IUT Villetaneuse
Université Sorbonne Paris Nord

BUT Science des données
Statistique Inférentielle (R2.08)

Année 2023-2024

Table des matières

Chapitre 0

Présentation de la ressource R2.08

La ressource R2.08 est une introduction à la statistique inférentielle (ou statistique déductive). Cette branche des statistiques regroupe les méthodes ayant pour objet d'obtenir des informations sur toute une population lorsque l'on a accès seulement à une partie de celle-ci. Elle s'appuie fortement sur la théorie des probabilités et les ressources antérieures concernant cette matière (R1.05 et R2.06) sont donc des prérequis.

Ce document contient l'essentiel du cours et des exercices qui seront étudiés en classe. Vous devez l'avoir avec vous à chaque séance. Les notions vues pendant une séance doivent être reprises lors d'un travail personnel et connues dès la séance suivante. Il est aussi recommandé de lire les démonstrations données dans le texte afin de mieux comprendre les idées. Enfin il est souhaitable de se munir d'une calculatrice scientifique, en particulier pour le chapitre 2.

Trois évaluations sont prévues :
- un contrôle court pour vérifier l'apprentissage des notions du chapitre 1 (séance 2 ou 3) ;
- un contrôle long en fin de chapitre 1 ;
- un contrôle long en fin de chapitre 2.

Contact : M. Bonino : bonino@sorbonne-paris-nord.fr

Chapitre 1

Estimation ponctuelle

Sommaire

1.1 Notion d'échantillon

Définition 1.1

Un échantillon de taille n (ou n-échantillon) d'une loi de probabilité $\mathbb{P}_*$ est une suite $X_1, X_2, \cdots, X_n$ de n variables aléatoires réelles (v.a.r.) indépendantes et suivant toutes la loi $\mathbb{P}_*$, c'est à dire que $\mathbb{P}_{X_i} = \mathbb{P}_*$ pour tout $i \in \{1, \cdots, n\}$. De façon abrégée, on dit que les X_i sont i.i.d. (indépendantes et identiquement distribuées) de loi $\mathbb{P}_*$.

- Implicitement, les variables aléatoires X_i sont toutes définies sur un même univers Ω (ainsi ce sont des fonctions $X_i : \Omega \to \mathbb{R}$) et $\mathbb{P}$ est une loi de probabilité sur Ω. Cet ensemble Ω et $\mathbb{P}$ peuvent être délicats à définir mathématiquement mais nous ne nous en préoccuperons pas dans ce cours. Nous aurons seulement besoin de considérer la loi de probabilité des X_i, c'est à dire $\mathbb{P}_{X_i}$ définie par $\mathbb{P}_{X_i}(E) = \mathbb{P}(X_i \in E)$ pour $E \subset X_i(\Omega)$.
- Intuitivement, les hypothèses de la Définition 1.1 signifient que connaître les valeurs prises par certaines X_i n'apporte pas d'information sur les valeurs prises par les autres (indépendance des X_i) et de plus que toutes les X_i prennent des valeurs données avec la même probabilité (les X_i suivent une loi commune, notée $\mathbb{P}_*$). Cette définition modélise des situations où l'on observe une certaine caractéristique sur n « individus » représentatifs d'une « population cible » dont la taille est grande par rapport à n. Par exemple :

- Les sondages : on interroge sur un sujet donné n personnes représentatives d'une grande population humaine et X_i est la réponse de la i^{eme} personne interrogée.

- Le contrôle qualité : pour s'assurer de la conformité d'une production industrielle à une norme (de poids, de robustesse,...), on prélève n objets issus de cette production et X_i est la mesure de la caractéristique étudiée sur le i^{eme} objet prélevé.

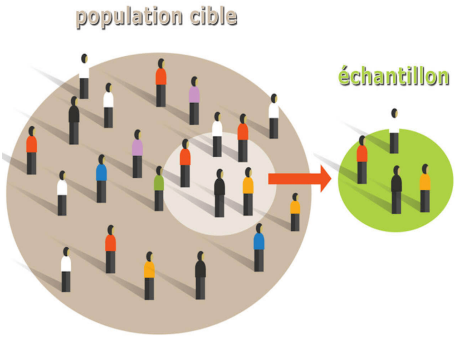


FIGURE 1.1 – Échantillon extrait d'une population

1.2 Estimateur

1.2.1 Généralités

Position du problème : On a une loi de probabilité $\mathbb{P}_*$ dont un paramètre θ (par exemple la moyenne ou la variance) est inconnu. À l'aide d'un échantillon de la loi $\mathbb{P}_*$, on calcule une valeur permettant d'estimer θ.

Définition 1.2

Soit θ un paramètre d'une loi de probabilité $\mathbb{P}_*$. Un estimateur de θ est une variable aléatoire $\widehat{\theta}_n$ définie en fonction d'un n-échantillon $X_1, X_2, \cdots, X_n$ de $\mathbb{P}_*$ dans le but d'obtenir une valeur approchée de θ.

Exemple 1.1 *L'estimateur habituel de l'espérance (ou moyenne) $\mathbb{E}(\mathbb{P}_*)$ est la* moyenne d'échantillon *(appelée aussi* moyenne empirique*), c'est à dire la la variable aléatoire*

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

1.2.2 Qualités d'un estimateur

On attend souvent des propriétés supplémentaires qui assurent que $\widehat{\theta}_n$ est un « bon » estimateur de θ. Dans ce cours, on se place dans la situation où $\theta \in \mathbb{R}$ (on peut en fait adapter ce qui suit au cas où $\theta \in \mathbb{R}^d$).

Estimateurs avec ou sans biais

Définition 1.3

Soit $\widehat{\theta}_n$ un estimateur du paramètre θ.

- *On dit que $\widehat{\theta}_n$ est sans biais si $\mathbb{E}(\widehat{\theta}_n) = \theta$ pour tout n ; dans le cas contraire, on parle d'estimateur biaisé.*
- *On dit que $\widehat{\theta}_n$ est asymptotiquement sans biais si $\lim_{n\to+\infty} \mathbb{E}(\widehat{\theta}_n) = \theta$.*

Convergence d'un estimateur

Il est aussi raisonnable d'attendre que le paramètre (inconnu) θ soit de mieux en mieux approché lorsque la taille n de l'échantillon $X_1, X_2, \cdots, X_n$ grandit. Il existe différentes notions de convergence traduisant cette idée. On se limite dans ce cours aux notions suivantes :

Définition 1.4

Soit $\widehat{\theta}_n$ un estimateur d'un paramètre θ.

- *On dit que $\widehat{\theta}_n$ converge presque sûrement (vers θ) si $\mathbb{P}(\lim_{n\to+\infty} \widehat{\theta}_n = \theta) = 1$.*
- *On dit que $\widehat{\theta}_n$ converge en moyenne quadratique (vers θ) si $\lim_{n\to+\infty} \mathbb{E}\big((\widehat{\theta}_n - \theta)^2\big) = 0$. Le nombre $\mathbb{E}\big((\widehat{\theta}_n - \theta)^2\big)$ est appelé risque quadratique moyen.*
- *On dit que $\widehat{\theta}_n$ est consistant, ou qu'il converge en probabilité (vers θ), si pour tout réel $a > 0$ on a $\lim_{n\to+\infty} \mathbb{P}(|\widehat{\theta}_n - \theta| \geq a) = 0$.*

La convergence en probabilité est la plus faible des trois notions de convergence ci-dessus. Plus précisément on a le résultat suivant :

Théorème 1.1

Soit $\widehat{\theta}_n$ un estimateur du paramètre θ.

- *Si $\widehat{\theta}_n$ converge presque sûrement alors il converge aussi en probabilité.*
- *Si $\widehat{\theta}_n$ converge en moyenne quadratique alors il converge aussi en probabilité.*

Pour expliquer en partie le Théorème 1.1, commencons par donner une inégalité classique en probabilités :

Proposition 1.1 (*Inégalité de Markov*)

Soit X une v.a.r. positive admettant une espérance. Alors on a

$$\forall \alpha > 0 \qquad \mathbb{P}(X \geq \alpha) \leq \frac{\mathbb{E}(X)}{\alpha}.$$

Preuve. Justifions d'abord cette inégalité quand X est discrète. On note $\mathcal{X} \subset \mathbb{R}^+$ l'ensemble des valeurs que peut prendre X. On a

$$\mathbb{E}(X) = \sum_{x \in \mathcal{X}} x \times \mathbb{P}(X = x) \;=\; \sum_{\substack{x \in \mathcal{X} \\ 0 \leq x < \alpha}} \underbrace{x \times \mathbb{P}(X = x)}_{\geq 0} \quad + \quad \sum_{\substack{x \in \mathcal{X} \\ x \geq \alpha}} \underbrace{x \times \mathbb{P}(X = x)}_{\geq \alpha \times \mathbb{P}(X = x)}$$

donc

$$\mathbb{E}(X) \geq \alpha \times \sum_{\substack{x \in \mathcal{X} \\ x \geq \alpha}} \mathbb{P}(X = x) \;=\; \alpha \times \mathbb{P}(X \geq \alpha)$$

puis on obtient l'inégalité voulue en divisant par $\alpha > 0$ de part et d'autre du dernier signe $\geq$.

Supposons maintenant que X est une variable aléatoire avec une densité de probabilité f. Le calcul est analogue au cas discret, en remplaçant les sommes ($\sum$) par des intégrales ($\int$). Puisque X est positive, la densité f est nulle sur $]-\infty, 0[$ et on a alors

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x f(x)\, dx \;=\; \int_0^{\alpha} \underbrace{x f(x)}_{\geq 0}\, dx \quad + \quad \int_{\alpha}^{+\infty} \underbrace{x f(x)}_{\geq \alpha f(x)}\, dx$$

donc

$$\mathbb{E}(X) \geq \alpha \times \int_{\alpha}^{+\infty} f(x)\, dx \;=\; \alpha \times \mathbb{P}(X \geq \alpha)$$

d'où le résultat. ■

Preuve partielle du Théorème 1.1. Nous admettrons le premier point. Pour le deuxième point, considérons un réel $a > 0$ quelconque et appliquons l'inégalité de Markov à la variable aléatoire $X = (\widehat{\theta}_n - \theta)^2$ et en prenant $\alpha = a^2$. On obtient, sous l'hypothèse que $\widehat{\theta}_n$ converge en moyenne quadratique :

$$0 \leq \mathbb{P}\big((\widehat{\theta}_n - \theta)^2 \geq a^2\big) \leq \frac{\mathbb{E}\big((\widehat{\theta}_n - \theta)^2\big)}{a^2} \underset{n \to +\infty}{\longrightarrow} 0\ .$$

donc aussi $\mathbb{P}\big((\widehat{\theta}_n - \theta)^2 \geq a^2\big) \underset{n \to +\infty}{\longrightarrow} 0$. Mais les évènements $\{(\widehat{\theta}_n - \theta)^2 \geq a^2\}$ et $\{|\widehat{\theta}_n - \theta| \geq a\}$ sont égaux (car $x^2 \geq a^2 \Leftrightarrow |x| \geq a$) donc ils ont la même probabilité, ce qui donne bien

$$\mathbb{P}\big(|\widehat{\theta}_n - \theta| \geq a\big) \underset{n \to +\infty}{\longrightarrow} 0\ .$$

■

On dispose du critère suivant qui est souvent utile pour montrer qu'un estimateur est convergent en moyenne quadratique.

Proposition 1.2

Soit $\widehat{\theta}_n$ un estimateur du paramètre θ. Alors $\widehat{\theta}_n$ converge en moyenne quadratique si et seulement si $\widehat{\theta}_n$ est asymptotiquement sans biais et $Var(\widehat{\theta}_n) \underset{n\to+\infty}{\longrightarrow} 0$.

Preuve. En utilisant les propriétés usuelles de la variance et de l'espérance, on a

$$Var(\widehat{\theta}_n) = Var(\widehat{\theta}_n - \theta) = \mathbb{E}\big((\widehat{\theta}_n - \theta)^2\big) - \big(\mathbb{E}(\widehat{\theta}_n - \theta)\big)^2 = \mathbb{E}\big((\widehat{\theta}_n - \theta)^2\big) - \big(\mathbb{E}(\widehat{\theta}_n) - \theta\big)^2$$

donc

$$\mathbb{E}\big((\widehat{\theta}_n - \theta)^2\big) = \underbrace{Var(\widehat{\theta}_n)}_{\geq 0} + \underbrace{\big(\mathbb{E}(\widehat{\theta}_n) - \theta\big)^2}_{\geq 0}$$

lorsque ces espérances et variances existent. En conséquence on a $\mathbb{E}\big((\widehat{\theta}_n - \theta)^2\big) \underset{n\to+\infty}{\longrightarrow} 0$ si et seulement si $Var(\widehat{\theta}_n) \underset{n\to+\infty}{\longrightarrow} 0$ et $\mathbb{E}(\widehat{\theta}_n) \underset{n\to+\infty}{\longrightarrow} \theta$. ■

Comparaison d'estimateurs

Il arrive souvent que l'on dispose de plusieurs estimateurs pour un même paramètre. On préfère en général les estimateurs (asymptotiquement) sans biais. Pour « départager » deux estimateurs sans biais, on peut comparer leurs variances :

Définition 1.5

Étant donnés deux estimateurs sans biais $\widehat{\theta}_n$ et $\widetilde{\theta}_n$ d'un même paramètre θ, on dit que $\widehat{\theta}_n$ est plus efficace que $\widetilde{\theta}_n$ si l'on a $Var(\widehat{\theta}_n) < Var(\widetilde{\theta}_n)$ pour tout n à partir d'un certain rang.

1.3 Estimateurs de la moyenne et de la variance

On considère dans ce paragraphe un échantillon $X_1, X_2, \cdots, X_n$ d'une loi de probabilité $\mathbb{P}_*$ et on s'intéresse à l'estimation de deux paramètres essentiels qui sont l'espérance (ou moyenne) μ et la variance σ^2 de $\mathbb{P}_*$ (en supposant qu'ils existent).

1.3.1 Estimateur de la moyenne

Définition 1.6

La moyenne d'échantillon (ou moyenne empirique) est la variable aléatoire

$$\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n}.$$

Proposition 1.3

Si la loi de probabilité $\mathbb{P}_*$ admet une espérance μ alors la variable aléatoire $\overline{X}_n$ est un estimateur sans biais de μ qui converge presque sûrement. Si de plus la loi $\mathbb{P}_*$ admet une variance σ^2 alors on a $Var(\overline{X}_n) = \dfrac{\sigma^2}{n}$ et $\overline{X}_n$ converge aussi en moyenne quadratique.

Preuve partielle. La linéarité de l'espérance donne $\mathbb{E}(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i)$. De plus, toutes les X_i suivent la loi $\mathbb{P}_*$ donc $\mathbb{E}(X_i) = \mu$ pour tout i et ainsi $\mathbb{E}(\overline{X}_n) = \frac{1}{n} \times n \times \mu = \mu$. Nous admettrons la convergence presque sûre de $\overline{X}_n$, qui est un théorème difficile de probabilités appelé loi (forte) des grands nombres. Supposons de plus que $\mathbb{P}_*$ admet une variance, notée σ^2. Comme les X_i sont indépendantes et suivent toutes la loi $\mathbb{P}_*$ on a

$$Var(X_1 + \cdots + X_n) = \sum_{i=1}^n \underbrace{Var(X_i)}_{=\sigma^2} = n\sigma^2$$

puis

$$Var(\overline{X}_n) = \frac{1}{n^2} \times Var(X_1 + \cdots + X_n) = \frac{\sigma^2}{n} \underset{n\to+\infty}{\longrightarrow} 0 \ .$$

La Proposition 1.2 montre alors que $\overline{X}_n$ converge en moyenne quadratique. ■


On ne confondra pas la moyenne (ou espérance) μ de la loi $\mathbb{P}_*$ avec la moyenne empirique $\overline{X}_n$. La première est un *nombre* (inconnu) alors que la deuxième est une *variable aléatoire* dont les valeurs observées permettent d'estimer μ.

1.3.2 Estimateur de la variance lorsque la moyenne est connue

Proposition 1.4

Si la loi de probabilité $\mathbb{P}_*$ admet une variance σ^2 et si sa moyenne μ est connue, alors la variable aléatoire

$$\overline{S_n^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$$

est un estimateur sans biais de σ^2 qui converge presque sûrement.

Preuve partielle. Nous admettrons la convergence. Le fait que $\overline{S_n^2}$ soit sans biais se vérifie de la façon suivante. On a $(X_i - \mu)^2 = X_i^2 + \mu^2 - 2\mu X_i$ pour chaque entier $i \in \{1, \cdots, n\}$ donc

$$\begin{aligned}
\overline{S_n^2} &= \frac{1}{n}\sum_{i=1}^n \left(X_i^2 + \mu^2 - 2\mu X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^n X_i^2 + \frac{1}{n}\underbrace{\sum_{i=1}^n \mu^2}_{=n\mu^2} + \frac{1}{n}\sum_{i=1}^n -2\mu X_i \\
&= \frac{1}{n}\sum_{i=1}^n X_i^2 + \mu^2 - 2\mu\frac{1}{n}\sum_{i=1}^n X_i. \\
&= \frac{1}{n}\sum_{i=1}^n X_i^2 + \mu^2 - 2\mu\overline{X}_n
\end{aligned}$$

Ensuite, en utilisant la linéarité de l'espérance, on obtient

$$\begin{aligned}
\mathbb{E}(\overline{S_n^2}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i^2) + \underbrace{\mathbb{E}(\mu^2)}_{=\mu^2} - 2\mu\underbrace{\mathbb{E}(\overline{X}_n)}_{=\mu} \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(Var(X_i) + \big(\mathbb{E}(X_i)\big)^2\Big) + \mu^2 - 2\mu^2 \\
&= \frac{1}{n}\underbrace{\sum_{i=1}^{n}(\sigma^2+\mu^2)}_{=n(\sigma^2+\mu^2)} - \mu^2 \\
&= \sigma^2
\end{aligned}$$

ce qui montre que $\overline{S_n^2}$ est un estimateur sans biais de σ^2. ■

1.3.3 Estimateurs de la variance lorsque la moyenne est inconnue

Définition 1.7

- *La variance d'échantillon (ou variance empirique) est la variable aléatoire*

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 = \big(\frac{1}{n}\sum_{i=1}^{n}X_i^2\big) - \overline{X_n}^2.$$

- *La variance d'échantillon corrigée est la variable aléatoire*

$$S_n^{2c} = \frac{n}{n-1}S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2.$$

(la lettre c dans la notation S_n^{2c} vaut pour le mot « corrigée »)

Proposition 1.5

Si la loi de probabilité $\mathbb{P}_*$ admet une variance σ^2 alors

- *La variable aléatoire S_n^2 est un estimateur biaisé et asymptotiquement sans biais de σ^2 ;*
- *La variable aléatoire S_n^{2c} est un estimateur sans biais de σ^2 ;*
- *Les estimateurs S_n^2 et S_n^{2c} convergent presque sûrement.*

Preuve partielle. Nous admettrons le troisième point de cette proposition (qui est une conséquence de la loi des grands nombres). Les deux premiers se démontrent de la façon suivante. Tout d'abord

on a $(X_i - \overline{X}_n)^2 = X_i^2 + \overline{X}_n^2 - 2\overline{X}_n X_i$ pour chaque entier $i \in \{1, \cdots, n\}$ donc

$$\begin{aligned} S_n^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 + \overline{X}_n^2 - 2\overline{X}_n X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \frac{1}{n}\underbrace{\sum_{i=1}^{n}\overline{X}_n^2}_{=n\overline{X}_n^2} + \frac{1}{n}\sum_{i=1}^{n} -2\overline{X}_n X_i \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \overline{X}_n^2 - 2\overline{X}_n \frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2. \end{aligned}$$

Ensuite on utilise à nouveau la linéarité de l'espérance :

$$\begin{aligned} \mathbb{E}(S_n^2) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \mathbb{E}(\overline{X}_n^2) \\ &= \frac{1}{n}\underbrace{\sum_{i=1}^{n}\left(Var(X_i) + \big(\mathbb{E}(X_i)\big)^2\right)}_{=n(\sigma^2+\mu^2)} - \left(Var(\overline{X}_n) + \big(\mathbb{E}(\overline{X}_n)\big)^2\right) \\ &= \sigma^2 + \mu^2 - (\frac{\sigma^2}{n} + \mu^2) \\ &= \frac{n-1}{n}\sigma^2 \neq \sigma^2 \end{aligned}$$

ce qui montre que S_n^2 est un estimateur biaisé de σ^2. Il est cependant asymptotiquement sans biais car $\lim_{n\to+\infty} \frac{n-1}{n}\sigma^2 = \sigma^2$.

Par définition $S_n^{2c} = \frac{n}{n-1}S_n^2$ donc $\mathbb{E}(S_n^{2c}) = \frac{n}{n-1}\mathbb{E}(S_n^2) = \sigma^2$ ce qui montre que S_n^{2c} est un estimateur sans biais de σ^2. ■

On ne confondra pas la variance σ^2 de $\mathbb{P}_*$ avec la variance empirique S_n^2 ni avec la variance empirique corrigée S_n^{2c}. La première est un *nombre* (inconnu) alors que S_n^2 et S_n^{2c} sont des *variables aléatoires* dont les valeurs observées permettent d'estimer σ^2.

1.4 Estimateurs et estimations

Il est important de bien distinguer les *estimateurs* et les *estimations* qu'ils fournissent. En statistiques, il est habituel de désigner les variables aléatoires par des lettres majuscules et les réalisations de ces mêmes variables aléatoires (c'est à dire les valeurs observées expérimentalement) par les lettres minuscules correspondantes. Par exemple, si X_1, X_2, X_3 est un échantillon de taille 3 d'une loi de probalité $\mathbb{P}_*$, on écrira $x_1 = 10$, $x_2 = 5$, $x_3 = 9$ pour dire que les valeurs observées expérimentalement pour ces trois variables aléatoires sont respectivement 10 ;5 ;9. Un *estimateur* de $\mathbb{E}(\mathbb{P}_*)$ est la moyenne d'échantillon $\overline{X}_3 = \frac{1}{3}(X_1 + X_2 + X_3)$. Une *estimation* de $\mathbb{E}(\mathbb{P}_*)$ est alors $\bar{x}_3 = \frac{1}{3}(10 + 5 + 9) = 8$. Un *estimateur* de $Var(\mathbb{P}_*)$ est la variance d'échantillon

corrigée $S_3^{2c} = \frac{1}{2}\big((X_1 - \overline{X}_3)^2 + (X_2 - \overline{X}_3)^2 + (X_3 - \overline{X}_3)^2\big)$. Une *estimation* de $Var(\mathbb{P}_*)$ est $s_3^{2c} = \frac{1}{2}\big((10-8)^2 + (5-8)^2 + (9-8)^2\big) = 7$.

1.5 La méthode des moments

1.5.1 Moments d'une variable aléatoire et d'une loi de probabilité

Définition 1.8

Soient $\mathbb{P}_*$ une loi de probabilité et X une variable aléatoire suivant la loi $\mathbb{P}_*$, c'est à dire telle que $\mathbb{P}_X = \mathbb{P}_*$. Pour $k \in \mathbb{N}^*$, le moment et le moment centré d'ordre k de X, et de $\mathbb{P}_*$, sont les nombres μ_k et m_k définis dans le tableau ci-dessous, à la condition que les sommes et les intégrales correspondantes convergent (sinon on dit que le moment d'ordre k n'existe pas).

	moment d'ordre k	*moment centré d'ordre k*
Si la loi $\mathbb{P}_*$ est discrète de support $\mathcal{X}$	$\mu_k = \mathbb{E}(X^k)$ $= \sum_{x\in\mathcal{X}} x^k \times \mathbb{P}_*(x)$	$m_k = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^k\Big)$ $= \sum_{x\in\mathcal{X}} \big(x - \mathbb{E}(X)\big)^k \times \mathbb{P}_*(x)$
Si la loi $\mathbb{P}_*$ a une densité f	$\mu_k = \mathbb{E}(X^k)$ $= \int_{-\infty}^{+\infty} x^k f(x)\,dx$	$m_k = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^k\Big)$ $= \int_{-\infty}^{+\infty} \big(x - \mathbb{E}(X)\big)^k f(x)\,dx$

TABLE 1.1 – Définition des moments

Remarque 1.1 *Avec les notations de la Définition 1.8 on a en particulier :*
$\mu_1 = \mathbb{E}(X)$;
$m_1 = 0$. *En effet* $m_1 = \mathbb{E}\big(X - \mathbb{E}(X)\big) = \mathbb{E}(X) - \mathbb{E}(X) = 0$;
$m_2 = Var(X)$;
La formule $Var(X) = \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2$ *s'écrit, en termes de moments,* $m_2 = \mu_2 - \mu_1^2$.

1.5.2 Moments empiriques

Définition 1.9

Soient $X_1, X_2, \cdots, X_n$ un n-échantillon d'une loi de probabilité $\mathbb{P}_*$ et $k \in \mathbb{N}^*$.

- *Le moment empirique d'ordre k est la variable aléatoire*

$$\widehat{\mu}_k(n) = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

- *Le moment empirique centré d'ordre k est la variable aléatoire*

$$\widehat{m}_k(n) = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X}_n)^k.$$

Remarque 1.2 - *Dans les notations $\widehat{\mu}_k(n)$ et $\widehat{m}_k(n)$, l'indice k est l'ordre du moment et n est la taille de l'échantillon.*
- Si les moments μ_k et m_k existent alors, d'après la loi forte des grands nombres, on a presque sûrement

$$\lim_{n\to+\infty} \widehat{\mu}_k(n) = \mu_k \quad et \quad \lim_{n\to+\infty} \widehat{m}_k(n) = m_k.$$

- On a $\widehat{\mu}_1(n) = \overline{X}_n$ (= la moyenne empirique) et $\widehat{m}_2(n) = S_n^2$ (= la variance empirique).
- La formule $S_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2$ s'écrit, en termes de moments empiriques :

$$\widehat{m}_2(n) = \widehat{\mu}_2(n) - \big(\widehat{\mu}_1(n)\big)^2.$$

Comme auparavant pour la moyenne et la variance, il est important de ne pas confondre les moments (centrés ou non) de la loi $\mathbb{P}_*$, qui sont des *nombres*, avec les moments empiriques correspondants qui sont des *variables aléatoires*.

1.5.3 Principe de la méthode des moments

Le principe de la méthode des moments est le suivant : si l'on dispose d'une formule exprimant θ en fonction de certains des moments μ_k ou m_k, alors on obtient un estimateur $\widehat{\theta}_n$ de θ en remplaçant dans la même expression les moments μ_k ou m_k par les moments empiriques $\widehat{\mu}_k(n)$ et $\widehat{m}_k(n)$ correspondant.

Cette méthode se justifie par la loi des grands nombres. Plus précisément, on peut donner l'énoncé suivant qui en est une conséquence et que nous admettrons.

Proposition 1.6

Soient θ un paramètre d'une loi de probabilité $\mathbb{P}_*$ et $X_1, X_2, \cdots, X_n$ un n-échantillon de cette loi $\mathbb{P}_*$. Si θ s'exprime en fonction des k premiers moments (éventuellement centrés) sous la forme $\theta = G(\mu_1, m_1, \cdots, \mu_k, m_k)$ où G est une fonction continue en $(\mu_1, m_1, \cdots, \mu_k, m_k)$ alors l'estimateur de θ défini par $\widehat{\theta}_n = G(\widehat{\mu}_1(n), \widehat{m}_1(n), \cdots, \widehat{\mu}_k(n), \widehat{m}_k(n))$ converge presque sûrement.

1.6 Méthode du maximum de vraisemblance

Position du problème : Un caractère d'une population suit une loi de probabilité $\mathbb{P}_*$ appartenant à une famille $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ de lois de probabilité, où l'ensemble Θ (appelé espace des paramètres) est inclus dans $\mathbb{R}^d$ pour un entier $d \geq 1$ (on se limitera le plus souvent dans ce cours au cas où $d = 1$). La valeur θ_0 du paramètre θ telle que $\mathbb{P}_* = \mathbb{P}_{\theta_0}$ est inconnue. On considère un échantillon $X_1, X_2, \cdots, X_n$ de la loi $\mathbb{P}_*$.

1.6.1 Cas d'une loi de probabilité discrète

On suppose ici que les lois de probabilité $\mathbb{P}_\theta$ (où $\theta \in \Theta$) sont des lois de probabilité discrètes ayant toutes le même support S.

Définition 1.10

- *Une fonction de vraisemblance est une fonction*

$$\begin{array}{ccc} \Theta & \to & \mathbb{R} \\ \theta & \mapsto & \mathbb{P}_\theta(\{x_1\}) \times \mathbb{P}_\theta(\{x_2\}) \times \cdots \times \mathbb{P}_\theta(\{x_n\}) \end{array}$$

où $x_1, x_2, \cdots, x_n$ sont des éléments donnés de S. Une telle fonction se note $L(\theta, x_1, x_2, \cdots, x_n)$, où θ est la variable et où $x_1, x_2, \cdots, x_n$ jouent le rôle de paramètres.

- *On appelle estimateur du maximum de vraisemblance tout estimateur $\widehat{\theta}_n$ de θ vérifiant*

$$L(\widehat{\theta}_n, X_1, X_2, \cdots, X_n) = \max_{\theta \in \Theta} L(\theta, X_1, X_2, \cdots, X_n).$$

Un tel estimateur $\widehat{\theta}_n$ n'existe pas toujours ou peut ne pas être défini de façon unique.

1.6.2 Cas d'une loi de probablité à densité

On suppose ici que chaque loi de probabilité $\mathbb{P}_\theta$ (où $\theta \in \Theta$) est une loi de probabilité sur $\mathbb{R}$ admettant une densité f_θ. Par analogie avec le cas discret, on adopte les définitions suivantes.

Définition 1.11

- *Une fonction de vraisemblance est une fonction*

$$\begin{array}{ccc} \Theta & \to & \mathbb{R} \\ \theta & \mapsto & f_\theta(x_1) \times f_\theta(x_2) \times \cdots \times f_\theta(x_n) \end{array}$$

où $x_1, x_2, \cdots, x_n$ sont des nombres réels donnés. Une telle fonction se note $L(\theta, x_1, x_2, \cdots, x_n)$, où θ est la variable et où $x_1, x_2, \cdots, x_n$ jouent le rôle de paramètres.

- *On appelle estimateur du maximum de vraisemblance tout estimateur $\widehat{\theta}_n$ de θ vérifiant*

$$L(\widehat{\theta}_n, X_1, X_2, \cdots, X_n) = \max_{\theta \in \Theta} L(\theta, X_1, X_2, \cdots, X_n).$$

1.7 Exercices

Exercice 1. Dans un jeu de hasard, le gain a trois valeurs possibles : 0, 1 ou 2. Les probabilités de ces différentes valeurs sont données par la loi de probabilité $\mathbb{P}_*$ suivante :

$$\mathbb{P}_*(\{0\}) = 4\theta,\ \mathbb{P}_*(\{1\}) = \theta,\ \mathbb{P}_*(\{2\}) = 1 - 5\theta$$

où θ est un paramètre inconnu. Afin de construire des estimateurs de θ, on considère un échantillon $X_1, X_2, \cdots, X_n$ de la loi $\mathbb{P}_*$ obtenu en jouant n fois à ce jeu.
1) Justifier que $\theta \in [0; \frac{1}{5}]$.
2) a) Montrer que $\mathbb{E}(\mathbb{P}_*) = 2 - 9\theta$.
b) En déduire un estimateur $\widehat{\theta}_n$ de θ en utilisant la méthode des moments.

c) L'estimateur $\widehat{\theta}_n$ est-il sans biais ?

d) Montrer que $Var(\mathbb{P}_*) = 17\theta - 81\theta^2$ puis en déduire $Var(\widehat{\theta}_n)$.

3) On note B_n la variable aléatoire égale au nombre de variables aléatoires parmi $X_1, X_2, \cdots, X_n$ qui prennent la valeur 1.

a) Quelle est la loi de probabilité de B_n ?

b) En déduire que la variable aléatoire $\widetilde{\theta}_n = \frac{B_n}{n}$ est un estimateur sans biais de θ.

c) Donner $Var(\widetilde{\theta}_n)$.

4) Trouver (au moins) un autre estimateur $\overline{\theta}_n$ de θ.

5) a) Les estimateurs $\widehat{\theta}_n$, $\widetilde{\theta}_n$ et $\overline{\theta}_n$ sont-ils convergents en moyenne quadratique ?

b) Lequel de ces trois estimateurs de θ est le meilleur ?

Exercice 2. Pour θ réel strictement négatif, on note $\mathbb{P}_*$ la loi de probabilité de support $[\theta, 0]$ qui admet pour densité la fonction f définie par

$$f(x) = \begin{cases} -\dfrac{4x^3}{\theta^4} & \text{si } \theta \leq x \leq 0; \\ 0 & \text{si } x < \theta \text{ ou } x > 0. \end{cases}$$

Le but de l'exercice est de trouver et d'étudier plusieurs estimateurs de θ. On considére pour cela un n-échantillon $X_1, X_2, \cdots, X_n$ de la loi $\mathbb{P}_*$.

1) a) Justifier que f est bien une densité de probabilité.

b) Montrer que $\mathbb{E}(\mathbb{P}_*) = \dfrac{4\theta}{5}$ et que $Var(\mathbb{P}_*) = \dfrac{2\theta^2}{75}$.

c) Utiliser la méthode des moments pour donner deux estimateurs $\widehat{\theta}_n$ et $\widehat{\widehat{\theta}}_n$ de θ, le premier en utilisant la valeur de $\mathbb{E}(\mathbb{P}_*)$ et le second celle de $Var(\mathbb{P}_*)$.

d) Montrer que $\widehat{\theta}_n$ est sans biais.

2) On considère un autre estimateur [1] de θ défini par $\widetilde{\theta}_n = \min\{X_1, X_2, \cdots, X_n\}$. Nous admettrons que $\widetilde{\theta}_n$ admet pour densité la fonction $f_{\widetilde{\theta}_n}$ ci-dessous (ceci sera établi à la question 4)) :

$$f_{\widetilde{\theta}_n}(x) = \begin{cases} -\dfrac{4n}{\theta^{4n}} x^{4n-1} & \text{si } \theta \leq x \leq 0 \\ 0 & \text{si } x < \theta \text{ ou } x > 0. \end{cases},$$

a) Montrer que

$$\mathbb{E}(\widetilde{\theta}_n) = \frac{4n\theta}{4n+1} \quad \text{et} \quad Var(\widetilde{\theta}_n) = \frac{2n\theta^2}{(2n+1)(4n+1)^2}.$$

b) L'estimateur $\widetilde{\theta}_n$ est-il sans biais ? asymptotiquement sans biais ?

c) Trouver une constante k_n (qui dépend de n) telle que la variable aléatoire $\overline{\theta}_n = k_n \times \widetilde{\theta}_n$ soit un estimateur sans biais de θ.

d) Calculer $Var(\overline{\theta}_n)$.

e) Lequel des deux estimateurs sans biais $\widehat{\theta}_n$ et $\overline{\theta}_n$ est le meilleur ?

3) Les estimateurs $\widehat{\theta}_n$, $\widetilde{\theta}_n$ et $\overline{\theta}_n$ sont-ils convergents en moyenne quadratique ?

4) a) Pour $x \in \mathbb{R}$, écrire l'évènement $\{\widetilde{\theta}_n > x\}$ en fonction des évènements $\{X_1 > x\}, \{X_2 > x\}, \cdots, \{X_n > x\}$.

b) En utilisant a), exprimer la fonction de répartition $F_{\widetilde{\theta}_n}$ de $\widetilde{\theta}_n$ à l'aide de celle de la loi $\mathbb{P}_*$.

1. Cet estimateur est en fait celui donné par la méthode du maximum de vraisemblance. Vous pourrez le démontrer si le paragraphe correspondant du cours a été traité.

c) Déduire du b) que

$$F_{\widetilde{\theta}_n}(x) = \begin{cases} 0 & \text{si } x < \theta\,; \\ 1 - \left(\dfrac{x}{\theta}\right)^{4n} & \text{si } \theta \leq x \leq 0; \\ 1 & \text{si } x > 0. \end{cases}$$

d) Retrouver en utilisant c) la densité $f_{\widetilde{\theta}_n}$ de $\widetilde{\theta}_n$ qui a été donnée à la question 2).

Exercice 3. Un bus passe régulièrement, toutes les k minutes, à l'un de ses arrêts, k étant un nombre positif inconnu que l'on souhaite estimer. En se présentant n fois au hasard à l'arrêt, on obtient un échantillon $X_1, \cdots, X_n$ de la loi de probabilité $\mathbb{P}_*$ décrivant le temps d'attente (considéré comme une grandeur continue) avant le passage du bus.

1) a) Quelle est la loi $\mathbb{P}_*$? Calculer son espérance et sa variance.

b) Trouver à l'aide de la méthode des moments deux estimateurs $\widehat{k}_n$ et $\widehat{\widehat{k}}_n$ de k, le premier en utilisant la valeur de $\mathbb{E}(\mathbb{P}_*)$ et le second celle de $Var(\mathbb{P}_*)$.

c) Vérifier que $\widehat{k}_n$ est sans biais.

2) On considère un autre estimateur [2] défini par $\widetilde{k}_n = \max\{X_1, X_2, \cdots, X_n\}$.

a) Pour $x \in \mathbb{R}$, écrire l'évènement $\{\widetilde{k}_n \leq x\}$ en fonction des évènements $\{X_1 \leq x\}, \{X_2 \leq x\}, \cdots, \{X_n \leq x\}$.

b) En utilisant a), exprimer la fonction de répartition $F_{\widetilde{k}_n}$ de $\widetilde{k}_n$ à l'aide de celle de la loi $\mathbb{P}_*$.

c) Déduire du b) l'expression de $F_{\widetilde{k}_n}$ puis une densité $f_{\widetilde{k}_n}$ de $\widetilde{k}_n$.

d) Montrer que $\widetilde{k}_n$ est biaisé mais asymptotiquement sans biais puis définir simplement à partir de $\widetilde{k}_n$ un nouvel estimateur $\overline{k}_n$ qui soit sans biais.

e) Les estimateurs $\widehat{k}_n$, $\widetilde{k}_n$ et $\overline{k}_n$ sont-ils convergents en moyenne quadratique ?

3) Parmi les estimateurs sans biais $\widehat{k}_n$ et $\overline{k}_n$, lequel est le plus efficace ?

4) Dix personnes se sont présentées au hasard, et indépendamment les unes des autres, à l'arrêt du bus et on mesuré le temps d'attente avant son arrivée. Les résultats de ces mesures sont données par le tableau suivant :

10 mn 26 sec	3 mn 57 sec	11 mn 6 sec	12 mn 16 sec	8 mn 14 sec
13 mn 10 sec	5 mn	9 mn 14 sec	6 mn 36 sec	2 mn 44 sec

Calculer les estimations de k fournies par les différents estimateurs rencontrés précédemment.

Exercice 4. On considère un échantillon $X_1, X_2, \cdots, X_n$ de la loi uniforme $\mathcal{U}([a, b])$, les deux paramètres a et b étant supposés inconnus.

1) Calculer l'espérance et la variance de la loi $\mathcal{U}([a, b])$ puis donner une expression de a et de b en fonction de cette espérance et de cette variance.

2) En déduire un estimateur $\widehat{a}_n$ de a et un estimateur $\widehat{b}_n$ de b par la méthode des moments.

Exercice 5. La vitesse des voitures a été mesurée en un point de contrôle sur un échantillon de 40 véhicules. Les résultats de ces mesures (en km/h) sont donnés par le tableau suivant.

124	105	94	93	106	101	90	91	118	115
100	96	112	109	89	102	95	89	88	108
101	96	105	107	100	101	91	114	98	101
103	92	103	94	95	115	97	105	99	103

2. C'est en fait l'estimateur donné par la méthode du maximum de vraisemblance

Donner une estimation de la vitesse moyenne des véhicules en ce point ainsi qu'une estimation non biaisée de la variance de la vitesse en ce même point.

Exercice 6. Partie I. Une régie de transports municipaux souhaite estimer la proportion $p \in]0,1[$ des habitants de la ville ayant utilisé les transports en commun au cours de la dernière semaine. Pour cela, elle commande un sondage où n personnes répondent (par oui ou non) à la question « Avez-vous emprunté au moins une fois les transports en commun de la ville au cours des sept derniers jours ? » On définit alors une suite de variables aléatoires $X_1, X_2, \cdots, X_n$ en posant

$$X_i = \begin{cases} 1 & \text{si la } i^{eme} \text{ personne interrogée répond « oui » ;} \\ 0 & \text{si la } i^{eme} \text{ personne interrogée répond « non » .} \end{cases}$$

Nous admettrons que, si les personnes interrogées sont représentatives de la population, les v.a.r. $X_1, X_2, \cdots, X_n$ peuvent être considérées comme indépendantes et suivant toute la même loi de de probabilité $\mathbb{P}_*$.

1) Quelle est cette loi de probabilité $\mathbb{P}_*$? Autrement dit, de quelle loi $\mathbb{P}_*$ la suite $X_1, X_2, \cdots, X_n$ est-elle un échantillon ?

2) $\overline{X}_n$ est-elle un estimateur sans biais de p ?

3) Donner $Var(\overline{X}_n)$ et en déduire que $\overline{X}_n$ est un estimateur de p qui converge en moyenne quadratique.

Partie II. La même régie de transports souhaite évaluer, grâce à un autre sondage, la proportion $p' \in]0,1[$ des voyageurs qui fraudent. Un problème dans une telle situation est que des individus questionnés pour le sondage pourraient ne pas répondre sincèrement à la personne qui les interroge. Pour contourner cette difficulté, on choisit n personnes représentatives des usagers des transports avec lesquelles on convient du procédé suivant. Chacune d'elles lance une pièce de monnaie. Le résultat du lancer n'est pas connu par le sondeur. Ensuite

- si le résultat est Pile alors la personne interrogée répond sincèrement à la question « Vous arrive-t-il de voyager sans titre de transport en règle ? ».

- si le résultat est Face alors la personne relance la pièce puis répond à la question « avez-vous obtenu Face au deuxième lancer ? ».

De cette façon, l'enquêteur ne peut pas savoir si les réponses « oui » ou « non » qu'il recueille concernent le comportement dans les transports des individus interrogés. On définit alors une suite de variables aléatoires $X_1, X_2, \cdots, X_n$ en posant

$$X_i = \begin{cases} 1 & \text{si la } i^{eme} \text{ personne interrogée répond « oui » ;} \\ 0 & \text{si la } i^{eme} \text{ personne interrogée répond « non » ;} \end{cases}$$

Ces variables aléatoires $X_1, X_2, \cdots, X_n$ peuvent être considérées comme indépendantes.

1) Quelle est la valeur prise par X_i dans les cas suivants ?

1er cas : la i^{eme} personne n'est pas toujours en règle et elle obtient Pile.

2eme cas : la i^{eme} personne est toujours en règle et elle obtient Face puis Face.

2) Justifier que $X_1, X_2, \cdots, X_n$ est un échantillon de la loi de Bernoulli de paramètre $\dfrac{p'}{2} + \dfrac{1}{4}$. Cette loi sera notée $\mathbb{P}'_*$ par la suite.

3) a) Exprimer p' en fonction de $\mathbb{E}(\mathbb{P}'_*)$ et en déduire un estimateur $\widehat{p'}_n$ de p' par la méthode des moments.

b) L'estimateur $\widehat{p'}_n$ est-il sans biais ?

c) Calculer $Var(\widehat{p'}_n)$ en en déduire que $\widehat{p'}_n$ est convergent en moyenne quadratique.

Exercice 7. On considère un échantillon $X_1, X_2, \cdots, X_n$ de la loi géométrique $\mathcal{G}(p)$ de paramètre

$p \in]0,1]$. On rappelle que

$$\mathbb{E}(\mathcal{G}(p)) = \frac{1}{p}, \quad Var(\mathcal{G}(p)) = \frac{1-p}{p^2}.$$

1) Trouver, avec le principe des moments et la formule donnant $\mathbb{E}(\mathcal{G}(p))$, un premier estimateur $\widehat{p}_n$ de p.
2) a) Trouver, à l'aide de la formule donnant $Var(\mathcal{G}(p))$, trois nombres a, b, c tels que $ap^2+bp+c = 0$.

b) Calculer les racines du polynôme $P(x) = ax^2 + bx + c$. En déduire une expression de p en fonction de a, b et c.

c) Utiliser b) et la méthode des moments pour donner un autre estimateur $\widetilde{p}_n$ de p.

Exercice 8. Calculer l'estimateur du maximum de vraisemblance pour
1) le paramètre p d'une loi binomiale $\mathcal{B}(k; p)$, le paramètre k étant connu.
2) le paramètre p d'une loi géométrique $\mathcal{G}(p)$.
3) le paramètre λ d'une loi exponentielle $\mathcal{E}(\lambda)$.
4) a) la moyenne μ d'une loi normale $\mathcal{N}(\mu; \sigma^2)$, la variance σ^2 étant connue.

b) la variance σ^2 d'une loi normale $\mathcal{N}(\mu; \sigma^2)$, la moyenne μ étant connue.
5) la borne inférieure a de l'intervalle $[a, b]$ pour la loi uniforme $\mathcal{U}([a; b])$, la borne supérieure b étant connue.

Chapitre 2

Estimation par intervalles de confiances

Sommaire

2.1 Généralités

Position du problème : On a à nouveau une loi de probabilité $\mathbb{P}_*$ dont un paramètre $\theta \in \mathbb{R}$ est inconnu. On utilise un n-échantillon $X_1, X_2, \cdots, X_n$ de la loi $\mathbb{P}_*$ pour construire cette fois une « fourchette » (c'est à dire un intervalle) contenant θ avec une forte probabilité.

Définition 2.1

Soient $\theta \in \mathbb{R}$ un paramètre d'une loi de probabilité $\mathbb{P}_*$ et $X_1, X_2, \cdots, X_n$ un échantillon de la loi $\mathbb{P}_*$. Un intervalle de confiance de niveau $c \in]0;1[$ pour le paramètre θ est un intervalle I_n défini en fonction de $X_1, X_2, \cdots, X_n$ et tel que $\mathbb{P}(\theta \in I_n) \geq c$.

2.2 Intervalles de confiance pour la moyenne

2.2.1 Intervalles de confiances et inégalité de Bienaymé-Tchebychev

Proposition 2.1 (*Inégalité de Bienaymé-Tchebychev*)

Soit X une v.a.r admettant une espérance $\mathbb{E}(X)$ et une variance $Var(X)$. Alors on a

$$\forall a > 0 \qquad \mathbb{P}\big(|X - \mathbb{E}(X)| \geq a\big) \leq \frac{Var(X)}{a^2}.$$

Preuve : Soit un réel $a > 0$. En appliquant l'inégalité de Markov vue au chapitre précédent (Proposition 1.1) à la variable aléatoire $(X - \mathbb{E}(X))^2$ et en prenant $\alpha = a^2$ on obtient

$$\mathbb{P}\Big(\big(X - \mathbb{E}(X)\big)^2 \geq a^2\Big) \leq \frac{\mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big)}{a^2}.$$

Or $\mathbb{E}\Big(\big(X-\mathbb{E}(X)\big)^2\Big) = Var(X)$; de plus les évènements $\{\big(X-\mathbb{E}(X)\big)^2 \geq a^2\}$ et $\{|X-\mathbb{E}(X)| \geq a\}$ sont égaux donc ont même probabilité, ce qui donne l'inégalité voulue. ■

Grâce à l'inégalité de Bienaymé-Tchebychev, on obtient le résultat suivant qui s'applique sans aucune hypothèse sur la loi $\mathbb{P}_*$ autre que l'existence de la variance. Cependant les intervalles de confiance qu'il fournit sont grands et donc souvent peu intéressants.

Proposition 2.2

Soit $X_1, X_2, \cdots, X_n$ un n-échantillon d'une loi de probabilité $\mathbb{P}_*$ d'espérance μ et de variance σ^2. Alors, pour tout réel $a > 0$, on a

$$\mathbb{P}\big(\overline{X}_n - a \leq \mu \leq \overline{X}_n + a\big) \geq 1 - \frac{\sigma^2}{na^2}.$$

Autrement dit, l'intervalle $I_n = [\overline{X}_n - a, \overline{X}_n + a]$ est un intervalle de confiance de niveau $1 - \dfrac{\sigma^2}{na^2}$ pour μ.

Preuve. En appliquant l'inégalité de Bienaymé-Tchebychev à la variable aléatoire $\overline{X}_n$ on obtient que $\mathbb{P}\big(|\overline{X}_n - \mathbb{E}(\overline{X}_n)| \geq a\big) \leq \dfrac{Var(\overline{X}_n)}{a^2}$ pour tout réel $a > 0$ et par suite

$$\mathbb{P}\big(|\overline{X}_n - \mathbb{E}(\overline{X}_n)| \leq a\big) \geq \mathbb{P}\big(|\overline{X}_n - \mathbb{E}(\overline{X}_n)| < a\big) = 1 - \mathbb{P}\big(|\overline{X}_n - \mathbb{E}(\overline{X}_n)| \geq a\big) \geq 1 - \frac{Var(\overline{X}_n)}{a^2}.$$

De plus, on sait que $\mathbb{E}(\overline{X}_n) = \mu$ et que $Var(\overline{X}_n) = \dfrac{\sigma^2}{n}$ ce qui donne $\mathbb{P}\big(|\overline{X}_n - \mu| \leq a\big) \geq 1 - \dfrac{\sigma^2}{na^2}$.
On conclut en observant que

$$|\overline{X}_n - \mu| \leq a \iff -a \leq \overline{X}_n - \mu \leq a \iff \overline{X}_n - a \leq \mu \leq \overline{X}_n + a.$$

■

2.2.2 Échantillons gaussiens de variance connue

Proposition 2.3 (*Z-intervalles*)

Soit $X_1, X_2, \cdots, X_n$ un n-échantillon gaussien, ce qui signifie que la loi de probabilité commune des X_i est une loi normale $\mathcal{N}(\mu; \sigma^2)$. Alors :

- *La variable aléatoire*
$$Z_n = \frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu)$$
suit la loi $\mathcal{N}(0;1)$.
- *Pour tout réel $c \in]0;1[$ l'intervalle*
$$I_n = \Big[\overline{X}_n - \frac{\sigma}{\sqrt{n}} \times q_{\frac{1+c}{2}} \,;\, \overline{X}_n + \frac{\sigma}{\sqrt{n}} \times q_{\frac{1+c}{2}}\Big],$$
où $q_{\frac{1+c}{2}}$ désigne le quantile d'ordre $\frac{1+c}{2}$ de la loi $\mathcal{N}(0;1)$, est un intervalle de confiance au niveau c pour l'espérance μ. Plus précisément :
$$\mathbb{P}\Big(\overline{X}_n - \frac{\sigma}{\sqrt{n}} \times q_{\frac{1+c}{2}} \leq \mu \leq \overline{X}_n + \frac{\sigma}{\sqrt{n}} \times q_{\frac{1+c}{2}}\Big) = c.$$

Preuve partielle. Le premier point résulte du fait que la somme de v.a.r. gaussiennes indépendantes est encore une v.a.r. gaussienne, ce que nous admettrons. Pour le deuxième point, notons F la fonction de répartition de la loi $\mathcal{N}(0;1)$. Pour tout $x \in \mathbb{R}$ on a $F(-x) = 1 - F(x)$ car la loi $\mathcal{N}(0;1)$ admet pour densité la fonction $f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ qui est paire (voir Figure 2.1 pour le cas où $x = 0,95$). On a donc

$$\mathbb{P}\big(-q_{\frac{1+c}{2}} \leq Z_n \leq q_{\frac{1+c}{2}}\big) = F\big(q_{\frac{1+c}{2}}\big) - F\big(-q_{\frac{1+c}{2}}\big) = \underbrace{F\big(q_{\frac{1+c}{2}}\big)}_{=\frac{1+c}{2}} - 1 + \underbrace{\Phi\big(q_{\frac{1+c}{2}}\big)}_{=\frac{1+c}{2}} = c.$$

On conclut en observant que

$$-q_{\frac{1+c}{2}} \leq Z_n \leq q_{\frac{1+c}{2}} \quad \iff \quad \overline{X}_n - \frac{\sigma}{\sqrt{n}} q_{\frac{1+c}{2}} \leq \mu \leq \overline{X}_n + \frac{\sigma}{\sqrt{n}} q_{\frac{1+c}{2}}.$$

■

Remarque 2.1

- *On rappelle que le quantile $q_{\frac{1+c}{2}}$ dans la Proposition 2.3 est le nombre réel vérifiant $F(q_{\frac{1+c}{2}}) = \frac{1+c}{2}$ où F est la fonction de répartition de la loi $\mathcal{N}(0;1)$. La figure 2.1 illustre la situation pour le cas où $c = 0,9$. Ce quantile s'obtient par lecture de table ou bien à l'aide de divers logiciels ou d'une calculatrice scientifique.*
- *L'intervalle de confiance de la Proposition 2.3 se calcule directement avec de nombreux logiciels de statistiques ainsi qu'avec la plupart des calculatrices scientifiques.*

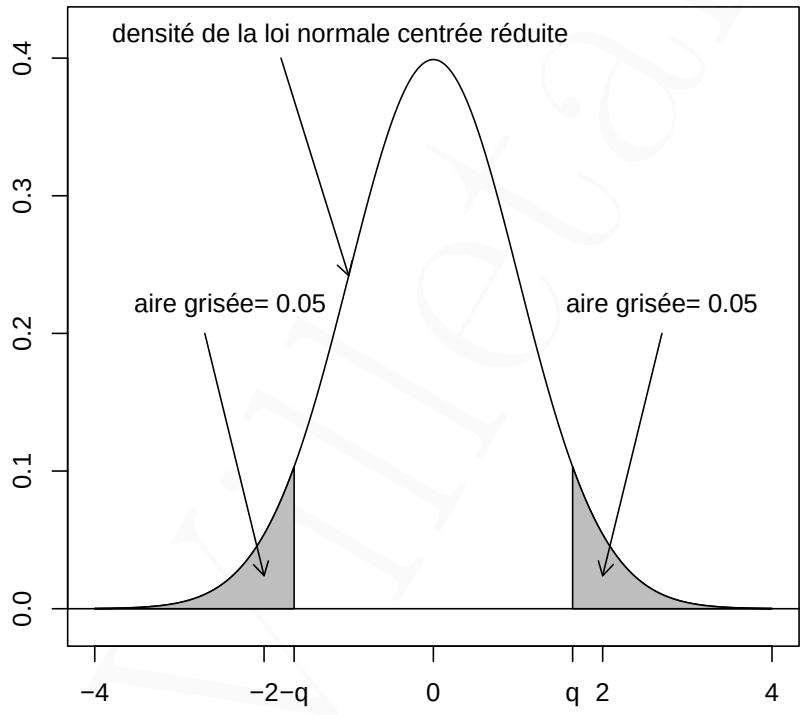


FIGURE 2.1 – Quantile q d'ordre 0,95 pour la loi $\mathcal{N}(0;1)$

2.2.3 Échantillons gaussiens de variance inconnue

La situation considérée ici est plus réaliste que celle du paragraphe 2.2.2 car il rare que l'on connaisse la variance d'une certaine quantité sur une population entière sans que l'on connaisse la moyenne de la même quantité sur la même population. Pour traiter cette situation, on a besoin d'une famille de lois de probabilités appelées lois de Student.

Définition 2.2 (*Lois de Student*)

La loi de Student à n degrés de liberté (ddl), où $n \in \mathbb{N}^*$, est la loi de probabilité de support $\mathbb{R}$ admettant pour densité la fonction f donnée par

$$\forall x \in \mathbb{R} \quad f(x) = \frac{1}{\sqrt{n\pi}} \times \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \times \frac{1}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}$$

où, pour tout réel $t > 0$, on définit $\Gamma(t) = \int_0^{+\infty} x^{t-1}e^{-x}dx$ (fonction Gamma d'Euler).

La figure ci-dessous donne l'allure de cette densité pour différentes valeurs du ddl n.

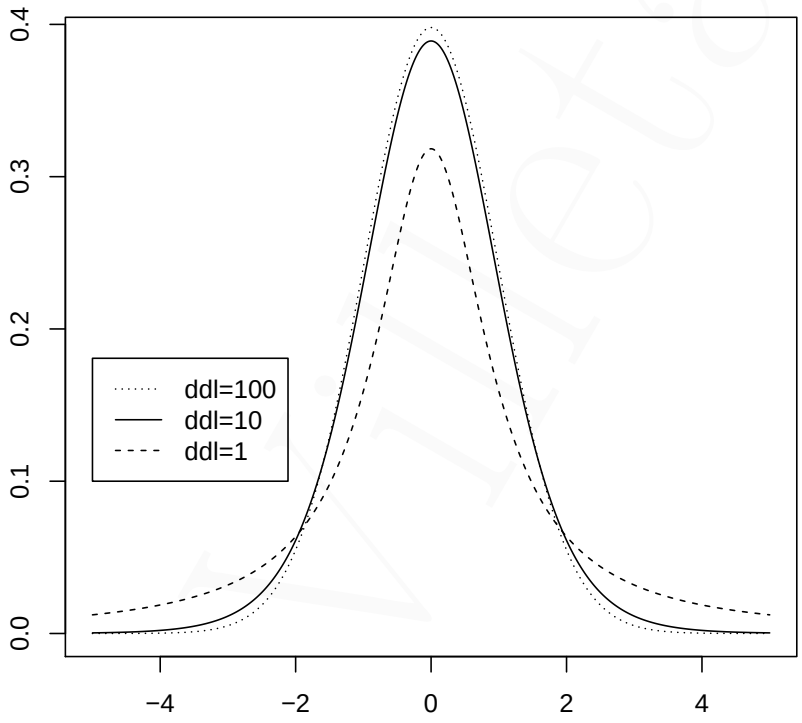


FIGURE 2.2 – Densités des lois de Student pour quelques valeurs du ddl

Remarque 2.2

- *En pratique, les calculs concernant les lois de Student se font avec des tables ou des logiciels.*
- *On peut montrer que la loi de Student n'admet pas d'espérance quand $n = 1$ et pas de variance quand $n \in \{1, 2\}$. Pour $n \geq 2$ son espérance est nulle et pour $n \geq 3$ sa variance vaut $n/(n-2)$.*
- *Pour n grand, la loi de Student à n ddl est très proche de la loi normale $\mathcal{N}(0; 1)$.*

Proposition 2.4 (*T-intervalles*)

Soit $X_1, X_2, \cdots, X_n$ un n-échantillon gaussien d'espérance μ et de variance σ^2.

- *La variable aléatoire*

$$T_n = \frac{\sqrt{n}}{\sqrt{S_n^{2c}}}(\overline{X_n} - \mu) = \frac{\sqrt{n-1}}{\sqrt{S_n^2}}(\overline{X_n} - \mu)$$

  suit une loi de Student à $n-1$ degrés de liberté.

- *Soient un réel $c \in]0,1[$. L'intervalle*

$$I_n = \Big[\overline{X}_n - \frac{\sqrt{S_n^{2c}}}{\sqrt{n}} \times q_{\frac{1+c}{2}} \,;\, \overline{X}_n + \frac{\sqrt{S_n^{2c}}}{\sqrt{n}} \times q_{\frac{1+c}{2}}\Big],$$

  où $q_{\frac{1+c}{2}}$ désigne quantile d'ordre $\frac{1+c}{2}$ de la loi de Student à $n-1$ ddl, est un intervalle de confiance pour la moyenne μ au niveau c. Plus précisément :

$$\mathbb{P}\Big(\overline{X}_n - \frac{\sqrt{S_n^{2c}}}{\sqrt{n}} \times q_{\frac{1+c}{2}} \leq \mu \leq \overline{X}_n + \frac{\sqrt{S_n^{2c}}}{\sqrt{n}} \times q_{\frac{1+c}{2}}\Big) = c.$$

Preuve partielle. Nous admettrons le premier point. Le deuxième point est très similaire à ce qui a été fait pour la Proposition 2.3. En effet la densité f donnée dans la Définition 2.2 pour la loi de Student est aussi une fonction paire donc, en notant cette fois Φ la fonction de répartition de la loi de Student à n ddl, on a à nouveau $\Phi(-x) = 1 - \Phi(x)$ pour tout $x \in \mathbb{R}$. On obtient alors

$$\mathbb{P}\big(-q_{\frac{1+c}{2}} \leq T_n \leq q_{\frac{1+c}{2}}\big) = \Phi\big(q_{\frac{1+c}{2}}\big) - \Phi\big(-q_{\frac{1+c}{2}}\big) = \underbrace{\Phi\big(q_{\frac{1+c}{2}}\big)}_{=\frac{1+c}{2}} - 1 + \underbrace{\Phi\big(q_{\frac{1+c}{2}}\big)}_{=\frac{1+c}{2}} = c.$$

On conclut en remarquant que, pour tout réel $q \geq 0$, on a

$$-q \leq T_n \leq q \quad \Longleftrightarrow \quad \overline{X}_n - \frac{\sqrt{S_n^{2c}}}{\sqrt{n}} q \leq \mu \leq \overline{X}_n + \frac{\sqrt{S_n^{2c}}}{\sqrt{n}},$$

■

Remarque 2.3 *L'intervalle de confiance de la Proposition 2.4 peut s'obtenir directement avec des logiciels statistiques ou des calculatrices scientifiques.*

2.2.4 Grands échantillons

Il arrive souvent que la loi de probabilité $\mathbb{P}_*$ dont on veut estimer la moyenne μ ne soit pas une loi de Gauss, et même que l'on ne sache pas de quel type de loi il s'agit. Dans ce cas, en admettant que $\mathbb{P}_*$ admette une variance σ^2, on peut tout de même obtenir des intervalles de confiance *asymptotiques* pour μ, c'est à dire valables lorsque l'échantillon utilisé $X_1, X_2, \cdots, X_n$ est de grande taille. Ceci est dû principalement au théorème de la limite centrale (déjà évoqué dans le cours de probabilité) qui affirme que, pour n grand, la loi de probabilité de la variable aléatoire $\frac{\sqrt{n}}{\sigma}\big(\overline{X}_n - \mu\big)$ est proche de la loi normale $\mathcal{N}(0;1)$. En s'appuyant sur ce résultat important, on peut énoncer :

Proposition 2.5

Soit $X_1, X_2, \cdots, X_n$ un n-échantillon d'une loi de probabilité $\mathbb{P}_*$ de moyenne μ et de variance σ^2. On suppose de plus que n est grand (en pratique on demande souvent $n \geq 30$).

- *Si la variance σ^2 est connue, alors l'intervalle*

$$I_n = \left[\overline{X}_n - \frac{\sigma}{\sqrt{n}} \times q_{\frac{1+c}{2}} \,;\, \overline{X}_n + \frac{\sigma}{\sqrt{n}} \times q_{\frac{1+c}{2}}\right],$$

où $q_{\frac{1+c}{2}}$ désigne le quantile d'ordre $\frac{1+c}{2}$ de la loi $\mathcal{N}(0;1)$, est approximativement un intervalle de confiance au niveau c pour l'espérance μ. Plus précisément :

$$\mathbb{P}\Big(\overline{X}_n - \frac{\sigma}{\sqrt{n}} \times q_{\frac{1+c}{2}} \leq \mu \leq \overline{X}_n + \frac{\sigma}{\sqrt{n}} \times q_{\frac{1+c}{2}}\Big) \simeq c.$$

- *Si la variance σ^2 n'est pas connue, le point précédant reste valable en remplaçant σ par son estimateur $\sqrt{S_n^2}$.*

2.3 Intervalle de confiance pour une proportion

Il s'agit ici de la situation où l'on veut estimer le paramètre p d'une loi de Bernoulli. Les variables aléatoires $X_1, X_2, \cdots, X_n$ de l'échantillon peuvent donc prendre seulement les valeurs 0 ou 1, avec $\mathbb{P}_{X_i}(0) = 1 - p$ et $\mathbb{P}_{X_i}(1) = p$.

Propriété 2.1

Lorsque les X_i prennent seulement les valeurs 0 ou 1, on a $S_n^2 = \overline{X}_n \times (1 - \overline{X}_n)$.

Preuve. Chaque X_i prend seulement les valeurs 0 ou 1 donc $X_i = X_i^2$ puis

$$(X_i - \overline{X}_n)^2 = X_i^2 - 2X_i \times \overline{X}_n + \overline{X}_n^2 = X_i - 2X_i \times \overline{X}_n + \overline{X}_n^2.$$

On obtient alors

$$\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \sum_{i=1}^{n}(X_i - 2X_i \times \overline{X}_n + \overline{X}_n^2) = \sum_{i=1}^{n} X_i - 2\overline{X}_n \times \sum_{i=1}^{n} X_i + n\overline{X}_n^2.$$

On en déduit, en multipliant par $\frac{1}{n}$, que

$$S_n^2 = \overline{X}_n - 2\overline{X}_n^2 + \overline{X}_n^2 = \overline{X}_n \times (1 - \overline{X}_n).$$

■

Proposition 2.6 (*Intervalle de confiance asympotique pour une proportion*)

Soient un réel $c \in]0;1[$ et un n-échantillon $X_1, X_2, \cdots, X_n$ de la loi de Bernoulli de paramètre $p \in [0;1]$. On suppose de plus que n est grand (en pratique on demande souvent $n \geq 30$). Alors l'intervalle

$$I_n = \Big[\overline{X}_n - \frac{\sqrt{\overline{X}_n \times (1-\overline{X}_n)}}{\sqrt{n}} \times q_{\frac{1+c}{2}} ; \overline{X}_n + \frac{\sqrt{\overline{X}_n \times (1-\overline{X}_n)}}{\sqrt{n}} \times q_{\frac{1+c}{2}}\Big],$$

où $q_{\frac{1+c}{2}}$ désigne le quantile d'ordre $\frac{1+c}{2}$ de la loi $\mathcal{N}(0;1)$, est approximativement un intervalle de confiance pour p au niveau c. Plus précisément :

$$\mathbb{P}\Big(\overline{X}_n - \frac{\sqrt{\overline{X}_n \times (1-\overline{X}_n)}}{\sqrt{n}} \times q_{\frac{1+c}{2}} \leq p \leq \overline{X}_n + \frac{\sqrt{\overline{X}_n \times (1-\overline{X}_n)}}{\sqrt{n}} \times q_{\frac{1+c}{2}}\Big) \simeq c.$$

Preuve. Comme la moyenne de la loi de Bernoulli de paramètre p est justement égale à p, il suffit d'appliquer le deuxième point de la Proposition 2.5 en tenant compte de la forme particulière de S_n^2 donnée par la Propriété 2.1. ■.

Remarque 2.4 *Il existe des méthodes permettant de construire des intervalles de confiance pour une proportion indépendamment du fait que l'échantillon $X_1, X_2, \cdots, X_n$ soit de grande taille ou non. L'exercice 3 montre une construction basée sur l'inégalité de Bienaymé-Tchebychev.*

Remarque 2.5 *Les intervalles de confiances donnés par les Propositions 2.2, 2.3, 2.4, 2.5 et 2.6 sont tous centrés sur la moyenne d'échantillon $\overline{X}_n$. Il est aussi possible de construire des intervalles de confiance n'ayant pas cette propriété.*

2.4 Intervalles de confiance pour la variance d'un échantillon gaussien

Le but de ce paragraphe est d'obtenir des intervalles de confiance pour la variance σ^2 d'une loi normale $\mathcal{N}(\mu;\sigma^2)$. On a besoin pour cela d'une famille de lois de probabilités appelées loi du khi-deux.

Définition 2.3 (*Lois du khi-deux*)

La loi du khi-deux à n degré de liberté (ddl), où $n \in \mathbb{N}^*$, est la loi de probabilité de support $\mathbb{R}^+$ admettant pour densité la fonction f donnée par

$$\forall x \in \mathbb{R} \quad f(x) = \begin{cases} \dfrac{1}{2^{n/2}\,\Gamma(n/2)} \times x^{\frac{n}{2}-1} \times e^{-x/2} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

où la fonction $\Gamma(t)$ est la même que dans la Définition 2.2.

La figure suivante montre l'aspect de ces densités pour quelques valeurs du ddl n.

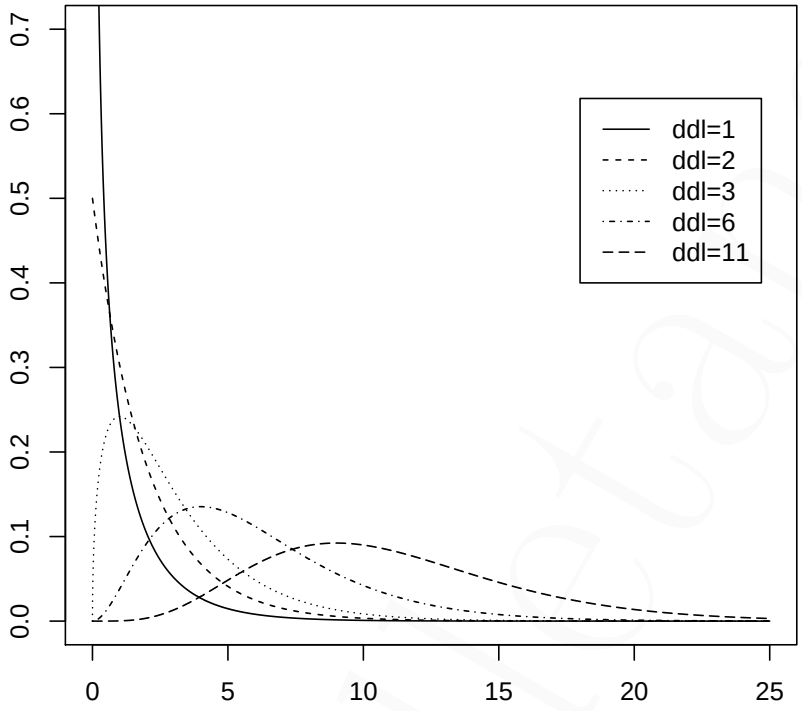


FIGURE 2.3 – Densités des lois du khi-deux pour quelques valeurs du ddl

Remarque 2.6
- *En pratique, les calculs concernant les lois du khi-deux se font avec des tables ou des logiciels.*
- *On peut démontrer que la loi χ^2_n a pour espérance n et pour variance $2n$.*

La proposition suivante, que nous admettrons, fait le lien entre les lois de Gauss, de Student et du khi-deux. Elle peut aussi servir de définition alternative aux lois de Student et du khi-deux.

Proposition 2.7

- *Si $X_1, X_2, \cdots, X_n$ est un échantillon de la loi $\mathcal{N}(0;1)$ alors la v.a.r. $\sum_{i=1}^n X_i^2$ suit la loi χ^2_n.*
- *Si X, Y sont deux v.a.r. indépendantes suivant respectivement la loi normale $\mathcal{N}(0;1)$ et la loi du χ^2_n alors la variable aléatoire $\dfrac{\sqrt{n}X}{\sqrt{Y}}$ suit la loi de Student à n ddl.*

2.4.1 Cas où la moyenne est connue

Proposition 2.8

Soient un réel $c \in]0,1[$ et $X_1, X_2, \cdots, X_n$ un échantillon de la loi $\mathcal{N}(\mu;\sigma^2)$.

- *La variable aléatoire*
$$\overline{K}_n = \sum_{i=1}^{n} \Big(\frac{X_i - \mu}{\sigma}\Big)^2 = \frac{n}{\sigma^2} \times \overline{S_n^2}$$
suit la loi du khi-deux à n ddl.
- *Si a, b dont deux réels vérifiant $0 < a \leq b$ et $\mathbb{P}(a \leq \overline{K}_n \leq b) = c$ alors l'intervalle*
$$I_n = \Big[\frac{n \times \overline{S_n^2}}{b}\,;\, \frac{n \times \overline{S_n^2}}{a}\Big]$$
est un intervalle de confiance au niveau c pour la variance σ^2. Plus précisément :
$$\mathbb{P}\Big(\frac{n \times \overline{S_n^2}}{b} \leq \sigma^2 \leq \frac{n \times \overline{S_n^2}}{a}\Big) = c.$$

Preuve. Les v.a.r. $\frac{X_i - \mu}{\sigma}$ sont aussi indépendantes et elles suivent toutes la loi normale $\mathcal{N}(0;1)$ donc la première partie de la Proposition 2.7 dit que $\overline{K}_n$ suit la loi χ^2_n. Pour le deuxième point, il suffit de remarquer que

$$a \leq \frac{n}{\sigma^2} \times \overline{S_n^2} \leq b \iff \frac{n \times \overline{S_n^2}}{b} \leq \sigma^2 \leq \frac{n \times \overline{S_n^2}}{a}.$$

■

2.4.2 Cas où la moyenne est inconnue

Proposition 2.9

Soient un réel $c \in]0;1[$ et un échantillon $X_1, X_2, \cdots, X_n$ de la loi $\mathcal{N}(\mu;\sigma^2)$.

- *La variable aléatoire*
$$K_n = \frac{(n-1) \times S_n^{2c}}{\sigma^2} = \frac{n \times S_n^2}{\sigma^2}$$
suit la loi du khi-deux à $n-1$ ddl.
- *Si a, b sont deux réels vérifiant $0 < a < b$ et $\mathbb{P}(a \leq K_n \leq b) = c$ alors l'intervalle*
$$I_n = \Big[\frac{(n-1) \times S_n^{2c}}{b}\,;\, \frac{(n-1) \times S_n^{2n}}{a}\Big] = \Big[\frac{n \times S_n^2}{b}\,;\, \frac{n \times S_n^2}{a}\Big]$$
est un intervalle de confiance au niveau c pour la variance σ^2. Plus précisément :
$$\mathbb{P}\Big(\frac{(n-1) \times S_n^{2c}}{b} \leq \sigma^2 \leq \frac{(n-1) \times S_n^{2c}}{a}\Big) = c.$$

Preuve partielle. On admettra le premier point. Pour le deuxième point, il suffit d'observer que

$$a \leq \frac{(n-1) \times S_n^{2c}}{\sigma^2} \leq b \iff \frac{(n-1) \times S_n^{2c}}{b} \leq \sigma^2 \leq \frac{(n-1) \times S_n^{2c}}{a}.$$

■

Remarque 2.7

- *Les nombres a et b dans les Propositions 2.8 et 2.9 ne sont pas uniques. On les choisit souvent de telle façon que $\mathbb{P}(\overline{K}_n \leq a) = \mathbb{P}(\overline{K}_n \geq b) = (1-c)/2$ et de même avec K_n au lieu $\overline{K}_n$. Autrement dit, on prend pour a le quantile d'ordre $(1-c)/2$ et pour b le quantile d'ordre $(1+c)/2$ de la loi du χ^2_n ou de la loi du χ^2_{n-1}. Voir la figure 2.4 pour le cas où $c = 0,9$ et pour la loi du χ^2_3. Ces quantiles s'obtiennent par lecture de table ou avec un logiciel.*
- *Pour avoir un intervalle de confiance pour σ^2 de la forme $I_n = [0; M]$ ou $I_n = [N; +\infty[$ avec $N > 0$, il suffit de reprendre ce qui précède avec, respectivement, $b = +\infty$ et $a = 0$.*

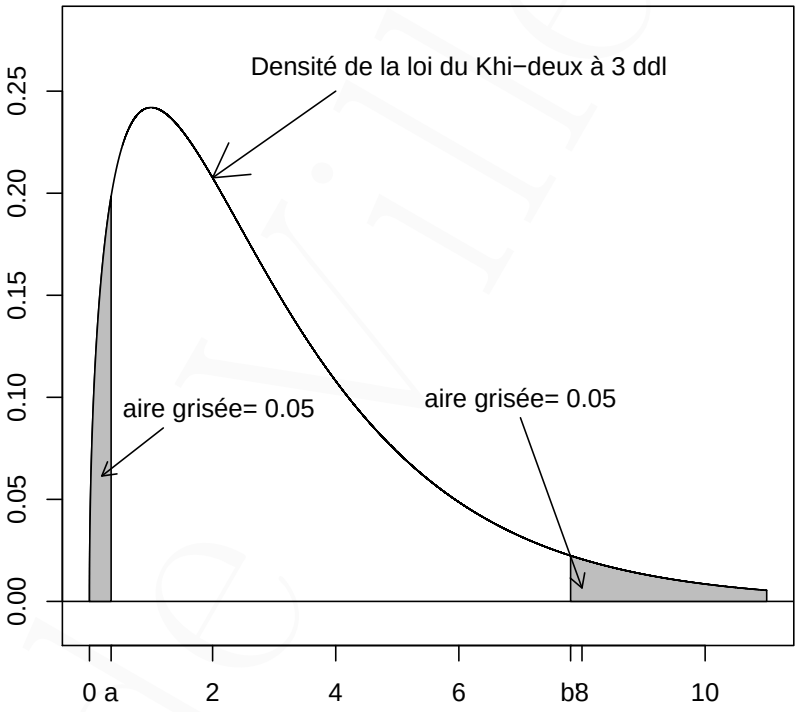


FIGURE 2.4 – Quantiles d'ordre 0,05 et 0,95 pour la loi du khi-deux à 3 ddl

2.5 Exercices

Exercice 1. On reprend les données de l'Exercice 5 du chapitre 1. La vitesse des voitures a été mesurée en un point de contrôle sur un échantillon de 40 véhicules. Les résultats de ces mesures (en km/h) sont donnés par le tableau suivant.

124	105	94	93	106	101	90	91	118	115
100	96	112	109	89	102	95	89	88	108
101	96	105	107	100	101	91	114	98	101
103	92	103	94	95	115	97	105	99	103

Construire un intervalle de confiance de niveau 95% pour la vitesse moyenne de passage des voitures en cet endroit.

Exercice 2. Un agriculteur exploite un verger de pommiers, tous de la même variété. La production d'un arbre est mesurée en nombre de kilogrammes de fruits produits pendant une année. La production moyenne des pommiers de ce verger est notée μ et la variance de leur production σ^2. Afin d'estimer μ, l'agriculteur décide de peser la quantité de fruits récoltée en un an sur douze arbres choisis au hasard dans son exploitation. Il obtient ainsi un échantillon $X_1, X_2, \cdots, X_{12}$ du rendement des pommiers du verger. On suppose de plus que cet échantillon est gaussien, autrement dit que les X_i suivent la loi $\mathcal{N}(\mu; \sigma^2)$. Les résultats des mesures sont donnés par le tableau suivant :

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
32.9	39.8	38.2	36.6	30.4	27.4	19.5	39.5	33.8	36.8	37.2	45.5

☞ Dans cet exercice, on donnera les résultats numériques en arrondissant convenablement à un chiffre après la virgule.

1) Dans cette question, on suppose la variance σ^2 connue : $\sigma^2 = 36$.

a) Donner une estimation ponctuelle de μ.

b) Construire un intervalle de confiance pour μ au niveau 0,9 en utilisant l'inégalité de Bienaymé-Tchebychev.

c) Construire un deuxième intervalle de confiance pour μ au niveau 0,9 en utilisant un résultat approprié du cours.

d) Parmi les deux intervalles obtenus au b) et au c), lequel est le plus intéressant ?

e) On reprend la méthode utilisée au c). La production de combien de pommiers faudrait-il mesurer (au minimum) pour obtenir un intervalle de confiance pour μ au niveau 0,9 et ayant une longueur inférieure à 4 ?

2) Dans cette question, on suppose la variance σ^2 inconnue.

a) Donner une estimation ponctuelle de σ^2 fournie par un estimateur non biaisé de σ^2.

b) Construire un intervalle de confiance au niveau 0,9 pour μ en utilisant un résultat approprié du cours.

c) Construire (au moins) deux intervalles de confiance au niveau 0,9 pour σ^2.

Exercice 3. Le champignon *Ceratocystis platani* est responsable d'une maladie des platanes, appelée chancre coloré du platane, qui nécessite l'abattage systématique des arbres atteints. Afin d'évaluer la proportion p de platanes contaminés par cette maladie dans une région donnée, on examine un échantillon de n de ces arbres et on définit des variables aléatoires $X_1, X_2, \cdots, X_n$ en posant :

$$X_i = \begin{cases} 1 & \text{si le } i^{eme} \text{ platane de l'échantillon est atteint par la maladie;} \\ 0 & \text{sinon.} \end{cases}$$

On suppose par la suite que $n = 225$ et que 43 des 225 platanes de l'échantillon sont atteints par le chancre coloré. ☞ Les bornes des intervalles de confiance demandés seront données en arrondissant convenablement à deux chiffres après la virgule.
1) Donner une estimation ponctuelle de p.
2) a) Quelle est la loi de probabilité commune des X_i ? En déduire la variance de la moyenne d'échantillon $\overline{X}_{225}$.

b) Vérifier que $p(1-p) \leq \frac{1}{4}$ pour tout $p \in [0,1]$.

c) Construire à l'aide de l'inégalité de Bienaymé-Tchebychev et du b) un intervalle de confiance de niveau 0,8 pour μ.
3) En utilisant un résultat approprié du cours, construire un nouvel intervalle de confiance (asymptotique) au niveau $0,8$ pour p. Comparer à celui obtenu au 2).

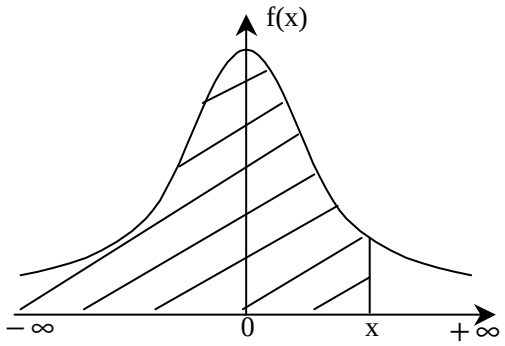
Exercice 4. On considère un échantillon $X_1, X_2, \cdots, X_n$ de la loi uniforme $\mathcal{U}([0,\theta])$ où $\theta > 0$ et on définit $\widetilde{\theta}_n = \max\{X_1, \cdots, X_n\}$.
1) Étant donné $c \in]0,1[$, déterminer x tel $\mathbb{P}(x \leq \widetilde{\theta}_n \leq \theta) = c$. ☞ Vous pourrez utiliser l'Exercice 3 du Chapitre 1.
2) En déduire un intervalle de confiance au niveau c pour le paramètre θ qui soit de la forme $[\widetilde{\theta}_n, k_n \times \widetilde{\theta}_n]$, où k_n est à préciser un nombre qui dépend de n.
3) Calculer l'intervalle obtenu au 2) en reprenant les valeurs numériques de l'Exercice 3 du Chapitre 1 et $c = 0,8$.

Chapitre 3

Annexe : tables de lois

Loi Normale centrée réduite

Probabilité de trouver une valeur inférieure à x.



$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2p}} e^{-\frac{u^2}{2}} du$$

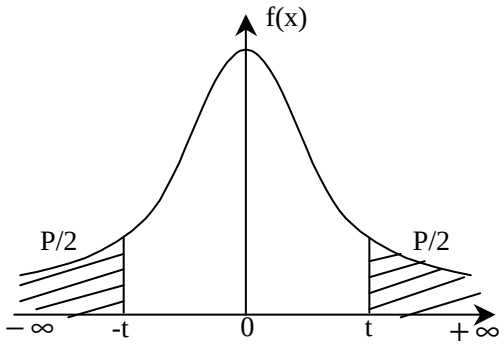
X	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

Table pour les grandes valeurs de x :

x	3	3,2	3,4	3,6	3,8	4	4,2	4,4	4,6	4,8
F(x)	0,99865003	0,99931280	0,99966302	0,99984085	0,99992763	0,99996831	0,99998665	0,99999458	0,99999789	0,99999921

Loi de Student

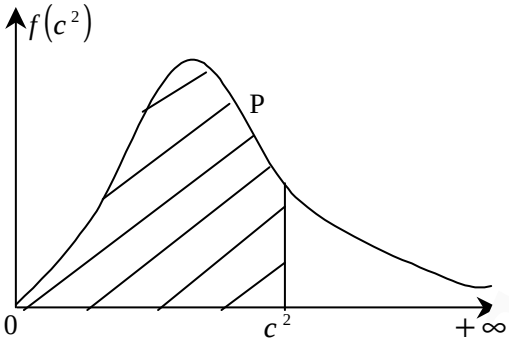
Valeurs de t ayant la probabilité P d'être dépassées en valeur absolue.



n \P	90%	80%	70%	60%	50%	40%	30%	20%	10%	5%	1%
1	0,1584	0,3249	0,5095	0,7265	1,0000	1,3764	1,9626	3,0777	6,3137	12,7062	63,6559
2	0,1421	0,2887	0,4447	0,6172	0,8165	1,0607	1,3862	1,8856	2,9200	4,3027	9,9250
3	0,1366	0,2767	0,4242	0,5844	0,7649	0,9785	1,2498	1,6377	2,3534	3,1824	5,8408
4	0,1338	0,2707	0,4142	0,5686	0,7407	0,9410	1,1896	1,5332	2,1318	2,7765	4,6041
5	0,1322	0,2672	0,4082	0,5594	0,7267	0,9195	1,1558	1,4759	2,0150	2,5706	4,0321
6	0,1311	0,2648	0,4043	0,5534	0,7176	0,9057	1,1342	1,4398	1,9432	2,4469	3,7074
7	0,1303	0,2632	0,4015	0,5491	0,7111	0,8960	1,1192	1,4149	1,8946	2,3646	3,4995
8	0,1297	0,2619	0,3995	0,5459	0,7064	0,8889	1,1081	1,3968	1,8595	2,3060	3,3554
9	0,1293	0,2610	0,3979	0,5435	0,7027	0,8834	1,0997	1,3830	1,8331	2,2622	3,2498
10	0,1289	0,2602	0,3966	0,5415	0,6998	0,8791	1,0931	1,3722	1,8125	2,2281	3,1693
11	0,1286	0,2596	0,3956	0,5399	0,6974	0,8755	1,0877	1,3634	1,7959	2,2010	3,1058
12	0,1283	0,2590	0,3947	0,5386	0,6955	0,8726	1,0832	1,3562	1,7823	2,1788	3,0545
13	0,1281	0,2586	0,3940	0,5375	0,6938	0,8702	1,0795	1,3502	1,7709	2,1604	3,0123
14	0,1280	0,2582	0,3933	0,5366	0,6924	0,8681	1,0763	1,3450	1,7613	2,1448	2,9768
15	0,1278	0,2579	0,3928	0,5357	0,6912	0,8662	1,0735	1,3406	1,7531	2,1315	2,9467
16	0,1277	0,2576	0,3923	0,5350	0,6901	0,8647	1,0711	1,3368	1,7459	2,1199	2,9208
17	0,1276	0,2573	0,3919	0,5344	0,6892	0,8633	1,0690	1,3334	1,7396	2,1098	2,8982
18	0,1274	0,2571	0,3915	0,5338	0,6884	0,8620	1,0672	1,3304	1,7341	2,1009	2,8784
19	0,1274	0,2569	0,3912	0,5333	0,6876	0,8610	1,0655	1,3277	1,7291	2,0930	2,8609
20	0,1273	0,2567	0,3909	0,5329	0,6870	0,8600	1,0640	1,3253	1,7247	2,0860	2,8453
21	0,1272	0,2566	0,3906	0,5325	0,6864	0,8591	1,0627	1,3232	1,7207	2,0796	2,8314
22	0,1271	0,2564	0,3904	0,5321	0,6858	0,8583	1,0614	1,3212	1,7171	2,0739	2,8188
23	0,1271	0,2563	0,3902	0,5317	0,6853	0,8575	1,0603	1,3195	1,7139	2,0687	2,8073
24	0,1270	0,2562	0,3900	0,5314	0,6848	0,8569	1,0593	1,3178	1,7109	2,0639	2,7970
25	0,1269	0,2561	0,3898	0,5312	0,6844	0,8562	1,0584	1,3163	1,7081	2,0595	2,7874
26	0,1269	0,2560	0,3896	0,5309	0,6840	0,8557	1,0575	1,3150	1,7056	2,0555	2,7787
27	0,1268	0,2559	0,3894	0,5306	0,6837	0,8551	1,0567	1,3137	1,7033	2,0518	2,7707
28	0,1268	0,2558	0,3893	0,5304	0,6834	0,8546	1,0560	1,3125	1,7011	2,0484	2,7633
29	0,1268	0,2557	0,3892	0,5302	0,6830	0,8542	1,0553	1,3114	1,6991	2,0452	2,7564
30	0,1267	0,2556	0,3890	0,5300	0,6828	0,8538	1,0547	1,3104	1,6973	2,0423	2,7500
40	0,1265	0,2550	0,3881	0,5286	0,6807	0,8507	1,0500	1,3031	1,6839	2,0211	2,7045
50	0,1263	0,2547	0,3875	0,5278	0,6794	0,8489	1,0473	1,2987	1,6759	2,0086	2,6778
60	0,1262	0,2545	0,3872	0,5272	0,6786	0,8477	1,0455	1,2958	1,6706	2,0003	2,6603
80	0,1261	0,2542	0,3867	0,5265	0,6776	0,8461	1,0432	1,2922	1,6641	1,9901	2,6387
100	0,1260	0,2540	0,3864	0,5261	0,6770	0,8452	1,0418	1,2901	1,6602	1,9840	2,6259
120	0,1259	0,2539	0,3862	0,5258	0,6765	0,8446	1,0409	1,2886	1,6576	1,9799	2,6174
200	0,1258	0,2537	0,3859	0,5252	0,6757	0,8434	1,0391	1,2858	1,6525	1,9719	2,6006
∞	0,1257	0,2533	0,3853	0,5244	0,6745	0,8416	1,0364	1,2816	1,6449	1,9600	2,5758

Loi du c^2

Valeur de c^2 ayant la probabilité P d'être dépassée.



ddl/P	0,5%	1,0%	2,5%	5,0%	10,0%	50,0%	90,0%	95,0%	97,5%	99,0%	99,5%
1	0,000	0,000	0,001	0,004	0,016	0,455	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	1,386	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	2,366	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	3,357	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	4,351	9,236	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	5,348	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	6,346	12,017	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	3,490	7,344	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	8,343	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	9,342	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	10,341	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	11,340	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,042	12,340	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	13,339	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	14,339	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	15,338	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	16,338	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	17,338	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	18,338	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	19,337	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	20,337	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	21,337	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	22,337	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	23,337	33,196	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	16,473	24,337	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	25,336	35,563	38,885	41,923	45,642	48,290
27	11,808	12,878	14,573	16,151	18,114	26,336	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	27,336	37,916	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	19,768	28,336	39,087	42,557	45,722	49,588	52,335
30	13,787	14,953	16,791	18,493	20,599	29,336	40,256	43,773	46,979	50,892	53,672
31	14,458	15,655	17,539	19,281	21,434	30,336	41,422	44,985	48,232	52,191	55,002
32	15,134	16,362	18,291	20,072	22,271	31,336	42,585	46,194	49,480	53,486	56,328
33	15,815	17,073	19,047	20,867	23,110	32,336	43,745	47,400	50,725	54,775	57,648
34	16,501	17,789	19,806	21,664	23,952	33,336	44,903	48,602	51,966	56,061	58,964
35	17,192	18,509	20,569	22,465	24,797	34,336	46,059	49,802	53,203	57,342	60,275

Lorsque $n > 30$ on peut admettre que la quantité $\sqrt{2c^2} - \sqrt{2n-1}$ suit une loi normale centrée réduite.