Mathématiques et Statistiques appliquées à la Gestion

Compléments et Exercices
BBA-1 (2020-2021)

Guillaume Metzler
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC EA3083, Lyon, France
guillaume.metzler@live.fr ou guillaume.metzler@univ-lyon2.fr

Résumé

Ce document est à destination des étudiants de l'*EM Lyon*. Il a pour objectif de répondre aux questions posées par les étudiants au cours des différentes séances. Il regroupe également des corrections de différents exercices proposés en cours. Si vous avez des points à signaler, ou des incompréhensions qui persistent, il ne faut pas hésiter à me les signaler. Ce document sera complété au fur et à mesure des séances.

Table des matières

1 Concernant le premier module

L'objectif de ce premier module est de motiver l'intérêt de l'usage des outils statistiques notamment pour faire de **l'inférence**, *i.e.* pour déduire des caractéristiques d'une **population** à partir d'un **échantillon**. Nous avons abordé la notion d'*estimation ponctuelle* d'un paramètre, c'est-à-dire la mesure de la valeur d'un paramètre relativement à un échantillon.

Ce cours était également l'occasion de faire des rappels sur la loi binomiale $\mathcal{B}(n, p)$ dont l'exemple le plus connu est la répétition de plusieurs lancers de pièces effectués de façon indépendante où p est la *probabilité de succès* et n désigne le nombre d'expériences effectuées. Pour rappel, la fonction de probabilité d'une variable aléatoire $X \sim \mathcal{B}(n, p)$ est définie par

$$\mathbb{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}.$$

Son espérance et sa variance sont respectivement égales à

$$\mathbb{E}[X] = np \quad \text{et} \quad Var(X) = np(1-p).$$

Nous avons ensuite vu que lorsque n est assez grand, typiquement lorsque $n \geq 30$, alors on peut approximer la loi binomiale par une loi normale de paramètres $\mu = np$ et de variance $\sigma^2 = np(1-p)$.

Nous avons enfin vu quelques définitions et propriétés de la loi normale que nous rappelons ci-dessous, ainsi que quelques rappels élémentaires en probabilités.

1.1 Quelques rappels

Rappels sur la loi normale. Une variable aléatoire $X \sim \mathcal{N}(\mu, \sigma^2)$, où μ désigne la moyenne de la loi normale et σ l'écart type (donc σ^2 représente la variance) admet pour densité de probabilité (ou densité tout court) la fonction f suivante

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

La fonction de répartition (fonction des probabilités cumulatives, en référence à *table des Z* se trouvant à la fin du premier cours) est notée F et est définie par

$$F(t) = \mathbb{P}[X \leq t] = \int_{-\infty}^{t} f(x)dx = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right) dx.$$

Ainsi le tableau des fréquences cumulatives ne fourni rien d'autre que les valeurs de la fonction de répartition F. Par exemple, pour $t = 1.38$, on a $F(1.38) = \mathbb{P}[X \leq 1.38] = 0.9162$. Elle correspond à l'aire sous la courbe rouge présentée en Figure 1.

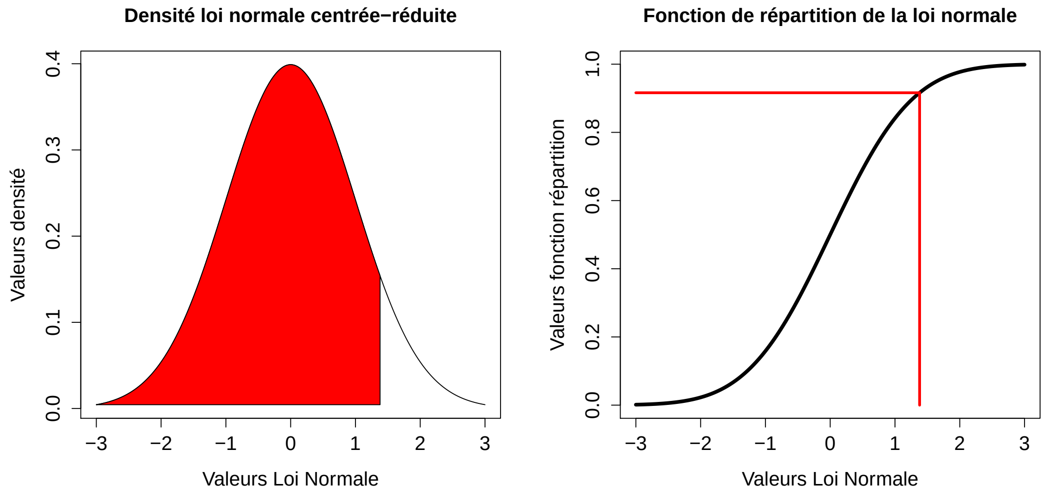

FIGURE 1 – Exemple représentant l'aire sous la courbe (en rouge) d'une loi normale centrée réduite pour $t = 1.38$ (valeur de t à lire sur l'axe des abscisses), à gauche. La figure de droite représente la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0, 1)$. L'axe des ordonnées renseigne directement sur la valeur de la fonction de répartition pour la valeur mentionnée en abscisse.

Quelques propriétés de loi normale. Une variable aléatoire X suivant une loi normale de moyenne μ et de variance σ^2 (*i.e.* d'écart-type σ) admet les propriétés suivantes :

- **Moyenne** $=$ **Médiane** $=$ **Mode** .
- **Symétrie :** pour tout nombre réel t : $\mathbb{P}[X \leq \mu - t] = \mathbb{P}[X \geq \mu + t]$.

On se rappelle également que pour ton nombre réel t, et pour n'importe quelle loi de la variable aléatoire X, nous avons

$$\mathbb{P}[X \leq t] = 1 - \mathbb{P}[X \geq t].$$

Si on reprend l'exemple de Figure 1 à gauche, cela veut dire que l'aire sous la courbe en rouge est égale à 1- l'aire sous la courbe en blanc.
Enfin, pour tout réel t_1 et t_2, nous avons

$$\mathbb{P}[t_1 \leq X \leq t_2] = \mathbb{P}[X \leq t_2] - \mathbb{P}[X \leq t_1].$$

On fini enfin par quelques chiffres importants concernant la loi normale :

- 95% des valeurs prise par une variable aléatoire $X \sim \mathcal{N}(\mu, \sigma^2)$ sont comprises dans l'intervalle $[\mu - 1.96\sigma, \mu + 1.96\sigma]$
- 99% des valeurs prise par une variable aléatoire $X \sim \mathcal{N}(\mu, \sigma^2)$ sont comprises dans l'intervalle $[\mu - 2.58\sigma, \mu + 2.58\sigma]$

Normalisation. Nous avons vu que nous disposons uniquement des probabilités de la forme $\mathbb{P}[Z \leq t]$ lorsque $Z \sim \mathcal{N}(0, 1)$, ce sont les éléments se trouvant dans la *table des Z.*

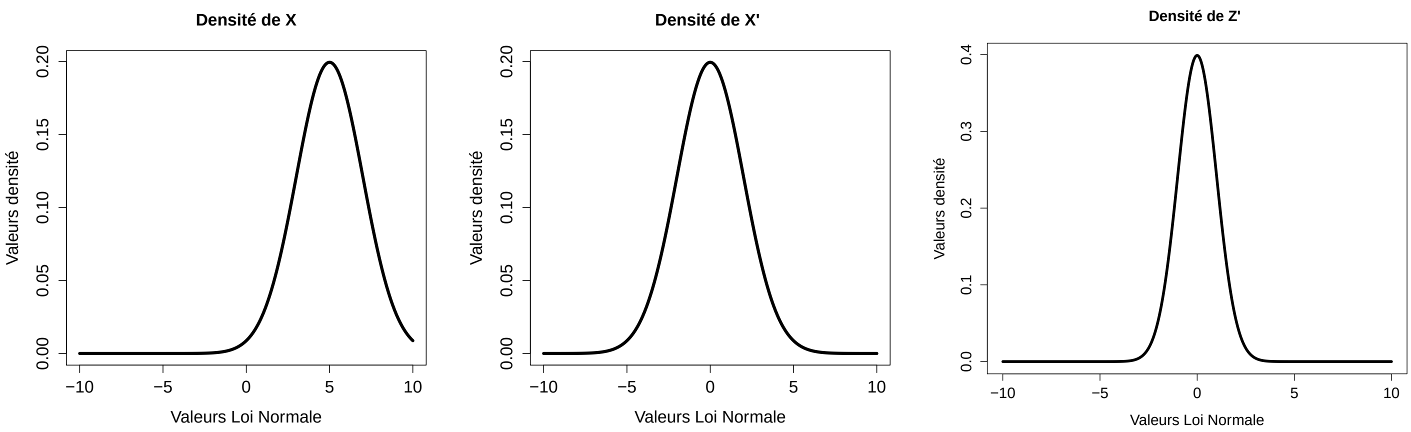

FIGURE 2 – Figures illustrant (de gauche à droite) les étapes de la normalisation. La première figure de gauche montre notre distribution (ou densité) initiale. La figure du milieu montre le recentrage de la gaussienne en 0. La figure de droite montre l'ajustement du facteur d'échelle (la réduction à un écart-type égal à 1) de notre distribution.

Ainsi, partant d'une variable aléatoire $X \sim \mathcal{N}(\mu, \sigma^2)$, il est donc intéressant de se ramener à une loi normale centrée-réduite si l'on cherche à estimer des probabilité de la forme :

$$\mathbb{P}[X \leq t] \quad \text{ou encore} \quad \mathbb{P}[t_1 \leq X \leq t_2].$$

Cela se fait en appliquant la transformation suivante

$$Z = \frac{X - \mu}{\sigma}.$$

Démonstration. On se rappelle des propriétés suivantes, pour tout nombre réel a, concernant l'**espérance** et la **variance** d'une variable aléatoire X

(i) $\mathbb{E}[X + a] = \mathbb{E}[X] + a$,
(ii) $\mathbb{E}[aX] = a\mathbb{E}[X]$,
(iii) $Var(a + X) = Var(X)$,
(iv) $Var(aX) = a^2 Var(X)$.

Ainsi, si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors d'après i), l'espérance de la variable aléatoire $X' = X - \mu$ est égale à 0, la variance reste elle inchangée d'après (iii) L'espérance de la variable aléatoire $Z = \dfrac{X'}{\sigma} = \dfrac{X - \mu}{\sigma}$ est également égale à 0 d'après (ii) et sa variance est égale à 1.

□

L'influence des différentes étapes de la transformation est illustrée en Figure 2.

Méthode. Pour pouvoir effectuer une lecture dans la *table des Z*, il est important de se ramener à des inégalités de la forme

$$\mathbb{P}[Z \leq t] \quad \text{où} \quad t \geq 0,$$

afin de pouvoir exploiter la table.

Exemple. Soit $X \sim \mathcal{N}(5,2)$ et déterminons la probabilité que notre variable aléatoire X prenne des valeurs plus petites que 1, *i.e.* $\mathbb{P}[X \leq 1]$.
La première étape consiste à normaliser

$$
\begin{aligned}
&\mathbb{P}[X \leq 1] \\
&\quad \downarrow \text{ on retranche la moyenne de } X \text{ puis on divisepar son écart-type de part et d'autre de l'inégalité} \\
&= \mathbb{P}\left[\frac{X-5}{2} \leq \frac{1-5}{2}\right], \\
&\quad \downarrow \text{ en simplifiant ce qui se trouve à droite l'inégalité} \\
&= \mathbb{P}\left[\frac{X-5}{2} \leq -2\right], \\
&\quad \downarrow \text{ on pose } Z = \frac{X-5}{2} \text{ et } Z \sim \mathcal{N}(0,1) \\
&= \mathbb{P}\left[Z \leq -2\right], \\
&\quad \downarrow \text{ symétrie de la loin normale centrée réduite} \\
&= \mathbb{P}\left[Z \geq 2\right], \\
&\quad \downarrow \text{ on utilise le fait que } \mathbb{P}[Z \leq t] = 1 - \mathbb{P}[Z \geq t] \\
&= 1 - \mathbb{P}\left[Z \leq 2\right], \\
&\quad \downarrow \text{ on cherche la valeur dans la table des Z} \\
&= 1 - 0.977, \\
&= 0.023.
\end{aligned}
$$

1.2 Exercice

Cet exercice, proposé lors du premier cours a bout objectif de vous faire travailler sur la lecture de la *table des Z*, i.e. sur des valeurs de la fonction de répartition de la loi normale.

Énoncé On se propose maintenant

a) étant donnée une variable aléatoire $X \sim \mathcal{N}(10, 50^2)$, de calculer la probabilité $\mathbb{P}[10 \leq X \leq 50]$.
b) étant donnée une variable aléatoire $X \sim \mathcal{N}(10, 20^2)$, de calculer la probabilité $\mathbb{P}[X \leq -5]$.

Correction

a) On commence par normaliser notre variable aléatoire afin de se ramener à une variable aléatoire Z suivant une loi normale centrée réduite puis on va ré-exprimer notre encadrement comme une différence de deux probabilité.

$$\begin{aligned}
\mathbb{P}[10 \leq X \leq 50] &= \mathbb{P}\left[\frac{10-10}{50} \leq \frac{X-10}{50} \leq \frac{50-10}{50}\right], \\
&= \mathbb{P}\left[0 \leq Z \leq 0.8\right], \\
&= \mathbb{P}\left[Z \leq 0.8\right] - \mathbb{P}\left[Z \leq 0\right], \\
&= 0.788 - 0.5, \\
&= 0.288.
\end{aligned}$$

b) Les étapes sont identiques à celles présentées en exemple.

$$\begin{aligned}
\mathbb{P}[X \leq -5] &= \mathbb{P}\left[\frac{X-10}{20} \leq \frac{-5-10}{20}\right], \\
&= \mathbb{P}\left[Z \leq -0.75\right], \\
&= \mathbb{P}\left[Z \geq 0.75\right], \\
&= 1 - \mathbb{P}\left[Z \leq 0.75\right], \\
&= 1 - 0.773, \\
&= 0.226.
\end{aligned}$$

1.3 Pour s'entraîner

Exercice. On considère que les âges dans une ville sont distribuées selon une loi normale $\mathcal{N}(\mu, \sigma^2)$, dont les valeurs de μ et σ dépendent de la ville. On considère qu'un individu est jeune s'il a moins de 25 ans (au secours ! je suis un vieux) et qu'il est vieux, s'il a plus de 50 ans (ah non ! je ne suis pas encore vieux, ouf !). Finalement un individu dont l'âge est compris entre 30 et 40 sera considéré comme étant dans la fleur de l'âge (mais je suis dans quelle catégorie moi ...).

On suppose que l'âge de la population de la ville de Strasbourg suit une loi normale de paramètres $\mu = 33$ et $\sigma = 20$ alors qu'à Lyon, la population suit une loi normale de paramètres $\mu = 35$ et $\sigma = 22$.

a) Quelles sont les limites de cette modélisation ?
b) Dans quelle ville pouvons nous trouver le plus de jeunes ?
c) A l'inverse, dans quelle ville trouve-t-on le plus de vieux ?
d) Finalement, dans quelle ville trouve-t-on le plus de personnes dans la fleur de l'âge ?

2 Concernant le deuxième module

Lors de cette deuxième séance nous sommes revenus sur l'intérêt que l'on pouvait avoir à étudier des **échantillons** pour **inférer** des informations sur notre **population**.

Cependant, il est légitime de penser qu'un échantillon seul ne permet de garantir que l'on dispose d'une bonne estimation de notre paramètre de population. Par exemple, si l'on dispose d'un autre échantillon, nous avons toutes les chances d'obtenir une estimation différente de notre paramètre de population. C'est ce que l'on appelle la **fluctuation d'échantillonnage**.

Dans ce cours, on va alors se servir des estimations effectuées sur plusieurs échantillons pour obtenir **non plus une estimation ponctuelle mais une estimation sous forme de distribution**. Notre estimation devient donc une **variable aléatoire** à laquelle on peut associer *une espérance*, *une variance* et une*fonction de probabilité.*

Une fois que l'on connait notre distribution d'échantillonnage, *i.e.* la distribution de notre estimateur, nous sommes en mesure d'établir des intervalles de confiance pouvant comprendre le paramètre de population.

Dans le cadre de ce module, on considère que le paramètre de population que l'on cherche à estimer est la moyenne μ. L'*estimateur* que l'on va utiliser pour "déterminer" μ est appelé **moment d'ordre** 1 ou **moyenne empirique** (empirique fait référence à l'évaluation du paramètre sur un échantillon et non sur la distribution) que l'on note $\bar{X}_n$ ou encore $\bar{X}$.

Dans cette section, on considère que l'écart-type σ des lois considérées est **connue**.

2.1 Quelques rappels

Estimateur de la moyenne et propriétés. Lorsque l'on cherche à estimer la moyenne d'une population, on se base sur la **moyenne empirique** notée $\bar{x}$, et définie par :

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Exemple : on interroge trois individus pour savoir combien de temps dans la journée ils consacrent au sport et on note x_1, x_2 et x_3 leur réponse. On peut estimer que le temps moyen qu'une personne consacre au sport dans sa journée est égal à

$$\bar{x}_3 = \frac{1}{3}\sum_{i=1}^{3} = \frac{x_1 + x_2 + x_3}{3}.$$

Soit maintenant une population suivant une certaine distribution ayant une espérance μ inconnue et que l'on cherche à estimer ainsi qu'une variance σ^2 connue. Nous avons vu que si l'on considère plusieurs échantillons de cette population, nous obtenons a priori plusieurs estimations différentes du paramètre de population μ à estimer.
En ce sens, notre n-estimateur (c'est-à-dire un estimateur reposant sur l'utilisation d'un échantillon aléatoire de taille n) de la moyenne peut alors être considéré comme une variable aléatoire $\bar{X}_n$ qui possède les propriétés suivantes :

(i) $\mathbb{E}[\bar{X}_n] = \mu$, *i.e.* l'espérance de notre estimateur de la moyenne est égale à la moyenne de la population,

(ii) $\sigma_{\bar{X}_n} = \dfrac{\sigma}{\sqrt{n}}$, *i.e.* l'écart type de notre estimateur de la moyenne est égal à l'écart-type de la population divisé par la racine carré de la taille de l'échantillon.

Démonstration. Soit X une variable aléatoire de paramètre μ inconnu et σ connu. On considère maintenant un échantillon de taille n obtenu par "tirages indépendants", les valeurs prisent suivent la même loi que X, ce sont en fait n copies de cette variable aléatoire, que l'on notera $X_1, X_2, ..., X_n$. Elle sont dites **indépendantes** et **identiquement distribuées** et ont donc même espérance (μ) et variance (σ^2) que X. Dans ce cas :

(i) On commence par déterminer l'espérance de $\bar{X}_n$ en utilisant la linéarité de l'espérance, *i.e.* $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. Ce qui nous donne :

$$\begin{aligned}
\mathbb{E}[\bar{X}_n] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right], \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i\right], \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu, \\
\mathbb{E}[\bar{X}_n] &= \mu.
\end{aligned}$$

(ii) On commence par déterminer la variance de $\bar{X}_n$ pour ensuite en déduire l'écart-type.

$$\begin{aligned}
Var(\bar{X}_n) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right), \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i), \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2, \\
Var(\bar{X}_n) &= \frac{\sigma^2}{n}.
\end{aligned}$$

On a donc $\sigma_{\bar{X}_n} = \sqrt{Var\left(\bar{X}_n\right)} = \dfrac{\sigma}{\sqrt{n}}$.

□

Remarque : l'utilisation de cet écart-type pour notre estimateur n'est valable que si la taille de l'échantillon est négligeable devant la taille de la population, *i.e.* si la taille de l'échantillon ne représente pas plus de 5% de la population. Dans le cas contraire, on utilisera

$$\sigma_{\bar{X}_n} = \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}}.$$

Le théorème central limite nous dit que la loi $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ est une bonne approximation de la loi de $\bar{X}_n$ lorsque n est suffisamment grand, *i.e.* lorsque $n \geq 30$.
De façon analogue, ce même théorème nous dit que la variable aléatoire $Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ converge vers une variable aléatoire $Z \sim \mathcal{N}(0,1)$.

Intervalle de confiance. Maintenant que l'on connait la loi (asymptotique) suivie par notre estimateur de la moyenne $\bar{X}_n$, il est tout fait possible, non plus d'effectuer des estimation ponctuelles de notre paramètre population mais des estimations par intervalle.

Une première conséquence du théorème central limite est que l'on peut avoir des informations quant aux valeurs obtenues par échantillonnage. Par exemple, comme le montre la Figure 3, 95% des valeurs moyenne obtenues par échantillonnage seront comprises dans l'intervalle

$$\left[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right],$$

qui sont représentées par la zone bleue dans le graphique.

Le paragraphe précédent énonce donc que

$$\mathbb{P}\left[\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}\right] = 95\%.$$

On peut cependant réécrire cet encadrement comme suit

$$\begin{aligned}
\mathbb{P}\left[\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}\right] &= \mathbb{P}\left[-1.96\frac{\sigma}{\sqrt{n}} \leq -\mu + \bar{x} \leq 1.96\frac{\sigma}{\sqrt{n}}\right], \\
&= \mathbb{P}\left[-\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right], \\
&= \mathbb{P}\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]
\end{aligned}$$

Ce résultat nous indique qu'une estimation de la moyenne par échantillonnage permet de construire un intervalle qui contiendra à la valeur du paramètre μ avec une probabilité égale à la précédente, cette probabilité est un **score de confiance** qui est donc associé à un **intervalle de confiance** qui est donc de la forme

$$\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

.

Le **score de confiance** est associé à un intervalle de confiance, à ce même intervalle intervalle de confiance on lui associe également **un risque** ou **une marge d'erreur**. Dans le cas précédent, le score de confiance associée à l'intervalle présenté était égal 95% et donc la marge d'erreur était de 5%.
On est bien sûr libre de choisir un risque plus faible et dans ce cas nous devons construire un

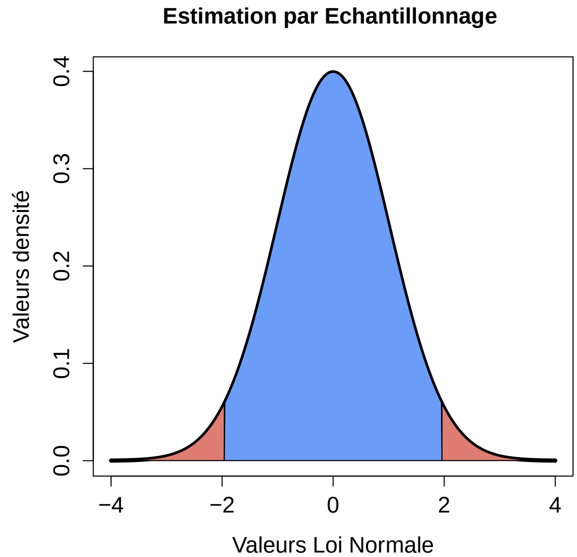

FIGURE 3 – Illustration des valeurs obtenues par échantillonnage. Dans la zone bleue se trouvent 95% des valeurs que l'on obtiendrait par échantillonnage et dans les zones rouges les 5% restants.

intervalle plus grand ou inversement.

Remarque : notez que depuis le début nous construisons des intervalles qui sont symétriques par rapport à la moyenne. On souhaite ici tirer profits de la symétrie de la loi normale.

De façon générale, on peut construire des intervalles de confiance avec n'importe quelle marge d'erreur α et donc des intervalles de confiance avec un score de confiance égal à $1-\alpha$, dont on illustre la taille en Figure 4 pour différentes valeurs de α.
Etant donnée la symétrie de la loi normale et tout comme nous construisons des intervalles symétriques par rapport à $\bar{x}$, la marge d'erreur est répartie équitablement de part et d'autre de la gaussienne.

Construction d'un intervalle de confiance. Etant donnée une marge d'erreur α, l'objectif est maintenant de déterminer les bornes supérieures et inférieures de notre intervalle de confiance. Or comme l'intervalle de confiance est symétrique par rapport à $\bar{x}$, il est suffisant de déterminer la borne supérieure (ou inférieure) afin d'en déduire l'autre.

Nous faisons le choix de rechercher la borne supérieure $\bar{x}_{sup}$, cette dernière est définie par la relation suivante :

$$\mathbb{P}[\bar{X} \geq \bar{x}_{sup}] = \frac{\alpha}{2}.$$

Cette dernière expression est équivalente à rechercher $\bar{x}_{sup}$ telle que

$$\mathbb{P}[\bar{X} \leq \bar{x}_{sup}] = 1 - \frac{\alpha}{2}.$$

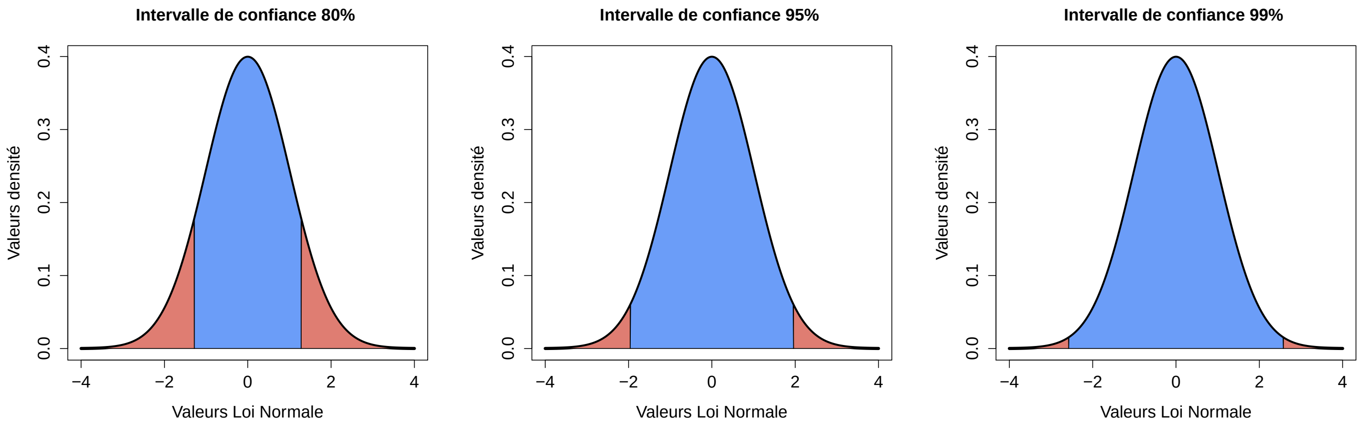

FIGURE 4 – Illustration de différents intervalles de confiance pour des marges d'erreur respectivement égales à $20, 5$ et 1% représentées en rouge sur les graphes.

Or, on se rappelle que la variable aléatoire $Z = \dfrac{\bar{X} - \bar{x}}{\dfrac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$. On va donc tirer profit de cette propriété pour déterminer $\bar{x}_{sup}$

$$\mathbb{P}[\bar{X} \leq \bar{x}_{sup}] = \mathbb{P}\left[Z \leq \frac{\bar{x}_{sup} - \bar{x}}{\dfrac{\sigma}{\sqrt{n}}}\right] = \mathbb{P}\left[Z \leq z_{1-\alpha/2}\right] = 1 - \frac{\alpha}{2}.$$

Il nous font donc déterminer la valeur $z_{1-\alpha/2}$ telle que $\mathbb{P}\left[Z \leq z_{1-\alpha/2}\right] = 1 - \dfrac{\alpha}{2}$. Comme nous l'avons fait lors du premier module, cette information va se trouver dans la *table des Z*. **Sauf que cette fois-ci on part de la probabilité connue $1 - \alpha/2$ afin de retrouver la valeur qui conduit à cette probabilité.**

Une fois la valeur de $z_{1-\alpha/2}$ déterminée on utilise la relation $z_{1-\alpha/2} = \dfrac{\bar{x}_{sup} - \bar{x}}{\dfrac{\sigma}{\sqrt{n}}}$ pour déterminer $\bar{x}_{sup}$, ce qui nous donne :

$$\bar{x}_{sup} = \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

De façon analogue, on obtiendra $\bar{x}_{inf}$:

$$\bar{x}_{inf} = \bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},$$

ce qui peut aussi s'écrire

$$\bar{x}_{inf} = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

en utilisant la relation $z_{1-\alpha/2} = -z_{\alpha/2}$ qui est une conséquence de la symétrie de la loi normale centrée réduite.

Remarque : Ce deuxième module consiste à effectuer le processus inverse de ce qui a été fait lors du premier module. Dans le premier module, nous devions déterminer la probabilité que notre variable aléatoire prenne une valeur plus petite qu'une valeur donnée à l'aide de la *table des Z*. Cette fois-ci, on connait la probabilité et il nous faut trouver, dans la t*able des Z*, la valeur à ne pas dépasser qui conduit à cette probabilité.

Exemple. On a tiré $10,000$ échantillons de parfum afin de mesurer la quantité de liquide présent dans les fioles. On souhaite vérifier que la moyenne de remplissage est toujours égale à 50.2 ml. Le volume moyen de parfum dans les fioles de nos échantillons est égal à 50.5 ml. On suppose en outre que le processus de remplissage suit une loi normale d'écart-type $\sigma = 2$.
Peut-on affirmer que la machine est correctement réglée avec un taux d'erreur de 10% ?

D'après l'énoncé, nous avons $n = 10,000, \bar{x} = 51, \sigma = 2$ et une marge d'erreur $\alpha = 0.1$ car on souhaite un intervalle de confiance de $1 - \alpha = 90\%$. On rappelle que notre intervalle de confiance est symétrique autour de $\bar{x}$ et vérifie

$$\mathbb{P}\left[\bar{x}_{inf} \leq \mu \leq \bar{x}_{sup}\right] = \mathbb{P}\left[\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha.$$

On se propose ici de déterminer la borne supérieure $\bar{x}_{sup}$. Par définition, $z_{1-\alpha/2}$ est la valeur, pour une variable centrée réduite Z, pour laquelle on a :

$$\mathbb{P}\left[Z \leq z_{1-\alpha/2}\right] = 1 - \frac{\alpha}{2}.$$

Dans notre cas $\alpha = 0.1$, donc $1 - \dfrac{\alpha}{2} = 0.95$. En cherchant dans la *table des Z*, on trouve que la valeur de $z_{1-\alpha/2}$ est égale à 1.64

D'après ce que nous avons vu précédemment, on a donc

$$\bar{x}_{sup} = \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 50.5 + 1.64 \times \frac{2}{\sqrt{10,000}} = 50.5 + 0.0328 \simeq 50.53.$$

In fine, notre intervalle de confiance est le suivant :

$$[50.47, 50.53].$$

Or $50.2 \notin [50.47, 50.53]$, on ne peut donc pas affirmer que notre machine est bien réglée avec une marge d'erreur de 10%.

2.2 Exercice

Énoncé On se propose maintenant de

a) donner un intervalle avec un niveau de confiance égale à 90% de la moyenne de notre population sachant que $\bar{x} = 50$ que l'écart type σ de notre population est connu et égal à 100 et que l'on dispose dispose d'un échantillon de taille $n = 100$.
b) de faire de même avec un niveau de confiance égale à 85%, sachant que $\bar{x} = 50, \sigma = 30$ et que l'on dispose dispose d'un échantillon de taille $n = 36$.

Correction. On reprend les mêmes étapes que celles effectuées dans l'exemple ci dessus pour les deux questions

a) On cherche un intervalle de confiance avec un score de confiance égal à $1-\alpha = 0.9$ donc la marge d'erreur est égale à $\alpha = 0.1$. Comme dans l'exemple on commence par cherche la valeur de $z_{1-\alpha/2}$ qui est ici égale à 1.64 comme dans l'exemple.
Ainsi la valeur de $\bar{x}_{sup}$ est

$$\bar{x}_{sup} = \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 50 + 1.64 \times \frac{10}{\sqrt{100}} = 50 + 1.64 \simeq 51.64,$$

et notre intervalle de confiance est alors défini par

$$[\bar{x}_{inf}, \bar{x}_{sup}] = \left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = [48.36, 51.64].$$

.

b) On cherche un intervalle de confiance avec un score de confiance égal à $1-\alpha = 0.85$ donc la marge d'erreur est égale à $\alpha = 0.15$. Comme dans l'exemple on commence par cherche la valeur de $z_{1-\alpha/2}$ qui est ici égale à 1.44.
Ainsi la valeur de $\bar{x}_{sup}$ est

$$\bar{x}_{sup} = \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 50 + 1.44 \times \frac{30}{\sqrt{36}} = 50 + 7.2 \simeq 57.2,$$

et notre intervalle de confiance est alors défini par

$$[\bar{x}_{inf}, \bar{x}_{sup}] = \left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = [42.8, 57.2].$$

.

2.3 Pour s'entraîner

Exercice 1. La durée de vie d'une ampoule, donnée en en heures, est représentée par une variable aléatoire X dont la distribution est supposée Normale avec un écart-type $\sigma = 400$, le paramètre de la moyenne μ est quant à lui inconnu.
Les mesures de la durée de vie d'un lot de 9 ampoules ont donné les résultats suivants :

$$2000; 1890; 3180; 1990; 2563; 2876; 3098; 2413; 2596.$$

a) Déterminer un intervalle de confiance pour la durée de vie moyenne d'une ampoule au niveau 90%
b) Peut-on affirmer, avec un risque d'erreur de 10% que la durée de vie moyenne d'une ampoule est égale à 2500 heures ?

Exercice 2. Une usine spécialisée dans la construction de câble souhaite vérifier la fiabilité de ses produits en évaluant la masse maximale que ses câbles peuvent supporter.
Pour cela, on modélise la masse maximale, en tonnes, supportée par un câble par une variable aléatoire X suivant une loi Normale de moyenne inconnue μ et d'écart-type $\sigma = 0.5$. Une étude a été effectuée sur un échantillon de 50 câbles. Il en ressort, en moyenne, que la charge maximale supportée par un câble est de 12.2 tonnes.

a) Déterminer un intervalle de confiance pour μ au niveau 0.99.
b) Peut-on affirmer que la machine, avec un risque d'erreur de 1% produit bien des câbles capables de supporter une masse d'au moins 11.5 tonnes ?
c) Déterminer la taille minimale de l'échantillon étudié pour que la longueur de l'intervalle de confiance au niveau 99% soit inférieure à 0.2.

3 Concernant le troisième module

Dans le précédent module, nous avons construit des intervalles de confiance pour l'estimateur de la moyenne $\bar{X}$ lorsque la variance σ^2 de la distribution de nos données (ou échantillons) était **connue**.

Nous avons alors vu que notre estimateur de la moyenne, lorsque la taille de l'échantillon est assez grande (> 30), était distribué normalement, et nous pouvions construire des intervalles de confiance de la forme

$$\left[\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

où, pour rappel, si $Z \sim \mathcal{N}(0,1)$, alors z_α est le nombre (ou quantile) vérifiant

$$\mathbb{P}[Z \leq z_\alpha] = \alpha.$$

Dans ce module, nous intéressons au cas où la variance de la distribution de nos données est **inconnue**.

3.1 Quelques rappels

Estimateurs. Nous cherchons à estimer des intervalles de confiance pour la valeur moyenne d'une population μ lorsque l'écart-type de la population σ est **inconnu**.
Les estimateurs de la moyenne et de l'écart-type sur un échantillon sont respectivement définis par

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Loi de Student A partir de ces estimateurs, on va **admettre** que la variable aléatoire T_{n-1} définie par

$$T_{n-1} = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}}$$

suit **une loi de Student à $n-1$ degrés de liberté**, cette loi s'exprime comme le quotient d'une loi normale centrée réduite par la racine carrée d'une loi du $\mathcal{X}^2$ (lire Khi-deux) à $n-1$ degré de liberté divisée par le nombre de degrés de liberté $n-1$.

La loi de Student, représentée en Figure 5, possède les propriétés suivantes :

- elle est symétrique par rapport à 0,
- $\mathbb{E}[T_{n-1}] = 0$ lorsque $n > 1$, et elle possède une espérance de forme *indéterminée* lorsque $n = 1$,
- $Var(X) = \dfrac{n}{n-2}$ lorsque $n > 2$.
- plus le nombre de degrés de liberté est important, plus la variance est faible.

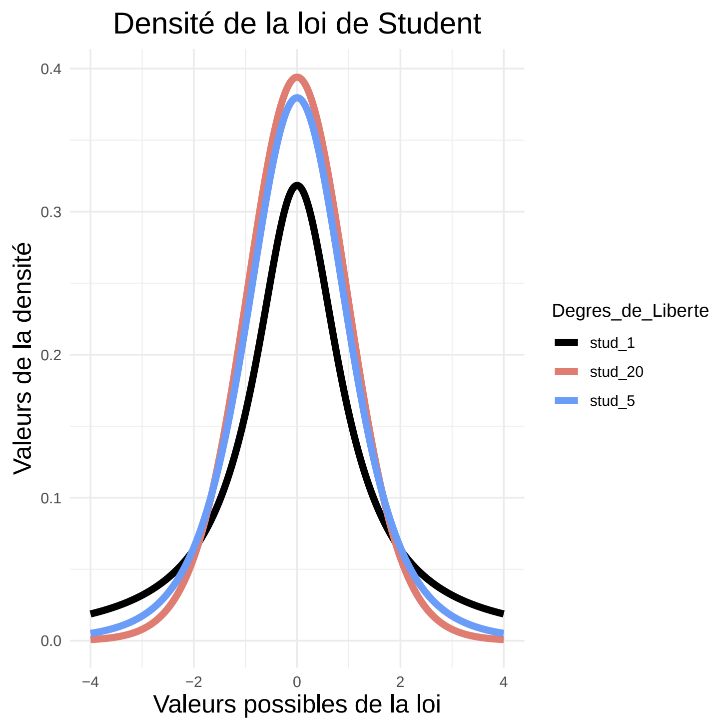

FIGURE 5 – Représentation de la loi de Student lorsque l'on fait varier le nombre de degrés de liberté

Les premier et dernier points montrent que lorsque n est assez grand, on peut approximer la loi de Student par une loi Normale Centrée Réduite.

Regardons maintenant comment construire des intervalles de confiance de niveau $1-\alpha$ pour notre moyenne.

Intervalle de confiance. Le processus de construction est le même que pour celui de la loi normale. D'après, ce que nous avons vu précédemment, nous pouvons affirmer qu'une proportion $1-\alpha$ des valeurs de la loi de Student se trouve dans l'intervalle $\left[t_{\alpha/2}; t_{1-\alpha/2}\right]$ ce qui veux dire que si T_{n-1} suit une loi de Student à $n-1$ degrés de liberté, alors

$$\mathbb{P}[t_{\alpha/2} \leq T_{n-1} \leq t_{1-\alpha/2}] = 1 - \frac{\alpha}{2}.$$

Remarques : en toute rigueur, je devrais noter $t_{n-1,1-\alpha/2}$ au lieu de $t_{1-\alpha/2}$. Je laisse cependant le soin au lecteur, à l'aide du contexte, de trouver lui même la valeur de n correspondante.
On prendra aussi garde au fait que **si l'on dispose d'un échantillon de taille n alors on est amené à considérer une loi de Student à $n-1$ degrés de liberté !**

En repartant de l'inégalité précédente, nous avons donc

$$\mathbb{P}\left[t_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}} \leq t_{1-\alpha/2}\right] = 1 - \frac{\alpha}{2}.$$

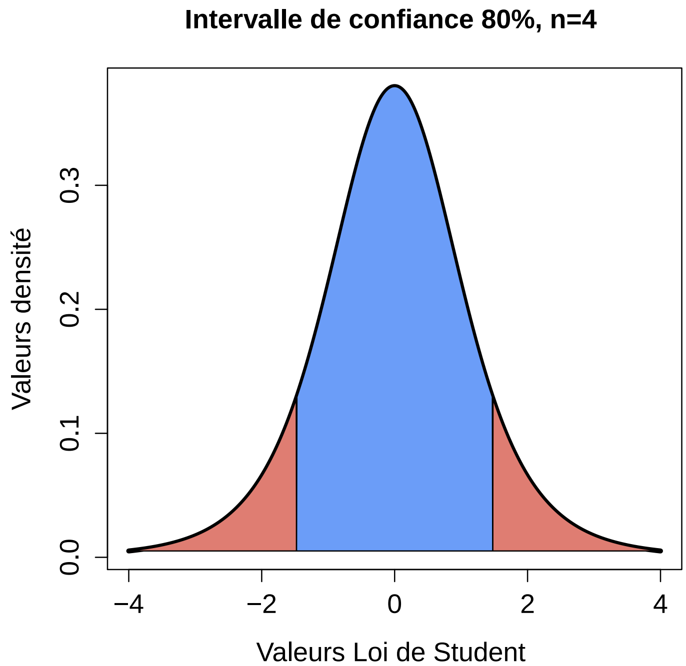

FIGURE 6 – Représentation de l'intervalle de confiance en bleu pour la loi de Student avec 4 degrés de liberté et pour un niveau de confiance $1-\alpha = 0.8$. On tire à nouveau profit de la symétrie pour répartir l'erreur de façon équitable à *gauche* et à *droite* de la distribution.

Ce qui veux dire qu'une proportion $1-\alpha$ des valeurs de la moyenne estimée à l'aide d'un échantillon se trouveront dans l'intervalle

$$\left[\mu + t_{\alpha/2}\sqrt{\frac{s^2}{n}} \leq \bar{x} \leq \mu + t_{1-\alpha/2}\sqrt{\frac{s^2}{n}}\right].$$

Mais cela signifie aussi qu'il y a une probabilité de $1-\alpha$ que la valeur μ se trouve dans l'intervalle

$$\left[\bar{x} + t_{\alpha/2}\sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}\sqrt{\frac{s^2}{n}}\right],$$

ce qui peut se montrer très facilement.
Les valeurs des quantiles de la loi de Student se lisent dans la table de cette loi et que vous pouvez trouver en Figure 8 en annexe de ce document. On représente également un exemple d'intervalle de confiance de niveau $1-\alpha = 0.8$ en Figure 6.

La méthode de construction d'un intervalle de confiance restant la même, je ne détaille pas la procédure dans cette section et je vous renvoie à la section précédente.

Exemple

3.2 Exercice

Énoncé. Une société de vente à distance de matériel informatique s'intéresse au nombre journalier de connexions sur son site internet. Sur une période de 10 jours, les nombres suivants ont été relevés :

$$759, 750, 755, 756, 761, 765, 770, 752, 760, 767.$$

On suppose que ces résultats sont indépendants et identiquement distribués selon une loi normale $\mathcal{N}(\mu, \sigma^2)$ dont les paramètres sont inconnus.
On admettra que, sur cet échantillon, $\bar{x} = 759.5$ et $s^2 = 46.73$.

a) Construire un intervalle de confiance pour μ avec les niveaux de confiance 0.90 et 0.99.
b) Quel niveau de confiance choisir pour avoir un intervalle de confiance deux fois plus étroit que celui obtenu avec une confiance de 0.9 ?
c) Sur combien de jours aurait-on dû relever le nombre de connexions pour que la longueur de l'intervalle de confiance, de niveau 95%, n'excède pas 1 (en supposant que les estimations de la moyenne et de la variance ne changent pas).

Correction. On note X la variable aléatoire représentant le nombre journaliers de connexions sur le site internet. Nous sommes dans le cas où la variance de notre distribution est inconnue, dans ce cas, la variable aléatoire T définie par

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} \simeq \mathcal{T}_{n-1},$$

i.e. peut être approximée par une loi de Student à $n-1$ degrés de liberté.
Dans ce cas, on a

Dans ce cas, un encadrement de la moyenne μ au niveau de confiance $1 - \alpha$, nous est donné par

$$\left[\bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

et vérifie

$$\mathbb{P}\left[\bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}\right] = 1 - \alpha.$$

a) On cherche un construire un intervalle de confiance de niveau 0.90, dans ce cas la marge d'erreur α est égale à 0.1. On doit donc chercher dans la table de la loi de Student, le quantile $t_{1-\alpha/2}$ d'ordre $1 - \alpha/2 = 0.95$, lorsque le nombre nombre de degrés de liberté est égal à $n - 1 = 9$. On a donc $t_{1-\alpha/2} = -t_{\alpha/2} = 1.833$ et notre intervalle de confiance $I_{\alpha=0.90}$ est donc, après calculs

$$\begin{aligned} I_{\alpha=0.90} &= \left[759.5 - 1.833\sqrt{\frac{46.73}{10}}; 759.5 + 1.833\sqrt{\frac{46.73}{10}}\right], \\ &= [755.538; 763.462]. \end{aligned}$$

Au niveau de confiance 0.99, on reprend exactement les même étapes que précédemment, à la seule différence que nous avons cette fois-ci $t_{1-\alpha/2} = 3.250$, ce qui nous donne un intervalle de confiance $I_{\alpha=0.90}$ égal à

$$I_{\alpha=0.99} = \left[759.5 - 3.25\sqrt{\frac{46.73}{10}}; 759.5 + 3.25\sqrt{\frac{46.73}{10}}\right],$$

$$= [752.47; 766.53] .$$

b) Pour avoir un niveau un intervalle de confiance deux fois plus étroits, il faut que la valeur de $t_{1-\alpha/2}$, pour ce nouveau niveau de confiance, soit deux fois plus petite que celle obtenu au niveau de confiance $1-\alpha = 0.9$, *i.e.* on cherche α telle que $t_{1-\alpha/2} = 1.833/2 = 0.9165$. Il nous faut maintenant chercher cette valeur dans la table en se concentrant sur la ligne correspondant à 9 degrés de liberté. On remarque que la valeur se trouve entre 0.75 et 0.9, pour déterminer une valeur, on va utiliser une interpolation linéaire :

$$1-\alpha/2 = 0.75 + \beta(0.9 - 0.75) \quad \text{où} \quad \beta = \frac{0.9165 - 0.703}{1.383 - 0.703}.$$

On obtient donc $1 - \alpha/2 = 0.797$ et on déduit donc que $\alpha = 0.41$, ce qui nous donne un niveau de confiance de 0.59.

c) On rappelle que la longueur de notre intervalle de confiance est égale à

$$2t_{1-\alpha/2}\frac{s}{\sqrt{n}}.$$

Pour un niveau de confiance égal à 0.95 et étant données les valeurs de la table de Student, on va considérer que $t_{1-\alpha/2} = 2$, (on pourrait prendre 1.96, comme approximation car n va être assez grand). A partir de cette valeur, il nous reste alors à déterminer n telle que

$$\begin{aligned} 2 \times 2 \times \frac{\sqrt{46.73}}{\sqrt{n}} &< 1. \\ 4 \times \sqrt{46.73} &< \sqrt{n}. \\ n &> 16 \times 46.73, \\ n &> 747.68. \end{aligned}$$

3.3 Pour s'entraîner

Exercice 1. Une entreprise fabrique des composants électroniques dont la durée de vie , exprimée en heure, est modélisée par une variable aléatoire X suivant une loi Normale.
Une série de 50 mesures ont été effectuées et ont donné la moyenne et l'écart-type suivants :

$$\bar{x} = 1200 \quad \text{et} \quad s = 200.$$

a) Donnez un intervalle de confiance de niveau 0.95 de cette durée de vie moyenne.
b) Quelle doit être la taille de l'échantillon pour que l'intervalle de confiance de niveau 0.95 de la durée de vie moyenne des composants ait une amplitude de 60 heures ?

Exercice 2. Une biologiste étudie un type d'algue qui attaque les plantes marines. La toxine contenue dans cette algue est obtenue sous forme d'une solution organique. Le biologiste mesure la quantité de toxine par gramme de solution qui est modélisée par une variable aléatoire X dont l'espérance μ et la variance σ^2 sont inconnues.
Il a obtenu les neuf mesures suivantes, exprimées en milligrammes :

$$1.2; 0.8; 0.6; 1.1; 1.2; 0.9; 1.5; 0.9; 1.0$$

a) Donnez un intervalle de confiance de niveau 0.90 de quantité de toxine..
b) Peut-on dire, avec un niveau de confiance de 0.80 que la quantité de toxine par gramme de solution μ est égale à 1.3 mg.

4 Concernant le quatrième module

4.1 Quelques rappels

Exemple. Le pourcentage obtenu à une question par une enquête est de 20%. On souhaite déterminer une marge d'erreur de 2.5% (on rappelle que la marge d'erreur est la moitié de la longueur de notre intervalle de confiance) dans 95% des cas. Quelle taille doit alors prendre notre échantillon ? Quelle serait notre marge d'erreur si la taille de notre échantillon était égale à 100 ?

Pour la première question de cet exemple, on se rappelle que la marge d'erreur est égale à

$$z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}.$$

On souhaite un intervalle de confiance de 0.95, on a donc $\alpha = 0.05$, on cherche donc la valeur de $z_{1-\alpha/2=0.975}$ dans la table de la loi normale centrée réduite et on trouve (ou on se souvient) que cette valeur est 1.96. Or on on souhaite :

$$\begin{aligned} z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} &= 0.025, \\ 1.96 \times \sqrt{\frac{0.2(1-0.2)}{n}} &= 0.025. \end{aligned}$$

On doit donc avoir

$$n = \frac{01.96\sqrt{0.2 \times 0.8}}{0.025} = 983.3,$$

ce qui signifie que n doit au moins être égal à 984.

Pour la deuxième question, nous repartons de la définition de la marge d'erreur et nous effectuons simplement le calcul en supposant que les proportions restent inchangées, ce qui nous donne

$$1.96 \times \sqrt{\frac{0.2 \times 0.8}{100}} = 0.0784.$$

Nous avons donc une marge d'erreur de 0.0784, soit un intervalle de confiance pour l'estimation de la proportion p de la forme

$$[0.2 - 0.0784; 0.2 + 0.0784] = [0.1216; 0.2784].$$

4.2 Exercices

Énoncé Lors d'un sondage précédant les élection présidentielles, 500 personnes ont été interrogées. Bien que ce ne soit pas le cas en pratique, on suppose, pour simplifier les calculs, que les 500 personnes constituent un échantillon indépendant et identiquement distribué (on note souvent cela *i.i.d.* de la population française. Sur les 500 personnes, 150 ont répondu vouloir vouloir voter pour le candidat C_1 et 140 pour le candidat C_2.

a) Donner une estimation ponctuelle des intentions de vote pour chaque candidat, sous la forme d'un pourcentage.
b) Donner un intervalle de confiance à 95% pour chacun des deux intentions de votes.
c) Peut-on prédire, qui de C_1 ou C_2 sera élu ?

Correction.

a) Les intentions de vote pour le candidat sont représentées par des variables aléatoires X_I suivant une loi de Bernoulli de paramètre $p1$, *i.e.* $\mathcal{B}(p_1)$. On suppose que les votes sont indépendants les uns des autres. L'ensemble des intentions de vote, $S_1 = X_1+X_2+...+X_n$; pour le candidat C_1, suit donc une loi binomiale $\mathcal{B}(n, p_1)$.
On rappelle que l'espérance d'une telle loi est égale à np_1. En terme de proportion, nous avons donc
$$\mathbb{E}\left[\frac{S_1}{n}\right] = p_1,$$
on choisit donc la proportion des gens ayant voté pour le candidat C_1 comme estimateur de p_1 et cet estimateur est égal à $f_1 = 150/500 = 0.3$. On peut faire de même pour le candidat C_2 pour lequel un estimateur de p_2 est donné par $f_2 = 140/500 = 0.28$.

b) On rappelle que notre estimateur de la moyenne (proportion) $F_1\ \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right)$. On peut donc obtenir un intervalle de confiance en commençant par centrer et réduire notre variable aléatoire F_1, ce qui nous donne
$$\mathbb{P}\left[z_{\alpha/2} \leq \frac{F_1 - p_1}{\sqrt{\frac{f_1(1-f_1)}{n}}} \leq z_{1-\alpha/2}\right] = 1 - \alpha,$$
où on aura pris soin d'utiliser notre estimateur de la variance de la variable aléatoire F_1. Ainsi pour une réalisation de la variable aléatoire $F_1(\omega) = f_1$, *i.e.* pour une proportion f_1 mesurée sur un échantillon, on a l'encadrement suivant de notre proportion théorique p_1 :
$$\mathbb{P}\left[f_1 + z_{\alpha/2}\sqrt{\frac{f_1(1-f_1)}{n}} \leq p_1 \leq f_1 + z_{1-\alpha/2}\sqrt{\frac{f_1(1-f_1)}{n}}\right] = 1 - \alpha.$$
On peut maintenant construire les intervalles de confiance pour chacune des deux intentions de votes :

- Pour le candidat C_1 avec un score de confiance égale à 0.95, on a

$$\begin{aligned} I_{\alpha=0.05} &= \left[f_1 + z_{\alpha/2}\sqrt{\frac{f_1(1-f_1)}{n}}; f_1 + z_{1-\alpha/2}\sqrt{\frac{f_1(1-f_1)}{n}}\right], \\ &= \left[0.3 - 1.96 \times \sqrt{\frac{0.3\times(1-0.3)}{500}}; 0.3 + 1.96 \times \sqrt{\frac{0.3\times(1-0.3)}{500}}\right], \\ &= [-0.2598; 0.3402]. \end{aligned}$$

- De même pour le candidat C_2, on a

$$
\begin{aligned}
I_{\alpha=0.05} &= \left[f_2 + z_{\alpha/2}\sqrt{\frac{f_2(1-f_2)}{n}}; f_2 + z_{1-\alpha/2}\sqrt{\frac{f_2(1-f_2)}{n}}\right], \\
&= \left[0.28 - 1.96 \times \sqrt{\frac{0.28 \times (1-0.28)}{500}}; 0.28 + 1.96 \times \sqrt{\frac{0.28 \times (1-0.28)}{500}}\right], \\
&= [-0.2406; 0.3193].
\end{aligned}
$$

c) Il sera très difficile de prévoir qui va gagner entre le candidat C_1 et le candidat C_2, les intervalles de confiance se chevauchent beaucoup trop pour que l'on puisse tirer une conclusion.
Remarque : on verra plus tard, avec la théorie des tests, que l'on pourra finir une réponse plus précise à cette question.

4.3 Pour s'entraîner

Exercice. La société SOAP veut estimer la taille du marché potentiel d'un nouveau produit de soin pour le corps auprès d'un public de femmes (c'est bien connu, les hommes ne prennent pas soin d'eux !)
Un sondage est effectué auprès de 40 femmes et 24 se disent satisfaites de ce nouveau produit.

a) La société déclare que plus d'une femme sur deux, avec un niveau de confiance de 90%, est satisfaite de son produit. Que penser de cette affirmation ?
b) Suite aux protestations des associations de consommateurs pour publicité mensongère, l'entreprise commande un second sondage auprès de 225 femmes, 150 se déclarent alors satisfaites du produit. L'entreprise doit-elle changer sa stratégie de communication suite à ce nouveau sondage ? On résonne toujours avec un intervalle de confiance de niveau 0.90 ?
c) L'entreprise souhaite à présent connaître plus précisément le nombre de femmes satisfaites par son produit, elle souhaite obtenir une proportion dans une fourchette de 5%

A Annexes au cours

Cette annexe regroupe des résultats qui sont utilisés en cours sans en donner un énoncé explicite. On présente également quelques lois de probabilités ainsi que leur propriétés ou encore les tables des lois classiques : *Loi normale centrée réduite*, *Loi de Student*, *Loi de Fisher* et *Loi du* $\mathcal{X}^2$ Tout ce qui figure dans cette annexe n'est pas à savoir dans le cadre de ce cours.

A.1 Fonctions de probabilités

Loi Binomiale $\mathcal{B}(n,p)$. La loi binomiale est la répétition de n expériences successives, identiques et indépendantes dont on dénombre seulement deux issues possibles. On note $p \in [0,1]$ la probabilité d'obtenir l'issue *favorable* ou la probabilité d'avoir un *succès*.

La fonction de probabilité de la loi binomiale est définie par

$$\mathbb{P}[X=k] = \binom{n}{k} p^k (1-p)^{n-k}.$$

Cette loi admet une espérance et une variance qui sont respectivement

$$\mathbb{E}[X] = np \quad \text{et} \quad Var(X) = np(1-p).$$

Loi Normale $\mathcal{N}(\mu,\sigma^2)$. La loi normale est caractérisée par sa moyenne (ou espérance) μ et par sa variance σ^2. Cela veut dire que la seule connaissance de ces deux paramètres permet de caractériser intégralement cette loi. Elle admet une densité f définie pour tout nombre réel $x \in \mathbb{R}$ par

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

Cette loi admet une espérance et une variance qui sont respectivement

$$\mathbb{E}[X] = \mu \quad \text{et} \quad Var(X) = \sigma^2.$$

La loi normale est caractérisée par sa symétrie autour de la valeur moyenne, i.e. pour tout réel t, on a :

$$\mathbb{P}[X \leq \mu - t] = \mathbb{P}[X \geq \mu + t].$$

Finalement, pour la loi normale, **la moyenne** est égale à **la médiane** qui est au égale **au mode**.

Loi du Khi-deux $\mathcal{X}_k^2$. Soient $X_1, X_2, ..., X_k$, k variables aléatoires indépendantes suivant une loi normale centrée réduite. Alors la la variable aléatoire X définie par $X = \sum_{i=1}^k X_i$ suit une loi du $\mathcal{X}^2$ à k degrés de liberté, notée $\mathcal{X}_k^2$.
La densité f de la variable aléatoire X est définie, pour tout $x \geq 0$, par

$$f(x,k) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)},$$

où

$$\begin{aligned} \Gamma : \mathbb{R}_+^\star &\to \mathbb{R}_+^\star, \\ x &\mapsto \int_0^{+\infty} t^{x-1} e^{-t} dt. \end{aligned}$$

Cette loi admet une espérance et une variance qui sont respectivement

$$\mathbb{E}[X] = k \quad \text{et} \quad Var(X) = 2k.$$

Loi de Student $\mathcal{T}_k$. Considérons X une variable aléatoire centrée réduite et U une variable aléatoire suivant une loi du $\mathcal{X}_n^2$, *i.e.* du Khi-deux à k degrés de libertés, indépendantes. Alors la variable aléatoire $T = \dfrac{X}{\sqrt{U/k}}$ suit une loi de Student à k degrés de liberté.
La densité f de la variable aléatoire T est définie par

$$f(x, k) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{(k+1)/2},$$

pour tout $k > 0$.

Cette loi admet une espérance lorsque $k > 1$ et une variance lorsque $k > 2$ qui sont respectivement égales à

$$\mathbb{E}[X] = 0 \quad \text{et} \quad Var(X) = \frac{k}{k-2}.$$

Loi de Fisher $\mathcal{F}_{k_1,k_2}$. Soient U_1 et U_2 deux variables aléatoires indépendantes suivant une loi du Khi-deux à respectivement k_1 et k_2 degrés de liberté. Alors la variable aléatoire $X = \dfrac{U_1/k_1}{U_2/d_2}$ suit une loi de Fisher, notée $F(k_1, k_2)$ (à k_1 et k_2 degrés de liberté).
La densité f de la variable aléatoire F est définie, pour tout $x \geq 0$, $k_1, k_2 > 0$, par

$$f(x, k_1, k_2) = \frac{1}{B\left(\frac{k_1}{2}, \frac{k_2}{2}\right)} \left(\frac{k_1}{k_2}\right)^{(k_1/2)} x^{(k_1/2)-1} \left(1 + \frac{k_1}{k_2} x\right)^{-(k_1+k_2)/2},$$

où

$$\begin{aligned} B : \mathbb{R}_+^\star \times \mathbb{R}_+^\star &\to \mathbb{R}_+^\star, \\ (x, y) &\mapsto \int_0^1 t^{x-1} (1-t)^{y-1} dt. \end{aligned}$$

Lorsque $k_2 > 2$, cette loi admet une espérance égale à

$$\mathbb{E}[F] = \frac{k_2}{k_2 - 2}.$$

Si $k_2 > 4$, alors elle admet également une variance égale à

$$Var(F) = \frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}.$$

A.2 Quelques résultats en probabilités

Théorème Central Limite. Soit $X_1, X_2, ..., X_n$ une suite de variables aléatoires réelles définies sur un espace probabilisé, **indépendantes** et **identiquement distribuées** suivant une même loi D **admettant une espérance (ou moyenne)** μ **et un écart-type** σ **non nul**.
Soit $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, alors :

$$Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n\to\infty]{loi} Z \sim \mathcal{N}(0,1),$$

i.e. la loi de la variable aléatoire Z_n telle que définie ci-dessus converge vers la loi normale centrée réduite.

A.3 Tables des Lois

La dernière partie de cette annexe regroupe quelques tables de lois qui seront utilisés dans ce cours. Le choix des tables présentés est bien évidemment non exhaustif et pourrait largement être complété.

$u_\alpha = \Phi^{-1}(\alpha)$ où $\alpha = \alpha_1 + \alpha_2$										
	α_2									
α_1	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.500	0.0000	0.0025	0.0050	0.0075	0.0100	0.0125	0.0150	0.0175	0.0201	0.0226
0.510	0.0251	0.0276	0.0301	0.0326	0.0351	0.0376	0.0401	0.0426	0.0451	0.0476
0.520	0.0502	0.0527	0.0552	0.0577	0.0602	0.0627	0.0652	0.0677	0.0702	0.0728
0.530	0.0753	0.0778	0.0803	0.0828	0.0853	0.0878	0.0904	0.0929	0.0954	0.0979
0.540	0.1004	0.1030	0.1055	0.1080	0.1105	0.1130	0.1156	0.1181	0.1206	0.1231
0.550	0.1257	0.1282	0.1307	0.1332	0.1358	0.1383	0.1408	0.1434	0.1459	0.1484
0.560	0.1510	0.1535	0.1560	0.1586	0.1611	0.1637	0.1662	0.1687	0.1713	0.1738
0.570	0.1764	0.1789	0.1815	0.1840	0.1866	0.1891	0.1917	0.1942	0.1968	0.1993
0.580	0.2019	0.2045	0.2070	0.2096	0.2121	0.2147	0.2173	0.2198	0.2224	0.2250
0.590	0.2275	0.2301	0.2327	0.2353	0.2378	0.2404	0.2430	0.2456	0.2482	0.2508
0.600	0.2533	0.2559	0.2585	0.2611	0.2637	0.2663	0.2689	0.2715	0.2741	0.2767
0.610	0.2793	0.2819	0.2845	0.2871	0.2898	0.2924	0.2950	0.2976	0.3002	0.3029
0.620	0.3055	0.3081	0.3107	0.3134	0.3160	0.3186	0.3213	0.3239	0.3266	0.3292
0.630	0.3319	0.3345	0.3372	0.3398	0.3425	0.3451	0.3478	0.3505	0.3531	0.3558
0.640	0.3585	0.3611	0.3638	0.3665	0.3692	0.3719	0.3745	0.3772	0.3799	0.3826
0.650	0.3853	0.3880	0.3907	0.3934	0.3961	0.3989	0.4016	0.4043	0.4070	0.4097
0.660	0.4125	0.4152	0.4179	0.4207	0.4234	0.4261	0.4289	0.4316	0.4344	0.4372
0.670	0.4399	0.4427	0.4454	0.4482	0.4510	0.4538	0.4565	0.4593	0.4621	0.4649
0.680	0.4677	0.4705	0.4733	0.4761	0.4789	0.4817	0.4845	0.4874	0.4902	0.4930
0.690	0.4959	0.4987	0.5015	0.5044	0.5072	0.5101	0.5129	0.5158	0.5187	0.5215
0.700	0.5244	0.5273	0.5302	0.5330	0.5359	0.5388	0.5417	0.5446	0.5476	0.5505
0.710	0.5534	0.5563	0.5592	0.5622	0.5651	0.5681	0.5710	0.5740	0.5769	0.5799
0.720	0.5828	0.5858	0.5888	0.5918	0.5948	0.5978	0.6008	0.6038	0.6068	0.6098
0.730	0.6128	0.6158	0.6189	0.6219	0.6250	0.6280	0.6311	0.6341	0.6372	0.6403
0.740	0.6433	0.6464	0.6495	0.6526	0.6557	0.6588	0.6620	0.6651	0.6682	0.6713
0.750	0.6745	0.6776	0.6808	0.6840	0.6871	0.6903	0.6935	0.6967	0.6999	0.7031
0.760	0.7063	0.7095	0.7128	0.7160	0.7192	0.7225	0.7257	0.7290	0.7323	0.7356
0.770	0.7388	0.7421	0.7454	0.7488	0.7521	0.7554	0.7588	0.7621	0.7655	0.7688
0.780	0.7722	0.7756	0.7790	0.7824	0.7858	0.7892	0.7926	0.7961	0.7995	0.8030
0.790	0.8064	0.8099	0.8134	0.8169	0.8204	0.8239	0.8274	0.8310	0.8345	0.8381
0.800	0.8416	0.8452	0.8488	0.8524	0.8560	0.8596	0.8633	0.8669	0.8705	0.8742
0.810	0.8779	0.8816	0.8853	0.8890	0.8927	0.8965	0.9002	0.9040	0.9078	0.9116
0.820	0.9154	0.9192	0.9230	0.9269	0.9307	0.9346	0.9385	0.9424	0.9463	0.9502
0.830	0.9542	0.9581	0.9621	0.9661	0.9701	0.9741	0.9782	0.9822	0.9863	0.9904
0.840	0.9945	0.9986	1.0027	1.0069	1.0110	1.0152	1.0194	1.0237	1.0279	1.0322
0.850	1.0364	1.0407	1.0450	1.0494	1.0537	1.0581	1.0625	1.0669	1.0714	1.0758
0.860	1.0803	1.0848	1.0893	1.0939	1.0985	1.1031	1.1077	1.1123	1.1170	1.1217
0.870	1.1264	1.1311	1.1359	1.1407	1.1455	1.1503	1.1552	1.1601	1.1650	1.1700
0.880	1.1750	1.1800	1.1850	1.1901	1.1952	1.2004	1.2055	1.2107	1.2160	1.2212
0.890	1.2265	1.2319	1.2372	1.2426	1.2481	1.2536	1.2591	1.2646	1.2702	1.2759
0.900	1.2816	1.2873	1.2930	1.2988	1.3047	1.3106	1.3165	1.3225	1.3285	1.3346
0.910	1.3408	1.3469	1.3532	1.3595	1.3658	1.3722	1.3787	1.3852	1.3917	1.3984
0.920	1.4051	1.4118	1.4187	1.4255	1.4325	1.4395	1.4466	1.4538	1.4611	1.4684
0.930	1.4758	1.4833	1.4909	1.4985	1.5063	1.5141	1.5220	1.5301	1.5382	1.5464
0.940	1.5548	1.5632	1.5718	1.5805	1.5893	1.5982	1.6072	1.6164	1.6258	1.6352
0.950	1.6449	1.6546	1.6646	1.6747	1.6849	1.6954	1.7060	1.7169	1.7279	1.7392
0.960	1.7507	1.7624	1.7744	1.7866	1.7991	1.8119	1.8250	1.8384	1.8522	1.8663
0.970	1.8808	1.8957	1.9110	1.9268	1.9431	1.9600	1.9774	1.9954	2.0141	2.0335
0.980	2.0537	2.0749	2.0969	2.1201	2.1444	2.1701	2.1973	2.2262	2.2571	2.2904
0.990	2.3263	2.3656	2.4089	2.4573	2.5121	2.5758	2.6521	2.7478	2.8782	3.0902

FIGURE 7 – Table des quantiles de la loi Normale Centrée Réduite $\mathcal{N}(0,1)$

$t_{\nu,\alpha}$								
	α							
ν	0.6	0.75	0.9	0.95	0.975	0.99	0.995	0.9995
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	3.373
1000	0.253	0.675	1.282	1.646	1.962	2.330	2.581	3.300

FIGURE 8 – Table des quantiles de la loi de Student $\mathcal{T}_k$