ministère éducation nationale jeunesse vie associative

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

éduSCOL

Ressources pour le lycée général et technologique

Ressources pour la classe terminale générale et technologique

Statistiques et probalité

Février 2012

Introduction

Le document ressource pour la partie du programme de la classe terminale « Probabilités et statistique » donne des éléments détaillés permettant aux professeurs de construire leur propre cours. Il ne s'agit pas d'un modèle reproductible tel quel mais d'un support théorique sur les notions introduites pour la première fois dans les programmes du secondaire.

Ces notions sont enseignées dans différents cursus de l'enseignement supérieur mais le point de vue adopté dans le programme de la classe terminale est assez différent.

Les fondements de théorie des probabilités indispensables pour comprendre les notions de statistique inférentielle présentes dans le programme sont développés aussi précisément que possible à ce niveau d'enseignement.

La *loi normale* est introduite en terminale S comme loi-limite d'une suite de variables aléatoires grâce au théorème de Moivre-Laplace. Bien qu'admis, ce théorème se visualise facilement grâce à des animations avec un logiciel de géométrie dynamique ou sur tableur et c'est sous cette forme que la loi normale doit être introduite en terminale ES.

La notion d'*intervalle de fluctuation* d'une variable aléatoire a été introduite en seconde et développée en première dans le cadre de la loi binomiale à l'aide de calculs sur tableur. Elle est enrichie par la notion d'*intervalle de fluctuation asymptotique* d'une variable aléatoire fréquence qui présente l'intérêt de pouvoir se déterminer par un simple calcul.

La notion d'*intervalle de confiance* pour une proportion est introduite grâce à l'intervalle de fluctuation asymptotique.

Tous les nouveaux items sont présentés avec des activités. Celles-ci sont souvent mises en œuvre sur calculatrices ou avec un algorithme. Des exemples d'exercices sont également proposés.

Un complément sur les lois uniforme et exponentielles est proposé, leur approche ayant été modifiée.

L'annexe 1 présente un historique du théorème de Moivre-Laplace en montrant que le concept de fluctuation d'une variable aléatoire autour de son espérance est apparu très tôt avec Jacques Bernoulli et a gagné en précision avec Moivre puis Laplace.

L'annexe 2 donne des compléments sur les lois normales, en particulier sur la fonction de répartition. Cette dernière n'est pas un attendu du programme mais est utilisée par les calculatrices pour les calculs de probabilités sur les lois normales.

L'annexe 3 propose une introduction à la théorie des sondages et donne quelques méthodes couramment utilisées.

L'annexe 4 donne le descriptif des fichiers tableurs, des animations et des algorithmes écrits dans différents langages (Algobox, Scilab, R,...) figurant dans le document. Tous ces fichiers sont téléchargeables. Une aide à la prise en main du logiciel R est également fournie.

L'annexe 5 donne une approche du calcul numérique d'une intégrale par la méthode de Monte-Carlo.

L'annexe 6 fournit des éléments de justification à propos de la notion de différence significative et du critère de disjonction des intervalles de confiance présenté dans le programme de la filière STI2D-STL. Ces éléments n'ont pas à être abordés avec les élèves.

Un document annexe propose une démonstration du théorème de Moivre-Laplace, élaborée de telle sorte que seuls des outils de terminale[1] y sont utilisés. Bien entendu cette démonstration n'est pas au programme mais le théorème de Moivre-Laplace en étant le socle théorique fondamental pour la partie probabilités, il a semblé intéressant d'en faire une proposition de démonstration.

Le théorème de Moivre-Laplace étant un cas particulier d'un théorème général connu sous le nom de théorème-limite central, une approche de ce théorème est proposée à partir de la loi des erreurs.

[1] À l'exception d'un changement de variable (linéaire) incontournable....

Table des matières

I. Variable centrée réduite

A. Comment centrer et réduire

Une variable aléatoire est dite centrée et réduite si son espérance est nulle et si son écart type vaut 1.

Soit X une variable aléatoire discrète d'espérance E(X) = m, de variance V(X) et d'écart type $\sigma = \sqrt{\mathrm{V}(X)}$ non nul.

- La variable aléatoire $(X - m)$ a une espérance nulle
- La variable aléatoire $Z = \dfrac{X - m}{\sigma}$ a une espérance nulle et une variance égale à 1, donc un écart type égal à 1.

Attention

L'écart-type d'une variable aléatoire suivant une loi binomiale ne fait pas partie des contenus mentionnés dans le programme des classes de première ES et L. Il convient donc, avant d'aborder le chapitre sur la loi normale en terminale, de l'introduire en lien avec l'écart-type d'une série statistique et d'en faire percevoir les effets dans le cadre d'une activité de simulation.

Si une variable X prend ses valeurs entre 0 et n, $(X - m)$ les prend entre $-m$ et $n - m$ donc $Z = \dfrac{X - m}{\sigma}$ les prend entre $-\dfrac{m}{\sigma}$ et $\dfrac{n - m}{\sigma}$. Si la variable aléatoire X est représentée par un diagramme en bâtons, on obtient la représentation de la variable $(X - m)$ par translation de vecteur $-m\,\vec{i}$ de ce diagramme. Puis on obtient la représentation de la variable aléatoire Z par « réduction » du nouveau diagramme. Les abscisses sur lesquelles sont construits les bâtons sont les valeurs de $\dfrac{X - m}{\sigma}$ et les hauteurs des bâtons sont les mêmes que celles obtenues pour la variable X, cela conduit à une concentration si $\sigma > 1$. Sur le graphique ci-dessous, on a à droite le diagramme en bâton d'une variable X, à gauche en clair le diagramme de $(X - m)$ et en plus foncé celui de Z.

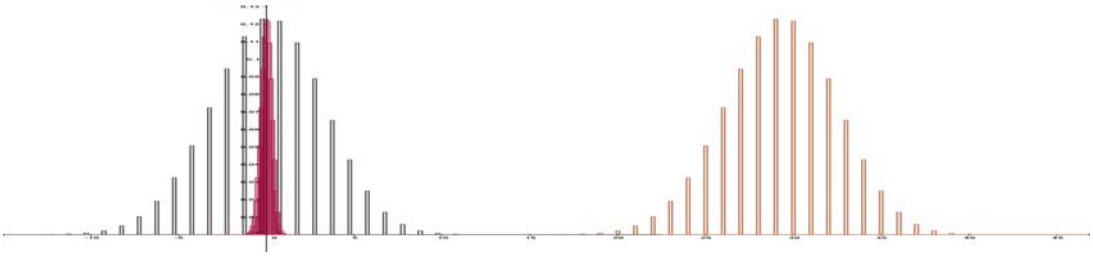


Figure 1 : Effet graphique du centrage et de la réduction sur une variable X suivant une loi $\mathcal{B}$ (45 ; 0,65)

Document associé : centrer et réduire une binomiale.ggb

B. Pourquoi centrer et réduire ?

Lorsqu'on passe de X à Z, on obtient une variable aléatoire dont les paramètres (espérance et variance) ne dépendent plus de ceux de X.

Rappel

Une variable aléatoire qui suit la loi binomiale $\mathcal{B}(n, p)$ peut s'interpréter comme un nombre de succès lors de la répétition de n expériences de Bernoulli indépendantes.

Soit X_n une variable aléatoire suivant la loi binomiale $\mathcal{B}(n, p)$; on a :

$E(X_n) = np$, $V(X_n) = np(1-p)$, et $\sigma(X_n) = \sqrt{np(1-p)}$.

La variable aléatoire $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$ a pour espérance 0 et pour variance 1, indépendantes de n et de p.

La variable aléatoire $F_n = \dfrac{X_n}{n}$ correspond à la proportion de succès, son espérance est p et sa variance est $\dfrac{p(1-p)}{n}$.

On constate que F_n a une espérance qui ne dépend pas de n et une variance qui diminue quand n augmente c'est-à-dire que les réalisations de F_n « ont tendance à se resserrer » autour de p lorsque n augmente. C'est cette concentration des valeurs les plus probables de F_n qui permettra d'améliorer la prise de décision à partir des observations.

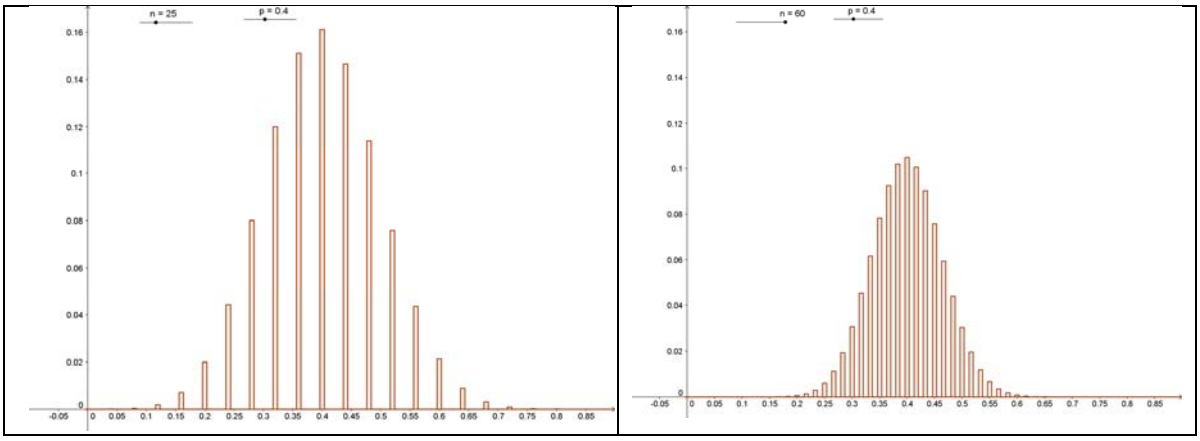


Figure 2 : Diagrammes en bâtons de F_n pour $n = 25$ et $n = 60$

Document associé : diagramme en bâtons de Fn.ggb

Sur les graphiques ci-dessus, on a représenté le diagramme en bâtons d'une variable $F_n = \dfrac{X_n}{n}$ où X_n suit la loi binomiale de paramètres 25 et 0,4 puis 60 et 0,4. Les valeurs prises par F_n sont entre 0 et 1 quel que soit n.

Le paragraphe suivant va permettre de constater que la variable Z_n « tend » vers une variable universelle indépendante de p. La connaissance de la loi de cette variable universelle permet de préciser la fluctuation de $\dfrac{X_n}{n}$ autour de son espérance p**.**

II. La loi normale centrée réduite

A. Activité : Introduction au théorème de Moivre-Laplace

Dans la représentation de la figure 3, on considère une variable aléatoire X_n suivant une loi binomiale $\mathcal{B}(n, p)$ et Z_n est la variable centrée réduite associée.

On prend deux valeurs $a = -1$ et $b = 2$ et on s'intéresse à $P(-1 \leq Z_n \leq 2)$.

Pour le cas visualisé ci-dessous, on a pris $n = 100$ et $p = 0{,}5$.

Donc $P(-1 \leq Z_{100} \leq 2) = P(45 \leq X_{100} \leq 60)$.

Les valeurs prises par Z_{100} quand $-1 \leq Z_{100} \leq 2$ sont de la forme $\dfrac{k - 50}{5}$ avec $45 \leq k \leq 60$.

L'idée est d'associer la loi (discrète) de Z_{100} à des aires de rectangles, comme on le fait pour l'histogramme d'une variable continue.

À chaque valeur de k on fait correspondre un rectangle vertical dont **l'aire** est égale à $\mathrm{P}(X_{100} = k) = \mathrm{P}(Z_{100} = \frac{k-50}{5})$ et dont la base est un segment de l'axe horizontal de longueur $\frac{1}{5}$, centré sur $\frac{k-50}{5}$ ($\frac{1}{5}$ étant l'écart entre deux valeurs consécutives prises par Z). La hauteur de ce rectangle est donc $5\mathrm{P}(X = k)$.

La réunion des rectangles obtenue pour $45 \leq k \leq 60$ a donc pour aire $\mathrm{P}(-1 \leq Z_{100} \leq 2)$.

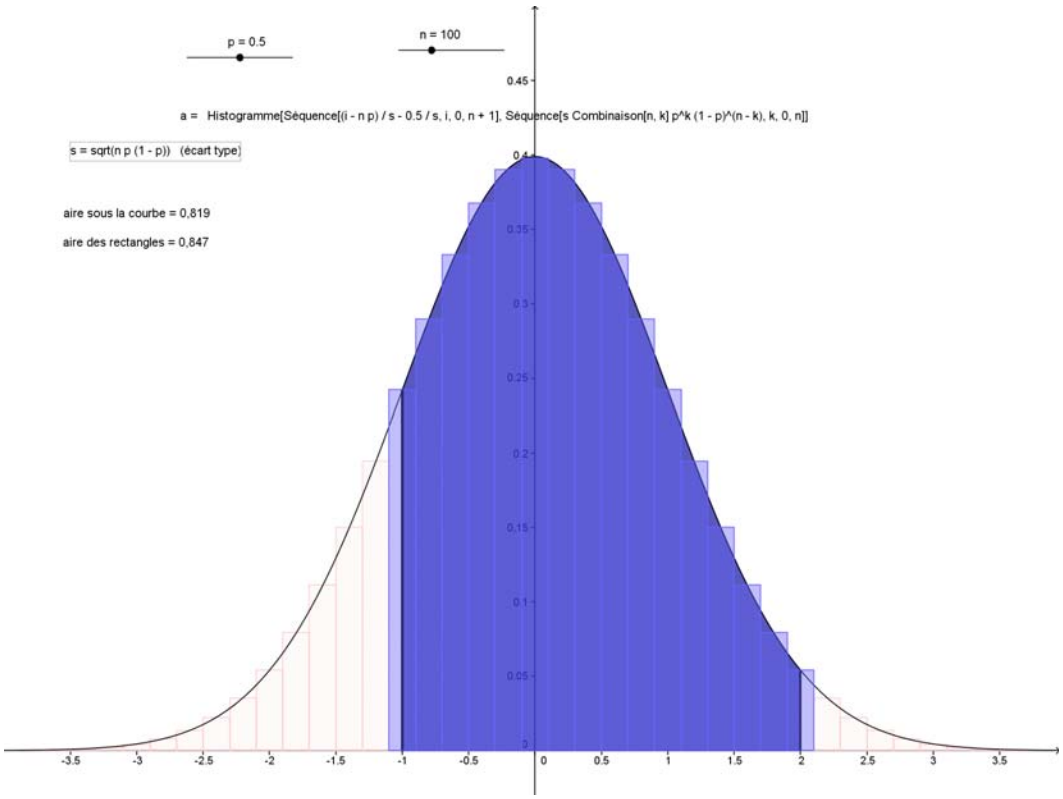


Figure 3 : Visualisation de P($a \leq Z_n \leq b$)

Document associé : binomiale et normale.ggb

Les bords supérieurs des rectangles font apparaître une courbe régulière et symétrique délimitant une aire qui est voisine de celle de la réunion des rectangles.

Le mathématicien Abraham de Moivre, protestant français émigré en Angleterre après la révocation de l'édit de Nantes (1685), a découvert que cette courbe est la courbe représentative de la fonction $x \mapsto \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}$. Le cours de terminale sur l'intégration permet d'écrire que l'aire située sous cette courbe vaut $\int_{-1}^{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x$. Pour comparer l'aire de la réunion des rectangles et celle sous la courbe, on peut remplir le tableau suivant :

k	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
$\mathrm{P}(X = k)$	0,048	0,058	0,067	0,073	0,078	0,080	0,078	0,073	0,067	0,058	0,048	0,039	0,030	0,022	0,016	0,010

La somme des aires des rectangles vaut 0,85 à 10^{-2} près et la valeur de $\int_{-1}^{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$, qu'on peut obtenir avec une calculatrice, est 0,82 à 10^{-2} près.

À partir de l'animation proposée, on constate que :

- Lorsque *n* devient grand, à *p* fixé, la largeur des rectangles est de plus en plus petite car elle vaut $\frac{1}{\sigma} = \frac{1}{\sqrt{np(1-p)}}$.
- L'aire correspondant à $P(Z_n \in [a,b])$ se rapproche de l'aire entre *a* et *b* sous une courbe fixe, qui est la courbe représentative de la fonction $x \mapsto \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

Exercice (TS)

Soit la fonction *g* définie par $g(x) = e^{-\frac{x^2}{2}}$.

1. Montrer que la fonction dérivée g' est minimale pour $x = 1$.
2. Montrer que la fonction $x \mapsto x + g(x)$ est croissante sur $[0, +\infty[$.
3. En déduire que si $0 \leq a \leq b$ alors $a - b \leq g(b) - g(a) \leq 0$ et que si $a \leq b \leq 0$ alors $0 \leq g(b) - g(a) \leq b - a$.
4. En déduire que pour tous réels *a* et *b* on a : $|g(b) - g(a)| \leq |b - a|$.

B. Théorème de Moivre-Laplace

Le résultat suivant est au programme de la classe de terminale S uniquement et il est admis.

Théorème

On suppose que, pour tout entier *n*, la variable aléatoire X_n suit une loi binomiale $\mathcal{B}(n, p)$.

On pose $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$, variable centrée et réduite associée à X_n .

Alors, pour tous réels *a* et *b* tels que $a < b$, on a : $\lim_{n \to +\infty} P(a \leq Z_n \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$.

Voici ce que dit Laplace à propos des travaux de Moivre :

« Moivre a repris dans son ouvrage [*The doctrine of Chances*] le théorème de Jacques Bernoulli sur la probabilité des résultats déterminés par un grand nombre d'observations[2].

Il ne se contente pas de faire voir, comme Bernoulli, que le rapport des événements qui doivent arriver approche sans cesse de celui de leurs possibilités respectives, il donne de plus une expression élégante et simple de la probabilité que la différence de ces deux rapports soit contenue dans des limites données. »

[2] Ce résultat est la loi des grands nombres**.** En seconde, on a donné une forme simplifiée de la loi des grands nombres, à savoir : la probabilité que la variable fréquence s'écarte de *p* diminue quand le nombre d'observations augmente**.**

L'annexe 1 donne des développements sur ce théorème fondamental.

C. La loi normale centrée réduite

Définition

Une variable aléatoire X suit la loi normale centrée réduite[3] notée $\mathcal{N}(0,1)$ si, pour tous réels a et b tels que $a < b$, on a :

$$P(a \leq X \leq b) = \int_a^b f(x)\mathrm{d}x = \int_a^b \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x .$$

La fonction f définie sur $\mathbb{R}$ par $\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}$ est appelée la fonction de densité de la loi $\mathcal{N}(0,1)$.

1. Premières propriétés

- f est continue sur $\mathbb{R}$.
- L'aire totale sous la courbe de f est égale à 1, elle représente la probabilité $P(X \in]-\infty,+\infty[)$.
- La fonction f est paire ; sa courbe représentative est donc symétrique par rapport à l'axe des ordonnées.
- L'aire sous la courbe sur $[0,+\infty[$ est égale à $\frac{1}{2}$.
- Pour tout réel u, $P(X \leq -u) = 1 - P(X \leq u)$.

  Sur la figure, où $u > 0$, les aires grisées sont égales en raison de la symétrie de la courbe représentative.

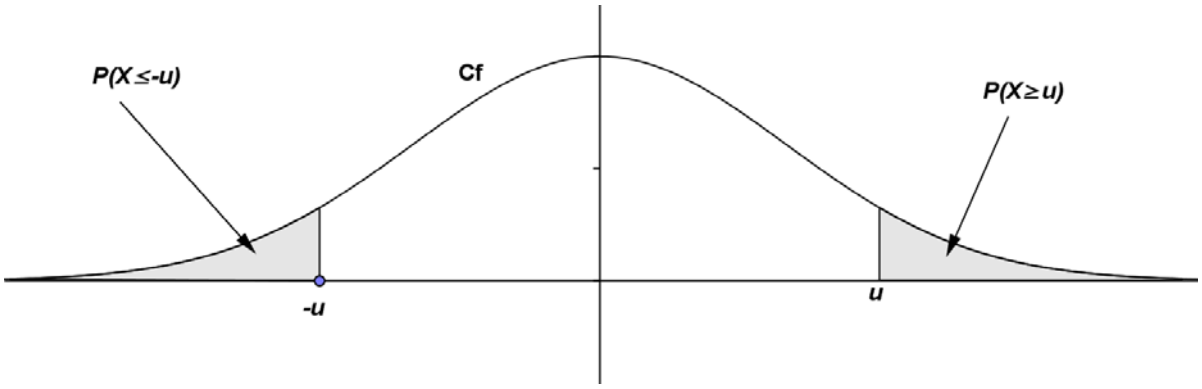


Figure 4: Représentation graphique de la fonction de densité de la loi normale centrée réduite

Théorème (au programme de terminale S)

Si X est une variable aléatoire suivant la loi normale $\mathcal{N}(0,1)$ alors, pour tout réel $\alpha \in]0,1[$, il existe un unique réel positif u_α tel que $P(-u_\alpha \leq X \leq u_\alpha) = 1 - \alpha$.

Démonstration (faisant partie des exigibles en terminale S).

Cette démonstration est intéressante car elle permet de réinvestir le cours sur les fonctions et l'intégration.

[3] Cette loi est également nommée "loi normale standard", en particulier dans les tableurs courants, mais cette dénomination ne figure pas au programme.

D'après la symétrie de la courbe, on a pour tout réel u positif,

$$\mathrm{P}(-u \leq X \leq u) = 2\mathrm{P}(0 \leq X \leq u) = 2\int_0^u f(x)\mathrm{d}x = 2H(u),$$

où H est la primitive de f sur $\mathbb{R}$ qui s'annule en 0. La fonction H est donc continue et strictement croissante sur $]0,+\infty[$. On a $\lim\limits_{u\to+\infty} H(u) = \frac{1}{2}$ puisque cela correspond à l'aire sous la courbe pour $u \in [0,+\infty[$, c'est-à-dire à $\mathrm{P}(X \geq 0)$.

La fonction $2H$ admet donc le tableau de variations et la courbe représentative ci-dessous :

t	0		$+\infty$
			1
$2H(t)$		↗	
	0		

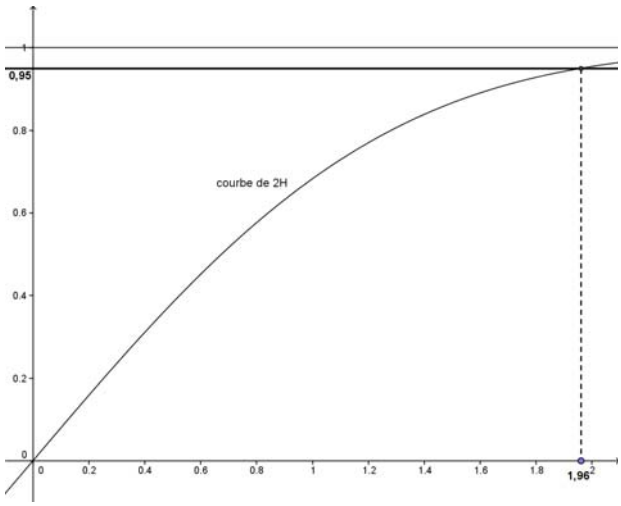
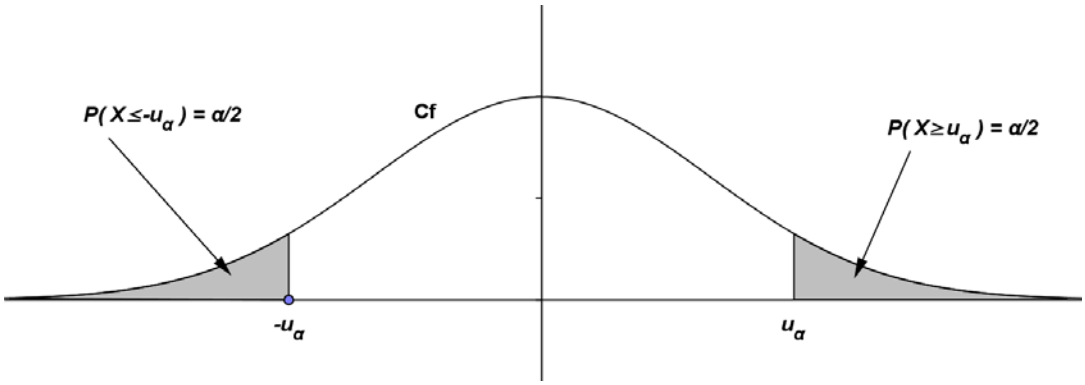


Figure 5 : courbe de la fonction $2H$

Pour tout réel α compris strictement entre 0 et 1, le réel $(1 - \alpha)$ est également compris strictement entre 0 et 1 et donc, d'après le corollaire du théorème des valeurs intermédiaires, il existe un unique réel u_α strictement positif tel que $2H(u_\alpha) = 1 - \alpha$ c'est-à-dire tel que $\mathrm{P}\left(-u_\alpha \leq X \leq u_\alpha\right) = 1 - \alpha$.



Il y a deux valeurs approchées très utilisées qu'il faut connaître :

$u_{0,05} \approx 1{,}96$ et $u_{0,01} \approx 2{,}58$ (à 10^{-2} près)

$u_{0,05}$ est le réel pour lequel $\mathrm{P}(-u_{0.05} \leq X \leq u_{0.05}) = 0{,}95$ et on a donc : $\mathrm{P}(-1{,}96 \leq X \leq 1{,}96) \approx 0{,}95$ de même, $\mathrm{P}(-2{,}58 \leq X \leq 2{,}58) \approx 0{,}99$.

Cela donne une idée de la répartition des valeurs de X. Environ 95% des réalisations de X se trouvent entre – 1,96 et 1,96.

2. Espérance d'une loi normale centrée réduite (uniquement en terminale S)

Selon la définition donnée dans le programme :

> Si X suit la loi $\mathcal{N}(0,1)$, alors l'espérance de X est définie par :
>
> $\mathrm{E}(X) = \lim\limits_{x\to-\infty}\int_x^0 t\,f(t)\mathrm{d}t + \lim\limits_{y\to+\infty}\int_0^y t\,f(t)\mathrm{d}t$.

(on fera le lien avec ce qui est vu avec les lois uniformes et exponentielles).

L'espérance d'une variable aléatoire X suivant la loi $\mathcal{N}(0,1)$ est nulle. En effet :

$$\int_0^y t\,f(t)\mathrm{d}t = \int_0^y \frac{1}{\sqrt{2\pi}} t\,\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t = \frac{1}{\sqrt{2\pi}}\int_0^y t\,\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t = \frac{1}{\sqrt{2\pi}}\left(1-\mathrm{e}^{-\frac{y^2}{2}}\right)$$

De même, $\int_x^0 t\,f(t)\mathrm{d}t = \frac{1}{\sqrt{2\pi}}\left(\mathrm{e}^{-\frac{x^2}{2}}-1\right)$

Par passage à la limite, on obtient $\mathrm{E}(X) = 0$.

La variance de X est définie par l'espérance du carré de l'écart entre X et son espérance soit $\mathrm{E}((X-\mathrm{E}(X))^2)$ et on admet qu'elle vaut 1.

On peut proposer le calcul de la variance en exercice, selon une méthode analogue à celle utilisée pour le calcul de l'espérance d'une loi exponentielle.

III. Lois normales

A. Généralités

On dispose d'un échantillon de 50 000 tailles (en cm) d'hommes adultes dont voici un résumé statistique et un histogramme[4]:

	Moyenne	Écart type	Nombre	Minimum	Maximum	Médiane	Interquartile
Tailles	175,0	8,0	50 000	145,1	208,5	175,0	10,8

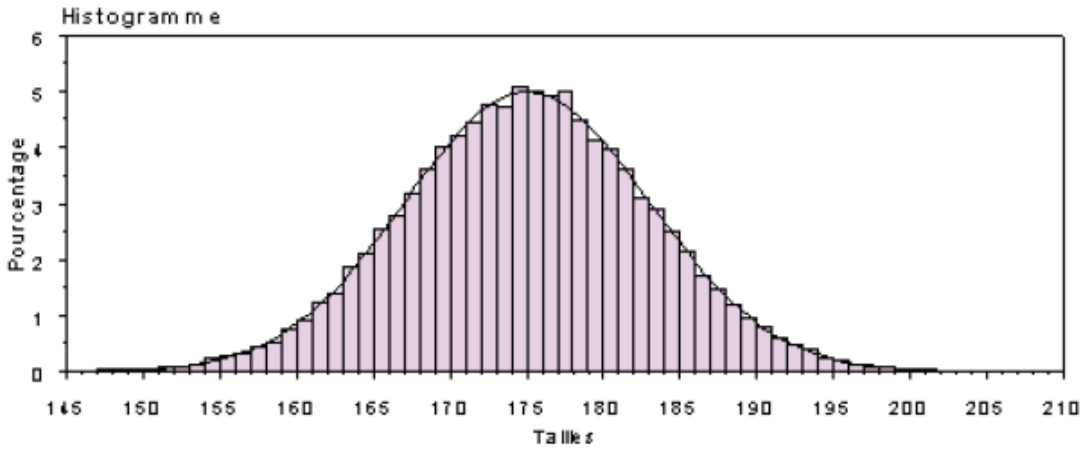


Figure 6 : Répartition des 50000 valeurs de la taille

Si on centre et réduit la variable « taille », l'histogramme obtenu présente une analogie évidente avec la figure 3[5] ; cela motive la définition suivante.

[4] Cet exemple est emprunté au document d'accompagnement publié en 2002.

Une variable aléatoire X suit une loi $\mathcal{N}(\mu, \sigma^2)$ si la variable aléatoire $\frac{X-\mu}{\sigma}$ suit la loi normale $\mathcal{N}(0,1)$.

L'espérance de X vaut μ et sa variance vaut σ^2. La notation $\mathcal{N}(\mu, \sigma^2)$ est justifiée à l'annexe 2.

Remarque

Il s'agit d'une loi à densité c'est-à-dire qu'il existe une fonction g définie sur $\mathbb{R}$ telle que, pour tous réels a et b vérifiant $a \leq b$, on a $P(a \leq X \leq b) = \int_a^b g(t)dt$. L'expression de la fonction de densité de X n'est pas au programme.

On peut constater que μ est à la fois l'espérance et la médiane[6] de X.

Exemple

La masse en kg des nouveaux nés à la naissance est une variable aléatoire qui peut être modélisée par une loi normale[7] de moyenne $\mu = 3{,}3$ et d'écart type $\sigma = 0{,}5$. La probabilité qu'un nouveau né pèse moins de 2,5 kg à la naissance est donc : $P(X < 2{,}5)$. La variable $Z = \frac{X-3{,}3}{0{,}5}$ suit la loi $\mathcal{N}(0,1)$.

On a alors : $P(X < 2{,}5) = P(Z < \frac{2{,}5-3{,}3}{0{,}5}) = P(Z < -1{,}6) = 1 - P(Z < 1{,}6) \approx 0{,}055$.

La probabilité cherchée est donc égale à 0,055 à 10^{-3} près.

On peut aussi obtenir directement la valeur de $P(X < 2{,}5)$.

On donne dans le paragraphe B la méthode pour obtenir cette valeur à la calculatrice.

Les intervalles « Un, deux, trois sigmas »

Les résultats suivants sont utilisés dans de nombreux contextes ; ils peuvent être visualisés sur la figure 7 ci-dessous :

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0{,}68 \text{ (à } 10^{-2} \text{ près)}$$

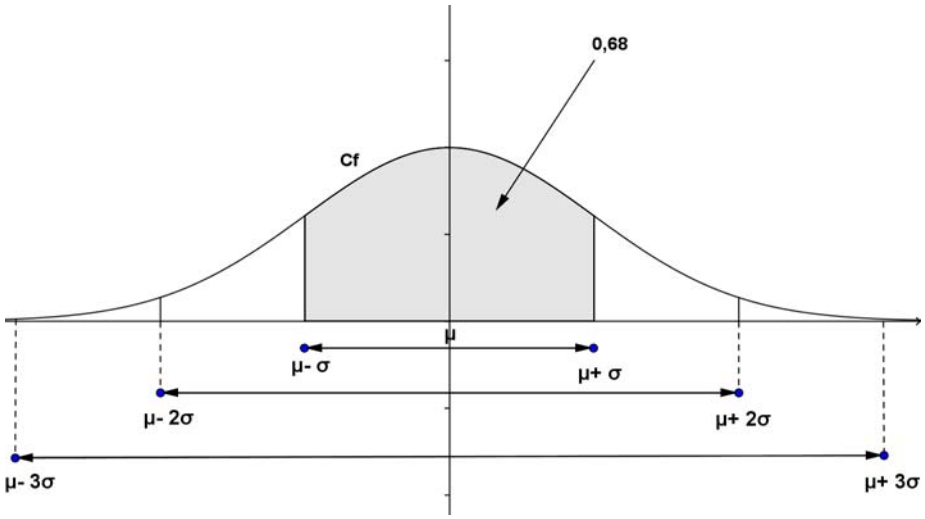
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0{,}95 \text{ (à } 10^{-2} \text{ près)}$$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0{,}997 \text{ (à } 10^{-3} \text{ près).}$$

[5] Il faut noter qu'il s'agit ici d'un histogramme car la variable « taille » est continue alors que sur la figure 3 les rectangles ne sont pas ceux d'un histogramme car la variable binomiale n'est pas continue.

[6] Un réel m est une médiane d'une variable aléatoire si $P(X \leq m) = 0{,}5$

[7] Le poids d'un nouveau né ne prend pas de valeurs négatives mais on peut vérifier que $P(X < 0)$ est négligeable de même que $P(X > 5)$.



Représentations graphiques montrant l'importance de la valeur de l'écart type σ

Courbes représentatives des densités de la loi normale $\mathcal{N}(0,1/4)$ en rouge (maximum voisin de 0,8), de la loi normale $\mathcal{N}(0,1)$ en bleu et de la loi normale $\mathcal{N}(0,4)$ en vert.

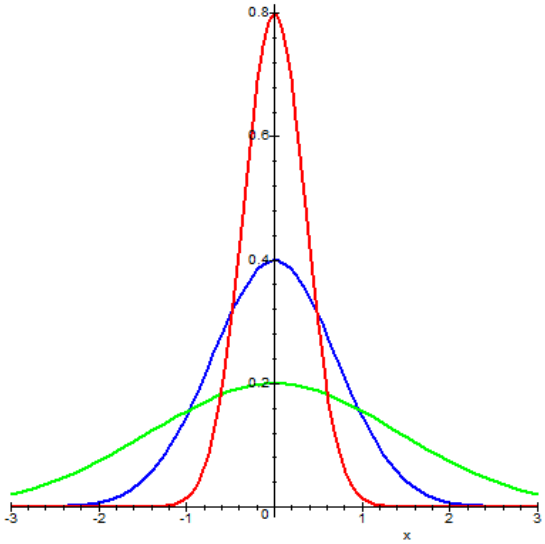


Figure 8: Influence de l'écart type

B. Exemples d'exercices

1. Montrer que si X suit la loi $\mathcal{N}(0,1)$, alors $-X$ suit la même loi.

2. La sélection chez les vaches laitières de race « Française Frisonne Pis Noir »

La production laitière annuelle en litres des vaches laitières de la race FFPN peut être modélisée par une variable aléatoire à densité X, de loi normale de moyenne $\mu = 6000$ et d'écart-type $\sigma = 400$. La fonction g désigne la fonction de densité de cette loi normale.

1° Afin de gérer au plus près son quota laitier (production maximale autorisée), en déterminant la taille optimale de son troupeau, un éleveur faisant naître des vaches de cette race souhaite disposer de certaines probabilités.

a) Calculer la probabilité qu'une vache quelconque de cette race produise moins de 5800 litres par an.

Solution

En utilisant calculatrices ou logiciels, on trouve : $P(X < 5800) \approx 0{,}3085$. Certaines calculatrices et logiciels de calcul numérique proposent une fonction dédiée à ce type de calcul (pnorm() dans R, normalFRép chez Texas (normaFrép pour fonction de répartition de la loi normale), menu Ncd chez Casio (Ncd pour Normal cumulative density). Il y a un faux ami : $P(X \leq x)$ qui est la fonction de répartition, en français est appelé "distribution function" en anglais, alors que notre fonction de distribution pour une variable discrète est classiquement $P(X = x)$.

Attention !

Les calculatrices ne fournissent pas $P(X \leq x)$ mais seulement $P(a \leq X \leq b)$.

Pour le calcul de $P(X \leq x)$ dans le cas où X suit une loi $\mathcal{N}(\mu, \sigma^2)$, la règle pratiquée est donc la suivante :

- Si $x \geq \mu$, on utilise $P(X \leq x) = 0{,}5 + P(\mu \leq X \leq x)$
- Si $x \leq \mu$, on utilise $P(X \leq x) = 0{,}5 - P(x \leq X \leq \mu)$.

Pour entrer les paramètres, il faut saisir les valeurs de μ et de σ (et non σ^2).

R	TEXAS(83Plus) et +	CASIO(35+) et +
répartition normale pré programmée pnorm(5800, mean = 6000, sd = 400, lower.tail = TRUE) ou pnorm(5800, 6000, 400) [1] 0.3085375 ou Complément pour l'enseignant : intégration numérique de la densité d'une loi normale de paramètres mu sigma. (-Inf signifie moins l'infini et Inf plus l'infini. $value signifie que l'on ne prend que la valeur numérique de l'objet résultat de la fonction integrate. La fonction gauss est la densité d'une loi de Gauss d'espérance mu et d'écart type sygma, g en est un cas particulier) gauss <- function(x, mu = moy, sigma = et){dnorm(x, mu, sigma)} moy <- 6000 ; et <- 400 integrate(gauss, -Inf, 5800)$value [1] 0.3085375	répartition normale pré programmée 0.5 - normalFRép(5800,6000,6000,400) 0.3085375 ou Complément pour l'enseignant : intégration numérique après changement de variable pour se ramener à la loi normale centrée réduite. intégrFonct(1/√(2Π)*e^(- t²/2), t,-5,(5800 - 6000)/400) 0.3085373	répartition normale pré programmée menu stat ▶ dist ▶ NORM ▶ Ncd ▶ Lower : 5800 ; Upper : 6000 σ : 400 ; μ : 6000. Normal C.D. prob = 0.19147 .5 - .19147 .30853 ou Complément pour l'enseignant : intégration numérique après changement de variable pour se ramener à la loi normale centrée réduite SET UP Integration : Simpson menu RUN ▶ OPTN ▶ CALC ▶ ∫dx(1/√(2Π)*e^(- x²/2), -5, (5800 - 6000)/400) 0.3085372

b) Calculer la probabilité qu'une vache quelconque de cette race produise entre 5900 et 6100 litres de lait par an.

Solution : $P(5900 < X < 6100) \approx 0{,}1974$.

c) Calculer la probabilité qu'une vache quelconque de cette race produise plus de 6250 litres par an.

Solution : $P(X > 6250) \approx 0{,}2660$.

2° Dans son futur troupeau, l'éleveur souhaite connaître :

a) la production maximale prévisible des 30% de vaches les moins productives du troupeau.

Il s'agit de déterminer la valeur x de X telle que $P(X < x) = 0{,}30$.

Réponse : $x \approx 5790$ litres de lait par an.

Certaines calculatrices et logiciels de calcul numérique proposent une fonction dédiée à ce type de calcul (qnorm() dans R pour normal quantile, FracNormale chez Texas pour fractiles de la loi normale, menu InN chez Casio pour loi normale «inverse».

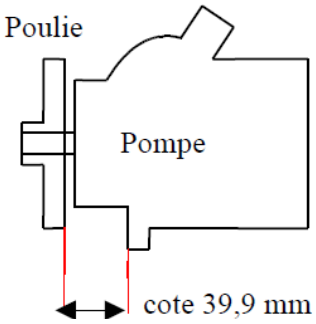
R	TEXAS(83Plus) et +	CASIO(35+) et +
répartition normale réciproque pré programmée qnorm(.30, 6000, 400) [1] 5790.24	répartition normale réciproque pré programmée FracNormale(.30,6000,400) 5790.24	répartition normale réciproque pré programmée menu stat ▶ dist ▶ NORM ▶ InvN ▶ Area :.3 σ : 400 ; μ : 6000. Inverse Normal x = 5790.2

b) la production minimale prévisible des 20% des vaches les plus productives.

Il s'agit de déterminer la valeur x de X telle que $P(X > x) = 0{,}20$.

Réponse : $x \approx 6336$ litres de lait par an.

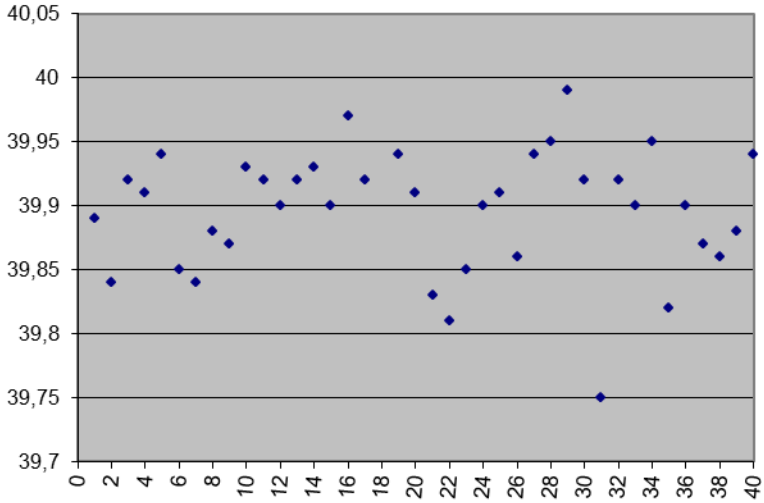
3. *Processus industriel*[8]



Le schéma ci-contre représente une pompe de direction assistée d'automobile. Le processus industriel étudié est une presse d'emmanchement de la poulie sur l'axe de la pompe. Les performances de la presse sont variables, cette variabilité ayant de nombreuses causes possibles : main d'œuvre, matériel, matière première.

Sur le schéma ci-contre est spécifiée par le constructeur une cote de 39,9 mm.

On a mesuré cette cote sur 40 ensembles poulie-pompe issus du processus de fabrication en série. Les variations sont représentées sur le graphique suivant :



1. Ce type de processus industriel induit la modélisation de la variable aléatoire « cote » par une variable suivant une loi normale $\mathcal{N}(\mu, \sigma^2)$[9].

Donner par lecture graphique une valeur estimée[10] de l'espérance μ et de l'écart-type σ à partir de la série des 40 valeurs. (*Réponse* : *environ 39,9 et 0,05*)

[8] Cet exemple est emprunté à la brochure IREM n° 112: Enseigner la statistique au lycée.

[9] On peut vérifier la validité d'un tel modèle par des tests de normalité, mais c'est hors de propos ici.

2. L'intervalle de tolérance pour cette cote est de 39,9 ± 0,15.

   Donner, à l'aide des 40 mesures effectuées, une valeur approchée de la probabilité que la variable cote soit dans cet intervalle. (*Réponse* : *environ 0,997*).

4. *Masse d'alerte pour cartes de contrôle*

Une coopérative produit du beurre en microplaquettes de 12,5g pour des collectivités et des chaînes hôtelières. Les microplaquettes sont conditionnées dans des boîtes de 40.

La masse des microplaquettes peut être modélisée par une variable aléatoire suivant une loi normale d'espérance $\mu = 12{,}5$ et de variance $\sigma^2 = 0{,}2^2$ et on admet que la variable aléatoire X égale à la masse d'une boîte de 40 microplaquettes suit alors une loi normale d'espérance $\mu = 500$ et de variance $\sigma^2 = 1{,}6$ (les notions relatives à la variance d'une somme de variables ne sont pas au programme, quelques notions sont abordées en annexe 2).

La boîte est jugée conforme si sa masse est comprise entre 496,2 g et 503,8 g (soit environ $500 \pm 3\sigma$).

1. Calculer la probabilité qu'une boîte prélevée aléatoirement en fin de chaîne de conditionnement soit non conforme. (*Réponse :* 0,003 *à* 10^{-3} *près)*
2. Pour contrôler le réglage de la machine, on détermine des poids d'alerte $\mu - h$ et $\mu + h$ tels que $P(\mu - h < X < \mu + h) = 0{,}99$. Ces poids d'alerte sont inscrits sur une carte de contrôle et correspondent à une marge de sécurité en lien avec des normes de conformité.

   Calculer les poids d'alerte.

   Solution

   Notons $Z = \dfrac{X - 500}{\sqrt{1{,}6}}$. Z suit une loi normale centrée réduite donc nous savons que $P(-2{,}58 < Z < 2{,}58) \approx 0{,}99$. Il ne reste plus, pour trouver $\mu - h$ et $\mu + h$**,** qu'à résoudre $\dfrac{\mu + h - 500}{\sqrt{1{,}6}} = 2{,}58$ et $\dfrac{\mu - h - 500}{\sqrt{1{,}6}} = -2{,}58$ ce qui donne $\mu + h \approx 503{,}3$ et $\mu - h \approx 496{,}7$.

Grâce à des échantillons prélevés en sortie de chaine ces masses d'alerte permettent de déceler des anomalies en temps réel.

5. *Réglage d'une machine d'embouteillage dans une coopérative*

Sur une chaîne d'embouteillage dans une brasserie, la quantité X (en cL) de liquide fournie par la machine pour remplir chaque bouteille de contenance 110 cL peut être modélisée par une variable aléatoire de loi normale de moyenne μ et d'écart-type $\sigma = 2$.

La législation impose qu'il y ait moins de 0,1% de bouteilles contenant moins d'un litre.

À quelle valeur de la moyenne μ doit-on régler la machine pour respecter cette législation?

Solution

Il s'agit déterminer la valeur de μ telle que $P(X < 100) < 0{,}001$. On détermine d'abord la valeur z (on dit aussi quantile) de la loi normale centrée réduite, telle que $P(Z < z) = 0{,}001$. On trouve (logiciels ou calculettes) $z \approx -3{,}09$. Comme $Z = (X - \mu) / 2$, il ne reste plus, pour trouver μ**,** qu'à résoudre $-3{,}09 = (100 - \mu) / 2$. On trouve $\mu \approx 106{,}18$.

[10] Il existe des méthodes d'estimation par intervalle de confiance de ces paramètres, mais ici il s'agit simplement d'une valeur empirique.

R	TEXAS(83Plus) et +	CASIO(35+) et +
qnorm(p) est la fonction qui permet de trouver t tel que P(T<t)≈p, T étant de loi normale (c'est la répartition normale réciproque pré programmée) qnorm(.001) [1] -3.090232	FracNormale() est la fonction qui permet de trouver t tel que P(T<t)≈p, T étant de loi normale (c'est la répartition normale réciproque pré programmée) FracNormale(.001,0,1) -3.0902323	InvN est le menu qui permet de trouver t tel que P(T<t)≈p T étant de loi normale (c'est la répartition normale réciproque pré programmée) menu stat ▶ dist ▶ NORM ▶ InvN ▶ Area :.001 σ : 1 ; μ : 0. Inverse Normal x = -3.0902

2° La contenance des bouteilles étant de 110 cL, quelle est alors la probabilité qu'une bouteille déborde lors du remplissage?

Solution : Avec $\mu \approx 106{,}18$, on obtient $P(X > 110) \approx 0{,}028$.

4° Le directeur de la coopérative veut qu'il y ait moins de 1% de bouteilles qui débordent au risque de ne plus suivre la législation.

a) Quelle est alors la valeur de μ?

Solution

Il s'agit cette fois de déterminer μ tel que $P(X > 110) < 0{,}01$. On trouve $\mu \approx 105{,}34$.

b) Quelle est dans les conditions de la question a) la probabilité que la bouteille contienne moins d'un litre?

Solution

Avec cette valeur de μ , on obtient $P(X < 100) \approx 0{,}0038$, ce qui est plus élevé que dans le cas précédent.

c) Déterminer μ et σ afin qu'il y ait moins de 0,1% de bouteilles de moins d'un litre ET moins de 1% de bouteilles qui débordent.

Solution

On cherche donc à déterminer les valeurs de μ et de σ de sorte que :

$$P(X < 100) < 0{,}001 \quad \text{et} \quad P(X > 110) < 0{,}01.$$

Les deux contraintes sur les probabilités fournissent les deux conditions suivantes.

On détermine d'abord la valeur z_{sup} de la loi normale centrée réduite telle que $P(Z > z_{sup}) = 0{,}01$. On trouve (logiciels ou calculettes) $z_{sup} \approx 2{,}33$.

On détermine ensuite la valeur z_{inf} telle que $P(Z < z_{inf}) = 0{,}001$. On trouve $z_{inf} \approx -3{,}09$.

Les deux contraintes se traduisent donc par les deux inégalités suivantes :

$$\frac{110-\mu}{\sigma} \geq 2{,}33 \qquad \text{et} \qquad \frac{100-\mu}{\sigma} \leq -3{,}09 \ .$$

On obtient donc un domaine de solutions et une discussion pourra être menée quant aux choix pertinents que le directeur de coopérative pourrait faire.

6. *Durée de vie d'un appareil*

La durée de vie d'un certain type d'appareil est modélisée par une variable aléatoire suivant une loi normale de moyenne et d'écart-type inconnus. Les spécifications impliquent que 80 % de la production des appareils ait une durée de vie entre 120 et 200 jours et que 5% de la production ait une durée de vie inférieure à 120 jours.

1. Quelles sont les valeurs de μ et σ^2 ?

2. Quelle est la probabilité d'avoir un appareil dont la durée de vie soit comprise entre 200 jours et 230 jours ?

Solution

1. On note X la variable durée de vie. Les spécifications se traduisent par :

$P(120 \leq X \leq 200) = 0{,}8$ et $P(X < 120) = 0{,}05$.

En notant toujours $Z = \dfrac{X - \mu}{\sigma}$ la variable centrée réduite, on obtient :

$P\left(\dfrac{120 - \mu}{\sigma} \leq Z \leq \dfrac{200 - \mu}{\sigma}\right) = 0{,}8$ et $P\left(Z < \dfrac{120 - \mu}{\sigma}\right) = 0{,}05$

En utilisant logiciel ou calculatrice, on obtient : $\mu = 120 + 1{,}65\ \sigma$ et $\mu = 200 - 1{,}04\ \sigma$.

La résolution du système donne : $\mu \approx 169$ et $\sigma^2 \approx 884$.

2. $P(200 \leq X \leq 230) = P(X \leq 230) - P(X \leq 200) \approx 0{,}13$

IV. Intervalle de fluctuation

A. Cas binomial

Soit X une variable suivant une loi $\mathcal{B}(n, p)$ et α un réel dans l'intervalle]0, 1[.

Dans un cadre général, tout intervalle $[a,b]$ tel que : $P(X \in [a,b]) \geq 1-\alpha$ peut être considéré comme un intervalle de fluctuation de X au seuil $1-\alpha$.

Ainsi l'intervalle $[0,n]$ est un intervalle de fluctuation évident au seuil 1 mais il est de toute évidence sans intérêt.

On peut chercher :

- celui qui a l'amplitude minimale (IF1)
- le plus petit intervalle centré autour de l'espérance np comme dans le théorème de Moivre-Laplace (IF2)
- celui qui symétrise les probabilités que X soit à l'extérieur, comme proposé dans le document ressource de première (IF3)
- Dans le programme de seconde, on donne un intervalle de fluctuation approché au seuil 0,95, valable sous certaines conditions, de la variable fréquence $F_n = \frac{X_n}{n}$: $\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}\right]$ (IF4)

À titre d'exemple voici les intervalles obtenus pour n = 100 et p =0,3 au seuil 0,95.

- IF (1) le plus petit : [22, 39] de probabilité 0,9502
- IF (2) centré sur 30 : [21, 39] de probabilité 0,9625
- IF(3) (première) : [21, 39] avec une probabilité inférieure à 0,025 que X soit à gauche et inférieure à 0,025 que X soit à droite de l'intervalle.
- IF(4) (seconde) : [20, 40] de probabilité 0,9710.

On peut vérifier que, pour une même valeur de p, ces différents intervalles sont de plus en plus proches lorsque n augmente.

B. Activité : recherche et utilisation d'un intervalle de fluctuation à l'aide d'un algorithme

Le responsable de la maintenance des machines à sous d'un casino doit vérifier qu'un certain type de machine est bien réglé sur une fréquence de succès de 0,06. Pour cela il veut établir un programme qui lui fournira, en fonction de n (nombre de coups joués) et de p (probabilité de succès), un intervalle de fluctuation, au seuil de 95%, de la fréquence de succès. Cela lui permettra de prendre la décision de régler chaque machine pour laquelle il aura observé, dans l'historique des jeux, une fréquence de succès se situant en dehors de cet intervalle de fluctuation.

1° Voici un exemple d'algorithme en Algobox et sa traduction dans le logiciel **R** permettant de déterminer l'intervalle de fluctuation d'une variable binomiale selon la méthode exposée dans le document ressource de première.

- On cherche le plus petit entier a pour lequel $P(X \leq a)$ est strictement supérieur à 0,025 et le plus petit entier b pour lequel $P(X \leq b)$ est supérieur ou égal à 0,975.
- Étant donné que a devient a + 1 en fin de « tant que », il faut faire afficher a – 1, et de même pour b.
- Avec Algobox, cet algorithme ne fonctionne que pour n < 70. Avec le logiciel **R** il n'y a pas cette limite. Le programme **R** fournit la proposition de décision en fonction de la valeur observée (kobs) du nombre de succès.

2° Lors du contrôle d'une machine, le technicien constate qu'elle a fourni 8 succès sur 65 jeux, soit une fréquence observée de succès d'environ 0,12. L'intervalle de fluctuation de la variable fréquence fourni par l'un des deux programmes précédents est [0,015 ; 0,123]. Bien que la fréquence observée de succès soit de 0,12, la règle de décision n'amène pas à remettre en question le réglage de la machine.

Si le même pourcentage de succès (0,12, kobs = 12) avait été observé sur 100 jeux, l'intervalle de fluctuation aurait été de [0,02 ; 0,11], ce qui aurait conduit à remettre en question le réglage de la machine. Le technicien aurait pris la décision de régler la machine.

Algorithme Algobox :

```
VARIABLES
  n EST_DU_TYPE NOMBRE
  p EST_DU_TYPE NOMBRE
  a EST_DU_TYPE NOMBRE
  b EST_DU_TYPE NOMBRE
  i EST_DU_TYPE NOMBRE
  Frep EST_DU_TYPE NOMBRE
DEBUT_ALGORITHME
  a PREND_LA_VALEUR 0
  b PREND_LA_VALEUR 0
  LIRE n
  LIRE p
  Frep PREND_LA_VALEUR 0
  TANT_QUE (Frep<=0.025) FAIRE
    DEBUT_TANT_QUE
    Frep PREND_LA_VALEUR 0
    POUR i ALLANT_DE 0 A a
      DEBUT_POUR
      Frep PREND_LA_VALEUR Frep+ALGOBOX_LOI_BINOMIALE(n,p,i)
      FIN_POUR
    a PREND_LA_VALEUR a+1
    FIN_TANT_QUE
  TANT_QUE (Frep<0.975) FAIRE
    DEBUT_TANT_QUE
    Frep PREND_LA_VALEUR 0
    POUR i ALLANT_DE 0 A b
      DEBUT_POUR
      Frep PREND_LA_VALEUR Frep+ALGOBOX_LOI_BINOMIALE(n,p,i)
      FIN_POUR
    b PREND_LA_VALEUR b+1
    FIN_TANT_QUE
  a PREND_LA_VALEUR a-1
  AFFICHER a
  b PREND_LA_VALEUR b-1
  AFFICHER b
FIN_ALGORITHME
```

 # Fontion R :

```
# IF binomial doc. ressour. 1ère : IF symétrique
(équilibré) en proba
# n est la taille de l'échantillon, p est la
probabilité de succès
# kobs est le nombre de succès observé dans
l'échantillon
# proba est le seuil de probabilité de l'intervalle
de fluctuation
# a est le plus petit entier tel que P(X <= a) >
0,025
# b est le plus petit entier tel que P(X <= b) >=
0,975
IFexact2 = function(n = 65, p = .06, kobs = 8, proba
= .95){
 a <- 0 ; b <- 0
 reparti1 <- pbinom(0:n, n, p, lower.tail = T)
 names(reparti1) <- 0:n
 pinf <- 0
 while(pinf <= (1 - proba) / 2){
   pinf <- pbinom(a, n, p, lower.tail = T)
   a <- a + 1                    }
 pinf <- 0
 while(pinf < (1 - (1 - proba) / 2)){
   pinf <- pbinom(b, n, p, lower.tail = T)
   b <- b + 1                        }
 probaab <- sum(dbinom((a - 1):(b - 1), n, p))
 if(kobs >= (a - 1) & kobs <= (b - 1)) {
   hypothese <- "ACCEPTÉE"              } else
{hypothese <-"REFUSÉE"}
#*******Affichage des résultats et des
graphiques*******
 cat("\nL'IF exact des comptages symétrique en proba
est :\n[",
   a - 1,",",b - 1,"] de probabilité :", probaab,
   "\n\nL'IF exact des proportions symétrique en
proba est :\n[",
   (a - 1) / n,",",(b - 1) / n,"]\n\n",
   "Hypothèse p théorique = ", p,
   ": confrontée à f observé =",kobs / n, " : ",
hypothese,"\n")
}

#----------Application et résultats:----
IFexact2(n = 50, p = 1/2, kobs = 19)
IFexact2(n = 65, p = .06, kobs = 8)
L'IF exact des comptages symétrique en proba
est :
[ 1 , 8 ] de probabilité : 0.9668145
L'IF exact des proportions symétrique en proba
est :
[ 0.01538462 , 0.1230769 ]
 Hypothèse p théorique =  0.06 : confrontée à
f observé = 0.1230769  :  ACCEPTÉE
```

Document associé : intervalle de fluctuation première.alg

Le théorème de Moivre-Laplace va permettre de donner un intervalle de fluctuation calculable directement, sous réserve que *n* soit assez grand. Comme il est obtenu grâce à une convergence, on le qualifie d'intervalle de fluctuation *asymptotique*.

C. Intervalle de fluctuation asymptotique

Théorème

Si la variable aléatoire X_n suit la loi $\mathcal{B}(n, p)$, avec p dans l'intervalle]0, 1[, alors pour tout réel α dans l'intervalle]0, 1[on a :

$$\lim_{n\to+\infty} P\left(\frac{X_n}{n} \in I_n\right) = 1-\alpha \text{ , où } I_n \text{ désigne l'intervalle } \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

et u_α désigne l'unique réel tel que $P(-u_\alpha \le Z \le u_\alpha) = 1-\alpha$ où Z suit la loi normale $\mathcal{N}(0,1)$.

Démonstration (exigible en terminale S**)**

D'après le théorème de Moivre-Laplace, on a $\lim_{n\to+\infty} P(-u_\alpha \le Z_n \le u_\alpha) = P(-u_\alpha \le Z \le u_\alpha)$

où $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$.

Or : $P(-u_\alpha \le Z_n \le u_\alpha) = P\left(np - u_\alpha\sqrt{np(1-p)} \le X_n \le np + u_\alpha\sqrt{np(1-p)}\right)$

$$= P\left(p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \le \frac{X_n}{n} \le p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right).$$

Application

Quand on sait qu'une suite converge vers une limite L, on peut considérer que pour n assez grand le terme de rang n constitue une approximation de L.

Ici, on inverse les rôles. On connaît la limite, mais pas les valeurs des termes de la suite. On admet donc que, sous certaines conditions, on peut approcher le terme de rang n de la suite $P\left(\frac{X_n}{n} \in I_n\right)$ par sa limite $1-\alpha$.

Ces conditions communément admises pour pratiquer l'approximation sont :

$n \ge 30, \quad np \ge 5, \quad n(1-p) \ge 5$.

L'intervalle $I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ est un intervalle de fluctuation « approché » de la variable fréquence $\frac{X_n}{n}$ au seuil $1-\alpha$.

La suite de terme général $P(\frac{X_n}{n} \in I_n)$ n'étant pas monotone, on ne peut pas savoir si la probabilité de l'intervalle est supérieure ou inférieure à la limite $1-\alpha$ (cf note[11]). Cette situation peut être illustrée à l'aide d'un tableur ou du logiciel **R**. Voici un exemple dans le cas où $p = \frac{1}{2}$ et $\alpha = 0,05$. Pour les valeurs de n entre 0 et 2000, on calcule la probabilité que la variable $\frac{X_n}{n}$ appartienne à l'intervalle I_n .

[11] Dans la pratique on parle de seuil $1-\alpha$, les écarts par rapport à cette limite étant minimes (voir fig 10).

On peut constater que le nuage de points obtenu a un aspect symétrique autour de la droite d'équation $y = 0{,}95$ et que lorsque n est grand les points se rapprochent de cette droite.

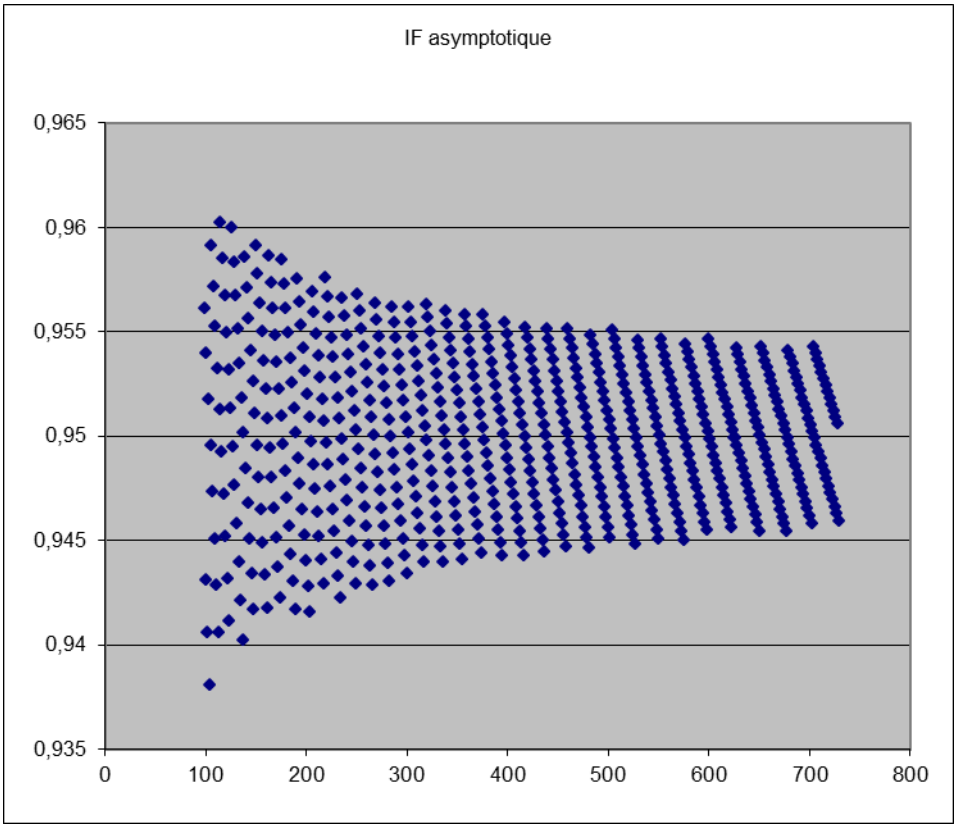


Figure 9: visualisation de la probabilité $P(\frac{X_n}{n} \in I_n)$

Lien vers : exploration intervalle de fluctuation asymptotique.xls

Définition

Un intervalle de fluctuation asymptotique de la variable aléatoire $F_n = \frac{X_n}{n}$ au seuil $1-\alpha$ est un intervalle déterminé à partir de p et de n et qui contient F_n avec une probabilité d'autant plus proche de $1-\alpha$ que n est grand. L'intervalle I_n du théorème précédent est donc un intervalle de fluctuation asymptotique de F_n au seuil $1-\alpha$.

Seul l'intervalle de fluctuation asymptotique au seuil de 95% est au programme des classes de terminale autre que la terminale S ; c'est celui qui est mis en œuvre dans l'exemple 1 ci-dessous.

Remarque

Quand $n \geq 30$, $np \geq 5$, $n(1-p) \geq 5$, il est courant de faire les calculs impliquant une variable binomiale en la remplaçant par une variable suivant une loi normale de mêmes espérance et variance.

Seul le programme de STI2D-STL mentionne cette pratique, qui ne doit donc pas être mise en œuvre dans les autres filières où tous les calculs de probabilités se font à la calculatrice en utilisant la loi exacte (au programme), quelle qu'elle soit.

Les calculs d'intervalles de fluctuation et d'intervalles de confiance se font avec les formules données dans le programme.

D. Exemples d'utilisation

Dans les exemples qui suivent, les tirages sont effectués sans remise. Toutefois, la taille des échantillons considérés étant faible par rapport à la taille de la population totale, on apparente les tirages à des tirages avec remise, correspondant alors à un schéma de Bernoulli et permettant d'appliquer les résultats théoriques précédents.

1. *Prise de décision*

On admet que dans la population d'enfants de 11 à 14 ans d'un département français le pourcentage d'enfants ayant déjà eu une crise d'asthme dans leur vie est de 13%.
Un médecin d'une ville de ce département est surpris du nombre important d'enfants le consultant ayant des crises d'asthme et en informe les services sanitaires. Ceux–ci décident d'entreprendre une étude et d'évaluer la proportion d'enfants de 11 à 14 ans ayant déjà eu des crises d'asthme.
Ils sélectionnent de manière aléatoire 100 jeunes de 11 à 14 ans de la ville.

La règle de décision prise est la suivante : si la proportion observée est supérieure à la borne supérieure de l'intervalle de fluctuation asymptotique au seuil de 95% alors une investigation plus complète sera mise en place afin de rechercher les facteurs de risque pouvant expliquer cette proportion élevée.

1) Déterminer l'intervalle de fluctuation asymptotique au seuil de 95% de la proportion de jeunes de 11 à 14 ans ayant eu une crise d'asthme dans un échantillon de taille 100. (*solution* : [0,06 ; 0,20])

2) L'étude réalisée auprès des 100 personnes a dénombré 19 jeunes ayant déjà eu des crises d'asthme. Que pouvez-vous conclure ?

Solution : la valeur 0,19 est à l'intérieur de l'intervalle de fluctuation asymptotique au seuil de 95%, On en conclut que la règle de décision choisie ne prévoit pas de réaliser une enquête supplémentaire.

3) Le médecin n'est pas convaincu par cette conclusion et déclare que le nombre de personnes interrogées était insuffisant pour mettre en évidence qu'il y avait plus de jeunes ayant eu des crises d'asthme que dans le reste du département.
Combien faudrait-il prendre de sujets pour qu'une proportion observée de 19% soit en dehors de l'intervalle de fluctuation asymptotique ?

Solution : il faut et il suffit que la borne supérieure de l'intervalle asymptotique de fluctuation soit inférieure à 0,19 ce qui équivaut à $0{,}13+1{,}96\times\dfrac{\sqrt{0{,}13\times 0{,}87}}{\sqrt{n}}<0{,}19$, soit $n>120$.
La taille doit donc être de 121 sujets au minimum si on souhaite mettre en évidence une proportion anormalement élevée dans la ville étudiée.

4) Représenter graphiquement la taille de l'échantillon nécessaire en fonction de la valeur *p*sup de la borne supérieure de l'intervalle de fluctuation au seuil de 95%.

Solution

L'expression de *n* en fonction de *p* sup est $n=\dfrac{1{,}96^2\times 0{,}13\times 0{,}87}{\left(p\,\text{sup}-0{,}13\right)^2}$.

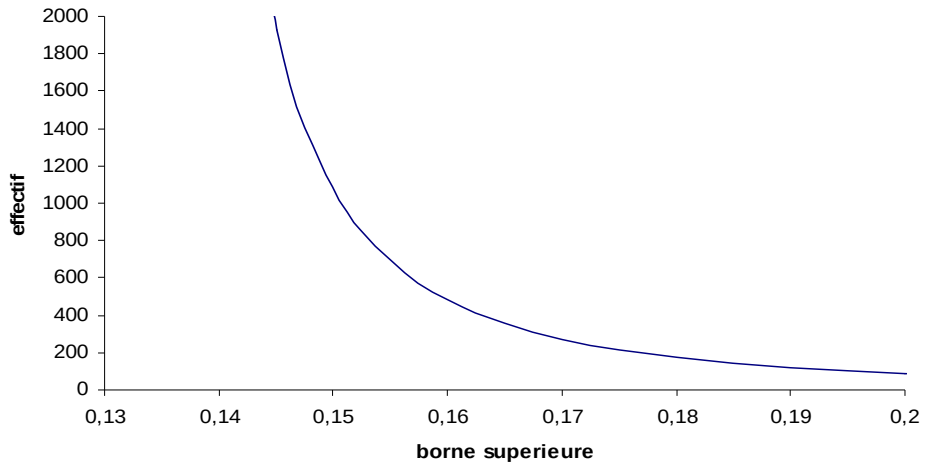


Figure 10 : Représentation de la taille nécessaire en fonction de la borne supérieure de l'intervalle de fluctuation asymptotique

2. *Problème de la surréservation (surbooking)*

Une compagnie aérienne possède des A340 (longs courriers) d'une capacité de 300 places.

Cette compagnie a vendu n billets pour le vol 2012.

La probabilité pour qu'un acheteur se présente à l'embarquement est p et les comportements des acheteurs sont indépendants les uns des autres.

On note X_n la variable aléatoire désignant le nombre d'acheteurs d'un billet se présentant à l'embarquement.

La compagnie cherche à optimiser le remplissage de l'avion en vendant éventuellement plus de places que la capacité totale de l'avion (surréservation ou surbooking) soit ici $n > 300$.

Comme il y a évidemment un risque que le nombre de passagers munis d'un billet se présentant à l'embarquement excède 300, la compagnie veut maîtriser ce risque.

1. Déterminer la loi de X_n.
2. On suppose que $0{,}5 \leq p \leq 0{,}95$. Écrire l'intervalle de fluctuation asymptotique I_n de $\frac{X_n}{n}$ au seuil de 0,95.
3. Montrer que si $I_n \subset \left[0, \frac{300}{n}\right]$ alors la probabilité que le nombre de passagers se présentant à l'embarquement excède 300 est proche de 0,05.
4. On cherche à déterminer la valeur de n maximale permettant de satisfaire la condition de l'inclusion $I_n \subset \left[0, \frac{300}{n}\right]$.

   a. Montrer que $I_n \subset \left[0, \frac{300}{n}\right] \Rightarrow pn + 1{,}96\sqrt{n}\sqrt{p(1-p)} - 300 \leq 0.$

   b. On pose $f(x) = px + 1{,}96\sqrt{x}\sqrt{p(1-p)} - 300$.

   Montrer qu'il existe un entier n_0 unique tel que si $n \leq n_0$ alors $f(n) \leq 0$ et si $n > n_0$ alors $f(n) > 0$.

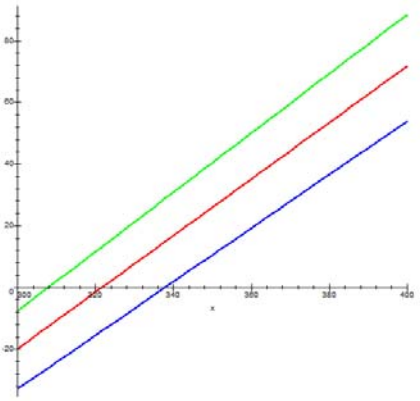
   c. Tracer la courbe représentative de f pour les valeurs $p = 0{,}85$; $p = 0{,}9$; $p = 0{,}95$.

   d. Déterminer à la calculatrice les valeurs de n_0 pour $p = 0{,}85$; $p = 0{,}9$; $p = 0{,}95$.

Solution

1. X_n suit une loi binomiale de paramètres n et p.
2. Comme $n > 300$ et $0,5 \leq p \leq 0,95$ on a $np \geq 5$ et $n(1-p) \geq 5$ on peut utiliser l'intervalle de fluctuation asymptotique au seuil de 0,95 :
$$I_n = \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \right].$$
3. Si $I_n \subset \left[0, \frac{300}{n}\right]$ alors $\mathrm{P}\left(X_n > 300\right) \leq \mathrm{P}\left(\frac{X_n}{n} \notin I_n\right)$.
Comme $\mathrm{P}\left(\frac{X_n}{n} \notin I_n\right) \approx 0,05$ alors on peut dire que $\mathrm{P}\left(X_n > 300\right)$ est proche également de 0,05 voire inférieur (l'événement $(X_n > 300)$ étant inclus dans la partie droite du complémentaire de I_n on pourrait vérifier avec le tableur que sa probabilité est en fait inférieure à 0,05 pour $n \geq 300$ et $0,5 \leq p \leq 0,95$).

4. a. $I_n \subset \left[0, \frac{300}{n}\right] \Rightarrow p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{300}{n} \Rightarrow np + 1,96\sqrt{n}\sqrt{p(1-p)} - 300 \leq 0$
b. En posant $y = \sqrt{x}$, on se ramène à une inéquation du second degré que l'on résout pour $x \geq 300$.
Les solutions de l'inéquation $f(x) \leq 0$ sont donc les réels de l'intervalle $\left[300, x_0\right]$ où
$$x_0 = \left(\frac{-1,96\sqrt{p(1-p)} + \sqrt{1200p + 1,96^2 p(1-p)}}{2p} \right)^2.$$
L'entier n_0 cherché est la partie entière de x_0.
c. $p = 0,85$ en bleu, $p = 0,9$ en rouge, $p = 0,95$ en vert.



d. Pour p =0,85 on trouve $n_0 = 337$,
Pour p = 0,9 on trouve $n_0 = 321$,
Pour p = 0,95 on trouve $n_0 = 307$.

3. *Echantillon représentatif d'une population pour un sondage*

La première partie de l'activité proposée page 29 peut être traitée dans ce cadre.
En vue de conduire une enquête sur certaines caractéristiques physiologiques d'une population, un échantillon de personnes a été sélectionné et on souhaite en conforter la représentativité.

E. Intervalle de fluctuation simplifie donné en seconde

On reprend les notations du paragraphe C. Dans le cas où $\alpha = 0{,}05$, on a $u_\alpha \approx 1{,}96$.

La fonction $p \mapsto p(1-p)$ admet un maximum pour $p = \frac{1}{2}$ égal à $\frac{1}{4}$.

On peut donc majorer $u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}$ par $\frac{1}{\sqrt{n}}$.

On en déduit que l'intervalle $J_n = \left[p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$ (approximation de l'intervalle I_n liée à l'approximation de $u_{0,05}$ par 1,96) est inclus dans l'intervalle $\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}} \right]$ et donc on a :

$$\mathrm{P}\left(\frac{X_n}{n} \in J_n\right) \leq \mathrm{P}\left(p - \frac{1}{\sqrt{n}} \leq \frac{X_n}{n} \leq p + \frac{1}{\sqrt{n}}\right)$$

Cette inégalité prouve que l'intervalle $\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}} \right]$ est un intervalle de fluctuation asymptotique à un seuil au moins égal à celui de l'intervalle J_n (proche de 0,95) et justifie le résultat énoncé en seconde sous une forme simplifiée, ne prenant pas en compte le caractère asymptotique.

Compte tenu du caractère asymptotique de l'intervalle de fluctuation $\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}} \right]$, il serait inexact d'affirmer que la probabilité que la variable aléatoire $\frac{X_n}{n}$ prenne ses valeurs dans cet intervalle est supérieure à 0,95 pour toute valeur de *n,* même lorsque les conditions usuelles d'approximation sont vérifiées. Ce point a déjà été clairement explicité dans le document ressource de la classe de première. Nous le reprenons ici.

On peut visualiser ci-dessous les valeurs des probabilités $\mathrm{P}\left(p - \frac{1}{\sqrt{n}} \leq \frac{X_n}{n} \leq p + \frac{1}{\sqrt{n}}\right)$ suivant les valeurs de *p* et de *n* et constater que le résultat énoncé en classe de seconde, s'il n'est pas tout à fait exact, fournit néanmoins en général une probabilité très proche de 0,95, ce qui justifie son utilisation dans la pratique.

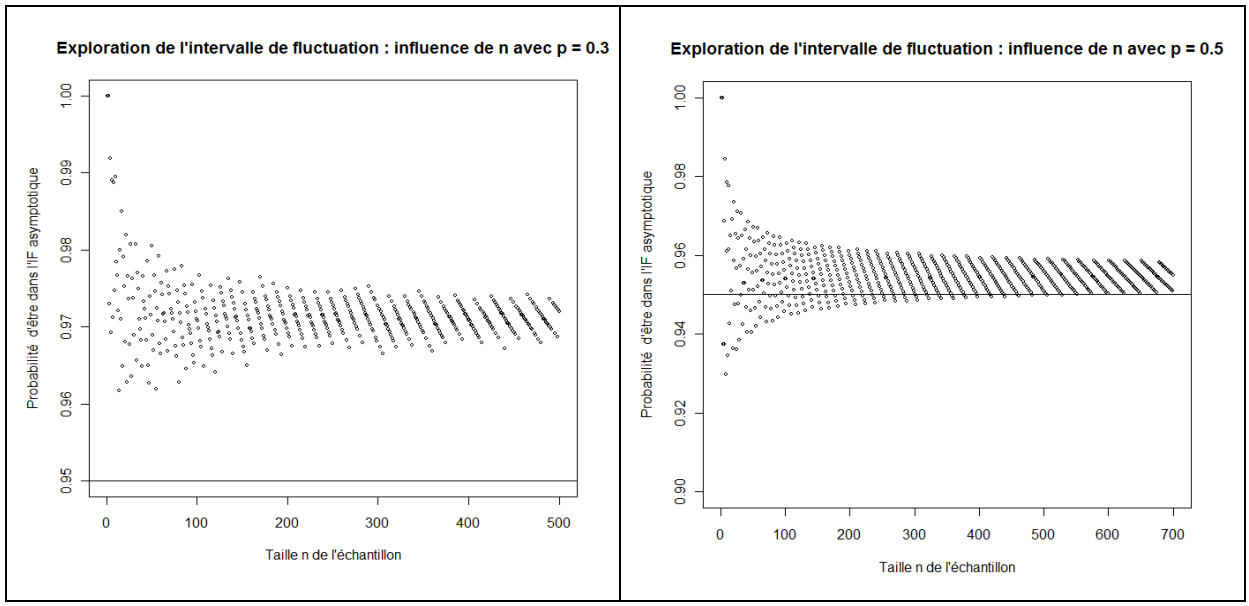


Figure 11 : Visualisation des probabilités de l'intervalle de fluctuation de seconde pour $p = 0,3$ (figure de gauche) et $p = 0,5$ (figure de droite).

Document associé : intervalle de fluctuation seconde.r

On peut constater que :

pour p =0,3 $\quad P(p-\frac{1}{\sqrt{n}} \le \frac{X_n}{n} \le p+\frac{1}{\sqrt{n}}) \ge 0{,}95$ semble vérifiée pour tout entier n,

pour p =0,5 $\quad P(p-\frac{1}{\sqrt{n}} \le \frac{X_n}{n} \le p+\frac{1}{\sqrt{n}}) \ge 0{,}95$ semble être vérifiée pour tout entier $n \ge 600$.

Cela conduit au résultat suivant :

Théorème

Si la variable aléatoire X_n suit la loi $\mathcal{B}(n, p)$ alors, pour tout p dans]0, 1[, il existe un entier n_0 tel que si $n \ge n_0$ alors $P(p-\frac{1}{\sqrt{n}} \le \frac{X_n}{n} \le p+\frac{1}{\sqrt{n}}) \ge 0{,}95$.

Démonstration

Pour une variable binomiale X_n de paramètres n et p, le théorème de Moivre-Laplace prouve que, en notant Z_n la variable centrée réduite associée à X_n, la limite de $a_n = P(-2 \le Z_n \le 2)$ est égale à $2P(Z \le 2) - 1$ où Z suit une loi $\mathcal{N}(0,1)$.

Or on a $L = 2P(Z \le 2) - 1 \ge 0{,}9544$.

Donc, pour $\varepsilon < 0{,}004$, si on considère l'intervalle ouvert $]L-\varepsilon, L+\varepsilon[$ contenant L, il existe un entier n_0 tel que si $n \ge n_0$, on a : $a_n \in]L-\varepsilon, L+\varepsilon[$ donc $a_n \ge 0{,}95$ puisque $L - \varepsilon \ge 0{,}9504$.

Or $a_n = P\left(p-\frac{2}{\sqrt{n}}\sqrt{p(1-p)} \le \frac{X_n}{n} \le p+\frac{2}{\sqrt{n}}\sqrt{p(1-p)}\right)$ ce qui donne, en majorant $p(1-p)$ par 1/4, un intervalle de fluctuation plus large donc de probabilité supérieure ou égale à a_n.

Donc pour tout entier $n \ge n_0$, on a : $P\left(p-\frac{1}{\sqrt{n}} \le \frac{X_n}{n} \le p+\frac{1}{\sqrt{n}}\right) \ge 0{,}95$.

Exemple d'activité

Selon la valeur de p, la valeur de n_0 peut varier considérablement.

Il est d'ailleurs difficile de déterminer avec certitude cette valeur de n_0. On peut cependant donner des valeurs de n_0 grâce à un algorithme de calcul.

P	0,35	0,36	0,37	0,38	0,39	0,4	0,41	0,42	0,43	0,44	0,45	0,46	0,47	0,48	0,49	0,5
n_0	31	30	36	64	56	81	90	120	143	209	271	288	304	399	399	529

On peut remarquer que la plus grande valeur de n_0 est atteinte pour $p = 1/2$. C'est effectivement pour cette valeur que la fluctuation est la plus importante puisque la variance est maximale pour cette valeur de p.

Algorithme Algobox :	**Programme SCILAB :**
Remarque : Cet algorithme ne permet d'obtenir n_0 que pour des valeurs de p entre 0 et 0,39 car Algobox ne calcule pas de valeurs avec la loi binomiale pour des valeurs de n supérieures à 70. Or pour $p \geq 0,4$ la valeur de n_0 est supérieure à 80. Pour le cas général il faut utiliser les logiciels **R** ou Scilab par exemple. Document associé : recherche du n0.alg	Remarque : La ligne 12 enlève 1 à la première valeur de n pour laquelle Fsup – Finf < 0,95. Or, on cherche la plus petite valeur de n_0 à partir de laquelle FSup-Finf ≥ 95 donc on doit faire afficher n+2. Document associé : recherche du n0.sce

```
VARIABLES
  n EST_DU_TYPE NOMBRE
  p EST_DU_TYPE NOMBRE
  Inf EST_DU_TYPE NOMBRE
  Sup EST_DU_TYPE NOMBRE
  Finf EST_DU_TYPE NOMBRE
  Fsup EST_DU_TYPE NOMBRE
  i EST_DU_TYPE NOMBRE
  j EST_DU_TYPE NOMBRE
  n0 EST_DU_TYPE NOMBRE
DEBUT_ALGORITHME
  LIRE p
  n PREND_LA_VALEUR 69
  Finf PREND_LA_VALEUR 0
  Fsup PREND_LA_VALEUR 1
  TANT_QUE (Fsup-Finf>=0.95) FAIRE
    DEBUT_TANT_QUE
    Inf PREND_LA_VALEUR floor(n*p-sqrt(n))
    Sup PREND_LA_VALEUR floor(n*p+sqrt(n))
    Finf PREND_LA_VALEUR 0
    Fsup PREND_LA_VALEUR 0
    POUR i ALLANT_DE 0 A Inf
      DEBUT_POUR
      Finf PREND_LA_VALEUR Finf+ALGOBOX_LOI_BINOMIALE(n,p,i)
      FIN_POUR
    POUR j ALLANT_DE 0 A Sup
      DEBUT_POUR
      Fsup PREND_LA_VALEUR Fsup+ALGOBOX_LOI_BINOMIALE(n,p,j)
      FIN_POUR
    n PREND_LA_VALEUR n-1
    FIN_TANT_QUE
  n0 PREND_LA_VALEUR n+2
  AFFICHER n0
FIN_ALGORITHME
```

```
1
2  p=input("p=");
3  n=1000;
4
5  Finf=0;
6  Fsup=1;
7  while Fsup-Finf>=0.95
8    inf=floor(n*p-sqrt(n));
9    sup=floor(n*p+sqrt(n));
10   Fsup=cdfbin("PQ",sup,n,p,1-p);
11   Finf=cdfbin("PQ",inf,n,p,1-p);
12   n=n-1;
13 end
14 disp(n+2);
15
```

Exemples d'exercices

1. Les enfants sont dits prématurés lorsque la durée gestationnelle est inférieure ou égale à 259 jours. La proportion de ces naissances est de 6%. Des chercheurs suggèrent que les femmes ayant eu un travail pénible pendant leur grossesse sont plus susceptibles d'avoir un enfant prématuré que les autres. Il est décidé de réaliser une enquête auprès d'un échantillon aléatoire de 400 naissances correspondant à des femmes ayant eu pendant leur grossesse un travail pénible. Les chercheurs décident a priori que si la proportion d'enfants nés prématurés dans cet échantillon est supérieure à la borne supérieure de l'intervalle de fluctuation asymptotique au seuil de 0,95 alors leur hypothèse sera acceptée. Finalement le nombre d'enfants prématurés est de 50. Quelle est donc la conclusion ?

Solution : Sous l'hypothèse que la proportion de prématurés dans l'échantillon est la même que dans la population générale, on détermine l'intervalle de fluctuation asymptotique au seuil 0,95.

$$\left[0,06-1,96\times\frac{\sqrt{0,06\times0,94}}{\sqrt{400}};0,06+1,96\times\frac{\sqrt{0,06\times0,94}}{\sqrt{400}}\right]=\left[0,037;0,083\right]$$

On calcule la valeur observée de proportion de prématurés dans l'échantillon et on obtient 0,125. Cette valeur n'appartient pas à l'intervalle de fluctuation asymptotique au seuil de 95%, donc avec la règle de décision choisie, on rejette l'hypothèse posée. Les chercheurs concluent donc que la proportion d'enfants prématurés est plus élevée chez les femmes ayant eu un travail pénible pendant leur grossesse.

2. 1) Vérifier que l'intervalle [–2,576 , 1,696] peut être considéré comme un intervalle de fluctuation au seuil de 95% d'une variable X suivant une loi $\mathcal{N}(0,1)$ (c'est-à-dire que $P(X \in I) \geq 0{,}95$).

2) Montrer qu'il existe une valeur a minimale telle que l'intervalle $[-a, a]$ soit un intervalle de fluctuation au seuil de 95% de X. En donner une valeur approchée à 10^{-2} près.

3) Montrer qu'il existe un unique réel b tel que $P(-2 \leq X \leq -2+b) = 0{,}95$.

Prouver que $b < a + 2$ où a est la valeur de la question 2).

Déterminer une valeur approchée de b à 10^{-2} près.

4) Montrer qu'il n'existe aucun réel c tel que $P(-1 \leq X \leq -1+c) = 0{,}95$.

Solution

1. Avec une calculatrice : $P(-2{,}576 \leq X \leq 1{,}696) \approx 0{,}95006 > 0{,}95$
2. $P(-a \leq X \leq a) = 2\int_0^a f(t)\mathrm{d}t = F(a)$

   On étudie la fonction F. Sa dérivée est $F'(x) = \frac{2}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} > 0$

   Donc F est strictement croissante de $[0, +\infty[$ vers $[0,1[$. Il existe donc un réel a unique tel que $F(a) = 0{,}95$.

   D'après ce qui a déjà été vu, a vaut environ 1,96.
3. $P(-2 \leq X \leq -2+x) = \int_{-2}^{-2+x} f(t)\mathrm{d}t = H(x)$.

   Cette fonction est strictement croissante car sa dérivée est égale à $\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(x-2)^2}$.

   Comme $H(4) = \int_{-2}^{2} f(t)\mathrm{d}t > F(1{,}96)$, il existe b unique tel que $H(b) = 0{,}95$.

   On a $0{,}95 = H(b) = \int_{-2}^{-a} f(t)\mathrm{d}t + 0{,}95 + \int_a^{-2+b} f(t)\mathrm{d}t$ donc $\int_a^{-2+b} f(t)\mathrm{d}t = -\int_{-2}^{-a} f(t)\mathrm{d}t < 0$

   donc $a > -2 + b$.

   A l'aide de la calculatrice, en utilisant la fonction qui à x fait correspondre $P(-2 \leq X \leq -2+x)$, on trouve $3{,}92 < b < 3{,}93$.

   On peut vérifier que l'intervalle $[-a, a]$ est plus court que l'intervalle $[-2, -2+b]$.
4. $\int_{-1}^{-1+c} f(t)\mathrm{d}t = \int_{-1}^{0} f(t)\mathrm{d}t + \int_0^{-1+c} f(t)\mathrm{d}t < \int_{-1}^{0} f(t)\mathrm{d}t + \frac{1}{2}$.

   Or $\int_{-1}^{0} f(t)\mathrm{d}t < 0{,}35$ donc $\int_{-1}^{-1+c} f(t)\mathrm{d}t < 0{,}85$ pour tout c.

Remarque

On peut démontrer plus généralement que l'intervalle de fluctuation au seuil de 95% centré en 0 est celui d'amplitude minimale.

V. Intervalle de confiance

A. Introduction

Il est souvent difficile pour des raisons à la fois financières et logistiques de pouvoir recueillir des données sur la population toute entière. Le plus souvent, on se contente de travailler sur un échantillon, c'est à dire une fraction ou sous-ensemble de cette population. Ceci présente bien sûr des avantages en termes de faisabilité et de coût, mais impose des contraintes pour que l'information recueillie au niveau de l'échantillon (estimation) soit la plus proche possible de celle de la population entière (paramètre). La démarche pratique est donc la suivante :

- on sélectionne un échantillon de la population que l'on étudie, on appelle cela l'échantillonnage.
- On vérifie, selon les cas, à partir d'intervalles de fluctuation que l'échantillon ainsi obtenu est « représentatif » de la population pour des critères qui sont connus dans la population.

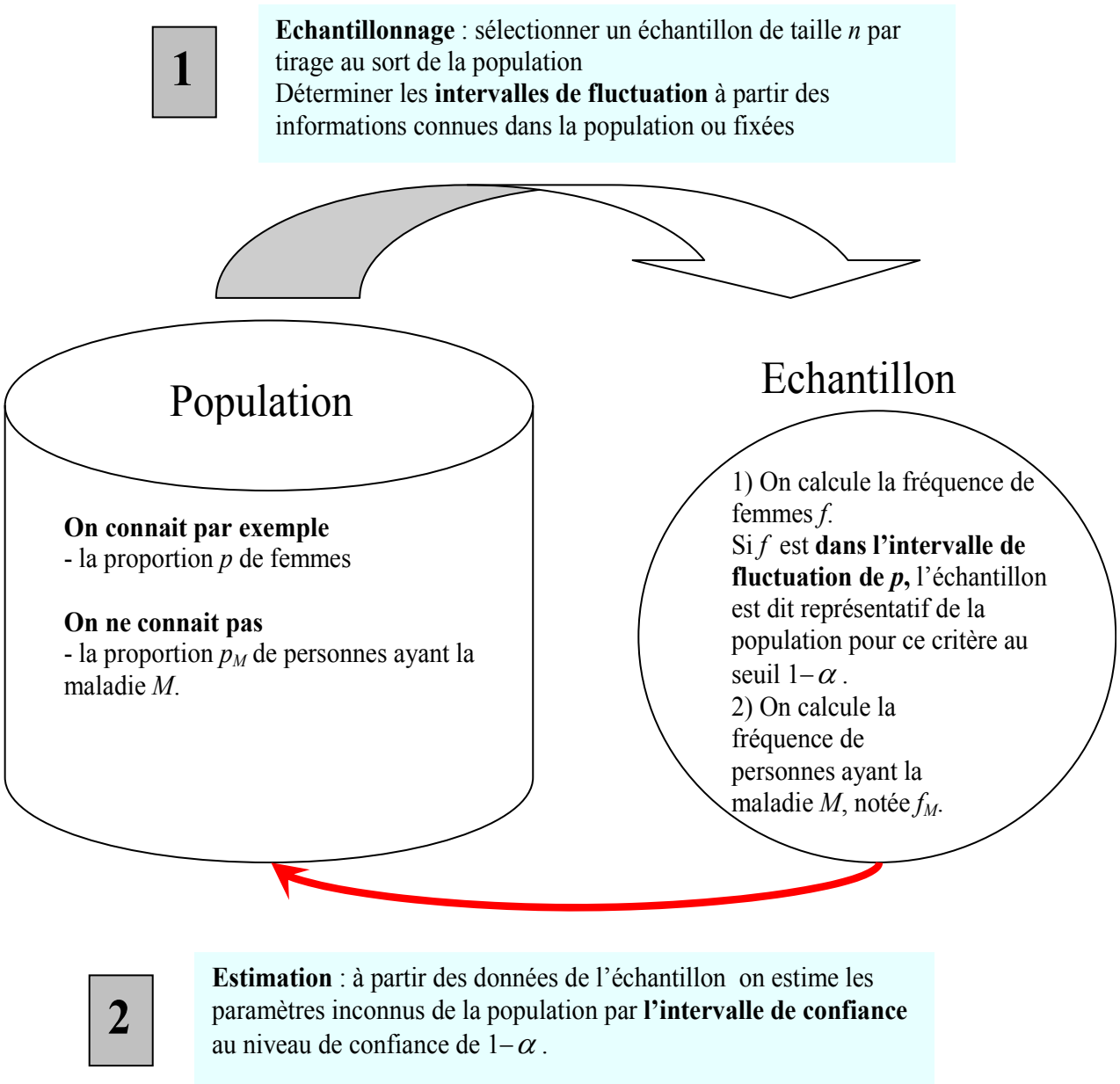


Figure 13 : Principe de l'échantillonnage

La notion d'échantillon représentatif est une question délicate, en particulier lorsqu'elle concerne des personnes dans le cadre d'un sondage. Elle l'est clairement moins lorsqu'il s'agit d'un échantillon de pièces dans une chaine de fabrication. Cette notion d'échantillon représentatif est évoquée ici afin de contextualiser un peu l'activité mais ne constitue en aucun cas un objectif du programme.

Il convient également de souligner que, dans les sondages, les tirages sont pour la plupart effectués sans remise mais peuvent s'apparenter à des tirages avec remise dès que la taille de l'échantillon est petite devant la taille de la population totale, ce qui est le cas dans les sondages classiques.

On peut d'ailleurs observer que dans le cas contraire, l'intérêt de ne questionner qu'un échantillon diminue.

L'activité qui suit propose une situation de sondage simplifiée qui ne correspond pas exactement aux techniques réelles de sondage. Quelques compléments d'informations sur les techniques de sondage et les questions qu'elles soulèvent figurent en annexe 3, à titre informatif.

Activité

On souhaite estimer la prévalence du surpoids dans une ville V, c'est-à-dire la proportion de personnes ayant une masse trop importante par rapport à leur taille. Pour cela 460 personnes ont été sélectionnées de manière aléatoire à partir de la liste des logements connue par la municipalité, c'est-à-dire que le fait d'avoir été sélectionné pour participer à l'étude est uniquement dû au hasard. On admet que cette procédure permet d'assimiler la sélection des personnes interrogées à un schéma de Bernoulli.

Un enquêteur s'est déplacé au sein de chaque logement après avoir convenu d'un rendez-vous afin de recueillir les informations nécessaires à l'enquête.

1° Dans un premier temps, l'enquêteur va s'assurer que l'échantillon est représentatif de la population qu'on étudie sur des informations qu'on peut vérifier et qui sont en lien avec le critère étudié. Dans le cas présent on peut connaître par exemple la proportion d'hommes et de femmes dans la population de la ville, ainsi que la répartition selon l'âge en demandant à la municipalité qui se référera aux informations du recensement. Parallèlement on peut comptabiliser le nombre d'hommes et de femmes dans l'échantillon ainsi que la répartition selon l'âge.

	Homme	Femme	Total
Echantillon	200	260	460

	< 60 ans	> 60 ans	Total
Echantillon	352	108	460

On sait que, dans la population, il y a 46% d'hommes et 20% de personnes de plus de 60 ans.

a) Déterminer l'intervalle de fluctuation asymptotique au seuil 0,95 de la variable aléatoire « proportion de femmes » dans un échantillon aléatoire de taille 460 sélectionné au sein de la population de cette ville.
b) Calculer la proportion de femmes dans l'échantillon et vérifier si cette valeur appartient à l'intervalle de fluctuation.
c) Déterminer l'intervalle de fluctuation asymptotique au seuil 0,95 de la variable aléatoire « proportion de personnes âgées de plus de 60 ans » dans un échantillon aléatoire de taille 460 sélectionné au sein de la population de cette ville.
d) Calculer la proportion de personnes de plus de 60 ans dans l'échantillon et vérifier si cette valeur appartient à l'intervalle de fluctuation.
e) Si pour chacune des variables, genre et âge, l'intervalle de fluctuation asymptotique au seuil de 95% contient la valeur de l'échantillon on considère que l'échantillon est représentatif de la population pour cette information. Quelle est donc la conclusion pour le cas étudié ici ?

2° La première étape de ce travail a donc été de sélectionner un échantillon qui soit accepté comme « représentatif » de la population. Ainsi les informations qui seront obtenues à partir de cet échantillon seront généralisables, avec un certain nombre de précautions, à l'ensemble de la population dont il est extrait. Dans le cas de l'étude présentée ici, on souhaite estimer la proportion de personnes en surpoids ; pour cela il est tout d'abord important de définir le surpoids. La définition du surpoids donnée par l'OMS (Organisation Mondiale de la Santé) est la suivante : une personne est considérée en surpoids si son IMC (Indice de masse corporelle) est supérieur à 25. L'IMC se calcule de la manière suivante : masse en kg/(taille en m)².

La proportion de personnes en surpoids dans l'échantillon étudié est de 29,5%. Comme il s'agit d'un calcul réalisé à partir des données d'un échantillon on sait que cette valeur ne correspond pas exactement à la valeur de la prévalence dans la population, car si nous avions pris un autre échantillon nous aurions obtenu une autre valeur. Pour cette raison il est nécessaire de communiquer un intervalle qui sera obtenu à partir des informations observées et pour lequel on puisse dire avec un « niveau de confiance » supérieur à 0,95 qu'il contient la vraie valeur de la prévalence du surpoids dans la ville. Si f est la fréquence observée dans l'échantillon une expression de cet intervalle, qui sera appelé intervalle de confiance, est $\left[f-\frac{1}{\sqrt{n}}, f+\frac{1}{\sqrt{n}}\right]$ où n est la taille de l'échantillon. Ce résultat, évoqué en classe de seconde, prend tout son sens en terminale et est démontré en terminale S.

Déterminer un intervalle de confiance au niveau de confiance de 95%.

Solution :

1° a) L'intervalle de fluctuation asymptotique au seuil de 95% est déterminé par :

$$\left[p-1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p+1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \approx [0{,}49\ ;\ 0{,}59].$$

b) La proportion de femmes dans l'échantillon est égale à 56,5%, cette valeur appartient à l'intervalle de fluctuation calculé ci-dessus.

c) On obtient avec un calcul analogue à celui de la question a) l'intervalle $[0{,}16\ ;\ 0{,}24]$

d) La proportion de plus de 60 ans dans l'échantillon est égale à 23,4%, cette valeur appartient à l'intervalle de fluctuation calculée ci-dessus.

e) On considère que l'échantillon observé est représentatif de la population pour les deux critères retenus (genre et âge).

La représentativité sur deux critères ne signifie évidemment pas la représentativité sur tous les critères et dans tous les cas, il est peu vraisemblable qu'un échantillon de 460 sujets soit représentatif pour tous les critères. Les résultats obtenus sur un échantillon ne peuvent pas remplacer les résultats exacts d'un recensement. Cependant la vérification précédente sur des critères importants permet de considérer que l'échantillon retenu est structuré comme la population étudiée, au regard de certains critères.

2° L'intervalle de confiance calculé au niveau de confiance de 95% est donc :

$$\left[0{,}295-\frac{1}{\sqrt{460}}; 0{,}295+\frac{1}{\sqrt{460}}\right] \approx [0{,}25; 0{,}34]$$

Cet intervalle fournit une estimation par intervalle de la prévalence du surpoids dans la ville étudiée.

B. Principe général de l'intervalle de confiance

Étant donné un paramètre *p*, ici une proportion inconnue, d'une population, la procédure d'estimation consiste à utiliser les informations recueillies dans un échantillon sélectionné de manière aléatoire pour obtenir une valeur de la variable aléatoire fréquence $F_n = \frac{X_n}{n}$ destinée à fournir une estimation de *p*. Mais on sait que cette estimation va varier d'un échantillon à l'autre, de par la **fluctuation d'échantillonnage,** autour de *p*. Il est donc nécessaire d'apprécier l'incertitude en fournissant une estimation par intervalle, appelé **intervalle de confiance de *p***. Cet intervalle est obtenu en fonction d'un coefficient lié au niveau de confiance que l'on accorde à cette estimation.

Lorsque $n \geq 30$ et $np \geq 5$ et $n(1-p) \geq 5$, la formule $\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}\right]$ fournit un intervalle de fluctuation de $F_n = \frac{X_n}{n}$ au seuil 0,95.

Supposons que *p* soit inconnu . On peut approcher *p* par la proportion *f* obtenue par les données de l'échantillon et déterminer un intervalle de confiance de *p* au niveau de confiance 0,95.

Selon le théorème du paragraphe IV-C, on sait que, pour *n* suffisamment grand, on a :

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

Comme $p - \frac{1}{\sqrt{n}} \leq F_n \leq p + \frac{1}{\sqrt{n}}$ équivaut à $F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}}$, on peut également écrire $\mathrm{P}\left(F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}}\right) \geq 0{,}95$, ce qui peut se traduire en disant que :

l'intervalle aléatoire $\left[F_n - \frac{1}{\sqrt{n}}, F_n + \frac{1}{\sqrt{n}}\right]$ a une probabilité au moins égale à 0,95 de contenir *p*.

À partir de l'intervalle aléatoire $\left[F_n - \frac{1}{\sqrt{n}}, F_n + \frac{1}{\sqrt{n}}\right]$ on obtient, en effectuant le tirage d'un échantillon, une *réalisation* de cet intervalle qui fournit alors un intervalle numérique de la forme $\left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}}\right]$.

Si l'on fait un très grand nombre de tirages, on sait que théoriquement on devrait[12] avoir pour au plus 5% d'entre eux des intervalles ne contenant pas la proportion inconnue *p*.

C. Définition

> Un intervalle de confiance pour une proportion *p* à un niveau de confiance $1-\alpha$ est la réalisation, à partir d'un échantillon, d'un intervalle aléatoire contenant la proportion *p* avec une probabilité supérieure ou égale à $1-\alpha$. Cet intervalle aléatoire est déterminé à partir de la variable aléatoire $F_n = \frac{X_n}{n}$ qui, à tout échantillon de taille *n,* associe la fréquence.

Le cas particulier où $1-\alpha = 0{,}95$ est le seul au programme.

[12] Il s'agit toujours d'un nombre fini de réalisations et il peut y avoir plus de 5% d'entre elles qui ne contiennent pas *p*.

***Remarque* 1**

En réalisant le tirage d'un échantillon, on obtient un intervalle de confiance de la forme $\left[f-\frac{1}{\sqrt{n}}, f+\frac{1}{\sqrt{n}}\right]$ de la proportion inconnue p à un niveau de confiance de 0,95.

Ainsi, à chaque tirage d'un échantillon, on obtient un intervalle de confiance différent.

***Remarque* 2**

Un intervalle de confiance étant un intervalle numérique, il est incorrect de conclure la détermination d'un intervalle de confiance par une phrase du type « p a une probabilité de 0,95 d'être entre $f-\frac{1}{\sqrt{n}}$ et $f+\frac{1}{\sqrt{n}}$ » car il n'y a plus d'aléatoire à ce stade. Il est en revanche convenable d'écrire :

« L'intervalle $\left[f-\frac{1}{\sqrt{n}}, f+\frac{1}{\sqrt{n}}\right]$ est un intervalle de confiance de la proportion inconnue p au niveau de confiance 0,95 ».

D. Intervalle de fluctuation ou intervalle de confiance : lequel utiliser ?

Règle générale

On utilise un intervalle de fluctuation lorsque la proportion p dans la population est **connue ou si l'on fait une hypothèse sur sa valeur.**

On utilise un intervalle de confiance lorsque l'on veut estimer une proportion **inconnue** dans une population.

***Exemple* 1**

Test de conformité d'une proportion : on veut déterminer si la proportion observée dans un échantillon est conforme à une valeur de référence connue dans la population.

Sous l'hypothèse que l'échantillon est issu d'un tirage aléatoire correspondant à un schéma de Bernoulli (tirage avec remise ou s'y apparentant), la variable fréquence F_n appartient à un intervalle de fluctuation avec une probabilité déterminée.

En fonction de l'appartenance ou non de la fréquence observée à cet intervalle, on peut prendre une *décision* concernant la conformité de l'échantillon.

Si les conditions d'utilisation sont réunies, on détermine l'intervalle de fluctuation asymptotique, sinon on a recours à un intervalle de fluctuation calculé avec la loi binomiale.

***Exemple* 2**

Estimation d'une proportion inconnue p grâce à un échantillon aléatoire

On se place dans le cas où l'échantillon comporte au moins 30 éléments afin de pouvoir utiliser l'intervalle de confiance au programme.

Si la fréquence observée f est telle que $nf \geq 5$ et $n(1-f) \geq 5$, on considère qu'on peut conclure qu'un intervalle de confiance de p au niveau de confiance 0,95 est $\left[f-\frac{1}{\sqrt{n}}, f+\frac{1}{\sqrt{n}}\right]$.

Le tableau suivant récapitule ce qui est au programme de chaque classe du lycée.

	Intervalle de fluctuation ***p connue***	**Intervalle de confiance** ***p inconnue***
SECONDE	$n \geq 25$ et $0,2 \leq p \leq 0,8$, seuil 95% $\left[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}\right]$	Sensibilisation
PREMIÈRE	Avec la loi binomiale	
TERMINALE	$n \geq 30$ et $np \geq 5$ et $n(1-p) \geq 5$ Asymptotique au seuil $1-\alpha$ $I_n = \left[p-u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}, p+u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$	Au niveau de confiance 95% $\left[f-\frac{1}{\sqrt{n}}, f+\frac{1}{\sqrt{n}}\right]$

En terminale autre que S, $\alpha = 0,05$ donc $u_\alpha = 1,96$.

E. Autre intervalle de confiance

Il existe d'autres manières de déterminer un intervalle de confiance d'une proportion.
Dans les commentaires du programme, il est signalé que dans d'autres champs disciplinaires on utilise l'intervalle $\left[f-1,96\frac{\sqrt{f(1-f)}}{\sqrt{n}}, f+1,96\frac{\sqrt{f(1-f)}}{\sqrt{n}}\right]$.

La justification de cet intervalle est hors programme.

Exemple

Pour un niveau de confiance de 0,95, on a $u_\alpha \approx 1,96$. Si sur un échantillon de taille 100 on observe une valeur de la fréquence égale à 0,44, l'intervalle de confiance de p au niveau 0,95 obtenu avec la formule précédente est [0,343 ; 0,537].

L'intervalle $\left[f-\frac{1}{\sqrt{n}}, f+\frac{1}{\sqrt{n}}\right]$ donne [0,34 ; 0,54].

F. Étude de la longueur de l'intervalle de fluctuation et conséquence pour l'intervalle de confiance

L'intervalle de fluctuation asymptotique $I_n = \left[p-u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}, p+u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ a pour longueur $2u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}$. Donc pour α et n fixés, la longueur de I_n varie comme $\sqrt{p(1-p)}$. Elle est donc maximale quand $p = \frac{1}{2}$ et d'autant plus faible que p est proche de 0 ou de 1.

Quelques valeurs de la longueur $2u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}$ pour $n = 1000$:

	$p = 0,1$	$p = 0,3$	$p = 0,4$	$p = 0,5$
$\alpha = 0,05$	0,037	0,057	0,061	0,062
$\alpha = 0,01$	0,049	0,075	0,08	0,082

Conséquence pour l'intervalle de confiance

Si on cherche à estimer par intervalle, au niveau de confiance 0,95, une valeur de p dont on sait qu'elle est plutôt proche de 0,5 (cas du second tour de l'élection présidentielle), on a un intervalle de confiance, appelé dans ce cas fourchette de sondage, d'amplitude proche de 0,06.

Si on cherche à estimer une valeur de p sans doute inférieure à 0,1 (cas des petits candidats du premier tour), on a une fourchette d'amplitude proche de 0,04.

On constate sur le tableau précédent que, n étant fixé, l'augmentation du niveau de confiance augmente simultanément la longueur de l'intervalle de confiance, ce qui est un résultat général facile à justifier (et à concevoir).

G. Détermination de la taille minimale de l'échantillon pour avoir une précision donnée

On étudie d'abord la taille minimale de l'échantillon pour avoir une longueur donnée a de l'intervalle de fluctuation pour un seuil ou un niveau de confiance fixé.

1) Avec l'intervalle asymptotique de seconde (donc $\alpha = 0{,}05$ et pour tout p)

On cherche n tel que. $\frac{2}{\sqrt{n}} \leq a$ ce qui équivaut à $n \geq \frac{4}{a^2}$.

Quelques valeurs :

Valeur de a	0,06	0,04	0,02	0,01
Valeur de n	1112	2500	10000	40000

Conséquence pour la taille de l'échantillon nécessaire pour obtenir une amplitude de l'intervalle de confiance fixée

On a : $\mathrm{P}\left(p - \frac{1}{\sqrt{n}} \leq \frac{X_n}{n} \leq p + \frac{1}{\sqrt{n}}\right) \geq 0{,}95 \Leftrightarrow \mathrm{P}\left(\frac{X_n}{n} - \frac{1}{\sqrt{n}} \leq p \leq \frac{X_n}{n} + \frac{1}{\sqrt{n}}\right) \geq 0{,}95$

L'amplitude de l'intervalle de fluctuation est évidemment la même que celle de l'intervalle de confiance.

Donc, avec un niveau de confiance de 0,95, pour obtenir un intervalle de confiance d'amplitude 0,06, il faut un échantillon de taille 1112 au moins.

2) Avec l'intervalle asymptotique $I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$

On cherche n tel que $2u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq a$ ce qui équivaut à $n \geq \frac{4{u_\alpha}^2 p(1-p)}{a^2}$.

Donnons quelques valeurs :

Pour $p = 0{,}5$

Valeur de a	0,06	0,04	0,02	0,01
Valeur de n si $\alpha = 0{,}05$	1067	2401	9604	38416
Valeur de n si $\alpha = 0{,}01$	1849	4161	16641	66664

Pour $p = 0{,}1$

Valeur de a	0,06	0,04	0,02	0,01
Valeur de n si $\alpha = 0{,}05$	385	865	3458	13830
Valeur de n si $\alpha = 0{,}01$	666	1498	5991	23964

H. Applications

1. Exemple de détermination d'un intervalle de confiance

Prenons un cas très classique : un sondage politique précédant le premier tour d'une élection présidentielle.
Le 18 avril 2002, l'institut IPSOS[13] effectue un sondage dans la population en âge de voter.
On constitue un échantillon de 1000 personnes (inscrites sur les listes électorales) que l'on suppose choisies ici de manière aléatoire. Ce n'est pas le cas en pratique (voir plus loin le paragraphe « sondages ») mais le principe reste le même que dans cet exemple.
Les résultats partiels en sont les suivants :
Sur les 1000 personnes
135 ont déclaré vouloir voter pour Jean-Marie Le Pen
195 ont déclaré vouloir voter pour Jacques Chirac
170 ont déclaré vouloir voter pour Lionel Jospin.
On peut déterminer trois intervalles de confiance au niveau de confiance de 95%[14] :
Jean-Marie Le Pen [0,135–0,032 ; 0,135+0,032] = [0,103 ; 0,167]
Jacques Chirac [0,195–0,032 ; 0,195+0,032]=[0,163 ; 0,227]
Lionel Jospin [0,170–0,032 ; 0,170+0,032]=[0,138 ; 0,202].
Donc la valeur unique en pourcentage donnée par l'institut est entachée d'une imprécision de +/–3 points. En examinant les trois intervalles trouvés, on peut a posteriori dire que le vrai résultat (16,9%,19,9%,16,2%) est compatible avec ceux-ci pour Jacques Chirac et Lionel Jospin car leurs résultats sont dans les intervalles correspondants. En revanche, le résultat de Jean-Marie Le Pen est légèrement supérieur à la borne supérieure de son intervalle de confiance (mais l'institut CSA lui donnait 14%, ce qui donne un intervalle [0,108 ; 0,172] qui contient son score réel).

2. Simulations

Le graphique ci-dessous donne 100 intervalles de confiance simulés au niveau de confiance 0,95 obtenus à partir de 100 échantillons de 50 individus extraits de la même population.

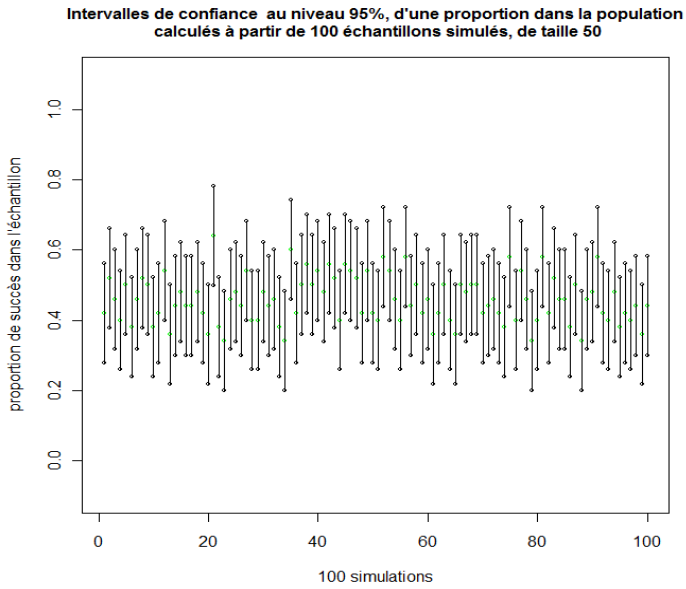


Figure 14

Document associé : intervalles de confiance simulés.r

[13] On peut consulter le site www.ipsos.fr/faq pour des détails sur les méthodes utilisées par cet institut.
[14] Pour chaque candidat, on applique la méthode précédente pour déterminer un intervalle de confiance de la proportion d'électeurs lui étant favorables.

On peut constater sur la figure 14 une fluctuation importante des bornes des intervalles de confiance numériques obtenus à chaque simulation.
Remarque : Les échantillons étant de taille 50, il y a exactement 51 valeurs possibles de la fréquence ce qui explique que l'on retrouve plusieurs fois les mêmes intervalles de confiance dès qu'on fait plus de 51 simulations.
La même simulation (figure 15) avec 100 intervalles de confiance simulés au seuil 0,95 obtenus à partir de 100 échantillons de 1000 individus extraits de la même population (la proportion inconnue est choisie aléatoirement à chaque série de 100 échantillons) fait apparaître une moindre fluctuation des bornes des intervalles.

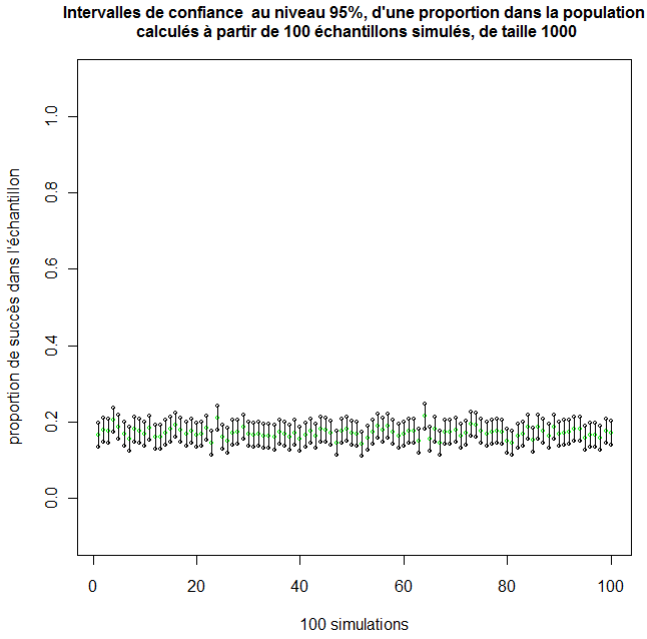
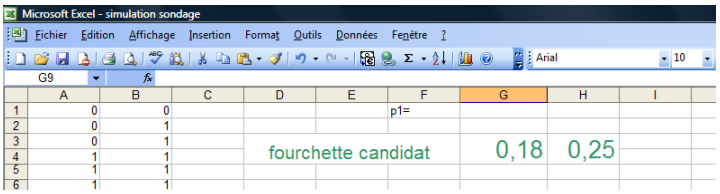


Figure 15

- Simulation simple d'un échantillon avec p inconnue.

Il s'agit de simuler des tirages d'échantillons dans une population où une proportion p est inconnue pour déterminer des intervalles de confiance de p au niveau de confiance 0,95.
On cache donc la valeur de p (qui peut être choisie au hasard) qui permet de faire ces simulations et on fait afficher les intervalles de confiance trouvés.



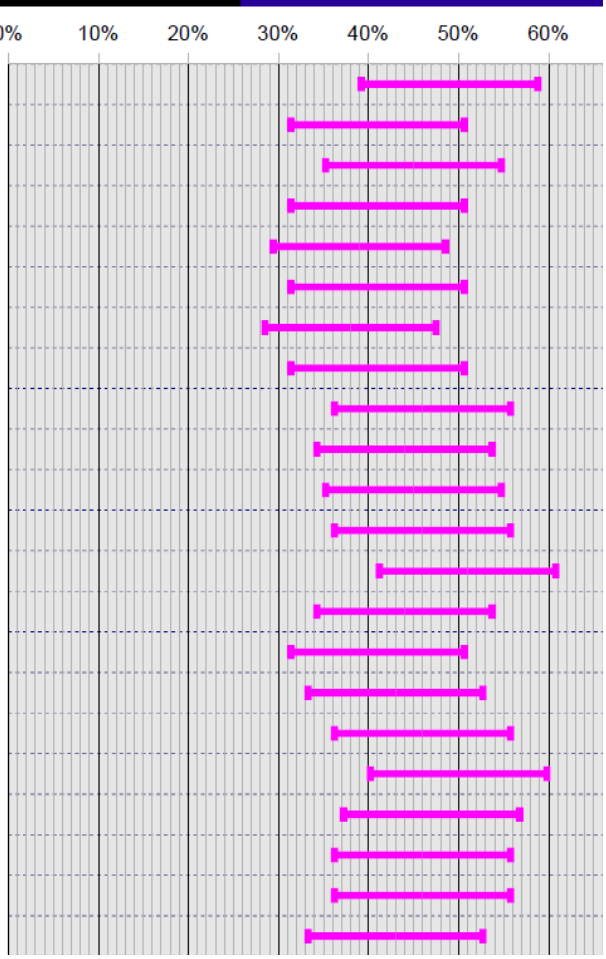
Document associé : simulation sondage.xls

La cellule G1 contient alea() ou un nombre masqué choisi entre 0 et 1.
Les cellules G3 et H3 contiennent les bornes inférieures et supérieures de l'intervalle de confiance.
L'appui sur F9 relance la simulation avec une nouvelle valeur de p.

- Simulation de plusieurs échantillons avec la même valeur de p inconnue.

On peut simuler sur tableur pour une proportion inconnue fixée un grand nombre de calculs d'intervalles de confiance à un niveau de confiance que l'on peut choisir.

niveau de confiance =	95 %		fréquence dans la population p =
échantillon	fréquence	fourchette	
1	49%	[39,20% ; 58,80%]	
2	41%	[31,36% ; 50,64%]	
3	45%	[35,25% ; 54,75%]	
4	41%	[31,36% ; 50,64%]	
5	39%	[29,44% ; 48,56%]	
6	41%	[31,36% ; 50,64%]	
7	38%	[28,49% ; 47,51%]	
8	41%	[31,36% ; 50,64%]	
9	46%	[36,23% ; 55,77%]	
10	44%	[34,27% ; 53,73%]	
11	45%	[35,25% ; 54,75%]	
12	46%	[36,23% ; 55,77%]	
13	51%	[41,20% ; 60,80%]	
14	44%	[34,27% ; 53,73%]	
15	41%	[31,36% ; 50,64%]	
16	43%	[33,30% ; 52,70%]	
17	46%	[36,23% ; 55,77%]	
18	50%	[40,20% ; 59,80%]	
19	47%	[37,22% ; 56,78%]	
20	46%	[36,23% ; 55,77%]	
21	46%	[36,23% ; 55,77%]	
22	43%	[33,30% ; 52,70%]	
23	56%	[46,27% ; 65,73%]	



Document associé : intervalles de confiance simulés-peignes.ods[15]

Exemples d'exercices

1. *Diagnostic de la jaunisse*

Un test de diagnostic rapide effectué sur des sujets ictériques (coloration jaune de la peau, des muqueuses -couche de cellules de protection recouvrant les organes creux en contact avec l'extérieur- et du blanc de l'œil –sclérotique-) doit permettre d'estimer si l'ictère est d'origine virale ou non, sans avoir besoin de faire des analyses longues et compliquées. Cependant il est important de pouvoir s'assurer que ce test est de bonne qualité c'est-à-dire qu'il doit pouvoir indiquer correctement si l'ictère est viral ou non. Il doit être capable d'identifier correctement le type d'ictère : il est positif chez les sujets dont l'ictère est viral et négatif sinon.

Une étude est effectuée sur 100 personnes ayant un ictère viral et 100 personnes ayant un ictère d'origine non virale.

Les résultats obtenus sont présentés dans le tableau ci-dessous

	Hépatite virale	Ictère d'origine non virale
Test positif	85	20
Test négatif	15	80

[15] Auteur du fichier : Stéphane Keller, LEGTA Louis Pasteur

a) Déterminer la proportion de sujets ayant un test positif parmi ceux ayant un ictère viral.

b) Déterminer un intervalle de confiance à 95% de la proportion de tests positifs lorsque l'ictère est viral. Cette proportion est appelée sensibilité du test diagnostic, c'est-à-dire la probabilité qu'une personne ayant un ictère viral réagisse au test. Un test diagnostic sera d'autant meilleur que la sensibilité est importante. (*réponse* : [0,75 ; 0,95]).

c) Déterminer la proportion de sujets ayant un test négatif parmi celles ayant un ictère non viral.

d) Déterminer un intervalle de confiance à 95% de la proportion de tests négatifs lorsque l'ictère est non viral. Cette proportion est appelée spécificité du test diagnostic, c'est-à-dire la probabilité qu'une personne ayant un ictère non viral ne réagisse pas au test. Un test diagnostic sera d'autant meilleur que la spécificité est importante.(*réponse* : [0,7 ; 0,9]).

2. *Dépistage de la bronchiolite*

Dans le but d'évaluer la prise en charge de la bronchiolite du nourrisson dans un hôpital de la région Aquitaine, une étude rétrospective a été mise en place.

1) Il est recommandé de coucher l'enfant de manière très inclinée (couchage en proclive) dans le cadre de la prise en charge de la bronchiolite. On évalue cette pratique à partir d'un échantillon de 134 dossiers. 106 des enfants ont été couchés en proclive.

   Déterminer un intervalle de confiance au niveau de confiance de 95% de la proportion d'enfants dont le couchage respecte la recommandation.

   Solution

$$\left[\frac{106}{134}-\frac{1}{\sqrt{134}};\frac{106}{134}+\frac{1}{\sqrt{134}}\right]\approx[0,70\,;0,88]$$

2) Une étude plus fine permet de comparer les pratiques entre les différents services ayant admis des enfants (cf. tableau 1).

   Tableau 1 : Répartition des cas suivant le type de services et le respect de la recommandation de couchage en proclive ; évaluation de la prise en charge de la bronchiolite en Aquitaine, une année donnée.

Couchage proclive	***En service des urgences***	***En service hospitalier***	***Total***
Oui	45	52	97
Non	29	8	37
Total	74	60	134

   a. Déterminer un intervalle de confiance au seuil de 95% de la proportion de couchage en proclive pour chaque type de service.
   Solution
   En service des urgences

$$\left[\frac{45}{74}-\frac{1}{\sqrt{74}};\frac{45}{74}+\frac{1}{\sqrt{74}}\right]=[0,492\,;0,724]$$

   En service hospitalier

$$\left[\frac{52}{60}-\frac{1}{\sqrt{60}};\frac{52}{60}+\frac{1}{\sqrt{60}}\right]=[0,738\,;0,996].$$

   b. (AP) Peut-on conclure selon vous au seuil de 95% que la pratique de couchage n'est pas identique selon le service ?

   Les deux intervalles de confiance n'ont pas d'intersection commune, on en conclut que les pratiques diffèrent entre les deux services.

Il s'agit là d'une règle assez répandue, même s'il en existe d'autres plus précises.

3 *Comparaison du taux de germination de semences de tomates de l'année avec celles de l'année précédente*.

Un maraîcher achète un lot de semences de tomates pour produire ses plants de tomate. Il lui reste des semences de l'année passée, dont il doit contrôler le taux de germination pour pouvoir les utiliser avec les autres. En effet, des taux de germination trop différents provoquent des trous dans les plates bandes de production, ce qui génère un coup de manutention plus élevé (il faut enlever les pots non germés avant de les conditionner). Il faut donc comparer les taux de germination des semences des deux années.

Une stratégie (il en existe d'autres, hors programme, mais qui peuvent faire l'objet d'une recherche) consiste à calculer et à comparer les intervalles de confiance des taux de germination (qui sont des proportions) des plants de l'année et de l'année précédente. Si les deux intervalles ne se recoupent pas, on peut conclure à une différence de taux de germination entre les semences des deux origines[16]. Il faudra alors les semer séparément.

Pour faire cette comparaison, le maraîcher prélève, aléatoirement dans les semences de l'année, un échantillon de 200 graines qu'il met à germer. Il constate que 185 graines germent.

Il prélève ensuite, aléatoirement dans les semences de l'année précédente, un échantillon de 200 graines qu'il met à germer. Il constate que 150 graines germent.

1. Déterminer un intervalle de confiance, au niveau de confiance de 95%, du taux de germination p_a du lot de semences de l'année.

Solution

$IC_{95\%} = [185/200 - 1/\sqrt{200}\ ;\ 185/200 + 1/\sqrt{200}\ ] \approx [0{,}925 - 0{,}071\ ;\ 0{,}925 + 0{,}071\ ]$

$\approx [0{,}85;0{,}99]$

2. Déterminer (par la même méthode qu'à la question a)) un intervalle de confiance au niveau 95%, du taux de germination p_b du lot de semences de l'année précédente.

3. Conclure.

Solution

Les deux intervalles sont disjoints, on peut donc conclure à une différence entre les taux de germination p_a et p_b au niveau de confiance 0,95.

Il est intéressant de noter que, sans connaître p_a et p_b, on dispose d'une méthode pour décider au niveau de confiance 95% que, si les intervalles de confiance sont disjoints, alors p_a et p_b sont différents.

Il existe d'autres méthodes d'estimation, mais quelle que soit la méthode utilisée, si elle est issue d'un échantillonnage aléatoire, la décision sera toujours entachée d'un risque d'erreur. Les méthodes utilisées assurent seulement la maîtrise de certains risques de se tromper.

[16] L'étude de cette problématique est suggérée en AP.

A. Loi uniforme

Le nouveau programme propose de définir la loi uniforme sur un intervalle $[a,b]$ quelconque.

Après avoir défini la loi uniforme sur $[0,1]$ à partir, par exemple, du choix au hasard d'un réel entre 0 et 1, on peut définir la loi uniforme sur $[a,b]$ en remarquant que pour que l'aire sous la courbe soit égale à 1, il faut et il suffit que la valeur de la constante soit $\frac{1}{b-a}$.

Une variable aléatoire X suit une loi uniforme sur l'intervalle $[a,b]$ si sa densité est la fonction f définie sur $[a,b]$ par : $f(x) = \frac{1}{b-a}$.

Espérance d'une variable aléatoire de loi uniforme sur $[a,b]$

L'espérance d'une variable aléatoire X suivant une loi uniforme sur $[a,b]$ est donnée par :

$$\mathrm{E}(X) = \int_a^b xf(x)\mathrm{d}x = \int_a^b \frac{x}{b-a}\mathrm{d}x = \frac{a+b}{2}.$$

On peut observer que la définition de l'espérance par la formule $\mathrm{E}(X) = \int_a^b xf(x)\mathrm{d}x$ prolonge celle de l'espérance d'une variable aléatoire discrète.

En effet, le terme $f(x)\mathrm{d}x$ peut s'interpréter comme l'aire d'un rectangle de côtés dx et $f(x)$, fournissant en quelque sorte la probabilité que la variable X prenne la valeur x. Dans ces conditions, l'intégrale $\int_a^b xf(x)\mathrm{d}x$ correspond à une « somme » de produits $x \times f(x)\mathrm{d}x$.

La figure 1 ci-dessous présente la situation dans le cas où $a = 0$ et $b = 1$.

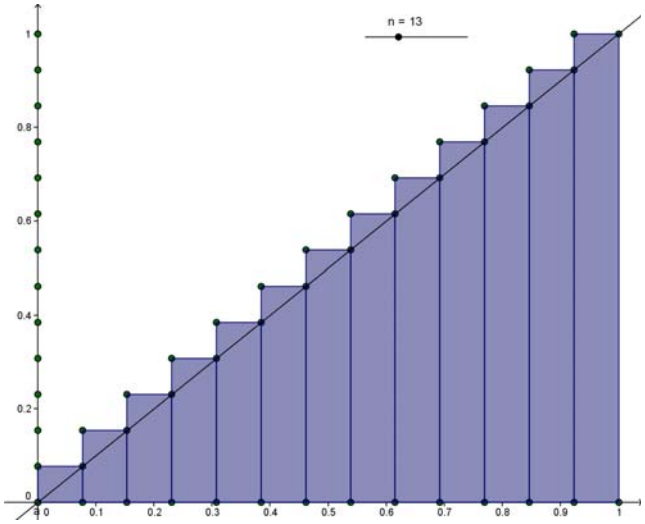


Figure 16

Document associé : espérance d'une variable uniforme.ggb

On a représenté les rectangles de base $\frac{1}{n}$ et de hauteur $\frac{k}{n}$, avec k entier variant de 1 à n.

La somme $S_n = \sum_{k=1}^{n} \frac{k}{n}\frac{1}{n}$ des aires de ces rectangles peut s'interpréter comme l'espérance d'une variable discrète équirépartie prenant les n valeurs $\frac{k}{n}$, pour k variant de 1 à n.

Elle vaut $\frac{n(n+1)}{2n^2}$ et a pour limite $\frac{1}{2}$.

Quand n tend vers l'infini, la somme des aires des rectangles tend vers l'aire située sous la droite d'équation $y = x$. On retrouve ainsi l'égalité $\int_0^1 xf(x)\mathrm{d}x = \frac{1}{2}$.

Exemples d'exercices

1. A partir de 7 heures le matin, les bus passent toutes les quinze minutes à un arrêt précis. Un usager se présente à cet arrêt entre 7h et 7h30. On fait l'hypothèse que l'heure exacte de son arrivée à cet arrêt, représentée par le nombre de minutes après 7h, est la variable aléatoire uniformément répartie sur l'intervalle [0, 30].

1) Quelle est la probabilité que l'usager attende moins de cinq minutes le prochain bus ?
2) Quelle est la probabilité qu'il attende plus de dix minutes ?

2. Partie A

Olivier vient tous les matins entre 7h et 7h 45 chez Karine prendre un café.

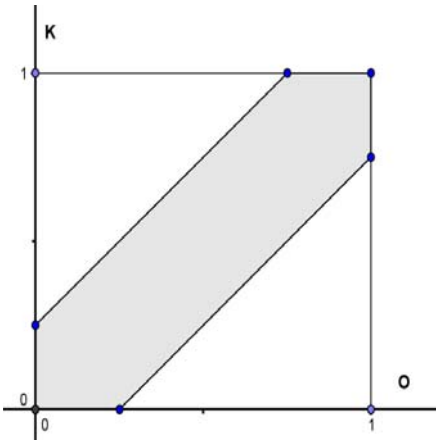
1) Sachant qu'Olivier ne vient jamais en dehors de la plage horaire indiquée et qu'il peut arriver à tout instant avec les mêmes chances, quelle densité peut-on attribuer à la variable aléatoire « heure d'arrivée d'Olivier » ?
2) Calculer la probabilité qu'Olivier sonne chez Karine :
   • Après 7h30 • Avant 7h10 • Entre 7h20 et 7h22 • A 7h30 exactement.

2. Partie B

Olivier et Karine décident de se retrouver au café de l'Hôtel de Ville entre 7h et 8h. Les instants d'arrivée d'Olivier et Karine sont assimilés à des variables aléatoires de loi uniforme sur $[0,1]$. Chacun attend un quart d'heure mais jamais au-delà de 8h. Quelle est la probabilité qu'ils se rencontrent ?

Éléments de solution

Pour la partie B, si on note O la variable aléatoire « instant d'arrivée d'Olivier » et K celle de Karine. La probabilité cherchée est $\mathrm{P}(|O-K| \leq \frac{1}{4})$; en utilisant une représentation graphique, cette probabilité est l'aire de la zone grisée ci-dessous, ensemble des points de coordonnées (x, y) du carré tels que $|x-y| \leq 0{,}25$. (On trouve $\frac{7}{16}$).



B. Lois exponentielles

Une variable aléatoire à densité X suit la loi exponentielle de paramètre $\lambda > 0$ si sa densité est la fonction f définie sur $[0,+\infty[$ par : $f(x) = \lambda e^{-\lambda x}$.

Pour tout intervalle $\langle c,d \rangle$ [17], on obtient : $P(X \in \langle c,d \rangle) = \int_c^d \lambda e^{-\lambda t} dt = e^{-\lambda c} - e^{-\lambda d}$.

En particulier, on obtient $P(X \leq a) = 1 - e^{-\lambda a}$.

L'espérance de X est la limite quand x tend vers $+\infty$ de $\int_0^x t\lambda e^{-\lambda t} dt$, on obtient $E(X) = \frac{1}{\lambda}$.

Pour effectuer le calcul de cette intégrale, on peut :

- chercher une primitive de la fonction $t \mapsto \lambda t e^{-\lambda t}$ sous la forme $(at+b)e^{-\lambda t}$ et déterminer ensuite a et b
- calculer la dérivée de la fonction g définie sur $[0,+\infty[$ par $g(t) = -te^{-\lambda t}$ et en utilisant le fait que $\int_0^x g'(t)dt = g(x)$, obtenir la valeur de l'intégrale $\int_0^x t\lambda e^{-\lambda t} dt$
- Expliquer éventuellement sur cet exemple le principe de l'intégration par parties, bien qu'il ne soit plus dans les capacités exigibles du programme.

On démontre qu'une variable aléatoire X suivant une loi exponentielle vérifie la propriété de durée de vie sans vieillissement, c'est-à-dire que, pour tous réels t et h positifs, $P_{X \geq t}(X \geq t+h) = P(X \geq h)$.

La réciproque de cette propriété n'est pas au programme.

[17] Cette notation désigne ici tous les types intervalles d'extrémités c et d où $c \leq d$.

Annexe 1 Introduction au théorème de Moivre-Laplace

L'objet de cette annexe 1 est de situer le théorème de Moivre-Laplace dans une perspective historique. Celle-ci permet de montrer l'évolution de la pensée probabiliste depuis Jacques BERNOULLI jusqu'à Pierre-Simon de LAPLACE qui donnera la preuve complète de ce théorème avec la rigueur possible à son époque.

La motivation commune à Bernoulli, Moivre et Laplace est de déterminer le plus finement possible la *fluctuation*[18] des valeurs prises par une variable aléatoire suivant une loi binomiale autour de son *espérance*. Il s'agissait ensuite d'utiliser *l'intervalle de fluctuation* obtenu pour *estimer* une probabilité inconnue, ce qui est la problématique moderne de *l'intervalle de confiance.*

Les énoncés des théorèmes sont donnés avec la formulation actuelle.

A. La loi des grands nombres de Jacques Bernoulli

Théorème de Bernoulli

On considère une variable aléatoire X_n suivant une loi binomiale $\mathcal{B}(n, p)$. On pose $F_n = \frac{X_n}{n}$.

Alors pour tout $\varepsilon > 0$ on a : $P\left(\left|F_n - p\right| \geq \varepsilon\right) \leq \frac{p(1-p)}{n\varepsilon^2}$.

Ce théorème est invoqué pour justifier l'approche fréquentiste de la notion de probabilité.

En effet, ce résultat liant fréquence et probabilité permet de donner une justification aux axiomes de la théorie générale (dite de Kolmogorov) par analogie avec les propriétés vues en statistiques.

La démonstration originale de Bernoulli, donnée dans son ouvrage *Ars conjectandi* publié à Bâle en 1713, fait appel avec beaucoup d'ingéniosité à la formule du binôme et aux propriétés des nombres $\binom{n}{k}$. Bernoulli est parfaitement conscient de la portée de son théorème comme le montre cet extrait de son ouvrage[19] :

« Mais pour que cela ne soit pas compris autrement qu'il ne convient, il faut bien noter ce qui suit ; je voudrais que le rapport entre les nombres de cas, que nous entreprenons de déterminer expérimentalement, ne fût pas pris de façon nette et sans partage (car ainsi c'est tout le contraire qui arriverait et il deviendrait d'autant moins probable de découvrir le vrai rapport qu'on ferait de plus nombreuses observations), mais je voudrais que le rapport fût admis avec une certaine latitude, c'est-à-dire compris entre une paire de limites, pouvant être prises aussi rapprochées qu'on voudra. »

On voit que le concept d'intervalle de confiance déduit d'un intervalle de fluctuation est déjà présent dans l'œuvre de Bernoulli.

Le théorème de Bernoulli est généralisé au dix-neuvième siècle par l'inégalité de Bienaymé-Tchebychev, après que les notions d'espérance et de variance auront été dégagées.

Inégalité de Bienaymé-Tchebychev

Soit (Ω , P) un univers probabilisé et X une variable aléatoire définie sur Ω possédant une variance V(X). On note E(X) son espérance. Alors pour tout ε > 0 on a :

[18] Les termes en italiques n'étaient pas utilisés par les mathématiciens de cette époque.
[19] Jacques Bernoulli , Ars Conjectandi, traduction de Robert Meunier, Irem de Rouen, 1987.

$$\mathrm{P}\left(\left|X-\mathrm{E}(X)\right|\geq\varepsilon\right)\leq\frac{\mathrm{V}(X)}{\varepsilon^2}.$$

Cette inégalité est intéressante pour donner tout son sens à la notion de *variance* dans un cadre plus général que celui du théorème de Bernoulli. D'après cette inégalité il apparaît clairement que plus la variance est petite, plus les fluctuations de X autour de son espérance sont faibles.

Remarque

Fondamentale du point de vue théorique, cette inégalité est insuffisante du point de vue des applications numériques car l'information sur la probabilité que F_n appartienne à l'intervalle de fluctuation $\left[p-\varepsilon, p+\varepsilon\right]$ est peu précise.

Exemple

$p=0,5 \quad \varepsilon=0,1 \quad n=100$ donnent un majorant de $\mathrm{P}(\left|F_n-p\right|\geq 0,1)$ égal à 0,25.

Or on sait que $\mathrm{P}\left(F_n\in\left[0,5-\frac{1}{\sqrt{100}};0,5+\frac{1}{\sqrt{100}}\right]\right)$ est voisin de 0,95 c'est-à-dire que $\mathrm{P}(\left|F_n-p\right|\geq 0,1)$ est voisin de 0,05.

C'est la recherche d'une meilleure précision qui a motivé le travail de Moivre puis de Laplace.

B. La démarche d'Abraham de Moivre

Abraham de Moivre est un protestant français, qui s'est exilé en Angleterre après la révocation en 1685 de l'Édit de Nantes. Il y rencontre James Stirling qui lui communique une précision importante sur la formule dite de *Stirling*, en réalité déjà présente dans les travaux de Moivre.

Dans son ouvrage *The Doctrine of chances* (1718), il met le calcul infinitésimal au service des probabilités. Cet ouvrage a été récemment traduit par les auteurs d'un document sur le théorème de Moivre-Laplace[20].

Le but d'A. de Moivre est d'évaluer $\mathrm{P}\left(\frac{n}{2}-\frac{1}{2}\sqrt{n}\leq X_n\leq\frac{n}{2}+\frac{1}{2}\sqrt{n}\right)$ où X_n suit une loi $\mathcal{B}(n, 1/2)$.

Il trouve 0,682688 comme valeur approchée en considérant n « infini ».

Or la limite de cette probabilité existe et vaut 0,682689492 environ.

De Moivre cherche ensuite à évaluer $\mathrm{P}\left(\frac{n}{2}-\sqrt{n}\leq X_n\leq\frac{n}{2}+\sqrt{n}\right)$. Il lui faut alors affiner sa technique et il utilise l'intégrale (au sens d'une aire) de la fonction $x\mapsto e^{-\frac{2x^2}{n}}$ apparue lors de l'évaluation de la somme des probabilités $\mathrm{P}(X_n=k)$ quand $\frac{n}{2}-\frac{1}{2}\sqrt{n}\leq k\leq\frac{n}{2}+\frac{1}{2}\sqrt{n}$.

Il parvient à la valeur approchée 0,95428 que l'on retrouvera plus loin.

La méthode de Moivre est un peu difficile à suivre, mais elle est esquissée dans la partie C avec une rédaction moderne.

En généralisant sans démonstration les résultats précédents au cas d'une loi $\mathcal{B}(n, p)$, il donne les éléments pour déterminer la probabilité d'un intervalle de fluctuation.

Il est intéressant de voir comment Moivre exploite son résultat.

En l'utilisant dans le sens direct, il en déduit que les fluctuations dues au hasard sont très limitées :

[20] *La loi des grands nombres, le théorème de De Moivre-Laplace,* D.Lanier, D.Trotoux,. http://www.math.ens.fr/culturemath/histoire%20des%20maths/pdf/LoidesGrandsNombres.pdf

« bien que le Hasard produise des Irrégularités, cependant les Rapports de Probabilités seront infiniment grands, et que avec l'avancement du Temps, ces Irrégularités n'auront aucune proportion avec le retour de l'Ordre qui résulte naturellement du DESSEIN ORIGINEL. »

En sens inverse, il retrouve le concept d'intervalle de confiance déjà esquissé par Bernoulli :

« *inversement*, si à partir d'Observations innombrables, nous trouvons que le Rapport des Evénements converge vers une quantité déterminée, comme le Rapport de P à Q ; alors nous concluons que ce Rapport exprime la Loi déterminée suivant laquelle l'Evénement se produit. »

Et enfin comme souvent à cette époque, il déduit de ce résultat mathématique une conviction religieuse :

« Et ainsi, si nous ne nous aveuglons pas nous-mêmes avec de la poussière métaphysique, nous seront conduits, d'une manière rapide et évidente, à la reconnaissance du grand CREATEUR et MAITRE de toutes choses ; *Lui-même toute sagesse, toute puissance* et *bonté*. »

C. Une approche du résultat de Moivre

On peut assez facilement comprendre comment apparaît la fameuse fonction dont la courbe a une forme de « cloche ». On a juste besoin de la formule dite « de Stirling ».

Prenons le cas où $p=\frac{1}{2}$ et donc $\mathrm{P}(X_n=k)=\binom{n}{k}\left(\frac{1}{2}\right)^n$, pour tout entier k compris entre 0 et n.

On pose $Z_n=\dfrac{X_n-\frac{n}{2}}{\frac{1}{2}\sqrt{n}}$ et on cherche le comportement de $\mathrm{P}(Z_n=x)$ quand n est grand, x étant fixé.

On a : $\mathrm{P}(Z_n=x)=P(X_n=\frac{1}{2}x\sqrt{n}+\frac{n}{2})$.

Comme X_n ne prend que des valeurs entières, $k=\frac{1}{2}x\sqrt{n}+\frac{n}{2}$ doit être entier.

On **fixe** x entier et on s'intéresse à la suite extraite de la suite $\left(\mathrm{P}(Z_n=x)\right)$ correspondant aux entiers n de la forme $n=(2m)^2$. On a alors $k=\frac{1}{2}x2m+2m^2=xm+2m^2$ qui est bien entier pour tout entier m.

On a $\binom{n}{k}=\dfrac{n!}{k!(n-k)!}$ et quand $m\to+\infty$ on a $k\to+\infty$ et $n-k\to+\infty$.

D'après la formule de Stirling, quand n est grand, $n!$ est équivalent à $n^{n+\frac{1}{2}}\mathrm{e}^{-n}\sqrt{2\pi}$ ce qui signifie que le quotient de $\dfrac{n!}{n^{n+\frac{1}{2}}\mathrm{e}^{-n}\sqrt{2\pi}}$ a pour limite 1 lorsque n tend vers $+\infty$.

n, k et $n-k$ étant grands, on peut considérer que :

$$\mathrm{P}(Z_n=x)\sim\frac{1}{\sqrt{2\pi}}\frac{n^{n+\frac{1}{2}}}{\left(\frac{n}{2}\right)^{n+1}\left(1+\frac{x}{\sqrt{n}}\right)^{\frac{x\sqrt{n}+n+1}{2}}\left(1-\frac{x}{\sqrt{n}}\right)^{\frac{-x\sqrt{n}+n+1}{2}}}\frac{1}{2^n}$$

$$\sim \frac{2}{\sqrt{n}}\frac{1}{\sqrt{2\pi}}\frac{1}{\left(1-\frac{x^2}{n}\right)^{\frac{n+1}{2}}}\frac{\left(1-\frac{x}{\sqrt{n}}\right)^{\frac{x\sqrt{n}}{2}}}{\left(1+\frac{x}{\sqrt{n}}\right)^{\frac{x\sqrt{n}}{2}}}$$

Or : $\left(1-\frac{x^2}{n}\right)^{\frac{n+1}{2}} \underset{n\to+\infty}{\to} \mathrm{e}^{-\frac{x^2}{2}}$ $\qquad$ $\left(1-\frac{x}{\sqrt{n}}\right)^{\frac{x\sqrt{n}}{2}} \underset{n\to+\infty}{\to} \mathrm{e}^{-\frac{x^2}{2}}$ $\qquad$ $\left(1+\frac{x}{\sqrt{n}}\right)^{\frac{x\sqrt{n}}{2}} \underset{n\to+\infty}{\to} \mathrm{e}^{-\frac{x^2}{2}}$

d'où finalement l'équivalent suivant :

$$\mathrm{P}(Z_n = x) \sim \frac{2}{\sqrt{n}}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}$$

valable pour x entier et n carré pair.

Remarque 1

Dans le cas de la suite extraite, on peut constater que $\mathrm{P}(Z_n = x)$ tend vers 0 quand n tend vers l'infini. Il reste à justifier que le résultat est valable pour tout x et pour la suite complète.

Remarque 2

Deux valeurs consécutives de Z_n sont distantes de $\frac{2}{\sqrt{n}}$ donc sur un intervalle de cette longueur ne se trouve qu'une valeur prise par Z_n. Si on pose $f(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}$, l'équivalent trouvé peut s'écrire $f(x)\Delta x$ et donc s'interpréter géométriquement comme l'aire d'un « petit » rectangle de base Δx et de longueur $f(x)$. Cela illustre la notion de densité.

Remarque 3

Numériquement pour $n = 100$ et $x = 1$, on obtient $k = 55$.

$\mathrm{P}(Z_{100} = 1) \approx 0{,}04847$ et l'équivalent vaut environ $0{,}04839$.

Remarque 4

L'aide apportée à Moivre par Stirling est la valeur de la constante égale à $\sqrt{2\pi}$ dans l'équivalent de n !.

D. Le théorème de Moivre-Laplace

Pierre-Simon de Laplace a été le premier à écrire un ouvrage exposant l'état des connaissances dans le domaine des probabilités. Il s'agit de *la théorie analytique des probabilités*[21] (1812). Dans ce texte, Laplace expose d'abord une série de résultats d'analyse (fonctions génératrices, transformée de Laplace...) qui lui permettent de démontrer des résultats de probabilités, et en particulier le théorème de Moivre-Laplace.

Laplace a des idées très précises sur les probabilités. Contrairement à Moivre, il ne cherche pas à prouver l'existence d'un *Grand créateur*, mais il cherche à approcher au mieux les lois qui régissent le monde dans lequel nous vivons. Il développe une vision très déterministe :

[21] Texte intégral disponible à l'adresse http://books.google.fr/books?id=6MRLAAAAMAAJ&printsec=frontcover&dq=Th%E9orie+analytique+des+probabilit%E9s+Laplace#v=onepage&q&f=false

« Nous devons donc envisager l'état présent de l'univers comme l'effet de son état antérieur, et comme la cause de celui qui va suivre. Une intelligence qui pour un instant donné connaîtrait toutes les forces dont la nature est animée et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ses données à l'analyse , embrasserait dans la même formule les mouvements des plus grands corps de l'univers et ceux du plus léger atome : rien ne serait incertain pour elle, et l'avenir comme le passé, serait présent à ses yeux. »

Ce n'est qu'au cours du vingtième siècle que cette vision déterministe sera remise en cause, en particulier par la physique quantique.

Concernant le théorème de Bernoulli, voici ce qu'il écrit :

« Ce théorème indiqué par le bon sens était difficile à démontrer par l'Analyse. Aussi l'illustre géomètre Jacques Bernoulli, qui s'en est occupé le premier, attachait-il une grande importance à la démonstration qu'il en a donnée. Le calcul des fonctions génératrices appliqué à cet objet, non seulement démontre avec fiabilité ce théorème, mais de plus il donne la probabilité que le rapport des événements observés ne s'écarte que dans certaines limites du vrai rapport de leurs possibilités respectives. »

« Moivre a repris dans son ouvrage [The Doctrine of Chances] *le théorème de Bernoulli sur la probabilité des résultats déterminés par un grand nombre d'observations. Il ne se contente pas de faire voir, comme Bernoulli, que le rapport des événements qui doivent arriver approche sans cesse de leurs possibilités respectives, il donne de plus une expression élégante et simple de la probabilité que la différence de ces deux rapports est contenues dans des limites données. »*

E. Convergence en loi

Ce paragraphe peut être réservé à une seconde lecture, son contenu dépassant nettement le niveau de la classe terminale.

Définition

Soient une suite de variables aléatoires réelles (X_n) et une variable aléatoire réelle X.

On note F_X (respectivement F_{X_n}) la fonction définie sur $\mathbb{R}$ par $F_X(x) = \mathrm{P}(X \leq x)$ (respectivement $F_{X_n}(x) = \mathrm{P}(X_n \leq x)$) appelée fonction de répartition de X (respectivement de X_n).

La suite (X_n) converge en loi vers X si, pour tout réel x où F_X est continue, on a : $\lim\limits_{n\to+\infty} F_{X_n}(x) = F_X(x)$.

La convergence en loi n'est pas la convergence des nombres $X_n(\omega)$ vers $X(\omega)$ mais la « convergence » des lois, et plus précisément la convergence simple, aux points de continuité de F_X, de la suite de fonctions F_{X_n} vers la fonction F_X . L'expression « la suite (X_n) » converge est donc un abus de langage, mais il est universellement pratiqué. En fait, si la suite (X_n) converge en loi vers X, alors elle converge en loi vers n'importe quelle variable aléatoire ayant la même loi que X.

Cas particulier

Dans le cas où toutes les variables sont à valeurs dans $\mathbb{N}$, la convergence en loi s'exprime par :

$\forall k \in \mathbb{N},\ \lim\limits_{n\to+\infty} \mathrm{P}(X_n = k) = \mathrm{P}(X = k)$.

Exemple

On considère une suite de variables aléatoires (X_n) suivant une loi binomiale B $(n, 1/n)$.

On démontre que : $\forall k \in \mathbb{N},\ \lim\limits_{n\to+\infty} \mathrm{P}(X_n = k) = \lim\limits_{n\to+\infty} \binom{n}{k} \left(\frac{1}{n}\right)^k \left(1-\frac{1}{n}\right)^{n-k} = e^{-1}\frac{1}{k!}$.

La loi limite est appelée loi de Poisson[22] de paramètre 1.

Comme pour la loi normale centrée réduite, cette loi est apparue comme loi limite.

Elle a ensuite été utilisée comme modèle dans divers domaines ; elle est appelée également loi des événements rares.

Annexe 2 Compléments sur les lois normales

A. Loi normale centrée réduite

Théorème

L'aire située entre la courbe représentative de la fonction f sur $\mathbb{R}$ définie par $f(t)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}$ et l'axe des abscisses est égale à 1.

On dit qu'une variable aléatoire X suit une loi normale centrée réduite $\mathcal{N}(0,1)$ si sa densité est la fonction $f(t)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}$.

On note Φ sa fonction de répartition c'est-à-dire la fonction définie sur $\mathbb{R}$ par :

$$\Phi(x)=\mathrm{P}(X\leq x)=\int_{-\infty}^{x} f(t)\mathrm{d}t\,.$$

Représentations graphiques :

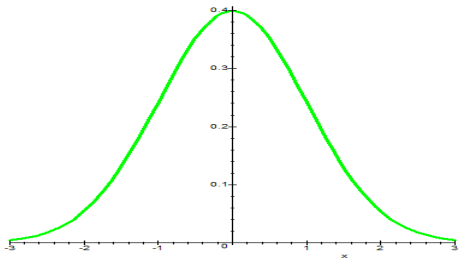


Figure 1 : $f(t)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}$

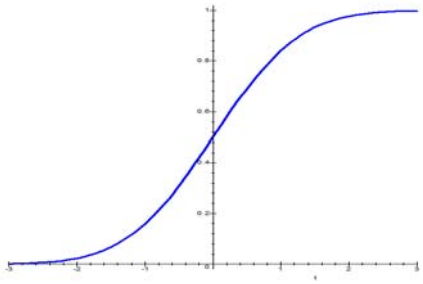


Figure 2 : $\Phi(x)=\mathrm{P}(X\leq x)=\int_{-\infty}^{x} f(t)\mathrm{d}t$

Remarque : Il faut noter que la fonction f n'a pas de primitive « explicite », c'est à dire qu'il est impossible de l'exprimer algébriquement avec les fonctions usuelles (polynômes, exponentielle, logarithme...). Pour cette raison, il a été établi des tables numériques (comme les tables de logarithmes). Avec les calculatrices, ces tables ont aujourd'hui perdu leur intérêt.

Quelques propriétés

P1. $\Phi(-x)=1-\Phi(x)$.

Visible graphiquement sur la figure 4 on peut aussi démontrer cette formule par changement de variable.

[22] Siméon-Denis Poisson (1781-1840) *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*

On en déduit que $\Phi(0) = \mathrm{P}(X \leq 0) = \mathrm{P}(X \geq 0) = \frac{1}{2}$.

Une variable suivant la loi $\mathcal{N}(0,1)$ a donc 0 pour médiane.

P2. $\Phi' = f$.

Il suffit d'écrire $\Phi(x) = \int_{-\infty}^{0} f(t)\mathrm{d}t + \int_{0}^{x} f(t)\mathrm{d}t = \frac{1}{2} + \int_{0}^{x} f(t)\mathrm{d}t$ pour constater que Φ est de classe C^1 sur $\mathbb{R}$ et que sa dérivée est f.

P3. Φ est une bijection strictement croissante de $\mathbb{R}$ dans]0,1[.

La stricte croissance et la continuité sont immédiates.

Les limites aux bornes sont 0 et 1[23] et elles ne sont pas atteintes du fait que Φ est strictement croissante.

B. Lois normales

Une variable aléatoire X suit une loi $\mathcal{N}(\mu, \sigma^2)$ si $\frac{X-\mu}{\sigma}$ suit la loi normale $\mathcal{N}(0,1)$.

Propriété de stabilité par addition et multiplication par un réel

Cette propriété est bien sûr hors programme en terminale puisque la somme de variables aléatoires n'y est pas abordée, ni la notion de variables indépendantes. Elle est cependant d'une très grande importance et justifie en particulier la notation $\mathcal{N}(\mu, \sigma^2)$.

Propriété

Si X suit une loi $\mathcal{N}(a,b^2)$ et que Y suit une loi $\mathcal{N}(c,d^2)$ et qu'elles sont indépendantes, alors leur somme $X + Y$ suit également une loi normale de paramètres $a + c$ et $b^2 + d^2$.

C'est cette propriété qui justifie la notation $\mathcal{N}(\mu, \sigma^2)$, à savoir que les variances s'additionnent (si les variables sont indépendantes) mais pas les écarts types.

Annexe 3 Approche simplifiée de la théorie des sondages

Dans la plupart des situations il est impossible d'interroger ou de recueillir des informations sur l'ensemble de la population ; pour cette raison on se contente le plus souvent d'un échantillon. Un échantillon correspond à un sous-ensemble de la population qui intéresse le responsable de l'étude.

Pour que les informations recueillies auprès de l'échantillon puissent permettre d'estimer des caractéristiques de la population il est important d'être rigoureux et d'utiliser des méthodes d'échantillonnage appropriées. Ces différentes méthodes sont présentées succinctement ci-dessous.

A. Qualités d'un échantillon permettant de répondre à une question posée

L'observation d'un échantillon ne permet pas de décrire avec certitude une population mais seulement d'estimer par intervalles de confiance les valeurs de certaines caractéristiques que l'on souhaite connaître dans cette population.

[23] En utilisant la définition d'une intégrale généralisée convergente.

« *Voir invoquer la « représentativité » dans un rapport d'enquête pour justifier de la qualité d'un sondage peut presque à coup sûr laisser soupçonner que l'étude a été réalisée dans une méconnaissance totale de la théorie de l'échantillonnage. Le concept de représentativité est aujourd'hui à ce point galvaudé qu'il est désormais porteur de nombreuses ambivalences. Cette notion, d'ordre essentiellement intuitif, est non seulement sommaire mais encore fausse et, à bien des égards, invalidée par la théorie …* »[24]

La première chose à préciser c'est qu'avec un échantillon on ne peut pas être représentatif de l'ensemble de la population sur toutes les caractéristiques, il est donc important de définir les caractéristiques qui intéressent les responsables de l'enquête.

Pour un statisticien, l'échantillon sera dit représentatif si on peut correctement estimer les paramètres d'intérêt de la population à partir de l'échantillon. Dans le cas contraire on parlera de biais d'échantillonnage. Pour pouvoir correctement estimer les paramètres, le statisticien n'a pas nécessairement besoin que l'échantillon soit une reproduction miniature de la population, par contre il a besoin que tous les profils de la population importants pour l'objectif de l'enquête soient représentés dans l'échantillon. Cela signifie donc que le plan d'échantillonnage utilisé dépendra de l'objectif de l'étude même si la population est la même.

La représentativité d'un échantillon nécessite que la procédure d'échantillonnage permette la constitution d'un sous-groupe recouvrant les caractéristiques qui peuvent influencer la valeur des paramètres que l'on veut estimer. La non-représentativité d'un échantillon peut par exemple être due à la sélection dans une base de sondage ne couvrant pas correctement la population.

Par exemple, supposons qu'on souhaite réaliser une enquête de prévalence d'une maladie A dans la population générale et qu'on sélectionne un échantillon à partir de la liste téléphonique (l'enquête devant se dérouler par appel téléphonique). Dans ce cas l'échantillon ne couvre pas correctement la population il y a un biais d'échantillonnage car les personnes qui répondront à l'enquête auront un téléphone et seront présentes à leur domicile, pour cette raison toutes les personnes qui seront hospitalisées à la date de l'enquête ne seront pas interrogées. Si les personnes atteintes de la maladie étudiée sont plus susceptibles de se rendre à l'hôpital, on risque de sous-estimer la prévalence de la maladie ou proportion de malades, en réalisant un échantillon comme proposé ci-dessus.

Dans tous les cas de figures on souhaite enquêter sur un nombre suffisant de sujets afin de pouvoir estimer correctement le paramètre de la population.

En principe la taille de l'échantillon est indépendante de la taille de la population que l'on veut étudier. Il faut interroger autant de personnes pour estimer avec la même précision le résultat de l'élection présidentielle en France, que l'élection du maire de Bordeaux.

La taille est en revanche fonction de la marge d'erreur (amplitude de l'intervalle de confiance) que l'on accepte de prendre et qui résulte inéluctablement du fait que l'estimation est issue d'un échantillon.

Un sondage peut être effectué de multiples façons que l'on regroupe en deux grandes familles : les sondages aléatoires, dits aussi probabilistes, et les sondages non aléatoires, dits aussi empiriques ou informels.

B. Echantillonnage non-probabiliste ou non aléatoire

Pour ce type de sondage la sélection des individus n'obéit plus au hasard mais est définie selon des critères de faisabilité, de ressemblance à la population et de critères subjectifs dépendant du choix des enquêteurs.

Les types de sondage satisfaisant aux critères de faisabilité ou de simplicité sont par exemple les échantillons de sujets volontaires (par exemple les enquêtes publiques préalables à la déclaration

[24] Y. Tillé (2001), Théorie des sondages : Échantillonnage et estimation en populations finies : cours et exercices, 284 pages, Paris, Dunod

d'utilité publique : les personnes qui le souhaitent prennent connaissance du projet et consignent leurs observations sur un registre d'enquête ouvert en mairie) ou les échantillons de convenance (par exemple on effectuera une enquête auprès de toutes les personnes qui viennent à la poste centrale de la ville V le mardi 4 septembre 2012).

Les types de sondage satisfaisant aux critères de ressemblance à la population sont appelés échantillonnage par choix raisonné. La méthode des quotas, qui est la méthode la plus utilisée parmi les sondages non aléatoires et dans les sondages d'opinion, fait partie de cette catégorie de sondage. Les enquêteurs doivent inclure un nombre donné d'individus présentant telle ou telle caractéristique dans des proportions voisines de celles de la population. Du moment que le quota est respecté, le mode de sélection des individus est laissé au libre choix de l'enquêteur. La méthode des quotas consiste à construire un échantillon qui soit une maquette, un modèle réduit de la population étudiée, en conservant les mêmes proportions. La plupart des sondages politiques effectués en France utilisent cette méthode.

La date cruciale pour l'histoire de l'échantillonnage est le mardi 3 novembre 1936, jour de la publication des résultats de l'élection présidentielle aux États-Unis. Le journal « Literary Digest » avait réalisé des sondages pré-électoraux, comme à leur habitude, par consultation individuelle d'électeurs (appelés « votes de paille » à cette époque). Cette méthode ne fait appel à aucune notion de représentativité, mais est réalisée sur un nombre important d'électeurs et jusqu'en 1936 elle donne des résultats tout à fait satisfaisants. Ce journal comme bien d'autres prédit alors l'élection de Lanton, mais finalement F.D. Roosevelt est élu. Seuls trois sondages l'avaient donné gagnant, tous réalisés par une méthode empirique appelée la méthode des quotas. Ce fut le début des grandes structures de sondages telles que la société de sondage Gallup aux Etats-Unis.

C. Echantillonnage probabiliste

Dans un plan d'échantillonnage aléatoire, tous les individus de la population ont une probabilité connue et non nulle d'être sélectionnés pour faire partie de l'échantillon. La sélection des individus constituant l'échantillon s'effectue par un plan d'échantillonnage à un ou plusieurs degrés et à chaque degré une procédure de tirage au sort est spécifiée ; il peut s'agir d'une procédure de sondage aléatoire simple, ou systématique, ou d'une procédure stratifiée, avec sélection équiprobable ou à probabilité proportionnelle à la taille. Logiquement seuls les sondages aléatoires permettent de fournir des estimations avec une précision donnée, c'est- à -dire avec un intervalle de confiance.

Les sélections aléatoires à partir d'une liste d'individus peuvent s'effectuer de différentes façons. Prenons l'exemple d'une enquête que l'inspection académique souhaiterait réaliser auprès des élèves des lycées d'un département afin d'étudier les difficultés scolaires rencontrées par ceux-ci. Il est impossible d'interroger la totalité des élèves et le souhait est de pouvoir obtenir des informations sur un échantillon représentatif de 500 élèves. Pour cette dernière raison il est décidé de sélectionner aléatoirement les élèves, mais plusieurs méthodes peuvent être proposées.

1) si la liste de tous les élèves est accessible de manière électronique on peut sélectionner 500 élèves dans la liste en utilisant par exemple un tableur, il y a plusieurs méthodes pour cela :
    a. créer pour chaque élève un nombre aléatoire suivant une loi uniforme, puis choisir de trier la liste en fonction de ce nombre aléatoire créé, cela revient à mélanger de façon aléatoire la liste. On sélectionne finalement les 500 premiers noms qui sont dans la liste triée. Cette méthode permet de réaliser une sélection simple sans remise.
    b. Numéroter tous les élèves de la liste, puis utiliser la fonction aléatoire du tableur pour sélectionner uniquement 500 nombres, les élèves correspondant à ces nombres seront sélectionnés. En appliquant cette méthode un nombre peut être sélectionné plusieurs fois. Cela revient donc à réaliser un échantillon avec remise.
    c. On peut aussi utiliser la méthode de sélection systématique, c'est-à-dire que si le nombre d'élèves est égale à N on tire au sort un nombre, noté d, entre 1 et N puis on

sélectionne de manière régulière sur la liste le $d + \frac{N}{500}$ énième élèves, si ce nombre dépasse le rang du dernier élève on reprend la liste au début.

Les trois méthodes présentées ci-dessus sont des sélections que l'on peut considérer équiprobables car chaque sujet a la même probabilité d'être sélectionné.

2) On peut souhaiter effectuer une enquête en face à face, c'est-à-dire qu'un enquêteur doit se déplacer pour interroger l'élève, il est donc important d'essayer de gérer le nombre de déplacements. Dans les méthodes proposées précédemment rien n'est contrôlé et l'enquêteur peut devoir traverser le département pour interroger un et un seul élève. Afin d'améliorer cela on peut décider de sélectionner un certain nombre d'établissements et de sélectionner un certain nombre d'élèves dans chaque établissement. On parlera alors de sondage à plusieurs degrés. Dans ce cas la sélection n'est pas toujours équiprobable.

   Exemple : supposons que 10 des 70 lycées soient sélectionnés et dans chaque lycée sélectionné on sélectionne 30 élèves. Dans ce cas la probabilité que l'élève A soit sélectionné est environ égale à 10/70 * 30/(nb d'élèves du lycée d'appartenance de l'élève A) , on remarque que cette probabilité dépend de la taille du lycée et donc non équiprobable.

3) On peut vouloir construire un échantillon représentant les lycées généraux et professionnels. Dans ce cas et afin de forcer cette représentativité, on commence par partager en deux paquets la liste : liste des lycées professionnels et liste des lycées généraux et on effectue un échantillon dans chacune des deux listes. On parle alors de sondage stratifié.

Annexe 4 Utilisation des Tice

A. Tableau des fichiers du document ressource Probabilités et Statistique du programme de Terminale.

Les textes en italique ou en italique vert concernent des fichiers non référencés de certaines figures du document principal ou des fichiers d'activités complémentaires n'apparaissant pas dans le document principal.

DOCUMENTS DE SYNTHÈSE	FICHIERS	FONCTIONS, PARAMÈTRES D'ANIMATION OU DE FONCTION ET DESCRIPTION **n (10 ; 60 ; 1 ; curs) signifie que l'on peut faire varier le paramètre n de 10 à 60 avec des pas de 1, à l'aide d'un curseur. curs peut être remplacé par bouton (à cliquer).**
Annexe 4	***InitiationR1.r***	*Démarrage rapide en R : installation et quelques exemples commentés, bibliographie.*
VI Figure 16	**espérance d'une variable uniforme.ggb**	**Animation GeoGebra** illustrant l'aire de **n (0 ; 40 ; curs)** rectangles de base **1/n**, de hauteur **k/n** avec **1 ≤ k ≤ n**. Convergence de la somme de l'aire de tous les rectangles vers l'aire sous la droite **y = x**.
I Figure 1	**centrer et réduire une binomiale.ggb**	**Animation GeoGebra:** superposition des diagrammes en barre de X v.a. binomiale de paramètres **n (10 ; 60 ; 1 ; curs)** et **p (0 ; 1 ; 0,01 ; curs),** de X – n×p, variable centrée et de Z = (X – n×p) / racine(n × p × (1 – p)), variable centrée et réduite.
I Figure 2	**diagramme en bâtons de Fn.ggb**	**Animation GeoGebra :** diagramme en bâton de la variable fréquence Fn = (Xn - p) / √(p×(1 – p) / n), pour **n (10 ; 60 ; 1 ; curs) et p (0 ; 1 ; 0,01 ; curs).**

DOCU-MENTS DE SYNTHÈSE	FICHIERS	FONCTIONS, PARAMÈTRES D'ANIMATION OU DE FONCTION ET DESCRIPTION **n (10 ; 60 ; 1 ; curs) signifie que l'on peut faire varier le paramètre n de 10 à 60 avec des pas de 1, à l'aide d'un curseur. curs peut être remplacé par bouton (à cliquer).**
II Figure 3	**binomiale et normale.ggb**	**Animation GeoGebra:** illustration du théorème de Moivre-Laplace : convergence de la suite de variables centrées réduites Zn, avec **n (10 ; 300 ; 1 ; curs)** et **p (0 ; 1 ; 0,01 ; curs)** vers la loi normale centrée réduite.
III Figure 7	***visualisation probas normales.ggb***	***Animation GeoGebra :*** *illustration des calculs de probabilité (calcul intégral d'aires) P(a < x < b), avec les lois normales de moyenne* ***m (0 ; 5 ; 0,1 ; curs)*** *et d'écart type* ***σ (0 ; 5 ; 0,1 ; curs)****, avec* ***a (-5 ; 5 ; 0,1 ; curs)*** *et* ***b (-5 ; 5 ; 0,1 ; curs)****.* *On peut donc faire la représentation graphique de la fonction de densité de la loi normale centrée réduite.* *Et visualiser les probabilités des intervalles mu ± sigma (m ± σ), mu ± 2sigma (m ± 2σ), mu ± 3sigma (m ± 3σ).*
III Figure 6	***TaillesHommes.r***	***tailleshommes(n = 50000, mu = 175, sigma = 8)*** ***Fonction en R : Simulation*** *d'un échantillon (une série statistique) de n (50 000) tailles d'hommes tirés d'une distribution normale de moyenne* ***mu*** *et d'écart type* ***sigma****. L'histogramme est tracé, quelques paramètres de la série sont calculés. On obtient d'autres séries que celle illustrée dans le document. n, mu et sigma sont des paramètres que l'on peut changer à volonté.*
III Figure 8	***influence de mu et sigma.ggb***	***Animation GeoGebra :*** *Influence de la moyenne et de l'écart type sur la forme de la courbe représentative de la densité de la loi normale de moyenne* ***m (0 ; 7 ; 0,1 ; curs)*** *et d'écart type* ***σ (0,3 ; 3 ; 0,1 ; curs)****.*
IV-B	**intervalle de fluctuation première.alg**	**Algorithme-programme Algobox** de calcul des deux bornes de l'intervalle de fluctuation binomial exact, selon une méthode du document ressource de 1ère. : Queues symétriques (équilibrées) en probabilité **n** est la taille de l'échantillon, **p** est la probabilité de succès, **a** et **b** sont les deux bornes de l'IF. **Attention :** limité à n<70
IVB	**intervalle de fluctuation première.r**	**IFexact2(n = 65, p = .06, kobs = 8, proba = .95)** **Fonction en R :** IF binomial exact, selon une méthode du document ressource de 1ère : Queues symétriques (équilibrées) en probabilité **n** est la taille de l'échantillon, **p** est la probabilité de succès, **kobs** est le nombre de succès observé dans l'échantillon **proba** est le seuil de l'intervalle de fluctuation. **a** est le plus petit entier tel que P(X <= a) > 0,025 ; **b** est le plus petit entier tel que P(X <= b) >= 0,975 Une conclusion est proposée quant à l'hypothèse testée.
IV-B *AutresAlgoDuDoc.pdf*	***IF_BinomialExact1.r***	***IFexact1(n = 65, p = .06, kobs = 8, proba = .95)*** ***Fonction en R :*** *IF binomial exact, selon une méthode du document ressource de 1ère : Queues symétriques (équilibrées) en probabilité* ***n*** *est la taille de l'échantillon,* ***p*** *est la probabilité de succès,* ***kobs*** *est le nombre de succès observé dans l'échantillon* ***proba*** *est le seuil de confiance de l'intervalle de fluctuation.* ***a*** *est le plus grand entier tel que P(X < a) ≤ 0,025 ;* ***b*** *est le plus petit entier tel que P(X > b) ≤ 0,025* *Une conclusion est proposée quant à l'hypothèse testée.*
IV C	**exploration de l'intervalle de fluctuation asymptotique.r**	**pIFasy2_1(n = 400, p= .5, proba = .95)** **Fonction en R** : illustration de l'évolution de la probabilité binomiale de l'intervalle de fluctuation asymptotique **IF2 : (p ± u_{proba}×racine(p×(1 – p) / n))** en fonction de **n** et **p**. **n** est la taille de l'échantillon, **p** est la probabilité de succès, **proba** est la valeur seuil de la probabilité de l'IF.

DOCUMENTS DE SYNTHÈSE	FICHIERS	FONCTIONS, PARAMÈTRES D'ANIMATION OU DE FONCTION ET DESCRIPTION **n (10 ; 60 ; 1 ; curs) signifie que l'on peut faire varier le paramètre n de 10 à 60 avec des pas de 1, à l'aide d'un curseur. curs peut être remplacé par bouton (à cliquer).**
IV C Figure 9	**exploration intervalle de fluctuation asymptotique.ods**	**Feuille de calcul tableur** permettant une visualisation des probabilités des intervalles de fluctuation asymptotiques de seconde et de terminale**.** **p** est la probabilité de succès. Pour celui de seconde, on peut conjecturer l'existence du seuil n_0 du paragraphe V-C-7.
IV C Figure 11	**intervalle de fluctuation seconde.r**	**pIFasy1_1(n = 700, p = .5, proba = .95)** **Fonction en R** : illustration de l'évolution de la probabilité binomiale de l'intervalle de fluctuation asymptotique **IF1 : (p ± 1 / racine(n))** en fonction de **n** et **p**. **n** est la taille de l'échantillon, **p** est la probabilité de succès, **proba** est la valeur seuil de la probabilité de l'IF.
IV E	**recherche du n0.alg**	**Algorithme-programme Algobox** de recherche, pour un **p** donné, de la plus petite valeur **n0** de **n** telle que la probabilité exacte que X appartienne à l'**IF1 (p ± 1 / racine(n))** soit au moins égale à 0,95. Valeurs de **n** au plus égales à 70, application numérique restreinte.
IV E	**recherche du n0.sce**	**Programme Scilab** de recherche, pour un **p** donné, la plus petite valeur **n0** de **n** telle que la probabilité exacte que X appartienne à l'**IF1 (p ± 1/racine(n))** soit au moins égale à 0,95.
AutresAlgo DuDoc.pdf	***nIFasy1_1.r***	***nIFasy1_1(nsup = 1000, probinf = .95)*** ***Fonction en R :*** *Pour les valeurs de **p** de 0,05 à 0,95, de 0,01 en 0,01, recherche la plus petite valeur n0 de **n** telle que la probabilité exacte que X appartienne à l'**IF1 (p □ 1 / racine(n))** soit au moins égale à probinf. Tableau des valeurs de n0 et graphique de n0 en fonction de p.*
V Figures 14 et 15	**intervalles de confiance simulés.r**	**simICdoc(n = 50, nbsim = 100, nbclass = 20)** **Fonction en R : Simulations** d'un peigne d'IC au niveau de confiance nominal de 0,95. nbclass est le nombre de classes de l'histogramme. La proportion de **p** dans la population est générée aléatoirement dans]0 ; 1[. Elle est affichée dans la console **R**.
	SimulICPropSimpl.r	***simIC(n = 50, nbsim = 100, nbclass = 20, moustache = 1.5)*** ***Fonction de R : Simulation*** *d'un peigne d'IC. nbclass est le nombre de classes de l'histogramme, moustache détermine la longueur des moustaches des boites. La proportion de p dans la population est générée aléatoirement dans]0 ; 1[. Le peigne est suivi de l'histogramme et de la boite à moustache de la distribution **simulée.***
V	**simulation sondage.xls**	**Simulation tableur** d'un intervalle de confiance d'une proportion **p** inconnue, calculé à partir d'un échantillon de taille 1000. **F9** pour refaire une autre simulation. On peut dévoiler ***p*** en mettant une couleur de police visible.
V	**intervalles de confiance simulés-peignes.ods**	**Simulation tableur** d'un peigne de **100** intervalles de confiance gaussiens d'une proportion **p (0% ; 100% ; 1% ; boutons)** au niveau de confiance **c (0% ; 100% ; 1% ; boutons)**. Pour chacun des **100** échantillons simulés, la feuille affiche f observé, l'intervalle de confiance (fourchette), VRAI si l'IC contient **p**, FAUX sinon, le pourcentage d'IC contenant **p**, et la représentation graphique de l'IC, en barre horizontale. On peut cacher ou faire afficher **p**. F9 pour lancer une nouvelle simulation de **100** échantillon de taille **n = 100**.
VI C Figure 17	***monte carlo.alg***	*Calcul approché par la méthode du « rejet » de l'intégrale de 0 à 1 de F1(x) (à saisir). Le nombre n de tirages est saisi en entrée. Graphique indiquant la surface atteinte par les tirages.*
VI C	***monte Carlo bis.alg***	*Calcul approché par la méthode de l'espérance de l'intégrale de 0 à 1 de F1(x) . Le nombre n de tirages est saisi en entrée.*
	commandes R.pdf	*Liste de commandes R.*
	Carte de référence R.pdf	*Fonctions vitales, outils de programmation sous R.*

QUELQUES EXEMPLES COMMENTÉS POUR DÉMARRER AVEC R
PROBABILITÉS, SIMULATIONS ET EXPLORATION DES SÉRIES SIMULÉES

I - INSTALLATION – MISE EN ROUTE

1° Installation

● **R** est un logiciel libre et gratuit téléchargeable à http://cran.univ-lyon1.fr, (site miroir) le site parent étant www.r-project.org. Il est multiplateforme, c'est à dire qu'il existe des versions qui tournent sous **linux, mac et windows**.

● Il existe quelques ouvrages et un grand nombre de sites en français, d'IUT, d'universités, d'organismes de recherche et d'écoles d'ingénieurs, traitant de l'utilisation de **R** (voir bibliographie).

● L'installation se fait très rapidement et simplement à partir du fichier exécutable téléchargé. Ce "package" de base est complet et permet d'effectuer tous les traitements statistiques courants (description, analyse exploratoire des données, probabilités, simulation, tests statistiques).

● L'utilisation de **R** peut se faire en ligne de commande, l'installation de base y suffit. On peut aussi utiliser certaines fonctionnalités de **R** sous forme classique de menus cliquables en français, il faut alors installer le package **"Rcmdr"**. Les commandes **R** correspondant à chaque menu sont affichées, ce qui facilite une première prise en main. La rédaction et la lecture des lignes de commande **R** sont grandement facilitées par l'utilisation d'un éditeur spécifique, **"Tinn-R"** (téléchargeable à http://sourceforge.net/projects/tinn-r) qui identifie toutes les commandes **R** et leurs paramètres et les colorie de façon différenciée pour en faciliter l'identification et l'utilisation.

● Dans les fichiers mis à disposition, figurent deux "Reference card" (**"Rrefcard2.pdf" et "ShortRefCard.pdf"**) qui contiennent les principales commandes **R** classées par thème.

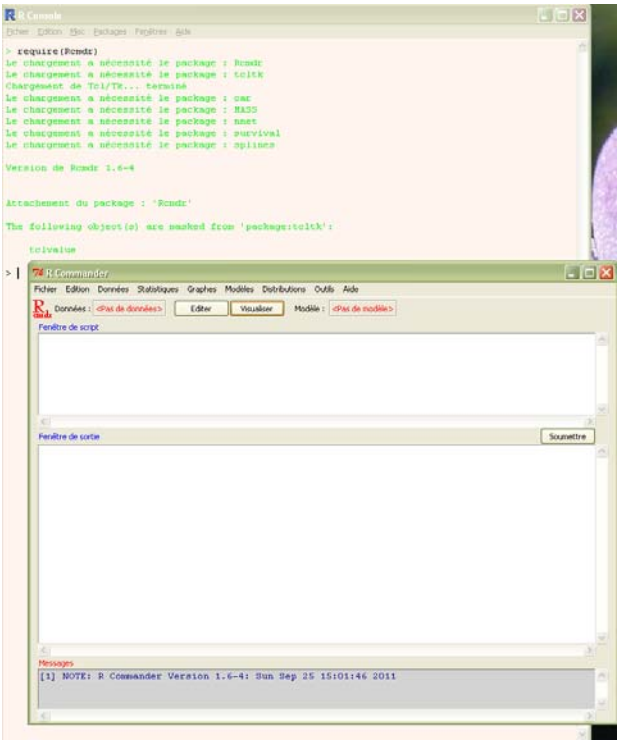
● La communauté des utilisateurs de **R** développent, pour les besoins des structures dans lesquelles ils travaillent ou des recherches engagées, des "packages" "agréés" par une "R-core-team", qu'ils mettent à disposition sur les sites et qui peuvent s'installer automatiquement. Il en existe plusieurs centaines. Deux sont mentionnés dans certains fichiers annexés au document ressource, qui sont **"lattice"** (graphismes avancés) et **"Hmisc"** (présentation avancée de résumés numériques). De même il existe un package (et un ouvrage en français) dédiés à l'analyse des données à la française, **"FactoMineR"**, développé par trois enseignants chercheurs de l'AgroCampus de Rennes.

2° Utilisation de l'interface avec menus à cliquer en français : "Rcmdr"

● On peut utiliser **R** en mode **menus à cliquer** en français. Le code de chaque commande sollicitée par les menus apparaît dans la "Fenêtre de script" est exécutée dans la "Fenêtre de sortie", qui affiche aussi les résultats. Cet affichage des commandes correspondant aux menus cliqués permet l'apprentissage progressif des codes des commandes **R**. C'est aussi un bon outil pour initier les élèves.

Il faut pour cela installer le package **Rcmdr :** Après avoir lancé **R** (RGui) et être connecté à internet, c'est automatique via le menu "Package – Installer le package"). Pour l'exécuter il faut cliquer le menu Package – Charger le package" ou taper **require(Rcmdr)** dans la console **R**.

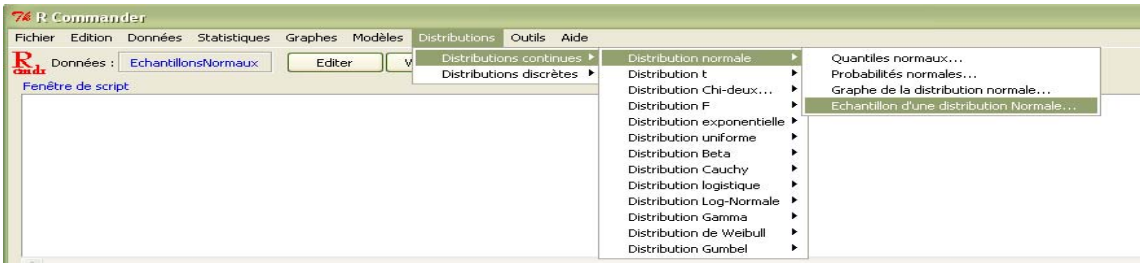
Voici un exemple simple de simulation d'une série de nombres tirés d'une loi normale centrée réduite, que l'on décrit ensuite par un tiges et feuille et un histogramme.

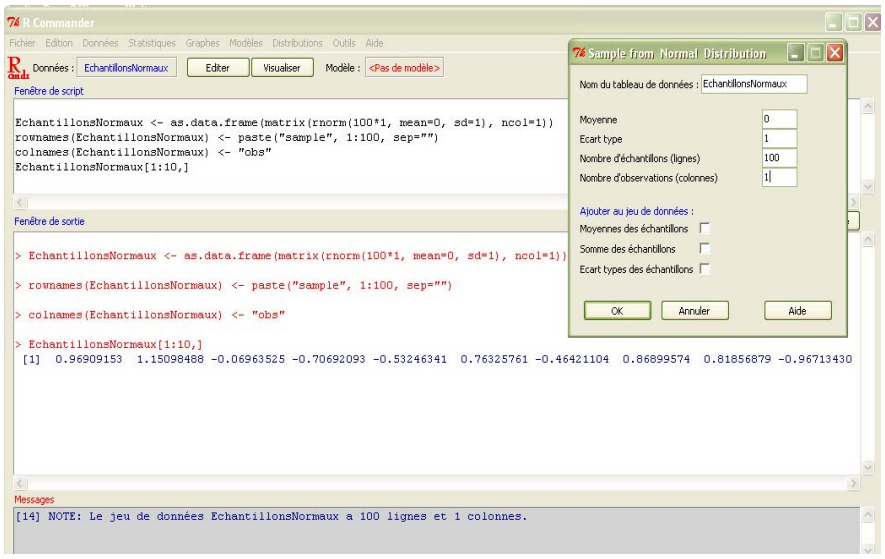


Chargement et exécution de Rcmdr

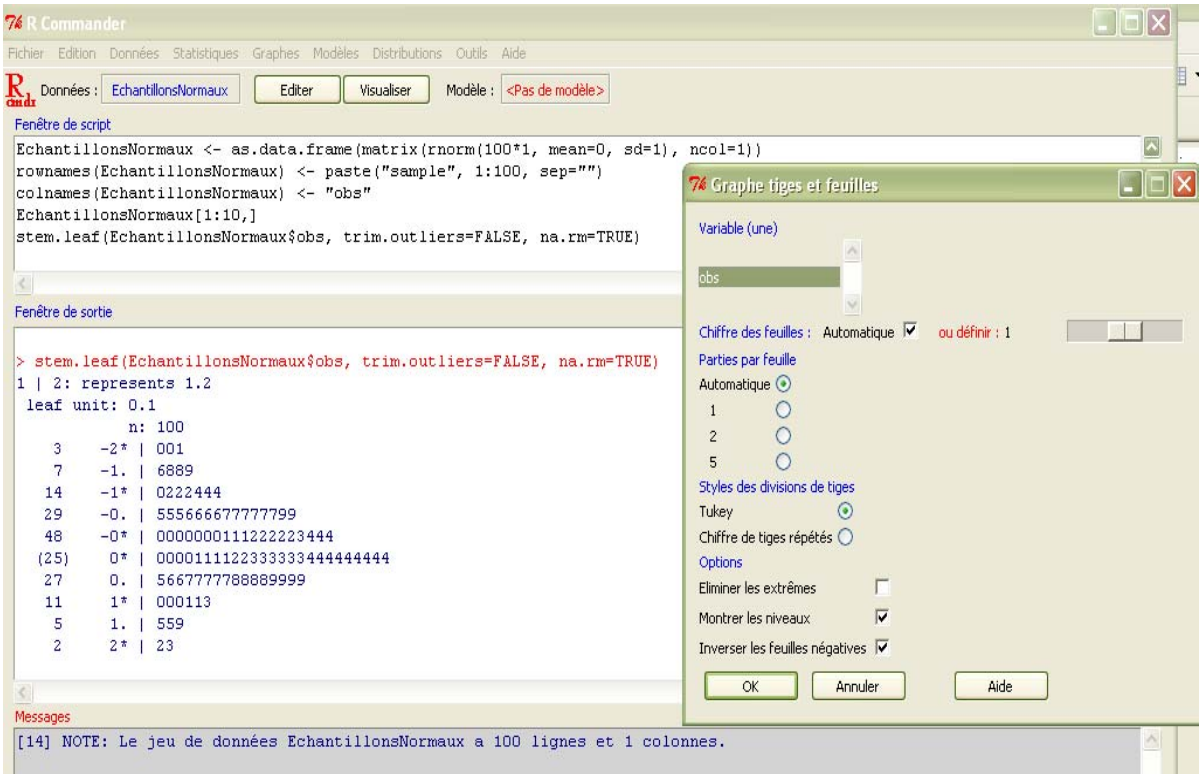
L'interface des menus à cliquer en français est lancée.

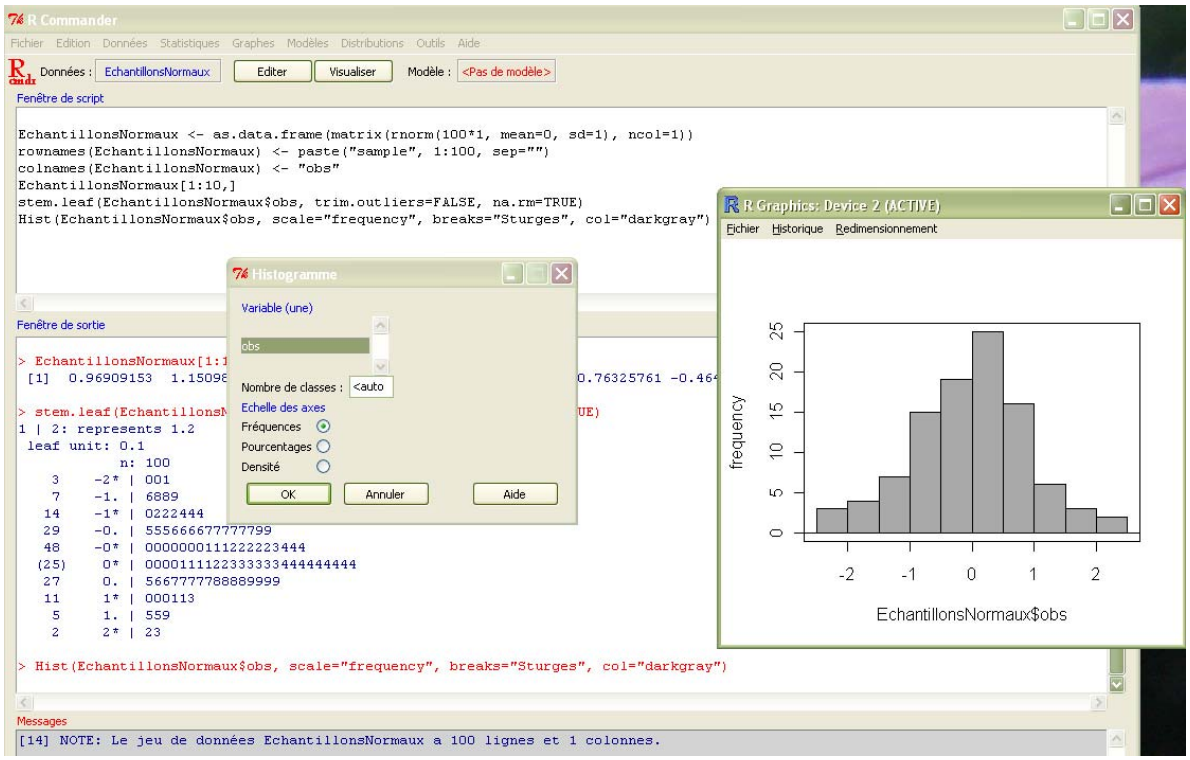
Utilisation directe : simulation de 100 nombres distribués selon la loi normale centré réduite et on fait afficher les 10 premières valeurs de la série simulée "EchantillonsNormaux".





Description de la série simulée, diagrammes en tiges et feuilles et histogramme :





3° Création et exécution des lignes de commandes

► Pour enchaîner les traitements ou programmer des fonctions **R**, il faut passer par les lignes de commande. On peut le faire directement dans la "console **R**", mais il est plus facile d'utiliser l'éditeur **Tinn-R** car il permet de saisir, d'enregistrer et d'utiliser les lignes de code que l'on

crée pour un traitement ou une fonction.

► On peut saisir et exécuter directement les lignes de commandes dans la console **R (R Console)**.

Pour accéder à la console **R**, il faut cliquer sur l'icône **R** créée lors de l'installation. Un fenêtre **RGui** (**G**raphic **u**ser **i**nterface) s'ouvre, qui contient, par défaut la console **R**, dans laquelle s'écrivent et s'exécutent les commandes et les fonctions **R**. La console **R** s'utilisera de préférence lorsque chaque ligne de commande est exécutée au fur et à mesure du traitement prévu. On exécute une ligne de commande en appuyant sur la touche entrée (valider). Une ligne peut comporter plusieurs commande séparées par des **";"**. Une commande peut s'écrire sur plusieurs lignes, des + apparaissent alors en début de ligne.

UTILISATION DIRECTE DE LA CONSOLE R

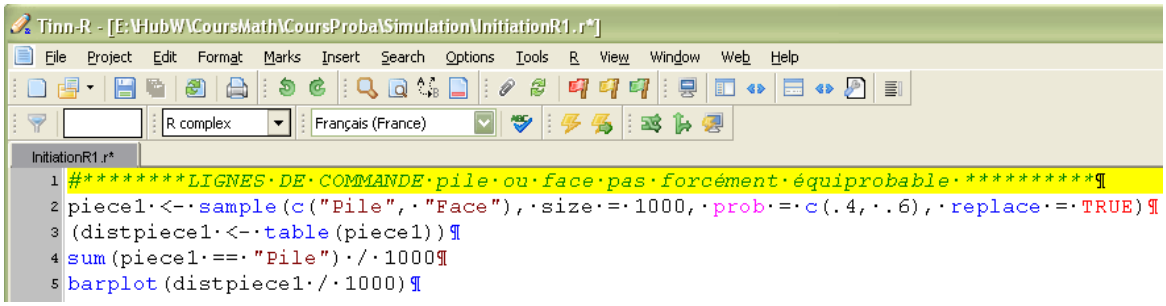
SAISIE DES LIGNES DE COMMANDE(S) ----->	**AFFICHAGE DU RÉSULTAT**

► Lignes de commandes saisies dans **Tinn-R** et exécutées dans la console **R**.

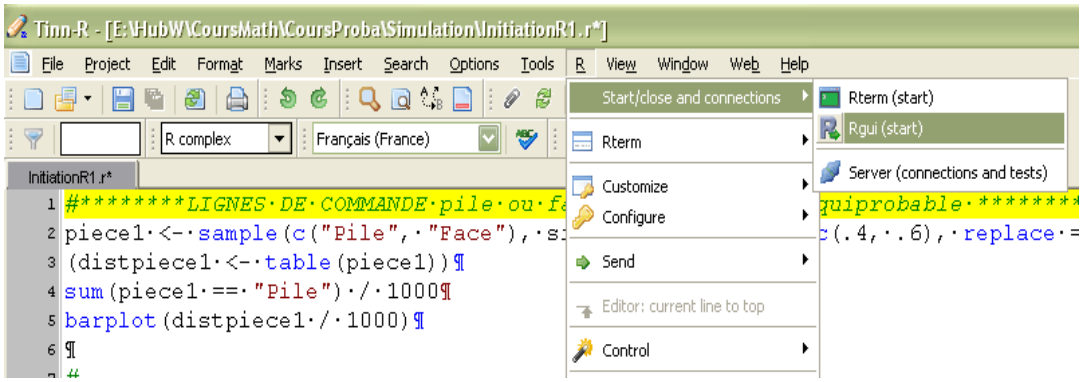
Lorsque l'on veut faire des traitement par lots (exécuter plusieurs lignes de commandes groupées) ou programmer des fonctions, l'utilisation de l'éditeur **Tinn-R** facilitera grandement l'écriture, la vérification et l'exécution des procédures ainsi créées.

La procédure classique consiste à suivre les étapes suivantes :

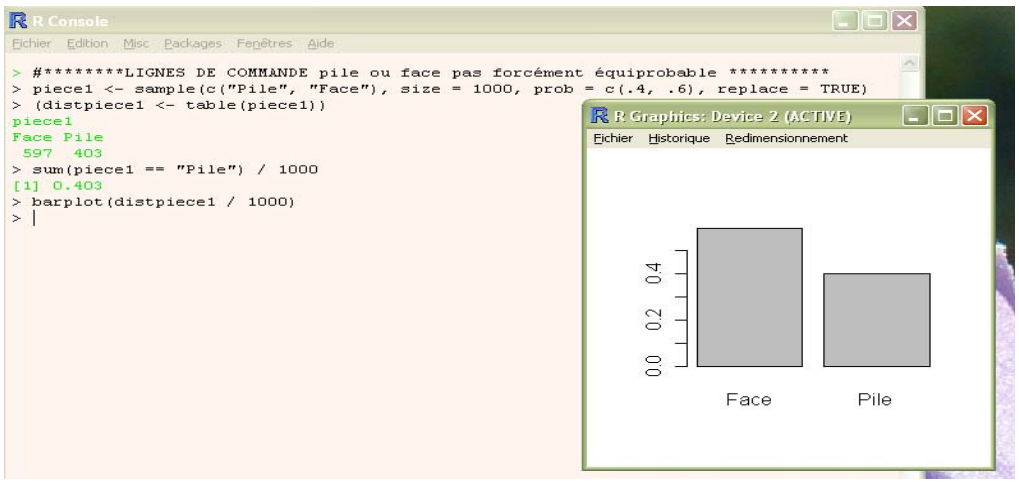
a Écriture du code dans **Tinn-R**.



b Lancer la console **R** (Rgui pour Graphic User Interface) à partir des menus **Tinn-R**.



c Copier coller les lignes de commande de **Tinn-R** dans la console **R**. Les lignes de commandes sont exécutées automatiquement et les résultats affichés, dans la console pour les résultats numériques, dans une ou plusieurs fenêtres graphiques, pour les graphiques.



3° Utilisation de fichiers contenant les lignes de commande ou les fonctions à exécuter

Pour utiliser les exemples fournis dans des fichiers, plusieurs cas peuvent se présenter.

► S'il s'agit de **lignes de commandes** fournies **dans le texte d'un fichier texte**, il suffit alors de sélectionner les lignes concernées et de les copier-coller dans la console **R** où elles seront automatiquement exécutées (sauf peut-être la dernière qu'il faudra valider) et les résultats seront affichés. Si on veut les modifier pour les adapter il faut soit les modifier dans le traitement de

texte ou bien passer par **Tinn-R**.
Par contre si c'est un fichier du type pdf il risque d'y avoir des problèmes causés par le fait que les fins de ligne sont transformés en fin de paragraphe. Il vaut mieux donc éviter, au moins dans la période d'apprentissage.
Il peut arriver, rarement, que la console **R** interprète mal certains caractères du traitement de texte, ayant un aspect visuel "normal". L'erreur est alors difficilement décelable, sauf en passant par **Tinn-R**.

- ► S'il s'agit de **lignes de commandes constituant une fonction**, fournies **dans le texte d'un fichier texte**, on procède comme précédemment en prenant bien soin de s'assurer que la dernière ligne a bien été validée (dans le cas contraire un + apparaît en début de ligne). La fonction vient d'être introduite en mémoire. Pour l'utiliser il suffit de saisir son **nom**, suivi sans espace de (). Ce **nom** figure obligatoirement en début du code. Dans ce cas, ce sont les valeurs des paramètre par défaut, indiquées dans la première ligne du code de la fonction, qui seront prises en compte. Pour utiliser d'autres valeurs, il suffit de les indiquer à l'intérieur des (). Exemple : **pileface()** réalise 2000 simulations du jet de deux pièces équilibrées (cf. le **II**). Pour en réaliser 6000, je saisis **pileface(6000)** ou **pileface(nbsim = 6000)**.
- ► Une bonne solution consiste à disposer des fichiers texte au format **Tinn-R** (extension **.r**) contenant les lignes de codes voulues. Un fichier **Tinn-R** est un simple fichier texte basique. Il peut contenir les lignes de code de une ou plusieurs procédures, les lignes de code de une ou plusieurs fonctions ou un mélange de lignes de procédures et de ligne de fonctions, comme par exemple dans le fichier **"InitiationR1.r"** qui contient les lignes de code des procédures et les lignes de codes des fonctions présentées dans les tableaux du **II**.
Il suffit alors de copier-coller les lignes de code que l'on veut exécuter.
- ► Une autre solution pertinente lorsqu'il s'agit d'utiliser une fonction, consiste à la faire lire et charger directement depuis le fichier source sur le disque dur. Prenons l'exemple de la fonction **IFexact1()**, dont les lignes de code sont dans le fichier **IF_BinomialExact1.r**. Il peut d'abord indiquer à **R** le dossier par défaut dans lequel se trouve le fichier à charger, en utilisant le menu « Fichier–Changer le répertoire courant ». Puis taper dans la console **R**, la commande **source("IF_BinomialExact1.r")**. Pour exécuter la fonction, il suffit ensuite de taper **Ifexact1()** ou, par exemple, **IFexact1(n = 150, p = .2, kobs = 25, proba = .95)**.

II - QUELQUES EXEMPLES SIMPLES COMMENTÉS

Convention typographique : Les lignes en orange contiennent les lignes de commande **R**. ***Les lignes en italique vert*** sont des parties de réponses de **R** (à ne pas coller dans la console). Les textes en turquoise ou **bleu clair** contiennent le code des fonctions **R**. **Les # commentaires sont en noir**, précédés de #. Les **mots en rouge sombre** sont les mots réservés aux commandes et fonctions internes de **R**.

`#**LIGNES DE COMMANDE pile ou face pas forcément équiprobable **`
`piece1 <- sample(c("Pile", "Face"), size = 1000,`
`  prob = c(.4, .6), replace = TRUE)`
`(distpiece1 <- table(piece1))`
`barplot(distpiece1 / 1000)`
`sum(piece1 == "Pile") / 1000`
	`<- est la commande d'affectation. C()`
	`créé un vecteur (au sens informatique).`
	`sample tire size(1000) fois avec remise`
	`dans l'ensemble {"Pile", "Face"}, avec`
	`une probabilité de 0,4 pour "Pile" et de`
	`0,6 pour "Face". Les n(1000) résultats`
	`obtenus sont mis dans le vecteur piece1.`
	`table(piece1) effectue le tri à plat`
	`(tableau des effectifs) de la série`
	`obtenue.`
	`barplot effectue le diagramme en barre.`
	`sum(piece1 == "Pile") compte le nombre`
	`de "Pile".`

`#**LIGNES DE COMMANDE pile ou face avec 2 pièces différentes **`
`piece1 <- sample(c("Pile", "Face"), size = 1000,`
`  prob = c(.4, .6), replace = T)`
`piece2 <- sample(c("Pile", "Face"), size = 1000,`
`  prob = c(.5, .5), replace = T)`
`deuxpieces <- paste(piece1,piece2, sep = "")`
`table(deuxpieces)`
`barplot(table(deuxpieces) / 1000)`
	`La pièce 1 est déséquilibrée, la pièce 2`
	`non.`
	
	`paste réuni deux à deux chacun des 1000`
	`résultats de piece1 et piece2, par`
	`exemple PileFace ...`

```r
#**FONCTION jet simultané de 2 pièces identiques équilibrées**
pileface <- function(nbsim = 2000){
 resultats <- rep(NA, 3)
 names(resultats) <- c("deuxpils", "deuxfaces", "autre")
 for(i in 1:nbsim){
 pieceA <- sample(c("Pile", "Face"), 1)
 pieceB <- sample(c("Pile", "Face"), 1)
 if(pieceA == "Pile" & pieceB == "Pile") {
         resultats[1] <- resultats[1] + 1} else {
          if(pieceA == "Face" & pieceB == "Face") {
                  resultats[2] <- resultats[2] + 1} else {
                   resultats[3] <- resultats[3] + 1      }
                                                 }
                  }
 print(resultats)
 print(resultats / nbsim)
 barplot(resultats / nbsim)
}
```

```
Fonction effectuant nbsim (2000) lancers
de deux pièces.
Paramètres et valeurs par défaut,début
du corps de fonction
Initialisation d'un "vecteur" à 3
composantes
nommer les 3 composantes du vecteur
début de la boucle des nbsim lancers
pieceA équilibrée
pieceB équilibrée
comptage des "Pile Pile"
comptage des "Face Face"
comptage des autres résultats.
Fin des tests et
des boucles
Affichage des résultats.


Fin du corps de fonction.
```

```r
#Le problème historique du grand duc de Toscane (Somme de 3 dés)
#****LIGNES DE COMMANDE Simulation GrandDuc****
de1 <- sample(c(1:6), 1000, replace = TRUE)
  (distde1 <- table(de1))
  barplot(distde1 / 1000)
de2 <- sample(c(1:6), 1000, replace = T)
  (distde2 <- table(de2))
  dev.new()
  barplot(distde2 / 1000)
de3 <- sample(c(1:6), 1000, replace = T)
  (distde3 <- table(de3))
  dev.new()
  barplot(distde3 / 1000)
de <- de1 + de2 + de3
(distde <- table(de))
  dev.new()
  barplot(distde / 1000)
  nbneuf <- sum(de == 9)
  nbdix <- sum(de == 10)
  cat("Fréquence des neuf =", nbneuf / 1000, "\n")
  cat("Fréquence des dix =", nbdix / 1000, "\n")
barplot(distde, xlab = "Somme des numéros des 3 faces",
 ylab = "Effectifs simulés",
 main = paste("Diagramme en barre de 1000 simulations\n du jet
de 3 dés équilibrés"))
```

```
1000 jets d'un dé équilibré, la série
des 1000 résultats est mise dans le
vecteur de1
tableau des effectifs de la série
obtenus
diagramme en barres

ouvre une nouvelle fenêtre graphique
même chose avec un autre dé équilibré,
la série des 1000 résultats est mise
dans le vecteur de2

même chose avec un autre dé équilibré,
la série des 1000 résultats est mise
dans le vecteur de3

somme, composante à composante des 3
vecteurs, les 1000 résultats sont mis
dans le vecteur de.

Tableau des effectifs de la série de,
diagramme en barres
comptage du nombre de neuf et du nombre
de 10.

affichage des résultats.
```

```r
#****FONCTION simulation Grand Duc********
simgrandduc <- function(nbsim=1000){
 de1 <- sample(c(1:6), nbsim, replace = TRUE)
 de2 <- sample(c(1:6), nbsim, replace = TRUE)
 de3 <- sample(c(1:6), nbsim, replace = TRUE)
 de <- de1 + de2 + de3
 distde <- table(de)
 print(distde)
 barplot(distde / nbsim)
}
```

```
La fonction effectue nbsim lancers de 3
dés. On additionne les résultats obtenus

tableau des effectifs des 1000 sommes
obtenues et leur diagramme en barres.
```

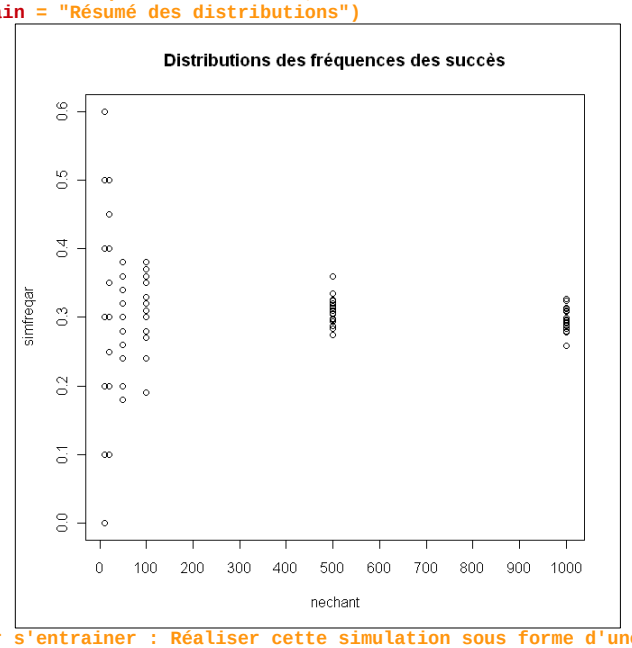
```r
#**LIGNES DE COMMANDE probabilité Grand Duc********
#**Somme des valeurs des faces obtenues en jetant 3 dés**
#***Calculer avec le modèle mathématique "exact"****
#****Construire l'univers correspondant à cette expérience***
serieS3de <- array(data = NA, dim = c(6, 6, 6))
 for(i in 1:6){
   for(j in 1:6){
     for(k in 1:6){
        serieS3de[i, j, k] <- i + j + k
                  }
              }
          }
 serieS3de
 (distS3de <- table(serieS3de))
 nbneuf <- sum(serieS3de == 9)
 nbdix <- sum(serieS3de == 10)
 cat("Probabilité de neuf =",nbneuf / 216,"\n")
 cat("Probabilité de dix =",nbdix / 216,"\n")
 dev.new()
 barplot(distS3de)
 graphics.off()
```

```
Initialisation d'un tableau de
dimension3

boucles imbriquées pour parcourir tous
les triplets possibles et générer
l'univers des résultats possibles : les
216 valeurs obtenues sont mises dans le
vecteur serieS3de.


Tableau des effectifs

Comptage du nombre de 9 et de10
calcul de la probabilité.
```

`#**Fonction probabilité Grand Duc********`
`probgranduc <- function(){`
`serieS3de <- array(data = NA, dim = c(6, 6, 6))`
`for(i in 1:6){`
`for(j in 1:6){`
`for(k in 1:6){`
`serieS3de[i, j, k] <- i + j + k`
`}`
`}`
`}`
`serieS3de`
`distS3de <- table(serieS3de)`
`nbneuf <- sum(serieS3de == 9)`
`nbdix <- sum(serieS3de == 10)`
`cat("Probabilité de neuf =",nbneuf / 216,"\n")`
`cat("Probabilité de dix =",nbdix / 216,"\n")`
`print(distS3de)`
`barplot(distS3de / 216)`
`}`
	Même chose sous la forme d'une fonction.

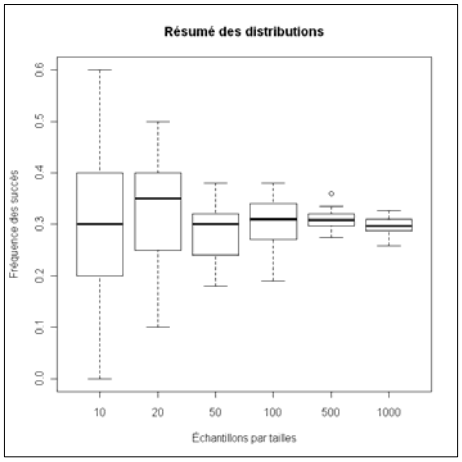
`######LIGNES DE COMMANDES-- CALCULS DE PROBABILITÉS ##########`
`#-------------- Loi binomiale ---------------`
`# Calcul de P(A <= X <= B), X étant une v.a. de distribution`
`# binomiale`
`# de paramètres n=100 et p=0,52.`
`# Les exemples choisis peuvent servir de base à une réflexion`
`# sur les différentes façons de déterminer un intervalle de`
`# fluctuation, à partir`
`# de l'exemple 1 (Monsieur Z du document d'inspection).`
`# P(42<=X<=62):`
`sum(dbinom(42:62, 100, .52))`
`# P(43<=X<=62):`
`sum(dbinom(43:62, 100, .52))`
`# P(42<=X<=61):`
`sum(dbinom(42:61, 100, .52))`
`# P(X<=41) ; P(X<=42) ; P(X<=43):`
`pbinom(41:43, 100, .52)`
`#----- Combinaisons et loi hypergéométrique --------`
`# Calcul de P(X=3) ; x=3 ; X de loi hypergéométrique de paramètres`
`# m = 3, n = 5, k = 4,`
`(proba <- choose(3, 3) * choose(5, 4-3) / choose(3+5, 4))`
`# Pour vérification :`
`(proba <- dhyper(x = 3, m = 3, n = 5, k = 4))`
	42:62 génère la suite des entiers de 42 à 62
	dbinom génère un vecteur des probabilités binomiales de P(X=42) à P(X=62). sum en fait la somme
	
	choose calcule les combinaisons
	
	dhyper calcule les probabilités hypergéométriques.

```
# **LIGNES DE COMMANDES ****** SIMULATIONS NUMÉRIQUES *****
# Illustration graphique de la loi des grands nombres :
# Lorsque n augmente, on observe les suites de distributions
# Les fréquences tendent vers une valeur limite : la probabilité
# Les écarts à cette valeur limite sont de plus en plus faibles
#
nechant <- rep(c(10, 20, 50, 100, 500, 1000),
               c(20, 20, 20, 20, 20, 20))
simfreqar <- c(rbinom(20, 10, .3)/10,
             rbinom(20, 20, .3)/20, rbinom(20, 50, .3)/50,
             rbinom(20, 100, .3)/100, rbinom(20, 500, .3)/500,
             rbinom(20, 1000, .3)/1000)
dev.off()
plot(nechant, simfreqar, xaxp = c(0, 1000, 10),
  main = "Distributions des fréquences des succès")
dev.new()
plot(as.factor(nechant), simfreqar,
  xlab = "Échantillons par tailles",
  ylab = "Fréquence des succès",
  main = "Résumé des distributions")
```



Pour s'entrainer : Réaliser cette simulation sous forme d'une fonction paramétrable...

On génère les tailles d'échantillons comme suit :
20 répétitions du nombre 10, 20 répétitions du nombre 20, …..., 20 répétitions du nombre 1000, mise dans nechant qui constituerons les abscisses des points à tracer.
20 nombres au hasard sont tirés dans une distribution binomiale(10, 0,3), 20 nombres au hasard sont tirés dans une distribution binomiale(20, 0,3), …, 20 nombres au hasard sont tirés dans une distribution binomiale(1000, 0,3), Les fréquences sont calculées en même temps.
Nuage de points des fréquences en fonction des tailles d'échantillons.
Résumé des séries de 20 valeurs (binomiales) sous formes de boites à moustaches.



Annexe 5 Méthode de Monte-Carlo [25]

La méthode de Monte-Carlo est une méthode probabiliste permettant le calcul approché d'intégrales (simples ou multiples) de fonctions quelle que soit leur régularité. C'est cette propriété qui explique son intérêt par rapport aux méthodes déterministes classiques.

Pour simplifier cette présentation, on suppose que l'on cherche à calculer $p = \int_0^1 f(x)\mathrm{d}x$ pour une fonction continue sur [0,1] à valeurs dans [0,1].

A. Méthode dite du « rejet »

Comme p est donc l'aire du domaine $D = \{(x,y) \in [0,1]^2 / y \leq f(x)\}$, une première méthode possible est de tirer aléatoirement un grand nombre de points du carré $[0,1]^2$ et de faire le quotient entre le nombre de points situés dans le domaine D et le nombre total de points.

Exemple

On peut voir sur la figure 17 ci-dessous le résultat graphique dans le cas de la fonction $f(x) = \frac{2}{3}(x^3 - x^2 + 1)$.

Avec $N = 10000$ points, une exécution de l'algorithme donne une valeur approchée de $\int_0^1 f(x)\mathrm{d}x$ égale à 0,6056.

La valeur exacte est $\frac{11}{18} \approx 0,6111$.

Code de l'algorithme

```
VARIABLES
    N EST_DU_TYPE NOMBRE
    x EST_DU_TYPE NOMBRE
    y EST_DU_TYPE NOMBRE
    i EST_DU_TYPE NOMBRE
    u EST_DU_TYPE NOMBRE
    aire EST_DU_TYPE NOMBRE
DEBUT_ALGORITHME
    LIRE N
    u PREND_LA_VALEUR 0
    POUR i ALLANT_DE 1 A N
        DEBUT_POUR
        x PREND_LA_VALEUR random()
        y PREND_LA_VALEUR random()
        SI (y<=F1(x)) ALORS
            DEBUT_SI
            u PREND_LA_VALEUR u+1
            TRACER_POINT (x,y)
            FIN_SI
        FIN_POUR
    AFFICHER u
    aire PREND_LA_VALEUR u/N
    AFFICHER aire
FIN_ALGORITHME
```

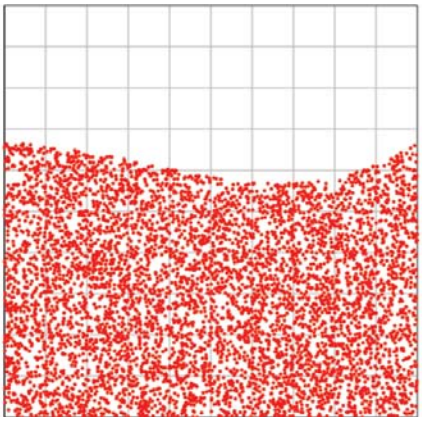
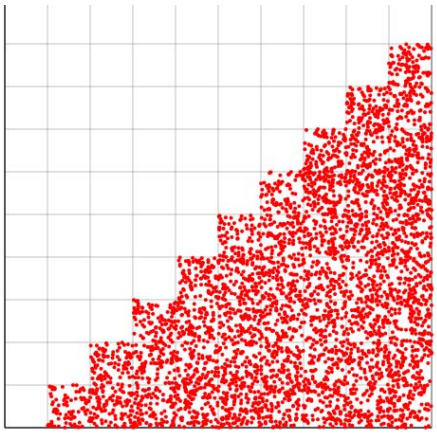


Figure 17 : tirage de 10000 points de $[0,1]^2$

[25] Ce paragraphe peut être réservé à une seconde lecture : son contenu n'est pas au programme mais peut être traité dans le cadre de l'accompagnement personnalisé en TS.



Voici ce que donne l'algorithme pour la fonction $f(x) = Ent(10x)/10$ où *Ent* désigne la partie entière.

On voit que la méthode fonctionne même avec des fonctions présentant des discontinuités. C'est son avantage sur les méthodes de calcul approché classiques (trapèzes, Simpson,...).

Document associé : monte carlo.alg

Justification

On considère deux variables indépendantes X et Y suivant la loi uniforme sur [0,1].

On admet que $\mathrm{P}\left(Y \le f(X)\right) = p$. L'exemple ci-dessous permet de vérifier cette propriété sur un cas particulier.

Si on considère $\left(X_1, Y_1\right), ..., \left(X_n, Y_n\right)$ n couples de variables aléatoires indépendantes suivant la loi uniforme sur [0,1], on a $\mathrm{P}\left(Y_k \le f(X_k)\right) = p$ pour tout $k \in \{1,..,n\}$.

Si S_n représente le nombre de couples $\left(X_k, Y_k\right)$ tels que $Y_k \le f(X_k)$, alors S_n suit une loi binomiale de paramètres *n* et *p*. La loi des grands nombres[26] permet d'affirmer que la suite $\left(\frac{S_n}{n}\right)$ converge en probabilité vers *p* c'est-à-dire que pour tout $\varepsilon > 0$, $\mathrm{P}\left(\left|\frac{S_n}{n} - p\right| \ge \varepsilon\right) \to 0$ quand *n* tend vers l'infini.

Si on génère avec un ordinateur un grand nombre de couples aléatoires $\left(X_k, Y_k\right)$, la proportion *f* de ces couples pour lesquels $Y_k \le f(X_k)$ fournit donc une valeur approchée de *p*.

De plus si $n \ge 30$ et $nf \ge 5$ et $n(1-f) \ge 5$ alors l'intervalle $\left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}}\right]$ est un intervalle de confiance de *p* au niveau 0,95.

Avec $n = 10000$ on obtient une précision de 0,01 avec une confiance de 0,95.

Exemple

On prend ici la fonction *f* définie par $f(x) = x^2$ et on pose $Z = f(X) - Y = X^2 - Y$ où les deux variables indépendantes X et Y suivent la loi uniforme sur [0,1].

On a $p = \mathrm{P}\left(Y \le X^2\right) = \mathrm{P}\left(Z \ge 0\right)$.

On admet que la densité de *Z* est la fonction *g* définie sur [-1,1] par $\begin{cases} g(x) = \sqrt{x+1} & si \quad x \in [-1,0] \\ g(x) = 1 - \sqrt{x} & si \quad x \in [0,1] \end{cases}$

Alors $\mathrm{P}\left(Z \ge 0\right) = \int_0^1 g(x)\mathrm{d}x = \frac{1}{3}$.

[26] Voir annexe 1.

B. Méthode de l'espérance

On admet le résultat suivant :

Si X est une variable aléatoire suivant une loi uniforme sur [0,1] et si f est une fonction continue sur [0,1] alors la variable aléatoire $Y = f(X)$ possède une espérance égale à $p = \int_0^1 f(x)\mathrm{d}x$.[27]

L'exemple ci-dessous donne une approche de ce résultat.

Si on considère n variables indépendantes $X_1, ..., X_n$ suivant une loi uniforme sur [0,1], alors la variable $\dfrac{\sum_{k=1}^{n} f(X_k)}{n}$ converge en probabilité vers p.

Exemple

On prend ici la fonction f définie par $f(x) = -\ln(1-x)$ pour $x \in [0,1[$ et on pose $Y = f(X) = -\ln(1-X)$ où X suit une loi uniforme sur [0,1].

On a $\mathrm{P}(Y \leq x) = \mathrm{P}\left(X \leq 1 - \mathrm{e}^{-x}\right) = 1 - \mathrm{e}^{-x}$ pour $x \in [0, +\infty[$.

Donc Y suit une loi exponentielle de paramètre 1. On sait alors que $\mathrm{E}(Y) = 1$.

Le calcul de l'intégrale $\int_0^1 -\ln(1-x)\mathrm{d}x$ est un exercice classique d'analyse.

```
VARIABLES
    x EST_DU_TYPE NOMBRE
    N EST_DU_TYPE NOMBRE
    S EST_DU_TYPE NOMBRE
    i EST_DU_TYPE NOMBRE
    aire EST_DU_TYPE NOMBRE
DEBUT_ALGORITHME
    S PREND_LA_VALEUR 0
    LIRE N
    POUR i ALLANT_DE 1 A N
        DEBUT_POUR
        x PREND_LA_VALEUR random()
        S PREND_LA_VALEUR S+F1(x)
        FIN_POUR
    aire PREND_LA_VALEUR S/N
    AFFICHER aire
FIN_ALGORITHME
```

Avec N=10000 et la fonction $f(x) = \dfrac{2}{3}(x^3 - x^2 + 1)$, une exécution de l'algorithme donne une valeur approchée de 0,6101.

Il peut être intéressant de comparer l'efficacité des deux méthodes.

On peut constater que la méthode de l'espérance est en général un peu plus précise.

Document associé : monte Carlo bis.alg

Remarque

La méthode de Monte-Carlo est couramment utilisée pour calculer des aires ou des volumes. Elle est plus aisée à mettre en œuvre que des méthodes déterministes.

[27] C'est un cas particulier d'un théorème de probabilité connu sous le nom de « théorème du transfert ».

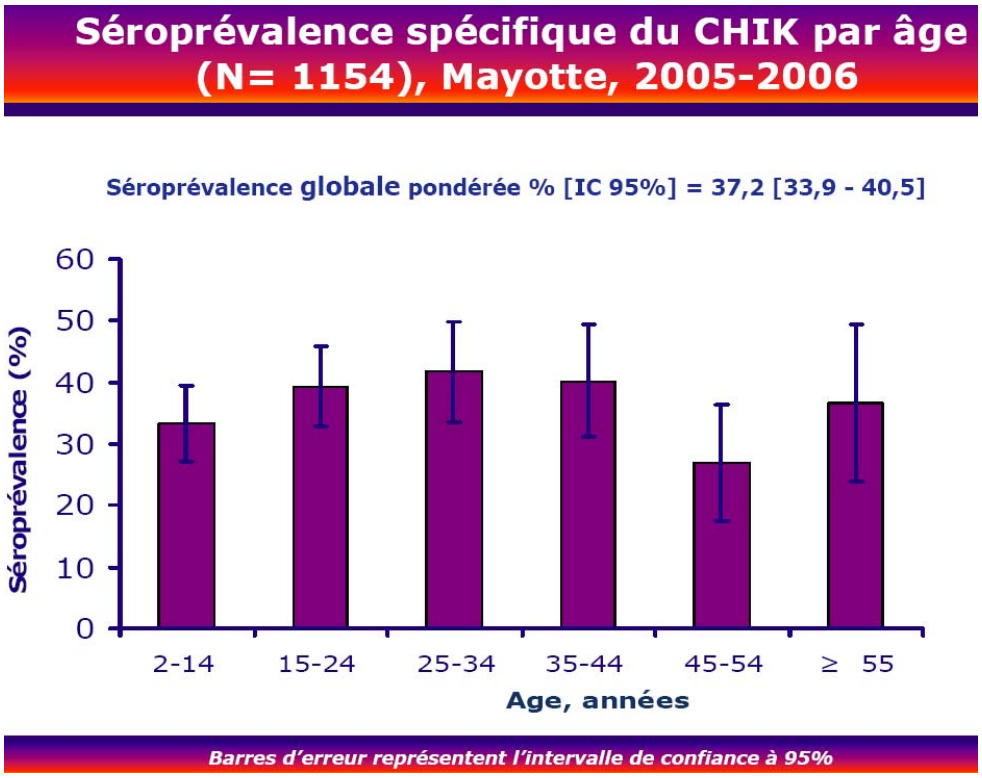
Annexe 6 Comparaison de deux frequences et difference significative

A. Une situation tres fréquente en sciences experimentales et en economie

Une situation très fréquente dans les démarches expérimentales est d'avoir à comparer deux séries de mesures, ou deux fréquences, pour étudier par exemple l'influence d'un facteur. On peut alors utiliser un test de comparaison (mais il s'agit alors d'une « boîte noire » dont on ne peut que difficilement justifier le fonctionnement au niveau des classes de terminales) ou, ce qui est souvent pratiqué dans les autres disciplines, comparer deux intervalles de confiance ou « barres d'erreurs ».

Exemple 1 : erreurs de mesure
Un document traitant de « l'estimation des incertitudes sur les erreurs de mesure » (Université de Strasbourg – Sciences physiques) indique qu'on peut « comparer des valeurs de façon très simple, par comparaison des segments d'incertitude » (il s'agit généralement d'intervalles de confiance à 95 %). « Les segments doivent avoir une partie commune ; dans le cas contraire, soit l'incertitude est trop faible (mauvaise évaluation de l'erreur), soit il y a un résultat erroné. Cette méthode est intéressante pour une comparaison globale de résultats expérimentaux, provenant par exemple d'expériences différentes. »

Exemple 2 : prévalence du chikungunya à Mayotte



Deux intervalles de confiance non disjoints : pas de différence significative.

Le paragraphe qui suit, bien que s'appuyant sur un certain nombre de résultats admis, a pour objectifs de donner des éléments de justification du critère retenu en terminale STI2D-STL pour juger d'une différence significative de deux proportions.

B. Comparaison de deux frequences

On souhaite comparer les proportions p_1 et p_2 d'un même caractère, dans deux populations distinctes, à partir de l'observation des fréquences f_1 et f_2 observées sur un échantillon de chacune des deux populations. La question posée est de savoir si la différence $f_1 - f_2$ est significative.

On suppose que les proportions des deux populations sont les mêmes : $p_1 = p_2$.
Sous cette hypothèse d'égalité des proportions des deux populations, la variable aléatoire $F_1 - F_2$, qui à chaque paire d'échantillons de taille n_1 et n_2, respectivement issus de chacune des deux populations, associe la différence $f_1 - f_2$ des fréquences observées, suit approximativement pour n_1 et n_2 assez grands, la loi normale $\mathcal{N}(0, \frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2})$.

Remarque
L'espérance de la variable $F_1 - F_2$ égale à la différence $p_1 - p_2$ est nulle compte tenu de l'hypothèse.
La variance de la variable $F_1 - F_2$ est égale à la somme des variances car les variances s'ajoutent si l'on suppose les variables F_1 et F_2 indépendantes.

Dans ces conditions, on peut déterminer l'intervalle de fluctuation de la variable $F_1 - F_2$ au seuil de 5%, d'où :

$$P\left(-1{,}96\sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}} \leq F_1 - F_2 \leq 1{,}96\sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}\right) = 0{,}95.$$

On conclut en disant que l'observation d'une différence $f_1 - f_2$, obtenue à partir des fréquences observées, vérifiant $|f_1 - f_2| > 1{,}96\sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}$ remet en question l'hypothèse $p_1 = p_2$ puisque avec l'hypothèse $p_1 = p_2$ cette situation n'a que 5% de chances de se produire.

C. Intersection de deux intervalles de confiance

Conformément aux notions présentées en classe de terminale, on peut déterminer à partir de l'observation f_1, un intervalle de confiance pour la proportion p_1 au niveau de confiance de 95% :

$$\left[f_1 - 1{,}96\sqrt{\frac{f_1(1-f_1)}{n_1}}, f_1 + 1{,}96\sqrt{\frac{f_1(1-f_1)}{n_1}}\right].$$

De même, on peut déterminer, à partir de l'observation f_2, un intervalle de confiance pour la proportion p_2 au niveau de confiance de 95 % :

$$\left[f_2 - 1{,}96\sqrt{\frac{f_2(1-f_2)}{n_2}}, f_2 + 1{,}96\sqrt{\frac{f_2(1-f_2)}{n_2}}\right].$$

On peut alors décider qu'il existe une « différence significative » entre f_1 et f_2 lorsque les intervalles de confiance précédents sont disjoints, c'est-à-dire lorsque :

$$|f_1 - f_2| > 1{,}96\left(\sqrt{\frac{f_1(1-f_1)}{n_1}} + \sqrt{\frac{f_2(1-f_2)}{n_2}}\right).$$

Si l'on compare ce critère de « différence significative » au précédent, on constate qu'il est plus « sévère » puisque :

$$\sqrt{\frac{f_1(1-f_1)}{n_1}} + \sqrt{\frac{f_2(1-f_2)}{n_2}} \geq \sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}.$$