Statistiques

G. Vinsard

20 mars 2013

Table des matières

Introduction

La statistique peut être vue comme un corps de disciplines apportant des réponses formalisées à la question de l'inférence (l'induction) de lois générales à partir d'observations particulières. Elle peut également être vue comme une série de techniques d'aide à la décision.

L'enseignement n'aborde pas ces aspects qui sont développés avec détails dans

> A. Desrosière, *La politique des grands nombres. Histoire de la raison statistique*, Paris, La découverte.

notamment dans le chapitre 3 où l'auteur fait apparaître le débat moderne entre holisme et individualisme comme la transposition sous une autre forme de celui qui opposait les réalisme des idées et le nominalisme au Moyen-Âge et cela en parcourant les cheminement des formes intermédaires aux XVIII^{e} et XIX^{e}siècles. D'autres chapitres montrent comment la science statistique s'est incorporée au cours du temps dans l'organisation de l'état. Pour une approche plus proche des pratiques quotidiennes de l'ingénieur on pourra lire avec profit

> M. Vessereau, *La statistique*, Puf., Que sais-je ?, 20^{e} édition, 1999 (1^{o} ed. 1947)

et bien entendu d'autres lectures peuvent être profitables.
Les lois de Gauss-Laplace, du χ^2, de Student-Fisher et de Snedecor sont utilisées mais les problèmes d'analyse numérique auxquels conduit leur manipulation doivent être considérés comme annexes. Le point de vue est qu'il existe toujours un logiciel capable de calculer les « Cumulative Distribution Functions » et « Probability Distribution Functions » ou à défaut les tables numériques qu'ils suppléent de plus en plus.

Ce point de vue suppose évidemment que ces notions de pdf (densité de probabilité) et cdf (fonction de répartition) soient comprises ; ce sont des pré-requis. De la même façon les approximations gaussienne ou poissonienne de la loi binomiale sont des pré-requis. Ainsi d'ailleurs que les éléments de probabilité que sont : l'espérance mathématique, la variance (et l'écart-type), l'inégalité de Bienaymé-Tchébychev et le théorème central-limit.

Ceci posé le contenu de l'enseignement est :

- l'identification de la moyenne et de l'écart-type d'un caractère quantitatif depuis un échantillon issu d'une population dans laquelle ce caractère est distribué suivant la loi de Gauss-Laplace ;
- les tests d'hypothèses simple et composite sur les valeurs de ces paramètres et dans ce cas ;
- la comparaison de deux populations gaussiennes ; le test des appariemments ;
- la régression linéaire ;
- le raccordement de données à une loi de probabilité.

C'est un contenu qui doit beaucoup au cours de statistiques de M Chambon à l'EMN dans les années 1980 dont le cours actuel de T. Verdel disponible

> `http://tice.inpl-nancy.fr/modules/unit-stat/`

est l'héritier le plus directement accessible.

1 Description de données, Inférence et sondage

1.1 Représentation de la diversité des valeurs d'un caractère dans une population

La circularité des définitions qui suivent est manifeste ; c'est hélas une difficulté qui n'est pas évitable [1] aussi procède-t-on ici par imprégnation : un texte est fourni dont la lecture devrait procurer un éclaicissement sur le sens des termes utilisés.

Un premier problème est de définir une représentation de la diversité des valeurs d'un caractère [2] d'une population [3]. Si le caractère est quantitatif [4] une telle représentation est founie par l'individu [5] moyen. Si la population a une taille [6] N et que les valeurs du caractère quantitatif sont

$$\overline{\begin{array}{cccccc} x_1 & \dots & x_n & \dots & x_N \end{array}} \tag{1}$$

l'individu moyen est doté de la valeur

$$\overline{x} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{2}$$

du caractère. Cet individu moyen n'est pas un individu membre de la population (si cela arrive, c'est purement fortuit) et d'autre part la valeur de son caractère n'est pas nécessairement une valeur possible pour le caractère étudié.

Par exemple si le caractère est le nombre d'appendice des individus (0 pour ceux qui ont été opérés de l'appendicite, 1 pour les autres) l'individu moyen peut être affecté de... disons 0.7 appendices.

Les caractères quantitatifs peuvent être discrets (le nombre d'appendice) ou continus (la hauteur des individus) : dans le cas continu l'individu moyen est doté d'une valeur de caractère qui a un sens mais ce n'est pas pour autant qu'il cesse d'être une abstraction née de la volonté de représentation de la diversité des valeurs de ce caractère dans une population.

Si par exemple la population est composée de trois enfants d'âges différents qui mesurent $1\ m$, $1.70\ m$ et $1.80\ m$, l'enfant moyen mesurera $1,30\ m$ et il est à parier que la confection de trois pantalons identiques adaptés à un individu d'$1.30\ m$ ne satisferait aucun de ces enfants. Et pourtant il est a priori plus économique de fabriquer trois pantalons identiques plutôt que trois différents.

Une autre paramètre de représentation de cette diversité de valeur d'un caractère est la médiane. La médiane est une valeur Q_2 telle que si les valeurs de caractère des individus sont triées

$$x_1 \leq x_2 \leq \ldots \leq x_N$$

alors si N est impair $Q_2 = x_{(N-1)/2}$ et si N est pair

$$x_1 \leq \ldots \leq x_{N/2} \leq Q_2 \leq x_{N/2+1} \leq \ldots \leq x_N$$

La médiane partage les individus en deux groupes : ceux dont la valeur de caractère est plus petite qu'elle et les autres.

1. Chercher ce qu'est le « trilemme d'Agrippa » pour avoir une idée sur la nature de ce genre de difficultés.
2. Un caractère est ce dont il s'agit d'analyser la distribution dans une population : par exemple la couleur des yeux dont les valeurs sont les couleurs possibles.
3. Une population est l'ensemble des individus comportant le caractère dont il s'agit d'analyser la distribution : par exemple les élèves de ce cours.
4. Un caractère quantitatif a des valeurs numériques : par exemple la hauteur (la taille mais « taille » est utilisée par ailleurs pour « nombre d'individus ») des individus composant la population des élèves de ce cours.
5. Un individu est un membre de la population.
6. « taille » est ici le mot utilisé pour donner le nombre d'individus de la population.

Cette notion de médiane s'étend avec les quartiles Q_1, Q_2 (la médiane) et Q_3 qui partagent la population en 4 groupes de taille égale. D'une façon plus générale encore il y a les quantiles qui partagent la population en un nombre de groupes quelconques mais telle que chacun des goupes comporte le même nombre d'individus.

Par exemple l'échelle de notation ECTS consiste à découper ceux des élèves qui valident un module en 5 groupes :

A	les 10%	meilleurs
B	25%	meilleurs parmi les suivants
C	30%	-
D	25%	-
E	10%	derniers

D'autre part un enseignant qui note à la française dispose d'une liste de note

$$\overline{\overline{\quad x_1 \quad \dots \quad x_n \quad \dots \quad x_N \quad}} \tag{3}$$

La conversion vers la notation ECTS utilise une variante des quantiles (pour les quantiles le nombre d'individu dans chacun des groupes reste le même) :

1. les nombres $N_A = N\ 10/100$, $N_B = N\ 25/100$, ..., $N_c = N\ 10/100$ sont calculés ;
2. ils sont arrondis à l'entier le plus proche (compte tenu de la contrainte $N_A+N_B+\ldots+N_E = N$) ;
3. les notes sont triées en ordre décroissant ;
4. la note 'quantile' séparant les groupes A et B est x_{AB} telle que

$$x_{N_A} < x_{AB} < x_{N_A+1}$$

5. mais bien évidemment des ennuis peuvent arriver dans les cas où certaines notes se répètent. Ce qui fait que ce type de conversion n'est pas si clair.

La moyenne et la médiane sont des paramètres indiquant la valeur à prendre s'il s'agit de réduire la population à un « individu moyen » pour un caractère donné. En ne considérant toujours que le caractère, des paramètres indiquant la diversité des valeurs qu'il peut prendre sont : l'écart-type

$$s = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})^2} = \sqrt{\left(\frac{1}{N}\sum_{n=1}^{N} x_n^2\right) - \overline{x}^2}$$

l'étendue

$$e_x = \max_{n\in[1,N]} x_n - \min_{n\in[1,N]} x_n$$

qui peut être introduit sous une forme moins sensible aux extrêmes comme

$$e_Q = Q_3 - Q_1$$

Ces paramètres caractérisent la dispersion de la diversité des valeurs. Plus ils sont grands par rapport à une valeur caractéristique des valeurs [7] plus les valeurs sont dispersées.

7. c'est une notion un peu floue qui ne prend de sens que sur des cas concrets

1.2 Inférence

En plus des questions de mise en forme et de classement des données, la statistique s'occupe d'inférer des lois générales à partir d'observations particulières. C'est une question dont la considération du « paradoxe des corbeaux d'Hempel » permet de mesurer la difficulté. Ce paradoxe peut se décomposer comme suit :

1. je veux connaître la couleur des corbeaux ;
2. j'observe donc des corbeaux (je sais reconnaître les corbeaux indépendamment de leur couleur) et je note que tous ceux que j'ai vu étaient noirs ;
3. j'infère alors que « tous les corbeaux sont noirs » ;
4. mais j'ai des doutes sur la vérité de la proposition, après tout je ne peux parler que des corbeaux que j'ai vu, pas des autres ;
5. alors j'essaie de théoriser et je me dis qu'on peut peut-être affecter à une proposition une valeur de vérité ; par exemple un nombre entre 0 et 1 ; pour 0 la proposition est fausse ; pour 1 elle est vraie ; et entre les deux elle est d'autant plus vraie qu'elle est proche de 1 ;
6. l'utilisation que je ferais de cette valeur de vérité est de décider que chaque fois que je vois un corbeau noir alors elle augmente ; et évidemment si je voyais un corbeau blanc elle se positionnerait à zéro pour y rester ;
7. j'examine maintenant cet essai de théorisation sur l'exemple des corbeaux ; pour formuler les choses plus précisément, je décide de placer l'ensemble des corbeaux dans l'ensemble des oiseaux ; et je remplace la proposition « tous les corbeaux sont noirs » par « pour tout oiseau, s'il est un corbeau alors il est noir » ;
8. l'intérêt de ce remplacement est que je suis maintenant dans les formulations canoniques de la logique ordinaire qui m'apprend que la proposition « pour tout oiseau, s'il est un corbeau alors il est noir » est logiquement équivalente à sa contraposée « pour tout oiseau, s'il n'est pas noir alors il n'est pas un corbeau » ;
9. si je raccorde maintenant ma théorisation de 5 au point 8, j'accepte que les valeurs de vérité d'une proposition et de sa contraposée soient identiques ; si donc je trouve un moyen de modifier la valeur de vérité de la contraposée, je modifie du même coup celle de la proposition première ;
10. il me semble maintenant que j'augmente la valeur de vérité de « pour tout oiseau, s'il n'est pas noir alors il n'est pas un corbeau » chaque fois que je vois un oiseau jaune et qui est un canari ; exactement comme j'augmente la valeur de vérité de « pour tout oiseau, s'il est un corbeau alors il est noir » chaque fois que je vois un corbeau noir ;
11. j'arrive donc au paradoxe : chaque fois que je vois un canari jaune, cela augmente la valeur de vérité que je porte à la proposition selon laquelle « pour tout oiseau, s'il est un corbeau alors il est noir » ; ce que je n'accepte pas !

Ce paradoxe se lève en attaquant le point le plus faible qui est la tentative de théorisation en 5. La définition d'une notion n'a pas nécessairement pour conséquence que cette notion permette d'interpréter les faits. Ici la notion de « valeur de vérité » d'une proposition conduit au résultat 11. Et si ce résultat est jugé aberrant il convient de retirer à la notion de « valeur de vérité » la propriété 6 d'être augmentée à chaque fois qu'on on voit en corbeau noir. Mais sans cette propriété, la notion devient inutile. En matière de concept comme ailleurs, ce qui est inutile est encombrant et on gagne toujours beaucoup à éviter de s'encombrer.

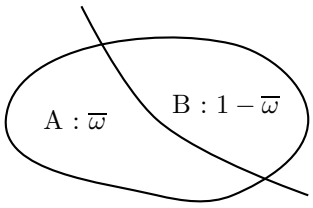
L'ennui c'est qu'on se demande bien comment faire autrement. Si on ne voit que des corbeaux noirs, on a envie de penser que « tous les corbeaux sont noirs » ; c'est à dire d'inférer une connaissance générale à partir d'observations particulières.

L'objet principal de la statistique est de pratiquer cette activité « inférer une connaissance générale à partir d'observations particulières » sans s'encombrer d'abstraction inutiles comme celle de « valeur de vérité ».

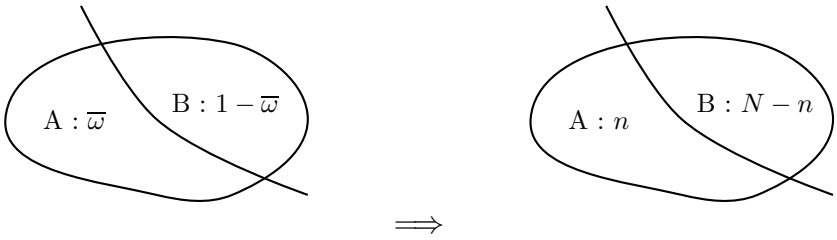
1.3 Un exemple emblématique : la prévision des résultats d'une élection

L'analyse statistique d'un résultat de sondage met en œuvre les concepts essentiels de cet enseignement. Il y a une population composée d'individus dont le caractère qu'on étudie a

pour valeurs est soit 'vote pour A' soit 'vote pour B' ; A et B étant deux candidats à une élection.



Il s'agit de savoir si $\overline{\omega}$ la proportion des individus de la population votant pour A est supérieure à $1/2$. Pour cela on décide d'un nombre N d'individus de la population à interroger : c'est un échantillon de taille N.



$\Longrightarrow$

Si on suppose qu'il n'y a pas de menteurs et que tous les individus acceptent de répondre, il y a donc un nombre n d'entre eux qui déclareront voter pour 'A' et les autres $N-n$ déclareront voter pour 'B'. La question est de déterminer si sur la foi de ces donnéees il est possible d'inférer le résultat de l'élection. Mais une première étape préalable est d'identifier la proportion $\overline{\omega}$.

Identification par l'inégalité de Bienaymé-Tchébychev

Le nombre n est une réalisation d'une variable aléatoire binomiale X_b de paramètre N et $\overline{\omega}$, i.e.

$$\mathbb{P}(X_b = n) = \mathcal{C}_N^n\, \overline{\omega}^n\, (1-\overline{\omega})^{N-n} \;\text{ et }\; \mathbb{E}(X_b) = N\, p\; ;\; \mathbb{V}ar(X_b) = N\, \overline{\omega}\, (1-\overline{\omega})$$

On dispose d'un premier résultat exploitable *ab initio* qui est l'inégalité de Bienaymé-Tchébychev

$$\mathbb{P}\left(|X_b - N\, \overline{\omega}| > \frac{\sqrt{N\, \overline{\omega}\, (1-\overline{\omega})}}{\sqrt{\alpha}} \right) < \alpha$$

Dans plus de $1-\alpha$ des cas le nombre $\tilde{n}$ d'électeurs de 'A' parmi un échantillon de N personnes satisfera à

$$|\tilde{n} - N\, \overline{\omega}| < \frac{\sqrt{N\, \overline{\omega}\, (1-\overline{\omega})}}{\sqrt{\alpha}} \text{ soit } N\, \overline{\omega} - \frac{\sqrt{N\, \overline{\omega}\, (1-\overline{\omega})}}{\sqrt{\alpha}} < \tilde{n} < N\, \overline{\omega} + \frac{\sqrt{N\, \overline{\omega}\, (1-\overline{\omega})}}{\sqrt{\alpha}}$$

L'affirmation que le nombre n effectivement obtenu satisfait à

$$N\, \overline{\omega} - \frac{\sqrt{N\, \overline{\omega}\, (1-\overline{\omega})}}{\sqrt{\alpha}} < n < N\, \overline{\omega} + \frac{\sqrt{N\, \overline{\omega}\, (1-\overline{\omega})}}{\sqrt{\alpha}}$$

est donc juste sauf dans moins de α des cas où elle est fausse.

Cette double inégalité s'inverse algébriquement pour fournir

$$\frac{n}{N} + \frac{1}{N}\frac{\frac{1}{2} - \frac{n}{N}}{\alpha + \frac{1}{N}} - \frac{1}{\sqrt{N}}\frac{\sqrt{\frac{1}{4N} + \alpha\frac{n}{N}\left(1 - \frac{n}{N}\right)}}{\alpha + \frac{1}{N}} < \overline{\omega} < \frac{n}{N} + \frac{1}{N}\frac{\frac{1}{2} - \frac{n}{N}}{\alpha + \frac{1}{N}} + \frac{1}{\sqrt{N}}\frac{\sqrt{\frac{1}{4N} + \alpha\frac{n}{N}\left(1 - \frac{n}{N}\right)}}{\alpha + \frac{1}{N}}$$

Et l'affirmation portant sur la quantité 'n' qui se trouve être dans un intervalle dépendant de $\overline{\omega}$ (sauf dans moins de α des cas) s'inverse logiquement pour devenir : sauf dans moins de α des cas où l'affirmation est fausse, $\overline{\omega}$ est tel que

$$\frac{n}{N}+\frac{1}{N}\frac{\frac{1}{2}-\frac{n}{N}}{\alpha+\frac{1}{N}}-\frac{1}{\sqrt{N}}\frac{\sqrt{\frac{1}{4N}+\alpha\frac{n}{N}\left(1-\frac{n}{N}\right)}}{\alpha+\frac{1}{N}}<\overline{\omega}<\frac{n}{N}+\frac{1}{N}\frac{\frac{1}{2}-\frac{n}{N}}{\alpha+\frac{1}{N}}+\frac{1}{\sqrt{N}}\frac{\sqrt{\frac{1}{4N}+\alpha\frac{n}{N}\left(1-\frac{n}{N}\right)}}{\alpha+\frac{1}{N}}$$

C'est une version affaiblie du programme d'inférence de loi générale à partir d'observations particulière illustré par le problème des corbeaux d'Hempel. Et elle n'aboutit à rien d'aberrant : si $\alpha = 0$ de manière que l'affimation soit toujours juste alors l'inégalité devient

$$0 < \overline{\omega} < 1$$

elle n'affirme rien : il est certain que la proportion de gens qui votent pour 'A' est comprise entre 0 et 1 (les inégalités strictes auraient pu être remplacées par des inégalités larges sans rien changer dès l'inégalité de Bienaymé-Tchébychev). Il n'est pas possible d'inférer sans risque.

L'inégalité de Bienaymé-Tchébychev est assez large : pour que la quantité

$$\frac{1}{\sqrt{N}}\frac{\sqrt{\frac{1}{4N}+\alpha\frac{n}{N}\left(1-\frac{n}{N}\right)}}{\alpha+\frac{1}{N}}$$

soit plus petite qu'un certain seuil disons 1% en acceptant un risque $\alpha = 0.05 = 5\%$ et dans le cas où $n/N \approx 1/2$ il faut que $N > 50000$. C'est très au delà des tailles usuelles d'échantillons dans les sondages et il est nécessaire pour les comprendre d'aller plus loin.

Identification par la loi de Gauss-Laplace

La loi de X_b est en fait connue, c'est la loi binomiale, il existe donc un nombre déterminé t_α tel que

$$\mathbb{P}\left(|X_b - N\,\overline{\omega}| > t_\alpha\,\sqrt{N\,\overline{\omega}\,(1-\overline{\omega})}\right) = \alpha \tag{4}$$

L'inégalité de Bienaymé-Tchébychev est maintenant remplacée par une égalité. Et la question est de trouver la relation entre α et t_α afin de refaire le raisonnement qui a été tenu.

Cette recherche passe déjà par l'approximation de la loi binomiale de X_b à une loi de Gauss-Laplace de paramètre $\mu = N\,\overline{\omega}$ et $\sigma = \sqrt{N\,\overline{\omega}\,(1-\overline{\omega})}$ (cf. éléments de probabilité en appendice)

$$\begin{aligned}\mathbb{P}(n_1 \leq X_b \leq n_2) &\approx \mathbb{P}\left(t_1 = \frac{\left(\frac{n_1}{N}-\overline{\omega}\right)}{\sqrt{N\,\overline{\omega}\,(1-\overline{\omega})}} \leq T \leq t_2 = \frac{\left(\frac{n_2}{N}-\overline{\omega}\right)}{\sqrt{N\,\overline{\omega}\,(1-\overline{\omega})}}\right)\\ &= 1/\sqrt{2\pi}\int_{t_1}^{t_2}\exp^{-x^2/2}\,dx\end{aligned}$$

où T est la variable aléatoire de Gauss-Laplace de moyenne nulle et d'écart-type unité, c'est à dire la loi normale centrée réduite.

L'égalité (4) peut alors être réécrite

$$\mathbb{P}\left(|X_b - N\,\overline{\omega}| > t_\alpha\,\sqrt{N\,\overline{\omega}\,(1-\overline{\omega})}\right) \approx \underbrace{2\,\frac{1}{2\pi}\int_{t_\alpha}^{\infty}\exp^{-x^2/2}\,dx}_{=\alpha}$$

et la reprise du raisonnement fait dans le paragraphe précédent permet d'affirmer que : sauf dans α des cas où l'affirmation est fausse, $\overline{\omega}$ est tel que

$$\frac{n}{N}+\frac{1}{N}\frac{\frac{1}{2}-\frac{n}{N}}{\frac{1}{t_\alpha^2}+\frac{1}{N}}-\frac{1}{\sqrt{N}}\frac{\sqrt{\frac{1}{4N}+\frac{1}{t_\alpha^2}\frac{n}{N}\left(1-\frac{n}{N}\right)}}{\frac{1}{t_\alpha^2}+\frac{1}{N}}<\overline{\omega}<\frac{n}{N}+\frac{1}{N}\frac{\frac{1}{2}-\frac{n}{N}}{\frac{1}{t_\alpha^2}+\frac{1}{N}}+\frac{1}{\sqrt{N}}\frac{\sqrt{\frac{1}{4N}+\frac{1}{t_\alpha^2}\frac{n}{N}\left(1-\frac{n}{N}\right)}}{\frac{1}{t_\alpha^2}+\frac{1}{N}}$$

où

$$\alpha = 2\,\frac{1}{2\pi}\,\int_{t_\alpha}^{\infty} \exp^{-x^2/2}\,dx$$

Cette formule est un peu compliquée. Si on pose

$$n/N = \overline{\omega}^*$$

et qu'on la développe par rapport à $1/N$ alors que N est supposé grand il vient

$$\overline{\omega}^* - t_\alpha\,\frac{\sqrt{\overline{\omega}^*\,(1-\overline{\omega}^*)}}{\sqrt{N}} + t_\alpha^2\,\frac{\overline{\omega}^* - \frac{1}{2}}{N} + \ldots < \overline{\omega} < \overline{\omega}^* + t_\alpha\,\frac{\sqrt{\overline{\omega}^*\,(1-\overline{\omega}^*)}}{\sqrt{N}} + t_\alpha^2\,\frac{\overline{\omega}^* - \frac{1}{2}}{N} + \ldots$$

soit encore en ne conservant que le terme en $1/\sqrt{N}$ qui est le moins petit dans le développement

$$\overline{\omega}^* - t_\alpha\,\frac{\sqrt{\overline{\omega}^*\,(1-\overline{\omega}^*)}}{\sqrt{N}} < \overline{\omega} < \overline{\omega}^* + t_\alpha\,\frac{\sqrt{\overline{\omega}^*\,(1-\overline{\omega}^*)}}{\sqrt{N}}$$

Si on considère de plus que les sondages n'ont d'intérêt que lorsque

$$\overline{\omega} \approx \frac{1}{2} \Longrightarrow \sqrt{\overline{\omega}^*\,(1-\overline{\omega}^*)} \approx \frac{1}{2} \quad (\overline{\omega} = 3/10 \Longrightarrow \sqrt{\overline{\omega}^*\,(1-\overline{\omega}^*)} \approx 46/100)$$

la formule se simplifie finalement comme

$$\overline{\omega}^* - \frac{t_\alpha}{2\,\sqrt{N}} < \overline{\omega} < \overline{\omega}^* + \frac{t_\alpha}{2\,\sqrt{N}}$$

Il reste à donner la correspondance numérique entre α et t_α qui se trouve dans les tables de Gauss mais dont il est bon de conserver quelques valeurs à l'esprit

α	0%	1%	5%	10%
t_α	∞	2.57	1.96	1.64

Pour un risque de 5%, en approximant 1.96 par 2, il est donc possible d'affirmer (avec donc un risque de se tromper de 5%) que

$$\frac{n}{N} - \frac{1}{\sqrt{N}} < \overline{\omega} < \frac{n}{N} + \frac{1}{\sqrt{N}}$$

Cette analyse est caractéristique de l'opération d'identification :

→ le paramètre $\overline{\omega}$ (la proportion d'électeurs de 'A') à identifier a une valeur déterminée mais qui n'est pas connue ;

→ cette valeur est placée dans un intervalle construit à partir d'une valeur n/N qui est la réalisation d'une variable aléatoire (et donc dont la valeur est certes connue pour l'échantillon choisi mais elle aurait été autre avec un autre échantillon) et de la considération d'un risque d'erreur α (qui est la probabilité pour que $\overline{\omega}$ n'appartienne pas à l'intervalle).

La valeur n/N s'appelle une estimation de $\overline{\omega}$ et l'intervalle s'appelle un intervalle de confiance au risque α (ici 5%).

Il est intéressant de noter qu'avec un risque nul ($t_\alpha = \infty$ on obtient un intervalle de confiance

$$-\infty < \overline{\omega} < \infty$$

qui ne donne aucune indication sur la valeur de $\overline{\omega}$. La demande de certitude absolue ne permet pas d'obtenir d'information (cf. les corbeaux d'Hempel).

Inférence du résultat de l'élection

L'intervalle de confiance peut être exploité directement pour l'inférence du résultat de l'élection. Mais ce qu'est véritablement un test statistique apparaît mieux en recommençant l'analyse mais en réutilisant les résultats qui ont conduit à l'intervalle de confiance.

Donc n individus d'un échantillon de taille N (issu d'une population dont la proportion d'électeurs de 'A' est $\overline{\omega}$) se déclarent électeurs de 'A'.

L'hypothèse qu'il s'agit de tester est *a priori*

$$H : \overline{\omega} > \frac{1}{2}$$

et la contre hypothèse est

$$\overline{H} : \overline{\omega} \leq \frac{1}{2}$$

Ce genre d'hypothèse est appelé hypothèse composite parce qu'elle ne porte pas sur une valeur simple de $\overline{\omega}$ mais sur toute une plage de valeur. Il est plus simple de commencer par expliquer le principe du test statistique sur un hypothèse simple avant de l'étendre au cas composite. Aussi commence-t-on par le test de

$$H' : \overline{\omega} = \delta$$

où $\delta \in [0, 1]$ est un nombre quelconque. Il faut remarquer que la contre-hypothèse

$$\overline{H'} : \overline{\omega} \neq \delta$$

est alors composite.

Test de l'hypothèse simple Le test d'hypothèse simple permet la réutilisation directe de l'intervalle de confiance de $\overline{\omega}$ au risque α ; celui ci est

$$\frac{n}{N} - \frac{t_\alpha}{2\sqrt{N}} < \overline{\omega} < \frac{n}{N} + \frac{t_\alpha}{2\sqrt{N}}$$

sauf dans α des cas où l'affirmation est fausse $\overline{\omega}$ est dans cet intervalle.

Et donc deux cas peuvent se produire : si d'abord

$$\delta \notin \left[\frac{n}{N} - \frac{t_\alpha}{2\sqrt{N}}, \frac{n}{N} + \frac{t_\alpha}{2\sqrt{N}}\right]$$

l'hypothèse H' est alors fausse sauf dans α des cas.

Et ensuite si

$$\delta \in \left[\frac{n}{N} - \frac{t_\alpha}{2\sqrt{N}}, \frac{n}{N} + \frac{t_\alpha}{2\sqrt{N}}\right]$$

cela ne signifie pas que H' soit vraie sauf dans α des cas. Si on le croyait il viendrait que $\forall\, q \in \mathbb{N}$ les hypothèses

$$H_q : \overline{\omega} = \delta + \frac{1}{1+q^2}\left(\frac{t_\alpha}{2\sqrt{N}} - \delta\right)$$

sont également vraies sauf dans α des cas. Ces hypothèses H_q sont des parties de la contre-hypothèse $\overline{H'}$ et donc exclusives de H', leur nombre infini ne permet pas d'affirmer que c'est seulement dans α des cas que H' serait fausse parce qu'elles seraient vraies.

Le cas pour lequel

$$\delta \in \left[\frac{n}{N} - \frac{t_\alpha}{2\sqrt{N}}, \frac{n}{N} + \frac{t_\alpha}{2\sqrt{N}}\right]$$

est celui où il n'est pas possible d'affirmer autre chose que : l'hypothèse H' n'a pas été infirmée (mise en défaut). Et alors on l'accepte par défaut.

La seule affirmation possible sur une hypothèse simple porte sur sa réfutation (comme l'hypothèse H' est fausse au risque α de la croire fausse alors qu'elle est juste), c'est l'aspect négatif des tests. Il n'est jamais possible de dire d'une hypothèse simple qu'elle est juste avec un certain risque de la croire juste alors qu'elle est fausse.

Pour appuyer cela on peut faire varier la valeur du risque. Si α devient petit, t_α devient grand et de plus en plus de nombre δ appartiennent à l'intervalle d'acceptation : à la limite $\alpha \to 0 \Longrightarrow t_\alpha \to \infty$ n'importe quel nombre δ dans l'hypothèse H' peut être accepté. Ce résultat conforte le principe (les corbeaux d'Hempel) selon lequel rien ne peut être affirmé avec un risque nul.

Test de l'hypothèse composite D'une part la considération de l'hypothèse composite change l'intervalle d'acceptation et d'autre part la remarque sur l'aspect négatif des tests montre que l'hypothèse qu'il faut tester pour savoir si le candidat 'A' sera ou non élu est plutôt la version stricte H'' de la contre-hypothèse H'

$$H'' : \overline{\omega} < 1/2$$

En reprenant l'approximation gaussienne de moyenne $\overline{\omega}$ et d'écart-type $1/(2\sqrt{N})$ de la variable aléatoire dont le rapport n/N est une réalisation, ce rapport sera dans l'intervalle

$$\left]-\infty, \overline{\omega} + \frac{t_{2\alpha}}{2\sqrt{N}}\right]$$

dans $1-\alpha$ des cas où la relation entre α et t_α est toujours

$$\alpha = 2\,\frac{1}{2\pi}\int_{t_\alpha}^{\infty} \exp^{-x^2/2}\,dx \Longleftrightarrow \alpha = \frac{1}{2\pi}\int_{t_{2\alpha}}^{\infty} \exp^{-x^2/2}\,dx$$

Si H'' est juste alors

$$\left]-\infty, \overline{\omega} + \frac{t_{2\alpha}}{2\sqrt{N}}\right] \subset \left]-\infty, \frac{1}{2} + \frac{t_{2\alpha}}{2\sqrt{N}}\right]$$

et donc

$$\left]\frac{1}{2} + \frac{t_{2\alpha}}{2\sqrt{N}}, \infty\right[\subset \left]\overline{\omega} + \frac{t_{2\alpha}}{2\sqrt{N}}, \infty\right[$$

Si n/N est dans l'intervalle où il ne peut être que dans α des cas lorsque $\overline{\omega} = 1/2$

$$\frac{n}{N} \in \left]\frac{1}{2} + \frac{t_{2\alpha}}{2\sqrt{N}}, \infty\right[$$

alors *a fortiori* il sera dans l'intervalle où il ne peut être que dans α des cas lorsque $\overline{\omega} < 1/2$

$$\frac{n}{N} \in \left]\overline{\omega} + \frac{t_{2\alpha}}{2\sqrt{N}}, \infty\right[$$

et donc il est possible de conclure que l'hypothèse H'' est fausse sauf dans α des cas : le candidat 'A' sera élu au risque α de se tromper dans l'affirmation.

Dans le cas où

$$\frac{n}{N} \notin \left]\frac{1}{2} + \frac{t_{2\alpha}}{2\sqrt{N}}, \infty\right[$$

on ne peut rien affirmer d'autre que l'hypothèse H" n'a pas été mise en défaut.

Évidemment en interchangeant 'A' et 'B' on voit que si

$$\frac{n}{N} \in \left]-\infty, \frac{1}{2} - \frac{t_{2\alpha}}{2\sqrt{N}}\right[$$

la conclusion peut être que le candidat 'B' sera élu au risque α de se tromper.

Valeurs numériques L'intervalle de confiance au risque α

$$\overline{\omega}^* - \frac{t_\alpha}{2\sqrt{N}} < \overline{\omega} < \overline{\omega}^* + \frac{t_\alpha}{2\sqrt{N}}$$

permet d'estimer la taille d'échantillon à choisir pour réaliser un sondage qui peut conduire à une prévisions raisonnablement fiable.

Pour une proportion $\overline{\omega}$ et un risque α donnés il faut que N soit tel que 1/2 n'appartienne pas à l'intervalle

	α= 1%	5%	10%
$\overline{\omega}$= 0.51	16512	9604	6724
0.52	4128	2401	1681
0.53	1834	1067	747

Les sondages font en général état d'un échantillon de taille 800.

Exercice type :

Je vais chercher le pain régulièrement à 10h30 le dimanche matin à la même boulangerie depuis dix ans. J'ai noté qu'il y a en moyenne 4 clients devant moi. Il me semble savoir (mais l'objectif n'est pas de le montrer) que le nombre de clients dans une boutique à date fixe peut être assimilé à une variable aléatoire de Poisson.

Aujourd'hui j'ai perdu la mémoire de toute chose à l'exclusion de ce qui précède. Puis-je conclure que je suis bien à ma boulangerie habituelle, qu'aujourd'hui est dimanche et qu'il est 10h30 si je suis dans une boulangerie et qu'il y a : a) 0 b) 1 c) 6 d) 15 clients devant moi.

Puisque cela fait 10 ans que j'ai constaté la loi indiquée elle l'a été plus de 520 dimanches et on suppose implicitement que la loi est exacte.

Premier traitement de l'exercice

Cas a

'Si j'étais dans la boulangerie, le dimanche, à 10 h 30 alors c'est dans 1.8 % des cas qu'il y aurait 0 clients au moment de mon arrivée.'

Si on ajoute que cette probabilité de 1.8 % est négligeable alors la phrase précédente peut être ré-écrite comme

'Si j'étais dans la boulangerie, le dimanche, à 10 h 30 alors il est certain qu'il ne peut y avoir 0 clients au moment de mon arrivée.'

Et alors comme il y en a effectivement 0 c'est que je ne suis pas dans la boulangerie, le dimanche, à 10 h 30.

Considérer les étapes : on forme une phrase juste d'après les données du problème et qui comporte une implication; on déforme cette phrase en une phrase approximativement juste (elle est juste à 1.8 % prés); et on prend brutalement la contraposée qui est la conclusion.

Cas b

On recopie le raisonnement du cas a pour obtenir

'Si j'étais dans la boulangerie, le dimanche, à 10 h 30 alors c'est dans 7.3 % des cas qu'il y aurait 1 clients au moment de mon arrivée.'

Il est plus difficile de négliger 7.3 % que 1.8 %. Si on le fait alors on obtient exactement la même conclusion que dans le cas a.

Si on ne le fait pas alors on n'obtient aucune conclusion possible. On peut être ou non dans la boulangerie, le dimanche, à 10 h 30.

Alors intervient l'idée selon laquelle est vrai ce qu'on ne peut pas nier, d'où la conclusion (qui ne repose sur rien) que je suis bien dans la boulangerie au jour et à l'heure dite.

Cas c

Idem au cas b, en plus convaincant pour accepter d'être dans la boulangerie au jour et à l'heure dite.

Cas d

Idem au cas a, en plus convaincant pour infirmer l'hypothèse selon laquelle je suis dans la boulangerie au jour et à l'heure dite.

Deuxième traitement de l'exercice

Quelque chose est un peu frustrant dans le premier traitement qui peut être illustré en changeant la loi de probabilité.

Par exemple : le dimanche, dans cette boulangerie, à 10 h 30 il y aurait équiprobabilité entre 0 et 99 clients et une probabilité nulle pour plus de clients.

Si on tient 1 % pour négligeable, pour n'importe quel nombre < 100 de clients on serait amené à nier être dans la boulangerie, ce qui n'est pas raisonnable.

Aussi il est préférable de commencer par donner une région d'acceptation pour être dans la boulangerie et seulement après effectuer les tests.

Dans le cas de la loi de Poisson de paramètre 4 on peut par exemple décider que dans $0.979 - 0.018 \approx 96$ % des cas on trouve de 1 à 7 clients ; puis décider que 4 % est négligeable.

Ainsi si on a de 1 à 7 clients on ne peut conclure d'être dans la boulangerie ; et dans tous les autres cas conclure qu'au risque 4 % de se tromper on n'y est pas.

Cet exemple pourrait faire penser que le traitement statistique d'un problème comporte une part subjective (voire même une escroquerie), celle qui est liée au choix de la valeur du risque, ou les modalités d'application des méthodes. C'est certain, mais comment mieux penser ?

$\lambda = \mathbf{0.1}$		
0	0.905	0.905
1	0.09	0.995

$\lambda = \mathbf{0.2}$		
0	0.819	0.819
1	0.164	0.982
2	0.016	0.999

$\lambda = \mathbf{0.4}$		
0	0.67	0.67
1	0.268	0.938
2	0.054	0.992

$\lambda = \mathbf{0.6}$		
0	0.549	0.549
1	0.329	0.878
2	0.099	0.977
3	0.02	0.997

$\lambda = \mathbf{0.8}$		
0	0.449	0.449
1	0.359	0.809
2	0.144	0.953
3	0.038	0.991

$\lambda = \mathbf{1.0}$		
0	0.368	0.368
1	0.368	0.736
2	0.184	0.92
3	0.061	0.981
4	0.015	0.996

$\lambda = \mathbf{2.0}$		
0	0.135	0.135
1	0.271	0.406
2	0.271	0.677
3	0.18	0.857
4	0.09	0.947
5	0.036	0.983
6	0.012	0.995

$\lambda = \mathbf{3.0}$		
0	0.05	0.05
1	0.149	0.199
2	0.224	0.423
3	0.224	0.647
4	0.168	0.815
5	0.101	0.916
6	0.05	0.966
7	0.022	0.988

$\lambda = \mathbf{4.0}$		
0	0.018	0.018
1	0.073	0.092
2	0.147	0.238
3	0.195	0.433
4	0.195	0.629
5	0.156	0.785
6	0.104	0.889
7	0.06	0.949
8	0.03	0.979
9	0.013	0.992

$\lambda = \mathbf{8.0}$		
2	0.011	0.014
3	0.029	0.042
4	0.057	0.1
5	0.092	0.191
6	0.122	0.313
7	0.14	0.453
8	0.14	0.593
9	0.124	0.717
10	0.099	0.816
11	0.072	0.888
12	0.048	0.936
13	0.03	0.966
14	0.017	0.983

$\lambda = \mathbf{10.0}$		
4	0.019	0.029
5	0.038	0.067
6	0.063	0.13
7	0.09	0.22
8	0.113	0.333
9	0.125	0.458
10	0.125	0.583
11	0.114	0.697
12	0.095	0.792
13	0.073	0.864
14	0.052	0.917
15	0.035	0.951
16	0.022	0.973
17	0.013	0.986

$\lambda = \mathbf{15.0}$		
7	0.01	0.018
8	0.019	0.037
9	0.032	0.07
10	0.049	0.118
11	0.066	0.185
12	0.083	0.268
13	0.096	0.363
14	0.102	0.466
15	0.102	0.568
16	0.096	0.664
17	0.085	0.749
18	0.071	0.819
19	0.056	0.875
20	0.042	0.917
21	0.03	0.947
22	0.02	0.967
23	0.013	0.981

$\lambda = \mathbf{20.0}$		
11	0.011	0.021
12	0.018	0.039
13	0.027	0.066
14	0.039	0.105
15	0.052	0.157
16	0.065	0.221
17	0.076	0.297
18	0.084	0.381
19	0.089	0.47
20	0.089	0.559
21	0.085	0.644
22	0.077	0.721
23	0.067	0.787
24	0.056	0.843
25	0.045	0.888
26	0.034	0.922
27	0.025	0.948
28	0.018	0.966
29	0.013	0.978

$\lambda = \mathbf{30.0}$		
20	0.013	0.035
21	0.019	0.054
22	0.026	0.081
23	0.034	0.115
24	0.043	0.157
25	0.051	0.208
26	0.059	0.267
27	0.066	0.333
28	0.07	0.403
29	0.073	0.476
30	0.073	0.548
31	0.07	0.619
32	0.066	0.685
33	0.06	0.744
34	0.053	0.797
35	0.045	0.843
36	0.038	0.88
37	0.031	0.911
38	0.024	0.935
39	0.019	0.954
40	0.014	0.968
41	0.01	0.978

$\lambda = \mathbf{60}$		
47	0.013	0.049
48	0.016	0.065
49	0.019	0.084
50	0.023	0.108
51	0.027	0.135
52	0.032	0.167
53	0.036	0.202
54	0.04	0.242
55	0.043	0.285
56	0.046	0.332
57	0.049	0.381
58	0.051	0.431
59	0.051	0.483
60	0.051	0.534
61	0.051	0.585
62	0.049	0.634
63	0.047	0.68
64	0.044	0.724
65	0.04	0.764
66	0.037	0.801
67	0.033	0.834
68	0.029	0.863
69	0.025	0.888
70	0.022	0.91
71	0.018	0.928
72	0.015	0.943
73	0.013	0.956
74	0.01	0.966

FIGURE 1 – Table de Poisson : pour les valeurs de λ données la première colonne est B, la deuxième $\frac{\lambda^N}{N!}\exp -\lambda$ et la troisième $\sum_{n=0}^{N} \lambda^n/n!\exp -\lambda$

2 Identification

Une *population* présentant un caractère *quantitatif* réparti suivant une *distribution* gaussienne ; un *échantillon* de *taille* N est prélèvé dans cette population et les valeurs numériques du caractère des individus composant cet échantillon sont reportées dans un tableau

$$\begin{array}{|cccccc|}\hline x_1 & \dots & x_n & \dots & x_N \\ \hline\end{array} \tag{5}$$

L'objectif est d'identifier à partir du tableau de données les valeurs des paramètres de cette distribution gaussienne qui sont : la moyenne μ et l'écart-type σ.

Les valeurs x_n du tableau peuvent être vues comme des réalisations d'une seule variable aléatoire de Gauss-Laplace X de moyenne μ et d'écart-type σ. Il est équivalent mais plus commode pour les notations d'introduire N variables aléatoires indépendantes et identiquement distribuées (selon la loi de Gauss-Laplace de paramètres μ et σ) notées X_n dont les x_n sont les réalisations une à une.

2.1 Écart-type connu *a priori*

La variable aléatoire

$$\overline{X} = \frac{1}{N}\sum_{n=1}^{N} X_n \tag{6}$$

est une variable aléatoire de Gauss-Laplace de moyenne μ et d'écart-type $\sigma/\sqrt{N}$ dont

$$\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{7}$$

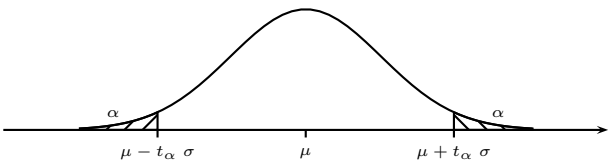
est une réalisation.

Si deux nombres, le *risque* $\alpha \in [0,1]$ et le seuil de ce risque t_α liés par

$$\int_{-t_\alpha}^{t_\alpha} \exp^{-t^2/2}/\sqrt{2\pi}\, dt = 1-\alpha \tag{8}$$

sont donnés alors la probabilité que $\overline{X}$ s'écarte de μ de plus de la quantité $t_\alpha\sigma/\sqrt{N}$ est

$$\mathbb{P}\left(|\overline{X}-\mu| > t_\alpha \frac{\sigma}{\sqrt{N}}\right) = \alpha \tag{9}$$



Ainsi $\overline{x}$ appartiendra à l'intervalle $\left[\mu - t_\alpha\frac{\sigma}{\sqrt{N}}, \mu + t_\alpha\frac{\sigma}{\sqrt{N}}\right]$ dans $1-\alpha$ des cas et il n'y appartiendra pas dans α des cas. Mais l'inégalité

$$\mu - t_\alpha\frac{\sigma}{\sqrt{N}} < \overline{x} < \mu + t_\alpha\frac{\sigma}{\sqrt{N}} \tag{10}$$

s'inverse en

$$\overline{x} - t_\alpha\frac{\sigma}{\sqrt{N}} < \mu < \overline{x} + t_\alpha\frac{\sigma}{\sqrt{N}} \tag{11}$$

Et donc μ appartiendra à l'intervalle $\left[\overline{x} - t_\alpha \frac{\sigma}{\sqrt{N}}, \overline{x} + t_\alpha \frac{\sigma}{\sqrt{N}}\right]$ dans $1-\alpha$ des cas et il n'y appartiendra pas dans α des cas.

Cette seconde formulation est le point de départ du procédé d'identification : l'échantillon est vraiment extrait de la population ; le caractère de chaque individu est mesuré ; la moyenne de ces caractères $\overline{x}$ est calculée ; et si σ est connu (hypothèse qui ne sera plus nécessaire plus loin) alors l'intervalle dans lequel il est probable (avec la probabilité $1-\alpha$)) que μ se situe sera

$$\left[\overline{x} - t_\alpha \frac{\sigma}{\sqrt{N}}, \overline{x} + t_\alpha \frac{\sigma}{\sqrt{N}}\right] \tag{12}$$

Cet intervalle porte le nom d'*intervalle de confiance* à $1-\alpha$ de la moyenne μ.

La valeur réelle de μ reste inconnue. Mais la valeur qui lui est accordée porte le nom d'*estimation*, c'est

$$\mu^* = \overline{x} \tag{13}$$

Cette dénomination vient de ce que $\overline{x}$ est une réalisation de la variable aléatoire $\overline{X}$; l'espérance mathématique de $\overline{X}$ est

$$\mathbb{E}(\overline{X}) = \mu \tag{14}$$

et une variation aléatoire dont l'espérance mathématique est une certaine quantité est appelée un estimateur de cette quantité, les réalisations de la variable aléatoire sont appelées des estimations de la quantité.

2.2 Moyenne connue *a priori*

On suppose que la moyenne μ est connue, dans ce cas on peut fabriquer la variable aléatoire

$$S^2 = \frac{1}{N}\sum_{n=1}^{N} (X_n - \mu)^2 \tag{15}$$

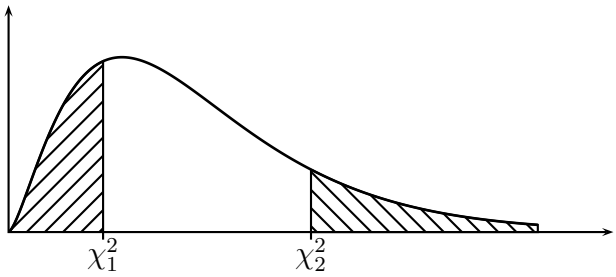
qui est telle que $N\, S^2/\sigma^2$ suive une loi du χ^2 à N degrés de liberté. On obtient une réalisation de S^2 par

$$s^2 = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu)^2$$

Si on donne trois nombres, le *risque* $\alpha \in [0,1]$, χ_1^2 et χ_2^2 tels que

$$\int_0^{\chi_1^2} \rho_{\chi^2}(u)\, du = \alpha/2 \;\text{ et }\; \int_{\chi_2^2}^{\infty} \rho_{\chi^2}(u)\, du = \alpha/2$$

où ρ_{χ^2} est la densité de probabilité de la loi du χ^2 à N degrés de liberté,



on a

$$\mathbb{P}\left(\chi_1^2 < \frac{NS^2}{\sigma^2} < \chi_2^2\right) = 1-\alpha \tag{16}$$

on peut dire alors que s^2 appartiendra à l'intervalle $\left[\frac{\chi_1^2\sigma^2}{N}, \frac{\chi_2^2\sigma^2}{N}\right]$ dans $1-\alpha$ des cas et qu'il n'y appartiendra pas dans α des cas.

Ou encore que σ appartiendra à l'intervalle $\left[\frac{\sqrt{N}\,s}{\chi_2}, \frac{\sqrt{N}\,s}{\chi_1}\right]$ dans $1-\alpha$ des cas et qu'il n'y appartiendra pas dans α des cas.

Et donc on obtient ainsi un intervalle de confiance à $1-\alpha$ de σ dont l'estimation est $\sigma^* = s$.

2.3 Moyenne et Écart-type simultanément inconnus

Maintenant la moyenne et l'écart-type ne sont pas connus *a priori*; il s'agit de les identifier simultanément.

Intervalle de confiance de l'écart-type

Dans ce cas, plutôt que (15) on forme la variable aléatoire ($\overline{X}$ définie par (6))

$$S^2 = \frac{1}{N}\sum_{n=1}^{N}(X_n - \overline{X})^2 \tag{17}$$

qui est telle que NS^2/σ^2 suit une loi du χ^2 à $N-1$ degrés de liberté. On obtient une réalisation de S^2 par $s^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})^2$. Si on donne trois nombres, le *risque* $\alpha \in [0,1]$, χ_1^2 tel que $\int_0^{\chi_1^2} \rho_{\chi^2}(u)\,du = \alpha/2$ et χ_2^2 tel que $\int_{\chi_2^2}^{\infty} \rho_{\chi^2}(u)\,du = \alpha/2$ où ρ_{χ^2} est la densité de probabilité de la loi du χ^2 à $N-1$ degrés de liberté, on a

$$\mathbb{P}\left(\chi_1^2 < \frac{NS^2}{\sigma^2} < \chi_2^2\right) = 1-\alpha \tag{18}$$

on peut dire alors que s^2 appartiendra à l'intervalle $\left[\frac{\chi_1^2\sigma^2}{N}, \frac{\chi_2^2\sigma^2}{N}\right]$ dans $1-\alpha$ des cas et qu'il n'y appartiendra pas dans α des cas.

Ou encore que σ appartiendra à l'intervalle $\left[\frac{\sqrt{N}\,s}{\chi_2}, \frac{\sqrt{N}\,s}{\chi_1}\right]$ dans $1-\alpha$ des cas et qu'il n'y appartiendra pas dans α des cas.

Et donc on obtient un intervalle de confiance à $1-\alpha$ de σ. Mais cette fois l'estimation σ^* n'est plus s, mais $\sigma^* = \frac{\sqrt{N}}{\sqrt{N-1}}s$.

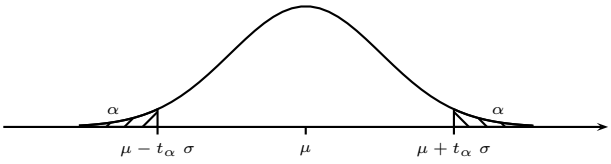
Intervalle de confiance de la moyenne

Plutôt que (6), on forme la variable aléatoire $T = \dfrac{\dfrac{\overline{X}-\mu}{\sigma/\sqrt{N}}}{\sqrt{\dfrac{1}{N-1}\displaystyle\sum_{n=1}^{N}\frac{(X_n-\overline{X})^2}{\sigma^2}}} = \dfrac{\overline{X}-\mu}{S/\sqrt{N-1}}$

qui suit une loi de Student-Fisher à $N-1$ degrés de liberté. Une réalisation de T est $t = \dfrac{\overline{x}-\mu}{s/\sqrt{N-1}} = \dfrac{\overline{x}-\mu}{\sigma^*/\sqrt{N}}$. Si on donne deux nombres, le risque $\alpha \in [0,1]$ et t_α liés par

$$\int_{-t_\alpha}^{t_\alpha} \rho_{SF}(t)\,dt = 1-\alpha$$

où ρ_{SF} est la densité de probabilité de Student-Fisher à $N-1$ degrés de liberté. Cette loi ressemble beaucoup à la loi de Gauss-Laplace



On a

$$\mathbb{P}\left(|T| < t_\alpha\right) = 1 - \alpha \tag{19}$$

on peut dire alors que $\overline{x}$ appartiendra à l'intervalle $\left[\mu - t_\alpha \frac{\sigma^*}{\sqrt{N}}, \mu + t_\alpha \frac{\sigma^*}{\sqrt{N}}\right]$ dans $1-\alpha$ des cas et qu'il n'y appartiendra pas dans α des cas.

Ou encore que μ appartiendra à l'intervalle $\left[\overline{x} - t_\alpha \frac{\sigma^*}{\sqrt{N}}, \overline{x} + t_\alpha \frac{\sigma^*}{\sqrt{N}}\right]$ dans $1-\alpha$ des cas et qu'il n'y appartiendra pas dans α des cas.

Et donc on obtient un intervalle de confiance à $1-\alpha$ de μ dont l'estimation est $\mu^* = \overline{x}$.

Exercice type : On veut caractériser la longueur des pattes avant droites de mygales; on suppose que ces longueurs sont distribuées suivant une loi de Gauss-Laplace. On mesure la longueur de la patte avant droite de 20 mygales et on obtient les valeurs (en centimètres)

2.3 5.3 7.0 7.5 7.8 8.1 8.7 10.7 10.7 10.9 11.4 11.6 11.7 11.9 12.5 13.1 13.5 13.7 13.7 14.3

Comment procéder ?

Solution La moyenne empirique est 10.32 et l'écart-type empirique 3.09. L'estimation de la moyenne μ de toutes les mygales sera donc $\mu^* = 10.31$ cm et l'estimation de l'écart-type σ sera $\sigma^* = 3.17$ cm (arrondi à la première décimale).

Et les intervalles de confiance à 5% (réparti bilatéralement) seront

$$3.09\sqrt{\frac{20}{32.9}} \leq \sigma \leq 3.09\sqrt{\frac{20}{8.91}}$$

soit

$$2.4 \leq \sigma \leq 4.6$$

et

$$10.31 - 2.093\frac{3.17}{\sqrt{20}} \leq \mu \leq 10.31 + 2.093\frac{3.17}{\sqrt{20}}$$

soit

$$8.8 \leq \mu \leq 11.8$$

ce qui signifie que c'est dans 5% des cas que les 20 mygales utilisées pour l'estimation formeraient un lot non typique des mygales en général et donc que μ et σ sortiraient de ces intervalles.

2.4 Tables de Gauss-Laplace, du χ^2, de Student-fisher

Gauss	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.437	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.475	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.483	0.4834	0.4838	0.4842	0.4846	0.485	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.489
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.492	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.494	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.496	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.497	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.498	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.499	0.499

Lecture : Si X est une variable aléatoire de Gauss-Laplace de moyenne μ et d'écart-type σ alors

$$\mathbb{P}\left(0 \leq \frac{X-\mu}{\sigma} \leq \underbrace{1.6}_{\text{1ier colonne}} + \underbrace{0.05}_{\text{1ier ligne}} = 1.65\right) = \underbrace{0.4505}_{\text{intersection ligne / colonne}}$$

et ça se pratique dans les deux sens.

FIGURE 2 – Table de Gauss

χ^2	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005
1	0.0	0.0	0.0	0.0	0.02	0.1	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.1	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.3	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.2	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.7	3.33	4.17	5.9	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.6	3.05	3.82	4.57	5.58	7.58	10.34	13.7	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.4	5.23	6.3	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.3
13	3.57	4.11	5.01	5.89	7.04	9.3	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.6	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.0	27.49	30.58	32.8
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.3	28.85	32.0	34.27
17	5.7	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.6	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.2	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.0
21	8.03	8.9	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.4
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.8
23	9.26	10.2	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.4	13.85	15.66	19.04	23.34	28.24	33.2	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.2	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.6	24.48	29.34	34.8	40.26	43.77	46.98	50.89	53.67

Lecture : Si S^2 est une variable aléatoire suivant une loi du χ^2 à 11 degrés de libertés alors

$$\mathbb{P}(0 \leq S^2 \leq 7.58) = 1 - 0.75 = 0.25$$

$$\mathbb{P}(7.58 \geq S^2) = 0.75$$

Si S^2 a un nombre N de degrés de liberté qui dépasse 30 degrés, on utilise une approximation comme celle qui consiste à dire que

$$\sqrt{2\ S^2} - \sqrt{2\ N - 1}$$

suit une loi normale centrée réduite.

FIGURE 3 – Table du χ^2

S.-F.	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0005
1	0.32492	1.0	3.077684	6.313752	12.7062	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.5407	5.84091	12.924
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.94318	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.306	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.36343	1.795885	2.20099	2.71808	3.10581	4.437
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.681	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.34503	1.76131	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.75305	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.015
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.25658	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.31946	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.68485	1.317836	1.710882	2.0639	2.49216	2.79694	3.7454
25	0.25606	0.68443	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75	3.646
∞	0.253347	0.67449	1.281552	1.644854	1.95996	2.32635	2.57583	3.29

Lecture : Si X est une variable aléatoire suivant une loi de Student-Fisher à 11 degrés de libertés alors

$$\mathbb{P}(-2.20099 \leq X \leq 2.20099) = 1 - 2 \times 0.025 = 0.95$$

$$\mathbb{P}(2.20099 \leq X) = 0.025$$

La ligne d'un nombre infini de degrés de liberté correspond la table de Gauss-Laplace.

FIGURE 4 – Table de Student-Fisher

3 Tests statistiques sur échantillons gaussiens

On dispose d'une population présentant un caractère quantitatif réparti suivant une distribution Gaussienne dont les paramètres moyenne μ et écart-type σ sont a priori inconnus. On prélève dans cette population un échantillon de taille N dont on mesure le caractère pour obtenir

$$\overline{\overline{\begin{array}{cccccc} x_1 & \dots & x_n & \dots & x_N \end{array}}} \quad \text{d'où} \quad \overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad s^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})^2 \tag{20}$$

Cette opération est appelé « expérience » dans la suite.

3.1 Test de valeurs pour la moyenne, écart-type connu *a priori*

On suppose que l'écart-type n'est pas inconnu mais au contraire connu.

Hypothèse simple

On souhaite tester l'hypothèse

$$H : \mu = a$$

où a est une valeur donnée.

Si cette hypothèse est vraie alors un risque α est associé à la valeur t_α telle que

$$\int_{-t_\alpha}^{t_\alpha} \exp^{-t^2/2} / \sqrt{2\pi}\, dt = 1 - \alpha$$

et, avant de faire toute expérience, on sait que si on donne un intervalle

$$I = \left[a - t_\alpha \sigma/\sqrt{N}, a + t_\alpha \sigma/\sqrt{N}\right]$$

alors :

$\rightarrow$ dans $1-\alpha$ des cas la valeur $\overline{x}$ de la moyenne des valeurs trouvées sera dans cet intervalle ;

$\rightarrow$ dans α des cas cette valeur $\overline{x}$ ne sera pas dans l'intervalle.

On fait l'expérience, deux situations peuvent alors se produire :

* $\overline{x} \notin I$; si on tire de ce fait que l'idée que l'hypothèse est fausse, on aura tort de le faire dans le α des cas où elle serait vraie bien que $\overline{x}$ soit en dehors de l'intervalle.
  Dans ce cas on conclut que l'hypothèse est fausse au risque α de se tromper (pour dire qu'il y a α des cas dans lesquels on se trompe effectivement).

* $\overline{x} \in I$, on a alors fait l'expérience pour rien ; il n'est pas possible de conclure quoi que ce soit et surtout pas que *l'hypothèse est vraie sauf dans α des cas.*
  Dans ce cas on conclut qu'en utilisant le risque α, on n'a pas pu mettre en évidence que l'hypothèse était fausse et donc on l'accepte par défaut.

Hypothèse composite

On souhaite tester l'hypothèse

$$H_0 : \mu < a$$

où a est une valeur donnée [8]

8. Rappelons qu'on appelle *hypothèse simple* une hypothèse qui définit complétement la densité de probabilité de la variable aléatoire sous-jacente aux analyses statistiques ; comme $\mu = a$ quand σ est supposée connue. On appelle *hypothèse composite* une hypothèse qui n'est pas simple ; comme $\mu < a$.

Si cette hypothèse est vraie alors un risque α associé à la valeur t_α tel que

$$\int_{-\infty}^{t_\alpha} \exp^{-t^2/2} / \sqrt{2\pi}\, dt = 1 - \alpha$$

est donné et, bis repetita placent, avant de faire toute expérience, on sait que si on donne un intervalle

$$I =]-\infty, a + t_\alpha \sigma/\sqrt{N}]$$

alors :

$\rightarrow$ dans plus de $1-\alpha$ des cas la valeur $\overline{x}$ de la moyenne des valeurs trouvées sera dans cet intervalle ;

$\rightarrow$ dans moins de α des cas cette valeur $\overline{x}$ ne sera pas dans l'intervalle.

On fait l'expérience, deux situations peuvent alors se produire :

* $\overline{x} \notin I$; si on tire l'idée que l'hypothèse est fausse, on aura tort de le faire dans le α des cas où elle serait vraie bien que $\overline{x}$ soit en dehors de l'intervalle.
  Dans ce cas on conclut que l'hypothèse est fausse au risque α de se tromper.

* $\overline{x} \in I$, on a alors fait l'expérience pour rien ; il n'est pas possible de conclure quoi que ce soit et surtout pas que *l'hypothèse est vraie sauf dans α des cas.*
  Dans ce cas on conclut qu'en utilisant le risque α, on n'a pas pu mettre en évidence que l'hypothèse était fausse et donc on l'accepte par défaut.

Une remarque

Avec une l'hypothèse composite H_0, on dispose également de sa négation

$$\overline{H}_0 = \{\mu > a\}$$

en comptant pour nul le cas $\mu = a$ exactement.

Pour tester H_0, on examine si $\overline{x}$ est plus ou moins grand que $a + t_\alpha\, \sigma/\sqrt{N}$; symétriquement et de façon indépendante, pour tester $\overline{H}_0$ on examine la position de $\overline{x}$ par rapport à $a - t_\alpha\, \sigma/\sqrt{N}$. Trois cas peuvent alors se produire

$\rightarrow$ $\overline{x} < a + t_\alpha\, \sigma/\sqrt{N}$: on conclut que H_0 est acceptée par défaut ; que $\overline{H}_0$ est refusée au risque α de se tromper ;

$\rightarrow$ $\overline{x} > a - t_\alpha\, \sigma/\sqrt{N}$: on conclut que $\overline{H}_0$ est acceptée par défaut ; que H_0 est refusée au risque α de se tromper ;

$\rightarrow$ $a - t_\alpha\, \sigma/\sqrt{N} < \overline{x} < a + t_\alpha\, \sigma/\sqrt{N}$: on conclut que H_0 et $\overline{H}_0$ sont toutes les deux acceptées par défaut.

La plage commune d'acceptation par défaut de l'hypothèse et de son contraire illustre que pour inférer quelque chose à partir d'observations il faut renoncer à la certitude absolue.

Exercice type : Il y a un certain dimorphisme sexuel chez les humains :

- la taille des mâles est distribuée suivant une loi normale de moyenne 1.70 m et d'écart type est de 0.05 m

- Les femelles sont plus petites.

On trouve un humain qui mesure a) 1.90 m b) 1.50 m c) 1.75 m; dans chaque cas est-ce plutôt un homme ou une femme ?

Solution Le risque 5 % réparti bilatéralement correspond à la valeur 1.96 donc 95 % des mâles mesurent entre $1.70 \pm 1.96 \times 0.05 = 1.60$ et 1.80 mètres.

Si l'individu mesure 1.90 m c'est seulement dans 5 % des cas que ce serait un mâle; on rejette donc l'hypothèse selon laquelle c'est un mâle avec 5 % chances de se tromper. Et on se trompe puisque les femelles sont plus petites et qu'il n'y a que des mâles et des femelles dans l'énoncé.

En fait il faut faire un test unilatéral : soit donc répartir le risque uniquement du côté des petites tailles, en gardant 5 % cela donne une borne inférieure de $1.70 - 1.65 \times 0.05 = 1.62$ et donc c'est seulement dans 5 % des cas qu'un mâle est plus petit que 1.62 m.

Pour le cas de 1.90 m on décide alors qu'il n'y a pas de raisons d'infirmer l'hypothèse selon laquelle cet individu serait un mâle.

Mais ça ne veut pas du tout dire qu'on a 5 % de chances de se tromper. D'ailleurs si les femelles mesurent 1.60 m avec un écart type de 0.05 alors 1.90 correspond à un risque quasiment nul : il y aurait donc beaucoup moins de 5 % de chances de ce tromper.

Si l'individu mesure 1.50 m en reprenant ce qui vient d'être fait, notamment le test unilatéral, on voit que 1.50 m est inférieur à 1.62 m et donc que c'est seulement dans 5 % des cas que cela arrive.

On peut alors affirmer qu'au risque 5 % de se tromper l'individu n'est pas un mâle, et donc puisqu'il n'y a que des mâles et des femelles chez les humains que c'est une femelle.

Il est possible d'être plus fin, 1.50 correspond à un risque de 0.004 % pour les mâles et on peut prendre ce risque quasiment nul pour affirmer que c'est une femelle.

Cela montre d'ailleurs les limites de ce type d'analyse dès qu'on s'écarte un peu des valeurs moyennes; en effet il est certain qu'il existe des mâles de 1m50 en proportion bien supérieure à ce qu'affirme le test statistique qui, rappelons le, suppose au départ que la population est gaussienne; ce n'est certainement pas vrai pour les valeurs très éloignées de la moyenne; il faudrait un modèle plus fin dont l'étude nécessite d'abord de maîtriser le paradigme gaussien (un peu pompeux mais efficace pour faire croire qu'on maîtrise un concept plutôt que des techniques logico-calculatoires) auquel on se limite ici.

Si l'individu mesure 1.65 En utilisant le risque 5 % on décide de ne pas infirmer l'hypothèse selon laquelle cet individu est un mâle.

Si on avait utilisé le risque 10% la borne de 1.62 serait devenue $1.7 - 1.28 \times 0.05 = 1.64$; on déciderait alors, avec le risque 10% de se tromper que cet individu est une femelle.

Plus on accepte de risque, plus il est possible de conclure de façon catégorique : 'dans 10% des cas j'ai tort mais c'est une femelle' contre 'j'ai utilisé un risque de 5 % (lâche !) et on conclut qu'on ne peut pas infirmer l'hypothèse selon laquelle l'individu est un mâle alors on l'accepte, par défaut.

3.2 Test de valeurs pour l'écart-type

Cette fois-ci on ne suppose plus l'écart-type connu mais on cherche à tester sur lui une hypothèse simple comme $\sigma = a$ ou composite comme $\sigma < a$. Rien ou presque ne change du raisonnement précédent.

Pour l'hypothèse simple

$$\sigma = a$$

on considère la variable aléatoire du χ^2 à $N-1$ degrés de liberté de densité ρ_{χ^2} dont $\chi^2 = Ns^2/a^2$ doit être une réalisation si l'hypothèse est vraie. On donne un risque α, on cherche les deux nombre χ_1^2 et χ_2^2 tels que

$$\int_0^{\chi_1^2} \rho_{\chi^2}(x)\, dx = \alpha/2 \text{ et } \int_{\chi_2^2}^{\infty} \rho_{\chi^2}(x)\, dx = \alpha/2$$

on introduit

$$I = [\chi_1^2, \chi_2^2]$$

et alors :

* si $Ns^2/a^2 \notin I$ on conclut que l'hypothèse $\sigma = a$ est fausse mais on affirmant cela on se trompe dans α des cas ;

* si $Ns^2/a^2 \in I$ on conclut qu'en utilisant le risque α on n'a pas pu mettre en évidence que l'hypothèse était fausse et donc qu'on l'accepte par défaut.

Pour $\sigma < a$ on considère la même variable aléatoire du χ^2 à $N-1$ degrés de liberté, on cherche le nombre χ_1^2 tel que $\int_{\chi_1^2}^{\infty} \rho_{\chi^2}(x)\, dx = \alpha$ et

* si $Ns^2/a^2 < \chi_1^2$ on accepte l'hypothèse par défaut ;

* si $Ns^2/a^2 > \chi_1^2$ on refuse l'hypothèse au risque α de se tromper.

Pour l'hypothèse composite

$$\sigma > a$$

on considère la même variable aléatoire du χ^2 à $N-1$ degrés de liberté, on cherche le nombre χ_1^2 tel que $\int_0^{\chi_1^2} \rho_{\chi^2}(x)\, dx = \alpha$ et

* si $Ns^2/a^2 > \chi_1^2$ on accepte l'hypothèse par défaut ;

* si $Ns^2/a^2 < \chi_1^2$ on refuse l'hypothèse au risque α de se tromper.

Remarque : pour ne pas se tromper de sens dans les cas d'inégalité il suffit de faire tendre a vers 0 ou l'∞ et de vérifier que $\sigma > 0$ ou $\sigma < \infty$ (qui sont quand même intrinséquement vraies) se trouve bien dans la zone d'acceptation de l'hypothèse.

3.3 Test de valeurs pour la moyenne, écart-type inconnu

Ce qui a été fait pour le cas de l'écart-type connu est à répéter en changeant σ en son estimation

$$\sigma^* = \sqrt{\frac{N}{N-1}}\, s$$

et « loi de Gauss-Laplace » par « loi de Student-Fisher à N-1 degrés de liberté.»

4 Comparaisons de populations

On connaît deux populations différentes qui possèdent en commun un caractère et on voudrait savoir si ces populations peuvent être considérées comme identiques vis-à-vis de ce caractère.

4.1 Deux populations gaussiennes

Le caractère des deux populations est quantitatif et réparti suivant de lois de Gauss-Laplace *a priori* différentes ; les moyennes et écart-types sont (μ_x, σ_x) et (μ_y, σ_y).

On dispose des échantillons $x_1, \ldots, x_N$ et $y_1, \ldots, y_P$ issus des deux populations.

Le paramètres (μ_x, σ_x) et (μ_y, σ_y) sont des quantités certaines mais inconnues qui sont estimées à partir des échantillons et donc on a obtenu les estimations μ_x^*, σ_x^*, μ_y^* et σ_y^* ; il s'agit maintenant de les comparer.

Pour cela on procède de la façon suivante :

1. on effectue un test pour décider s'il est possible d'affirmer que $\sigma_x = \sigma_y$

2. si le résultat du test est négatif alors les deux populations sont différentes, le travail est fini ; 2-bis) dans le cas contraire on effectue un test pour décider s'il est possible d'affirmer que $\mu_x = \mu_y$.

Comparaison des écart-types

En introduisant les variables aléatoires dont les caractères des individus des échantillons sont les réalisations (X_n pour x_n et Y_n pour y_n) on calcule

$$K^2 = \frac{1}{\sigma_x^2} \sum_{n=1}^{N} (X_n - \overline{X})^2 \quad \text{et} \quad L^2 = \frac{1}{\sigma_y^2} \sum_{p=1}^{P} (Y_n - \overline{Y})^2 \tag{21}$$

avec

$$\overline{X} = \frac{1}{N} \sum_{n=1}^{N} X_n \quad \text{et} \quad \overline{Y} = \frac{1}{P} \sum_{p=1}^{P} Y_p \tag{22}$$

K^2 et L^2 suivent des lois du χ^2 à $N-1$ et $P-1$ degrés de liberté dont les réalisations sont

$$(N-1)\ \frac{\sigma_x^{*2}}{\sigma_x^2} \quad \text{et} \quad (P-1)\ \frac{\sigma_y^{*2}}{\sigma_y^2}$$

Donc

$$F = \frac{K^2/(N-1)}{L^2/(P-1)} \tag{23}$$

suit une loi de Snedecor à $N-1$ et $P-1$ degrés de liberté dont

$$\frac{\sigma_y^2}{\sigma_x^2}\ \frac{\sigma_x^{*2}}{\sigma_y^{*2}}$$

est une réalisation.

Les choses sont en place pour le test de l'hypothèse

$$H : \sigma_x = \sigma_y$$

Déjà on donne un risque α ; on détermine F_1 et F_2 tels que

$$\int_0^{F_1} \rho_S(u)\, du = \int_{F_2}^{\infty} \rho_S(u)\, du = \alpha/2$$

où ρ_S est la densité de Snedecor à $N-1$ et $P-1$ degrés de liberté.

Si $\sigma_x = \sigma_y$ alors ce n'est que dans α des cas que $\sigma_x^{*2}/\sigma_y^{*2}$ ne sera pas dans l'intervalle

$$[F_1, F_2]$$

et donc

$\rightarrow$ si $\sigma_x^{*2}/\sigma_y^{*2} \notin [F_1, F_2]$ on conclut que l'hypothèse $\sigma_x = \sigma_y$ est fausse au risque α de se tromper ;

$\rightarrow$ si $\sigma_x^{*2}/\sigma_y^{*2} \in [F_1, F_2]$ on accepte l'hypothèse par défaut.

Comparaison des moyennes

On suppose que le test de comparaison des écart-types a permis d'accepter (par défaut) l'hypothèse selon laquelle les écart-types sont égaux. Il donc faut maintenant comparer les moyennes mais auparavant on doit fabriquer une nouvelle estimation de l'écart-type commune aux deux échantillons.

Pour cela on exploite que si $\sigma_x = \sigma_y = \sigma$ alors $K^2 + L^2$ suit une loi du χ^2 à $N+P-2$ degrés de liberté dont une réalisation est $((N-1)\sigma_x^{*2} + (P-1)\sigma_y^{*2})/\sigma^2$ et par conséquent la nouvelle estimation σ^* de l'écart-type commun aux deux échantillons sera

$$\sigma^{*2} = \frac{(N-1)\sigma_x^{*2} + (P-1)\sigma_y^{*2}}{N+P-2} \tag{24}$$

D'autre part $\overline{X} - \overline{Y}$ suit une loi de Gauss-Laplace de

$$\text{moyenne : } \mu_x - \mu_y \text{ et d'écart-type : } \sigma\sqrt{\frac{1}{N} + \frac{1}{P}}$$

d'où on retire que si l'hypothèse $\mu_x = \mu_y$ est vraie alors

$$\frac{\dfrac{\overline{X} - \overline{Y}}{\sigma\sqrt{\frac{1}{N} + \frac{1}{P}}}}{\sqrt{\dfrac{K^2 + L^2}{\sigma^2\,(N+P-2)}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{K^2 + L^2}}\sqrt{\frac{N+P-2}{\frac{1}{N} + \frac{1}{P}}} \tag{25}$$

suit une loi de Student-Fisher à $N+P-2$ degrés de liberté dont la réalisation est

$$r_{sf} = \frac{\mu_x^* - \mu_y^*}{\sigma^*\sqrt{\frac{1}{N} + \frac{1}{P}}}$$

.

On donne un risque α ; on calcule t_α tel que $\int_{-t_\alpha}^{t_\alpha} \rho_{sf}(u)\, du = 1 - \alpha$ (ρ_{sf} la densité de Student-Fisher à $N+P-2$ degrés de liberté) et alors

$\rightarrow$ si $|r| > t_\alpha$ on peut dire que l'hypothèse $\mu_x = \mu_y$ est infirmée au risque α de se tromper ;

$\rightarrow$ si $|r_{sf}| < t_\alpha$ on accepte l'hypothèse $\mu_x = \mu_y$ par défaut.

Ce qui permet de conclure sur la question globale de savoir si le caractère est réparti de façon identique sur les deux populations considérées.

Exercice type : Le chef d'un service d'assistance téléphonique veut contrôler le travail de deux collaborateurs de l'entreprise. Ce travail consiste à répondre aux questions des clients; et le chef considère que le nombre d'appel téléphonique auquel un collaborateur a répondu dans la journée est une bonne mesure du travail qu'il a effectué.

Le chef dispose pour chacun des deux collaborateurs de données; pour chacun, il a noté : le nombre de jours N qui ont servi à la mesure; la moyenne estimée μ^*; l'écart-type estimé σ^* :

$$\begin{array}{llll} N_x & \mu_x^* & \sigma_x^* & \text{pour Xavier} \\ N_y & \mu_y^* & \sigma_y^* & \text{pour Yvette} \end{array}$$

Il n'a pas oublié de vérifier qu'on ne pouvait pas refuser que la loi de cette quantité journalière de travail soit décrite par une loi de Gauss-Laplace.

La question du chef est : puis-je mettre en évidence que ces deux collaborateurs ont une capacité de travail différente ?

Solution

Premier cas

	N	μ^*	σ^*
Yvette	25	103	22
Xavier	13	94	38

D'abord le chef teste la possibilité que Xavier et Yolande appartiennent à une même population gaussienne dont l'écart-type soit la quantité certaine mais inconnue σ.

Si cela est alors $(22/38)^2$=0.33 est une réalisation d'une loi de Snedecor à 25−1=24 et 13−1=12 degrés de liberté.

Si le chef choisit le risque 5% qu'il coupe en deux parties de 2.5% alors F_2 est défini par (ρ la densité de Snedecor)

$$\int_{F_2}^{\infty} \rho(x)dx=0.025$$

il lit directement dans la table à 2.5% $F2$=3.02 (à l'intersection des 24 (au numérateur et 12 au dénominateur degrés de liberté). F_1 est défini par

$$\int_{0}^{F_1} \rho(x)dx=0.025$$

mais il n'y a pas de table permettant d'obtenir directement cette valeur.

Par contre le chef sait que

$$\mathcal{P}(F \leq F_1)=\mathcal{P}(1/F \geq 1/F_1)=\int_{1/F_1}^{\infty} \rho(x)dx$$

et donc il peut lire la valeur $1/F_1$; toujours sur la table à 2.5% mais cette fois ci. à l'intersection des 12 (au numérateur) et 24 (au dénominateur) degrés de liberté F_1=1/2.54=0.39.

Donc le chef pourra conclure qu'au risque 5% de se tromper Xavier et Yvette ne font pas partie d'une même population gaussienne.

Second cas

	N	μ^*	σ^*
Yvette	25	103	22
Xavier	13	94	31

Comme on trouve que $(22/31)^2$=0.5 est situé entre F_1=0.39 et F_2=3.02, le chef ne pourra pas conclure à la différence entre les écart-types.

Le chef calcule donc une meilleure estimation de l'écart-type comme

$$\sigma^*=\sqrt{\frac{(25-1)(22)^2+(13-1)(31)^2}{25+31-2}}=25.36$$

Et il calcule

$$\frac{103-94}{25.36\sqrt{\frac{1}{25}+\frac{1}{13}}}=1.03$$

Avec le risque 5% réparti bilatéralement, si Xavier et Yvette appartiennent à la même population alors ce nombre doit être en valeur absolu inférieur au nombre donnée par la table de Student-Fisher pour 36 degrés de liberté, cela dégénère en Gauss et donc le chef connaît parfaitement le nombre : c'est 1.96.

Comme $-1.96 < 1.03 < 1.96$, le chef, malgré toute la rancoeur que cela lui apporte, ne peut conclure à une différence entre le travail de Xavier et Yvette.

Alors le chef décide de prendre un risque de 30% de se tromper et là il peut conclure et tirer de sa conclusion une action néfaste à l'un des collaborateurs qui, s'il ne connaît pas suffisamment de statistique ne saura pas contrer les arguments du chef.

4.2 Comparaison de proportions

Les deux populations ont maintenant un caractère qualitatif à deux valeurs dont la proportion d'une des valeurs est $\overline{\omega}_1$ pour l'une des populations et $\overline{\omega}_2$ pour l'autre.

Les variables aléatoires X et Y représentent le nombre de fois où on obtient cette valeur par un tirage d'un échantillon de taille N et P dans les deux populations.

Ces variables aléatoires (binomiales) ont un couple moyenne/écart-type $(N\ \overline{\omega}_1, \sqrt{N\ \overline{\omega}_1(1-\overline{\omega}_1)})$ et
$(P\ \overline{\omega}_2, \sqrt{P\ \overline{\omega}_2(1-\overline{\omega}_2)})$ qui ne dépend que d'un seul paramètre la proportion $\overline{\omega}_1$ et $\overline{\omega}_2$.

Si $\overline{\omega}_1$ et $\overline{\omega}_2$ ne sont ni très petits ni très grand on peut accepter l'approximation selon laquelle X et Y suivent des lois de Gauss-Laplace.

Égalité des proportions

Si $\overline{\omega}_1 = \overline{\omega}_2 = \overline{\omega}$ alors $X/N - Y/P$ est une variable aléatoire de Gauss-Laplace de moyenne nulle et d'écart-type

$$\sqrt{\overline{\omega}(1-\overline{\omega})\ (\frac{1}{N}+\frac{1}{P})} \tag{26}$$

On choisit deux échantillons de taille N et P issus des deux populations ; on obtient des estimations de proportion $\overline{\omega}_1^*$ et $\overline{\omega}_2^*$.

De plus (avec une certaine légèreté) on exploite le fait que $\sqrt{\overline{\omega}(1-\overline{\omega})}$ varie assez peu on peut approximer $\overline{\omega}$ par son estimation

$$\overline{\omega}^* = \frac{N\overline{\omega}_1^* + P\overline{\omega}_2^*}{N+P} \tag{27}$$

ce qui n'est possible, répétons-le, que si $\overline{\omega}_1$ et $\overline{\omega}_2$ ne sont ni grand ni petits.

Le test revient à choisir un risque α, calculer t_α tel que $\int_{-t_\alpha}^{t_\alpha} \rho_g(u)\, du = 1-\alpha$ (ρ_g la densité de la loi de Gauss-Laplace centrée réduite) et alors

→ si $|\overline{\omega}_1^* - \overline{\omega}_2^*| > t_\alpha \sqrt{\overline{\omega}^*(1-\overline{\omega}^*)\,(\frac{1}{N}+\frac{1}{P})}$ on peut dire que l'hypothèse selon laquelle les proportions de la valeur du caractère considéré dans les deux populations sont égales est infirmée au risque α de se tromper

→ sinon cette hypothèse est acceptée par défaut.

Inégalité des proportions

On peut également vouloir affirmer que $\overline{\omega}_1 > \overline{\omega}_2$; pour cela, au prix d'une contorsion logique, on utilise quand même l'approximation d'une proportion commune $\overline{\omega}^*$ et on fait fait un test monolatéral.

On calcule t_α tel que $\int_{t_\alpha}^{\infty} \rho_g(u)\, du = \alpha$ et alors

→ si $\overline{\omega}_1^* - \overline{\omega}_2^* < t_\alpha \sqrt{\overline{\omega}^*(1-\overline{\omega}^*)\,(\frac{1}{N}+\frac{1}{P})}$ on peut dire que l'hypothèse $\overline{\omega}_1 > \overline{\omega}_2$ est infirmée au risque α de se tromper ;

→ sinon cette hypothèse est acceptée par défaut.

4.3 Test des appariements

La comparaison de deux populations portait sur les valeurs moyennes obtenues par une analyse pratiquée sur chaque population indépendamment l'une de l'autre.

Les test des appariements suppose que les deux populations sont issues d'une seule population ; que la différence entre elles porte sur un facteur qui a une valeur différente v_0 et v_1 dans l'un et l'autre cas ; et que les valeurs du caractère testé sur un individu soumis à l'une et l'autre valeurs du facteur sont appariées. Soit

facteur	individu No 1		individu No N
v_0	x_1	$\dots$	x_N
v_1	y_1		y_N

Cela permet de considérer la série

$x_1 - y_1$	$\dots$	$x_N - y_N$

comme une suite de réalisation de variables aléatoires de Gauss-Laplace de moyenne μ et écart-type σ inconnus.

Et la question 'le facteur a-t-il une influence ?' se formalise par le test de l'hypothèse

$$H : \mu = 0$$

De même la question 'le facteur augmente-t-il la valeur du caractère testé ?' par le test de l'hypothèse

$$H : \mu > 0$$

choses déjà vues.

Exemple illustratif : Par exemple 'porter ou non un nouveau maillot de bain en peau de requin' sur le temps mis pour parcourir 400 mètres en brasse.

On fait alors effectuer par N nageurs ce trajet deux fois, avec le nouveau maillot de bain et avec l'ancien, et on note les temps mis par les nageurs sous la forme de N couples $(x_1, y_1), \dots, (x_N, y_N)$.

On suppose que la série de nombres $x_1 - y_1, \dots x_N - y_N$ est la réalisation d'une variable aléatoire de Gauss-Laplace d'écart-type σ inconnu et de moyenne μ également inconnue.

On teste alors l'hypothèse $\mu < 0$: si elle est rejetée c'est que le nouveau maillot de bain ne diminue pas le temps mis pour parcourir 400 mètres en brasse ; si elle est acceptée on adopte la conclusion inverse.

On peut tester aussi l'hypothèse $\mu > 0$: si elle est rejetée c'est que le nouveau maillot de bain n'augmente pas le temps mis pour parcourir 400 mètres en brasse ; si elle est acceptée on adopte la conclusion inverse.

4.4 Tables de Snedecor

Ces tables de Snedecor mettent en correspondance le nombre F_α tel que le rapport

$$\frac{\chi^2/ddl}{\chi'^2/ddl'}$$

ait une probabilité α de dépasser ce nombre F_α avec les degrés ddl (lus dans la première rangée) et ddl' (lus dans la première colonne).

10%	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.86	49.5	53.59	55.83	57.24	58.2	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.0	62.26	62.53	62.79	63.06	63.33
2	8.53	9.0	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.2	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.9	3.87	3.84	3.83	3.82	3.8	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.4	3.37	3.34	3.32	3.3	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.1
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.9	2.87	2.84	2.82	2.8	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.7	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.5	2.46	2.42	2.4	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.3	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.2	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.3	2.27	2.25	2.21	2.17	2.12	2.1	2.08	2.05	2.03	2.0	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.1	2.06	2.04	2.01	1.99	1.96	1.93	1.9
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.2	2.16	2.14	2.1	2.05	2.01	1.98	1.96	1.93	1.9	1.88	1.85
14	3.1	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.1	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.8
15	3.07	2.7	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.9	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.1	2.06	2.03	2.0	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.2	2.13	2.08	2.04	2.0	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.4	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.7	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.0	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.9	1.86	1.81	1.76	1.73	1.7	1.67	1.64	1.6	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.8	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.1	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.7	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.5
27	2.9	2.51	2.3	2.17	2.07	2.0	1.95	1.91	1.87	1.85	1.8	1.75	1.7	1.67	1.64	1.6	1.57	1.53	1.49
28	2.89	2.5	2.29	2.16	2.06	2.0	1.94	1.9	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.5	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.5	1.46
40	2.84	2.44	2.23	2.09	2.0	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.6	1.54	1.51	1.48	1.44	1.4	1.35	1.29
120	2.75	2.35	2.13	1.99	1.9	1.82	1.77	1.72	1.68	1.65	1.6	1.54	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.3	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.6	1.55	1.49	1.42	1.38	1.34	1.3	1.24	1.17	1.0

FIGURE 5 – Table de Snedecor pour le risque 10 %

5%	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.51	19.0	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.5
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.7	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.0	5.96	5.91	5.86	5.8	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.5	4.46	4.43	4.4	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	4.0	3.94	3.87	3.84	3.81	3.77	3.74	3.7	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.3	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.9	2.86	2.83	2.79	2.75	2.71
10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.84	2.77	2.74	2.7	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.9	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.4
12	4.75	3.89	3.49	3.26	3.11	3.0	2.91	2.85	2.8	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.3
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.6	2.53	2.46	2.42	2.38	2.34	2.3	2.25	2.21
14	4.6	3.74	3.34	3.11	2.96	2.85	2.76	2.7	2.65	2.6	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59	2.54	2.48	2.4	2.33	2.29	2.25	2.2	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.2	2.96	2.81	2.7	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.1	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.9	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39	2.35	2.28	2.2	2.12	2.08	2.04	1.99	1.95	1.9	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.1	2.05	2.01	1.96	1.92	1.87	1.81
22	4.3	3.44	3.05	2.82	2.66	2.55	2.46	2.4	2.34	2.3	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.8	2.64	2.53	2.44	2.37	2.32	2.27	2.2	2.13	2.05	2.0	1.96	1.91	1.86	1.81	1.76
24	4.26	3.4	3.01	2.78	2.62	2.51	2.42	2.36	2.3	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.6	2.49	2.4	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.9	1.85	1.8	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.2	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.2	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.7	2.55	2.43	2.35	2.28	2.22	2.18	2.1	2.03	1.94	1.9	1.85	1.81	1.75	1.7	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.0	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.0	3.15	2.76	2.53	2.37	2.25	2.17	2.1	2.04	1.99	1.92	1.84	1.75	1.7	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.5	1.43	1.35	1.25
∞	3.84	3.0	2.6	2.37	2.21	2.1	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.0

FIGURE 6 – Table de Snedecor pour le risque 5 %

2.5%	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647	800	864	900	922	937	9482	957	963	968	977	985	993	997	1001	1006	1010	1014	1018
2	38.51	39.0	39.17	39.25	39.3	39.33	39.36	39.37	39.39	39.4	39.41	39.43	39.45	39.46	39.47	39.47	39.48	39.49	39.5
3	17.44	16.04	15.44	15.1	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.9
4	12.22	10.65	9.98	9.6	9.36	9.2	9.07	8.98	8.9	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.6	6.23	5.99	5.82	5.7	5.6	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.9	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.9	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.2	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.3	4.2	4.1	4.0	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.2	4.1	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.5	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.2	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.0	2.94	2.88
12	6.55	5.1	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.0	3.77	3.6	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.6
14	6.3	4.86	4.24	3.89	3.66	3.5	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.2	4.76	4.15	3.8	3.58	3.41	3.29	3.2	3.12	3.06	2.96	2.86	2.76	2.7	2.64	2.58	2.52	2.46	2.4
16	6.12	4.69	4.08	3.73	3.5	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.5	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.1	3.01	2.93	2.87	2.77	2.67	2.56	2.5	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.9	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.2	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.08
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.8	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.7	2.6	2.5	2.39	2.33	2.27	2.21	2.14	2.08	2.0
23	5.75	4.35	3.75	3.41	3.18	3.02	2.9	2.81	2.73	2.67	2.57	2.47	2.36	2.3	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.7	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.3	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.1	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.8	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.0	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.9	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.2	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.2	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.9	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.8	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.8	3.23	2.89	2.67	2.52	2.39	2.3	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.0

FIGURE 7 – Table de Snedecor pour le risque 2.5 %

1%	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999	5403	5624	5763	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6365
2	98.5	99.0	99.17	99.25	99.3	99.33	99.36	99.37	99.39	99.4	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.5
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.05	26.87	26.69	26.6	26.5	26.41	26.32	26.22	26.12
4	21.2	18.0	16.69	15.98	15.52	15.21	14.98	14.8	14.66	14.55	14.37	14.2	14.02	13.93	13.84	13.74	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.2	9.11	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.1	7.98	7.87	7.72	7.56	7.4	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.2	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.8	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.4	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.2	5.06	4.94	4.85	4.71	4.56	4.4	4.33	4.25	4.16	4.08	4.0	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.4	4.25	4.1	4.02	3.94	3.86	3.78	3.69	3.6
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.5	4.39	4.3	4.16	4.01	3.86	3.78	3.7	3.62	3.54	3.45	3.36
13	9.07	6.7	5.74	5.2	4.86	4.62	4.44	4.3	4.19	4.1	3.96	3.82	3.66	3.59	3.51	3.42	3.34	3.26	3.16
14	8.86	6.52	5.56	5.04	4.7	4.46	4.28	4.14	4.03	3.94	3.8	3.66	3.5	3.43	3.35	3.27	3.18	3.09	3.0
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.0	3.9	3.8	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.2	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.1	3.02	2.93	2.84	2.75
17	8.4	6.11	5.18	4.67	4.34	4.1	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.0	2.92	2.84	2.75	2.65
18	8.28	6.01	5.09	4.58	4.25	4.01	3.84	3.7	3.6	3.51	3.37	3.23	3.08	3.0	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.5	4.17	3.94	3.76	3.63	3.52	3.43	3.3	3.15	3.0	2.92	2.84	2.76	2.67	2.58	2.49
20	8.1	5.85	4.94	4.43	4.1	3.87	3.7	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.7	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.4	3.31	3.17	3.03	2.88	2.8	2.72	2.64	2.55	2.46	2.36
22	7.94	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.5	2.4	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.3	3.21	3.07	2.93	2.78	2.7	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.9	3.67	3.5	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.4	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.7	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.82	2.66	2.58	2.5	2.42	2.33	2.23	2.13
27	7.68	5.49	4.6	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.2	2.1
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.9	2.75	2.6	2.52	2.44	2.35	2.26	2.17	2.06
29	7.6	5.42	4.54	4.04	3.72	3.5	3.33	3.2	3.09	3.0	2.87	2.73	2.57	2.5	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.7	3.47	3.3	3.17	3.07	2.98	2.84	2.7	2.55	2.47	2.39	2.3	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.8	2.66	2.52	2.37	2.29	2.2	2.11	2.02	1.92	1.8
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.5	2.35	2.2	2.12	2.03	1.94	1.84	1.73	1.6
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.04	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.64	4.61	3.78	3.32	3.02	2.8	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.7	1.59	1.47	1.32	1.0

FIGURE 8 – Table de Snedecor pour le risque 1 %

5 Régression linéaire

5.1 La taille des pères et les fils

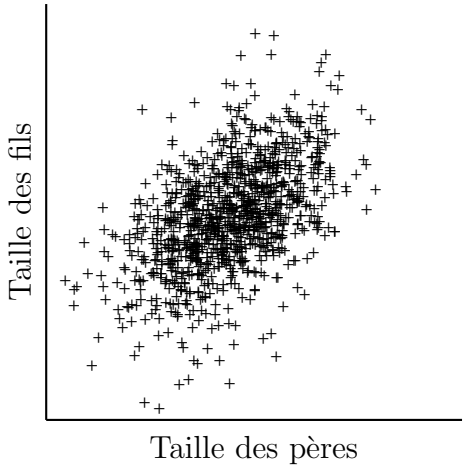
L'expression « régression linéaire » a été inventée par Galton pour expliquer que d'une part la distribution des tailles restait la même d'une génération à l'autre et que d'autre part la taille des parents influait sur celle des enfants. Pour simplifier il considérait la moyenne de la taille du père et de la mère qui fournit un individu unique concevant ses fils en quelque sorte par parthénogenèse.

Le problème est loin d'être trivial. Il y a les variables aléatoires :

– X la taille d'un individu de la génération des pères ;

– Y la taille d'un individu de la génération des fils ;

et si on connait les filiations il est possible de reporter sur un graphe les couples du caractère taille chez les pères et les fils



C'est un « nuage de points » (expression consacrée) dont l'examen fait apparaître que les petits (ou grands) pères ont des fils petits (ou grands) ; mais ces points ne sont pas arrangés le long d'une même courbe qui mettrait en évidence une liaison déterminée entre les deux variables aléatoires

$$Y = f(X) \text{ où } f \text{ est une fonction } ad\ hoc$$

Les valeurs de la variable aléatoire Y sont d'une part expliquées par celles de X et d'autre part ont un caractère aléatoire propre indépendant de X, c'est à dire que la liaison est

$$Y = f(X) + \Sigma$$

où Σ est une variable aléatoire sur laquelle des hypothèses doivent êtres faites. Celles-ci sont que Σ est une variable aléatoire de Gauss-Laplace indépendante de X, centrée

$$\mathbb{E}(\Sigma) = 0 \quad ; \quad \mathbb{V}ar(\Sigma) = \sigma$$

et dont la valeur σ de l'écart-type est à déterminer.

La fonction f pourrait être quelconque mais une hypothèse supplémentaire est faite de la choisir comme une fonction affine, c'est à dire

$$Y = \alpha\, X + \beta + \Sigma$$

où les coefficients α et β sont à déterminer.

Avant de chercher à déterminer ces coefficients σ, α et β, il est possible d'établir un résultat surprenant d'où vient l'origine de ce nom de « régression. » Si on fait l'hypothèse que la distribution de taille des pères est la même que celle des fils alors

$$\mathbb{E}(X) = \mathbb{E}(Y) = \mu \quad ; \quad \mathbb{V}ar(X) = \mathbb{V}ar(Y) = \sigma'^2$$

où μ et σ' sont les paramètres de cette distribution commune aux pères et aux fils. Et donc

$$\mu = \alpha\ \mu + \beta$$

ce qui fournit une relation *a priori* entre α et β. Mais il vient également

$$\sigma'^2 = \alpha^2\ \sigma'^2 + \sigma^2 \Longrightarrow (1 - \alpha^2)\ \sigma'^2 = \sigma^2$$

Il est nécessaire que (on suppose implicitement que $0 < \alpha$)

$$\alpha < 1$$

et donc si la taille x_n d'un père est connue, la variable aléatoire de la taille de ses fils sera

$$Y_n = \alpha\ x_n + \beta + \Sigma = \mu + \alpha\ (x_n - \mu) + \Sigma$$

Elle a une espérance mathématique

$$\mathbb{E}(Y_n) = \mu + \alpha\ (x_n - \mu) = x_n + (1 - \alpha)\ (\mu - x_n)$$

qui est telle que

$$\left\{ \begin{array}{lll} \mathbb{E}(Y_n) < x_n & \text{si} & x_n > \mu \\ \mathbb{E}(Y_n) > x_n & - & x_n < \mu \end{array} \right.$$

La moyenne de la taille des fils d'un père plus grand que la moyenne régresse vers la moyenne de l'espèce (et inversement elle progresse).

Ce résultat ne dit rien des phénomènes qui maintiennent la stationnarité des caractéres d'une espèce mais il montre que si cette stationnarité est constatée alors ces phénomènes ont pour effet d'atténuer la diversité sur les extrêmes lors de la reproduction. Sinon il y aurait incompatibilité entre l'hypothèse de stationnarité et celle d'indépendance de la variabilité du caractère lors de la reproduction d'un individu à caractère donné et de la variabilité des caractères dans l'espèce.

5.2 Formalisation du problème de régression linéaire

Abandonnons le problème historique pour un problème générique formulé de façon un peu différente. On dispose de deux séries de mesures liées (y_n et x_n sont mesurés simultanément)

$$\begin{array}{|cccccc|} \hline x_1 & \dots & x_n & \dots & x_N \\ y_1 & \dots & y_n & \dots & y_N \\ \hline \end{array} \Longrightarrow \left\{ \begin{array}{l} \overline{x} = \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} x_n \quad ; \quad \overline{x^2} = \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} x_n^2 \quad ; \quad \overline{x\ y} = \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} x_n\ y_n \\ \overline{y} = \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} y_n \quad ; \quad \overline{y^2} = \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} y_n^2 \end{array} \right. \tag{28}$$

et on affirme qu'il y a une liaison affine entre x et y (le modèle) soit

$$y = \alpha\ x + \beta \tag{29}$$

l'objectif est de déterminer α et β au sens statistique du terme : i.e. déterminer les intervalles de confiances de ces quantités certaines mais inconnues à partir des données.

Des hypothèses sont nécessaires. d'abord les x_n sont des quantités certaines (et connues) et ensuite les y_n sont des réalisations d'une variable aléatoire $Y(x)$ qui suit la loi de Gauss-Laplace de moyenne et d'écart-type

$$\alpha\ x + \beta \quad \text{et} \quad \sigma$$

donc la moyenne des y_n dépend des x_n mais pas l'écart-type.

Pour fixer les choses on n'a pas

$$Y(x) = \alpha\ x + \beta$$

où serait alors la variable aléatoire dans le second membre ? mais

$$Y(x) = \alpha\ x + \beta + \Sigma$$

où Σ est une variable aléatoire de Gauss-Laplace de moyenne nulle et d'écart-type σ. Cette dernière hypothèse suppose que l'écart entre la prévision du modèle

$$\alpha\ x_n + \beta$$

et la valeur

$$y_n$$

est représentée par une variable aléatoire de Gauss-Laplace centrée et d'écart-type σ ne dépendant pas de x_n : c'est la traduction de ce qu'on entend par « erreur additive. »

5.3 La droite des moindres carrés

Il ne s'agit pas d'expliquer la méthode pratique des moindres carrés [9] qui est supposée connue et qu'on rappelle ici juste pour qu'il n'y ait pas d'ambiguïté sur les intentions.

Si on fabrique la fonction

$$\begin{array}{rccl} f\ : & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (u,v) & \longrightarrow & f(u,v) = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}(y_n - u\ x_n - v)^2 \end{array} \tag{30}$$

qui se réécrit

$$f(u,v) = (u\ v)\begin{pmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} - 2(u\ v)\begin{pmatrix} \overline{x\ y} \\ \overline{y} \end{pmatrix} + \overline{y^2}$$

on voit que la valeur de l'argument (u,v) pour laquelle cette fonction correspond à un graphe qui « passe au mieux » par l'ensemble des couples (x_n, y_n) est (α^*, β^*) celle qui minimise cette fonction et donc solution de la condition nécessaire d'extréma

$$\begin{pmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{pmatrix}\begin{pmatrix} \alpha^* \\ \beta^* \end{pmatrix} = \begin{pmatrix} \overline{x\ y} \\ \overline{y} \end{pmatrix} \implies \begin{pmatrix} \alpha^* \\ \beta^* \end{pmatrix} = \frac{1}{\overline{x^2} - \overline{x}^2}\begin{pmatrix} 1 & -\overline{x} \\ -\overline{x} & \overline{x^2} \end{pmatrix}\begin{pmatrix} \overline{x\ y} \\ \overline{y} \end{pmatrix} \tag{31}$$

La droite des moindres carrées est donc $y = \alpha\ x + \beta$ et les « moindres carrés » dont il s'agit sont la somme des carrés des distances (verticales) entre valeurs prédites par la droite $\alpha\ x_n + \beta$ et mesurée y_n.

Cette présentation est très commode, notamment elle se généralise sans peine au cas de la régression polynomiale

$$y = a_0 + a_1\ x + a_2\ x^2 + \ldots + a_P\ x^P$$

ou encore au cas d'une dépendance linéaire multiple

$$y = a_0 + a_1\ x_1 + a_2\ x_2 + \ldots + a_P\ x_P$$

où les x_p, $p = 1 \ldots P$ seraient des variables indépendantes. Mais elle est insuffisante pour un traitement statistique des données et c'est pourquoi il est nécessaire de la revisiter de ce point de vue.

9. certes commune de nos jours mais inventée quand même simultanément par Gauss et Legendre...

5.4 Le traitement statistique

Ce traitement statistique consiste à introduire N variables aléatoires Y_n dont les y_n sont des réalisations ; puis à construire à partir de ces variables aléatoires Y_n et des x_n trois autres A, B et E qui seront des estimateurs de α, β et σ, c'est à dire directement

$$\mathbb{E}(A) = \alpha \; ; \; \mathbb{E}(B) = \beta$$

et pour σ

$$\mathbb{E}\left(\frac{E^2}{N-2}\right) = \sigma^2$$

Cela étant fait les outils de construction des intervalles de confiance des estimations α^*, β^* et σ^* seront en place.

L'analyse du problème est un peu longue à écrire aussi se contente-t-on ici d'en donner le résultat :

nom		définition	loi de	moyenne	écart-type	réalisation
$\overline{Y}$	=	$\frac{1}{N}\sum_{n=1}^{N} Y(x_n)$	Gauss-Laplace	$\alpha\,\overline{x}+\beta$	$\frac{\sigma}{\sqrt{N}}$	$\overline{y}$
E'^2	=	$\frac{1}{\sigma^2}\sum_{n=1}^{N}(Y(x_n)-(\alpha\,x_n+\beta))^2$	χ^2 à N ddls	N	$\sqrt{2N}$	-
A	=	$\frac{\frac{1}{N}\sum_{n=1}^{N}(Y(x_n)-\overline{Y})\,(x_n-\overline{x})}{\overline{x^2}-\overline{x}^2}$	Gauss-Laplace	α	$\frac{\sigma}{\sqrt{N}\sqrt{\overline{x^2}-\overline{x}^2}}$	$\alpha^* = \frac{\overline{x\,y}-\overline{x}\,\overline{y}}{\overline{x^2}-\overline{x}^2}$
B	=	$\overline{Y}-\overline{x}\,A$	Gauss-Laplace	β	$\frac{\sigma}{\sqrt{N}}\frac{\overline{x^2}}{\overline{x^2}-\overline{x}^2}$	$\beta^* = \frac{\overline{y}\,\overline{x^2}-\overline{x}\,\overline{x\,y}}{\overline{x^2}-\overline{x}^2}$
E^2	=	$\frac{1}{\sigma^2}\sum_{n=1}^{N}(Y(x_n)-(A\,x_n+B))^2$	χ^2 à $N-2$ ddls	$N-2$	$\sqrt{2(N-2)}$	$e^2 = \frac{1}{\sigma^2}\sum_{n=1}^{N}(y_n-\alpha^* x_n-\beta^*)^2$
$\tilde{A}$	=	$\frac{A-\alpha}{\frac{\sigma}{\sqrt{N}\sqrt{\overline{x^2}-\overline{x}^2}}}\Big/\sqrt{\frac{E^2}{N-2}}$	Student-Fisher à N-2 ddls	0	$\sqrt{\frac{N-2}{N-4}}$	$\sqrt{\overline{x^2}-\overline{x}^2}\,\frac{\alpha-\alpha^*}{\sigma^*/\sqrt{N}}$
$\tilde{B}$	=	$\frac{B-\beta}{\frac{\sigma\sqrt{\overline{x^2}}}{\sqrt{N}\sqrt{\overline{x^2}-\overline{x}^2}}}\Big/\sqrt{\frac{E^2}{N-2}}$	Student-Fisher à N-2 ddls	0	$\sqrt{\frac{N-2}{N-4}}$	$\frac{\sqrt{\overline{x^2}-\overline{x}^2}}{\sqrt{\overline{x^2}}}\,\frac{\beta-\beta^*}{\sigma^*/\sqrt{N}}$

Les propriétés sont que

- A et E^2 sont indépendantes (sinon on ne pourrait pas dire que $\tilde{A}$ et suit une loi de Student-Fisher) ;
- B et E^2 sont également indépendantes ;
- mais A et B ne sont pas indépendantes ; par contre A et $\overline{Y}$ le sont.

Les intervalles de confiance

Les choses sont ainsi en place. Tout d'abord on calcule α^*, β^* puis $(\sigma^2\;e^2)$ à partir des données et des formules du tableau ; avec un risque bilatéral α' (prime pour ne pas confondre avec α la pente de la régression) on obtient l'intervalle de confiance et l'estimation de σ par

$$\chi_1^2 \leq e^2 = \frac{1}{\sigma^2}\sum_{n=1}^{N}(y_n-\alpha^* x_n-\beta^*)^2 \leq \chi_2^2 \quad ; \quad \sigma^* = \sqrt{\frac{(\sigma^2\;e^2)}{N-2}} \tag{32}$$

où χ_1 et χ_2 sont trouvés par la loi du χ^2 à $N-2$ degrés de liberté.

Ensuite on utilise $\tilde{A}$ et $\tilde{B}$ pour trouver l'intervalle de confiance de α et β : on donne un risque α' à partir duquel on trouve le seuil bilatéral $t_{\alpha'}$ avec la loi de Student-Fisher à $N-2$ degrés de liberté pour obtenir que dans $1-\alpha'$ des cas

$$\alpha^* - t_{\alpha'} \frac{\sigma^*}{\sqrt{N}\sqrt{\overline{x^2} - \overline{x}^2}} \leq \alpha \leq \alpha^* + t_{\alpha'} \frac{\sigma^*}{\sqrt{N}\sqrt{\overline{x^2} - \overline{x}^2}} \tag{33}$$

et

$$\beta^* - t_{\alpha'} \frac{\sigma^* \sqrt{\overline{x^2}}}{\sqrt{N}\sqrt{\overline{x^2} - \overline{x}^2}} \leq \beta \leq \beta^* + t_{\alpha'} \frac{\sigma^* \sqrt{\overline{x^2}}}{\sqrt{N}\sqrt{\overline{x^2} - \overline{x}^2}} \tag{34}$$

Test de dépendance des deux variables

Les résultats précédents permettent de fabriquer un test de dépendance des deux séries de variables x et y.

Si les variables aléatoires sous-jacentes aux deux séries de variables sont indépendantes alors le coefficient α devrait être nul et il suffit donc de tester l'hypothèse $\alpha = 0$.

Pour cela, si on donne un risque α' il suffit de regarder si 0 est ou non dans l'intervalle de confiance :

$\rightarrow$ il l'est : l'hypothèse d'indépendance est acceptée par défaut ;

$\rightarrow$ il ne l'est pas : on affirme qu'au risque α' de se tromper l'hypothèse est infirmée et que les deux variables ne sont pas indépendantes.

(On peut souligner ici la fiabilité du raisonnement statistique : supposons qu'on ait de façon parfaitement déterministe $y = x^2$ et que les points x_n soient répartis de façon symétrique par rapport à la valeur 0 ; dans ce cas $\alpha^* = 0$ par construction et le test de dépendance conclut invariablement à l'acceptation par défaut de l'indépendance des deux variables qui sont pourtant liées. Mais comme ce n'est qu'une acceptation par défaut, il n'y a pas de contradiction.)

Coefficient de corrélation

La question de la dépendance de deux variables est souvent traitée avec le coefficient de corrélation r tel que

$$r = \frac{\overline{x\,y} - \overline{x}\,\overline{y}}{\sqrt{\overline{y^2} - \overline{y}^2}\,\sqrt{\overline{x^2} - \overline{x}^2}}$$

qui se réécrit

$$r = \frac{\sqrt{\overline{x^2} - \overline{x}^2}}{\sqrt{\overline{y^2} - \overline{y}^2}}\,\alpha^*$$

Ce coefficient de corrélation est la pente de la droite adimensionnée par le rapport des écart-types empiriques des données.

Il faudrait introduire la variable aléatoire dont il est une réalisation et ce faisant expliquer ce qu'on appelle l'analyse de la variance. Ce ne sera pas fait ici.

La plage de variation de la droite des moindres carrés

La droite de régression est

$$y = \alpha\,\xi + \beta$$

mais α et β ne sont qu'imparfaitement connus : la seule chose que l'on sait d'eux c'est qu'ils appartiennent aux intervalles de confiances (33) et (34). De cette façon la quantité $y = \alpha\ \xi + \beta$ est elle-aussi imparfaitement connue.

On pourrait calculer un encadrement de $\alpha\ \xi + \beta$ à partir de (33) et (34) mais l'hypothèse implicite serait que A et B sont indépendantes. Or elles ne le sont pas. Il est donc nécessaire de reprendre le problème à la base et de chercher directement l'intervalle de confiance de $\alpha\ \xi + \beta$ à partir de la variable aléatoire

$$A\ \xi + B$$

qui est un estimateur de $\alpha\ \xi + \beta$ et dont l'estimation est $\alpha^*\ \xi + \beta^*$.

Tous calculs faits on trouve que cet intervalle de confiance est

$$(\alpha^*\ \xi + \beta^*) - t_{\alpha'}\ \frac{\sigma^*}{\sqrt{N}}\ \sqrt{\frac{(\xi - \overline{x})^2}{\overline{x^2} - \overline{x}^2} + 1} < \alpha\ \xi + \beta < (\alpha^*\ \xi + \beta^*) + t_{\alpha'}\ \frac{\sigma^*}{\sqrt{N}}\ \sqrt{\frac{(\xi - \overline{x})^2}{\overline{x^2} - \overline{x}^2} + 1}$$

où $t_{\alpha'}$ est le seuil de la loi de Student-Fisher à $n - 2$ degrés de liberté.

Les bornes de cet intervalle dépendent de ξ : elle sont minimales pour $\xi = \overline{x}$ et croissent avec l'écart $\xi - \overline{x}$. Cela montre que plus on s'éloigne de la valeur $\xi = \overline{x}$ moins les prévisions de la droite des moindres carrés sont précises.

L'intervalle de confiance d'une observation

L'intervalle de confiance de la quantité $\alpha\ \xi + \beta$ n'est pas celui d'une observation faite pour la valeur ξ. En effet la variable aléatoire qui correspond à ce cas est

$$Y(\xi) = \alpha\ \xi + \beta + \Sigma$$

et il est nécessaire de prendre en compte Σ. De fait cet intervalle de confiance est

$$(\alpha^*\ \xi + \beta^*) - t_{\alpha'}\ \frac{\sigma^*}{\sqrt{N}}\ \sqrt{\frac{(\xi - \overline{x})^2}{\overline{x^2} - \overline{x}^2} + 1 + N} < y(\xi) < (\alpha^*\ \xi + \beta^*) + t_{\alpha'}\ \frac{\sigma^*}{\sqrt{N}}\ \sqrt{\frac{(\xi - \overline{x})^2}{\overline{x^2} - \overline{x}^2} + 1 + N}$$

où $t_{\alpha'}$ est le seuil de la loi de Student-Fisher à $n - 2$ degrés de liberté.

Soit, pour N grand et ξ proche de $\overline{x}$

$$(\alpha^*\ \xi + \beta^*) - t_{\alpha'}\ \sigma^* < y(\xi) < (\alpha^*\ \xi + \beta^*) + t_{\alpha'}\ \sigma^*$$

c'est à dire seulement celui que la présence de Σ apporte.

Exercice type : Les données sans doute historiques de la taille des fils en fonction de celles des pères (en inches) sont disponibles

`http://springer.bme.gatech.edu/Ch16.Reg/Regdat/pearson.dat`

Rangées dans une liste `peres_et_fils` qui est injectée dans la commande (sous `Maxima` avec le package `stat`)

```
z:simple\_linear\_regression(peres\_et\_fils,conflevel=0.95);
```

elles fournissent la réponse :

```
              |             SIMPLE LINEAR REGRESSION
              |  model = 0.51409303862366 x + 33.88660435405404
              |           correlation = 0.5013383111729
              |          v_estimation = 5.936804125670635
(%o16)        | b_conf_int = [0.4610188067073, 0.56716727054002]  <- 'b est ce qu'on appelle a
              |         hypotheses = H0: b = 0 ,H1: b # 0              et vice-versa'
              |           statistic = 19.00617589043929
              |          distribution = [student_t, 1076]
              |                   p_value = 0.0
```

le problème posé est de dépouiller ces résultats. La ligne `model` donne les estimations α^* et β^* des coefficients α et β du modèle, ce sont

$$\alpha^* = 0.51409303862366 \quad ; \quad \beta^* = 33.88660435405404$$

La ligne `correlation` donne le coefficient de corrélation, donc

$$r = \frac{\overline{x\,y} - \overline{x}\,\overline{y}}{\sqrt{\overline{y^2} - \overline{y}^2}\,\sqrt{\overline{x^2} - \overline{x}^2}} = 0.5013383111729$$

La ligne `v_estimation` est l'estimation σ^* de la variance σ, donc

$$\sigma^* = \sqrt{\frac{\sum_{n=1}^{N}(y_n - \alpha^*\,x_n - \beta^*)^2}{N-2}} = 5.936804125670635$$

La ligne `b conf_int` est l'intervalle de confiance au risque 5%
(puisque l'argument de `simple_linear_regression` contient `conflevel=0.95`) du coefficient α, donc dans 95% des cas

$$\alpha^* - t_{\alpha'}\frac{\sigma^*}{\sqrt{N}\sqrt{\overline{x^2} - \overline{x}^2}} = 0.4610188067073 < \alpha < \alpha^* + t_{\alpha'}\frac{\sigma^*}{\sqrt{N}\sqrt{\overline{x^2} - \overline{x}^2}} = 0.56716727054002$$

La ligne `hypotheses = H0: b = 0 ,H1: b # 0` indique que l'hypothèse testée est

$$H0 : \alpha = 0$$

la contre hypothèse est

$$H1 : \alpha \neq 0$$

La ligne `distribution = [student_t, 1076]` indique la distribution utilisée pour tester $H0$; c'est Student-Fisher à 1076 degrés de liberté (peres_et_fils contient 1078 valeurs) c'est à dire Gauss-Laplace.
La ligne `statistic` donne la valeur que devrait prendre t_α dans la distribution utilisée pour que 0 puisse appartenir à l'intervalle de confiance.
La ligne `p value` donne la probabilité qui correspond à cette valeur. Le logiciel trouve 0 (mais c'est une valeur très faible en deçà de la précision utilisée pour les calculs) et donc l'hypothèse est rejettée très largement.

Tout cela peut apparaître comme un peu cryptique ; mais l'usage fait disparaître la première impression.

6 Raccordement à une loi de probabilité

On dispose d'une population présentant un caractère quantitatif réparti suivant une certaine distribution. On prélève dans cette population un échantillon de taille N dont on mesure le caractère pour obtenir

$$\begin{array}{|cccccc|}\hline x_1 & \dots & x_n & \dots & x_N \\ \hline\end{array} \quad \text{d'où} \quad \overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad s^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})^2$$

et on cherche à faire un test pour décider si le caractère mesuré dans cette population est réparti suivant une loi donnée. Pour fixer les idées cette loi sera supposée gaussienne.

6.1 Le principe du test

La loi donnée est gaussienne, elle est représenté par une variable aléatoire X qui dépend donc de paramètres μ et σ qu'il faudra estimer. Supposons tout d'abord que ces paramètres soient connus.

Si on donne un intervalle $I = [a, b]$ alors la variable aléatoire fonction de X

$$\mathcal{B}_I = \Pi_I(u) \text{ avec } \Pi_I(u) = \begin{cases} 1 & \text{si} \qquad u \in I \\ 0 & \text{sinon} \end{cases}$$

est une variable aléatoire de Bernoulli de paramètre

$$\overline{\omega}_I = \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \rho(t)\, dt$$

et la variable aléatoire qui compte le nombre de fois où un individu issu de échantillon tombe dans l'intervalle I

$$B_I = \sum_{n=1}^{N} \mathcal{B}_I(X_n)$$

est une variable binomiale de paramètres N et $\overline{\omega}_I$.

Si donc on connaissait avec certitude les paramètres μ et σ, on pourrait faire un test analogue à celui des sondages pour décider de l'hypothèse que la population est gaussienne.

Mais d'une part on ne connait pas μ et σ et d'autre part l'utilisation d'un seul intervalle est assez sous optimale. D'où les aménagements qui suivent.

6.2 Le test de gaussienn... itude

On commence par découper l'axe réel en *classes*, c'est à dire en intervalles

$$\mathbb{R} = \bigcup_{p=0}^{P} [b_n, b_{n+1}] \text{ avec } b_0 = -\infty,\ b_{P+1} = \infty,\ p < q \Longrightarrow b_p < b_q$$

Si la population est gaussienne de paramètres μ et σ alors les $P+1$ variables aléatoires B_p qui représentent le nombre d'individus d'un échantillon de taille N issu de cette population qu'on trouve dans l'intervalle $[b_p, b_{p+1}]$ sont binomiales.

Soit, pour $p = 0, \dots, P$:

$$\mathbb{P}(B_p = n) = \mathcal{C}_N^n \overline{\omega}_p^n (1 - \overline{\omega}_p)^{N-n}$$

où

$$\omega_p = \int_{(b_p-\mu)/\sigma}^{(b_{p+1}-\mu)/\sigma} \rho(t)\, dt \quad (\text{si } \rho(t) = \exp^{-t^2/2}/\sqrt{2\pi})$$

Si on fait l'approximation gaussienne pour chacune des variables B_p, on arrive à les considérer comme $P+1$ variables aléatoires de Gauss-Laplace de moyenne $\mu_p = N\overline{\omega}_p$ et d'écart-type $\sigma_p = \sqrt{N\overline{\omega}_p(1-\overline{\omega}_p)}$

Soit, pour $p = 0, \ldots, P$: $\mathbb{P}(n_1 \leq B_p < n_2) \approx \displaystyle\int_{(n_1-\mu_p)/\sigma_p}^{(n_2-\mu_p)/\sigma_p} \rho(t)\, dt.$

Donc on voit que

$$\tilde{K}^2 = \sum_{p=0}^{P} \frac{(B_p - N\overline{\omega}_p)^2}{N\overline{\omega}_p(1-\overline{\omega}_p)}$$

est la somme de $P+1$ carrés de variable aléatoire de Gauss-Laplace centrées et réduites.

Ce serait une variable aléatoire du χ^2 à $P+1$ degrés de libertés si les B_p étaient indépendantes, ce qu'elles ne sont pas parce que leur somme est connue, elle vaut exactement N.

Un deuxième empêchement vient de ce que les paramètres μ et σ^2 dont dépendent les B_p ne sont en fait pas connus ; il faut les remplacer par leurs estimateurs

$$\overline{B} \text{ et } \frac{N}{N-1}\,(\overline{B^2} - \overline{B}^2)$$

dont les estimations sont

$$\mu^* = \overline{x} \text{ et } \sigma^{*2} = \frac{N}{N-1}\, s^2$$

mais alors il n'est pas sûr a priori que les variables B_p ainsi transformées restent des variables de Gauss-Laplace.

Comme on le voit il y aurait une analyse probabiliste des choses à mener sérieusement ! Ici on se contentera du résultat qui est que la nouvelle variable

$$K^2 = \sum_{p=0}^{P} \frac{(B_p - N\overline{\omega}_p)^2}{N\overline{\omega}_p}$$

suit bien une loi du χ^2 mais à

$$\underbrace{P+1}_{\text{nombre de classes}} - \underbrace{1}_{\text{somme connue}} - \underbrace{2}_{\text{nombre de paramètres identifiés}}$$

degrés de liberté. C'est la somme du rapport des carrés des écart entre effectifs effectifs et effectifs théoriques rapportée aux effectifs théoriques.

Pratiquement, on choisit les $P+1$ classes puis on forme le tableau

$[b_0, b_1]$	$\ldots$	$[b_p, b_{p+1}]$	$\ldots$	$[b_P, b_{P+1}]$	$\leftarrow$ classes
n_0	$\ldots$	n_p	$\ldots$	n_P	$\leftarrow$ effectifs effectifs
$N\overline{\omega}_0$	$\ldots$	$N\overline{\omega}_p$	$\ldots$	$N\overline{\omega}_P$	$\leftarrow$ effectifs théoriques

$$K^2 = \sum_{p=0}^{P} \frac{(n_p - N\overline{\omega}_p)^2}{N\overline{\omega}_p}$$

On donne un risque (monolatéral) α à partir duquel on trouve χ^2 par

$$\int_0^{\chi^2} \rho_{\chi^2}(x)\, dx = 1 - \alpha$$

où ρ_{χ^2} est la densité de probabilité de la loi du χ^2 à $P+1-1-2$ degrés de liberté et

- si $K^2 > \chi^2$ alors l'hypothèse selon laquelle le caractère de la population est distribué suivant une loi gaussienne est réfutée au risque α de se tromper ;
- dans le cas contraire elle est acceptée par défaut.

Remarques

Ce test fait l'approximation gaussienne dans chacune des classes même quand les effectifs y sont faibles ; on voit donc une certaine imprécision ;

Suivant le nombre de classes choisi et la forme qu'on leur donne (équidistantes sauf évidemment la première et la dernière), le test peut conduire à des conclusions différentes.

Bien que ce test ait été particularisé à l'hypothèse que le caractère de la population est distribué suivant une loi de Gauss-Laplace, il peut aisément être adapté à une autre loi.

Si la loi en question est de densité $\rho(x;\alpha_1,\ldots,\alpha_M)$ qui dépend de M paramètres,

- on identifie les α_m, ce qui conduit à des estimations α_m^* ;
- donne les classes comme cela a été fait dans le cas gaussien, on calcule les proportions par $\overline{\omega}_p = \int_{b_p}^{b_{p+1}} \rho(x;\alpha_1^*,\ldots,\alpha_M^*)$
- on fait le test avec ces proportions comme dans le cas gaussien mais en retenant que le nombre de degrés de liberté est : (nombre de classes) - 1 - (nombre de paramètres identifiés).

Par exemple on peut tester si un caractère est réparti suivant une loi uniforme sur l'intervalle $[a,b]$.

Histogramme et vocabulaire

L'exposé de ce test introduit les classes qui sont la structure de l'histogramme qui sera présenté en séance.

Quelques éléments de vocabulaire sont : on appelle *étendue* de l'échantillon la quantité $\max_{n=1\ldots N} x_n - \min_{n=1\ldots N} x_n$; *médiane* la quantité $x_{1/2}$ telle si $I_{1/2}^-$ est l'ensemble des x_n inférieurs à $x_{1/2}$ et $I_{1/2}^+$ celui des x_n supérieurs à $x_{1/2}$ alors le nombre d'élément de ces deux ensembles soit $N/2$; *quartiles* $x_{1/4}$ et $x_{3/4}$ les médianes de $I_{2/2}^-$ et $I_{1/2}^+$.

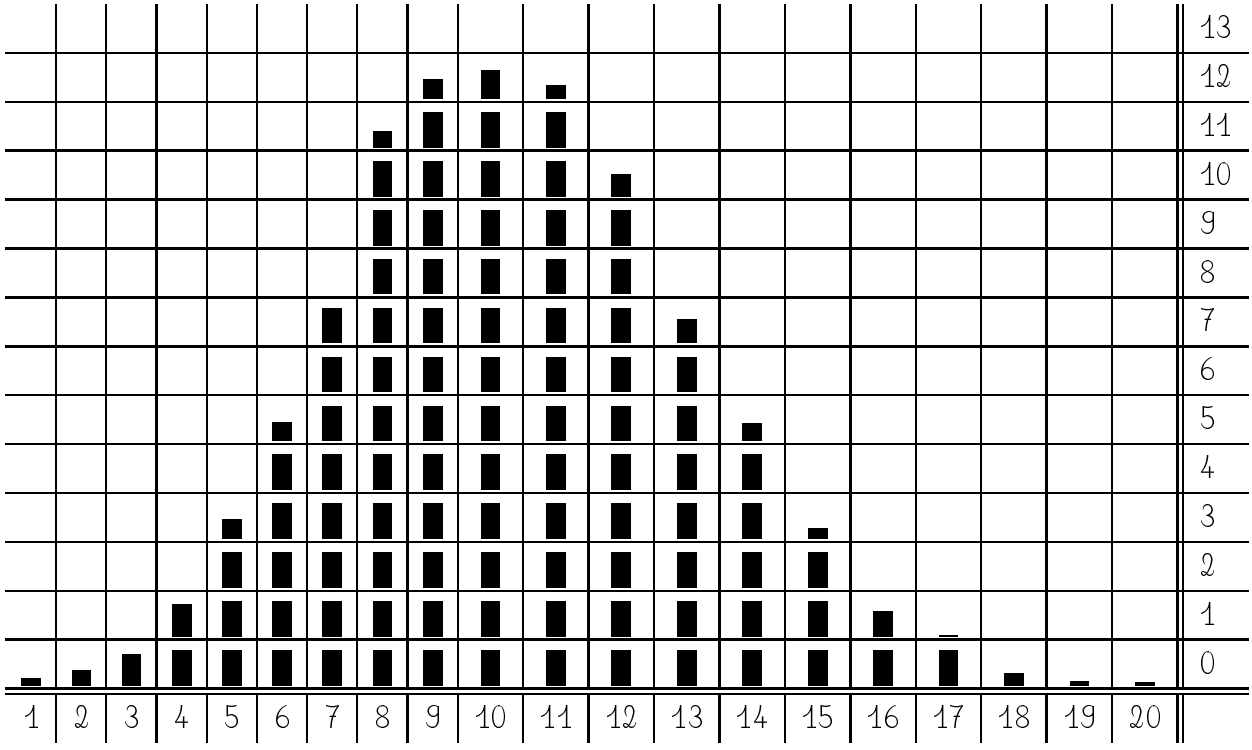
On appelle aussi *centre des classes* les quantités $c_n = (b_n + b_{n+1})/2$ pour $n = 1,\ldots,P-1$.

Exercice type : On veut vérifier que la longueur des pattes avant droite des mygales sont réparties suivant une Loi de Gauss-Laplace. On demande comment procéder.

D'abord dans le cas où on dispose d'un grand nombre de mesures ; ensuite dans le cas où on ne dispose que d'un petit nombre de mesures.

Solution pour le cas où on dispose d'un grand nombre de mesures On a mesuré les pattes de 10000 mygales (les valeurs remplissent 11 pages) ; voici l'histogramme obtenu en utilisant les classes

$$[-\infty, 3/2] \;;\; [3/2, 5/2] \;;\; \ldots \;;\; [39/2, \infty]$$



en abscisse on porte les centres des classes et en ordonnée la fréquence avec laquelle on trouve une valeur de la liste de données située dans la classe.

Les paramètres de ces données sont

Nombre de mesures	10000
$\overline{x}$	9.95
s	2.99
valeur maximale	20.0
valeur minimale	0.2
quartile 1/4	7.9
quartile 1/2 (médiane)	9.9
quartile 3/4	12.0

La question est alors : peut-on supposer que la distribution est gaussienne ?

Pour y répondre qualitativement on peut dire : a) l'histogramme a une forme 'en cloche' caractéristique ; b) la médiane et la moyenne coïncident presque, ce qui est une condition nécessaire (mais pas suffisante) pour que la distribution soit symétrique comme l'est une gaussienne (ainsi d'ailleurs qu'un nombre infini de distributions).

Ces éléments ne permettent pas de refuser d'emblée que la distribution soit gaussienne aussi fait-on l'effort de faire le test quantitatif décrit dans la notice.

D'abord on identifie les paramètres

$$\mu^* = 9.95 \quad \sigma^* = \sqrt{\tfrac{1000}{999}}2.99 = 2.9901\ldots \approx 3$$

(on fait $\sigma^* = 3$ parce qu'il ne faut pas être ridicule)

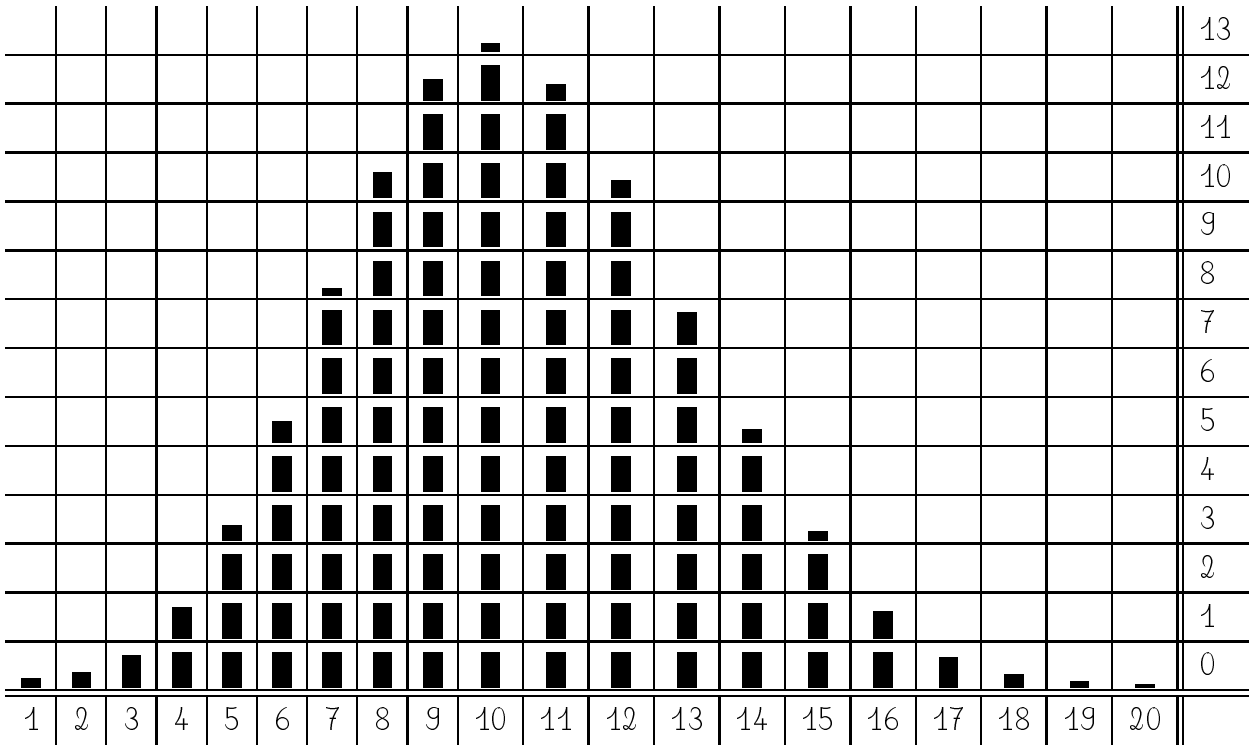
Ensuite on calcule les effectifs théoriques qui correspondent à la classe utilisée et à la loi de Gauss-Laplace de moyenne μ^* et d'écart-type σ^*, soit donc si

$$\rho(x) = \frac{\exp -\frac{1}{2}\left(\frac{x-\mu^*}{\sigma^*}\right)^2}{\sqrt{2\pi}\sigma^*}$$

les quantités

$$\overline{\omega}_0 = \int_{-\infty}^{3/2} \rho(x)dx \;\; ; \;\; \overline{\omega}_1 = \int_{3/2}^{5/2} \rho(x)dx \;\; ; \;\; \ldots \overline{\omega}_{19} = \int_{39/2}^{\infty} \rho(x)dx$$

qu'on porte sur un histogramme *théorique*



qui ressemble fort à l'histogramme empirique.

Il reste maintenant à calculer la quantité ξ, on dispose du tableau (les deux premières lignes ont servi à tracer les histogrammes et la troisième est calculée à partir des deux premières)

18	43	89	193	355	552	799	1147	1254	1281	1237	1064	767	549	328
24	41	93	189	344	561	820	1074	1260	1323	1246	1050	793	537	325
1.5	0.1	0.17	0.08	0.35	0.14	0.54	4.96	0.03	1.33	0.07	0.19	0.85	0.27	0.03

171	102	34	10	7	N_p
177	86	37	15	7	$N\overline{\omega}_p$
0.2	2.98	0.24	1.67	0.0	$(N_p - N\overline{\omega}_p)^2/N\overline{\omega}_p$

On obtient donc

$$\xi = 15.70$$

et, en prenant le risque 5%, on doit comparer cette valeur à celle de χ^2 au dessus de laquelle il ne reste que 5% de cas. Le nombre de degrés de liberté est $20-1-2$ (le nombre de classes auquel on retire forfaitairement 1 et le nombre de paramètres qu'il a fallut identifier, soit deux).

Cette valeur est $\chi^2 = 26.6 > 15.60$; on ne peut donc refuser l'hypothèse selon laquelle la distribution des longueurs de pattes de mygales est normale. On l'accepte par défaut.

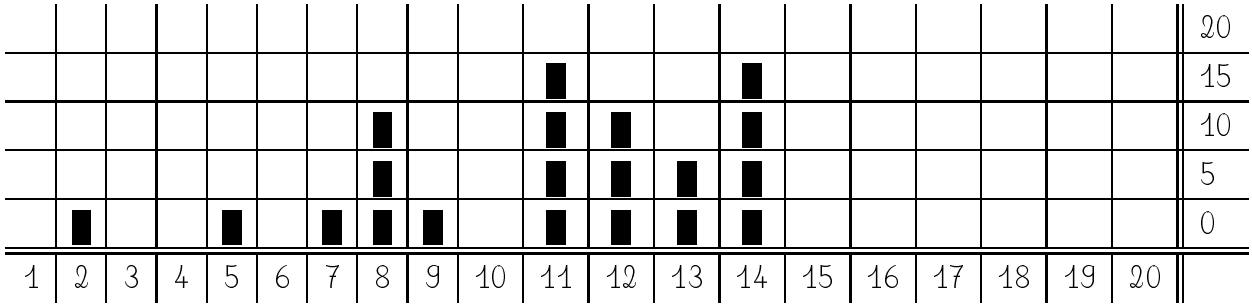
Dans ce cas on peut même dire quelque chose de plus : si on cherche le risque qui correspond à 15.60, on trouve 13.5 pour $\alpha = 70\%$. Pour refuser l'hypothèse selon laquelle la distribution est gaussienne il faudrait accepter un risque de 70% de se tromper. C'est plus qu'à pile ou face.

Solution dans le cas où on dispose d'un petit nombre de mesures On ne voit pas comment un individu seul, et même une petite équipe, pourrait mesurer 10000 longueurs de pattes de mygales. Alors en général on tente de conclure avec beaucoup moins de mesures. Et cela apporte cette difficulté supplémentaire que les histogrammes sont beaucoup moins clairs.

Si on reprend la liste de valeurs

2.3 5.3 7.0 7.5 7.8 8.1 8.7 10.7 10.7 10.9 11.4 11.6 11.7 11.9 12.5 13.1 13.5 13.7 13.7 14.3

et qu'on tente de faire un premier histogramme avec les 20 classes précédentes on trouve



qui est un peu loin d'une courbe en cloche. D'ailleurs quantitativement

Nombre de mesures	20
$\overline{x}$	10.32
s	3.09
valeur maximale	14.3
valeur minimale	2.3
quartile 1/4	8.1
quartile 1/2 (médiane)	11.4
quartile 3/4	13.1

la médiane et la moyenne sont écartées de 1/3 d'écart-type et les quartiles 1/4 et 3/4 ne sont pas symétriques par rapport à la médiane.

C'est peu prometteur. Mais pourtant on peut essayer de faire comme si on était sûr d'avoir une distribution gaussienne. On trouve donc

$$\mu^* = 10.32 \quad \sigma^* = \sqrt{\tfrac{20}{19}}3.09 = 3.17$$

Ensuite on calcule les effectifs théoriques qui correspondent à la classe utilisée et à la loi de Gauss-Laplace de moyenne μ^* et d'écart-type σ^*, soit donc si

$$\rho(x) = \frac{\exp -\frac{1}{2}\left(\frac{x-\mu^*}{\sigma^*}\right)^2}{\sqrt{2\pi}\sigma^*}$$

les quantités

$$\overline{\omega}_0 = \int_{-\infty}^{3/2} \rho(x)dx \;\; ; \;\; \overline{\omega}_1 = \int_{3/2}^{5/2} \rho(x)dx \;\; ; \;\; \ldots \overline{\omega}_{19} \int_{39/2}^{\infty} \rho(x)dx$$

Il reste maintenant à calculer la quantité ξ, on fait un tableau analogue au précédent

à ceci près qu'on va autoriser les valeurs théoriques à ne plus être entières

et cela pour éviter des divisions par 0 dans le calcul de ξ.

0	1	0	0	1	0	1	3	1
0.048525	0.081634	0.185393	0.377130	0.687177	1.121580	1.639751	2.147401	2.519055
0.05	10.33	0.19	0.38	0.14	1.12	0.25	0.34	0.92
0	4	3	2	4	0	0	0	0
2.646989	2.491476	2.100638	1.586480	1.073262	0.650373	0.353023	0.171642	0.074751
2.65	0.91	0.39	0.11	7.98	0.65	0.35	0.17	0.07
0	0							
0.029151	0.014569							
0.03	0.01							

On obtient alors

$$\xi = 27.04$$

qu'il faut toujours comparer à 26.6 si on choisit un risque de 5%.

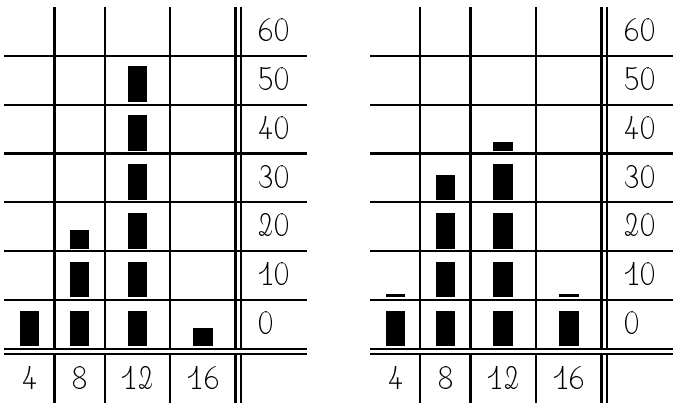
La conclusion est alors qu'on rejette l'hypothèse selon laquelle la distribution est gaussienne avec un risque de 5% de se tromper.

Mais quand même, un doute peut venir d'avoir fait 20 classes alors qu'on avait seulement 20 valeurs; et cela oblige à considérer des effectifs théoriques non entier.

Alors on peut tenter de réduire le nombre de classes. Et tant qu'a réduire ce nombre on essaie de ne pas avoir de classes vides. Par exemple on utilise

$$[-\infty, 6] \;\; ; \;\; [6, 10] \;\; ; \;\; [10, 14]; \;\; [14, \infty]$$

Il vient alors pour les histogrammes empirique et théorique



Ils ne se ressemblent pas trop, mais si on continue avec le calcul de la valeur de ξ : d'abord le tableau

2	5	12	1
2	7	8	2
0.0	0.57	2.0	0.5

qui fournit la valeur

$$\xi = 3.07$$

qui est à comparer avec la valeur du χ^2 à $4-1-2=1$ degrés de liberté qui est 12.706.

On conclut donc qu'on ne peut refuser que les données soient des réalisations de loi de Gauss-Laplace ; et qu'alors on accepte cela par défaut.

D'ailleurs pour pouvoir refuser l'hypothèse il aurait fallu prendre un risque 30% de se tromper ; ce qui déjà un risque assez grand.

Commentaires sur la solution Les valeurs utilisées ont été générées par un générateur de nombres au hasard gaussien de moyenne 10 et d'écart-type 3.

On peut voir que dans le cas où il y a un grand nombre de données les estimations ne sont pas très bonnes ; mais on peut aussi voir que dans le cas où il y a un petit nombre de données les estimations ne sont pas très mauvaises.

6.3 Comparaison d'une distribution avec une loi gaussienne connue

On a mesuré N valeurs d'un caractère numérique. On a donc trouvé μ^*, σ^* et on n'a pas d'éléments pour refuser que cette distribution de valeurs puisse être gaussienne.

D'autre part on connaît les paramètres μ et σ d'une loi de Gauss-Laplace et on soupçonne que la distribution de valeurs précédente correspond à cette loi.

On souhaite donc tester l'hypothèse : 'les valeurs mesurées correspondent à la loi de Gauss-Laplace de moyenne μ et d'écart-type σ'.

Si c'est le cas la quantité

$$\chi^2 = \frac{\frac{N-1}{N}\sigma^{*2}}{\sigma^2}$$

est la réalisation d'une loi du χ^2 à $N-1$ degrés de liberté.

On donne donc un risque α, on trouve les valeurs χ_1^2 et χ_2^2 telles que si $\rho(x)$ est la densité de probabilité de la loi du χ^2 à $N-1$ degrés de liberté alors

$$\int_0^{\chi_1^2} \rho(x)dx = \alpha/2 \quad ; \quad \int_{\chi_1^2}^{\infty} \rho(x)dx = \alpha/2$$

Si la valeur χ^2 est dans l'intervalle $[\chi_1^2, \chi_2^2]$ alors on ne peut rien conclure

et surtout pas que c'est dans α des cas on se trompe

et donc on accepte par défaut que la distribution corresponde à une loi de Gauss-Laplace d'écart-type σ.

Sinon on conclut qu'au risque α de se tromper la distribution ne correspond pas à une loi de Gauss-Laplace d'écart-type σ.

Une remarque 'χ^2 appartient ou non à $[\chi_1^2, \chi_2^2]$' est équivalent à

$$\sigma^* \sqrt{\frac{N-1}{\chi_2^2}} \leq \sigma \leq \sigma^* \sqrt{\frac{N-1}{\chi_1^2}}$$

est vrai ou faux ; ce n'est rien d'autre que l'intervalle de confiance de l'estimation de σ.

Si on conclut que les écart-types estimé et théorique sont incompatible, l'étude s'arrête ici.

Dans le cas contraire on examine si μ peut ou non appartenir à l'intervalle de confiance de l'estimation μ^* de μ ; et pour cela on utilise la valeur σ certaine et connue ; et donc on peut prendre la loi de Gauss-Laplace plutôt que Student-Fisher.

Si donc

$$\mu \in [\mu^* - t_{\alpha/2}\sigma/\sqrt{N}, \mu^* + t_{\alpha/2}\sigma/\sqrt{N}]$$

on accepte par défaut que la distribution corresponde à une loi de Gauss-Laplace de moyenne μ et d'écart-type σ.

Sinon on on affirme que ce n'est pas le cas mais qu'on se trompe dans α des cas.

Exercice type :

La contribution de Mendel Les pois (c'est la plante) produisent des graines (c'est le pois). Il existe des pois qui produisent des graines lisses ; il existe des pois qui produisent des graines ridées.

Si on croise un pois produisant des graines lisses avec un pois produisant des graines ridées alors on obtient dans les gousses de graines :

$$\begin{cases} 1/4 & \text{de graines ridées} \\ 3/4 & \text{de graines lisses} \end{cases}$$

Si on plante les graines ridées et qu'on croise les pois obtenus entre eux, on n'obtient que des graines ridés. Et cela se répète génération après génération.

Si on plante les graines lisses et qu'on croise les pois obtenus entre eux, on obtient dans les gousses de graines :

$$\begin{cases} 1/4 & \text{de graines ridées} \\ 3/4 & \text{de graines lisses} \end{cases}$$

et les ridées redonneront des ridées ; les lisses cette même proportion de lisses et de ridées.

Si chaque pois possède un couple de caractères dont les éléments sont L (pour lisse) ou R (pour ridés) et que la reproduction consiste à prendre les deux couples de chacun des pois et à former un nouveau couple de caractères pour le nouveau pois alors :

$$\begin{aligned} (R,R) + (R,R) &\longrightarrow (R,R) \\ (L,L) + (L,L) &\longrightarrow (L,L) \\ (L,R) + (L,R) &\longrightarrow \tfrac{1}{4}(L,L) + \tfrac{1}{2}(L,R) + \tfrac{1}{4}(R,R) \end{aligned}$$

en supposant l'équirépartition dans la formation des couples.

Si les couples (R,R) conduisent à des graines ridés ; les couples (L,L) à des graines lisses ; et les couples (L,R) encore à des graines lisses alors on explique les observations précédentes.

Les couples (*a priori* fictifs) forment le génotype ; le fait qu'une graine soit lisse ou ridée, le phénotype.

Dire que le couple (L,R) conduit à un phénotype lisse revient à dire que l'élément R est récessif alors que l'élément L est dominant (les mots sont de Mendel).

Voici donc la contribution de Mendel, qui a examiné 7 lots de graines de seconde génération (donc qui doivent être pour 3/4 lisses et 1/4 ridées).

Ses résultats sont, en notant n le nombre de graines lisses

E.	N°	n observées	n théoriques (3/4 N°)	E.	N°	n obs.	n thé. (3/4 N°)
1	7324	5474	5493	5	580	428	435
2	8023	6022	6017	6	858	651	644
3	929	705	697	7	1064	787	789
4	1181	882	886				

Et Mendel conclut qu'il *n'a pas pu mettre en évidence d'écarts notables entre son explication théorique et l'examen des faits.*

Statistique Un des objectifs de la statistique en général est de tenter d'être plus précis (quantitativement parce que logiquement c'est pas possible) dans les conclusions que Mendel ne l'a fait (ce qui évidemment n'enlève pas à Mendel sa stature de géant sur les épaules duquel nains que nous sommes devons-nous nous jucher afin de voir plus loin).

Dans ce cas, l'approximation gaussienne de la loi binomiale est certainement très bonne et donc l'hypothèse est que 'le nombre de graines lisse suit une loi de Gauss-Laplace de moyenne 3/4 et d'écart-type $\sqrt{3N}/4$'.

Les intervalles de confiance à 5% des données (on utilise la proportion estimée pour calculer l'intervalle de confiance comme dans le paragraphe 'estimation de proportion') sont

E.	N°	n observées	I. C.	E.	N°	n obs.	I. C.
1	7324	5474	$[0.74, 0.76]$	5	580	428	$[0.7, 0.77]$
2	8023	6022	$[0.74, 0.76]$	6	858	651	$[0.73, 0.79]$
3	929	705	$[0.73, 0.79]$	7	1064	787	$[0.71, 0.77]$
4	1181	882	$[0.72, 0.77]$				

Et pour ces 7 cas on conclut ne pas pouvoir refuser que les proportions de graines lisses soient de 3/4.

Cependant il faut un risque démesuré de 92% pour commencer à refuser l'hypothèse selon laquelle cette proportion soit de 3/4. Et c'est là que le statisticien Fisher intervient.

Too good to be true Fisher constate que l'accord entre les distributions théorique et empirique est vraiment très bon. Et que c'est peut-être *too good to be true*.

Pour mettre en évidence cela il fait l'hypothèse que les 7 nombres de graines lisses suivent une loi Gaussienne de moyenne 3/4 et d'écart-type $\sqrt{3N}/4$; il en déduit que la quantité

$$\chi^2 = \frac{(5474 - 5493)^2}{3 \times 7324/4} + \ldots + \frac{(1064 - 787)^2}{3 \times 1064/4}$$

est une réalisation d'une loi du χ^2 à 7 degrés de libertés.

Il trouve

$$\chi^2 = 1.494$$

et remarque que si $\rho(x)$ est la densité de probabilité de la loi du χ^2 à 7 degrés de liberté alors

$$\int_0^{1.69} \rho(x)dx = 0.025$$

donc que c'est seulement dans 2.5% des cas qu'on trouvera des nombres de graines lisses qui correspondent à cette plage.

Or 1.494 est dans la plage. Fisher conclut qu'il a 2.5% de chances de se tromper mais que Mendel n'a pas pu trouver des valeurs aussi déterministes.

C'est *too good to be true*. Mendel a triché.

Que penser ? Mendel n'est pas n'importe qui. Et le raisonnement statistique n'exclut pas finalement pas qu'il appartienne aux 2.5% des cas dans lesquels on se trompe en pensant que qu'il a triché.

En fait Fisher a sélectionné plus de résultats d'expériences et il arrive à la conclusion que c'est seulement dans 7 cas pour 100000 que Mendel aurait pu arriver à des résultats aussi bons.

On peut observer que la proportion de géants de la dimension de Mendel dans l'humanité toute entière est bien inférieure à 7/100000 (sinon il y en aurait en moyenne 7 à Nancy intra-muros ; ce qui ne semble pas être le cas) ; et donc le raisonnement statistique trouve dans un tel cas une limite d'application.

On pourra trouver une analyse de la question ici `http://perso.ensem.inpl-nancy.fr/Gerard.Vinsard/enseignement_2006/doc/escroquerie/Escroquerie_scientifique.pdf`

7 Problème de synthèse

7.1 Position du problème

Qu'est ce qu'une note ?

Une note est une valeur comprise entre 0 et 20. La notation d'un ensemble d'élève est donc la mise en correspondance entre les noms de ces élèves et une note.

La confection d'une notation est réalisée par un procédé qui ressemble à celui-ci : (1) un texte d'examen qui comporte des questions est réalisé ; (2) les élèves sont invités à répondre à ces questions, pour cela il sont disposés dans une même salle et disposent du même temps pour répondre aux questions ; (3) un correcteur examine chaque copie d'élève et décide d'un nombre qui est affecté à la copie et qui constitue la note.

La note est censé mesurer la valeur de l'élève dans la discipline qui lui correspond.

Mais, bien évidemment, cette mesure est sujette à caution. Par exemple un élève peut être perturbé par tout un ensemble de phénomènes extérieurs à son rapport avec la discipline (il a faim, il a froid, il a soif ...) lors de l'épreuve d'examen.

De plus la correction est également sujette à caution. Le correcteur peut avoir des difficultés à lire certaines écritures, ne pas aimer les fautes de syntaxe, et aussi avoir faim, froid, ...

Ce bref examen de ce qu'est une note met en évidence que la note obtenue mesure pour une part la valeur de l'élève dans la discipline et pour une autre part la somme d'un nombre trop important de causes pour qu'on puisse réellement les expliciter toutes.

Comme la note contient beaucoup de causes extérieures à la valeur de l'élève, on est tenté de supposer que certaines d'entre ces causes conduisent à augmenter la note idéale (celle qui correspond effectivement à la valeur de l'élève) alors que d'autres conduisent au contraire à diminuer cette note idéale ; et que l'effet global de toutes ces causes est nul.

Avec cette hypothèse, la note obtenue réellement correspond à la note idéale.

Qu'est ce que le score moyen ?

Le score moyen est la moyenne de toutes les notes. Si on croit que les causes qui conduisent à un écart entre note moyenne et idéale agissent de façon similaires dans chaque examen, alors une même cause d'écart jouera autant de fois qu'il y a d'examens pour un même élève.

Par exemple, sans qu'il en ait bien conscience d'ailleurs, si un certain élève est au meilleur de sa forme intellectuelle quand il a soif et qu'il a oublié de boire avant une épreuve il obtiendra une note plutôt meilleure à cette épreuve que s'il avait bu.

Inversement, s'il a bu avant une épreuve, il obtiendra une note inférieure à celle qu'il aurait eu s'il avait soif.

De la même façon qu'on a supposé que dans une seule épreuve la multiplicité des causes expliquant l'écart entre la note obtenue et la note réelle faisait qu'*en moyenne* leur effet était nul, on suppose ici que la multiplicité des épreuves élimine les effets d'une même cause dans la comparaison entre la note moyenne idéale et la note moyenne obtenue.

La position du problème

Il y a tellement d'incertitude sur ce qu'est exactement une note, qu'on doit largement pouvoir attaquer un tableau de notes en utilisant les méthodes de la statistique ; et c'est l'objectif de ce problème.

7.2 Les données

Un élève pense qu'il y a une certaine incohérence dans la notation de l'examen correspondant à la matière Soins-aux-Créatures-Magiques : il décide alors de mettre en évidence cette

incohérence d'une façon irréfutable.

D'abord il collecte des données.

Données de Soins-Aux-Créatures-Magiques

La correction de l'examen a livré la série de 119 notes suivantes

10 13 9 6 6.5 19.5 12 12 9 20 11.5 20 13.5 20 13 14 7.5 18 12.5 17 16 11.5 4 10 11 5 12.5 8.5 6 17 16.5 9 10.5 14 11 10.5 7.5 9 13 5.5 13.5 5 9 12 9.5 9 14.5 14 16 16 14 14 12 14.5 9.5 10 10 13 14 9 13 14.5 6 12 10 2.5 12.5 11 5 9.5 20 15 10 9 17.5 8 8 8 8.5 6.5 10 10 10 9.5 18.5 9 6 4.5 8 11 7 6.5 11 6 15 9.5 6 12 7 15 10.5 17.5 14.5 9 8.5 5 9.5 9 20 18 3 15.5 9 18 14 7 8.5 14.5 13.5

Ensuite il calcule quelques nombres utiles

nombre de valeurs	119	minimum	2.5	médiane	10.5
moyenne	11.17647058823528	maximum	20.0	quartile 1/4	8.5
écart-type (empirique)	4.115245398042475			quartile 3/4	14

Puis ces nombres n'étant pas très parlant il tente quelques représentations dans lesquels : la première colonne est la classe, la seconde le centre de la classe, la troisième le nombre de notes dans la classe, la quatrième une visualisation de ce nombre sous forme d'histogramme, la cinquième le nombre (fractionnaire) qui est obtenu en utilisant la loi théorique gaussienne (de moyenne et écart-type les estimations de ceux-ci), la sixième l'histogramme correspondant à ce nombre, la septième la différence entre les nombres de notes dans la classe empirique et théorique, et enfin la dernière une représentation graphique de ce nombre.

36 classes :

Intervalle	C.	E.	Distribution E.	G.	Distribution G.	E-G	Distribution E-G
$]-\infty, 2.75]$	2.5	1	.	2.47	..	-1	−
$]2.75, 3.25]$	3	1	.	0.81	.	0	
$]3.25, 3.75]$	3.5	0		1.02	.	-1	−
$]3.75, 4.25]$	4	1	.	1.27	.	0	
$]4.25, 4.75]$	4.5	1	.	1.56	..	-1	−
$]4.75, 5.25]$	5	4	....	1.88	..	2	++
$]5.25, 5.75]$	5.5	1	.	2.24	..	-1	−
$]5.75, 6.25]$	6	6	.....	2.62	...	3	+++
$]6.25, 6.75]$	6.5	3	...	3.03	...	0	
$]6.75, 7.25]$	7	4	....	3.45	...	1	+
$]7.25, 7.75]$	7.5	2	..	3.87	....	-2	−−
$]7.75, 8.25]$	8	4	....	4.27	....	0	
$]8.25, 8.75]$	8.5	4	....	4.66	.....	-1	−
$]8.75, 9.25]$	9	12	.....	5.0	.....	7	+++++++
$]9.25, 9.75]$	9.5	6	.....	5.29	.....	1	+
$]9.75, 10.25]$	10	8	.....	5.51	.....	2	++
$]10.25, 10.75]$	10.5	3	...	5.66	.....	-3	−−−
$]10.75, 11.25]$	11	5	.....	5.74	.....	-1	−
$]11.25, 11.75]$	11.5	2	..	5.72	.....	-4	−−−−
$]11.75, 12.25]$	12	6	.....	5.63	.....	0	
$]12.25, 12.75]$	12.5	3	...	5.45	.....	-2	−−
$]12.75, 13.25]$	13	5	.....	5.21	.....	0	
$]13.25, 13.75]$	13.5	3	...	4.9	.....	-2	−−
$]13.75, 14.25]$	14	7	.....	4.55	.....	2	++
$]14.25, 14.75]$	14.5	5	.....	4.16	....	1	+
$]14.75, 15.25]$	15	3	...	3.74	....	-1	−
$]15.25, 15.75]$	15.5	1	.	3.32	...	-2	−−
$]15.75, 16.25]$	16	3	...	2.91	...	0	
$]16.25, 16.75]$	16.5	1	.	2.51	...	-2	−−
$]16.75, 17.25]$	17	2	..	2.13	..	0	
$]17.25, 17.75]$	17.5	2	..	1.78	..	0	
$]17.75, 18.25]$	18	3	...	1.47	.	2	++
$]18.25, 18.75]$	18.5	1	.	1.2	.	0	
$]18.75, 19.25]$	19	0		0.96	.	-1	−
$]19.25, 19.75]$	19.5	1	.	0.76	.	0	
$]19.75, +\infty[$	20	5	.....	0.31		5	+++++

$$\sum_{r=1}^{36} \frac{(E_r - G_r)^2}{G_r} = 104.29490787967532$$

19 classes :

Intervalle	C.	E.	Distribution E.	G.	Distribution G.	E-G	Distribution E-G
$]-\infty, 2.5]$	2	1	.	2.13	..	-1	−
$]2.5, 3.5]$	3	1	.	1.63	..	-1	−
$]3.5, 4.5]$	4	2	..	2.56	...	-1	−
$]4.5, 5.5]$	5	5	.....	3.77	....	1	+
$]5.5, 6.5]$	6	9	.....	5.25	.....	4	+ + ++
$]6.5, 7.5]$	7	6	.....	6.89	.....	-1	−
$]7.5, 8.5]$	8	8	.....	8.54	.....	-1	−
$]8.5, 9.5]$	9	18	.....	9.98	.....	8	+ + + + + + ++
$]9.5, 10.5]$	10	11	.....	11.01	.....	0	
$]10.5, 11.5]$	11	7	.....	11.45	.....	-4	− − −−
$]11.5, 12.5]$	12	9	.....	11.24	.....	-2	−−
$]12.5, 13.5]$	13	8	.....	10.4	.....	-2	−−
$]13.5, 14.5]$	14	12	.....	9.08	.....	3	+ + +
$]14.5, 15.5]$	15	4	....	7.49	.....	-3	− − −
$]15.5, 16.5]$	16	4	....	5.82	.....	-2	−−
$]16.5, 17.5]$	17	4	....	4.27	....	0	
$]17.5, 18.5]$	18	4	....	2.95	...	1	+
$]18.5, 19.5]$	19	1	.	1.93	..	-1	−
$]19.5, +\infty[$	20	5	.....	0.67	.	4	+ + ++

$$\sum_{r=1}^{19} \frac{(E_r - G_r)^2}{G_r} = 45.35648437093211$$

11 classes :

Intervalle	C.	E.	Distribution E.	G.	Distribution G.	E-G	Distribution E-G
$]-\infty, 1.0]$	0	0		0.82	.	-1	−
$]1.0, 3.0]$	2	2	..	2.03	..	0	
$]3.0, 5.0]$	4	6	.....	5.19	.....	1	+
$]5.0, 7.0]$	6	14	.....	10.54	.....	3	+ + +
$]7.0, 9.0]$	8	22	.....	17.03	.....	5	+ + + + +
$]9.0, 11.0]$	10	22	.....	21.87	.....	0	
$]11.0, 13.0]$	12	16	.....	22.31	.....	-6	− − − − −−
$]13.0, 15.0]$	14	18	.....	18.1	.....	0	
$]15.0, 17.0]$	16	7	.....	11.67	.....	-5	− − − − −
$]17.0, 19.0]$	18	6	.....	5.98	.....	0	
$]19.0, +\infty[$	20	6	.....	1.52	..	4	+ + ++

$$\sum_{r=1}^{11} \frac{(E_r - G_r)^2}{G_r} = 20.354778080601072$$

5 classes :

Intervalle	C.	E.	Distribution E.	G.	Distribution G.	E-G	Distribution E-G
$]-\infty, 2.5]$	0	1	.	2.13	..	-1	−
$]2.5, 7.5]$	5	23	.....	20.1	.....	3	+ + +
$]7.5, 12.5]$	10	53	.....	52.21	.....	1	+
$]12.5, 17.5]$	15	32	.....	37.06	.....	-5	− − − − −
$]17.5, +\infty[$	20	10	.....	5.55	.....	4	+ + ++

$$\sum_{r=1}^{5} \frac{(E_r - G_r)^2}{G_r} = 5.291510721334771$$

Données de l'ensemble des examens

Il recueille maintenant les moyennes de l'ensemble des examens, soit la liste de notes

8.66 12.56 11.03 7.19 10.58 11.58 11.19 13.28 10.15 14.23 11.6 13.62 10.9 13.49 12.77 12.7 9.94 12.84 10.2 11.52 11.17 12.76 9.13 10.58 10.53 9.07 12.53 9.4 8.5 10.75 11.51 9.74 11.75 11.47 12.15 11.33 10.41 9.85 11.81 7.77 9.03 9.88 10.22 10.17 9.45 9.85 10.43 12.21 10.69 11.5 11.29 13.17 8.56 12.59 11.02 10.96 13.73 12.39 12.9 10.32 11.49 8.75 7.86 12.03 10.59 8.54 12.19 11.19 9.41 11.83 15.92 12.67 9.42 11.65 11.82 10.6 10.61 11.12 10.02 10.44 11.4 11.55 11.82 9.96 12.99 12.56 9.11 9.79 10.56 9.74 8.88 8.46 9.26 11.25 12.14 11.43 9.1 9.87 9.76 11.62 10.65 16.22 13.49 8.63 12.38 9.67 11.76 10.56 13.39 11.13 9.13 14.13 10.66 13.71 13.21 13.33 12.89 14.13 12.3

Puis, sachant que le coefficient de Soins-Aux-Créatures-Magiques est de 0.75 alors que la somme des coefficients de toutes les matières est de 6.7, il déduit les notes de Soins-aux-Créatures-Magiques de cette liste et arrive alors à la liste

8.87 12.5 11.29 7.34 11.09 10.58 11.09 13.44 10.29 13.5 11.61 12.82 10.57 12.67 12.74 12.54 10.25 12.19 9.91 10.83 10.56 12.92 9.78 10.65 10.47 9.58 12.53 9.51 8.82 9.96 10.88 9.83 11.91 11.15 12.29 11.43 10.78 9.96 11.66 8.06 8.47 10.5 10.37 9.94 9.44 9.96 9.92 11.98 10.02 10.93 10.95 13.07 8.13 12.35 11.21 11.08 14.2 12.31 12.76 10.49 11.3 8.03 8.09 12.03 10.66 9.3 12.15 11.21 9.97 12.12 15.41 12.38 9.35 11.98 11.1 10.93 10.94 11.51 10.21 10.94 11.58 11.75 12.05 10.02 12.3 13.01 9.5 10.46 10.88 9.58 9.12 8.71 9.04 11.91 11.78 11.67 9.49 9.6 10.11 11.19 10.67 16.06 13.36 8.58 12.87 10.26 12.04 10.76 12.56 10.26 9.9 13.96 10.87 13.17 13.11 14.13 13.44 14.08 12.15

qui correspond donc à 5.95 coefficients. il fait l'hypothèse que ces notes correspondent à celles qui auraient été obtenues en faisant la moyenne de $5.95/0.75 = 7.9333... \approx 8$ examens de même coefficient 0.75 que celui de Soins-aux-Créatures-Magiques.

L'élève calcule de nouveau quelques valeurs utiles

nombre de valeurs	119	minimum	7.34	médiane	10.94
moyenne	11.113733493397353	maximum	16.06	quartile 1/4	9.96
écart-type (empirique)	1.5846623680490644			quartile 3/4	12.19

puis de nouveau fait quelques représentations

19 classes :

Intervalle	C.	E.	Distribution E.	G.	Distribution G.	E-G	Distribution E-G
$]-\infty, 7.25]$	7	0		0.9	.	-1	−
$]7.25, 7.75]$	7.5	1	.	1.15	.	0	
$]7.75, 8.25]$	8	4	....	2.22	..	2	++
$]8.25, 8.75]$	8.5	3	...	3.9	....	-1	−
$]8.75, 9.25]$	9	4	....	6.19	.....	-2	−−
$]9.25, 9.75]$	9.5	9	....	8.92	....	0	
$]9.75, 10.25]$	10	15	....	11.65	....	3	+++
$]10.25, 10.75]$	10.5	14	....	13.8	....	0	
$]10.75, 11.25]$	11	19	....	14.82	....	4	++++
$]11.25, 11.75]$	11.5	9	....	14.43	....	-5	−−−−−
$]11.75, 12.25]$	12	12	....	12.74	....	-1	−
$]12.25, 12.75]$	12.5	11	....	10.2	....	1	+
$]12.75, 13.25]$	13	8	....	7.4	....	1	+
$]13.25, 13.75]$	13.5	4	....	4.87	.....	-1	−
$]13.75, 14.25]$	14	4	....	2.91	...	1	+
$]14.25, 14.75]$	14.5	0		1.57	..	-2	−−
$]14.75, 15.25]$	15	0		0.77	.	-1	−
$]15.25, 15.75]$	15.5	1	.	0.34		1	+
$]15.75, +\infty[$	16	1	.	0.1		1	+

$$\sum_{r=1}^{19} \frac{(E_r - G_r)^2}{G_r} = 19.896581193090828$$

10 classes :

Intervalle	C.	E.	Distribution E.	G.	Distribution G.	E-G	Distribution E-G
$]-\infty, 7.5]$	7	1	.	1.38	.	0	
$]7.5, 8.5]$	8	5	.....	4.6	.....	0	
$]8.5, 9.5]$	9	10	....	12.5	....	-2	−−
$]9.5, 10.5]$	10	28	.....	23.16	.....	5	+++++
$]10.5, 11.5]$	11	28	.....	29.28	.....	-1	−
$]11.5, 12.5]$	12	23	.....	25.26	.....	-2	−−
$]12.5, 13.5]$	13	17	....	14.87	....	2	++
$]13.5, 14.5]$	14	5	.....	5.97	.....	-1	−
$]14.5, 15.5]$	15	1	.	1.64	..	-1	−
$]15.5, +\infty[$	16	1	.	0.24		1	+

$$\sum_{r=1}^{10} \frac{(E_r - G_r)^2}{G_r} = 5.10092522596905$$

Croisement des deux séries de données

L'élève décide maintenant de représenter l'une des séries de données en fonction de l'autre. Il y a deux possibilités.

Les notes de Soins-aux-Créatures-Magiques en abscisse— Quelques nombres utiles

$$N = 119 \qquad \overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n = 11.176470588235293$$

$$\overline{y} = \frac{1}{N}\sum_{n=1}^{N} y_n = 11.113733493397358 \qquad \overline{xy} = \frac{1}{N}\sum_{n=1}^{N} x_n y_n = 127.25119906786247$$

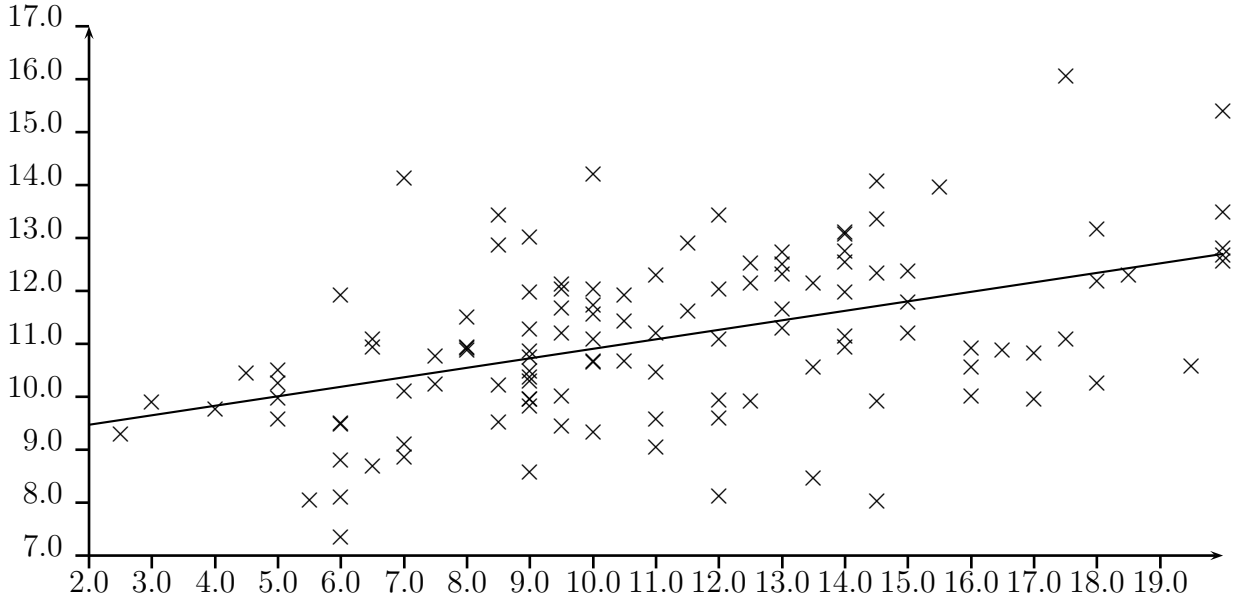
$$\overline{x^2} = \frac{1}{N}\sum_{n=1}^{N} (x_n)^2 = 141.84873949579833 \qquad \overline{y^2} = \frac{1}{N}\sum_{n=1}^{N} (y_n)^2 = 126.0262269829731$$

$$\mathrm{Cov}xy = \overline{xy} - \overline{xy} = 3.038883553421414 \qquad \mathrm{Var}x = \overline{x^2} - \overline{x}^2 = 4.1152453980424735$$

$$\mathrm{Var}y = \overline{y^2} - \overline{y}^2 = 1.5846623680490564 \qquad a = \mathrm{Cov}xy/\mathrm{Var}x = 0.17944137269619131$$

$$b = \overline{y} - a\overline{x} = 9.108212269145811 \qquad r = \mathrm{Cov}xy/\sqrt{\mathrm{Var}x\mathrm{Var}y} = 0.46599534266441645$$

Ensuite le tracé du nuage de points et de la droite $y = ax + b$



Les notes de Soins-aux-Créatures-Magiques en ordonnées Quelques nombres utiles

$N = 119$ $\quad \overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n = 11.113733493397358$

$\overline{y} = \frac{1}{N}\sum_{n=1}^{N} y_n = 11.176470588235293$ $\quad \overline{xy} = \frac{1}{N}\sum_{n=1}^{N} x_n y_n = 127.25119906786247$

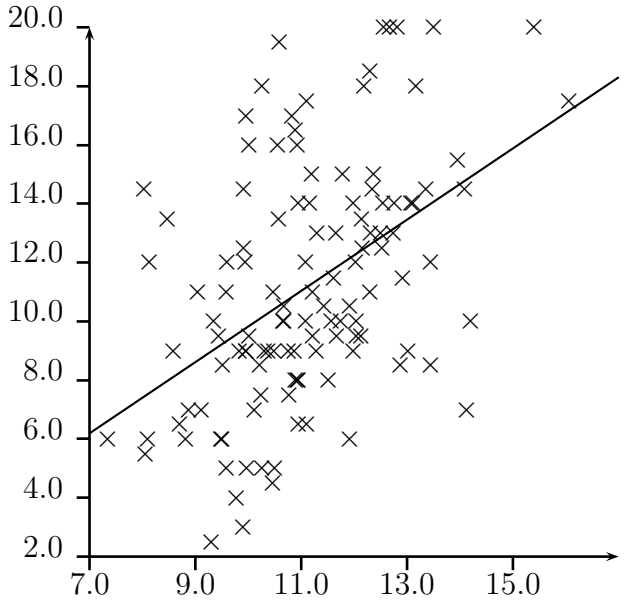
$\overline{x^2} = \frac{1}{N}\sum_{n=1}^{N} (x_n)^2 = 126.0262269829731$ $\quad \overline{y^2} = \frac{1}{N}\sum_{n=1}^{N} (y_n)^2 = 141.84873949579833$

$\mathrm{Cov}xy = \overline{xy} - \overline{xy} = 3.038883553421414$ $\quad \mathrm{Var}x = \overline{x^2} - \overline{x}^2 = 1.5846623680490564$

$\mathrm{Var}y = \overline{y^2} - \overline{y}^2 = 4.1152453980424735$ $\quad a = \mathrm{Cov}xy/\mathrm{Var}x = 1.2101538018915077$

$b = \overline{y} - a\overline{x} = -2.272856252008484$ $\quad r = \mathrm{Cov}xy/\sqrt{\mathrm{Var}x\mathrm{Var}y} = 0.46599534266441645$

Ensuite le tracé du nuage de points et de la droite $y = ax + b$



Le tracé des résidus L'élève se dit qu'il pourrait avoir besoin de visualiser $y_n - ax_n - b$ en fonction de x_n dans les deux cas précédents alors il le fait.

Quelques nombres utiles

$N = 119$ $\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n = 11.176470588235293$

$\overline{y} = \frac{1}{N}\sum_{n=1}^{N} y_n = -2.4406247e-15$ $\overline{xy} = \frac{1}{N}\sum_{n=1}^{N} x_n y_n = -7.221860e-14$

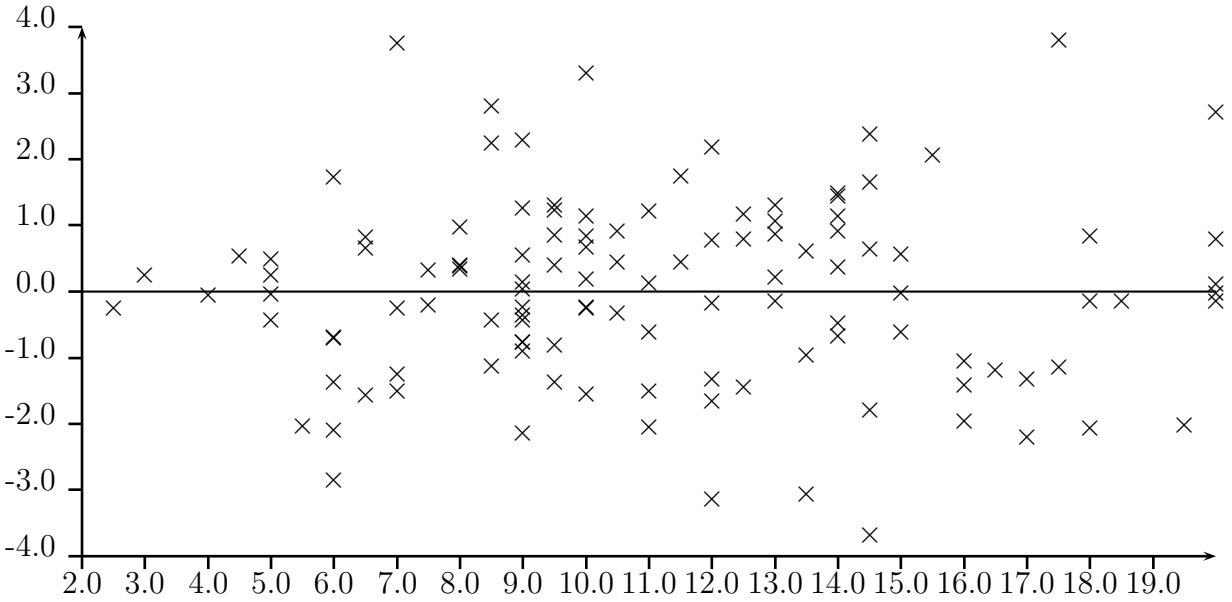
$\overline{x^2} = \frac{1}{N}\sum_{n=1}^{N} (x_n)^2 = 141.84873949579833$ $\overline{y^2} = \frac{1}{N}\sum_{n=1}^{N} (y_n)^2 = 1.9658533844210533$

$\mathrm{Cov}xy = \overline{xy} - \overline{xy} = -4.49410e-14$ $\mathrm{Var}x = \overline{x^2} - \overline{x}^2 = 4.1152453980424735$

$\mathrm{Var}y = \overline{y^2} - \overline{y}^2 = 1.4020889359883892$ $a = \mathrm{Cov}xy/\mathrm{Var}x = -2.65369875e-15$

$b = \overline{y} - a\overline{x} = 2.7218361e-14$ $r = \mathrm{Cov}xy/\sqrt{\mathrm{Var}x\mathrm{Var}y} = -7.78882e-15$

Ensuite le tracé du nuage de points et de la droite $y = ax + b$



Quelques nombres utiles

$N = 119$ $\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n = 11.113733493397358$

$\overline{y} = \frac{1}{N}\sum_{n=1}^{N} y_n = -2.328296285e-14$ $\overline{xy} = \frac{1}{N}\sum_{n=1}^{N} x_n y_n = -3.1960242e-13$

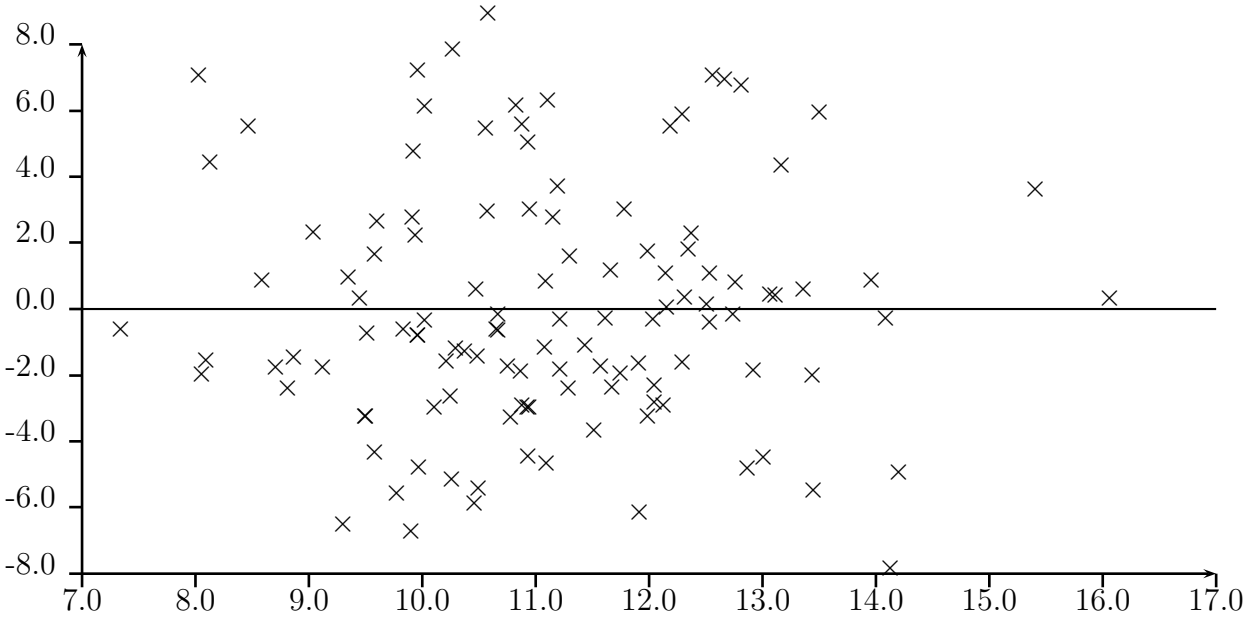
$\overline{x^2} = \frac{1}{N}\sum_{n=1}^{N} (x_n)^2 = 126.0262269829731$ $\overline{y^2} = \frac{1}{N}\sum_{n=1}^{N} (y_n)^2 = 13.257728200431364$

$\mathrm{Cov}xy = \overline{xy} - \overline{xy} = -6.0841776e-14$ $\mathrm{Var}x = \overline{x^2} - \overline{x}^2 = 1.5846623680490564$

$\mathrm{Var}y = \overline{y^2} - \overline{y}^2 = 3.641116339864927$ $a = \mathrm{Cov}xy/\mathrm{Var}x = -2.422860447e-14$

$b = \overline{y} - a\overline{x} = 2.4598729e-13$ $r = \mathrm{Cov}xy/\sqrt{\mathrm{Var}x\mathrm{Var}y} = -1.054461164e-14$

Ensuite le tracé du nuage de points et de la droite $y = ax + b$



7.3 Les questions qu'on peut poser

1. Il est commode de manipuler des distributions gaussiennes. Les distributions de Soins-Aux-Créatures-Magiques et de l'ensemble des examens peuvent-elles être considérées comme gaussiennes ?
   (a) L'élève doit formuler l'hypothèse à partir de laquelle il est possible de conclure, en utilisant un risque à choisir ainsi que celle des données précédentes qui permettent de répondre à la question ;
   (b) Il doit examiner tous les cas de mises en classes ;
   (c) dans les cas ou la distribution ne serait pas gaussienne, il lui faut identifier ce qui fait qu'on ne la trouve pas gaussienne et constater que ce n'est rien que de très normal ;
2. En admettant que ces distributions soient gaussiennes, il doit tester l'hypothèse selon laquelle il n'y a aucune différence entre les étudiants.
   (i.e. la note obtenue par un étudiant quelconque dans une matière quelconque est la réalisation d'une variable aléatoire gaussienne de moyenne μ et d'écart-type σ)
3. *A contrario* de la question précédente, il se demande s'il y a une relation de dépendance entre les notes de Soins-Aux-Créatures-Magiques et les autres notes ? Il doit donc formuler l'hypothèse correspondant à la question ; puis conclure.
4. Il lui est enfin possible de se demander si on peut trouver une liaison autre qu'affine entre les notes de Soins-Aux-Créatures-Magiques et les autres notes ?

7.4 Éléments de réponses aux questions

Les distributions sont-elles gaussiennes ?

On peut le savoir à partir de la quantité calculée à la suite du tracé des distributions.
Analysons d'abord Soins-aux-Créatures-Magiques :

Soins-aux-Créatures-Magiques

1. Pour la décomposition en 36 classes, il faut que 104.29 soit une réalisation possible d'une loi du χ^2 à $36 - 1 - 2 = 33$ degrés de liberté. Le tableau donné ne va pas jusque là et il faudrait utiliser l'approximation donnée pour les plus grands nombres que 30, voici un complément (pour éviter l'approximation indiquée)

χ^2	0.25	0.1	0.05	0.025	0.01	0.005
31	35.8870758	41.42173582	44.98534328	48.23188959	52.19139483	55.00270388
32	36.97298212	42.58474508	46.19425952	49.48043774	53.48577183	56.32811496
33	38.05752897	43.74517956	47.39988392	50.72508007	54.77553976	57.64844525
34	39.14077897	44.90315752	48.60236736	51.96599519	56.06090874	58.96392587
35	40.22278992	46.05878843	49.80184957	53.20334854	57.34207343	60.2747709

où on voit qu'on peut affirmer qu'avec un risque inférieur à 0.5 % de se tromper les notes de Soins-aux-Créatures-Magiques ne sont pas réparties suivant une loi gaussienne.

2. Pour la décomposition en 19 classes, il faut que 45.36 soit une réalisation possible d'une loi du χ^2 à $19 - 1 - 2 = 16$ degrés de liberté.

   On est dans le même cas que précédemment.

3. Pour la décomposition en 11 classes, il faut que 20.35 soit une réalisation possible d'une loi du χ^2 à $11 - 1 - 2 = 8$ degrés de liberté.

   Là, en utilisant un risque de 0.5 % on peut accepter (par défaut) l'hypothèse. Mais il parait plus sage de choisir un risque de 1 % et de la refuser (en se trompant dans 1 % des cas).

4. Pour la décomposition en 5 classes, il faut que 5.29 soit une réalisation possible d'une loi du χ^2 à $5 - 1 - 2 = 2$ degrés de liberté.

   On utilise maintenant un risque de 10 % pour refuser l'hypothèse ; c'est peut-être mieux de l'accepter.

On remarque donc que moins on fait de classes, plus il est facile de conclure que la répartition des notes est gaussienne. C'est normal, si il n'y a pas assez de classes on ne sépare plus les notes et on ne peut plus rien dire (et donc on accepte tout par défaut). À l'opposé, si on fait beaucoup de classes (par rapport à N), on finit par systématiquement refuser que toute distribution puisse être gaussienne.

Ici on peut remarquer quelque chose : si on choisit comme raisonnable le cas des 11 classes, et qu'on compare (visuellement) les distributions actuelles et théoriques on remarque que les 5 notes 20 sont clairement en excès. Si on enlève ces notes alors la quantité

$$\sum_{r=1}^{11} \frac{(E_r - G_r)^2}{G_r} = 20.354778080601072$$

devient

$$\sum_{r=1}^{11} \frac{(E_r - G_r)^2}{G_r} = 20.354778080601072 - \frac{(6 - 1.52)^2}{1.52^2} = 20.35 - 13.20 = 7.15$$

qui entre largement dans la zone d'acceptation que les distribution des notes soit gaussienne.

Ensemble des notes L'ensemble des notes est clairement gaussien dans les deux cas traités. C'est normal si on fait l'hypothèse qu'une note obtenue par un élève, en plus d'être assez fluctuante, mesure une qualité de l'élève qui est commune à toute les notes.

Alors en supposant que la partie fluctuante de la note est une variable aléatoire, la moyenne des notes bénéficient de la tendance qu'a toute moyenne à devenir gaussienne.

Soin-aux-Créature-Magiques est-il différent

Avant de répondre à cette question, faisons d'abord une expérimentation.

Expérimentation

Soit une variable aléatoire gaussienne de moyenne $\mu = 11$ et d'écart-type $\sigma = 4$ dont on observe des réalisations.

On commence d'abord par calculer 119 réalisations qu'on place dans une liste appelée Soins-aux-Créatures-Magiques ; puis on calcule 119 fois des séries de 8 réalisations dont on fait la moyenne et on place ces moyennes dans une liste appelée Moyenne-des-examens.

D'abord Soins-aux-Créatures-Magiques

nombre de valeurs	119	minimum	0.0	médiane	10.8
moyenne	10.931428571428558	maximum	20.0	quartile 1/4	8.12
écart-type (empirique)	3.904445706948866			quartile 3/4	13.52

et

Intervalle	C.	E.	Distribution E.	G.	Distribution G.	E-G	Distribution E-G
$]-\infty, 0.5]$	0	1	.	0.46		1	+
$]0.5, 1.5]$	1	1	.	0.5		1	+
$]1.5, 2.5]$	2	0		0.91	.	-1	−
$]2.5, 3.5]$	3	2	..	1.58	..	0	
$]3.5, 4.5]$	4	1	.	2.55	...	-2	−−
$]4.5, 5.5]$	5	3	...	3.87	....	-1	−
$]5.5, 6.5]$	6	8	.....	5.5	.....	3	+++
$]6.5, 7.5]$	7	9	.....	7.32	.....	2	++
$]7.5, 8.5]$	8	10	.....	9.14	.....	1	+
$]8.5, 9.5]$	9	7	.....	10.7	.....	-4	−−−−
$]9.5, 10.5]$	10	12	.....	11.74	.....	0	
$]10.5, 11.5]$	11	11	.....	12.07	.....	-1	−
$]11.5, 12.5]$	12	12	.....	11.64	.....	0	
$]12.5, 13.5]$	13	11	.....	10.51	.....	0	
$]13.5, 14.5]$	14	11	.....	8.9	.....	2	++
$]14.5, 15.5]$	15	4	....	7.07	.....	-3	−−−
$]15.5, 16.5]$	16	5	.....	5.26	.....	0	
$]16.5, 17.5]$	17	4	....	3.67	....	0	
$]17.5, 18.5]$	18	5	.....	2.4	..	3	+++
$]18.5, 19.5]$	19	1	.	1.47	.	0	
$]19.5, +\infty[$	20	1	.	0.48		1	+

$$\sum_{r=1}^{21} \frac{(E_r - G_r)^2}{G_r} = 11.70101289053418$$

Puis Moyenne-des-examens

nombre de valeurs	119	minimum	7.05	médiane	10.82
moyenne	10.911008403361338	maximum	14.19	quartile 1/4	10.0
écart-type (empirique)	1.3642410362437465			quartile 3/4	11.86

et

Intervalle	C.	E.	Distribution E.	G.	Distribution G.	E-G	Distribution E-G
$]-\infty, 7.5]$	7	1	.	0.76	.	0	
$]7.5, 8.5]$	8	4	....	3.91	....	0	
$]8.5, 9.5]$	9	14	.....	13.36	.....	1	+
$]9.5, 10.5]$	10	28	.....	27.44	.....	1	+
$]10.5, 11.5]$	11	32	.....	33.83	.....	-2	−−
$]11.5, 12.5]$	12	26	.....	25.06	.....	1	+
$]12.5, 13.5]$	13	9	.....	11.15	.....	-2	−−
$]13.5, +\infty[$	14	5	.....	2.5	...	2	++

$$\sum_{r=1}^{8} \frac{(E_r - G_r)^2}{G_r} = 3.1526085547884755$$

On constate que, comme c'était prévisible, l'écart-type de Soin-aux-Créatures-Magiques est proche de 4 alors que celui de Moyenne-des-examens est proche de $4/\sqrt{8} = \sqrt{2} = 1.414$.

Tout est dans tout Si maintenant on forme l'hypothèse que n'importe quel étudiant peut obtenir dans n'importe quelle matière une note qui est la réalisation d'une variable aléatoire gaussienne de moyenne μ et d'écart-type σ, alors la distribution des notes devra avoir la même allure que dans l'expérimentation précédente.

On peut formaliser l'hypothèse en remarquant qu'il s'ensuit qu'alors les deux distributions de notes doivent être l'une :

- une réalisation de 119 notes d'une même variable aléatoire gaussienne de moyenne μ et d'écart-type σ (quantités certaines mais inconnues) ;
- l'autre une realisation de 119 notes d'une même variable aléatoire de moyenne μ et d'écart-type $\sigma/\sqrt{8}$.

Par conséquent doit ramener l'écart-type de l'ensemble des notes à un écart-type qui aurait été obtenu pour une seule note, soit donc

$$1.58\ldots \times \sqrt{8} = 4.46$$

Comme l'écart-type de Soin-aux-créatures magiques est de 4.11 . . ., le carré de leur rapport est de 1.21 qui en prenant un risque de 5% réparti bilatéralement doit être comparé (on décide que $119 - 1 = 118 \approx 120$) avec les deux valeurs $F_1 = 1/1.27 = 0.79$ et $F_2 = 1.27$.

On ne peut pas mettre en évidence que Soin-aux-Créatures-Magique ait un écart-type différent de celui des autres notes.

Il reste ensuite à améliorer l'écart-type en regroupant les deux écart-types, on trouve 4.3, et le test sur les moyennes porte sur le nombre

$$\frac{10.9 - 11.1}{4.3\sqrt{\frac{2}{119}}} = -0.55$$

Ce nombre est à comparer en valeur absolue à 1.96 (pour un risque de 5% bilatéral).

On ne peut donc pas dire que l'écart entre les moyennes est significatif ; et donc l'élève échoue dans sa tentative de montrer que Soin-aux-Créatures-Magiques est une discipline dont les notes ne correspondent pas à l'ensemble des examens.

7.5 Les trois groupes

L'école de l'élève comporte trois groupes d'élèves qu'on peut nommer Gryfondor, Serdeigle et Serpentar.

Les notes de moyenne générale obtenues par ces trois groupes sont : Pour les Gryfondor

nombre de valeurs	14	minimum	9.03	médiane	11.83
moyenne	11.694285714285712	maximum	14.13	quartile 1/4	10.96
écart-type (empirique)	1.5549768683888814			quartile 3/4	12.89

Intervalle	C.	E.	Distribution E.	G.	Distribution G.	E-G	Distribution E-G
$]8.5, 9.5]$	9	2	..	1.22	.	1	+
$]9.5, 10.5]$	10	1	.	2.0	..	-1	−
$]10.5, 11.5]$	11	2	..	3.11	...	-1	−
$]11.5, 12.5]$	12	4	....	3.35	...	1	+
$]12.5, 13.5]$	13	3	...	2.48	..	1	+
$]13.5, 14.5]$	14	2	..	1.27	.	1	+

$$\sum_{r=1}^{7} \frac{(E_r - G_r)^2}{G_r} = 2.0593732057921263$$

Puis pour Serdeigle

nombre de valeurs	56	minimum	7.77	médiane	10.69
moyenne	11.017142857142847	maximum	16.22	quartile 1/4	10.02
écart-type (empirique)	1.695270431740821			quartile 3/4	11.81

Intervalle	C.	E.	Distribution E.	G.	Distribution G.	E-G	Distribution E-G
$]7.5, 8.5]$	8	3	...	2.84	...	0	
$]8.5, 9.5]$	9	6	.....	6.55	.....	-1	−
$]9.5, 10.5]$	10	10	.....	10.84	.....	-1	−
$]10.5, 11.5]$	11	20	.....	12.88	.....	7	+ + + + + + +
$]11.5, 12.5]$	12	6	.....	10.97	.....	-5	− − − − −
$]12.5, 13.5]$	13	7	.....	6.7	.....	0	
$]13.5, 14.5]$	14	2	..	2.94	...	-1	−
$]14.5, 15.5]$	15	0		0.92	.	-1	−
$]15.5, 16.5]$	16	2	..	0.21		2	++

$$\sum_{r=1}^{11} \frac{(E_r - G_r)^2}{G_r} = 24.107233043918573$$

Et enfin pour Serpentar

nombre de valeurs	49	minimum	7.19	médiane	11.25
moyenne	11.075306122448977	maximum	14.13	quartile 1/4	9.85
écart-type (empirique)	1.6457863352790045			quartile 3/4	12.3

Intervalle	C.	E.	Distribution E.	G.	Distribution G.	E-G	Distribution E-G
]6.5, 7.5]	7	1	.	0.77	.	0	
]7.5, 8.5]	8	1	.	2.2	..	-1	−
]8.5, 9.5]	9	9	.....	5.44	.....	4	+ + ++
]9.5, 10.5]	10	9	.....	9.45	.....	0	
]10.5, 11.5]	11	7	.....	11.57	.....	-5	− − − − −
]11.5, 12.5]	12	11	.....	9.97	.....	1	+
]12.5, 13.5]	13	9	.....	6.05	.....	3	+ + +
]13.5, 14.5]	14	2	..	2.58	...	-1	−

$$\sum_{r=1}^{9} \frac{(E_r - G_r)^2}{G_r} = 6.562148689974997$$

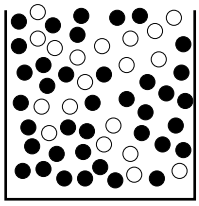
Question

Les 3 groupes d'élèves sont-ils homogènes ? on utilisera les risques 5 %, 10 % puis 20 %.

8 Éléments de probabilité

8.1 Les lois premières

Bernoulli On donne une urne remplie de $\mathcal{N}$ boules dont $\mathcal{N}\overline{\omega}$ sont noires et $\mathcal{N}(1-\overline{\omega})$ sont blanches



couleur	noir	blanche
probabilité d'obtenir une boule de cette couleur	$\overline{\omega}$	$1-\overline{\omega}$

Pour décrire cette situation on introduit une variable aléatoire X dont on connaît la forme de la loi de probabilité, c'est la loi de Bernoulli. Cette loi est complétement définie dès qu'on connaît la valeur de son paramètre : la proportion $\overline{\omega}$.

Binomial On tire avec remise un nombre N de boules et on appelle x le nombre de boules noires qu'on a obtenu. Pour décrire cette situation on introduit la variable aléatoire X dont la loi est binomiale. C'est une loi discrète avec un nombre fini de valeurs dont la table est

$X=$	0	$\dots$	n	$\dots$	N
$\mathbb{P}(X=n)$	$\overline{\omega}^N$	$\dots$	$\underbrace{\frac{N!}{n!(N-n)!}}_{=\mathcal{C}_N^n}\overline{\omega}^n(1-\overline{\omega})^{N-n}$	$\dots$	$(1-\overline{\omega})^N$

et qui possède deux paramètres N et $\overline{\omega}$.

Hypergéométrique On tire sans remise un nombre N de boules et on appelle x le nombre de boules noires qu'on a obtenu. Pour décrire cette situation il faut introduire la variable aléatoire X dont la loi est hypergéométrique.

Mais pour éviter trop de considérations calculatoires, on acceptera que si le nombre $\mathcal{N}$ est assez grand devant N alors les tirages sans remise ne modifient pas sensiblement la composition de l'urne et donc cette loi se ramène alors à la loi binomiale.

Poisson Toujours avec le modèle d'urne, la probabilité de l'aléa binomial s'écrit

$$\mathcal{C}_N^n\,\overline{\omega}^n(1-\overline{\omega})^{N-n}=\frac{N(N-1)\dots(N-(n-1))}{n(n-1)\dots 1}\frac{(N\overline{\omega})^n}{N^n}\frac{(1-\frac{N\overline{\omega}}{N})^N}{(1-\frac{N\overline{\omega}}{N})^n} \tag{35}$$

Si on considère que n et $N\overline{\omega}$ restent finis (non nuls et non infinis) et que N tend vers l'infini alors cette probabilité devient

$$\frac{(N\overline{\omega})^n}{n(n-1)\dots 1}\exp -N\overline{\omega}=\frac{\lambda^n}{n!}\exp -\lambda \tag{36}$$

où $\lambda=N\overline{\omega}$.

C'est la probabilité de la loi de Poisson qui est une loi discrète avec un nombre infini de valeurs de table

$X=$	0	$\dots$	n	$\dots$
$\mathbb{P}(X=n)$	$\exp^{-\lambda}$	$\dots$	$\frac{\lambda^n}{n!}\exp^{-\lambda}$	$\dots$

Cette loi peut approximer la loi binomiale dans le cas où N est grand alors que $\overline{\omega}$ est petit de manière que $N\overline{\omega}$ ne soit ni grand ni petit.

Gauss-Laplace Encore avec le modèle d'urne, si on fait des tirages de N boules avec remise, si Y est la variable aléatoire représentant le nombre de boules noires, si N est grand, que $\overline{\omega}$ est de l'ordre de $1/2$ et qu'on s'intéresse à des quantités comme

$$\mathbb{P}(n_1 \leq Y \leq n_2) = \sum_{n=n_1}^{n_2} \mathcal{C}_N^n \, \overline{\omega}^n (1-\overline{\omega})^{N-n} \tag{37}$$

où $0 \ll n_1 < n_2 \ll N$ on a

$$\mathbb{P}(n_1 \leq Y \leq n_2) \approx \frac{1}{\sqrt{2\pi}\sqrt{\frac{\overline{\omega}(1-\overline{\omega})}{N}}} \int_{n_1/N}^{n_2/N} \exp\left(-\frac{(x-\overline{\omega})^2}{2\dfrac{\overline{\omega}(1-\overline{\omega})}{N}} \right) \quad dx \tag{38}$$

Cette approximation rend intéressant l'introduction de la variable aléatoire X de loi continue

$$\rho(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp^{-(x-\mu)^2/2\sigma} \tag{39}$$

qui s'appelle la loi de Gauss-Laplace ou encore loi normale de paramètres μ et σ ; on peut écrire

$$\mathbb{P}(n_1 \leq Y \leq n_2) = \mathbb{P}\left(x_1 = \frac{n_1}{N} \leq X \leq x_2 = \frac{n_2}{N} \right) = \int_{x_1}^{x_2} \rho(x;\mu,\sigma)\, dx \tag{40}$$

et même, si on introduit la variable aléatoire T de densité $\rho_0(t) = \exp^{-t^2/2} \ /\sqrt{2\pi}$ qui s'appelle la loi normale centré réduite, par le jeu du changement de variable $x = (t-\mu)/\sigma$ on obtient

$$\mathbb{P}(n_1 \leq Y \leq n_2) = \mathbb{P}\left(x_1 = \frac{n_1}{N} \leq X \leq x_2 = \frac{n_2}{N} \right) = \mathbb{P}\left(t_1 = \frac{\left(\frac{n_1}{N} - \overline{\omega}\right)}{\sigma} \leq T \leq t_2 = \frac{\left(\frac{n_2}{N} - \overline{\omega}\right)}{\sigma} \right) \tag{41}$$

soit

$$\mathbb{P}(n_1 \leq Y \leq n_2) = \mathbb{P}\left(t_1 = \frac{\left(\frac{n_1}{N} - \overline{\omega}\right)}{\sigma} \leq T \leq t_2 = \frac{\left(\frac{n_2}{N} - \overline{\omega}\right)}{\sigma} \right) = 1/\sqrt{2\pi} \int_{t_1}^{t_2} \exp^{-x^2/2} \, dx \tag{42}$$

l'intégrant ne comporte plus aucuns paramètres, ceux-ci sont reportés dans les bornes de l'intégrale et cela facilite l'analyse numérique nécessaire à l'exécution des calculs effectifs.

8.2 Sur les variables aléatoires

Espérance, Variance, Écart-type Si on donne une variable aléatoire X quantitative, continue de densité $\rho(x)$ définie sur tout l'axe $\mathbb{R}$ (si elle ne l'est pas on complète la densité pour qu'elle le soit en associant 0 à toutes les valeurs pour lesquels elle ne l'était pas) alors

$$\mu = \mathbb{E}(X) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} x\rho(x)\, dx \ ; \ \sigma^2 = \mathbb{V}\text{ar}(X) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} (x-\mu)^2 \rho(x)\, dx \tag{43}$$

μ est l'espérance mathématique ou la moyenne, σ^2 est la variance, σ est l'écart-type de la variable aléatoire X.

L'espérance mathématique ou la variance peuvent ne pas exister (par exemple avec une variable aléatoire de Cauchy) mais on supposera ici que ces quantités existent toujours.

Indépendance de deux variables aléatoires Si on donne deux variables aléatoires quantitatives X et Y, elle sont dites indépendante si

$$\mathbb{P}(x_1 \leq X \leq x_2 \text{ et } y_1 \leq Y \leq y_2) = \mathbb{P}(x_1 \leq X \leq x_2)\, \mathbb{P}(y_1 \leq Y \leq y_2) \tag{44}$$

ce qui n'empêche pas qu'elles puissent avoir la même loi.

Une conséquence importante de l'indépendance de deux variables aléatoire est qu'elles sont non-corrélées

$$\mathbb{E}(X\, Y) = \mathbb{E}(X)\, \mathbb{E}(Y) \tag{45}$$

N variables aléatoires identiques Si on donne N variables aléatoires $X_1, \ldots, X_N$ identiques et indépendantes, c'est à dire de même densité ρ, de moyenne μ et d'ecart-type σ on peut avec elles définir une nouvelle variable aléatoire, la moyenne

$$\overline{X} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} X_n \tag{46}$$

dont les moyenne et écart-type sont par calcul direct μ et $\sigma/\sqrt{N}$

Inégalité de Bienaymé-Tchébychev Si X est une variable aléatoire de moyenne μ et d'écart-type σ on a

$$\mathbb{P}(|X - \mu| > k) < \frac{\sigma^2}{k^2} \tag{47}$$

Si on donne N variables aléatoires $X_1, \ldots, X_N$ identiques et indépendantes et qu'on applique l'inégalité de Bienaymé-Tchébychev à $\overline{X}$, on obtient

$$\mathbb{P}(|\overline{X} - \mu| > k) < \frac{\sigma^2}{N\, k^2} \tag{48}$$

d'où on tire la loi des grands nombres qui revient à dire que quand on fait la moyenne $\overline{X}$ d'un très grand nombre N de variable aléatoire alors cette variable aléatoire $\overline{X}$ n'est presque plus une variable aléatoire ; elle devient le simple nombre μ.

Théorème central-limit Si on donne N variables aléatoires $X_1, \ldots, X_N$ identiques et indépendantes de moyenne μ et d'écart-type σ et qu'on en fait la moyenne $\overline{X}$ alors cette moyenne tend vers une loi de Gauss-Laplace de moyenne μ et d'écart-type $\sigma/\sqrt{N}$.

8.3 Pandemonium de Gauss

On donne N variables aléatoires $X_1, \ldots, X_N$ indépendantes, identiques de moyenne μ et d'écart-type σ et suivant une loi de Gauss-Laplace. $\overline{X} = 1/N \sum_{n=1}^{N} X_n$ suit alors une loi de Gauss-Laplace de moyenne μ et d'écart-type $\sigma/\sqrt{N}$.

Loi du χ^2 La variable aléatoire

$$\chi^2 \stackrel{\text{def}}{=} \sum_{n=1}^{N} \frac{(X_n - \mu)^2}{\sigma^2} \tag{49}$$

suit une loi du χ^2 à N degrés de liberté ; dont la définition est d'être la somme du carré de N variables aléatoires de Gauss-Laplace centrées et réduites.

La densité de la loi du χ^2 à N degrés de liberté est

$$\rho_N(x) = \frac{x^{\frac{N}{2}-1}\exp^{-x/2}}{\displaystyle\int_0^\infty x^{\frac{N}{2}-1}\exp^{-x/2}\,dx} \text{ pour } x \geq 0 \text{ sinon } 0 \tag{50}$$

sa moyenne et son écart-type sont N et $\sqrt{2N}$

La variable aléatoire

$$S^2 = \sum_{n=1}^{N} \frac{(X_n - \overline{X})^2}{\sigma^2} \tag{51}$$

suit une loi du χ^2 à $N-1$ degrés de liberté. $N-1$ parce que les termes de la somme ne sont des carrés de variables aléatoire de Gauss-Laplace qui sont centrées mais pas réduites et de toutes façon pas indépendante les unes des autres.

Loi de Student-Fisher On donne une variable aléatoire X_0 suivant une loi de Gauss-Laplace centrée réduite, supplémentaire indépendante de $X_1, \ldots, X_N$, la variable aléatoire

$$T \stackrel{\text{def}}{=} \frac{X_0}{\sqrt{\dfrac{1}{N}\displaystyle\sum_{n=1}^{N}\frac{(X_n-\mu)^2}{\sigma^2}}} \tag{52}$$

suit une loi de Student-Fisher à N degrés de liberté ; dont la définition est le rapport entre une variable aléatoire de Gauss-Laplace centrée et réduite par la racine carré d'une variable aléatoire suivant une loi du χ^2 divisée par son nombre de degrés de liberté et indépendante de la variable aléatoire du numérateur.

La densité de la loi de Student-Fisher à N degrés de liberté est

$$\rho(x) = \frac{1}{(x^2+N)^{\frac{N+1}{2}}\displaystyle\int_{-\infty}^{\infty}\frac{dx}{(x^2+N)^{\frac{N+1}{2}}}} \tag{53}$$

sa moyenne est 0 (elle est centrée) et son écart-type $\sqrt{N/(N-2)}$ (et donc la loi de Student-Fisher n'est intéressante pratiquement que pour $N > 2$).

La variable aléatoire

$$T = \frac{\dfrac{\overline{X}-\mu}{\sigma/\sqrt{N}}}{\sqrt{\dfrac{1}{N-1}\displaystyle\sum_{n=1}^{N}\frac{(X_n-\overline{X})^2}{\sigma^2}}} = \frac{\overline{X}-\mu}{\frac{1}{\sqrt{N}}\sqrt{\dfrac{1}{N-1}\displaystyle\sum_{n=1}^{N}(X_n-\overline{X})^2}} \tag{54}$$

suit une loi de Student-Fisher à $N-1$ degrés de libertés. $N-1$ parce que $\sum_{n=1}^{N}\frac{(X_n-\overline{X})^2}{\sigma^2}$ suit une loi du χ^2 à $N-1$ degrés de liberté et que $\frac{\overline{X}-\mu}{\sigma/\sqrt{N}}$ lui est indépendante.

Loi de Snedecor On donne deux variables aléatoires S^2 et Q^2 suivant les lois du χ^2 à N et P degrés de liberté, la variable aléatoire

$$F = \frac{S^2/N}{Q^2/P} \tag{55}$$

suit une loi de Snedecor à N et P degrés de liberté.

La densité de la loi de Snedecor à N (numérateur) et P (dénominateur) degrés de liberté est

$$\rho(x) = \frac{x^{P-1}}{(1+x^2)^{\frac{N+P}{2}}} \frac{1}{\displaystyle\int_0^\infty \frac{x^{P-1}dx}{(1+x^2)^{\frac{N+P}{2}}}} \tag{56}$$

sa moyenne est $P/(P-2)$ (pour $P > 2$) et son écart-type $\sqrt{\frac{2P^2(N+P-2)}{N(P-2)^2(P-4)}}$ pour $P > 4$. On évitera donc de l'utiliser si $P \leq 4$.

Une chose utile à noter est que

$$\begin{aligned} \mathbb{P}(f_1 \leq F \leq f_2) &= \mathbb{P}\left(f_1 \leq \frac{S^2/N}{Q^2/P} \leq f_2\right) = \mathbb{P}\left(\frac{1}{f_2} \leq \frac{Q^2/P}{S^2/N} \leq \frac{1}{f_2}\right) \\ &= \mathbb{P}(1/f_1 \leq G \leq 1/f_2) \end{aligned} \tag{57}$$

où G est une variable aléatoire de Snedecor à P et N degrés de liberté.