travaux pratiques STATISTIQUES

T.P. Excel : LOI BINOMIALE ET INTERVALLE DE CONFIANCE



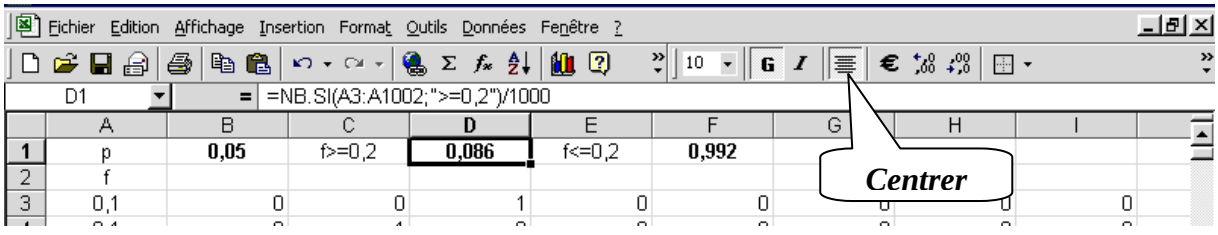
On considère une production dont la fréquence p d'éléments défectueux est inconnue.
On prélève dans cette production un échantillon de $n = 10$ pièces et on observe la fréquence f d'éléments défectueux de cet échantillon, qui peut être considéré comme prélevé au hasard et avec remise.
On cherche à estimer p par un intervalle de confiance avec le coefficient de confiance de 95%, mais la taille n de l'échantillon est trop petite pour pouvoir approcher par une loi normale la loi de la variable aléatoire F qui à tout échantillon de taille n associe la fréquence de ses éléments défectueux.

I – APPROCHE STATISTIQUE PAR SIMULATION

Dans un échantillon de taille $n = 10$ on observe 2 éléments défectueux. Donc $f = 0,2$.
On cherche, à l'aide de $f = 0,2$, à donner une estimation de p par un intervalle de confiance avec le coefficient de confiance 95%.

1 – Simulation pour *p* connu

Imaginons d'abord que l'on connaisse la valeur de p, soit $p = 0,05$ par exemple, et simulons 1000 échantillons de taille 10.
Ouvrir un fichier ***Excel***.
En A1 écrire p et en B1 entrer la valeur 0,05.
En A2 écrire f puis en B3 entrer la ***formule*** =ENT(ALEA()+$B$1)
Cliquer sur la cellule B3 et approcher le pointeur de la souris du coin inférieur droit. Quant le pointeur se transforme en une croix noire, ***recopier vers la droite*** en gardant le bouton gauche de la souris enfoncé, jusqu'en K3. Les 1 obtenus indiquent les éléments défectueux dans un premier échantillon de taille 10.
En A3 entrer la ***formule*** =SOMME(B3:K3)/10 qui donne la valeur de f pour cet échantillon.
Sélectionner les cellules de A3 à K3 puis ***recopier vers le bas*** jusqu'à la ligne 1002 pour simuler 1000 échantillons.
Cliquer sur le 1 de la première ligne puis sur l'icône ***Centrer***.
Centrer de même la colonne A.



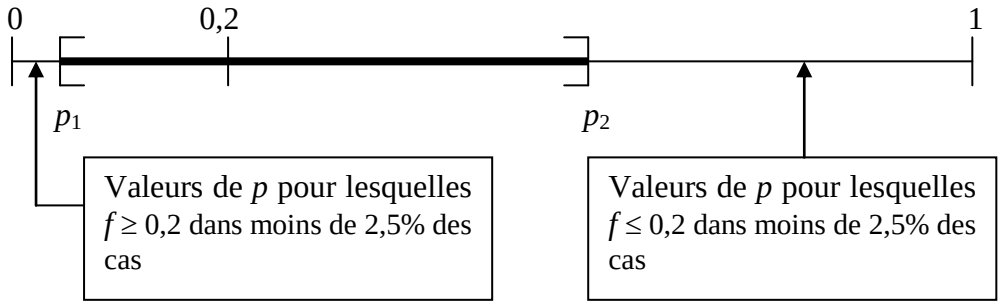
En C1 écrire f>=0,2 et entrer en D1 la ***formule*** =NB.SI(A3:A1002;">=0,2")/1000

qui correspond à la fréquence des valeurs de f supérieures ou égales à 0,2 sur les 1000 échantillons que vous avez simulés.
En E1 écrire f<=0,2 puis entrer en F1 la ***formule*** donnant la fréquence des valeurs de f inférieures ou égales à 0,2 sur les 1000 échantillons que vous avez simulés.

– Compléter la feuille réponse.

2 – Recherche expérimentale d'un intervalle de confiance

En réalité p est inconnu et la fréquence du seul échantillon prélevé est $f = 0,2$.
A partir de la fréquence de cet échantillon, nous allons déterminer les extrémités p_1 et p_2 de l'intervalle de confiance de p au coefficient de confiance 95% de la façon suivante :



– Changer en B2 les valeurs de p de façon à compléter la feuille réponse.

II – UTILISATION DE LA LOI BINOMIALE

1 – Intervalle de confiance pour p avec $f = 0,2$

Cliquer sur l'onglet ***Feuil2***.

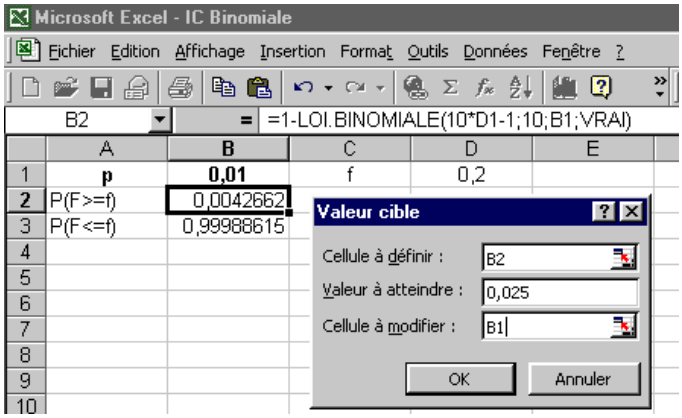
B2 | = | =1-LOI.BINOMIALE(10*D1-1;10;B1;VRAI)

	A	B	C	D	E
1	**p**	**0,01**	f	0,2	
2	P(F>=f)	0,0042662			
3	P(F<=f)	0,99988615			

Ecrire en A1 p, puis entrer en B1 sa valeur, par exemple 0,01.
Ecrire en C1 f, puis entrer en D1 sa valeur 0,2.
En A2 écrire P(F>=f).

Soit F la variable aléatoire qui associe à tout échantillon de taille 10 la fréquence f de ses éléments défectueux. La variable aléatoire $10F$ correspond au nombre d'éléments défectueux de l'échantillon et suit la loi binomiale $\mathcal{B}(10, p)$.
On a $P(F \geq 0,2) = 1 - P(F \leq 0,1) = 1 - P(10F \leq 1)$.
En B2 entrer la ***formule*** =1 – LOI.BINOMIALE(10*D1 – 1;10;B1;VRAI)
La valeur logique VRAI indique qu'il s'agit d'un calcul cumulé.
En A3 écrire P(F<=f) puis en B3 la ***formule*** =LOI.BINOMIALE(10*D1;10;B1;VRAI)

On va chercher à déterminer la valeur p_1 de p telle que $P(F \geq 0,2) \approx 0,025$.



Cliquer dans la cellule B2 puis dans le menu ***Outils*** choisir ***Valeur cible…*** puis compléter la boîte de dialogue comme ci-contre.
Faire ***OK***.
Recopier sur la feuille réponse la valeur p_1 obtenue, arrondie à 10^{-2} près.

Déterminer la valeur p_2 de p telle que $P(F \leq 0{,}2) \approx 0{,}025$ en utilisant l'***Outil Valeur cible…*** à partir de la cellule B3.

– Compléter la feuille réponse.

2 – Intervalle de confiance pour p avec $f = 0{,}4$

On suppose cette fois que l'échantillon prélevé amène 4 éléments défectueux sur les 10 prélevés. Modifier la cellule D1 puis utiliser l'***Outil Valeur cible…***

– Compléter la feuille réponse.

– FEUILLE REPONSE

NOMS : *.....*

I – APPROCHE STATISTIQUE PAR SIMULATION

1 – Simulation pour *p* connu

Pour $p = 0{,}05$ quelle est, parmi les 1000 échantillons que vous avez simulés, la fréquence de ceux pour lesquels $f \geq 0{,}2$? .....
En A2 remplacer la valeur de p par 0,01.
Pour $p = 0{,}01$ quelle est, parmi les 1000 échantillons que vous avez simulés, la fréquence de ceux pour lesquels $f \geq 0{,}2$? .....
Effectuer d'autres simulations en appuyant sur la touche ***F9***. Lorsque $p = 0{,}01$ a-t-on beaucoup de chances d'observer sur un échantillon une fréquence $f = 0{,}2$? .....

2 – Recherche expérimentale d'un intervalle de confiance

En B1 entrer pour p la valeur 0,2 puis faire plusieurs fois F9.
$p = 0{,}2$ se situe-t-il avant ou après p_1 ? .....
.....
Entrer en B1 pour p la valeur 0,03 puis 0,04 , faire plusieurs fois ***F9*** puis cocher la réponse apportée par vos simulations : θ $0{,}03 \leq p_1$ θ $0{,}03 \geq p_1$ θ $0{,}04 \leq p_1$ θ $0{,}04 \geq p_1$.
Entrer en B1 pour p la valeur 0,54 puis 0,57 , faire plusieurs fois ***F9*** puis cocher la réponse apportée par vos simulations : θ $0{,}54 \leq p_2$ θ $0{,}54 \geq p_2$ θ $0{,}57 \leq p_2$ θ $0{,}57 \geq p_2$.

II – UTILISATION DE LA LOI BINOMIALE

1 – Intervalle de confiance pour *p* avec *f* = 0,2

Indiquer l'intervalle de confiance $[p_1 , p_2]$ pour p avec le coefficient de confiance 95% obtenu pour p à partir de la valeur $f = 0{,}2$.
.....
.....
Le centre de cet intervalle est-il la fréquence $f = 0{,}2$ de l'échantillon prélevé ?
.....

2 – Intervalle de confiance pour *p* avec *f* = 0,4

Indiquer l'intervalle de confiance $[p_1 , p_2]$ pour p avec le coefficient de confiance 95% obtenu pour p à partir de la valeur $f = 0{,}4$.
.....
.....

Corrigé du TP Excel
"ESTIMATION D'UNE FREQUENCE PAR UN INTERVALLE DE CONFIANCE A L'AIDE D'UNE LOI BINOMIALE"

I – APPROCHE STATISTIQUE PAR SIMULATION

1 – Simulation pour *p* connu

	A	B	C	D	E	F	G	H	I
1	p	**0,01**	f>=0,2	**0,003**	f<=0,2	**1**			
2	f								
3	0,1	0	0	0	1	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0

Lorsque $p = 0{,}01$ on observe que l'on a $f \geq 0{,}2$ dans moins de 1% des cas, ce qui rend peu probable l'observation de $f = 0{,}2$.

2 – Recherche expérimentale d'un intervalle de confiance

	A	B	C	D	E	F	G	H	I
1	p	**0,02**	f>=0,2	**0,015**	f<=0,2	**0,997**			
2	f								
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0

	A	B	C	D	E	F	G	H	I
1	p	**0,03**	f>=0,2	**0,052**	f<=0,2	**0,998**			
2	f								
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0

D'après ces simulations, on a, en observant la valeur de la cellule D1 par rapport à 2,5% : $0{,}02 \leq p_1 \leq 0{,}03$.

	A	B	C	D	E	F	G	H	I
1	p	**0,54**	f>=0,2	**0,992**	f<=0,2	**0,028**			
2	f								
3	0,5	0	0	1	0	1	1	0	0
4	0,6	0	1	0	1	0	1	1	1

	A	B	C	D	E	F	G	H	I
1	p	**0,57**	f>=0,2	**0,998**	f<=0,2	**0,018**			
2	f								
3	0,7	1	1	1	0	1	0	1	1
4	0,5	1	1	1	1	0	0	0	0

D'après ces simulations, on a, en observant la valeur de la cellule F1 par rapport à 2,5% : $0{,}54 \leq p_2 \leq 0{,}57$.

II – UTILISATION DE LA LOI BINOMIALE

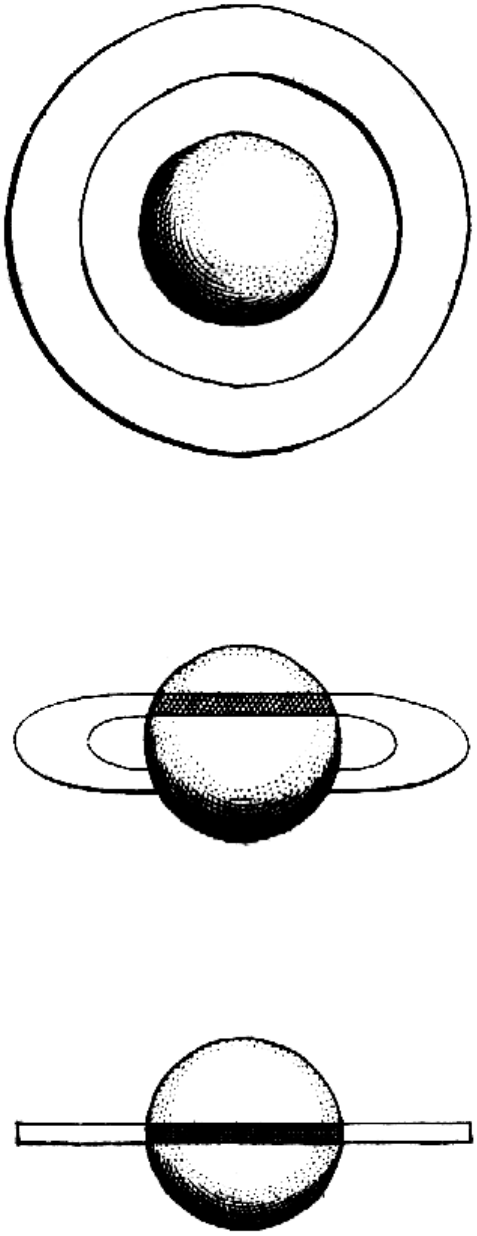
1 – Intervalle de confiance pour *p* avec *f* = 0,2

On trouve [0,025 ; 0,556] comme intervalle de confiance de p avec le coefficient de confiance 95% à partir de $f = 0{,}2$.
La valeur observée $f = 0{,}2$ n'est pas le centre de cet intervalle.

2 – Intervalle de confiance pour *p* avec *f* = 0,4

On trouve [0,122 ; 0,738] comme intervalle de confiance de p avec le coefficient de confiance 95% à partir de $f = 0{,}4$.

Différentes positions de Saturne avec son anneau.



Epreuves corrigées de B.T.S.

Annales du B.T.S. : ECHANTILLONNAGE ET ESTIMATION

DISTRIBUTION D'ECHANTILLONNAGE D'UNE FREQUENCE

1 – Analyses biologiques 1999

On décide de construire un test qui, à la suite des contrôles sur un échantillon de 50 sportifs prélevé au hasard, permette de décider si, au seuil de signification de 10 %, le pourcentage de sportifs susceptibles d'être contrôlés positifs est de $p = 0{,}02$.
Soit F la variable aléatoire qui, à tout échantillon aléatoire (supposé non exhaustif) de 50 sportifs contrôlés, associe le pourcentage de sportifs contrôlés positivement.

On suppose que F, suit la loi normale $\mathrm{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ où $p = 0{,}02$ et $n = 50$.

Déterminer, le réel positif a tel que $P(p - a \leq F \leq p + a) = 0{,}9$.

DISTRIBUTION D'ECHANTILLONNAGE D'UNE MOYENNE

2 – Construction navale 1998 (Carte de contrôle aux moyennes)

Une entreprise produit en grande série des bagues de verrouillage utilisées pour la fabrication de raccords rapides. Ces bagues, en alliages, sont obtenues sur une presse automatique. Plusieurs outillages sont en service suivant les alliages utilisés et les dimensions à obtenir.
On suppose que la variable aléatoire X qui, à toute bague prélevée au hasard dans la fabrication de l'entreprise, associe son diamètre intérieur exprimé en millimètres, suit la loi normale de moyenne m = 24,20 et d'écart type $\sigma = 0{,}045$.

1° La procédure de contrôle est basée sur des prélèvements à intervalles réguliers, d'échantillons de 20 bagues. La production est suffisamment importante pour qu'on puisse assimiler un prélèvement de 20 bagues à un prélèvement aléatoire avec remise.
On nomme $\overline{X}$ la variable aléatoire qui à tout prélèvement de 20 bagues associe la moyenne des diamètres intérieurs des 20 bagues.
a - Quelle est la loi de probabilité suivie par $\overline{X}$?
Quels sont les paramètres de cette loi ?
b - Déterminer un intervalle $[Ls_1 , Ls_2]$ centré en m = 24,20 et tel que la variable aléatoire $\overline{X}$ prenne une valeur dans cet intervalle avec la probabilité 0,95.
(Les nombres Ls_1 et Ls_2 s'appellent les limites de surveillance).
c - Déterminer un intervalle $[Lc_1 , Lc_2]$ centré en m = 24,20 et tel que la variable aléatoire $\overline{X}$ prenne une valeur dans cet intervalle avec la probabilité 0,99.

(Les nombres Lc_1 et Lc_2 s'appellent les limites de contrôle).

2° La procédure de contrôle prévoit qu'après chaque prélèvement, on mesure les diamètres intérieurs des 20 bagues obtenues, on calcule leur moyenne $\overline{x}$ et
- si $\overline{x}$ n'appartient pas à l'intervalle $[Lc_1 , Lc_2]$ on procède immédiatement aux réglages,
- si $\overline{x}$ appartient à l'un des intervalles $[Lc_1 , Ls_1]$ ou $[Ls_2 , Lc_2]$ on prélève immédiatement un autre échantillon (procédure d'alerte).
Sur un prélèvement de 20 bagues, on obtient les mesures suivantes :

diamètre x_i	[24,09 ; 24,11]	[24,11 ; 24,13]	[24,13 ; 24,15]	[24,15 ; 24,17]	[24,17 ; 24,19]
effectif n_i	1	2	3	4	3
diamètre x_i	[24,19 ; 24,21]	[24,21 ; 24,23]	[24,23 ; 24,25]	[24,25 ; 24,27]	[24,27 ; 24,29]
effectif n_i	2	2	1	1	1

Quelle décision doit-on prendre au vu de ce prélèvement ?

3 – D'après Groupement B 2000

Dans cette question, on veut contrôler la moyenne μ de l'ensemble des diamètres, en mm, des pieds de boulons constituant un stock très important.

On note Y la variable aléatoire qui, à chaque boulon tiré au hasard dans le stock, associe le diamètre, en mm, de son pied.
La variable aléatoire Y suit la loi normale de moyenne inconnue μ et d'écart type $\sigma = 0,1$.

On désigne par $\overline{Y}$ la variable aléatoire qui, à chaque échantillon aléatoire de 100 boulons prélevé dans un stock, associe la moyenne des diamètres des pieds de ces 100 boulons (le stock est assez important pour que l'on puisse assimiler ces prélèvements à des tirages avec remise).

a) Justifier que, sous l'hypothèse que la moyenne de la variable aléatoire Y est 10, la variable aléatoire $\overline{Y}$ suit la loi normale de moyenne 10 et d'écart type 0,01.

b) Déterminer le nombre réel positif h tel que $P(10 - h \leq \overline{Y} \leq 10 + h) = 0,95$.

4 – Plastiques et composites 1998 (recherche de la taille de l'échantillon)

Un atelier produit en grande série des pièces cylindriques.
On désigne par X la variable aléatoire associant, à chaque pièce tirée au hasard dans la production, son diamètre x, en millimètres.
On suppose que X suit la loi normale de moyenne 12,50 et d'écart type 0,02.
On note $\overline{X}$ la variable aléatoire qui, à tout échantillon de n pièces pris au hasard et avec remise dans la production, associe la moyenne des diamètres des n pièces.
On note P_n la probabilité que cette variable aléatoire $\overline{X}$ appartienne à l'intervalle :
[12,495 ; 12,505].

On suppose n assez grand et on rappelle que dans ce cas $\overline{X}$ suit approximativement la loi normale de moyenne 12,50 et d'écart type $\frac{0{,}02}{\sqrt{n}}$.

En utilisant cette loi, déterminer la taille minimale n de l'échantillon pour que la probabilité P_n soit supérieure ou égale à 0,97.

5 – Conception et réalisation de carrosseries 1998 (recherche de *n*)

Une machine perce un trou dans une patte de fixation de bouclier avant.
La variable aléatoire qui, à chaque patte de fixation associe le diamètre du trou mesuré en millimètres suit la loi normale de moyenne $\mu = 16{,}56$ et d'écart type $\sigma = 0{,}19$.
On prélève dans la production un échantillon de taille n.

Calculer n pour que la moyenne des diamètres sur les pièces prélevées appartienne à l'intervalle [16,4 ; 16,72] avec la probabilité 0,95.

ESTIMATION D'UNE FREQUENCE

6 – Informatique de gestion 1998

Dans un pays voisin, on doit bientôt élire le président de la république. Afin d'apprécier ses chances, le candidat A fait procéder à un sondage un mois avant la date du scrutin.
On tire au hasard 900 personnes dans l'ensemble de tous les électeurs (compte tenu du nombre total d'électeurs, le tirage peut être assimilé à un tirage avec remise).
Sur ces 900 personnes, 435 ont déclaré voter pour le candidat A.
1° Donner une estimation ponctuelle, à 10^{-2} près, de la proportion p d'électeurs favorables au candidat A.
2° On note F_n la variable aléatoire qui, à tout échantillon de n électeurs, associe la proportion p_n d'électeurs favorables à A. (On sait que F_n, suit approximativement une loi normale de moyenne p et d'écart type $\sqrt{\frac{p(1-p)}{n}}$).
On considère alors le sondage précédent (n = 900 et 435 personnes sur 900 sont favorables à A). Donner, pour l'estimation de p, un intervalle de confiance avec le coefficient de confiance 95 %. Les bornes de cet intervalle seront données à 10^{-2} près.
On sait que, dans ce cas, on commet une erreur faible en remplaçant, dans l'écart type, $\sqrt{p(1-p)}$ par $\sqrt{p_n(1-p_n)}$.
3° Un organe de presse désire publier le résultat du sondage. Au vu des résultats précédents, la diffusion de l'intervalle de confiance peut-elle intéresser les lecteurs ? Pourquoi ?

7 – Constructions métalliques 1999

Une société fabrique des poutrelles mécaniques.
On cherche à estimer le pourcentage p, inconnu, des poutrelles défectueuses de la production.
On admet que la variable aléatoire F, qui, à tout échantillon de 100 poutrelles prises au hasard dans la production (prélèvement assimilé à un tirage avec remise), associe le pourcentage de poutrelles défectueuses de cet échantillon, suit la loi normale N (p, $\sqrt{\frac{p(1-p)}{n}}$)
On prélève un tel échantillon dans la production. On constate qu'il contient 3 poutrelles défectueuses parmi les 100.

1) Donner, à partir de cet échantillon, une estimation ponctuelle f du pourcentage p.

2) Déterminer une estimation de p par un intervalle de confiance centré en f avec un coefficient de confiance égal à 90 %. Pour les bornes de cet intervalle, on donnera des valeurs approchées comportant deux décimales.

8 – Opticien lunetier 1999

Un opticien vend beaucoup d'appareils photographiques.
L'opticien s'intéresse aux appareils autofocus avec zoom. Il veut estimer par un intervalle de confiance le pourcentage p d'acheteurs d'appareils autofocus avec zoom dans sa clientèle.

1) Dans un échantillon de 100 clients, 60 achètent un tel appareil.
a) Donner l'estimation ponctuelle de p fournie par cet échantillon.
b) Donner une estimation de p par un intervalle de confiance avec le coefficient de confiance 95 %.

2)Déterminer la taille n, n étant supérieur à 30, d'un échantillon de clients pour qu'un intervalle de confiance de p, centré sur l'estimation ponctuelle trouvée précédemment soit [0,557 ; 0,643] avec le coefficient de confiance 90 %.

ESTIMATION D'UNE MOYENNE

9 – Groupement C 2000 (écart type connu)

On s'intéresse au diamètre de pièces d'un certain type fabriquées dans une entreprise.
On admet que la variable aléatoire *X* qui, à chaque pièce prélevée au hasard, associe son diamètre exprimé en millimètres, suit une loi normale de moyenne 250 et d'écart type 2.

Après un certain temps de fonctionnement de la machine, pour vérifier le bien fondé de l'hypothèse précédente, on s'intéresse à la moyenne des diamètres des pièces produites.
Pour cela, on étudie un échantillon de 100 pièces prises au hasard et avec remise dans la production.
La moyenne $\bar{x}$ des diamètres des pièces de cet échantillon est égale à 249,7.
On suppose que la variable aléatoire $\overline{X}$ qui, à tout échantillon de 100 pièces prélevées au hasard et avec remise, associe la moyenne des diamètres de ces pièces suit une loi normale de **moyenne inconnue** μ et d'écart type $\frac{2}{\sqrt{100}}$.

Au vu de l'échantillon, déterminer un intervalle de confiance centré en $\bar{x}$ de la moyenne μ avec le coefficient de confiance 95%.

10 – Groupement D 1999 (estimation ponctuelle de l'écart type)

Une entreprise fabrique des pots de peinture.
Elle les fait livrer habituellement par lots de 20 pots ou de 100 pots. On se propose d'étudier les variations de la quantité d'un certain produit A contenu dans chaque pot.
On a contrôlé le dosage du produit A à la sortie de deux chaînes de fabrication.
Deux échantillons de 100 pots ont été analysés ; l'un provient de la chaîne n° l, l'autre de la chaîne n°2.
Le tableau suivant donne la répartition de l'échantillon de la chaîne n° l en fonction de la masse de produit A exprimée en grammes.

m (en g)	[100, 102[	[102, 104[	[104, 106[	[106, 108[	[108, 110[	[110, 112[	[112, 114[	[114, 116[
Effectifs	1	3	25	32	27	6	4	2

On donne des valeurs approchées de la moyenne m_2 et de l'écart type s_2 de l'échantillon fabriqué par la chaîne n° 2 : $m_2 = 107$ et $s_2 = 2$ (en grammes).
Dans les questions 1 et 2 les valeurs seront arrondies au dixième le plus proche.

l° En prenant les centres des classes, calculer une valeur approchée de la moyenne m_1 et de l'écart type s_1 de l'échantillon issu de la chaîne n° l.

2° En considérant les résultats obtenus dans la première question, donner les estimations ponctuelles :
- des quantités moyennes μ_1 et μ_2 de produit A pour les productions de ces deux chaînes,
- des écarts types σ_1 et σ_2 correspondants.

11 – Chimiste 1999 (estimation ponctuelle de l'écart type)

Deux laboratoires A et B fabriquent des tubes à essai et les conditionnent dans des paquets. Tous les paquets contiennent le même nombre de tubes.

1. On note X_1 la variable aléatoire prenant pour valeur le nombre de tubes défectueux par paquet provenant de l'entreprise A. Sur un échantillon aléatoire de 49 paquets provenant du laboratoire A les nombres des tubes défectueux par paquet sont les suivants :

7	5	5	4	4	4	9	7	9	2	7	8	7	8	4
4	9	10	5	10	6	4	5	6	1	2	5	7	8	0
6	0	1	5	2	0	5	2	3	3	4	1	3	10	1
0	10	2	7											

Calculer une valeur approchée à 10^{-2} près de la moyenne m_1 et de l'écart type s_1 de cet échantillon. On admet dans la suite de cet exercice qu'une estimation ponctuelle $\hat{\mu}_1$ de la moyenne μ_1 de la variable aléatoire X_1 est 4,84 et qu'une estimation ponctuelle $\hat{\sigma}_1$ de l'écart type σ_1 de X_1 est 2,99.

2. On note X_2 la variable aléatoire prenant pour valeur le nombre de tubes défectueux par paquet provenant de l'entreprise B. Sur un échantillon aléatoire de 64 paquets provenant de l'entreprise B on a obtenu une moyenne m_2 de 3,88 tubes défectueux et un écart type s_2 de 1,45.
En déduire une estimation ponctuelle $\hat{\mu}_2$ de la moyenne μ_2 de la variable aléatoire X_2 et une estimation ponctuelle $\hat{\sigma}_2$ de l'écart type σ_2 de X_2.

12 – Groupement B 1999 (écart type connu)

Une entreprise de matériel pour l'industrie produit des modules constitués de deux types de pièces : P_1 et P_2.
Dans cette question on s'intéresse au diamètre des pièces P_2.
Soit $\overline{X}$ la variable aléatoire qui, à tout échantillon de 60 pièces P_2 prélevées au dans la production de la journée considérée, associe la moyenne des diamètres des pièces de cet échantillon.
La production est assez importante pour qu'on puisse assimiler tout prélèvement à un tirage avec remise.
On suppose que $\overline{X}$ suit la loi normale de moyenne μ et d'écart type $\frac{\sigma}{\sqrt{60}}$ avec $\sigma = 0{,}084$.
On mesure le diamètre, exprimé en centimètres, de chacune des 60 pièces P_2 d'un échantillon choisi au hasard et avec remise dans la production d'une journée.
On constate que la valeur approchée arrondie à 10^{-3} près de la moyenne $\overline{x}$ de cet échantillon est $\overline{x} = 4{,}012$.

a) A partir des informations portant sur cet échantillon, donner une estimation ponctuelle, à 10^{-3} près, de la moyenne μ du diamètre des pièces P_2 produites pendant cette journée.

b) Déterminer un intervalle de confiance centré en $\overline{x}$ de la moyenne μ des diamètres des pièces P_2 produites pendant la journée considérée, avec le coefficient de confiance 95%.

c) On considère l'affirmation suivante : "la moyenne μ est obligatoirement entre 3,991 et 4,033". Peut-on déduire de ce qui précède qu'elle est vraie ?

13 – Bâtiment 1998 (écart type inconnu)

Une usine adhérente de "l'ADETS" (Association pour le développement du treillis soudé) fabrique en grande série différents types de treillis soudés pour la construction.
L'usine produit des fils de 6 mètres de long utilisés pour la fabrication de panneaux de treillis soudé de type "903".
On mesure le diamètre exprimé en millimètres de chacun des 40 fils d'un échantillon choisi au hasard et avec remise dans la production d'une journée.
On constate que les valeurs approchées arrondies, à 10^{-3} près, de la moyenne $\bar{x}$ et de l'écart type s des diamètres pour cet échantillon sont $\bar{x} = 3{,}512$ et $s = 0{,}095$.

1° Proposer une estimation ponctuelle, à 10^{-3} près, de la moyenne μ et de l'écart type σ du diamètre des fils produits pendant cette journée.

2° Soit $\bar{X}$ la variable aléatoire qui, à tout échantillon de taille n = 40 prélevé au hasard et avec remise, associe la moyenne des diamètres des fils de cet échantillon. On suppose que $\bar{X}$ suit la loi normale $N\left(\mu , \frac{\sigma}{\sqrt{40}}\right)$. On prend pour valeur de σ l'estimation ponctuelle obtenue au 1°.
a) Déterminer un intervalle de confiance centré en $\bar{x}$ de la moyenne μ de la population, avec le coefficient de confiance 95 % .
b) Le nombre μ appartient-il à l'intervalle de confiance obtenu au a) ? Expliquer.
Le nombre μ a-t-il plus de chances d'être inférieur à $\bar{x}$ que d'être supérieur à $\bar{x}$? Expliquer.

14 – Opticien lunetier 2000

On s'intéresse à la longueur de l'ensemble des pièces produites par une machine le 6 octobre 1999. Elles sont assemblées par lots de 50. Un lot pris au hasard est considéré comme un échantillon de la fabrication. Le tableau suivant décrit la distribution des longueurs des pièces de ce lot.

Longueur des pièces*	Nombre de pièces
[107 ; 108[	1
[108 ; 109[	6
[109 ; 110[	14
[110; 111[	20
[111 ; 112[	9

*exprimée en mm

1 . On suppose que la longueur d'une pièce est égale à la valeur centrale de la classe dans laquelle elle est répertoriée. Calculer la moyenne $\bar{x}$ des longueurs des pièces de l'échantillon, ainsi que son écart type s . On retiendra pour $\bar{x}$ et s leurs valeurs arrondies au centième. Le détail des calculs n'est pas demandé.

2. Déduire des résultats obtenus à la question précédente, une estimation ponctuelle μ de la moyenne des longueurs des pièces de la fabrication.

3. On admet que la variable aléatoire $\overline{L}$ qui, à chaque échantillon de 50 pièces, associe la moyenne des longueurs de pièces de cet échantillon, suit la loi normale $\mathcal{N}(\mu, \frac{1}{\sqrt{50}})$.

Estimer par un intervalle de confiance la moyenne des longueurs des pièces de la fabrication avec le coefficient de confiance 0,9.

15 – Productique textile 1999

Dans les parties A et B on étudie la charge de rupture d'un fil de "soie discontinue" stocké sur un enrouleur.

A) Pour cette partie, les valeurs numériques seront arrondies au millième.

On a fait un essai dynamométrique sur un fil de soie discontinue stocké sur un enrouleur.
Les 12 éprouvettes de fil lestées sont prélevées au hasard sur la partie extérieure de l'enroulement.
On a obtenu les résultats suivant :

numéro de l'éprouvette	1	2	3	4	5	6	7	8	9	10	11	12
charge de rupture en newtons	2,4	2,4	2,1	2,4	2,3	2,3	1,8	2,3	2,5	2,1	1,9	2,1

On note X la variable aléatoire qui, à chaque éprouvette prélevée au hasard sur la partie extérieure de l'enroulement, associe la charge de rupture en newtons de cette éprouvette. Déduire, du tableau ci-dessus, une estimation ponctuelle de l'espérance mathématique μ et de l'écart type σ de la variable aléatoire X.

B) 1) Une autre série de 35 mesures dynamométriques d'éprouvettes prélevées au hasard sur le fond de l'enroulement et exprimées en newtons a donné les résultats suivants regroupés par classes de même amplitude :

classes	[1,6 ; 1,7]	]1,7 ; 1,8]	]1,8 ; 1,9]	]1,9 ; 2,0]	]2,0 ; 2,1]	]2,1 ; 2,2]	]2,2 ; 2,3]	]2,3 ; 2,4]
effectifs	2	3	5	9	8	4	3	1

Calculer la moyenne $\bar{y}$ de l'échantillon des 35 éprouvettes.

2) On note Y la variable aléatoire qui, à chaque éprouvette prélevée au hasard sur le fond de l'enroulement, associe la charge de rupture, exprimée en newtons, de cette éprouvette.
On note respectivement μ_1 et σ_1 la moyenne et l'écart type de la variable aléatoire Y.
On note $\overline{Y}$ la variable aléatoire qui, à tout échantillon de taille 35 prélevé au hasard dans le fond de l'enroulement (tirage assimilé à un tirage avec remise), associe la moyenne des 35 éprouvettes prélevées.

On admet que la variable aléatoire $\overline{Y}$ suit la loi normale $\mathcal{N}(\mu_1 ; \frac{\sigma_1}{\sqrt{35}})$.

On suppose σ_1 connu : $\sigma_1 = 0{,}17$.

Déterminer un intervalle de confiance au niveau 99% de la moyenne μ_1 de la charge de rupture d'une éprouvette prélevée au fond de l'enroulement.

16 – Agro-équipement 1999

9 – agro-equipement 99

Un atelier fabrique des joints d'un certain modèle, utilisés dans la construction de moteurs. Dans cet exercice on s'intéresse à la durée de vie de tels joints pendant l'état de marche d'un moteur.
On prélève au hasard, dans l'ensemble des joints fabriqués, un échantillon de 64 unités. La mesure de la durée de vie des joints de cet échantillon a fourni les résultats suivants :

Durée de vie en heures	[500 ; 900[	[900 ; 1100[	[1100 ; 1300[	[1300 ; 1500[	[1500 ; 1700[	[1700 ; 2100[
Nombre de joints	2	9	13	25	11	4

1) On fait l'approximation suivante : dans chaque classe, tous les éléments son placés au centre de la classe. Calculer alors la moyenne et l'écart type de cette série statistique (donner la valeur exacte de la moyenne et une valeur approchée à 10^{-1} près de l'écart type).

2) On s'intéresse maintenant à l'ensemble des joints fabriqués ; leurs durées de vie constitue une série statistique de moyenne μ **inconnue** et d'écart type σ estimé à 260 heures.

On désigne par $\overline{X}$ la variable aléatoire qui, à tout échantillon de 64 joints, prélevé au hasard dans la production, associe la moyenne des durées de vie de ces joints. Ce tirage peut être assimilé à un tirage avec remise.
On rappelle que $\overline{X}$ suit la loi normale $\mathrm{N}\left(\mu, \frac{260}{\sqrt{64}}\right)$.
En utilisant l'échantillon précédent, déterminer un intervalle de confiance, pour la moyenne μ relative à l'ensemble des joints, avec le coefficient de confiance 95 %.

17 – Chimiste 2000

Dans la fabrication de comprimés effervescents, il est prévu que chaque comprimé doit contenir 1625 mg de bicarbonate de sodium. Afin de contrôler la fabrication de ces médicaments, on a prélevé un échantillon de 150 comprimés et on a mesuré la quantité de bicarbonate de sodium pour chacun d'eux. Les résultats obtenus sont résumés dans le tableau suivant :

Classes	[1610, 1615[	[1615, 1620[	[1620, 1625[	[1625, 1630[	[1630, 1635[
Effectifs	7	8	42	75	18

1) En convenant que les valeurs mesurées sont regroupées au centre de chaque classe, calculer une valeur approchée à 10^{-2} près de la moyenne $\bar{x}$ et de l'écart type *s* de cet échantillon.

2) A partir des résultats obtenus pour cet échantillon, assimilé à un échantillon non exhaustif, donner les estimations ponctuelles $\hat{\mu}$ et $\hat{\sigma}$ de la moyenne μ et de l'écart type σ de la quantité de bicarbonate de sodium dans la population (formée de l'ensemble de tous les comprimés fabriqués et supposée très grande).

Dans la question suivante on prendra pour valeur de σ son estimation $\hat{\sigma}$.

3) On appelle $\overline{X}$ la variable aléatoire qui, à tout échantillon de taille *n* = 150 associe la quantité moyenne de bicarbonate de sodium de cet échantillon.

a) La loi de $\overline{X}$ peut-elle être approchée par une loi classique ? Si oui, laquelle ? Donner ses paramètres ?

b) Déterminer un intervalle de confiance de la quantité moyenne de bicarbonate de sodium dans la population avec le coefficient de confiance 95 %.

Calculer l'amplitude de cet intervalle et interpréter le résultat.

c) Quelle devrait être la taille minimum de l'échantillon prélevé pour connaître avec le coefficient de confiance 95 % la quantité moyenne de bicarbonate de sodium dans la population à 1 mg près

18 – Groupement D 2000 (σ inconnu et "petite" taille ! Recherche du seuil)

On étudie le résultat de la pesée d'un objet de masse *m* (exprimée en grammes).

On admet que la variable aléatoire *X* qui prend comme valeurs les résultats (exprimés en grammes) de la pesée d'un même objet donné suit la loi normale de moyenne μ et d'écart type σ..

PARTIE B
Dans cette partie, on suppose que μ et σ sont inconnus.
On a relevé dans le tableau suivant les résultats de 10 pesées d'un même objet :

masse en grammes	72,20	72,24	72,26	72,30	72,36	72,39	72,42	72,48	72,50	72,54

Les résultats seront arrondis au centième le plus proche.

1) Calculer la moyenne et l'écart type de cet échantillon.

2) En déduire des estimations ponctuelles de la moyenne μ et d'écart type σ de la variable *X*.

3) Dans la suite, on admet que la variable aléatoire qui à tout échantillon de 10 pesées associe la moyenne de ces pesées suit une loi normale. En prenant pour écart type la valeur estimée en 2), donner un intervalle de confiance au seuil de 5% de la moyenne μ.

4) L'écart type de l'appareil de pesée, mesuré à partir de nombreuses études antérieures, est en réalité, pour un objet ayant environ cette masse, de 0,08. Dans cette question, on prend donc $\sigma = 0{,}08$

Donner un intervalle de confiance au seuil de 5 % de la moyenne μ.
Déterminer α (à l'unité près) pour que au seuil de α %, un intervalle de confiance de μ soit [72,31 ; 72,43].

* Remarque : Le programme des STS exclut l'étude d'échantillons de taille inférieure à 30 lorsque l'écart type est inconnu (l'utilisation de la loi de *Student* est alors préférable !).

1 – Analyses biologiques 99

On appelle F la variable aléatoire qui, à tout échantillon de 50, associe le pourcentage de sportifs contrôlés positivement.
On admet que sous l'hypothèse $p = 0,02$, F suit la loi normale $\mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ où $p = 0,02$ et $n = 50$.
Donc F suit $\mathcal{N}$ (0,02 ; 0,0198).
On pose $T = \frac{F - 0,02}{0,0198}$ alors T suit la loi normale centrée réduite.
Cherchons un réel a positif tel que :
$P(0,02 - a \leq F \leq 0,02 + a) = 0,9$,

$P(-\frac{a}{0,0198} \leq T \leq \frac{a}{0,0198}) = 0,9$

$2\,P(T \leq \frac{a}{0,0198}) - 1 = 0,9$,

$P(T \leq \frac{a}{0,0198}) = 0,95$,

Par lecture inverse de la table de la loi normale centrée réduite on a $\frac{a}{0,0198} = 1,645$

donc $a = 1,645 \times 0,0198$, $a = 0,0326$.
Remarque : n'étant pas dans les conditions habituelles d'approximation de la loi binomiale par la loi normale ($np > 5$), celle-ci semble peu adéquate.

2 – Construction navale 98

1° La variable aléatoire X suit la loi normale de moyenne 24,20 et d'écart type 0,045, nous savons que la variable $\overline{X}$ suit alors la loi normale de moyenne
E($\overline{X}$) = 24,20 et d'écart type
$\sigma' = \frac{0,045}{2\sqrt{5}} \approx 0,01006$ donc $\sigma' = 0,01$ à 10^{-4} près.

2° La variable aléatoire $\overline{X}$ suit la loi normale $\mathcal{N}$ (24,2 ; 0,01), la variable aléatoire $T = \frac{\overline{X} - 24,2}{0,01}$ suit la loi normale centrée réduite $\mathcal{N}$ (0, 1).
P(24,2 – a ≤ $\overline{X}$ ≤ 24,2 + a) = 0,95 équivaut à

P($-\frac{a}{0,01} \leq T \leq \frac{a}{0,01}$) = 0,95 et à

$2\,\pi(\frac{a}{0,01}) - 1 = 0,95$ et à $\frac{a}{0,01} = 1,96$,
a = 0,0196, $[Ls_1, Ls_2] \approx [24,180, 24,220]$.

2° P($-\frac{a}{0,01} \leq T \leq \frac{a}{0,01}$) = 0,99 et à

$2\,\pi(\frac{a}{0,01}) - 1 = 0,99$ et à $\frac{a}{0,01} = 2,575$,
a = 0,02575 $[Lc_1, Lc_2] \approx [24,174, 24,226]$.

3° On trouve avec une calculatrice $\overline{x} = 24,178$,
$\overline{x} \in [Lc_1, Ls_1]$ donc on doit prélever immédiatement un autre échantillon.

3 – Groupement B 2000

1° Si Y suit la loi normale $\mathcal{N}$ (10 ; 0,1) ; la variable aléatoire $\overline{Y}$ suit la loi normale $\mathcal{N}$ (10 ; $\frac{0,1}{\sqrt{100}}$), donc $\overline{Y}$ suit la loi normale $\mathcal{N}$ (10 ; 0,01) .

2° $\overline{Y}$ suit la loi normale $\mathcal{N}$ (10 ; 0,01) donc la variable aléatoire $U = \frac{\overline{Y} - 10}{0,01}$ suit la loi normale centrée, réduite $\mathcal{N}$ (0, 1).
$P(10 - h \leq \overline{Y} \leq 10 + h) =$
$P(\frac{10 - h - 10}{0,01} \leq \overline{Y} \leq \frac{10 + h - 10}{0,01})$,

$P(10 - h \leq \overline{Y} \leq 10 + h) = P(-\frac{h}{0,01} \leq \overline{Y} \leq \frac{h}{0,01})$,

$P(10 - h \leq \overline{Y} \leq 10 + h) = 2\,\Pi(\frac{h}{0,01}) - 1$

$P(10 - h \leq \overline{Y} \leq 10 + h) = 0,95$ si et seulement si
$2\,\Pi(\frac{h}{0,01}) - 1 = 0,95$ qui est équivalent à

$\Pi(\frac{h}{0,01}) = 0,975$, $\Pi((1,96) = 0,975$

donc $\frac{h}{0,01} = 1,96$, $h = 0,0196$, $h \approx 2$.

4 – Plastiques et composites 98

La variable aléatoire $\overline{X}$ qui, à tout échantillon aléatoire non exhaustif de taille n associe sa moyenne, suit la loi normale $\mathcal{N}$ (12,5 ; $\frac{0,02}{\sqrt{n}}$),

la variable aléatoire $T = \dfrac{\overline{X} - 12{,}5}{\dfrac{0{,}02}{\sqrt{n}}}$ suit la loi normale centrée réduite N (0,1).

P(12,495 ≤ $\overline{X}$ ≤ 12,505) ≥ 0,97 équivaut à

$P(-\dfrac{0{,}005}{0{,}02}\sqrt{n} \le T \le \dfrac{0{,}005}{0{,}02}\sqrt{n}) \ge 0{,}97$ et à

$\pi(0{,}25\sqrt{n}) \ge 0{,}985$, $0{,}25\sqrt{n} \ge 2{,}17$ d'où

$\sqrt{n} \ge \dfrac{2{,}17}{0{,}25}$, $\sqrt{n} \ge 8{,}68$,

$n \ge 75{,}34$, la valeur minimale de n est donc 76.

5 – Conception et réalisation de carrosseries 98

La variable aléatoire X suit la loi normale N (16,56 ; 0,19). La variable aléatoire $\overline{X}$ qui, à tout échantillon aléatoire non exhaustif de taille n associe sa moyenne, suit la loi normale N (16,56 ; $\dfrac{0{,}19}{\sqrt{n}}$), la variable aléatoire

$T = \dfrac{\overline{X} - 16{,}56}{0{,}19}\sqrt{n}$ suit la loi normale centrée réduite N (0,1).

P(16,4 ≤ $\overline{X}$ ≤ 16,72) = 0,95 équivaut à

$P(-\dfrac{0{,}16}{0{,}19}\sqrt{n} \le T \le \dfrac{0{,}16}{0{,}19}\sqrt{n}) = 0{,}95$ et à

$2\pi(\dfrac{0{,}16}{0{,}19}\sqrt{n}) - 1 = 0{,}95$ et à

$\pi(\dfrac{0{,}16}{0{,}19}\sqrt{n}) = 0{,}975$,

$\dfrac{0{,}16}{0{,}19}\sqrt{n} = 1{,}96$ d'où

$\sqrt{n} = \dfrac{1{,}96 \times 0{,}19}{0{,}16}$, $\sqrt{n} \approx 2{,}33$

$n = 5{,}41$, la valeur minimale de n est donc 6.

6 – Informatique de gestion 1998

1° Une estimation ponctuelle de p est $\dfrac{435}{900}$ fréquence obtenue dans l'échantillon prélevé, donc $p = 0{,}48$ à 10^{-2} près, une estimation ponctuelle de $\sqrt{p(1-p)}$ est $\sqrt{0{,}48(1-0{,}48)} \approx 0{,}49959$ d'où une estimation ponctuelle de l'écart type à 10^{-2} près : $\sigma_F = \dfrac{0{,}5}{30}$, $\sigma_F = 0{,}017$ à 10^{-3} près .

2° Un intervalle de confiance avec le coefficient de confiance de 95% correspondant à cet échantillon est : $[f - 1{,}96\ \sigma_F ; f + 1{,}96\ \sigma_F] =$

$[\dfrac{435}{900} - 1{,}96 \times 0{,}017 ; \dfrac{435}{900} + 1{,}96 \times 0{,}017]$.

On obtient [0,45 ; 0,52].

3° La partie la plus étendue de l'intervalle se trouve inférieure à 0,5 soit inférieure à 50 %, on peut donc estimer que le candidat A a peu de chance d'être élu.

7 – Constructions métalliques 99

a) Une estimation ponctuelle de p est $f = \dfrac{3}{100}$, $f = 0{,}03$ pourcentage obtenu dans l'échantillon.

b) Une estimation par intervalle de confiance de p avec le coefficient de confiance de 90% correspondant à cet échantillon est approximativement :

$[f - t_\alpha\sqrt{\dfrac{f(1-f)}{n}}\ ,\ f + t_\alpha\sqrt{\dfrac{f(1-f)}{n}}\ ]$

$2\ \pi(t_\alpha) - 1 = 0{,}90$ équivaut à $\pi(t_\alpha) = \dfrac{1{,}90}{2}$,

$t_\alpha = 1{,}645$, $f = 0{,}03$ et $n = 100$. On obtient :

$[0{,}03 - 1{,}645\sqrt{\dfrac{0{,}03 \times 0{,}97}{100}} ; 0{,}03 + 1{,}645\sqrt{\dfrac{0{,}03 \times 0{,}97}{100}}\ ]$,

$I_C = [0{,}00 ; 0{,}06]$ à 10^{-2} près .

Remarque : la condition d'approximation $nf > 5$ n'étant pas vérifiée, on peut plutôt consulter l'abaque de la page 75 qui fournit pour $n = 100$ et $f = 0{,}3$, l'intervalle : [0,005 ; 0,08].

8 – Opticien lunetier 1999

1) a) Le pourcentage de clients est $f = \dfrac{60}{100}$, $f = 0{,}6$. Une estimation ponctuelle de p est 0,06.

b) Une estimation par intervalle de confiance de p avec le coefficient de confiance de 95% correspondant à cet échantillon est approximativement :

$[f - t_\alpha\sqrt{\dfrac{f(1-f)}{n}}\ ,\ f + t_\alpha\sqrt{\dfrac{f(1-f)}{n}}\ ]$

$2\ \pi(t_\alpha) - 1 = 0{,}95$ équivaut à $\pi(t_\alpha) = \dfrac{1{,}95}{2}$,

$t_\alpha = 1{,}96$, $f = 0{,}6$ et $n = 100$. On obtient :

$[0{,}6 - 1{,}96\sqrt{\dfrac{0{,}6 \times 0{,}4}{100}} ; 0{,}6 + 1{,}96\sqrt{\dfrac{0{,}6 \times 0{,}4}{100}}\ ]$,

$I_C = [0{,}504 ; 0{,}696]$ à 10^{-3} près .

2) Si on conserve 0,6 comme valeur de f et comme estimation de p, un intervalle de confiance de p

avec le coefficient de confiance de 90 % correspondant à un échantillon de taille n est :

$I_C = [0{,}6 - 1{,}645\sqrt{\frac{0{,}6\times 0{,}4}{n}}\ ;\ 0{,}6 + 1{,}645\sqrt{\frac{0{,}6\times 0{,}4}{n}}\ ]$, si cet intervalle est égal à [0,557 ; 0,643] alors

$0{,}643 - 0{,}557 = 2 \times 1{,}645 \times \frac{\sqrt{0{,}24}}{\sqrt{n}}$,

$\sqrt{n} = \frac{2\times 1{,}645\times\sqrt{0{,}24}}{0{,}086}$, $\quad \sqrt{n} \approx 18{,}74$,

$n = 351$.

9 – Groupement C 2000

On désigne par $\overline{X}$ la variable aléatoire qui à tout échantillon non exhaustif de 100 pièces associe la moyenne des diamètres de ces 100 pièces,

$\overline{X}$ suit la loi normale $\mathrm{N}(\mu\ ;\ \frac{2}{\sqrt{100}})$.

$I = [\ \bar{x} - t\ \frac{2}{\sqrt{100}}\ ,\ \bar{x} + t\frac{2}{\sqrt{100}}\ ]$ où $\bar{x}$ est la moyenne d'un échantillon aléatoire non exhaustif.
Si le coefficient de confiance est $2\ \pi(t) - 1 = 0{,}95$, alors $\pi(t) = 0{,}975$ et $t \approx 1{,}96$.

$I = [249{,}7 - 1{,}96 \times 0{,}2\ ;\ 249{,}7 + 1{,}96 \times 0{,}2]$,

$I = [249{,}308\ ;\ 250{,}092]$ à 10^{-3} près.

10 – GROUPEMENT D 99

1° Avec une calculatrice on obtient, en grammes
$m_1 = 107{,}5$ et $s_1 = 2{,}5$

2° a) Pour estimation de la moyenne μ de la population on prend la moyenne m de l'échantillon, donc μ_1 et μ_2 sont estimés respectivement par 107,5 et 107.

b) Pour estimation ponctuelle de l'écart type σ de la population on prend $s\sqrt{\frac{n}{n-1}}$ donc

σ_1 est estimé par $2{,}5\sqrt{\frac{100}{99}}$, $\sigma_1 \approx 2{,}5$ et

σ_2 est estimé par $2\sqrt{\frac{100}{99}}$, $\sigma_2 \approx 2$.

11 - CHIMISTE 99

1. $m_1 = 4{,}84$; $s_1 = 2{,}96$ à 10^{-2} près.

2. $m_2 = 3{,}88$; $s_2 = 1{,}45$ alors μ_2 est estimé par $\hat{\mu}_2 = 3{,}88$ et σ_2 par $\hat{\sigma}_2 \approx 1{,}46$.

12 – GROUPEMENT B 99

a) Pour estimation de la moyenne μ de la population on prend la moyenne $\bar{x}$ de l'échantillon, donc on estime μ par 4,012 à 10^{-3} près.

b) La variable aléatoire $\overline{X}$ suit la loi normale $\mathrm{N}\ (\mu, \frac{\sigma}{\sqrt{n}})$; donc la variable aléatoire $\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ suit la loi normale centrée, réduite $\mathrm{N}\ (0, 1)$.

L'intervalle $[\ \bar{x} - 1{,}96\ \frac{\sigma}{\sqrt{n}}\ ;\ \bar{x}\ + 1{,}96\frac{\sigma}{\sqrt{n}}\ ]$ est, par définition, l'intervalle de confiance de la moyenne μ de la population avec le coefficient de confiance 95 %. D'après a) $\bar{x}\ = 4{,}012$; $\sigma = 0{,}084$; $n = 60$.
Avec ces valeurs numériques, l'intervalle de confiance de μ avec le coefficient de confiance 95% est [3,991 ; 4,033].
c) D'après l'approche fréquentiste des probabilités, si on prélevait un très grand nombre de tels échantillons, environ 95 pour 100 d'entre eux contiendraient la moyenne inconnue μ de la population. En fait on n'en prélève qu'un seul et on ne peut pas savoir si celui-ci contient ou non le nombre μ, *mais la méthode mise en œuvre permet d'obtenir un intervalle contenant μ dans 95 cas sur 100* donc, la moyenne μ n'appartient pas nécessairement à l'intervalle de confiance.

13 – BÂTIMENT 98

1° μ estimé par $\bar{x} = 3{,}512$,

σ estimé par $\sqrt{\frac{40}{39}}\ s = 0{,}096$ à 10^{-3} près.

2° a) Un intervalle de confiance de la moyenne μ de la population est

$[\ \bar{x} - 1{,}96\ \frac{\sigma}{\sqrt{n}}\ ;\ \bar{x}\ + 1{,}96\ \frac{\sigma}{\sqrt{n}}\ ] = [3{,}482\ ;\ 5{,}542]$.

b) D'après l'approche fréquentiste des probabilités si on prélevait un très grand nombre d'échantillons, environ 95 % d'entre eux contiendraient la moyenne inconnue μ.
En fait on n'en prélève qu'un seul et on ne peut pas savoir si celui-ci contient ou non le nombre μ.

c) On ne peut également savoir si μ a plus de chances de se trouver près du centre que d'une extrémité de cet intervalle.

14 – OPTICIEN LUNETIER 2000

1. La calculatrice donne pour moyenne et écart type de l'échantillon : $\bar{x} = 110{,}1$ et $s \approx 0{,}9798$**.**
A 10^{-2} près on a $\bar{x} = 110{,}1$ et $s \approx 0{,}98$.

2. Une estimation ponctuelle de la moyenne est $\mu = \bar{x} = 110{,}1$

3. Une estimation par intervalle de confiance de μ avec le coefficient de confiance de 90% correspondant à cet échantillon est approximativement :

I = $[\ \bar{x} - t_\alpha \frac{1}{\sqrt{50}}\ ;\ \bar{x} + t_\alpha \frac{1}{\sqrt{50}}\ ]$ avec $\bar{x} = 110{,}1$

$2\,\pi(t_\alpha) - 1 = 0{,}90$ équivaut à $\pi(t_\alpha) = \frac{1{,}90}{2}$,

$t_\alpha = 1{,}645$ d'où l'intervalle :

I = $[110{,}1 - 1{,}645\ \frac{1}{\sqrt{50}}\ ;\ 110{,}1 + 1{,}645\ \frac{1}{\sqrt{50}}\ ]$,

I = [109,87 ; 110,33]

15 – PRODUCTIQUE TEXTILE 99

A) La calculatrice donne, à 10^{-3} près, pour moyenne et écart type de l'échantillon $\bar{x} = 2{,}217$ et $s = 0{,}207$.

Une estimation ponctuelle de la moyenne est $\mu = \bar{x} = 2{,}22$ une estimation ponctuelle de l'écart type est $0{,}207 \times \sqrt{\frac{12}{11}}$ soit $\sigma = 0{,}22$ à 10^{-2} près.

B) 1) La moyenne de l'échantillon de 35 éprouvettes est $\bar{y} = 1{,}99$ à 10^{-2} près.

2) Une estimation ponctuelle de la moyenne μ_1 de la variable aléatoire Y est 1,99.

Si le coefficient de confiance est $2\,\pi(t) - 1 = 0{,}99$, alors $\pi(t) = 0{,}995$ et $t \approx 2{,}575$.

On donne $\sigma_1 = 0{,}17$ avec $\bar{y} = 1{,}99$ et $n = 35$, on obtient : $[1{,}99 - 2{,}575 \frac{0{,}17}{\sqrt{35}}\ ;\ 1{,}99 + 2{,}575 \frac{0{,}17}{\sqrt{35}}\ ]$.

Donc un intervalle de confiance de μ avec le coefficient de confiance 95 % est [1,91 ; 2,06] .

16 – AGRO-EQUIPEMENT 99

1) La calculatrice donne pour moyenne et écart type de l'échantillon : $\bar{x} = 1346{,}875$ et $s = 256{,}2$ à 10^{-1} près.

2) On rappelle que $\bar{X}$ suit une loi normale de moyenne μ inconnue et d'écart type $\frac{260}{\sqrt{64}}$

Si le coefficient de confiance est $2\,\pi(t) - 1 = 0{,}95$, alors $\pi(t) = 0{,}975$ et $t \approx 1{,}96$.

I = $[\ \bar{x} - t\ \frac{\sigma}{\sqrt{n}}\ ,\ \bar{x} + t \frac{\sigma}{\sqrt{n}}\ ]$ où $\bar{x}$ est la moyenne d'un échantillon aléatoire non exhaustif, n la taille de l'échantillon et σ l'écart type de la population.

I = $[1346{,}875 - 1{,}96 \frac{260}{8}\ ;\ 1346{,}875 + 1{,}96\ \frac{260}{8}\ ]$,

I = [1283,15 ; 1410,55] à 10^{-2} près.

17 – CHIMISTE 2000

1) La calculatrice donne, à 10^{-2} près, pour moyenne et écart type de l'échantillon $\bar{x} = 1625{,}47$ et $s = 4{,}66$.

2) Une estimation ponctuelle de la moyenne est $\hat{\mu} = \bar{x} = 1625{,}47$ une estimation ponctuelle de l'écart type est $4{,}66 \times \sqrt{\frac{150}{149}}$ soit à 10^{-2} près $\hat{\sigma} = 4{,}67$.

3) a) La variable aléatoire $\overline{X}$ suit la loi normale $\mathrm{N}\ (\mu, \frac{\sigma}{\sqrt{n}})$ avec μ inconnu et l'écart type estimé à 10^{-2} près par $\hat{\sigma} = 4{,}67$ d'où $\overline{X}$ suit approximativement la loi normale $\mathrm{N}\ (\mu, \frac{4{,}67}{\sqrt{150}})$ soit la loi normale $\mathrm{N}\ (\mu, 0{,}38)$.

b) Un intervalle de confiance de la moyenne μ est I = $[\ \bar{x} - t \times 0{,}38\ ,\ \bar{x} + t \times 0{,}38]$ où $\bar{x}$ est la moyenne d'un échantillon aléatoire non exhaustif.

Si le coefficient de confiance est $2\,\pi(t) - 1 = 0{,}95$, alors $\pi(t) = 0{,}975$ et $t \approx 1{,}96$.

I = [1625,47 – 1,96× 0,38 ; 1625,47 + 1,96 × 0,38],

I = [1624,72 ; 1626,22] à 10^{-2} près.

L'amplitude de l'intervalle est $2 \times 1{,}96 \times 0{,}38 \approx 1{,}5$ pour diminuer cette amplitude il faut augmenter la taille des échantillons.

c) Si l'amplitude de l'intervalle de confiance est 1 alors $2 \times 1{,}96 \times \frac{4{,}67}{\sqrt{n}} = 1$,

$\sqrt{n} \approx 2 \times 1{,}96 \times 4{,}67$,

$n = 356$.

18 – GROUPEMENT D 2000

B – 1° La calculatrice donne la moyenne de l'échantillon $\bar{x} \approx 72{,}369$ soit $\bar{x} = 72{,}37$ à 10^{-2} près et l'écart type de l'échantillon $s \approx 0{,}11121$ soit $s = 0{,}11$ à 10^{-2} près.

2° Une estimation ponctuelle de la moyenne de la population μ est 72,37.
Une estimation ponctuelle de l'écart type de la population σ est $0{,}1112\sqrt{\frac{10}{9}} \approx 0{,}1172$ soit 0,12 à 10^{-2} près.

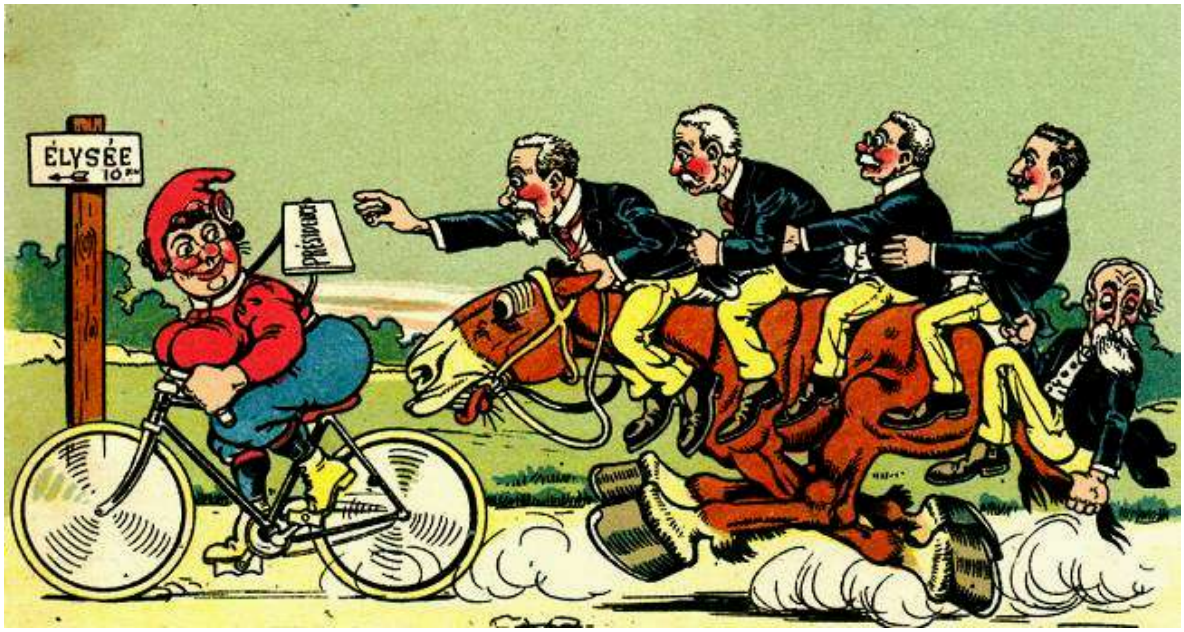
3° Un intervalle de confiance de la moyenne μ est :
$I = [\ \bar{x} - t\ \frac{0{,}12}{\sqrt{10}}\ ,\ \bar{x} + t\frac{0{,}12}{\sqrt{10}}\ ]$ où $\bar{x}$ est la moyenne d'un échantillon aléatoire non exhaustif.
Si le coefficient de confiance est $2\ \pi(t) - 1 = 0{,}95$, alors $\pi(t) = 0{,}975$ et $t \approx 1{,}96$.
$I = [72{,}37 - 1{,}96 \times \frac{0{,}12}{\sqrt{10}}\ ,\ 72{,}37 + 1{,}96 \times \frac{0{,}12}{\sqrt{10}}\ ]$,
$I = [72{,}30\ ;\ 72{,}44]$ à 10^{-2} près.

4) Si $I = [72{,}31\ ;\ 72{,}43]$ on a $t \times \frac{0{,}08}{\sqrt{10}} = 0{,}06$ d'où $t = \frac{0{,}06\sqrt{10}}{0{,}08}$, $t \approx 2{,}37$ la table du formulaire donne
$2\pi(2{,}37) - 1 = 2 \times 0{,}9911 - 1$,
$2\pi(2{,}37) - 1 = 0{,}9822$, donc à une unité près le coefficient de confiance est 0,98. Au risque de 2% un intervalle de confiance de μ est [72,31 ; 72,43].

Supplément à la séance n°2
Échantillonnage et estimation

ÉCHANTILLONNAGE

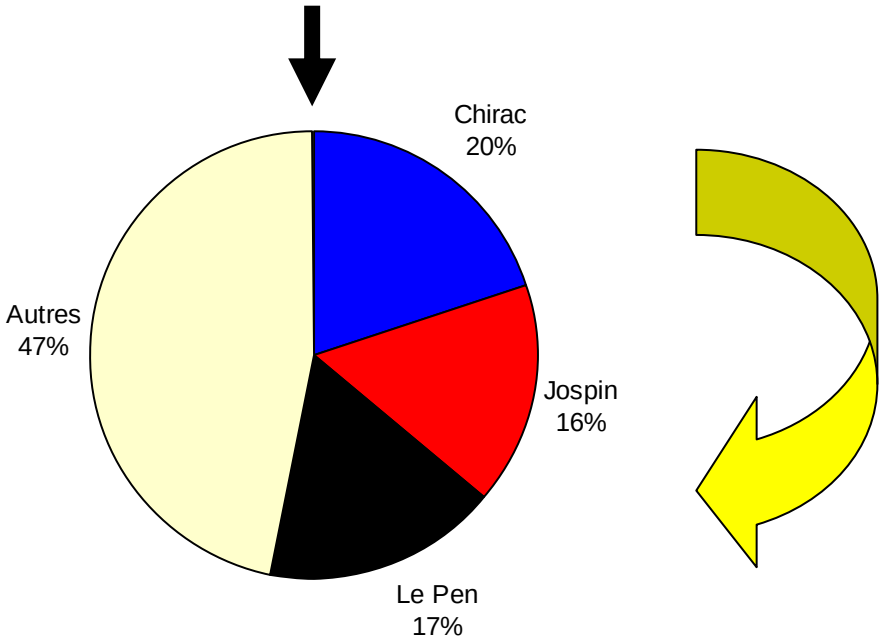
Un exemple utile au citoyen : l'élection présidentielle de 2002 (épisode 1)



L'attitude de l'opinion vis à vis des sondages est souvent sans nuance : on leur prête des pouvoirs de prédiction qu'ils n'ont pas (en omettant souvent de fournir les « fourchettes ») et (ou) on déclare qu'ils se trompent 9 fois sur 10.
On peut mettre en parallèle le dernier sondage publié par BVA et effectué sur 1000 électeurs (Jacques Chirac 19 %, Lionel Jospin 18 %, Jean-Marie Le Pen 14 %) avec le résultat du premier tour (Jacques Chirac 19,88 %, Lionel Jospin 16,18 %, Jean-Marie Le Pen 16,86 %) et poser la question « le sondage est-il faux ? ».
Si le sondage avait la prétention de prévoir exactement les résultats, ou ne serait-ce que l'ordre des candidats, il serait bien sûr faux. Mais cette prétention n'est pas scientifique, comme nous l'apprend l'observation des fluctuations d'échantillonnage.
Dans un sondage politique, tout se passe comme si on tirait au sort les 1000 électeurs. En fait, essentiellement parce qu'un véritable tirage au hasard revient trop cher (il est très coûteux de devoir interroger précisément la personne qui a été tirée au sort et pas une autre), on a recourt à d'autres méthodes, comme celle des quotas, mais avec une qualité équivalente au tirage totalement aléatoire.

« *Avec la méthode des quotas, il n'existe pas de loi mathématique permettant de déterminer la marge d'erreur d'un sondage, en pratique toutefois, on considère que la marge d'erreur des sondages par quotas est égale, voire inférieure à celle des sondages aléatoires.* »
Jean-François Doridot, directeur du département opinion d'Ipsos *Le Monde* 17/03/02.

On peut donc dire que dans la situation précédente, simuler un sondage « bien fait » consiste à faire tourner 1000 fois une roue de loterie partagée en quatre secteurs de 20 % (Jacques Chirac), 16 % (Lionel Jospin), 17 % (Jean-Marie Le Pen) et 47 % (autres candidats) correspondant à l'état de l'opinion le jour de l'élection.



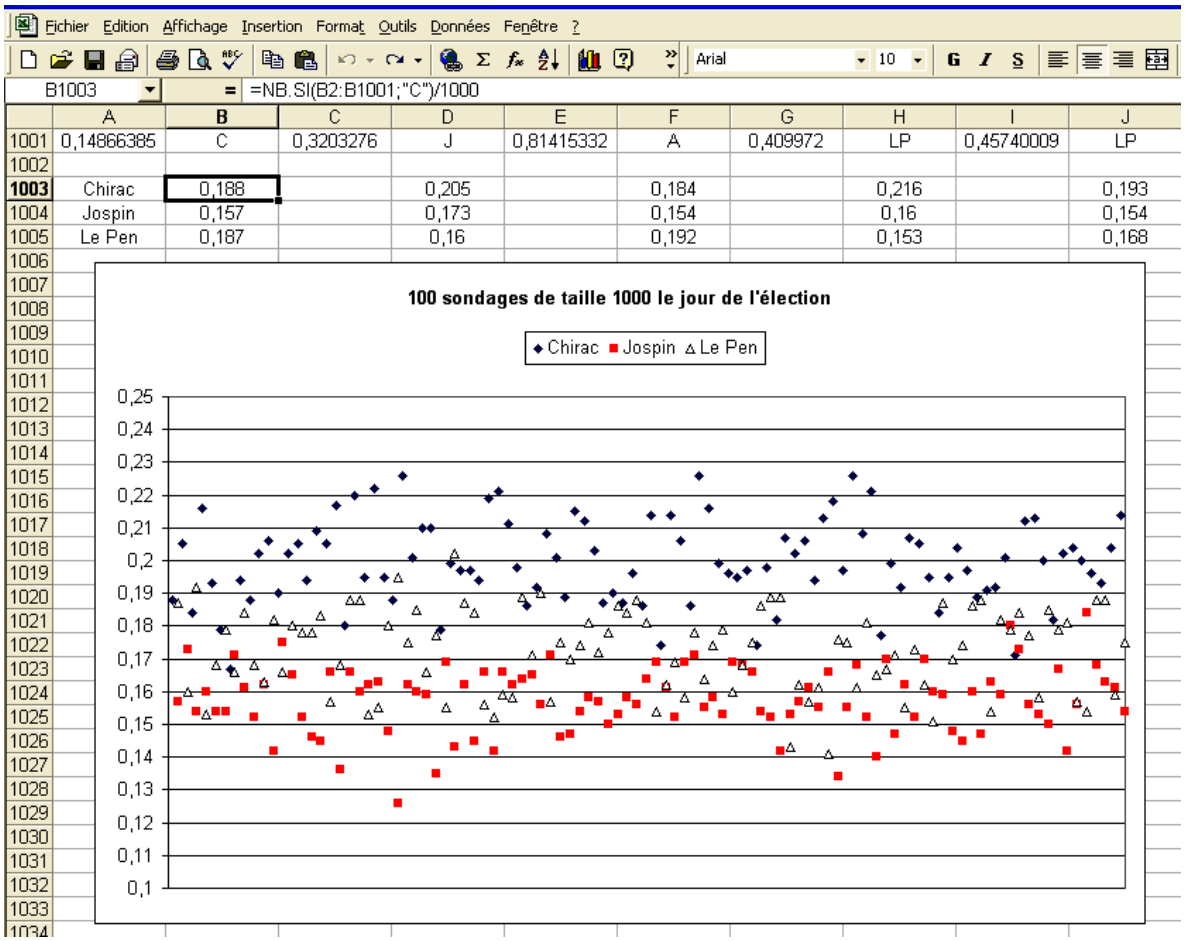
La simulation d'une telle roue de loterie est facile à mettre en place avec un tableur.
Le générateur de nombres aléatoires fournit un nombre « au hasard » dans l'intervalle [0, 1[. Il suffit alors de partager cet intervalle proportionnellement aux pourcentages des secteurs de la roue de loterie, ce qui peut être demandé aux élèves.
Ainsi, l'instruction =ALEA() entrée en cellule A1
et l'instruction =SI(A1<=0,2;"Chirac";SI(A1<=0,36;"Jospin";SI(A1<=0,53;"LePen";"Autre"))) entrée en cellule B1 simule le sondage d'un électeur.
Il suffit de recopier ces instructions 1000 fois vers le bas pour simuler un sondage de 1000 personnes puis 100 fois vers la droite pour simuler 100 sondages.

Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?

Arial

B2 = =SI(A2<=0,2;"C";SI(A2<=0,36;"J";SI(A2<=0,53;"LP";"A")))

	A	B	C	D	E	F	G
1	alea	simul 1					
2	0,75155518	A					
3	0,30898485	J					
4	0,22239056	J					
5	0,28795551	J					
6	0,45922784	LP					
7	0,00980021	C					
8	0,44341701	LP					
9	0,74180146	A					
10	0,05361272	C					
11	0,78839653	A					
12	0,07754578	C					
13	0,76070321	A					
14	0,12234491	C					



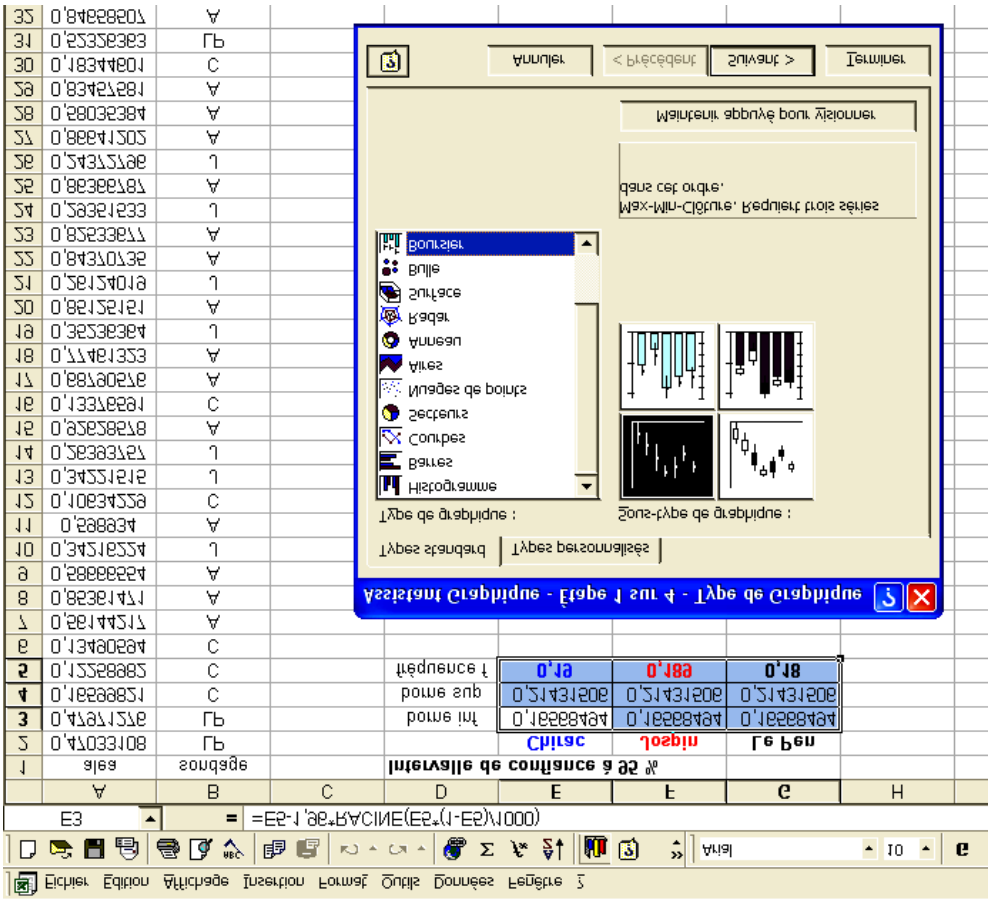
D'une certaine façon tous ces sondages sont « corrects » et l'observation du nuage de points correspondant aux fréquences des trois candidats sur 100 sondages suffit à prendre conscience des fluctuations dues au hasard.

ESTIMATION

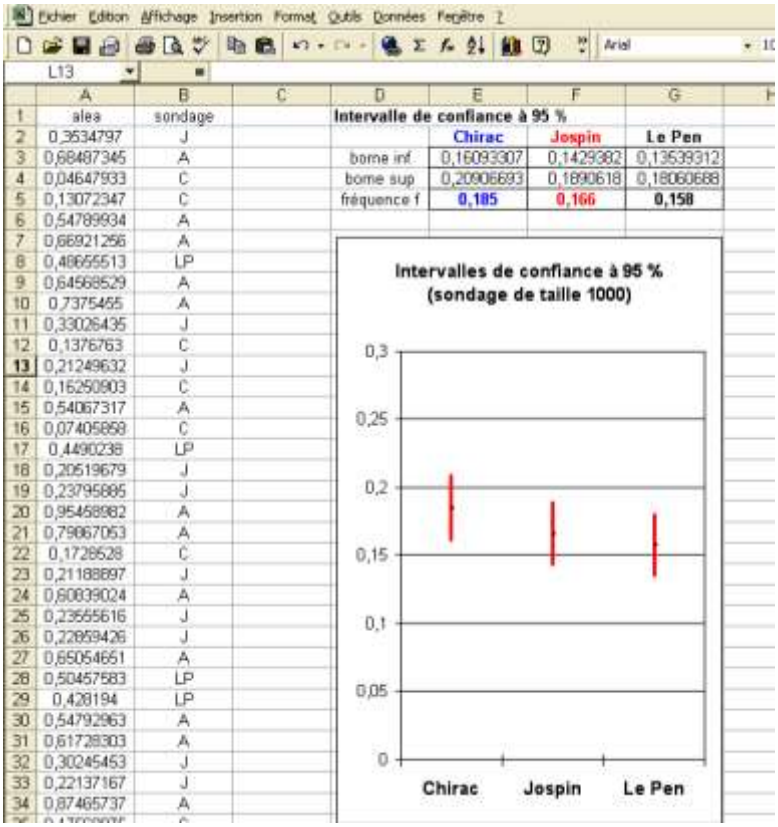
1 – Un exemple utile au citoyen : l'élection présidentielle de 2002 (épisode 2)

On reprend l'exemple du premier tour de 2002.
Sur la simulation d'un sondage de taille 1000 le jour de l'élection (« sorti des urnes »), on peut calculer les intervalles de confiance à 95 % pour chacun des trois candidats :

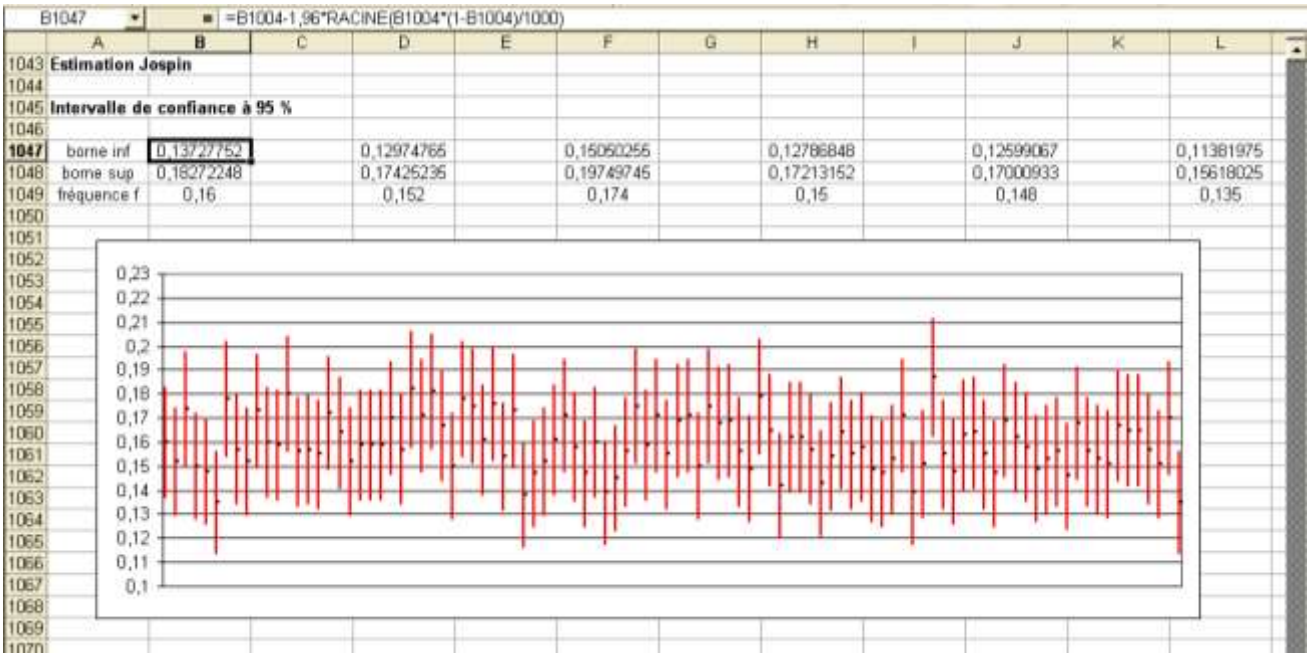
On peut représenter ces trois intervalles de confiance en utilisant le « diagramme boursier » du tableur. Il faut pour cela avoir, dans cet ordre, la borne inférieure, la borne supérieure et le centre de l'intervalle de confiance.
Sur l'image d'écran suivante, on a d'abord entré en E5 la formule
=NB.SI(B2:B1001;"C")/1000 puis en E3 la formule
=E5-1,96*RACINE(E5*(1-E5)/1000) et en E4 la formule
=E5+1,96*RACINE(E5*(1-E5)/1000) .



En faisant F9, on obtient un autre sondage.



On peut également, pour un seul candidat, représenter simultanément les intervalles de confiances calculés sur 100 sondages, ci-dessous dans le cas de Lionel Jospin.

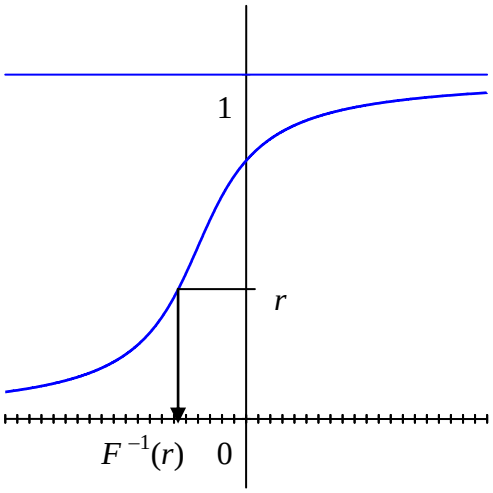


On constate alors qu'environ 95 % de ces intervalles de confiance recouvrent la valeur obtenue par le candidat le jour de l'élection : pour Lionel Jospin, environ 95 % des intervalles de confiance coupent la droite d'équation $y = 0{,}16$.

Simulation d'une distribution continue par inversion de sa fonction de répartition

Cette méthode est générale et elle est praticable lorsque l'expression analytique de la fonction de répartition est simple (cas de la loi exponentielle) ou que l'on dispose d'un moyen de calcul approché (cas de la loi normale).
La méthode repose sur le résultat suivant :

> **Si X est une variable aléatoire réelle, de fonction de répartition F continue, strictement croissante, alors la variable aléatoire $Y = F(X)$ est uniformément distribuée sur [0, 1].**



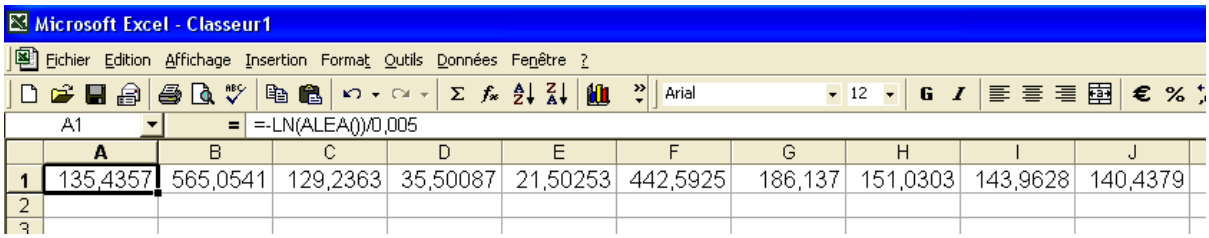
Ainsi, si l'on tire n nombres au hasard r_i, parmi des nombres uniformément répartis sur [0, 1], un échantillon de la distribution de X sera donné par $F^{-1}(r_i)$.

Distribution exponentielle

Une variable aléatoire T de loi exponentielle de paramètre λ a une fonction de répartition F définie sur $[0 , +\infty[$ par : $F(t) = P(T \leq t) = \int_0^t f(t)\, dt = \int_0^t \lambda\, e^{-\lambda t}\, dt = 1 - e^{-\lambda t}$.

Pour simuler les réalisations de T, il suffit, d'après la méthode précédente, de calculer des valeurs $-\frac{1}{\lambda} \ln(1 - r_i)$ où les valeurs r_i sont données par le générateur de nombres aléatoires.
Et comme les nombres $1 - r_i$ sont aussi uniformément distribués sur [0, 1], on peut dire que les réalisations d'une variable aléatoire suivant la loi exponentielle de paramètre λ sont simulées par l'instruction : `= – LN(ALEA()) / λ` .

Exemple : simulation de la loi $\mathcal{E}$ ($\lambda = 0{,}005$) dont l'espérance est $\frac{1}{0{,}005} = 200$.

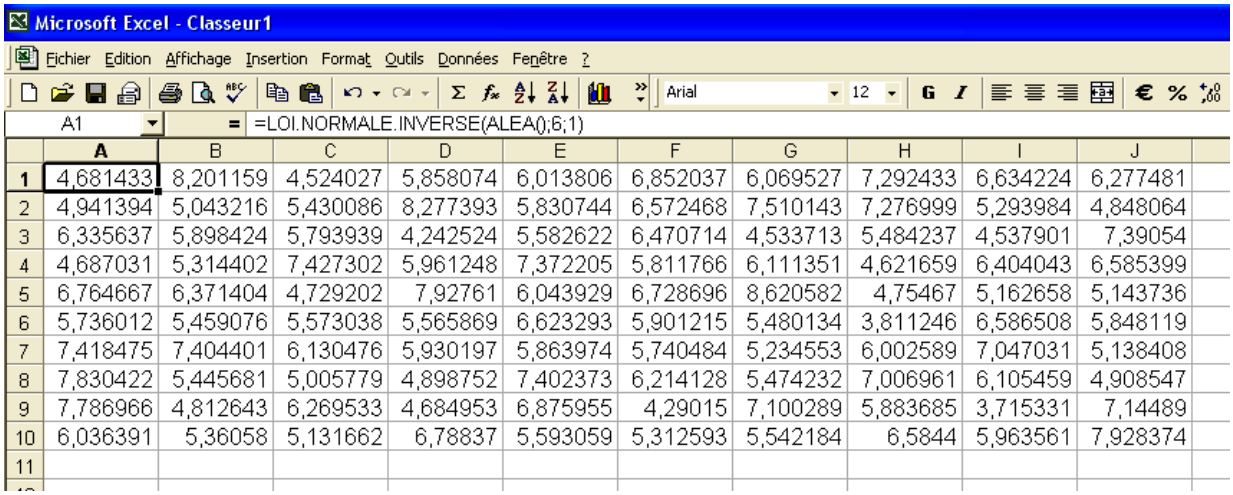


	A	B	C	D	E	F	G	H	I	J
1	135,4357	565,0541	129,2363	35,50087	21,50253	442,5925	186,137	151,0303	143,9628	140,4379
2										

Distribution normale

On peut accéder sur le tableur à la fonction de répartition inverse F^{-1} d'une variable aléatoire de loi normale de moyenne μ et d'écart type σ par la fonction :
=LOI.NORMALE.INVERSE(t ; μ ; σ)

Exemple : simulation d'un échantillon de taille 100 extrait d'une population de loi normale de moyenne $\mu = 6$ et d'écart type $\sigma = 1$:



	A	B	C	D	E	F	G	H	I	J
1	4,681433	8,201159	4,524027	5,858074	6,013806	6,852037	6,069527	7,292433	6,634224	6,277481
2	4,941394	5,043216	5,430086	8,277393	5,830744	6,572468	7,510143	7,276999	5,293984	4,848064
3	6,335637	5,898424	5,793939	4,242524	5,582622	6,470714	4,533713	5,484237	4,537901	7,39054
4	4,687031	5,314402	7,427302	5,961248	7,372205	5,811766	6,111351	4,621659	6,404043	6,585399
5	6,764667	6,371404	4,729202	7,92761	6,043929	6,728696	8,620582	4,75467	5,162658	5,143736
6	5,736012	5,459076	5,573038	5,565869	6,623293	5,901215	5,480134	3,811246	6,586508	5,848119
7	7,418475	7,404401	6,130476	5,930197	5,863974	5,740484	5,234553	6,002589	7,047031	5,138408
8	7,830422	5,445681	5,005779	4,898752	7,402373	6,214128	5,474232	7,006961	6,105459	4,908547
9	7,786966	4,812643	6,269533	4,684953	6,875955	4,29015	7,100289	5,883685	3,715331	7,14489
10	6,036391	5,36058	5,131662	6,78837	5,593059	5,312593	5,542184	6,5844	5,963561	7,928374
11										

Des sujets récents de BTS

Groupement B 2006

Une entreprise fabrique des chaudières de deux types :

– des chaudières dites « à cheminée »,
– des chaudières dites « à ventouse ».

On considère un échantillon de 100 chaudières prélevées au hasard dans un stock important. Ce stock est assez important pour qu'on puisse assimiler ce tirage à un tirage avec remise.
On constate que 94 chaudières sont sans aucun défaut.

1° Donner une estimation ponctuelle de la fréquence inconnue p des chaudières de ce stock qui sont sans aucun défaut.

2° Soit F la variable aléatoire qui, à tout échantillon de 100 chaudières prélevées au hasard et avec remise dans ce stock, associe la fréquence des chaudières de cet échantillon qui sont sans aucun défaut.

On suppose que F suit la loi normale de moyenne p et d'écart type $\sqrt{\dfrac{p(1-p)}{100}}$, où p est la fréquence inconnue des chaudières du stock qui sont sans aucun défaut.
Déterminer un intervalle de confiance de la fréquence p avec le coefficient de confiance 95 %. Arrondir les bornes à 10^{-2}.

3° On considère l'affirmation suivante : « la fréquence p est obligatoirement dans l'intervalle de confiance obtenu à la question 2° ».
Est-elle vraie ? (On ne demande pas de justification.)

Éléments de réponse

1° $f = 0{,}94$, donc une estimation ponctuelle de p est $p = 0{,}94$. 0,5 point

2° Un intervalle de confiance est :

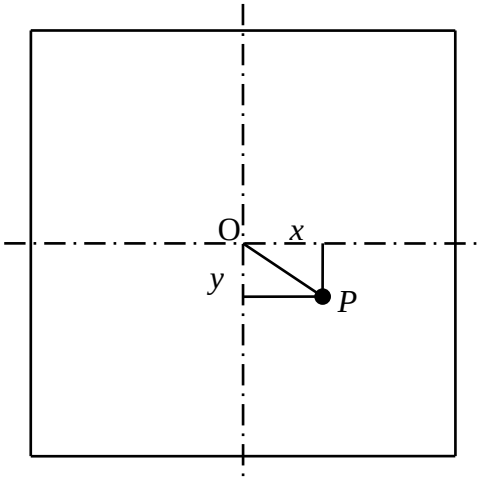
$$I = \left[f - t\sqrt{\frac{f(1-f)}{n-1}},\ f + t\sqrt{\frac{f(1-f)}{n-1}} \right]$$

avec $f = 0{,}94$; $n = 100$ et $t = 1{,}96$.
$\boldsymbol{I \approx [0{,}89\ ;\ 0{,}99]}$. 1,5 point

3° Non, la fréquence p n'appartient pas obligatoirement à I. 0,5 point

Groupement B 2004 – Nouvelle Calédonie

Une machine doit percer des pièces métalliques carrées en leur centre, noté O sur la figure. En réalité le centre du trou circulaire ainsi réalisé est un point* P *proche du point O et variable d'une pièce à l'autre.
On note (x, y) les coordonnées du point* P *dans le repère orthonormal de la figure, où l'unité est le millimètre.
On note z la distance OP.



On souhaite avoir une estimation de la moyenne μ_2 des distances O*P* sur la production d'une journée.
Sur un échantillon de 64 pièces prélevées au hasard et avec remise dans la production de cette journée, on constate que la moyenne $\bar{z}$ et l'écart type s_2 des distances O*P* en mm sont, arrondis à 10^{-2}, $\bar{z} = 2{,}01$ et $s_2 = 1{,}04$.

1° A partir des informations portant sur cet échantillon, donner une estimation ponctuelle de la moyenne μ_2 et de l'écart type σ_2 des longueurs O*P* pour la production de la journée ; arrondir à 10^{-2}.

2° Soit $\overline{Z}$ la variable aléatoire qui, à tout échantillon de 64 pièces prélevées au hasard et avec remise dans la production de la journée, associe la moyenne des longueurs O*P* de cet échantillon.

On suppose que $\overline{Z}$ suit la loi normale de moyenne inconnue μ_2 et d'écart type $\frac{1{,}05}{\sqrt{64}}$.

Déterminer un intervalle de confiance centré en $\bar{z}$ de la moyenne μ_2 des longueurs O*P* pour la production, avec le coefficient de confiance 95%. (On arrondira les bornes de l'intervalle à 10^{-2}).

3° On considère l'affirmation suivante : "la moyenne μ_2 est obligatoirement dans l'intervalle de confiance obtenu à la question précédente".
Cette affirmation est-elle vraie ? (Donner la réponse sans explication).

Éléments de réponse

1° Estimation ponctuelle de μ_2 : **2,01** .
Estimation ponctuelle de σ_2 : $\sqrt{\frac{64}{63}} \times 1{,}04 \approx \mathbf{1{,}05}$. — 1 point

2° $[\,2{,}01 - 1{,}96 \times \frac{1{,}05}{\sqrt{64}}\ ;\ 2{,}01 + 1{,}96 \times \frac{1{,}05}{\sqrt{64}}\,] \approx \mathbf{[1{,}75\ ;\ 2{,}27]}$. — 1,5 point

3° Non. — 0,5 point

Groupement B 2004

Une entreprise fabrique, en grande quantité, des tiges métalliques cylindriques pour l'industrie. Leur longueur et leur diamètre sont exprimés en millimètres.

Dans cet exercice, les résultats approchés sont à arrondir à 10^{-2}.

Dans cette question on s'intéresse au diamètre des tiges.

Soit $\overline{D}$ la variable aléatoire qui, à tout échantillon de 50 tiges prélevées au hasard et avec remise dans la production d'une journée, associe la moyenne des diamètres des tiges de cet échantillon.

On suppose que $\overline{D}$ suit la loi normale de moyenne inconnue μ et d'écart type $\frac{\sigma}{\sqrt{50}}$ avec $\sigma = 0{,}19$.

On mesure le diamètre, exprimé en millimètres, de chacune des 50 tiges d'un échantillon prélevé au hasard et avec remise dans la production de la journée considérée.

On constate que la valeur approchée arrondie à 10^{-2} de la moyenne $\bar{x}$ des diamètres des tiges de cet échantillon est $\bar{x} = 9{,}99$.

1° A partir des informations portant sur cet échantillon, donner une estimation ponctuelle de la moyenne μ des diamètres des tiges produites dans cette journée.

2° Déterminer un intervalle de confiance centré sur $\bar{x}$ de la moyenne μ des diamètres des tiges produites pendant la journée considérée, avec le coefficient de confiance 95 %.

3° On considère l'affirmation suivante : " la moyenne µ est obligatoirement dans l'intervalle de confiance obtenu à la question 2° ".
Est-elle vraie ? (On ne demande pas de justification).

Éléments de réponse

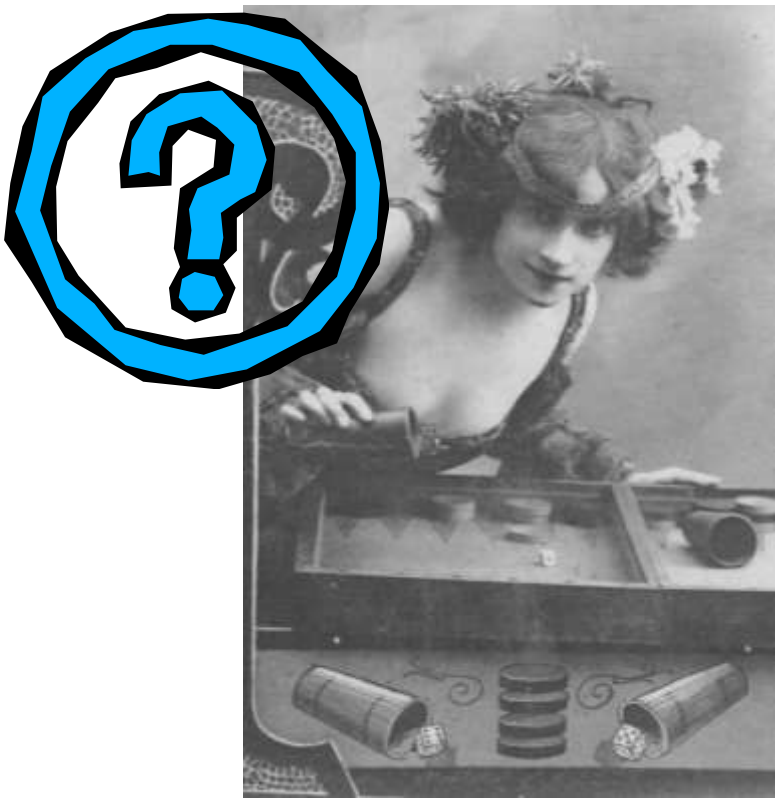
1° $\mu = \mathbf{9{,}99}$. — 0,5 point

2° $[\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}} \ ; \ \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}]$

$= [9{,}99 - 1{,}96\times\frac{0{,}19}{\sqrt{50}} \ ; \ 9{,}99 - 1{,}96\times\frac{0{,}19}{\sqrt{50}}] = [\mathbf{9{,}94} \ ; \ \mathbf{10{,}04}]$. — 1 point

3° Non. — 0,5 point

Séance 3 :
TESTS D'HYPOTHESES



"*La seule certitude que j'ai, c'est d'être dans le doute.*"
Pierre DESPROGES - 1986.

Il y a deux débouchés "naturels" à l'étude statistique des fluctuations d'échantillonnage.
Le premier est celui de l'estimation. Dans ce cas, on n'a aucun a priori sur le (ou les) paramètre(s) à estimer sur la population. On construira alors un intervalle de confiance généralement centré sur la fréquence f (ou la moyenne $\bar{x}$) calculée sur l'échantillon.
Dans bien des situations, on a une idée a priori de la valeur p_0 (ou μ_0) que devrait avoir la fréquence p (ou la moyenne μ) 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).

Il reviendra à ***Jerzy NEYMAN*** (1894-1981) et ***Egon PEARSON*** (1895-1980, fils de Karl) de proposer vers 1930 ***une démarche de décision universellement admise***.
Ils ont remarqué que, dans la plupart des situations, les deux hypothèses, H_0 et H_1 en concurrence, 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.

J. Neyman

E. Pearson

I – UN EXEMPLE D'INTRODUCTION

"La statistique est comparable à un fusil chargé qui, en des mains inexpérimentées, peut amener à de graves accidents."

Cheysson, cité par March dans *"Les principes de la méthode statistique"* – 1930.

Les notions d'erreur de seconde espèce et de puissance d'un test ne sont pas au programme de BTS. Les choix de construction du test ne sont pas demandés à l'examen. Il est hors de question de faire un cours, ou d'exiger des connaissances sur ces sujets. Cependant ces notions sont essentielles à la compréhension d'un test (en particulier l'erreur de 2nde espèce), on les abordera donc ici dans un cadre concret et particulièrement signifiant.
Le programme indique d'ailleurs :

"On soulignera que la décision prise, rejet ou acceptation, dépend des choix faits a priori par l'utilisateur : choix de l'hypothèse nulle, choix du seuil de signification."

Prenons un exemple : un laboratoire affirme, qu'à la différence de ses concurrents, dont le produit est efficace à 80%, le sien l'est à 90%. Deux tests sont envisageables selon le point de vue.

- Test fabriquant :

H_0 : $p = 0{,}9$ test unilatéral à gauche.
Ce test ne rejettera l'hypothèse selon laquelle le nouveau produit est meilleur ($p = 0{,}9$) que si le résultat de l'échantillon va significativement dans l'autre sens.

- Test client :

H_0 : $p = 0{,}8$ test unilatéral à droite.
Le client privilégie l'hypothèse $p = 0{,}8$. Il n'est prêt à changer d'avis que si les résultats de l'échantillon sont significativement meilleurs.
C'est certainement au niveau de l'utilisation des tests statistiques qu'une incompréhension des principes conduira le plus à des réponses inappropriées. On voit là le rôle formateur que nous avons à jouer par rapport à une utilisation "presse bouton" des logiciels.

On se placera, pour l'exemple qui suit, 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.

Il s'agit de faire comprendre les éléments essentiels suivants :

- La ***construction du test*** doit se faire ***avant*** la prise d'échantillon. Elle doit faire l'objet d'un ***protocole*** sur lequel se mettent d'accord les deux partis en présence, ici professeurs et élèves, dans les relations commerciales, vendeur et acheteur. D'où la nécessité d'une ***normalisation*** des tests (voir les normes de l'AFNOR pour l'industrie). C'est un non sens statistique de construire le test après le prélèvement de l'échantillon. On pourrait conclure ce que l'on veut !
- Les ***erreurs*** sont inévitables (à la différence des autres domaines des mathématiques, il s'agit ici d'évaluer et d'accepter les risques). Ces erreurs sont de deux types et les ***choix*** effectués pour la construction du test correspondent à un ***compromis*** entre la maîtrise des risques α et β et la taille *n* de l'échantillon (coût du contrôle).
  Bien qu'hors programme des BTS, ce sont des enjeux importants du test, et on peut les faire comprendre, sans entrer dans une étude systématique.
- L'erreur de 1ère espèce α étant la plus facile à maîtriser (β ne peut être déterminée que si l'on connaît les lois de probabilité sous H_1), c'est sur elle, et l'hypothèse H_0, que sera construit le test.
  Cela conduit à ***privilégier H_0*** : si la forme de la région d'acceptation dépend de la nature de H_1 (bilatéral ou unilatéral), ses limites ne dépendent que de H_0 (à partir de laquelle, on les calcule). Les deux hypothèses ne jouent donc pas un rôle symétrique.
- Il faut mettre en évidence les différentes ***étapes d'un test***, dont le plan est :

  1. ***Construction du test*** :
     a - ***Choix des hypothèses*** H_0 et H_1 (test bilatéral ou unilatéral).
     Ce choix est dirigé par l'énoncé. Celui de α et *n* est imposé à l'examen.
     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, selon le résultat observé.

NORMES AFNOR

5.30 TEST STATISTIQUE
Procédure basée sur une fonction des observations (c'est à dire une statistique) d'un ou de plusieurs échantillons et conduisant à rejeter, avec un certain risque d'erreur, une hypothèse généralement appelée "hypothèse nulle".

5.39 TEST BILATERAL
Test pour lequel l'hypothèse nulle est rejetée si la statistique utilisée prend une valeur située hors d'un intervalle déterminé.
Souvent l'intervalle est choisi de telle manière que la probabilité de rejet, lorsque l'hypothèse nulle est vraie, soit également partagée de part et d'autre de celui-ci. On dit alors que le test est "*symétrique*".

5.40 TEST UNILATERAL
Test pour lequel l'hypothèse nulle est rejetée si la statistique utilisée prend une valeur inférieure (supérieure) à une valeur donnée.

5.36 HYPOTHESE ALTERNATIVE
Hypothèse que l'on oppose à l'hypothèse nulle.

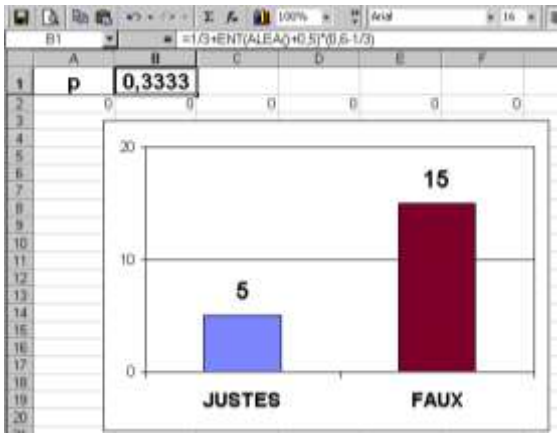
5.34 ERREUR DE PREMIERE ESPECE
Erreur commise lorsqu'on rejette l'hypothèse nulle alors que celle-ci est vraie. La probabilité d'une telle erreur s'appelle "risque de première espèce".

Remarques : des détails de vocabulaire.
• Comme il y a deux types d'erreurs, il y a deux risques possibles :
Dans le test du fabriquant (ou du vendeur) où l'hypothèse H_0 correspond à "la fabrication est conforme" :
"*On rejette à tort H_0 au risque de α %*" (***risque du vendeur***, qui se trouve alors lésé car on rejette sa production alors qu'elle est conforme, pas de chance, l'échantillonnage a fait ressortir les mauvaises pièces).
"*On accepte à tort H_0 au risque de β %*" (***risque du client***, pas de chance pour lui car l'échantillonnage n'a pas mis en évidence le fait que la livraison n'est pas conforme).
Il n'est donc pas correct de dire, comme on le lit dans certains sujets d'examen, que l'on "accepte H_0 au risque α". Dans le cas extrême d'un test à 0%, on accepte H_0 les yeux fermés, mais pas au risque de 0% ! ***Parler plutôt de seuil, c'est moins risqué*** !
• La terminologie d'hypothèse "nulle" vient peut être du fait que son acceptation implique généralement une action nulle (pas de modification) ou signifie un effet nul (dans le cas d'un médicament en médecine par exemple). En tous cas elle n'est pas "nulle" au sens que pourraient lui donner les élèves : c'est l'hypothèse généralement privilégiée.

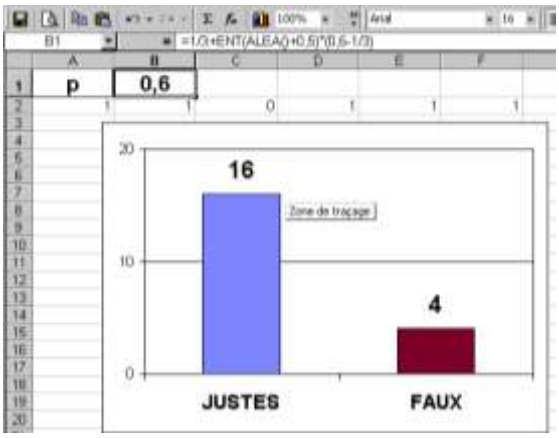
L'expérimentation, par ***simulation*** du test, permet, en ressentant le hasard, de mieux vivre les enjeux. On propose, dans le TD joint en annexe, une simulation sur calculatrices, plus pratique dans le cadre habituel de la classe, mais l'analogue sur Excel est plus visuel.
On prend aléatoirement comme valeur de p, 1/3 ou 0,60 (avec une probabilité 0,5 dans chaque cas). On simule alors un QCM de 20 questions indépendantes, avec chacune trois propositions, où la probabilité de bonne réponse est, à chaque question, p. Le test (l'examen) consiste, pour le professeur, à détecter si l'étudiant répond au hasard. Il suffit, sur Excel, de faire F9, pour avoir une autre simulation d'un QCM. On observe alors un nombre non négligeable d'erreurs de 2ème espèce (on considère qu'on a répondu au hasard alors que c'est faux), particulièrement injustes.

⇨ ***Voir en annexe le T.D. : "Introduction aux tests statistiques"*** pour mieux comprendre la situation.

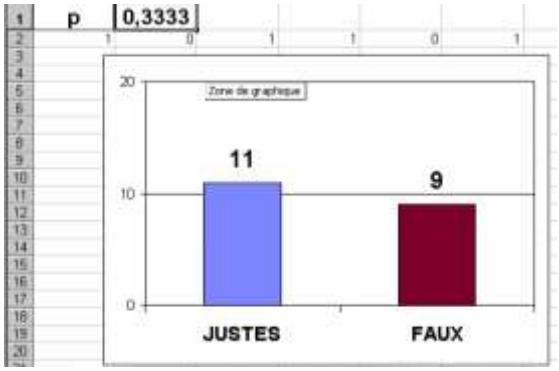
Il y a quatre cas possibles : étudiant répondant avec p = 1/3 (au hasard) et recalé, étudiant répondant avec p = 0,60 et accepté, étudiant répondant avec p = 1/3 et accepté (erreur 1), étudiant répondant avec p = 0,60 et recalé (erreur 2).



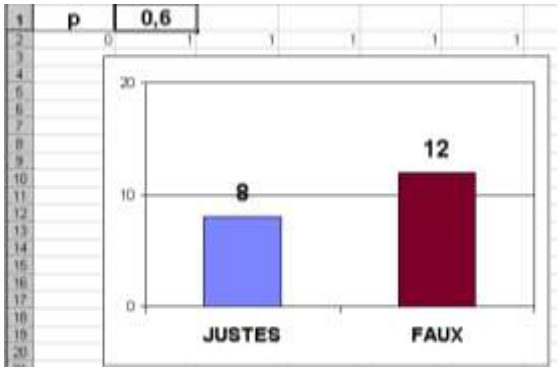
L'étudiant répond au hasard (p = 1/3)



L'étudiant ne répond pas au hasard (p = 0,60)

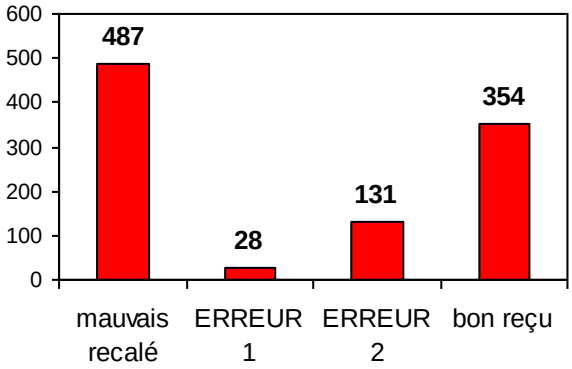


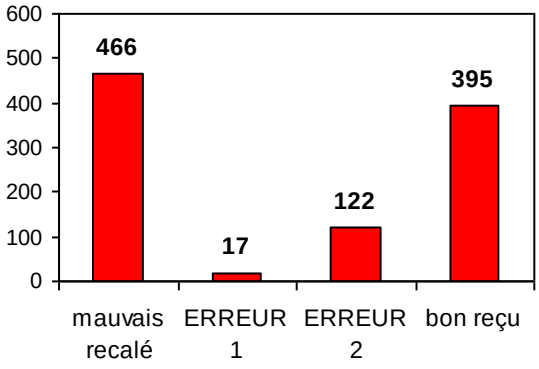
Erreur de 1ère espèce



Erreur de 2nde espèce

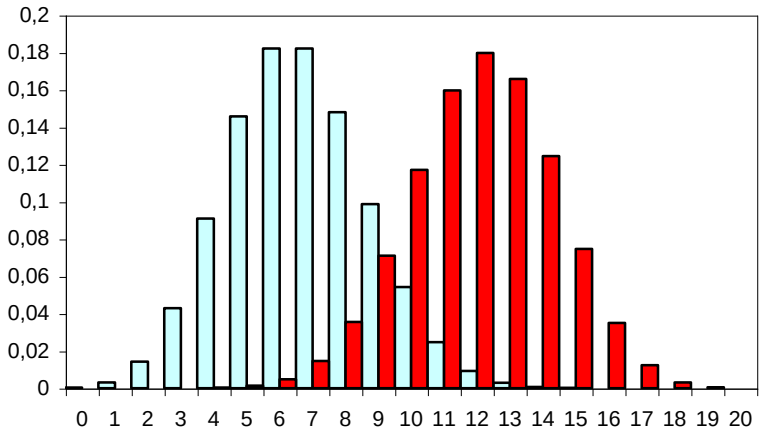
On peut faire des statistiques, en simulant 1000 QCM de 20 questions et en comptabilisant les quatre cas possibles :





Il suffit de faire F9, pour avoir aussitôt une autre simulation de 1000 QCM. On voit bien les niveaux de chaque erreur.

Les probabilités théoriques d'observation des erreurs de 1ère et 2nde espèce sont ici respectivement $\frac{1}{2}\alpha \approx 1{,}9$ % et $\frac{1}{2}\beta \approx 12{,}2$ % (en effet les probabilités d'avoir $p = \frac{1}{3}$ et $p = 0{,}6$ sont 0,5 chacune).



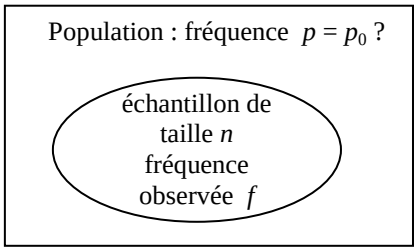
Soit x le nombre de bonnes réponses au QCM. Lorsqu'on est recalé pour $x \leq 10$, l'erreur de 1ère espèce (seuil du test) correspond aux rectangles clairs 11, 12, 13 etc. et l'erreur de 2nde espèce (pour $p = 0{,}60$) aux rectangles foncés 10, 9, 8, 7, 6, 5 etc.

On peut ensuite modifier la zone d'acceptation (la barre de l'examen) pour en évaluer aussitôt l'impact : le passage de $x \leq 10$ à $x \leq 8$ pour être recalé à l'examen, diminue visiblement l'erreur de 2nde espèce, en augmentant l'erreur de 1ère espèce (seuil du test).

II - TESTS D'UN PARAMETRE (OU PARAMETRIQUES)

1 - TEST D'UNE FREQUENCE

a) Position du problème



On s'interroge sur la fréquence p d'un phénomène dans une population. Mais, au lieu de chercher à estimer p, en ignorant tout (cas de "l'estimation"), on souhaite le comparer à une valeur fixée p_0, attendue.
Après construction d'un test, il s'agit de voir si la valeur f observée sur un ***échantillon*** est ***vraisemblable, avec l'hypothèse nulle*** "$p = p_0$".
A la ***différence de l'estimation***, où l'on construit un intervalle de confiance à partir de f (centré sur f et qui fluctue selon l'échantillon), on construit ici une zone d'acceptation de l'hypothèse nulle à partir de p_0, fixée avant le prélèvement de l'échantillon (centrée sur p_0 dans le cas bilatéral).
Les tests intervenant dans les relations commerciales (vendeur/acheteur), contrôles de qualité, des ***normes*** (voir ci-contre) définissent (pour chaque type de produit) les conditions du test et les règles de décision.

b) Plan du test d'une fréquence

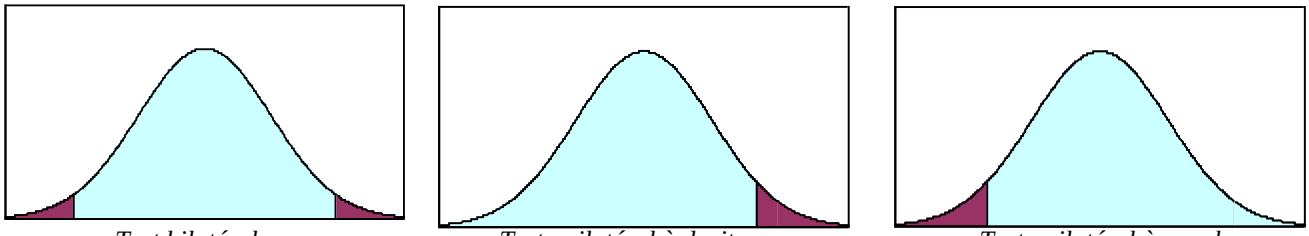
1. ***Construction du test*** :
a - ***Choix des hypothèses*** :
H_0 : " $p = p_0$".
H_1 : " $p \neq p_0$" (test bilatéral) ou
H_1 : " $p > p_0$" (unilatéral à droite) ou H_1 : " $p < p_0$" (unilatéral à gauche).
b - ***Détermination de la région critique au seuil α*** :
Sous l'hypothèse H_0, la variable aléatoire F, qui à chaque échantillon de taille n associe la fréquence observée, suit approximativement, pour n assez grand, la loi normale $N\left(p_0 , \sqrt{\frac{p_0(1-p_0)}{n}}\right)$. On calcule alors le réel positif h tel que :
$P(p_0 - h \leq F \leq p_0 + h) = 1 - \alpha$ (cas bilatéral) ou
$P(F \leq p_0 + h) = 1 - \alpha$ (cas unilatéral à droite) ou
$P(p_0 - h \leq F) = 1 - \alpha$ (cas unilatéral à gauche).
c - ***Enoncé de la règle de décision***.
On prélève un échantillon de taille n, pour lequel on calcule la fréquence f du phénomène.
La règle de décision est de trois types.

NORMES AFNOR

5.35 SEUIL DE SIGNIFICATION
Valeur du risque de première espèce lorsque l'hypothèse nulle est une hypothèse simple.
Notation : α .

5.37 ERREUR DE SECONDE ESPECE
Erreur commise lorsqu'on ne rejette pas l'hypothèse nulle alors que celle-ci est fausse. La probabilité d'une telle erreur s'appelle "risque de seconde espèce". Dans un test paramétrique, ce risque dépend de la valeur vrai du paramètre (ou des paramètres).
Notation : β .

5.38 PUISSANCE D'UN TEST
Complément à l'unité du risque de seconde espèce.



Test bilatéral *Test unilatéral à droite* *Test unilatéral à gauche*

Les courbes sont centrées sur p_0. En foncé, les régions critiques (rejet de H_0).

T Si $f \notin [\, p_0 - h\, ,\, p_0 + h]$, on rejette H_0 au risque α (cas bilatéral) ;
ou $f \notin]-\infty\, ,\, p_0 + h]$ (unilatéral à droite)
ou $f \notin [\, p_0 - h\, ,\, +\infty[$ (unilatéral à gauche).
T Sinon, on accepte H_0 au seuil α.

2. ***Utilisation du test*** :
Prélèvement d'un échantillon et prise de décision, selon le résultat observé.

⇨ ***Voir les annales de BTS*** .

Remarques : Dans le ***cas unilatéral***, les élèves ont parfois des ***difficultés*** à comprendre
- que l'hypothèse alternative n'est pas le contraire de l'hypothèse nulle ;
- que, pour tester $p > p_0$, ce qui nous intéresse, on parte de hypothèse nulle $p = p_0$ (changement de point de vue par rapport au test bilatéral où l'hypothèse nulle est bien ce que l'on cherche à tester) ;
- que, pour tester $p > p_0$, on recherche $P(F \leq p_0 + h)$.

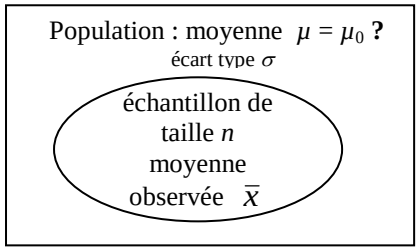
c) Utilisation des fonctions de la calculatrice

CASIO Graph 80	TI 83	SHARP EL 9600
STAT TEST Z 1-P prop≠p_0 <p_0 >p_0 choix bilatéral ou non p_0 : entrer p x : n: entrer nombre succès et taille échant.	STAT TESTS 5:1-PropZTest p_0 : entrer p_0 x : n: entrer nb succès et taille échant. prop≠p_0 <p_0 >p_0 choix bilatéral ou non	STAT E TEST 17 InputStats ENTER 10 Ztest1prop

Les calculatrices affichent le nombre $z = \dfrac{f - p_0}{\sqrt{\dfrac{f(1-f)}{n}}}$, ainsi que la probabilité $P\,(T \leq -|z|) + P\,(T \geq |z|)$ dans le cas bilatéral.
C'est cette zone qui est ombrée sur la courbe de densité de la loi N (0;1). Si sa surface est supérieure au seuil de rejet, l'hypothèse H_0 est acceptée.
Ainsi si la probabilité calculée est 0.0662 par exemple, l'hypothèse H_0 est acceptée au seuil de rejet de 5% mais rejetée au seuil de 10%. La probabilité affichée est donc le *seuil limite d'acceptation de l'hypothèse nulle*.

2 - TEST D'UNE MOYENNE

a) Position du problème



On s'interroge sur la moyenne μ d'une population. Mais, au lieu de chercher à estimer μ, en ignorant tout (cas de "l'estimation"), on souhaite le comparer à une valeur fixée μ_0, attendue.
Après construction d'un test, il s'agit de voir si la moyenne $\bar{x}$ observée sur un ***échantillon*** est ***vraisemblable, avec l'hypothèse nulle*** "$\mu = \mu_0$".
A la ***différence de l'estimation***, où l'on construit un intervalle de confiance à partir de $\bar{x}$ (centré sur $\bar{x}$), on construit ici une zone d'acceptation de l'hypothèse nulle à partir de μ_0 (centrée sur μ_0 dans le cas bilatéral).

b) Plan du test d'une moyenne lorsque σ est connu ou n grand

1. ***Construction du test*** :
a - ***Choix des hypothèses*** :
H_0 : " $\mu = \mu_0$".
H_1 : " $\mu \neq \mu_0$" (test bilatéral) ou
H_1 : " $\mu > \mu_0$" (unilatéral à droite) ou H_1 : " $\mu < \mu_0$" (unilatéral à gauche).
b - ***Détermination de la région critique au seuil α*** :
Sous l'hypothèse H_0, la variable aléatoire $\overline{X}$, qui à chaque échantillon de taille n associe la moyenne observée, suit approximativement, pour n assez grand, la loi $N\left(\mu_0 , \frac{\sigma}{\sqrt{n}}\right)$ (lorsque σ est inconnu, on le remplace, pour $n \geq 30$, par son estimation s, calculée à partir de l'écart type s_n de l'échantillon $s = \sqrt{\frac{n}{n-1}}\, s_n$).

On calcule alors le réel positif h tel que :
$P(\mu_0 - h \leq \overline{X} \leq \mu_0 + h) = 1 - \alpha$ (cas bilatéral) ou
$P(\overline{X} \leq \mu_0 + h) = 1 - \alpha$ (cas unilatéral à droite) ou $P(\mu_0 - h \leq \overline{X}) = 1 - \alpha$ (cas unilatéral à gauche).
c - ***Enoncé de la règle de décision***.
On prélève un échantillon de taille n, dont calcule la moyenne $\bar{x}$.
T Si $\bar{x} \notin [\mu_0 - h , \mu_0 + h]$, on rejette H_0 au risque α (cas bilatéral) ;
ou $\bar{x} \notin]-\infty , \mu_0 + h]$ (unilatéral à droite)
ou $\bar{x} \notin [\mu_0 - h , +\infty[$ (unilatéral à gauche).
T Sinon, on accepte H_0 au seuil α.

2. ***Utilisation du test*** :
Prélèvement d'un échantillon et prise de décision, selon le résultat observé.

⇨ ***Voir annales de BTS .***

Remarques : Le fait que σ soit ***inconnu***, conduit parfois à des ***énoncés ambigus***, où l'on prélève l'échantillon avant la construction du test, de façon à estimer l'écart type.

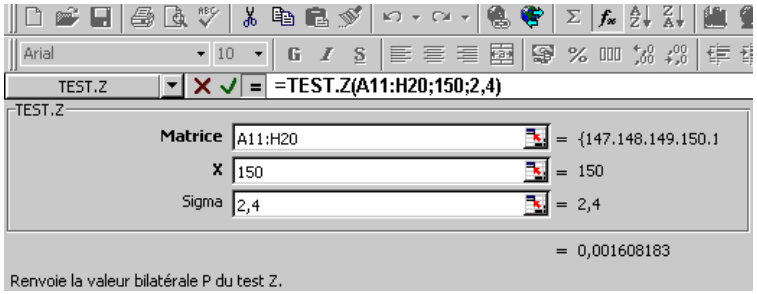
c) Utilisation des fonctions de la calculatrice ou de l'ordinateur

Sur calculatrices :

CASIO Graph 80	TI 83	SHARP EL 9600
STAT TEST Z 1-S Data: Var μ: ≠μ₀ <μ₀ >μ₀ choix bilatéral ou unilat. μ₀: entrer μ σ : entrer σ pop. ou estimé $\bar{x}$: entrer $\bar{x}$ échantillon n : entrer taille échantillon CALC ou DRAW Affichage de z et de la probabilité associée	STAT TESTS 1:Z-Test Inpt: Stats μ_0 : entrer μ_0 $\bar{x}$: entrer $\bar{x}$ échantillon σ : entrer σ pop. ou son estimation *s* n : entrer taille échantillon μ: =μ₀ <μ₀ >μ₀ choix bilatéral ou unilat. Calculate ou Draw Affichage de z et de la probabilité associée	STAT E TEST 17 InputStats ENTER 08 Ztest1samp

La calculatrice affiche le nombre $z = \dfrac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$, réalisation sur l'échantillon de la variable aléatoire $T = \dfrac{\bar{X} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$, ainsi que *le seuil limite d'acceptation de l'hypothèse nulle* (la probabilité $P(T \leq -|z|) + P(T \geq |z|)$ dans le cas bilatéral).

• Sur Excel :



La fonction "TEST.Z" permet un test bilatéral d'une moyenne, à partir des données d'un échantillon.

Dans la boîte de dialogue, renseigner chaque rubrique comme suit.

Matrice : cellules contenant les mesures de l'échantillon.
x : valeur de la moyenne μ.
Sigma : valeur σ de l'écart type de la population.
La valeur affichée est le *seuil limite d'acceptation de l'hypothèse nulle*.

d) Cas ou σ est inconnu et n petit : test de Student

Lorsque *n* est assez petit, et que σ est inconnu, le test adapté est basé sur la loi de *Student* ("T-Test" des calculatrices).

"Exemples d'utilisation du test de Student (cas des petits échantillons)"
"Ce TP n'est à réaliser qu'en liaison avec les enseignants des disciplines professionnelles et seulement si, dans celles-ci, ces procédures sont utilisées."
"Aucune connaissance à son sujet n'est exigible dans le cadre du programme de mathématiques."

TP5 du module "Statistique inférentielle".

William Sealy Gosset – alias *Student* – (1876 – 1937) est le précurseur des statisticiens industriels. Il fit toute sa carrière dans les brasseries *Guinness*, délaissant les possibilités qui lui furent offertes d'une carrière universitaire. Considéré par les brasseurs comme l'un des leurs, occupant ses loisirs à la statistique en vue de l'amélioration de la production, ses échanges avec les statisticiens universitaires étaient parfois vus d'un mauvais oeil par ses employeurs. Ceci explique le surnom de *Student* utilisé pour dénommer la loi dont il est à l'origine. En effet, la société *Guinness* l'autorisa à publier ses articles à condition qu'il use au choix, du pseudonyme "Pupil" ou "Student". *Gosset* choisit le second.

Considérons, qu'au sein d'une production répartie selon une loi normale de moyenne μ et d'écart type σ, on prélève au hasard un échantillon de taille n (petit). On note X_i la variable aléatoire qui, au $i^{ème}$ tirage, associe son résultat. On suppose que les X_i sont indépendantes, de même loi normale $\mathrm{N}(\mu, \sigma)$.

On note encore $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Lorsque $\mu = \mu_0$, la variable aléatoire $\overline{X}$ - μ_0 suit la loi normale $\mathrm{N}\left(0, \frac{\sigma}{\sqrt{n}}\right)$ qui n'est pas connue car ***σ est inconnu***. En statistique il est tentant, quand on a un paramètre σ inconnu de le remplacer par son estimation.

On sait que la variable aléatoire $S_{n-1}{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ est un estimateur sans biais de σ^2, $E(S_{n-1}{}^2) = \sigma^2$ (et fournit, pour n assez grand, une "bonne" estimation ponctuelle de σ^2). L'idée de *Student* a été d'introduire la variable aléatoire $T = \sqrt{n}\,\frac{\overline{X} - \mu_0}{S_{n-1}} = \sqrt{n-1}\,\frac{\overline{X} - \mu_0}{S_n}$.

Il a montré que si $\mu = \mu_0$ alors T suit une loi de probabilité, ***indépendante de σ***, que l'on peut calculer et appelée ***loi de Student à n – 1 degrés de liberté***.

1. ***Construction du test*** :

a - ***Choix des hypothèses*** :

H_0 : " $\mu = \mu_0$".

H_1 : " $\mu \neq \mu_0$" (test bilatéral) ou " $\mu > \mu_0$" (unilatéral à droite) ou " $\mu < \mu_0$" (à gauche).

b - ***Détermination de la région critique au seuil α*** :

Sous l'hypothèse H_0, la variable aléatoire $T = \sqrt{n-1}\,\frac{\overline{X} - \mu_0}{S_n}$,

où $S_n{}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$, suit la loi de *Student* à $n - 1$ degrés de liberté.

La table donne alors le réel positif h tel que :

$P(-h \leq T \leq h) = 1 - \alpha$ (cas bilatéral) ou

$P(T \leq h) = 1 - \alpha$ (cas unilatéral à droite) ou $P(-h \leq T) = 1 - \alpha$ (cas unilatéral à gauche).

c - ***Enoncé de la règle de décision***.
On prélève un échantillon de taille *n*, dont calcule la moyenne $\bar{x}$ et l'écart type s_n.
On calcule $t = \sqrt{n-1}\,\dfrac{\bar{x} - \mu_0}{s_n}$.
T Si $t \notin [-h\,,\,h]$, on rejette H_0 au risque α (cas bilatéral) ;
ou $t \notin]-\infty\,,\,h]$ (unilatéral à droite)
ou $t \notin [-h\,,\,+\infty[$ (unilatéral à gauche).
T Sinon, on accepte H_0 au seuil α.

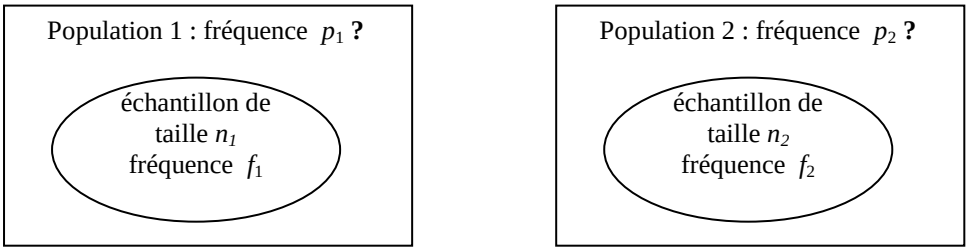
2. ***Utilisation du test*** :
Prélèvement d'un échantillon et prise de décision, selon le résultat observé.

Exemples :
• En prélevant avec remise un échantillon de taille $n = 15$ dans une population normale de moyenne μ, on souhaite tester l'hypothèse $H_0 : \mu = 30$ contre $H_1 : \mu \neq 30$ au seuil de 5%.
Construction du test : pour $\alpha = 0{,}05$ et $\nu = 14$, la valeur critique fournie par la table (voir page 80) est $h = 2{,}145$.
Utilisation du test : on prélève un échantillon de taille 15 avec remise sur lequel la moyenne est $\bar{x} = 37{,}2$ et l'écart type $s_{15} = 6{,}2$.
On a $t = \dfrac{37{,}2 - 30}{6{,}2}\sqrt{14} \approx 4{,}35$. On constate que $4{,}35 > 2{,}145$, on rejette donc l'hypothèse H_0 au seuil de 5%.
• Même exemple que le précédent dans le cadre unilatéral à droite :
On teste $H_0 : \mu = 30$ contre $H_1 : \mu > 30$ au seuil de 5%.
Sur la table de la page 80 (construite pour le cas bilatéral) on recherche, en raison de la symétrie de la loi de *Student*, la probabilité $P = 0{,}10$ ($P/2 = 0{,}05$) et on obtient la valeur critique $h = 1{,}761$. L'hypothèse H_0 est rejetée.

III - TESTS DE COMPARAISON (OU DE LA DIFFERENCE SIGNIFICATIVE)

1 - COMPARAISON DE DEUX FREQUENCES

a) Position du problème



On souhaite comparer les fréquences p_1 et p_2 d'un même phénomène, dans deux populations.
Après construction d'un test, il s'agit de voir si la différence $f_1 - f_2$ des fréquences observées sur un ***échantillon*** de chaque population est ***vraisemblable, avec l'hypothèse nulle*** "$p_1 = p_2$" (a-t-on une ***différence significative*** ?).

b) Plan du test de comparaison de deux fréquences

1. ***Construction du test*** :
a - ***Choix des hypothèses*** :
H_0 : " $p_1 = p_2$".
H_1 : " $p_1 \neq p_2$" (test bilatéral) ou
H_1 : " $p_1 > p_2$" (unilatéral à droite) ou H_1 : " $p_1 < p_2$" (unilatéral à gauche).
b - ***Détermination de la région critique au seuil α*** :
Sous l'hypothèse H_0, la variable aléatoire $D = F_1 - F_2$, qui à chaque paire d'échantillons de taille n_1 et n_2, respectivement issus des populations 1 et 2, associe la différence $f_1 - f_2$ des fréquences observées, suit approximativement, pour n_1 et n_2 assez grands, la loi $N\left(0, \sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}\right)$ (les variances s'ajoutent si l'on suppose F_1 et F_2 indépendantes).

Remarque : sous l'hypothèse H_0 on suppose l'égalité des fréquences dans les deux populations, certains prennent alors comme écart type de D l'expression $\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ où $f = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$.

On calcule alors le réel positif h tel que :
$P(-h \leq D \leq h) = 1 - \alpha$ (cas bilatéral) ou $P(D \leq h) = 1 - \alpha$ (cas unilatéral à droite) ou $P(-h \leq D) = 1 - \alpha$ (cas unilatéral à gauche).
c - ***Enoncé de la règle de décision***.
On prélève un échantillon de taille n_1 dans la population 1, puis un échantillon de taille n_2 dans la population 2, pour lesquels on calcule la différence des fréquences observées $f_1 - f_2$.
T Si $f_1 - f_2 \notin [-h, h]$, on rejette H_0 au risque α (cas bilatéral) ;

ou $f_1 - f_2 \notin]-\infty , h]$ (unilatéral à droite)
ou $f_1 - f_2 \notin [-h , +\infty[$ (unilatéral à gauche).
T Sinon, on accepte H_0 au seuil α.

2. ***Utilisation du test*** :
Prélèvement des deux échantillons et prise de décision, selon le résultat observé.

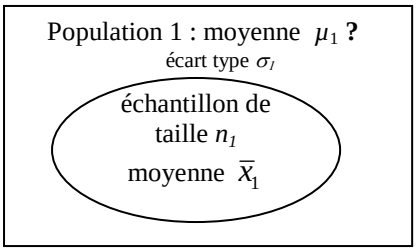
⇨ ***Voir annales de BTS***.

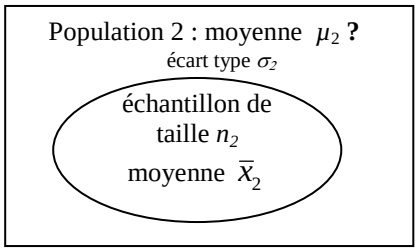
c) Utilisation des fonctions de la calculatrice

CASIO Graph 80	TI 83	SHARP EL 9600
STAT TEST Z 2-P p_1: ≠ p_2 <p_2 >p_2 choix bilatéral ou non entrer nombre de succès et tailles des échantillons Affichage comme les tests précédents	STAT TESTS 6:2-PropZTest entrer nombre de succès et tailles des échantillons p_1: ≠ p_2 <p_2 >p_2 choix bilatéral ou non Affichage comme les tests précédents	STAT E TEST 17 InputStats ENTER 11 Ztest2prop

2 - COMPARAISON DE DEUX MOYENNES

a) Position du problème





On souhaite comparer les moyennes μ_1 et μ_2 de deux populations.
Après construction d'un test, il s'agit de voir si la différence $\overline{x}_1 - \overline{x}_2$ des moyennes observées sur un ***échantillon*** de chaque population est ***vraisemblable, avec l'hypothèse nulle*** "$\mu_1 = \mu_2$".

b) Plan du test de comparaison de deux moyennes (grands échantillons ou écarts types connus)

1. ***Construction du test*** :
a - ***Choix des hypothèses*** :
H_0 : " $\mu_1 = \mu_2$".
H_1 : " $\mu_1 \neq \mu_2$" (test bilatéral) ou
H_1 : " $\mu_1 > \mu_2$" (unilatéral à droite) ou H_1 : " $\mu_1 < \mu_2$" (unilatéral à gauche).
b - ***Détermination de la région critique au seuil α*** :
Sous l'hypothèse H_0, la variable aléatoire $D = \overline{X_1} - \overline{X_2}$, qui à chaque paire d'échantillons de taille n_1 et n_2, respectivement issus des populations 1 et 2, associe la différence $\overline{x}_1 - \overline{x}_2$ des moyennes observées, suit approximativement, pour n_1 et n_2 assez grands, la loi $N\left(0 , \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$ (les variances s'ajoutent si l'on suppose $\overline{X_1}$ et $\overline{X_2}$ indépendantes).
On calcule alors le réel positif h tel que :
$P(-h \leq D \leq h) = 1 - \alpha$ (cas bilatéral) ou
$P(D \leq h) = 1 - \alpha$ (cas unilatéral à droite) ou $P(-h \leq D) = 1 - \alpha$ (cas unilatéral à gauche).

c - ***Enoncé de la règle de décision***.
On prélève un échantillon de taille n_1 dans la population 1, dont calcule la moyenne $\bar{x}_1$, puis, un échantillon de taille n_2 dans la population 2, pour lequel la moyenne vaut $\bar{x}_2$.
T Si $\bar{x}_1 - \bar{x}_2 \notin [-h, h]$, on rejette H_0 au risque α (cas bilatéral) ;
ou $\bar{x}_1 - \bar{x}_2 \notin]-\infty, h]$ (unilatéral à droite)
ou $\bar{x}_1 - \bar{x}_2 \notin [-h, +\infty[$ (unilatéral à gauche).
T Sinon, on accepte H_0 au seuil α.

2. ***Utilisation du test*** :
Prélèvement des deux échantillons et prise de décision, selon le résultat observé.

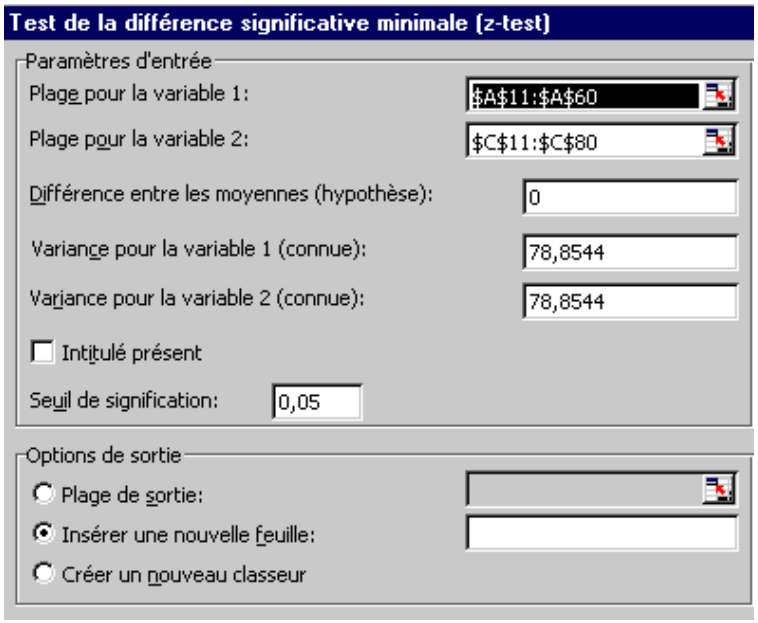
⇨ ***Voir annales de BTS***.

c) Utilisation des fonctions de la calculatrice ou de l'ordinateur

• Sur calculatrices :

CASIO Graph 80	TI 83	SHARP EL 9600
STAT TEST Z 2-S Data : Var μ_1: ≠ μ_2 <μ_2 >μ_2 choix bilatéral ou unilatéral Entrer écart types (ou estimation) des pop. moyennes et tailles des échantillons. Affichage analogue aux tests précédents	STAT TESTS 3:2-SampZTest Inpt: Stats Entrer moyennes et tailles des échantillons, écart types (ou estimation) des populations μ_1: ≠ μ_2 <μ_2 >μ_2 choix bilatéral ou uni. Pooled : NO Affichage analogue aux tests précédents	STAT E TEST 17 InputStats ENTER 09 Ztest2samp

• Sur Excel :



"L'*utilitaire d'analyse"* (qu'il faut éventuellement installer) permet l'accès à l'*outil* :
"*Z-test de la différence significative minimale*".

IV – TESTS D'AJUSTEMENT

"Exemples d'utilisation de la droite de Henry, du test du χ^2 *"*
"Ce TP n'est à réaliser qu'en liaison avec les enseignants des disciplines professionnelles et seulement si, dans celles-ci, ces procédures sont utilisées."
"Aucune connaissance à son sujet n'est exigible dans le cadre du programme de mathématiques."

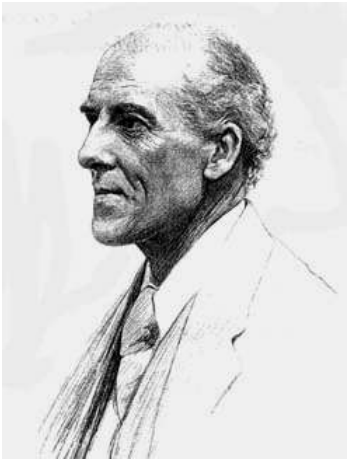
TP5 du module "Statistique inférentielle".

Les tests d'ajustement ont pour but de vérifier qu'un échantillon provient ou non d'une variable aléatoire de distribution donnée (connue).
Une première méthode, empirique, peut consister à comparer la forme de l'histogramme des fréquences observées aux histogrammes théoriques (dans le cas discret) ou au profil des fonctions de densité (dans le cas continu) des différents modèles possibles. Cette méthode peut déjà permettre d'éliminer certains modèles mais la qualité de l'ajustement n'est pas même quantifiée.

1 – PROCEDURES D'AJUSTEMENT LINEAIRE

Dans bien des cas, une transformation fonctionnelle (un changement de variable), on dit aussi une anamorphose (effet de perspective en peinture), permet de ramener l'ajustement à une régression linéaire selon les moindres carrés. On se contentera même parfois d'un ajustement linéaire "au jugé" par utilisation d'un papier fonctionnel.
C'est le cas de l'ajustement à une ***loi exponentielle*** (papier semi-logarithmique) ou à une ***loi de Weibull*** (papier d'*Alan Plait*) que nous envisagerons dans le cadre de la fiabilité.
L'idée est la même dans le cadre d'un ajustement à une ***loi normale*** selon la procédure de la ***droite de Henry*** (papier gausso-arithmétique), déjà vue.
Ces procédures sont simples à mettre en œuvre mais ont le défaut de ne pas quantifier les risques d'erreurs lors de la prise de décision, ce que permet en revanche la procédure des tests d'hypothèses. Le test du khi 2 est exposé ci-dessous, le ***test de Kolmogorov***, plus adapté pour tester une distribution continue (normalité par exemple), est exposé à la fin de la 4ème séance.

2 – LE TEST DU KHI-DEUX



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…

Par définition, la ***loi du khi-deux (ou chi-deux) à n degrés de liberté*** est la loi suivie par la somme *S* des carrés de *n* variables aléatoires indépendantes de loi normale centrée réduite.

Soit une variable aléatoire discrète ou discrétisée, c'est à dire divisée en k classes de probabilités $p_1, \ldots, p_k$.
Soit un échantillon de taille n de cette variable aléatoire, fournissant pour chaque classe des effectifs $x_1, \ldots, x_k$.
Il s'agit de comparer ces effectifs aux valeurs théoriques :

Classes	Effectifs observés	Effectifs théoriques
1	x_1	$t_1 = np_1$
…	…	…
k	x_k	$t_k = np_k$

Il faut décider d'un critère d'adéquation des observations par rapport au modèle théorique. De façon classique, on choisira l'écart quadratique réduit, noté χ^2_{obs} et valant :

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

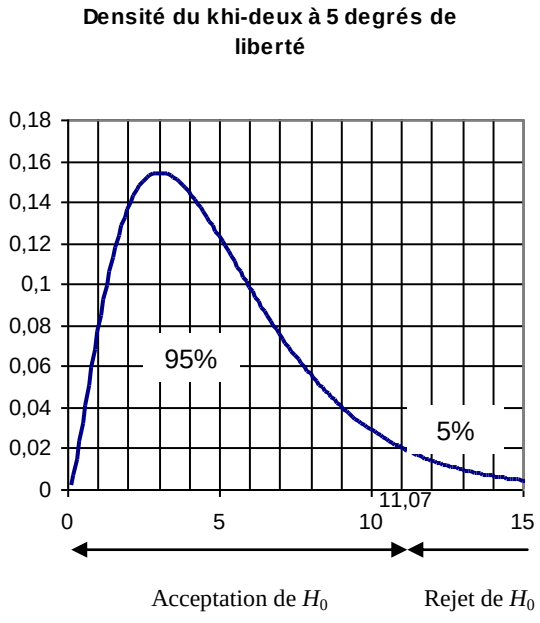
De grandes valeurs de χ^2_{obs} rendraient le modèle suspect.
Pour étudier la variabilité de ce critère, introduisons la variable aléatoire $T = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$ avec $\sum_{i=1}^{k} X_i = n$ où X_i est la variable aléatoire correspondant à l'effectif de la $i^{ème}$ classe.
K. Pearson a démontré que la loi de T est approximativement, pour n grand, une loi du χ^2 à $k - 1$ degrés de liberté (on peut noter que la relation ci-dessus fait que la valeur de X_k est déterminée dès que les valeurs de $X_1, \ldots, X_{k-1}$ sont connues).

La loi du khi-deux est tabulée.
Sur *Excel*, la fonction LOI.KHIDEUX(valeur t ; degrés de libertés) fournit la probabilité $P(T > t)$ et la fonction KHIDEUX.INVERSE(probabilité p ; degrés de liberté) renvoie la valeur t telle que $P(T > t) = p$.

Construction d'un test du khi-deux

• *Choix des hypothèses* :
H_0 : pour tout $1 \le i \le 6$, la probabilité que X prenne une valeur dans la classe i est p_i.
H_1 : il existe i tel que la probabilité précédente diffère de p_i.

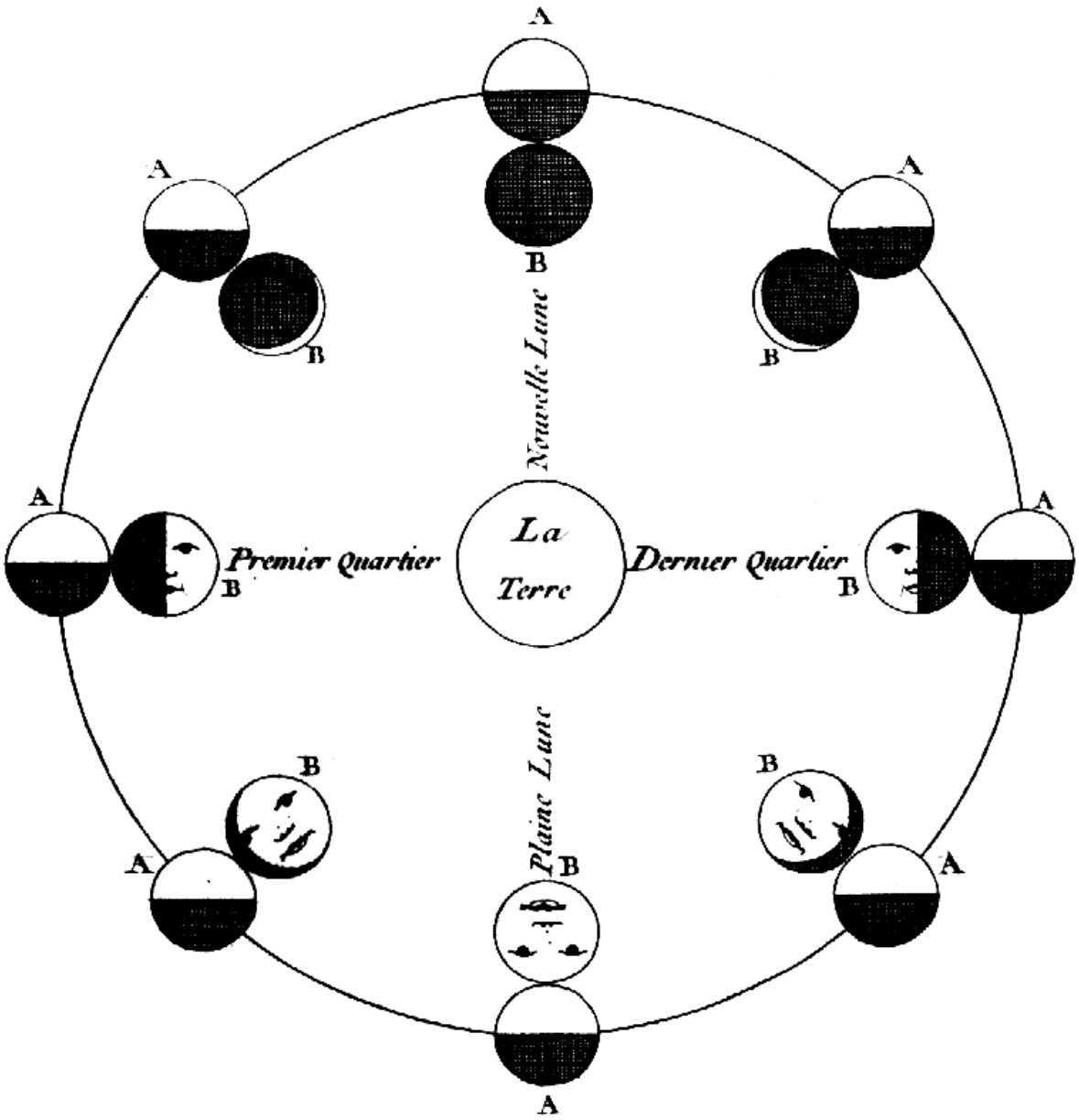
• *Détermination de la région critique* :
Si H_0 est vraie, la variable aléatoire $T = \sum_{i=1}^{k} \frac{(X_i - t_i)^2}{t_i}$ suit approximativement la loi du khi-deux à $k - 1$ degrés de liberté.
On recherche sur une table le réel t tel que, $P(T > t) = \alpha$.
D'où la zone d'acceptation de H_0 au seuil α : $[0, t]$.



• *Règle de décision* :
Soit χ^2_{obs} l'écart quadratique réduit obtenu entre les effectifs observés et les effectifs théoriques. Si $\chi^2_{obs} \leq t$ on accepte H_0 au seuil de 5%. Si $\chi^2_{obs} > t$ on rejette H_0.

Remarques :
• Si l'on utilise l'échantillon pour estimer indépendamment j paramètres de la loi testée, le degré de liberté du khi-deux devient : $k - 1 - j$. Par exemple $k - 3$ pour une loi normale où l'on estime μ et σ à l'aide de $\bar{x}$ et s_{n-1}.
• La loi de T n'est qu'approchée et on donne la condition $np_i > 5$ pour l'effectif de chaque classe (sinon on regroupe les classes à effectif trop faible). C'est pour vérifier cette condition que l'on pratique le test du khi-deux sur les effectifs et non sur les fréquences.

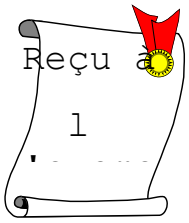
⇨ ***Voir le TP Excel : NORMALITE D'UNE PRODUCTION – TEST DU KHI 2***.


A
A
B
A
B
B
Nouvelle Lune
A
Premier Quartier
B
La
Terre
Dernier Quartier
A
B
B
B
A
Pleine Lune
B
A

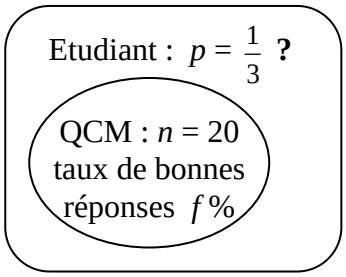
TRAVAUX DIRIGES

T.P. Calculatrices : INTRODUCTION AUX TESTS STATISTIQUES

Dans un lycée syldave, les professeurs, exaspérés par le manque de travail d'une partie des étudiants de BTS, décident d'établir un examen de passage à la fin du premier semestre (les mœurs syldaves sont assez rudes ...).
L'examen se présentera sous la forme d'un QCM de 20 questions indépendantes. A chaque question, trois réponses sont proposées, dont une seule est exacte. Un étudiant n'ayant fourni aucun travail, répondra au hasard et donc, correctement, avec une probabilité $p = \frac{1}{3}$ à chaque question.



L'objectif des professeurs est de recaler ce type d'étudiant, avec une probabilité d'environ 95 %.
Pour cela il faut définir la barre d'acceptation *avant* l'épreuve, de sorte que les étudiants souscrivent au ***protocole*** ("règles du jeu") de l'examen.



On peut considérer ce QCM comme un ***test statistique*** devant permettre de détecter si l'étudiant qui le passe répond au hasard ($p = \frac{1}{3}$). Le QCM est un ***échantillon aléatoire*** non exhaustif de ses réponses.

I - CONSTRUCTION DU TEST

1) Choix des hypothèses

On teste l'hypothèse H_0 : " $p = \frac{1}{3}$ " (appelée "***hypothèse nulle***"), contre l'***hypothèse alternative*** H_1 : " $p > \frac{1}{3}$ " (c'est un test "***unilatéral***").
L'hypothèse nulle correspond à un étudiant répondant au hasard. L'hypothèse alternative doit "au contraire" correspondre à un étudiant qui a travaillé. Pourquoi ne prend-on pas " $p \neq \frac{1}{3}$ " ? .....
.....

2) Calcul de la zone d'acceptation de H_0

On suppose que H_0 est vraie : l'étudiant répond au hasard. On désigne par X la variable aléatoire qui, à chaque étudiant de ce type, associe le nombre de ses bonnes réponses au QCM.
Quelle est le loi de X (justifier) ? .....
.....
.....

A l'aide de la table ci-contre, déterminer le nombre k de bonnes réponses tel que $P(X \leq k)$ soit le plus proche possible de 95 %.

.....

.....

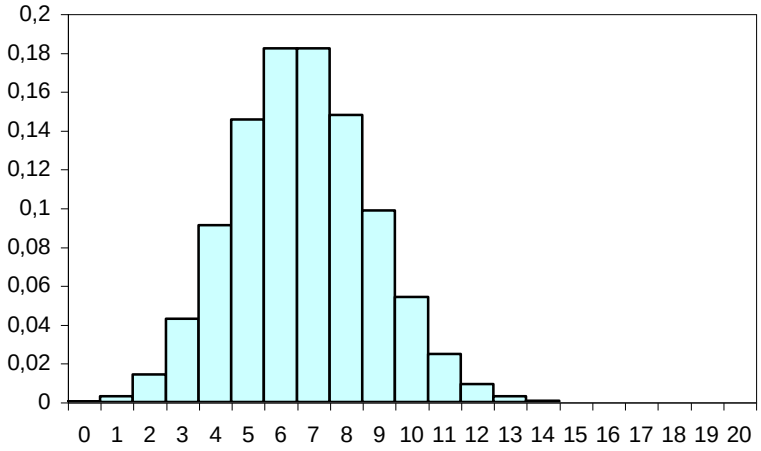
Quand le nombre de bonnes réponses est inférieur ou égal à k, on acceptera H_0 .

S'il est strictement supérieur à k, on supposera que l'étudiant a travaillé et l'on rejettera H_0, avec un ***risque*** de rejet à tort de : $\alpha = P(X > k)$.

Quel est, ici, le risque 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



3) Règle de décision

Enoncer la ***règle de décision*** de l'examen.

.....

.....

II - UTILISATION DU TEST ET ERREURS

1) Expérimentation du test

Le programme suivant choisit aléatoirement une valeur de p : avec une chance sur deux, $p = \frac{1}{3}$ ou $p = 0{,}60$ (cas d'un étudiant ayant moyennement travaillé). Puis, il simule le passage de l'examen et affiche le nombre x de réponses correctes ainsi que la valeur de p.

CASIO	TI 82 - 83	TI 89 - 92
1÷3+Int(Ran# +.5)(.6-1÷3) → P↵ 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 ?

.....

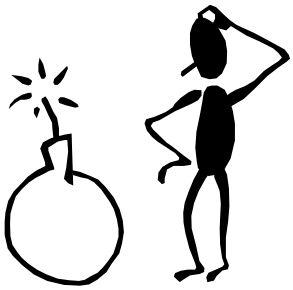
2) Les erreurs

Décision / Réalité	H_0 acceptée	H_1 acceptée
H_0 vraie	$1 - \alpha$	α ERREUR DE 1ère ESPECE
H_1 vraie	β ERREUR DE 2ème 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

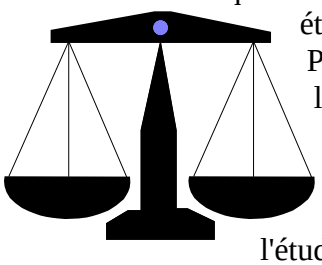
3) L'erreur de 2nde espèce

Un étudiant se manifeste alors. Il est sérieux et travailleur, mais, mal assuré, il perd souvent une partie de ses moyens à l'examen. Il estime cependant sa probabilité de bien répondre à une question à $p = 0{,}6$.
"C'est pas juste ! Bien qu'ayant une probabilité de bonne réponse de 60 %, j'ai une chance sur quatre d'être recalé !"
Vérifier l'affirmation de cet étudiant, qui craint d'être victime d'une erreur de 2ème espèce (utiliser la table ci-contre, des valeurs cumulées de la loi $\mathcal{B}(20 ; 0{,}60)$).

n = 20	p = 0,60
k	P(X <= k)
0	1,09951E-08
1	3,40849E-07
2	5,04126E-06
3	4,7345E-05
4	0,000317031
5	0,001611525
6	0,006465875
7	0,021028927
8	0,056526367
9	0,127521246
10	0,244662797
11	0,404401275
12	0,584107062
13	0,749989328
14	0,874401027
15	0,949048047
16	0,984038837
17	0,996388528
18	0,999475951
19	0,999963438
20	1

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

Il faut avouer que ce n'est pas très moral vis à vis de cet étudiant.



Pour diminuer le risque de 2ème espèce, l'étudiant propose de baisser la barre d'admission à 8 : si le nombre x de bonnes réponses est tel que $x \leq 7$, l'étudiant est recalé, si $x \geq 8$, l'étudiant est reçu.

Quel est, dans ces conditions, le risque β de 2ème espèce, pour un étudiant tel que $p = 0{,}60$? .....
.....
Mais que devient le risque α d'admettre un étudiant n'ayant pas travaillé ? .....
.....
.....

III - TEST DE 100 QUESTIONS

Les professeurs jugeant ce risque de première espèce inacceptable, décident, pour diminuer β sans augmenter α, de proposer un QCM de 100 questions.

1) Construction du test

• ***Choix des hypothèses :***

On teste l'hypothèse H_0 : " $p = \frac{1}{3}$ " , contre H_1 : " $p > \frac{1}{3}$ " .

• ***Calcul de la zone d'acceptation de H_0, au seuil α de 5 % :***

On suppose que H_0 est vraie : $p = \frac{1}{3}$. On désigne par X la variable aléatoire qui, à chaque étudiant de ce type, associe le nombre de ses bonnes réponses au QCM. On sait que X suit la loi $\mathcal{B}(100 , \frac{1}{3})$. Pour simplifier les calculs, on approche la loi de X par une loi normale.

Quels en sont les paramètres ? .....

On note $F = \frac{1}{100}X$, la variable aléatoire correspondant aux fréquences des bonnes réponses à un QCM. En supposant que F suive une loi normale, quels en sont les paramètres, sous l'hypothèse H_0 ? .....

Déterminer le réel h tel que, sous l'hypothèse H_0, $P(F \leq h) = 0{,}95$.

.....

.....

.....

• ***Règle de décision :***

.....

.....

2) Simulation

Reprendre le programme précédent, en remplaçant **20** par **100**.

Comparer l'efficacité de ce Q.C.M. au précédent (observer la fréquence des erreurs de première et seconde espèce).

.....

.....

.....

.....

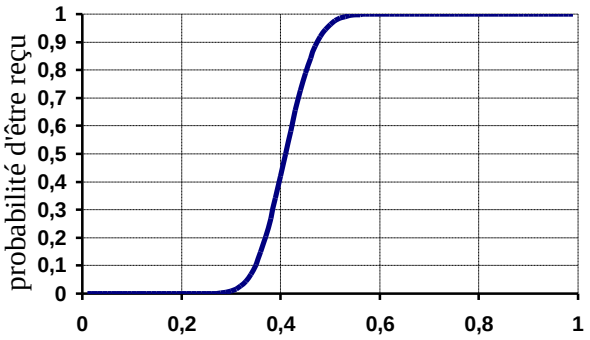
POUR ALLER PLUS LOIN ...

Puissance du test

Plaçons nous maintenant du point de vue de l'étudiant, pour lequel $p = p_0$ et qui se préoccupe de sa probabilité d'être reçu : $1 - \beta(p_0) = P(F > 0{,}41 \mid p = p_0)$.

Montrer que, si $p = p_0$, on a $1 - \beta(p_0) = P(F > 0{,}41) = 1 - \Pi\left(10 \times \frac{0{,}41 - p_0}{\sqrt{p_0(1-p_0)}}\right)$, où Π est la fonction de répartition de la loi $\mathcal{N}(0\ ,\ 1)$. .....

.....



La fonction $p \mapsto 1 - \beta(p)$ se nomme "***puissance***" du test et est représentée ci-contre.

Lire sur le graphique, la probabilité d'être reçu au nouvel examen, d'un étudiant tel que $p = 0{,}40$, puis tel que $p = 0{,}60$. Confirmer par le calcul.

Est-ce raisonnable ?

.....

.....

.....

.....

.....

Corrigé des travaux dirigés "INTRODUCTION AUX TESTS STATISTIQUES"

I – CONSTRUCTION DU TEST

1) Le cas $p < 1/3$ n'est pas envisagé (que dire d'un étudiant cherchant à répondre faux ?). La forme de la région de rejet dépend de H_1 qui correspond uniquement à $p > 1/3$ (l'étudiant ne répond pas au hasard).

2) Si H_0 est vraie, on a $p = 1/3$. Répondre au Q.C.M. est alors la répétition de 20 expériences aléatoires indépendantes, avec deux issues possibles (bonne réponse avec $p = 1/3$, ou mauvaise réponse) et où X associe le nombre de bonnes réponses. La variable aléatoire X suit donc la loi binomiale $\mathcal{B}(20, 1/3)$.

Avec la table fournie, on constate que $k = 10$, avec $P(X \leq 10) \approx 0{,}96$.

Le risque α est donc $\alpha \approx 4\%$.

3) Règle de décision

Soit x le nombre de bonnes réponses au Q.C.M.,

- si $x \leq 10$ alors H_0 est acceptée et l'étudiant est RECALE,
- si $x \geq 11$ alors H_0 est refusée et l'étudiant est ADMIS.

II – UTILISATION DU TEST ET ERREURS

1) La simulation permet de "vivre" les aléas du hasard, et d'observer les deux types d'erreurs.

2) On observe deux types d'erreurs de décision.

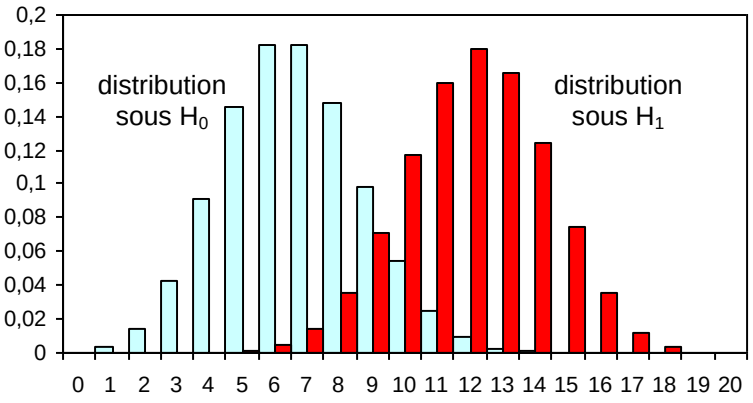
L'erreur de première espèce correspond à l'étudiant qui n'a pas travaillé et est reçu à l'examen.

L'erreur de seconde espèce correspond à l'étudiant qui a travaillé et est recalé.

3) L'erreur de seconde espèce :

Si H_1 est vraie, on a $p = 0{,}6$ et la variable aléatoire X suit alors la loi binomiale $\mathcal{B}(20 ; 0{,}60)$.

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 $\mathcal{B}(20 ; 0{,}60)$, $\beta \approx 2\ \%$.

Mais alors, d'après la table de la loi $\mathcal{B}(20 ; 1/3)$, $\alpha \approx 100 - 66 = 34\ \%$!

III – TEST DE 100 QUESTIONS

1) Construction du test

Sous l'hypothèse H_0, on approche la loi $\mathcal{B}$ (100, 1/3) de la variable aléatoire X par la loi normale $\mathcal{N}\left(\frac{100}{3}, \sqrt{100}\times\frac{1}{3}\times\frac{2}{3}\right)$.

La variable aléatoire $F = \frac{1}{100}X$ suit alors approximativement la loi $\mathcal{N}\left(\frac{1}{3}, \frac{\sqrt{2}}{30}\right)$.

On pose $T = \dfrac{F-\frac{1}{3}}{\frac{\sqrt{2}}{30}}$, qui suit la loi $\mathcal{N}$ (0, 1). On a $P(F \le h) = P\left(T \le \left(h-\frac{1}{3}\right)\times\frac{30}{\sqrt{2}}\right)$ = 0,95 d'où,

d'après la table de la loi normale centrée réduite, $h = \frac{1}{3} + 1{,}645\frac{\sqrt{2}}{30} \approx 0{,}41$.

La règle de décision du test de 100 questions, au seuil de 5 % est donc :

Soit x le nombre de bonnes réponses. Si $x \le 41$, le candidat est recalé. Si $x \ge 42$, le candidat est admis.

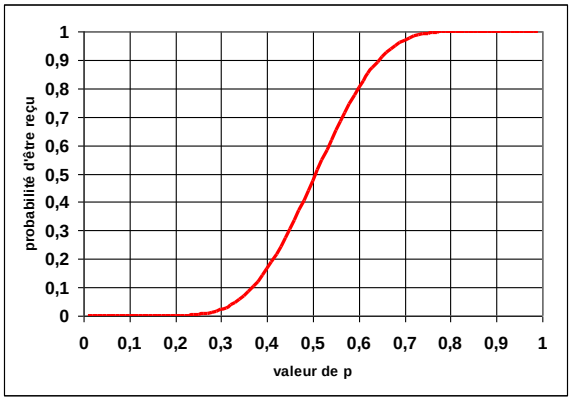
2) Puissance du test

Sous l'hypothèse $p = p_0$, la variable aléatoire F suit approximativement la loi normale $\mathcal{N}\left(p_0, \sqrt{\frac{p_0(1-p_0)}{100}}\right)$. On a donc $P(F > 0{,}41) = P\left(T > \frac{0{,}41-p_0}{\sqrt{p_0(1-p_0)}}\times 10\right)$ et la probabilité d'être reçu est donc bien $1 - \beta(p_0) = P(F > 0{,}41) = 1-\Pi\left(10\times\frac{0{,}41-p_0}{\sqrt{p_0(1-p_0)}}\right)$.

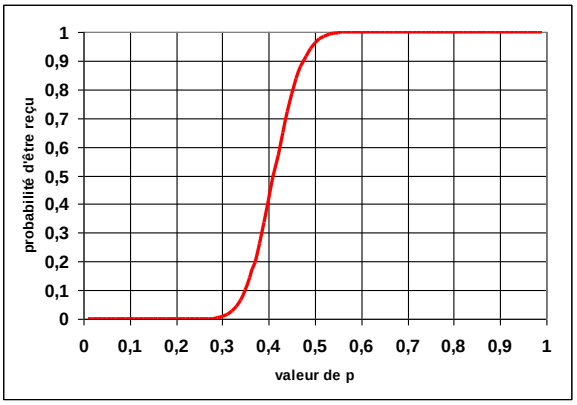
Pour un étudiant tel que la probabilité de bonne réponse est p = 0,4 , la probabilité d'être reçu à l'examen est $1 - \beta(0{,}4) = 1 - \Pi(0{,}204) \approx 0{,}42$.

Pour un étudiant tel que la probabilité de bonne réponse est p = 0,6 , la probabilité d'être reçu à l'examen est $1 - \beta(0{,}6) = 1 - \Pi(-3{,}88) \approx 0{,}99995$.

Ces résultats sont "raisonnables", c'est à dire conformes à ce que l'on attend d'un examen. Ce test est donc "puissant" en ce sens que son pouvoir de discrimination est important.

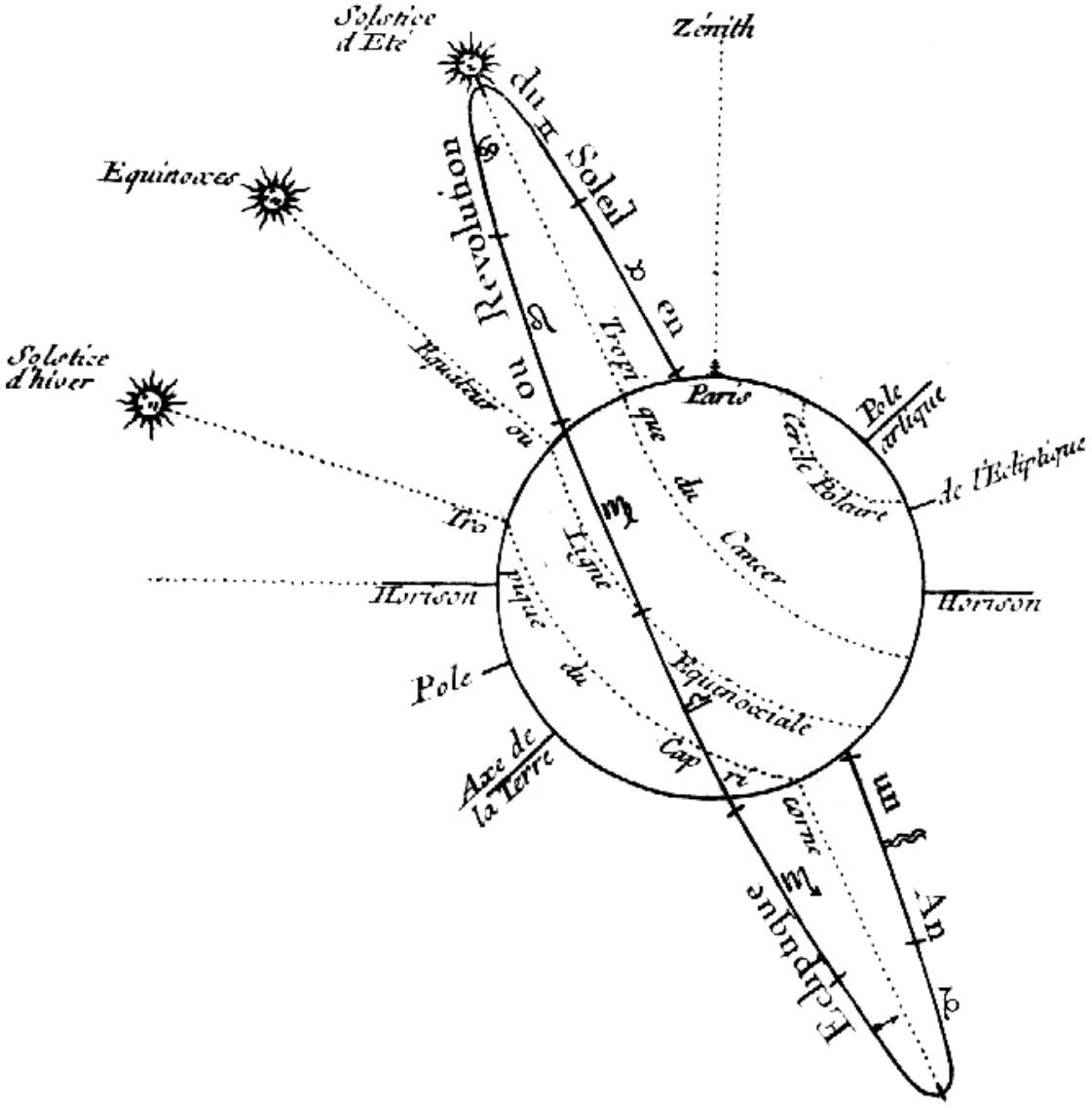


n = 20



n = 100

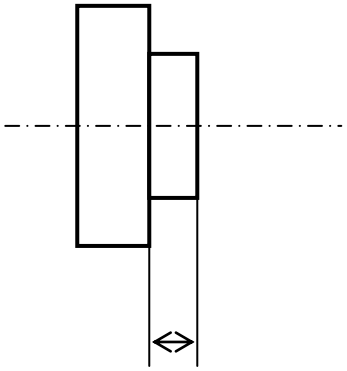
Comparaison des courbes de puissance des tests au seuil α de 5 %, pour n = 20 questions et pour n = 100 questions.


Solstice d'Eté
Zénith
Equinoxes
Solstice d'hiver
Paris
Pole arctique
Horison
Pole
Axe de la Terre
Equateur
Tropique du Cancer
Ligne équinoxiale
Cercle Polaire
Ecliptique
Révolution du Soleil en un An

Travaux pratiques STATISTIQUES

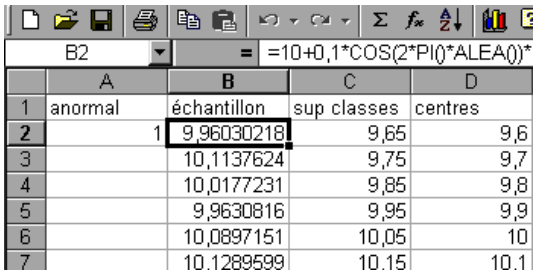
T.P. Excel : NORMALITE D'UNE PRODUCTION TEST DU KHI-DEUX

On souhaite placer sous contrôle statistique la fabrication d'un galet (sorte de roulette) pour la cote de 10 mm correspondant à la figure ci-contre.
On souhaite tester si la production se fait, pour cette cote, selon la loi normale N (10 ; 0,1).
Pour cela, on prélèvera dans la production un échantillon aléatoire de taille 100. Vu la quantité produite, on pourra considérer que ce prélèvement est effectué avec remise.



I – SIMULATION D'UN ECHANTILLON DE 100 GALETS EXTRAIT D'UNE PRODUCTION

Ouvrir *Excel*.
En **A1** écrire : anormal.
En **A2** entrer la ***formule*** : =ENT(ALEA()+0,5)
Cette formule permettra qu'une fois sur deux, la production simulée ne soit pas de distribution normale.



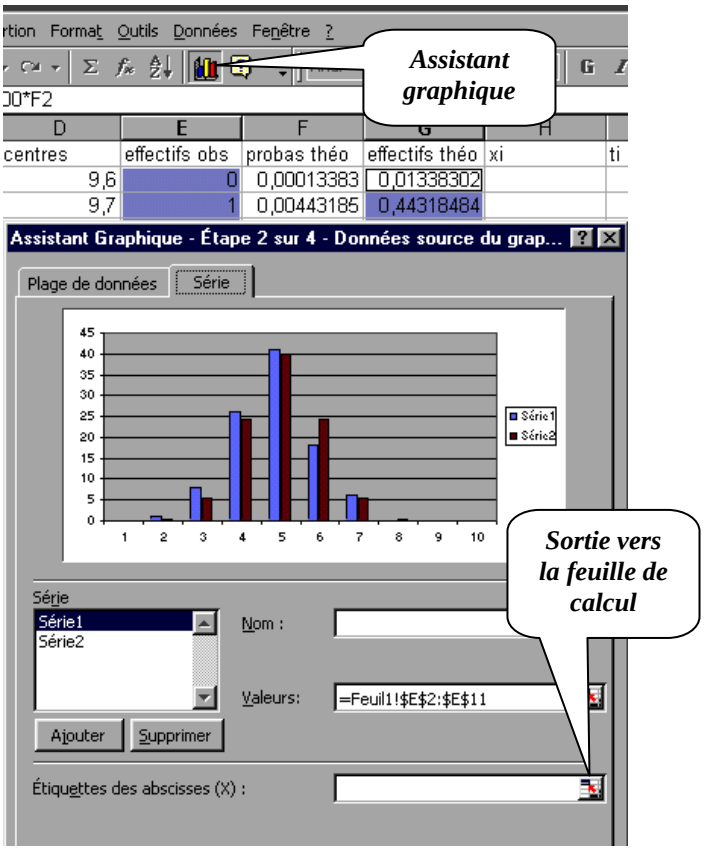
En **B1** écrire : échantillon.
En **B2** entrer la ***formule*** :
=10+0,1*COS(2*PI()*ALEA())*RACINE(–2*LN(ALEA()))+$A$2*0,3*ENT(ALEA()+0,1)
La première partie de cette formule simule une réalisation selon la loi N (10 ; 0,1) alors que la seconde partie, lorsque A2 contient la valeur 1, introduit une perturbation.
Cliquer dans **B2** puis, lorsque le pointeur de la souris s'est transformé en croix noire, glisser pour ***recopier vers le bas*** jusqu'en **B101**.
Vous allez regrouper les résultats de l'échantillon en 10 classes.
En **C1** écrire : sup classes.
En **C2** entrer la borne supérieure de la première classe, soit : 9,65.
En **C3** entrer la ***formule*** : =C2+0,1 puis ***recopier vers le bas*** jusqu'en **C11**.
En **D1** écrire : centres.
En **D2** entrer la ***formule*** : =C2 – 0,05 puis ***recopier vers le bas*** jusqu'en **D11**.
En **E1** écrire : effectifs obs.
De façon a compter les effectifs de chaque classe, ***sélectionner*** (croix blanche) les cellules de **E2** à **E11** puis, sans appuyer sur Entrée, taper la ***formule*** :
=FREQUENCE(B2:B101;C2:C11)
puis valider en appuyant simultanément sur ***CTRL MAJUSCULE ENTREE***.
On doit comparer les effectifs observés à ceux, théoriques, que donnerait la loi normale N (10 ; 0,1).
En **F1** écrire : probas théo.
En **F2** entrer la ***formule*** : =0,1*LOI.NORMALE(D2;10;0,1;FAUX)

puis ***recopier vers le bas*** jusqu'en **F11**.
En **G1** écrire : effectifs théo.
En **G2** entrer la ***formule*** : =100*F2 puis ***recopier vers le bas*** jusqu'en **G11**.

– **Compléter la feuille réponse**.

II – COMPARAISON DES HISTOGRAMMES

Une première façon d'étudier la normalité de la production est de comparer les histogrammes observés et théoriques.
Sélectionner les cellules de **E2** à **E11** ainsi que (appuyer sur la touche ***CTRL***) de **G2** à **G11** puis cliquer sur l'icône de l'***Assistant graphique***.
Etape 1/4 :
Choisir ***Histogramme*** et cliquer sur ***Suivant***.
Etape 2/4 :
Dans l'onglet ***Série***, à la rubrique ***Etiquette des abscisses (X) :*** sortir vers la feuille de calcul (en cliquant sur l'icône) pour ***sélectionner*** les centres des classes, puis revenir dans l'Assistant graphique par l'icône analogue.
Cliquer sur ***Terminer***.



– **Compléter la feuille réponse**.

II – TEST DU KHI-DEUX

Vous allez regrouper les classes à trop faible effectif.
En **H1** écrire : xi.
En **H4** entrer la ***formule*** : =E2+E3+E4
En **H5** entrer la ***formule*** : =E5 et ***recopier vers le bas*** jusqu'en **H7**.
En **H8** entrer la ***formule*** : =SOMME(E8:E11)
En **I1** écrire : ti.
En **I4** entrer la ***formule*** : =G2+G3+G4
En **I5** entrer la ***formule*** : =G5 et ***recopier vers le bas*** jusqu'en **I7**.
En **I8** entrer la ***formule*** : =SOMME(G8:G11)

La procédure du test du khi-deux

Soit X la variable aléatoire qui à tout galet pris au hasard dans la production associe sa cote en mm.
On considère un échantillon de taille 100 de cette variable aléatoire, fournissant pour chaque classe des effectifs $x_1, \ldots, x_6$.
Il s'agit de comparer ces effectifs aux valeurs théoriques.

Centre de classe	Effectifs observés	Effectifs théoriques
9,7	x_1	t_1
9,9	x_2	t_2
…	…	…
10,1		
10,35	x_5	t_5

Il faut décider d'un critère d'adéquation des observations par rapport au modèle théorique. De façon classique, on choisira l'*écart quadratique réduit*, noté χ^2_{obs} et valant $\chi^2_{obs} = \sum_{i=1}^{5} \frac{(x_i - t_i)^2}{t_i}$. De grandes valeurs de χ^2_{obs} rendraient le modèle suspect.

En **J1** écrire : écart.
En **J4** entrer la ***formule*** : =(I4-H4)^2/I4 puis ***recopier vers le bas*** jusqu'en **J8**.
En **I10** écrire : khi 2 obs.
En **J10** entrer la ***formule*** : =SOMME(J4:J8)

Pour étudier la variabilité de ce critère, on introduit la variable aléatoire $T = \sum_{i=1}^{5} \frac{(X_i - t_i)^2}{t_i}$

avec $\sum_{i=1}^{5} X_i = 100$.

On montre que la loi de T est approximativement une loi connue sous le nom de khi 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 $X_1, \ldots, X_4$ sont connues).
La loi du khi-deux est tabulée. Sur *Excel* :
La fonction KHIDEUX.INVERSE(probabilité p ; degrés de liberté) renvoie la valeur t telle que $P(T > t) = p$.
En **I11** écrire : limite accept.
En **J11** entrer la ***formule*** : =KHIDEUX.INVERSE(0,05;4)
En **I13** écrire : test khi 2.
En **J13** entrer la ***formule*** : =SI(J10<J11;"NORMAL";"ANORMAL")
En **I14** écrire : réalité.
En **J14** entrer la ***formule*** : =SI(A2=0;"NORMAL";"ANORMAL")
Faire plusieurs fois ***F9*** pour renouveler les simulations.

– **Compléter la feuille réponse**.

– FEUILLE REPONSE

NOMS : *.....*

I – SIMULATION D'UN ECHANTILLON DE 100 GALETS EXTRAIT D'UNE PRODUCTION

1) Quelle est l'amplitude des 10 classes où l'on regroupe les données de l'échantillon ?

.....

2) On considère la variable aléatoire X qui, à chaque galet pris au hasard dans la production, associe sa cote en mm. On suppose que X suit la loi normale $\mathcal{N}(10 ; 0{,}1)$ de fonction de densité f. On a alors $P(9{,}55 \leq X \leq 9{,}65) = \int_{9{,}55}^{9{,}65} f(x)\,dx$.

Justifier que cette probabilité vaut environ $0{,}1 \times f(9{,}6)$ (c'est ainsi qu'elle est calculée en F2). .....
.....
.....

II – COMPARAISON DES HISTOGRAMMES

Comment, en analysant les histogrammes, peut-on déceler que la production ne se fait pas selon une loi normale ?

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

II – TEST DU KHI-DEUX

Construction d'un test du khi-deux au seuil de 5%

• *Choix des hypothèses* :
H_0 : pour tout $1 \leq i \leq 5$, la probabilité que X prenne une valeur dans la classe i est p_i.
H_1 : il existe i tel que la probabilité précédente diffère de p_i.

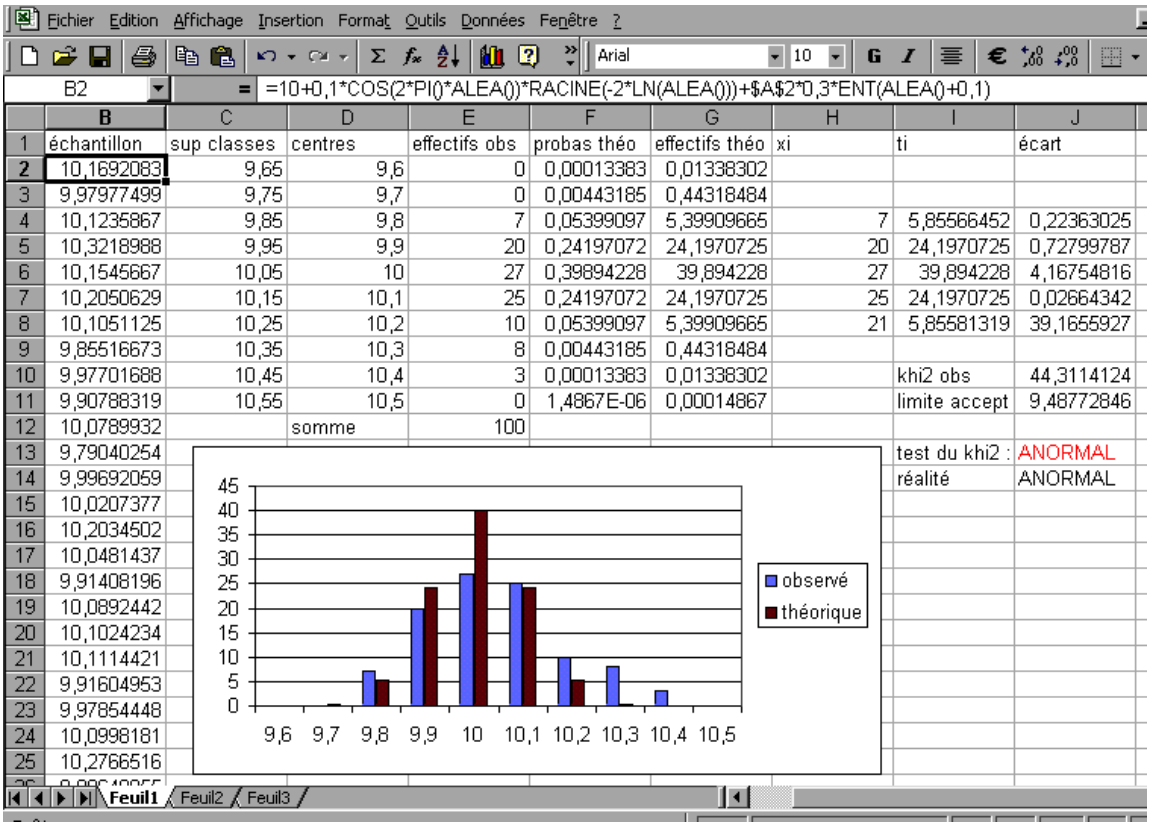
• *Détermination de la région critique* : Si H_0 est vraie, la variable aléatoire $T = \sum_{i=1}^{5} \frac{(X_i - t_i)^2}{t_i}$, où X_i est la variable aléatoire correspondant à l'effectif de la $i^{\text{ème}}$ classe, suit approximativement la loi du khi-deux à 4 degrés de liberté.
Le réel t tel que, $P(T > t) = \alpha$ est donné par *Excel* : on a $t \approx$ ……
D'où la zone d'acceptation de H_0 au seuil α : ……

• *Règle de décision* : Soit χ^2_{obs} l'écart quadratique réduit obtenu entre les effectifs observés et les effectifs théoriques. Le nombre χ^2_{obs} est contenu dans la cellule …….
Si $\chi^2_{\text{obs}} \leq t$ on accepte H_0 au seuil de 5%. Si $\chi^2_{\text{obs}} > t$ on rejette H_0.

• *Utilisation du test* : Observez-vous des erreurs de décision ? .....
.....
En **J11** remplacer 0,05 par 0,10. A quoi cela correspond-il ? .....
Qu'observe-t-on sur de nombreuses simulations ? .....

Eléments de réponse pour l'activité "NORMALITE D'UNE PRODUCTION – TEST DU KHI 2"

Un exemple de décision correcte :



Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?

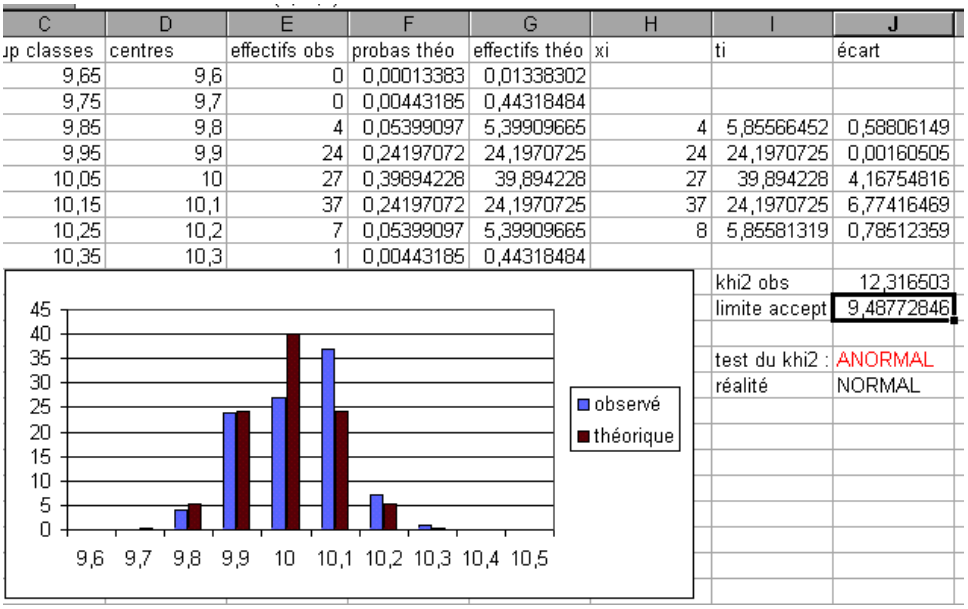
Arial 10

B2 =10+0,1*COS(2*PI()*ALEA())*RACINE(-2*LN(ALEA()))+$A$2*0,3*ENT(ALEA()+0,1)

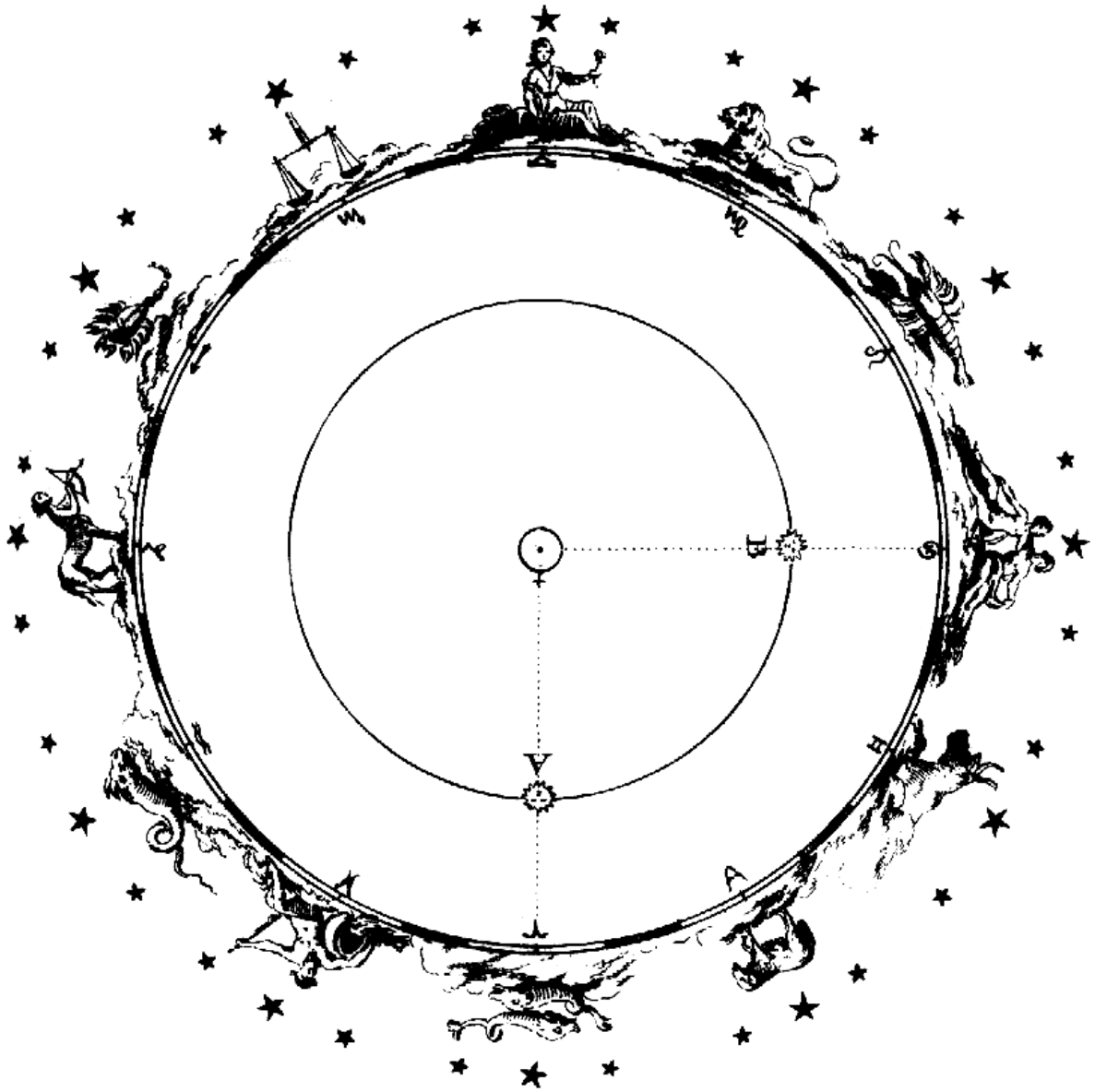
	B	C	D	E	F	G	H	I	J
1	échantillon	sup classes	centres	effectifs obs	probas théo	effectifs théo	xi	ti	écart
2	10,1692083	9,65	9,6	0	0,00013383	0,01338302			
3	9,97977499	9,75	9,7	0	0,00443185	0,44318484			
4	10,1235867	9,85	9,8	7	0,05399097	5,39909665	7	5,85566452	0,22363025
5	10,3218988	9,95	9,9	20	0,24197072	24,1970725	20	24,1970725	0,72799787
6	10,1545667	10,05	10	27	0,39894228	39,894228	27	39,894228	4,16754816
7	10,2050629	10,15	10,1	25	0,24197072	24,1970725	25	24,1970725	0,02664342
8	10,1051125	10,25	10,2	10	0,05399097	5,39909665	21	5,85581319	39,1655927
9	9,85516673	10,35	10,3	8	0,00443185	0,44318484			
10	9,97701688	10,45	10,4	3	0,00013383	0,01338302		khi2 obs	44,3114124
11	9,90788319	10,55	10,5	0	1,4867E-06	0,00014867		limite accept	9,48772846
12	10,0789932		somme	100					
13	9,79040254							test du khi2 :	ANORMAL
14	9,99692059							réalité	ANORMAL
15	10,0207377								
16	10,2034502								
17	10,0481437								
18	9,91408196								
19	10,0892442								
20	10,1024234								
21	10,1114421								
22	9,91604953								
23	9,97854448								
24	10,0998181								
25	10,2766516								

Feuil1 Feuil2 Feuil3

Au seuil de 5%, les erreurs de première espèce se produisent, sous H_0, dans 5% des cas, donc ici 2,5% des simulations (on est sous H_0 dans 50% des cas).
Une erreur de première espèce :



C	D	E	F	G	H	I	J
up classes	centres	effectifs obs	probas théo	effectifs théo	xi	ti	écart
9,65	9,6	0	0,00013383	0,01338302			
9,75	9,7	0	0,00443185	0,44318484			
9,85	9,8	4	0,05399097	5,39909665	4	5,85566452	0,58806149
9,95	9,9	24	0,24197072	24,1970725	24	24,1970725	0,00160505
10,05	10	27	0,39894228	39,894228	27	39,894228	4,16754816
10,15	10,1	37	0,24197072	24,1970725	37	24,1970725	6,77416469
10,25	10,2	7	0,05399097	5,39909665	8	5,85581319	0,78512359
10,35	10,3	1	0,00443185	0,44318484			
						khi2 obs	12,316503
						limite accept	9,48772846
						test du khi2 :	ANORMAL
						réalité	NORMAL



Epreuves corrigées de B.T.S.

Annales du B.T.S. : TESTS D'HYPOTHESES

TEST D'UNE FREQUENCE

1 – Analyses biologiques 1999

Les statistiques ont permis d'établir qu'en période de compétition la probabilité, pour un sportif pris au hasard, d'être déclaré positif au contrôle antidopage est égale à 0,02.
On décide de construire un test qui, à la suite des contrôles sur un échantillon de 50 sportifs prélevé au hasard, permette de décider si, au seuil de signification de 10 %, le pourcentage de sportifs contrôlés positifs est de $p = 0{,}02$.

a) *Construction du test bilatéral :*
Soit F la variable aléatoire qui, à tout échantillon aléatoire (supposé non exhaustif) de 50 sportifs contrôlés, associe le pourcentage de sportifs contrôlés positivement.

On suppose que F, suit la loi normale $\mathrm{N}\left(p, \sqrt{\dfrac{p(1-p)}{n}}\right)$ où $p = 0{,}02$ et $n = 50$.

Enoncer une hypothèse nulle H_0 et une hypothèse alternative H_1 pour ce test bilatéral. Déterminer, sous l'hypothèse H_0, le réel positif a tel que $P(p - a \leq F \leq p + a) = 0{,}9$. Enoncer la règle de décision du test.

b) *Utilisation du test :*
Dans l'échantillon E deux contrôles antidopage ont été déclarés positifs. En appliquant la règle de décision du test à cet échantillon assimilé à un échantillon aléatoire non exhaustif, peut-on conclure au seuil de risque 10 % que l'échantillon observé est représentatif de l'ensemble de la population sportive ?

2 – Biochimiste 1994

On effectue des contrôles d'alcoolémie d'automobilistes dans une région donnée un jour donné. Les statistiques permettent d'établir que la probabilité qu'un automobiliste choisi au hasard dans les conditions précédentes présente un contrôle positif est 2 % .
Dans une ville donnée de cette région, on effectue au hasard 200 contrôles de taux d'alcoolémie d'automobilistes, dans les conditions précédentes.
Les taux d'alcoolémie mesurés ont donné les résultats suivants :

Taux d'alcoolémie T_i en g/l	[0;0,2[	[0,2;0,4[	[0,4;0,6[	[0,6;0,8[	[0,8;1[	[1;1,2[	[1,2;1,5[	[1,5;2[
Effectifs	42	78	52	19	4	2	2	1

1° En choisissant pour valeurs observées les centres de classes, calculer les valeurs approchées de la moyenne $\bar{x}$ et de l'écart type s du taux d'alcoolémie de cet échantillon.

2° Un contrôle est considéré comme positif si le taux d'alcoolémie mesuré est supérieur ou égal à 0,8 g/l. Calculer la fréquence (proportion) de contrôles positifs de cet échantillon.

3° On désigne par *F* la variable aléatoire qui, à tout échantillon de 200 contrôles, associe la proportion des contrôles positifs de cet échantillon. On assimile ces échantillons à des échantillons non exhaustifs et on admet que *F* suit la loi normale $N\left(p\ ,\ \sqrt{\frac{p(1-p)}{200}}\right)$ où $p = 0{,}02$ est la proportion de contrôles positifs dans la population de cette région.
Construire un test qui permette de décider, au seuil de 5 %, si l'échantillon précédent est représentatif de l'ensemble des contrôles de la région.

3 – Groupement D 2001 (Test unilatéral)

Un magicien prétend qu'il peut souvent deviner à distance la couleur d'une carte tirée au hasard d'un jeu de cartes bien battu et comportant des cartes de deux couleurs différentes en nombre égal.
On appelle *p* la probabilité que le magicien donne une réponse juste (succès) lors d'un tirage.

Si le magicien est un imposteur on a $p = \frac{1}{2}$, sinon $p > \frac{1}{2}$.

On appelle F la variable aléatoire qui, à tout échantillon de taille *n*, associe la fréquence des succès obtenus par le magicien au cours des *n* tirages d'une carte. On admet que *F* suit la loi normale de moyenne inconnue *p* et d'écart type $\sqrt{\frac{p(1-p)}{n}}$ et on choisit $n = 100$.

On construit un **test unilatéral** permettant de détecter, au risque de 5%, si le magicien est un imposteur.

On choisit comme hypothèse nulle $H_0 : p = \frac{1}{2}$, et comme hypothèse alternative $H_1 : p > \frac{1}{2}$.

1. Calculer, sous l'hypothèse H_0, le réel positif *h* tel que $P\left(F \leq \frac{1}{2} + h\right) = 0{,}95$.

2. Enoncer la règle de décision du test.

3. Sur un échantillon de taille 100, le magicien a obtenu 64 succès. Peut-on considérer, au risque de *5%*, que le magicien est un imposteur ?

4 – Etude et économie de la construction 1998

Une entreprise de bâtiment a constaté qu'un certain nombre de mitigeurs thermostatiques, posés par elle, avait un mauvais fonctionnement. Ce mauvais fonctionnement est dû à une pièce cylindrique montée sur cette catégorie de mitigeur. L'entreprise, pose 304 mitigeurs.
La variable aléatoire F qui, à tout échantillon de 304 pièces, associe la fréquence de défauts est une variable aléatoire qui suit la loi normale $N\ \ \left(p,\ \sqrt{\frac{p(1-p)}{304}}\right)$.

La production est suffisamment importante pour que l'on puisse assimiler tout échantillon de 304 pièces à 304 tirages aléatoires et indépendants.

a) Construire un **test unilatéral** permettant d'accepter ou de refuser l'hypothèse selon laquelle, au seuil de 5%, $p = 0{,}05$. Pour cela :
- on choisira pour hypothèse $H_0 : p = 0{,}05$ et pour hypothèse $H_1 : p > 0{,}05$;
- on déterminera le réel positif a tel que sous l'hypothèse H_0 : $P(F \leq a) = 0{,}95$;
- on déterminera la région critique au seuil de 5 % ;
- on énoncera la règle de décision.

b) On sait qu'il y a 18 défauts sur 304 pièces. Utiliser le test précédent pour conclure si, au seuil de 5 %, l'on accepte ou refuse l'affirmation : $p = 0{,}05$.

5 – Maintenance 1993

Une machine fabrique des tiges en grande série. La production étant importante on assimile tout prélèvement d'échantillon à un prélèvement avec remise.
La machine est supposée bien réglée quand la proportion de pièces acceptables est supérieure ou égale à 90 %. Pour contrôler le réglage de la machine on construit un test permettant de décider si, au seuil de 5 %, la machine est bien réglée et on prélève de temps en temps des échantillons aléatoires de 150 tiges.

a) Construction du **test unilatéral** :
Soit F la variable aléatoire qui à tout échantillon aléatoire de 150 tiges associe le pourcentage de tiges acceptables dans cet échantillon.
On choisit pour hypothèse nulle $H_0 : p = 90\ \%$,
et pour hypothèse alternative $H_1 : p < 90\ \%$.
Déterminer le nombre réel h tel que sous l'hypothèse H_0 $P(F > h) = 0{,}95$.
Enoncer la règle de décision de ce test.

b) Utilisation du test :
On prélève un échantillon aléatoire de 150 tiges. On trouve 22 tiges défectueuses.
Quel est le pourcentage de tiges acceptables de cet échantillon ?
En appliquant la règle de décision du test à cet échantillon non exhaustif, peut-on conclure, au seuil de 5 %, que la machine est bien réglée ?

TEST DE COMPARAISON DE DEUX FREQUENCES

6 – Comptabilité Gestion Nouvelle Calédonie 1994

Les nouveaux modèles de téléviseurs " 70 cm ", que va fabriquer une usine, sont de deux types : modèle (1) et modèle (2). Une enquête préalable à la fabrication, réalisée auprès de 400 ménages de la population S des ménages des "quartiers sud" de la ville V, indique qu'entre les deux modèles de téléviseurs, 63 % préfèrent le modèle (1). La même enquête, réalisée auprès de 500 ménages de la population N
des ménages des "quartiers nord" de la ville, indique que 67 % préfèrent le modèle (1).

On note F_S la variable aléatoire qui, à tout échantillon de 400 ménages pris au hasard et avec remise dans la population *S*, associe la proportion de ménages de cet échantillon qui préfèrent le modèle (1).
On note F_N la variable aléatoire qui, à tout échantillon de 500 ménages pris au hasard et avec remise dans la population *N*, associe la proportion de ménages de cet échantillon qui préfèrent le modèle (1).
On suppose que la loi de la variable $D = F_S - F_N$ est approximativement une loi normale de moyenne $p_S - p_N$ inconnue et d'écart type 0,032 (p_S et p_N étant les pourcentages de préférence dans les populations *S* et *N*).

Construire, puis mettre en oeuvre un test permettant de décider s'il y a une différence significative, au seuil 5 %, entre les pourcentages de préférence issus des deux échantillons de l'enquête préalable.

On peut ajouter pour aider les candidats:
- L'hypothèse H_0 est donnée par $p_S = p_N$, énoncer l'hypothèse alternative H_1.
- Déterminer l'intervalle $[-a; a]$ tel que, sous l'hypothèse H_0 , $P(-a \leq D \leq a) = 0{,}95$
- Enoncer la règle de décision du test.
- Utiliser ce test avec les deux échantillons de l'énoncé et conclure.

On peut également construire un test unilatéral de comparaison.

TEST D'UNE MOYENNE

7 – Groupement C 2001

Un client réceptionne une commande. Il prélève un échantillon de 125 billes choisies au hasard et avec remise dans le lot reçu et constate que le diamètre moyen est égal à 25,1.
Pour les billes fabriquées par l'entreprise, la variable aléatoire *X* qui prend pour valeurs leurs diamètres suit une loi normale d'écart type 0,44.
L'entreprise s'est engagée à ce que la moyenne des diamètres des billes fournies soit de 25.
Le client décide de construire un test bilatéral permettant de vérifier l'hypothèse selon laquelle le diamètre des billes du lot reçu est de 25.

1) Quelle est l'hypothèse nulle H_0 ? Quelle est l'hypothèse alternative H_1 ?

2) On désigne par $\overline{X}$ la variable aléatoire qui, à tout échantillon de 125 billes, prises au hasard et avec remise, associe la moyenne des diamètres obtenus.
a) Donner sous l'hypothèse nulle la loi de $\overline{X}$. En préciser les paramètres.
b) Déterminer le nombre *a* tel que $P(25 - a < \overline{X} < 25 + a) = 0{,}95$
c) Énoncer la règle de décision du test.

3) Au vu de l'échantillon, au risque de 5 %, que peut conclure le client sur le respect de l'engagement de l'entreprise ?

8 – Domotique 1998

On fabrique des pièces en série. Soit X la variable aléatoire qui, à toute pièce prise au hasard dans la production, associe sa cote exprimée en mm. On sait que X suit une loi normale d'écart type $\sigma = 2{,}4$ mm, mais on a des doutes sur l'espérance mathématique μ de X.
Afin de vérifier l'espérance mathématique de X, on prélève un échantillon de 50 pièces. On assimile tout échantillon de 50 pièces à un échantillon aléatoire non exhaustif. Sur le tableau suivant on trouve la distribution des mesures des cotes arrondies à l'entier le plus proche :

Cotes	147	148	149	150	151	152	153	154
Effectifs	2	3	5	10	9	9	8	4

Calculer la moyenne $\bar{x}$ de cet échantillon. On ne calculera pas son écart type.
On suppose que la variable aléatoire $\overline{X}$ qui à tout échantillon de 50 pièces prélevées au hasard associe la moyenne des cotes des pièces de cet échantillon suit la loi normale $\mathrm{N}\left(\mu ; \frac{2{,}4}{\sqrt{50}}\right)$.
On veut construire un test bilatéral pour permettre d'accepter ou de rejeter, au risque 5 %, l'hypothèse selon laquelle l'espérance mathématique μ de X est 150 mm.
On prend comme hypothèse nulle H_0 : "$\mu = 150$",
et comme hypothèse alternative H_1 : "$\mu \neq 150$".

a) Trouver un réel positif h tel que, sous l'hypothèse H_0 :
$P(150 - h \leq \overline{X} \leq 150 + h) = 0{,}95$
b) Enoncer la règle de décision.
c) Utiliser ce test avec l'échantillon des 50 pièces proposé dans cet énoncé et conclure.

9 – Systèmes constructifs bois et habitat 1999

Une machine fabrique des barres en grande série. On veut vérifier le bon réglage de la machine.

On appelle L la variable aléatoire qui prend pour valeur la longueur d'une barre. On admet que L suit une loi normale de moyenne μ et d'écart type 1.

Dans le cas où la machine est bien réglée, la moyenne μ est égale à 1000.
On appelle $\overline{L}$ la variable aléatoire qui prend pour valeur la moyenne des longueurs des 100 barres d'un échantillon aléatoire.

1° On construit un test d'hypothèse bilatéral permettant de vérifier le bon réglage de la machine au seuil de 2 %.
a) Donner l'hypothèse nulle H_0 et l'hypothèse alternative H_1 .
b) Sous l'hypothèse H_0, quelle est la loi de $\overline{L}$?
c) Déterminer la région critique et énoncer la règle de décision relative à ce test.

2° a) On a regroupé par classes les longueurs des barres d'un échantillon :

Longueur	[997 ; 999[	[999 ; 1001[	[1001 ; 1003[
Quantité	14	75	11

Déterminer une valeur approchée de la moyenne $\bar{l}$ de la longueur des barres prélevées, en supposant que dans chaque classe tous les éléments sont situés au centre.
b) Au vu de cet échantillon, que peut-on conclure quant au réglage de la machine ?

10 – Groupement C 1999

Un lycée achète son papier pour photocopieur à une entreprise. On appelle X la variable aléatoire qui à chaque feuille, prise au hasard, associe son épaisseur en microns. Le fabricant spécifie que X suit la loi normale de moyenne 110 et d'écart type 3.
Le lycée met à l'épreuve les affirmations du fabricant concernant la moyenne de la variable aléatoire X. On suppose que l'écart type est connu et égal à 3. Pour cela, il étudie un échantillon de 1000 feuilles prises au hasard dans une livraison. L'étude de l'épaisseur de ces feuilles donne, en microns, une moyenne de 109,9.

1° On effectue un test d'hypothèse bilatéral ; préciser quelle est l'hypothèse nulle H_0 et l'alternative H_1.

2° On désigne par $\overline{X}$ la variable aléatoire qui à tout échantillon de 1000 feuilles tirées au hasard et avec remise associe la moyenne des épaisseurs des feuilles de cet échantillon. Quelle est, sous l'hypothèse H_0, la loi de probabilité de $\overline{X}$?

3° Au vu de l'échantillon étudié, peut-on admettre que la moyenne est 110 ? Faire un test au seuil de 10 %.

11 – Comptabilité et Gestion 1999

Un fabricant de vêtements de sport et de loisirs commercialise directement une partie de sa production.
Le fabricant désire savoir si la campagne promotionnelle entreprise a permis de modifier le montant moyen des achats.
Il réalise une enquête auprès d'un échantillon de 50 client choisis au hasard et avec remise ; la dépense moyenne pour cet échantillon est de 597 F.
Soit $\overline{Y}$ la variable aléatoire qui, à tout échantillon de 50 clients pris au hasard et avec remise, associe moyenne de leurs achats.
On rappelle que $\overline{Y}$ suit la loi normale de moyenne μ et d'écart type $\frac{\sigma}{\sqrt{50}}$ (on prendra pour σ la valeur 195).

1ère rédaction
Construire *un **test bilatéral*** permettant d'accepter ou de refuser, au seuil de signification de 5 % l'hypothèse selon laquelle le montant moyen des achats a été modifié.

Pour répondre à cette question :
choisir pour hypothèse nulle H_0 : $\mu = 550$ et pour hypothèse alternative H_1 : $\mu \neq 550$,
déterminer la région critique au seuil de signification de 5%,
énoncer la règle de décision,
utiliser le test avec l'échantillon précédent et conclure.

2ème rédaction
Construire *un **test unilatéral*** permettant d'accepter ou de refuser, au seuil de signification de 5 % l'hypothèse selon laquelle le montant moyen des achats a été augmenté.
Pour répondre à cette question :
choisir pour hypothèse nulle H_0 : $\mu = 550$ et pour hypothèse alternative H_1 : $\mu > 550$,
déterminer le nombre réel positif h tel que $P(\overline{Y} \leq h) = 0{,}95$, en déduire la région critique au seuil de signification de 5%,
énoncer la règle de décision,
utiliser le test avec l'échantillon précédent et conclure.

Dans l'épreuve de l'examen on aurait dû préciser si le test est unilatéral ou bilatéral.

12 – Traitement des matériaux 1999

Un client commande un lot de pièces dont on lui annonce que la moyenne des épaisseurs de nickel déposé est 25 microns. Ce client veut vérifier cette affirmation et mesure les épaisseurs de nickel d'un échantillon de 30 pièces prélevées avec remise dans ce lot. Il obtient les résultats suivants :

Epaisseurs en microns	[22 ; 22,5[	[22,5 ; 23[	[23 ; 23,5[	[23,5 ; 24[	[24 ; 24,5[
Effectifs	1	1	2	3	5
Epaisseurs en microns	[24,5 ; 25[	[25 ; 25,5[	[25,5 ; 26[	[26 ; 26,5[	[26,5 ; 27[
Effectifs	6	4	4	2	2

1° Calculer la moyenne $\overline{x_e}$ et l'écart type σ_e, de cet échantillon à 10^{-2} près.
Les calculs intermédiaires ne sont pas demandés.

2° En déduire que l'écart type estimé du lot entier est, à 10^{-2} près, de 1,12.

3° Soit $\overline{X}$ la variable aléatoire qui, à tout échantillon de taille $n = 30$ prélevé au hasard et avec remise dans ce lot, associe l'épaisseur moyenne de nickel déposé sur les 30 pièces prélevées. On suppose que $\overline{X}$ suit la loi normale de paramètres μ et $\frac{\sigma}{\sqrt{n}}$ où μ et σ sont la moyenne et l'écart type des épaisseurs de nickel en microns déposé sur les pièces du lot.

1ère rédaction
a) On prend $\sigma = 1, 12$. Construire un **test bilatéral** permettant d'accepter ou de refuser ce lot, au seuil de 5%, vu le cahier des charges qui impose $\mu = 25$.
b) Doit-on, à partir de l'échantillon indiqué dans la question 1, accepter ou refuser ce lot en utilisant ce test ?

2ème rédaction

a) On prend σ = 1, 12. Construire un **test unilatéral** permettant d'accepter ou de refuser ce lot, au seuil de 5%, vu le cahier des charges qui impose $\mu > 25$.

b) Doit-on, à partir de l'échantillon indiqué dans la question 1, accepter ou refuser ce lot en utilisant ce test ?

Dans cette question, on aurait pu préciser si le test est unilatéral ou bilatéral.

13 – Comptabilité et Gestion 1990

Avant d'engager une campagne publicitaire, la direction de l'hypermarché vous demande de construire un test unilatéral qui, au vu des chiffres d'affaires journaliers des trente jours ouvrables suivant cette campagne, permettra de décider si, au seuil de signification 5%, la moyenne des chiffres d'affaires journaliers a augmenté, c'est-à-dire dépassé 1,5 million de francs, à la suite de cette campagne publicitaire.

1° *Construction du **test unilatéral** :*

On note μ la moyenne *inconnue* de la nouvelle population des chiffres d'affaires journaliers obtenus après la campagne publicitaire et on suppose que l'écart type de cette population est 0,3 million de francs.

Soit Z la variable aléatoire qui, à tout échantillon aléatoire, non exhaustif, de trente chiffres d'affaires journaliers de cette nouvelle population, associe la moyenne de ceux-ci. On suppose que Z suit la loi normale de moyenne μ et d'écart type $\frac{0{,}3}{\sqrt{30}}$

a) Choisir une hypothèse nulle H_0 et une hypothèse alternative H_1 pour ce test *unilatéral* .

b) Déterminer le nombre réel h tel que, sous l'hypothèse H_0, on ait : $P(Z \leq h) = 0{,}95$.

c) Enoncer la règle de décision de ce test.

2° *Utilisation de ce test :*

Les chiffres d'affaires journaliers pendant les trente jours ouvrables suivant la campagne publicitaire sont donnés par le tableau suivant :

chiffre d'affaires	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2	2,1	2,2	2,3
Nombre de jours	1	2	2	8	8	2	2	1	2	1	0	1

a) Calculer la moyenne des chiffres d'affaires journaliers pendant ces trente jours.

b) En appliquant la règle de décision du test à cet échantillon de trente chiffres d'affaires journaliers que l'on assimile à un échantillon aléatoire non exhaustif, peut-on conclure, au seuil de signification 5 % qu'à la suite de la campagne publicitaire la moyenne des chiffres d'affaires journaliers a dépassé 1,5 millions de francs ?

TEST DE COMPARAISON DE DEUX MOYENNES

14 – Chimiste 1999

Deux laboratoires A et B fabriquent des tubes à essai et les conditionnent dans des paquets. Tous les paquets contiennent le même nombre de tubes.

1. On note X_1 la variable aléatoire prenant pour valeur le nombre de tubes défectueux par paquet provenant de l'entreprise A. Sur un échantillon aléatoire de 49 paquets provenant du laboratoire A les nombres des tubes défectueux par paquet sont les suivants :

7	5	5	4	4	4	9	7	9	2	7	8	7	8	4
4	9	10	5	10	6	4	5	6	1	2	5	7	8	0
6	0	1	5	2	0	5	2	3	3	4	1	3	10	1
0	10	2	7											

Calculer une valeur approchée à 10^{-2} près de la moyenne m_1 et de l'écart type s_1 de cet échantillon. On admet dans la suite de cet exercice qu'une estimation ponctuelle $\hat{\mu}_1$ de la moyenne μ_1 de la variable aléatoire X_1 est 4,84 et qu'une estimation ponctuelle $\hat{\sigma}_1$ de l'écart type σ_1 de X_1 est 2,99.

2. On note X_2 la variable aléatoire prenant pour valeur le nombre de tubes défectueux par paquet provenant de l'entreprise B. Sur un échantillon aléatoire de 64 paquets provenant de l'entreprise B on a obtenu une moyenne m_2 de 3,88 tubes défectueux et un écart type σ_2 de 1,45.
En déduire une estimation ponctuelle $\hat{\mu}_2$ de la moyenne μ_2 de la variable aléatoire X_2 et une estimation ponctuelle $\hat{\sigma}_2$ de l'écart type σ_2 de X_2.

3. On se propose de construire un test d'hypothèse pour comparer les qualités de production des laboratoires A et B.
On note $\overline{X_1}$ la variable aléatoire prenant pour valeur le nombre moyen de tubes défectueux par paquet dans des échantillons aléatoires de 49 paquets de la production du laboratoire A.
On note $\overline{X_2}$ la variable aléatoire prenant pour valeur le nombre moyen de tubes défectueux par paquet dans des échantillons aléatoires de 64 paquets de la production du laboratoire B.

3.1 Le nombre d'observations étant important, on admet que les lois de probabilité de $\overline{X_1}$ et $\overline{X_2}$ peuvent être approchées par des lois normales.
Exprimer la moyenne et l'écart type de chacune de ces variables aléatoires en fonction de ceux de X_1 et de X_2. Dans toute la suite, on considère donc que $\overline{X_1}$ et $\overline{X_2}$ sont deux variables aléatoires indépendantes suivant une loi normale.

3.2 On note D la variable aléatoire telle que : $D = \overline{X_1} - \overline{X_2}$.
Quelle est la loi de probabilité de D ? Déterminer la moyenne et l'écart type de D. Justifier.

4. Dans cette question, on admet que D suit la loi normale $\mathcal{N}(\mu_1 - \mu_2 ; 0{,}46)$.
On pose pour hypothèse nulle $H_0 : \mu_1 = \mu_2$ et pour hypothèse alternative $H_1 : \mu_1 \neq \mu_2$.

4.1 Calculer, sous l'hypothèse H_0, les nombres h et k tels que
$P(-h < D < h) = 0{,}99$ et $P(-k < D < k) = 0{,}95$.

4.2 Enoncer la règle de décision relative à ce test lorsqu'on choisit un seuil de signification de 1 %, puis de 5 %.

4.3 Peut-on conclure après examen des échantillons donnés dans les questions 1 et 2 que la différence des moyennes observées est significative au seuil de risque de 1 % ? Au seuil de risque de 5 % ?

15 – Groupement D 1999 (Test de comparaison unilatéral !)

Une entreprise fabrique des pots de peinture.
On se propose d'étudier les variations de la quantité d'un certain produit A contenu dans chaque pot.
On a contrôlé le dosage du produit A à la sortie de deux chaînes de fabrication.
Deux échantillons de 100 pots ont été analysés ; l'un provient de la chaîne n° l, l'autre de la chaîne n° 2.
Le tableau suivant donne la répartition de l'échantillon de la chaîne n° l en fonction de la masse de produit A exprimée en grammes.

m (en g)	[100, 102[	[102, 104[	[104, 106[	[106, 108[	[108, 110[	[110, 112[	[112, 114[	[114, 116[
Effectifs	1	3	25	32	27	6	4	2

On donne des valeurs approchées de la moyenne m_2 et de l'écart type s_2 de l'échantillon fabriqué par la chaîne n° 2 : $m_2 = 107$ et $s_2 = 2$ (en grammes).

Dans les questions 1 et 2 les valeurs seront arrondies au dixième le plus proche.

l° En prenant les centres des classes, calculer une valeur approchée de la moyenne m_l et de l'écart type s_1 de l'échantillon issu de la chaîne n° l.

2° En considérant les résultats obtenus dans la première question, donner les estimations ponctuelles :
• des quantités moyennes μ_1 et μ_2 de produit A pour les productions de ces deux chaînes,
• des écarts types σ_1 et σ_2 correspondants.

3° On se propose de tester si la différence des moyennes observées dans les deux échantillons est due à des fluctuations d'échantillonnage ou si la chaîne de fabrication n° l produit des pots contenant davantage de produit A que la chaîne n° 2.
On note $\overline{X_1}$ la variable aléatoire qui à tout échantillon aléatoire de 100 pots provenant de la chaîne n° 1 associe la quantité moyenne de produit A dans cet échantillon.
On note $\overline{X_2}$ la variable aléatoire qui à tout échantillon aléatoire de 100 pots provenant de la chaîne n° 2 associe la quantité moyenne de produit A dans cet échantillon.
On admettra que :
• $\overline{X_1}$ suit une loi normale de paramètres μ_1 et $\frac{\sigma_1}{10}$;
• $\overline{X_2}$ suit une loi normale de paramètres μ_2 et $\frac{\sigma_2}{10}$;
• $\overline{X_1}$ et $\overline{X_2}$ sont des variables aléatoires indépendantes ;

• D = $\overline{X_1} - \overline{X_2}$ suit une loi normale.

On choisit l'hypothèse nulle H_0 : «$\mu_1 = \mu_2$» contre l'hypothèse alternative H_1 : «$\mu_1 > \mu_2$».

a) Calculer la variance de la variable aléatoire *D*. On appelle $\sigma(D)$ son écart type.
Vérifier que $\sigma(D) \approx 0{,}32$.
b) Calculer au centième le plus proche le réel *a* tel que, sous H_0, $P(D \leq a) = 0{,}99$.
c) Au vu des résultats observés, l'hypothèse nulle H_0 est-elle acceptée ou rejetée (au seuil de 1 %) ?

Corrigé des exercices d'épreuves de B.T.S.

1 – Analyses biologiques 99

a) Construisons un test bilatéral permettant, à la suite du prélèvement au hasard d'un échantillon de $n = 50$ sportifs, de tester au seuil de 10% l'hypothèse H_0 selon laquelle le pourcentage de sportifs contrôlés positivement est $p = 0,02$.

L'hypothèse alternative étant $H_1 : p \neq 0,02$.

On appelle F la variable aléatoire qui, à tout échantillon de 50, associe le pourcentage de sportifs contrôlés positivement.

On admet que sous l'hypothèse H_0, F suit la loi normale $\mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ où $p = 0,02$ et $n = 50$.

donc F suit $\mathcal{N}$ (0,02 ; 0,0198).

(***Remarque*** : cette approximation est discutable, dans la mesure où la condition $np > 15$ n'est pas réalisée).

Déterminons la région critique :

On pose $T = \frac{F - 0,02}{0,0198}$ alors T suit la loi normale centrée réduite.

Cherchons un réel a positif tel que :

$P(0,02 - a \leq F \leq 0,02 + a) = 0,9$,

$P(-\frac{a}{0,0198} \leq T \leq \frac{a}{0,0198}) = 0,9$

$2\,P(T \leq \frac{a}{0,0198}) - 1 = 0,9$,

$P(T \leq \frac{a}{0,0198}) = 0,95$.

Par lecture inverse de la table de la loi normale centrée réduite on a $\frac{a}{0,0198} = 1,645$

donc $a = 1,645 \times 0,0198 = 0,0326$.

Ainsi la région d'acceptation du test est

I = [0,02 – 0,0326 ; 0,02 + 0,0326] soit

I = [– 0,0124 ; 0,0526].

La région critique est donc $\mathbb{R}$ – I

Enoncé de la règle de décision:

On calcule dans un échantillon aléatoire, supposé non exhaustif, de taille 50, le pourcentage f d'individus sportifs contrôlés positivement.

Si $f \in$ I on accepte H_0 et l'échantillon observé est représentatif de l'ensemble de la population sportive au seuil de risque de 10%

Si $f \notin$ I on rejette H_0 et on accepte H_1. Dans ce cas l'échantillon observé n'est pas représentatif de l'ensemble de la population sportive au seuil de risque de 10%.

b) Application du test :

Dans l'échantillon E , 2 contrôles antidopage ont été déclarés positifs sur 50 donc $f = \frac{2}{50} = 0,04$, $f \in$ I

Par conséquent H_0 est acceptée et l'échantillon observé est représentatif de l'ensemble de la population sportive au risque de 10%.

2 - Biochimiste 94

1° $\bar{x} = 0,386$ $s = 0,249$ à 10^{-3} près.

2° $f = 0,045$.

3° Construction du test :

Choix de $H_0 : p = 0,02$;

Choix de $H_1 : p \neq 0,02$

Détermination de la région critique :

Sous H_0 on a $p = 0,02$, F suit la loi normale $\mathcal{N}\left(p, \sqrt{\frac{p(1p)}{200}}\right)$, F suit la loi $\mathcal{N}$ (0,02 ; 0,0099).

La variable aléatoire : $T = \frac{F - p}{0,0099}$ associée à F, suit la loi normale $\mathcal{N}$ (0,1). $P(p - a \leq F \leq p + a) =$ 0,95 équivaut à $P(-\frac{a}{0,0099} \leq T \leq \frac{a}{0,0099}) = 0,95$ et à

$2\pi(\frac{a}{0,0099}) - 1 = 0,95$ et à $\frac{a}{0,0099} = 1,96$,

$a = 0,0194$ à 10^{-4} près.

P(0,0006 ≤ F ≤ 0,0394) = 0,95.

Règle de décision :

On calcule dans un échantillon de contrôles de taux d'alcoolémie, supposé non exhaustif de taille 200, le pourcentage f de contrôles positifs.

si $f \in$ [0,0006 ; 0,0394] on accepte H_0 .

si $f \notin$ [00006 ; 0,0394] on rejette H_0 et on accepte H_1.

Utilisation du test :

On a trouvé $f = 0,045$, $f \notin$ [0,0006 ; 0,0394] on rejette H_0 et on accepte H_1.

On conclut, au seuil de signification 5 %, que l'échantillon n'est pas représentatif de l'ensemble des contrôles de la région.

3 – Groupement D 2001

Construction du test :

Choix de $H_0 : p = 0,5$;

Choix de $H_1 : p > 0,5$.

Détermination de la région critique :

Sous H_0 on a $p = 0,5$, F suit la loi normale N ($p, \sqrt{\frac{0,5 \times 0,5}{100}}$), F suit la loi N (0,5 ; 0,05).

La variable aléatoire : $T = \frac{F - 0,5}{0,05}$ associée à F, suit la loi normale N (0,1).

$P(F \leq 0,5 + h) = 0,95$ équivaut à $P(T \leq \frac{h}{0,05}) = 0,95$ et à $\Pi(\frac{h}{0,05}) = 0,95$ et à $\frac{h}{0,05} = 1,645$,

$h = 0,05 \times 1,645$, $h = 0,082$ à 10^{-3} près.

$P(F \leq 0,582) = 0,95$ à 10^{-3} près.

Règle de décision :
On calcule dans un échantillon de taille 100 le pourcentage f de succès.
si $f \leq 0,582$ on accepte H_0 .
si $f > 0,582$ on rejette H_0 et on accepte H_1.

Utilisation du test :

On a $f = \frac{64}{100}$, $f = 0,64$, $f > 0,582$ donc on rejette H_0 et on accepte H_1.
On conclut, au seuil de signification 5 %, que p est significativement supérieur à 0,5, on peut considérer, au risque de 5 %, que le magicien n'est pas un imposteur.

4 – Etudes et économie de la construction

Construction du test :
Choix de $H_0 : p = 0,05$;
Choix de $H_1 : p > 0,05$.

Détermination de la région critique :
Sous H_0 on a $p = 0,05$, F suit la loi normale
N ($p, \sqrt{\frac{p(1-p)}{304}}$), F suit la loi N (0,05 ; 0,0125).

La variable aléatoire : $T = \frac{F - p}{0,0125}$ associée à F, suit la loi normale N (0,1).

$P(F \leq a) = 0,95$ équivaut à $P(T \leq \frac{a - 0,05}{0,0125}) = 0,95$ et à $\pi(\frac{a - 0,05}{0,0125}) = 0,95$ et à $\frac{a - 0,05}{0,0125} = 1,645$,

$a = 0,05 + 0,0125 \times 1,645$, $a = 0,0705$ à 10^{-4} près.

$P(F \leq 0,07) = 0,95$ à 10^{-3} près.

Règle de décision :
On calcule dans un échantillon de 304 pièces le pourcentage f de défauts.
si $f \leq 0,07$ on accepte H_0 .
si $f > 0,07$ on rejette H_0 et on accepte H_1.

Utilisation du test :

On a trouvé $f = \frac{18}{304}$, $f \approx 0,059$, $f \leq 0,07$ donc on accepte H_0. On conclut, au seuil de signification 5 %, que l'on accepte l'affirmation $p = 0,05$.

5 - Maintenance 93

a) Construction du test unilatéral :
Hypothèse nulle $H_0 : p = 0,9$,
Hypothèse alternative $H_1 : p < 0,9$.
Détermination de la région critique :
Sous l'hypothèse H_0 , F suit la loi normale
N (0,9 ; $\sqrt{\frac{0,9 \times 0,1}{150}}$). La variable $T = \frac{F - 0,9}{\sqrt{\frac{0,9 \times 0,1}{150}}}$
suit la loi normale centrée réduite N (0 ; 1).
Donc sous H_0 : $P(F > h) = 0,95$ équivaut à
$P(T > \frac{h - 0,9}{\sqrt{\frac{0,9 \times 0,1}{150}}}) = 0,95$. Or $P(T > -1,645) = 0,95$,
donc $\frac{h - 0,9}{\sqrt{\frac{0,9 \times 0,1}{150}}} = -1,645$,

$h = 0,9 - 1,645 \sqrt{\frac{0,9 \times 0,1}{150}}$, $h \approx 0,8597$,

$P(F > 0,8597) = 0,95$
Règle de décision :
On prélève un échantillon aléatoire non exhaustif de 150 tiges, on calcule le pourcentage f de tiges acceptables de cet échantillon :
Si $f \geq$ 85,97 % on accepte H_0 , on rejette H_1.
Si $f <$ 85,97 % on rejette H_0 . on accepte H_1.

b) Utilisation du test :
Le pourcentage de tiges acceptables est :
$f = 85,33$ %
$f <$ 85,97 % on rejette H_0 on accepte H_1 .
Au seuil de 5%, on conclut que la machine n'est pas bien réglée.

6 - Comptabilité et gestion Nouvelle Calédonie 94

On sait que la variable aléatoire $D = F_S - F_N$ suit la loi normale N ($p_S - p_N$; 0,032).

Construction du test :
Choix de l'hypothèse nulle $H_0 : p_S = p_N$.
L'hypothèse alternative H_1 *est:* $p_S \neq p_N$.
Détermination de la région critique :

Sous H_0, D suit la loi normale $\mathcal{N}$ (0 ; 0,032) et $\frac{D}{0{,}032}$, suit la loi normale $\mathcal{N}$ (0, 1).

Quel que soit le nombre réel $t \geq 0$, on a :

$P(-t \leqslant \frac{D}{0{,}032} \leqslant t) = 2\Pi(t) - 1$.

La condition $2\Pi(t) - 1 = 0{,}95$ donne $t = 1{,}96$.
On en déduit : $P(-0{,}06272 \leqslant D \leqslant 0{,}06272) = 0{,}95$.

Énoncé de la règle de décision du test :
Dans la population S (respectivement N), on prélève un échantillon aléatoire non exhaustif de taille $n_S = 400$ (respectivement $n_N = 500$), on calcule la proportion f_S de ménages de cet échantillon qui préfèrent le modèle (1) (respectivement f_N).
Soit $d = f_S - f_N$.
Si $d \in [-0{,}06272 ; 0{,}06272]$, on accepte H_0.
Si $d \notin [-0{,}06272 ; 0{,}06272]$, on rejette H_0, on accepte H_1.

Utilisation du test :
Les échantillons prélevés ont donné $f_S = 0{,}63$ et $f_N = 0{,}67$.
On a donc $d = -0{,}04$, $d \in [-0{,}06272 ; 0{,}06272]$, on accepte H_0 et par suite, on accepte l'hypothèse, au seuil de signification 5 %, que les pourcentages dans deux quartiers n'ont pas de différence significative.

7 – Groupement C 2001

1° Choix de H_0 : $\mu = 25$, choix de H_1 : $\mu \neq 25$.
2° a) Sous H_0 , X suit la loi normale $\mathcal{N}$ (25 ; 0,44), $\overline{X}$ suit la loi normale de moyenne $\mu = 25$ et d'écart type $\frac{\sigma}{n} = \frac{0{,}44}{\sqrt{125}}$ donc $\overline{X}$ suit la loi normale $\mathcal{N}(25 ; \frac{0{,}44}{\sqrt{125}}) \approx \mathcal{N}(25 ; 0{,}039)$.

b) $T = \frac{\overline{X} - 25}{0{,}039}$ suit la loi normale centrée réduite $\mathcal{N}$ (0, 1), $P(25 - a \leq \overline{X} \leq 25 + a) = 0{,}95$ équivaut à $2\Pi(\frac{a}{0{,}039}) - 1 = 0{,}95$ et à $\Pi(\frac{a}{0{,}039}) = 0{,}975$,

$\frac{a}{0{,}039} = 1{,}96$, $a = 0{,}0764$ à 10^{-4} près.

$P(24{,}9236 \leq \overline{X} \leq 25{,}0764) = 0{,}95$.
Si H_0 est vraie on a 95% de chances de prélever un échantillon aléatoire dont la moyenne appartient à l'intervalle I = [24,9236 ; 25,0764] soit 5% de chance que cette moyenne soit à l'extérieur de I.
La région critique est l'extérieur de l'intervalle I = [24,9236 ; 25,0764].
c) Règle de décision :
On prélève un échantillon aléatoire, non exhaustif, de taille 150.
On calcule sa moyenne $\bar{x}$,
si $\bar{x} \in I$ on accepte H_0 et on rejette H_1 ;
si $\bar{x} \notin I$ on accepte H_1 et on rejette H_0.

3° Utilisation du test :
$\bar{x} = 25{,}1$ qui n'appartient pas à l'intervalle [24,9236 ; 25,0764].
On rejette l'hypothèse H_0 : $\mu = 25$, on accepte l'hypothèse H_1 : $\mu \neq 25$.
Le client peut conclure, au risque 5%, que l'entreprise ne respecte pas l'engagement .

8 – Domotique 1998

Avec la calculatrice on trouve $\bar{x} = 151$.
a) Sous H_0 , X suit la loi normale $\mathcal{N}$ (150 ; 2,4) , $\overline{X}$ suit la loi normale $\mathcal{N}(150 ; \frac{2{,}4}{\sqrt{50}})$.

$T = \frac{\overline{X} - 150}{2{,}4 / \sqrt{50}}$ suit la loi normale centrée réduite $\mathcal{N}$ (0, 1), $P(150 - h \leq \overline{X} \leq 150 + h) = 0{,}95$ équivaut à $2\,\pi(\frac{h\sqrt{50}}{2{,}4}) - 1 = 0{,}95$ et à $\pi(\frac{h\sqrt{50}}{2{,}4}) = 0{,}975$, $\frac{h\sqrt{50}}{2{,}4} = 1{,}96$, $h = \frac{1{,}96 \times 2{,}4}{\sqrt{50}}$,
$h = 0{,}665$ à 10^{-3} près.

b) $P(149{,}335 \leq \overline{X} \leq 150{,}665) = 0{,}95$.
Si H_0 est vraie on a 95% de chances de prélever un échantillon aléatoire dont la moyenne appartient à l'intervalle I = [149,335 ; 150,665] soit 5% de chance que cette moyenne soit à l'extérieur de I.
La région critique est l'extérieur de l'intervalle I = [149,335 ; 150,665].
On prélève un échantillon aléatoire, non exhaustif, de taille 50.
On calcule sa moyenne $\bar{x}$,
si $\bar{x} \in I$ on accepte H_0 et on rejette H_1 ;
si $\bar{x} \notin I$ on accepte H_1 et on rejette H_0.
c) Utilisation du test :
$\bar{x} = 151$, $\bar{x} \notin [149{,}335 ; 150{,}665]$ on rejette H_0.
On conclut, au seuil de 5%, que la moyenne des cotes des pièces de la production n'est pas 150 mm.

9 – SCBH 99

l° a) L'hypothèse nulle H_0 est : $\mu = 1000$ et l'alternative H_1 est $\mu \neq 1000$.

b) Sous l'hypothèse H_0 : μ = 1000, la variable aléatoire $\overline{L}$ suit la loi normale N (1000, $\frac{1}{\sqrt{100}}$).

c) $\overline{L}$ suit la loi normale N (1000 ; 0,1) donc la variable aléatoire $U = \frac{\overline{L} - 1000}{0,1}$ suit la loi normale centrée, réduite N (0, 1).

$P(1000 - h \leq \overline{L} \leq 1000 + h) = P(\frac{-h}{0,1} \leq U \leq \frac{h}{0,1})$,

$P(1000 - h \leq \overline{L} \leq 1000 + h) = 2\,\Pi(\frac{h}{0,1}) - 1$,

Donc $P(1000 - h \leq \overline{L} \leq 1000 + h) = 0,98$ si et seulement si $2\,\Pi(\frac{h}{0,1}) - 1 = 0,98$, qui est équivalent à $2\,\Pi(\frac{h}{0,1}) = 1,98$, $\Pi(\frac{h}{0,1}) = 0,99$,

Dans le formulaire on trouve $\frac{h}{0,1} \approx 2,32$,

P(999,768 ≤ $\overline{L}$ ≤ 1000,232) = 0,98.
La région critique est l'extérieur de l'intervalle :
I = [999,768 ; 1000,232].
Règle de décision : On prélève au hasard et avec remise un échantillon de 100 barres et on calcule la moyenne $\overline{l}$ des diamètres des épaisseurs des feuilles de cet échantillon.
Si $\overline{l}$ appartient à l'intervalle [999,768 ; 1000,232], on accepte H_0 au seuil de 2 %.
Sinon on rejette H_0 et on accepte H_1 à ce même seuil.
2° Pour l'échantillon de l'énoncé $\overline{l}$ = 999,94 $\overline{l}$ est compris entre 999,768 et 1000,232, au seuil de 2 %, on accepte H_0, on rejette H_1 : on conclut, au seuil de 2 %, que la machine est bien réglée.

10 – Groupement C 99

l° L'hypothèse nulle H_0 est : μ = 110 et l'alternative H_1 est $\mu \neq 110$.

2° X suit la loi normale N (μ, 3) ; la variable aléatoire $\overline{X}$ suit la loi normale N (μ, $\frac{3}{\sqrt{1000}}$) où μ est la moyenne inconnue des épaisseurs des feuilles.
Sous l'hypothèse H_0 : μ = 110, la variable aléatoire $\overline{X}$ suit la loi normale N (110, $\frac{3}{\sqrt{1000}}$).

3° $\overline{X}$ suit la loi normale N (110, $\frac{3}{\sqrt{1000}}$) donc la variable aléatoire $U = \frac{\overline{X} - 110}{\frac{3}{\sqrt{1000}}}$ suit la loi normale centrée, réduite N (0, 1). $P(110 - h \leq \overline{X} \leq 110 + h) = P(\frac{-h\sqrt{1000}}{3} \leq \frac{\overline{X} - 110}{3}\sqrt{1000} \leq \frac{h\sqrt{1000}}{3})$,

$P(110 - h \leq \overline{X} \leq 110 + h) =$
$P(\frac{-h\sqrt{1000}}{3} \leq U \leq \frac{h\sqrt{1000}}{3})$,

$P(110 - h \leq \overline{X} \leq 110 + h) = 2\,\Pi(\frac{h}{3}\sqrt{1000}) - 1$

$P(110 - h \leq \overline{X} \leq 110 + h) = 0,90$ si et seulement si $2\,\Pi(\frac{h}{3}\sqrt{1000}) - 1 = 0,90$ qui est équivalent à

$\Pi(\frac{h}{3}\sqrt{1000}) = 0,95$, $\Pi(1,645) = 0,95$ donc $\frac{h}{3}\sqrt{1000} = 1,645$, h = 0, 15606.

h = 0, 156 est le nombre réel positif tel que
$P(110 - h \leq \overline{X} \leq 110 + h) = 0,90$.
La région d'acceptation est I = [109,844 ; 110,156].
Règle de décision :
On prélève au hasard et avec remise un échantillon de 1000 feuilles et on calcule la moyenne $\overline{x}$ des diamètres des épaisseurs des feuilles de cet échantillon.
Si $\overline{x}$ appartient à l'intervalle [109,844 ; 110,156], on accepte H_0 au seuil de 10 %.
Sinon on rejette H_0 et on accepte H_1 à ce même seuil.
Pour l'échantillon observé $\overline{x}$ = 109,9 est compris entre 109,844 et 110,156 on accepte H_0 au seuil de 10 % : on conclut, au seuil de 10 %, que la moyenne μ des épaisseurs des feuilles est égale à 110 microns.

11 – Comptabilité et gestion 99

1ère rédaction

a) L'hypothèse nulle H_0 est : μ = 550 et l'alternative H_1 est $\mu \neq 550$.

b) Sous l'hypothèse H_0 : μ = 550, la variable aléatoire $\overline{Y}$ suit la loi normale N (550, $\frac{195}{\sqrt{50}}$).

La variable aléatoire $U = \dfrac{\overline{Y} - 550}{\dfrac{195}{\sqrt{50}}}$ suit la loi normale $\mathcal{N}$ (0, 1). $P(550 - h \leq \overline{Y} \leq 550 + h) = P(\dfrac{-h}{195}\sqrt{50} \leq U \leq \dfrac{h}{195}\sqrt{50})$,

$P(550 - h \leq \overline{Y} \leq 550 + h) = 2\,\Pi(\dfrac{h}{195}\sqrt{50}) - 1$,

$P(550 - h \leq \overline{Y} \leq 550 + h) = 0{,}95$ si et seulement si $2\,\Pi(\dfrac{h}{195}\sqrt{50}) - 1 = 0{,}95$ qui est équivalent à

$\dfrac{h}{195}\sqrt{50} = 1{,}96$, $h \approx 54{,}051$.

$P(495{,}949 \leq \overline{Y} \leq 604{,}051) = 0{,}95$.
La région d'acceptation est I = [496 ; 604].

Règle de décision : On prélève au hasard et avec remise un échantillon de 50 clients et on calcule la moyenne $\overline{x}$ des achats des clients de cet échantillon.
Si $\overline{x}$ appartient à [496 ; 604] on accepte H_0 au seuil de 5 %. Sinon on rejette H_0 et on accepte H_1 à ce même seuil.
Pour l'échantillon observé $\overline{x}$ = 597 est compris entre 496 et 604 on accepte H_0 au seuil de 5 % : on conclut, au seuil de 5 %, que la moyenne μ des dépenses est égale à 550 F.

2ème rédaction
a) L'hypothèse nulle H_0 est : $\mu = 550$ et l'alternative H_1 est $\mu \geq 550$.
b) Sous l'hypothèse H_0 : $\mu = 550$, la variable aléatoire $\overline{Y}$ suit la loi normale $\mathcal{N}$ (550, $\dfrac{195}{\sqrt{50}}$).

La variable aléatoire $U = \dfrac{\overline{Y} - 550}{\dfrac{195}{\sqrt{50}}}$ suit la loi normale centrée réduite $\mathcal{N}$ (0, 1).

$P(\overline{Y} \leq 550 + h) = P(U \leq \dfrac{h}{195}\sqrt{50})$,

$P(\overline{Y} \leq 550 + h) = \Pi(\dfrac{h}{195}\sqrt{50})$,

$P(550 - h \leq \overline{Y} \leq 550 + h) = 0{,}95$ si et seulement si $2\,\Pi(\dfrac{h}{195}\sqrt{50}) - 1 = 0{,}95$ qui est équivalent à

$\dfrac{h}{195}\sqrt{50} = 1{,}645$, $h \approx 45{,}364$.

$P(\overline{Y} \leq 595{,}364) = 0{,}95$.
La région critique est l'ensemble des valeurs inférieures à 595,364.

Règle de décision : On prélève au hasard et avec remise un échantillon de 50 clients et on calcule la moyenne $\overline{x}$ des achats des clients de cet échantillon.
Si $\overline{x} \leq 595{,}364$ on accepte H_0 au seuil de 5 %. Sinon on rejette H_0 et on accepte H_1 à ce même seuil.
Pour l'échantillon observé $\overline{x}$ = 597 est supérieur à 595,364 on rejette H_0 au seuil de 5 % : on conclut, au seuil de 5 %, que la moyenne μ des dépenses est égale à 550 F.

12 – Traitement des matériaux 99

Première rédaction :
a) Construction du test :
Choix de H_0 : $\mu = 25$;
choix de H_1 : $\mu \neq 25$.
Sous H_0, X suit la loi normale $\mathcal{N}$ (25 ; 1,12), $\overline{X}$ suit la loi normale $\mathcal{N}$ (25 ; $\dfrac{1{,}12}{\sqrt{30}}$), $T = \dfrac{\overline{X} - 25}{1{,}12/\sqrt{30}}$ suit la loi normale centrée réduite $\mathcal{N}$ (0, 1).
$P(25 - h \leq \overline{X} \leq 25 + h) = 0{,}95$ équivaut à
$P(-\dfrac{h\sqrt{30}}{1{,}12} \leq T \leq \dfrac{h\sqrt{30}}{1{,}12})$ et à

$2\,\Pi(\dfrac{h\sqrt{30}}{1{,}12}) - 1 = 0{,}95$,

et à $\Pi(\dfrac{h\sqrt{30}}{1{,}12}) = 0{,}975$, $\dfrac{h\sqrt{30}}{1{,}12} = 1{,}96$,

$h = \dfrac{1{,}96 \times 1{,}12}{\sqrt{30}}$, $h = 0{,}4008$ à 10^{-4} près.

$P(24{,}60 \leq \overline{X} \leq 25{,}40) = 0{,}95$ à 10^{-2} près.
Si H_0 est vraie on a 95% de chances de prélever un échantillon aléatoire dont la moyenne appartient à l'intervalle I = [24,60 ; 25,40] soit 5% de chance que cette moyenne soit à l'extérieur de I.
La région critique est l'extérieur de l'intervalle I = [24,60 ; 25,40].

Règle de décision :
On prélève un échantillon aléatoire, non exhaustif, de taille 30.
On calcule sa moyenne $\overline{x}$,
si $\overline{x} \in$ I on accepte H_0 et on rejette H_1 ;
si $\overline{x} \notin$ I on accepte H_1 et on rejette H_0.

b) Utilisation du test :
$\overline{x}$ = 24,75, $\overline{x} \in$ [24,60 ; 25,40] on accepte H_0. On accepte le lot, au seuil de 5%.

Deuxième rédaction :
a) Construction du test :

Choix de H_0 : $\mu = 25$;
choix de H_1 : $\mu < 25$.
Sous H_0 , X suit la loi normale $\mathcal{N}$ (25 ; 1,12) , $\overline{X}$ suit la loi normale $\mathcal{N}$ (25 ; $\frac{1,12}{\sqrt{30}}$), $T = \frac{\overline{X} - 25}{1,12 / \sqrt{30}}$
suit la loi normale centrée réduite $\mathcal{N}$ (0, 1).
$P(\overline{X} > 25 - h) = 0,95$ équivaut à
$P(T > -\frac{h\sqrt{30}}{1,12}) = 0,95$ et à $\Pi(\frac{h\sqrt{30}}{1,12}) = 0,95$
et à $\frac{h\sqrt{30}}{1,12} = 1,645$, $h = \frac{1,645 \times 1,12}{\sqrt{30}}$,
$h = 0,03364$ à 10^{-4} près, $P(\overline{X} > 24,66) = 0,95$.
Si H_0 est vraie on a 95% de chances de prélever un échantillon aléatoire dont la moyenne soit supérieure à 24,66 soit 5% de chance que cette moyenne soit inférieure à 24,66 l'extérieur de I.
La région critique est l'extérieur de l'intervalle I = [24,60 ; 25,40].
Régle de décision :
On prélève un échantillon aléatoire, non exhaustif, de taille 30.
On calcule sa moyenne $\bar{x}$,
si $\bar{x} \geq 24,66$ on accepte H_0 et on rejette H_1 ;
si $\bar{x} < 24,66$ on accepte H_1 et on rejette H_0.
b) Utilisation du test :
$\bar{x} = 24,75$, $\bar{x} \geq 24,66$ on accepte H_0. On accepte le lot, au seuil de 5%.

13 - Comptabilité et gestion 90

1° a) *Choix de H_0* : la moyenne des chiffres d'affaires journaliers de l'hypermarché après la campagne publicitaire est μ = 1,5 (million de francs).
Choix de H_1 : $\mu > 1,5$.

b) *Détermination de la région critique* :
Sous l'hypothèse H_0 , on a $\mu = 1,5$, donc Z suit la loi normale $\mathcal{N}$ (1,5 ; 0,3) et la variable aléatoire centrée réduite T associée à Z définie par :
$T = \frac{\sqrt{30}}{0,3}(Z - 1,5)$ suit la loi $\mathcal{N}$ (1, 0).
Par suite, $P(Z \leq h) = 0,95$ équivaut successivement à
$P(T \leq \frac{\sqrt{30}}{0,3}(h - 1,5)) = 0,95$;
$\pi(\frac{\sqrt{30}}{0,3}(h - 1,5)) = 0,95$; $\frac{\sqrt{30}}{0,3}(h - 1,5) = 1,645$,
d'où $h = 1,590$.

c) On prélève un échantillon non exhaustif de taille 30 dans la population des chiffres d'affaires journaliers obtenus après la campagne publicitaire. On calcule la moyenne μ de cet échantillon.
Si $\mu \leq 1,590$ on accepte H_0 , et on rejette H_0.
Si $\mu > 1,590$ on rejette H_0 , et on accepte H_1.

2° a) $\mu \approx 1,623$.
b) $\mu > 1,590$ on rejette H_0 , et on accepte H_1.
On conclut, au seuil de signification 5%, qu'à la suite de la campagne publicitaire la moyenne des chiffres d'affaires journaliers a augmenté, c'est-à-dire dépassé 1,5 million de francs.

14 - Chimiste 99

1. $m_1 = 4,84$; $s_1 = 2,96$ à 10^{-2} près.
2. $m_2 = 3,88$; $s_2 = 1,45$ alors μ_2 est estimé par $\hat{\mu}_2 = 3,88$ et σ_2 par $\hat{\sigma}_2 \approx 1,46$.

3.1 $E(\overline{X_1}) = \mu_1$ et $E(\overline{X_2}) = \mu_2$;
$\sigma(\overline{X_1}) = \frac{\sigma_1}{\sqrt{49}}$, $\sigma(\overline{X_1}) = \frac{\sigma_1}{7}$ et
$\sigma(\overline{X_2}) = \frac{\sigma_2}{\sqrt{64}}$, $\sigma(\overline{X_2}) = \frac{\sigma_2}{8}$
3.2 $\overline{X_1}$ et $\overline{X_2}$ sont deux variables aléatoires indépendantes suivant une loi normale donc $D = \overline{X_1} - \overline{X_2}$ suit la loi normale de moyenne inconnue $\mu_1 - \mu_2$ inconnue et d'écart type $\sqrt{\frac{\sigma_1^2}{49} + \frac{\sigma_2^2}{64}}$.
4.1 *Construction du test :*
H_0 : $\mu_1 = \mu_2$ ou $\mu_1 - \mu_2 = 0$;
H_1 : $\mu_1 \neq \mu_2$.
Détermination de la région critique :
Sous H_0, D suit la loi normale $\mathcal{N}$ (0 ; 0,46) .
La variable $T = \frac{D}{0,46}$ suit la loi normale $\mathcal{N}$ (0, 1).
$P(-h < D < h) = 0,99$ équivaut à
$P(\frac{h}{0,46} < T < \frac{h}{0,46}) = 0,99$; $\frac{h}{0,46} = 2,58$
soit $h \approx 1,19$ et $P(-1,19 < D < 1,19) = 0,99$.
L'extérieur de l'intervalle I = [- 1,19 ; 1,19] est la région critique du test au risque de 1 %.
En procédant de même au seuil 5%, $\frac{k}{0,46} = 1,96$, $k \approx 0,90$ et $P(-0,90 < D < 0,90) \approx 0,95$.
L'extérieur de l'intervalle I = [- 0,90 ; 0,90] est la région critique du test au risque de 5 %.
4.2 *Règle de décision :*
On prélève deux échantillons aléatoires non exhaustifs on calcule leurs moyennes m_1 , m_2 et $m_1 - m_2$
Si $m_1 - m_2 \in$ I on accepte H_1 .
Si $m_1 - m_2 \notin$ I on accepte H_1 on rejette H_0.
4.3 *Utilisation du test :* $m_1 - m_2 = 0,96$

Au seuil de 1% on a 0,96 $\in$ [– 1,19 ; 1,19] on accepte H_0 .
Au seuil de 5%, 0,96 $\notin$ [– 0,9 ; 0,9] on rejette H_0 et on accepte H_1.
On conclut, au seuil de signification 1% qu'il n'y a pas de différence significative entre les moyennes, mais au seuil de 5% on ne peut pas accepter cette hypothèse.

15 – Groupement D 99

l° Avec une calculatrice on obtient, en grammes
$m_1 = 107,5$ et $s_1 = 2,5$
Pour effectuer les calculs on a supposé que, dans chacune des classes, tous les éléments étaient placés au centre. Les nombres ainsi obtenus ne sont donc que des valeurs approchées de m_1 *et* s_1.

2° a) Pour estimation de la moyenne μ de la population on prend la moyenne m de l'échantillon, donc μ_1 et μ_2 sont estimés respectivement par 107,5 et 107.
b) Pour estimation ponctuelle de l'écart type σ de la population on prend $s\sqrt{\frac{n}{n-1}}$ donc σ_1 est estimé par $2,5\sqrt{\frac{100}{99}}$, $\sigma_1 \approx 2,5$ et σ_2 est estimé par $2\sqrt{\frac{100}{99}}$, $\sigma_2 \approx 2$.

3° a) $V(D) = V(\overline{X_1}) + V(\overline{X_2})$,
$V(D) \approx 0,25^2 + 0,2^2$, $V(D) \approx 0,1025$.
D'où $\sigma(D) = \sqrt{V(D)}$, $\sigma(D) \approx 0,32$.
b) Sous l'hypothèse H_0, D suit approximativement la loi normale $\mathcal{N}$ (0 ; 0,32) donc la variable aléatoire $U = \frac{D}{0,32}$ suit approximativement la loi normale centrée, réduite $\mathcal{N}$ (0, 1).

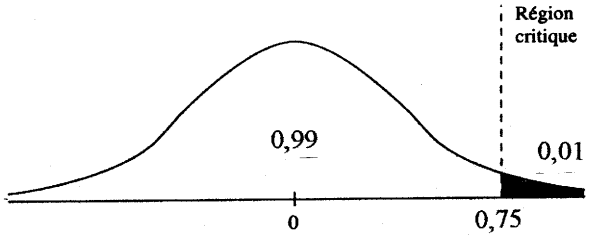
$P(D \leq a) = 0,99$ équivaut à $P(U \leq \frac{a}{0,32}) = 0,99$,

c'est à dire $\Pi(\frac{a}{0,32}) = 0,99$,

d'après la table de la loi normale $\mathcal{N}$ (0, 1),

$\frac{a}{0,32} \approx 2,33$, $a \approx 0,75$. $P(D \leq 0,75) \approx 0,99$.

c) On prélève au hasard et avec remise un échantillon de 100 pots provenant de la chaîne n° 1 et on détermine la quantité moyenne m_1 de produit A, en grammes.
On fait de même avec la chaîne n° 2, on obtient une moyenne m_2 , on calcule $d = m_1 - m_2$.
Règle de décision :
Si $d > 0,75$ on rejette H_0, on accepte H_1 au seuil de 1 %.
Si $d \leq 0,75$ on accepte H_0 au seuil de 1 %.



Pour les deux échantillons observés, $m_1 = 107,5$ et $m_2 = 107$
donc $m_1 - m_2 = 0,5$, $d = 0,5$.
$d \leq 0,75$ on accepte H_0 au seuil 1 %.
Au seuil de 1 %, la fabrication n° 1 ne produit pas des pots contenant davantage de produit A que la chaîne n° 2.

Supplément à la séance n°3

1 – Retour à Aamjiwnaang

Sources : Science et Vie février 2006 – Environnemental Health Perspectives octobre 2005 (article en ligne).

Les données proviennent d'une étude effectuée au Canada et montrant une différence (très) significative du sex-ratio à la naissance (déficit de garçons) sur une population exposée à une pollution chimique. Dans ce cas particulier, l'inquiétude provient du fait que bien que ces industries canadiennes respectent les normes, une exposition prolongée à de faibles doses de polluants puisse avoir un impact sanitaire mesurable.
Le « sex-ratio » est le rapport du nombre de garçons à celui des filles à la naissance. Il est habituellement de 105 garçons pour 100 filles.
Dans la réserve indienne d'Aamjiwnaag, située au Canada, il est né entre 1999 et 2003, n =132 enfants dont 46 garçons.
Question statistique : la fréquence des garçons observée à Aamjiwnaag pour la période 1999-2003 présente-t-elle une « différence significative » avec $p = 0{,}512$?

Construction d'un test unilatéral au risque de 1 %

On privilégie l'hypothèse selon laquelle le sex-ratio est « normal » c'est-à-dire que la probabilité d'avoir un garçon est $p = 0{,}512$ et on ne rejettera cette hypothèse que si l'observation est « significativement inférieure à p », c'est-à-dire si la fréquence obtenue sur l'échantillon est inférieure au seuil correspondant à 1 % des échantillons sous l'hypothèse $p = 0{,}512$.

• Choix des hypothèses :
H_0 : .....
H_1 : .....

• Détermination de la zone de rejet de H_0 :

Soit F la variable aléatoire correspondant à la fréquence des garçons sur un échantillon aléatoire de 132 naissances (on évitera dans ce contexte d 'évoquer des « tirages avec remise », mais il s'agit bien de ce modèle).
On suppose que F suit approximativement une loi normale.
Sous l'hypothèse H_0, indiquer les paramètres de cette loi (on arrondira à 10^{-3}).

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

Déterminer le réel h positif tel que, sous l'hypothèse H_0, $P(F > h) = 0{,}99$ (on arrondira à 10^{-3}).

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

- Règle de décision du test :

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

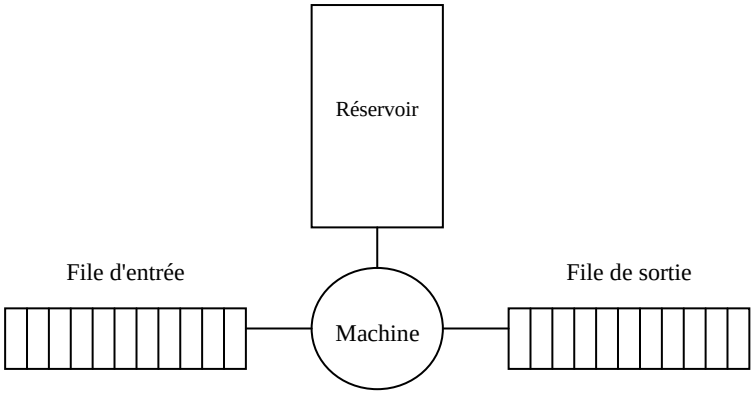
Mise en oeuvre du test

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

2 – Des sujets récents de BTS

Groupement B 2003 (moyenne – bilatéral)

Dans une usine du secteur de l'agroalimentaire, une machine à embouteiller est alimentée par un réservoir d'eau et par une file d'approvisionnement en bouteilles vides, selon le schéma ci-contre.
L'exercice consiste en une étude statistique du bon fonctionnement de ce système.



Test d'hypothèse

Pour contrôler le bon fonctionnement de la machine, on construit un test d'hypothèse bilatéral sur la moyenne, test qui sera mis en œuvre toutes les heures.
Pour une production d'une heure, la variable aléatoire Z qui, à toute bouteille prise au hasard dans cette production associe sa contenance en litres suit la loi normale de moyenne μ et d'écart type $\sigma = 0{,}01$. Dans cette question la moyenne μ est inconnue.
On désigne par $\overline{Z}$ la variable aléatoire qui, à chaque échantillon aléatoire de 100 bouteilles prélevé dans cette production d'une heure, associe la moyenne des contenances des bouteilles de cet échantillon (la production pendant une heure est assez importante pour que l'on puisse assimiler ces prélèvements à des tirages avec remise).
On considère que la machine est bien réglée lorsque $\mu = 1{,}5$.
L'hypothèse nulle est H_0 : "$\mu = 1{,}5$".
L'hypothèse alternative est H_1 : "$\mu \neq 1{,}5$".
Le seuil de signification du test est fixé à 0,05.

a) Justifier le fait que, sous l'hypothèse nulle H_0, $\overline{Z}$ suit la loi normale de moyenne 1,5 et d'écart type 0,001.
b) Sous l'hypothèse nulle H_0, déterminer le nombre réel h positif tel que :
$P(1{,}5 - h \leq \overline{Z} \leq 1{,}5 + h) = 0{,}95$.
c) Enoncer la règle de décision permettant d'utiliser ce test.
d) On prélève un échantillon de 100 bouteilles et on observe que, pour cet échantillon, la moyenne des contenances est $\bar{z} = 1{,}495$.
Peut-on, au seuil de 5%, conclure que la machine est bien réglée ?

Éléments de réponse

a) Sous H_0 la variable aléatoire $\overline{Y}$ suit la loi normale de moyenne 1,5 et d'écart type $\frac{0{,}01}{\sqrt{100}}$.

b) $\frac{h}{0{,}001} \approx 1{,}96$ d'où $h \approx 0{,}002$.

c) Zone d'acceptation de H_0 :
[1,498 ; 1,502].
d) La moyenne observée $\bar{y}$ n'appartient pas à l'intervalle d'acceptation.
H_0 est rejetée, au risque de 5%.

Groupement B 2005 (moyenne – bilatéral)

Une usine fabrique, en grande quantité, des rondelles d'acier pour la construction. Leur diamètre est exprimé en millimètres.

Dans cet exercice, sauf indication contraire, les résultats approchés sont à arrondir à 10^{-2}.

On se propose de construire un test d'hypothèse pour contrôler la moyenne μ de l'ensemble des diamètres, en millimètres, de rondelles constituant une grosse livraison à effectuer.
On note X_2 la variable aléatoire qui, à chaque rondelle prélevée au hasard dans la livraison, associe son diamètre.
La variable aléatoire X_2 suit la loi normale de moyenne inconnue μ et d'écart type $\sigma = 0{,}17$.
On désigne par $\overline{X_2}$ la variable aléatoire qui, à chaque échantillon aléatoire de 100 rondelles prélevé dans la livraison, associe la moyenne des diamètres de ces rondelles (la livraison est assez importante pour que l'on puisse assimiler ces prélèvements à des tirages avec remise).

L'hypothèse nulle est $H_0 : \mu = 90$. Dans ce cas la livraison est dite conforme pour le diamètre.
L'hypothèse alternative est $H_1 : \mu \neq 90$.
Le seuil de signification du test est fixé à 0,05.

1° Enoncer la règle de décision permettant d'utiliser ce test en admettant, sous l'hypothèse nulle H_0, le résultat suivant qui n'a pas à être démontré :
$P\,(89{,}967 \leq \overline{X_2} \leq 90{,}033) = 0{,}95$.

2° On prélève un échantillon de 100 rondelles dans la livraison et on observe que, pour cet échantillon, la moyenne des diamètres est $\bar{x} = 90{,}02$.
Peut-on, au seuil de risque de 5%, conclure que la livraison est conforme pour le diamètre ?

Éléments de réponse

1° Règle de décision :
- Soit $\bar{x}$ la moyenne calculée sur un échantillon de 100 rondelles prélevées au hasard.
- Si $\bar{x} \in [89{,}967 ; 90{,}033]$ la livraison est considérée comme conforme(au seuil de 5%).
- Sinon, la livraison est considérée comme non conforme (au

risque de 5%). 1 point

2° La livraison est considérée conforme, au seuil de 5%. 0,5 point

BTS Chimiste 2005 – fréquences – comparaison – unilatéral (Étude critique de l'énoncé)

Une entreprise fabrique des appareils de mesure qui doivent satisfaire à un cahier des charges.
L'entreprise met en place un nouveau dispositif sensé améliorer la fiabilité des appareils produits. Deux chaînes de fabrication sont mises en service : la chaîne n° 1, sans nouveau dispositif et la chaîne n° 2 avec le nouveau dispositif. Afin de tester l'hypothèse selon laquelle le nouveau dispositif *améliore* de manière significative la fiabilité des appareils produits, on a prélevé de manière aléatoire 200 appareils à la sortie de chacune des deux chaînes de fabrication.
Un pourcentage p_1 (resp. p_2) d'appareils issus de la chaîne n° 1 (resp. n° 2) ont fonctionné parfaitement pendant les trois premiers mois.

1.
a) Expliquer pourquoi on met en place un test unilatéral.
b) On prend pour hypothèse nulle H_0 : $p_1 = p_2$. Préciser l'hypothèse H_1 alternative qui va être opposée à l'hypothèse H_0.

On note F_1 (resp. F_2) la variable aléatoire qui à chaque échantillon de taille 200 provenant de la chaîne n° 1 (resp. n° 2) associe la fréquence f_1 (resp. f_2) d'appareils ayant parfaitement fonctionné pendant trois mois.
Sur les deux échantillons prélevés, on a obtenu des valeurs observées qui sont : $f_1 = 87\%$ et $f_2 = 93\%$.
On note $D = F_2 - F_1$.
Sous l'hypothèse nulle, les deux chaînes sont censées produire le même pourcentage p d'appareils conformes et la loi suivie par D (*celle que l'on adopte*) est la loi normale $\mathcal{N}\left(0 ; \sqrt{\frac{p(1-p)}{200} + \frac{p(1-p)}{200}}\right)$.
On prend $p = 0{,}9$ car $\left[p = \frac{f_1 + f_2}{2} \right]$.

2. Préciser les paramètres de la loi suivie par D.

3. Si α est le seuil de risque, on désigne par h_α le réel positif tel que : $P(D \leq h_\alpha) = 1 - \alpha$.
a) On suppose dans cette question que $\alpha = 0{,}01$.
Déterminer la valeur arrondie au centième de h_α.
Enoncer la règle de décision du test.
Conclure quant à l'efficacité présumée du nouveau dispositif au seuil de risque 0,01.
b) On suppose dans cette question que $\alpha = 0{,}05$.
Déterminer h_α.
Enoncer la règle de décision du test.
Conclure quant à l'efficacité présumée du nouveau dispositif au seuil de risque 0,05.

Le BTS Chimiste 2005 a proposé (exercice 1, partie B) un test de comparaison de deux fréquences qui nous a semblé cumuler plusieurs difficultés, selon une présentation pouvant prêter à confusion, en particulier sur la différence entre une fréquence observée sur un échantillon et une fréquence inconnue sur une population, ainsi que sur la signification du seuil de risque. Les professeurs présents ont d'ailleurs noté que seule la partie de l'énoncé concernant le test d'hypothèse était accompagnée d'un corrigé détaillé, voulant sans doute contribuer, de façon louable, à éclairer les correcteurs sur le sujet, sans toutefois y parvenir complètement.

Les difficultés de cet exercice sont les suivantes :

– Le test est **unilatéral**, situation *asymétrique* assez délicate.

– Il porte sur les fréquences, cas plus rare dans les sujets d'examen et qui nécessite une **estimation de l'écart type** de la variable aléatoire correspondant à la différence observée, qui est assez difficile, et *éludée* par l'énoncé,

– **Deux seuils de risque** sont proposés, l'un pour lequel l'hypothèse nulle est acceptée, l'autre pour lequel elle est refusée, sans que l'on ait bien *conscience des enjeux*, et le corrigé est assez troublant à cet égard.

Compte-tenu du choix qui a été fait, de nous présenter le cas le plus difficile au programme de ce BTS, une grande rigueur dans la rédaction aurait permis d'y voir plus clair. Nous faisons quelques suggestions en italiques.

Le test mis en place étudie la « fiabilité des appareils produits ». Il aurait été utile de **définir** dès le départ ce que l'on entendait par là, puisqu'il s'agit du caractère étudié. Par exemple :

« Un appareil est considéré comme fiable lorsqu'il a correctement fonctionné pendant les 3 premiers mois ».

C'est du moins ce que l'on comprend d'après la suite de l'énoncé.

Il faudrait ensuite définir clairement les deux populations dans lesquelles s'effectuera l'échantillonnage. L'énoncé affirme seulement :

« …, on a prélevé de manière aléatoire 200 appareils à la sortie de chacune des deux chaînes de fabrication.

Un pourcentage p_1 (resp. p_2) d'appareils issus de la chaîne n°1 (resp. n°2) ont fonctionné parfaitement pendant les trois premiers mois. »

A la lecture de cette phrase, on pourrait imaginer que les pourcentages p_1 et p_2 ont été observés, peut-être sur les échantillons. Il n'en est rien. Il faut comprendre que ces pourcentages p_1 et p_2 sont **inconnus** (ce qui n'est pas dit !) malgré le passé du « ont fonctionné ». Bien sûr, on ne voit pas pourquoi, si p_1 et p_2 étaient connus, on s'amuserait à les tester. Il est bien entendu essentiel de comprendre que ces « pourcentages » correspondent aux deux populations étudiées et qu'ils sont inconnus. Ils joueront le rôle de probabilités de « succès » dans l'échantillonnage. On aurait pu dire :

« On étudie les productions des deux chaînes pendant un mois [par exemple]. *On note p_1 (resp. p_2) le pourcentage, **inconnu**, des appareils considérés comme fiables dans la production de la chaîne n° 1 (resp. n° 2).*

On prélève au hasard (la production est suffisamment importante pour assimiler ce prélèvement à un tirage avec remise) 200 appareils dans la production de la chaîne n°1 durant ce mois, dont on étudie le bon fonctionnement pendant trois mois. On note f_1 la fréquence observée des appareils n'ayant pas connu de panne pendant trois mois sur cet échantillon. On procède de même avec la production de la chaîne n° 2, en notant f_2 la fréquence observée des appareils considérés comme fiables sur un échantillon de taille 200. »

On examine ensuite « la » loi de la variable aléatoire $D = F_2 - F_1$ sous l'hypothèse nulle. Le texte énonce là quelques certitudes alors que l'on ne va pas utiliser la « vraie » loi de D (même sous l'hypothèse nulle), fondée sur des lois binomiales avec un paramètre p inconnu, mais que l'on va procéder « au mieux ». (Le texte parle à cet endroit d'appareils « conformes », c'est une coquille, il s'agit d'appareils « fiables »).
Sous l'hypothèse nulle, on a $p_1 = p_2 = p$ et la variable aléatoire $D = F_2 - F_1$ suit approximativement (ce n'est pas dit dans l'énoncé) la loi normale de moyenne nulle et d'**écart type inconnu** (on oublie de le dire) $\sqrt{\frac{p(1-p)}{200} + \frac{p(1-p)}{200}}$, puisque Var ($D$) = Var ($F_1$) + Var ($F_2$). Reste, pour pouvoir construire le test, à estimer la valeur de cet écart type.
On ne peut pas écrire, comme c'est fait dans l'énoncé :

« On prend $p = 0{,}9$ car $\left[p = \frac{f_1 + f_2}{2} \right]$ ».

Même si H_0 est vraie, il n'y a bien sûr aucune raison d'observer des fréquences f_1 et f_2 telle que $p = \frac{f_1 + f_2}{2}$. L'écriture $p = \frac{f_1 + f_2}{2}$ entretient la **confusion** entre les fréquences f_1 et f_2 observées et la fréquence p inconnue, même sous l'hypothèse nulle.
Il serait préférable d'écrire :

« On estimera la valeur p inconnue par $0{,}9 = \frac{0{,}87 + 0{,}93}{2}$ et on supposera donc que D a pour écart type $\sqrt{\frac{0{,}9 \times 0{,}1}{200} + \frac{0{,}9 \times 0{,}1}{200}}$. »

Ce n'est qu'une supposition, et même sous H_0, on ne connaît pas l'écart type de D.

La dernière question de l'exercice (la question 3.) pose le problème du choix, a priori, du seuil de risque α. On fait faire à l'élève deux tests, l'un avec $\alpha = 0{,}01$ qui conduit à accepter H_0, l'autre avec $\alpha = 0{,}05$, qui conduit à rejeter H_0. Il est certes important de faire constater que selon le choix de α, la réponse du test peut-être différente, le professeur ayant la charge, en formation, d'en expliquer les enjeux. L'explication donnée par le corrigé dans le cas de l'acceptation de H_0 pour $\alpha = 0{,}01$ laisse cependant perplexe :

> « aucune raison de refuser H_0 le nouveau dispositif ne semble pas avoir amélioré la fiabilité des appareils. Mais le risque de se tromper est très faible ; les calculs étant effectués sous l'hypothèse H_0, on n'en doute pas a priori. »

Que veut-on dire par là ? Quand on accepte H_0 (comme ici), le « **risque de se tromper** » est inconnu, il n'est pas « très faible ». En fait, ce risque inconnu est d'autant plus élevé que α est, comme ici, petit. Ce qu'il faudrait dire, c'est qu'avec $\alpha = 0{,}01$, le risque de rejeter à tort H_0 est très faible. Ce qui veut dire qu'on tient a priori à conserver H_0, peut-être parce que le nouveau procédé de fabrication est très onéreux, et qu'on ne rejettera H_0 que si la différence observée est très significative. C'est peut-être ce qu'entend le corrigé quand il dit de l'hypothèse H_0 qu'on « n'en doute pas a priori ».

Groupement B 2007 (moyenne – bilatéral)

Une usine fabrique des ventilateurs en grande quantité. On s'intéresse à trois types de pièces : l'axe moteur, appelé pièce de type 1, les pales, appelées pièce de type 2 et le support, appelé pièce de type 3.
Une importante commande de pièces de type 3 est passée auprès d'un sous-traitant. La hauteur du support doit être de 400 millimètres.
On se propose de construire un test d'hypothèse bilatéral pour contrôler, au moment de la livraison, la moyenne μ de l'ensemble des hauteurs, en millimètres, des pièces de type 3.
On note Z la variable aléatoire qui, à chaque pièce de type 3 prélevée au hasard dans la livraison associe sa hauteur.
La variable aléatoire Z suit la loi normale de moyenne inconnue μ et d'écart type $\sigma = 5$.
On désigne par $\overline{Z}$ la variable aléatoire qui, à chaque échantillon aléatoire de 100 pièces de type 3 prélevé dans la livraison, associe la moyenne des hauteurs des pièces de cet échantillon. La livraison est assez importante pour que l'on puisse assimiler ces prélèvements à des tirages avec remise.
L'hypothèse nulle est $H_0 : \mu = 400$.
L'hypothèse alternative est $H_1 : \mu \neq 400$.
Le seuil de signification du test est fixé à 0,05.

1° Sous l'hypothèse H_0, on admet que la variable aléatoire Z suit la loi normale de moyenne 400 et d'écart type 0,5.
Déterminer sous cette hypothèse le nombre réel h positif tel que :
$P(400 - h \leq \overline{Z} \leq 400 + h) = 0{,}95$.

2° En déduire la règle de décision permettant d'utiliser ce test.

3° On prélève un échantillon aléatoire de 100 pièces dans la livraison reçue et on observe que, pour cet échantillon, la moyenne des hauteurs des pièces est $\bar{z} = 399{,}12$.
Peut-on, au seuil de 5 %, conclure que la livraison est conforme pour la hauteur ?

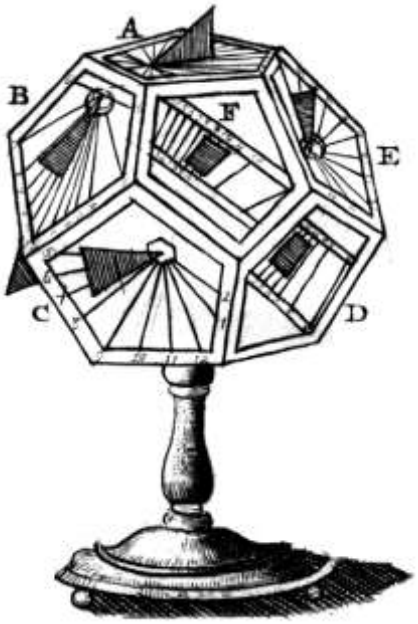
Éléments de réponse

C. 1° $h = \mathbf{0{,}98}$. 1 point

2° On prélève au hasard et avec remise, un échantillon de 100 pièces dans la livraison et on calcule la moyenne $\bar{z}$ des hauteurs des pièces de cet échantillon.
La règle de décision est :
- Si $\bar{z}$ appartient à l'intervalle [399,02 ; 400,98], on accepte H_0 au seuil de 0,05.
- Sinon on rejette H_0 et on accepte H_1. 1 point

3° $\bar{z} = 399{,}12$ appartient à l'intervalle d'acceptation.
On accepte l'hypothèse H_0 : la livraison est considérée comme conforme pour la hauteur. 0,5 point

Séance 4 :
FIABILITE
LOI EXPONENTIELLE
LOI DE WEIBULL



"Carpe diem."
Conseil épicurien.

La fiabilité est l'étude statistique des ***durées de vie*** (temps de bon fonctionnement) ou des défaillances. Le terme "fiabilité" est un néologisme introduit dans les années 1960 pour traduire le terme anglais "reliability". Elle a comme spécificités, d'une part une modélisation (des durées de vie) par des variables aléatoires positives (la loi normale n'est plus "reine"), d'autre part une interprétation et un vocabulaire particulier : la fonction de répartition est la fonction de "survie", ou fonction de "fiabilité", la "fonction de hasard" correspondant au "taux de mortalité" ou au "taux d'avarie" joue un rôle central.
Les mesures de durée de vie se sont longtemps limitées à l'étude de la durée de vie humaine, pour des applications démographiques et dans le domaine des assurances (assurances vies et rentes viagères). Ce n'est qu'avec une industrie standardisée, au XX^{e} siècle, que l'on s'est intéressé à d'autres types de durée de vie. Avant la production en série, chaque pièce était faite "sur mesure" et cette individualisation empêchait le traitement statistique. Cependant, le mot "fiabilité" n'est pas encore mentionné dans le "*Grand Larousse du XX^{e} siècle*" édité en 1930.
Au delà de la fiabilité industrielle apparaissent de nouvelles applications, notamment dans le domaine biomédical (survie de patients atteints de cancers…) ou économique (modélisation de la durée de vie des entreprises ou de la durée de séjour au chômage)[5].

[5] Renseignements donnés par Droesbeke, Fichet, Tassi dans *"Analyse statistique des durées de vie"* – Economica 1989.

Nous commencerons par une étude statistique de la durée de vie humaine (parce que nous sommes tous concernés). Ce biais culturel pour aborder la fiabilité n'est bien sûr pas celui que l'on suit avec les étudiants de B.T.S. pour lesquels ce module est au programme.

I - EXEMPLE D'INTRODUCTION : Etude statistique des tables de mortalité

1 - La première table de mortalité connue

En 1662, apparaît dans les *Observations naturelles et politiques sur les bulletins de mortalité de la ville de Londres*, la première table de mortalité jamais construite. La voici reproduite ci-contre (Source : *"Les mathématiques sociales"* - Dossier *Pour la Science* de juillet 1999 - article de Hervé Le Bras).

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

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

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

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

Considérons le second tableau correspondant aux survivants. Dans le vocabulaire de la fiabilité, il s'agit de la *fonction de fiabilité*, notée $t \mapsto R(t)$ (fiabilité se dit "*reliability*" en anglais) où t est le temps.

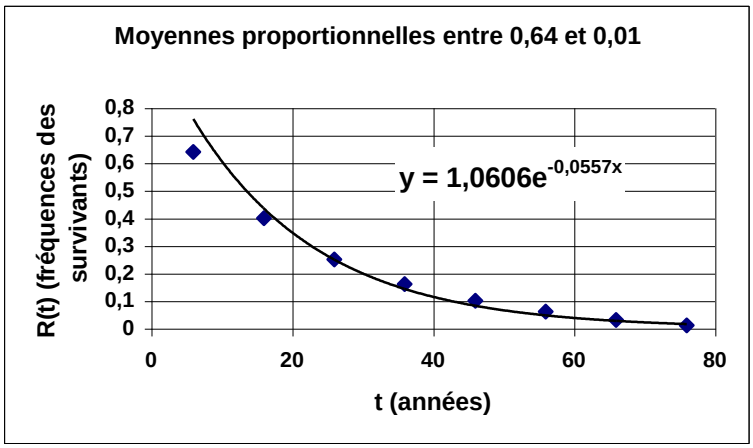
Il s'agit d'une *modélisation*. A partir de l'observation qu'au bout de 6 ans, il y a 64 % de survivants et 1 % au bout de 76 ans, on a recherché une *suite géométrique* pour les pourcentages de survivants à chaque décennie intermédiaire.

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

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

Hervé Le Bras précise que la notion de taux de mortalité n'apparaîtra que bien plus tard et analyse cela par la conception, que l'on avait à l'époque, de la mort (déterminée par Dieu, ou les astres...).

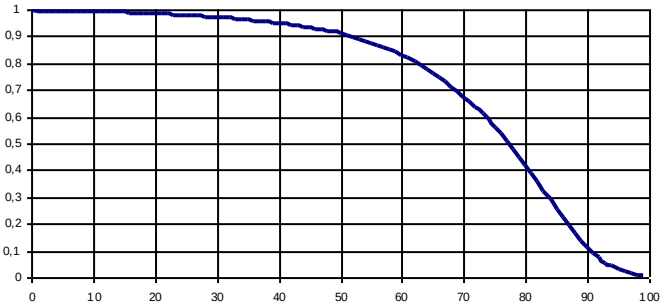
Compte-tenu de ce qu'était la mortalité infantile, on avait cependant écarté les 6 premières années de vie. Cela conduit à rechercher une fiabilité "de type exponentielle", correspondant au hasard total ("coups du sort").



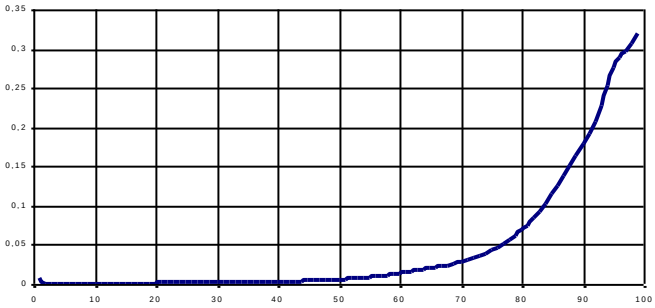
Voici ci-contre, sur ***Excel***, l'ajustement exponentiel correspondant à cet exemple.

2 - La mortalité en France selon le recensement de 1992

A partir du tableau de données (page suivante), on peut construire les courbes suivantes :



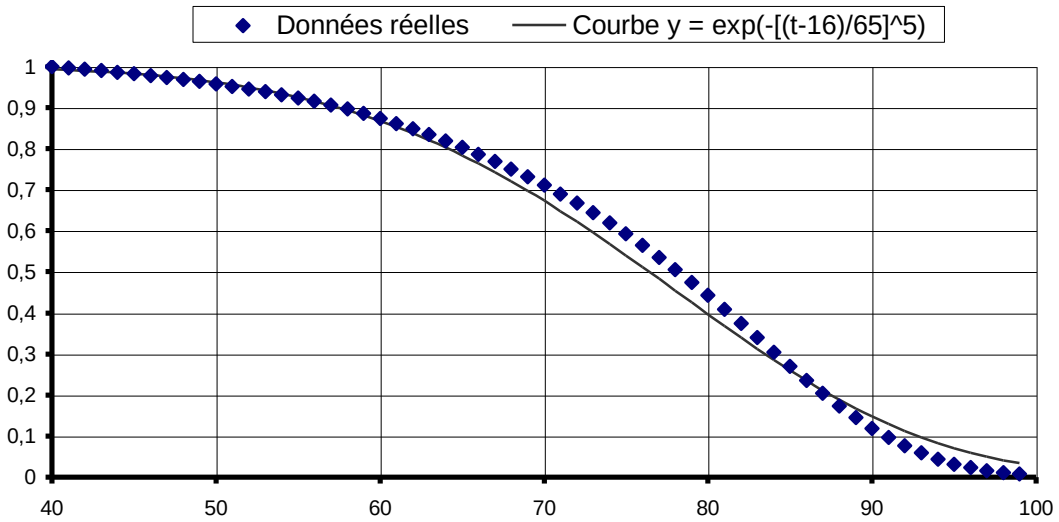
Fréquence des survivants $t \mapsto R\,(t)$



Taux de mortalité $t \mapsto \lambda(t)$

On constate que la courbe des survivants n'a pas la même allure que lorsque la mortalité est supposée constante. Sur le second graphique, le taux de mortalité, après une légère décroissance - mortalité infantile - est nettement croissant au delà de 40 ans ($\lambda(n)$ est le rapport du nombre de décès durant l'intervalle $[n\ ,\ n+1]$, au nombre de survivants à la date n : $\lambda(n) = \dfrac{R(n) - R(n+1)}{R(n)}$). Lorsque le taux de mortalité n'est pas constant, on cherchera à le modéliser sous forme d'une fonction puissance, ce qui conduira aux lois de *Weibull*.

Dans le cas des survivants ("fiabilité $R(t)$") après 40 ans, la loi de *Weibull* conduit au modèle suivant



age ti	survivants S(ti)	R(ti)	Taux de mortalité en %	age ti	survivants S(ti)	R(ti)	Taux de mortalité en %
0	100000	1		50	90889	0,90889	0,59
1	99264	0,99264	0,74	51	90322	0,90322	0,62
2	99204	0,99204	0,06	52	89717	0,89717	0,67
3	99166	0,99166	0,04	53	89075	0,89075	0,72
4	99138	0,99138	0,03	54	88390	0,8839	0,77
5	99113	0,99113	0,03	55	87633	0,87633	0,86
6	99092	0,99092	0,02	56	86819	0,86819	0,93
7	99073	0,99073	0,02	57	85949	0,85949	1,00
8	99054	0,99054	0,02	58	85024	0,85024	1,08
9	99036	0,99036	0,02	59	84020	0,8402	1,18
10	99019	0,99019	0,02	60	82930	0,8293	1,30
11	99001	0,99001	0,02	61	81757	0,81757	1,41
12	98982	0,98982	0,02	62	80496	0,80496	1,54
13	98960	0,9896	0,02	63	79156	0,79156	1,66
14	98935	0,98935	0,03	64	77722	0,77722	1,81
15	98904	0,98904	0,03	65	76216	0,76216	1,94
16	98864	0,98864	0,04	66	74636	0,74636	2,07
17	98809	0,98809	0,06	67	72986	0,72986	2,21
18	98736	0,98736	0,07	68	71244	0,71244	2,39
19	98633	0,98633	0,10	69	69400	0,694	2,59
20	98515	0,98515	0,12	70	67441	0,67441	2,82
21	98388	0,98388	0,13	71	65417	0,65417	3,00
22	98249	0,98249	0,14	72	63301	0,63301	3,23
23	98108	0,98108	0,14	73	61080	0,6108	3,51
24	97971	0,97971	0,14	74	58703	0,58703	3,89
25	97830	0,9783	0,14	75	56195	0,56195	4,27
26	97684	0,97684	0,15	76	53558	0,53558	4,69
27	97535	0,97535	0,15	77	50827	0,50827	5,10
28	97382	0,97382	0,16	78	47975	0,47975	5,61
29	97218	0,97218	0,17	79	45014	0,45014	6,17
30	97041	0,97041	0,18	80	41965	0,41965	6,77
31	96856	0,96856	0,19	81	38811	0,38811	7,52
32	96669	0,96669	0,19	82	35539	0,35539	8,43
33	96482	0,96482	0,19	83	32239	0,32239	9,29
34	96287	0,96287	0,20	84	28914	0,28914	10,31
35	96077	0,96077	0,22	85	25623	0,25623	11,38
36	95853	0,95853	0,23	86	22409	0,22409	12,54
37	95621	0,95621	0,24	87	19338	0,19338	13,70
38	95381	0,95381	0,25	88	16437	0,16437	15,00
39	95131	0,95131	0,26	89	13734	0,13734	16,44
40	94865	0,94865	0,28	90	11282	0,11282	17,85
41	94581	0,94581	0,30	91	9112	0,09112	19,23
42	94278	0,94278	0,32	92	7217	0,07217	20,80
43	93952	0,93952	0,35	93	5571	0,05571	22,81
44	93598	0,93598	0,38	94	4168	0,04168	25,18
45	93222	0,93222	0,40	95	3019	0,03019	27,57
46	92825	0,92825	0,43	96	2150	0,0215	28,78
47	92400	0,924	0,46	97	1511	0,01511	29,72
48	91933	0,91933	0,51	98	1046	0,01046	30,77
49	91429	0,91429	0,55	99	712	0,00712	31,93

(Source : INSEE)

II – LOI EXPONENTIELLE

1 – Première approche : pannes à taux d'avarie constant

Cette approche correspond à l'étude des temps de bon fonctionnement d'un matériel à taux d'avarie constant.

Taux d'avarie et fonction de défaillance

On désigne par $\boldsymbol{T}$ la variable aléatoire qui, à tout matériel tiré au hasard, associe sa **durée de vie** ou temps de bon fonctionnement avant défaillance (TBF).

On se place dans le cas où T est une variable aléatoire de type continu prenant ses valeurs dans]0;+∞[et possédant une **densité** de probabilité $\boldsymbol{f}$.

La fonction de répartition F de la variable aléatoire T est telle que, pour tout $t \geq 0$, $\boxed{F(t) = P(T \leq t) = \int_0^t f(x)dx}$. $\boldsymbol{F}$ est appelée **fonction de défaillance** (F comme "*failure*").

On note, pour tout $t \geq 0$, $\boxed{R(t) = P(T > t) = 1 - F(t)}$. $\boldsymbol{R}$ est appelée **fonction de fiabilité** (R comme "*reliability*").

La question maintenant est celle du choix d'une loi pour la variable aléatoire T, conforme aux observations statistiques et modélisant la fiabilité du matériel.

Ce choix dépend de l'aspect du **taux d'avarie**, défini statistiquement par :

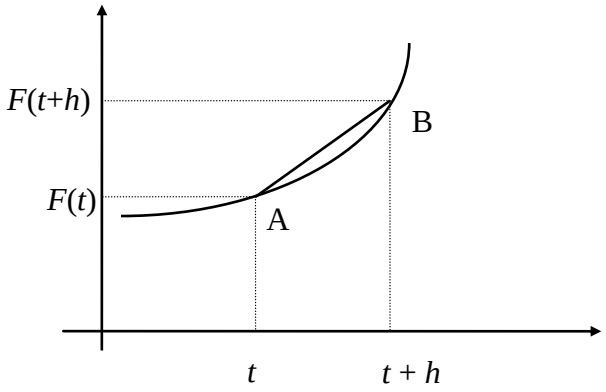
$$\boxed{\frac{\text{nombre de défaillants au cours d'une période}}{\text{nombre de survivants au début de la période}}}.$$

On définira, d'un point de vue probabiliste, le taux d'avarie moyen par unité de temps entre t et $t+h$ par : $\lambda_{[t;t+h]} = \dfrac{F(t+h) - F(t)}{1 - F(t)} \times \dfrac{1}{h}$.

Pour définir le taux d'avarie instantané à l'instant t, la démarche est analogue à celle suivie pour définir la vitesse instantanée.

Le **taux d'avarie** instantané à l'instant t est alors défini par :

$$\lambda(t) = \frac{1}{1 - F(t)} \lim_{h \to 0} \frac{F(t+h) - F(t)}{h} = \frac{F'(t)}{1 - F(t)}.$$



Taux d'avarie constant

On suppose que le taux d'avarie est constant. La fonction de défaillance F vérifie alors, pour $t \geq 0$,

$\lambda = \dfrac{F'(t)}{1 - F(t)}$ d'où $\ln(1 - F(t)) = -\lambda t + k$ puis $F(t) = 1 + Ce^{-\lambda t}$.

Sachant qu'il n'y a pas de défaillance avant l'instant $t = 0$, on doit avoir $F(0) = 0$.

La solution particulière F vérifiant la condition $F(0) = 0$ est alors définie par :

$\boxed{F(t) = 1 - e^{-\lambda t} \quad \text{pour } t \geq 0}$, fonction de répartition de la **loi exponentielle** $\mathcal{E}(\lambda)$ de paramètre λ.

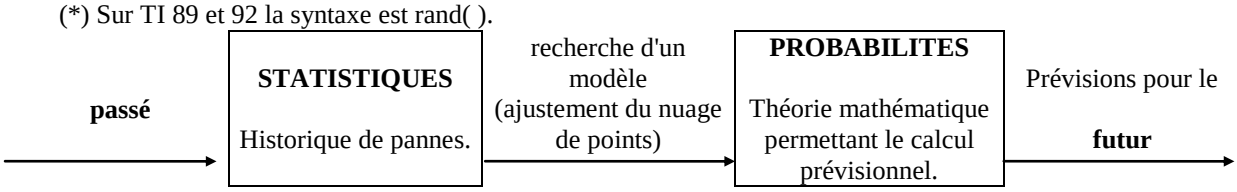
Simulation d'un historique de pannes

A titre d'expérimentation, on peut simuler un historique de pannes à taux d'avarie constant.

Une pièce est supposée avoir un taux d'avarie par heure de 0,007. Chaque heure, l'instruction ***int(rand + 0.007)*** donne 0 s'il n'y a pas eu de panne durant l'heure et donne 1 (avec la probabilité 0,007) s'il y a eu panne durant l'heure. Le temps de bon fonctionnement correspond au temps d'attente de la valeur 1, ce qui est simulé par le programme suivant.

Commentaires	CASIO sans instruction While	CASIO	T.I. 80 - 81 (sans instruction While)	T.I. 82 83 85 89 92(T)
I : compteur du temps de bon fonctionnement (en heures).	-1 → I ↵ Lbl 1 ↵ I + 1 → I ↵ Int (Ran# + 0.007) = 0 ⇒ Goto 1 ↵ I	0 → I ↵ While Int(Ran# + 0.007) = 0 ↵ I + 1 → I ↵ WhileEnd ↵ I	:-1 → I :Lbl 1 :I + 1 → I *: If int (rand+0.007) = 0* :Goto 1 :Disp I	:0 → I :While int(rand + 0.007) = 0 :I + 1 → I :End :Disp I

(*) Sur TI 89 et 92 la syntaxe est rand( ).

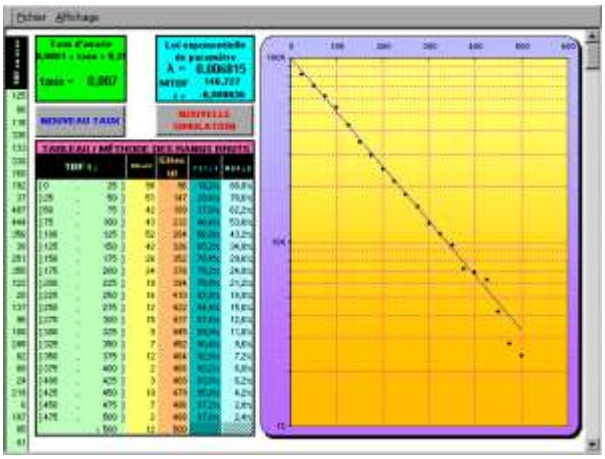


La mise en commun des simulations nous a permis de faire un historique sur 200 TBF :

TBF (h)	25	50	75	100	125	150	175	200	225	250	275	300	325	350
$R(t_i)$ %	64	61	56	50	43	35	29	24	18	15	12	9	8	5

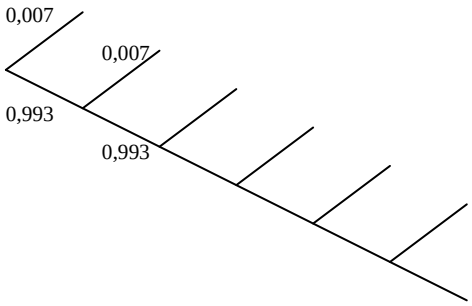
On obtient, sur **papier semi-log**., avec les résultats précédents, un nuage de points à l'alignement très convenable sur le point de coordonnées (t = 0 , 100%), sauf pour le dernier.

Ci-contre, simulation analogue et visualisation de l'ajustement exponentiel sur Excel.



Modélisation

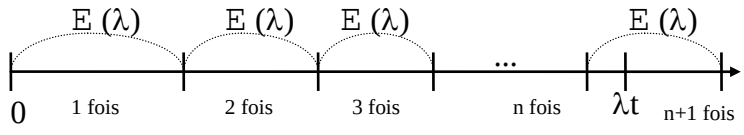
Le temps étant discret dans la simulation (saut d'heure en heure), on simule une loi géométrique plutôt qu'une loi exponentielle.



Loi géométrique (temps d'attente du premier succès dans un schéma de *Bernoulli*)

		T : temps d'attente d'une panne	X : nombre de pannes sur [0 ; t = 286]
Processus de **Poisson**	Temps continu	$\mathcal{E}$ (λ = 0,007)	$\mathcal{P}$ ($\lambda t \approx 2$)
Schéma de **Bernoulli**	Temps discret	$\mathcal{G}$ (p = 0,007) avec $P(T=k) = p(1-p)^{k-1}$	$\mathcal{B}$ (286 ; 0.007)

2- Seconde approche : temps d'attente d'un succès dans un processus de Poisson



De façon plus générale, la loi exponentielle apparaît correspondre au temps d'attente entre deux succès d'un **processus de Poisson** (pannes à taux d'avarie constant mais aussi appels à un standard téléphonique,

arrivées à un guichet de banque, arrivées de véhicules...). Un processus de Poisson possède trois caractéristiques : il est • **sans mémoire**, • **à rythme contant**, • **correspondant à des événements rares**.

On peut suivre ici l'énoncé du ***BTS Maintenance aéronautique 94*** qu'il est intéressant de proposer aux élèves.
⇨ ***Voir également les annales du B.T.S.***

On considère un atelier comprenant de nombreuses machines identiques fonctionnant indépendamment.

a) On appelle X_t la variable aléatoire qui, à chaque machine choisie au hasard, associe le nombre de pannes pendant l'intervalle de temps [0, t]. On admet que X_t suit la **loi de *Poisson*** de paramètre λt.
Déterminer la probabilité que X_t soit nulle .

On a $P(X_t = 0) = \mathrm{e}^{-\lambda t}$.

b) On appelle T la variable aléatoire prenant comme valeur, pour chaque machine choisie au hasard, l'instant de la première panne.
En utilisant le a), déterminer $R(t) = P(T > t)$.

On a $R(t) = P(T > t) = P(X_t = 0) = \mathrm{e}^{-\lambda t}$.

c) En déduire, pour $t \geq 0$, $F(t) = P(T \leq t)$ puis $f(t) = F'(t)$.

On obtient $F(t) = 1 - R(t) = 1 - \mathrm{e}^{-\lambda t}$.

Donc $f(t) = \lambda\, \mathrm{e}^{-\lambda t}$.

d) a étant 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 $I(a) = \frac{1}{\lambda} + e^{-\lambda a}(-a - \frac{1}{\lambda})$.

• Calculer la limite de $I(a)$ quand a tend vers $+\infty$. Que représente cette limite ?

On constate que $\lim\limits_{a \to +\infty} I(a) = \frac{1}{\lambda}$. Pour une variable aléatoire continue, ce calcul fournit la valeur de l'espérance.

On a ainsi montré que, si T suit la loi exponentielle $\mathcal{E}(\lambda)$, l'**espérance** des temps de bon fonctionnements (*Mean Time Between Failures*) vaut **$\boxed{\text{MTBF} = E(T) = 1/\lambda}$.**

e) t et h étant deux nombres réels strictement positifs, démontrer que la probabilité conditionnelle $P(T > t+h \mid T > t)$ ne dépend que de h (et pas de t). On dit que T est *"sans mémoire"*.

On a $P(T > t+h \mid T > t) = \frac{P(T > t+h)}{P(T > t)} = \frac{R(t+h)}{R(t)} = e^{-\lambda h}$ qui ne dépend que de h.

(On peut ne pas traiter cette question avec les élèves, le conditionnement reste un sujet délicat.)

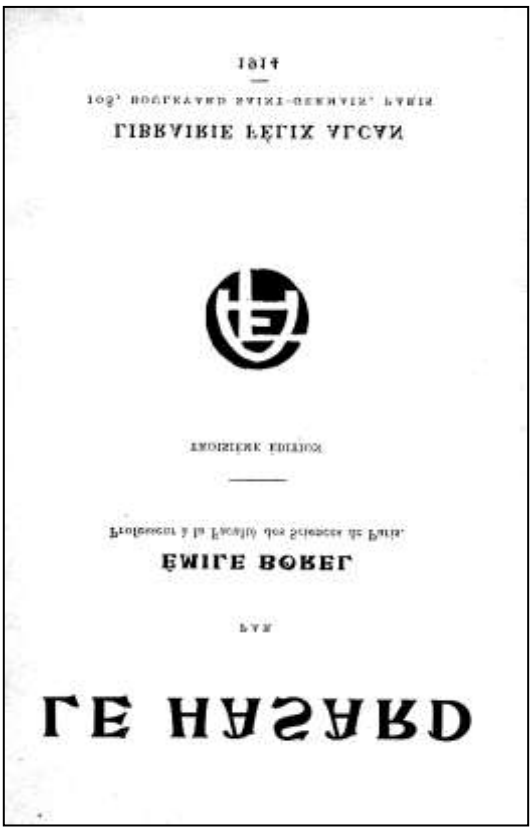
3 – L'exemple de la désintégration atomique

Le programme de terminale S applicable à la rentrée 2002 contient, à propos de la loi exponentielle, les éléments suivants :

"Contenus :
Exemples de lois continues : - 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."

On peut, à ce propos, étudier les extraits suivants du livre d'*Emile Borel*, "*Le hasard*" (édition de 1914), contemporain des découverte de *Pierre* et *Marie Curie*.
L'exercice peut également être proposé en B.T.S.

Pierre Curie (1859-1906) épousa en 1895 une étudiante d'origine polonaise, *Marie Sklodowska* (1867-1934). Leur collaboration aboutit à la découverte du polonium et du radium, ce qui leur valut de recevoir en 1903, en commun avec *Henri Becquerel*, le prix Nobel de physique. *Marie Curie*, restée veuve, isola le radium pur et en détermina la masse atomique. Elle reçut, à ce titre, le prix Nobel de chimie en 1911.



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).

— 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.)

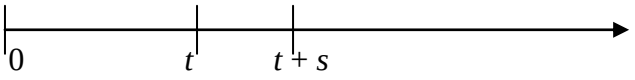
e^{-hy}. Si l'on mesure un grand nombre de distances (c'est-à-dire d'intervalles de temps écoulés entre deux émissions successives), on peut vérifier l'accord entre ce résultat et l'expérience. Cette vérification a été faite d'une manière satisfaisante[1], rendant par suite très vraisemblable le point de départ du calcul. L'étude des scintillations conduit aussi à des résultats très satisfaisants.

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 : $\dfrac{P(t \leq T \leq t + s)}{P(T > t)}$.
Montrer que ce rapport peut s'écrire $\dfrac{F(t + s) - F(t)}{1 - F(t)}$.

2) On désigne par "taux moyen de désintégration par unité de temps entre t et $t + s$ " la quantité $\dfrac{F(t + s) - F(t)}{1 - F(t)} \times \dfrac{1}{s}$ et par "taux instantané de désintégration au temps t " la limite $h(t) = \lim\limits_{s \to 0} \dfrac{F(t + s) - F(t)}{1 - F(t)} \times \dfrac{1}{s}$.
Montrer que $h(t) = \dfrac{F'(t)}{1 - F(t)}$.

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 $t \in \mathbb{R}$, $\dfrac{F'(t)}{1 - F(t)} = h$. 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 $f(t)$.

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 partie, que $I(a) = \dfrac{1}{h} + e^{-ha}\left(-a - \dfrac{1}{h}\right)$.

7) Déterminer l'espérance de T (temps "moyen" d'attente d'une désintégration) : $E(T) = \lim\limits_{a \to +\infty} I(a)$.

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)

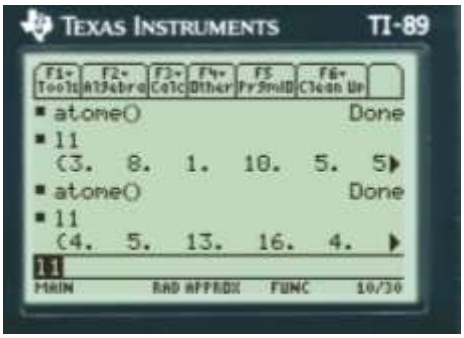
Comparer le temps moyen $\bar{x}$ obtenu sur 200 temps simulés avec $E(T) = \frac{1}{h} = \frac{1}{0{,}07}$.

Corrigé de l'exercice : "Désintégration atomique"

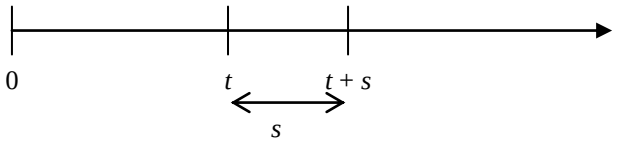
1) $P(T > t)$ correspond au pourcentage (théorique) d'atomes non désintégrés au temps t.
$P(t \leq T \leq t + s)$ correspond au pourcentage (théorique) d'atomes se désintégrant durant l'intervalle de temps $[t, t + s]$.
On a ensuite $P(T > t) = 1 - P(T \leq t) = 1 - F(t)$ et

$$P(t \leq T \leq t + s) = \int_t^{t+s} f(x)dx = \int_0^{t+s} f(x)dx - \int_0^t f(x)dx = F(t + s) - F(t).$$

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) : $1/h \approx 14{,}3$.



Remarque : 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 est davantage celle utilisée en physique (voir les manuels de physique).

4 – Détermination pratique de λ

a - Par changement de variable et régression linéaire

Comme $R(t) = e^{-\lambda t} \Leftrightarrow \ln R(t) = -\lambda t$, lorsque le nuage de points (t_i , ln $R(t_i)$) est correctement ajusté (selon la méthode des moindres carrés) par la droite d'équation $y = a\,t + b$ avec $b \approx 0$, on peut considérer que la variable aléatoire T , correspondant aux temps de bon fonctionnement, suit la loi exponentielle de paramètre $\lambda = -a$.

Remarque à propos du calcul des fréquences observées :
Les fréquences observées de matériels survivants $R(t_i)$ correspondent a priori à $R(t_i) = 1 - \frac{n_i}{n}$ où n_i est le défaillances à l'instant t_i (méthode des ***rangs bruts***).
Pour de petits échantillons ($n \leq 50$), on préfère, selon la méthode des ***rangs moyens*** de *Johnson*, estimer $R(t_i)$ par $R(t_i) = 1 - \frac{n_i}{n+1}$ (cela évite d'avoir $R(t_n) = 0$ alors qu'avec un échantillon plus grand, on aurait probablement encore au moins un matériel en fonctionnement au temps t_n).

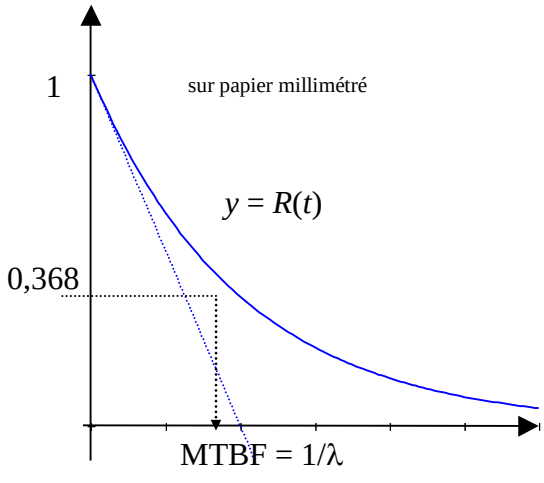
Quand $n < 20$, on prend la méthode des ***rangs médians*** de *Johnson* où $R(t_i) = 1 - \frac{n_i - 0{,}3}{n + 0{,}4}$

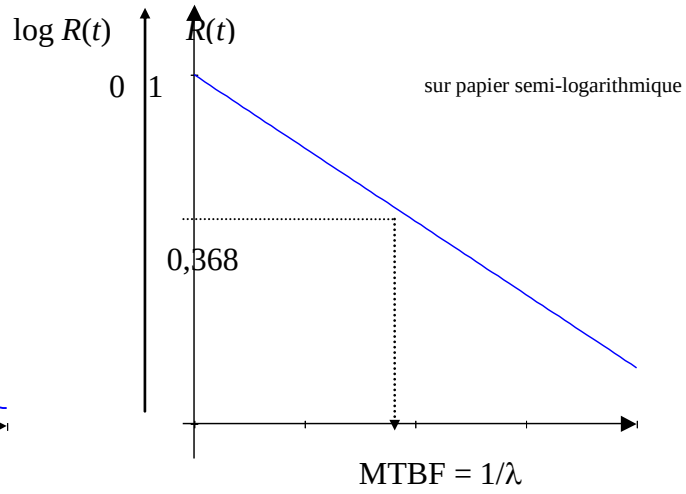
(cette méthode n'est pas utilisée à l'épreuve de mathématique).

⇨ **Exemple : voir BTS Systèmes constructifs bois et habitats 98 .**

b - Ajustement graphique, sur papier semi-logarithmique

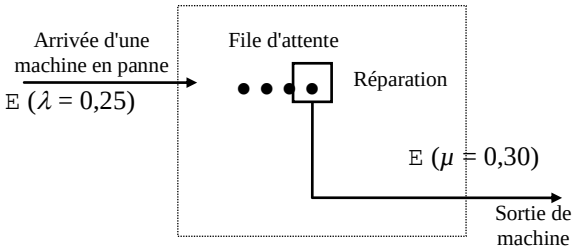
Sur papier semi-logarithmique, on a, en abscisses une échelle arithmétique pour le temps et, en ordonnées, une échelle logarithmique (log décimal) pour la fiabilité :
On sait que $R(t) = e^{-\lambda t}$ d'où $\ln R(t) = -\lambda t$ et $\log R(t) = \frac{\ln R(t)}{\ln 10} = -\frac{\lambda}{\ln 10} t$.
On obtient donc une représentation linéaire de la fiabilité.
Déterminer la pente d'une droite sur du papier semi-logarithmique n'est pas pratique. On obtiendra λ en considérant la MTBF (1/λ) : $R(t) = e^{-\lambda t} \Rightarrow R(1/\lambda) = e^{-1} \approx 0{,}368$.



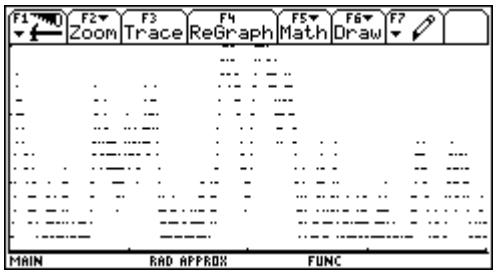


⇨ **Exemple : voir BTS Domotique 96.**

5 – Application de la loi exponentielle à l'étude des files d'attente



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 la machine en réparation) sur l'intervalle de temps [0 ; ≈2000 h].



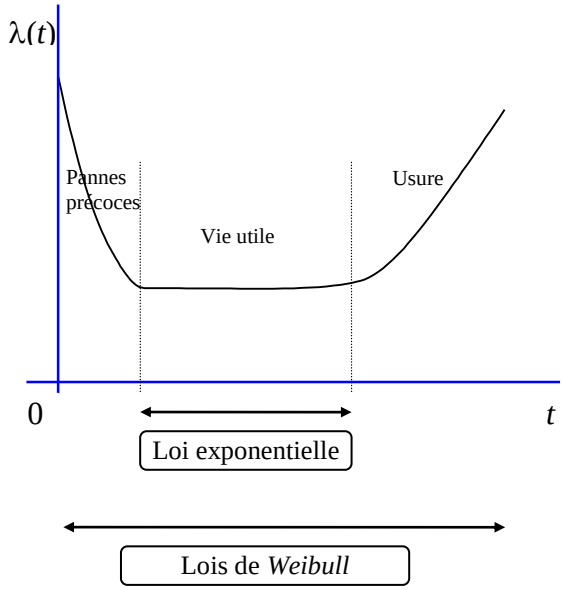
CASIO	T.I.
ViewWindow 0,2000,500,0,15,5 ↵	:0 → Xmin
? → L ↵	:2000 → Xmax
? → M ↵	:500 → Xscl
0 → B ↵	:0 → Ymin
1 → Q ↵	:15 → Ymax
(-ln Ran#) ÷ L → C ↵	:5 → Yscl
Plot C,Q ↵	:PlotsOff
(-ln Ran#) ÷ M → R ↵	:ClrDraw
(-ln Ran#) ÷ L → A ↵	:Input L
Lbl 1 ↵	:Input M
A > R ⇒ Goto 3 ↵	:0 → B
R - A → R ↵	:1 → Q
Lbl 2 ↵	:(-ln (rand)) / L → C
C + A → C ↵	:Pt-On (C,Q)
B + QA → B ↵	:(-ln (rand)) / M → R
Q + 1 → Q ↵	:(-ln (rand)) / L → A
(-ln Ran#) ÷ L → A ↵	:Lbl 1
Goto 4 ↵	:If A > R
Lbl 3 ↵	:Goto 3
A - R → A ↵	:R - A → R
C + R → C ↵	:Lbl 2
B + QR → B ↵	:C + A → C
Q - 1 → Q ↵	:B + Q×A → B
(-ln Ran#) ÷ M → R ↵	:Q + 1 → Q
Lbl 4 ↵	:(-ln (rand)) / L → A
Plot C , Q ↵	:Goto 4
C ≥ 2000 ⇒ Goto 5 ↵	:Lbl 3
Q = 0 ⇒ Goto 2 ↵	:A - R → A
Goto 1 ↵	:C + R → C
Lbl 5 ↵	:B + Q×R → B
B ÷ C	: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

III - LOI DE WEIBULL

Le nom de ***Wallodi Weibull*** (1887 – 1979) est attaché à celui de la fiabilité. D'origine suédoise, *Weibull* travailla comme inventeur (roulements à billes, marteau électrique...) et ingénieur conseil dans de nombreuses sociétés suédoises ou allemandes, par exemple chez *SAAB*. Il s'intéressa aux problèmes de résistance des matériaux, en particulier à ceux de fatigue et de rupture des tubes à vide. C'est dans ce cadre qu'apparaît en 1939 pour la première fois la distribution de *Weibull*. Mais l'article qui eut le plus d'influence fut publié en 1951 dans le *"Journal of Applied Mechanics"* sous le titre *"A Statistical Distribution Function of Wide Applicability"* où sont décrit sept cas d'utilisation de la distribution de *Weibull*. En effet, l'intérêt de cette distribution, outre ses propriétés analytiques satisfaisantes, est de permettre un bon ajustement d'une grande variété de problèmes de durée de vie.

1 - Aspect général des taux d'avarie

On constate expérimentalement, que pour la plupart des matériels, la courbe représentative du taux d'avarie (instantané) en fonction du temps, a la forme suivante :



Courbe en baignoire

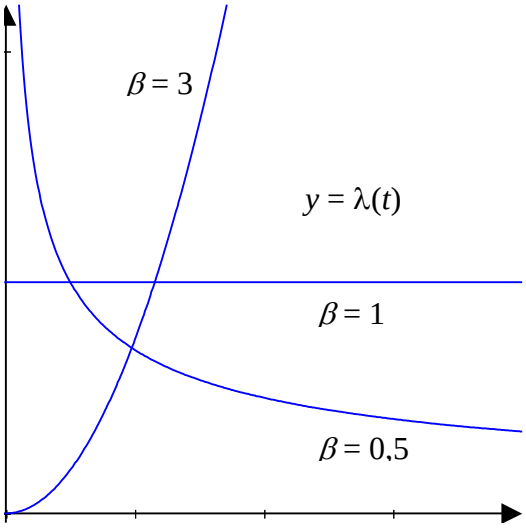
Lorsque λ est constant, on prendra pour T la loi exponentielle.
Lorsque λ varie, on cherchera un modèle parmi les lois de *Weibull*.

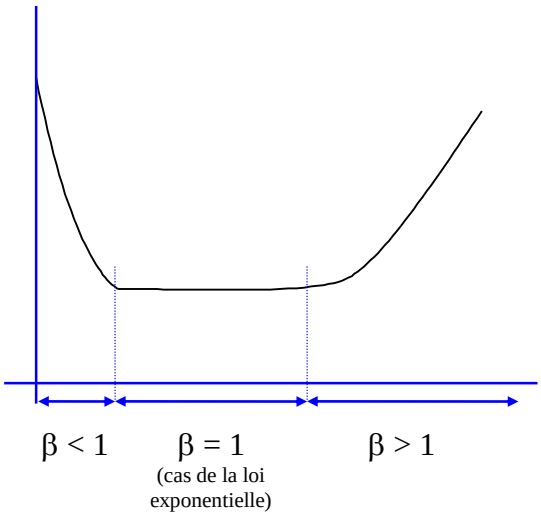
2 - Loi de Weibull

Pour couvrir tous les cas où le taux d'avarie λ(t) varie avec le temps, *Weibull* a choisi pour λ une fonction dépendant de trois paramètres : γ ; β et η .
La variable aléatoire T qui, à tout matériel tiré au hasard, associe son temps de bon fonctionnement avant défaillance, suit la ***LOI DE WEIBULL*** *de paramètres* $\boldsymbol{\gamma}$ ***;*** $\boldsymbol{\beta}$ ***;*** $\boldsymbol{\eta}$ *lorsque le taux d'avarie est :* $\lambda(t) = \frac{\beta}{\eta}\left(\frac{t-\gamma}{\eta}\right)^{\beta-1}$ avec $t > \gamma$, $\beta > 0$, $\eta > 0$.

Rôle des différents paramètres :

- β est le "***paramètre de forme***" : son choix permet d'ajuster l'allure de λ à la forme expérimentale de l'évolution du taux d'avarie :





- γ (paramètre de position) fixe l'origine de l'étude ($t > \gamma$).
- η (paramètre d'échelle) permet, lorsque $\beta = 1$, d'avoir $\lambda = 1/\eta$, constante arbitraire.

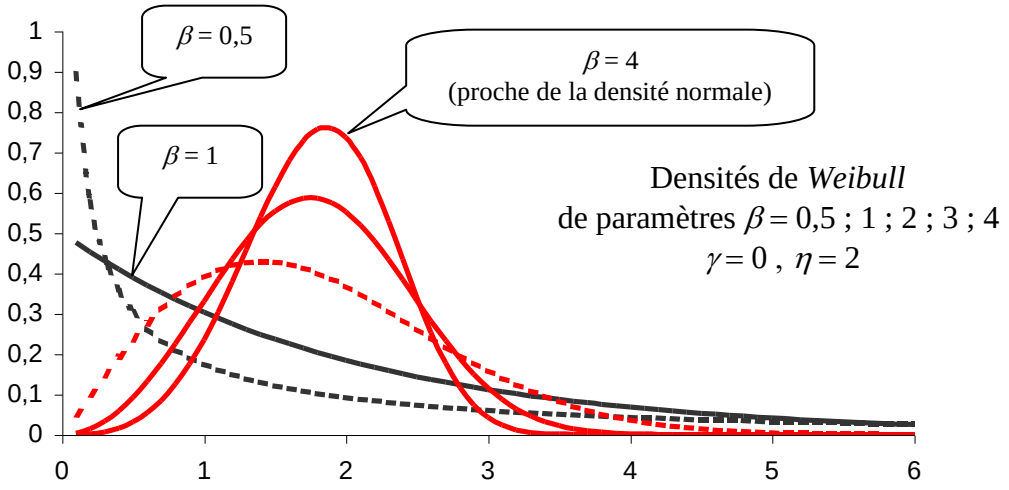
| FONCTION DE FIABILITE, DE DEFAILLANCE, DENSITE :

On a vu que, par définition, $\lambda(t) = \dfrac{F'(t)}{1-F(t)}$ d'où $\displaystyle\int_\gamma^t \lambda(x)\,dx = \left[-\ln(1-F(x))\right]_\gamma^t = -\ln R(t)$.

Donc, $R(t) = e^{-\int_\gamma^t \lambda(x)\,dx} = e^{-\frac{\beta}{\eta}\int_\gamma^t \left(\frac{x-\gamma}{\eta}\right)^{\beta-1} dx}$.

Ainsi, lorsque T suit la loi de *Weibull* de paramètres γ, β, η, on a :

$$\boxed{R(t) = e^{-\left(\frac{t-\gamma}{\eta}\right)^\beta} \quad F(t) = 1 - e^{-\left(\frac{t-\gamma}{\eta}\right)^\beta} \quad f(t) = \frac{\beta}{\eta}\left(\frac{t-\gamma}{\eta}\right)^{\beta-1} e^{-\left(\frac{t-\gamma}{\eta}\right)^\beta}}.$$



| MTBF :

$E(T) = \displaystyle\int_\gamma^{+\infty} t f(t)\,dt = \eta \int_0^{+\infty} x^{\frac{1}{\beta}} e^{-x} dx + \gamma = \eta\Gamma\left(1+\frac{1}{\beta}\right) + \gamma$ où Γ est la fonction d'Euler.

Puis on peut calculer $\sigma(T) = \eta B$ avec $B = \sqrt{\Gamma\left(1+\frac{2}{\beta}\right) - \left(\Gamma\left(1+\frac{1}{\beta}\right)\right)^2}$

En conclusion, $\boxed{\text{MTBF} = E(T) = \eta A + \gamma \ ; \ \sigma(T) = \eta B}$, où A et B proviennent d'un calcul d'intégrale et sont donnés par la table du formulaire.

3 - Détermination des paramètres de la loi de Weibull

a) Par régression linéaire

Pour se ramener à un ajustement linéaire, on doit passer deux fois au logarithme :

$$R(t) = e^{-\left(\frac{t-\gamma}{\eta}\right)^\beta} \Leftrightarrow -\ln(R(t)) = \left(\frac{t-\gamma}{\eta}\right)^\beta \Leftrightarrow \ln[-\ln(R(t))] = \beta\ln(t-\gamma) - \ln(\eta^\beta).$$

On détermine γ de sorte que le nuage de points de coordonnées $x_i = \ln(t_i - \gamma)$ et $y_i = \ln(-\ln R(t_i))$ soit correctement ajusté par une droite d'équation $y = ax + b$.

On a alors $\ln(-\ln R(t)) = a\ln(t-\gamma) + b$ c'est à dire $\ln R(t) = -(t-\gamma)^a \times e^b$

soit $R(t) = e^{-\left(\frac{t-\gamma}{e^{-b/a}}\right)^a}$.

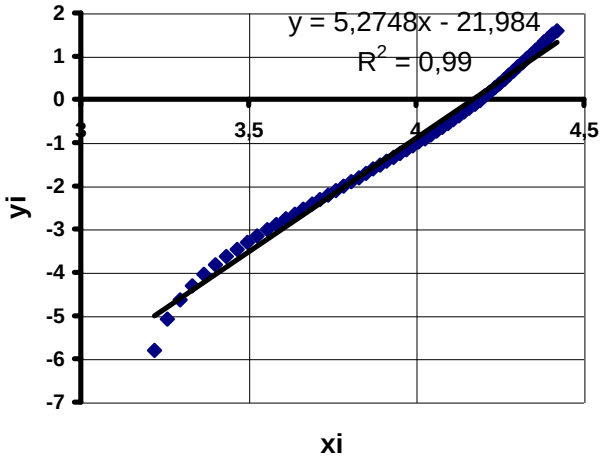
On a ainsi un ajustement par la loi de *Weibull* de paramètres γ ; $\beta = a$; $\eta = e^{-b/a}$.

Exemple: Dans le cas de la mortalité après 40 ans, on a, sur *Excel*, choisi $\gamma = 16$ de sorte que, dans l'ajustement linéaire du nuage (x_i , y_i) le coefficient de corrélation r soit le plus

proche possible de 1. Le coefficient β est alors donné par la pente a de de la droite de régression, et le coefficient η par $e^{-b/a}$: $\beta \approx 5$ et $\eta \approx 65$.

On peut alors calculer la MTBF, qui représente l'espérance de vie d'un français de 40 ans, soit un âge de 75 ans.

Paramètres	
r	0,99500975
γ	16
β	5,2747657
η	64,5713411
MTBF	75,4702922
σ	12,971447



⇨ **Exemple : BTS Groupement B 2000 spécialité Maintenance industrielle.**

b) Par ajustement graphique sur le "papier de Weibull"

Cas où $\gamma = 0$:

On suppose dans ce qui suit que $\gamma = 0$ (origine des défaillances à $t = 0$), ce qui est le cas le plus fréquent à l'examen.

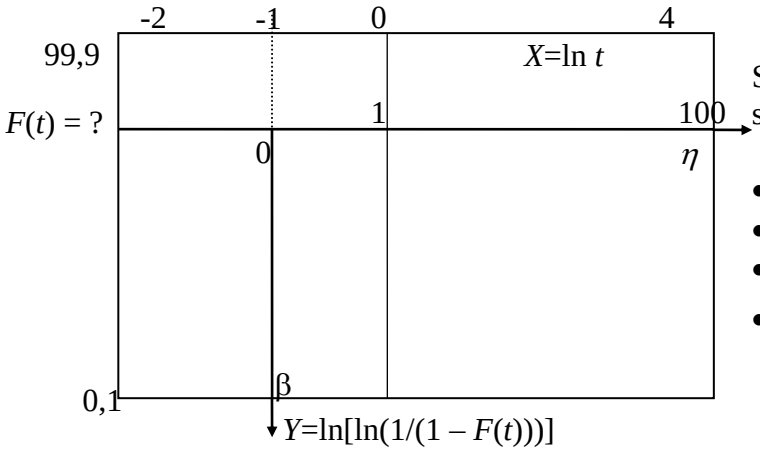
Il existe dans ce cas un papier spécial, dit "papier de *Weibull*" ou "d'*Allan Plait*" qui transforme le tracé de la fonction de défaillance F en une droite.

De $R(t) = e^{-\left(\frac{t}{\eta}\right)^{\beta}}$, on déduit $\ln R(t) = -\left(\frac{t}{\eta}\right)^{\beta}$ donc $\ln \frac{1}{R(t)} = \left(\frac{t}{\eta}\right)^{\beta}$,

puis $\ln(\ln \frac{1}{R(t)}) = \beta[\ln t - \ln \eta]$ d'où $\ln(\ln \frac{1}{1-F(t)}) = \beta[\ln t - \ln \eta]$.

En posant $Y = \ln(\ln \frac{1}{1-F(t)})$ et $X = \ln t$, on obtient l'équation d'une droite :

$Y = \beta X - \beta \ln \eta$.



Sur le papier de *Weibull*, on distingue les axes suivants :

- Sur l'axe noté "η" : le temps t.
- Sur l'axe "marge supérieure" : $X = \ln t$.
- Sur l'axe "marge gauche" : $F(t)$ en %.
- Sur l'axe noté "β" : $Y = \ln(\ln \frac{1}{1-F(t)})$.

On peut demander aux élèves de retrouver le pourcentage $F(t)$ correspondant à l'intersection de l'axe "η " avec la marge gauche : la condition $Y = 0$ conduit à $\ln(\ln\frac{1}{1-F(t)}) = 0$ c'est à dire $F(t) = 1 - e^{-1} \approx 63{,}2$ %.

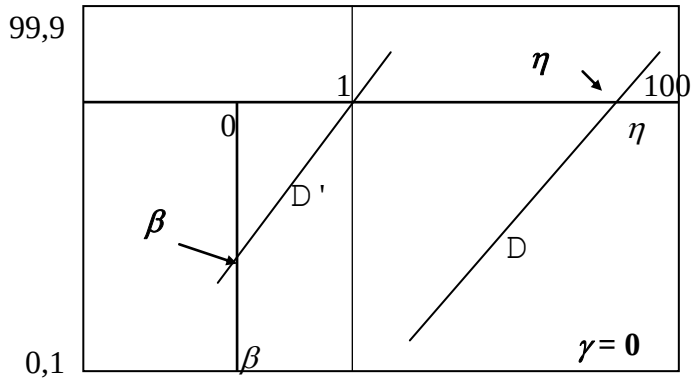
Si, lorsque l'on reporte sur ce papier les points M(t_i ; $F(t_i)$), on obtient un "bon" alignement selon une droite D , on en déduit que $\gamma = 0$.

Cette droite a pour équation $Y = \beta X - \beta \ln \eta$, avec $X = \ln t$ et $Y = \ln(\ln \frac{1}{1-F(t)})$.

Pour le point de D d'ordonnée $Y = 0$, on a $X = \ln t = \ln \eta$. L'échelle de l'axe noté "η " étant logarithmique, on en déduit que la <u>valeur du paramètre η</u> est donnée par l'intersection de D avec l'axe noté "η".

Le paramètre β est le coefficient directeur de la droite D . Si l'on considère la parallèle D ' à D passant par le point de coordonnées ($X = 0$, $Y = 0$), D ' a pour équation : $Y = \beta X$.

Pour $X = -1$, on a donc $Y = -\beta$. En graduant l'axe des valeurs de β "vers le bas", <u>la valeur du paramètre β</u> pourra se lire directement à l'intersection de la droite D ' avec l'axe noté "β ".



D'où la procédure suivante :

- Reporter les points (t_i;$F(t_i)$).
- Tracer une droite D d'ajustement du nuage de points (si $\gamma = 0$).
- Reporter la parallèle D ' à D à partir du point d'abscisse 1 de l'axe noté "η ".
- Repérer sur les axes correspondants les valeurs des paramètres η et β .

⇨ **Exemple BTS Maintenance 98.**

Cas où $\gamma \neq 0$:

Lorsque $\gamma \neq 0$, les points reportés sur le papier de *Weibull* sont voisins d'une courbe non rectiligne.

Lorsque $\gamma > 0$, on peut lire une valeur très approximative de γ en prolongeant la courbe "vers le bas", en effet on a $\lim_{t \to \gamma} F(t) = 0$ (pas de défaillance) d'où $\lim_{t \to \gamma} Y = -\infty$.

On prend comme TBF : $t_i - \gamma$ et, avec ce changement de variable, on se ramène au cas précédent où "$\gamma = 0$".

⇨ **Exemple BTS Maintenance Nouvelle Calédonie 95** ($\gamma = 60$)**.**

IV – ESTIMATION DES PARAMETRES D'UNE LOI EXPONENTIELLE OU DE WEIBULL

Ce paragraphe constitue un approfondissement des précédents et n'est pas "indispensable en première lecture".

Les modules "*Statistique inférentielle*" et "*Fiabilité*" des programmes de mathématiques en BTS parus en 2001 insistent sur les outils statistiques, souvent sophistiqués, utilisés par les logiciels industriels spécialisés que certains de nos étudiants peuvent être amenés à rencontrer (en particulier lors de stages). Il s'agit alors pour le professeur de mathématiques, d'avoir le recul nécessaire pour éclairer ces élèves sur la signification de ces méthodes.

Extraits du programme :
"Les méthodes statistiques sont aujourd'hui largement utilisées dans les milieux économique, social ou professionnel. Des ***procédures*** *plus ou moins* ***élaborées*** *sont mises en œuvre […]. Des* ***logiciels spécialisés*** *exécutent automatiquement les calculs, suivant les normes AFNOR ou ISO."*
"L'objectif essentiel […] est d'amener les étudiants à prendre du recul vis-à-vis des méthodes utilisées."
*"On montrera que l'utilisation du papier de Weibull permet d'obtenir une estimation des paramètres de cette loi, à partir de la fonction de répartition empirique. (****L'utilisation de logiciels ad hoc donne directement une estimation optimale des mêmes paramètres et permet, en outre, d'obtenir un intervalle de confiance****)."*

L'objectif de ce chapitre est d'apporter quelques éléments théoriques concernant l'estimation selon le maximum de vraisemblance et l'obtention d'intervalles de confiance qui en découle.

1 – LA METHODE DU MAXIMUM DE VRAISEMBLANCE

Cette méthode a été développée par *R. A. Fischer* dès sa première publication statistique en 1912. On peut cependant en trouver trace dans les travaux, au XVIII^{e} siècle, de *J. H. Lambert* et *Daniel Bernoulli*, puis de *C. F. Gauss* dans le cadre de la détermination de la loi normale. La popularité de cette méthode vient du fait de son applicabilité à une large variété de modèles statistiques, permettant des investigations là où d'autres méthodes ne permettent pas d'accéder.

On considère une variable aléatoire continue X , de fonction de densité f , laquelle dépend d'un paramètre θ à estimer. L'estimation de θ se fera à partir d'un échantillon de n réalisations indépendantes de X : soit x_1, x_2, …, x_n ces observations. L'idée consiste à estimer θ par la valeur rendant la plus probable l'observation des valeurs x_1, x_2, …, x_n (qui ont réellement été obtenues sur l'échantillon).
On note X_i la variable aléatoire fournissant le $i^{\text{ème}}$ élément de l'échantillon. Les X_i sont donc indépendantes, de même loi que X. Ces variables aléatoires étant continues, on ne calcule pas exactement la probabilité $P(\bigcap_i (X_i = x_i))$, qui est nulle, mais la probabilité $P(\bigcap_i (X_i \in [x_i, x_i + dx_i]))$.

En vertu de l'indépendance des X_i , on a :

$$P(\bigcap_i \left(X_i \in [x_i, x_i + dx_i]\right)) = \prod_{i=1}^{n} P(X_i \in [x_i, x_i + dx_i]) = \prod_{i=1}^{n} f(x_i, \theta) dx_i .$$

On recherchera ainsi θ tel que la "***vraisemblance***" $\boxed{L = \prod_{i=1}^{n} f(x_i, \theta)}$ soit maximale (L pour "*likelihood*"), ce qui revient à maximiser $\ln L = \sum_{i=1}^{n} \ln f(x_i, \theta)$.

Une condition nécessaire est donc que l'estimation de θ soit solution de $\boxed{\frac{\partial}{\partial \theta} \ln L = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln f(x_i, \theta) = 0}$ nommée ***équation de vraisemblance***.

2 – ESTIMATION DU PARAMETRE D'UNE LOI EXPONENTIELLE PAR LA METHODE DU MAXIMUM DE VRAISEMBLANCE

Dans ce qui suit on considère que la variable aléatoire T associant son temps de bon fonctionnement à tout matériel d'un certain type pris au hasard, suit une loi exponentielle de paramètre λ. On cherche à estimer la valeur de λ d'après les temps de bon fonctionnement (t_1, …, t_n) obtenus sur un échantillon aléatoire de taille n.

Estimation du paramètre λ par le maximum de vraisemblance

On a ici comme densité de probabilité pour T, $f(t) = \lambda e^{-\lambda t}$ et l'on cherche λ maximisant la quantité $\ln L = \sum_{i=1}^{n} \ln\left(\lambda e^{-\lambda t_i}\right) = n \ln \lambda - \lambda \sum t_i$.

L'équation de vraisemblance s'écrit donc $\frac{\partial}{\partial \lambda} \ln L = 0 \Leftrightarrow \frac{n}{\lambda} - \sum t_i = 0$.

L'estimation de λ obtenue selon la méthode du maximum de vraisemblance sera donc $\frac{n}{\sum t_i}$.

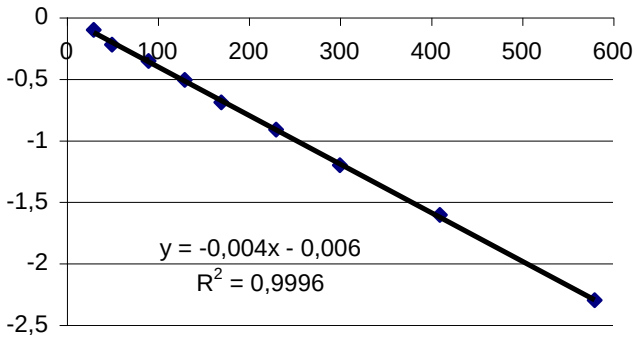
Exemple

Un échantillon aléatoire de 9 temps de bon fonctionnement a donné les valeurs suivantes :
30 ; 50 ; 90 ; 130 ; 170 ; 230 ; 300 ; 410 ; 580.
L'estimation de λ fournie par le maximum de vraisemblance est donc :

$\lambda \approx \frac{9}{30 + ... + 580} \approx 0{,}0045.$

Remarque :
La méthode d'estimation par régression linéaire selon les moindres carrés, fondée sur l'équivalence $R(t) = e^{-\lambda t} \Leftrightarrow \ln R(t) = y = -\lambda t$, fournit, comme estimation de λ, la valeur 0,004 (voir les calculs ci-après, réalisés selon la méthode des "rangs moyens" où les fréquences empiriques $R(t_i)$ sont obtenues en divisant l'effectif des survivants par $n + 1$ au lieu de n).

ti	R(ti)	yi = ln(R(ti))
30	0,9	-0,10536052
50	0,8	-0,22314355
90	0,7	-0,35667494
130	0,6	-0,51082562
170	0,5	-0,69314718
230	0,4	-0,91629073
300	0,3	-1,2039728
410	0,2	-1,60943791
580	0,1	-2,30258509



Remarque : par cette méthode, la droite d'ajustement ne passe pas nécessairement par le point (0, 0), ce qui est dû aux variations aléatoires.

L'estimation fournie par la méthode du maximum de vraisemblance n'est pas nécessairement optimale. Ainsi l'estimateur correspondant au maximum de vraisemblance pour la loi exponentielle, la variable aléatoire $\Lambda = \frac{n}{\sum T_i}$, n'est pas sans biais (la variable aléatoire T_i est celle qui associe son temps de bon fonctionnement au matériel numéro i testé, sur les n matériels testés indépendamment jusqu'à défaillance). En effet, on montre que $E(\Lambda) = \frac{n}{n-1}\lambda$.

⇨ Ainsi le maximum de vraisemblance a tendance, "en moyenne" à surestimer un peu λ (surtout sur de petits échantillons).

En revanche, $E\left(\frac{1}{\Lambda}\right) = E\left(\frac{1}{n}\sum T_i\right) = \frac{1}{\lambda}$ ainsi $\frac{1}{\Lambda}$ est un estimateur sans biais de la MTBF, estimée alors par $\frac{1}{n}\sum t_i = \bar{t}$.

L'avantage de l'estimateur $\Lambda = \frac{n}{\sum T_i}$ donné par le maximum de vraisemblance est qu'il permet de construire un intervalle de confiance pour λ, ce que ne permet pas l'estimation par régression linéaire.

Intervalle de confiance pour λ

Plutôt que de considérer la variable aléatoire $\Lambda = \frac{n}{\sum T_i}$ dont la loi ne fait pas partie des lois usuelles, on étudie la variable aléatoire $2\lambda \sum_1^n T_i$.

Lorsque T_i suit la loi exponentielle de paramètre λ, la variable aléatoire λT_i suit une loi gamma de paramètre 1 et la variable aléatoire $2\lambda T_i$ suit une loi du khi 2 à 2 degrés de liberté. Les T_i étant indépendantes, la variable aléatoire $2\lambda \sum_1^n T_i$ suit une loi du khi 2 à $2n$ degrés de liberté[6].

[6] Pour les propriétés liant ces lois, on peut consulter : SAPORTA – "Probabilités, analyse des données et statistique" – Ed. Technip – pages 40-41.

On peut alors déterminer un *intervalle de probabilité* pour la variable aléatoire $2\lambda \sum_1^n T_i$ en utilisant la table du khi 2.

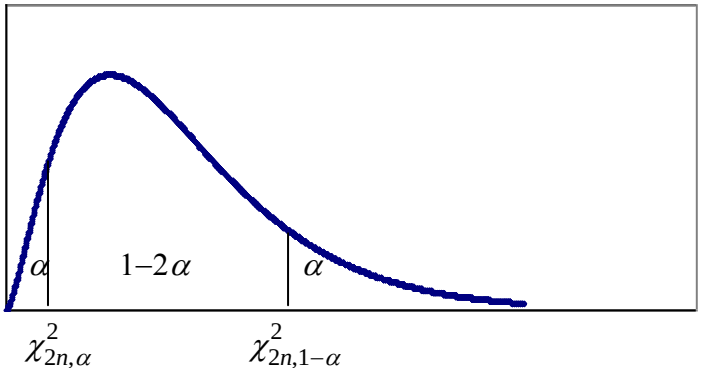
On recherche dans la table les nombres $\chi^2_{2n,\alpha}$ et $\chi^2_{2n,1-\alpha}$ tels que :

$$P(\chi^2_{2n,\alpha} \le 2\lambda \sum T_i \le \chi^2_{2n,1-\alpha}) = 1 - 2\alpha$$

ce qui équivaut à

$$P\left(\frac{\chi^2_{2n,\alpha}}{2\sum T_i} \le \lambda \le \frac{\chi^2_{2n,1-\alpha}}{2\sum T_i} \right) = 1 - 2\alpha .$$

densité du khi 2 à 2*n* degrés de liberté



Pour les temps de bon fonctionnement t_i obtenus sur l'échantillon de taille n, on prendra comme ***intervalle de confiance*** de λ au coefficient de confiance $1 - 2\alpha$:

$$\left[\frac{\chi^2_{2n,\alpha}}{2\sum t_i} , \frac{\chi^2_{2n,1-\alpha}}{2\sum t_i} \right].$$

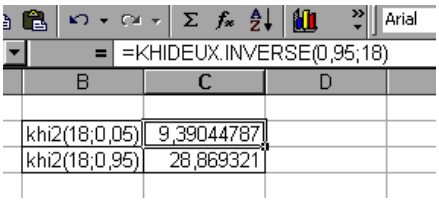
Exemple

On reprend l'exemple précédent, pour lequel $\sum t_i = 1990$.

La table du khi 2 (ou la fonction KHIDEUX.INVERSE d'Excel) donne, pour un χ^2 à $2n = 18$ degrés de liberté :

$\chi^2_{18\,;\,0,05} \approx 9{,}390$ et $\chi^2_{18\,;\,0,95} \approx 28{,}869$

(attention, sur Excel, on indique la probabilité "à droite" et non "à gauche").



On aura donc comme intervalle de confiance pour λ au niveau de confiance de 90 % :
[0,0023 ; 0,0073] .

3 – ESTIMATION DES PARAMETRES D'UNE LOI DE WEIBULL PAR LA METHODE DU MAXIMUM DE VRAISEMBLANCE

Dans ce qui suit on considère que la variable aléatoire T associant son temps de bon fonctionnement à tout matériel d'un certain type pris au hasard, suit une loi de *Weibull*. On suppose que $\gamma = 0$ et on cherche à estimer les valeurs de β et η d'après les observations d'un échantillon aléatoire de taille n.

Estimation du paramètre β

On a dans ce cas la densité $f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\eta}\right)^{\beta}}$.

Temporairement, on pose $\mu = \frac{1}{\eta^{\beta}}$ de sorte que $f(t) = \beta\,\mu\,t^{\beta-1}e^{-\mu\,t^{\beta}}$.

On a $\ln L = \sum_{i=1}^{n} \ln f(t_i) = \sum_{i=1}^{n} \left(\ln \beta + \ln \mu + (\beta - 1) \ln t_i - \mu t_i^{\beta} \right)$.

D'où $\ln L = n \ln \beta + n \ln \mu + (\beta - 1) \sum_{i=1}^{n} \ln t_i - \mu \sum_{i=1}^{n} t_i^{\beta}$.

L'équation de vraisemblance conduit à annuler les dérivées partielles, par rapport à β et μ, de l'expression précédente :

$$\begin{cases} \frac{\partial}{\partial \beta} \ln L = 0 \\ \frac{\partial}{\partial \mu} \ln L = 0 \end{cases} \Leftrightarrow \begin{cases} \frac{n}{\beta} + \sum \ln t_i - \mu \sum (\ln t_i)\, t_i^{\beta} = 0 \\ \frac{n}{\mu} - \sum t_i^{\beta} = 0 \end{cases}.$$

On recherche donc l'estimation de β vérifiant : $\boxed{\frac{n}{\beta} + \sum \ln t_i - \frac{n}{\sum t_i^{\beta}} \sum (\ln t_i) t_i^{\beta} = 0}$.

On n'a donc pas une formule explicite de l'estimateur (la variable aléatoire permettant l'estimation) de β selon le maximum de vraisemblance. L'estimation sera alors obtenue par approximations successives.

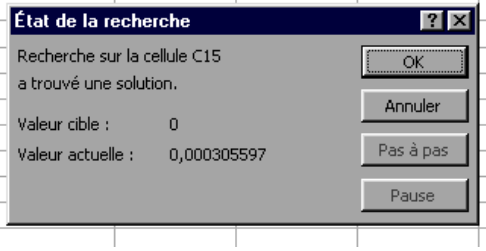
Exemple

Un échantillon aléatoire de 11 temps de bon fonctionnement a donné les valeurs t_i du tableau suivant. On ajuste la loi de T à une loi de *Weibull* de paramètre $\gamma = 0$.

Sur Excel, on peut utiliser l'Outils "Valeur cible…" pour rechercher une valeur de β rendant la quantité $\frac{n}{\beta} + \sum \ln t_i - \frac{n}{\sum t_i^{\beta}} \sum (\ln t_i) t_i^{\beta}$ proche de 0. On peut ensuite affiner cette recherche par tâtonnement en modifiant le contenu de la cellule contenant l'estimation.

C15 = =(11/B2)+C13-(11/D13)*E13

	A	B	C	D	E	F	G
1	ti	estimation béta	ln ti	ti^béta	ti^béta * ln ti		
2	14	3,460119054	2,63905733	9241,435882	24388,6791		
3	19		2,94443898	26585,18582	78278,4574		
4	24		3,17805383	59661,64397	189607,916		
5	26		3,25809654	78700,27363	256413,089		
6	29		3,36729583	114834,181	386680,659		
7	31		3,4339872	144640,1999	496692,596		
8	36		3,58351894	242656,5043	869564,179		
9	40		3,68887945	349396,3002	1288880,83		
10	43		3,76120012	448740,5619	1687803,05		
11	46		3,8286414	566682,8322	2169625,35		
12	48		3,87120101	656590,8707	2541795,24		
13		sommes :	37,5543706	2697729,989	9989730,05		
14							
15	équation de	vraisemblance :	0,0003056				
16							
17							
18							
19							
20							
21							
22							
23							



On peut aussi obtenir une estimation de β selon le maximum de vraisemblance par approximations successives, par exemple en partant d'une valeur β_0 donnée et de la formule de récurrence suivante :

$$\beta_{k+1} = \frac{1}{\dfrac{\sum t_i^{\beta_k} \ln t_i}{\sum t_i^{\beta_k}} - \dfrac{\sum \ln t_i}{n}}.$$

Cela qui conduit (voir le tableau ci-contre), après quelques itérations, à estimer β par la valeur 3,46.

β_k
2
4,82285736
2,98984586
3,74064906
3,33082585
3,52862579
3,42662084
3,47755519
3,45169721
3,46471624
3,45813378
3,46145485
3,45977746
3,46062421
3,46019665
3,46041251
3,46030353
3,46035855
3,46033077
3,4603448
3,46033771
3,46034129

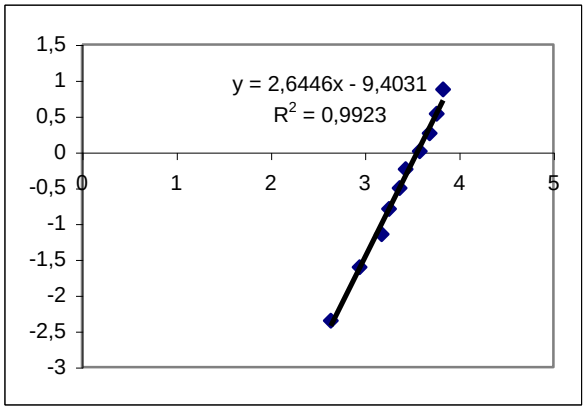
On utilise également, dans ce contexte[7], la méthode de *Newton* d'approximations successives.

Remarque :
La méthode d'estimation ("habituelle") par régression linéaire selon les moindres carrés (ou le papier de *Weibull*), basée sur l'équivalence

$$R(t) = e^{-\left(\frac{t}{\eta}\right)^{\beta}} \Leftrightarrow y = \beta x - \beta \ln \eta$$

où l'on a posé $x = \ln t$ et $y = \ln[- \ln R(t)]$, conduit aux résultats ci-dessous :

t_i	$F(t_i)$	$R(t_i)$	$x_i=\ln(t_i)$	$y_i=\ln(-\ln R(t_i))$
14	0,09090909	0,90909091	2,63905733	-2,35061866
19	0,18181818	0,81818182	2,94443898	-1,60609005
24	0,27272727	0,72727273	3,17805383	-1,14427809
26	0,36363636	0,63636364	3,25809654	-0,79410601
29	0,45454545	0,54545455	3,36729583	-0,50065122
31	0,54545455	0,45454545	3,4339872	-0,23767695
36	0,63636364	0,36363636	3,58351894	0,01153414
40	0,72727273	0,27272727	3,68887945	0,26181256
43	0,81818182	0,18181818	3,76120012	0,53341735
46	0,90909091	0,09090909	3,8286414	0,87459138
48	1	0	3,87120101	#NOMBRE!



Cette méthode, où les valeurs de $F(t_i)$ sont calculées selon les rangs bruts, donne donc comme estimation de β la valeur 2,6.

L'estimation par maximum de vraisemblance présentant un biais, sensible sur de petits échantillons. On peut d'ailleurs affiner le résultat précédent en utilisant la méthode des "rangs médians", où $R(t_i) = 1 - \dfrac{n_i - 0{,}3}{n + 0{,}4}$ (n_i étant le nombre de défaillant), la régression linéaire conduit alors à estimer β par la valeur 2,88 .
Remarque : quand on n'a pas d'expression analytique pour un estimateur (comme c'est le cas ici pour celui de β par le maximum de vraisemblance), la recherche des propriétés de

[7] Pour l'utilisation du maximum de vraisemblance en fiabilité, on pourra consulter : C. Cocozza-Thivent – "Processus stochastiques et fiabilité des systèmes" Ed. Springer 1997.

l'estimateur peut se faire par simulation. On pourrait ainsi corriger le biais de l'estimateur de β par le maximum de vraisemblance.
L'avantage de la méthode du maximum de vraisemblance sera, ici encore, d'obtenir des estimateurs à partir desquels on peut déduire des intervalles de confiance. C'est ce que nous montrons plus loin pour le paramètre η , mais les calculs permettant d'obtenir des intervalles de confiance sur β sont ici trop compliqués.
Les tables de *Johnson* indiquent qu'au niveau de confiance de 90%, pour un échantillon de taille n = 11, il faut prévoir une "erreur" de l'ordre de 35%. Ainsi pour $\hat{\beta} \approx 2{,}6$ on obtient l'intervalle [1,69 ; 3,51] et pour $\hat{\beta} \approx 3{,}46$ l'intervalle [2,24 ; 4,67] .
Mann, *Shafer* et *Singpurwalla* ont, plus précisément, déterminé des formules analytiques et des tables de calcul, utilisant la distribution du khi deux[8].

Estimation du paramètre η par le maximum de vraisemblance

Reprenons les résultats donnés précédemment par la méthode du maximum de vraisemblance. On a comme condition nécessaire : $\frac{n}{\mu} - \sum t_i^{\beta} = 0$ où $\mu = \frac{1}{\eta^{\beta}}$.

Si l'on suppose que β est connu, on a donc une estimation de η par :

$\frac{1}{\eta^{\beta}} = \frac{n}{\sum t_i^{\beta}}$ d'où $\boxed{\eta = \left(\frac{\sum t_i^{\beta}}{n} \right)^{\frac{1}{\beta}}}$.

On a ainsi une expression explicite de l'estimateur $\hat{\eta} = \left(\frac{\sum T_i^{\beta}}{n} \right)^{\frac{1}{\beta}}$ du maximum de vraisemblance (à la différence de $\hat{\beta}$) ce qui facilite l'obtention d'intervalles de confiance (en supposant β connu).

Exemple

On reprend l'exemple précédent, avec β = 2,6.
L'estimateur du maximum de vraisemblance fournit comme estimation de η la valeur : 34,98.

Remarque :
La méthode d'estimation par régression linéaire donnait :
2,6446 ln η = 9,4031 d'où $\eta \approx 35{,}01$.

t_i	*ti* ^ 2,6	estim. η
14	954,845032	34,979425
19	2112,3183	
24	3877,47698	
26	4774,53086	
29	6342,12496	
31	7542,93567	
36	11127,2156	
40	14633,7617	
43	17661,1096	
46	21046,0386	
48	23508,6244	

[8] Voir C. Mascovici, J. C. Ligeron – "Utilisation des techniques de fiabilité en mécanique" Ed. Lavoisier.

Intervalle de confiance pour η

Par rapport à ce que fournit l'estimation par régression linéaire, l'avantage de l'estimateur $\hat{\eta} = \left(\frac{\sum T_i^{\beta}}{n} \right)^{\frac{1}{\beta}}$ obtenu, lorsque β est connu, par le maximum de vraisemblance, est qu'il permet d'obtenir un intervalle de confiance.

On a $\hat{\eta}^{\beta} = \frac{1}{n} \sum T_i^{\beta}$ et l'on va voir que les variables aléatoires $Y_i = T_i^{\beta}$ suivent une loi exponentielle (situation déjà envisagée au paragraphe précédent).

En effet, si T suit une loi de *Weibull* de paramètres $\gamma = 0$, β et η, sa fonction de répartition est donnée par $F(t) = 1 - e^{-\left(\frac{t}{\eta}\right)^{\beta}}$.

On a alors, pour l'expression de la fonction de répartition de la variable aléatoire T^{β} :

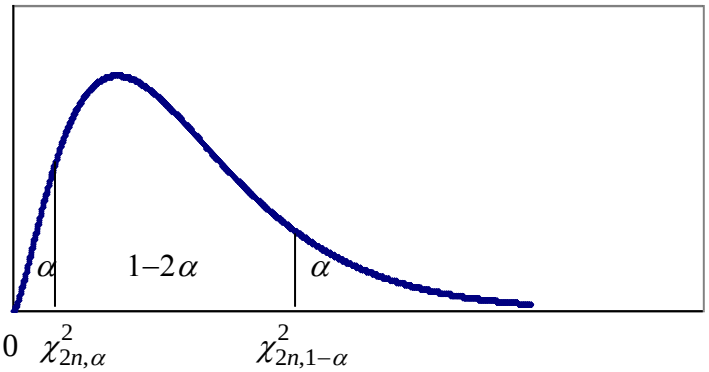
$P(T^{\beta} \leq y) = P(T \leq y^{1/\beta}) = F(y^{1/\beta}) = 1 - e^{-\frac{y}{\eta^{\beta}}}$. Ce qui montre que T^{β} suit une loi exponentielle de paramètre $\lambda = \frac{1}{\eta^{\beta}}$.

Lorsque Y_i suit la loi exponentielle de paramètre λ, la variable aléatoire λY_i suit une loi gamma de paramètre 1. Les Y_i étant indépendantes, $\sum_1^n \lambda Y_i$ suit une loi gamma de paramètre n et $2\lambda \sum_1^n Y_i$ suit une loi du khi 2 à $2n$ degrés de liberté.

On peut alors déterminer un *intervalle de probabilité* pour la variable aléatoire $2\lambda \sum_1^n Y_i$ en utilisant la table du khi 2 :

$$P(\chi^2_{2n,\alpha} \leq 2\lambda \sum Y_i \leq \chi^2_{2n,1-\alpha}) = 1 - 2\alpha$$

densité du khi 2 à 2*n* degrés de liberté



ce qui équivaut à

$$P\left(\frac{\chi^2_{2n,\alpha}}{2\sum Y_i} \leq \lambda \leq \frac{\chi^2_{2n,1-\alpha}}{2\sum Y_i} \right) = 1 - 2\alpha .$$

Revenons alors aux paramètres de *Weibull* avec $\lambda = \frac{1}{\eta^{\beta}}$ et $Y_i = T_i^{\beta}$:

On a donc

$$P\left(\frac{2\sum T_i^{\beta}}{\chi^2_{2n,1-\alpha}} \leq \eta^{\beta} \leq \frac{2\sum T_i^{\beta}}{\chi^2_{2n,\alpha}} \right) = 1 - 2\alpha ,$$

c'est à dire $P\left(\left(\frac{2\sum T_i^{\beta}}{\chi^2_{2n,1-\alpha}} \right)^{1/\beta} \leq \eta \leq \left(\frac{2\sum T_i^{\beta}}{\chi^2_{2n,\alpha}} \right)^{1/\beta} \right) = 1 - 2\alpha$.

Pour les temps de bon fonctionnement t_i obtenus sur l'échantillon de taille n et compte tenu de l'égalité $\sum t_i^{\beta} = n\hat{\eta}^{\beta}$, on prendra comme ***intervalle de confiance*** de η au coefficient de confiance $1 - 2\alpha$: $\boxed{\left[\hat{\eta}\left(\frac{2n}{\chi^2_{2n,1-\alpha}}\right)^{1/\beta} \quad , \quad \hat{\eta}\left(\frac{2n}{\chi^2_{2n,\alpha}}\right)^{1/\beta}\right]}$.

Exemple

On reprend l'exemple précédent, pour lequel on avait $n = 11$ temps de bon fonctionnement avec $\beta = 2{,}6$ et $\hat{\eta} \approx 34{,}98$.

Donnons un intervalle de confiance pour η au niveau de confiance $1 - 2\alpha = 90\ \%$:

$$\left[34{,}98\left(\frac{22}{\chi^2_{22\,;\,0{,}95}}\right)^{1/2{,}6} \quad , \quad 34{,}98\left(\frac{22}{\chi^2_{22\,;\,0{,}05}}\right)^{1/2{,}6}\right].$$

La table du khi 2 donne, pour un χ^2 à 22 degrés de liberté :

$\chi^2_{22\,;\,0{,}05} \approx 12{,}338$ et $\chi^2_{22\,;\,0{,}95} \approx 33{,}924$.

On aura donc comme intervalle de confiance pour η à 90 % : [29,61 ; 43,70] .

Intervalle de confiance pour la MTBF et la fiabilité

Dans le cas où l'on suppose β connu, on obtient facilement des intervalles de confiance pour la M.T.B.F. et la fiabilité, à partir de l'intervalle de confiance $[\eta_1\ ,\ \eta_2]$ calculé pour η.

On sait que la M.T.B.F. est donnée par $E(T) = \eta A$ (lorsque $\gamma = 0$) où $A = \Gamma(1 + \frac{1}{\beta}) = \int_0^{+\infty} x^{1/\beta} e^{-x} dx$ est donné par une table du formulaire du BTS.

On aura donc comme intervalle de confiance pour la MTBF : $\boxed{[A\eta_1\ ,\ A\eta_2]}$.

De même, puisque la fiabilité est donnée par $R(t) = \exp\left(-\left(\frac{t}{\eta}\right)^{\beta}\right)$, on aura comme intervalle de confiance pour la fiabilité à l'instant t : $\boxed{\left[\exp\left(-\left(\frac{t}{\eta_1}\right)^{\beta}\right), \exp\left(-\left(\frac{t}{\eta_2}\right)^{\beta}\right)\right]}$.

Exemple

On reprend les données précédentes où $\beta = 2{,}6$; $\eta_1 = 29{,}61$ et $\eta_2 = 43{,}70$.

La table du formulaire du BTS donne pour $\beta = 2{,}6$ la valeur $A \approx 0{,}8882$ d'où un intervalle de confiance à 90 % (dans le cas où l'on considère β connu) pour la MTBF : [26,2 ; 38,9].

Dans le cas β inconnu, il existe un abaque (norme NF X 06-501) selon la taille n de l'échantillon et le niveau de confiance.

Pour la fiabilité à l'instant $t = 25$ (par exemple), on obtient comme intervalle de confiance à 90 % (en supposant $\beta = 2{,}6$) : [0,525 ; 0,792].

V – TEST D'ADEQUATION AU MODELE DE LA LOI EXPONENTIELLE OU DE WEIBULL

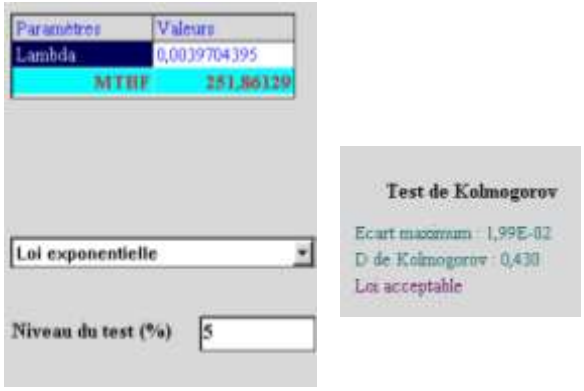
Ce paragraphe constitue un approfondissement et n'est pas "indispensable en première lecture". La question de l'adéquation à un modèle est particulièrement mise en lumière dans les modules "*Statistique inférentielle*" et "*Fiabilité*" des programmes de BTS parus en 2001, où l'on insiste sur les apports des logiciels spécialisés.

Extraits du programme :
"On soulignera que la validité d'une méthode statistique est liée à l'adéquation entre la réalité et le modèle la représentant."
"[…] Dans le cadre de cette liaison, on pourra donner quelques exemples d'autres procédures que celles figurant au programme de mathématiques […], en privilégiant les aspects qualitatifs, mais aucune connaissance à leur sujet n'est exigible dans le cadre de ce programme."
"Des logiciels spécialisés exécutent automatiquement ces calculs, suivant les normes AFNOR ou ISO."

L'objectif de ce paragraphe est d'apporter quelques éléments théoriques concernant le test de *Kolmogorov* mis en œuvre dans les logiciels de fiabilité que les étudiants sont susceptible de rencontrer dans la pratique, en particulier lors des stages. Le professeur de mathématiques pourra alors éclairer les étudiants sur la signification de ces tests.

1 – LA QUESTION DE L'ADEQUATION D'UN MODELE AUX OBSERVATIONS

Il y a plusieurs degrés de sophistication dans l'étude de l'adéquation d'un modèle au données statistiques empiriques. Le premier degré est l'ajustement graphique, "au jugé", utilisant un papier fonctionnel : si les points correspondant aux temps de bon fonctionnement observés sont pratiquement alignés sur le papier semi-logarithmique (respectivement sur le papier de *Weibull*), on considèrera que le modèle de la loi exponentielle (respectivement d'une loi de *Weibull* de paramètre $\gamma = 0$) est adapté. Cette procédure, encore largement utilisée dans l'industrie, est très simple (elle nécessite peu de calculs) mais la qualité de l'ajustement n'est aucunement quantifiée.
Un deuxième degré consiste à intégrer dans les procédures précédentes un ajustement linéaire selon la méthode des moindres carrés. Le coefficient de corrélation linéaire est alors un témoin numérique de la qualité de l'ajustement. En revanche, les risques d'erreur que l'on prend en choisissant tel ou tel modèle ne peuvent être quantifiés en termes de probabilité.
Le dernier degré consiste à mettre en œuvre un test statistique d'hypothèse qui permettra d'évaluer au moins le risque d'erreur de première espèce. Ces procédures sont, du point de vue théorique, très élaborées, mais on les trouve proposées dans les logiciels spécialisés, où tous les calculs sont pris en charge.
Il s'agit alors de comprendre la signification des paramètres affichés par le logiciel[9]



[9] L'image correspond au logiciel "FiabExpert" distribué par la société Knowllence (www.knowllence.com).

(paramètres à choisir ou résultats calculés), de façon à effectuer les choix en connaissance de cause. Sur l'exemple, on doit choisir le "niveau du test" (il s'agit bien sûr du risque de première espèce), puis sont affichés (ici pour un modèle de loi exponentielle) "l'écart maximum" et "D" indiquant si la loi est "acceptable" pour le test mis en œuvre.
Il serait bon de savoir à quoi tout cela correspond.

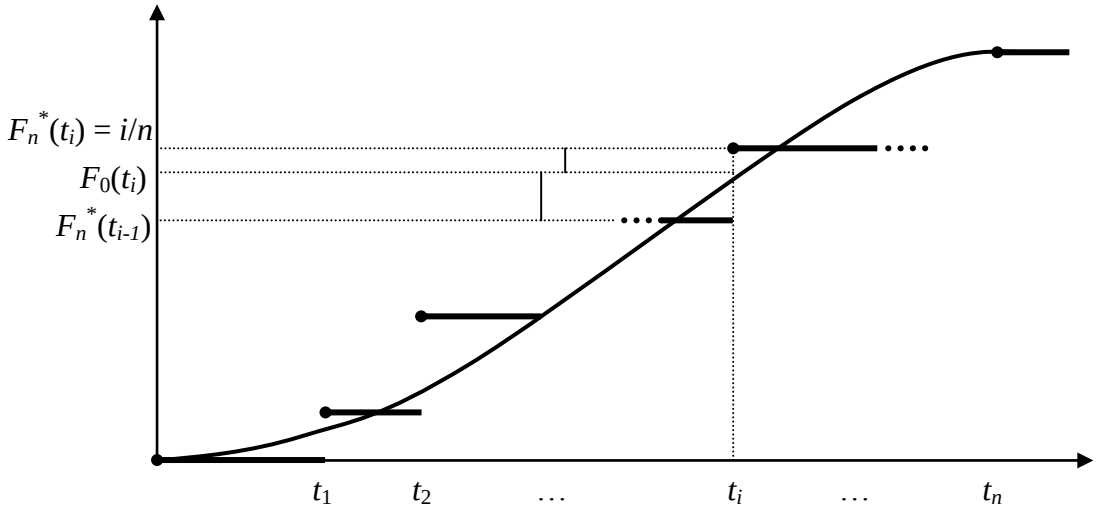
Dans ce qui suit on considère la variable aléatoire *T* associant son temps de bon fonctionnement à tout matériel d'un certain type pris au hasard. On possède un échantillon aléatoire de *n* réalisations indépendantes de *T*, c'est à dire *n* temps de bon fonctionnement $t_1, \ldots, t_n$. On recherche la loi de *T* représentant ces observations.

2 – TEST DE KOLMOGOROV

Ce test s'applique à une variable aléatoire de fonction de répartition continue. Il est par exemple utilisé pour tester la normalité d'une distribution. On l'appliquera ici au problème de l'adéquation aux lois exponentielle et de *Weibull*. Ce test est dû, dans le cas (comme ici) d'un échantillon, à *Kolmogorov* en 1933, et étendu par *Smirnov* au cas de la comparaison entre deux échantillons, en 1939.

Principe du test

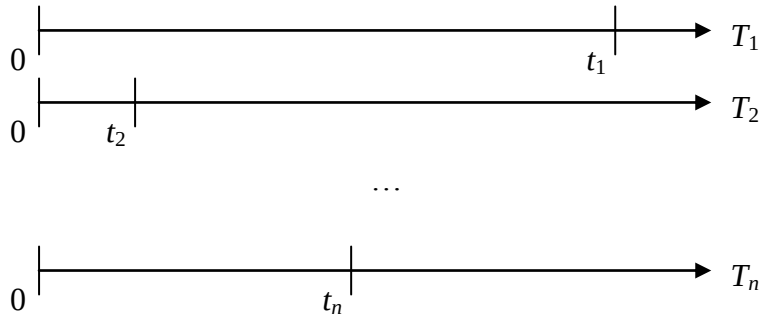
Soit *T* une variable aléatoire de fonction de répartition continue *F*. Le test est fondé sur la distance D_n entre la ***fonction de répartition empirique*** F_n^* observée et la ***fonction de répartition théorique*** F_0 (continue) de la variable aléatoire *T*, sous l'hypothèse H_0 : $F = F_0$ du modèle testé.
On suppose que les *n* valeurs observées de la variable aléatoire *T* (les *n* temps de bon fonctionnement, dans le cadre de la fiabilité) sont ordonnés : $t_1 \le t_2 \le \ldots \le t_n$.



La fonction F_n^* est constante par intervalles, définie par $F_n^*(t) = \begin{cases} 0 \text{ si } t < t_1 \\ \dfrac{i}{n} \text{ si } t \in [t_i\,,\, t_{i+1}[\\ 1 \text{ si } t \ge t_n \end{cases}$.

La distance utilisée est $D_n = \sup_t \left| F_n^*(t) - F_0(t) \right|$ qui est atteinte en l'un des points de discontinuité de F_n^* : $D_n = \max_i \left(\left| F_0(t_i) - F_n^*(t_{i-1}) \right|, \left| F_0(t_i) - F_n^*(t_i) \right| \right)$.

Soit T_1, …, T_n les variables aléatoires indépendantes de même loi que T fournissant les n temps de bon fonctionnement : on peut considérer que l'on teste n matériels identiques, fonctionnant indépendamment jusqu'à défaillance, et que T_i est la variable aléatoire correspondant au temps de bon fonctionnement du $i^{\text{ème}}$ matériel (ici les t_i ne sont pas ordonnés).



On considère la "fonction" aléatoire dont la fonction de répartition empirique (observée) est une réalisation : $\tilde{F}_n^*(t) = \frac{1}{n} \text{ card}\{i / T_i \le t\}$ (pour chaque valeur de $t \ge 0$, $\tilde{F}_n^*(t)$ est une variable aléatoire).

Soit la variable aléatoire $D_n = \sup_t \left| \tilde{F}_n^*(t) - F_0(t) \right|$. On montre que la loi de D_n est indépendante de F_0 et, pour n grand, *Kolmogorov* a démontré le théorème limite suivant :

Sous l'hypothèse que F_0 est la fonction de répartition continue de T, on a :

$$P(\sqrt{n}\, D_n < y) \xrightarrow[n \to +\infty]{} \sum_{k \in Z} (-1)^k e^{-2k^2 y^2} = K(y)$$ (fonction K de *Kolmogorov*).

La loi de D_n est tabulée (même pour n petit – voir page suivante), ce qui permet la détermination de la zone d'acceptation de H_0 pour un certain seuil de risque :

La construction et mise en œuvre du test au seuil de risque α (première espèce) est la suivante :

- Choix des hypothèses :

  H_0 : $F = F_0$ contre H_1 : $F \ne F_0$.
- Recherche de la zone d'acceptation de H_0 :

  Sous l'hypothèse H_0, les probabilités $P(D_n < \delta)$ correspondent à celles de la table de *Kolmogorov-Smirnov*. On y recherche donc la valeur d_n telle que $P(D_n < d_n) = 1 - \alpha$.

  La zone d'acceptation de H_0 sera donc $[0, d_n[$.
- Règle de décision :

  Sur un échantillon de taille n on calcule :

  $$D_n = \max_i \left(\left| F_0(t_i) - F_n^*(t_{i-1}) \right|, \left| F_0(t_i) - F_n^*(t_i) \right| \right).$$

  Si $D_n < d_n$ on accepte l'hypothèse H_0 au seuil α.

  Si $D_n \ge d_n$ on refuse l'hypothèse H_0 au risque α de se tromper.

TABLE DE KOLMOGOROV-SMIRNOV

Valeurs d_n telles que $P = P(D_n < d_n)$

n	$P = .80$	$P = .90$	$P = .95$	$P = .98$	$P = .99$
1	.90000	.95000	.97500	.99000	.99500
2	.68377	.77639	.84189	.90000	.92929
3	.56481	.63604	.70760	.78456	.82900
4	.49265	.56522	.62394	.68887	.73424
5	.44698	.50945	.56328	.62718	.66853
6	.41037	.46799	.51926	.57741	.61661
7	.38148	.43607	.48342	.53844	.57581
8	.35831	.40962	.45427	.50654	.54179
9	.33910	.38746	.43001	.47960	.51332
10	.32260	.36866	.40925	.45662	.48893
11	.30829	.35242	.39122	.43670	.46770
12	.29577	.33815	.37543	.41918	.44905
13	.28470	.32549	.36143	.40362	.43247
14	.27481	.31417	.34890	.38970	.41762
15	.26588	.30397	.33760	.37713	.40420
16	.25778	.29472	.32733	.36571	.39201
17	.25039	.28627	.31796	.35528	.38086
18	.24360	.27851	.30936	.34569	.37062
19	.23735	.27136	.30143	.33685	.36117
20	.23156	.26473	.29408	.32866	.35241
21	.22617	.25858	.28724	.32104	.34427
22	.22115	.25283	.28087	.31394	.33666
23	.21645	.24746	.27490	.30728	.32954
24	.21205	.24242	.26931	.30104	.32286
25	.20790	.23768	.26404	.29516	.31657
26	.20399	.23320	.25907	.28962	.31064
27	.20030	.22898	.25438	.28438	.30502
28	.19680	.22497	.24993	.27942	.29971
29	.19348	.22117	.24571	.27471	.29466
30	.19032	.21756	.24170	.27023	.28987
31	.18732	.21412	.23788	.26596	.28530
32	.18445	.21085	.23424	.26189	.28094
33	.18171	.20771	.23076	.25801	.27677
34	.17909	.20472	.22743	.25429	.27279
35	.17659	.20185	.22425	.25073	.26897
36	.17418	.19910	.22119	.24732	.26532
37	.17188	.19646	.21826	.24404	.26180
38	.16966	.19392	.21544	.24089	.25843
39	.16753	.19148	.21273	.23786	.25518
40	.16547	.18913	.21012	.23494	.25205
41	.16349	.18687	.20760	.23213	.24904
42	.16158	.18468	.20517	.22941	.24613
43	.15974	.18257	.20283	.22679	.24332
44	.15796	.18053	.20056	.22426	.24060
45	.15623	.17856	.19837	.22181	.23798
46	.15457	.17665	.19625	.21944	.23544
47	.15295	.17481	.19420	.21715	.23298
48	.15139	.17302	.19221	.21493	.23059
49	.14987	.17128	.19028	.21277	.22828
50	.14840	.16959	.18841	.21068	.22604

TABLE DE KOLMOGOROV-SMIRNOV

Valeurs d_n telles que $P = P(D_n < d_n)$

n	$P = .80$	$P = .90$	$P = .95$	$P = .98$	$P = .99$
51	.14697	.16796	.18659	.20864	.22386
52	.14558	.16637	.18482	.20667	.22174
53	.14423	.16483	.18311	.20475	.21968
54	.14292	.16332	.18144	.20289	.21768
55	.14164	.16186	.17981	.20107	.21574
56	.14040	.16044	.17823	.19930	.21384
57	.13919	.15906	.17669	.19758	.21199
58	.13801	.15771	.17519	.19590	.21019
59	.13686	.15639	.17373	.19427	.20844
60	.13573	.15511	.17231	.19267	.20673
61	.13464	.15385	.17091	.19112	.20506
62	.13357	.15263	.16956	.18960	.20343
63	.13253	.15144	.16823	.18812	.20184
64	.13151	.15027	.16693	.18667	.20029
65	.13052	.14913	.16567	.18525	.19877
66	.12954	.14802	.16443	.18387	.19729
67	.12859	.14693	.16322	.18252	.19584
68	.12766	.14587	.16204	.18119	.19442
69	.12675	.14483	.16088	.17990	.19303
70	.12586	.14381	.15975	.17863	.19167
71	.12499	.14281	.15864	.17739	.19034
72	.12413	.14183	.15755	.17618	.18903
73	.12329	.14087	.15649	.17498	.18776
74	.12247	.13993	.15544	.17382	.18650
75	.12167	.13901	.15442	.17268	.18528
76	.12088	.13811	.15342	.17155	.18408
77	.12011	.13723	.15244	.17045	.18290
78	.11935	.13636	.15147	.16938	.18174
79	.11860	.13551	.15052	.16832	.18060
80	.11787	.13467	.14960	.16728	.17949
81	.11716	.13385	.14868	.16626	.17840
82	.11645	.13305	.14779	.16526	.17732
83	.11576	.13226	.14691	.16428	.17627
84	.11508	.13148	.14605	.16331	.17523
85	.11442	.13072	.14520	.16236	.17421
86	.11376	.12997	.14437	.16143	.17321
87	.11311	.12923	.14355	.16051	.17223
88	.11248	.12850	.14274	.15961	.17126
89	.11186	.12779	.14195	.15873	.17031
90	.11125	.12709	.14117	.15786	.16938
91	.11064	.12640	.14040	.15700	.16846
92	.11005	.12572	.13965	.15616	.16755
93	.10947	.12506	.13891	.15533	.16666
94	.10889	.12440	.13818	.15451	.16579
95	.10833	.12375	.13746	.15371	.16493
96	.10777	.12312	.13675	.15291	.16408
97	.10722	.12249	.13606	.15214	.16324
98	.10668	.12187	.13537	.15137	.16242
99	.10615	.12126	.13469	.15061	.16161
100	.10563	.12067	.13403	.14987	.16081
$n > 100$	$1.073/\sqrt{n}$	$1.223/\sqrt{n}$	$1.358/\sqrt{n}$	$1.518/\sqrt{n}$	$1.629/\sqrt{n}$

Exemple 1 (loi exponentielle)

On suppose que l'on dispose des $n = 9$ temps de bon fonctionnement t_i ordonnés suivants (d'après BTS maintenance 1995) :
30, 50, 90, 130, 170, 230, 300, 410, 580.
On recherche si une modélisation par la loi exponentielle est envisageable. Une estimation (par régression linéaire), sous cette hypothèse, du coefficient λ a conduit à la valeur $\lambda = 0{,}0040$ (méthode des rangs moyens).

On effectue un test de *Kolmogorov* au seuil α de 5%.

• Choix des hypothèses :

H_0 : pour tout $t \geq 0$, $P(T \leq t) = F_0(t) = 1 - e^{-0{,}004\,t}$
contre H_1 : il existe $t \geq 0$ tel que $P(T \leq t) \neq F_0(t)$.

• Recherche de la zone d'acceptation de H_0 au seuil de 5%.

On recherche dans la table de *Kolmogorov-Smirnov* la valeur d_n, pour $n = 9$, telle que, sous l'hypothèse H_0, $P(D_9 < d_9) = 0{,}95$. On trouve $d_9 = 0{,}43001$.

• On calcule avec les 9 temps de bon fonctionnement de l'échantillon la valeur de $D_9 = \max_i \left(\left| F_0(t_i) - F_n^*(t_{i-1}) \right|, \left| F_0(t_i) - F_n^*(t_i) \right| \right)$ où la fonction de répartition empirique sera calculée selon la méthode des rangs moyens.

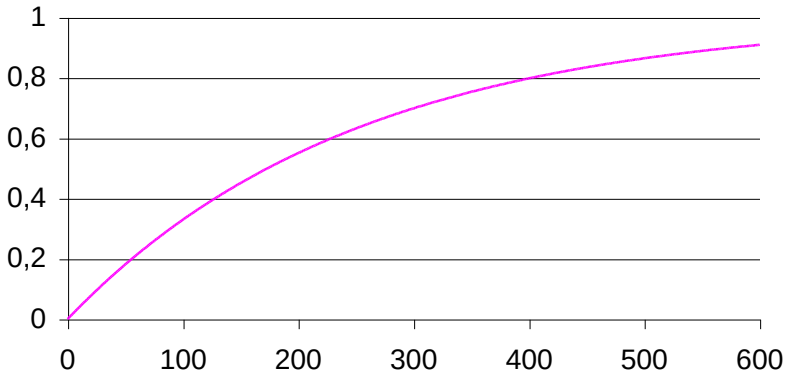
On a obtenu sur Excel :

t_i	F_9*(t_i) rangs moyens	$F_0(t_i)$	$\left\|F_0(t_i) - F_9^*(t_i)\right\|$	$\left\|F_0(t_i) - F_9^*(t_{i-1})\right\|$
30	0,1	0,11307956	0,01307956	
50	0,2	0,18126925	0,01873075	0,08126925
90	0,3	0,30232367	0,00232367	0,10232367
130	0,4	0,40547945	0,00547945	0,10547945
170	0,5	0,49338301	0,00661699	0,09338301
230	0,6	0,60148096	0,00148096	0,10148096
300	0,7	0,69880579	0,00119421	0,09880579
410	0,8	0,80601996	0,00601996	0,10601996
580	0,9	0,90172641	0,00172641	0,10172641
			D_9	0,10601996

L'écart $D_9 \approx 0{,}106$ calculé sur l'échantillon est inférieur à $d_9 = 0{,}43001$.
On accepte donc, au seuil de risque de 5%, l'hypothèse H_0 selon laquelle T suit une loi exponentielle de paramètre 0,004.

Remarque :

On a représenté ici les fonctions de répartition empiriques F_9^* (en escaliers) et théorique F_0.



Le logiciel FiabExpert affiche les résultats ci-contre, pour un test de Kolmogorov à 5%.
Ce qui est nommé "D de Kolmogorov" est la valeur limite pour l'acceptation de H_0, fournie par la table.

Paramètres	Valeurs
Lambda	0,0039704395
MTBF	251,86129

Test de Kolmogorov

Ecart maximum : 1,99E-02
D de Kolmogorov : 0,430
Loi acceptable

Ce qui est nommé "Ecart maximum" est la valeur $\max\left|F_9^*(t_i) - F_0(t_i)\right|$ calculée avec λ = 0,00397 comme le montre la vérification suivante effectuée sur Excel : cette valeur est 0,019959771.

C2 = =LOI.EXPONENTIELLE(A2;0,00397;VRAI)

	A	B	C	D
1	ti	F*(ti) rangs moyens	F(ti)	abs(F*(ti)-F(ti))
2	30	0,1	0,11228098	0,012280976
3	50	0,2	0,18004023	**0,019959771**
4	90	0,3	0,3004374	0,000437403
5	130	0,4	0,40315629	0,003156295
6	170	0,5	0,49079266	0,009207339
7	230	0,6	0,59872167	0,001278331
8	300	0,7	0,69608281	0,003917195
9	410	0,8	0,80361927	0,003619269
10	580	0,9	0,90000149	1,49069E-06
11				

Le logiciel semble ainsi sous estimer l'écart D_9 entre les fonctions de répartition empirique et théorique, du moins lorsque la fonction de répartition empirique F_n^* est définie comme constante par morceaux. De ce fait, elle a tendance à s'éloigner de F après chaque valeur t_i , comme on le voit sur la figure(on pourrait aussi la définir comme affine par morceaux).

Il semble que, dans la pratique, on prenne plus souvent $D_n = \max_i\left(\left|F_0(t_i) - F_n^*(t_i)\right|\right)$ plutôt que $D_n = \max_i\left(\left|F_0(t_i) - F_n^*(t_{i-1})\right|, \left|F_0(t_i) - F_n^*(t_i)\right|\right)$ qui , du fait de la définition de la fonction de répartition empirique en escaliers, a tendance à surestimer l'écart au modèle.

Exemple 2 (loi de Weibull)

Prenons l'exemple donné à l'épreuve du BTS Maintenance Nouvelle Calédonie 1995.
On dispose des 10 temps de bon fonctionnement suivants :
71 ; 78 ; 84 ; 90 ; 96 ; 104 ; 110 ; 120 ; 130 ; 145 .
En utilisant la méthode des rangs moyens et le papier de *Weibull*, on a estimé que la variable aléatoire T suit une loi de *Weibull* de paramètres $\gamma = 60$; $\beta = 1{,}6$ et $\eta = 50$.

On effectue un test de *Kolmogorov* au seuil α de 5%.
• Choix des hypothèses :
H_0 : pour tout $t \geq 0$, $P(T \leq t) = F_0(t) = 1 - e^{-\left(\frac{t-60}{50}\right)^{1,6}}$
contre H_1 : il existe $t \geq 0$ tel que $P(T \leq t) \neq F_0(t)$.
• Recherche de la zone d'acceptation de H_0 au seuil de 5%.
On recherche dans la table de *Kolmogorov-Smirnov* la valeur d_n, pour $n = 10$, telle que, sous l'hypothèse H_0, $P(D_{10} < d_{10}) = 0{,}95$. On trouve $d_{10} = 0{,}40925$.
• On calcule avec les 10 temps de bon fonctionnement de l'échantillon la valeur de $D_{10} = \max_i \left(\left| F_0(t_i) - F_n^*(t_{i-1}) \right|, \left| F_0(t_i) - F_n^*(t_i) \right| \right)$ où la fonction de répartition empirique sera obtenue selon la méthode des rangs moyens.
On a obtenu sur Excel :

ti	F*(ti) rangs moyens	F_0(ti)	abs(F*(ti)-F_0(ti))	abs(F*(ti-1)-F_0(ti))
71	0,090909091	0,084871086	0,006038005	
78	0,181818182	0,17718359	0,004634592	0,086274499
84	0,272727273	0,265833596	0,006893677	0,084015414
90	0,363636364	0,357001671	0,006634693	0,084274398
96	0,454545455	0,446335677	0,008209778	0,082699313
104	0,545454545	0,557372856	0,01191831	0,102827401
110	0,636363636	0,632120559	0,004243078	0,086666013
120	0,727272727	0,73781915	0,010546423	0,101455514
130	0,818181818	0,819709771	0,001527953	0,092437044
145	0,909090909	0,903413909	0,005677	0,085232091
			max :	max :
			0,01191831	**0,102827401**

L'écart $D_{10} \approx 0{,}103$ calculé sur l'échantillon est inférieur à $d_{10} = 0{,}40925$.
On accepte donc, au seuil de risque de 5%, l'hypothèse H_0 selon laquelle T suit une loi de *Weibull* de paramètres $\gamma = 60$; $\beta = 1{,}6$ et $\eta = 50$.

On donne ci-contre l'affichage du logiciel FiabExpert sur cet exemple.
De même que pour l'exemple précédent, seul l'écart $\max \left| F_{10}^*(t_i) - F_0(t_i) \right|$ est pris en compte, ce qui semble correspondre à la pratique en entreprise.

Paramètres	Valeurs
Beta	1,6210059
Eta	50,923402
Gamma	59,03511
MTBF	104,63855

Test de Kolmogorov
Ecart maximum : 1,29E-02
D de Kolmogorov : 0,409
Loi acceptable

Travaux pratiques STATISTIQUES

T.P. Excel : AJUSTEMENT A UNE LOI EXPONENTIELLE OU A UNE LOI DE WEIBULL

Objectifs

- Utiliser la régression linéaire selon les moindres carrés pour ajuster une distribution observée à un modèle exponentiel ou de *Weibull*.
- Calculer une estimation des paramètres caractéristiques de la loi.

I – AJUSTEMENT A UNE LOI EXPONENTIELLE

(D'après BTS Maintenance 1995.)

Une équipe a relevé durant une année les temps de fonctionnement, en heures, entre deux réglages consécutifs d'une machine de conditionnement et a obtenu les temps de bon fonctionnement, rangés en ordre croissant, suivants :
30 ; 50 ; 90 ; 130 ; 170 ; 230 ; 300 ; 410 ; 580 .
On souhaite, à partir de cet historique, modéliser les durées de bon fonctionnement par une loi exponentielle.

1 – REGRESSION LINEAIRE

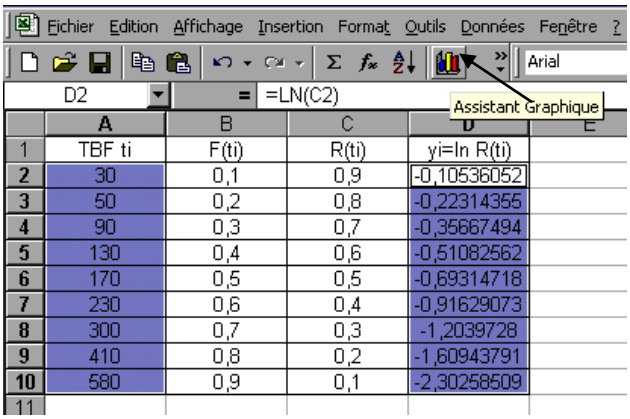
Lancer Excel.
Préparer le tableau ci-contre.

	A	B	C	D
1	TBF ti	F(ti)	R(ti)	yi=ln R(ti)
2	30			
3	50			
4	90			
5	130			
6	170			
7	230			
8	300			
9	410			
10	580			
11				

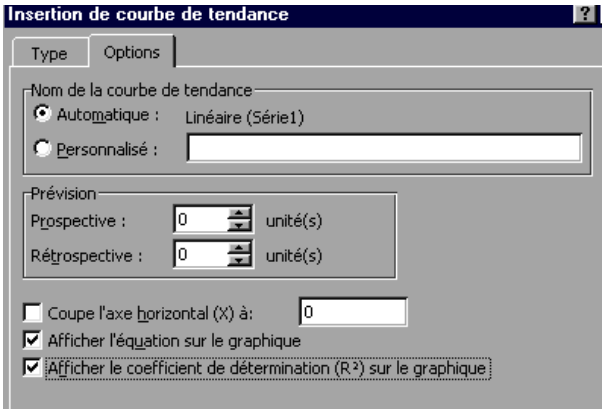
On va déterminer les fréquences de défaillances $F(t_i)$ par la méthode des rangs moyens :
En ***B2*** entrer la valeur 0,1.
En ***B3*** entrer la ***formule*** = B2 + 0,1
Recopier vers le bas (on approche le pointeur de la souris du coin inférieur droit de la cellule B2, quant celui-ci se transforme en croix noire, on glisse, en appuyant sur le bouton gauche de la souris) jusqu'en ***B10***.
En ***C2*** entrer la ***formule*** = 1 – B2
Recopier vers le bas jusqu'en ***C10***.
En ***D2*** entrer la ***formule*** = LN(C2)
Recopier vers le bas jusqu'en ***D10***.

Lorsque la variable aléatoire T correspondant au temps de fonctionnement suit une loi exponentielle de paramètre λ, on a $R(t) = e^{-\lambda t}$.
En prenant le logarithme, on a l'équivalence : $R(t) = e^{-\lambda t} \Leftrightarrow \ln R(t) = -\lambda t$.
Si le modèle exponentiel est adapté, on devrait donc avoir des points (t_i , y_i) , avec $y_i = \ln R(t_i)$, pratiquement alignés sur la droite d'équation $y = -\lambda t$.
Vous allez procéder à un ajustement linéaire selon la méthode des moindres carrés, sur le nuage de points (t_i , y_i) .

Sélectionner les cellules de ***A2*** à ***A10*** puis, en appuyant en même temps sur la touche ***CTRL***, ***sélectionner*** les cellules de ***D2*** à ***D10***. Cliquer alors sur l'icône ***Assistant Graphique***.
A l'étape 1/4, choisir ***Nuage de points*** puis cliquer sur ***Terminer***.
Avec le bouton droit de la souris, cliquer sur la légende du graphique et choisir ***Effacer***.



	A	B	C	D
1	TBF ti	F(ti)	R(ti)	yi=ln R(ti)
2	30	0,1	0,9	-0,10536052
3	50	0,2	0,8	-0,22314355
4	90	0,3	0,7	-0,35667494
5	130	0,4	0,6	-0,51082562
6	170	0,5	0,5	-0,69314718
7	230	0,6	0,4	-0,91629073
8	300	0,7	0,3	-1,2039728
9	410	0,8	0,2	-1,60943791
10	580	0,9	0,1	-2,30258509
11				



Avec le bouton droit de la souris, cliquer sur un point du graphique et choisir ***Ajouter une courbe de tendance…***
Dans l'onglet ***Type*** choisir ***Linéaire***. Dans l'onglet ***Options*** cocher ***Afficher l'équation sur le graphique*** et ***Afficher le coefficient de détermination (R^2) sur le graphique***.

– ***Compléter la feuille réponse***.

2 – ESTIMATION DES PARAMETRES

L'ajustement linéaire précédent incite à choisir le modèle exponentiel.
En déterminer le paramètre sur la feuille réponse.

– ***Compléter la feuille réponse***.

II – AJUSTEMENT A UNE LOI DE WEIBULL

(D'après BTS Maintenance Nouvelle Calédonie 1995.)

Une machine fabrique des pièces cylindriques en grande série.
Le parc de l'atelier comporte 10 machines fonctionnant dans les mêmes conditions.
Afin d'étudier la fiabilité de ces machines, on relève le nombre de jours de bon fonctionnement avant la première défaillance. Les résultats sont :
110 ; 104 ; 78 ; 145 ; 130 ; 90 ; 120 ; 96 ; 71 ; 84.
On désigne par *T* la variable aléatoire qui, à tout machine de ce type, associe sa durée de vie.

1 – REGRESSION LINEAIRE ET ESTIMATION DE γ

Cliquer sur l'onglet ***Feuil2*** puis préparer le tableau ci-dessous :

A1 = TBF ti

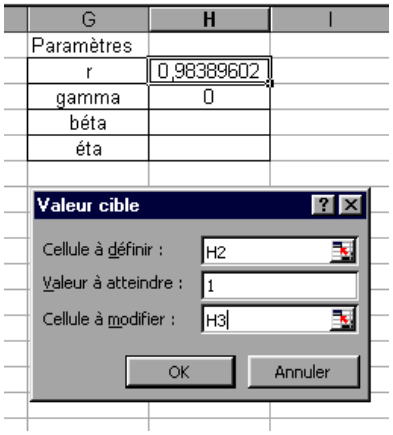
	A	B	C	D	E	F	G	H
1	TBF ti	F(ti)	R(ti)	ln(ti - g)	ln(-ln(R(ti))		Paramètres	
2	71						r	
3	78						gamma	
4	84						béta	
5	90						éta	
6	96							
7	104							
8	110							
9	120							
10	130							
11	145							

Dans la colonne B, on calcule les fréquences de défaillance selon la méthode des rangs moyens.
En ***B2*** entrer la ***formule*** = 1/11
En ***B3*** entrer la ***formule*** = B2 + 1/11 puis ***recopier vers le bas*** jusqu'en ***B11***.
En ***C2*** entrer la ***formule*** = 1 – B2 puis ***recopier vers le bas*** jusqu'en ***C11***.
On va provisoirement donner au paramètre γ la valeur 0.
En ***H3*** entrer la valeur 0.
En ***D2*** entrer la ***formule*** = LN(A2 – H$3) , puis ***recopier vers le bas*** jusqu'en ***D11*** (le symbole $ permet de conserver la référence 3 lors du recopiage).
En ***E2*** entrer la ***formule*** = LN(– LN(C2)) , puis ***recopier vers le bas*** jusqu'en ***E11***.

Lorsque T suit une loi de *Weibull* de paramètres γ, β, η, on a $R(t) = e^{-\left(\frac{t-\gamma}{\eta}\right)^{\beta}}$.
Or, en prenant deux fois le logarithme, on a l'équivalence :

$$R(t) = e^{-\left(\frac{t-\gamma}{\eta}\right)^{\beta}} \Leftrightarrow \ln(-\ln(R(t))) = \beta\ln(t-\gamma) - \beta\ln(\eta) \Leftrightarrow y = \beta x - \beta\ln(\eta)$$

en posant $y = \ln(-\ln(R(t)))$, calculés dans la colonne E, et $x = \ln(t-\gamma)$, calculés dans la colonne D.
On est donc amené à rechercher γ de sorte à avoir une régression linéaire "correcte" de y en x.
En ***H2*** entrer la ***formule*** = COEFFICIENT.CORRELATION(E2:E11;D2:D11)
Il s'agit de déterminer γ de sorte que ce coefficient de corrélation soit le plus proche possible de 1.



Dans le menu ***Outils*** choisir ***Valeur cible…***
Remplir la boîte de dialogue ainsi :
Cellule à définir : H2
Valeur à atteindre : 1
Cellule à modifier : H3
Puis cliquer sur ***OK***.

Excel à recherché une valeur de γ de sorte à améliorer l'alignement des points de coordonnées (x_i, y_i).
Arrondir au mieux le résultat en prenant pour γ , dans la cellule ***H3***, une valeur entière.

– ***Compléter la feuille réponse***.

2 – ESTIMATION DE β ET η – CALCUL DE LA M.T.B.F.

Le paramètre de forme β correspond à la pente de la droite de régression de y en x.
En ***H4*** entrer la ***formule*** =PENTE(E2:E11;D2:D11)

Le paramètre η est tel que : $b = -\beta \ln \eta \Leftrightarrow \eta = e^{-\frac{b}{\beta}}$.
En ***H5*** entrer la ***formule*** =EXP(– ORDONNEE.ORIGINE(E2:E11;D2:D11)/H4)

On sait que la MTBF est alors donnée par MTBF = $\gamma + \eta A$ où le nombre A est fourni par le formulaire de l'examen pour la valeur de β trouvée. Cette valeur A est également donnée par l'instruction EXP(LNGAMMA(1+1/β)) d'Excel.
En ***G7*** taper : MTBF .
En ***H7*** entrer la ***formule*** = H3 + H5*EXP(LNGAMMA(1+1/H4))

– ***Compléter la feuille réponse***.

– FEUILLE REPONSE

NOMS : *.....*

I – AJUSTEMENT A UNE LOI EXPONENTIELLE

1 – REGRESSION LINEAIRE

Quelle est l'équation de régression linéaire $y = a\,t + b$ affichée par Excel ?

.....

En déduire, à l'aide cette équation, une expression de $R(t)$ sous la forme $R(t) = e^{a\,t + b}$.

.....

Calculer le coefficient de corrélation linéaire r (c'est ici l'opposé de la racine du coefficient de détermination R^2) : .....

Comment peut-on interpréter la valeur de r ? .....

2 – ESTIMATION DES PARAMETRES

Donner une valeur approchée à 10^{-2} près de e^{-b} : .....

En prenant comme valeur approchée $e^{-b} = 1$, donner l'expression de $R(t)$:

.....

En déduire qu'on peut considérer que T suit une loi exponentielle.

En donner le paramètre λ :

.....

.....

Calculer la M.T.B.F. :

.....

II – AJUSTEMENT A UNE LOI DE WEIBULL

1 – REGRESSION LINEAIRE ET ESTIMATION DE γ

Quelle est la valeur entière de γ pour laquelle le coefficient de corrélation linéaire r est le plus proche de 1 ?

On trouve $\gamma =$ ……

Quelle est alors la valeur de r ? .....

En déduire qu'on peut considérer que T suit une loi de *Weibull*.

.....

.....

2 – ESTIMATION DE β ET η – CALCUL DE LA M.T.B.F.

Les estimations obtenues sont les suivantes :

$\beta =$ ……

$\eta =$ ……

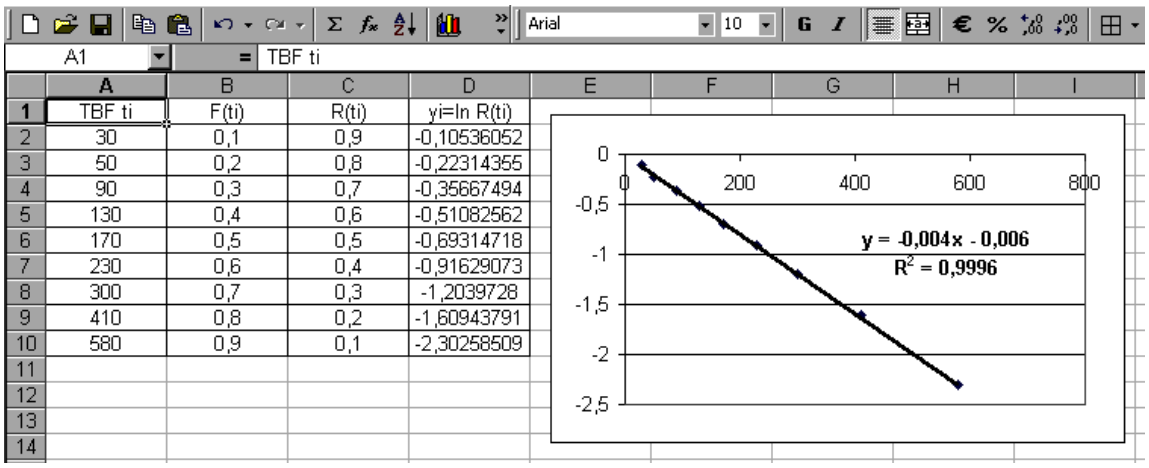
M.T.B.F. = …….

Corrigé de l'activité EXCEL "AJUSTEMENT A UNE LOI EXPONENTIELLE OU A UNE LOI DE WEIBULL"

I – AJUSTEMENT A UNE LOI EXPONENTIELLE

1 – REGRESSION LINEAIRE

On obtient les résultats suivants :



A1 = TBF ti

	A	B	C	D
1	TBF ti	F(ti)	R(ti)	yi=ln R(ti)
2	30	0,1	0,9	-0,10536052
3	50	0,2	0,8	-0,22314355
4	90	0,3	0,7	-0,35667494
5	130	0,4	0,6	-0,51082562
6	170	0,5	0,5	-0,69314718
7	230	0,6	0,4	-0,91629073
8	300	0,7	0,3	-1,2039728
9	410	0,8	0,2	-1,60943791
10	580	0,9	0,1	-2,30258509
11				
12				
13				
14				

D'après l'ajustement linéaire, on a $R(t) = e^{-0,004\,t - 0,006}$.
Le coefficient de corrélation linéaire est $r \approx -0,9998$.
Comme r est, en valeur absolue, très proche de 1, on peut dire que le nuage de point est pratiquement aligné.

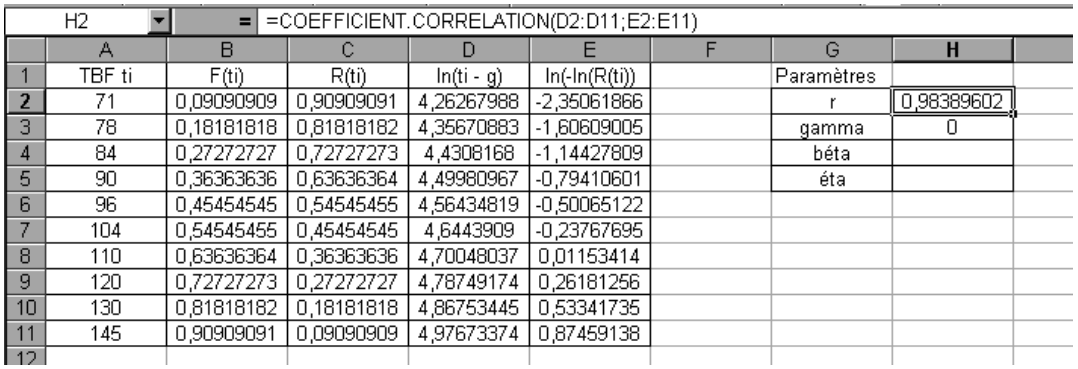
2 – ESTIMATION DES PARAMETRES

On a $e^{-0,006} \approx 0,99$.
En prenant $e^{-0,006} = 1$, on a $R(t) = e^{-0,004\,t}$.
On peut donc considérer que T suit la loi exponentielle de paramètre $\lambda = 0,004$.
On a M.T.B.F. $= 1/\lambda = 1/0,004 \approx 250$ heures.

II – AJUSTEMENT A UNE LOI DE WEIBULL

1 – REGRESSION LINEAIRE ET ESTIMATION DE γ

Pour $\gamma = 0$, on obtient $r \approx 0,98$.



H2 = =COEFFICIENT.CORRELATION(D2:D11;E2:E11)

	A	B	C	D	E	F	G	H
1	TBF ti	F(ti)	R(ti)	ln(ti - g)	ln(-ln(R(ti))		Paramètres	
2	71	0,09090909	0,90909091	4,26267988	-2,35061866		r	0,98389602
3	78	0,18181818	0,81818182	4,35670883	-1,60609005		gamma	0
4	84	0,27272727	0,72727273	4,4308168	-1,14427809		béta	
5	90	0,36363636	0,63636364	4,49980967	-0,79410601		éta	
6	96	0,45454545	0,54545455	4,56434819	-0,50065122			
7	104	0,54545455	0,45454545	4,6443909	-0,23767695			
8	110	0,63636364	0,36363636	4,70048037	0,01153414			
9	120	0,72727273	0,27272727	4,78749174	0,26181256			
10	130	0,81818182	0,18181818	4,86753445	0,53341735			
11	145	0,90909091	0,09090909	4,97673374	0,87459138			
12								

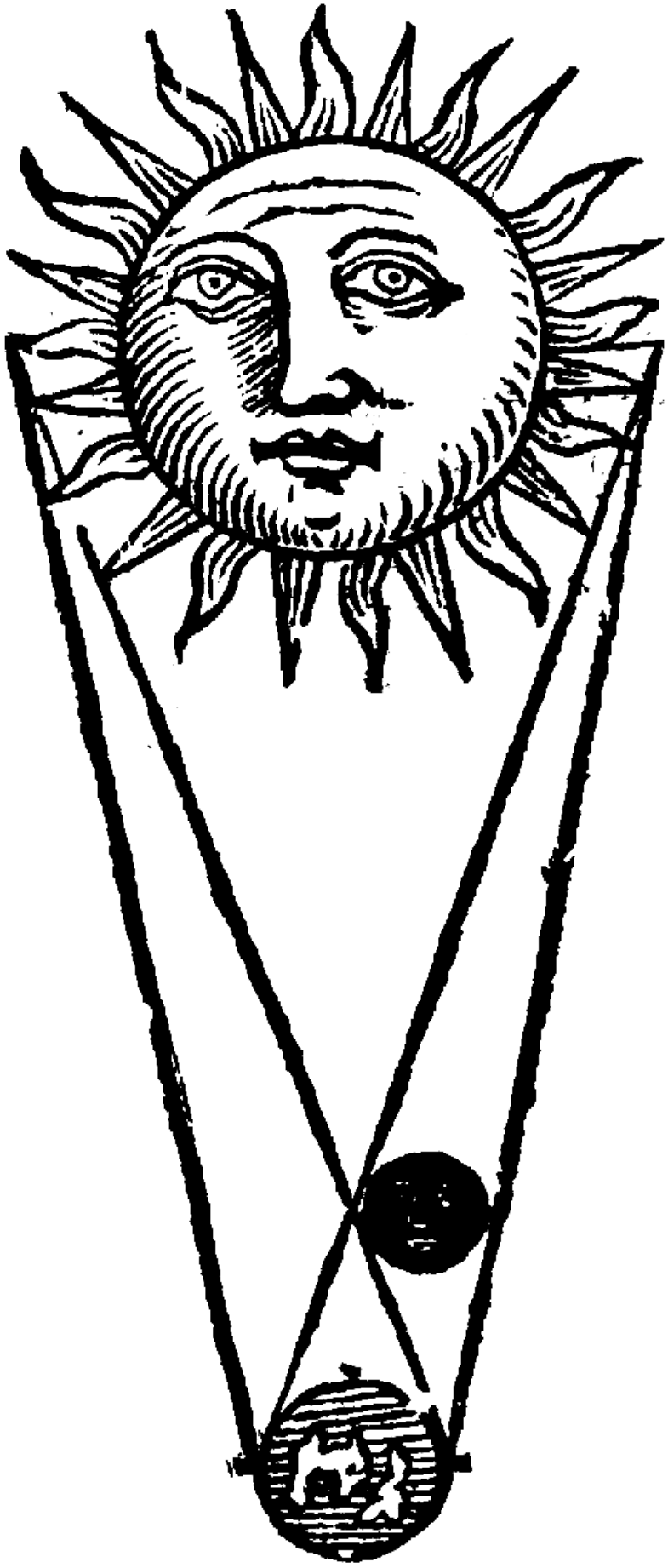
La valeur optimale de γ est $\gamma = 60$. On a alors $r \approx 0,9997$.
Comme r est très proche de 1, on peut en déduire qu'il est raisonnable de considérer que T suit une loi de *Weibull* de paramètre $\gamma = 60$.

2 – ESTIMATION DE β ET η – CALCUL DE LA M.T.B.F.

Les résultats obtenus sont les suivants :

H7 = =H3+H5*EXP(LNGAMMA(1+1/H4))

	A	B	C	D	E	F	G	H
1	TBF ti	F(ti)	R(ti)	ln(ti - g)	ln(-ln(R(ti))		Paramètres	
2	71	0,09090909	0,90909091	2,39789527	-2,35061866		r	0,99974507
3	78	0,18181818	0,81818182	2,89037176	-1,60609005		gamma	60
4	84	0,27272727	0,72727273	3,17805383	-1,14427809		béta	1,56665662
5	90	0,36363636	0,63636364	3,40119738	-0,79410601		éta	49,8672964
6	96	0,45454545	0,54545455	3,58351894	-0,50065122			
7	104	0,54545455	0,45454545	3,78418963	-0,23767695		**MTBF**	**104,800741**
8	110	0,63636364	0,36363636	3,91202301	0,01153414			
9	120	0,72727273	0,27272727	4,09434456	0,26181256			
10	130	0,81818182	0,18181818	4,24849524	0,53341735			
11	145	0,90909091	0,09090909	4,44265126	0,87459138			
12								



Epreuves corrigées de B.T.S.

Annales du B.T.S. : LOI EXPONENTIELLE LOI DE WEIBULL

LOI EXPONENTIELLE

1 – Systèmes constructifs bois et habitat 1998

Un système déclenche automatiquement, en fonction de la température et de la vitesse du vent, l'ouverture et la fermeture d'un store extérieur protégeant la vitrine d'un magasin.
Une étude statistique a permis d'obtenir les valeurs suivantes de la fonction de fiabilité $t \to R(t)$ du système où t désigne le nombre de jours depuis l'installation de celui-ci.

t	90	200	360	500	750	1000	1200	1500
$R(t)$	0,89	0,76	0,61	0,51	0,36	0,25	0,19	0,13
$\ln[R(t)]$								

a) Reproduire et compléter le tableau précédent. On donnera les valeurs décimales arrondies à 10^{-3} de $\ln[R(t)]$.

b) Tracer le nuage de points correspondant à la série statistique $(t, \ln[R(t)])$. En abscisse, 100 jours seront représentés par 1 centimètre et, en ordonnée, 1 unité sera représentée par 5 centimètres.

c) Donner une équation de la droite d'ajustement $y = m\,t + p$ obtenue par la méthode des moindres carrés et tracer celle-ci sur le graphique précédent. Les valeurs décimales arrondies à 10^{-6} près des coefficients m et p peuvent être obtenues directement à l'aide d'une calculatrice.
Montrer qu'en utilisant cette équation on obtient : $R(t) = 1{,}003413 \times e^{-0{,}001374\,t}$.

d) On admet pour la suite que $R(t) = e^{-0{,}001374\,t}$; ainsi la fiabilité du système suit une loi exponentielle.
Donner une valeur approchée à 10^{-6} près du taux d'avarie du système, puis la MTBF (moyenne du temps de bon fonctionnement) au jour près.

e) Calculer, à 10^{-2} près, la probabilité de voir le système tomber en panne pendant l'année de garantie, c'est à dire avant 365 jours.

2 – Informatique de gestion 1999

On considère une production de composants d'un certain type. On admet que la variable aléatoire T qui, tout composant tiré au hasard dans la production, associe sa durée de vie t, exprimée en heures, suit une loi exponentielle de paramètre λ.

1.a) On note R la fonction de fiabilité ; donner l'écriture de $R(t)$ en fonction de λ et de t.
b) On admet que $R(600) = 0{,}93$. Calculer la valeur exacte de λ. Donner sa valeur décimale arrondie à 10^{-5} près, lue sur la calculatrice.

2. On prendra dans cette question : $\lambda = 0,0001$.
a) Déterminer la MTBF de T (moyenne des temps de bon fonctionnement).
b) Calculer, à 10^{-2} près, la probabilité $P(T > 1500)$.

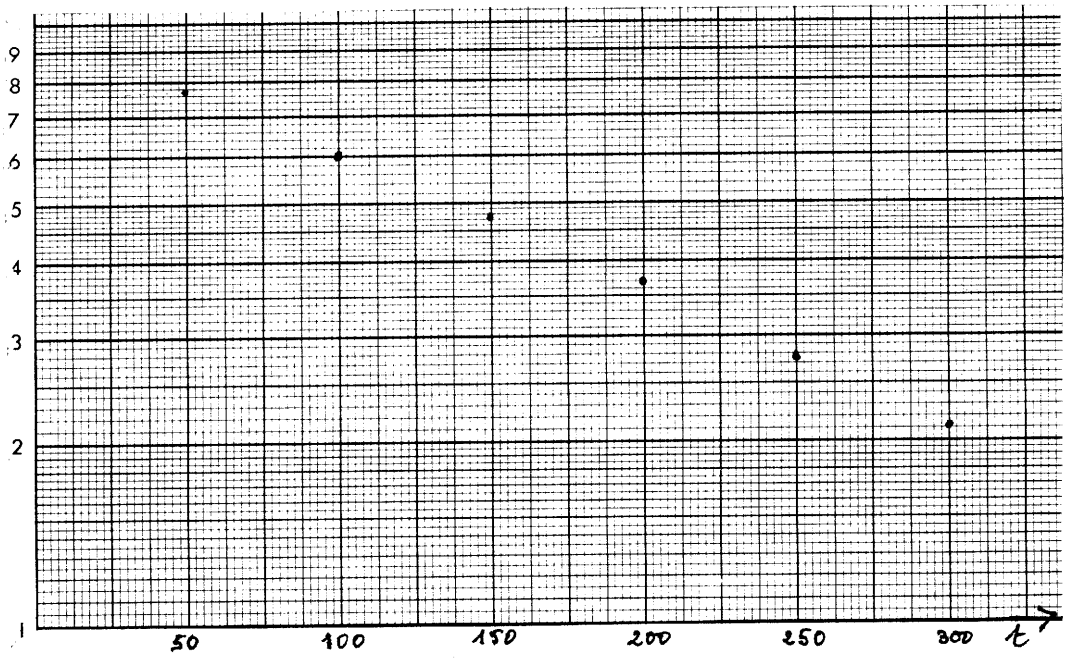
3 – Domotique 1996

On s'intéresse à un type de constituant fonctionnant dans un appareil, en moyenne, trois heures par jour. On note T la variable aléatoire qui, à tout constituant aléatoire, associe sa durée de vie en jour et on note pour $t > 0$, $R(t) = P(T > t)$.
L'étude des temps de bon fonctionnement d'un lot de constituants fabriqués par une machine a permis de placer 6 points de coordonnées $(t_i, R(t_i))$ sur du papier semi-logarithmique, voir annexe.

1° La variable T suit une loi exponentielle de paramètre λ :
- Expliquer pourquoi.
- Estimer λ en se servant du graphique.
- En déduire $R(t)$ en fonction de t.

2°A l'aide du graphique, déterminer successivement la probabilité qu'un constituant fonctionne :
- plus de 100 jours ;
- moins de 50 jours.
Retrouver ces résultats par le calcul, en utilisant l'expression de $R(t)$ trouvée au 1°.

ANNEXE :



4 – Maintenance 1995

L'équipe de maintenance a relevé durant une année les temps de fonctionnement, en heures, entre deux réglages consécutifs d'une machine de conditionnement et a obtenu les temps de bon fonctionnement, rangés en ordre croissant, suivants :

30 ; 50 ; 90 ; 130 ; 170 ; 230 ; 300 ; 410 ; 580 .

1° A l'aide de la méthode des rangs moyens compléter le tableau suivant :

TBF t_i	$F(t_i)$	$R(t_i)$	$y_i = \ln R(t_i)$
.....	.....	.....	.....

2° A l'aide de la calculatrice, déterminer une équation $y = at + b$ de la droite d'ajustement des valeurs de y à celles de t, ainsi que le coefficient de corrélation entre t et y.

3° En prenant pour valeurs approchées : $a = -0{,}004$ et $b = 0$ donner l'expression de $R(t)$. En déduire la loi suivie par la variable aléatoire mesurant la durée de bon fonctionnement. Calculer la MTBF .

4° Déterminer par le calcul la périodicité de réglage systématique basée sur une fiabilité de 80 % .

5 – Domotique 1993

La commande d'un portail automatique est composée de trois éléments : une commande manuelle à infrarouges type plip, un récepteur et un vérin électrique.
Pour étudier la fiabilité du système on a relevé le nombre de jours de bon fonctionnement avant panne et on a obtenu le tableau suivant :

t_i	100	250	450	600	750	900	1100	1400
$R(t_i)$	0,86	0,69	0,51	0,41	0,32	0,26	0,19	0,12

où t_i désigne le nombre de jours et $R(t_i)$ la valeur prise par la fonction de fiabilité à la date t_i.

a) Tracer le nuage de points $M_i(t_i ; R(t_i))$ sur du papier semi-logarithmique.

b) On admet que la fiabilité du système suit une loi exponentielle.
En déduire graphiquement la MTBF, puis le taux d'avarie.

c)Calculer la probabilité de voir le système tomber en panne durant l'année de garantie.

6 – *Utiliser le papier semi-log ou la calculatrice que préférez-vous ?*

On a relevé, durant une période de 1500 heures la durée de vie de 24 éléments identiques, mis en service à la même heure.
On a obtenu les résultats suivants :

Durée de vie en heures	[0, 100]	]100, 200]	]200, 300]	]300, 400]	]400, 500]
Nombre de défaillants	5	4	3	3	2
Durée de vie en heures	[500, 600]	]600, 750]	]750, 1000]	]1000, 1500]	
Nombre de défaillants	1	2	2	2	

1° En utilisant la méthode des rangs moyens compléter le tableau :

TBF t_i	$F(t_i)$ en %	$R(t_i)$ en %	$y_i = \ln R(t_i)$
.....	.....	.....	.....

2° Tracer le nuage de points $M(t_i ; R(t_i))$ sur du papier semi-logarithmique, en déduire que la variable aléatoire qui mesure la durée de vie des éléments suit une loi exponentielle. Déterminer graphiquement la MTBF.

3° En déduire le paramètre et l'écart type de cette loi exponentielle ainsi que l'expression de $R(t)$.

4° Déterminer graphiquement et par le calcul le pourcentage d'éléments de ce type encore en service au bout de 900 heures et à quel instant t_0 la fiabilité d'un élément est de 50 % .
Etant donné qu'à l'instant t_0 , la fiabilité d'un élément est de 50 %, déterminer, à cet instant t_0 , celle d'un système de deux éléments montés en " parallèle " si l'on admet que les deux éléments fonctionnent de façon indépendante.

5° On monte en série 2 éléments. Montrer que la variable aléatoire qui mesure la durée de vie de ce type de système suit une loi exponentielle dont on déterminera le paramètre. (On admettra encore l'indépendance du fonctionnement des deux éléments).

6° A l'aide de la calculatrice, déterminer une équation $y = a\,t + b$ de la droite d'ajustement des valeurs de y à celles de t, ainsi que le coefficient de corrélation entre t et y .
En déduire l'expression de $R(t)$ et comparer le paramètre de cette loi avec celui obtenu au 3°.

7 – Groupement C 1999 (Analyse)

On appelle f la fonction définie sur $\mathbb{R}$ par : $f(t) = e^{-\frac{t^3}{10^9}}$.
1°a) Démontrer que f est une fonction décroissante.
Déterminer sa limite en $+\infty$ et interpréter géométriquement ce résultat.
b) Déterminer la limite de f en $-\infty$.

c) Tracer soigneusement la courbe représentative de f dans un repère orthogonal pour t variant de 0 à 1500 (échelle : 1 cm pour 100 unités sur l'axe des abscisses et 10 cm pour une unité sur l'axe des ordonnées).

2°a) Résoudre algébriquement dans $\mathbb{R}$ l'équation $f(t) = 0{,}5$; donner la valeur exacte de la solution, puis sa valeur approchée arrondie à l'unité.
b) En déduire l'ensemble des solutions de l'inéquation $f(t) < 0{,}5$.

3° On appelle T la variable aléatoire associant à toute machine d'un certain type sa durée, en heures, de fonctionnement sans panne.
On admet que, pour t réel positif ou nul, $f(t)$ représente la probabilité que T soit supérieur à t ainsi $P(T > t) = f(t)$.
a) Calculer la probabilité qu'une telle machine fonctionne plus de 1000 heures sans panne.
b) Pourquoi peut-on affirmer qu'il y a plus de neuf chances sur dix qu'une telle machine fonctionne sans panne plus de 400 heures ?

8 – Maintenance Nouvelle Calédonie 1996 (Analyse)

1) Calcul d'intégrales
Calculer en fonction du nombre réel positif t, les intégrales suivantes :

a) $F(t) = \dfrac{1}{200}\displaystyle\int_0^t \mathrm{e}^{-0{,}005x}\,\mathrm{d}x$;

$J(t) = \dfrac{1}{200}\displaystyle\int_0^t x\,\mathrm{e}^{-0{,}005x}\,\mathrm{d}x$;

$K(t) = \dfrac{1}{200}\displaystyle\int_0^t x^2\,\mathrm{e}^{-0{,}005x}\,\mathrm{d}x$.

(On pourra utiliser des intégrations par parties pour calculer $J(t)$ et $K(t)$).

2) Interprétation en probabilités
Soit la fonction f définie par : $\begin{cases} f(x) = 0 & \text{pour} \quad x < 0, \\ f(x) = \dfrac{1}{200}\mathrm{e}^{-0{,}005x} & \text{pour} \quad x = 0. \end{cases}$

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
a) Montrer que la fonction de fiabilité associée à variable aléatoire T est définie sur $[0, +\infty[$ par $R(t) = \mathrm{e}^{-0{,}005\,t}$.
b) Calculer à 10^{-3} près : $P(T > 300)$, $P(T > 100)$ et $P(100 < T \leq 300)$.
c) Calculer la valeurs entière approchée de t_0 telle que $P(T \leq t_0) = 0{,}1$.

9 – Maintenance et exploitation des matériels aéronautiques 1994

Un exemple de loi exponentielle

A) λ étant un nombre réel strictement positif on considère la fonction f définie sur $\mathbb{R}$ de la manière suivante : $\begin{cases} \text{si } x<0 \text{ alors } f(x)=0 \\ \text{si } x=0 \text{ alors } f(x)=\lambda\, e^{-\lambda x}. \end{cases}$

Soit X la variable aléatoire réelle qui admet la fonction f pour densité de probabilité ; on dit que X suit la loi exponentielle de paramètre λ.

1°Déterminer la fonction de répartition F de la variable aléatoire X définie sur [0, +∞[par $F(x) = \int_0^x f(t)\,dt$ et représenter graphiquement F sur [0, +∞[.

2° a étant un nombre réel positif fixé on note : $I(a) = \int_0^a t\, f(t)\,dt$.

a) Démontrer à l'aide d'une intégration par parties que : $I(a) = \frac{1}{\lambda} + e^{-\lambda a}(-a - \frac{1}{\lambda})$.

b) Calculer la limite de $I(a)$ quand a tend vers +∞. Que représente cette limite ?

3° t et h étant deux nombres réels strictement positifs démontrer que la probabilité conditionnelle $P(X > t + h / X > t)$ ne dépend que de h.
On dit que X est sans "mémoire".

B) 1° On appelle Y_t la variable aléatoire qui, à toute période de durée t exprimée en heures, associe le nombre d'appels reçus par un standard téléphonique pendant cette période de durée t. On admet que Y_t suit la loi de Poisson de paramètre λt.
Déterminer la probabilité que Y_t soit nulle.

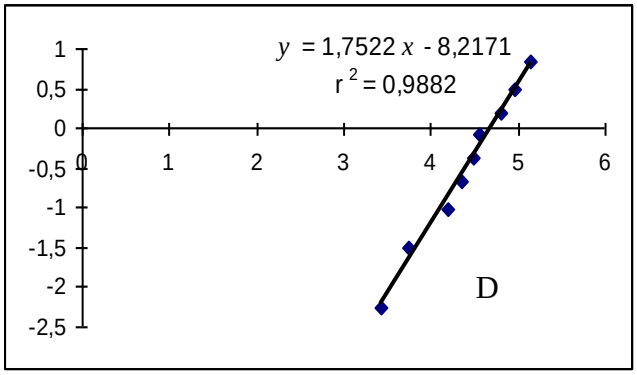
2°Ayant choisi une origine des temps on appelle X la variable aléatoire prenant pour valeur le temps d'attente du premier appel.
a) Exprimer en fonction de λ et de t la probabilité $P(X > t)$, puis la probabilité $P(X \leq t)$.
b) En déduire que X suit une loi exponentielle.

LOI DE WEIBULL

10 – Maintenance 2000

Maintenance du système électronique des armoires de contrôle :
Le service de maintenance préconise, pour les armoires de contrôle, des interventions préventives (par changement de certains éléments électroniques). La période de ces interventions sera déterminée à partir d'un historique de pannes d'une armoire de contrôle choisie au hasard.
Les neuf premiers temps de bon fonctionnement (en jours) de cette armoire de contrôle sont les suivants (rangés en ordre croissant) : 31 ; 42 ; 67 ; 77 ; 89 ; 95 ; 122 ; 144 ; 173 .
Soit T la variable aléatoire qui, à toute armoire de contrôle, associe son temps de bon fonctionnement. On cherche à ajuster la loi de T à une loi de Weibull.
A l'aide d'un tableur, on a obtenu le tableau et le graphique ci-dessous, où $F(t_i)$ et $R(t_i)$ correspondent respectivement à défaillance et à la fiabilité au temps t_i (selon la méthode des rangs moyens) :

t_i	$F(t_i)$	$R(t_i)$	$x_i=\ln(t_i)$	$y_i=\ln[-\ln R(t_i)]$
31	0,1	0,9	3,43398720	-2,25036733
42	0,2	0,8	3,73766962	-1,49993999
67	0,3	0,7	4,20469262	-1,03093043
77	0,4	0,6	4,34380542	-0,67172699
89	0,5	0,5	4,48863637	-0,36651292
95	0,6	0,4	4,55387689	-0,08742157
122	0,7	0,3	4,80402104	0,18562676
144	0,8	0,2	4,96981330	0,47588500
173	0,9	0,1	5,15329159	0,83403245



Sur le graphique ci-dessus, figure la droite de régression D de y en x, obtenue par la méthode des moindres carrés, avec son équation dans un repère orthogonal, ainsi que le carré r^2 du coefficient de corrélation linéaire.

On admet que $\boxed{R(t) = e^{-\left(\frac{t}{\eta}\right)^{\beta}} \text{ équivaut à } y = \beta x - \beta \ln \eta}$,
où l'on a posé $x = \ln t$ et $y = \ln[-\ln R(t)]$.

1. Déduire des informations précédentes les résultats ci-dessous :
a) Le nuage des points de coordonnées (x_i , y_i) est correctement ajusté par cette droite D ;
b) On peut considérer que T suit une loi de Weibull de paramètre $\gamma = 0$;
c) On peut prendre, pour les deux autres paramètres, $\beta = 1{,}75$ (arrondi au centième) et $\eta = 109$ (arrondi à l'unité).
(On pourra utiliser l'équivalence encadrée ci-dessus.)
2. Déterminer, par le calcul, la périodicité d'interventions préventives basée sur une fiabilité de 80%.

11 – Maintenance 1998

Un distributeur automatique élabore du jus d'orange en mélangeant de l'eau et du concentré d'orange.

Une étude de fiabilité de ce type de distributeur a permis d'obtenir le tableau suivant où t_i est le temps de bon fonctionnement exprimé en jours et $F(t_i)$ le pourcentage cumulé de distributeurs hors d'usage à l'instant t_i.

t_i (en jours)	46	48	55	60	64	68	80
$F(t_i)$ (en %)	12,5	25	37,5	50	62,5	75	87,5

1° Placer les points de coordonnées $(t_i, F(t_i))$ sur le papier de Weibull fourni en annexe.
Expliquer pourquoi cette distribution peut être ajustée par une loi de Weibull dont on déterminera les paramètres. Calculer la MTBF.

2° Déterminer par **le calcul** et **graphiquement** la périodicité d'un entretien systématique reposant sur une **fiabilité** de 90 % .

12 – Maintenance 1997

Un tour à commande numérique fabrique en grande série des cylindres.
Dans un premier temps le service de maintenance décide d'établir une carte de contrôle qui permette de décider du moment où il est nécessaire de régler le tour.
Dans un deuxième temps le service de maintenance décide d'utiliser cette carte de contrôle pour relever les temps écoulés entre deux réglages successifs du tour et obtenir ainsi un fichier historique permettant de déterminer une périodicité de réglage systématique basée sur une fiabilité de 90 % .

Détermination de la périodicité de réglage systématique .
L'équipe de maintenance a relevé pendant plusieurs mois les temps de fonctionnement, en heures, entre deux réglages consécutifs du tour et a établi le tableau suivant :

t_I (en heures)	10	11	12	13	14	15	16	17	18	19	20
$F(t_I)$ (en %)	10	15	23	30	40	50	63	74	84	92	95

où t_I représente le temps de fonctionnement entre deux réglages consécutifs et $F(t_I)$ le pourcentage cumulé de réglages effectués avant le temps t_I.

1° Placer les points de coordonnées $(t_I, F(t_I))$ sur le papier de Weibull fourni en annexe .
Expliquer pourquoi cette distribution peut être ajustée par une loi de Weibull dont on déterminera les paramètres. Donner l'expression de $R(t)$.

2° Calculer la MTBF et la probabilité de ne pas avoir de réglage à faire avant cette MTBF.

3° Déterminer **graphiquement** et par **le calcul** la périodicité de réglage systématique basée sur une fiabilité de 90 % .

13 – Maintenance Nouvelle Calédonie 1995

Une machine fabrique des pièces cylindriques en grande série .
Le parc de l'atelier comporte 10 machines fonctionnant dans les mêmes conditions.
Afin d'étudier la fiabilité de ces machines, on relève le nombre de jours de bon fonctionnement avant la première défaillance. Les résultats sont :

110 ; 104 ; 78 ; 145 ; 130 ; 90 ; 120 ; 96 ; 71 ; 84 .

On désigne par T la variable aléatoire qui, à toute machine de ce type , associe sa durée de vie.

1° En utilisant la méthode des rangs moyens et à l'aide du papier graphique de Weibull fourni en annexe,
a) vérifier que la variable aléatoire T suit une loi de Weibull de paramètre $\gamma = 60$;
b) déterminer les deux autres paramètres de cette loi ;
c) déterminer à quel instant t_0 la fiabilité d'une telle machine est de 70 %.

2° Vérifier par le calcul le résultat obtenu dans la question 1) c) .

3° Calculer la M.T.B.F .

14 – Maintenance Nouvelle Calédonie 1998

La société qui vous emploie utilise une pièce OS 117 sur une de ses machines.
Le fabricant des pièces OS 117 affirme que leur durée de vie moyenne est de 600 heures.
Vous devez contrôler l'affirmation du fabricant et étudier la fiabilité de ces pièces car leur coût est important et le délai de livraison est d'un mois.
Vous possédez l'historique de panne des 36 dernières pièces déjà utilisées.
On désigne par X la variable aléatoire qui, à chaque pièce OS 117, prélevée au hasard dans la production, associe sa durée de vie exprimée en heures.

Dans un deuxième temps vous décidez d'étudier à l'aide de l'historique précédent la fiabilité des pièces OS 117. Cela vous a permis de constater que la variable aléatoire X suit approximativement la loi de Weibull de paramètres γ = 100 heures, β = 3 et η = 500 heures.

1° Déterminer la MTBF et l'écart type de cette variable aléatoire, donner l'expression de $R(t)$.

2° Déterminer par le calcul, à quel instant t, la pièce OS 117 a une fiabilité égale à 0,9..

3° La pièce OS 117 fonctionne 16 heures par jour (y compris les jours fériés). Dès la défaillance de cette pièce on la remplace par la seule pièce en stock et on commande immédiatement une autre pièce. Le délai de livraison est de 30 jours.

Quelle est la probabilité que la pièce OS 117 tombe en panne avant l'arrivée de la pièce de rechange ?

15 – Maintenance 1994

On étudie la durée de vie d'un certain type de composants électriques fabriqués par une usine.
On désigne par T la variable aléatoire qui à chaque composant, prélevé au hasard dans la production , associe sa durée de vie exprimée en mois.
Après une étude statistique, on admet que T suit la loi de Weibull de paramètres : $\gamma = 0$; $\beta = 2{,}4$; $\eta = 50$.

1° Représenter sur le papier de Weibull fourni en annexe la fonction de défaillance F correspondante.

2° Déterminer graphiquement à 1% près les probabilités des événements suivants :
a) " la durée de vie d'un composant est inférieure à 10 mois " ;
b) " la durée de vie d'un composant est comprise entre 10 mois et 50 mois ".

3° Déterminer par le calcul, puis à l'aide du graphique, le temps au bout duquel un composant doit être changé, sachant que sa probabilité de survie doit rester supérieure à 90%.
Comparer les deux résultats.

4° Un système (S) est constitué de deux composants du type précédent, montés en série et fonctionnant de manière indépendante (le système (S) est donc défaillant dès qu'un de ses composants l'est). Déterminer le temps au bout duquel (S) doit être changé, sachant que la probabilité de survie de (S) doit rester supérieure à 90%.

16 – Maintenance Nouvelle Calédonie 1995

Une usine fabrique des engrenages. Le service de maintenance a relevé leurs durées de vie en usure accélérée. les résultats sont consignés dans le tableau ci-dessous :

Durée de vie (heures)	250	350	400	510	550	600	750	800
$F(t_i)$ en pourcentages	3	11	18	40	50	63	91	96

$F(t_i)$ est le pourcentage d'engrenages hors service à la date t_i.

1° A l'aide du papier de Weibull justifier que la variable aléatoire qui prend pour valeurs la durée de vie des engrenages peut être ajustée par une loi de Weibull. Déterminer les paramètres de cette loi.

2° a) Calculer la M.T.B.F. et l'écart type de cette loi de Weibull.
b) Déterminer par le calcul au bout de combien de temps, 5% des engrenages sont défectueux. Vérifier graphiquement le résultat obtenu.

3°Une transmission mécanique comporte une série de trois engrenages identiques dont les durées de vie suivent la loi précédente et fonctionnent de façon indépendante.
Quelle est la probabilité que la durée de vie d'un tel système soit au moins de 300 heures ?

17 – *Utiliser le papier de Weibull ou la calculatrice que préférez-vous ?*

A) Une usine utilise 19 machines de même modèle. L'étude du bon fonctionnement en heures, avant la première panne de chacune de ces 19 machines, a permis d'obtenir l'historique suivant :

TBF en heures	[0, 250]	]250, 450]	]450, 600]	]600, 800]	]800, 1100]	]1100,1400]
Nombre de pannes	1	2	2	3	4	4

Trois autres machines ont fonctionné correctement au moins jusqu'à la date : 1400 heures.

1° Déterminer à l'aide du papier de Weibull les paramètres de la loi de Weibull ajustant cette distribution. Donner l'expression de $R(t)$.

2° Calculer la MTBF et l'écart type de cette loi.

3° Déterminer graphiquement puis par le calcul la périodicité d'un entretien systématique basé sur une fiabilité de 0,9.

4° Déterminer graphiquement et par le calcul, la probabilité qu'une machine de ce type fonctionne plus de 2000 heures sans panne.

5° Donner l'expression du taux d'avarie que peut-on dire de sa représentation graphique ?.

6° On pose $u_i = \ln t_i$ et $v_i = \ln(-\ln R(t_i))$
a) Montrer par la méthode des moindres carrés, à l'aide de la calculatrice, que les points $M_i(u_i, v_i)$ peuvent être ajustés par une droite d'équation $v = a\,u + b$.
Donner le coefficient de corrélation entre u et v.
b) En prenant des valeurs approchées à 10^{-1} près des coefficients montrer que l'équation peut s'écrire : $v = 2\,u - 14$.
En déduire une expression de $R(t)$ sous la forme : $R(t) = e^{-(\frac{t}{\eta})^{\beta}}$

B) On modifie l'énoncé de la partie A de la manière suivante :
Une usine utilise 19 machines de même modèle. L'étude du bon fonctionnement en heures, avant la première panne de chacune de ces 19 machines, a permis d'obtenir l'historique suivant :

TBF en heures	[1000, 1250]	]1250, 1450]	]1450, 1600]	]1600, 1800]	]1800, 2100]	]2100, 2400]
Nombre de pannes	1	2	2	3	4	4

Trois autres machines ont fonctionné correctement au moins jusqu'à la date : 2400 heures.

1° Placer sur le même papier de Weibull que celui de la partie A les points $N_i\,(t_i\,,\,F(t_i))$, vérifier que $\gamma = 1000$. Donner l'expression de $R(t)$.

2° Calculer la MTBF et l'écart type de cette loi.

3° Déterminer graphiquement puis par le calcul la périodicité d'un entretien systématique basé sur une fiabilité de 0,9.

4° Déterminer graphiquement et par le calcul, la probabilité qu'une machine de ce type fonctionne plus de 2000 heures sans panne.

5° Donner l'expression du taux d'avarie que peut-on dire de sa représentation graphique ?

18 – Maintenance 1997 (Analyse : influence du paramètre de forme)

La variation du paramètre de forme β permet d'ajuster une grande quantité de distributions expérimentales à une loi de Weibull. On désigne par f_β la fonction de densité et par F_β la fonction de défaillance de la variable aléatoire T suivant une telle loi.
On désigne par C_β la courbe représentative de f_β dans un repère orthogonal $(O ; \vec{i}, \vec{j})$ les unités graphiques étant de 5 cm sur l'axe des abscisses et de 10 cm sur l'axe des ordonnées.
Dans cet exercice on considère les deux cas $\beta = 1$ et $\beta = 2$ lorsque les deux autres paramètres de la loi de Weibull sont $\eta = 1$ et $\eta = 0$.

1° <u>Etude du cas $\beta = 1$:</u>

On rappelle que la fonction f_1 et la fonction F_1 sont alors définies sur [0, +∞ [par :

$f_1(t) = e^{-t}$ et $F_1(t) = \int_0^t f_1(x)\, dx$.

a) Construire la courbe C_1 dans le repère $(O ; \vec{i}, \vec{j})$.
On ne demande pas l'étude de la fonction f_1.
b) Calculer, en fonction de t , l'intégrale $F_1(t)$ et en déduire la probabilité $P(T \leq 1)$ à 10^{-3} près.

c) Calculer $I(t) = \int_0^t x\, f_1(x)\, dx$ **.** (On pourra effectuer une intégration par parties) .

Déterminer l'espérance mathématique $E(T) = \lim_{t \to +\infty} I(t)$.

2° <u>Etude du cas $\beta = 2$:</u>

On rappelle que la fonction f_2 et la fonction F_2 sont alors définies sur [0, +∞ [par :

$f_2(t) = 2t\, e^{-t^2}$ et $F_2(t) = \int_0^t f_2(x)\, dx$.

a) On admet que $\lim_{t \to +\infty} f_2(t) = 0$. Etudier le sens de variation de f_2.

b) En posant $u = -t^2$ et en utilisant le développement limité à l'ordre 1 de e^u au voisinage de 0, démontrer que le développement limité à l'ordre 3 de f_2 au voisinage de 0 est défini par : $f_2(t) = 2t - 2t^3 + t^3 \varepsilon(t)$ avec $\lim_{t \to +\infty} \varepsilon(t) = 0$.

En déduire une équation de la tangente D à C_2 au point d'abscisse 0, et la position relative de D et de C_2 au voisinage de ce point.

c) Construire la tangente D et la courbe C_2 dans le repère $(O ; \vec{i}, \vec{j})$ utilisé au 1°.

d) Démontrer que $F_2(t) = 1 - e^{-t^2}$ et en déduire la valeur de t telle que $P(T \leq t) = 0,05$.

Corrigé des exercices d'épreuves de BTS

1 – SYSTEME CONSTRUCTIFS BOIS ET HABITATS 98

a)

t	90	200	360	500	750	1000	1200	1500
$R(t)$	0,89	0,76	0,61	0,51	0,36	0,25	0,19	0,13
$\ln[R(t)]$	-0,117	-0,274	-0,494	-0,673	-1,022	-1,386	-1,661	-2,040

b)

c) A l'aide de la calculatrice on obtient :
$r = -0,999858$ à 10^{-6} près.
Une équation de la droite de régression de y en t est : $y = -0,001374\,t - 0,003407$.
On en déduit à 10^{-6} près :
$R(t) = e^{-0,001374\,t - 0,003407}$,
$R(t) = e^{-0,001374\,t}\ e^{-0,003407}$,
$R(t) = 1,003413 \times e^{-0,001374\,t}$.

d) On admet $R(t) = e^{-0,001374\,t}$.
Nous sommes en présence d'une loi exponentielle donc le taux d'avarie est

$\lambda = \mathbf{0,001374}$ et MTBF $= \frac{1}{\lambda}$

MTBF $= \frac{1}{0,001374}$, MTBF $\approx$ **728 jours** .

e) La probabilité de tomber en panne avant 365 jours est $F(365) = 1 - R(365)$,
$F(365) = 1 - e^{-0,001374 \times 365}$, $\boldsymbol{F(365) \approx 0,3944}$.

2 – INFORMATIQUE DE GESTION 99

1) a) $R(t) = e^{-\lambda t}$.
b) $R(600) = 0,93$ équivaut à $e^{-600\lambda} = 0,93$
et à $-600\,\lambda = \ln 0,93$, $\lambda = -\frac{\ln 0,93}{600}$,
$\lambda = 0,00012$ à 10^{-5} près.

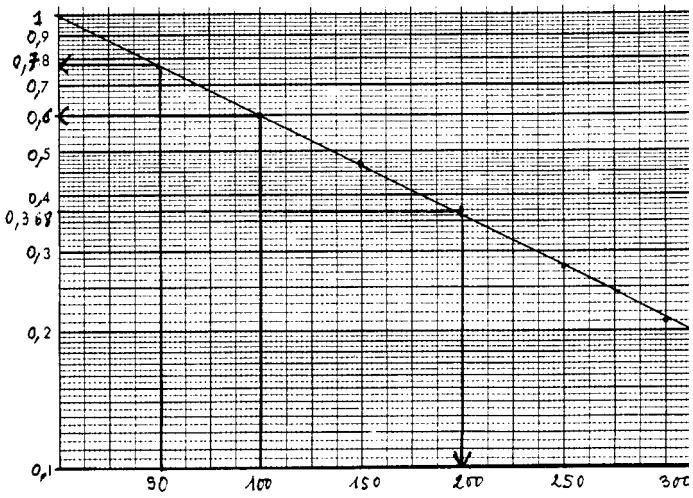
2) a) $R(t) = e^{-0,0001\,t}$, MTBF $= \frac{1}{\lambda}$,
MTBF $= \frac{1}{0,0001}$, MTBF = 10 000 heures.
b) $P(T > 1500) = R(1500)$,
$P(T > 1500) = 0,86$ à 10^{-2} près.

3 - DOMOTIQUE 96

1° Sur le papier semi-logarithmique les points sont sensiblement alignés, la droite d'ajustement passe par le point de coordonnées (0, 1). La droite de régression de $\ln R(t)$ en t a une équation de la forme : $\ln R(t) = a\,t + b$ avec pour $t = 0$, $R(t) = 1$ donc $b = 0$, $R(t) = e^{at}$, l'approximation par une loi exponentielle est donc justifiée.
Si λ désigne le paramètre de cette loi exponentielle $R(t) = e^{-\lambda t}$, $R(\frac{1}{\lambda}) = e^{-1}$ donc $R(\frac{1}{\lambda}) = 0,368$, d'après le graphique, $R(t) = 36,8$ % pour $t \approx 200$ jours, $\frac{1}{\lambda} = 200$, $\lambda \approx \mathbf{0,005}$, $\boldsymbol{R(t) = e^{-0,005\,t}}$.

2° La probabilité qu'un constituant fonctionne :
- plus de 100 jours est $P(T > 100) = R(100)$;
- moins de 50 jours est $P(T < 50) = 1 - R(50)$.

Graphiquement on trouve : $\boldsymbol{R(100) \approx 0,60}$;
$1 - R(50) \approx 1 - 0,78$ soit $\approx$ **0,22**.
Par le calcul : $R(100) = e^{-0,005 \times 100}$,
$R(100) = e^{-0,5}$, $\boldsymbol{R(100) \approx 0,606.}$
$1 - R(50) = 1 - e^{-(0,005)(50)}$,
$1 - R(50) = e^{-0,25}$, $1 - R(50) \approx$ **0,221**.



4 - MAINTENANCE 95

1°

t_i	$F(t_i)$	$R(t_i)$	$\ln R(t_i)$
30	0,1	0,9	– 0,10536
50	0,2	0,8	– 0,22314
90	0,3	0,7	– 0,356675
130	0,4	0,6	– 0,510825
170	0,5	0,5	– 0,69315
230	0,6	0,4	– 0,91629
300	0,7	0,3	– 1,20397
410	0,8	0,2	– 1,609438
580	0,9	0,1	– 2,302585

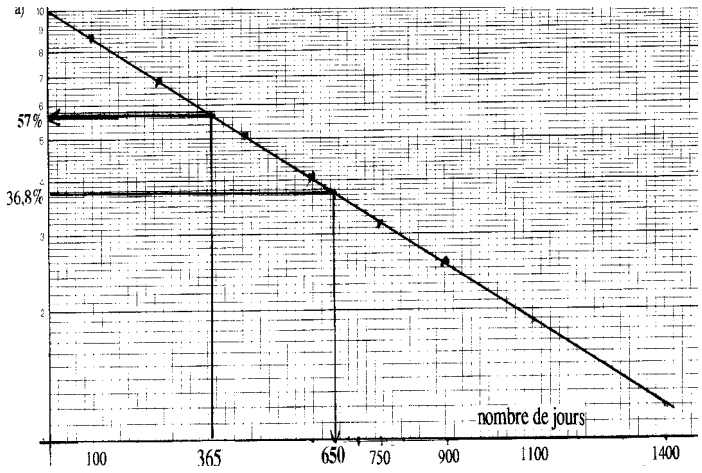
2° A l'aide de la calculatrice on obtient :
$r = -0{,}999813$.
On obtient une équation de la droite de régression de y en t :
$y = -3{,}953.\ 10^{-3}\ t - 0{,}00599$.
On en déduit :
$\ln [R(t)] = -0{,}004\ t - 0{,}006$ à 10^{-3} près .
$R(t) = e^{-0{,}004\ t - 0{,}006}$.

3° On prend 1 comme valeur approchée de $e^{-0{,}006}$
donc $R(t) = e^{-0{,}004\ t}$.
Nous sommes en présence d'une loi exponentielle donc MTBF $= \dfrac{1}{\lambda}$

MTBF $= \dfrac{1}{0{,}004} =$ **250 h** .

4° La date t_0 telle que $R(t_0) = 0{,}80$ est telle que
$e^{-0{,}004\ t_0} = 0{,}80$, $-0{,}004\ t_0 = \ln 0{,}80$,
$t_0 \approx 55{,}8$ heures .
La périodicité d'un entretien systématique est donc d'environ 56 heures.

5 - DOMOTIQUE 93

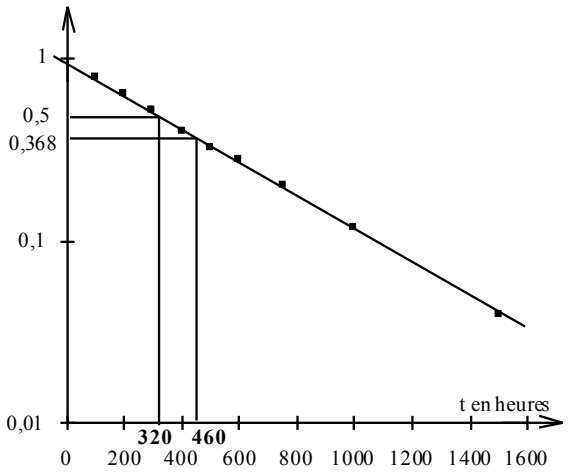


Le nuage de points $M(t_i;R(t_i))$ peut être ajusté par une droite passant par le point de coordonnées (0 ; 1) .
b) La variable aléatoire qui mesure la durée de vie des pièces de ce type suit une loi exponentielle.
La MTBF ≈ 640 donc le paramètre est $\lambda \approx 0{,}0015625$ et $R(t) = e^{-0{,}00156\ t}$.

c) $R(365) \approx 0{,}565$, $F(365) = 1 - R(365) \approx 0{,}435$.
La probabilité de voir le système tomber en panne durant l'année de garantie est donc environ 43,5%. (graphiquement : $R(365) \approx 56\%$).

6 - 1°

t_i	100	200	300	400	500	600	750	1000	1500
$\sum n_i$	5	9	12	15	17	18	20	22	24
$F(t_i)$ %	20	36	48	60	68	72	80	88	96
$R(t_i)$%	80	64	52	40	32	28	20	12	4



2° Le nuage de points $M(t_i, R(t_i))$ peut être ajusté par une droite passant par le point de coordonnées (0;1) , la variable aléatoire qui mesure la durée de vie des pièces suit donc une loi exponentielle.
Sur la droite le point d'ordonnée 36,78% a pour abscisse $t \approx 460$ donc **MTBF ≈ 460 heures.**

3° Le paramètre est $\lambda \approx 1/460$ $\lambda \approx 0{,}0022$ et $R(t) = e^{-t/460}$, $\boldsymbol{R(t) = e^{-0{,}0022\ t}}$.

4° On trouve graphiquement $R(900) \approx 14$ % et par le calcul $R(t) = e^{-900/460} \approx$ **0,138.** Pour $R(t_0) = 0{,}5$ on trouve sur le graphique $t_0 \approx 320$ h.
Par le calcul $R(t_0) = 0{,}5$ équivaut à $e^{-t_0/460} = 0{,}5$,
$\dfrac{t_0}{460} = -\ln 0{,}5$, $t_0 = 460 \ln 2$, $t_0 \approx 318{,}84$, $\boldsymbol{t_0 \approx 319.}$
Si $R(t_0) = 0{,}5$ alors $F(t_0) = 0{,}5$.
Le système de 2 éléments montés en parallèle est défaillant si les deux éléments sont défaillants.
Les deux éléments fonctionnant de façon indépendante, la probabilité que le système soit défaillant est : $F_S(t_0) = (0{,}5)^2 = 0{,}25$.
La fiabilité du système à l'instant t_0 est donc
$R_S(t_0) = 1 - F_S(t_0)$ $R_S(t_0) =$ **0,75.**

5° Le système S ' de deux éléments montés en série fonctionne à la date t si les deux éléments fonctionnent à la date t . Les deux éléments fonctionnant de façon indépendante, la fiabilité à la date t est $R_{S'}(t) = R(t)^2$
$R_{S'}(t) = (e^{-t/460})^2$, $\boldsymbol{R_{S'}(t) = e^{-t/230}}$.
La fiabilité diminue , la MTBF n'est plus alors que de **230 heures.**

6° On trouve en prenant les valeurs approchées :
$R(t) = e^{-0{,}002\ t}$, réponse sensiblement identique à celle trouvée au 3°.

7 – GROUPEMENT C 99

1° a) $f'(t) = -\dfrac{3t^2}{10^9}\,e^{-\frac{t^3}{10^9}}$, pour tout réel $t \neq 0$,

$e^{-\frac{t^3}{10^9}} > 0$ et $\dfrac{3t^2}{10^9} > 0$ donc pour tout $t \neq 0$ $f'(t) < 0$ et $f'(0) = 0$,

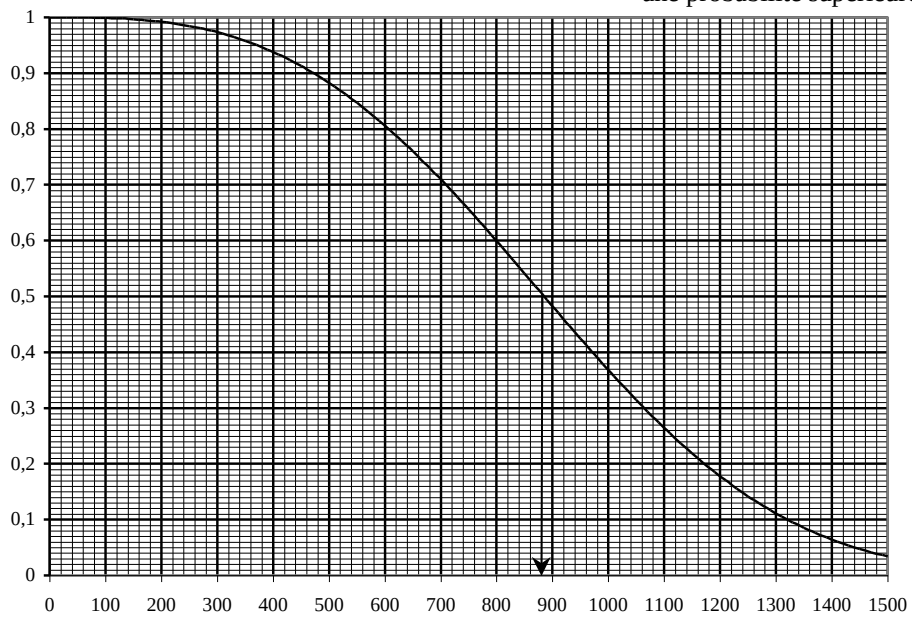
f est strictement décroissante sur $\mathbb{R}$.

$\lim\limits_{t\to+\infty} -\dfrac{t^3}{10^9} = -\infty$, $\lim\limits_{t\to+\infty} e^{-\frac{t^3}{10^9}} = 0$ puisque $\lim\limits_{u\to-\infty} e^{u} = 0$ donc $\lim\limits_{t\to+\infty} f(t) = 0$.

L'axe des abscisses est asymptote à la courbe représentative de f.

b) $\lim\limits_{t\to-\infty} -\dfrac{t^3}{10^9} = +\infty$, $\lim\limits_{t\to-\infty} e^{-\frac{t^3}{10^9}} = +\infty$ puisque $\lim\limits_{u\to+\infty} e^{u} = +\infty$ donc $\lim\limits_{t\to-\infty} f(t) = +\infty$.

c)



3° a) Les équations suivantes sont équivalentes dans $\mathbb{R}$: $f(t) = 0{,}5$; $e^{-\frac{t^3}{10^9}} = 0{,}5$; $-\dfrac{1}{10^9}t^3 = \ln 0{,}5$,

$t^3 = 10^9 \ln 2$, $t = 10^3 \times \sqrt[3]{\ln 2}$, $t \approx 885$.

b) f est strictement décroissante sur $\mathbb{R}$,

$f(10^3 \times \sqrt[3]{\ln 2}) = 0{,}5$ donc $f(t) < 0{,}5$ équivaut à $t > 10^3 \times \sqrt[3]{\ln 2}$, l'ensemble des solutions de l'inéquation est donc l'intervalle $]10^3 \times \sqrt[3]{\ln 2}\,, +\infty[$.

4° a) La probabilité qu'une telle machine fonctionne plus de 1000 heures est $P(T > 1000) = f(1000)$, or

$f(1000) = e^{-\frac{10^9}{10^9}}$, $P(T > 1000) = e^{-1}$,

$P(T > 1000) \approx 0{,}368$.

b) La probabilité qu'une telle machine fonctionne plus de 400 heures est $P(T > 400) = f(400)$.

Il s'agit de vérifier que $f(400) > 0{,}9$ or

$f(400) = e^{-\frac{400^3}{10^9}}$, $f(400) \approx 0{,}93$ donc on a bien une probabilité supérieure à 0,9.

8 - MAINTENANCE Nouvelle Calédonie **96**

1° a) $F(t) = \int_0^t 0{,}005\,e^{-0{,}005x}\,dx$,

$F(t) = [e^{-0{,}005\,x}]_0^t$, $\boldsymbol{F(t) = 1 - e^{-0{,}005\,t}}$.

b) $J(t) = \int_0^t 0{,}005x\,e^{-0{,}005x}\,dx$.

On intègre par parties :

$J(t) = [-x\,e^{-0{,}005\,x}]_0^t + 200\,F(t)$,

$\boldsymbol{J(t) = -t\,e^{-0{,}005\,t} - 200\,e^{-0{,}005\,t} + 200.}$

c) $K(t) = \int_0^t 0{,}005x^2\,e^{-0{,}005x}\,dx$.

On intègre par parties :

$K(t) = [-x^2\,e^{-0{,}005\,x}]_0^t + 400\,J(t)$,

$\boldsymbol{K(t) = -t^2\,e^{-0{,}005\,t} - 400\,t\,e^{-0{,}005\,t} - 80000\,e^{-0{,}005\,t} + 80000.}$

2° $I = \lim_{t\to+\infty} -e^{-0{,}005\,t}$, $\lim_{t\to+\infty} e^{-0{,}005\,t} = 0$, $\boldsymbol{I = 1.}$

$E(T) = \lim_{t\to+\infty} (-t\,e^{-0{,}005\,t} - 2\,e^{-0{,}005\,t} + 200)$,

$\boldsymbol{E(T) = 200}$.

$E(T^2) = \lim_{t\to+\infty} (-t^2\,e^{-0{,}005\,t} - 400\,t\,e^{-0{,}005\,t} + 80000\,e^{-0{,}005t} + 80000)$.

$\boldsymbol{K = 80000}$.

$V(T) = E(T^2) - [E(T)]^2 = K - 200^2$,

$\boldsymbol{V(T) = 40000,\quad \sigma(T) = \sqrt{V(T)} = 200.}$

3° a) En fiabilité F est la fonction de défaillance, R définie sur $[0, +\infty[$ par $R(t) = 1 - F(t)$ est la fonction de fiabilité.

D'après ce qui précède $F(t) = 1 - e^{-0{,}005\,t}$ donc $R(t) = e^{-0{,}005\,t}$, la variable aléatoire T suit donc une loi exponentielle de paramètre $\lambda = 0{,}005$.

b) $P(T > 300) = R(300) = e^{-0{,}005 \times 300}$,

$P(T > 300) = e^{-1{,}5}$,

$\boldsymbol{P(T > 300) = 0{,}223\ à\ 10^{-3}\ près.}$

$P(T \le 100) = 1 - e^{-0{,}005 \times 100}$,

$P(T \le 100) = 1 - e^{-0{,}5}$,

$\boldsymbol{P(T \le 100) = 0{,}393\ à\ 10^{-3}\ près.}$

$P(100 < T \le 300) = F(300) - F(100)$,

$\boldsymbol{P(100 < T \le 300) =}$ **0,384** à 10^{-3} près.

c) $P(T \le t_0) = 0{,}1$ équivaut à $F(t_0) = 0{,}1$ et à $R(t_0) = 0{,}9$ et à $e^{-0{,}005\,t_0} = 0{,}9$ d'où :

$-0{,}005\,t_0 = \ln 0{,}9$, $t_0 = -200 \ln 0{,}9$, $\boldsymbol{t_0 \le 21.}$

9 - MAINTENANCE ET EXPLOITATION DES MATERIELS AERONAUTIQUES 94

A) 1° $F(x) = [e^{-\lambda.t}]_0^x$, $\boldsymbol{F(x) = 1 - e^{-\lambda x}}$.

2° a) On intègre par parties en posant :

$$\begin{cases} u(t) = t \\ v'(t) = \lambda e^{-\lambda.t} \end{cases} \quad \begin{cases} u'(t) = 1 \\ v(t) = -e^{-\lambda.t} \end{cases}$$

$I(a) = [-te^{-\lambda.t}]_0^a - \int_0^a -e^{-\lambda.t}\,dt$,

$I(a) = [-t\,e^{-\lambda.t} - \frac{e^{-\lambda.t}}{\lambda}]_0^a$,

$I(a) = \frac{1}{\lambda} + e^{-\lambda a}(-a - \frac{1}{\lambda})$.

b) $\lim_{a\to+\infty} e^{-\lambda a} = 0$, $\lim_{a\to+\infty} a\,e^{-\lambda a} = 0$,

donc $\lim_{a\to+\infty} \boldsymbol{I(a)} = \frac{1}{\lambda} = \boldsymbol{E(X).}$

3° $P(X > t) = e^{-\lambda t}$, $P(X > t + h) = e^{-\lambda(t+h)}$,

$\boldsymbol{P(X > t + h / X > t) = e^{-\lambda h}}$.

B) 1° $P(Y_t = k) = \frac{(\lambda t)^k e^{\lambda.t}}{k!}$,

$P(Y_t = 0) = \frac{(\lambda t)^0 e^{\lambda.t}}{0!}$, $\boldsymbol{P(Y_t = 0) = e^{-\lambda.t}.}$

2° a) $P(X > t) = P(Y_t = 0)$ donc $P(X > t) = e^{-\lambda.t}$,

$\boldsymbol{P(X \le t) = 1 - e^{-\lambda.t}.}$

b) $P(X < t) = R(t) = e^{-\lambda.t}$ donc X suit la loi exponentielle de paramètre λ.

10 - MAINTENANCE 2000

1. a) Le coefficient de corrélation linéaire valant environ 0,994 , il est très proche de 1, ce qui justifie l'ajustement linéaire du nuage $(x_i\ ;\ y_i)$ par la droite D (on constate par ailleurs graphiquement le bon alignement des points sur la droite de régression).

b) D'après l'équivalence donnée, l'alignement des points de coordonnées $(x_i\ ,\ y_i)$ implique que l'expression de $R(t)$ correspond à une loi de Weibull de paramètre $\gamma = 0$.

c) D'après les valeurs obtenues par le tableur,

$\beta = 1{,}7522$ (coefficient directeur de la droite de régression) et

$\beta \ln \eta = 8{,}2171$ donne

$\eta = e^{\frac{8{,}2171}{1{,}7522}}$ d'où $\eta \approx 109$.

On prendra $\gamma = 0$; $\beta = 1{,}75$ et $\eta = 109$.

2. On résout $R(t) = 0{,}80$ qui équivaut à

$e^{-\left(\frac{t}{109}\right)^{1,75}} = 0{,}8$ et à

$\frac{t}{109} = (-\ln 0{,}8)^{\frac{1}{1,75}}$ d'où $t \approx 46{,}3$.

On prévoira donc une intervention préventive tous les 46 jours.

11 - MAINTENANCE 98

1° Sur la graphique les points sont sensiblement alignés on est donc en présence d'une loi de Weibull de paramètre $\gamma = 0$.

On trouve graphiquement $\eta = 66$ et $\beta = 4{,}2$.

MTBF = $\eta A + \gamma$, on trouve dans le formulaire pour $\beta = 4{,}2$ et $\eta = 66$: $A = 0{,}9089$ d'où

MTBF ≈ 59,99 jours , **MTBF ≈ 60 jours**.

2° $R(t_0) = 0{,}9$ équivaut à $F(t_0) = 0{,}10$, on lit sur le papier de Weibull $t_0 =$ **38jours**.

$R(t) = e^{-(\frac{t}{66})^{4,2}}$, $e^{-(\frac{t_0}{66})^{4,2}} = 0{,}9$ équivaut à

$-(\frac{t_0}{66})^{4,2} = \ln 0{,}9$ et à $\frac{t_0}{66} = (-\ln 0{,}9)^{\frac{1}{4,2}}$ et à

12 - MAINTENANCE 97

1° Sur la graphique les points sont sensiblement alignés on est donc en présence d'une loi de Weibull de paramètre $\gamma = 0$, on trouve graphiquement :

$\eta = 16$ et $\beta = 5$ donc $R(t) = e^{-(t/16)^5}$.

2° MTBF = $\eta A + \gamma$ on trouve dans le formulaire pour $\beta = 5$: $A = 0{,}9182$, **MTBF ≈ 14,7 heures,** par le calcul $R(14{,}7) \approx 0{,}52$.

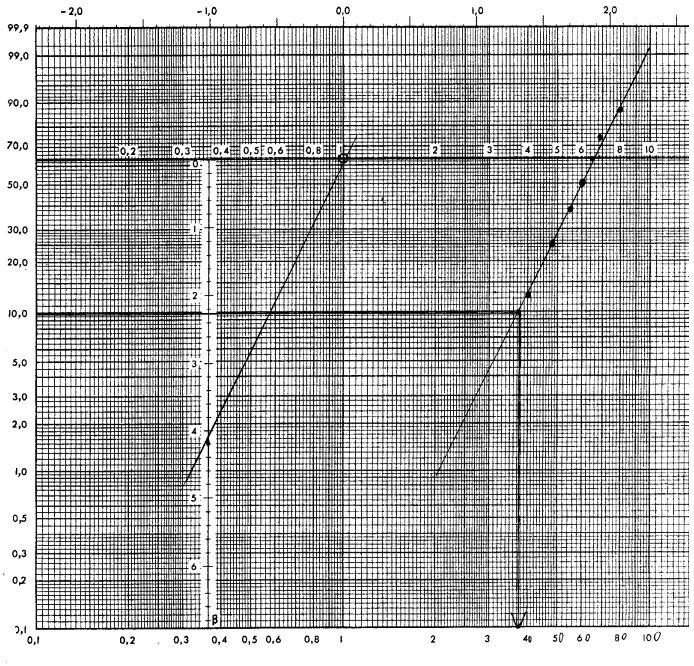
3° $R(t_0) = 0{,}9$ équivaut à $F(t_0) = 0{,}10$, on lit sur le papier de Weibull $t_0 =$ **10 heures.**

Par le calcul $R(t_0) = 0{,}9$ équivaut à $e^{-(t_0/16)^5} = 0{,}9$

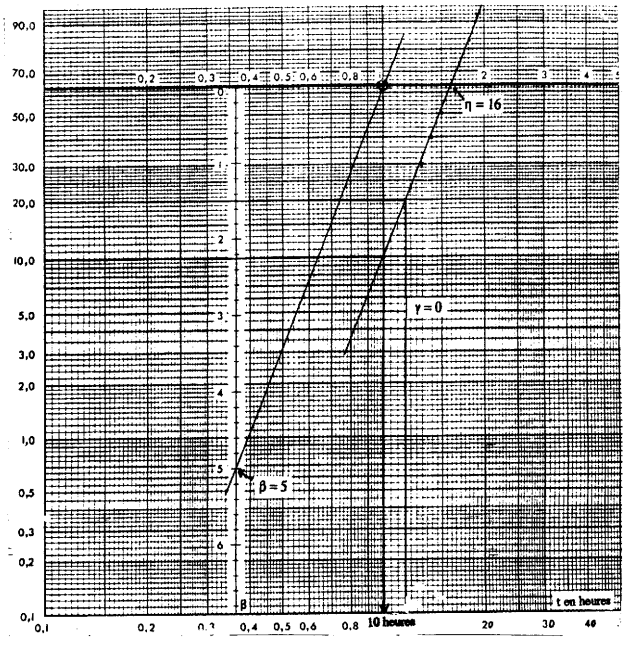
et à $-(\frac{t_0}{16})^5 = 0{,}9$ et à $\frac{t_0}{16} = (-\ln 0{,}9)^{0,2}$

et à $t_0 \approx 10{,}2$ heures. La périodicité d'un réglage systématique basée sur une fiabilité de 90 % est donc de **10 heures**.

$t_0 \approx$ **38,6 jours.**



La périodicité d'un entretien systématique basé sur une fiabilité de 90 % est donc de **38 jours**.



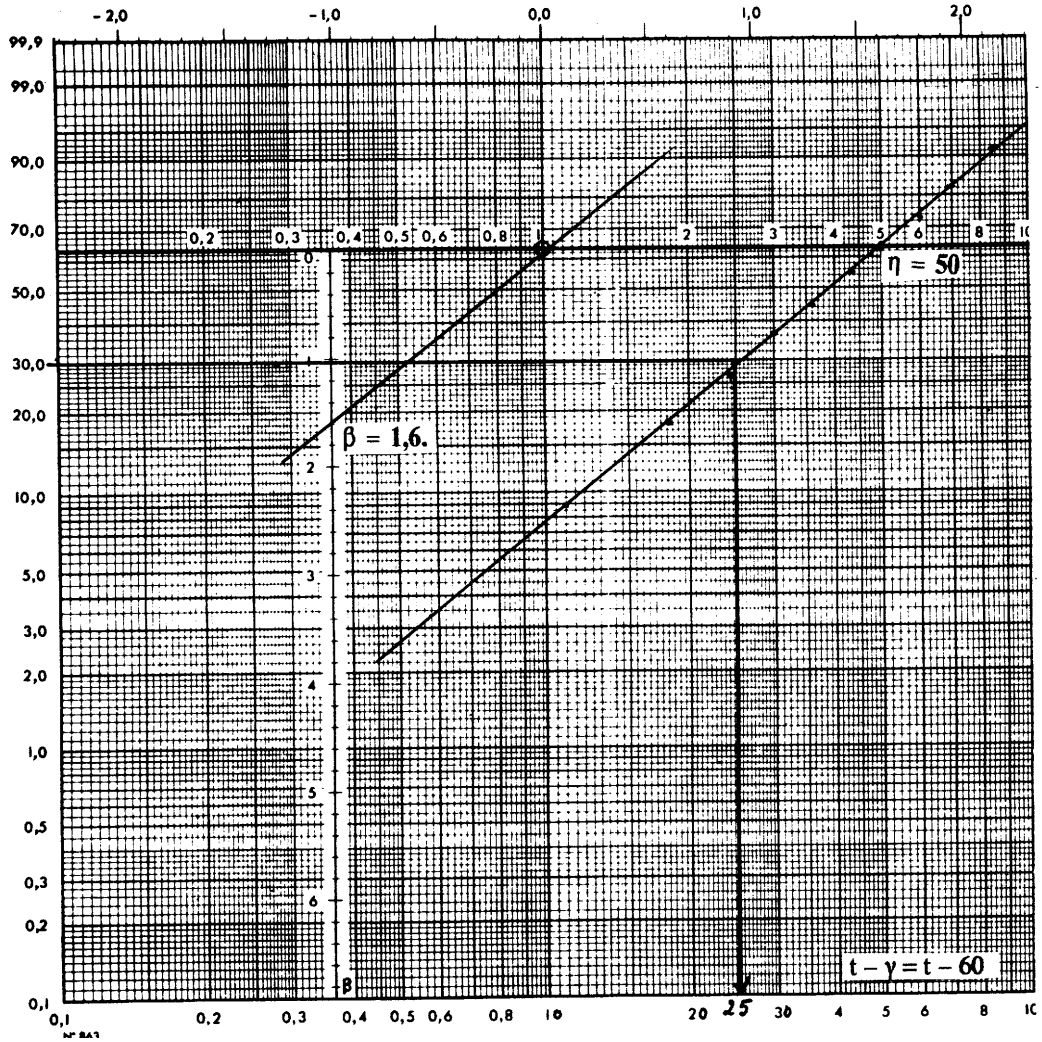
13 - MAINTENANCE NOUMEA 95

1° a)

Durée de vie t_i en jours	71	78	84	90	96	104	110	120	130	145
Effectifs cumulés N_i	1	2	3	4	5	6	7	8	9	10
$F(t) = \frac{N_i}{n+1} = \frac{N_i}{11}$	9,09	18,18	27,27	36,36	45,45	54,55	63,64	72,73	81,82	90,91
$t - \gamma = t - 60$	11	18	24	30	36	44	50	60	70	85

Sur le papier de Weibull on place les points de coordonnées M_i $(t_i - 60 , F(t_i))$.

Les points sont pratiquement alignés on a donc bien $\gamma = 60$.



b) On trouve $\eta = 50$ **et** $\beta = 1,6$.

c) A l'instant t_0 on a $R(t_0) = 0,7$ donc $F(t_0) = 0,3$ soit 30 % .

On lit sur le papier de Weibull $t_0 - 60 = 25$ donc $t_0 = 25 + 60$ soit t_0 **= 85 jours.**

2° $R(t_0) = 0,7$ donc $e^{-(\frac{t_0-60}{50})^{1,6}} = 0,7$,

$[\frac{t_0-60}{50}]^{1,6} = -\ln 0,7$, $t_0 = 60 + 50\,(-\ln 0,7)^{\frac{1}{1,6}}$,

t_0 **= 86,3 jours.**

3° MTBF = $A\,\eta + \gamma$ or $\beta = 1,6$ donc $A = 0,8966$, MTBF = 104,83 donc **MTBF = 105** jours.

14 – MAINTENANCE NOUVELLE CALEDONIE 98

1° Le formulaire donne pour $\beta = 3$: $A = 0,8930$, $\gamma = 100$ donc MTBF = $\eta A + \gamma$ = **546,5 h**.

Le formulaire donne pour $\beta = 3$: $B = 0,325$, $\eta = 500$ donc l'écart type $\sigma = \eta B$ = **162,5 h.**

$$R(t) = \exp\left[-\left(\frac{t-100}{500}\right)^3\right]$$

2° Si $R(t) = 0{,}9$ alors $-\left(\frac{t-100}{500}\right)^3 = \ln 0{,}9$ qui équivaut à $\frac{t-100}{500} = (-\ln 0{,}9)^{1/3}$ et à

$t = 100 + 500(-\ln 0{,}9)^{1/3}$ d'où $t \approx$ **436,15 heures.**

3° Le délai avant l'arrivée de la pièce de rechange est $16 \times 30 = 480$ heures.
La probabilité de défaillance avant 480 heures est $F(480) = 1 - R(480)$.

$$F(480) = \exp\left[-\left(\frac{480-100}{500}\right)^3\right],$$

F(480) = 1 – 0,6447, $F(480) \approx$ **0,355**. La probabilité de rupture de stock est donc **35,5 %.**

15 - MAINTENANCE 94

1°

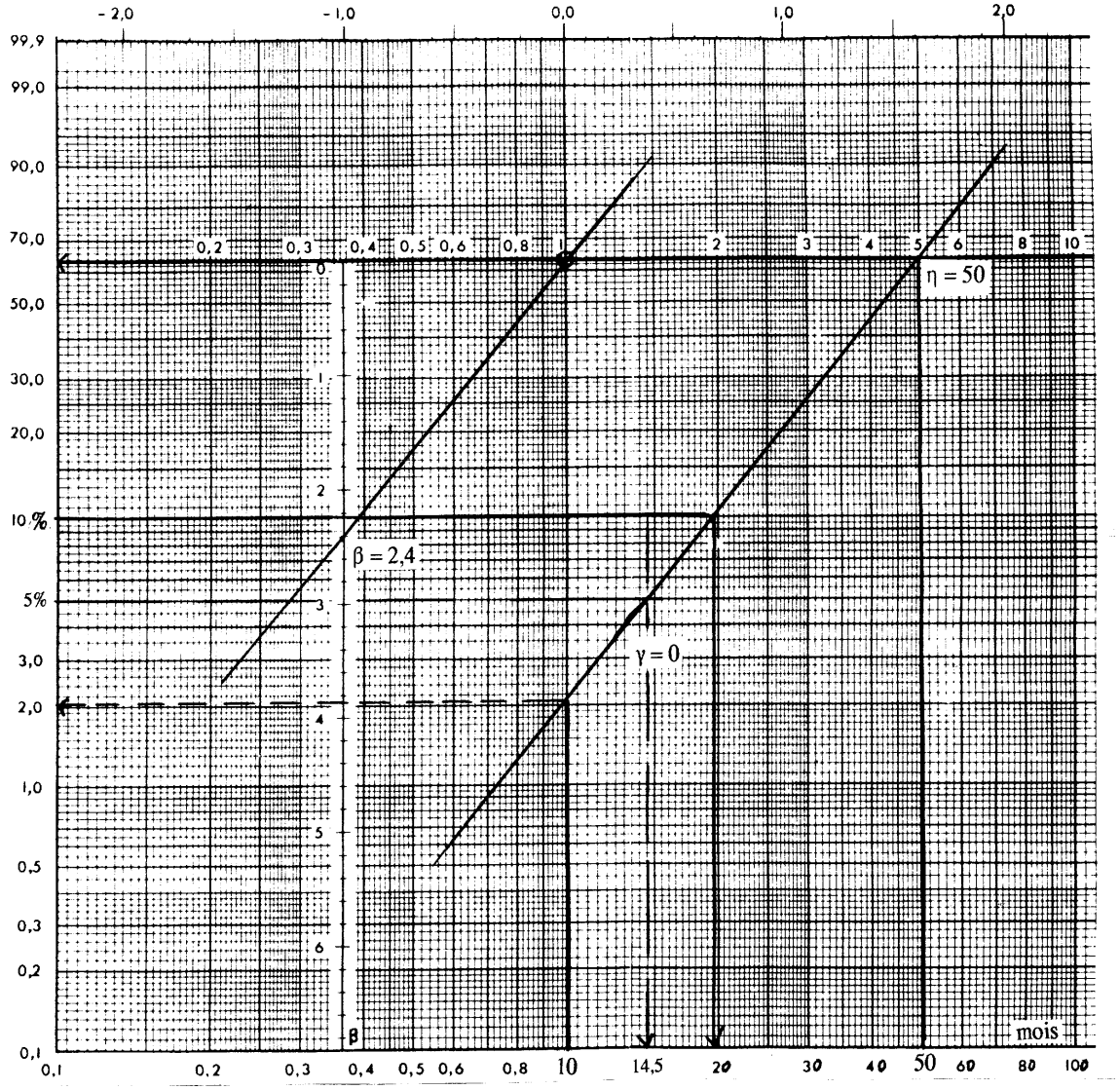
2° a) Graphiquement : $P(T < 10) = F(10) \approx$ **2%** .
b) Graphiquement : $P(10 < T < 50) = F(50) - F(10) \approx$ **61%.**

3° On cherche t_0 tel que $R(t_0) > 0{,}9$ donc $F(t_0) < 0{,}1$; $R(t_0) = e^{-(\frac{t_0}{50})^{2,4}}$ donc $e^{-(\frac{t_0}{50})^{2,4}} \geq 0{,}9$,

$-(\frac{t_0}{50})^{2,4} \geq \ln(0{,}9)$, donc $t_0 \leq 50\,(-\ln(0{,}9))^{1/2,4}$

le calcul donne $t_0 \leq 19{,}57$ mois .
Graphiquement $\mathbf{t_0 \leq 20}$ **mois** , résultats identiques à un mois près .

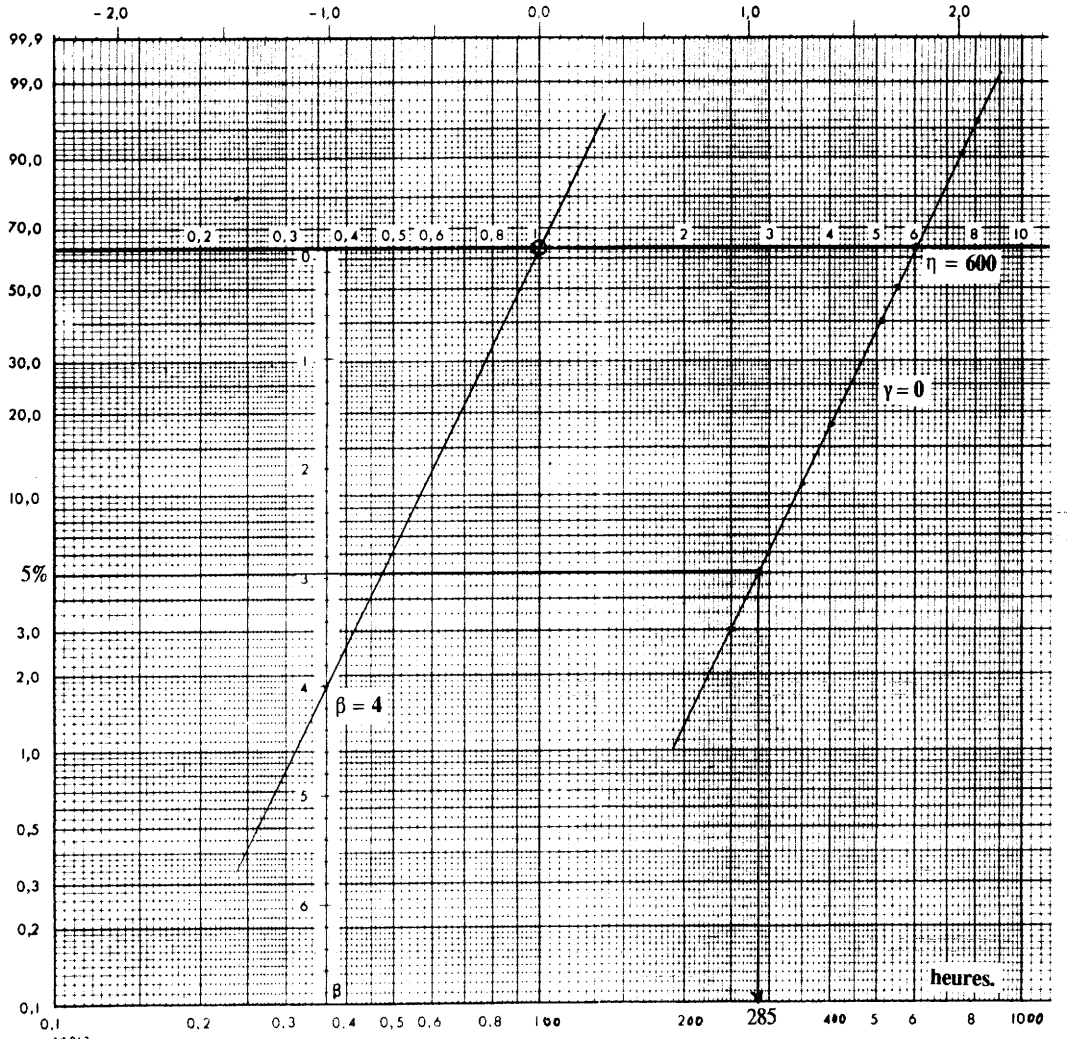
4° Soit T_1 et T_2 les variables aléatoires qui mesurent les durées de vie respectives de deux composants de (S) tirés au hasard. Soit $R_1(t)$ et $R_2(t)$ les fiabilités respectives des deux composants et soit $R_S(t)$ la fiabilité du système .
Lorsque le système est constitué par le montage en série de deux composants $R_S(t) = P(T_1 \geq t \text{ et } T_2 \geq t)$.
La durée de vie du 1er composant est indépendante du 2ième et les variables T_1 et T_2 suivent la même loi que T donc $R_S(t) = [P(T \geq t)]^2 = [R(t)]^2$, donc



$R_S(t_0) > 0{,}9$ équivaut à $[R(t_0)]^2 > 0{,}9$ et à $R(t_0) > (0{,}9)^{1/2}$, $R(t_0) > 0{,}95$. Par le calcul (comme au 3°) on trouve $t_0 \approx 14{,}66$ mois.
$P(T < t_0) < 0{,}05$, $F(t_0) < 5\%$, graphiquement
$t_0 \approx$ 14,5 mois.

16 - MAINTENANCE NOUMEA 93

1° On place sur le papier de Weibull les points de coordonnées $(t_i , F(t_i))$.



On constate que ces points sont alignés ce qui justifie que la variable aléatoire qui mesure la durée de bon fonctionnement peut être ajustée par une loi de Weibull de paramètre $\gamma = 0$.
On trouve $\eta = 600$ **heures** et $\beta = 4$.

2° a) Le formulaire donne la MTBF = $A\eta + \gamma$ et l'écart type $\sigma = B\eta$.
Pour $\beta = 4$, on lit dans la table $A = 0{,}9064$ et $B = 0{,}254$, donc MTBF = (0,9064)(600) + 0 ,
MTBF = 543,84 heures.
$\sigma = 0{,}254 \times 600$ **σ = 152,4 heures.**

b) Ici $R(t) = e^{-(\frac{t}{600})^4}$ si 5% des engrenages sont défectueux au bout de t heures, alors t est solution de $R(t) = 0{,}95$ donc $e^{-(\frac{t}{600})^4} = 0{,}95$,

$-(\frac{t}{600})^4 = \ln(0{,}95)$, $t = 600\,(-\ln 0{,}95)^{0{,}25}$,

t_0 **= 285,5 heures** .

Vérification graphique :
t_0 est l'abscisse du point d'ordonnée 5, situé sur la droite obtenue à la question 1 .On lit $t_0 \approx 285$ *h*.

3° La probabilité qu'un engrenage fonctionne au moins 300 heures est R(300) ≈ 0,9394.
La transmission fonctionne si et seulement si les trois engrenages sont en état après 300 heures; d'autre part ces engrenages fonctionnent de façon indépendante, donc la probabilité que la durée de vie du système soit au moins de 300 heures est $p \approx (0{,}9394)^3$, soit $\boldsymbol{p \approx 0{,}83}$.

17 – Papier ou calculatrice ?
A – 1°

t_i	250	450	600	800	1100	1400
$F(t_i)$ en %	5	15	25	40	60	80

Le nuage de points est presque rectiligne, on en déduit $\gamma = 0$, la droite d'ajustement D coupe l'axe des abscisses au point d'abscisse $\eta = 1100$, la parallèle à D passant par le point d'abscisse 1 coupe l'axe donnant les valeurs de β tel que $\beta = 2$
$R(t) = e^{-(\frac{t}{1100})^2}$.

2° Pour $\beta = 2$ on trouve dans la table $A = 0{,}8862$
$B = 0{,}463$, MTBF = $\eta A + \gamma$ donc **MTBF ≈ 975 h**
$\sigma = \eta B$ **σ ≈ 509 heures** .

3° Déterminons la périodicité d'un entretien systématique basé sur une fiabilité de 90 % :
Graphiquement : $F(t)$ =10% , on trouve sur le papier de Weibull $t = 360$ heures.
Par le calcul : $e^{-(t/1100)^2} = 0{,}9$ équivaut à
$(\frac{t}{1100})^2 = -\ln 0{,}9$ et à $t = 1100\,(-\ln 0{,}9)^{1/2}$
$t \approx$ 357 heures. Il faut faire un entretien systématique au moins toutes les **360** heures.

4° *Graphiquement :* $F(2000) = 96\%$,
$R(2000) = 1 - F(2000)$, $R(2000) =$ **4%.**
Par le calcul : $e^{-(2000/1100)^2} \approx$ **3,7%.**

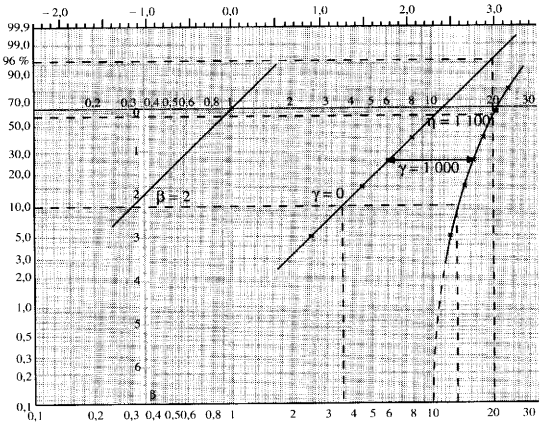
5° $\lambda(t) = \frac{2}{1100}(\frac{t}{1100})^1$,
$\lambda(t) = 1{,}65.\ 10^{-6}\ t$, le taux d'avarie est faible, sa coube représentative est une droite.

B 1° On obtient les mêmes valeurs de $F(t)$ les points sont donc déplacés d'un écart de 1000 h ,
$F(1250) = 5\%$, $F(1450) = 15\%$ etc...

2° MTBF = $\eta A + \gamma$, MTBF ≈ 975 + 1000,
MTBF = **1975 h**,
$\sigma = \eta B$ $\sigma \approx$ **509** heures.

3° *Graphiquement :* $F(t) = 10\%$ pour $t = 1360$ h.
Par le calcul : $(\frac{t-1000}{1100})^2 = -\ln 0{,}9$ et à
$t = 100 + 1100\,(-\ln 0{,}9)^{1/2}$,
$t \approx$ **1357** h.

4° *Graphiquement :* $F(2000) = 57\%$,
$R(2000) = 43\%$.
Calcul : $\boldsymbol{R(2000) \approx 43{,}7\ \%}$.



5° $\lambda(t) = \frac{2}{1100}(\frac{t-1000}{1100})$, la courbe représentative est une droite qui ne passe plus par l'origine.

18 - MAINTENANCE 97

1° $f_1(t) = e^{-t}$
a) Voir graphique.
b) $F_1(t) = \int_0^t f_1(x)\,dx$, $F_1(t) = \int_0^t e^{-x}\,dx = [-e^{-x}]_0^t$,
$\boldsymbol{F_1(t) = 1 - e^{-t}}$.
$P(T \le 1) = 1 - e^{-1}$, $\boldsymbol{P(T \le 1) = 0{,}632}$ à 10^{-3} près.

c) $I(t) = \int_0^t x\, f_1(x)\,dx = \int_0^t x\, e^{-x}\,dx$, en posant $u(x) = x$, $v'(x) = e^{-x}$ on a $u'(x) = 1$, $v(x) = -e^{-x}$, on obtient :

$I(t) = [-x\, e^{-x}]_0^t - \int_0^t -e^{-x}\,dx$,

$I(t) = -t\, e^{-t} - [e^{-x}]_0^t$,

$I(t) = -t\, e^{-t} - e^{-t} + 1$,

de $\lim_{t\to+\infty} e^{-t} = 0$, et de $\lim_{t\to+\infty} t\, e^{-t} = 0$, on déduit :

E(T) = $\lim_{t\to+\infty}$ I(t) = 1 .

2° a) $f_2(t) = 2t\, e^{-t^2}$,

$f_2'(t) = 2\, e^{-t^2} - 4t^2\, e^{-t^2}$,

$f_2'(t) = 2\, e^{-t^2}(1 - 2t^2)$,

$f_2'(t)$ est du signe de $1 - 2t^2$,

f_2 définie sur $[0 ; +\infty[$, $f_2'(t) = 0$ pour $t_0 = \frac{\sqrt{2}}{2}$,

$f(t_0) \approx 0{,}857$.

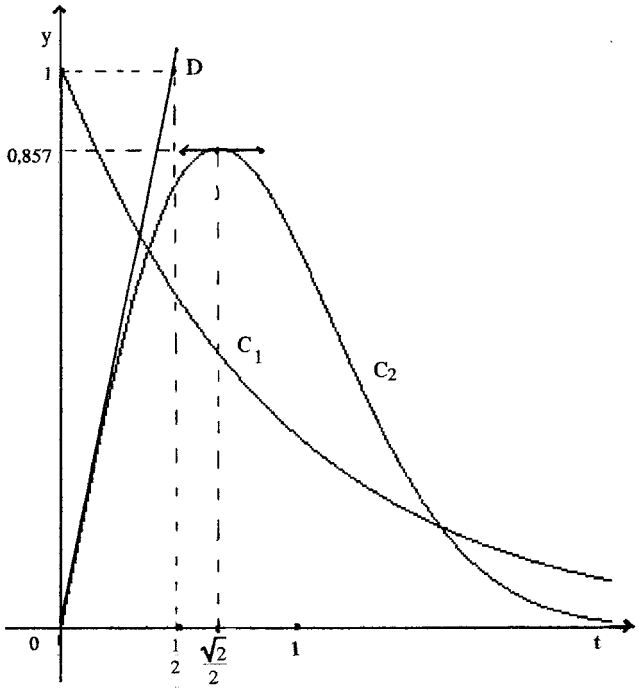
t	0		t_0		$+\infty$
$f_2'(t)$		+	0	–	
$f_2(t)$	0	↗	$f(t_0)$	↘	0

b) Le développement limité de e^u à l'ordre 1, au voisinage de 0 est : $e^u = 1 + u + u\,\varepsilon_1(u)$ avec $\lim_{u\to0} \varepsilon_1(u) = 0$, le développement limité à l'ordre 2 de e^{-t^2} est : $1 - t^2 + t^2\,\varepsilon_2(t)$ avec $\lim_{t\to0} \varepsilon_2(t) = 0$, le développement limité à l'ordre 3 de $f_2(t)$ au voisinage de 0 est : $f_2(t) = 2t(1 - t^2) + t^3\,\varepsilon_3(t)$,

$f_2(t) = 2t - 2t^3 + t^3\,\varepsilon_3(t)$ avec : $\lim_{t\to0} \varepsilon_3(t) = 0$.

On en déduit que l'équation de la tangente D à C_2 au point d'abscisse nulle est $\boldsymbol{y = 2t}$.

La position de C_2 par rapport à D, au voisinage de O, dépend du signe de : $f_2(t) - 2t = -2t^3 + t^3\,\varepsilon_3(t)$ qui est du signe contraire de t, donc C_2 est au dessous de D au voisinage de l'origine.



c) Voir graphique ci-après :

d) $F_2(t) = \int_0^t f_2(x)\,dx$, $F_2(t) = \int_0^t 2x\, e^{-x^2}\,dx$,

$(e^{-x^2})' = -2x\, e^{-x^2}$ donc $F_2(t) = [-e^{-x^2}]_0^t$,

$F_2(t) = -e^{-t^2} + 1$, $F_2(t) = 1 - e^{-t^2}$.

$P(T \le t) = 0{,}05$ équivaut à $F_2(t) = 0{,}05$ et à

$e^{-t^2} = 0{,}95$ et à $t^2 = -\ln 0{,}95$, $t = \sqrt{-\ln 0{,}95}$,

$t \approx 0{,}22648$, $t = 0{,}226$ à 10^{-3} près.

Supplément à la séance n°4

1 – Un exercice de terminale S sur la loi exponentielle

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

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 $e^{-\lambda t} \in]0, 1]$.

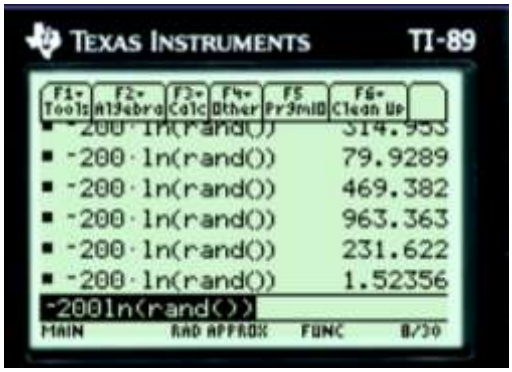
4) Si $t < 0$ alors $f(t) = 0$ car T est à valeurs dans $[0, +\infty[$.

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 !) :



2 – Test du khi-deux

Les tests d'ajustement ont pour but de vérifier si un échantillon donné peut raisonnablement ou non provenir des réalisations indépendantes d'une variable aléatoire de distribution connue.
Une première méthode, empirique, peut consister à comparer la forme de l'histogramme des fréquences observées aux histogrammes théoriques, dans le cas discret, ou au profil des fonctions de densité, dans le cas continu, des différents modèles possibles. Cette méthode n'est pas ridicule et peut déjà permettre d'éliminer certains modèles mais la qualité de l'ajustement n'est pas même quantifiée.
On franchit une première étape lorsque l'on peut quantifier la qualité de l'ajustement à l'aide d'un coefficient de régression. Dans bien des cas en effet, une transformation fonctionnelle, un changement de variable, permet de ramener l'ajustement à une régression linéaire selon les moindres carrés. C'est ainsi que procède le tableur lorsque, sur une courbe en « nuage de points », on demande une « courbe de tendance ». Cette procédure est simple à mettre en œuvre mais a le défaut de ne pas quantifier les risques d'erreurs lors de la prise de décision, ce que permet en revanche la procédure des tests d'hypothèse.
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…

Le cas des petits pois de Mendel

Nous exposons tout d'abord la procédure du test du khi-deux sur un exemple tiré de l'histoire.
Paradoxalement, l'absence de variabilité peut être aussi suspecte que ses débordements et permet également de détecter statistiquement des fraudes. C'est ainsi que Ronald Fisher examina les données expérimentales qui permirent à Gregor Johann Mendel (1822-1884) d'étayer sa théorie de l'hérédité.
Dans la plupart des cas, les résultats expérimentaux de Mendel étaient étonnamment proches de ceux prévus par sa théorie. Fisher montra que Mendel, sous l'hypothèse que sa théorie était exacte, et compte tenu de la variabilité naturelle des expériences, ne pouvait observer des résultats si proches des valeurs théoriques, qu'avec une probabilité inférieure à 0,00004 , autant dire humainement impossible. On peut ainsi 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.

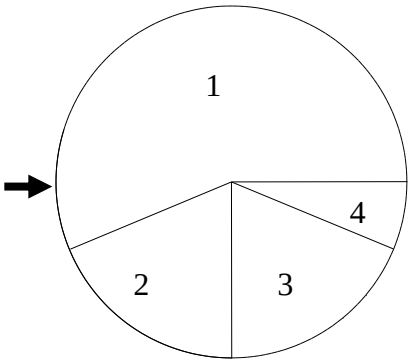
La théorie de Mendel prévoit que le croisement de petits pois « jaunes et ronds » avec des petits pois « verts et anguleux » donnera naissance à quatre nouvelles variétés, dans les proportions ci-dessous.

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	$p_4 = \frac{1}{16}$

Mendel prétend avoir réalisé 556 observations sur les résultats de ces croisements, et comparé ses observations aux valeurs théoriques attendues. Le tableau suivant indique les résultats qu'il présente.

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	$x_4 = 32$	$t_4 = 34{,}75$

L'adéquation est presque parfaite, un peu trop peut-être. 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'agirait 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 sur le tableur.
La formule =ENT(1+16*ALEA()) 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 de petits pois, qu'entre 10 et 12, il s'agit de l'espèce 2, etc.
Sur une feuille de calcul, on a entré en A2 la formule =ENT(1+16*ALEA()) et en B2 la formule =SI(A2<10;1;SI(A2<13;2;SI(A2<16;3;4))) qui indique la catégorie de petit pois correspondante.
Après avoir sélectionné les cellules A2 et B2, on les a recopiées vers le bas jusqu'à la ligne 557 pour simuler les observations qu'aurait pu faire Mendel, en supposant son modèle héréditaire parfaitement conforme à la réalité.
Les effectifs de chacune des quatre catégories sont comptabilisés dans la colonne F. La cellule F2 contient par exemple la formule =NB.SI(B2:B557;1) .

La différence entre les effectifs observés et les effectifs « théoriques » est mesurée en utilisant « l'écart quadratique réduit », c'est à dire $\frac{(x_i - t_i)^2}{t_i}$. On pondère en effet selon la valeur t_i de l'effectif théorique car en dehors du cas de l'équirépartition, les écarts absolus ne sont pas comparables.
La cellule G2 contient ainsi la formule (F2-E2)^2/E2 . La somme des écarts quadratiques réduits est effectuée dans la cellule G8. C'est la quantité que l'on nomme « khi-deux observé », $\chi^2_{obs} = \sum_{i=1}^{4} \frac{(x_i - t_i)^2}{t_i}$.

B2 =SI(A2<10;1;SI(A2<13;2;SI(A2<16;3;4)))

	A	B	C	D	E	F	G
1	aléas	types de petis pois		type	effectif théorique t	effectif observé x	(x-t)^2/t
2	16	4		1	**312,75**	**303**	0,30395683
3	9	1		2	**104,25**	**123**	3,37230216
4	5	1		3	**104,25**	**100**	0,17326139
5	6	1		4	**34,75**	**30**	0,64928058
6	13	3		somme	556	556	
7	6	1					
8	2	1				chi-deux observé	**4,49880096**
9	3	1				chi-deux Mendel	**0,47**
10	10	2					
11	7	1					
12	10	2					
13	10	2					
14	13	3					
15	2	1					

De grandes valeurs de χ^2_{obs} rendraient le modèle de Mendel suspect. La simulation montrée ici fournit $\chi^2_{obs} \approx 4{,}5$ alors que les expériences de Mendel conduisent à $\chi^2_{obs} \approx 0{,}47$.
La simulation, en appuyant plusieurs fois sur la touche F9, montre qu'un résultat aussi bon que celui de Mendel (sous l'hypothèse que le modèle est correct) est assez rare.
Pour préciser cela, calculons quelques probabilités.
On peut 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 du modèle de Mendel, suivent des lois binomiales de paramètre $n = 556$ et $p = p_i$ avec $p_1 = \frac{9}{16}$, $p_2 = \frac{3}{16}$, $p_3 = \frac{3}{16}$, $p_4 = \frac{1}{16}$.
Pour étudier la variabilité du critère du khi-deux, introduisons la variable aléatoire $T = \sum_{i=1}^{4} \frac{(X_i - t_i)^2}{t_i}$ avec $\sum_{i=1}^{4} X_i = 556$.
En notant n le nombre total de données (on a ici $n = 556$), l'effectif théorique de la classe i est $t_i = np_i$.
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 $P(T > t)$.

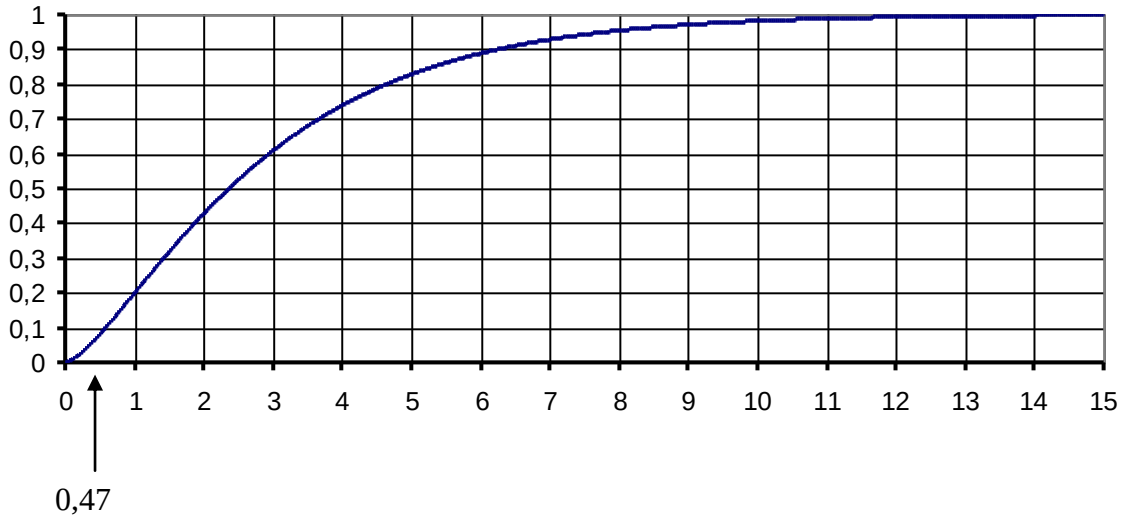
B48 = =LOI.KHIDEUX(A48;3)

	A	B	C	D
46	0,45	0,92973057	0,07026943	
47	0,46	0,92758713	0,07241287	
48	**0,47**	**0,92543108**	**0,07456892**	
49	0,48	0,92326283	0,07673717	
50	0,49	0,92108281	0,07891719	

Cette fonction permet alors d'obtenir
$P(T > 0{,}47) \approx 0{,}925$ d'où
$P(T \leq 0{,}47) \approx 0{,}075$.
Mendel avait donc environ 7,5 % de chances d'obtenir de si bons résultats.

A titre d'illustration, pour situer le résultat de Mendel, la fonction de répartition $t \mapsto F(t) = P(T \leq t)$ pour T suivant la loi du khi-deux à 3 degrés de liberté est représentée ci-dessous.

Fonction de répartition du khi-deux à 3 degrés de liberté



On veut bien croire en la chance de Mendel, 7,5 % c'est rare mais pas trop. Le problème est qu'en raison de l'indépendance des différentes sortes d'expériences, Fisher a pu additionner les différents χ^2_{obs} et aboutir ainsi à $\chi^2_{obs} < 42$, avec 84 degrés de liberté.

=1-LOI.KHIDEUX(42;84)

C	D
3,53507E-05	

C'est la probabilité que cet évènement se produise qui est de l'ordre de 0,00004 , comme on le constate sur le tableur.
Une enquête sur les manuscrits de Mendel a, par la suite, montré que certains résultats d'expérience avaient été grattés et « corrigés »...

Une présentation du test du khi-deux

Par définition, la loi du khi-deux à n degrés de liberté est la loi suivie par la somme des carrés de n variables aléatoires indépendantes de loi normale centrée réduite.
Soit une variable aléatoire discrète ou discrétisée, c'est à dire divisée en k classes de probabilités $p_1, \ldots, p_k$.
Soit un échantillon de taille n de cette variable aléatoire, fournissant pour chaque classe des effectifs $x_1, \ldots, x_k$.
Il s'agit de comparer ces effectifs aux valeurs théoriques :

Classes	Effectifs observés	Effectifs théoriques
1	x_1	$t_1 = np_1$
…	…	…
k	x_k	$t_k = np_k$

Il faut décider d'un critère d'adéquation des observations par rapport au modèle théorique.
De façon classique (et parce qu'on aboutit ainsi à une loi « connue »), on choisit l'écart quadratique réduit, noté χ^2_{obs} et valant :

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

De grandes valeurs de χ^2_{obs} rendraient le modèle suspect.

Pour étudier la variabilité de ce critère, on considère la variable aléatoire $T = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$ avec $\sum_{i=1}^{k} X_i = n$ où X_i est la variable aléatoire correspondant à l'effectif de la i^{e} classe.
Karl Pearson a démontré que la loi de T est approximativement, pour n grand, une loi du χ^2 à $k - 1$ degrés de liberté (on peut noter que la relation ci-dessus fait que la valeur de X_k est déterminée dès que les valeurs de $X_1, \ldots, X_{k-1}$ sont connues).

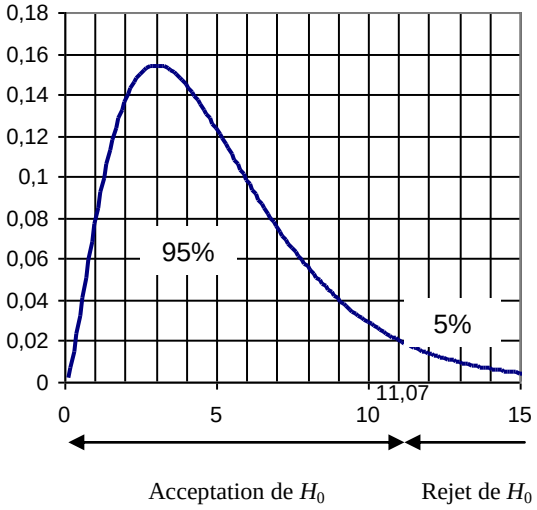
Sur Excel, la fonction LOI.KHIDEUX(valeur t ; degrés de libertés) fournit la probabilité $P(T > t)$ et la fonction KHIDEUX.INVERSE(probabilité p ; degrés de liberté) renvoie la valeur t telle que $P(T > t) = p$.

La construction d'un test du khi-deux se fait selon les étapes suivantes.

• *Choix des hypothèses* :
H_0 : pour tout $1 \leq i \leq 6$, la probabilité que X prenne une valeur dans la classe i est p_i.
H_1 : il existe i tel que la probabilité précédente diffère de p_i.

• *Détermination de la région critique* :
Si H_0 est vraie, la variable aléatoire $T = \sum_{i=1}^{k} \frac{(X_i - t_i)^2}{t_i}$ suit approximativement la loi du khi-deux à $k - 1$ degrés de liberté.

Densité du khi-deux à 5 degrés de liberté



On recherche sur une table le réel t tel que, $P(T > t) = \alpha$.
D'où la zone d'acceptation de H_0 au seuil α : [0, t].

• *Règle de décision* :

Soit χ^2_{obs} l'écart quadratique réduit obtenu entre les effectifs observés et les effectifs théoriques.

Si $\chi^2_{\text{obs}} \leq t$ on accepte H_0 au seuil de 5%.

Si $\chi^2_{\text{obs}} > t$ on rejette H_0.

Remarques :

• Si l'on utilise l'échantillon pour estimer indépendamment j paramètres de la loi testée, le degré de liberté du khi-deux devient : $k - 1 - j$. Par exemple $k - 3$ pour une loi normale où l'on estime μ et σ à l'aide de $\bar{x}$ et s_{n-1}.

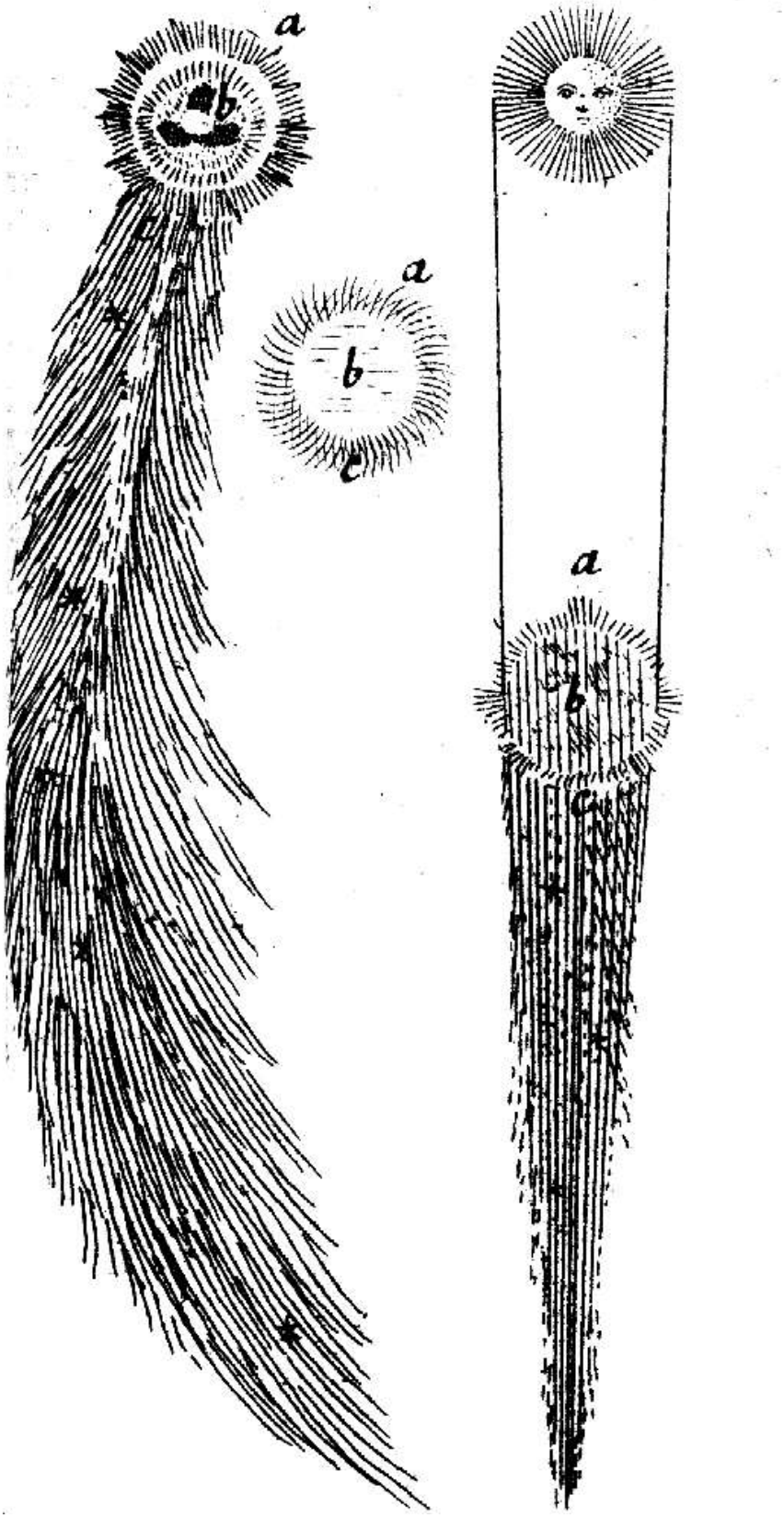
• La loi de T n'est qu'approchée et on donne la condition $np_i > 5$ pour l'effectif de chaque classe (sinon on regroupe les classes à effectif trop faible). C'est pour vérifier cette condition que l'on pratique le test du khi-deux sur les effectifs et non sur les fréquences.

Lien entre le khi-deux et le critère utilisé en terminale pour l'équidistribution

Le test du khi-deux s'appuie sur la variable aléatoire $T = \sum_{i=1}^{k} \frac{(X_i - t_i)^2}{t_i}$ où X_i est la variable aléatoire correspondant à l'effectif de la classe i et t_i l'effectif théoriquement prévu par le modèle pour cette classe.
Dans le cas d'un modèle équiréparti, comme on l'envisage en terminale ES et S (chapitre précédent), les effectifs théoriques sont tous égaux à $t_i = \frac{n}{k}$ avec k le nombre de classes et n le nombre total d'observations. Les différentes classes ayant toutes le même poids théorique, il est inutile de pondérer l'écart quadratique. De plus, pour n assez grand, la condition d'approximation par la loi du khi-deux sera satisfaite, sans que l'on ait à surveiller si un effectif est trop faible. Pour ces deux raisons, les programmes de terminales envisagent donc comme critère d'adéquation, la somme des écarts quadratiques sur les fréquences. C'est à dire que le « test » de terminale est fondé sur la variable aléatoire $D^2 = \sum_{i=1}^{k} (F_i - \frac{1}{k})^2$ où F_i est la variable aléatoire correspondant à la fréquence de la classe i.
Le lien entre les variables D^2 et T est facile à faire.

On a : $kn D^2 = \sum_{i=1}^{k} \frac{k}{n} \times n^2 (F_i - \frac{1}{k})^2 = \sum_{i=1}^{k} \frac{(n F_i - \frac{n}{k})^2}{\frac{n}{k}}$.

Ainsi, on a $k n D^2 = T$.


a
b
a
b
c
a
b
c

REFERENCES

• **Eléments théoriques au niveau lycée et B.T.S. :**

"***L'induction statistique au lycée illustrée par le tableur***" - P. Dutarte - Ezd. Didier.

"***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.

• **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***"

"***Statistique et citoyenneté***"

• **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.