Mathématiques et Statistiques appliquées à la Gestion

Compléments de cours et Exercices BBA-1 (2021-2022)

Guillaume Metzler

Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC EA3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Résumé

L'objectif de ce cours est de présenter le principe de l'**inférence statistique**, i.e. comment déduire des informations sur une population à partir de la seul connaissance d'un échantillon.

Dans un premier temps, nous verrons comment, à partir des grandeurs estimer sur un échantillon, avoir des garanties sur ces estimations à l'échelle de notre population. Pour répondre à cet objectif, nous commençons par rappeler la notion d'**estimation ponctuelle** et nous verrons comment, à l'aide de la notion de fluctuations d'échantillonnages, passer d'un estimateur ponctuel à un estimateur qui est une variable aléatoire suivant une certaine loi et ainsi **construire des intervalles de confiance** pour nos paramètres inconnus. Ainsi, nous verrons comment construire de tels intervalles et donc faire de telles estimations dans différents cas ou et pour différents paramètres. La deuxième partie de ce cours est consacrée aux tests statistiques pour laquelle l'usage des intervalles de confiance précédemment construits va se révéler fort utile.

Le document propose un résumé de chaque séance ainsi que des exemples et exercices corrigés pour vous aider à pratiquer et assimiler les notions et méthodes abordées en cours. Il propose également des exercices non corrigés si vous souhaitez vous entrainer d'avantage. Ce document se veut évolutif afin de répondre à vos questions mais aussi pour corriger les coquilles ou le manque de précision dans certaines parties.

Table des matières

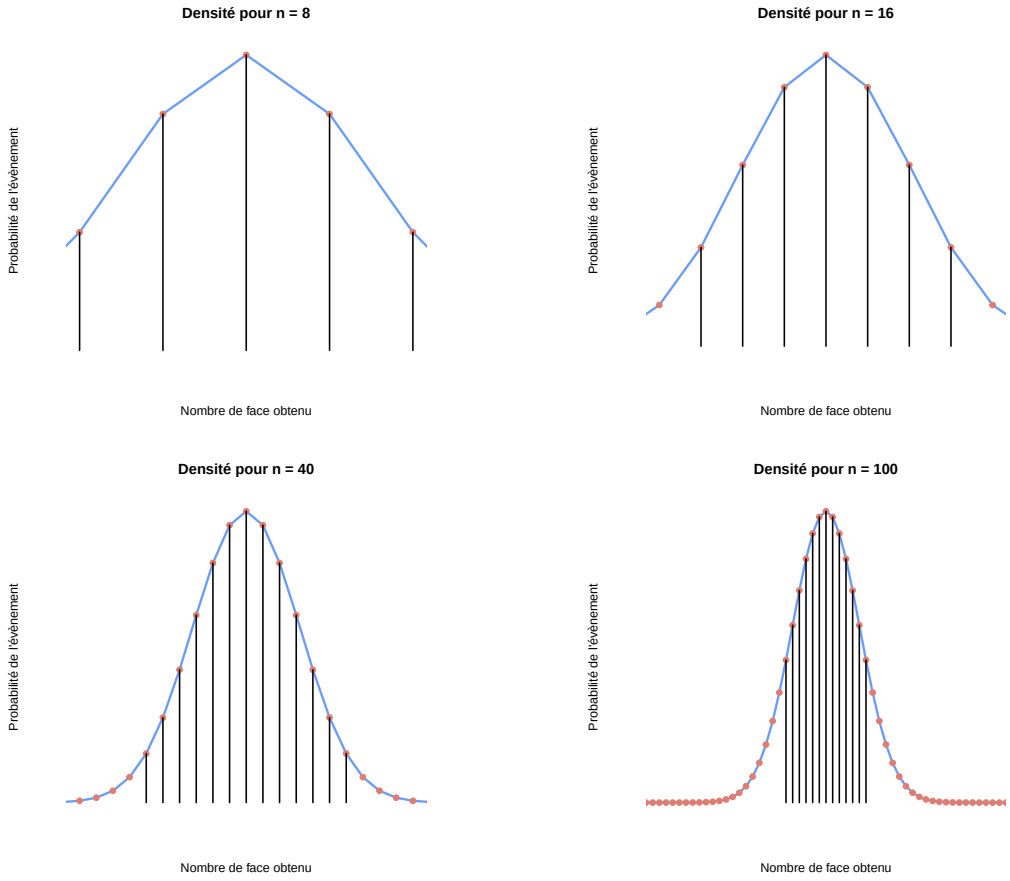

FIGURE 1 – Illustration de la densité de la loi binomiale de paramètres (n, p) pour différentes valeurs de n : $8, 16, 40, 100$

1 Estimation ponctuelle : Loi Binomiale et Loi Normale

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 $\sigma^2 = np(1-p)$.

En effet, prenons un exemple simple avec lancer de pièce non truquée et on note p cette probabilité, on a donc $p = 0.5$. On note X la variable aléatoire associé à l'évènement : *obtenir face*, alors $\mathbb{P}[X = 1] = 0.5 = p$. La probabilité d'obtenir pile est la même.

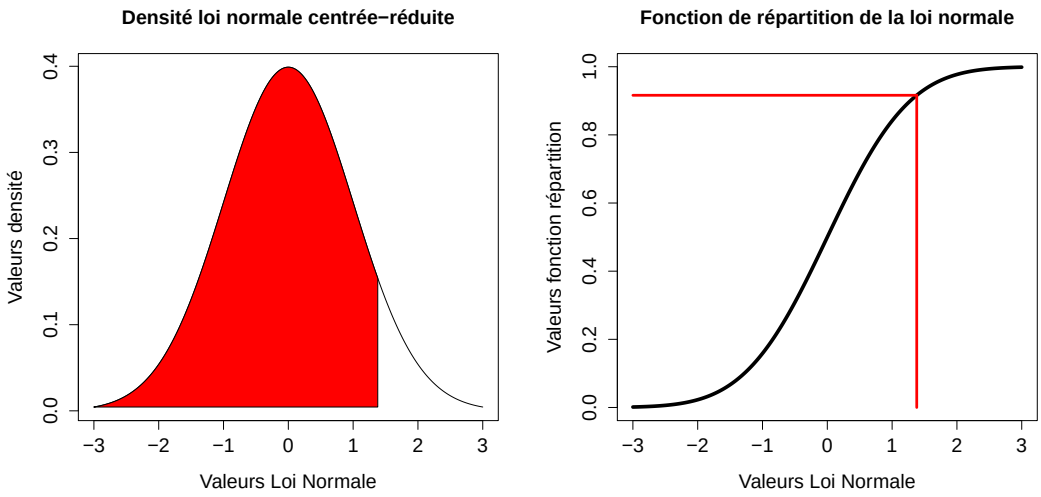

FIGURE 2 – Exemple représentant l'aire sous la courbe (en rouge) d'une loi normale centrée réduite pour $t = 1.38$ (valeur de t à lire sur l'axe des abscisses), à gauche. La figure de droite représente la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0, 1)$. L'axe des ordonnées renseigne directement sur la valeur de la fonction de répartition pour la valeur mentionnée en abscisse.

Considérons maintenant X_n la variable aléatoire consistant à répéter ce lancer de pièce n fois. X_n suit donc une loi binomiale de paramètres (n, p). La Figure 1 illustre la densité de la loi binomiale en fonction de la valeur de n. On remarque que plus la valeur de n augmente, plus **la densité de la loi binomiale** se rapproche de **la densité de la loi normale**.

Enfin, nous finissons par 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 2.

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 2 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*.
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 3.

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.

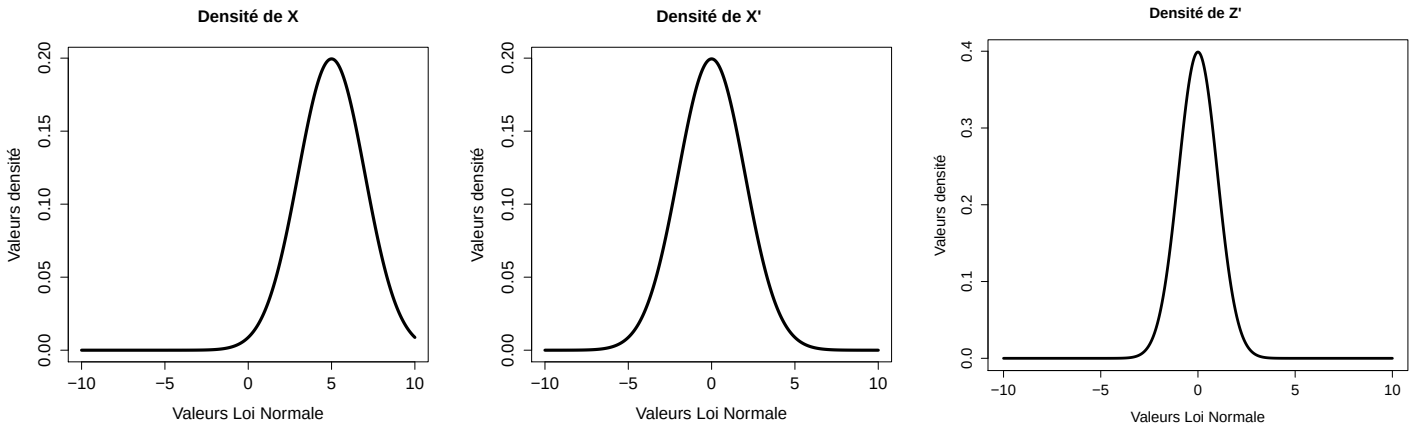

FIGURE 3 – Figures illustrant (de gauche à droite) les étapes de la normalisation. La première figure de gauche montre notre distribution (ou densité) initiale. La figure du milieu montre le recentrage de la gaussienne en 0. La figure de droite montre l'ajustement du facteur d'échelle (la réduction à un écart-type égal à 1) de notre distribution.

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

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

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 A retenir

En théorie et en pratique

Deux types de variables aléatoires :

- **Les lois discrètes :** comme la loi Binomiale $\mathcal{B}(n,p)$.
  Dans ce cas on peut calculer la probabilité qu'une telle variable aléatoire prenne une valeur précise.
- **Les lois continues :** comme la Normale $\mathcal{N}(\mu,\sigma)$.
  Dans ce cas, **la probabilité qu'une telle loi prenne une valeur précise est toujours nulle !**

  En revanche, on pourra toujours calculer la probabilité qu'une variable aléatoire X, distribuée selon une loi normale par exemple, prenne ses valeurs dans un intervalle $[t_1,t_2]$:

  $$\mathbb{P}[t_1 < X < t_2].$$

  Cette **probabilité** est alors **l'aire sous la courbe** représentant la densité de la loi.

La fonction $F(t) = \mathbb{P}[X \leq t]$ est appelée **fonction de répartition** de la loi de X. Pour tout t, $F(t)$ donne la probabilité que la variable aléatoire X prenne des valeurs plus petites que t.

En pratique si on cherche à déterminer $\mathbb{P}[X \leq t]$ où X est distribuée selon une loi normale de paramètres μ et σ, *i.e.* $X \sim \mathcal{N}(\mu,\sigma)$ on se ramène toujours à **une loi normale centrée et réduite** Z en posant

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

On a alors l'égalité suivante

$$\mathbb{P}[X \leq t] = \mathbb{P}\left[\frac{X-\mu}{\sigma} \leq \frac{t-\mu}{\sigma}\right] = \mathbb{P}\left[Z \leq \frac{t-\mu}{\sigma}\right].$$

On cherche ensuite la probabilité dans la table de la loi.
Enfin pour toute autre situation, si Z est une loi normale centrée et réduite, on se rappellera que :

- $\mathbb{P}[Z \geq t] = 1 - \mathbb{P}[Z \leq t]$, (**cette relation est toujours vraie**)
- $\mathbb{P}[Z \geq t] = \mathbb{P}[Z \leq -t]$,
- $\mathbb{P}[t_1 \leq Z \leq t_2] = \mathbb{P}[Z \leq t_2] - \mathbb{P}[Z \leq t_1]$.

1.4 Pour s'entraîner

Exercice 1. On considère que les âges dans une ville sont distribuées selon une loi normale $\mathcal{N}\left(\mu, \sigma^2\right)$, 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 ?

Exercice 2. Mon opérateur de téléphonie mobile m'assure que 95% des SMS que j'envoie seront transmis en moins d'une minute.

a) Quelle est la probabilité qu'un SMS envoyé en moins d'une minute ?
b) J'envoie chaque jour 4 SMS. Quelle est la probabilité que le nombre de SMS arrivés en moins d'une minute soit : $0, 1, 2, 3$ et 4 ?
c) Le week-end, j'envoie cette fois-ci 40 SMS par jour. Proposez une modélisation pour le nombre de SMS arrivés en mois d'une minute.
d) Quelle est la probabilité pour que le dimanche, au moins la moitié de mes SMS arrive en mois d'une minute ?

Exercice 3. On suppose que la distance en mètres parcourue par un javelot lancé par un athlète A suit une loi normale. Au cours d'un entraînement, on constate qu'exactement 10% des javelots atteignent plus de 75 mètres et exactement 25% des javelots atteignent moins de 50 mètres.
Calculer la longueur moyenne parcourue par un javelot ainsi que l'écart-type de cette longueur.

2 Estimation d'une espérance : cas où l'écart-type est connu

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 peut ensuite déterminer la variance de $\bar{X}_n$ pour ensuite en déduire l'écart-type. Pour cela, on utilisera les deux propriétés suivantes :

- pour tout nombre réel $a, Var[aX] = a^2 Var[X]$,
- si X et Y sont deux variables aléatoires indépendantes, alors $Var[X+Y] = Var[X] + Var[Y]$.

En utilisant successivement ces deux propriétés, nous avons :

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

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 $\mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$ 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 $Z_n = \dfrac{\bar{X}_n - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ 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 4, 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 d'erreur** Dans le cas précédent, le score de confiance associée à l'intervalle présenté était égal 95% et donc le risque d'erreur était de 5%.
On est bien sûr libre de choisir un risque plus faible et dans ce cas nous devons construire un 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 5 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.

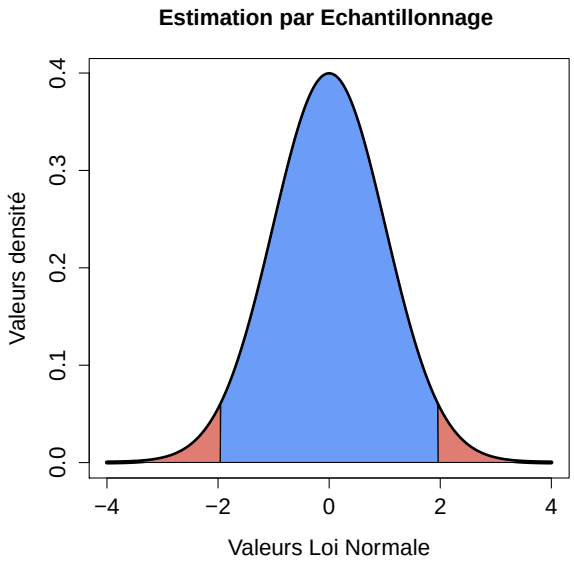

FIGURE 4 – Illustration des valeurs obtenues par échantillonnage. Dans la zone bleue se trouvent 95% des valeurs que l'on obtiendrait par échantillonnage et dans les zones rouges les 5% restants.

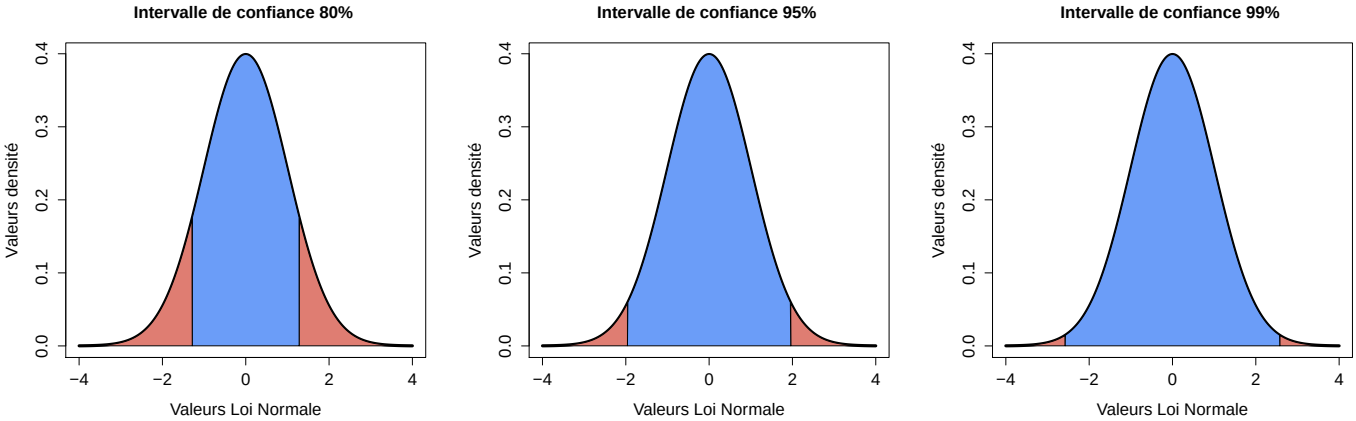

FIGURE 5 – Illustration de différents intervalles de confiance pour des marges d'erreur respectivement égales à $20, 5$ et 1% représentées en rouge sur les graphes.

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

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

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.

Définition 1 (Marge d'erreur)**.** *Lorsque l'on effectue de l'estimation par intervalle, on définit la marge d'erreur de notre intervalle de confiance par*

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Cette quantité représente l'écart au paramètre de population recherché. Plus notre marge d'erreur est grande, plus nous avons de risque d'obtenir des estimations éloignées de notre paramètres de population. Idéalement on cherche à obtenir des intervalles de confiance dont la marge d'erreur est faible, pour se faire on peut utiliser de grands échantillons, i.e. de grandes valeurs de n.

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 $1 - \frac{\alpha}{2} = 0.95$. 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 A retenir

En théorie et en pratique

On considère un échantillon de n mesures notées $x_1, \cdots, x_n$, alors l'estimateur de la moyenne $\bar{x}$ est donné par

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_{n-1} + x_n}{n}.$$

Cette estimateur de la moyenne $\bar{X}$ est une **variable aléatoire** dont la distribution dépend du contexte. Dans le cas où l'**on connaît l'écart-type** σ de la distribution et que les données sont issues d'une **distribution normale ou que notre échantillon est de taille** n **supérieure à** 30, alors

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \simeq Z \sim \mathcal{N}(0, 1),$$

où μ est le paramètre inconnu que l'on cherche à estimer.
Intervalle de confiance (symétrique !, mais on peut aussi rencontrer des intervalles de confiance non symétriques) de niveau $1 - \alpha$ pour la moyenne μ

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

où z_{alpha} est le quantile d'ordre α de la loi normale centrée réduite, *i.e.* c'est la valeur pour laquelle une variable aléatoire Z suivant une loi normale centrée réduite vérifie :

$$P[Z \leq z_\alpha] = \alpha.$$

On peut aussi dire qu'une proportion $1 - \alpha$ des estimations de la moyenne $\bar{x}$ tombent dans l'intervalle

$$\left[\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Pour donner un intervalle de confiance de niveau $1 - \alpha$ sur un paramètre inconnu comme la moyenne μ **dans le cas où l'écart-type de la distribution** σ **est connu**, on doit

1. donner une estimation de la valeur moyenne $\bar{x}$ à partir des données
2. vérifier la taille n de notre échantillon
3. déterminer la valeur de $z_{1-\alpha/2}$
4. calculer les bornes de l'intervalle de confiance à partir des informations précédentes

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

Si on souhaite vérifier si une machine est au norme (on connaît la valeur de référence μ) on pourra regarder si $\bar{x}$ se trouve dans l'intervalle

$$\left[\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

On procédera de la même façon pour la construction de cet intervalle.

2.4 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.

Exercice 3. On suppose que le poids d'un nouveau né est une variable normale d'écart-type égal à 0.5 kg. Le poids moyen des 49 enfants nés au mois de janvier 2004 dans l'hôpital de Charleville-Mézières a été de 3,6 kg.

a) Déterminer un intervalle de confiance à 95% pour le poids moyen d'un nouveau né dans cet hôpital.
b) Quel serait le niveau de confiance d'un intervalle de longueur 0.1 kg centré en 3.6 pour ce poids moyen ?

Exercice 4. 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 suivant une loi normale dont l'espérance μ et la variance σ^2 sont inconnues.
Il a obtenu les neuf mesures suivantes, exprimées en milligrammes :

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

On admet que l'écart-type associé à cet échantillon est égal à $s = 0.26$.

a) Calculer la moyenne $\bar{x}$ associée à cet échantillon.
b) Donnez un intervalle de confiance de niveau 0.90 de quantité de toxine..
c) Peut-on dire, avec un niveau de confiance de 0.80 que la quantité μ de toxine par gramme de solution est égale à 1.3 mg.

3 Estimation d'une espérance : cas où l'écart-type est inconnu

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 i**nconnu**.
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_k définie par

$$T_k = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{k}}}$$

suit **une loi de Student à k 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 k degré de liberté divisée par le nombre de degrés de liberté k.

La loi de Student, représentée en Figure 6, possède les propriétés suivantes :

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

Les premier et dernier points montrent que lorsque k 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.

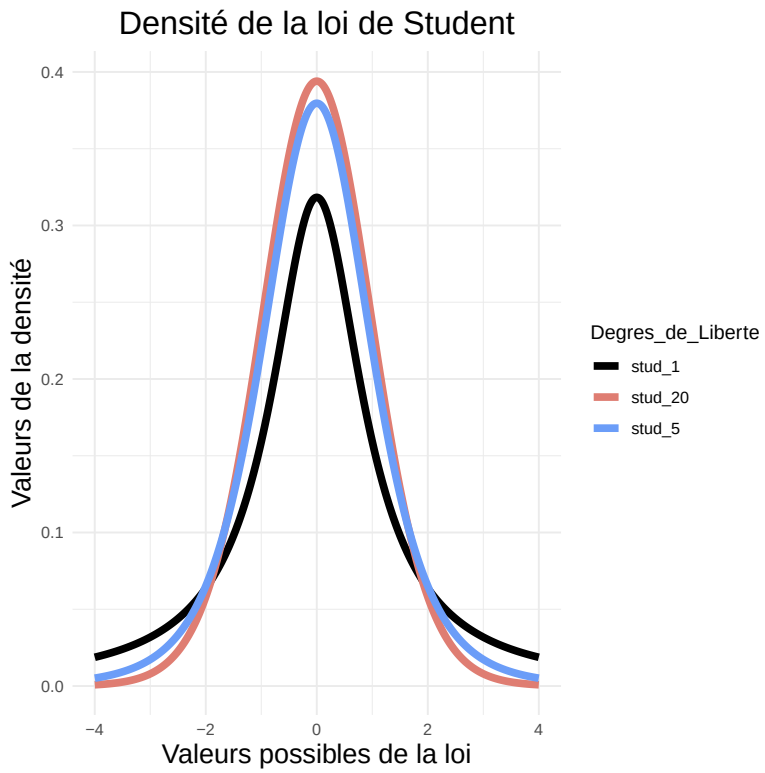

FIGURE 6 – Représentation de la loi de Student lorsque l'on fait varier le nombre de degrés de liberté

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_k suit une loi de Student à k degrés de liberté, alors

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

Remarques :
(i) En toute rigueur, je devrais noter $t_{k,1-\alpha/2}$ 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 k correspondante lorsqu'il fait mention de $t_{1-\alpha/2}$.
(ii) 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 !**
(iii) On se rappelle que lorsque **la distribution est symétrique**, comme cela est le cas pour la loi Normale centrée réduite ou la loi de Student, **nous avons** $t_{\alpha/2} = -t_{1-\alpha/2}$.

En repartant de l'inégalité précédente, nous avons donc

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

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

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

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

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

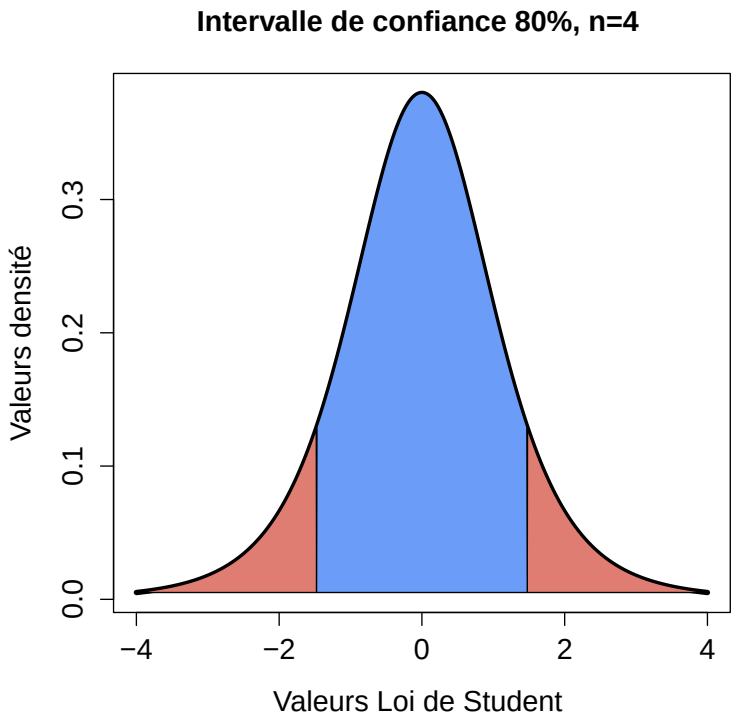

FIGURE 7 – Représentation de l'intervalle de confiance en bleu pour la loi de Student avec 4 degrés de liberté et pour un niveau de confiance $1-\alpha = 0.8$. On tire à nouveau profit de la symétrie pour répartir l'erreur de façon équitable à *gauche* et à *droite* de la distribution.

ce qui 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 18 en annexe de ce document. On représente également un exemple d'intervalle de confiance de niveau $1-\alpha = 0.8$ en Figure 7.

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.

On retrouve la même définition de *marge d'erreur* que ce nous avions dans la section précédente, adaptée cette fois-ci au présent contexte.

Définition 2 (Marge d'erreur)**.** *Lorsque l'on effectue de l'estimation par intervalle, on définit la marge d'erreur de notre intervalle de confiance par*

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

Cette quantité représente l'écart au paramètre de population recherché, i.e. la moyenne.

Exemple. Un laboratoire pharmaceutique souhaite étudier la fiabilité d'un automate qui est chargé de remplir des boîtes contenant une préparation médicale. On souhaite vérifier que chaque boîte contient une masse $\mu = 86.65$ grammes de cette préparation. On suppose que le remplissage des boîtes est distribué selon une normale de paramètres μ et σ^2 inconnus.
Pour vérifier le bon réglage de sa machine avec un risque d'erreur de 0.1, le laboratoire a prélevé un échantillon de 1000 boîtes ; sur cet échantillon, on a obtenu une masse moyenne de $\bar{x} = 86g$ et un écart-type $s = 0.12g$.

Nous sommes dans le cas où l'écart-type de notre population, ici la variabilité de masse dû à l'automate, notre estimateur de la moyenne suit donc **une loi de Student dont le degré de liberté**

est égal à l'échantillon moins un, *i.e.* 999.
L'intervalle de confiance $I_{1-\alpha}$ de niveau $1-\alpha = 0.9$ est défini par :

$$
\begin{aligned}
&\downarrow \text{définition d'un intervalle de confiance avec un score } 1-\alpha \\
I_{1-\alpha=0.9} &= \left[\bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}; \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}\right], \\
&\downarrow \text{Si } 1-\alpha = 0.9 \text{ alors } \alpha = 0.1, \text{ donc } \alpha/2 = 0.05 \text{ et } 1-\alpha/2 = 0.95. \\
&\downarrow \text{On remplace } \bar{x}, s \text{ et } n \text{ par les valeurs de l'énoncé.} \\
&= \left[86 + t_{0.05}\frac{0.12}{\sqrt{1000}}; 86 + t_{0.95}\frac{0.12}{\sqrt{1000}}\right], \\
&\downarrow \text{On cherche le quantile d'ordre 0.95 dans la table de student avec 999 degrés de liberté,} \\
&\downarrow \text{donc on va la ligne df= 1000 (on approxime) et la colonne } t_{0.95}. \\
&= \left[86 - 1.646 \times \frac{0.12}{\sqrt{1000}}; 86 + t_{0.95}\frac{0.12}{\sqrt{1000}}\right], \\
&\downarrow \text{Application numérique.} \\
I_{1-\alpha=0.9} &= [85.99; 86.01]
\end{aligned}
$$

Le laboratoire pharmaceutique peut donc conclure à un mauvais réglage de l'automate et devrait donc augmenter la valeur moyenne de remplissage de leur automate.

Bonus : essayez de faire de même en approximant votre loi de Student par une loi Normale centrée réduite. Cette approximation est licite car le nombre de degrés de liberté est suffisamment grand.

3.2 Exercices

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 $s^2 = 42.06$.

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, 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

$$\begin{aligned} I_{1-\alpha=0.90} &= \left[759.5 - 1.833\sqrt{\frac{42.06}{10}}; 759.5 + 1.833\sqrt{\frac{42.06}{10}}\right], \\ &= [755.74; 763.26]. \end{aligned}$$

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

$$\begin{aligned} I_{1-\alpha=0.99} &= \left[759.5 - 3.25\sqrt{\frac{42.06}{10}}; 759.5 + 3.25\sqrt{\frac{42.06}{10}}\right], \\ &= [752.83; 766.17]. \end{aligned}$$

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

$$\begin{aligned} 2 \times 2 \times \frac{\sqrt{42.06}}{\sqrt{n}} &< 1. \\ 4 \times \sqrt{42.06} &< \sqrt{n}. \\ n &> 16 \times 42.06, \\ n &> 672.96. \end{aligned}$$

Énoncé. Un ensemble de correcteurs souhaitent étudier la moyenne. obtenue à une épreuve d'Anglais et à une épreuve de Français-Philosophie d'un concours d'accès à l'Ecole Centrale de Lyon. Sur les $10,000$ copies à corriger, ils en corrigent 100 prisent aléatoirement pour chacune des épreuves. On suppose que les notes sont distribuées normalement

a) Sur les 100 copies sélectionnées, ils ont obtenu une moyenne de 10 avec un écart-type de 7 à l'épreuve d'Anglais. Avec quel risque d'erreur peut-on affirmer que la moyenne obtenue par les étudiants à ce concours sera inférieure à 11 ?
b) Sur les 100 copies sélectionnées, ils ont obtenu une moyenne de 6 avec un écart-type de 3 à l'épreuve de Français-Philosophie. Avec quel risque d'erreur peut-on affirmer que la moyenne obtenue par les étudiants à ce concours sera supérieure à 7 ?

Corrigé. Il s'agit ici de manipuler des intervalles de confiance dont la borne inférieure u supérieure nous est donnée.

a) Pour cette première question, il nous est demandé de déterminer le risque de se tromper en disant que la moyenne à cette épreuve soit inférieure à 11. Deux possibilités s'offrent à nous (i) on peut calculer le niveau de confiance associé à l'intervalle $[-\infty; 11]$ (*i.e.* que notre moyenne soit bien plus petite que 11) dans ce cas, le risque se défini comme étant $1 - \mathbb{P}[\mu \in [-\infty; 11]]$ **sinon** (ii) on peut directement calculer le risque que la moyenne soit plus grande que 11, *i.e.* $\mathbb{P}[\mu \in [11; +\infty]]$. On se concentre uniquement sur le cas (i).

   Nous sommes dans le cas où l'écart-type de la population est inconnu, donc notre intervalle de confiance est défini par la loi de Student ici à $n-1$ soit 99 degrés de libertés. Habituellement, on le construit de façon symétrique, cela veut dire que l'on répartit l'erreur aux deux extrémités de la distribution. Ce qui nous donne un intervalle de la forme :

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

   Mais dans le cas présent, l'erreur est représenté par l'intervalle $[11; +\infty]$**, donc notre intervalle de confiance est** $[-\infty; 11]$ **et est donc de la forme**

$$\left[-\infty; \bar{x} + t_{1-\alpha}\frac{s}{\sqrt{n}}\right].$$

   Le fait que l'erreur ne se trouve que d'un seul côté de la distribution fait que l'on travaille avec un terme en α **et plus en** $\alpha/2$.
   On doit donc maintenant chercher la valeur de $t_{1-\alpha}$ telle que

$$\bar{x} + t_{1-\alpha}\frac{s}{\sqrt{n}} = 11 \;\Leftrightarrow\; t_{1-\alpha} = \frac{11 - \bar{x}}{\frac{s}{\sqrt{n}}} = \frac{11 - 10}{\frac{7}{\sqrt{100}}} = \frac{10}{7} = 1.429.$$

   Il ne nous reste plus qu'à chercher la valeur de $1 - \alpha$ dans la table de la loi de Student correspondante à la valeur de quantile 1.429. La valeur que l'on trouvera nous donnera la probabilité d'appartenir à l'intervalle de confiance et cette dernière est égale à 0.925 (on a pris la valeur moyenne pour les deux valeurs les proches de 1.429).

   On est donc sûr à 92.5% **que la moyenne des étudiants sera inférieure à** 11**, on a donc un risque de** 7.5% **de se tromper.**
b) On procède exactement de la même façon pour l'épreuve de Français-Philosophie. Cette fois-ci, on veut savoir avec quel risque on peut affirmer que la moyenne obtenue par les étudiants sera supérieure à 7. Notre intervalle de confiance sera donc de la forme

$$\left[\bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}; +\infty\right] = \left[\bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n}}; +\infty\right] = [7; +\infty].$$

Et notre risque correspond donc à la probabilité d'appartenir à l'intervalle

$$\left[-\infty; \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}\right] = [-\infty; 7]\,.$$

On va cette fois-ci faire le choix de calculer directement le risque, contrairement à la question précédente. On doit donc maintenant chercher la valeur de $t_{1-\alpha}$ telle que

$$\bar{x} + t_{1-\alpha}\frac{s}{\sqrt{n}} = 7 \;\Leftrightarrow\; t_{1-\alpha} = \frac{7-\bar{x}}{\frac{s}{\sqrt{n}}} = \frac{7-6}{\frac{3}{\sqrt{100}}} = \frac{10}{3} = 3.333.$$

En cherchant la valeur de $1-\alpha$ dans la table de la loi de Student, on trouve que $1-\alpha = 0.9995$ (si on prend la valeur la plus proche), *i.e.* **on est presque sûr de se tromper en affirmant que la moyenne des étudiants à cette épreuve de concours sera supérieure à** 7, on a en effet une probabilité très proche de 1 qu'elle soit plus petite que 7.

3.3 A retenir

En théorie et en pratique

On considère un échantillon de n mesures notées $x_1, \cdots, x_n$, alors les estimateurs de la moyenne $\bar{x}$ et de la variance s^2 sont donnés par

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_{n-1} + x_n}{n}.$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

Cette estimateur de la moyenne $\bar{X}$ est une **variable aléatoire** dont la distribution dépend du contexte. Dans le cas où l'**on ne connaît pas** l'écart-type σ de la distribution et que les données sont issues d'une **distribution normale ou que notre échantillon est de taille n supérieure à** 30, alors

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \simeq T \sim \mathcal{T}_{n-1},$$

où μ est le paramètre inconnu que l'on cherche à estimer.
Intervalle de confiance (symétrique !, mais on peut aussi rencontrer des intervalles de confiance non symétriques) de niveau $1-\alpha$ pour la moyenne μ

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

où t_{alpha} est le quantile d'ordre α de la loi de Student à $n-1$ degrés de liberté, *i.e.* c'est la valeur pour laquelle une variable aléatoire T suivant une loi de Student à $n-1$ degrés de liberté vérifie :

$$P[T \leq t_\alpha] = \alpha.$$

On peut aussi dire qu'une proportion $1-\alpha$ des estimations de la moyenne $\bar{x}$ tombent dans l'intervalle

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

Pour donner un intervalle de confiance de niveau $1-\alpha$ sur un paramètre inconnu comme la moyenne μ **dans le cas où l'écart-type de la distribution σ est inconnu**, on doit

1. donner une estimation de la valeur moyenne $\bar{x}$ à partir des données
2. donner une estimation de l'écart-type s à partir des données
3. vérifier la taille n de notre échantillon
4. déterminer la valeur de $t_{1-\alpha/2}$
5. calculer les bornes de l'intervalle de confiance à partir des informations précédentes

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

Si on souhaite vérifier si une machine est au norme (on connaît la valeur de référence μ) on pourra regarder si $\bar{x}$ se trouve dans l'intervalle

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

On procédera de la même façon pour la construction de cet intervalle.

3.4 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.

Exercice 3. Un laboratoire pharmaceutique commercialise des sachets de bicarbonate de soude. L'ensachage est fait par une machine automatique en grande série. On appelle X la variable aléatoire qui à un sachet associe sa masse, exprimée en grammes. On considère que X suit la loi normale de moyenne μ inconnue et d'écart type σ inconnu On prélève au hasard et sans remise un échantillon de 85 sachets. La moyenne des masses de ces 85 sachets est 40 g et on a mesuré un écart-type de 0.1 g

a) Construire un intervalle de confiance pour la moyenne μ de niveau de confiance $1 - \alpha = 0.99$.
b) On suppose que le réglage optimale de la machine est de paramètre $\mu = 40.01$ g. Peut-on dire que la machine chargée de l'ensachage est bien réglée ?
c) Quelle doit être la taille minimale de l'échantillon pour que la taille de notre intervalle de confiance n'excède pas 0.01 ?

4 Estimation d'une proportion

Dans les deux derniers modules nous avons abordé la notion de fluctuation d'échantillonnage afin de construire des intervalles de confiance pour l'estimation d'une espérance d'une population μ. Nous avons comment construire de tels intervalles dans le cas où l'écart type de la distribution de la population est **connu**. Dans un tel cas, nous avons que l'estimation de la moyenne $\bar{X}$ suivait une loi normale $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ et l'intervalle de confiance de niveau $1-\alpha$ de la moyenne est alors donné par :

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

On a également vu comment construire de tels intervalles lorsque l'écart-type de la population est inconnu. Dans un tel cas, nous avons que l'estimation de la moyenne $\bar{X}$ suivait une loi de Student $\mathcal{T}_{n-1}$, où $n-1$ est le nombre de degrés de liberté de la loi et va dépendre de la taille n de l'échantillon.

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

L'estimation par intervalle de confiance d'un paramètre ne s'effectue cependant que sur l'espérance d'une distribution, on peut également appliquer ce type de méthode pour chercher à **estimer des proportions**, ce qui se révèle très utile en pratique lorsque l'on cherche à prédire les résultats d'une élection dans le cadre d'élections (Sondages IFOP)

4.1 Quelques rappels

Commençons d'abord par rappeler ce qu'est une proportion. On donne ci-dessous une définition de la proportion comme une estimation ponctuelle sur un échantillon donné.

Définition 3 (Proportion)**.** *Pour un échantillon donné de taille n, on définit la proportion $\bar{p}$ comme étant le ratio le nombre d'éléments n_x dans l'échantillon qui possèdent cette caractéristique et la taille de l'échantillon n. Plus formellement*

$$\bar{p} = \frac{n_x}{n}.$$

Le fait qu'un individu **possède ou non** (côté binaire) une caractéristique précise peut être modélisé par une variable aléatoire $X \sim \mathcal{B}(p)$ ou $\mathcal{B}(1,p)$ où le paramètre p est inconnu et représente la probabilité qu'un individu ait la caractéristique en question.

Considérons maintenant un échantillon $X_1, X_2, ..., X_n$ indépendants et ayant la même loi que $X \sim \mathcal{B}(p)$, alors l'estimateur de la proportion défini par

$$\bar{P} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

suit une loi admettant une espérance et une variance définies par

$$\mathbb{E}\left[\bar{P} = p\right] \quad \text{et} \quad Var(X) = \frac{p(1-p)}{n}.$$

En outre on a $n\bar{P} \sim \mathcal{B}(n,p)$ comme somme de n variables aléatoires indépendantes de Bernoulli de paramètre p.

Or nous avons vu en Section 1 que si notre échantillon était suffisamment grand, nous pouvions approximer la loi Binomiale par une loi Normale, *i.e.* lorsque n est assez grand

$$n\bar{P} \sim \mathcal{N}\left(np, np(1-p)\right),$$

où np et $np(1-p)$ désignent respectivement l'espérance et la variance de la loi binomiale $\mathbb{B}(n,p)$. On peut donc effectuer l'approximation suivante pour notre variable aléatoire $\bar{P}$

$$\frac{\bar{P}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1),$$

où nous avons simplement retranché la moyenne et divisé par l'écart-type pour se ramener à une loi Normale centrée et réduite.

Partant de cette approximation nous sommes à nouveau capable de construire des intervalles à un niveau $1-\alpha$ de confiance sur notre paramètre inconnu p. Plus précisément, nous sommes capables d'estimer avec quelle probabilité notre valeur obtenu par échantillonnage appartiendra dans un intervalle de confiance donné, i.e. nous avons

$$\mathbb{P}\left[p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq \bar{p} \leq p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right] = 1-\alpha.$$

A partir de ce résultat, nous pouvons également affirmer que nous avons une probabilité de $1-\alpha$ que notre paramètre inconnu appartienne à l'intervalle

$$\left[\bar{p} + z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}; \bar{p} + z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right].$$

L'écart à la vraie valeur de population est appelé marge d'erreur et a toujours la même signification que celle présentée dans les section précédentes, seule son expression change.

Définition 4 (Marge d'erreur)**.** *Lorsque l'on effectue de l'estimation par intervalle d'une proportion p, on définit la marge d'erreur de notre intervalle de confiance par*

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

La valeur du paramètre p étant ici inconnue, on utilisera l'estimation de cette valeur sur notre échantillon pour déterminer la marge d'erreur.

A partir de là, nous sommes capables de déterminer une valeur de n, *i.e.* la taille de minimale de notre échantillon afin que notre intervalle de confiance sur notre proportion soit suffisamment précis (voir exemples ci-dessous). Si l'on souhaite avoir une marge d'erreur égale à M, alors notre échantillon doit être de taille

$$n = \frac{z_{1-\alpha/2}^2 \times p(1-p)}{M^2}.$$

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 :

$$z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad = \quad 0.025,$$

$$1.96 \times \sqrt{\frac{0.2(1-0.2)}{n}} \quad = \quad 0.025.$$

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 A retenir

En théorie et en pratique

On considère un échantillon de n mesures notées $x_1, \cdots, x_n$, un estimateur de la proportion $\bar{p}$ (ou f) est donné par

$$\bar{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_{n-1} + x_n}{n},$$

où les x_i prennent les valeurs 0 ou 1.
Cette estimateur de la proportion $\bar{p}$ est une **variable aléatoire** à nouveau. Asymptotiquement, lorsque la taille de l'échantillon est suffisamment grande et vérifie $n\bar{p} \geq 5$ et $n(1-\bar{p}) \geq 5$

$$\frac{\bar{p} - p}{\sqrt{\bar{p}(1-\bar{p})/n}} \simeq Z \sim \mathcal{N}(0,1),$$

où p est le paramètre inconnu que l'on cherche à estimer.
Intervalle de confiance (symétrique !, mais on peut aussi rencontrer des intervalles de confiance non symétriques) de niveau $1-\alpha$ pour la proportion p

$$\left[\bar{p} - z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}; \bar{p} + z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right] = \left[\bar{p} + z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}; \bar{p} + z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right],$$

où z_{alpha} est le quantile d'ordre α de la loi de Normale centrée-réduite, *i.e.* c'est la valeur pour laquelle une variable aléatoire Z suivant une loi Normale centrée-réduite vérifie :

$$P[Z \leq z_\alpha] = \alpha.$$

On peut aussi dire qu'une proportion $1-\alpha$ des estimations de la proportion $\bar{p}$ tombent dans l'intervalle

$$\left[p - t_{1-\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}; p + t_{1-\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}\right].$$

Pour donner un intervalle de confiance de niveau $1-\alpha$ sur un paramètre inconnu comme la proportion p, on doit

1. donner une estimation de la proportion $\bar{p}$ à partir des données
2. vérifier la taille n de notre échantillon
3. déterminer la valeur de $z_{1-\alpha/2}$
4. calculer les bornes de l'intervalle de confiance à partir des informations précédentes

$$\left[\bar{p} - z_{1-\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}; \bar{p} + z_{1-\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}\right]$$

Si on souhaite vérifier si une machine est au norme (on connaît la valeur de référence μ) on pourra regarder si $\bar{x}$ se trouve dans l'intervalle

$$\left[p - z_{1-\alpha/2}\sqrt{p(1-p)/n}; p + z_{1-\alpha/2}\sqrt{p(1-p)/n}\right].$$

On procédera de la même façon pour la construction de cet intervalle.

4.4 Pour s'entraîner

Exercice 1. 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%

Exercice 2. On veut étudier la proportion p de gens qui vont au cinéma chaque mois. On prend donc un échantillon de taille $n = 100$. Soit N le nombre de personnes dans l'échantillon qui vont au cinéma mensuellement.

a) Quelle est la loi de N ? Par quelle loi peut-on l'approcher et pourquoi ? (voir cours) ? En déduire une approximation de la loi $F = N/n$?
b) On observe une proportion de f de gens qui vont chaque mois au cinéma. Donner la forme d'un intervalle de confiance pour p, de niveau de confiance $1 - \alpha$.
c) En déduire un intervalle de confiance sur p sachant que $f = 0.1$ aux niveaux de confiance

   (i) $1 - \alpha = 0.9$
   (ii) $1 - \alpha = 0.95$
   (iii) $1 - \alpha = 0.98$

Exercice 3. Un laboratoire pharmaceutique met en place un test pour estimer l'efficacité d'un nouveau médicament contre les migraines. Deux groupe de 125 patients souffrant de migraines considérés comme des échantillons aléatoire participe à ce test. On administre aux patients du groupe A le nouveau médicament, alors que les patients du groupe B reçoivent un placebo (sans principe actif). Au bout de quatre jours de traitement , 75 patients du groupe A et 65 patients du groupe B déclarent ressentir une diminution de l'intensité de leur migraine.

a) Déterminer les intervalles de confiance au niveau de confiance 0.95 des proportions de patients déclarant ressentir une diminution de l'intensité de leur migraine dans chaque échantillon.
b) Ces intervalles de confiance permettent-ils, au niveau de confiance 0.95, de considérer que le médicament soit plus efficace que le placebo ?
c) Quelle devrait-être la taille minimale de chaque échantillon pour que, avec des proportions identiques à celles observées précédemment, les résultats confirment l'efficacité du médicament, au niveau de confiance 0.95.

5 Théorie des tests

Dans les précédentes sections, nous avons vu comment construire des intervalles de confiance dans le cadre de l'estimation par intervalle à un niveau de confiance $1-\alpha$ donné. Nous avons étudié cette méthode pour différentes situations

- Pour l'estimation de la moyenne
  - lorsque la variance σ^2 est connue : nous sommes alors passés par la **loi Normale** pour construire de tels intervalles
  - lorsque la variance σ^2 est inconnue : nous avons construit les intervalles de confiance à l'aide de la **loi de Student**
- Pour l'estimation de proportion : nous avons à nouveau utilisé la loi normale pour construire nos intervalles de confiance

La construction de ces intervalles de confiance permet de déterminer les valeurs les plus probables que l'on pourrait obtenir par échantillonnage mais aussi de déterminer, à un niveau de confiance donné, l'intervalle de valeurs possibles pour nos paramètres inconnus de populations.

Dans cette section, nous allons voir comment fournir des réponses plus précises à des questions que nous étions par exemple posé lorsque l'on cherche à prédire le résultat à une élection par exemple ou encore de vérifier si le réglage sur ujne machine est le bon. Pour ce faire, nous abordons la notion de **test statistiques** qui ont nous permettre de répondre à une question sur nos paramètres inconnus à partir des informations issues d'un ou plusieurs échantillons.

Dans cette section, nous aborderons uniquement les **tests dits paramétriques**, ceux dont le but est de fournir des réponses quant aux **valeurs prises par les paramètres d'une loi**.

5.1 Quelques rappels

Exemple introductif . Avant de présenter le formalisme des tests, reprenons le cas du réglage d'un automate sur une chaîne montage où le but est de savoir si l'automate est ou non bien réglé. Imaginons que l'on connaisse le bon réglage de la machine μ_0 et que l'on dispose d'un échantillon sur lequel nous mesurons une certaine valeur $\bar{x}$ et que l'on connaît la variabilité pour l'exécution de la tache de notre automate, *i.e.* nous connaissons σ^2.

Notre objectif est de savoir si, le réglage actuel de notre automate μ (a priori inconnu !) est le bon ou non, *i.e.* si sa valeur est égale à notre valeur de référence μ_0. Savoir si oui ou non notre automate reviens à **tester**

$$H_0 : \mu = \mu_0 \quad \text{contre} \quad H_1 : \mu \neq \mu_0.$$

Dit autrement, le réglage actuel de la machine est-il proche du réglage de référence μ_0 ? Dans l'énoncé ci-dessus H_0 et H_1 sont appelées des **hypothèses** et n**otre objectif est de savoir quelle est l'hypothèse la plus réaliste, au sens probabiliste et statistique du terme.**

Le processus de "vérification" est très proche de ce nous avons vu pour établir nos intervalles de confiance. Si l'on considère toujours l'exemple de notre automate, nous avons vu que sous l'hypothèse H_0, *i.e.* lorsque l'on suppose que la machine est correctement réglée, alors la variable aléatoire Z définie par

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

suit une loi Normale centrée réduite (nous verrons plus tard que cette variable aléatoire Z est aussi appelée **Statistique de test**).

Pour savoir si notre hypothèse H_0 est vraie, il faudrait que notre espérance μ ne soit pas trop éloignée de la valeur idéale (ou de référence) μ_0. Pour cela, il faudrait que les valeurs prises par nos

échantillons soient comprises dans un intervalle de valeurs centré autour de la valeur de référence μ_0 avec une certaine probabilité $1-\alpha$. Si nous devions reformuler cela, il faudrait donc que les valeurs obtenus par échantillonnage se trouvent, dans 95% des cas dans l'intervalle :

$$\left[\mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

En pratique, nous ne disposons que d'un nombre restreint d'échantillons pour ne pas dire que d'**un seul échantillon**. Ainsi, dans le cas où l'estimation $\bar{x}$ sur notre échantillon n'appartient à cet intervalle, on décide de **rejeter l'hypothèse** H_0 au risque d'erreur α. Cela veut dire que le risque de rejeter H_0 à tort est égal à α. Dans le cas contraire on ne rejette pas H_0.

Remarque importante : on remarque que l'on ne rejette pas l'hypothèse H_0 si la valeur obtenue par échantillonnage est bien comprise dans l'intervalle de confiance calculé.

Formalisme. Un test statistique et donc un processus qui va permettre de juger de la validité d'une hypothèse étant données les observations effectuées sur un échantillon. En pratique on va considérer deux **hypothèses**

Définition 5. *Hypothèse Une hypothèse est une supposition faite sur les valeurs prises par les paramètres d'une loi (lorsque l'on effectue des tests paramétriques). On considérera toujours deux hypothèses : l'hypothèse H_0 également appelée **hypothèse nulle** et qui sera l'hypothèse utilisée pour effectuer le test. Elle est opposée à l'hypothèse H_1, appelée **hypothèse alternative**.*

Les différents tests, lorsque l'on cherche à tirer une conclusion quant à l'espérance d'une population, peuvent prendre les formes suivantes :

- $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$, ce type de test sera appelé un test **bilatéral**, on va chercher à estimer l'écart entre μ_0 et la valeur obtenu par échantillonnage, *i.e.* la valeur de $|\bar{x} - \mu_0|$
- $H_0 : \mu \leq \mu_0$ contre $H_1 : \mu > \mu_0$, ce type de test sera appelé un test **unilatéral**, on souhaite vérifier que μ ne prenne pas des valeurs plus grandes que μ_0
- $H_0 : \mu \geq \mu_0$ contre $H_1 : \mu < \mu_0$, on cherchera cette fois-ici à vérifier que μ ne prenne pas des valeurs plus petites que μ_0

Lorsque l'on effectue un test d'hypothèse, il y a généralement 4 issues possibles que l'on résume dans la Table 1, à chaque décision prise peut également être associé un score de confiance ou un risque. Par exemple, lorsque l'on effectue un test sous l'hypothèse H_0 et que celle-ci est vraie, il y a deux issues possibles

- on ne rejette pas l'hypothèse H_0, dans ce cas on peut avoir une confiance de $1-\alpha$ dans le test effectué
- on rejette l'hypothèse H_0 alors que celle-ci est vraie, c'est une erreur de première espèce. Cela arrive avec un risque de α que l'on appelle **risque de première espèce**

Exemple : un chercheur souhaite comparer l'efficacité de deux médicaments en effectuant un test différentes personnes. Il considère comme hypothèse nulle H_0 le fait que les médicaments aient la même efficacité contre H_1, ils n'ont pas la même efficacité.
Dans ce cas, une erreur de première espèce survient si le chercheur rejette l'hypothèse nulle et conclut que les deux médicaments sont différents alors qu'en fait ils ne le sont pas. Cette erreur n'est fondamentalement pas très grave car en réalité les deux médicaments sont tout aussi efficaces, il n'y a donc pas de risque pour le patient. En revanche, si une erreur de deuxième espèce survient, *i.e.* si le chercheur conserve H_0 alors que H_1 est vraie, cela peut être lourd de conséquences pour certains patients. Cette erreur est donc plus grave.

Décision \ Vérité	H_0	H_1
H_0	Conclusion correcte	Erreur de seconde espèce
H_1	Erreur de première espèce	Conclusion correcte

Décision \ Vérité	H_0	H_1
H_0	Confiance $1-\alpha$	Risque β
H_1	Risque α	Puissance $1-\beta$

TABLE 1 – Nom des différentes issues en fonction de la décision prise ainsi que de la vérité. La deuxième table présente les niveaux de confiance, risque, puissance pour chacune des situations.

Remarque : lorsque l'on fait un test en statistique, on cherche pas à vérifier la validité d'une hypothèse, *i.e.* l'issu d'un test statistique permet de conclure au rejet ou non de l'hypothèse H_0, mais on ne dit pas que l'on **accepte l'hypothèse** H_0

Comme on a pu le voir dans l'exemple introductif (et comme nous l'avons fait pour la construction des intervalles de confiance), nous serons toujours amenés à étudier les valeurs prises par une variable aléatoire de référence et construite en fonction du contexte (connaissance ou non de l'espérance et ou variance de notre population).

Test et Intervalle de confiance. On va se concentrer sur le cas où l'on cherche à tester la moyenne μ lorsque l'on connaît la variance σ^2 associée à une population afin d'illustrer les deux types de tests : **bilatéral** et **unilatéral**. Nous rappelons les deux méthodes dont nous disposons permettant de conduire au rejet ou non de l'hypothèse H_0, en se basant sur les intervalles de confiance

- **Test Bilatéral :**
  On effectue ce test lorsque l'on effectue le test d'hypothèses

  $$H_0 : \mu = \mu_0 \quad \text{contre} \quad H_1 : \mu \neq \mu_0.$$

  On rappelle que dans ce cas, un intervalle de confiance de niveau $1-\alpha$, sous l'hypothèse H_0 est donné par

  $$\left[\mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

  On souhaite vérifier que les valeurs prises par nos échantillons ne sont pas trop éloignée s de la valeur de référence μ_0, si ce n'est pas le cas, on rejette l'hypothèse H_0. La Figure 8 (gauche) illustre les zones de rejets et de non rejet de l'hypothèse H_0 en fonction des valeurs $\bar{x}$ prises par notre échantillon. Ainsi notre hypothèse H_0 est rejetée si elle se trouve dans l'une des deux zones rouges. Elle est conservée dans le cas contraire, *i.e.* si la valeur prise par notre statistique de test est bien comprise entre les deux quantiles définies par le niveau de confiance $1-\alpha$ (ou encore par le risque d'erreur α) de notre test.

  $$\bar{x} \in \left[\mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \mu_0 + z_{1-\alpha_2}\frac{\sigma}{\sqrt{n}}\right]$$

  ou, de façon équivalente, si

  $$\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \in \left[z_{\alpha/2}; z_{1-\alpha/2}\right].$$

- **Test Unilatéral à gauche :**
  On effectue ce test lorsque l'on effectue le test d'hypothèses

  $$H_0 : \mu = \mu_0 \text{ ou } \mu \geq \mu_0 \quad \text{contre} \quad H_1 : \mu < \mu_0.$$

Dans ce cas, on souhaite vérifier que les valeurs obtenues par échantillonnage **ne sont pas inférieures** de la valeur de référence μ_0. Dans le cas d'un test unilatéral **à gauche**, notre intervalle de confiance de niveau $1-\alpha$ est défini par

$$\left[\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}; +\infty\right].$$

La Figure 8 (milieu) illustre les zones de rejets et de non rejet de l'hypothèse H_0 en fonction des valeurs $\bar{x}$ prises par notre échantillon. Ainsi notre hypothèse H_0 est rejetée si elle se trouve dans la zone rouge. Elle est conservée dans le cas contraire, *i.e.* si la valeur prise par notre statistique de test est bien comprise entre les deux quantiles définies par le niveau de confiance $1-\alpha$ (ou encore par le risque d'erreur α) de notre test.

$$\bar{x} \in \left[\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}; +\infty\right]$$

ou, de façon équivalente, si

$$\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \in [z_\alpha; +\infty].$$

- **Test Unilatéral à droite :**
  On effectue ce test lorsque l'on effectue le test d'hypothèses

$$H_0 : \mu = \mu_0 \text{ ou } \mu \leq \mu_0 \quad \text{contre} \quad H_1 : \mu > \mu_0.$$

Dans ce cas, on souhaite vérifier que les valeurs obtenues par échantillonnage **ne sont pas supérieures** de la valeur de référence μ_0. Dans le cas d'un test unilatéral **à droite**, notre intervalle de confiance de niveau $1-\alpha$ est défini par

$$\left[-\infty; \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right].$$

La Figure 8 (droite) illustre les zones de rejets et de non rejet de l'hypothèse H_0 en fonction des valeurs $\bar{x}$ prises par notre échantillon. Ainsi notre hypothèse H_0 est rejetée si elle se trouve dans la zone rouge. Elle est conservée dans le cas contraire, *i.e.* si la valeur prise par notre statistique de test est bien comprise entre les deux quantiles définies par le niveau de confiance $1-\alpha$ (ou encore par le risque d'erreur α) de notre test.

$$\bar{x} \in \left[-\infty; \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right]$$

ou, de façon équivalente, si

$$\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \in [-\infty; z_{1-\alpha}].$$

Test et p-value . Jusqu'à présent, nous avons simplement exploités des outils reposant sur les intervalles de confiance pour conclure ou non au rejet d'une hypothèse H_0 à un seuil de risque α fixé. Cependant, il existe une méthode permettant d'apporter des informations plus précises quant au risque d'erreur que l'on est susceptible de commettre dans le cas où l'on rejette l'hypothèse H_0, c'est ce que l'on appelle **la p-value**.

Définition 6 (p-value.)**.** *Soit une Statistique de Test U distribuée selon une loi $\mathcal{L}$ et notons $\bar{u}$ la valeur prise par notre statistiques de test pour un échantillon donné (ex : $\bar{u} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ dans le cas d'un test sur la moyenne lorsque σ est connu). La p-value est alors la probabilité que notre variable aléatoire U (i.e. notre statistique, prenne une valeur plus "improbable"), i.e.*

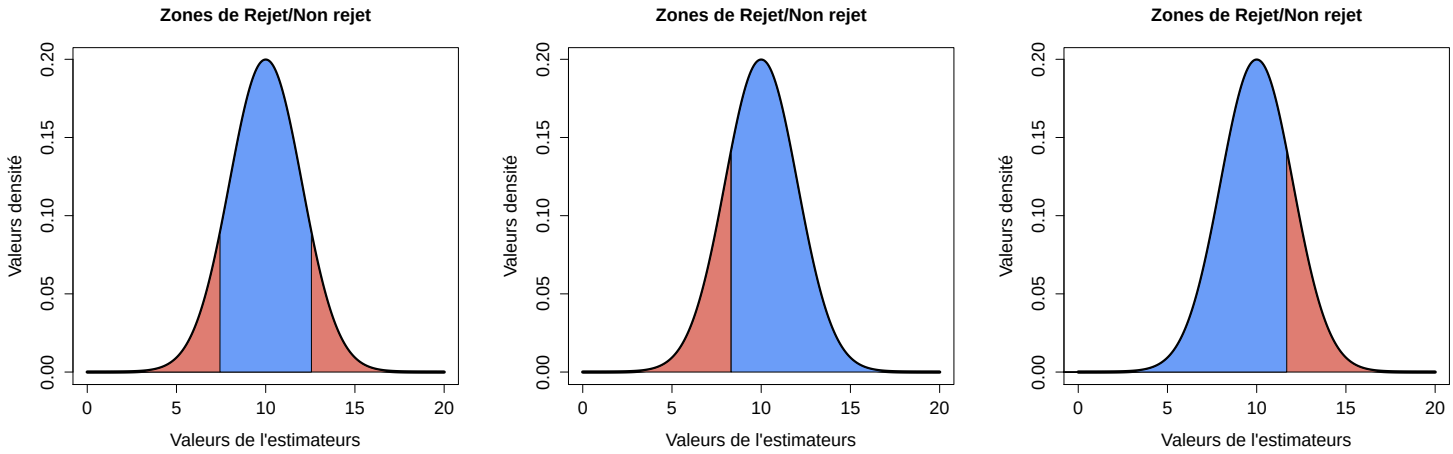

FIGURE 8 – Graphique représentant les zones de rejet (pour un test bilatéral ou unilatéral à gauche ou droite) ou non de l'hypothèse H_0 en fonction des valeurs prises par notre estimateur sur un échantillon avec une loi $\mathcal{N}(10, 4)$. La zone bleue correspond à une zone de non rejet de l'hypothèse H_0 alors que la zone rouge correspond à des zones de rejet de l'hypothèse H_0 à un risque d'erreur $\alpha = 0.2$.

- *pour un test bilatéral :*
$$2 \times \mathbb{P}\left[U \geq |\bar{u}|\right],$$
- *pour un test unilatéral à gauche :*
$$\mathbb{P}\left[U \leq \bar{u}\right],$$
- *pour un test unilatéral à droite :*
$$\mathbb{P}\left[U \geq \bar{u}\right].$$

Cette p-value peut-être déduite de la table des quantiles (ou table des probabilités) associés à la loi de U. Elle est ensuite comparée au risque d'erreur α, si la p-value est plus petite que le risque d'erreur α, cela signifie que le risque pris en rejetant H_0 est inférieur au risque seuil α que l'on s'était fixé, on peut donc rejeter H_0 *a priori* sans craintes. Dans le cas contraire, on conserve H_0.

La p-value nous donne donc une information plus précise que les méthodes précédentes qui reposent sur l'utilisation d'intervalles de confiance. **En effet, elle nous renseigne sur la probabilité que l'on a de rejeté une hypothèse à tort étant donnée l'observation effectuée.**

Résumé des étapes d'un test d'hypothèses. Pour résumer les différentes étapes d'un test statistiques ou d'un test d'hypothèses.

1) Définir l'hypothèse nulle H_0 ainsi que l'hypothèse alternative H_1
2) Fixer un risque d'erreur α, qui servira de seuil de décision
3) Déterminer la loi de la Statistique de test U sous l'hypothèse H_0, c'est-à-dire la variable aléatoire ainsi que sa loi qui seront utilisées pour conclure au rejet ou non de H_0.
4) Déterminer la valeur de la statistique de test u en fonction des grandeurs estimées à l'aide d'un échantillon.
5) Conclure au rejet de H_0 (plusieurs choix s'offrent à vous) en choisissant l'une des méthodes suivantes. On prendra le cas particulier où l'on effectue un test bilatéral mais les autres cas s'obtiennent de façon analogue.

   Si z n'appartient pas l'intervalle de de confiance de niveau $1 - \alpha$ associé à la loi suivie par notre statistique de test U
$$\left[u_{\alpha/2};\ u_{1-\alpha/2}\right]$$

OU si la valeur $\bar{x}$ mesurée sur votre échantillon n'appartient à l'intervalle de confiance associée, qui est de la forme

$$\left[\mu_0 + u_{\alpha/2}\sigma_{\bar{x}};\ \mu_0 + u_{1-\alpha/2}\sigma_{\bar{x}}\right]$$

OU calculer la p-value associée à la valeur prise par votre statistique de test et la comparer au risque d'erreur α. Si la p-value est plus petite que α on rejette alors l'hypothèse H_0.

Quelques statistiques de tests. Cette dernière section présente les différentes statistiques de test que vous êtes susceptibles de rencontrer jusqu'à présent. On laisse le soin au lecteur de construire les intervalles de confiance dans les différents cas mais aussi en fonction du test effectué (bilatéral ou unilatéral). Ces dernières sont issues des lois de probabilités étudiées dans les sections précédentes.

- **Test sur l'espérance μ lorsque σ est connu**

  C'est le cas que nous avons présenté un peu plus tôt et que nous étudié en Section 2, la statistique de test U employée est définie par

  $$U = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}},$$

  où $\bar{X}$ est un estimateur de la moyenne sur un échantillon. La statistique de test U est distribuée selon **une loi Normale centrée réduite** $\mathcal{N}(0,1)$, où de façon équivalente $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

- **Test sur l'espérance μ lorsque σ est inconnu**

  Nous avons vu en Section 3 comment construire des intervalles de confiance dans ce cas là à l'aide d'un estimateur de la variance s^2 sur notre échantillon. La statistique de test U employée est définie par

  $$U = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}},$$

  Dans le cas présent, U suit **une loi de Student à $n-1$ degrés de liberté** où n représente la taille de l'échantillon.

- **Test sur la proportion p**

  C'est le cas étudié en Section 4. La statistique de test U employée ici est définie par

  $$U = \frac{\bar{P} - p}{\sqrt{\frac{p(1-p)}{n}}},$$

  où $\bar{P}$ est l'estimateur d'une proportion dans un échantillon de taille n et vérifiant, pour rappel, $n\bar{P} \sim \mathcal{B}(n,p)$. p étant inconnu, la variance de la loi est estimée à l'aide de la valeur de $\bar{P}$ prise sur un échantillon, *i.e.* on remplace $\sqrt{\frac{p(1-p)}{n}}$ par $\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$.
  La variable aléatoire U **suit une loi Normale centrée réduite** $\mathcal{N}(0,1)$, ou, de façon équivalente, $\bar{P} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$ lorsque n est assez grand (conséquence du théorème centrale limite)

Exemple. Reprenons notre exemple des automates chargés de remplir des boîtes avec un certaine contenance. On souhaite vérifier que la machine est correctement réglée avec une marge d'erreur de 5%, *i.e.* si on suppose que le remplissage d'une boîte suit une loi Normale $\mathcal{N}(\mu, \sigma^2)$, on souhaite vérifier que $\mu = \mu_0 = 86.5$g. On suppose en outre que la variance σ^2 est inconnue. Pour faire cette vérification, on sélectionne 100 boîtes aléatoirement sur lequel le poids moyen de remplissage $\bar{x} = 86$g et la variabilité de remplissage est $s^2 = 6.25$g.

Pour répondre à cette question, il y a plusieurs solutions possibles. Dans tous les cas, il s'agit de formuler le test d'hypothèses suivant :

— $H_0 : \mu = \mu_0 = 86.5$
— $H_1 : \mu \neq \mu_0$

Nous sommes dans le cas où l'on cherche à estimer, ou faire un test, sur une espérance dans le cas où la variance est inconnue. La statistique de test U à considérer suit donc une loi de Student à $n - 1$ soit 99 degrés de liberté.

On a donc 2 possibilités pour répondre à cette question : (i) construire l'intervalle de confiance comme nous avons déjà pu le faire plus tôt et vérifier que la valeur théorique μ_0 appartient bien à cet intervalle **ou** (ii) calculer la p-value associée à la statistique de test qui nous renseignera sur la probabilité que l'on a de rejeter H_0 et la comparer au seuil de risque fixé $\alpha = 0.05$.

(i) Nous avons vu précédemment que l'intervalle de confiance est défini par

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

Dans notre cas, nous avons $t_{1-\alpha/2} = t_{0.975} = 1.984$, ce qui nous conduit à l'intervalle de confiance suivant :

$$[85.504; 86.496].$$

On en conclut donc que l'on peut rejeter l'hypothèse H_0 au risque d'erreur de 5%. Nous allons maintenant voir que la deuxième solution nous fournira une réponse un peu plus précise.

(ii) Notre statistique de test U suit une loi de Student à 99 degrés de liberté, la valeur u prise par la statistique de test sur l'échantillon en question est :

$$u = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{86 - 86.5}{\sqrt{\frac{6.25}{100}}} = -2.$$

Nous sommes dans le cas d'**un test bilatéral**, donc la p-value est définie par

$$\mathbb{P}[|U| \geq |u|] = 2\mathbb{P}[U \geq |u|] \quad \text{par symétrie de la loi de Student.}$$

On doit donc estimer $2\mathbb{P}[U \geq 2] = 2 \times (1 - \mathbb{P}[U \leq 53.03]) = 0.048$. La p-value est inférieure à 0.05, on peut donc rejeter l'hypothèse H_0 avec un risque d'erreur égal à 0.048.

Vous pourriez tester différentes valeurs de s^2, par exemple, pour $s^2 = 10$, pouvez rejeter l'hypothèse H_0 ? Quelle est la probabilité de rejeter l'hypothèse H_0 à tort ?

5.2 Exercice

Énoncé. Des études en psychologies du développement ont montré qu'à l'âge de 12 mois, 50% des bébés "normaux" marchent. On souhaite mener une étude sur les retards de développement des bébés prématurés. On teste l'hypothèse que les bébés prématurés marchent plus tardivement que les bébés normaux. On observe une population de 80 bébés prématurés. A 12 mois, 35 de ces 80 bébés marchent.

a) Faut-il réaliser un test unilatéral ou bilatéral ?
b) Peut-on valider, au seuil de risque α de 5%, l'hypothèse de recherche ?
c) Quelle est la p-value associée dans ce cas là ?

Corrigé. Dans cet exercice on souhaite faire le test d'hypothèses suivant :

— H_0 : les bébés prématurés marchent au même moment que les bébés normaux, *i.e.* $p = p_0 = 50\%$ ou $p \geq p_0$
contre
— H_1 : les bébés prématurés marchent plus tardivement $p < p_0$.

a) A la lumière de l'énoncé, on doit donc faire un test unilatéral à gauche, on souhaite donc vérifier que les valeurs obtenues par échantillonnages ne sont pas inférieures à p_0.
b) On se propose de regarder si la valeur obtenue par échantillonnage est dans notre intervalle de confiance. S'agissant d'un test unilatéral à gauche, notre intervalle de confiance est de la forme

$$\left[p_0 - z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}; +\infty\right] = \left[0.5 - z_{1-0.05}\sqrt{\frac{0.5\times(1-0.5)}{80}}; +\infty\right] = [0.39; +\infty]\,.$$

La valeur $\bar{p} = \dfrac{35}{80} = 0.4375$ se trouvant dans cet intervalle, on ne peut pas conclure au rejet de l'hypothèse H_0 donc on ne peut pas valider l'hypothèse des chercheurs au seuil de risque 5%.
c) Dans le cas présent, la statistique de test est définie par la variable aléatoire

$$Z = \frac{\bar{F} - p_0}{\sqrt{\dfrac{\bar{p}(1-\bar{p})}{n}}} \sim \mathcal{N}(0,1).$$

La valeur de la statistique de test est égale à $z = \dfrac{0.4375 - 0.5}{\sqrt{\dfrac{0.4375(1-0.4375)}{80}}} = -1.13$ et la p-value associée est 0.13.

Énoncé. Pour étudier un nouvel alliage métallique, on a soumis un échantillon aléatoire de 16 tiges aux essais pour obtenir les résistances suivantes en kg/cm^2 :

$$1895, 1920, 1886, 1890, 1864, 1880, 1875, 1915, 1850, 1927, 1910, 1912, 1886, 1903, 1854, 1880$$

. On suppose que la résistance est distribuée normalement.

a) Donner un intervalle de confiance de niveau 95% de la résistance moyenne de la rupture.
b) Avant l'introduction de ce nouvel alliage la résistance moyenne à la rupture des tiges était de 1840 kg/cm^2. Que peut-on conclure des essais effectués avec le nouvel alliage ?
c) Formuler un test d'hypothèse répondant à la question précédente et calculer la p-value.

Corrigé L'exercice propose déjà de construire un intervalle de confiance puis de faire un test d'hypothèse sur la résistance de tiges métalliques.

a) Pour construire notre intervalle de confiance, nous aurons besoin de deux choses, de la moyenne ainsi que de la variance **débiaisée** évaluée sur notre échantillon. En effet, nous ne connaissons ni la moyenne, ni la variance (ou écart-type) à l'échelle de la population. On rappelle que ces deux quantités sont définies par :

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

Numériquement, on obtient les valeurs $\bar{x} = 1890.44$ et $s^2 = (23.09)^2$. Notre intervalle de confiance de niveau 95% se construit à l'aide de la loi de Student, car nous ne connaissons pas la variance de la population. Plus précisément nous devrons utiliser la loi de Student à $16 - 1 = 15$ degrés de liberté. Ce qui nous donne un intervalle de confiance de niveau $1 - \alpha$:

$$I_{1-\alpha} = \left[\bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n}}; \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}\right].$$

Le quantile d'ordre 0.975 d'une loi de Student à 15 degrés de liberté est égal à 2.13, d'où

$$I_{0.95} = [1878.14; 1902.73].$$

b) On peut conclure au fait que ce nouvel alliage a augmenté la résistance des tiges métalliques car la valeur 1840 ne se trouvant pas dans l'intervalle de confiance. Elle même plus petite que la borne inférieure de notre intervalle de confiance.

c) On souhaite effectuer le test suivant :
- H_0 La résistance des tiges métalliques est de 1840 kg/cm^2, *i.e.* $\mu = \mu_0 = 1840$
- H_0 La résistance des tiges métalliques est supérieure à 1840 kg/cm^2, *i.e.* $\mu > \mu_0 = 1840$.

Cette formulation suggère donc que l'on va effectuer un test unilatéral supérieur (la zone de rejet va se trouver à droite). On nous demande ici de calculer la p-value associée à notre test, qui, comme pour la construction de l'intervalle de confiance, repose sur la statistique de test de Student T et est définie par

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}.$$

Sous l'hypothèse H_0, elle devient :

$$T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}.$$

A l'aide des valeurs de notre échantillon, la valeur de notre statistique de test est :

$$t = \frac{1890.44 - 1840}{\frac{23.09}{\sqrt{16}}} = 8.74$$

La p-value associée est définie comme $\mathbb{P}(T > t) = 1 - \mathbb{P}(T < t)$ or $\mathbb{P}(T < t) \simeq 0.999$, donc notre p-value est proche de 0. On peut donc rejeter l'hypothèse H_0 et confirmer que la résistance des tiges métalliques a bien augmenter avec ce nouvel alliage.

5.3 A retenir

En théorie et en pratique

Nature du test et zone de rejet
Soit m une quantité statistique que l'on cherche à tester. On rappelle alors la forme des **zones de rejet** de l'hypothèse H_0 selon **la nature de l'hypothèse alternative** H_1.

$$H_0 : m = m_0 \quad \text{contre} \quad H_1 : m \neq m_0.$$

Test bilatéral pour lequel la région de rejet est défini par

$$\left[-\infty; u_{\alpha/2}\right] \cup \left[u_{\alpha/2}; \infty\right]$$

et la $p-$value est donnée par $2P[U \leq |u|]$, où U est une variable aléatoire ayant la même loi que la statistique de test et u est la valeur de la statistique de test sur l'échantillon

$$H_0 : m = m_0 \text{ ou } m \leq m_0 \quad \text{contre} \quad H_1 : m > m_0.$$

Test unilatéral à droite ou **supérieur** pour lequel la région de rejet est défini par

$$[u_{1-\alpha}; \infty]$$

et la $p-$value est donnée par $P[U \geq u]$, où U est une variable aléatoire ayant la même loi que la statistique de test et u est la valeur de la statistique de test sur l'échantillon

$$H_0 : m = m_0 \text{ ou } m \geq m_0 \quad \text{contre} \quad H_1 : m < m_0.$$

Test unilatéral à gauche ou **inférieur** pour lequel la région de rejet est défini par

$$[-\infty; u_{\alpha}]$$

et la $p-$value est donnée par $P[U \leq u]$, où U est une variable aléatoire ayant la même loi que la statistique de test et u est la valeur de la statistique de test sur l'échantillon.

On rejette l'hypothèse H_0 si la $p-$value est plus petite que le risque d'erreur α que l'on se fixe pour effectuer le test.

Quel test utiliser ?

- Pour un test sur **la moyenne lorsque l'écart-type est connu**, on effectuera un test basé sur **la loi normale centrée réduite**.
- Pour un test sur **la moyenne lorsque l'écart-type est inconnu**, on effectuera un test basé sur **la loi de Student** dont le degré de liberté dépend de la taille de l'échantillon étudié.
- Pour un test sur **la proportion**, on effectuera un test basé sur **la loi normale centrée réduite**.

Etapes pour effectuer un test

1) Définir l'hypothèse nulle H_0 ainsi que l'hypothèse alternative H_1
2) Fixer un risque d'erreur α, qui servira de seuil de décision
3) Déterminer la loi de la Statistique de test U sous l'hypothèse H_0, c'est-à-dire la variable aléatoire ainsi que sa loi qui seront utilisées pour conclure au rejet ou non de H_0.
4) Déterminer la valeur de la statistique de test u en fonction des données de échantillon.
5) Conclure au rejet (ou non) de l'hypothèse H_0 (plusieurs choix s'offrent à vous).

5.4 Pour s'entraîner

Exercice 1. Un ingénieur risque crédit, employé dans une société spécialisée dans le crédit à la consommation, veut vérifier l'hypothèse selon laquelle la valeur moyenne des mensualités des clients de son portefeuille est de 200 euros. Un échantillon aléatoire de 144 clients, prélevé aléatoirement dans la base de données, donne une moyenne empirique $\bar{x} = 193.74$ et une estimation non biaisée de l'écart-type $s = 48.24$.

a) Quelles sont les hypothèses statistiques associées à la problématique de l'ingénieur et quel type de test faut-il utiliser pour l'aider à prendre une décision correcte d'un point de vue statistique ?
b) Peut-il conclure au risque de 0.05, que la valeur moyenne postulée des remboursements est correcte ?
c) Faites le schéma des régions de rejet et de non rejet de l'hypothèse H_0 en y notant les valeurs critiques calculées à la question précédente.
d) Représenter la p-value associée à ce test. Que vaut-elle ?
e) En utilisant la p-value, quelle aurait été la réponse à la question b) pour un risque de première espèce $\alpha = 0.1$?

Exercice 2. Une machine produit des tiges métalliques dont la longueur nominale est égale à 8.30 cm. Les fluctuations de longueurs dues au procédé de fabrication correspondent à un écart-type de 0.6 cm. Sur la base d'un échantillon aléatoire de taille 100, on veut tester si la machine est bien réglée. La moyenne des longueurs mesurées sur l'échantillon est de $8.57cm$.

a) Faut-il réaliser un test bilatéral ou unilatéral ? Conclure au seuil de risque α de $0.05, 0.01$ et 0.001.
b) Quel serait la conclusion avec un échantillon de taille 20 uniquement avec la même moyenne mesurée ?

Exercice 3. Un négociant en vin s'intéresse à la contenance des bouteilles d'un producteur soupçonné par certains clients de frauder. Il souhaite s'assurer que cette contenance respecte bien en moyenne la limite inférieure légale de 75 cl. A cet effet, il mesure le contenu de 10 bouteilles prises au hasard et obtient les valeurs suivantes :

$$73.2; 72.6; 74.5; 75; 75.5; 73.7; 74.1; 75.8; 74.8; 75$$

On suppose que le processus de remplissage suit une loi Normale $\mathcal{N}(\mu, \sigma^2)$ où $\sigma = 1$.

a) Déterminer la valeur moyenne de remplissage $\bar{x}$ sur cet échantillon.
b) Formulez clairement le choix de notre hypothèse nulle H_0 et de votre hypothèse alternative H_1.
c) Quel type de test faut-il utiliser pour l'aider à prendre une décision correcte d'un point de vue statistique ? On précisera s'il s'agit d'un test bilatéral ou unilatéral (inférieur ou supérieur) ainsi que la loi suivie par la Statistique de test. Justifiez votre réponse.
d) Le négociant peut-il conclure, au risque d'erreur de 1%, que le producteur respecte bien la limite inférieure légale de 75 cl ?
e) Le négociant veut pouvoir détecter, une probabilité élevée (99%), une contenance moyenne d'au moins 74.8 cl tout en gardant un test au risque d'erreur de 1%. Que doit-il faire ?

Exercice 4. On estime que 60% de la population française regarde régulièrement des séries en streaming (de façon légale ou non).
Une enquête a été lancée par un institut de sondage au près d'étudiants d'une université Lyonnaise

afin de savoir quelle proportion d'étudiants regardent des séries en streaming.
Sur les 2500 étudiants interrogés, 2000 d'entre eux affirment regarder régulièrement des séries en streaming sur leur ordinateur. Que peut-on affirmer concernant les étudiants vis-à-vis de la population française en générale

a) Donner une estimation ponctuelle de la proportion d'étudiants regardant des séries en streaming.
b) Dans un premier temps, on souhaite savoir si la proportion d'étudiants regardant des séries en streaming est bien la même qu'à l'échelle de la population.
   Sur quelle quantité va porter votre test statistique ? Formulez clairement le choix de notre hypothèse nulle H_0 et de votre hypothèse alternative H_1.
c) Quel type de test faut-il utiliser pour l'aider à prendre une décision correcte d'un point de vue statistique ? On précisera s'il s'agit d'un test bilatéral ou unilatéral (inférieur ou supérieur) ainsi que la loi suivie par la Statistique de test. Justifiez votre réponse.
d) Peut-on conclure au risque d'erreur de 5% que la proportion est différente ?
e) Quel test aurait-il été plus judicieux d'effectuer dans ce cas là ? Formulez les nouvelles hypothèses et conclure pour un risque d'erreur de 5%.

6 Comparaison de Moyennes

La section précédente était consacrée à la présentation des tests d'hypothèses en formulant une hypothèse *nulle* H_0, qui est l'hypothèse sur laquelle sera construite notre test, *versus* l'hypothèse *alternative* H_1. Nous avons vu comment définir un test d'hypothèses, la notion de statistique de test selon ce que l'on souhaite "tester" ainsi que la notion de p-value qui permet de définir le risque de première espèce, *i.e.* la probabilité de rejeter l'hypothèse H_0 à tort.Nous avons ensuite formulé des test d'hypothèses dans le cas où nous souhaitions étudier si la *moyenne* d'échantillon était bien conforme à une valeur de référence, nous avons aussi étudié cela dans le cas d'une *proportion*.

Jusqu'à présent, nous ne disposions que des informations relatives à un seul échantillon, mais il peut parfois se révéler utile de comparer plusieurs populations pour savoir si elles présentent ou non les mêmes caractéristiques. Par exemple on pourrait étudier la numération des cellules selon deux protocoles expérimentales différents et chercher à voir **si le nombre de cellules observées est bien significativement différent d'un protocole à un autre**. D'un point de vue Marketing, nous pourrions par exemple regarder si les gens dépensent en moyenne la même somme d'argent lorsqu'ils partent en vacances pendant la période estivale que lors de vacances pendant la période hivernale. On va donc s'intéresser à **la comparaison de moyenne entre deux populations** qui seront représentées par deux échantillons, mais nous avons deux cas possibles :

- dans le cas numération des cellules, nos deux cultures différentes impliquent que l'on effectue des mesures sur des populations différentes ou **indépendantes**, *i.e.* nos mesures effectuées proviennent bien de deux populations différentes

- dans le cas de notre analyse marketing, on pourrait interroger un même groupe de personnes qui part en vacances en été **et** en hiver. Dans ce les mesures effectuées proviennent d'**un même échantillon de personnes mais à deux instants différents**. On parle alors d'é**chantillons dépendants ou appariés**.

Dans toute cette section, on va supposer que nos données sont issues d'une distribution gaussienne ou que nos échantillons sont de tailles suffisantes (en général > 30).

6.1 Quelques rappels

Cas des échantillons indépendants. C'est le cas où les mesures effectuées sont bien faits sur deux populations d'individus différentes. Un autre exemple pourrait consister à comparer les moyennes obtenus par les étudiants des groupes c et d à cet enseignement afin de voir s'il y a, d'un point de vue statistique, une différence de niveau entre les deux groupes.

Lorsque l'on dispose des moyennes issues de deux échantillons indépendants et que l'on souhait tester si les moyennes sont identiques ou non, on ne va plus chercher à comparer les valeurs obtenues sur les échantillons à une valeur théorique comme nous l'avons fait dans la section précédente. Cette fois-ci, on va étudier la différence entre les deux moyennes et donc formuler le test d'hypothèses suivant :

— $H_0 : \mu_1 - \mu_2 = 0$, les moyennes de nos échantillons sont égales
contre
— $H_1 : \mu_1 \neq \mu_0$ si on souhaite effectuer un test bilatéral
$H_1 : \mu_1 < \mu_0$ ou $\mu_1 > \mu_2$ dans le cas d'un test unilatéral

En étudiant la différence des moyennes, cela revient donc à supposer que la différence étudiée suit une loi Normale centrée, donc de moyenne nulle.

Considérons maintenant deux échantillons :

- un échantillon de taille n_1 issu d'une première population et de moyenne μ_1

- un échantillon de taille n_2 issu d'une autre population **indépendante** de la première et de moyenne μ_2

Il faut maintenant distinguer deux cas, comme nous l'avons fait pour les intervalles de confiance, (i) le cas où les variances sont connues et (ii) le cas où les variances sont inconnues.

(i) Cas où la variance des distributions des deux populations σ_1^2 et σ_2^2 sont connues

La variance associée à l'estimateur de la différence des moyennes est égale à la somme des variances sur chaque échantillon, *i.e.*

$$\sigma_{\bar{x}_1-\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

où σ_1^2 et σ_2^2 sont les variances respectives de nos deux échantillons. Cela découle du résultat suivant : pour deux variables aléatoires X et Y **indépendantes** : $Var(X+Y) = Var(X) + Var(Y)$.
La statistique de test utilisée est définie par la variable aléatoire :

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

La variable aléatoire Z suit une loi Normale centrée réduite $\mathcal{N}(0,1)$. Donc pour tester si nos deux moyennes sont égales, pour un test bilatéral par exemple et sous l'hypothèse $H_0 : \mu_1 = \mu_2$, la statistique de test prend la valeur

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

In fine, pour savoir on rejette l'hypothèse H_0 au risque d'erreur α si

$$z \notin \left[z_{\alpha/2}; z_{1-\alpha/2}\right].$$

(ii) Cas où la variance des distributions des deux populations σ_1^2 et σ_2^2 sont inconnues.

Le principe reste exactement le même, on considère simplement la statistique de test suivante

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

où s_1^2 et s_2^2 sont les estimations de la variance sur chaque population obtenues à l'aide d'un échantillon. L'estimateur T suit une loi de Student dont le degré de liberté est défini par l'approximation d'Aspin-Welch. Le nombre de degrés de liberté est l'entier le plus proche de

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}.$$

In fine, on rejette l'hypothèse H_0 au risque d'erreur α si

$$t \notin \left[t_{\alpha/2}; t_{1-\alpha/2}\right].$$

Remarque : ce test également appelé *test t de Welch*

Cas des échantillons dépendants ou appariés. C'est le cas où les mesures effectuées sont bien faits sur une même population mais à deux instants différents. Un autre exemple que nous pourrions citer est l'étude de l'impact d'un régime sur la masse des personnes, on procède donc à l'étude de deux échantillons un consistant à la prise de mesure avant le régime et un autre à la prise de mesures après ou pendant le régime.

Dans ce cas là on dispose de deux échantillons de même taille n (vu que l'on étudie les mêmes personnes mais à deux instants différents) :

$$(x_1, x_2, x_3, ..., x_{n-1}, x_n) \quad \text{et} \quad (y_1, y_2, y_3, ..., y_{n-1}, y_n).$$

Le test va porter sur le seul échantillon défini comme la différence des mesures effectuées sur les deux échantillons, *i.e.*

$$(x_1 - y_1, x_2 - y_2, x_3 - y_3, ..., x_{n-1} - y_{n-1}, x_n - y_n).$$

On considère D la variable aléatoire qui traduit les distribution d'échantillonnage de la différence qui possède une espérance μ.Si on ne connait pas la variance de la distribution de notre population, ce qui est sera toujours le cas pour nous, on considère $\bar{x}$ comme étant la moyenne estimée sur l'échantillon "différence" et on note s l'écart-type associé à ce même échantillon "différence".

Le test d'hypothèses consiste à étudier la nullité de la moyenne de notre population, *i.e.* on formule les hypothèses suivantes

- $H_0 : \mu = 0$, la moyenne de l'échantillon différence est nulle
  contre
- $H_1 : \mu \neq 0$ si on souhaite effectuer un test bilatéral
  $H_1 : \mu < 0$ ou $\mu > 0$ dans le cas d'un test unilatéral

La statistique considérée est alors définie par

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

suit une loi de Student à $n - 1$ degrés de liberté. Dans le cas d'un test bilatéral, on rejette alors l'hypothèse nulle H_0 si la valeur de la statistique de test, au risque d'erreur α sous l'hypothèse H_0, définie par

$$t = \frac{\bar{x}}{\sqrt{\frac{s^2}{n}}}$$

se trouve en dehors de l'intervalle

$$\left[t_{\alpha/2}; t_{1-\alpha/2}\right].$$

Exemple. On s'intéresse à la taille de des jeunes enfants. En prélevant un échantillon au sein d'une école primaire, 41 garçons et 61 filles dans des classes de CP ont été mesurés. La taille moyenne des garçons de cet échantillon est $\bar{x}_1 = 107$ cm avec un écart-type $v_1 = 8$ cm. La taille moyenne des filles est de $\bar{x}_2 = 104$ cm avec un écart-type $v_2 = 9$ cm. On supposera que nos deux échantillons sont gaussiens. Peut-on affirmer, au risque d'erreur $\alpha = 0.05$, que les garçons sont plus grand que les filles ?

Il s'agit d'un cas où l'on chercher à **comparer des moyennes** pour lesquelles on dispose de deux estimations obtenus sur **des échantillons indépendants**. Dans le cas présent, **on ne connait pas la variance associée à chaque population** donc le test statistique que nous utiliserons sera basé sur **la loi de Student**.

On va chercher à tester l'hypothèse H_0 : les garçons et les filles ont la même taille, *i.e.* $\mu_1 = \mu_2$ où μ_1 désigne la taille moyenne des garçons et μ_2 désigne la taille moyenne des filles, contre l'hypothèse alternative H_1 : les garçons sont en moyenne plus grands que les filles, *i.e.* $\mu_1 > \mu_2$. On va donc effectuer un t**est unilatéral supérieur** ou **test unilatéral à droite**.
La statistique de test est définie par

$$T_d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

La statistique de test T_d est distribuée selon une loi de Student à d degrés de liberté où le nombre de degrés de liberté est donné par l'approximation d'Aspin-Welch :

$$d = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}} = \frac{\left(\frac{8^2}{41} + \frac{9^2}{61}\right)^2}{\frac{8^4}{41^2(41-1)} + \frac{9^4}{61^2(61-1)}} = 31.99 \to 32.$$

Donc une loi de Student à 47 degrés de libertés.
La valeur de la statistique de test t nous est donnée par

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{107 - 104}{\sqrt{\frac{8^2}{41} + \frac{9^2}{61}}} = 1.765.$$

On rappelle que l'on effectue un test unilatéral supérieur, donc la p-value est définie par :

$$p\text{-value} = \mathbb{P}[T > 1.765] = 1 - \mathbb{P}[T < 1.765] = 1 - 0.956 = 0.044.$$

Notre p-value est plus petite que le risque α considéré, on peut donc rejeter l'hypothèse H_0 et on peut donc affirmer, qu'en moyenne, les garçons sont plus grands que les filles.

6.2 Exercice

Énoncé. On désire tester l'effet d'un régime alimentaire. On a pesé 10 individus, avant et après le régime. Voici les poids obtenus sur les différents individus, on suppose que les données sont distribuées selon une loi gaussienne.

Avant	67	83	158	78	87	58	79	63	69	72
Après	66	84	121	76	82	58	77	64	68	70

Peut-on dire que le régime a eu un effet, au risque d'erreur $\alpha = 0.05$?

Corrigé. Nous sommes dans le cas où l'on cherche à comparer des moyennes sur des données issues d'une même population mais à deux instants différents, *i.e.* **les deux échantillons dont l'on dispose sont donc appariés/dépendants.**
On commence donc par calculer l'échantillon basé sur la différence entre les deux échantillons *Avant* - *Après*

Avant	67	83	158	78	87	58	79	63	69	72
Après	66	84	121	76	82	58	77	64	68	70
Différence	1	-1	37	2	5	0	2	-1	1	2

Cet échantillon Différence se caractérise par une moyenne μ que l'on va chercher à comparer à 0. Afin de déterminer si le régime a eu un effet ou non, on formule les hypothèses H_0 : le régime n'a pas eu d'effet, *i.e.* $\mu = 0$ contre l'hypothèse H_1 : le régime a eu un effet, *i.e.* $\mu \neq 0$. Il s'agit donc d'un test bilatéral.
La statistique de test utilisée est définie par

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

qui suit une loi de Student à $n-1$ soit 10 degrés de liberté.

Pour calculer la valeur de notre statistique de test, nous devons commencer par estimer la valeur moyenne $\bar{x}$ et l'écart-type s sur notre échantillon. On rappelle que

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

Sur notre échantillon nous avons $\bar{x} = 4.8$ et $s = 11.448$, la valeur de notre statistique de test est donc $t = 1.326$.

On rappelle que l'on effectue un test bilatéral, donc la p-value est définie par :

$$p\text{-value} = 2\mathbb{P}[T > 1.326] = 2(1 - \mathbb{P}[T < 1.326]) = 2(1 - 0.891) = 0.218.$$

La p-value est plus grande que le risque d'erreur α fixé à 0.05, on ne peut donc pas rejeter l'hypothèse H_0.

6.3 A retenir

En théorie et en pratique

On doit cette fois-ci comparer deux échantillons S_1 et S_2 pour tirer des conclusions (*e.g.* efficacité d'un produit, meilleur stratégie de vente). On peut être confronté aux tests suivants :
le **test bilatéral** qui se formule par

$$H_0 : m_1 = m_2 \text{ ou } m_1 - m_2 = 0 \quad \text{contre} \quad H_1 : m_1 \neq m_2.$$

le **test unilatéral à droite** ou **supérieur** que l'on formule par

$$H_0 : m_1 - m_2 \leq 0 \text{ ou } m_1 - m_2 = 0 \quad \text{contre} \quad H_1 : m_1 - m_2 > 0.$$

le **test unilatéral à gauche** ou **inférieur** défini par

$$H_0 : m_1 - m_2 \geq 0 \text{ ou } m_1 - m_2 = 0 \quad \text{contre} \quad H_1 : m_1 - m_2 < 0.$$

Cas des échantillons indépendants : il n'existe aucun lien entre les deux séries de mesures.

- **On connaît la variance des deux distributions**
  La statistique de test employée U suit **une loi normale centrée-réduite** sous H_0 et sa valeur est donnée par

$$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

- **On ne connaît pas la variance des deux distributions**
  On estime la variance de l'échantillon global par

$$s^2 = s_1^2/n_1 + s_2^2/n_2.$$

  La statistique de test employée U suit **une loi de Student** dont le nombre de degré de liberté est donné par l'approximation d'Aspin-Welch sous H_0 et sa valeur est donnée par :

$$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

Cas des échantillons dépendants ou appariés : les mesures sont liées, *i.e.* ce sont des mesures répétées sur les mêmes individus.
On étudie **l'échantillon différence** $D = S_1 - S_2$ pour supprimer les liens entre les données. On estime ensuite moyenne $\bar{x}$ et variance s^2 de cet échantillon D par les formules usuelles.
La statistique de test employée U suit **une loi normale centrée-réduite** sous H_0 et sa valeur est donnée par

$$u = \frac{\bar{x}}{\sqrt{s^2/n}}.$$

---

1) Identifier l'indépendance ou non des échantillons
2) Définir l'hypothèse nulle H_0 ainsi que l'hypothèse alternative H_1
3) Fixer un risque d'erreur α, qui servira de seuil de décision
4) Déterminer la loi de la Statistique de test U sous l'hypothèse H_0, c'est-à-dire la variable aléatoire ainsi que sa loi qui seront utilisées pour conclure au rejet ou non de H_0
5) Déterminer la valeur de la statistique de test u en fonction des données de échantillon
6) Conclure au rejet (ou non) de l'hypothèse H_0 (plusieurs choix s'offrent à vous)

6.4 Pour s'entraîner

Exercice 1. On désire tester l'effet d'un médicament censé réduire le taux de cholestérol. On a mesure de la taux de cholestérol (g/l) chez 10 patients, avant la prise de ce médicament et une semaine après la prise de ce médicament. Les résultats obtenus sont résumés dans la table ci-dessous :

Avant	0.1	0.2	0.15	0.3	0.34	0.16	0.09	0.24	0.17	0.29
Après	0.8	0.18	0.12	0.2	0.3	0.21	0.12	0.16	0.17	0.22

Peut-on dire, au risque d'erreur $\alpha = 0.1$, que le laboratoire n'est pas un truand ?

Exercice 2. On souhaite mesurer l'influence de l'alcool sur le temps de réaction au volant. Sur un échantillon aléatoire de 30 chauffeurs, le temps de réaction a été observé en laboratoire avec et sans consommation d'alcool (les 30 chauffeurs ont été répartis aléatoirement). Les temps de réactions en secondes ont été rapportés dans le tableau suivant :

Sans	0.68	0.64	0.68	0.82	0.58	0.80	0.72	0.65	0.84	0.73	0.65	0.59	0.78	0.67	0.65
Avec	0.73	0.62	0.76	0.92	0.68	0.87	0.77	0.70	0.88	0.79	0.72	0.80	0.78	0.86	0.78

Peut on affirmer qu'il y a une influence de l'alcool sur le temps de réaction ($\alpha = 5\%$) ?

Exercice 3. Pour déterminer le poids moyen d'un épis de blé appartenant à deux variétés, on procède à 10 pesées pour chaque variété. On donne ci-après les moyennes et variances empiriques des deux échantillons des variétés.

$$\bar{x}_1 = 1.707g, \quad \bar{x}_2 = 1.685g, \quad s_1^2 = 4.329, \quad s_2^2 = 1.827.$$

Au risque d'erreur $\alpha = 0.1$, peut-on dire que les deux variétés de blé ont deux masses différentes ?

Exercice 4. Au sein d'une université, on souhaite déterminer si le niveau en mathématiques entre des étudiants des filières *Biologie* et les étudiants de la filière **Physique** et le même. Pour effectuer ce test, on soumet 50 étudiants de chaque filière à une épreuve. On estime que l'écart-type des notes obtenues par les biologiste en général est $\sigma_1 = 6$ et que l'écart-type des notes obtenues par les physiciens est $\sigma_2 = 4$. Les notes moyennes obtenus par les deux groupes de candidats sont les suivantes

$$\bar{x}_1 = 12, \qquad \text{et} \qquad \bar{x}_2 = 13.5.$$

Au risque d'erreur $\alpha = 0.01$, peut-on dire que le niveau des physiciens est meilleur que le niveau des biologiste en mathématique ?

Exercice 5. Le service des Ressources Humaines souhaite savoir si l'entreprise applique bien la nouvelle législation en vigueur concernant la rémunération des hommes et des femmes à travail égal (poste et compétences). L'entreprise étant une grand multinationale, elle s'intéresse plus particulièrement à la catégorie ouvriers pour son étude, catégorie la plus représentée dans son entreprise.
Pour cela elle a étudié les revenus annuels (en milliers d'euros) de 54 salariés femmes et de 69 salariés hommes, les résultats de l'étude sont présentés dans la table ci-dessous

	Homme	Femme
Moyenne	$\bar{x}_1 = 24.5$	$\bar{x}_2 = 22.3$
Ecart-type	$s_1 = 2.1$	$s_2 = 2.3$
Médiane	$med_1 = 25$	$med_2 = 24$

Peut-on dire que l'entreprise respecte la nouvelle législation en vigueur, *i.e.* est-ce qu'elle respecte bien l'égalité de salaire homme-femme pour un même poste et à compétences égales ? Conclure au risque d'erreur de 1%.

Exercice 6. Des chercheurs en psychomotricité étudie l'influence des grossesses prématurées sur le développement des fonctions motrices d'un enfant. En outre, ils cherchent à montrer que les bébés issus d'un accouchement prématurée marchent plus tardivement que les bébés nés à termes.
Pour leur étude, ils ont relevé l'âge moyen en mois, auquel le bébé effectue ses premiers pas sur plus de 210 bébés normaux et 154 bébés nés après un accouchement prématuré, l'étude à conduit aux chiffres suivants :

	Prématurés	Non prématurés
Moyenne	$\bar{x}_1 = 13$	$\bar{x}_2 = 14$
Ecart-type	$s_1 = 2$	$s_2 = 2.5$
Médiane	$med_1 = 11$	$med_2 = 13$

Les chercheurs peuvent-ils conclure, au risque d'erreur de 5% que les bébés prématurés marchent plus tardivement que les bébés issus d'une grossesse menée à terme ?

Exercice 7. Une enquête a été réalisée par l'INSEE sur le salaire net moyen dans les différentes régions de la France Métropolitaine au cours des années 2016 et 2017 Afin de simplifier la lecture, nommerons ces régions R_1 jusqu'à R_{13}. On présente ci-dessous les résultats obtenus au cours de ces deux périodes pour un nombre limité de département.

Région / Période	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}	R_{13}
Horaire Net 2016	18.7	13.11	12.56	13.60	12.8	12.12	13.1	12.90	13.31	11.5	13.89	14.17	12.53
Horaire Net 2017	18.8	13.61	13.12	13.51	13.63	13.47	13.35	13.20	13.33	13.80	14.40	14.27	12.82
Diff 2017 - 2016	0.10	0.50	0.56	-0.09	0.83	1.35	0.25	0.30	0.02	2.30	0.51	0.10	0.29

$$\bar{x}_{2016} = 13.40, \quad \bar{x}_{2017} = 13.94, \quad \bar{x}_{diff} = 0.54,$$
$$s_{2016} = 1.74, \quad s_{2017} = 1.52, \quad s_{diff} = 0.65.$$

On va supposer que la populations étudiée dans les deux cas est la même et qu'elle est de taille infinie. Nous souhaitons déterminer si la moyenne du taux de chômage en France métropolitaine a évolué ou non entre les deux périodes.
Au risque d'erreur de 10%, peut-on dire que le salaire moyen de la population française a évolué entre 2016 et 2017

7 Analyse de Variances (ANOVA)

Dans la section précédente nous avons chercher à comparer des moyennes dans le cas où nous disposions de deux échantillons ou groupes indépendants. Nous avons vu que nous pouvions tester si les moyennes de la mesure effectuée sur les deux groupes sont significativement différentes ou non à l'aide d'un test de Student.

Dans cette section, nous allons chercher à généraliser cela à plusieurs groupes, *i.e.* lorsque l'on cherche à comparer les moyennes de plus de deux groupes. Cela peut se révéler intéressant dans certaines marketing lorsque l'on à déterminer sur quelle catégorie d'individus consacrée ses efforts afin de booster ses ventes ou son chiffre d'affaire. Il n'est alors pas rare de devoir segmenter la population dans des groupes disjoints afin d'identifier si les dépenses moyennes varient d'un groupe à un autre.

Nous allons donc voir comment faire cela à l'aide d'une Analyse de Variances (ANOVA) et quel est le test à utiliser.

7.1 Quelques rappels

Généralités Plusieurs situations peuvent nous amener à comparer les moyennes par exemple lorsque l'on cherche à étudier la moyenne des dépenses en fonction du lieu de vacances ou du type de vacances (ski, mer, montagne, hotel club, ...). Il s'agit donc de tester si les dépenses moyennes sont indépendantes du type de vacances.
Cela conduit à effectuer le test d'hypothèse suivant :

$$H_0 : \mu_i = \mu \; \forall i \quad \text{v.s.} \quad H_1 : \exists \; i, j \text{ tels que } \mu_i \neq \mu_j$$

On illustre en Figure 9 un exemple où les distributions respectent l'hypothèse H_0 et H_1 respectivement. Comme le montre ces graphiques, quelques hypothèses sont cependant nécessaires pour permettre d'effectuer une analyse de variance :

- avoir un échantillon de taille suffisamment importante dans chaque groupe ($n_j \geq 30$), cela permet de garantir que les estimateurs de la moyenne sont distribuées selon une loi gaussienne. Si les données de base sont déjà issues d'une loi normale, alors la taille de l'échantillon importe peu.
- on va supposer que les variances sont identiques dans chacun des groupes, c'est ce que l'on appelle l'hypothèse **d'homoscédasticité**.

Dans ce qui suit, on notera Y la variable aléatoire dont on cherche à savoir si la moyenne varie selon le groupe sur lequel elle est étudiée.
et on va considérer que l'on étudie K groupes pour lesquelles on dispose d'échantillons de taille n_k et on notera N la taille de l'échantillon total, *i.e.* $N = \sum_{k=1}^{K} n_k$. Les observations associées seront donc notées $y_{i,k}$ pour désigner le $i-$ème exemple du groupe k.

Etude de la variance. La question est maintenant de savoir comment mettre en avant une différence ou non entre les valeurs moyennes des différents groupes. La terme "Analyse de variances" devrait nous mettre la puce à l'oreille et nous indique qu'il va certainement falloir étudier **des** variances mais lesquelles ...

Jusqu'à présent, nous avons défini la variance relativement à une variable aléatoire ou à un échantillon, mais que se passe-t-il si l'on dispose d'un échantillon pour chaque groupe i ? On peut en fait étudier deux "natures" de variances :

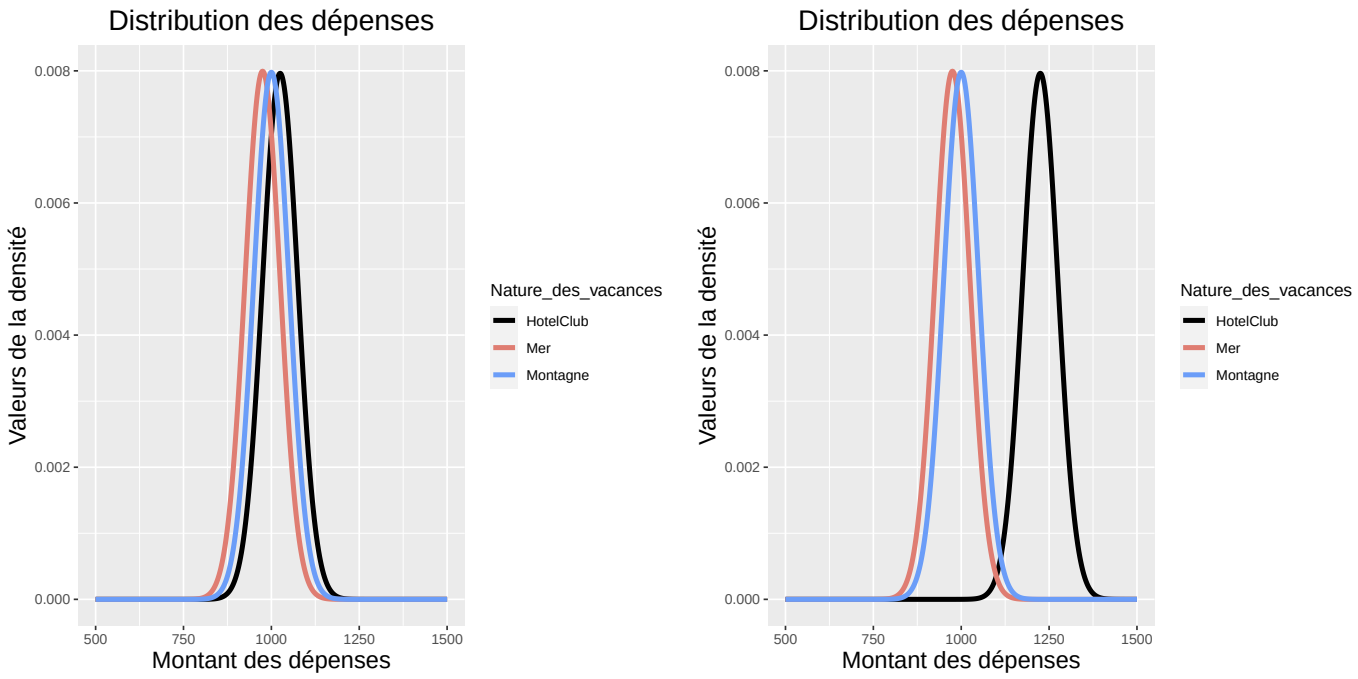

FIGURE 9 – Représentation graphique des dépenses en fonction du lieu de vacance. La figure de gauche illustre le cas où la dépense moyenne est identique d'un groupe à l'autre, *i.e.* elle illustre l'hypothèse H_0. La figure de droite illustre le cas où l'hypothèse H_1 est retenue, *i.e.* la moyenne d'au moins l'un des groupe est différente des autres.

- la variance inter-classe, notée SCE_{facteur} (que l'on appelle également la somme des carrés des erreurs dues au facteur du modèle, dans un autre contexte).
  C'est donc une variance **entre les différentes classes** qui est définie par :

$$SCE_{\text{facteur}} = \sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2,$$

  où $\bar{y}_k = \dfrac{1}{n_k}\sum_{i=1}^{n_k} y_{i,k}$ désigne la moyenne évaluée sur l'échantillon du groupe k et $\bar{y} = \dfrac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k} y_{i,k}$ est la moyenne globale (donc sur l'ensemble des échantillons).
  On peut voir ce terme comme une **variance pondérée des moyennes**.

- la variance intra-classe, notée $SCE_{\text{résidu}}$ (que l'on appelle également la somme des carrés des erreurs dues qui n'est pas du au facteur dans un autre contexte).
  Il s'agit cette fois-ci de calculer des variances internes à chaque groupe/échantillon. Cette variance intra-classe est définie par

$$SCE_{\text{résidu}} = \sum_{k=1}^{K}(n_k - 1)s_k^2,$$

  où $s_k^2 = \dfrac{1}{n_k - 1}\sum_{i=1}^{n_k}$ est la **variance débiaisée** du groupe/échantillon k.
  On peut donc voir cette variance intra-classe comme une **moyenne pondérée des variances**.

Pour la petite histoire, les deux variances présentées précédemment sont en fait liées par la relation suivante :

$$SCE_{\text{total}} = SCE_{\text{facteur}} + SCE_{\text{résidu}},$$

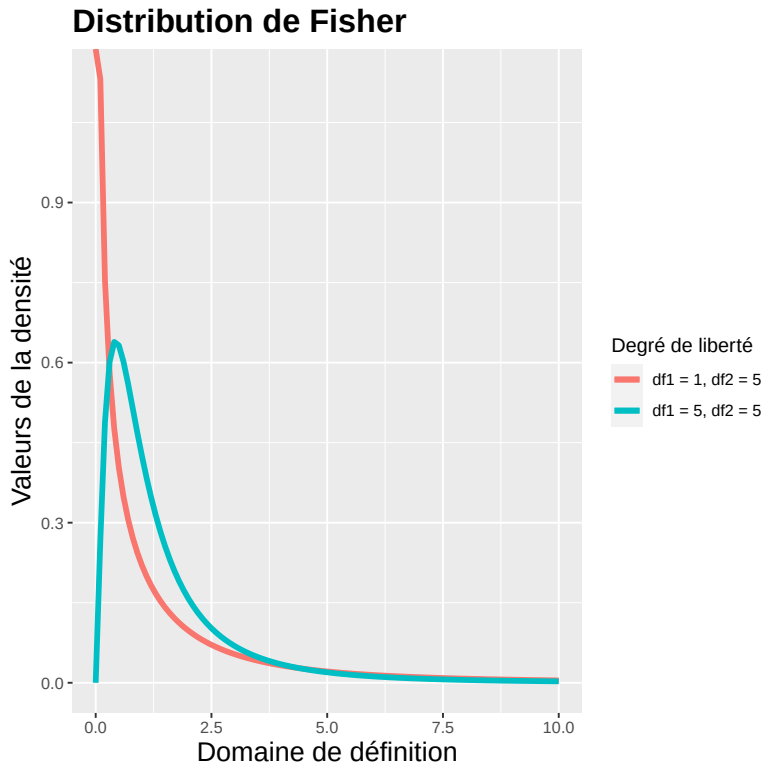

FIGURE 10 – Représentation de la distribution de Fisher pour deux configurations différentes, en rouge avec 1 et 5 degrés de liberté et en bleu avec 5 et 5 degrés de liberté.

où $SCE_{\text{total}} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_{i,k} - \bar{y})^2$ n'est rien d'autre que la variance totale de l'ensemble des groupes. La relation précédente n'est rien d'autre qu'une décomposition de cette variance.

Test statistique. Mais revenons à notre test, on rappelle que l'on souhaite étudier si les moyennes entre les différents groupes est différente ou non. Pour cela, on utilisera la statistique de test définie par

$$F = \frac{\dfrac{SCE_{\text{facteur}}}{K-1}}{\dfrac{SCE_{\text{résidu}}}{n-K}} \sim F(K-1, N-K).$$

Cette statistique de test est distribuée selon une loi de Fisher à $K-1$ et $N-K$ degrés de liberté. Vous notez au passage que cette loi de probabilité dépend de deux paramètres. De plus amples informations sont données sur cette loi en Annexe du document. Une représentation graphique de la loi de Fisher est donnée en Figure 10.

Pour déterminer si on rejette ou non l'hypothèse H_0 on procède ici à un **test unilatéral supérieur**, *i.e.* on va comparer la valeur de la statistique de test F au quantile d'ordre $1-\alpha$ d'une loi de Fisher à $K-1$ et $N-K$ degrés de liberté :

- on rejette H_0 si $f \geq F_{1-\alpha}(K-1, N-K)$,
- on ne rejette pas H_0 sinon.

Nous pourrions également calculer la p-value et la comparer au risque de première espèce α.

	Montagne	Mer	Hôtel-Club
Moyenne : $\bar{y}_k$	1 000	980	1050
Variance : s_k^2	100	150	75

TABLE 2 – Tableau de données concernant les dépenses effectuées par les voyageurs selon le type de séjour.

Dans ce cas la p-value est donnée par

$$\mathbb{P}\left[F(K-1, N-K) \geq F\right].$$

Exemple. Prenons un petit exemple pour illustrer cette notion en reprenant notre exemple des vacances, on souhaite savoir si le type de vacances : *montagne*, *mer* ou *hôtel club* a une influence sur les dépenses des vacanciers pendant leur séjour.
Pour cela, une enquête a été réalisée sur un total de $N = 600$ personnes : $n_k = 200$ par type de vacances. Les données collectées ont été traitées et sont résumées dans la Table 2 et les distributions sont représentées sur la Figure 11 gauche.

A partir de ce tableau de données, nous devons donc calculer nos deux termes de variances, ce que nous avons appelé *la moyenne des variances* et *la variance des moyennes*. Ici le calcul sera plus simple car les échantillons ont tous la même taille entre les différents groupes.
Nous devons donc :

- calculer la *variance des moyennes, i.e.* SCE_{facteur}, pour cela on commence par évaluer notre moyenne globale $\bar{y}$ qui est égale à

$$\bar{y} = \frac{1}{K}\sum_{k=1}^{K} \bar{y}_k = \frac{1000 + 980 + 1050}{3} = 1010.$$

  On a donc

$$\begin{aligned} SCE_{\text{facteur}} &= \sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2, \\ &= 200 \times \left((1000-1010)^2 + (980-1010)^2 + (1050-1010)^2\right), \\ &= 520000. \end{aligned}$$

- calculer la *moyenne des variances, i.e.* $SCE_{\text{résidu}}$:

$$\begin{aligned} SCE_{\text{résidu}} &= \sum_{k=1}^{K} (n_k - 1)s_k^2, \\ &= 199 \times \left(100^2 + 150^2 + 75^2\right), \\ &= 7586875. \end{aligned}$$

Ce qui bous donne une statistique de test de Fisher f égale à :

$$F = \frac{\dfrac{SCE_{\text{facteur}}}{K-1}}{\dfrac{SCE_{\text{résidu}}}{n-K}} = \frac{\dfrac{520000}{3-1}}{\dfrac{7586875}{600-3}} = 20.46.$$

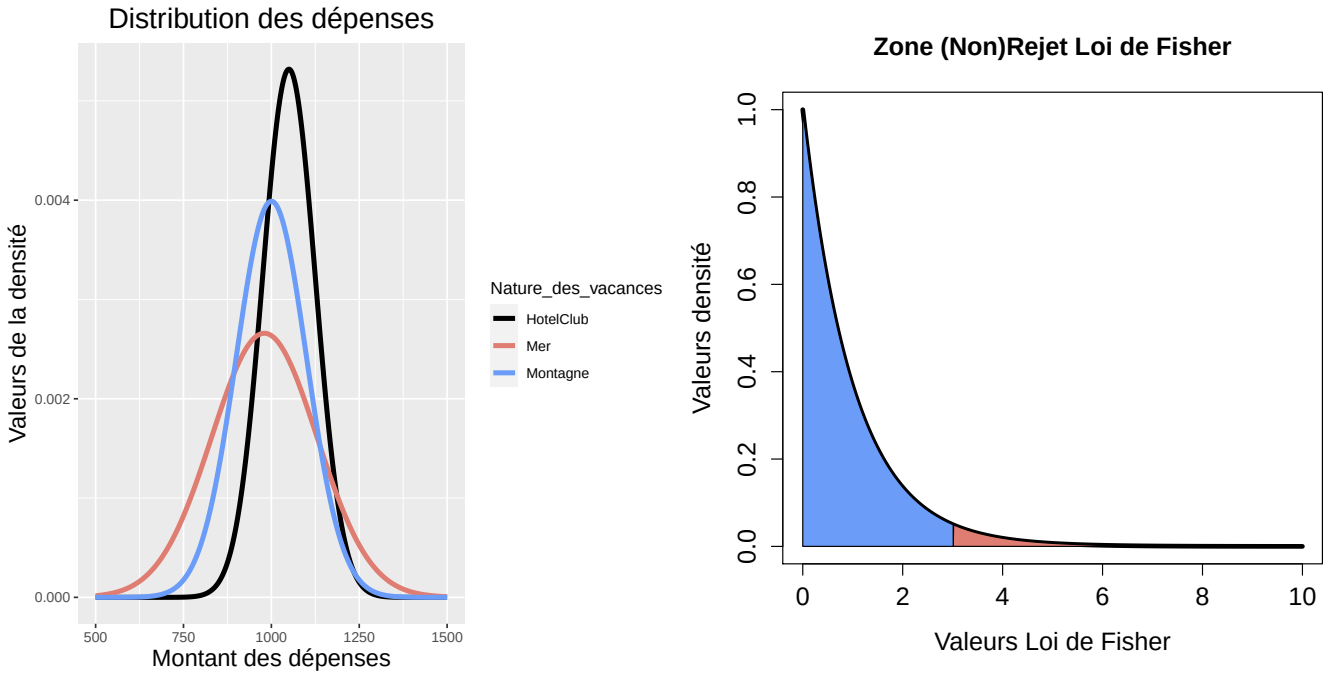

FIGURE 11 – A gauche la distribution des dépenses selon la nature des vacances. A droite, on représente la loi de Fisher à 2 et 597 degrés de liberté ainsi que la zone de rejet de l'hypothèse H_0 en rouge et de non rejet de l'hypothèse H_0 en bleue.

Or le quantile de d'ordre 0.95 d'une loi de Fisher à 2 et 597 degrés de liberté est égal à $F_{1-\alpha}(2, 597) = 3.01$. On représente la densité de loi de Fisher en Figure 11 (droite) pour visualisation du résultat.
On rejette donc l'hypothèse H_0 et on peut conclure que les dépenses ne sont pas les mêmes selon le type de séjour, *i.e.* que la nature du séjour a une influence sur le montant des dépenses effectuées par les vacanciers.

Pour finir. Notez une chose très importante, l'analyse de variances ne vous permettra pas de dire quel groupe a une moyenne significativement différente des autres, ni même combien de groupes ont des moyennes qui sont différentes d'au moins un autre groupe. **Vous pourrez seulement conclure qu'il existe un groupe dont la moyenne est différente d'au moins un autre groupe.** Mais cela est déjà suffisant pour conclure qu'un facteur externe à une influence sur les valeurs observées, *i.e.* que la variable quantitative étudiée est dépendante de la variable qualitative.

7.2 Pour s'entraîner

Exercice 1. Un data scientist d'une société d'assurance est chargé d'étudier l'impact d'une campagne de publicité réalisée dans 7 régions dans lesquelles la société est déjà implantée. Pour ceci, il a extrait de la base de données, pour un certain nombre d'agents généraux de chaque région, le nombre de nouveaux clients récoltés :

Région	1	2	3	4	5	6	7
Nb d'agents généraux	9	7	7	6	7	6	6
Nb moyen de nouveaux clients	26.88	22.34	19.54	18.95	27.17	25.87	25.72
Variance du nb de nouveaux clients	13.54	12.59	12.87	13.42	13.17	12.56	12.64

L'ingénieur statisticien décide alors de réaliser une analyse de variance afin de tester si le facteur

région a une influence sur le nombre de nouveaux clients récoltés. On appelle X_k^i le nombre de nouveaux clients du $i-$ème agent général de la région k. Soit n_k le nombre d'agents généraux de la région k, et K le nombre de régions ($K = 7$). Nous supposons que les variables aléatoires X_k^i sont normales, de moyenne μ_k et de variance σ.

Dans la suite on posera

$$\bar{X}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} X_k^i \quad \text{et} \quad \bar{X} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} X_k^i \quad \text{et} \quad N = \sum_{k=1}^{K} n_k.$$

1. Rappeler les hypothèses de l'ANOVA, sont-elles vérifiées ici ?
2. Formuler les hypothèses H_0 et H_1.
3. Interpréter $\bar{X}_k$ et $\bar{X}$.
4. On rappelle que la variance d'un échantillon peut se décomposer de la façon suivate

   $$SCE_{\text{totale}} = SCE_{\text{facteur}} + SCE_{\text{résidu}}.$$

   Dit autrement, que la variance totale de l'échantillon est égale à la variance inter-classe plus la variance intra-classe.

   (a) Rappeler la définition des différentes quantités précédentes et les calculer.

   (b) Déterminer si la campagne de publicité a eu le même impact dans toutes les régions. Conclure au risque de première espèce $\alpha = 0.05$. Pour cette question, on ne demande pas de calculer la p-value mais plutôt à se comparer à la valeur critique égale à 2.33 : quantile d'ordre $1 - \alpha$ d'une loi de Fisher à $K - 1$ et $N - K$ degrés de liberté.

8 Corrélations entres variables qualitatives et test du $\mathcal{X}^2$

Jusqu'à présent, nous avons essentiellement travaillé avec des variables quantitatives qui pouvaient représenter une grandeur physique. L'étude effectuée à la section précédente a montré une esquisse de l'étude de corrélations entre *une variables quantitative et une variable qualitative.*
On se propose maintenant de poursuivre dans cette direction en étudiant les éventuelles corrélations entre deux variables qualitatives, *i.e.* deux variables ne prenant comme valeurs que des modalités comme la variable *groupe sanguin d'un individu.*

8.1 Quelques rappels

L'analyse de corrélations entre deux variables qualitatives repose sur un test du Khi^2 que l'on notera aussi $\mathcal{X}^2$. Le nom de ce test doit vous faire passer à une loi statistique que l'on a déjà rencontré plutôt dans ce document ! Pour la petite histoire, le test du $\mathcal{X}^2$ est un test qui permet de comparer plusieurs **distributions** pour savoir si ces dernières sont identiques ou non ! Finalement, nous pourrions voir l'analyse de corrélations entre deux variables qualitatives comme une étude de la distribution d'une variable en fonction des valeurs prises par une autre variables (les probabilités étudiées sont donc proportions).
Regardons sur un exemple comment tout cela fonctionne.

Contexte et exemple. Nous allons considérer deux variables aléatoires X et Y dont les valeurs possibles x_i et y_i sont des **modalités** (comme le groupe sanguin), donc deux variables **qualitatives**.

Prenant un exemple dans lequel on souhaite étudier les variables suivantes :

- X : groupe sanguin qui peut prendre les valeurs x_i suivantes : *A, B, AB* ou *O* (on ignore ici les rhésus pour simplifier l'exemple. Normalement une telle information est très importante dans un contexte médical !)
- Y : lieu de résidence qui peut prendre les valeurs y_i suivantes : *Paris* ou *Lyon.*

La question dont on souhaite apporter une réponse est la suivante : **est-ce que la distribution des groupes sanguins est la même quelque soit la ville étudiée ?** Cela peut se reformuler de la façon suivante : *est-ce que les deux variables X et Y sont indépendantes ?*

Imaginons que l'on dispose d'un échantillon pour effectuer cette étude et pour chaque poche de sang on note le groupe sanguin (qui peut donc prendre 4 modalités ou valeurs différentes) et la ville d'origine de la poche de sang (qui peut donc prendre 2 modalités différentes). La première chose que l'on va faire est de faire un *tableau croisé* (que l'on appellera aussi *matrice ou tableau de contingence*) dans lequel on va simplement noter les informations issues de notre échantillon. Dans notre exemple, nous aurons donc le tableau avec 2 lignes et 4 colonnes ci-dessous :

Ville \ Groupe Sanguin	A	B	AB	O
Paris	234	176	354	87
Lyon	116	84	146	43

Comment faire maintenant pour conclure à partir de ces chiffres ? C'est l'objectif de notre prochaine section où nous allons présenter la méthodes générale. Mais l'idée c'est de comparer des **distributions** dont les **probabilités** correspondent aux **proportions** dans les différentes villes étudiées.

Un peu de théorie. Supposons que l'on dispose deux variables aléatoires X qui prend K_X modalités et une variable Y qui prendrait alors K_Y modalités. Afin de savoir si les variables X et Y sont indépendantes on dispose d'un échantillon de n.

L'échantillon nous amène à étudier la table de contingence suivante à laquelle on ajoute quelques informations qui nous renseignent sur le nombre d'exemples appartenant à chaque groupe (on l'ajoute à la fin des lignes et colonnes du tableau).

X \ Y	modalité 1	modalité 2	$\cdots$	modalité K_Y	total
modalité 1	$n_{1,1}$	$n_{1,2}$	$\cdots$	n_{1,K_Y}	$n_{1,.}$
modalité 2	$n_{1,1}$	$n_{1,2}$	$\cdots$	n_{1,K_Y}	$n_{2,.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
modalité K_X	$n_{K_X,1}$	$n_{K_X,2}$	$\cdot$	n_{K_X,K_Y}	$n_{K_X,.}$
total	$n_{.,1}$	$n_{.,2}$	$\cdots$	$n_{.,K_Y}$	n

où :

- $n_{i,j}$ désigne le nombre d'individus ayant la modalité i pour la variable X et la modalité j pour la variable Y
- pour tout $i = 1, ..., K_X$ on a $n_{i,.} = \sum_{j=1}^{K_Y} n_{i,j}$ qui représente le nombre total d'individus ayant la modalité i pour la variable X
- pour tout $i = 1, ..., K_Y$ on a $n_{.,j} = \sum_{i=1}^{K_X} n_{i,j}$ qui représente le nombre total d'individus ayant la modalité j pour la variable Y
- $n = \sum_{i=1}^{K_X} \sum_{j=1}^{K_Y} n_{i,j}$ est le nombre total d'individus.

Ce premier tableau peut faire peur en première approche mais à la section suivante cela deviendra beaucoup plus concret lorsque l'on retournera à l'étude de notre exemple.

La question est de savoir ce que l'on va faire de ce tableau ... il faut garder à l'esprit que l'on souhaite étudier l'indépendance entre les deux variables X et Y, ce qui, si on reprend notre vocabulaire en terme de distribution et de probabilités, signifie que **la proportion d'individus ayant la modalité i pour la variable X est la même quelque soit la valeur de la variable Y.**

Ce que l'on va alors faire, c'est comparer les valeurs observées $n_{i,j}$ à des valeurs théoriques $c_{i,j}$ que l'on devrait observer si nous avons bien indépendance entre les deux variables. La statistique de test que l'on étudiera sera alors une "distance" d^2entre les valeurs observées $n_{i,j}$ et les valeurs théoriques sous l'hypothèse d'indépendance $c_{i,j}$ qui est définie par :

$$d^2 = \sum_{i=1}^{K_X} \sum_{j=1}^{K_Y} \frac{(n_{i,j} - c_{i,j})^2}{c_{i,j}} = \frac{\left(n_{i,j} - \dfrac{n_{i,.}}{n_{.,j}}\right)^2}{\dfrac{n_{i,.}}{n_{.,j}}}.$$

Test Statistique. D'un point de vue formel on posera les hypothèses suivantes :

- H_0 : les deux distributions sont identiques, *i.e.* les variables aléatoires X et Y sont indépendantes.
- H_1 : les variables X et Y sont corrélées

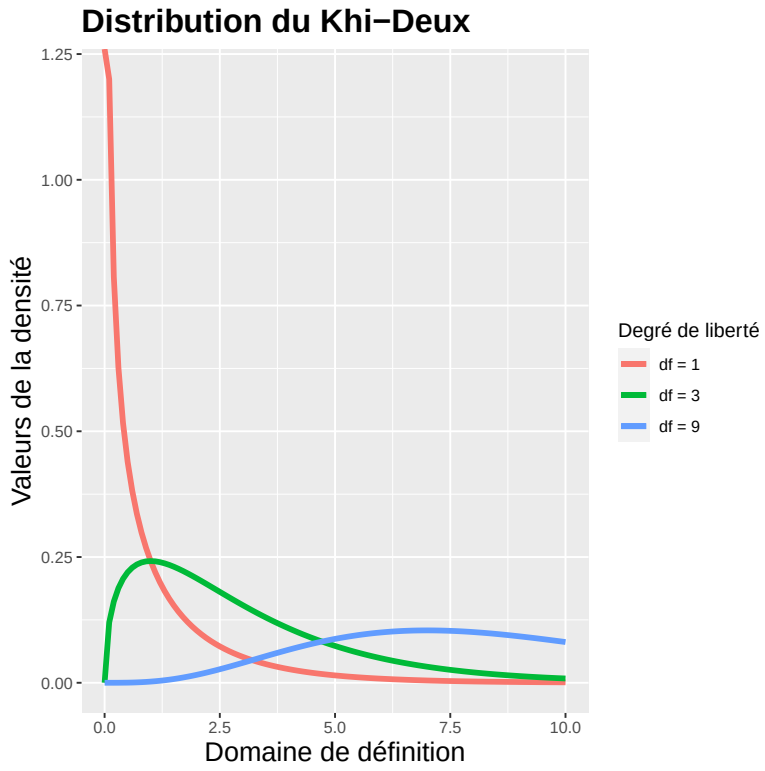

FIGURE 12 – Représentation de la distribution du $\mathcal{X}^2$ à 1 degré de liberté (en rouge), à 3 degrés de liberté (en vert) et à 9 degrés de liberté (en bleu).

On dressera ensuite notre table de contingence et on calculera la statistique de test d^2 sous l'hypothèse H_0. Le test statistique consistera à rejeter l'hypothèse H_0 au risque de première espace α si

$$d^2 > \mathcal{X}^2_{1-\alpha}\left((K_X - 1) \times (K_Y - 1)\right),$$

donc si la statistique de test d^2 est plus grande que la quantile d'ordre $1 - \alpha$ d'une loi du $\mathcal{X}^2$ à $(K_X - 1) \times (K_Y - 1)$ degrés de liberté (des exemples de distribution de loi du $\mathcal{X}^2$ sont présentés en Figure 12). On rappelle que K_X et K_Y représentent le nombre de degrés de liberté des variables X et Y respectivement.

Important : pour effectuer ce test, il faut cependant vérifier que tous les effectifs théoriques $c_{i,j} = \dfrac{n_{i,\cdot}}{n_{\cdot,j}}$ sont supérieurs à 5. Dans le cas contraire il faudra effectuer un autre test mais que l'on étudiera pas ici.

Remarque : dans le cas d'un test du $\mathcal{X}^2$ on remarque que ce dernier prend la forme d'un test *unilatéral supérieur*, *i.e.* la zone de rejet se trouve sur la partie droite de notre distribution.

Retour à notre exemple. On commence par formuler les hypothèses de travail pour notre test statistique :

- H_0 : le groupe sanguin est indépendant de la Ville, *i.e.* on observe la même proportion d'individus possédant un groupe sanguin donné peu importe la ville.
- H_1 : les deux variables étudiées sont corrélées.

On peut maintenant repartir de notre table de contingence à laquelle on ajoute le nombre total d'individus possédant les modalités d'une variable donnée.

Ville \ Groupe Sanguin	A	B	AB	O	total
Paris	234	176	354	87	851
Lyon	116	84	146	43	389
total	350	260	500	130	1240

Il nous faut maintenant calculer les effectifs théoriques $c_{i,j}$ sous l'hypothèse H_0.

On pourra utiliser la relation $c_{i,j} = \frac{n_{i,\cdot}}{n_{\cdot,j}}$ et on prendra soin de vérifier que ces effectifs théoriques sont bien supérieurs à 5.
On dresse ci-dessous la table des effectifs théoriques sous l'hypothèse H_0 :

Ville \ Groupe Sanguin	A	B	AB	O
Paris	240	178	343	89
Lyon	110	82	157	41

Prenons le cas des individus avec le groupe sanguin A. Si les deux variables *Groupe Sanguin* et *Ville* sont indépendantes, nous devrions avoir la même proportion (par rapport aux nombres d'individus dans chaque ville) d'individus qui possèdent le groupe sanguin A à Lyon et Paris.

On dénombre 350 individus avec le groupe sanguin A. On compte 851 individus à Paris et 389 à Lyon, donc notre échantillon comporte une proportion de $\frac{851}{851+389}$ d'individus Parisien et une proportion de $\frac{389}{851+389}$ d'individus Lyonnais. Sous l'hypothèse H_0 nous devrions donc avoir :

- $\frac{851}{851+389} \times 350 = 240$ Parisien avec le groupe sanguin A et
- $\frac{389}{851+389} \times 350 = 110$ Lyonnais avec le groupe sanguin A.

On peut refaire le même raisonnement avec les groupes sanguins B, AB et O. On vérifie que tous nos effectifs théoriques sont bien supérieurs à 5. On peut alors calculer notre statistique de test d^2 qui est ici égale à 1.815 et qui est distribuée selon une loi du $\mathcal{X}^2$ à 3 degrés de liberté. Or le quantile d'ordre 0.95 d'une loi du $\mathcal{X}^2$ à 3 degrés de liberté est égal à $\mathcal{X}^2_{0.95}(3) = 7.81$.

On ne rejette donc pas l'hypothèse H_0 et on peut donc conclure que la distribution des groupes sanguins est indépendante de la ville étudiée.

8.2 Pour s'entraîner

Exercice 1. Au sein d'une université, nous avons interrogé un panel de 325 étudiants en deuxième année de licence issus de trois filières différentes afin d'estimer la proportion de réussite au sein des différentes filières. Les résultats de cette enquête sont présentés dans la table ci-dessous :

Réussite \ Filière	Eco-Gestion	Anthropologie	Marketing
Oui	96	34	145
Non	22	12	16

Peut-on affirmer que la réussite de l'étudiant est indépendant de la filière dans laquelle il se trouve (au risque d'erreur de 10%) ?

Indication : le quantile d'ordre 0.9 d'une loi du $\mathcal{X}^2$ à 2 degrés de liberté est égal à 4.60.

Exercice 2. Sur 2000 personnes interrogées à Lyon, 1040 affirment utiliser régulièrement les transports en commun. Sur 1500 interrogées à Paris, 915 affirment la même chose. Est-ce que ces résultats permettent de soutenir que les Lyonnais et les Parisiens utilisent autant les transports en commun (risque de 5%) ?

Indication : le quantile d'ordre 0.95 d'une loi du $\mathcal{X}^2$ à 1 degré de liberté est égal à 3.84.

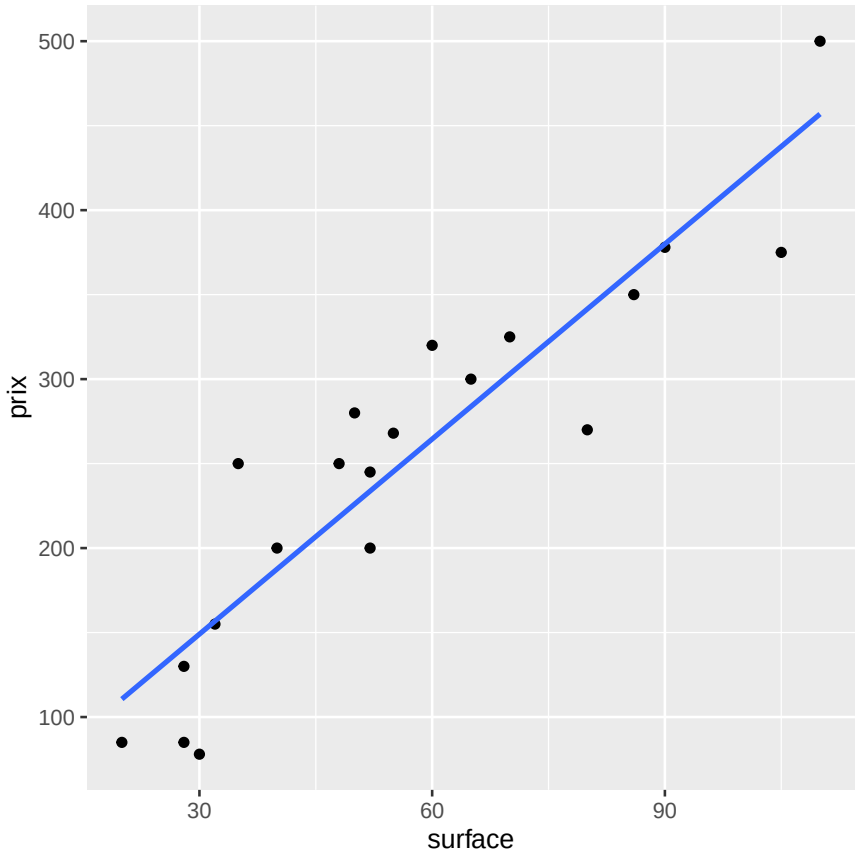

FIGURE 13 – Exemple d'un modèle de régression. On cherche ici à estimer le prix d'un bien immobilier en fonction de la surface de ce bien.

9 Modèles linéaires et corrélations entre variables quantitatives

Cette dernière partie aborde le dernier type de corrélation que nous rencontrerons dans le cadre de ce cours, à savoir, la corrélation entre *deux variables quantitatives réelles*. Jusqu'à présent, nous avons que nous pouvions tester l'indépendance ou la corrélation entre deux deux variables qualitatives avec un test du $\mathcal{X}^2$ et l'indépendance entre une variable qualitative et une variable quantitative avec un test de Student (dans le cas où la variable qualitative prend exactement deux modalités) ou une ANOVA (Analyse de Variance) sinon.

L'étude de la corrélation entre deux variables quantitatives implique l'étude de ce que l'on appelle un modèle linéaire dont un exemple est donné en Figure 13. Il s'agit ici de trouver la droite qui approxime le mieux les données et cette droite est très riche en information car elle va permettre :

1. d'estimer le prix en fonction de la surface du bien, ce qui est couramment fait lorsque l'on cherche à estimer la valeur de son bien, on regarde son prix au m^2.
2. cette droite va également nous permettre de quantifier l'intensité de la relation entre les deux variables étudiées, ici *surface* et *prix*.

Le premier point ne serait pas traité ici, il est beaucoup plus technique mais il n'est pas sans lien avec le deuxième point : l'étude de la corrélation entre deux variables.

Pour la petite histoire ce type de droite s'appelle un *modèle linéaire gaussien simple* et il suppose que **la variable à expliquer** Y (le prix) s'exprime comme une fonction affine de la variable explicative

(la surface) X. Le modèle s'écrit ainsi de la façon suivante :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{où} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

où ε_i est une variable aléatoire distribuée selon une loi gaussienne centrée et avec une certaine variance σ^2 inconnue. Cette variable aléatoire représente un bruit dans les données au moment de l'estimation de Y_i, du prix. C'est une variabilité présente dans les données et qui ne sera pas forcément explicable par les données. Enfin, ce terme peut également se voir comme une sorte d'erreur d'estimation.

9.1 Quelques rappels

Des critères empiriques de dépendance. Retournons maintenant sur notre étude de corrélation entre deux variables et commençons par regarder comment définir la corrélation entre deux variables quantitative, cela se fait à l'aide de la notion de *covariance*.

Définition 7. *Soient X et Y deux variables aléatoires réelles. La covariance des variables aléatoires X et Y, notée $Cov(X, Y)$ est un nombre réel défini par :*

$$Cov(X, Y) = \mathbb{E}\left[(\mathbb{X} - \mathbb{E}[\mathbb{X}])(\mathbb{Y} - \mathbb{E}[\mathbb{Y}])\right].$$

En particulier, nous avons $Cov(X, X) = Var(X)$,i.e. lorsque $X = Y$ on retrouve la définition de variance.

Supposons maintenant que l'on dispose d'un échantillon $\{(x_i, y_i)\}_{i=1}^n$, alors on peut estimer une valeur empirique de cette covariance par la relation :

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

La covariance est donc un nombre réel qui va permettre de décrire la tendance qu'ont les données à être supérieure à leur moyenne. On la retrouve très souvent quand il s'agit de quantifier la relation entre deux variables quantitatives. Par exemple, dans le cas d'un modèle linéaire, cette quantité se retrouve directement dans le *coefficient directeur* de la droite de régression. Elle présente cependant un inconvénient majeur, ses valeurs dépendent fortement de l'ordre de grandeur des données utilisées, ainsi des variables ayant une covariance élevée n'est pas forcément synonyme d'une forte dépendance entre les deux variables.

Corrélation linéaire. Pour pallier à cela, on va s'intéresser à une autre quantité que l'on appelle le *coefficient de corrélation linéaire de Pearson* (ou coefficient de corrélation linéaire).

Définition 8. *Soient deux variables aléatoires réelles X et Y. Le coefficient de corrélation linéaire de Pearson $\rho(X, Y)$ est défini par :*

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\mathbb{E}\left[(\mathbb{X} - \mathbb{E}[\mathbb{X}])(\mathbb{Y} - \mathbb{E}[\mathbb{Y}])\right]}{\sqrt{Var(X)Var(Y)}}.$$

Il s'agit d'un nombre réel compris entre -1 et 1 qui va permettre de quantifier la dépendance entre deux variables. A nouveau si' l'on dispose d'un échantillon $\{(x_i, y_i)\}_{i=1}^n$, alors on peut estimer une valeur empirique de ce coefficient par :

$$\rho(X, Y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}}.$$

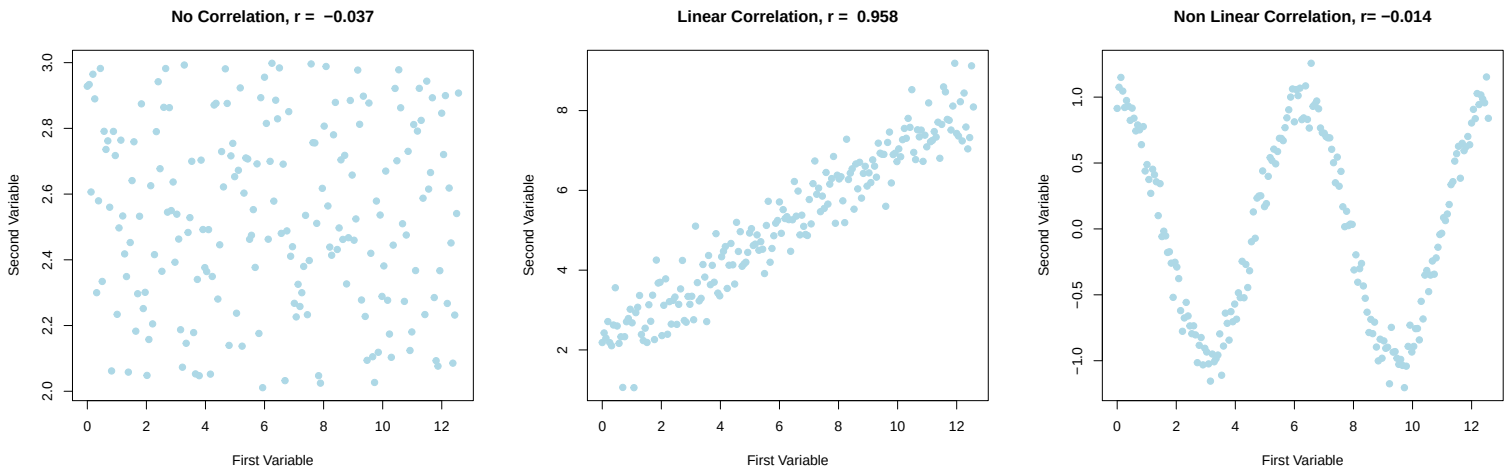

FIGURE 14 – Exemple de différentes corrélations avec les coefficients de corrélation associés : absence de corrélation - corrélation linéaire - corrélation non linéaire.

Plusieurs choses sont à formuler concernant cette définition, notée d'abord l'emploi du terme **linéaire**. Cela sous-entend deux choses :

1. ce coefficient permet de quantifier la dépendance entre les deux variables étudiées uniquement si cette dépendance est linéaire
2. la dépendance entre deux variables peut ne pas être linéaire, comme on peut le voir en Figure 14.

Le coefficient de corrélation s'interprète de la façon suivante selon ses valeurs :

- si $\rho(X, Y) < 0$: cela signifie que des valeurs croissantes de la variable X implique des valeurs décroissantes dans la variable Y. Nous aurons donc une droite décroissante, *i.e.* avec une pente négative,
- si $\rho(X, Y) > 0$: cela signifie que des valeurs croissantes de la variable X implique des valeurs croissantes dans la variable Y. Nous aurons donc une droite croissante, *i.e.* avec une pente positive,
- si $\rho(X, Y) \simeq 0$: cela signifie que les deux variables ne présentent pas de dépendance linéaire.

Des données peuvent présentées une certaine dépendance, notamment sur la figure la plus à droite, pour autant le coefficient de corrélation linéaire peut-être nul. Dans ce cas, il faudrait trouver d'autres critères pour mesurer cette corrélation.

Enfin, une dernière remarque concernant le coefficient de corrélation linéaire, plus ce dernier est proche de 1 ou -1 plus le lien entre les deux variables est fort ! Mais attention, une valeur proche de 1 ne signifie qu'il y a nécessairement une corrélation forte entre les deux variables étudiées.

Prenons un exemple trivial ou l'on cherche à estimer la longueur des cheveux des individus en fonction de leur taille. Pour cela on dispose d'un certain échantillon sur lequel a été effectué une série de mesure et les données sont représentées en Figure 15 à gauche.

D'après ce graphique, plus l'individu est grand plus ses cheveux sont cours, on a donc une corrélation linéaire assez forte de l'ordre de -0.6 comme le montre la valeur du coefficient de corrélations linéaire sur ces données. Enfin a priori ... en effet, parfois une corrélation entre deux variables est due à des facteurs externes qui ne sont pas pris en compte dans l'étude, dans le cas présent le sexe de l'individu.
En effet, on sait bien qu'en moyenne les femmes sont plus petites que les hommes et que ces dernières

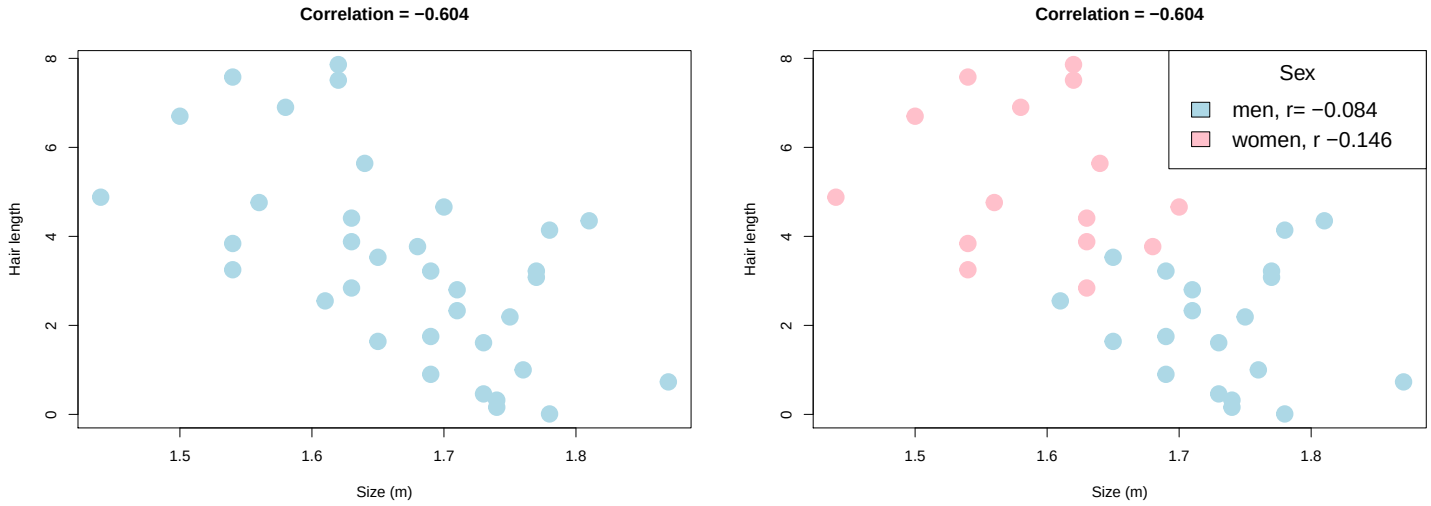

FIGURE 15 – Exemple de corrélation induite par un facteur exogène/extérieur et qui n'est pas pris en compte dans l'étude de la dépendance entre les deux variables.

ont souvent les cheveux plus longs que les hommes. On se propose alors d'étudier cette même dépendance entre les deux variables mais en tenant compte cette fois-ci du sexe de l'individu dans le calcul du coefficient de corrélation, les résultats sont présentés sur la Figure 15 à droite. On voit cette fois-ci que sur chacun des groupes, le coefficient de corrélation linéair est proche de 0 et qu'il n'y a donc pas de lien explicite entre les deux variables étudiées.

Nous avons vu comment caractériser l'intensité de la corrélation entre deux variables, il serait maintenant intéressant de voir si cette corrélation est significative d'un point de vue statistique.

Une étude statistique : test Ce que nous allons donc chercher à faire c'est savoir si la corrélation entre deux variables quantitatives est significative ou non. Pour cela, on va regarder si le coefficient de corrélation linéaire ρ est significativement différent de 0. On formule donc le test :

$$H_0 : \rho = 0 \quad \text{v.s.} \quad H_1 : \rho \neq 0.$$

La statistique de test que l'on va alors étudiée est la suivante :

$$T = \frac{\hat{\rho} - \rho}{\sqrt{\dfrac{1-\hat{\rho}^2}{n-2}}},$$

qui est distribuée selon une loi du Student à $n-2$ degrés de liberté. Sous l'hypothèse H_0, cette statistique de test devient

$$t = \frac{\hat{\rho}}{\sqrt{\dfrac{1-\hat{\rho}^2}{n-2}}}.$$

Ainsi, pour savoir si l'on rejette ou non l'hypothèse H_0, nous avons deux possibilités. S'agissant d'un test bilatéral, nous pouvons :

1. déterminer la valeur critique et la comparer à la valeur de la statistique de test. Ainsi, on rejette l'hypothèse H_0 si
$$|t| > T_{1-\alpha/2}(n-2),$$
i.e. si en valeur absolue la statistique de test prend des valeurs plus grandes que la quantile d'ordre $1-\alpha/2$ d'une loi de Student à $n-2$ degrés de liberté.

2. calculer la p-value et la comparer au risque de première espèce α, *i.e.* faire la comparaison

$$2 \times \mathbb{P}[T_{n-2} > |t|].$$

Si on rejette l'hypothèse H_0, alors on suppose que le coefficient de corrélation linéaire est significativement différent de 0. Ainsi on conclura que les deux variables sont linéairement dépendantes.

Regardons plus en détail cela sur un exemple afin de rendre les choses plus claires.

Exemple. On considère le jeu de données présenté dans la table ci-dessous qui représente les notes obtenus à un candidat à un examen en fonction de son âge. Notre objectif est de savoir s'il existe un lien entre ces deux variables. On va supposer que l'âge du candidat est une variable quantitative continue.

Âge	18	20	32	25	38	45	32	25	34	56	62	30
Note	10	8	9	9	5	4	6	7	6	4	2	8

Dans un premier temps, on va donc calculer notre coefficient de corrélation linéaire, puis nous chercherons à savoir si ce dernier est significativement différent de 0.

Le détail des calculs n'est pas fourni ici, et les résultats peut être facilement obtenu avec un tableau Excel par exemple. Mais nous aurons besoin de calculer la moyenne et variance des variables $X = Note$ et $Y = Age$, ces éléments sont donnés dans la table ci-dessous :

$\bar{X}$	$\bar{Y}$	$Var(X)$	$Var(Y)$	$Cov(X,Y)$	$\rho(X,Y)$
34.08	6.5	196.91	5.91	-29.71	-0.91

Il nous faut maintenant vérifier que ce coefficient est significativement différent de 0. On va donc tester l'hypothèse H_0 selon laquelle $\rho = \rho(X,Y) = 0$. Comme d'habitude, nous effectuerons notre test au risque d'erreur $\alpha = 0.05$.
On rappelle, que sous cette hypothèse, la statistique de test employée t est distribuée selon une loi de Student à $n-2$ degrés de libertés (soit 10 ici) et qu'elle est définie par :

$$t = \frac{\hat{\rho}}{\sqrt{\frac{1-\hat{\rho}^2}{n-2}}} = \frac{-0.91}{\sqrt{\frac{1-0.91^2}{12-2}}} = -6.94$$

Il ne nous reste plus qu'à utiliser la table de Student et chercher la valeur critique correspondante. On va donc chercher la valeur du quantile d'ordre $1-\alpha/2 = 0.975$ de la loi de Student à 10 degrés de liberté et cette valeur est égale à 0.228 d'après la table fournie en annexe.

On conclut donc au rejet de l'hypothèse H_0, ainsi les variables *Note* et *Age* sont corrélées.

On souhaite maintenant aller un peu plus loin et trouver le modèle linéaire qui décrit le mieux le lien entre ces deux variables. Si l'on dispose de l'âge de l'individu il est certainement possible de prédire la note obtenue à l'examen vu qu'il existe un lien linaire entre ces deux variables.
Regardons comment faire cela.

Vers un modèle de prédiction Nous allons revenir à notre modèle linéaire simple gaussien présenté en introduction de cette section

$$Y = \beta_0 + \beta_1 X + \varepsilon_i, \quad \text{où} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

où $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

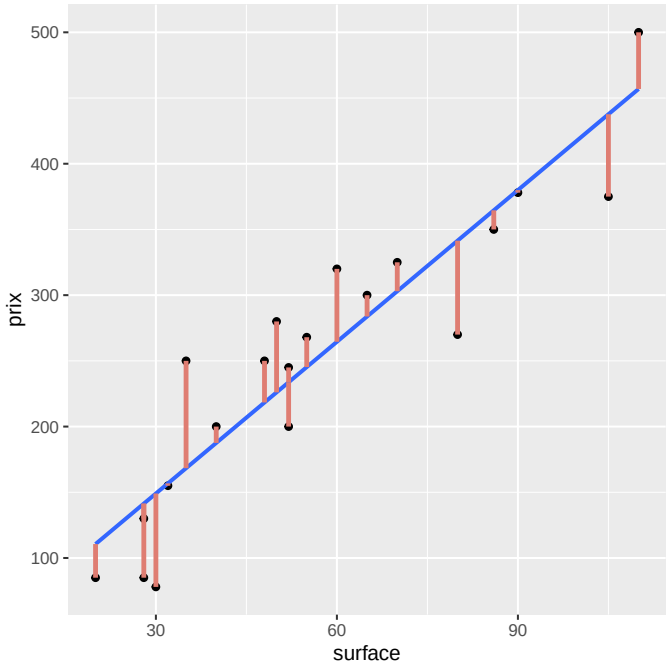

FIGURE 16 – Représentation des résidus de la régression linéaire ou modèle linéaire après apprentissage du modèle.

Ce modèle est dit *simple* car on cherche à estimer la valeur de Y en fonction d'une seule variable explicative X. On parlerait de modèle multiple dans le cas où l'on cherche à prédire la valeur de Y en fonction de plusieurs variables explicatives.
En outre, le modèle est dit gaussien car, l'erreur étant gaussienne et si l'on dispose d'une observation x_i de la variable aléatoire X alors on a

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Cela signifie que la valeur à prédire Y_i est une variable aléatoire car ε_i est une variable aléatoire. Plus précisément cette variable aléatoire est gaussienne : $Y_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$. Donc en particulier $y_i = \mathbb{E}[Y_i] = \beta_0 + \beta_1 x_i$. Pourquoi cette dernière valeur ? Car c'est en en fait elle qui va être utilisée comme prédiction de la variable aléatoire Y_i en fonction de l'observation x_i.

On voit donc que pour prédire la valeur de y_i, prédiction que l'on notera $\hat{y}_i$, il nous faut déterminer les valeurs de β_0 et β_1 qui approximent le mieux nos données, *i.e.* la meilleure *droite de régression* et cela se fait en cherchant les valeurs de paramètres qui minimisent la quantité suivante :

$$\min_{\beta_0,\beta_1} \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad = \quad \min_{\beta_0,\beta_1} \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 .$$

Si on reprend notre modèle cela revient donc à minimiser l'erreur de notre modèle, ce que l'on appelle aussi la somme des résidus quadratiques du modèle. Un résidu $\varepsilon_i = (\hat{y}_i - y_i)$ n'est rien d'autre que l'écart entre la valeur observée y_i et la valeur prédite $\hat{y}_i$. On peut en voir une représentation graphique en Figure 16.

Proposition 1. *Soit $\{(x_i, y_i)\}_{i=1}^n$ un jeu de données où tous les exemples sont indépendants et identiquement distribués alors la droite de régression de la forme $y_i = \beta_0 + \beta_1 x_i$ qui explique ou approximent le mieux les données est donnée par :*

$$\beta_0 = \bar{y} - \frac{Cov(x,y)}{Var(x)} \times \bar{x} \quad et \quad \beta_1 = \frac{Cov(x,y)}{Var(x)}.$$

Ce résultat montre que la covariance qui traduit une certaine corrélation entre les variables étudiées joue un rôle important dans le calcul de ces coefficients. En admettant ce résultat nous sommes alors capables de prédire la valeur de la variable réponse y_i en fonction de l'observation x_i.

Pour aller plus loin. Jusqu'à présent, nous avons étudier le modèle linéaire pour chercher à estimer une variable quantitative en fonction d'une autre variable quantitative uniquement.
Cependant, le modèle linéaire ne se limite pas à la seul étude des variables quantitatives, il est par exemple possible de chercher à estimer la valeur prise par une variable quantitative en fonction de **variables qualitatives** en considérant un modèle de la forme :

$$Y_{ij} = \beta_0 + \beta_j + S_{ij},$$

où seules les S_{ij} sont des variables aléatoires gaussiennes qui représentent une information non explicable par le modèle, on suppose en général que la variabilité est la même d'un groupe à l'autre et pour tous les individus. Les autres termes ont la signification suivante :

- Y_{ij} est la valeur de la variable réponse de l'individu i appartenant au groupe j
- β_0 est un **effet fixe** ou **une valeur moyenne** propre à l'ensemble des individus i indépendamment du groupe auxquels ils appartiennent
- β_j est un **effet fixe** ou **une valeur moyenne** propre au groupe j.

Si on remonte un peu plus loin dans ces notes, on se rend compte que l'on a déjà rencontré ce type de situation plus tôt, lorsque l'on cherchait à voir si la moyenne entre différents groupes d'individus est différente ou non ... lorsque l'on faisait de l'ANOVA !
En fait, l'Analyse de Variances n'est rien d'autre qu'un modèle linéaire. Vous vous souvenez que dans l'ANOVA, nous cherchions à savoir les moyennes entre les différents groupes sont différentes. Avec ce type de modèles, cela revient à tester si les paramètres β_j sont tous égaux à 0, si c'est le cas alors, on peut réécrire le modèle précédent

$$Y_{ij} = \beta_0 + S_{ij}.$$

Ce qui signifie que la moyenne des différents groupes est bien la même et égale à β_0.

Enfin, vous vous souvenez de la décomposition de la variance effectuée dans le cadre de l'ANOVA ... et bien nous pourrions effectuer une même décomposition dans le cas d'un modèle linéaire, ce qui permettrait de définir ce qu'est un bon modèle linéaire pour faire de la prédiction ! Mais on s'éloigne du sujet ...

9.2 Pour s'entraîner

Exercice 1. Au cours d'une étude, des chercheurs tentent de savoir si le poids du père de famille est corrélée au poids de leur fils aîné. Pour cela, ils ont interrogé 15 familles et ont mesuré le poids en kg des 15 pères de famille et de leur fils aîné.

Père	78	87	56	67	59	90	97	65	78	85	71	68	56	70	78
Fils	65	67	60	56	45	81	84	60	71	67	65	59	55	62	74

1. Quel reproche peut-on faire à cet étude ?
2. Quelle est la tendance générale ? Que peut-on dire de l'évolution du poids des fils ainés en fonction du poids de leur père ?
3. Déterminer le coefficient de corrélation linéaire ρ.
4. Peut-on dire que le poids du fils ainé est corrélé au poids du père de famille ? Conclure au risque de première espèce $\alpha = 0.05$.

5. (Bonus) On note Y la variable aléatoire décrivant le poids du fils aîné et X celle du père de famille. Déterminer les coefficients de la droite de régression

$$Y = \beta_0 + \beta_1 X$$

et représenter les données et cette droite à l'aide d'Excel.

Exercice 2. Nous souhaitons exprimer la hauteur y (en pieds) d'un arbre d'une essence donnée en fonction de son diamètre x (en pouces) à 1m30 du sol. Pour ce faire, nous avons mesuré 20 couples (diamètre,hauteur) et effectué les calculs suivants : $\bar{x} = 4.53$ et $\bar{y} = 8.65$ et

$$\frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})^2\,, \quad \frac{1}{20}\sum_{i=1}^{20}(y_i - \bar{y})^2 \quad \text{et} \quad \frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})\,(y_i - \bar{y})\,.$$

1. Déterminer le coefficient de corrélation et interpréter cette valeur.
2. Peut-on dire que la corrélation entre le diamètre de l'arbre et sa hauteur est significative ? Conclure au risque de première espèce $\alpha = 0.05$.
3. (Bonus) Déterminer les coefficients de la relation linéaire $y = \beta_0 + \beta_1 x$ à l'aide des données fournies.

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 $\mathcal{X}^2_k$. 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 $X = \sum_{i=1}^{k} X_i^2$ suit une loi du $\mathcal{X}^2$ à k degrés de liberté, notée $\mathcal{X}^2_k$.
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 $\mathcal{X}^2_n$, *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

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

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

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

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 $\bar{X}_n = \dfrac{1}{n}\sum_{i=1}^n X_i$, alors :

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

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.

$\Phi(x) = \mathbb{P}(X \leq x)$ où $X \sim \mathcal{N}(0,1)$ et $x = x_1 + x_2$

	x_2									
x_1	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.60	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.70	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.20	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.30	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.40	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.50	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.60	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.70	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.80	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.90	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.00	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.10	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.20	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.30	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.40	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.50	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.60	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.70	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.80	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.90	1	1	1	1	1	1	1	1	1	1

FIGURE 17 – 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 18 – Table des quantiles de la loi de Student $\mathcal{T}_k$

A.4 Examen Session 2020

Mathématiques et Statistiques appliquées à la Gestion

Etude de Cas

BBA-1 (2020-2021)

Résumé

Cet examen se présente comme étude de cas où vous serez plus ou moins guidés dans les questions. Cette étude abordera une très grande partie des notions vus en cours, y compris l'analyse de variance ou encore le test d'indépendance du $\mathcal{X}^2$ étudiés lors de la septième séance.
Pour l'évaluation, une grande importance sera accordée à la rédaction de vos réponses, *i.e.* une bonne réponse sans justification ne vous donnera que très peu de points. Pour cela, selon la nature de la question vous penserez à donner les détails suivants :

- **intervalle de confiance** : vous préciserez la loi sur laquelle se base la construction de votre intervalle de confiance en précisant le contexte
- **test d'hypothèse** : vous formulerez les hypothèses H_0 et H_1, la nature du test ainsi que la statistique de test employée et sa loi. Sauf mention explicite, vous pourrez conclure en construisant un intervalle de confiance ou en calculant la p-value p

Enfin, vous rédigerez votre étude sur le support de votre choix (papier, word, open office, pages, ...) et vous déposerez votre travail sur la Dropbox prévue à cet effet **en format pdf**. Le barème est donné à titre indicatif.
Les données sont présentées à la fin de ce document et sont également disponibles sous la forme d'un tableur afin de vous faciliter les calculs.

Remarques : il n'est pas nécessaire de détailler les applications numériques pour le calcul des moyennes et variances (ou écart-types), vous pouvez directement donner le résultat. Dans le calcul des p-value ou des quantiles u_α, vous utiliserez les valeurs plus proches dans le cas où la valeur recherchée ne figure pas dans les tables.

Etude des employées d'une entreprise et législation

Pour cette nouvelle année 2021, un grand groupe industriel national qui possèdent près de 200 entreprises dans l'hexagone souhaite effectuer une étude concernant ses salariés. Le grand siège situé à Paris, quartier de la Défense, a mandaté son service de Management et Ressources Humaines pour cette étude. Il souhaite essentiellement deux choses (i) étudier le niveau de rémunération de ses salariés ainsi que l'évolution de ce salaire et (ii) le PDG attachant énormément d'importance à l'égalité Homme-Femme, il souhaite savoir si cette égalité est respectée dans son entreprise conformément à la loi n°2014-873 du 4 août 2014.

Pour des raisons financières mais également humaines, le siège ne peut effectuer cette étude sur l'ensemble de ces entreprises, elle décide donc de mener son enquête auprès de trois entreprises situées dans la partie sud de la France : *Marseille, Bordeaux* et *Nice*. Chacune de ces entreprises comptent 15 salariés pour lesquels nous avons relevé les rémunérations annuelles nettes en 2019 et en 2020, le sexe du salarié ainsi que sa fonction (technicien ou ouvrier). Les données sont présentées en dernière page.

Pour des soucis de modélisation, nous supposerons que les salaires de 2019 et 2020 sont distribuées selon une loi normale dont les paramètres sont pour le moment inconnus.

I) Salaire, évolution et recrutement (*12 points*)

La première mission des services Management et Ressources Humaines est de vérifier la situation de ses salariés concernant leur rémunération. L'étude menée l'année précédente a conclu que la rémunération annuelle en 2019 suivait une loi normale de moyenne $\mu = 20,000$ et d'écart-type $\sigma = 2,000$.

- Quelles sont les limites de cette modélisation ?
- Peut-on confirmer que l'étude menée au cours de l'année 2020, sur les revenus 2019, est correcte ? On construira un intervalle de confiance de niveau $1 - \alpha = 0.9$.
- Selon cette même étude qu'elle était la proportion de salarié gagnant au moins $1,2$ fois le SMIC en 2019 ? On utilisera le fait que le SMIC annuel net s'élevait à 14 400 euros à cette période.

L'étude étant en cours, on ne connaît pas encore la distribution exacte des rémunérations des salariés pour l'année 2020, en revanche les chargés de l'étude font l'hypothèse raisonnable selon laquelle les données suivent une loi normale comme l'année précédente. En revanche, pour l'année 2020, l'écart-type σ est inconnu de même que la moyenne μ.

- Donnez une estimation au niveau de confiance $1 - \alpha = 0.90$ du paramètre de la moyenne μ pour la distribution des salaires.

Afin d'éviter de créer des émeutes au sein de ses entreprises, le groupe souhaite que le recrutement des nouveaux salariés se fasse à un niveau rémunération ne dépassant pas celui des 5% des rémunérations les plus faibles en 2020.

- Quel est le salaire net annuel qu'il devra alors verser à un nouveau salarié ?

Le groupe a fait de grands bénéfices au cours de l'année 2019. Suite à ces excellents résultats, le PDG avait promis une hausse de rémunération a ses salariés au cours de l'année 2020.

- Peut-on dire que le PDG a respecté ses engagements au risque $\alpha = 0.05$?

Enfin, pour initier l'étude sur la notion d'égalité, le groupe souhaite vérifier que le niveau de rémunération moyen, en 2020, est le même peu importe la localité de l'entreprise.

- Est-ce que les salariés ont le même niveau de rémunération à *Marseille, Bordeaux* et *Nice* ? Conclure au risque $\alpha = 0.05$.

II) Conformité à la législation (*8 points*)

Pour leur deuxième mission, les salariés mandatés doivent plus particulièrement vérifier que l'égalité Homme-Femme est bien respectée dans les entreprises du groupe.
Plus particulièrement, une loi du 22 décembre 1972 pose le principe de l'égalité de rémunération entre les hommes et les femmes, bien que ce principe soit ancestrale, force est de constater que ces inégalités persistent même près de 50 ans après la promulgation de cette loi.

- Peut-on affirmer, au risque de première espèce $\alpha = 0.05$ que les hommes et les femmes de cette entreprise ont le même niveau de rémunération ? On se basera sur les chiffres de l'année 2020.

Quel test aurait-il été plus judicieux d'effectuer ? (il n'est pas demandé de faire ce test).

Enfin, le 13 juillet 1983, la loi Roudy établit le principe d'égalité professionnelle entre les femmes et les hommes qui garantit les droits d'accès aux hommes et aux femmes aux même fonctions. Notre PDG souhaite être exemplaire en ce domaine en ayant la même proportion d'hommes et de femmes dans les différentes fonctions.

- Peut-on dire que l'entreprise compte autant d'hommes que de femmes ? Conclure au risque de première espace $\alpha = 0.9$.

- Est-il vrai de dire que les femmes et les hommes de cette entreprise sont égaux d'un point vue professionnel, *i.e.* est-ce que l'on dénombre autant de techniciens/ouvriers hommes que femmes ? Conclure au risque de première espèce $\alpha = 0.05$.

Données

Marseille			
Sexe	Fonction	Salaire 2019	Salaire 2020
homme	ouvrier	15 867	18 774
femme	ouvrier	18 828	21 381
homme	technicien	18 282	27 801
homme	ouvrier	16 613	21 759
homme	technicien	19 728	20 468
homme	ouvrier	19 723	27 326
homme	ouvrier	19 694	24 768
homme	ouvrier	19 379	25 669
homme	ouvrier	18 681	23 758
homme	technicien	19 653	22 380
homme	ouvrier	20 197	22 649
homme	technicien	15 317	22 758
homme	ouvrier	24 436	22 276
homme	technicien	19 382	25 962
homme	technicien	16 397	26 484

Bordeaux			
Sexe	Fonction	Salaire 2019	Salaire 2020
homme	technicien	18 567	25 404
femme	ouvrier	21 893	22 050
femme	technicien	19 819	21 993
homme	ouvrier	20 822	25 548
femme	ouvrier	21 861	22 453
homme	technicien	18 237	22 116
femme	ouvrier	19 405	21 001
homme	ouvrier	17 388	24 645
femme	ouvrier	19 589	21 233
femme	ouvrier	21 810	19 985
homme	ouvrier	21 048	25 544
homme	technicien	19 952	24 646
femme	ouvrier	17 659	20 316
homme	technicien	19 318	23 096
femme	ouvrier	18 235	22 866

Nice			
Sexe	Fonction	Salaire 2019	Salaire 2020
femme	ouvrier	17 147	19 426
femme	ouvrier	16 973	21 348
femme	technicien	20 210	17 442
homme	ouvrier	22 469	21 291
homme	technicien	18 300	22 565
homme	technicien	21 398	20 153
femme	technicien	21 048	20 464
femme	ouvrier	18 907	19 515
homme	technicien	21 773	21 185
femme	technicien	14 609	19 023
homme	ouvrier	18 745	23 592
femme	technicien	16 366	20 101
femme	ouvrier	16 146	17 965
homme	ouvrier	20 759	20 912
homme	ouvrier	19 788	21 361

TABLE 3 – Jeu de données relatif à l'entreprise pour les 3 villes étudiées : Marseille, Bordeaux et Nice. Pour chaque ville on dispose d'informations relatives au Sexe, à la Fonction ainsi que les revenus annuels 2019 et 2020 des salariés.

A.5 Maths pour l'examen à l'ordinateur

Signification	Symbole mathématique	Equivalent sur machine
Moyenne distribution	μ	mu
Ecart-type distribution	σ	sigma
Variance distribution	σ^2	sigma^ 2
Ecart-type échantillon	s	s
Variance échantillon	s^2	s^2
Risque d'erreur	α	alpha
Valeurs de références	p_0 (proportion) ou μ_0 (moyenne)	p0 ou mu0
Estimateur de la moyenne	$\bar{x}$	xbar ou barx
Estimateur de la proportion	$\bar{p}$	barp ou pbar
Hypothèses	H_0 et H_1	H0 et H1
Racine carré	$\sqrt{x}$	sqrt(x) ou rcd(x)
Division	$\frac{x}{y}$	x/y
Quantile d'ordre $1-\alpha/2$ (loi normale)	$z_{1-\alpha/2}$	z_{1-alpha/2}
Quantile d'ordre $1-\alpha/2$ (loi student)	$t_{1-\alpha/2}$	t_{1-alpha/2}
Probabilité	$\mathbb{P}[U < u_\alpha]$	P(U < u_{alpha})
Exemple pour l'écriture d'un intervalle de confiance de niveau $1-\alpha$	$\left[\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}};\ \bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$	[barx ± u_{1-alpha/2} x sigma/sqrt(n)]