Stage "Statistique et probabilités de la seconde à la terminale"

Michèle Boyer, Chaput Brigitte, Bernadette Denys, Christophe Hache, Bernard Parzysz, Jacqueline Mac, Brigitte Sotura

IREM PARIS 7

Documents pour la formation des enseignants

n° 11
mars 2008

Stages « Statistique et probabilités de la seconde à la terminale »

Groupe « Statistique » de l'IREM de Paris Diderot – Paris 7
Michèle Boyer, Brigitte Chaput, Bernadette Denys, Christophe Hache, Bernard Parzysz, Jacqueline Mac Aleese, Brigitte Sotura

ISSN : 2102-488X

INSTITUT DE RECHERCHE SUR L'ENSEIGNEMENT DES MATHÉMATIQUES
UNIVERSITÉ PARIS DIDEROT

Stages : Statistique et Probabilités de la seconde à la terminale

Introduction générale

Le groupe "STAT" de l'IREM de Paris 7 anime depuis 2000 des stages inscrits aux Plans Académiques de Formation des académies de Paris, Créteil et Versailles. Ces stages ont pour but d'éclairer les enseignants de lycée général sur les programmes dans ces deux domaines : probabilités et statistique, qui ne font pas toujours partie de la formation initiale des enseignants encore maintenant. Nous leur offrons des compléments de formation théorique. Ces apports peuvent être adaptés à l'usage ultérieur que les professeurs vont en avoir dans leurs classes : les probabilités et statistique ne sont pas développées dans le cadre de la théorie de Lebesgue et certains théorèmes importants, dont on travaille le sens, sont admis. Nous leur proposons aussi des séances de travail sur les outils TICE pour mettre en oeuvre des simulations aléatoires.

Nous avons rassemblé ici certains documents distribués lors des stages. Ce ne sont pas des documents directement utilisables avec des élèves : ils comportent la plupart une partie assez théorique mais aussi une partie plus adaptable en classe. Il est bien évident que ce qui se passe en classe est déterminant pour que les élèves s'approprient les notions dont il est question dans ce cahier : un exemple en est donné en partie F.

Ce cahier comporte 6 parties principales :

- un texte sur "fluctuation, échantillonnage, fourchette" (partie A): après un exposé simplifié de la théorie probabiliste et statistique sous-jacente, nous nous plaçons dans la perspective des programmes de seconde pour évoquer des activités possibles à ce niveau.

- un texte sur "données biologiques et données gaussiennes" (partie B) : ces notions sont explicitement au programme de première des sections littéraires et recouvrent une démarche statistique complexe que nous mettons en lumière. Nous finissons par une proposition de travail à partir d'un énoncé suggéré dans les annexes des programmes.

- un texte sur "tests statistiques" (partie C) : nous commençons par décrire sur des exemples simples cette démarche statistique un peu déroutante pour des non spécialistes pour arriver à une formalisation plus précise. Cela permet d'appliquer la démarche à la situation de test d'adéquation à une loi équirépartie (figurant au programme des terminales scientifiques) ainsi qu'à la situation de test d'indépendance (test du χ^2), test si souvent utilisé qu'il est difficile de ne pas en parler dans une telle formation. Nous avons illustré cette démarche par un exemple issu d'une situation expérimentée en seconde.

- une réflexion sur loi normale et loi binomiale (partie D) : ou comment on peut visualiser l'approximation d'un histogramme de loi binomiale par la densité normale (ou courbe en cloche).

- un compte-rendu d'expérience menée en seconde en 2005-2006 (partie E) : partant de la question "comment peut-on connaitre la composition d'une bouteille opaque, fermée et ne

pouvant pas être ouverte, contenant 5 boules de deux couleurs?", nous avons élaboré un scénario pour faire travailler les élèves sur la notion de hasard, leur faire prendre conscience de la fluctuation d'échantillonnage plus ou moins importante selon la taille de l'échantillon, leur faire visualiser aussi la convergence de la fréquence empirique vers la proportion..., tout cela entrant dans le cadre du programme de seconde.

- des fiches de travail pour calculatrices et tableur (partie F) : ce sont des documents proposant des activités de simulation, destinés aux enseignants pour une appropriation de ces techniques et qui nécessitent une adaptation pour une exploitation en classe.

- enfin, une bibliographie sommaire (partie G) : articles de contenus théoriques, de réflexions pédagogiques ou didactiques, sites internet de ressources...

Partie A
Echantillonnage, fluctuation, fourchette : savoirs savants, savoirs enseignés, savoirs cachés

Introduction

Dans une perspective de prise de décision concernant la protection d'une espèce, ou la prévision de production de véhicules ou la prévision d'un résultat politique important ou un pari sur un gain, la personne concernée s'intéresse à un caractère d'une population bien déterminée : être bagué ou non dans une population d'oiseaux, changer ou non de voiture pour une classe de ménages, voter oui ou non à un référendum, gagner ou perdre à un jeu de pile ou face ...

On ne peut pas examiner toute la population d'oiseaux, on ne peut interroger tous les ménages ni tous les électeurs (ou alors c'est le vote lui-même), il n'y a pas de population fixée dans un jeu de pile ou face ... Comment faire pour "évaluer" ses chances ou la proportion de ménages ayant l'intention de changer de véhicule ?

Que veut dire "évaluer" ? il y a nécessairement un passage à une formalisation. Laquelle ?

A partir de quoi "évaluer" : sondage ? échantillonnage ? recensement ?

Quel est l'intérêt d'échantillonner quand on connait la population ou quand on connait la pièce qui sert à jouer ? Quel est l'intérêt d'étudier un modèle théorique ou d'approfondir la connaissance d'un modèle théorique ?

Une telle étude est-elle nécessaire à la connaissance de l'enseignant même s'il ne peut la transmettre à ses élèves ? Savoir enseigné, savoir savant, savoir caché, savoir accessible...

Dans ce type de situation, un des objectifs du programme de seconde est l'observation, la compréhension et l'utilisation du théorème suivant : "lors de n répétitions de lancer de pièce, dans la plupart des cas, la fréquence observée de pile est une valeur approchée à $\frac{1}{\sqrt{n}}$ de la probabilité p de tomber sur pile". Quelles mathématiques cache cet énoncé très naïf ?

I Un modèle mathématique dans le cas simple d'une situation aléatoire à deux issues possibles (savoirs savants et cachés):

1) On s'intéresse donc à un caractère prenant deux valeurs, notées par convention 0 et 1. Lors de n tirages avec remise dans la population donnée ou lors de n répétitions de lancer de la pièce, on observe une fréquence F_n de 1 : quel est le lien entre F_n et la vraie proportion p dans la population (accessible si l'examen de la population entière est réalisé...) ou la vraie probabilité p de tomber sur pile (qui ne sera jamais accessible)?

   A chaque tirage ou à chaque lancer on associe un résultat X_i, prenant la valeur 0 ou 1. **On sait que chaque X_i a une probabilité p de prendre la valeur 1.**

   Dans la modélisation proposée ici, on fait l'**hypothèse très forte que les répétitions ou les tirages n'ont pas d'influence les uns sur les autres** : il y a indépendance de la famille $(X_1, X_2, ..., X_n)$. (Remarque : cette hypothèse n'est pas réalisée quand on fait un tirage exhaustif dans une population réelle, mais on peut prouver que si la taille du tirage est très faible par rapport à la taille de la population on se rapproche de ces conditions).

   En explorant mathématiquement ce modèle, on prouve des inégalités, on montre des théorèmes. En voici quelques uns, utilisés sans le dire, en seconde.

2) **Théorèmes limites : LFGN, TCL**

a) **Loi faible des grands nombres** : c'est un théorème de convergence, en un certain sens, d'une suite de variables aléatoires.

Enoncés naïfs :

- la plupart du temps, la fréquence converge vers la probabilité quand le nombre d'observations augmente.

- la probabilité que l'écart entre la fréquence et la probabilité soit grand tend vers 0.

Enoncé mathématique :

Si (X_n) est une suite de variables de Bernoulli, indépendantes et de paramètre p et si on note $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ la moyenne empirique des variables, alors pour tout a strictement positif, la probabilité que l'écart entre $\bar{X}_n$ et p dépasse a (en valeur absolue) tend vers 0 quand n tend vers l'infini:

$$\lim_{n\to\infty} \mathbf{P}(|\bar{X}_n - p| > a) = 0$$

b) **Théorème central limite** : c'est aussi un théorème de convergence d'une suite de variables aléatoires.

Enoncés naïfs :

- quand n est grand, la moyenne empirique $\bar{X}_n$ suit à peu près la loi normale de moyenne p et de variance $\frac{p(1-p)}{n}$.

- quand n est grand, la moyenne empirique centrée et réduite suit à peu près la loi normale centrée réduite.

Enoncé mathématique :

Si (X_n) est une suite de variables de Bernoulli, indépendantes et de paramètre p, et si on note $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ la moyenne empirique des variables, alors, pour tout a et tout b, réels finis ou non, la probabilité que la variable centrée réduite $Y_n = \sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}}$ soit dans $[a, b]$ tend vers la probabilité qu'une variable de loi normale centrée réduite soit dans $[a, b]$:

$$\lim_{n\to\infty} \mathbf{P}(\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \in [a, b]) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

3) **Mathématiques utilisées pour démontrer la LFGN :**

a) Inégalité de Markov : si X est une variable aléatoire positive de moyenne m, alors pour tout a strictement positif la probabilité que X dépasse a est majorée par $\frac{m}{a}$:

$$\mathbf{P}(X > a) \leq \frac{m}{a}$$

b) Inégalité de Bienaymé Tchébychev : si X est une variable aléatoire de moyenne m et de variance σ^2, alors pour tout a strictement positif la probabilité que l'écart entre X et m dépasse a est majorée par $\frac{\sigma^2}{a^2}$:

$$\mathbf{P}(|X - m| > a) \leq \frac{\sigma^2}{a^2}$$

c) Moyenne et variance de la somme de n variables aléatoires indépendantes : c'est la somme des moyennes ou des variances des variables.

d) On applique BT à la variable $F_n = \bar{X}_n$, de moyenne p et de variance $\frac{p(1-p)}{n}$.

4) **Mathématiques utilisées pour démontrer le TCL :**

Cela peut se montrer à la main, avec des outils simples d'analyse : somme de Riemann, formule de Stirling, développements limités, continuité uniforme, intégrales de Wallis... (cf le petit livre de Emmanuel Lesigne : "Pile ou face : une introduction aux théorèmes limites du calcul des probabilités" chez Ellipses)

Sinon, cela utilise des outils de théorie de la mesure et la notion de fonction caractéristique d'une loi de probabilité.

5) Cas où on n'a pas besoin du passage à la limite :

Ces théorèmes sont utilisés, dans le cadre de notre problèmatique, pour évaluer la probabilité que l'écart entre $\bar{X}_n$ et p soit plus grand qu'un nombre donné a.

Il y a des modélisations où le passage à la limite est inutile car on sait calculer exactement la probabilité que l'écart entre la moyenne empirique et la vraie moyenne dépasse un seuil fixé. C'est le cas quand on observe une série de mesures, modélisées chacune par une variable de loi normale de moyenne m et variance σ^2. Alors la moyenne empirique suit la loi normale de moyenne m et de variance $\frac{\sigma^2}{n}$ et la moyenne empirique centrée réduite suit exactement la loi normale centrée réduite !

La loi normale centrée réduite a été tabulée (calculs analytiques), ce qui permet de donner une valeur approchée de $\mathbf{P}(\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \in [a, b])$.

II Des probabilités vers la statistique inférentielle, toujours dans le cas simple d'une expérience aléatoire à deux issues

alors que la démarche en seconde est l'inverse : des stat observées, non montrées, vers les proba !

1) **Qu'est-ce qu'un modèle statistique ?**

Dans les situations envisagées en statistique, l'expérience est connue, une modélisation aléatoire par un espace probabilisable (ensemble des résultats possibles, famille des évènements) est proposée mais on ne connaît pas la probabilité régissant cette expérience, contrairement au paragraphe précédent où le modèle était entièrement connu (espace des résultats, évènements, probabilité) et le résultat de l'expérience (non faite) inconnu. On indexera donc maintenant la probabilité par un paramètre p : $\mathbf{P}_p$, paramètre p à déterminer plus ou moins précisément à l'aide du résultat observé de l'expérience. Dans l'exemple à deux issues possibles du paragraphe précédent, le paramètre est la probabilité p de l'issue 1.

2) **Raisonnement statistique :**

Plage de normalité d'une variable aléatoire de loi connue :

Si p est le paramètre régissant l'expérience, alors on sait que nF_n suit la loi binomiale de paramètres n et p et pour tout α, on peut trouver a tel que F_n a la probabilité $1 - \alpha$ au moins de se trouver dans l'intervalle $[p - a, p + a]$.

(ou plus généralement : on peut trouver des triplets (n, a, α), taille d'échantillon, largeur d'intervalle et risque, tel que F_n a la probabilité $1-\alpha$ au moins de se trouver dans l'intervalle $[p-a, p+a]$ ou tel que F_n a la probabilité α au plus de se trouver en dehors de l'intervalle $[p-a, p+a]$).

Quand α vaut 5% par exemple, on peut appeler la plage $[p-a, p+a]$ "plage de normalité" de la variable F_n : cela signifie que, en tirant au hasard un nombre selon la loi de F_n, il y a une probabilité de 95% que la valeur observée soit dans cette plage.

Intervalle de confiance pour le paramètre inconnu certain :

On "retourne" les inégalités pour dire : si p était la vraie valeur, avec probabilité $1-\alpha$, p se trouverait dans l'intervalle aléatoire $[F_n - a, F_n + a]$. On observe F_n et on conclut qu'une plage vraisemblable de valeurs de p est cet intervalle.

Une fois l'observation faite, F_n n'est plus aléatoire mais connu, soit $F_{n,obs}$, et on ne peut plus parler de la probabilité que la vraie valeur p soit dans l'intervalle $[F_{n,obs}-a, F_{n,obs}+a]$. La vraie valeur p, toujours inconnue et qui le sera toujours dans la vraie vie, est dans cet intervalle ou ne l'est pas.

3) **Liens entre la taille n d'échantillon, la largeur a de l'intervalle et le risque α de ne pas s'y trouver:**

Comment jouer sur ces différents nombres ?

a) **Liens exacts :** recherche des bornes de l'intervalle de manière exacte à l'aide d'abaques. Lire l'article "Avant le référendum", publié sur le site "culturemath" à destination des enseignants de lycée (http://dma.ens.fr/culturemath)

b) **Liens approchés grossiers**, avec la majoration de BT (améliorables si on connait d'autres majorations plus élaborées ou la loi exacte)

exemple 1 : $p = 0.3, a = 0.1$, $\mathbf{P_{0.3}}(|F_n - 0.3| > 0.1) \leq \frac{21}{n}$

Cette inégalité n'apprend rien si $n = 20$; elle dit qu'il y a au plus une chance sur 2 d'observer une fréquence en dehors de $[0.2, 0.4]$ si $n = 42$, moins d'une chance sur 10 d'observer une fréquence en dehors de $[0.2, 0.4]$ si $n = 210$.

exemple 2 : $p = 0.3, n = 100$, $\mathbf{P_{0.3}}(|F_{100} - 0.3| > a) \leq \frac{0.21}{100a^2}$

Cette inégalité dit qu'il y a moins d'une chance sur 5 pour observer une fréquence en dehors de $[0.2, 0.4]$ ($a = 0.1$) ; elle n'apprend rien sur la probabilité d'observer une fréquence en dehors de $[0.29, 0.31]$ ($a = 0.01$, le majorant est 21 !).

Cette inégalité de BT n'est pas assez précise, elle permet de montrer une convergence mais elle ne contrôle pas bien la vitesse à laquelle celle-ci a lieu. Cela permet d'avoir une idée du lien entre ces trois nombres : taille de l'échantillon, largeur de la plage, probabilité que F_n soit dans l'intervalle $[p-a, p+a]$. Cela permet aussi de poser les questions : quelle précision a choisir, quelle majoration (α) de risque choisir, combien de répétitions n faire ?

c) **Liens approchés plus fins** mais asymptotiques avec le TCL :

Ce théorème permet de remplacer la majoration par une valeur approchée de la probabilité :

$$\mathbf{P}_p(\sqrt{n}\frac{|F_n - p|}{\sqrt{p(1-p)}} > a) \simeq \mathbf{P}(|Z| > a)$$

où Z suit la loi normale centrée réduite.
Si $a = 1.96$, $\mathbf{P}(|Z| > a) = 5\%$; si $a = 1.65$, $\mathbf{P}(|Z| > a) = 10\%$; si $a = 2.57$, $\mathbf{P}(|Z| > a) = 1\%$.

On en déduit les plages pour F_n en fonction de p, puis en résolvant les inégalités en p les plages pour p en fonction de F_n :

On peut encore simplifier ces inégalités en remplaçant p par F_n dans le produit $p(1-p)$, car on montre que

$$\lim_{n\to\infty} \mathbf{P}_p(\sqrt{n}\frac{F_n - p}{\sqrt{F_n(1-F_n)}} \in [a,b]) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

c'est à dire que, sous $\mathbf{P}_p$, la suite de variables $\sqrt{n}\frac{F_n-p}{\sqrt{F_n(1-F_n)}}$ converge en loi vers une variable aléatoire normale centrée réduite.

4) **Que peut-on dire aux élèves de seconde ?**

a) Dans le modèle probabiliste où p est connu, on a acquis une connaissance sur des relations entre ce paramètre p et le comportement de certaines variables aléatoires, en démontrant des théorèmes. Pour une expérience donnée (lancer de pièce inconnue, tir dans une cible ou en dehors, toute situation où il y a seulement deux issues possibles), répétée de manière indépendante, modélisée par une variable de Bernoulli de paramètre p inconnu, les variables vont être effectivement observées et prendre des valeurs précises. On va lire en sens inverse les relations démontrées initialement, c'est à dire on va aller des variables observées vers le paramètre et non du paramètre vers les variables à observer.
Comment les élèves peuvent-ils accéder à cette connaissance sur le modèle à paramètre connu ? La théorie est trop complexe et le programme suggère donc de l'aborder expérimentalement :

b) **à p connu**, on expérimente un lancer de pièce équilibrée sur un tableur ou une calculatrice en faisant plusieurs fois la même chose, en faisant varier la taille de l'échantillon (n) ... et on fait des constats : variabilité de F_n à n fixé, convergence de F_n pour n croissant, fréquence du nombre de F_n dans un intervalle fixé à l'avance...
Il est entendu alors que la pièce est équilibrée et que le mécanisme permettant de faire l'expérience produit correctement ce hasard.

c) **à p inconnu de l'élève**, mais connu du professeur, pour répondre à la question précise : "quelle est la valeur de p ?", l'élève reproduit la procédure précédente et expérimente (utilisant le mécanisme précédent déréglé par le professeur de manière connue de lui seul) pour obtenir une ou plusieurs valeurs de F_n et utilise l'argumentation théorique vue plus haut : il obtient un intervalle de la forme $[F_{n,obs} - a, F_{n,obs} + a]$ ou plusieurs et peut alors dire des propriétés sur p, faire des conjectures sur ce paramètre.

d) **à p inconnu de tous** (la vraie vie !), on ne peut pas faire de simulation ! Donc, soit on fait un recensement de la population étudiée (quelquefois ou souvent impossible à réaliser), soit on fait des sondages répétés dans la population étudiée pour obtenir une plage expérimentale de valeurs possibles (cela demande du temps et des moyens - voir le traitement de l'adéquation en terminale), soit on fait un seul sondage et on utilise la théorie précédente pour faire des conjectures et prendre des décisions, avec des risques d'erreur.

III Documents, articles, ...:

- le CD des programmes de mathématiques, il y a tout dedans : les programmes de seconde à la terminale, les onze fiches de stat, les commentaires des programmes, des sites internet...
- le site http:// dma.ens.fr/culturemath : on y trouve des dossiers sur des thèmes mathématiques à l'intention des profs
- le site SEL (statistique en ligne) : www.inrialpes.fr/sel
- Article "Jeu de pile ou face ", de l'Irem de Lille sur "culturemath".

Il s'agit d'évaluer la probabilité de gagner dans un jeu à deux à la règle un peu complexe. La théorie permet de calculer effectivement la valeur mais sous une forme non close (somme d'une série !)

Que peut-on faire explorer aux élèves ?

- Article "Equirépartition d'une suite de nombres " de T. Chomette et F. Boucekkine sur "culturemath" :

quelle est la loi du premier chiffre de 2^n dans son écriture décimale, est-ce une loi uniforme sur $\{1,2,3,4,5,6,7,8,9\}$?

Les mathématiques sous-jacentes sont complexes, que peut-on faire explorer aux élèves de seconde ?

Partie B

Traitement de données biologiques : plages de normalité, données gaussiennes

I Un exemple de démarche suivie pour l' "élaboration d'une courbe "normale" d'évolution du diamètre bipariétal foetal (mesuré in utero) en fonction du terme (exprimé en semaines)"

1) **Problématique :**

Il y a des intérêts médicaux dans la détermination de cette mensuration foetale (suivi du développement foetal, adéquation avec les mesures maternelles, détermination de l'âge foetal, recherche d'anomalies céphaliques...). Dans les années 75-80, plusieurs courbes de référence étaient publiées mais ne permettaient pas, dans le service NDBS, une interprétation très juste des mesures effectuées, sans doute pour plusieurs raisons : différence de populations, d'appareillage, de méthodes d'examen et de mesure...Les échographistes ont donc pris l'initiative d'élaborer une courbe "normale", ou de référence, pour leur propre recrutement.

2) **Travail préliminaire sur les données**

Dans une première exploitation des dossiers médicaux disponibles, il a été établi une courbe à partir de 3 255 valeurs. Puis, après réflexion, certaines valeurs, correspondant à des mesures sur des patientes n'ayant pas accouché dans le service, ou à des mesures faites au tout début de la pratique de l'échographie, ou encore à des mesures liées à des grossesses de terme imprécis ou pathologiques, ont été éliminées.

Il est donc resté 2074 mesures , pour 1169 patientes de grossesse normale, classées par terme (semaine en cours).

3) **Détermination d'une "plage de normalité" pour chaque terme**

a) Il s'avère impossible d'utiliser moyenne et écart type dans la mesure où les répartitions empiriques ne sont pas symétriques et ne permettent pas d'utiliser une hypothèse de normalité (au sens probabiliste) de la distribution théorique sous-jacente.

b) Donc, la méthode des quantiles a été utilisée : il est proposé comme limite inférieure de la "normale" le 10-ième percentile (ou premier décile) et comme limite supérieure de la "normale" le 90-ième percentile ; entre ces deux limites on trouve 80% des valeurs observées.

c) Amélioration des bornes :

a) Pour les valeurs définitives, on a procédé à l'**élimination des valeurs "outliers"** selon le critère de Dixon (si $(x_1, x_2, ...x_n)$ sont les n valeurs observées rangées par ordre croissant, on considère x_1 (resp. x_n) comme "outlier" si $\frac{x_2-x_1}{x_n-x_1} > \frac{1}{3}$ (resp. $\frac{x_n-x_{n-1}}{x_n-x_1} > \frac{1}{3}$))

b) On a même déterminé des **intervalles de confiance à 95% pour ces limites** grâce à la modélisation suivante :
on appelle (X_i) la suite de v.a. indépendantes dont on a une observation $(x_1, ...x_n)$, on note q_α le quantile d'ordre α de la loi théorique des (X_i) et $Y_i = 1_{X_i \leq q_\alpha}$. Y_i suit

la loi de Bernoulli de paramètre α, $S_n = \sum_{i=1}^{n} Y_i$ suit la loi binomiale de paramètres n et α.
S_n, nombre de valeurs observées inférieures à q_α, est un estimateur du rang du quantile d'ordre α dans l'échantillon de taille n ; à partir d'une plage de normalité pour S_n on déduit un intervalle de confiance pour q_α.

c) **Plage de normalité pour une v.a. $\mathcal{B}(n, \alpha)$:**
c'est un intervalle $[a, b[$ dans $\mathbf{N}$ tel que $\sum_{k=a}^{b-1} \binom{n}{k} \alpha^k (1-\alpha)^{n-k} \geq 0.95$.
Les observations x_a et x_b constituent les bornes de l'intervalle de confiance pour q_α.
Il faut disposer d'abaques ou de moyens de calculs pour déterminer ces entiers a et b.
Restent aussi à faire des interpolations ou des calculs de rangs car on peut observer plusieurs fois les mêmes valeurs.

II Démarche générale

1) **La question**

A partir d'une sous-population (ou échantillon) d'une population donnée, échantillon sur lequel est faite une mesure quantitative (taille, poids, taux de cholestérol, nombre de plaquettes,...), on veut déterminer les valeurs "normales", c'est à dire un intervalle dans lequel une valeur mesurée sur un individu quelconque de la population a de grandes chances de se trouver.

2) Pour cela, on fait implicitement une modélisation probabiliste en disant que x_i est la valeur observée d'une variable aléatoire X_i, liée au i-ième individu de l'échantillon. On suppose que ces X_i sont des v.a. indépendantes et de même loi, inconnue.

Une plage de normalité pour une loi donnée connue sur $\mathbf{R}$ est un intervalle $[a, b]$ tel que $\mathbf{P}(a \leq X \leq b) \geq s$, où le seuil s, fixé par le praticien ou le statisticien, vaut 80%, 90% ou 95%. Plus s est grand, plus la plage $[a, b]$ est large.

Pour un seuil fixé s, il y a bien sûr une infinité de façons de déterminer des couples (a, b) solutions : on peut aussi bien choisir des intervalles bornés (et un autre critère, celui de la longueur de l'intervalle, permet de selectionner une plage) ou des intervalles non bornés (dits unilatères). L'utilisateur de ces plages de normalité choisit leur forme en fonction de l'interprétation donnée à l'observation d'une trop grande valeur ou d'une trop petite ..., interprétation en relation avec la ou les pathologies suspectées.

3) Mais, dans la plupart des cas, la loi sous-jacente n'est pas connue et il faut donc l'estimer. C'est à partir de l'estimation que seront déterminées les plages de normalité.

Comment estimer la loi de la variable mesurée ?

On peut faire une hypothèse de normalité et alors cette loi est déterminée par les deux paramètres que sont la moyenne et la variance, qu'il faut donc estimer : les estimateurs empiriques sont classiquement utilisés. Alors on sait que 95% des valeurs observables sont comprises dans l'intervalle $[\widehat{m} - 2\widehat{\sigma}, \widehat{m} + 2\widehat{\sigma}]$ (cet intervalle est le plus court pour le seuil 95%). C'est l'intervalle "standard" utilisé.

Ou cette hypothèse n'est pas réaliste. Il faut en faire d'autres plus adaptées au problème ou alors utiliser des techniques de statistique non paramétrique comme la méthode des quantiles vue dans l'exemple précédent.

4) **Pourquoi parler si souvent de données gaussiennes (ou normales) ?**

a) Si la quantité mesurée peut être considérée comme la somme d'un grand nombre de petits effets indépendants de même nature, alors cette quantité sera approximativement gaussienne. Ceci est justifié par le **théorème central limite**, dont l'énoncé est le suivant :

TCL : si Y_n est une suite de v.a. indépendantes de même loi de carré intégrable, de moyenne m et de variance σ^2 et si $X_n = \sum_{i=1}^{n} Y_i$, alors la suite de v.a. centrées réduites $\frac{X_n - nm}{\sigma\sqrt{n}}$ converge en loi vers une v.a. normale centrée réduite,

ou bien, dit de manière plus pragmatique, pour n grand, la fonction de répartition de X_n est proche de celle d'une loi normale de moyenne nm et de variance $n\sigma^2$.

Par exemple, on peut sans doute considérer que la taille d'un enfant est la somme des petits accroissemants quotidiens...C'est peut être plus difficile pour certaines variables biologiques.

b) Sur quel critère décider qu'une variable suit une loi gaussienne ?
Il y a bien sûr des tests statistiques d'adéquation : le premier connu est le test du χ^2 portant sur l'adéquation de lois discrètes, qui ne marche donc pas là directement mais qui peut s'adapter si on fait des classes.
L'allure d'une répartition empirique permet-elle d'avoir une idée ? Il faut se méfier car plusieurs distributions théoriques ont des graphes de densité "en cloche" : la loi normale bien sûr, mais aussi la loi de Cauchy, la double exponentielle (ou première loi de Laplace, même si elle est un peu pointue en l'origine !)...

III Que se passe-t-il dans les laboratoires de biologie ?

1) - il y a des domaines où les normes sont OMS et ne dépendent donc ni des réactifs, ni des populations concernées ; dans d'autres domaines, les normes ou valeurs attendues ou valeurs de référence sont données par les fabricants de réactifs (par exemple en hormonologie, pour les marqueurs tumoraux, en enzymologie...) et ces normes peuvent être adaptées à une population donnée.

- en général, ces normes sont établies à partir d'un échantillonnage (plus ou moins nombreux) et sont de la forme $[\widehat{m}-2\widehat{\sigma}, \widehat{m}+2\widehat{\sigma}]$, donc comme si les données étaient gaussiennes et pour une plage de normalité à 95%, c'est à dire que une observation normale sur 20 peut tomber en dehors de cette plage. Dans quelques cas, ces normes sont déterminées par la méthode des quantiles.

2) Quelques exemples :

a) Dans un test quantitatif permettant la détermination immunoenzymatique des produits de dégradation de la fibrine dans le plasma humain, le fournisseur écrit : dans un étude réalisée sur 200 plasmas venant de donneurs de sang, 96% des valeurs trouvées sont inférieures à 500ng/ml

b) Dans la détermination quantitative de la thyroxine libre dans le sérum, détermination utilisant les systèmes automatisés de chimiluminescence ACS:180, le fournisseur indique : 388 échantillons prélevés sur des sujets apparemment en bonne santé ont été dosés et les valeurs de référence sont pour les euthyroidiens 0.89-1.76 ng/dl, pour les hypothyroidiens moins de 0.89 et pour les hyperthyroidiens plus de 1.76

c) Pour les dosages radioimmunologiques par compétition de l'aldostérone sérique, le fournisseur du réactif donne des résultats obtenus sur une population normale ayant une alimentation sodée normale de 47 individus (sujets couchés) : valeurs extrêmes 19-117 pg/ml, moyenne =66 + ou - 22 pg/ml

(ce même fournisseur donne des résultats obtenus sur 8 sujets !)

Aucun de ces fournisseurs ne précise la méthode utilisée pour calculer ces valeurs attendues, de référence, normales...Tous précisent que c'est au laboratoire d'analyses d'établir ses propres valeurs de référence.

IV Travail à partir d'un paragraphe de l'annexe sur les données gaussiennes

1) **Enoncé :**

Dans l'annexe officielle sur les données gaussiennes (classe de première des séries générales), on trouve ces lignes :

"n personnes choisies au hasard dans une population de gens en parfaite santé subissent quatre examens médicaux indépendants : on constate alors qu'environ une personne sur cinq a au moins un examen qui sort de la plage de normalité !"

2) **Comment justifier théoriquement cette constatation empirique ?**

Solution :

Si X, Y, Z, T sont les résultats à ces quatre examens, l'hypothèse suggérée est l'indépendance de ces 4 v.a.. Si on nomme I_x, I_y, I_z et I_t les plages de normalité à 95% de ces 4 mesures, l'événement (X ou Y ou Z ou T sort de la plage correspondante) a pour probabilité, (en passant par le complémentaire) :

$$\begin{aligned} 1-\mathbf{P}(X\in I_x, Y\in I_y, Z\in I_z, T\in I_t) &= 1-\mathbf{P}(X\in I_x)\mathbf{P}(Y\in I_y)\mathbf{P}(Z\in I_z)\mathbf{P}(T\in I_t) \\ &= 1-0.95^4 \\ &= 0.196 \end{aligned}$$

en raison de l'indépendance. Donc cette probabilité est de l'ordre de 20%, ce qui est annoncé.

3) **Mais quel est le rôle de n ?**

Il sert à justifier le "environ" de l'énoncé que le paragraphe précédent ne justifie pas. Le théorème de la loi des grands nombres et le théorème central limite permettent de dire qu'il y a convergence, quand n tend vers l'infini, de la fréquence observée de l'événement considéré vers sa probabilité théorique (dans la modélisation choisie).

4) Cette solution est-elle accessible aux élèves concernés ?

5) **Comment simuler cette situation ?**

Il est dit dans cette annexe :

"il suffit de disposer de huit chiffres au hasard, de les prendre 2 par 2 pour fabriquer quatre nombres compris entre 0 et 99, puis de compter 1 si l'un au moins de ces quatre nombres est supérieur ou égal à 95 et 0 sinon. La proportion de 1 est de l'ordre de 1/5."

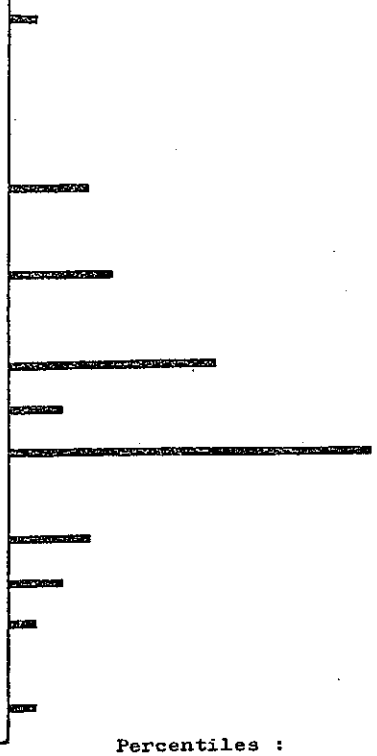
Justifier cette affirmation.

Proposer une simulation plus simple, à l'aide seulement de la fonction alea (ou random) d'une calculatrice, donnant une loi uniforme sur $[0, 1[$.

15ème semaine 26

39 valeurs

V	N	R
2	1	1
2,1	.	
2,2	.	
2,3	.	
2,4	3	3
2,5	.	
2,6	4	6,5
2,7	.	
2,8	8	12,5
2,9	2	17,5
3	14	25,5
3,1	.	
3,2	3	34
3,3	2	36,5
3,4	1	38
3,5	.	
3,6	1	39



Percentiles :

10ème = 2,45 (2 - 2,65)

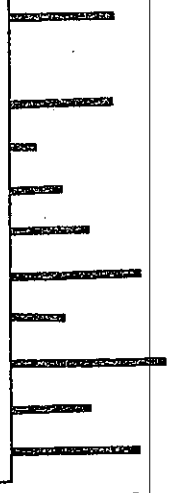
50ème = 2,93

90ème = 3,26 (3,16 - 3,6)

14ème semaine 2

35 valeurs

V	N	R
2	4	2,5
2,1	.	
2,2	4	6,5
2,3	1	9
2,4	2	10,5
2,5	3	13
2,6	5	17
2,7	2	20,5
2,8	6	24,5
2,9	3	29
3	5	33



Percentiles :

10ème = 2,05 (1,93 - 2,21)

50ème = 2,61

90ème = 2,94 (2,9 - 3,05)

31ème semaine 42

74 valeurs

V	N	R
6,6	1	1
...		
6,8	.	
6,85	1	2
6,9	.	
7	1	3
...		
7,3	2	4,5
7,4	1	6
7,5	2	7,5
7,6	6	11,5
7,7	3	16
7,8	4	19,5
7,9	4	23,5
7,95	1	26
8	14	33,5
8,1	6	43,5
8,2	7	50
8,25	1	54
8,3	4	56,5
8,4	9	63
8,5	1	68
8,6	2	69,5
8,7	1	71
8,8	1	72
8,9	.	
9	1	73
9,1	.	
9,2	1	74

Percentiles :

10ème = 7,49 (7 - 7,61)

50ème = 8,04

90ème = 8,47 (8,39 - 8,7)

30ème semaine

72 valeurs

V	N	R
6,8	2	1,5
6,9	.	
7	2	3,5
7,1	.	
7,2	6	7,5
7,3	.	
7,4	3	12
7,5	2	14,5
7,6	9	20
7,7	6	27,5
7,8	14	37,5
7,9	4	46,5
8	17	57
8,1	.	
8,2	2	66,5
8,3	3	69
8,4	2	71,5

Percentiles :

10ème = 7,19 (6,95 - 7,4)

50ème = 7,79

90ème = 8,16 (8,08 - 8,42)

38ème semaine

190 valeurs

V	N	R
8	1	1
8,1	.	
8,2	.	
8,3	..	
8,4	1	2
8,5	5	5
8,6	3	9
8,7	6	13,5
8,8	10	21,5
8,9	8	30,5
8,95	3	36
9	21	48
9,1	18	67,5
9,2	25	89
9,3	19	111
9,35	2	121,5
9,4	18	131,5
9,5	10	145,5
9,55	2	151,5
9,6	13	159
9,65	1	166
9,7	3	168
9,8	11	175
9,9	2	181,5
10	8	186,5

Percentiles :

10ème = 8,77 (8,64 - 8,86)

50ème = 9,25

90ème = 9,74 (9,64 - 9,88)

39ème semaine

164 valeurs

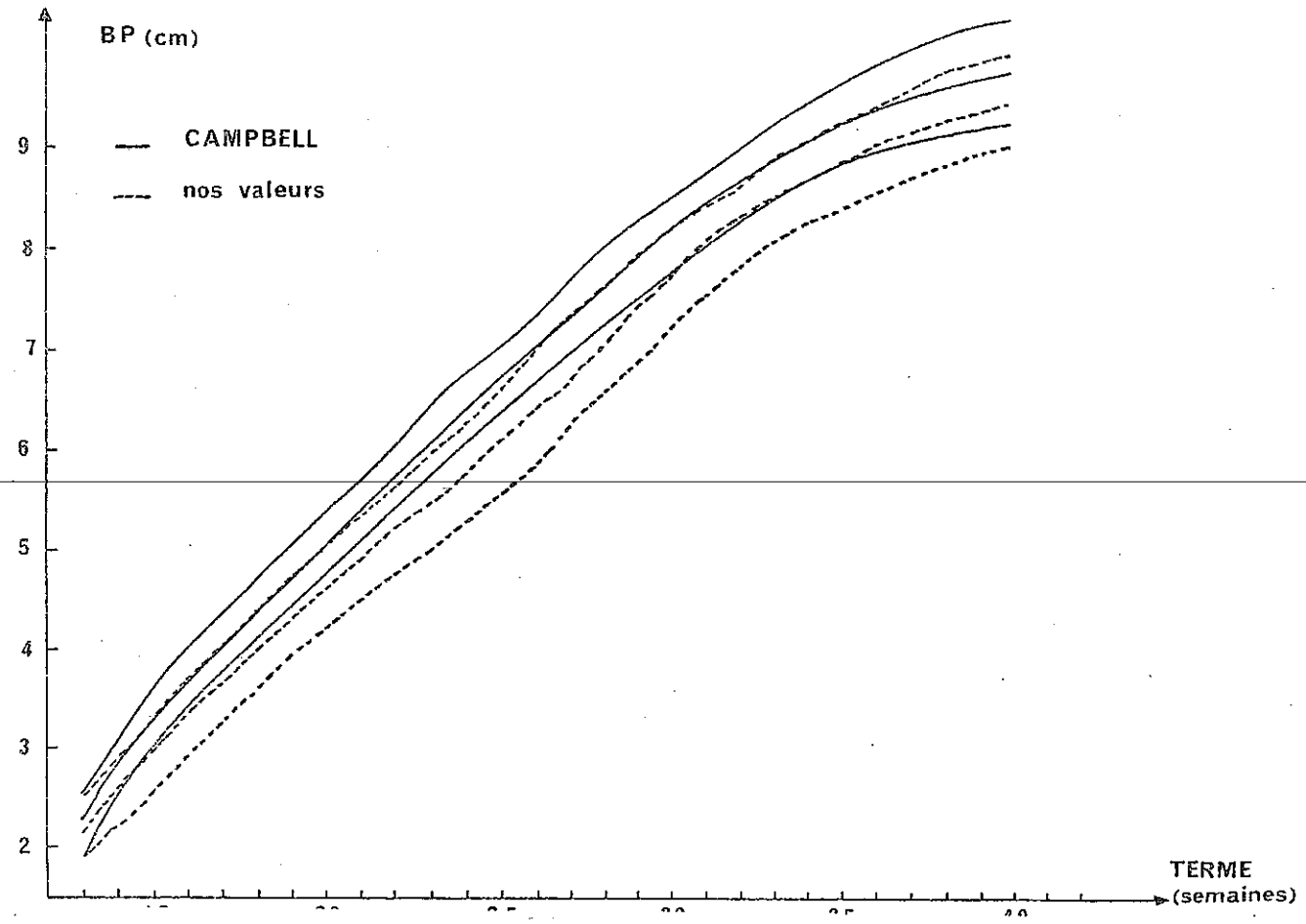
V	N	R
8,5	1	1
8,6	4	3,5
8,7	1	6
8,8	5	9
8,9	4	13,5
9	23	27
9,1	13	45
9,15	2	52,5
9,2	19	63
9,3	15	80
9,35	2	88,5
9,4	24	101,5
9,45	1	114
9,5	8	118,5
9,55	1	123
9,6	19	133
9,7	5	145
9,75	2	148,5
9,8	6	152,5
9,9	.	
10	7	159
10,1	.	
10,2	2	163,5

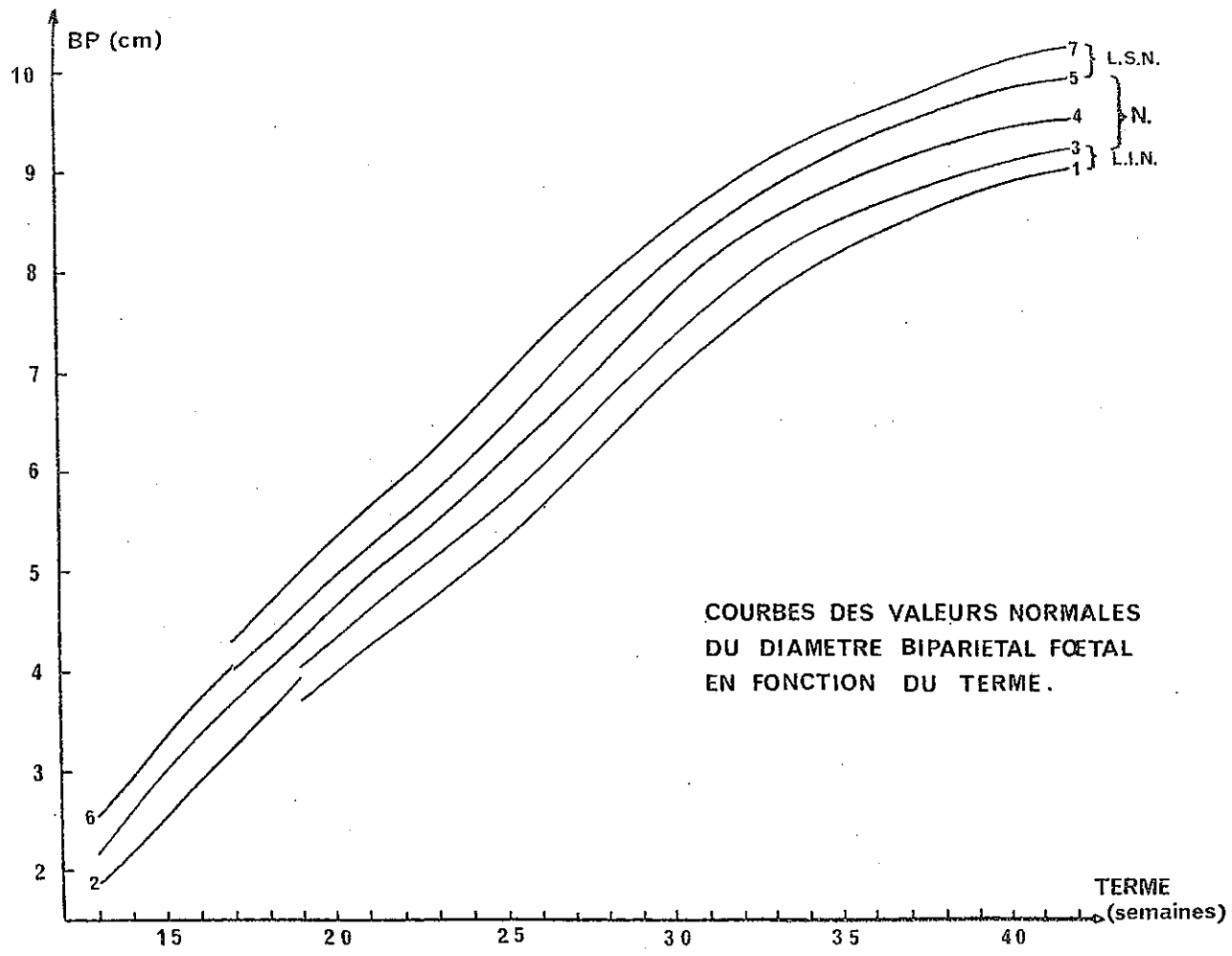
Percentiles :

10ème = 8,92 (8,8 - 8,98)

50ème = 9,31

90ème = 9,74 (9,67 - 9,91)

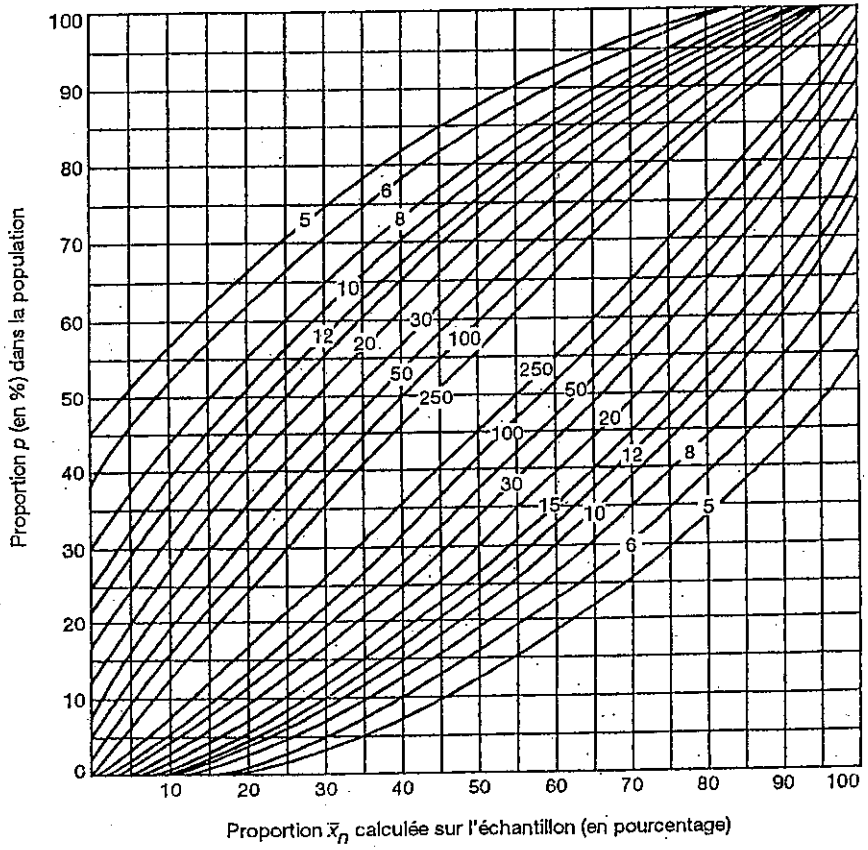




COURBES DES VALEURS NORMALES
DU DIAMETRE BIPARIETAL FŒTAL
EN FONCTION DU TERME.

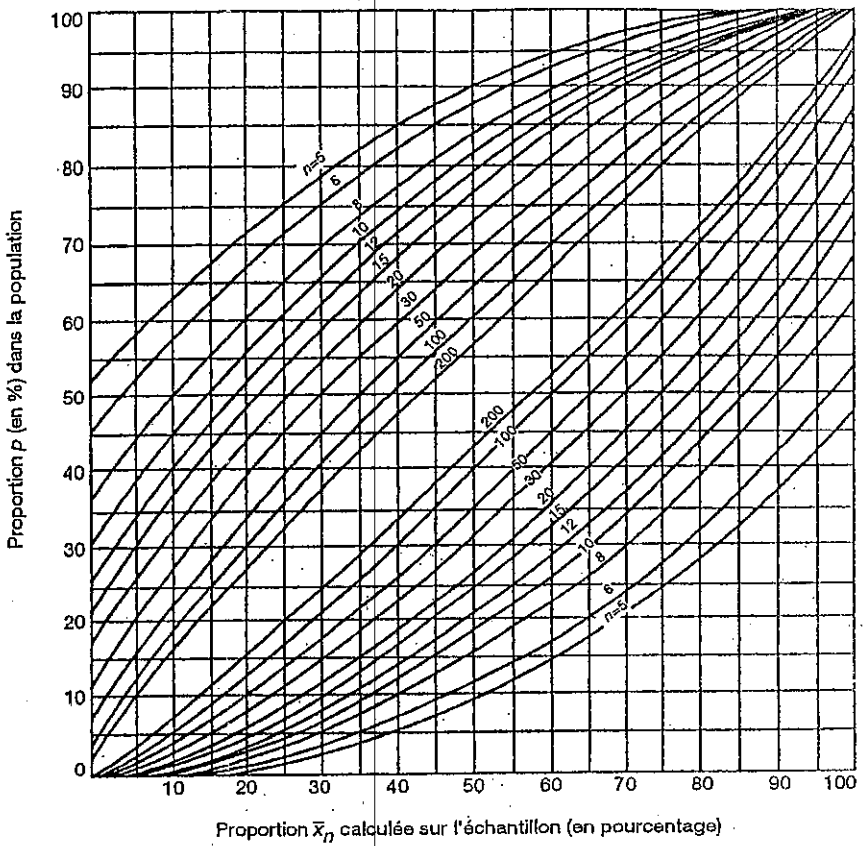
ABAQUE 1 : INTERVALLES DE CONFIANCE POUR UNE PROPORTION p

Intervalle bilatéral au niveau de confiance 0,90
et
Intervalles unilatéraux au niveau de confiance 0,95



ABAQUE 2 : INTERVALLES DE CONFIANCE POUR UNE PROPORTION p

Intervalle bilatéral au niveau de confiance 0,95
et
Intervalles unilatéraux au niveau de confiance 0,975



Lecture :

En fonction de la proportion empirique $\bar{x}_n$ (en abscisse) et de la taille (n) de l'échantillon dont on dispose, l'abaque indique, en ordonnée, les deux bornes d'un intervalle bilatéral au niveau 0,90 , ou la borne (inférieure ou supérieure) d'un intervalle unilatéral au niveau 0,95.

Exemple :

Sur un échantillon de taille $n = 15$, on observe une proportion empirique $\bar{x}_n = 60\%$.

- Les bornes de l'intervalle bilatéral sont 35% et 82% environ.
- Les intervalles unilatéraux correspondants sont [0;82%] et [35%;1].

Effectif de l'échantillon	Pourcentage observé									
	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %	45 %	50 %
10		0 – 45	1 – 50	3 – 56	5 – 60	7 – 65	9 – 70	12 – 74	15 – 78	19 – 81
20	0 – 25	1 – 32	3 – 38	6 – 44	9 – 49	12 – 54	15 – 59	19 – 64	23 – 68	27 – 73
30	0 – 20	2 – 27	5 – 33	8 – 39	11 – 44	15 – 49	19 – 54	23 – 59	27 – 64	31 – 69
40	1 – 17	3 – 24	6 – 30	9 – 36	13 – 41	17 – 47	21 – 52	25 – 57	29 – 62	34 – 66
50	1 – 15	3 – 22	6 – 28	10 – 34	14 – 39	18 – 45	22 – 50	26 – 55	31 – 60	36 – 64
60	1 – 14	4 – 21	7 – 27	11 – 32	15 – 38	19 – 43	23 – 48	28 – 53	32 – 58	37 – 63
70	1 – 13	4 – 20	8 – 26	11 – 31	15 – 37	20 – 42	24 – 47	28 – 52	33 – 57	38 – 62
80	1 – 12	4 – 19	8 – 25	12 – 30	16 – 36	20 – 41	25 – 46	29 – 52	34 – 57	39 – 61
90	2 – 12	5 – 18	8 – 24	12 – 30	16 – 35	21 – 41	25 – 46	30 – 51	34 – 56	39 – 61
100	2 – 11	5 – 18	9 – 24	13 – 29	17 – 35	21 – 40	26 – 45	30 – 50	35 – 55	40 – 60
150	2 – 10	6 – 16	10 – 22	14 – 27	18 – 33	23 – 38	27 – 43	32 – 48	37 – 53	42 – 58
200	2 – 9	6 – 15	10 – 21	15 – 26	19 – 32	24 – 37	28 – 42	33 – 47	38 – 52	43 – 57
500	3 – 7	8 – 13	12 – 18	17 – 24	21 – 29	26 – 34	31 – 39	36 – 44	41 – 49	46 – 54
1 000	4 – 7	8 – 12	13 – 17	18 – 23	22 – 28	27 – 33	32 – 38	37 – 43	42 – 48	47 – 53
2 000	4 – 6	9 – 11	13 – 17	18 – 22	23 – 27	28 – 32	33 – 37	38 – 42	43 – 47	48 – 52

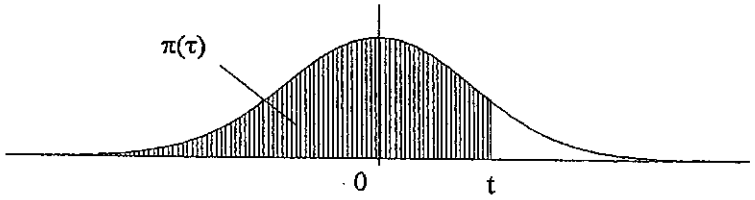
Table 1 : intervalle de confiance à 95 % d'un pourcentage
Exemple : sur 100 sujets, on a observé 10 cas positifs soit $p_o = 10$ %. La table indique que le pourcentage théorique est compris dans l'intervalle 5 % - 18 % (au risque 5 %).
Quand le pourcentage observé dépasse 50 %, on travaille sur le pourcentage complémentaire.

Effectif de l'échantillon	Pourcentage observé									
	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %	45 %	50 %
10		0 – 54	1 – 60	1 – 65	2 – 69	4 – 74	6 – 77	8 – 81	10 – 84	13 – 87
20	0 – 32	1 – 39	2 – 45	4 – 51	6 – 56	8 – 61	11 – 66	13 – 70	18 – 74	22 – 78
30	0 – 25	1 – 32	3 – 38	5 – 44	8 – 50	11 – 55	15 – 60	19 – 65	22 – 69	26 – 74
40	0 – 21	2 – 28	4 – 35	7 – 41	10 – 46	13 – 51	17 – 57	21 – 61	25 – 66	29 – 71
50	0 – 19	2 – 26	5 – 32	8 – 38	11 – 44	15 – 49	19 – 54	23 – 59	27 – 64	32 – 68
60	1 – 17	3 – 24	5 – 30	9 – 36	12 – 42	16 – 47	20 – 52	24 – 57	29 – 62	33 – 67
70	1 – 16	3 – 23	6 – 29	9 – 35	13 – 40	17 – 46	21 – 51	25 – 56	30 – 61	34 – 66
80	1 – 15	3 – 22	6 – 28	10 – 34	14 – 39	18 – 45	22 – 50	26 – 55	31 – 60	35 – 65
90	1 – 14	4 – 21	7 – 27	10 – 33	14 – 38	18 – 44	23 – 49	27 – 54	32 – 59	36 – 64
100	1 – 14	4 – 20	7 – 26	11 – 32	15 – 38	19 – 43	23 – 48	28 – 53	32 – 58	37 – 63
150	2 – 12	5 – 18	8 – 24	12 – 30	16 – 35	21 – 41	25 – 46	30 – 51	35 – 56	39 – 61
200	2 – 10	5 – 17	9 – 23	13 – 28	18 – 34	22 – 39	27 – 44	31 – 49	36 – 54	41 – 59
500	3 – 8	7 – 14	11 – 20	16 – 25	20 – 30	25 – 36	30 – 41	34 – 46	39 – 51	44 – 56
1 000	3 – 7	8 – 13	12 – 18	17 – 23	22 – 29	26 – 34	31 – 39	36 – 44	41 – 49	46 – 54
2 000	4 – 6	8 – 12	13 – 17	18 – 22	23 – 28	27 – 33	32 – 38	37 – 43	42 – 48	47 – 53

Table 2 : intervalle de confiance à 99 % d'un pourcentage
Exemple : sur 100 sujets, on a observé 10 cas positifs soit $p_o = 10$ %. La table indique que le pourcentage théorique est compris dans l'intervalle 4 % - 20 % (au risque 1 %).
Quand le pourcentage observé dépasse 50 %, on travaille sur le pourcentage complémentaire.

Loi normale

Le tableau donne les valeurs de $\pi(t) = P(X \leq t)$, si X suit la loi normale centrée réduite.



t	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Grandes valeurs de t

t	3	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	4	4,5
$\pi(t)$	0,998650	0,999032	0,999313	0,999517	0,999663	0,999767	0,999841	0,999892	0,999928	0,999968	0,999997

Partie C

Tests d'hypothèses, tests d'adéquation : savoirs cachés du programme de terminales S et ES

Introduction

Dans quelles situations utilise-t-on ce mot "test" ?
Que trouve-t-on dans les livres ?

a) Définitions du dictionnaire Larousse en 5 volumes :

- essai d'un produit, d'un appareil pour vérifier son action, son fonctionnement,

- toute circonstance qui permet d'éprouver, de mesurer quelque chose

- manoeuvre, épreuve, réaction biologique, sérologique ou chimique pratiquée à des fins diagnostiques

- test-match

- test statistique ou test d'hypothèse : méthode permettant, à partir d'observations d'un ou de plusieurs échantillons d'une population, d'accepter ou de rejeter, avec un certain risque d'erreur, une hypothèse portant sur la population ou sur la loi de probabilité choisie pour représenter celle-ci.

b) Dans un livre d'introduction à la statistique (Morgenthaler), en début de chapitre sur les tests avant la formalisation :

"Les tests statistiques sont utilisés pour répondre aux questions liées aux paramètres, par exemple pour savoir si oui ou non une espérance est plus petite qu'une certaine constante. Ces réponses doivent être apportées sur la base de mesures soumises à des erreurs aléatoires."

c) Dans un livre de la collection "statistique et probabilités appliquées", dont l'objectif est de rendre accessibles les fondements théoriques de la statistique à un public de niveau deug, ingénieur, chercheur appliqué (Lejeune), en introduction au chapitre "Tests d'hypothèses paramétriques" :

"Les tests constituent une approche décisionnelle de la statistique inférentielle. Un tel test a pour objet de décider sur la base d'un échantillon si une caractéristique de la population répond ou non à une certaine spécification que l'on appelle hypothèse. Ces spécifications peuvent avoir diverses provenances : normes imposées, affirmations faites par un tiers, valeurs cruciales de paramètres de modèles..."

C'étaient trois citations allant du plus commun au plus mathématique.

Prenons trois exemples de situations :
- exemple 1 : dans une étude de pollution atmosphérique, on s'intéresse à la quantité de particules de diamètre inférieur à 2.5 microns, dans la mesure où ces particules peuvent pénétrer nos poumons et provoquer une intoxication si la quantité respirée est supérieure à un certain seuil.
- exemple 2 : le responsable d'une chaîne de production de compteurs électriques veut savoir si les appareils produits sont bien calibrés, c'est à dire si la vitesse de rotation est la bonne (ceci est du contrôle de fabrication).
- exemple 3 : avant d'entrer dans un jeu de hasard, un joueur souhaite savoir si la pièce de monnaie utilisée est équilibrée.

I Elaboration d'une démarche mathématique pour proposer une réponse

Dans ces 3 situations, il s'agit donc de savoir si une caractéristique, une spécification est satisfaite.

Dans chacune, une observation (quantité de particules de diamètre inférieur à 2.5 microns mesurée dans un volume d'air donné à un endroit donné, vitesse de rotation du compteur électrique, côté de la pièce obtenu) a un résultat aléatoire.

On peut donc penser utiliser une modélisation probabiliste et les techniques en découlant pour trouver des réponses argumentées aux questions posées.

Il faut alors faire un travail de modélisation probabiliste de la situation, proposer un modèle où la probabilité dépend d'un paramètre inconnu et où l'hypothèse faite, la question posée, se traduit sur ce paramètre :

-exemple 1 : la quantité de particules mesurée est une variable aléatoire de moyenne m. La question est : a-t-on $m \leq Norme$?

-exemple 2 : la vitesse de rotation du compteur est une variable aléatoire de moyenne m. La question est : a-t-on $m = m_0$, où m_0 est la vitesse calibrée sur la chaîne de production ?

-exemple 3 : le résultat d'un lancer de pièce est pile ou face et la probabilité qu'elle a de tomber sur pile est inconnue valant p. L'hypothèse se traduit par : a-t-on $p = \frac{1}{2}$?

Comment prendre une décision ?

L'idée est de faire plusieurs observations indépendantes du même phénomène et de trouver une méthode pour juger si l'écart entre les données récoltées et ce qui est attendu quand l'hypothèse est satisfaite est significatif. Il faut donc **choisir** une **mesure de cet écart** : plus l'écart sera grand, moins l'hypothèse paraîtra vraisemblable. Un écart "grand" (cette appréciation est à préciser) conduira donc à la décision de rejeter l'hypothèse.

Mais il y a un bémol : il se pourrait que l'hypothèse soit satisfaite et que l'écart soit quand même grand. Cela arrive-t-il souvent ?

Il faut répondre à la question plus précise de **signification d'une grande valeur de l'écart**, et pour cela, évaluer la chance d'obtenir un grand écart quand cette hypothèse est satisfaite : il faut connaître la façon dont cet écart varie sous l'hypothèse, c'est à dire connaître la loi de probabilité de cet écart sous l'hypothèse testée.

Traitons l'exemple de la pièce :

- l'observation du lancer de la pièce est modélisée par une variable aléatoire X de loi de Bernoulli, de paramètre p : $P_p(X = 1) = p, P_p(X = 0) = 1 - p$. C'est le modèle probabiliste , ou plutôt statistique choisi.

- l'hypothèse : "la pièce est équilibrée" se ramène donc à l'hypothèse sur la valeur p de la probabilité qu'a la pièce de tomber sur pile : $p = \frac{1}{2}$.

- **on choisit** comme **observation** 8 lancers successifs indépendants de la pièce et on ne retient de l'observation que le nombre de fois où la pièce est tombée sur pile, soit S. A quelle valeur de S s'attend-on ? plutôt autour de 4.

On choisit pour mesurer **l'écart** entre l'observation et ce qui est attendu quand l'hypothèse est vraie le nombre $|S-4|$ et **on choisit** de dire que l'écart est grand s'il est supérieur ou égal à 3, c'est à dire si S vaut $0, 1, 7$ ou 8.

Par exemple, avec ces choix, l'observation de la valeur 1 pour S conduit à la décision : la pièce n'est pas équilibrée ! et l'observation de la valeur 4 pour S conduit à la décision : la pièce est équilibrée !

- mais ces décisions peuvent être erronées : quels sont les risques qu'elles le soient ?

La probabilité d'observer ces valeurs extrêmes de S sous l'hypothèse "$p = \frac{1}{2}$" se calcule facilement :

en effet, sous cette hypothèse, la variable aléatoire S suit la loi binomiale de paramètre 8 et $\frac{1}{2}$ et donc la probabilité qu'elle prenne ces 4 valeurs est $\frac{1}{2^8}(\binom{8}{0} + \binom{8}{1} + \binom{8}{7} + \binom{8}{8}) = \frac{9}{128} = 0.07$.

L'observation des valeurs $0, 1, 7, 8$ est peu probable si l'hypothèse "$p = \frac{1}{2}$" est vraie : il y a 7 chances sur 100 d'observer une de ces valeurs, alors qu'il y a 93 chances sur 100 d'observer une valeur comprise entre 2 et 6.

On peut se tromper en prenant la décision d'accepter l'équilibre après l'observation d'une valeur de S comprise entre 2 et 6 car la pièce peut ne pas être équilibrée ; si p est la probabilité qu'a la pièce de tomber sur pile, la probabilité de cette erreur de décision est $\sum_{k=2}^{6}\binom{8}{k}p^k(1-p)^{8-k}$, qu'on ne peut évaluer que si on se donne la valeur de p.

Par exemple, si $p = \frac{3}{4}$, cette probabilité d'erreur de décision vaut environ 0.63 : si la pièce a 3 chances sur 4 de tomber sur pile, il y a une probabilité de 63% d'accepter l'équilibre qand on observe entre 2 et 6 piles sur 8 lancers, c'est à dire de prendre une décision erronée.

Bilan du travail fait :

a) identification de la chose certaine inconnue, sur laquelle on pose une question.

b) modélisation de la situation par un modèle probabiliste connu mais de paramètres inconnus.

c) traduction de la question posée en a) dans le modèle.

d) choix d'un écart

e) choix des valeurs extrêmes de cet écart

f) calcul de l'écart pour l'expérience faite et prise de décision

g) calcul de la loi de probabilité de l'écart, évaluation des risques d'erreur des décisions prises et retour sur le choix des valeurs extrêmes

Il faut remarquer qu'avant l'expérience, il y a un **inconnu aléatoire** : c'est le résultat de l'expérience et il y a un **inconnu certain** : c'est la loi qui régit l'expérience (loi qui peut appartenir à une famille paramétrée simple comme la famille de lois de Bernoulli de paramètre p dans $[0, 1]$). Cet inconnu certain est toujours inconnu après l'expérience et on cherche à le cerner à partir du résultat de l'expérience.

Qu'est-ce qui aurait changé si on avait observé quatre lancers seulement ?

S est toujours le nombre de pile obtenus et l'écart est $|S-2|$, considéré anormalement grand s'il est supérieur à 2, c'est à dire si S vaut 0 ou 4.

L'erreur consistant à décider que la pièce n'est pas équilibrée alors qu'elle l'est a pour probabilité
$\frac{1}{2^4}(\binom{4}{0}+\binom{4}{4}) = 0.125$;

L'erreur consistant à décider que la pièce est équilibrée alors qu'elle a la probabilité $\frac{3}{4}$ de tomber sur pile a pour probabilité
$\binom{4}{1}(\frac{3}{4})^1(\frac{1}{4})^3 + \binom{4}{2}(\frac{3}{4})^2(\frac{1}{4})^2 + \binom{4}{3}(\frac{3}{4})^3(\frac{1}{4})^1 = 0.68$

Qu'est-ce qui aurait changé si on avait observé seize lancers ?

S est toujours le nombre de pile obtenus et l'écart est $|S-8|$, considéré anormalement grand s'il est supérieur ou égal à 5, c'est à dire si S vaut $0, 1, 2, 3$ ou $13, 14, 15, 16$.

Le premier risque vaut alors
$\frac{1}{2^{16}}(\binom{16}{0}+\binom{16}{1}+\binom{16}{2}+\binom{16}{3}+\binom{16}{13}+\binom{16}{14}+\binom{16}{15}+\binom{16}{16}) = 0.021$;

Le deuxième risque, si la probabilité de tomber sur pile est $\frac{3}{4}$, vaut
$\sum_{k=4}^{12}\binom{16}{k}(\frac{3}{4})^k(\frac{1}{4})^{16-k} = 0.595$.

On peut remarquer qu'avec les règles de décision choisies les risques de première et de deuxième espèce sont plus petits quand on a observé un plus grand nombre de lancers et que le risque de deuxième espèce est grand par rapport au risque de première espèce.

Que devient cette démarche dans les exemples 1 et 2?

Les étapes a), b) et c) sont déjà précisées. Pour l'étape d), on peut proposer comme écart la valeur absolue de la différence entre la moyenne empirique des observations (si on fait n mesures) et la moyenne théorique attendue, qui est *Norme* ou m_0. L'utilisateur, connaissant bien le champ d'application, peut donner des valeurs extrêmes liées à sa pratique ; sinon il faut réaliser le point g) pour les définir. Pour cette étape g), il est nécessaire de faire des hypothèses supplémentaires sur la loi de la variable observée pour aller plus loin dans les calculs !

II Exemple du graphologue

La formulation de cet exercice est celle qui est habituelle dans les cours de statistique. Où retrouve-t-on les points mis en évidence dans la partie précédente ?

1. Enoncé

Pour embaucher un graphologue, le chef du personnel d'une grosse entreprise envisage de faire un test. Il propose à chaque candidat d'identifier l'écriture d'un médecin et celle d'un avocat présentées par paire et cela pour 12 paires d'écritures. Le candidat sera embauché s'il identifie correctement l'écriture du médecin et celle de l'avocat pour au moins 9 paires.

(a) Déterminer la probabilité d'embaucher un incompétent en supposant que c'est un candidat qui répond au hasard. Cette hypothèse se traduit quantitativement par la phrase : la probabilité qu'il reconnaisse correctement une paire d'écritures donnée est $p = \frac{1}{2}$.

(b) Pensant que cette probabilité est trop forte, le chef du personnel change son critère et exige 10 bonnes réponses pour pouvoir embaucher un candidat.

Déterminer maintenant la probabilité d'embaucher un incompétent.

(c) Cette probabilité est trop faible. Le chef du personnel envisage donc de proposer 13 paires d'écritures et non plus 12.

Déterminer dans cette nouvelle situation la nombre minimal de bonnes réponses à exiger pour que la probabilité d'embaucher un incompétent soit inférieure à 0.05.

(d) Après s'être placé du point de vue du chef de personnel qui embauche, plaçons-nous du point de vue du chômeur qui cherche à être embauché : cette personne estime savoir reconnaître une paire d'écritures dans 85% des cas. Il sait qu'il sera embauché s'il fournit au moins 10 bonnes réponses sur les 13 paires proposées.

Déterminer la probabilité qu'il ne soit pas embauché, c'est à dire qu'il fournisse moins de 10 bonnes réponses.

2. Solutions à cet exercice de graphologue

(a) Suivons les étapes mises en évidence au paragraphe précédent :

Etape a) : elle est dite "le candidat est-il incompétent ?"

Etape b) : chaque réponse du candidat à l'identification d'une paire d'écritures est modélisée par une variable aléatoire de loi de Bernoulli de paramètre p (la valeur 1 correspondant à une réponse exacte) ; on suppose de plus que les réponses successives sont indépendantes et que les variables associées sont indépendantes de même loi.

Etape c) : la question se traduit par "p vaut-il 1/2 ?" ($p = \frac{1}{2}$ signifie que le candidat répond au hasard). C'est l'hypothèse, dite nulle, que l'on notre H_0.

Etapes d) et e) : le recruteur a donné sa règle de décision à partir du nombre S de réponses exactes du candidat : si ce nombre est supérieur au seuil 9, il décide que le candidat ne répond pas au hasard et décide donc de le recruter : il rejette l'hypothèse nulle.

Etape g) : Si H_0 est vraie, la variable aléatoire S, somme de 12 variables aléatoires de Bernoulli indépendantes de paramètre $\frac{1}{2}$ suit la loi binomiale de paramètres 12 et $\frac{1}{2}$.
C'est à ce niveau là qu'est posée la question 1) : la probabilité sous H_0 que S soit supérieur ou égal à 9 se calcule bien à partir du tableau suivant de la loi de S :

0	1	2	3	4	5	6	7	8	9	10	11	12
.000	.003	.016	.054	.121	.193	.226	.193	.121	.054	.016	.003	.000

On obtient : $\mathbf{P}_{0.5}(S \geq 9) \approx 0.073$. Le recruteur a à peu près 7 chances sur 100 de se tromper en affirmant qu'un candidat est compétent alors qu'en répondant au hasard celui-ci a donné au moins 9 réponses exactes sur 12 questions, donc il a 7 chances sur 100 de recruter un tel candidat.

(b) Seul le seuil change et la probabilité d'embaucher un incompétent devient : $\mathbf{P}_{0.5}(S \geq 10) \approx 0.0192$. Le risque devient sans doute acceptable pour le recruteur !

(c) Le recruteur change l'observation puisqu'il propose 13 paires d'écritures. S'il adopte une règle de décision du même type qu'avant, en notant S' le nombre de réponses exactes du candidat, il cherche le seuil $s_{0.05}$ tel que $\mathbf{P}_{0.5}(S' \geq s_{0.05})$ soit au plus égal à 0.05.
On utilise le tableau de la loi de S', qui suit la loi binomiale de paramètres 13 et $\frac{1}{2}$:

0	1	2	3	4	5	6	7	8	9	10	11	12	13
.000	.002	.009	.035	.087	.157	.209	.209	.157	.087	.035	.009	.002	.000

On trouve $\mathbf{P}_{0.5}(S \geq 10) = 0.046$.
Le recruteur choisit donc de ne pas qualifier un candidat d'incompétent s'il donne au moins 10 bonnes réponses sur 13 questions et donc accepte de le recruter.

(d) Le candidat sait qu'il a une probabilité de donner la bonne réponse dans 85% des cas, c'est à dire $p = 0.85$. Il sait qu'il sera refusé, à tort, s'il donne moins de 10 bonnes réponses. La probabilité pour que le recruteur prenne cette décision erronée, pénalisant le candidat, est $\mathbf{P}_{0.85}(S' < 10)$ et se calcule grâce au tableau suivant de la loi de S' :

0	1	2	3	4	5	6	7	8	9	10	11	12	13
.0000	.000	.000	.000	.000	.000	.001	.006	.027	.084	.190	.294	.277	.121

On trouve $\mathbf{P}_{0.85}(S' < 10) = 0.118$. Le risque n'est pas très grand.

III Un peu de formalisation et de vocabulaire, sur des savoirs cachés de Terminale

Un test de validité d'hypothèse est un mécanisme permettant de confirmer ou d'infirmer une hypothèse par l'observation d'un échantillon. C'est une fonction définie de l'ensemble des valeurs possibles de l'observation vers l'ensemble des décisions qui sont d_0 = "accepter l'hypothèse" et d_1 = "refuser l'hypothèse". C'est une variable aléatoire.

1) **Formulation des hypothèses :**

- L'hypothèse testée, ou hypothèse nulle (notée H_0) est celle que l'observateur croit vraie et qu'il ne rejettera que si elle est contradictoire avec les résultats de l'expérience. En général, elle se traduit par une égalité ou des inégalités portant sur un paramètre du modèle de la situation étudiée.

Elle correspond à une hypothèse de prudence . Par exemples :

pour le chef de personnel, embaucher quelqu'un qui répond au hasard n'est pas pertinent et cela ne doit pas arriver ou alors extrêmement rarement : H_0 correspond donc à "le candidat est incompétent" ;

pour un chef de production, affirmer que sa production est mauvaise si elle est bonne le conduit à engager des frais inutiles : H_0 correspond donc à "la chaîne de montage est bien réglée" ;

pour le joueur qui s'engage dans des paris, affirmer que la pièce a plus de chance de tomber sur pile que sur face est très risqué et peut le conduire à la ruine : H_0 correspond donc à "la pièce est équilibrée",...

- L'hypothèse alternative (notée H_1) est une hypothèse estimée a priori peu probable mais possible. On parle d'alternative simple si elle s'exprime par une égalité sur un paramètre, hypothèse multiple sinon.

Quand H_1 n'est pas le contraire de H_0, le test est dit non contradictoire.

On parle de test unilatéral si H_0 est de la forme $\theta \leq \theta_0$ et H_1 de la forme $\theta > \theta_0$, de test bilatéral si H_0 est de la forme $\theta = \theta_0$ et H_1 de la forme $\theta \neq \theta_0$.

2) **Statistique du test :**

C'est une variable aléatoire (fonction de l'observation ou de l'échantillon observé) à partir de laquelle est définie la règle de décision. Cela correspond à l'écart de la partie I, la règle de décision étant de ne pas rejeter H_0 si l'écart est inférieur à un seuil et de la rejeter sinon.

Dans l'exemple 3 traité, c'est la variable S, nombre de pile observé dans la série de 8 lancers.

Dès qu'on connaît le paramètre du modèle, on connaît la loi de cette statistique : dans l'exemple, S suit la loi binomiale de paramètres 8 et p si p est la probabilité qu'a la pièce de tomber sur pile.

3) **Règle de décision, région de rejet de H_0** (ou région critique du test)

Une règle de décision associée à une statistique de test T consiste à rejeter H_0 si T est supérieure à un seuil s : l'ensemble des échantillons tels que T vérifie cette condition est appelée région de rejet ou région critique.

Dans l'exemple du graphologue, H_0 est "$p = \frac{1}{2}$" où p représente la probabilité du candidat à reconnaître une paire d'écritures et où cette valeur $\frac{1}{2}$ correspond à une incompétence ; la statistique du test est le nombre N de paires reconnues et la région de rejet est l'ensemble des observations telles que N soit supérieur ou égal à 9.

4) **Deux types d'erreur :**

Toute décision peut être erronée :

- prendre la décision d_1, c'est à dire rejeter H_0 alors que H_0 est vraie : c'est l'erreur de première espèce

- prendre la décision d_0, c'est à dire accepter H_0 alors que H_1 est vraie : c'est l'erreur de deuxième espèce.

5) **Risques, niveau de signification :**

Ces erreurs sont aléatoires et on veut mesurer la probabilité qu'elles arrivent. Ce sont les risques d'erreurs :

- risque de première espèce, qui mesure la probabilité de commettre l'erreur de première espèce, et qui vaut dans le cas où H_0 est une hypothèse simple $E_{H_0}(d_1)$, ou dans l'exemple du graphologue $P_{0.5}(N \geq 9)$. On appelle niveau du test ou niveau de signification la valeur de ce risque.

- risque de deuxième espèce, qui mesure la probabilité de commettre l'erreur de deuxième espèce. Dans le cas du graphologue et du candidat qui a une probabilité 0.85 de reconnaître les écritures, accepter H_0 alors que H_0 est fausse (c'est à dire ne pas recruter le candidat alors qu'il est compétent) a pour probabilité $P_{0.85}(N < 9)$.

6) **Détermination de la région de rejet à partir d'un niveau de signification fixé :**

L'idéal est de minimiser ces deux risques simultanément, ce qui est impossible. On choisit donc de contrôler le risque de première espèce. On se donne une valeur α "petite".

On décide de construire une règle de décision à partir d'un écart T de la façon suivante :
si T est supérieur à s_α, on rejette H_0.
Le seuil s_α est déterminé par la contrainte $P_{H_0}(T > s_\alpha) = \alpha$.

Après l'expérience, on calcule la valeur de T, soit T_{obs} et on la compare à ce seuil s_α : si T_{obs} est supérieur à s_α, on rejette l'hypothèse et si T_{obs} est inférieur à s_α, on ne peut pas la rejeter au niveau de signification donné α.

7) **Tableau récapitulatif des risques :**

Le risque de première espèce est noté α, le risque de deuxième espèce est noté $\gamma = 1 - \beta$. On cherchera, le premier fixé, à minimiser le second ou encore à maximiser son complément à 1 (qui représente la probabilité de rejeter H_0 quand elle est fausse), soit maximiser β : ce nombre s'appelle la puissance du test.

	H_0 est vraie	H_0 est fausse
acceptation de H_0	$1 - \alpha$	γ
rejet de H_0	α	$\beta = 1 - \gamma$

Le niveau du test α, comme la statistique T, est au choix de l'utilisateur. Le problème du statisticien est alors de proposer des règles de décision donnant un risque de seconde espèce le moins élevé possible pour un risque de première espèce fixé.

V Exemple de "la bouteille"

1) **Rappel de la situation** (inspirée de l'expérience de Brousseau faite en CM2, dans les années 75)

   Une bouteille opaque contient 5 boules de deux couleurs différentes, noir et blanc. Comment savoir ce qu'elle contient, sans l'ouvrir bien sûr ?

   Il y a donc quatre compositions possibles : 1N4B, 2N3B, 3N4B, 4N1B et donc quatre valeurs possibles de la proportion p de boules noires : $0.2, 0.4, 0.6, 0.8$.

   L'idée est de retourner un certain nombre de fois cette bouteille pour voir une seule boule apparaître au niveau du goulot et de décider de la proportion en fonction de la proportion observée de boules noires.

2) **Les étapes :**

   Etape a : la chose inconnue est p.

   Etape b : on modélise chaque issue d'un retournement par une variable de Bernoulli, valant 1 si c'est noir qui apparaît et 0 sinon. On suppose que les retournements sont indépendants. Le paramètre de cette loi est p.

   Etape c : on veut savoir si p vaut 0.2. On teste donc l'hypothèse $H_0 : p = 0.2$. L'alternative est multiple et s'écrit : $H_1 = (p \in \{0.4, 0.6, 0.8\})$

   Etape d : on choisit l'écart le plus naturel : $|F_n - 0.2|$ où F_n représente la fréquence empirique d'apparition de noir. nF_n suit une loi binomiale de paramètres le nombre n de retournements et p.

3) **Choix de la règle de décision a priori :** on rejette H_0 si la fréquence empirique est en dehors de l'intervalle $]0.1, 0.3]$.

   Calculs des risques des deux espèces :

   a) avec un échantillon de taille n=10 :

      a) Risque de première espèce = probabilité que la fréquence F_{10} soit en dehors de $]0.1, 0.3]$ avec $p = 0.2$: $\alpha_{10} = 0.4966835$
      b) Risques de deuxième espèce = probabilités que la fréquence F_{10} soit dans $]0.1, 0.3]$ avec $p = 0.4 ou 0.6 ou 0.8$

$$\gamma_{10,0.4} = 0.3359232, \quad \gamma_{10,0.6} = 0.05308416, \quad \gamma_{10,0.8} = 0.00086016$$

      Le risque de première espèce n'est pas acceptable, ni celui de seconde espèce pour $p = 0.4$.

   b) avec un échantillon de taille n=100 :

      a) Risque de première espèce = probabilité que la fréquence F_{100} soit en dehors de $]0.1, 0.3]$ avec $p = 0.2$: $\alpha_{100} = 0.01175572$

b) Risques de deuxième espèce = probabilité que la fréquence F_{100} soit dans $]0.1, 0.3]$ avec $p = 0.4$ ou 0.6 ou 0.8

$$\gamma_{100,0.4} = 0.02478282, \quad \gamma_{100,0.6} = 1.251478\ 10^{-9}, \quad \gamma_{100,0.8} = 4.796711\ 10^{-27}$$

Tous les risques sont acceptables, inférieurs à 5%, niveau de signification habituellement utilisé.

4) **Choix de la régle de décision en fonction d'un risque de première espèce donné**, risque acceptable de rejet à tort de l'hypothèse $p = 0.2$:

On choisit la région de rejet de la même forme qu'en 1),soit de la forme $(F_n \notin]0.2 - k, 0.2 + k])$, mais avec des bornes à déterminer pour que le risque d'erreur de première espèce soit inférieur à 0.05 :

a) avec un échantillon de taille n=10 :

a) Risque de l'erreur de première espèce valant 0.05 au plus :
on cherche donc le plus petit c tel que $\mathbf{P}_{0.2}(Bin(10, 0.2) \notin [2 + c, 2 + c]) \leq 0.05$.
Avec la distribution donnée par le logiciel "R", on trouve : $c = 2.5$ et la région de rejet est $(F_{10} \notin]0.05, 0.45])$, le niveau exact est : $\alpha_{10} = 0.032793498$.

b) Risques des erreurs de deuxième espèce : c'est donc la probabilité d'accepter $p = 0.2$ alors que le vrai p est 0.4 ou 0.6 ou 0.8.

$$\gamma_{10,0.4} = 0.6270566, \quad \gamma_{10,0.6} = 0.1661338, \quad \gamma_{10,0.8} = 0.00636928$$

On remarque que le risque de deuxième espèce est très grand pour $p = 0.4$, valeur de p la plus proche de celle de l'hypothèse nulle, à peine acceptable pour $p = 0.6$ et acceptable pour $p = 0.8$.

b) avec un échantillon de taille 100 :

a) Risque de l'erreur de première espèce valant 0.05 au plus :
On cherche le plus petit c tel que $\mathbf{P}_{0.2}(Bin(100, 0.2) \notin [20 - c, 20 + c]) \leq 0.05$.
Avec la distribution donnée par le logiciel "R", on trouve : $c = 9$ et la région de rejet est $F_{100} \notin]0.11, 0.29]$, le niveau exact est $\alpha_{100} = 0.04534896$.

b) Risques d'erreurs de deuxième espèce : c'est donc la probabilité d'accepter $p = 0.2$ alors que le vrai p est 0.4 ou 0.6 ou 0.8.

$$\gamma_{100,0.4} = 0.01477532, \quad \gamma_{100,0.6} = 3.464215\ 10^{-10}, \quad \gamma_{100,0.8} = 5.039385\ 10^{-28}$$

On remarque que tous ces risques sont acceptables, inférieurs au seuil 0.05 accepté pour celui de première espèce.

Comme dans la solution précédente (choix a priori), on remarque que plus p est éloigné de 0.2, plus le risque de deuxième espèce est faible.

IV Tests d'adéquation, tests du χ^2

1) Problèmatique spécifique de l'adéquation :

Face à une situation expérimentale où les résultats issus de répétitions indépendantes d'une même épreuve sont aléatoires, on pose la question suivante : les observations faites sont-elles compatibles avec l'hypothèse que la loi régissant l'épreuve vérifie une caractéristique donnée.

Par exemple, dans le cas du joueur et de la pièce, l'observation de 8 lancers, donnant des fréquences empiriques d'obtention de pile et de face, correspond-elle à la probabilité théorique d'obtention de pile et de face d'une pièce équilibrée ?

Exemple encore plus pratique : il y a eu 1 pile et 7 faces sur 8 lancers ; peut-on dire qu'avec cette pièce il y a une chance sur deux d'avoir pile et une chance sur deux d'avoir face ? Il faut comparer les deux lois sur $\{0,1\}$: la loi empirique $\{\frac{7}{8},\frac{1}{8}\}$ et la loi théorique $\{\frac{1}{2},\frac{1}{2}\}$.

Dans la liste des étapes mises en évidence en partie I, les points a,b et c sont déjà mis en place et en particulier la modélisation probabiliste est faite implicitement.

Il faut par contre choisir un écart : on choisit classiquement la distance dite du χ^2, que j'expliciterai un peu plus loin de manière générale.

Sur l'exemple de la pièce, cela donne l'écart suivant : $T = \frac{(S-4)^2}{4} + \frac{(8-S-4)^2}{4}$, qui, en gros, somme les carrés des écarts entre les effectifs observés et les effectifs théoriques si l'hypothèse est vraie.

Il faut ensuite décider quand les valeurs de cet écart sont grandes. Soit on le décide arbitrairement, soit on essaie de limiter le risque de première espèce.

Remarque : Dans le cas de la pièce, $T = \frac{(S-4)^2}{2}$ est le carré de l'écart choisi en partie I. Tester l'adéquation de la loi observée à la loi $\{\frac{1}{2},\frac{1}{2}\}$ ou tester l'égalité à $\frac{1}{2}$ du paramètre du modèle choisi en I revient au même. C'est un changement de vocabulaire seulement.

2) Exemple du dé pipé ou non :

Un joueur utilise un dé et se demande s'il est pipé ou non. Pour répondre à cette question, il fait un grand nombre de lancers et au vu des résultats obtenus, il doit prendre une décision : le dé est pipé ou le dé est non pipé.

Suivons la démarche vue en I :

-l'étape a) est dite : "le dé est-il équilibré ?".

-l'étape b) consiste à modéliser le résultat d'un lancer de dé par une variable aléatoire prenant ses valeurs dans $\{1,2,...,6\}$ et de loi $(p_1,p_2,...,p_6)$ inconnue (le paramètre $p = (p_1,p_2,...,p_6)$ est un vecteur de $\mathbf{R^6}$ de coordonnées positives de somme 1).

-l'étape c) consiste à interpréter le fait que le dé soit non pipé en l'égalité à $(\frac{1}{6},\frac{1}{6},...,\frac{1}{6})$ du paramètre p. L'hypothèse H_0 s'écrit : $p = (\frac{1}{6},\frac{1}{6},...,\frac{1}{6})$.

- l'étape d) utilise la répartition empirique observée après n lancers de ce dé. Si on note N_j le nombre de fois où la face j est sortie, cette répartition empirique est $(\frac{N_1}{n},\frac{N_2}{n},...,\frac{N_6}{n}) =$

$(F_{1,n}, F_{2,n}, ..., F_{6,n})$. C'est un vecteur aléatoire dont on observe une valeur quand on fait réellement les lancers de dés et dont la loi dépend du paramètre p.

On choisit comme écart :

$$Ecart_n = \sum_{j=1}^{6} \frac{(N_j - \frac{1}{6}n)^2}{\frac{1}{6}n} = 6n \sum_{j=1}^{6} (F_{j,n} - \frac{1}{6})^2$$

Reste à préciser quand cet écart est grand : on décide de lier ce seuil au risque de première espèce et donc de le déterminer à l'aide de la loi de l'écart quand l'hypothèse H_0 est vraie. Pour connaître cette loi, soit on procède comme ce qui est proposé en terminale, soit on connaît un peu plus de statistique théorique.

3) **Ce que la théorie dit :**

C'est compliqué et utilise la notion de convergence en loi (un peu comme le théorème central limite qui affirme que la suite de variables aléatoires $Y_n = \sqrt{n}\frac{\bar{X}_n - m}{\sigma}$ converge vers une variable aléatoire normale centrée réduite) :

La variable $Ecart_n$ dépend du nombre n de lancers de dés faits. Si le paramètre p vaut $(\frac{1}{6}, \frac{1}{6}, ..., \frac{1}{6})$, la suite de variables aléatoires $Ecart_n$ converge en loi vers une variable aléatoire dont la loi est la loi du χ^2 à 5 degrés de liberté . C'est à dire que dans de bonnes conditions, on peut approcher la loi de $Ecart_n$ par celle du χ^2 à 5 degrés de liberté, loi qui est tabulée. La lecture de la table permet de déterminer le seuil s_α tel que la probabilité que l'écart dépasse ce seuil soit α. Par exemple, si $\alpha = 10\%$, on lit $s_{0.10} = 9.236$ et si $\alpha = 5\%$, $s_{0.05} = 11.070$.

La règle de décision sera donc : au niveau de signification 10%, on rejette l'équilibre du dé si on observe un écart supérieur à 9.236.

4) **Ce qu'on propose en terminale :**

Ce qui précède est hors de portée des élèves de terminale.

Par contre, ils savent utiliser les méthodes de simulation.

Le programme propose donc de remplacer la loi théorique de l'écart par la loi empirique obtenue à partir de la simulation d'un grand nombre N d'écarts. Chaque écart est calculé à partir de la simulation de n lancers de dés équilibrés.

Cette dernière série de N valeurs observées a une distribution de fréquence dont on peut caractériser les paramètres de tendance centrale et de dispersion, en particulier les quantiles. On peut déterminer ainsi le nombre réel $s_{0.10}$, 9-ème décile tel que 90% de la série observée d'écarts soit inférieure à ce seuil $s_{0.10}$ ou encore le réel $s_{0.05}$, 95-ème centile, tel que 95% de cette série soit inférieure à ce seuil.

Et on poursuit le raisonnement : si le dé du joueur est équilibré, la valeur de l'écart peut être considérée comme un élément supplémentaire à la série simulée et il y a une probabilité de 90% qu'elle soit inférieure au seuil $s_{0.10}$ que l'on vient de déterminer.

La règle de décision que l'on adoptera sera la suivante : si la valeur observée de l'écart est supérieure au seuil $s_{0.10}$, on affirme que le dé est pipé ; sinon, on on ne peut pas rejeter le fait qu'il soit équilibré.

Ces deux décisions peuvent être entachées d'erreur : la probabilité d'affirmer que le dé est pipé alors qu'il ne l'est pas est de 10% si le seuil est $s_{0.10}$, 5% si le seuil est $s_{0.05}$. C'est le risque de première espèce. On ne dit rien sur le risque de deuxième espèce, qui n'est pas calculé.

On ne parle pas clairement en terminale des risques pris quand on agit de la sorte. Il faudrait dire par exemple : "au niveau de signification 10%, on rejette l'équilibre du dé" ou "on ne peut rejeter l'équilibre du dé".

5) **Exemple du test d'indépendance** (adéquation avec estimation de paramètres):

A partir d'un exemple : lors d'une enquête sociologique portant sur les étudiants de l' université XXX, on se demande s'il existe un lien entre le mode de logement et le sexe d'un étudiant. On veut tester l'indépendance entre ces deux caractères décrivant un étudiant.

L'enquête n'est pas exhaustive ; on considère d'une part que chaque étudiant interrogé fournit une observation du couple "(mode de logement,sexe)" $= (X, Y)$, observation au résultat aléatoire et d'autre part que ces étudiants ont des comportements indépendants entre eux. La loi de probabilité de ce couple est inconnue et on traduit la question posée dans cette modélisation probabiliste : la loi du couple (X, Y) est-elle le produit des lois marginales de X et de Y ? c'est cela l'hypothèse H_0.

ou encore, y-a-t-il adéquation entre la répartition empirique observée du couple (X, Y) et la répartition théorique s'il y avait indépendance ?

L'idée est d'utiliser la démarche du paragraphe précédent. Mais se pose le problème du calcul des effectifs théoriques, puisqu'on ne connait pas les lois de X et de Y sous l'hypothèse nulle.

On va alors les estimer par les lois marginales empiriques et choisir comme écart entre la distribution observée du couple (X, Y) et sa distribution théorique estimée sous H_0 la variable suivante :

$$T = \sum_{i\in\{1,2\},j\in\{1,..,4\}} \frac{(N_{ij} - n\frac{N_{i.}}{n}\frac{N_{.j}}{n})^2}{n\frac{N_{i.}}{n}\frac{N_{.j}}{n}}$$

où sur les n étudiants interrogés, N_{ij} est le nombre d'étudiants ayant répondu la modalité i à la variable sexe et la modalité j à la variable "mode de logement", $N_{i.}$ le nombre d'étudiants ayant répondu la modalité i à "sexe" et $N_{.j}$ celui des étudiants ayant répondu j à "mode de logement".

On montre que, si l'hypothèse H_0 est vraie, cette variable aléatoire T converge en loi quand n tend vers l'infini vers une variable aléatoire de loi du χ^2 à $(2-1)(4-1) = 3$ degrés de liberté. On peut alors définir une règle de décision dont le risque de première espèce a une valeur donnée.

Exemple numérique :

Les modes de logement repérés sont : seul, en famille, en couple, autres et le sexe a pour modalité : masculin, féminin.

	seul	en famille	en couple	autres	tout mode
féminin	12	11	14	12	49
masculin	15	6	9	14	44
tout sexe	27	17	23	26	93

Tableau des effectifs théoriques estimés sous l'hypothèse d'indépendance:

	seul	en famille	en couple	autres	tout mode
féminin	14.2	9	12.1	13.7	49
masculin	12.8	8	10.9	12.3	44
tout sexe	27	17	23	26	93

Calcul de l'écart :

$$T_{obs} = \frac{2.2^2}{14.2} + \frac{2^2}{9} + \frac{1.9^2}{12.1} + \frac{1.7^2}{13.7} + \frac{2.2^2}{12.8} + \frac{2^2}{8} + \frac{1.9^2}{10.9} + \frac{1.7^2}{12.3}$$

On trouve $T_{obs} = 2.74$.

D'après la théorie, ce T suit (asymptotiquement) la loi du χ^2 à 3 ddl et la probabilité que T dépasse la valeur 7.8 est de 5%, celle qu'il dépasse 6.2 est de 10%.

T_{obs} est inférieur à ces deux seuils. L'observation permet de prendre la décision suivante : au niveau de signification 10%, on ne peut rejeter l'indépendance entre le mode de logement et le sexe .

Conclusion générale :
Mise en place d'une procédure pour prendre une décision dans une situation aléatoire.

a) identification de la chose certaine inconnue, sur laquelle on pose une question.

b) modélisation de la situation par un modèle probabiliste connu mais de paramètres inconnus.

c) traduction de la question posée en a) dans le modèle choisi.

d) choix d'un écart entre ce qui est observé aléatoire et ce qui est attendu sous l'hypothèse.

e) choix des valeurs extrêmes de cet écart.

f) calcul de l'écart pour l'expérience faite et prise de décision.

g) calcul de la loi de probabilité de l'écart, évaluation des risques d'erreur des décisions prises et retour sur le choix des valeurs extrêmes.

Il y a une relation entre niveau de risque accepté ou acceptable et grande valeur de l'écart : c'est le praticien (chef du personnel, chef de production, joueur, sociologue, décideur en général...) qui prend la décision et choisit en fin de compte le risque acceptable.

Partie D

Loi binomiale, courbe en cloche et tableur (B.Parzysz mars 07)

Commençons par rappeler que, étant donné un entier non nul n et un réel p compris entre 0 et 1, on dit qu'une variable aléatoire X_n suit la loi binomiale de paramètres n et p, notée $B(n;p)$, lorsqu'elle prend les valeurs entières k comprises entre 0 et n avec les probabilités $p_k = \binom{n}{k} p^k (1-p)^{n-k}$. Une telle variable a pour espérance $m = np$ et pour variance $\sigma^2 = np(1-p)$.

Les tableurs-grapheurs fournissent à la demande un diagramme en bâtons de cette distribution, et on peut alors constater que, pour n "assez grand" (comme on dit), les sommets des dits bâtons dessinent une "courbe en cloche" assez régulière. Cette notion de "courbe en cloche" est habituellement associée à la densité de la loi normale, mais :

- d'une part, il ne suffit pas que la représentation graphique d'une fonction infiniment dérivable sur $\mathbf{R}$ ait une allure de cloche pour que cette fonction soit la densité d'une variable distribuée normalement
- et, d'autre part, la loi binomiale est une loi discrète tandis que la loi normale est une loi "à densité".

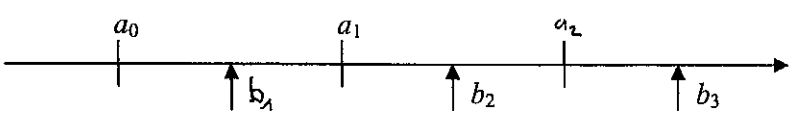
Ce qui conduit à se poser la question suivante : au-delà de l'opposition discret / continu, et comme le suggère la représentation graphique, la distribution binomiale a-t-elle quelque chose à voir avec la loi normale ?

La réponse théorique se trouve bien sûr dans le fameux théorème central limite, qui affirme en particulier que, étant donné une variable X_n de loi binomiale $B(n;p)$, la variable centrée réduite associée, soit $Y_n = \frac{X_n - np}{\sqrt{np(1-p)}}$, converge en loi vers la loi normale centrée réduite lorsque n tend vers l'infini. Mais le tableur-grapheur permet, de façon plus concrète, d'une part de visualiser cette convergence, et d'autre part d'estimer sa " qualité ".

Prenons à titre d'exemple le cas $n = 100$ et $p = 0,2$ (d'où $m = 20$ et $\sigma = 4$).
Une variable X_{100} de loi $B(100;0,2)$ prend toute valeur entière k comprise entre 0 et 100 avec la probabilité p_k indiquée plus haut. La variable centrée réduite Y_{100} associée à X_{100} prend alors les valeurs $a_k = \frac{k-20}{4}$ avec les mêmes probabilités p_k.

Pour comparer le diagramme en bâtons de la distribution de probabilité de Y_{100} avec la densité d'une variable aléatoire N de loi normale centrée réduite, nous allons "discrétiser" la densité de celle-ci de la façon suivante :

Les 101 valeurs a_k sont régulièrement espacées entre $a_0 = -5$ et $a_{100} = 20$, avec un pas de 0,25. Notons b_k le milieu de l'intervalle $[a_{k-1}; a_k]$ et cherchons, pour chacune des variables Y_{100} et N, la probabilité qu'elle a de prendre les valeurs de l'intervalle $[b_k; b_{k+1}[$.



Remarque : pour recouvrir $\mathbf{R}$ *tout entier, posons en outre* $b_0 = -\infty$ *et* $b_{101} = +\infty$.

Soit maintenant k un entier compris entre 0 et 100 :
- pour Y_{100}, on a $P(b_k \leq Y_{100} < b_{k+1}) = p_k$,
- pour N, on a $P(b_k \leq N < b_{k+1}) = F(b_{k+1}) - F(b_k)$, où F est la fonction de répartition de la loi normale centrée réduite.

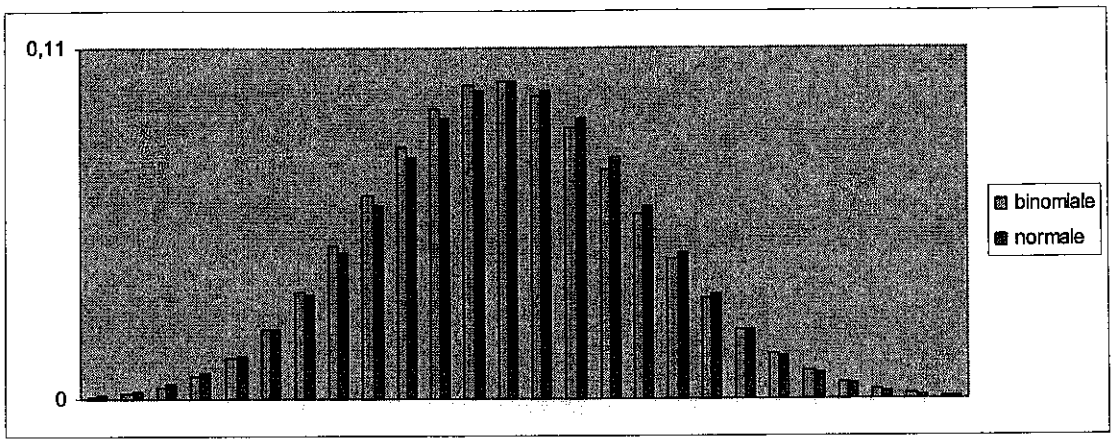
Il s'agit donc de comparer ces deux probabilités pour toutes les valeurs de k.

Remarque : la fonction $k \longrightarrow p_k$ n'est autre que la distribution de probabilité de la variable binomiale X_{100}, tandis que la fonction $k \longrightarrow F(b_{k+1}) - F(b_k)$ est la distribution de probabilité de la variable $4N + 20$, où N est la variable normale centrée réduite discrétisée comme indiqué ci-dessus.

Du fait que la distribution de probabilité de la loi binomiale, ainsi que la fonction de répartition de la loi normale, font partie des fonctions statistiques du logiciel, le tableur nous fournit rapidement les valeurs souhaitées (voir le tableau ci-dessous, qui a été tronqué en ne conservant que les valeurs de k pour lesquelles la probabilité n'est pas "négligeable")

k	a_k	b_k	binomiale	normale	% écart
8	-3	-3,125	0,0006	0,0011	48,9
9	-2,75	-2,875	0,0015	0,0023	36,1
10	-2,5	-2,625	0,0034	0,0044	24,3
11	-2,25	-2,375	0,0069	0,008	14,2
12	-2	-2,125	0,0128	0,0136	6,24
13	-1,75	-1,875	0,0216	0,0217	0,47
14	-1,5	-1,625	0,0335	0,0325	3,22
15	-1,25	-1,375	0,0481	0,0457	5,1
16	-1	-1,125	0,0638	0,0605	5,52
17	-0,75	-0,875	0,0789	0,0752	4,86
18	-0,5	-0,625	0,0909	0,0878	3,48
19	-0,25	-0,375	0,0981	0,0964	1,7
20	0	-0,125	0,0993	0,0995	0,18
21	0,25	0,125	0,0946	0,0964	1,93
22	0,5	0,375	0,0849	0,0878	3,35
23	0,75	0,625	0,072	0,0752	4,28
24	1	0,875	0,0577	0,0605	4,56
25	1,25	1,125	0,0439	0,0457	4,05
26	1,5	1,375	0,0316	0,0325	2,59
27	1,75	1,625	0,0217	0,0217	0,02
28	2	1,875	0,0141	0,0136	3,88
29	2,25	2,125	0,0088	0,008	9,38
30	2,5	2,375	0,0052	0,0044	16,8
31	2,75	2,625	0,0029	0,0023	26,7
32	3	2,875	0,0016	0,0011	39,6
33	3,25	3,125	0,0008	0,0005	56,5

Le pourcentage de l'écart relatif entre les deux probabilités (colonne de droite du tableau) montre qu'il est assez faible pour les valeurs de k situées autour de 20, c'est-à-dire de l'espérance de X_{100}, et correspond donc aux valeurs les plus probables de la loi binomiale. Par contre, pour les valeurs plus éloignées, il devient vite important. Mais, comme ces valeurs ont une faible probabilité, l'impact visuel n'est pas très perceptible, comme le montre la juxtaposition des deux diagrammes en bâtons (voir graphique 1 ci-dessous).



Graphique 1 : $n = 100$ et $p = 0.2$

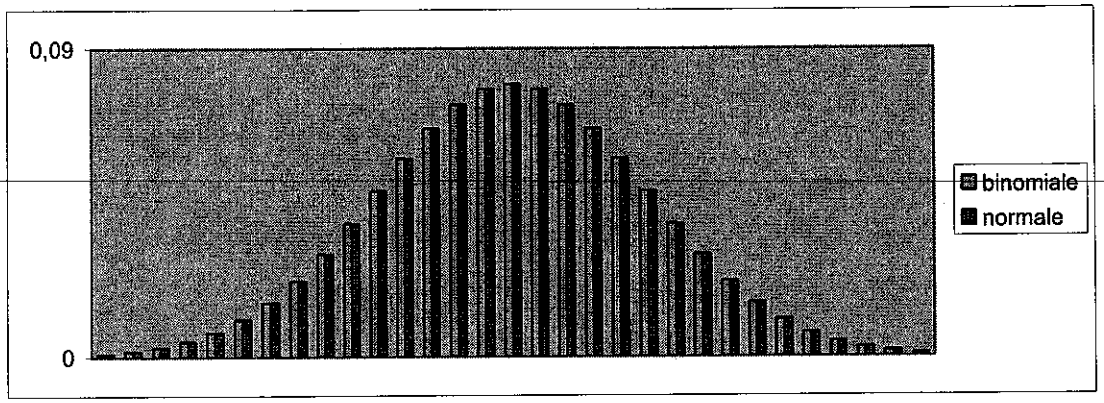
Au-delà de cette ressemblance visuelle, on peut également chercher à se faire une idée de l'influence des deux paramètres n et p sur la "proximité visuelle" des deux diagrammes.

1. **Cas n = 100 et p = 0,5.**

   On a cette fois $m = 50$ et $\sigma = 5$. Le tableau tronqué ci-dessous montre que les écarts relatifs sont nettement plus faibles que lorsque $p = 0,2$: on a une plage de 31 valeurs autour de 50 dans laquelle l'écart relatif est inférieur à 4% (contre 5 valeurs lorsque $p = 0,2$). Le graphique 2 confirme cette plus grande "proximité" des deux diagrammes.

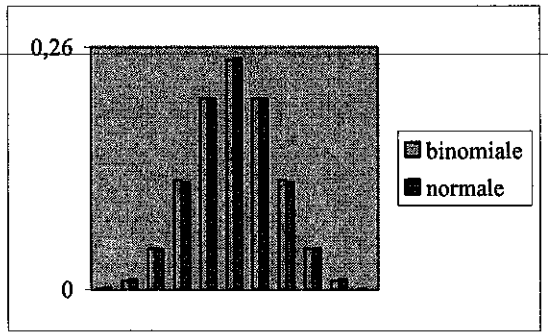
k	a_k	b_k	binomiale	normale	% écart
35	-3	-3,1	0,0009	0,0009	3,82
36	-2,8	-2,9	0,0016	0,0016	2,59
37	-2,6	-2,7	0,0027	0,0027	1,63
38	-2,4	-2,5	0,0045	0,0045	0,92
39	-2,2	-2,3	0,0071	0,0071	0,41
40	-2	-2,1	0,0108	0,0109	0,08
41	-1,8	-1,9	0,0159	0,0158	0,13
42	-1,6	-1,7	0,0223	0,0222	0,23
43	-1,4	-1,5	0,0301	0,03	0,25
44	-1,2	-1,3	0,039	0,0389	0,22
45	-1	-1,1	0,0485	0,0484	0,17
46	-0,8	-0,9	0,058	0,0579	0,09
47	-0,6	-0,7	0,0666	0,0666	0,02
48	-0,4	-0,5	0,0735	0,0736	0,03
49	-0,2	-0,3	0,078	0,0781	0,07
50	0	-0,1	0,0796	0,0797	0,08
51	0,2	0,1	0,078	0,0781	0,07
52	0,4	0,3	0,0735	0,0736	0,03
53	0,6	0,5	0,0666	0,0666	0,02
54	0,8	0,7	0,058	0,0579	0,09
55	1	0,9	0,0485	0,0484	0,17
56	1,2	1,1	0,039	0,0389	0,22
57	1,4	1,3	0,0301	0,03	0,25
58	1,6	1,5	0,0223	0,0222	0,23
59	1,8	1,7	0,0159	0,0158	0,13
60	2	1,9	0,0108	0,0109	0,08
61	2,2	2,1	0,0071	0,0071	0,41
62	2,4	2,3	0,0045	0,0045	0,92
63	2,6	2,5	0,0027	0,0027	1,63
64	2,8	2,7	0,0016	0,0016	2,59
65	3	2,9	0,0009	0,0009	3,82

Graphique 2 : $n = 100$ et $p = 0.5$



2. **Cas n =10 et p = 0,5.**

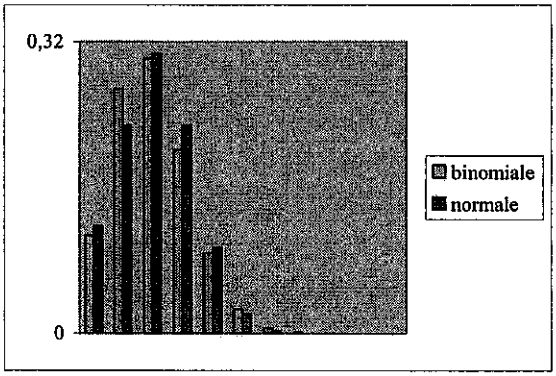
Le tableau et le graphique 3 ci-dessous font apparaître une grande proximité des deux distributions, malgré la faible valeur de n : hormis les 4 valeurs extrêmes (de probabilité inférieure à 1%), l'écart relatif ne dépasse pas 1%.



k	a_k	b_k	binomiale	normale	% écart
0	-3,162	-2,846	0,001	0,0022	56
1	-2,53	-2,214	0,0098	0,0112	13
2	-1,897	-1,581	0,0439	0,0435	1
3	-1,265	-0,949	0,1172	0,1145	2,4
4	-0,632	-0,316	0,2051	0,2045	0,3
5	0	0,316	0,2461	0,2482	0,8
6	0,632	0,949	0,2051	0,2045	0,3
7	1,265	1,581	0,1172	0,1145	2,4
8	1,897	2,214	0,0439	0,0435	1
9	2,53	2,846	0,0098	0,0112	13
10	3,162	3,479	0,001	0,0003	287

3. **Cas n = 10 et p = 0,2.**

Le tableau ci-dessous et le graphique 4 montrent cette fois des différences assez sensibles puisque, à l'exception de la valeur $k = np(= 2)$, l'écart relatif n'est jamais inférieur à 6%.



k	a_k	b_k	binomiale	normale	% écart
0	-1,581	-1,186	0,1074	0,1178	8,9
1	-0,791	-0,395	0,2684	0,2285	17
2	0	0,395	0,302	0,3074	1,7
3	0,791	1,186	0,2013	0,2285	12
4	1,581	1,976	0,0881	0,0938	6,1
5	2,372	2,767	0,0264	0,0212	24
6	3,162	3,558	0,0055	0,0026	108
7	3,953	4,348	0,0008	0,0002	336
8	4,743	5,139	7E-05	7E-06	995
9	5,534	5,929	4E-06	1E-07	4E+03
10	6,325	6,72	1E-07	9E-12	1E+07

En conclusion, ces quelques exemples montrent (sans démontrer) que, s'il y a bien convergence de la loi binomiale vers la loi normale, cette convergence est d'autant plus rapide que la probabilité p est voisine de $0,5$. Par comparaison avec les autres, les graphiques 1 et 4 et les tableaux qui leur sont associés nous fournissent une clé d'explication : ils montrent en effet que, pour $p = 0,2$, la distribution binomiale n'est pas symétrique par rapport à la valeur la plus probable, np (pour toutes les valeurs numériques choisies, np est entier), au contraire de la distribution normale. Au contraire, pour $p = 0,5$, la distribution binomiale, comme la distribution normale, est symétrique par rapport à la valeur np : on a en effet $p_k = p_{n-k}$ car $\binom{n}{k} = \binom{n}{n-k}$ et $p = 1 - p$. Le logiciel, de par ses deux fonctionnalités complémentaires (tableur et grapheur), permet, d'une part de calculer rapidement les distributions de la loi binomiale et de la loi normale "discrétisée", et d'autre part de faire voir ces distributions grâce aux diagrammes en bâtons, ce qui permet de les comparer de façon qualitative (grapheur) et quantitative (tableur). Ainsi, même si l'on ne dispose pas des outils théoriques permettant de démontrer la convergence de la loi binomiale $B(n;p)$ vers la loi normale, on peut s'en persuader grâce à l'outil informatique, et même se faire une idée de l'influence des paramètres n et p dans la rapidité de cette convergence.