1 Légende

Pas de symbole : à lire.
⋆ : Passage utile mais pas obligatoire pour la compréhension de la suite.
⋆⋆ : Passage difficile mais pas obligatoire pour la compréhension de la suite.
★ : passage très important.

2 Chapitre 1 : variables aléatoires-Lois de probabilité

2.1 Rappels et approfondissement

2.1.1 Ensemble dénombrable et familles sommables

Ensemble dénombrable : Ensemble E tel qu'il existe une fonction f de $\mathbb{N}$ vers E vérifiant :

- Pour tout $x \in E$, il existe un $n \in \mathbb{N}$ tel que $x = f(n)$ (on dit que f est surjective).
- Pour tout $n \in \mathbb{N}$ et $m \in \mathbb{N}$ tels que $n \neq m$, on a $f(n) \neq f(m)$ (on dit que f est injective).

Une fonction à la fois injective et surjective est dite "bijective".
Les ensembles $\mathbb{N}$, $\mathbb{Z}$ et $\mathbb{Q}$ sont dénombrables, mais $\mathbb{R}$ ne l'est pas.
⋆**Famille sommable :** Soit $(u_k)_{k\in\mathbb{N}}$ une suite de réels (famille). On dit que la famille $(u_k)_{k\in\mathbb{N}}$ est sommable si et seulement si pour tout σ fonction bijective de $\mathbb{N}$ sur $\mathbb{N}$, les suites $\sum_{k=0}^{n} u_{\sigma(k)}$ ont même limite (éventuellement infinie) lorsque n tend vers l'infini. La limite commune est alors notée $\sum_{k\in\mathbb{N}} u_k$.

On a :

- Si les réels u_k sont **tous positifs**, alors $(u_k)_{k\in\mathbb{N}}$ est sommable.
- Si les réels u_k sont **de signe quelquonque**, alors $(u_k)_{k\in\mathbb{N}}$ est sommable si et seulement si $(|u_k|)_{k\in\mathbb{N}}$ est sommable et $\sum_{k\in\mathbb{N}} |u_k| < +\infty$.

La définition des familles sommables se généralise aux autres espaces dénombrables. Soit E un espace dénombrable et f une bijection de $\mathbb{N}$ dans E. La famille $(u_k)_{k\in E}$ est sommable si et seulement si la famille $(u_{f(k)})_{k\in\mathbb{N}}$ est sommable, sa somme est alors $\sum_{k\in\mathbb{N}} u_{f(k)}$ et est notée $\sum_{k\in E} u_k$

2.1.2 Probabilités et tribus

Sur un ensemble dénombrable :
★ : Soit E un ensemble dénombrable, une probabilité sur E est une famille de réels positifs $(p_k)_{k\in E}$ telle que $\sum_{k\in E} p_k = 1$.

Toute partie d'un ensemble dénombrable étant dénombrable, on peut étendre la probabilité $(p_k)_{k\in E}$ en une fonction P de l'ensemble des parties de E (noté 2^E) vers $[0,1]$ par :

$$P(A) = \sum_{k\in A} p_k.$$

L'ensemble des parties dont on peut "mesurer" la probabilité est l'ensemble de toutes les parties de E, la probabilité d'une partie de E est donnée par $P(A)$.
★ : **Sur un ensemble quelquonque :**
Dans le cas précédent où E est dénombrable, la fonction P et l'ensemble des parties 2^E vérifient les propriétés :

1. $P(E) = 1$.
2. $P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n)$ pour tout $(A_n)_{n\in\mathbb{N}}$ famille de parties deux à deux disjointes.
3. $\emptyset \in 2^E$.
4. Si $A \in 2^E$ alors son complémentaire $\bar{A} \in 2^E$.
5. Si $(A_n)_{n\in\mathbb{N}}$ famille d'éléments de 2^E, alors $\bigcup A_n \in 2^E$.

Soit maintenant E un ensemble quelquonque, on appelera tribu un sous-ensemble $\mathcal{T}$ de 2^E (donc un sous-ensemble de parties de E) vérifiant :

1. $\emptyset \in \mathcal{T}$.
2. Si $A \in \mathcal{T}$ alors son complémentaire $\bar{A} \in \mathcal{T}$.
3. Si $(A_n)_{n\in\mathbb{N}}$ famille d'éléments de $\mathcal{T}$, alors $\bigcup A_n \in \mathcal{T}$.

On appelera probabilité une fonction définie sur une tribu $\mathcal{T}$ à valeurs dans $[0,1]$ vérifiant :

1. $P(E) = 1$.
2. $P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n)$ pour tout $(A_n)_{n\in\mathbb{N}}$ famille d' élements de $\mathcal{T}$ deux à deux disjointes.

★ : On pourrait penser naïvement que pour n'importe quel ensemble E, on peut mesurer la probabilité de n'importe quelle partie de E et donc de poser $\mathcal{T} = 2^E$, cependant même dans le cas réel ce n'est pas le cas. Un contre-exemple, non obligatoire à connaître est donné dans le paragraphe suivant.
★ : **Sur $\mathbb{R}$:**
$\mathbb{R}$ n'étant pas dénombrable, on ne peut pas définir la notion de probabilité à partir d'une famille (non dénombrable) $(p_k)_{k\in\mathbb{R}}$. En effet, les sommes infinies ne peuvent être définies pour un espace non dénombrable.
Sur $\mathbb{R}$, une probabilité continue est définie à partir d'une fonction F dite de répartition vérifiant :

1. F est continue et croissante.
2. $\lim_{x\to-\infty} F(x) = 0$ et $\lim_{x\to+\infty} F(x) = 1$.

On définit alors la probabilité des intervalles par :

$$P([a,b]) = P(]a,b]) = P([a,b[) = P(]a,b[) = F(b) - F(a), \tag{1}$$

pour $a \leq b$. En particulier, lorsque F est dérivable de dérivée f (ou plutôt si l'ensemble des points où ce n'est pas dérivable est au plus dénombrable), on a :

$$P([a,b]) = \int_a^b f(x)dx, \tag{2}$$

f est appelée **densité de probabilité**, c'est une fonction positive (car F est croissante) et vérifiant :

$$\int_{-\infty}^{+\infty} f(x)dx = 1. \tag{3}$$

$\star\star$: Lorsqu'une partie A s'écrit comme une réunion dénombrable d'intervalles deux à deux disjoints $A = \bigcup_{n\in\mathbb{N}} [a_n, b_n]$, on prolonge P par :

$$P(A) = \sum_{n\in\mathbb{N}} P([a_n, b_n]). \tag{4}$$

Afin de pouvoir raisonner comme dans le cas dénombrable, on se pose la question de savoir si toute partie de $\mathbb{R}$ peut s'écrire comme réunion dénombrable d'intervalles. Si cela était le cas, alors P vérifierait :

$$P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n), \tag{5}$$

pour tout $(A_n)_{n\in\mathbb{N}}$ famille d'élements de $2^{\mathbb{R}}$ deux à deux disjoints, et alors P serait une probabilité sur l'ensemble des parties de $\mathbb{R}$, on pourrait alors évaluer la probabilité de toutes les parties de $\mathbb{R}$.
Cependant, ce n'est pas le cas. En effet, considérons la fonction de répartition F définie sur $\mathbb{R}$ par :

$$F(x) = \begin{cases} 0 \text{ si } x \leq -1 \\ \frac{1}{3}(x+1) \text{ si } x \in [-1,2] \\ 1 \text{ si } x \geq 2 \end{cases} \tag{6}$$

La probabilité définie par F vérifie en particulier : pour toute partie V de $[0,1]$ telle que l'on peut définir la probabilité et pour tout $r \in [-1,1]$, on a $P(V+r) = P(V)$.
Supposons que l'on puisse étendre P à toutes les parties de $\mathbb{R}$. Pour tout $x \in [0,1]$, on définit l'ensemble $\bar{x} = \{y \in [0,1] : x - y \in \mathbb{Q}\}$. (Attention : si $x \neq x'$, il se peut que $\bar{x} = \bar{x'}$ lorsque $x - x' \in \mathbb{Q}$). Parmi les axiomes fondateurs des mathématiques, l'axiome du choix dit que de toute famille d'ensembles disjoints, on peut construire un ensemble en ne prenant qu'un seul élément dans chaque ensemble disjoint.
En utilisant l'axiome du choix, on construit l'ensemble V en prenant un seul élément dans chaque $\bar{x}$. L'ensemble V vérifie en particulier :

– $V \subset [0,1]$.
– Pour tout $x \in [0,1]$, il existe un unique $y \in V$ tel que $x - y \in \mathbb{Q}$.
– Pour tout couple (x, y) d'éléments distincts de V, $x - y \notin \mathbb{Q}$.

Considérons l'ensemble :

$$A = \bigcup_{r \in [-1,1] \cap \mathbb{Q}} (V + r). \tag{7}$$

Soit $x \in (V + r) \cap (V + r')$, alors il existe deux éléments y et y' de V tels que $x = y + r = y' + r'$, ce qui entraîne $y - y' = r' - r \in \mathbb{Q}$ et donc $y = y'$ d'où $r = r'$. Ainsi l'ensemble A est réunion dénombrable de parties disjointes.
Supposons tout d'abord que $P(V) = 0$, alors d'après l'hypothèse comme quoi on peut mesurer la probabilité de tous les ensembles, on en déduit $P(A) = 0$. Soit $x \in [0,1]$, alors il existe $y \in V$ tel que $x - y = r \in \mathbb{Q}$ de plus $r \in [-1,1]$ d'où $x \in V + r \subset A$. Ainsi, $[0,1] \subset A$ donc $P([0,1]) = 0$. Cependant, d'après la définition de P, on a $P([0,1]) = \frac{1}{3}$ ainsi $\frac{1}{3} = 0$ ce qui est contradictoire.
Supposons alors que $P(V) > 0$, alors par l'invariance par translation de P, $P(A) = +\infty$, ce qui est contradictoire avec le fait que P est une probabilité.
On en déduit alors que l'on ne peut mesurer la probabilité de V.
★ : D'après le contre-exemple précédent (qui n'est pas à retenir), on a vu qu'il existe des parties de $\mathbb{R}$ pour lesquelles on ne peut définir la probabilité. Mais quelles sont alors les parties de $\mathbb{R}$ pour lesquelles on peut définir la probabilité ? Ce sont :

1. Les intervalles :

$$P([a,b]) = P([a,b[) = P(]a,b]) = P(]a,b[) = \int_a^b f(x)dx, \tag{8}$$

2. Complémentaire d'un intervalle :

$$P(\mathbb{R} \backslash [a,b]) = 1 - P([a,b]). \tag{9}$$

3. Si $A = \bigcup_{n \in \mathbb{N}} A_n$ est une famille dénombrable d'ensembles disjoints telle que A_n soit un intervalle ou complémentaire d'intervalle :

$$P(A) = \sum_{n \in \mathbb{N}} P(A_n). \tag{10}$$

Plus généralement, on définit la tribu borélienne $\mathcal{B}_{\mathbb{R}}$ comme la plus petite tribu contenant les intervalles. Les éléments de $\mathcal{B}_{\mathbb{R}}$ sont des réunions dénombrables d'ensembles vérifiant les points 1. 2. et 3. ainsi que de complémentaires d'ensembles vérifiant 1. (complémentaire d'intervalle) 2. (intervalle) 3. (intersection dénombrable de complémentaires d'intervalles).

2.1.3 ★ Variables aléatoires et construction de variables aléatoires

On définit :
Ω population.

$\mathcal{A}$ une tribu sur Ω : l'ensemble des évênements.
P une probabilité définie sur $\mathcal{A}$.
Soit E un ensemble (ensemble de valeurs) et $\mathcal{T}$ une tribu sur E. Une fonction X de Ω dans E est une variable aléatoire (à valeurs dans E) si et seulement si pour tout $A \in \mathcal{T}$, l'ensemble $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$ appartient à $\mathcal{A}$.
On notera $X \in A$ l'ensemble $X^{-1}(A)$ et la loi de X sera donnée par la probabilité des évênements $P(X \in A)$ pour tout $A \in \mathcal{T}$.
★**En général et même tout le temps :** Ω et $\mathcal{A}$ sont "abstraits" et seront supposés de façon à ce que X soit une variable aléatoire. Seul E ensemble des valeurs de X, $\mathcal{T}$ ensemble des évênements de E et P définissant la loi de X seront pour nous intéressants.
★ **Exemples :**

- Construction variable de Bernouilli de paramètre $p \in [0,1]$, notée $\mathcal{B}(1,p)$. $E = \{0,1\}$ qui est fini, donc l'ensemble des évênements est la tribu, ensemble des parties de E : $2^E = \{\emptyset, \{0\}, \{1\}, E\}$. Sur $(\Omega, \mathcal{A})$, on définit P de façon à ce que $P(X=0) = 1-p$ et $P(X=1) = p$.
- Construction variable suivant loi binômiale de paramètre $p \in [0,1]$ et $n \in \mathbb{N}$, notée $\mathcal{B}(n,p)$. $E = \{0, ..., n\}$ qui est fini, donc l'ensemble des évênements est la tribu, ensemble des parties de E : 2^E (il y a 2^n évênements). Sur $(\Omega, \mathcal{A})$, on définit P de façon à ce que $P(X=k) = C_n^k p^k (1-p)^{n-k}$ pour tout $0 \leq k \leq n$.
- Construction variable suivant loi de Poisson de paramètre $\lambda > 0$. $E = \mathbb{N}$ qui est dénombrable, donc l'ensemble des évênements est la tribu 2^E. Sur $(\Omega, \mathcal{A})$, on définit P de façon à ce que $P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$.
- Construction variable suivant loi normale de moyenne μ et variance σ^2, notée $\mathcal{N}(\mu, \sigma^2)$. $E = \mathbb{R}$, donc l'ensemble des évênements est la tribu borélienne, notée $\mathcal{B}_{\mathbb{R}}$. Sur $(\Omega, \mathcal{A})$, on définit P de façon à ce que :

$$P(a < X < b) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^b \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx.$$

2.1.4 ★ Propriétés de variables aléatoires

Indépendance : Des variables aléatoires $X_1, \ldots, X_n$ définies sur $(\Omega, \mathcal{A}, P)$ et à valeurs respectivements dans $(E_1, \mathcal{T}_1), \ldots, (E_n, \mathcal{T}_n)$ sont indépendantes si pour tout $A_1 \in \mathcal{T}_1, \ldots, A_n \in \mathcal{T}_n$, on a :

$$P((X_1 \in A_1) \cap \ldots \cap (X_n \in A_n)) = P(X_1 \in A_1) \times \ldots \times P(X_n \in A_n). \quad (11)$$

Ce qu'il faut comprendre est que la valeur d'un des X_j n'influence pas la loi des autres X_j.
Réalisation d'une variable aléatoire : C'est l'image $X(\omega)$. ω est "tiré au hazard", plus $P(X \in A)$ est grand et plus il y aura de chance que $X(\omega)$ soit dans A. Une réalisation=une expérience.
Echantillon d'une variable aléatoire : Plusieurs réalisations (indépendantes)

de la variable aléatoire X. Echantillon de taille n=n expériences. On peut considérer un échantillon comme un vecteur aléatoire $(X_1, \ldots, X_n)$, où $X_1, \ldots, X_n$ sont indépendantes et de même loi que X, $X_1, \ldots, X_n$ est appelé "échantillon indépendant identiquement distribué" (i.i.d).
Erreur à ne SURTOUT PAS COMMETTRE : $(X, \ldots, X)$ n'est surtout pas un échantillon de X car pour un ω, toutes les composantes de $(X(\omega), \ldots, X(\omega))$ ont même valeur $X(\omega)$, cela voudrait dire par exemple que toutes les personnes d'un groupe (choisi au hazard) ont exactement même taille.
A partir d'un échantillon, on peut estimer la moyenne et la variance par : la moyenne empirique et la variance empirique.

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j, \tag{12}$$

$$S_n = \frac{1}{n}\sum_{j=1}^{n} X_j^2 - \overline{X}_n^2. \tag{13}$$

★ : Ce sont des variables aléatoires, en effet $\overline{X}_n$ et S_n sont des fonctions de ω. Si on refait encore n nouvelles expériences, on obtiendra une autre valeur.
Lorsque n devient grand, $\overline{X}_n$ et S_n se rapprochent des vrais moyenne (ou espérance) et variance.
Cas discret, vrais moyenne et variance : Soit X variable aléatoire de loi donnée par $P(X = k) = p_k$.

$$E(X) = \sum_{k\in\mathbb{N}} k p_k. \tag{14}$$

$$V(X) = \sum_{k\in\mathbb{N}} k^2 p_k - \mu^2. \tag{15}$$

Cas continu, vrais moyenne et variance : Soit X variable aléatoire à valeurs réelles de densité f :

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx. \tag{16}$$

$$V(X) = \int_{-\infty}^{+\infty} x^2 f(x) dx - \mu^2. \tag{17}$$

★ : **La moyenne d'une somme de variables aléatoires est la somme des moyennes,** plus exactement :

$$E(a_1X_1 + \ldots + a_nX_n) = a_1E(X_1) + \ldots + a_nE(X_n).$$

★ : **La variance d'une somme de variables aléatoires INDEPENDANTES est la somme des variances,** plus exactement :

$$V(a_1X_1 + \ldots + a_nX_n) = a_1^2V(X_1) + \ldots + a_n^2V(X_n).$$

ERREUR A NE SURTOUT PAS COMMETTRE : $V(X-Y)$ n'est pas égal à $V(X)-V(Y)$, en effet $V(X-Y) = V(X+(-1)Y) = V(X)+(-1)^2V(Y) = V(X)+V(Y)$.

2.1.5 ★ : Propriétés de la moyenne empirique

★ : Si $(X_1, \ldots, X_n)$ est un échantillon iid de loi de moyenne μ et de variance σ^2, alors :

$$E(\overline{X}_n) = \mu \quad (18)$$

$$V(\overline{X}_n) = \frac{\sigma^2}{n}. \quad (19)$$

Plus n est grand, plus la variance est faible et donc pour plusieurs jeux de n expériences, les valeurs de la moyenne empirique différeront peu et seront proches de μ.

2.1.6 ★ : Propriétés des lois normales

★ : **Loi centrée-réduite** : Fonction de répartition notée ϕ tabulée pour les valeurs positives. Ainsi $\phi(2.65)$ se lit dans la table (2.6 en ligne et 0.05 en colonne). Si $x < 0$, on utilise $\phi(x) = 1 - \phi(-x)$. De plus $P(X > x) = 1 - \phi(x)$.
★ : **Loi normale de moyenne μ et variance σ^2** : Si X suit une loi normale de moyenne μ et variance σ^2, alors $Z = \frac{X-\mu}{\sigma}$ suit une loi normale centrée-réduite. On a alors $P(X < x) = \phi\left(\frac{x-\mu}{\sigma}\right)$ et bien sûr $P(X > x) = 1 - \phi(x)$.
Somme de normales : Si $X_1, \ldots, X_n$ sont **indépendantes** de loi normales respectives $\mathcal{N}(\mu_1, \sigma_1^2), \ldots, \mathcal{N}(\mu_n, \sigma_n^2)$, aloes la somme $a_1X_1 + \ldots + a_nX_n$ est une loi normale de moyenne et variance égales à :

$$\begin{aligned} \mu &= a_1\mu_1 + \ldots + a_n\mu_n \\ \sigma^2 &= a_1^2\sigma_1^2 + \ldots + a_n^2\sigma_n^2. \end{aligned} \quad (20)$$

★ : **Moyenne empirique d'un échantillon d'une loi normale :** Si $(X_1, \ldots, X_n)$ est un échantillon iid de $\mathcal{N}(\mu, \sigma^2)$, alors la moyenne empirique $\overline{X}_n$ suit une loi normale $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ **MEME LORSQUE n EST PETIT.**

2.1.7 Théorème limite-central et approximation par loi normale

★ : **Convergence en loi, cas discret :** Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires à valeurs dans $\mathbb{N}$ et X une variable aléatoire à valeurs dans $\mathbb{N}$. On dit que X_n converge **en loi** vers X si pour tout $k \in \mathbb{N}$, on a :

$$\lim_{n\to+\infty} P(X_n = k) = P(X = k). \quad (21)$$

★ : **Convergence en loi, cas continu :** Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}$ et X une variable aléatoire à valeurs dans $\mathbb{R}$. Soit

F_n fonction de répartition de X_n et F fonction de répartition de X. On dit que X_n converge **en loi** vers X si pour tout $x \in \mathbb{R}$, on a :

$$\lim_{n \to +\infty} F_n(x) = F(x). \tag{22}$$

★ : **Théorème limite-central (TLC), version somme** : Soit $X_1, \ldots, X_n$ un échantillon iid de loi ayant une moyenne μ et une variance σ^2 et soit Y une variable aléatoire suivant une loi normale $\mathcal{N}(n\mu, n\sigma^2)$. Alors la suite $(X_1 + \ldots + X_n - Y)_{n \in \mathbb{N}}$ converge en loi vers 0.
A retenir : Sous les hypothèses du TLC, à savoir indépendance, même loi, existence d'une variance, la somme peut être approximée par une normale de moyenne $n\mu$ et $n\sigma^2$ lorsque n est grand (supérieur à 30).
★ : **Théorème limite-central (TLC), version moyenne empirique** : Soit $X_1, \ldots, X_n$ un échantillon iid de loi ayant une moyenne μ et une variance σ^2 et soit Y une variable aléatoire suivant une loi normale $\mathcal{N}(\mu, \frac{\sigma^2}{n})$. Alors la suite $(\overline{X}_n - Y)_{n \in \mathbb{N}}$ converge en loi vers 0.
A retenir : Sous les hypothèses du TLC, à savoir indépendance, même loi, existence d'une variance, la moyenne empirique peut être approximée par une normale de moyenne μ et $\frac{\sigma^2}{n}$ lorsque n est GRAND (supérieur à 30).

2.2 Exercice 1

1. Variable aléatoire définie sur $(\Omega, \mathcal{A})$. Ensemble des valeurs est $E = \{0, 1\}$, 0 si mal rempli, 1 si bien rempli. L'ensemble des valeurs étant fini, alors l'ensemble des évênements est la tribu 2^E : ensemble des parties de E. On définit sur $(\Omega, \mathcal{A})$, la probabilité P telle que $P(Y = 0) = 1 - p = 0.06$ et $P(Y = 1) = 0.94$
2. La loi de Y est la loi de Bernouilli de paramètre 0.94, notée $\mathcal{B}(1, 0.94)$.
3. **ATTENTION :** Il est demandé de calculer et pas de donner la moyenne et la variance donc ne pas utiliser le formulaire et donner directement la réponse.
   On est dans le cas discret avec $p_0 = 0.06$, $p_1 = 0.94$ et $p_k = 0$ pour $k \geq 2$, d'où :

$$\begin{aligned} E(Y) &= 0.06 \times 0 + 0.94 \times 1 = 0.94. \\ E(Y^2) &= 0.06 \times 0^2 + 0.94 \times 1^2 = 0.94. \\ V(Y) &= E(Y^2) - E(Y)^2 = 0.94 - 0.94^2 = 0.94 \times 0.06 = 0.0564. \end{aligned}$$

4. $X = \sum_{j=1}^{n} Y_j$, où $Y_1, \ldots, Y_n$ sont indépendantes de même loi de Bernouilli $\mathcal{B}(1, 0.94)$. X est à valeurs dans $F = \{0, \ldots, n\}$ (ATTENTION DE NE PAS L'APPELER E DEJA UTILISE), F est fini, donc l'ensemble des évênements est l'ensemble des parties de F. X est une variable aléatoire définie sur $(\Omega, \mathcal{A})$ comme somme de variables aléatoires.

5. La somme de variables de Bernouilli indépendantes, de même paramètre, suit une loi binômiale, notée $\mathcal{B}(n, 0.94)$. On a :

$$P(X = k) = C_n^k 0.94^k \times 0.06^{n-k},$$

pour $0 \leq k \leq n$.
L'espérance d'une somme est la somme des espérance (ATTENTION : même si cela n'est pas indépendant : ajouter une hypothèse inutile = pas la totalité des points.), on a donc :

$$E(X) = 0.94 \times n.$$

La variance d'une somme de variables indépendantes est la somme des variances (ATTENTION : oubli de l'indépendance : conséquence plus grave que ajout d'hypothèse inutile), ainsi :

$$V(X) = 0.0564 \times n.$$

6. Pour $n = 5$, $\mathcal{P}(X = 0) = 0.06^5 \simeq 7.76 \times 10^{-7}$ (aucun dossier).
$P(X = 5) = 0.94^5 \simeq 0.74$ (tous les dossiers).
$P(X > 3) = P(X = 4) + P(X = 5) = C_5^4 0.06 \times 0.94^4 + 0.74 \simeq 0.97$.
$P(2 < X < 4) = \mathcal{P}(X = 3) = C_5^3 0.94^3 \times 0.06^2 \simeq 0.029$.
7. (a) Si $n = 100$, on remarque que X est somme de variables indépendantes de même loi ayant une variance. On a $n \geq 30$. Utilisant le théorème limite-central, X peut être approximé par une normale de moyenne et de variance données par :

$$\begin{aligned} E(X) &= 0.94 \times n = 94. \\ V(X) &= 0.0564 \times n = 5.64. \end{aligned}$$

ATTENTION : Les moyenne et variance n'ont aucune raison de différer de celles trouvées à la question 5).

(b) La variable aléatoire $Z = \frac{X-94}{\sqrt{5.64}} = \frac{X-94}{2.38}$ suit approximativement une loi normale centrée-réduite.
$\mathcal{P}(X < 95) \simeq \mathcal{P}(Z < 0.42) = \phi(0.42) = 0.6628$.
$\mathcal{P}(X > 90) = 1 - \mathcal{P}(X < 90) \simeq 1 - \mathcal{P}(Z < -\frac{4}{2.38}) = 1 - \mathcal{P}(Z < -1.68) = 1 - (1 - \phi(1.68)) = \phi(1.68) = 0.9535$.
$\mathcal{P}(90 < X < 95) = \mathcal{P}(-1.68 < Z < 0.42) = \phi(0.42) - (1 - \phi(1.68)) = 0.6163$.
$\mathcal{P}(85 < X < 90) = \mathcal{P}(-3.79 < Z < -1.68) = \phi(3.79) - \phi(1.68)$.

2.3 Exercice 2

1. Population=main d'oeuvre d'une entreprise.
Variable étudiée=résultats au test.

2. (a)

	Population	Echantillon
taille	-	25
moyenne	150	154.7
écart-type	10	12.3
	Vrais paramètres	Estimation des paramètres (empirique et aléatoire)

(b) **Rappel important :** Lorsque $(X_1, \ldots, X_n)$ est un échantillon de variables aléatoires indépendantes de loi normale de moyenne μ (vraie moyenne) et écart-type σ (vrai écart-type), noté $\mathcal{N}(\mu, \sigma)$ (ou $\mathcal{N}(\mu, \sigma^2)$ si on paramétrise par la variance), alors la moyenne empirique $\overline{X_n} = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} X_j$ est une **variable aléatoire** qui suit une loi normale de moyenne μ et d'écart-type $\frac{\sigma}{\sqrt{n}}$ (ou variance $\frac{\sigma^2}{n}$).
$\overline{X_{25}}$ suit une loi normale de moyenne 150 et d'écart-type $\frac{10}{\sqrt{25}} = 2$.
(**Attention :** 154.7 n'est surtout pas la moyenne de $\overline{X_{25}}$ mais sa réalisation ! ! !).

(c) On cherche $P(\overline{X_{25}} > 154.7)$, posons $Z = \dfrac{\overline{X_{25}} - 150}{2}$ qui suit une loi normale centrée-réduite. On a alors :

$$P(\overline{X_{25}} > 154.7) = P(Z > \frac{4.7}{2}) = P(Z > 2.35) = 1 - \underbrace{\phi(2.35)}_{\text{se lit dans la table}}.$$
$$= 1 - 0.9906 = 0.0094.$$

Attention : A l'examen, je ne veux surtout pas voir de $P(2.35)$ (qui n'a aucun sens) à la place de $P(Z > 2.35)$.

3. Même question que pour 2. avec une taille plus grande.
Attention : Utiliser votre esprit critique, si la taille augmente, on se rapproche de la vraie moyenne et donc la probabilité d'observer une valeur plus grande que 154.7 devrait être encore plus petite.
Ici $\overline{X_{36}}$ suit loi normale de moyenne 150 et d'écart-type $\frac{10}{\sqrt{36}} = \frac{5}{3}$.
Posons $Z = \dfrac{3\left(\overline{X_{36}} - 150\right)}{5}$, on trouve :

$$P(\overline{X_{36}} > 154.7) = P(Z > \frac{4.7 \times 3}{5}) = P(Z > 2.82) = 1 - \underbrace{\phi(2.82)}_{\text{se lit dans la table}}.$$
$$= 1 - 0.9976 = 0.0024.$$

2.4 Exercice 3

1. **Attention :** La moyenne empirique (comme tout estimateur) étant une variable aléatoire, elle possède une moyenne (vraie moyenne) et un écart-type (vrai écart-type).

Rappel important : Si $(X_1, \ldots, X_n)$ est un échantillon de variables aléatoires indépendantes de même moyenne μ (vraie moyenne) et d'écart-type σ (vrai écart-type), alors la vraie moyenne de la moyenne empirique est μ et le vrai écart-type de la moyenne empirique est $\dfrac{\sigma}{\sqrt{n}}$.

taille de l'échantillon	moyenne empirique		
	loi	moyenne	écart-type
$n = 4$	?	60	5
$n = 8$	?	60	$\frac{5\sqrt{2}}{2}$
$n = 32$	normale	60	$\frac{5\sqrt{2}}{4}$
$n = 100$	normale	60	1

2. Lorsque n est grand, supérieur à 30, on peut considérer que la moyenne empirique suit une loi normale. Ici, $n = 100$, la moyenne empirique $\overline{X_{100}}$ suit une loi normale de moyenne 60 et écart-type 1. Posons $Z = \overline{X_{100}} - 60$, alors :

$$\begin{aligned} P(\overline{X_{100}} < 56) &= 1 - \underbrace{\phi(4)}_{\text{se lit dans la table}} . \\ &= 1 - 0.99997 = 0.00003. \end{aligned}$$

3 Chapitre 2, Estimation ponctuelle et méthode du maximum de vraisemblance

3.1 Rappels et approfondissement

3.1.1 ★ Statistiques, convergence et estimateurs

★ : **Convergence en proba :** Une suite de variables aléatoires $(X_n)_{n\geq 0}$ converge en probabilité vers X si pour tout $\epsilon > 0$, $\lim\limits_{n\to+\infty} \mathcal{P}(|X_n - X| > \epsilon) = 0$.

★ : **Loi des grands nombres :** Si $(X_n)_{n\geq 1}$ est une suite de variables aléatoires de même loi, ayant une moyenne μ, indépendantes, alors la moyenne empirique :

$$\frac{1}{n}\sum_{k=1}^{n} X_k,$$

converge en probabilité vers μ.

★ : **On a vu :** La moyenne (resp. variance) sur échantillon appelées aussi moyenne et variance empirique sont différentes de la moyenne (resp. variance) sur population qui sont les vrais paramètres. La moyenne et variance sur échantillon convergent en probabilité vers les paramètres théoriques : on dit que ce sont des estimateurs de la moyenne et de la variance.

Définition 1. *Soit $(X_1, \ldots, X_n)$ échantillon iid de loi donnée par une densité $y \to p(y;\theta)$ dépendant d'un paramètre réel θ.*

Une statistique est une variable aléatoire de la forme $T_n = \varphi_n(X_1, \ldots, X_n)$, c'est un estimateur (convergent) de θ s'il converge en probabilité vers θ lorsque n tend vers l'infini.
Un estimateur est sans biais si $E(T_n) = \theta$.

Proposition 1 (Inégalité de Bienaymé-Tchebytchev)**.** *Soit $(X_1, \ldots, X_n)$ échantillon iid de loi dépendant d'un paramètre θ et soit $T_n = \varphi_n(X_1, \ldots, X_n)$ une statistique. On a :*

$$\mathcal{P}(|T_n - \theta| > \epsilon) \leq \frac{E(|T_n - \theta|^2)}{\epsilon^2}, \tag{23}$$

et si le risque quadratique $\lim_{n \to +\infty} E(|T_n - \theta|^2) = 0$, alors T_n converge en proba vers θ et donc T_n est un estimateur de θ.

★ : **Cas d'une statistique sans biais :** Si $E(T_n) = \theta$ (sans biais), le risque quadratique est égal à la variance $V(T_n)$. Pour savoir si T_n converge, on regarde la limite de $V(T_n)$.
★ : **Qualité d'un estimateur :** Plus le risque quadratique est faible, meilleur est l'estimateur.
Lorsque l'estimateur est sans biais, plus la variance est faible, meilleur est l'estimateur.

3.1.2 ★ Vraisemblance

D'une variable aléatoire discrète : Soit Y une variable aléatoire discrète dont $P(Y = k) = p_k = p_k(\theta)$ est une fonction d'un paramètre θ (inconnu). La vraisemblance de Y est une fonction qui à θ associe la variable aléatoire $p_Y(\theta)$. On note $L_\theta(Y)$ la valeur de la vraisemblance en θ. (L comme "likelihood").
Exemple : Si Y suit une loi de Poisson de paramètre $\lambda > 0$, alors $p_k(\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$. D'où la vraisemblance est la fonction qui à λ associe $L_\theta(Y) = \frac{\lambda^Y}{Y!} e^{-\lambda}$.
D'une variable aléatoire continue : Soit Y une variable aléatoire contnue de densité $y \to p(y; \theta)$ dépendant d'un paramètre θ. La vraisemblance de Y est une fonction qui à θ associe la variable aléatoire $p(Y, \theta)$. On note $L_\theta(Y)$ la valeur de la vraisemblance en θ. (L comme "likelihood").
D'un échantillon : Soit $(Y_1, \ldots, Y_n)$ un échantillon iid de loi dépendant d'un paramètre θ, la vraisemblance est la fonction qui à θ associe $L_\theta(Y_1) \times \ldots \times L_\theta(Y_n)$.

3.1.3 ★ Log-vraisemblance

D'une variable aléatoire : C'est la fonction qui à θ associe $\log L_\theta(Y)$.
D'un échantillon : C'est la fonction qui à θ associe $\log L_\theta(Y_1) + \ldots + \log L_\theta(Y_n)$.

3.1.4 ★ Score et propriétés du score

Existe si la log-vraisemblance est **dérivable** et que l'ensemble des paramètres est un **ouvert** de $\mathbb{R}$. **D'une variable aléatoire :** C'est la fonction qui à θ associe $\frac{d}{d\theta} \log L_\theta(Y)$.

D'un échantillon : C'est la fonction qui à θ associe $\frac{d}{d\theta}\log L_\theta(Y_1) + \ldots + \frac{d}{d\theta}\log L_\theta(Y_n)$.
★ : **Support de la loi :** Valeurs prises par la variable aléatoire. Si le support ne dépend pas du paramètre θ, alors :

$$E\left(\frac{d}{d\theta}\log L_\theta(Y)\right) = 0. \tag{24}$$

On en déduit alors que la variance du score est égale à :

$$E\left[\left(\frac{d}{d\theta}\log L_\theta(Y)\right)^2\right], \tag{25}$$

quantité appelée **information de Fisher de la variable aléatoire** Y et notée $I_Y(\theta)$.
Si de plus, la log-vraisemblance est **deux-fois dérivable**, alors l'information de Fisher est égale à :

$$I_Y(\theta) = -E\left[\frac{d^2}{d\theta^2}\log L_\theta(Y)\right]. \tag{26}$$

3.1.5 ★ : Information de Fisher d'un échantillon et inégalité de Cramer-Rao

L'information de Fisher d'un échantillon $(Y_1, \ldots, Y_n)$, notée $I_{Y_1,\ldots,Y_n}(\theta)$, est définie comme :

$$I_{Y_1,\ldots,Y_n}(\theta) = E\left[\left(\frac{d}{d\theta}\log L_\theta(Y_1) + \ldots + \frac{d}{d\theta}\log L_\theta(Y_n)\right)^2\right]. \tag{27}$$

Si le support ne dépend pas de θ, alors $I_{Y_1,\ldots,Y_n}(\theta)$ est égale à la variance du score de l'échantillon et donc :

$$I_{Y_1,\ldots,Y_n}(\theta) = I_{Y_1}(\theta) + \ldots + I_{Y_n}(\theta) = nI_Y(\theta). \tag{28}$$

Sous la condition du support et si l'information de Fisher est non nulle, on a également l'inégalité de Cramer-Rao : soit T_n un estimateur sans-biais de θ, alors :

$$V(T_n) \geq \frac{1}{I_{Y_1,\ldots,Y_n}(\theta)}. \tag{29}$$

Interprétation : Si l'information de Fisher est très petite, alors $V(T_n)$ est très grande et donc aucun estimateur sans biais estime correctement. Par contre, si l'information de Fisher est très grande, on ne peut pas conclure car on a pu utiliser un estimateur de mauvaise qualité. Un estimateur **sans biais** de "bonne qualité" est un estimateur **efficace**, c'est à dire tel que :

$$V(T_n) = \frac{1}{I_{Y_1,\ldots,Y_n}(\theta)}. \tag{30}$$

$\star$: **Pour ceux qui souhaitent aller plus loin.** T_n est également une variable aléatoire, on peut donc définir son information de Fisher, notée $I_{T_n}(\theta)$. Lorsque le support de la loi des $Y_1, \ldots, Y_n$ ne dépend pas de θ, on a alors :

$$I_{T_n}(\theta) \leq I_{Y_1,\ldots,Y_n}(\theta),$$

ce qui signifie que l'échantillon $(Y_1, \ldots, Y_n)$ porte toujours plus d'information sur θ que tout estimateur fonction des $(Y_1, \ldots, Y_n)$. Un estimateur conservant l'information et donc tel que :

$$I_{T_n}(\theta) = I_{Y_1,\ldots,Y_n}(\theta),$$

s'appelle un **estimateur exhaustif**.
$\star\star$: **Pour ceux qui souhaitent aller encore plus loin.** Nous avons vu que si on utilise des estimateurs de la forme $T_n = \varphi_n(Y_1, \ldots, Y_n)$, l'information disponible sur le paramètre est limitée par l'information de Fisher $I_{Y_1,\ldots,Y_n}$. Si celle-ci est faible, même si on utilise des estimateurs exhaustifs, on risque de récupérer peu d'information sur le paramètre, ce qui se traduit par un fort désordre (forte entropie) des estimations. Afin d'avoir une valeur plus fiable d'estimation, on peut incorporer de l'information a priori. Cela signifie que l'on connaît la valeur du paramètre avec une certaine imprécision au lieu de ne pas le connaître du tout. Cette connaissance a priori est modélisée en considérant que le paramètre lui-même suit une loi de probabilité : on entre dans le monde des **statistiques bayésiennes**.

3.1.6 ★ Maximum de vraisemblance

Méthode pour obtenir un estimateur : Soit $(Y_1, \ldots, Y_n)$ un échantillon iid de loi dépendant d'un paramètre θ, soit θ_0 la vraie valeur du paramètre. Pour $(y_1, \ldots, y_n) = (Y_1(\omega), \ldots, Y_n(\omega))$ réalisation de l'échantillon, le maximum de vraisemblance est le paramètre θ maximisant :

$$L_\theta(y_1, \ldots, y_n) = p(y_1; \theta) \times \ldots \times p(y_n; \theta). \tag{31}$$

La fonction logarithme étant une fonction croissante, cela revient à maximiser :

$$\log L_\theta(y_1, \ldots, y_n) = \log p(y_1; \theta) + \ldots + \log p(y_n; \theta). \tag{32}$$

★ : **Pour être sûr de ne pas se tromper :** commencer par calculer $\log p(y_j; \theta)$ sur la densité et faire la somme ensuite.
Le maximum de vraisemblance peut être obtenu (pas toujours) en résolvant l'équation de vraisemblance :

$$\frac{d}{d\theta} \log L_\theta(y_1, \ldots, y_n) = 0. \tag{33}$$

On note $\hat{\theta}_{\mathrm{MV}}(y_1, \ldots, y_n)$ la solution obtenue pour la réalisation $(y_1, \ldots, y_n)$. Finalement, l'estimateur du maximum de vraisemblance est la variable aléatoire :

$$\omega \in \Omega \to \hat{\theta}_{\mathrm{MV}}(Y_1(\omega), \ldots, Y_n(\omega)), \tag{34}$$

que l'on note $\hat{\theta}_{\mathrm{MV}}(Y_1, \ldots, Y_n)$.

3.2 Exercice 1

1. On a :
$$\overline{X}_{n_1} = \frac{1}{n_1}\sum_{k=1}^{n_1} X_k, \tag{35}$$
où $(X_1, \ldots, X_{n_1})$ variables indépendantes de loi de Bernouilli $P(X_k = 1) = p_1$. On a donc :
$$\begin{aligned} E(\overline{X}_{n_1}) &= \frac{1}{n_1}\sum_{k=1}^{n_1} E(X_k) \\ &= \frac{1}{n_1}\sum_{k=1}^{n_1} p_1 \\ &= p_1. \end{aligned} \tag{36}$$
il est sans biais.
$\overline{X}_{n_2}$ et $\dfrac{\overline{X}_{n_1} + \overline{X}_{n_2}}{2}$ sont également sans biais lorsque $p_1 = p_2 = p$.
2. **Meilleur des trois :** celui ayant le plus petit risque quadratique. Comme les estimateurs sont sans biais, le risque quadratique est égal à la variance. Les variables X_k étant indépendantes on en déduit :
$$\begin{aligned} V(\overline{X}_{n_1}) &= \frac{1}{n_1}p(1-p). \\ V(\overline{X}_{n_2}) &= \frac{1}{n_2}p(1-p). \end{aligned} \tag{37}$$
Les variables $\overline{X}_{n_1}$ et $\overline{X}_{n_2}$ étant indépendantes, donc :
$$V\left(\frac{\overline{X}_{n_1} + \overline{X}_{n_2}}{2}\right) = \frac{n_1 + n_2}{4n_1n_2}p(1-p).$$
Si $n_1 = n_2 = n$, alors $\overline{X}_{n_1}$ et $\overline{X}_{n_2}$ ont même variance et :
$$V\left(\frac{\overline{X}_{n_1} + \overline{X}_{n_2}}{2}\right) = \frac{1}{2n}p(1-p).$$
donc le meilleur estimateur est $\dfrac{\overline{X}_{n_1} + \overline{X}_{n_2}}{2}$.
Si $n_1 < n_2$, alors $V(\overline{X}_{n_1}) > V(\overline{X}_{n_2})$. Il reste alors à comparer $\dfrac{\overline{X}_{n_1} + \overline{X}_{n_2}}{2}$ avec $\overline{X}_{n_2}$. On a :$V\left(\dfrac{\overline{X}_{n_1} + \overline{X}_{n_2}}{2}\right) \leq V(\overline{X}_{n_2})$ si et seulement si :
$$\frac{n_1 + n_2}{4n_1n_2} \leq \frac{1}{n_2}.$$

ce qui est équivalent à :

$$n_1 + n_2 \leq 4n_1,$$

donc, sous la condition $n_1 < n_2$, $\dfrac{\overline{X}_{n_1} + \overline{X}_{n_2}}{2}$ est le meilleur si et seulement si $3n_1 \geq n_2$. (Il est cependant toujours meilleur que $V(\overline{X}_{n_1})$ car $3n_2 > n_1$.)
Au final :

- Si $n_1 = n_2$, $\overline{X}_{n_1}$ et $\overline{X}_{n_2}$ ont même qualité et $\dfrac{\overline{X}_{n_1} + \overline{X}_{n_2}}{2}$ est le meilleur.
- Si $n_1 < n_2$ et si $3n_1 = n_2$, $\overline{X}_{n_2}$ et $\dfrac{\overline{X}_{n_1} + \overline{X}_{n_2}}{2}$ ont même qualité et sont meilleurs que $\overline{X}_{n_1}$.
- Si $n_1 < n_2$ et si $3n_1 > n_2$, alors $\dfrac{\overline{X}_{n_1} + \overline{X}_{n_2}}{2}$ est meilleur que $\overline{X}_{n_2}$ qui est lui même meilleur que $\overline{X}_{n_1}$.
- Si $n_1 < n_2$ et si $3n_1 < n_2$, alors $\overline{X}_{n_2}$ est meilleur que $\dfrac{\overline{X}_{n_1} + \overline{X}_{n_2}}{2}$, lui-même meilleur que $\overline{X}_{n_1}$.

En inversant les rôles de n_1 et n_2 (on a le droit, ils sont quelquonques), on trouve aussi :

- Si $n_2 < n_1$ et si $3n_2 = n_1$, $\overline{X}_{n_1}$ et $\dfrac{\overline{X}_{n_1} + \overline{X}_{n_2}}{2}$ ont même qualité et sont meilleurs que $\overline{X}_{n_2}$.
- Si $n_2 < n_1$ et si $3n_2 > n_1$, alors $\dfrac{\overline{X}_{n_1} + \overline{X}_{n_2}}{2}$ est meilleur que $\overline{X}_{n_1}$ qui est lui même meilleur que $\overline{X}_{n_2}$.
- Si $n_2 < n_1$ et si $3n_2 < n_1$, alors $\overline{X}_{n_1}$ est meilleur que $\dfrac{\overline{X}_{n_1} + \overline{X}_{n_2}}{2}$, lui-même meilleur que $\overline{X}_{n_2}$.

3.3 Exercice 2

1. $Z = \sum_{j=1}^{n} X_j$, où X_j vaut 1 si le foyer j possède une machine à laver, 0 sinon. X_j suit une loi de Bernouilli de paramètre p, ainsi Z suit une loi binômiale $\mathcal{B}(n, p)$. La moyenne des X_j étant égale à p, on peut estimer p en utilisant la moyenne empirique des X_j. Ainsi :

$$\bar{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j,$$

est un estimateur de p. On a $E(\bar{X}_n) = p$, il est donc sans biais et $V(\bar{X}_n) = \frac{p(1-p)}{n}$ qui tend vers 0 en l'infini, ce qui confirme la convergence de l'estimateur.

2. La suite $(\bar{X}_n)_{n\in\mathbb{N}}$ converge en probabilité vers p, ainsi la suite $(T_n)_{n\in\mathbb{N}}$, où

$T_n = \bar{X}_n(1-\bar{X}_n)$ converge en probabilité vers $p(1-p)$. On a :

$$\begin{aligned} E(T_n) &= E(\overline{X}_n) - E(\overline{X}_n^2) \\ &= \frac{E(Z)}{n} - \frac{E(Z^2)}{n^2} \\ &= \frac{E(Z)}{n} - \frac{V(Z)+E(Z)^2}{n^2} \\ &= \frac{np}{n} - \frac{np(1-p)+(np)^2}{n^2} \\ &= p(1-p) - \frac{p(1-p)}{n} \\ &= \frac{n-1}{n}p(1-p). \end{aligned}$$

donc l'estimateur est biaisé.

3. $R_n = \dfrac{n}{n-1}T_n$ est un estimateur sans biais de $p(1-p)$.
4. Condition pour que U soit un estimateur : les suites $\bar{X}_n$ et $\bar{Y}_m$ convergent en probabilité vers p, donc U converge en probabilité vers $ap+bp = (a+b)p$. U est donc un estimateur de p si et seulement si $a+b=1$. Sous la condition $a+b=1$, $E(U) = aE(\bar{X}_n) + bE(\bar{Y}_m) = ap+bp = (a+b)p = p$, U est donc sans biais.
   U étant sans biais, sa qualité est évaluée en observant sa variance. $\bar{X}_n$ et $\bar{Y}_m$ étant indépendantes, on a :

$$\begin{aligned} V(U) &= a^2V(\bar{X}_n) + b^2V(\bar{Y}_m). \\ &= (\frac{a^2}{n} + \frac{b^2}{m})p(1-p). \\ &= \frac{ma^2+nb^2}{mn}p(1-p). \end{aligned}$$

   On doit donc minimiser $(a,b) \to ma^2+nb^2$ sous la contrainte $a+b=1$. En remplaçant b par $1-a$, ceci est équivalent à minimiser $f : a \to ma^2 + n(1-a)^2 = (m+n)a^2 - 2na + n$. Ceci est un polynôme de degré 2 dont le coefficient de plus grand degré $(m+n)$ est positif, ainsi la minimisation est équivalente à résoudre $f'(a) = 0$, soit $2(m+n)a - 2n$ qui donne $a = \frac{n}{m+n}$ et donc $b = \frac{m}{m+n}$.
   Ainsi $U = \dfrac{1}{m+n}(n\bar{X}_n + m\bar{Y}_m)$ est un estimateur sans biais meilleur que $\bar{X}_n$ et $\bar{Y}_m$.

3.4 Exercice 3

1. ★ **ATTENTION** : On vous demande de calculer, donc ne pas répondre directement en faisant un copié-collé du formulaire.

REFLEXE A AVOIR : vous avez une loi continue, ce sont des intégrales. Dans :

$$E(X_i) = \int_0^{+\infty} \frac{1}{\theta} x e^{-\frac{x}{\theta}} dx,$$

on effectue une intégration par partie (IPP).
★ RAPPEL SUR L'IPP :

$$\int_a^b u(x)v'(x)dx = u(b)v(b) - u(a)v(a) - \int_a^b u(x)v'(x)dx.$$

ATTENTION DE NE PAS OUBLIER dx dans l'intégrale car $\int_a^b f(x)$ signifie $f(a) + f(b)$ ce qui n'a absolument rien à voir.
Ainsi, dans :

$$E(X_i) = \int_0^{+\infty} \frac{1}{\theta} x e^{-\frac{x}{\theta}} dx,$$

on pose $u(x) = x$, $v'(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$, d'où $u'(x) = 1$ et $v(x) = -e^{-\frac{x}{\theta}}$, d'où :

$$\begin{aligned} \int_0^{+\infty} \frac{1}{\theta} x e^{-\frac{x}{\theta}} dx &= \lim_{n \to +\infty} -n e^{-\frac{n}{\theta}} + \int_0^n e^{-\frac{x}{\theta}} dx \\ &= \int_0^{+\infty} e^{-\frac{x}{\theta}} dx. \end{aligned}$$

Comme la fonction f est une densité de probabilité, on en déduit que :

$$\int_0^{+\infty} e^{-\frac{x}{\theta}} dx = \theta \int_0^{+\infty} f(x)dx = \theta.$$

Ainsi $E(X_i) = \theta$.
Pour la variance, on calcule :

$$E(X_i^2) = \int_0^{+\infty} \frac{1}{\theta} x^2 e^{-\frac{x}{\theta}} dx.$$

Pour l'intégration par parties, on pose $u(x) = x^2$ et $v'(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$, d'où $u'(x) = 2x$ et $v(x) = -e^{-\frac{x}{\theta}}$, d'où :

$$\begin{aligned} \int_0^{+\infty} \frac{1}{\theta} x^2 e^{-\frac{x}{\theta}} dx &= \lim_{n \to +\infty} -n^2 e^{-\frac{n}{\theta}} + 2 \int_0^n x e^{-\frac{x}{\theta}} dx \\ &= 2 \int_0^{+\infty} x e^{-\frac{x}{\theta}} dx \\ &= 2\theta E(X_i) = 2\theta^2. \end{aligned}$$

On en déduit que $V(X_i) = E(X_i^2) - E(X_i)^2 = \theta^2$.
Remarque : De part les expressions de l'espérance et de la variance, on en déduit deux façons d'estimer θ. Soit on utilise moyenne empirique qui donne un estimateur de θ ou bien la racine carrée de la variance empirique qui nous donne un autre estimateur de θ.

2. Fonction de vraisemblance de l'échantillon : produit des fonctions de vraisemblance prises en chaque X_i, d'où :

$$L_\theta(X_1, \ldots, X_n) = \frac{1}{\theta^n} e^{-\frac{1}{\theta}\sum_{i=1}^n X_i}.$$

3. La log-vraisemblance (logarithme de la vraisemblance) est :

$$\log L_\theta(X_1, \ldots, X_n) = -n\log(\theta) - \frac{1}{\theta}\sum_{i=1}^n X_i.$$

L'espace des paramètres $\Theta =]0, +\infty[$ est un ouvert. La log-vraisemblance est bien dérivable sur l'espace des paramètres $\Theta =]0, +\infty[$ et sa dérivée, appelée score, est :

$$\frac{d}{d\theta}\log L_\theta(X_1, \ldots, X_n) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^n X_i.$$

L'estimateur du maximum de vraisemblance $\hat{\theta}_{MV}$ doit satisfaire :

$$-\frac{n}{\hat{\theta}_{MV}} + \frac{1}{\hat{\theta}^2_{MV}}\sum_{i=1}^n X_i = 0,$$

soit :

$$-n\hat{\theta}_{MV} + \sum_{i=1}^n X_i = 0,$$

d'où :

$$\hat{\theta}_{MV} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n.$$

C'est la moyenne empirique. La moyenne de la moyenne empirique est toujours égale à la moyenne de la variable aléatoire étudiée (cad X_i), soit $E(\bar{X}_n) = E(X_i) = \theta$, l'estimateur est donc sans biais.
La variance de la moyenne empirique vaut $V(\bar{X}_n) = \frac{V(X_i)}{n} = \frac{\theta^2}{n}$ qui tend vers 0 lorsque n tend vers l'infini, l'estimateur est donc convergent.
Pour connaître l'efficacité, on doit calculer l'information de Fisher.
★ : Le support de la loi des X_i ; c'est à dire l'ensemble des valeurs prises par X_i ne dépend pas de θ. De plus, la log-vraisemblance est deux fois dérivable sur Θ, ainsi l'information de Fisher est donnée par :

$$I_{X_1,\ldots,X_n} = -E\left[\frac{d^2}{d\theta^2}\log L_\theta(X_1, \ldots, X_n)\right].$$

La dérivée seconde de la log-vraisemblance est égale à :

$$\frac{d^2}{d\theta^2}\log L_\theta(X_1, \ldots, X_n) = \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^n X_i,$$

et donc l'information de Fisher est égale à $I_{X_1,\ldots,X_n}(\theta) = \frac{n}{\theta^2}$. La variance de $\bar{X}_n$ étant l'inverse de l'information de Fisher, il est donc efficace.

3.5 Exercice 5

1. La loi de Poisson est une loi discrète définie par $p_k = P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$ pour tout $k \in \mathbb{N}$. La vraisemblance de la variable aléatoire X est donc :

$$L_\lambda(X) = \frac{\lambda^X}{X!}e^{-\lambda},$$

et donc la vraisemblance de l'échantillon est :

$$L_\lambda(X_1, \ldots, X_n) = \frac{\lambda^{X_1+\ldots+X_n}}{X_1! \times \ldots \times X_n!}e^{-n\lambda}.$$

La log-vraisemblance est :

$$\log L_\lambda(X_1, \ldots, X_n) = \left(\sum_{j=1}^{n} X_j\right) \times \log\lambda - \sum_{j=1}^{n} \log(X_j!) - n\lambda.$$

L'espace des paramètres est $]0, +\infty[$ qui est ouvert et la log-vraisemblance est dérivable sur cet espace, sa dérivée vaut :

$$\frac{d}{d\lambda} \log L_\lambda(X_1, \ldots, X_n) = \frac{1}{\lambda}\sum_{j=1}^{n} X_j - n.$$

Ainsi l'estimateur du maximum de vraisemblance $\hat{\lambda}_{MV}$ doit vérifier :

$$\frac{1}{\hat{\lambda}_{MV}}\sum_{j=1}^{n} X_j - n = 0.$$

On en déduit que :

$$\hat{\lambda}_{MV} = \frac{1}{n}\sum_{j=1}^{n} X_j = \bar{X}_n,$$

qui est la moyenne empirique.

2. La moyenne des X_j est la moyenne d'une loi de Poisson de paramètre λ, elle vaut donc λ, ainsi la moyenne de $\bar{X}_n$ est λ, l'estimateur est donc sans biais.
La variance des X_j est la variance d'une loi de Poisson de paramètre λ, elle vaut donc λ, ainsi $V(\bar{X}_n) = \frac{\lambda}{n}$, l'estimateur est donc convergent.

3. Le support de la loi de Poisson est $\mathbb{N}$ qui ne dépend pas de λ, de plus la log-vraisemblance est deux-fois dérivable. L'information de Fisher est alors égale à l'opposé de la moyenne de la dérivée seconde de la log-vraisemblance. La dérivée seconde de la log-vraisemblance étant égale à :

$$\frac{d^2}{d\lambda^2} \log L_\lambda(X_1, \ldots, X_n) = -\frac{1}{\lambda^2}\sum_{j=1}^{n} X_j,$$

on en déduit que l'information de Fisher est égale à $I_{X_1,\ldots,X_n}(\lambda) = \frac{n}{\lambda}$. L'estimateur $\bar{X}_n$ est alors efficace (sa variance étant égale à l'inverse de l'information de Fisher).

3.6 Exercice 6

1. Pour tout $x \in [0,1]$, $1-x \geq 0$ donc f est positive. De plus, la dérivée de la fonction $x \to (1-x)^{\frac{1}{\theta}}$ est la fonction $x \to -\frac{1}{\theta}(1-x)^{\frac{1}{\theta}-1}$, ainsi **une** primitive de $\frac{1}{\theta}(1-x)^{\frac{1}{\theta}-1}$ est donnée par la fonction $x \to -(1-x)^{\theta}$. On en déduit :

$$\begin{aligned}\int_{-\infty}^{+\infty} f(x)dx &= \int_0^1 \frac{1}{\theta}(1-x)^{\frac{1}{\theta}-1}dx \\ &= -\underbrace{(1-1)^{\frac{1}{\theta}}}_{0} - [-\underbrace{(1-0)^{\frac{1}{\theta}}}_{1}] \\ &= 1.\end{aligned}$$

   f est bien une densité de probabilité.

2. Lavraisemblance de l'échantillon est donnée par :

$$L_\theta(X_1,\ldots,X_n) = \frac{1}{\theta^n} \times (1-X_1)^{\frac{1}{\theta}-1} \times \ldots \times (1-X_n)^{\frac{1}{\theta}-1}.$$

3. La log-vraisemblance est donnée par :

$$\log L_\theta(X_1,\ldots,X_n) = -n\log(\theta) + \left(\frac{1}{\theta}-1\right)\sum_{j=1}^{n}\log(1-X_j).$$

   L'espace des paramètres est $]0,+\infty[$ qui est un ouvert et la log-vraisemblance est dérivable de dérivée :

$$\frac{d}{d\theta}\log L_\theta(X_1,\ldots,X_n) = -\frac{n}{\theta} - \frac{1}{\theta^2}\sum_{j=1}^{n}\log(1-X_j).$$

   L'estimateur du maximum de vraisemblance doit satisfaire :

$$-\frac{n}{\hat{\theta}_{MV}} - \frac{1}{\hat{\theta}_{MV}^2}\sum_{j=1}^{n}\log(1-X_j) = 0,$$

   donc est donné par :

$$\hat{\theta}_{MV} = -\frac{1}{n}\sum_{j=1}^{n}\log(1-X_j),$$

   qui est la moyenne empirique des $(-\log(1-X_1),\ldots,-\log(1-X_n))$.

4. ★ : **ATTENTION :** Avant de calculer une fonction de répartition, regarder d'abord les valeurs prises par la variable aléatoire, comme ça pas d'erreur.
   X prend ses valeurs dans $[0,1]$ donc $1-X$ prend également ses valeurs dans $[0,1]$ ainsi $\log(1-X)$ prend ses valeurs dans $]-\infty,0]$, ainsi Z prend

ses valeurs dans $[0, +\infty[$. On en déduit que pour $z \leq 0$, $P(Z \leq z) = 0$. Soit alors $z \geq 0$, on a :

$$\begin{aligned} P(Z \leq z) &= P(-\log(1-X) \leq z) \\ &= P(\log(1-X) \geq -z) \\ &= P(1-X \geq e^{-z}) \\ &= P(X \leq 1-e^{-z}) \\ &= \int_0^{1-e^{-z}} \frac{1}{\theta}(1-x)^{\frac{1}{\theta}-1} dx \\ &= -[1-(1-e^{-z})]^{\frac{1}{\theta}} - [-(1-0)^{\frac{1}{\theta}}] \\ &= 1-e^{-\frac{z}{\theta}}, \end{aligned}$$

dont la dérivée est la fonction qui à z associe $\frac{1}{\theta}e^{-\frac{z}{\theta}}$ si z positif, 0 sinon. Ainsi Z suit une loi exponentielle de paramètre θ.

5. La moyenne de Z est θ, donc la moyenne de l'estimateur (moyenne empirique d'un échantillon de Z) est également θ, ainsi l'estimateur est sans biais.
6. La variance de Z est θ^2 donc la variance de l'estimateur est $\frac{\theta^2}{n}$. Le support de X est $[0,1]$ qui ne dépend pas de θ et la log-vraisemblance est deux fois dérivable de dérivée seconde égale à :

$$\frac{d^2}{d\theta^2} \log L_\theta(X_1, \ldots, X_n) = \frac{n}{\theta^2} + \frac{2}{\theta^3} \sum_{j=1}^{n} \log(1-X_j).$$

L'information de Fisher est alors égale à :

$$\begin{aligned} I_{X_1,\ldots,X_n}(\theta) &= -\frac{n}{\theta^2} + \frac{2n\theta}{\theta^3} \\ &= \frac{n}{\theta^2}. \end{aligned}$$

L'estimateur est donc efficace.

4 Chapitre 3 : Estimation par intervalle de confiance

4.1 Rappels de cours

4.1.1 Principe général

Soit $T_n = \varphi_n(X_1, \ldots, X_n)$ estimateur ponctuel d'un paramètre θ, on cherche intervalle (aléatoire) $I_{1-\alpha}$ contenant T_n tel que $P(\theta \in I_{1-\alpha}) = 1-\alpha$, α est le risque. **ATTENTION :** La forme de l'intervalle ne doit dépendre en aucun cas d'un paramètre inconnu.

★ : **Estimateur ponctuel :** Une variable aléatoire T_n fonction des variables aléatoires $X_1, \ldots, X_n$, soit $T_n = \varphi_n(X_1, \ldots, X_n)$. Une variable aléatoire X est

une **fonction** de Ω (muni d'une tribu $\mathcal{A}$) dans E (espace des valeurs muni d'une tribu $\mathcal{T}$) telle que pour tout $A \in \mathcal{T}$, l'ensemble $(X \in A) = \{\omega \in \Omega : X(\omega) \in A\}$ est un élément de $\mathcal{A}$. L'ensemble Ω est l'univers (ensemble de tous les possibles).
Une **estimation ponctuelle** est une **réalisation** de T_n, c'est à dire une image $T_n(\omega) = \varphi_n(X_1(\omega), \ldots, X_n(\omega))$, on la note t_n ou $\hat{\theta}_n$ (si le paramètre à estimer est noté θ) et $X_j(\omega)$ est noté x_j. Ainsi $\hat{\theta}_n = \varphi_n(x_1, \ldots, x_n)$. ω représente un tirage au hasard, et $(x_1, \ldots, x_n)$ sont les mesures obtenues, on effectue ensuite l'estimation ponctuelle sur ces mesures.
★ : **Estimateur par intervalle :** L'estimateur est un intervalle aléatoire $I_{1-\alpha}$, c'est à dire une fonction de Ω dans l'ensemble des intervalles de $\mathbb{R}$.
De manière analogue, **l'intervalle de confiance observé** est une réalisation de $I_{1-\alpha}$, cette réalisation dépend de la valeur de l'estimation ponctuelle.

4.1.2 ★ : Quelques lois utiles

Définition 2 (Loi du Chi-deux)**.** *On appelle loi du Chi-deux à n degrés de liberté, notée $\mathcal{X}^2_n$, la loi de la variable aléatoire :*

$$R_n = \sum_{j=1}^{n} X_j^2, \tag{38}$$

où $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes suivant une loi normale centrée-réduite.
La densité d'une loi $\mathcal{X}^2_n$ a pour support $\mathbb{R}^+$ et est donnée par :

$$\begin{array}{rcl} f_{\mathcal{X}^2_n} : \mathbb{R} & \to & \mathbb{R}^+ \\ r & \to & \left\{ \begin{array}{l} \dfrac{1}{\Gamma\left(\dfrac{n}{2}\right) 2^{\frac{n}{2}}} r^{\frac{n}{2}-1} \exp\left(-\dfrac{r}{2}\right) \ \textit{si } r \geq 0 \\ 0 \ \textit{sinon.} \end{array} \right. \end{array},$$

où Γ est la fonction eulérienne donnée par :

$$\Gamma(a) = \int_0^{+\infty} x^{a-1} \exp(-x) dx,$$

pour tout $a > 0$.

Remarque sur la fonction eulérienne : La fonction Γ vérifie les propriétés suivantes :

- Pour tout entier n non nul, $\Gamma(n) = (n-1)!$.
- $\Gamma\left(\dfrac{1}{2}\right) = \sqrt{\pi}$.
- Pour tout $a > 1$, on a $\Gamma(a) = (a-1)\Gamma(a-1)$.

Cette remarque nous permet de calculer facilement la fonction Γ aux valeurs de la forme $\dfrac{k}{2}$. Par exemple,

$$\Gamma\left(\frac{7}{2}\right) = \frac{5}{2} \times \Gamma\left(\frac{5}{2}\right) = \frac{5}{2} \times \frac{3}{2} \times \Gamma\left(\frac{3}{2}\right) = \frac{5}{2} \times \frac{3}{2} \times \frac{1}{2} \times \Gamma\left(\frac{1}{2}\right) = \frac{15\sqrt{\pi}}{8}.$$

En exercice : Calculer $\Gamma\left(\frac{11}{2}\right)$, en déduire une expression plus explicite de la densité du Chi-deux à 11 degrés de liberté.

Définition 3 (Loi de Student)**.** *Soit Z et R_n deux variables aléatoires indépendantes de loi respective normale centrée-réduite et du Chi-deux à n degrés de liberté.*
Alors la loi de :

$$T_n = \frac{Z}{\sqrt{\frac{R_n}{n}}},$$

est appelée "loi de Student" à n degrés de liberté.
La densité d'une loi de Student à n degrés de liberté est donnée par :

$$t \in \mathbb{R} \to \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \times \frac{1}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}.$$

★ ***Lecture dans la table :*** *Pour α donné en colonne et le degré de liberté donné en ligne, la table donne le quantile d'ordre $1-\frac{\alpha}{2}$.*

Définition 4 (Loi de Fisher)**.** *La loi de Fisher à n et m degrés de liberté, notée $\mathcal{F}(n,m)$ est la loi de la variable aléatoire :*

$$F_{n,m} = \frac{m}{n} \times \frac{R_n}{R_m},$$

où R_n et R_m sont indépendantes suivant respectivement des lois du Chi-deux à n et m degrés de liberté.
La densité d'une loi de Fisher a pour support $\mathbb{R}^+$ et est donnée par :

$$\begin{array}{rcl} f_{\mathcal{F}(n,m)} : \mathbb{R} & \to & \mathbb{R}^+ \\ t & \to & \left\{ \begin{array}{c} \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \times \left(\frac{n}{m}\right)^{\frac{n}{2}} \times \frac{t^{\frac{n}{2}-1}}{\left(1+\frac{nt}{m}\right)^{\frac{n+m}{2}}} \ si\ r \geq 0 \\ 0\ sinon. \end{array} \right. \end{array},$$

4.1.3 ★ : Lois des estimateurs ponctuels des paramètres d'une loi normale $\mathcal{N}(\mu, \sigma^2)$

Soit $(X_1, \ldots, X_n)$ échantillon iid d'une loi normale $\mathcal{N}(\mu, \sigma^2)$.
Estimation de la moyenne : Dans le cas où la moyenne μ est inconnue, celle-ci est estimée par la moyenne empirique :

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j.$$

Pour déduire la loi de $\overline{X}_n$, nous utilisons la proposition suivante :

Proposition 2 (Stabilité des lois normales)**.** *Soient $X_1, \ldots, X_n$ des variables aléatoires* ***indépendantes*** *suivant les lois respectives $\mathcal{N}(\mu_1, \sigma_1^2), \ldots, \mathcal{N}(\mu_n, \sigma_n^2)$ et $\lambda_1, \ldots, \lambda_n$ des nombres réels, alors la somme $\sum_{j=1}^{n} \lambda_j X_j$ suit une loi normale de moyenne $\mu = \sum_{j=1}^{n} \lambda_j \mu_j$ et de variance $\sigma^2 = \sum_{j=1}^{n} \lambda_j^2 \sigma_j^2$.*

⋆⋆ **: Remarque sur la stabilité :** Nous disons que deux densités de probabilité ont la même forme si les variables aléatoires correspondantes X et Y ne différent uniquement par un facteur de translation et d'échelle, c'est à dire si il existe μ et σ tels que $Y = \sigma X + \mu$. On dit qu'une loi est stable si pour tout échantillon iid $(X_1, \ldots, X_n)$ de cette loi et tout n-uplet de réels $\lambda_1, \ldots, \lambda_n$, la densité de $\sum_{j=1}^{n} \lambda_j X_j$ est de la même forme que la densité des X_j. Il est démontré que la loi normale est la seule loi stable à variance finie, c'est à dire telle que :

$$\int_{-\infty}^{\infty} x^2 f(x) dx < +\infty.$$

L'intérêt des lois stables résident dans la généralisation du théorème limite-central au cas où les variables aléatoires étudiées n'ont pas de variance finie. On a vu que dans le cas où nous avons un échantillon iid de variables aléatoires de variance finie, la somme des variables aléatoires tend vers une loi normale. Dans le cas où la variance est infinie, introduisant α le plus grand nombre entre 0 et 2 tel que :

$$\int_{-\infty}^{\infty} |x|^{\alpha} f(x) dx < +\infty,$$

la somme converge en loi vers l'unique loi stable telle que :

$$\begin{aligned} \int_{-\infty}^{\infty} |x|^{\beta} f(x) dx &< +\infty, \text{ si } \beta < \alpha, \\ \int_{-\infty}^{\infty} |x|^{\beta} f(x) dx &= +\infty, \text{ si } \beta > \alpha. \end{aligned} \tag{39}$$

Une telle loi est appelée **loi alpha-stable**. Malgré le résultat intéressant concernant le théorème limite-central, l'étude des lois alpha-stables est très délicate, notamment car nous ne pouvons, sauf dans des cas particuliers, exprimer leur densité.
Enfin, la loi de la moyenne empirique est donnée par la proposition suivante :

Proposition 3 (Loi de la moyenne empirique)**.** *Soit $(X_1, \ldots, X_n)$ un échantillon iid de la loi normale $\mathcal{N}(\mu, \sigma^2)$, alors la moyenne empirique $\overline{X_n}$ suit une loi normale $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.*

Démonstration. En exercice : démontrez le en utilisant la stabilité des lois normales. □

Estimation de la variance : Dans le cas où la moyenne μ est connue et seule la variance σ^2 est inconnue, la variance est estimée par :

$$\Sigma_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2. \tag{40}$$

On a la proposition suivante :

Proposition 4 (Loi de la variance empirique, moyenne connue)**.** *Soit $(X_1, \ldots, X_n)$ un échantillon iid de la loi normale $\mathcal{N}(\mu, \sigma^2)$, alors la variance empirique Σ_n^2 est un estimateur sans biais de la variance. De plus $\frac{n\Sigma_n^2}{\sigma^2}$ suit une loi du Chi-deux à n degrés de liberté.*

Démonstration. Montrons tout d'abord que l'estimateur est sans biais. On a :

$$E(\Sigma_n^2) = \frac{1}{n}\sum_{j=1}^{n} E\left[(X_j - \mu)^2\right].$$

Par définition de la variance, on a : $E\left[(X_j - \mu)^2\right] = V(X_j) = \sigma^2$, d'où le résultat.
Montrons maintenant que $\frac{n\Sigma_n^2}{\sigma^2}$ suit une loi du Chi-deux à n degrés de liberté. Soit $Z_j = \frac{X_j - \mu}{\sigma}$, alors $Z_1, \ldots, Z_n$ est un échantillon iid de loi normale centrée-réduite. On en déduit que la somme $Z_1^2 + \ldots + Z_n^2$ suit une loi du Chi-deux à n degrés de liberté. Du fait que cette somme soit égale à $\frac{n\Sigma_n^2}{\sigma^2}$, on en déduit le résultat. □

Dans le cas où la moyenne est inconnue, la variance empirique peut être estimée par :

$$S_n^2 = \frac{1}{n}\sum_{j=1}^{n}\left(X_j - \overline{X}_n\right)^2. \tag{41}$$

On a la proposition suivante :

Proposition 5. *L'estimateur de la variance S_n^2 est biaisé de biais $\frac{n-1}{n}$.*

Démonstration. De $\left(X_j - \overline{X}_n\right)^2 = X_j^2 - 2X_j\overline{X}_n + \overline{X}_n^2$, on en déduit :

$$S_n^2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2 - \overline{X}_n^2.$$

Du fait que $E(X_j^2) = \mu^2 + \sigma^2$ et $E\left(\overline{X}_n^2\right) = \mu^2 + \frac{\sigma^2}{n}$, on déduit :

$$\begin{aligned} E\left[S_n^2\right] &= \mu^2 + \sigma^2 - \left(\mu^2 + \frac{\sigma^2}{n}\right) \\ &= \left(1 - \frac{1}{n}\right)\sigma^2 \\ &= \frac{n-1}{n} \times \sigma^2. \end{aligned}$$

□

Ainsi, l'estimateur sans biais de la variance est donné par :

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}_n\right)^2. \tag{42}$$

Nous avons la proposition suivante :

Proposition 6. *Soit $(X_1, \ldots, X_n)$ un échantillon iid de la loi normale $\mathcal{N}(\mu, \sigma^2)$, alors les variables aléatoires $\overline{X}_n$ et $\hat{S}_n^2$ sont indépendantes. De plus, la loi de $\frac{(n-1)\hat{S}_n^2}{\sigma^2}$ suit une loi du Chi-deux à $n-1$ degrés de liberté.*

Remarque : Les variables aléatoires $\overline{X}_n$ et S_n^2 sont également indépendantes et $\frac{nS_n^2}{\sigma^2}$ suit une loi du Chi-deux à $n-1$ degrés de liberté. En effet, $\frac{nS_n^2}{\sigma^2} = \frac{(n-1)\hat{S}_n^2}{\sigma^2}$.
Avant de démontrer la proposition, nous avons besoin de la définition et des lemmes suivants :

Définition 5 (Matrice orthogonale)**.** *Une matrice carrée P de taille $n \times n$ est dite orthogonale si :*

$$P \times P^* = P^* \times P = I,$$

où P^* est la transposée de P et I est la matrice identité.

Lemme 1 (Conservation du produit scalaire)**.** *Si P est une matrice orthogonale de taille $n \times n$ et si u et v sont deux vecteurs colonne de taille n, alors le produit scalaire $(Pu).(Pv)$ est égal au produit scalaire $u.v$.*

Lemme 2 (Invariance sphérique des lois normales)**.** *Soit $(Z_1, \ldots, Z_n)$ un échantillon iid d'une loi normale centrée-réduite, soit $\vec{Z}$ le vecteur colonne dont la $j^{ième}$ composante est Z_j et soit $\vec{\tilde{Z}} = P\vec{Z}$. Notant $\tilde{Z}_j$ la $j^{ième}$ composante de $\vec{\tilde{Z}}$, alors $(\tilde{Z}_1, \ldots, \tilde{Z}_n)$ est également un échantillon iid de la loi normale centrée-réduite.*

Preuve de la proposition. Soit $(Z_1, \ldots, Z_n)$ un échantillon iid de la loi normale centrée-réduite et soit :

$$\overline{Z}_n = \frac{1}{n}\sum_{j=1}^{n} Z_j.$$

Montrons tout d'abord que $\sum_{j=1}^{n}(Z_j - \overline{Z}_n)^2$ suit une loi du Chi-deux à $n-1$ degrés de liberté. Soit P une matrice orthogonale de taille $n \times n$ dont la dernière ligne est formée des valeurs $\frac{1}{\sqrt{n}}$, c'est à dire :

$$P = \begin{pmatrix} \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \end{pmatrix}.$$

On définit l'échantillon iid $(\tilde{Z}_1, \ldots, \tilde{Z}_n)$ de la loi normale centrée-réduite au travers du vecteur colonne :

$$\begin{pmatrix} \tilde{Z}_1 \\ \vdots \\ \tilde{Z}_n \end{pmatrix} = P \times \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}.$$

Le résultat du produit matriciel nous donne $\tilde{Z}_n = \sqrt{n}\overline{Z}_n$ (**En exercice :** vérifier le). De plus, comme P est une matrice orthogonale, on a conservation du produit scalaire, ainsi :

$$\begin{pmatrix} \tilde{Z}_1 \\ \vdots \\ \tilde{Z}_n \end{pmatrix} \cdot \begin{pmatrix} \tilde{Z}_1 \\ \vdots \\ \tilde{Z}_n \end{pmatrix} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} \cdot \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix},$$

on en déduit alors que :

$$\sum_{j=1}^{n} \tilde{Z}_j^2 = \sum_{j=1}^{n} Z_j^2.$$

Ainsi :

$$\begin{aligned} \sum_{j=1}^{n}(Z_j - \overline{Z}_n)^2 &= \sum_{j=1}^{n} Z_j^2 - n\overline{Z}_n^2 \\ &= \sum_{j=1}^{n} \tilde{Z}_j^2 - n\overline{Z}_n^2 \\ &= \sum_{j=1}^{n} \tilde{Z}_j^2 - \tilde{Z}_n^2 \\ &= \sum_{j=1}^{n-1} \tilde{Z}_j^2, \end{aligned}$$

qui est indépendant de $\tilde{Z}_n^2$ et donc de $\overline{Z}_n$ et suit une loi du Chi-deux à $n-1$ degrés de liberté (somme de $n-1$ carrés de variables normales centrées-réduites). Soit maintenant, $(X_1, \ldots, X_n)$ un échantillon iid de la loi normale $\mathcal{N}(\mu, \sigma^2)$. On

pose $Z_j = \dfrac{X_j - \mu}{\sigma}$. On a alors $\overline{Z}_n = \dfrac{\overline{X}_n - \mu}{\sigma}$. De plus, $\displaystyle\sum_{j=1}^{n}(Z_j - \overline{Z}_n)^2$ suit une loi du Chi-deux à $n-1$ degrés de liberté et est indépendante de $\overline{Z}_n$ et donc de $\overline{X}_n$. On montre finalement que $\displaystyle\sum_{j=1}^{n}(Z_j - \overline{Z}_n)^2 = \frac{(n-1)\hat{S}_n^2}{\sigma^2}$ et on en déduit le résultat. □

4.1.4 ★ : Cas de la loi normale : intervalle de confiance pour la moyenne

Dans tous les cas, nous choisissons comme estimateur ponctuel de la moyenne, la moyenne empirique donnée par :

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j,$$

où $(X_1, \ldots, X_n)$ est un échantillon iid de $\mathcal{N}(\mu, \sigma^2)$. **Même dans le cas non gaussien**, $\bar{X}_n$ est un estimateur **sans biais** de la moyenne et sa variance est $\dfrac{\sigma^2}{n}$. Dans le cas gaussien, la moyenne empirique suit de plus une loi normale. On peut tout de même approximer par une loi normale lorsque $n \geq 30$.

(a) Cas où la variance est connue :

Du fait que $\bar{X}_n$ suive une loi normale de moyenne μ et d'écart-type $\frac{\sigma}{\sqrt{n}}$, alors :

$$Z = \sqrt{n} \times \frac{\bar{X}_n - \mu}{\sigma},$$

suit une loi normale centrée-réduite. Soit $\alpha \in]0,1[$, on doit chercher z tel que :

$$P\left(\left|\sqrt{n} \times \frac{\bar{X}_n - \mu}{\sigma}\right| \leq z\right) = 1 - \alpha.$$

On en déduit que $z = z_{1-\frac{\alpha}{2}}$ (faire un dessin pour s'en convaincre) est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée-réduite.

De plus $\left|\sqrt{n} \times \frac{\bar{X}_n - \mu}{\sigma}\right| \leq z_{1-\frac{\alpha}{2}}$ est équivalent à $\mu \in \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right]$.

L'intervalle de confiance au risque α est donné par :

$$I_{1-\alpha} = \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right].$$

(b) Cas où la variance est inconnue :

On a toujours $\mu \in \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right]$ avec une probabilité $1 - \alpha$. Cependant, l'intervalle en question dépend du paramètre inconnu σ. L'idée est alors de remplacer dans l'expression de l'intervalle σ par un estimateur de σ.

Un estimateur sans biais de σ^2 est donné par :

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2.$$

Du fait que la moyenne empirique $\bar{X}_n$ suit une loi normale de moyenne μ et variance $\frac{\sigma^2}{n}$, alors $\bar{X}_n - \mu = \frac{\sigma}{\sqrt{n}} Z$, où Z suit une loi normale centrée-réduite. De plus, d'après précédemment, $\hat{S}_n = \frac{\sigma}{\sqrt{n-1}}\sqrt{R_{n-1}}$, où R_{n-1} suit une loi du Chi-deux à $n-1$ degrés de liberté et est indépendante de Z.
Au final :

$$\sqrt{n} \times \frac{\bar{X}_n - \mu}{\hat{S}_n} = \frac{Z}{\sqrt{\frac{R_{n-1}}{n-1}}},$$

suit une loi de Student à $n-1$ degrés de liberté.
On cherche alors le quantile $t_{1-\frac{\alpha}{2}}(n-1)$ de la loi de Student à $n-1$ degrés de liberté, qui est tel que :

$$P\left(\left|\sqrt{n} \times \frac{\bar{X}_n - \mu}{\hat{S}_n}\right| \leq t_{1-\frac{\alpha}{2}}(n-1)\right) = 1-\alpha.$$

L'intervalle de confiance au risque α est alors donné par :

$$I_{1-\alpha} = \left[\bar{X}_n - \frac{\hat{S}_n}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1), \bar{X}_n + \frac{\hat{S}_n}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)\right].$$

4.1.5 ★ : Cas de la loi normale : intervalle de confiance pour la variance

(a) : Cas moyenne connue Un estimateur sans biais de la variance est donné par :

$$\Sigma_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2.$$

De même $R_n = \frac{n}{\sigma^2}\Sigma_n^2$ suit une loi du Chi-deux à n degrés de liberté (somme de n carrés de loi normale). Soit $\alpha \in [0,1]$, on cherche a et b tels que :

$$P\left[a \leq \frac{n}{\sigma^2}\Sigma_n^2 \leq b\right] = 1-\alpha,$$

a et b peuvent être donnés par $a = r_{\frac{\alpha}{2}}(n)$ et $b = r_{1-\frac{\alpha}{2}}(n)$ quantiles d'ordre $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de la loi du Chi-deux à n degrés de liberté. De plus, $r_{\frac{\alpha}{2}}(n) \leq \frac{n}{\sigma^2}\Sigma_n^2 \leq r_{1-\frac{\alpha}{2}}(n)$ est équivalent à $\frac{n\Sigma_n^2}{r_{1-\frac{\alpha}{2}}(n)} \leq \sigma^2 \leq \frac{n\Sigma_n^2}{r_{\frac{\alpha}{2}}(n)}$, d'où l'intervalle de confiance au risque α pour la variance est :

$$I_{1-\alpha} = \left[\frac{n\Sigma_n^2}{r_{1-\frac{\alpha}{2}}(n)}, \frac{n\Sigma_n^2}{r_{\frac{\alpha}{2}}(n)}\right].$$

(b) : Cas moyenne inconnue Un estimateur sans biais de la variance est donné par :

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2.$$

Dans ce cas, comme il a été montré précédemment, $R_{n-1} = \frac{n-1}{\sigma^2}\hat{S}_n^2$ suit une loi du Chi-deux à $n-1$ degrés de liberté. Soit $\alpha \in [0,1]$, on cherche a et b tels que :

$$P\left[a \leq \frac{n-1}{\sigma^2}\hat{S}_n^2 \leq b\right] = 1-\alpha,$$

a et b peuvent être donnés par $a = r_{\frac{\alpha}{2}}(n-1)$ et $b = r_{1-\frac{\alpha}{2}}(n-1)$ quantiles d'ordre $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de la loi du Chi-deux à $n-1$ degrés de liberté. De plus, $r_{\frac{\alpha}{2}}(n-1) \leq \frac{n-1}{\sigma^2}\hat{S}_n^2 \leq r_{1-\frac{\alpha}{2}}(n-1)$ est équivalent à $\frac{(n-1)\hat{S}_n^2}{r_{1-\frac{\alpha}{2}}(n-1)} \leq \sigma^2 \leq \frac{(n-1)\hat{S}_n^2}{r_{\frac{\alpha}{2}}(n-1)}$, d'où l'intervalle de confiance au risque α pour la variance est :

$$I_{1-\alpha} = \left[\frac{(n-1)\hat{S}_n^2}{r_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)\hat{S}_n^2}{r_{\frac{\alpha}{2}}(n-1)}\right].$$

4.1.6 Cas général

On s'intéresse au cas où $(X_1, \ldots, X_n)$ n'est pas forcément issu d'une loi normale et le paramètre θ (considéré comme réel) n'est pas obligatoirement la moyenne ou la variance. Soit $\hat{\theta}_{n,MV}$ l'estimateur du maximum de vraisemblance de θ et $I_{X_1,\ldots,X_n}(\theta)$ l'information de Fisher de l'échantillon, alors par le théorème central-limite, si n est grand, on peut considérer :

$$\sqrt{I_{X_1,\ldots,X_n}(\hat{\theta}_{n,MV})} \times \left(\hat{\theta}_{n,MV} - \theta\right),$$

comme suivant une loi normale centrée-réduite. Si $z_{1-\frac{\alpha}{2}}$ désigne le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée-réduite, alors l'intervalle de confiance de θ au risque α est donné par :

$$I_{1-\alpha} = \left[\hat{\theta}_{n,MV} - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{I_{X_1,\ldots,X_n}(\hat{\theta}_{n,MV})}}, \hat{\theta}_{n,MV} + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{I_{X_1,\ldots,X_n}(\hat{\theta}_{n,MV})}}\right].$$

4.2 Exercices

4.2.1 Exercice 1

1. $E(\overline{X}_n) = p$ et $V(\overline{X}_n) = \frac{p(1-p)}{n}$.
2. S'en déduit de la question précédente.
3. Pour n grand, la loi de $\sqrt{n} \times \frac{\bar{X}_n - p}{\sqrt{p(1-p)}}$ est une loi normale centrée-réduite. La variance $p(1-p)$ est cependant inconnue.
   (a) Remplaçant $p(1-p)$ par son estimation $\bar{x}_n(1-\bar{x}_n) = 0.64 \times 0.36 = 0.2304$, on trouve finalement que $10 \times \frac{\bar{X}_n - p}{0.48}$ suit une loi normale centrée-réduite. On cherche a tel que $P(|Z| \leq a) = 0.95$ soit $P(-a \leq$

$Z \leq a) = 0.95$, soit $P(Z \leq a) - P(Z \leq -a) = 0.95$, soit $2P(Z \leq a) - 1 = 0.95$, a quantile d'ordre $\frac{1.95}{2} = 0.975$, ainsi $a = 1.96$ et donc $p \in [0.54592, 0.73408]$ avec un indice de confiance de 95%.

(b) Remplaçant $p(1-p)$ par 0.25, estimant donc l'écart-type par 0.5, on trouve $[0.542, 0.738]$.

Pour l'exercice 1, c'est beaucoup mieux (d'un point de vue mathématique) si on utilise la technique de l'exercice 2.

4.2.2 Exercice 2

Considérons la variable aléatoire X qui prend deux valeurs 1 (si boule blanche) et 0 (sinon). Cette variable suit une loi de Bernouilli de paramètre inconnu p. Soit $(X_1, \ldots, X_n)$ un échantillon iid de la loi de Bernouilli $\mathcal{B}(1, p)$, la vraisemblance de l'échantillon est donnée par :

$$L_p(X_1, \ldots, X_n) = p^{\sum_{j=1}^n X_j}(1-p)^{n-\sum_{j=1}^n X_j},$$

sa log-vraisemblance est :

$$\log L_p(X_1, \ldots, X_n) = \sum_{j=1}^n X_j \log(p) + \left(n - \sum_{j=1}^n X_j\right) \log(1-p),$$

et son score (obtenu en dérivant la log-vraisemblance par rapport au paramètre) est :

$$\frac{1}{p(1-p)} \sum_{j=1}^n X_j - \frac{n}{1-p}.$$

On en déduit que l'estimateur de la proportion de boules blanches est la moyenne empirique donnée par :

$$\bar{X}_n = \frac{1}{n} \sum_{j=1}^n X_j.$$

L'information de Fisher de l'échantillon en p est la variance du score qui est égale à $\frac{n}{p(1-p)}$. Comme $n = 300 \geq 30$, on en déduit que :

$$Z = \sqrt{\frac{n}{\bar{X}_n(1-\bar{X}_n)}}(p - \bar{X}_n),$$

suit approximativement une loi normale centrée-réduite.
On doit chercher z tel que :

$$P(|Z| < z) = 0.95,$$

z est alors le quantile d'ordre $1 - \frac{\alpha}{2} = 0.975$, soit $z_{0.975} = 1.96$. L'intervalle de confiance au risque 0.05 est donc :

$$I_{0.95} = \left[\bar{X}_n - \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \times 1.96, \bar{X}_n + \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \times 1.96\right].$$

Pour une proportion observée égale à $\bar{x}_n = \frac{83}{300} = 0.27666666$, la réalisation de l'intervalle de confiance est :

$$I_{0.95}(\bar{x}_n) = [0.23, 0.33].$$

4.2.3 Exercice 3

1. $\bar{x}_n = 4920$ et $s_n^2 = 51200$ et sans biais $\hat{s}_n^2 = 57600$.
2. $\sqrt{n}\frac{\bar{X}_n - m}{\hat{S}_n} = T_{n-1}$ où T_{n-1} loi de Student à $n-1$ degrés de liberté de densité :
$$t \to \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\Gamma\left(\frac{n-1}{2}\right)} \times \frac{1}{(1+\frac{t^2}{n-1})^{\frac{n}{2}}} \tag{43}$$
3. (a) Pour m : moyenne empirique.
Pour σ^2 :
$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2,$$
   (b) $\bar{x}_n = 4920$ et $\hat{s}_n^2 = 57600$.
   (c) Variance également inconnue, $\sqrt{n}\frac{\bar{X}_n - m}{\hat{S}_n} = T_{n-1}$. On cherche a tel que $P(|T_{n-1}| \leq a) = 0.95$. a est le quantile d'ordre $\frac{1.95}{2} = 0.975$ de la loi de Student à $n-1 = 8$ degrés de liberté. Lecture table Student (lire $\alpha = 0.05$ en colonne) donne $a = 2.306$ et donc $m \in [\bar{x}_n - 2.306 \times \frac{\hat{s}_n}{\sqrt{n}}, \bar{x}_n + 2.306 \times \frac{\hat{s}_n}{\sqrt{n}}] = [4735, 5104]$ avec un indice de confiance de 95%.
   (d) On cherche a tel que $P(|T_{n-1}| \leq a) = 0.90$, a quantile d'ordre $\frac{1.9}{2} = 0.95$ de la loi de Student à $n-1 = 8$ degrés de liberté. On trouve $a = 1.8595$, soit $m \in [\bar{x}_n - 1.8595 \times \frac{\hat{s}_n}{\sqrt{n}}, \bar{x}_n + 1.8595 \times \frac{\hat{s}_n}{\sqrt{n}}] = [4771, 5068]$ avec 90%.

4.2.4 Exercice 4

1. Estimation de m : $\bar{x}_n = 19.72$, l'estimateur $\bar{X}_n$ suit loi normale de moyenne m et variance σ^2. $\frac{\sqrt{n}(\bar{X}_n - m)}{\sigma}$ loi normale centrée-réduite. σ inconnu, $\frac{\sqrt{n}(\bar{X}_n - m)}{\hat{S}_n}$ suit loi de Student à 9 degrés de liberté. $\hat{s}_n^2 = 0.6773333$. On cherche a quantile d'ordre 0.95 dans la table de Student (correspondant à $\alpha = 0.1$), c'est $a = 1.8331$. D'où $m \in [\bar{x}_n - 1.8331 \times \frac{\hat{s}_n}{\sqrt{n}}, \bar{x}_n + 1.8331 \times \frac{\hat{s}_n}{\sqrt{n}}] = [19.24, 20.19]$.
2. La loi de $R_{n-1} = (n-1) \times \frac{\hat{S}_n^2}{\sigma^2}$ est chi-deux à $n-1 = 9$ degrés de liberté. On cherche a et b tels que :
$$P(a \leq R_9 \leq b) = 0.9,$$
a est le quantile d'ordre 0.05 de la Chi-deux à 9 degrés de liberté (**ATTENTION AU PIEGE :** selon la table, cela peut se lire avec la valeur $1 - 0.05 = 0.95$

en colonne. Dans ce type de table, la valeur en colonne est décroissante de gauche à droite.), soit $a = 3.325$ et b est le quantile d'ordre 0.95 (**MEME REMARQUE :** lire 0.05 en colonne dans certaine table), soit $b = 16.919$. On en déduit alors l'intervalle de confiance au risque 0.1 :

$$\sigma^2 \in [\frac{(n-1)\hat{s}_n^2}{b}, \frac{(n-1)\hat{s}_n^2}{a}] = [0.36, 1.83].$$

3. $Z = \frac{\sqrt{n}(\bar{X}_n - m)}{\sqrt{0.69}}$ suit une loi normale centrée-réduite. On cherche z tel que :

$$P(|Z| \leq z) = 0.9,$$

soit $z = z_{0.95} = 1.65$ quantile d'ordre 0.95 de la loi normale centrée-réduite. L'inégalité $|Z| \leq 1.65$ est équivalente à :

$$m \in \left[\overline{X}_n - \frac{\sqrt{0.69}}{\sqrt{n}} \times 1.65, \overline{X}_n + \frac{\sqrt{0.69}}{\sqrt{n}} \times 1.65\right].$$

Pour $n = 10$ et $\overline{x}_n = 19.72$, on en déduit l'intervalle de confiance au risque 0.1 :

$$I_{0.9} = [19.29, 20.15].$$

4.2.5 Exercice 5

1. Pour calculer l'espérance, on utilise la formule :

$$E(X) = a \int_0^{+\infty} x e^{-ax} dx.$$

Premier réflexe à avoir : Est-ce que l'on reconnaît une primitive ? Dans notre cas, nous ne reconnaissons pas de primitive.
Deuxième réflexe à avoir : On utilise alors une intégration par partie.
Si on n'arrive pas à utiliser une intégration par parties : C'est que le calcul de l'intégrale en question est **hors-programme** et dans ce cas on ne vous demande pas de la calculer. Il existe en effet d'autres méthodes pour calculer une intégrale. Par exemple, le théorème de Fubini-Tonelli (utilisé notamment pour montrer que la densité de la loi normale est bien une densité de probabilité), la propriété du score (utilise le fait que sous les conditions que vous connaissez, la moyenne du score est nulle), la formule des résidus, le théorème de Parseval (permet de calculer une intégrale à partir d'une somme dont on connaît la valeur et vis-versa) et certainement plein d'autres.
Dans notre cas : on fait une intégration par parties (heureusement, sinon, on ne vous aurait pas demandé de calculer cette intégrale). On pose alors $u(x) = x$ (u doit être facile à dériver) et $v'(x) = ae^{-ax}$ (v' doit être

facile à intégrer). Ainsi $u'(x) = 1$ et $v(x) = -e^{-ax}$. On a alors :

$$\begin{aligned} E(X) &= [u(x)v(x)]_0^{+\infty} - \int_0^{+\infty} u'(x)v(x)dx \\ &= 0 + \int_0^{+\infty} e^{-ax}dx \\ &= \left[-\frac{1}{a}e^{-ax}\right]_0^{+\infty} \\ &= \frac{1}{a}. \end{aligned}$$

Pour calculer la variance, on calcule d'abord $E(X^2)$, soit :

$$E(X^2) = a\int_0^{+\infty} x^2 e^{-ax}dx.$$

On effectue également une intégration par parties en posant $u(x) = x^2$ et $v'(x) = ae^{-ax}$, ainsi $u'(x) = 2x$ et $v(x) = -e^{-ax}$. On a alors :

$$E(X^2) = \underbrace{\left[-x^2e^{-ax}\right]_0^{+\infty}}_{0} + 2\int_0^{+\infty} xe^{-ax}dx.$$

On reconnait que :

$$\int_0^{+\infty} xe^{-ax}dx = \frac{1}{a}E(X) = \frac{1}{a^2},$$

ainsi $V(X) = \dfrac{2}{a^2} - \dfrac{1}{a^2} = \dfrac{1}{a^2}$.

2. Fonction de vraisemblance :

$$L_a(X_1, \ldots, X_n) = a^n e^{-a\sum_{j=1}^n X_j}.$$

La log-vraisemblance de l'échantillon est alors :

$$\log L_a(X_1, \ldots, X_n) = n\log(a) - a\sum_{j=1}^n X_j.$$

Nous avons toutes les propriétés de régularité, à savoir :
– Le support de la loi est $\mathbb{R}^+$ et est indépendant de a.
– L'ensemble des valeurs du paramètre est l'intervalle ouvert $]0, +\infty[$.
– La log-vraisemblance est dérivable deux-fois.
La dérivée première de la log-vraisemblance (score) est :

$$\frac{n}{a} - \sum_{j=1}^n X_j,$$

ainsi, en résolvant :

$$\frac{n}{a} - \sum_{j=1}^{n} X_j = 0,$$

on en déduit l'estimateur du maximum de vraisemblance de a :

$$T_n = \frac{n}{\sum_{j=1}^{n} X_j} = \frac{1}{\overline{X}_n}.$$

3. La moyenne de la moyenne empirique est égale à la moyenne de la variable en question donc $E(\overline{X}_n) = \frac{1}{a}$. De plus, on a $V(\overline{X}_n) = \dfrac{V(X)}{n} = \dfrac{1}{na^2}$, ainsi $\overline{X}_n$ est bien un estimateur sans biais et convergent de $\frac{1}{a}$.
4. Pour avoir la loi asymptotique, on va utiliser le fait que :

$$Z = \sqrt{I_{X_1,\ldots,X_n}(T_n)}(a - T_n),$$

suit asymptotiquement une loi normale centrée-réduite. L'information de Fisher en a est la variance du score, ainsi :

$$I_{X_1,\ldots,X_n}(a) = V\left(\sum_{j=1}^{n} X_j\right) = \frac{n}{a^2},$$

on en déduit :

$$I_{X_1,\ldots,X_n}(T_n) = n\overline{X}_n^2.$$

On montre alors facilement que :

$$Z = \sqrt{n} \times (a\overline{X}_n - 1).$$

On en déduit le résultat.

5. On cherche z tel que :

$$P(|Z| \leq z) = 0.95, \tag{44}$$

c'est le quantile d'ordre 0.975, soit $z_{0.975} = 1.96$. $\sqrt{n} \times \left|a\overline{X}_n - 1\right| \leq 1.96$ est équivalent à :

$$-\frac{1.96}{\sqrt{n}} \leq a\overline{X}_n - 1 \leq \frac{1.96}{\sqrt{n}},$$

qui est également équivalent à :

$$\frac{1}{\overline{X}_n} \times \left(1 - \frac{1.96}{\sqrt{n}}\right) \leq a \leq \frac{1}{\overline{X}_n} \times \left(1 + \frac{1.96}{\sqrt{n}}\right).$$

Pour $n = 100$ et $\overline{x}_n = 312.8$, on en déduit l'intervalle de confiance pour a au niveau 0.95 :

$$I_{0.95}(a) = [0.0026, 0.0038].$$

4.2.6 Exercice 6

Dans cet exercice, nous ne connaissons ni la moyenne ni la variance. On supposera pour simplifier que 397 est l'écart-type corrigé (racine carrée de la variance sans biais). Notons μ la moyenne, σ l'écart-type, $\overline{X}_n$ la moyenne empirique et $\hat{S}_n^2$ l'estimateur sans biais de la variance. **ATTENTION :** on ne vous dit pas que la loi est une loi normale, ainsi :

$$\sqrt{n} \times \frac{(\overline{X}_n - m)}{\hat{S}_n},$$

ne suit pas une loi de Student.
Cependant, $n = 45 \geq 30$, ainsi on peut considérer que :

$$Z = \sqrt{n} \times \frac{(\overline{X}_n - m)}{\hat{S}_n},$$

suit une loi normale centrée-réduite (**ATTENTION :** ce n'est pas approximé par une loi de Student).
On cherche z tel que :

$$P(|Z| < z) = 0.9,$$

soit $z = z_{0.95} = 1.65$. L'intervalle de confiance est de la forme :

$$I_{0.9} = \left[\overline{X}_n - \frac{\hat{S}_n}{\sqrt{n}} \times 1.65, \overline{X}_n + \frac{\hat{S}_n}{\sqrt{n}} \times 1.65 \right],$$

sa réalisation pour $n = 45$, $\overline{x}_n = 988.15$ et $\hat{s}_n = 397$ est :

$$I_{0.9} = [890.5, 1085.8].$$

5 Chapitre 4 : Test de Neymann-Pearson, de Student et de Fisher

5.1 Rappels de cours

5.1.1 $\star$: Test de Neymann-Pearson, approche bayésienne

On considère une variable aléatoire X et deux hypothèses H_0 et H_1 telles que la loi de X diffère selon l'hypothèse. On observe un échantillon $(x_1, \ldots, x_n)$, réalisation de l'échantillon iid $(X_1, \ldots, X_n)$ de la variable aléatoire X, le test de Neymann-Pearson consiste alors à décider quelle hypothèse est **a posteriori** la plus probable.
Pour cela, on se donne :

- Les probabilités **a priori** des hypothèses $\pi_0 = P(H_0)$ et $\pi_1 = P(H_1)$ telles que $\pi_0 + \pi_1 = 1$. Ces probabilités a priori représentent l'idée que l'on a sur les hypothèses avant toute expérience.

– L'attache aux données : les deux densités de X : $x \to p(x|H_0)$ et $x \to p(x|H_1)$.

On en déduit les deux vraisemblances $p(x_1, \ldots, x_n|H_0) = \prod_{j=1}^{n} p(x_j|H_0)$ et $p(x_1, \ldots, x_n|H_1) = \prod_{j=1}^{n} p(x_j|H_1)$.

La probabilité **a posteriori** de chacune des hypothèses est alors donnée par la règle de **Bayes** :

$$P(H_i|x_1, \ldots, x_n) = \frac{\pi_i \times p(x_1, \ldots, x_n|H_i)}{\pi_0 \times p(x_1, \ldots, x_n|H_0) + \pi_1 \times p(x_1, \ldots, x_n|H_1)}, \qquad (45)$$

pour $i = 0, 1$.
On décide alors l'hypothèse la plus probable **a posteriori**, ainsi on décidera H_0 si :

$$P(H_0|x_1, \ldots, x_n) > P(H_1|x_1, \ldots, x_n).$$

Remarque : Lorsque $\pi_0 = \pi_1 = \frac{1}{2}$, on décide H_0 si et seulement si $p(x_1, \ldots, x_n|H_0) > p(x_1, \ldots, x_n|H_1)$, H_0 est alors l'hypothèse **maximisant la vraisemblance** parmi les deux hypothèses.
Le dénominateur figurant dans la formule de Bayes (45) ne dépend pas de l'hypothèse, ainsi décider l'hypothèse H_0 est équivalent à :

$$\pi_0 \times p(x_1, \ldots, x_n|H_0) > \pi_1 \times p(x_1, \ldots, x_n|H_1), \qquad (46)$$

ce qui est également équivalent à :

$$\frac{p(x_1, \ldots, x_n|H_0)}{p(x_1, \ldots, x_n|H_1)} > \frac{\pi_1}{\pi_0}. \qquad (47)$$

$(x_1, \ldots, x_n)$ étant réalisation de la variable aléatoire $(X_1, \ldots, X_n)$, alors $\frac{p(x_1, \ldots, x_n|H_0)}{p(x_1, \ldots, x_n|H_1)}$ est réalisation de la variable aléatoire $\frac{p(X_1, \ldots, X_n|H_0)}{p(X_1, \ldots, X_n|H_1)}$.
On définit les risques de première espèce α et de deuxième espèce :

$$\alpha = P(\text{décider} H_1|H_0), \qquad (48)$$
$$\beta = P(\text{décider} H_0|H_1). \qquad (49)$$

Ceux-ci sont respectivement égaux à :

$$\alpha = P\left(\frac{p(X_1, \ldots, X_n|H_0)}{p(X_1, \ldots, X_n|H_1)} < \frac{\pi_1}{\pi_0}|H_0\right), \qquad (50)$$
$$\beta = P\left(\frac{p(X_1, \ldots, X_n|H_0)}{p(X_1, \ldots, X_n|H_1)} > \frac{\pi_1}{\pi_0}|H_1\right). \qquad (51)$$

Remarque : $P\left(\frac{p(X_1,\ldots,X_n|H_0)}{p(X_1,\ldots,X_n|H_1)} < \frac{\pi_1}{\pi_0}|H_0\right)$ est la probabilité que $\frac{p(X_1,\ldots,X_n|H_0)}{p(X_1,\ldots,X_n|H_1)} < \frac{\pi_1}{\pi_0}$ lorsque X a pour densité $x \to p(x|H_0)$.
★ **:** Des dernières formulations des risques de première espèce et seconde espèce, on remarque que l'on peut se donner implicitement les probabilités a priori en imposant le risque de première espèce (resp. de seconde espèce), c'est cette approche qui est utilisée dans les exercices.

5.1.2 ★ : Décider, accepter et rejeter

Attention : Décider l'hypothèse H_i ne signifie pas accepter l'hypothèse H_i, cela signifie seulement que l'on rejete l'autre hypothèse H_{1-i}. On acceptera H_i seulement si H_i et H_{1-i} sont contraires.

5.1.3 ⋆ : Test de Neymann-Pearson dans le cas normal, approche bayesienne

Soit X une variable aléatoire, on souhaite tester les deux hypothèses :
- H_0 : X suit une loi normale de moyenne μ_0 et de variance σ_0^2.
- H_1 : X suit une loi normale de moyenne μ_1 et de variance σ_1^2.

Soit $(x_1,\ldots,x_n)$ un échantillon observé, réalisation d'un échantillon iid de variables aléatoires $(X_1,\ldots,X_n)$. On a :

$$\begin{aligned} p(x_0,\ldots,x_n|H_0) &= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma_0^n}\exp\left(-\frac{1}{2\sigma_0^2}\sum_{j=1}^n(x_j-\mu_0)^2\right) \\ p(x_0,\ldots,x_n|H_1) &= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma_1^n}\exp\left(-\frac{1}{2\sigma_1^2}\sum_{j=1}^n(x_j-\mu_1)^2\right). \end{aligned}$$

On montre alors que décider H_0 est équivalent à :

$$\frac{\sigma_1^n}{\sigma_0^n}\times\exp\left(-\frac{1}{2\sigma_0^2}\sum_{j=1}^n(x_j-\mu_0)^2+\frac{1}{2\sigma_1^2}\sum_{j=1}^n(x_j-\mu_1)^2\right) > \frac{\pi_1}{\pi_0}, \qquad (52)$$

également équivalent à :

$$n\log\left(\frac{\sigma_1}{\sigma_0}\right)-\frac{1}{2\sigma_0^2}\sum_{j=1}^n(x_j-\mu_0)^2+\frac{1}{2\sigma_1^2}\sum_{j=1}^n(x_j-\mu_1)^2 > \log\left(\frac{\pi_1}{\pi_0}\right). \qquad (53)$$

Considérons le cas $\sigma_0=\sigma_1=\sigma$.
Si on suppose $\mu_0>\mu_1$, en développant (53), décider H_0 est alors équivalent à :

$$\bar{x}_n > \frac{\sigma^2\log\left(\frac{\pi_1}{\pi_0}\right)}{(\mu_0-\mu_1)n}+\frac{\mu_0+\mu_1}{2}. \qquad (54)$$

En exercice : Démontrer la formule précédente. Quelle serait la règle de décision si $\mu_0 < \mu_1$?

Sous H_0, la moyenne empirique a pour moyenne μ_0 et variance $\frac{\sigma^2}{n}$, ainsi le risque de première espèce est égal à :

$$\alpha = P\left(Z < \frac{\sigma \log\left(\frac{\pi_1}{\pi_0}\right)}{(\mu_0 - \mu_1)\sqrt{n}} + \frac{\sqrt{n}(\mu_1 - \mu_0)}{2\sigma}\right), \tag{55}$$

où Z suit une loi normale centrée-réduite.

Sous H_1, la moyenne empirique a pour moyenne μ_1 et variance $\frac{\sigma^2}{n}$, ainsi le risque de seconde espèce est égal à :

$$\beta = P(Z > \frac{\sigma \log\left(\frac{\pi_1}{\pi_0}\right)}{(\mu_0 - \mu_1)\sqrt{n}} + \frac{\sqrt{n}(\mu_0 - \mu_1)}{2\sigma}). \tag{56}$$

En exercice : Quels seraient les risques de première et seconde espèce si $\mu_0 < \mu_1$?

Important : Privilégier a priori l'hypothèse H_0 consiste à se donner une loi a priori telle que $\pi_0 > \pi_1$, on montre que cela est équivalent à $\beta > \alpha$.

Ainsi, si on ne veut pas privilégier une hypothèse par rapport à une autre, on se donne $\pi_0 = \pi_1 = \frac{1}{2}$. La règle de décision se simplifie en :

$$\bar{x}_n > \frac{\mu_0 + \mu_1}{2}, \text{ si } \mu_0 > \mu_1,$$
$$\bar{x}_n < \frac{\mu_0 + \mu_1}{2}, \text{ si } \mu_0 < \mu_1,$$

5.1.4 ★ Test de Neymann-Pearson dans le cas normal (variance commune et connue), approche simplifiée

A retenir : Pour un test de Neymann-Pearson du type :

- $H_0 : \mu = \mu_0$.
- $H_1 : \mu = \mu_1$,

la règle de décision est :

On décide H_0 si :

- $\bar{x}_n > c$ si $\mu_0 > \mu_1$.
- $\bar{x}_n < c$ si $\mu_0 < \mu_1$.

Pour déterminer c, il existe deux grandes méthodes :

1. Bayesienne : on se donne les probabilités a priori π_0 et π_1, dans ce cas $c = \frac{\sigma^2 \log\left(\frac{\pi_1}{\pi_0}\right)}{(\mu_0 - \mu_1)n} + \frac{\mu_0 + \mu_1}{2}$.
   Cas particulier : ne pas privilégier aucune des hypothèses est équivalent à $c = \frac{\mu_0 + \mu_1}{2}$.

2. A risque constant : on impose un risque de première espèce α.
Dans le cas où $\mu_0 > \mu_1$, on cherche c tel que :

$$\alpha = P(\overline{X}_n < c|H_0).$$

Sous H_0, $\overline{X}_n$ suit une loi normale de moyenne μ_0 et variance $\frac{\sigma^2}{n}$, ainsi $Z = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma}\times$ suit loi normale centrée-réduite. On cherche z_α tel que :

$$\alpha = P(Z < z_\alpha),$$

ainsi $c = \mu_0 + \frac{\sigma}{\sqrt{n}} \times z_\alpha$, où z_α est le quantile d'ordre α de la normale centrée-réduite.
On peut remarquer que la règle de décision est alors équivalente à : décider H_0 si $\sqrt{n}\frac{\overline{x}_n - \mu_0}{\sigma} > z_\alpha$.
Dans le cas où $\mu_0 < \mu_1$, on cherche c tel que :

$$\alpha = P(\overline{X}_n > c|H_0),$$

c'est équivalent à chercher $z_{1-\alpha}$ tel que :

$$\alpha = P(Z > z_{1-\alpha}),$$

où Z suit une loi normale centrée-réduite. La règle de décision devient : on décide H_0 si $\sqrt{n}\frac{\overline{x}_n - \mu_0}{\sigma} < z_{1-\alpha}$.
En exercice : Sous la condition $\mu_0 > \mu_1$, montrer que la loi a priori du test à risque constant est la loi de Bernouilli de paramètre

$$p = \frac{\exp\left(\frac{\sqrt{n}(\mu_0-\mu_1)z_\alpha}{\sigma} + \frac{n(\mu_0-\mu_1)^2}{2\sigma^2}\right)}{1+\exp\left(\frac{\sqrt{n}(\mu_0-\mu_1)z_\alpha}{\sigma} + \frac{n(\mu_0-\mu_1)^2}{2\sigma^2}\right)}.$$

donnée par :

$$\begin{aligned}\pi_0 &= \frac{1}{1+\exp\left(\frac{\sqrt{n}(\mu_0-\mu_1)z_\alpha}{\sigma} + \frac{n(\mu_0-\mu_1)^2}{2\sigma^2}\right)}.\\ \pi_1 &= \frac{\exp\left(\frac{\sqrt{n}(\mu_0-\mu_1)z_\alpha}{\sigma} + \frac{n(\mu_0-\mu_1)^2}{2\sigma^2}\right)}{1+\exp\left(\frac{\sqrt{n}(\mu_0-\mu_1)z_\alpha}{\sigma} + \frac{n(\mu_0-\mu_1)^2}{2\sigma^2}\right)}.\end{aligned}$$

Quelle est la loi a priori dans le cas où $\mu_0 < \mu_1$?

★ **: TRES IMPORTANT ! ! ! ! ! ! Dans la pratique :** Dans les exercices, vous n'aurez à utiliser seulement deux méthodes :

1. Soit on ne privilégie aucune hypothèse, donc $c = \frac{\mu_0 + \mu_1}{2}$ (μ_0 et μ_1 sont bien sûr donnés dans l'énoncé), le risque de première espèce est alors calculé par :

$$\begin{aligned} \alpha &= P\left(Z < \frac{c-\mu_0}{\sigma} \times \sqrt{n}\right) \text{ si } \mu_0 > \mu_1 \\ \alpha &= P\left(Z > \frac{c-\mu_0}{\sigma} \times \sqrt{n}\right) \text{ si } \mu_0 < \mu_1, \end{aligned}$$

et le risque de seconde espèce est calculé par :

$$\begin{aligned} \beta &= P\left(Z > \frac{c-\mu_1}{\sigma} \times \sqrt{n}\right) \text{ si } \mu_0 > \mu_1 \\ \beta &= P\left(Z < \frac{c-\mu_1}{\sigma} \times \sqrt{n}\right) \text{ si } \mu_0 < \mu_1. \end{aligned}$$

2. Soit on effectue un test à risque de première espèce constant. On se donne α dans l'énoncé et on calcule c tel que :

$$\begin{aligned} \alpha &= P\left(Z < \frac{c-\mu_0}{\sigma} \times \sqrt{n}\right) \text{ si } \mu_0 > \mu_1 \\ \alpha &= P\left(Z > \frac{c-\mu_0}{\sigma} \times \sqrt{n}\right) \text{ si } \mu_0 < \mu_1. \end{aligned}$$

Le risque de seconde espèce est alors calculé par :

$$\begin{aligned} \beta &= P\left(Z > \frac{c-\mu_1}{\sigma} \times \sqrt{n}\right) \text{ si } \mu_0 > \mu_1 \\ \beta &= P\left(Z < \frac{c-\mu_1}{\sigma} \times \sqrt{n}\right) \text{ si } \mu_0 < \mu_1. \end{aligned}$$

Si $\alpha > \beta$, cela signifie que l'on a privilégié H_1, si $\alpha < \beta$, on a privilégié H_0.

ATTENTION A NE PAS SE FAIRE PIEGER LE JOUR DE L'EXAMEN : On peut aussi utiliser une variante, imposer le risque de seconde espèce au lieu d'imposer le risque de première espèce. Ce type de test est appelé test à taux de fausses alarmes constant (TFAC) et est beaucoup plus utilisé dans l'industrie que le test à risque de première espèce constant. Dans ce cas, on vous donne β dans l'énoncé et vous devez chercher c tel que :

$$\begin{aligned} \beta &= P\left(Z > \frac{c-\mu_1}{\sigma} \times \sqrt{n}\right) \text{ si } \mu_0 > \mu_1 \\ \beta &= P\left(Z < \frac{c-\mu_1}{\sigma} \times \sqrt{n}\right) \text{ si } \mu_0 < \mu_1. \end{aligned}$$

Ensuite, vous calculez le risque de première espèce.

5.1.5 ★ : PARAGRAPHE TRES IMPORTANT Tests de la valeur de la moyenne, de la valeur de la variance, de comparaison de moyenne et de variance

Ce paragraphe traite de tous les tests que vous aurez à utiliser dans les exercices. Il peut être lu indépendamment des paragraphes précédents. Dans ce paragraphe, il n'y a aucune démonstration mais seulement les méthodes à appliquer dans les différentes situations.

(a) Introduction

Soit $(X_1, \ldots, X_n)$ un échantillon iid d'une loi normale de moyenne m inconnue et d'écart-type σ connu. Nous avons vu que pour un test du type :

- $H_0 : m = m_0$.
- $H_1 : m = m_1$.

La stratégie de décision est de la forme : on décide H_0 si :

$$\overline{x}_n > c,$$

si $m_0 > m_1$. Cette stratégie de décision est en fait équivalente à :

$$\sqrt{n} \times \frac{\overline{X}_n - m_0}{\sigma} > z,$$

où $z = \sqrt{n} \times \dfrac{c - m_0}{\sigma}$. Ainsi, une manière plus simple de procéder est de chercher z à partir du risque de première espèce α tel que :

$$\begin{aligned} \alpha &= P\left(\text{décider } H_1 | H_0\right) \\ &= P\left(\sqrt{n} \times \frac{\overline{X}_n - m_0}{\sigma} \leq z | H_0\right). \end{aligned}$$

Sous H_0, $\sqrt{n} \times \dfrac{\overline{X}_n - m_0}{\sigma}$ suit une loi normale centrée-réduite, ainsi z est le quantile d'ordre α de la loi centrée-réduite.

La variable aléatoire $\sqrt{n} \times \dfrac{\overline{X}_n - m_0}{\sigma}$ est appelée "statistique du test" et ne dépend que de l'hypothèse nulle, elle doit vérifier le fait que sous H_0, sa loi ne dépend pas de paramètres inconnus.

Quant à la forme de la règle de décision, elle dépend de l'hypothèse alternative (nous le verrons dans les exemples suivants).

Finalement, le seuil z dépend du risque de première espèce.

(b) Test de la valeur de la moyenne, écart-type connu

Dans ce type de test, nous observons la réalisation d'une loi normale de moyenne m inconnue et d'écart-type σ connu. L'hypothèse nulle de ce test est :

$H_0 : m = m_0$,

où m_0 est donné dans l'énoncé.

Pour ce type d'hypothèse nulle, la statistique du test est obligatoirement :

$$Z = \sqrt{n} \times \frac{\overline{X}_n - m_0}{\sigma}.$$

Sous l'hypothèse H_0, la statistique Z suit une loi normale centrée-réduite.
La forme de la règle de décision dépend de l'hypothèse alternative :

- Pour $H_1 : m \neq m_0$:
  On décide H_0 si :
  $$\sqrt{n} \times \frac{|\overline{x}_n - m_0|}{\sigma} \leq z,$$
  où z est fonction du risque de première espèce α :
  $$\alpha = P\left(\sqrt{n} \times \frac{|\overline{X}_n - m_0|}{\sigma} > z | H_0\right).$$
  z est donc le quantile d'ordre $1 - \dfrac{\alpha}{2}$ de la loi normale centrée-réduite.
  Quant au risque de seconde espèce, il est calculé pour une valeur particulier de $m_1 \neq m_0$ et vaut :
  $$\beta(m_1) = P\left(\sqrt{n} \times \frac{|\overline{X}_n - m_0|}{\sigma} \leq z | m = m_1\right).$$
- Pour $H_1 : m = m_1$ avec $m_1 < m_0$, ou $H_1 : m < m_0$:
  On décide H_0 si :
  $$\sqrt{n} \times \frac{\overline{x}_n - m_0}{\sigma} \geq z,$$
  où z est fonction du risque de première espèce α :
  $$\alpha = P\left(\sqrt{n} \times \frac{\overline{X}_n - m_0}{\sigma} \leq z | H_0\right).$$
  z est donc le quantile d'ordre α de la loi normale centrée-réduite.
  Si $H_1 : m = m_1$, le risque de seconde espèce vaut :
  $$\beta = P\left(\sqrt{n} \times \frac{\overline{X}_n - m_0}{\sigma} \geq z | H_1\right).$$
  Si $H_1 : m < m_0$: le risque de seconde espèce est calculé pour une valeur particulière $m_1 < m_0$ et vaut :
  $$\beta(m_1) = P\left(\sqrt{n} \times \frac{\overline{X}_n - m_0}{\sigma} \geq z | m = m_1\right).$$
- Pour $H_1 : m = m_1$ avec $m_1 > m_0$, ou $H_1 : m > m_0$:
  On décide H_0 si :
  $$\sqrt{n} \times \frac{\overline{x}_n - m_0}{\sigma} \leq z,$$
  où z est fonction du risque de première espèce α :
  $$\alpha = P\left(\sqrt{n} \times \frac{\overline{X}_n - m_0}{\sigma} \geq z | H_0\right).$$

z est donc le quantile d'ordre $1-\alpha$ de la loi normale centrée-réduite.
Si $H_1 : m = m_1$, le risque de seconde espèce vaut :

$$\beta = P\left(\sqrt{n} \times \frac{\overline{X}_n - m_0}{\sigma} \leq z | H_1\right).$$

Si $H_1 : m > m_0$: le risque de seconde espèce est calculé pour une valeur particulière $m_1 > m_0$ et vaut :

$$\beta(m_1) = P\left(\sqrt{n} \times \frac{\overline{X}_n - m_0}{\sigma} \leq z | m = m_1\right).$$

(c) Test de la valeur de la moyenne, écart-type inconnu
ATTENTION : Ce type de test n'est possible que pour des variables suivant des lois normales même si n est supérieur à 30. Si n est supérieur à 30 et que les variables ne suivent pas des lois normales, on utilisera le test précédent en remplaçant l'écart-type σ par son estimation.
L'hypothèse nulle associée à ce test est toujours :
$H_0 : m = m_0$,
pour des données issues d'un échantillon d'une loi normale de moyenne m et d'écart-type σ. Mais, cette fois-ci, l'écart-type est inconnu.
Dans ce cas, la statistique du test est :

$$T_{n-1} = \sqrt{n} \times \frac{\overline{X}_n - m_0}{\hat{S}_n},$$

où $\hat{S}_n^2$ est l'estimateur sans biais de la variance.
Sous H_0, T_{n-1} suit une loi de Student à $n-1$ degrès de liberté.
Quant à la règle de décision, elle dépend de l'hypothèse alternative H_1 :

- Pour $H_1 : m \neq m_0$:
  On décide H_0 si :
  $$\sqrt{n} \times \frac{|\overline{x}_n - m_0|}{\sigma} \leq t,$$
  où t est fonction du risque de première espèce α :
  $$\alpha = P\left(\sqrt{n} \times \frac{\left|\overline{X}_n - m_0\right|}{\sigma} > t | H_0\right).$$
  t est donc le quantile d'ordre $1 - \dfrac{\alpha}{2}$ de la loi de Student à $n-1$ degrés de liberté. (le calcul du risque de seconde espèce est par contre hors-programme).
- Pour $H_1 : m = m_1$ avec $m_1 < m_0$, ou $H_1 : m < m_0$:
  On décide H_0 si :
  $$\sqrt{n} \times \frac{\overline{x}_n - m_0}{\sigma} \geq t,$$

où t est fonction du risque de première espèce α :

$$\alpha = P\left(\sqrt{n} \times \frac{\overline{X}_n - m_0}{\sigma} \leq t | H_0\right).$$

t est donc le quantile d'ordre α de la loi de Student à $n-1$ degrés de liberté.

– Pour $H_1 : m = m_1$ avec $m_1 > m_0$, ou $H_1 : m > m_0$:
On décide H_0 si :

$$\sqrt{n} \times \frac{\overline{x}_n - m_0}{\sigma} \leq t,$$

où t est fonction du risque de première espèce α :

$$\alpha = P\left(\sqrt{n} \times \frac{\overline{X}_n - m_0}{\sigma} \geq t | H_0\right).$$

t est donc le quantile d'ordre $1-\alpha$ de la loi de Student à $n-1$ degrés de liberté.

(d) Test de la valeur de la variance, moyenne connue

ATTENTION : Ce type de test n'est possible que pour des variables suivant des lois normales même si n est supérieur à 30.

On observe la réalisation d'un échantillon iid $(X_1, \ldots, X_n)$ de la loi normale de moyenne connue m et de variance inconnue σ^2.

Dans ce type de test, l'hypothèse nulle est :

$H_0 : \sigma = \sigma_0$.

La statistique associée à ce test est :

$$R_n = \frac{n\Sigma_n^2}{\sigma_0^2},$$

où :

$$\Sigma_n^2 = \frac{1}{n}\sum_{j=1}^{n} (X_j - m_0)^2.$$

Σ_n^2 est un estimateur sans biais de la variance.

Sous H_0, R_n suit une loi du Chi-deux à n degrés de liberté.

De même, la règle de décision dépend de l'hypothèse alternative :

– Pour $H_1 : \sigma > \sigma_0$: on décide H_0 si :

$$\frac{n\sigma_n^2}{\sigma_0^2} \leq c,$$

où c est fonction du risque de première espèce :

$$\alpha = P\left(\frac{n\Sigma_n^2}{\sigma_0^2} > c | H_0\right).$$

Ainsi, c est le quantile d'ordre $1-\alpha$ de la loi du Chi-deux à n degrés de liberté.

Quant au risque de seconde espèce, il est calculé pour une valeur particulière $\sigma_1 > \sigma_0$:

$$\begin{aligned}\beta(\sigma_1) &= P\left(\frac{n\Sigma_n^2}{\sigma_0^2} \leq c|\sigma = \sigma_1\right) \\ &= P\left(\frac{n\Sigma_n^2}{\sigma_1^2} \leq \frac{\sigma_0^2 \times c}{\sigma_1^2}|\sigma = \sigma_1\right),\end{aligned}$$

en utilisant le fait que $\frac{n\Sigma_n^2}{\sigma_1^2}$ suit une loi du Chi-deux à n degrés de liberté sous $\sigma = \sigma_1$.

– Pour $H_1 : \sigma < \sigma_0$: on décide H_0 si :

$$\frac{n\sigma_n^2}{\sigma_0^2} \geq c,$$

où c est fonction du risque de première espèce :

$$\alpha = P\left(\frac{n\Sigma_n^2}{\sigma_0^2} \leq c|H_0\right).$$

Ainsi, c est le quantile d'ordre α de la loi du Chi-deux à n degrés de liberté.
Quant au risque de seconde espèce, il est calculé pour une valeur particulière $\sigma_1 < \sigma_0$:

$$\begin{aligned}\beta(\sigma_1) &= P\left(\frac{n\Sigma_n^2}{\sigma_0^2} \geq c|\sigma = \sigma_1\right) \\ &= P\left(\frac{n\Sigma_n^2}{\sigma_1^2} \geq \frac{\sigma_0^2 \times c}{\sigma_1^2}|\sigma = \sigma_1\right),\end{aligned}$$

en utilisant le fait que $\frac{n\Sigma_n^2}{\sigma_1^2}$ suit une loi du Chi-deux à n degrés de liberté sous $\sigma = \sigma_1$.

(e) Test de la valeur de la variance, moyenne inconnue

ATTENTION : Ce type de test n'est possible que pour des variables suivant des lois normales même si n est supérieur à 30.

Pour le même type de test, dans le cas où la moyenne m est inconnue, la statistique utilisée est :

$$R_{n-1} = \frac{(n-1) \times \hat{S}_n^2}{\sigma_0^2},$$

où :

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}_n\right)^2$$

est l'estimateur sans biais de la variance.

Sous H_0, R_{n-1} suit une loi du Chi-deux à $n-1$ degrés de liberté.

De même, la règle de décision dépend de l'hypothèse alternative :

– Pour $H_1 : \sigma > \sigma_0$: on décide H_0 si :

$$\frac{(n-1)\hat{s}_n^2}{\sigma_0^2} \leq c,$$

où c est fonction du risque de première espèce :

$$\alpha = P\left(\frac{(n-1) \times \hat{S}_n^2}{\sigma_0^2} > c | H_0\right).$$

Ainsi, c est le quantile d'ordre $1-\alpha$ de la loi du Chi-deux à $n-1$ degrés de liberté.
Quant au risque de seconde espèce, il est calculé pour une valeur particulière $\sigma_1 > \sigma_0$:

$$\begin{aligned}\beta(\sigma_1) &= P\left(\frac{(n-1)\hat{S}_n^2}{\sigma_0^2} \leq c | \sigma = \sigma_1\right) \\ &= P\left(\frac{(n-1)\hat{S}_n^2}{\sigma_1^2} \leq \frac{\sigma_0^2 \times c}{\sigma_1^2} | \sigma = \sigma_1\right),\end{aligned}$$

en utilisant le fait que $\dfrac{(n-1)\hat{S}_n^2}{\sigma_1^2}$ suit une loi du Chi-deux à $n-1$ degrés de liberté sous $\sigma = \sigma_1$.

– Pour $H_1 : \sigma < \sigma_0$: on décide H_0 si :

$$\frac{(n-1)\hat{s}_n^2}{\sigma_0^2} \geq c,$$

où c est fonction du risque de première espèce :

$$\alpha = P\left(\frac{(n-1)\hat{S}_n^2}{\sigma_0^2} \leq c | H_0\right).$$

Ainsi, c est le quantile d'ordre α de la loi du Chi-deux à $n-1$ degrés de liberté.
Quant au risque de seconde espèce, il est calculé pour une valeur particulière $\sigma_1 < \sigma_0$:

$$\begin{aligned}\beta(\sigma_1) &= P\left(\frac{(n-1)\hat{S}_n^2}{\sigma_0^2} \geq c | \sigma = \sigma_1\right) \\ &= P\left(\frac{(n-1)\hat{S}_n^2}{\sigma_1^2} \geq \frac{\sigma_0^2 \times c}{\sigma_1^2} | \sigma = \sigma_1\right),\end{aligned}$$

en utilisant le fait que $\dfrac{(n-1)\hat{S}_n^2}{\sigma_1^2}$ suit une loi du Chi-deux à $n-1$ degrés de liberté sous $\sigma = \sigma_1$.

(f) Test de comparaison de moyennes, écart-types connus
On observe deux réalisations de deux échantillons $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$. L'échantillon $(X_1, \ldots, X_n)$ est issu d'une loi normale de moyenne inconnue m_X et d'écart-type connu σ_X et l'échantillon $(Y_1, \ldots, Y_n)$ est issu d'une loi normale de moyenne inconnue m_Y et d'écart-type connu σ_Y.
L'hypothèse nulle de ce test est :
$H_0 : m_X = m_Y$.
Pour un tel test, la statistique (dépendant uniquement de l'hypothèse nulle) est :

$$Z = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}.$$

Sous l'hypothèse nulle H_0, Z suit une loi normale centrée-réduite (**En exercice : vérifiez-le**). De même, la stratégie de décision dépend de l'hypothèse alternative :

- Pour $H_1 : m_X \neq m_Y$: on décide H_0 si :

$$\frac{|\overline{x}_n - \overline{y}_m|}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \leq z,$$

où z est donné à partir du risque de première espèce :

$$\alpha = P\left(\frac{\left|\overline{X}_n - \overline{Y}_m\right|}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \geq z | H_0\right). \tag{57}$$

z est donc le quantile d'ordre $1 - \dfrac{\alpha}{2}$ de la loi normale centrée-réduite.
Le risque de seconde espèce dépend de la valeur $\delta = |m_X - m_Y|$ et est donné par :

$$\beta(\delta) = \phi\left(z + \frac{\delta}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}\right) + \phi\left(z - \frac{\delta}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}\right) - 1,$$

où ϕ est la fonction de répartition de la loi normale centrée-réduite.

- Pour $H_1 : m_X > m_Y$: on décide H_0 si :

$$\frac{\overline{x}_n - \overline{y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \leq z,$$

où z est donné à partir du risque de première espèce :

$$\alpha = P\left(\frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \geq z | H_0\right). \tag{58}$$

z est donc le quantile d'ordre $1-\alpha$ de la loi normale centrée-réduite.
De même, le risque de seconde espèce dépend de la valeur $\delta = m_X - m_Y > 0$ et vaut :

$$\begin{aligned}\beta(\delta) &= P\left(\frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \leq z | m_X - m_Y = \delta\right) \\ &= 1 - \phi\left(\frac{\delta}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} - z\right).\end{aligned}$$

– Pour $H_1 : m_X < m_Y$: le raisonnement est le même que précédemment en inversant les rôles de X et de Y (**Attention à bien inverser les rôles**).

(g) Test de comparaison de moyennes, écart-type commun et inconnu : test de Student

Cette fois-ci, on considère que $\sigma_X = \sigma_Y = \sigma$ mais σ est inconnu.

Remarque : Il faut faire TRES ATTENTION que ce type de test n'est vrai que si la variable en question est gaussienne (même si n est supérieur à 30). Lorsque n est supérieur à 30 pour des variables non normales, on utilisera le test précédent en remplaçant σ_X et σ_Y par leur estimation (TCL).

La statistique de ce test, associée à l'hypothèse nulle $H_0 : m_X = m_Y$ est :

$$T_{n+m-2} = \frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \times \frac{\overline{X}_n - \overline{Y}_m}{\hat{S}},$$

où $\hat{S}^2$ est l'intra-variance empirique donnée par :

$$\hat{S}^2 = \frac{(n-1)\hat{S}_X^2 + (m-1)\hat{S}_Y^2}{n+m-2},$$

où $\hat{S}_X^2$ et $\hat{S}_Y^2$ sont respectivement les estimateurs sans biais de la variance à partir des échantillons $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$.

En exercice : Montrer que $\hat{S}^2$ est également un estimateur sans biais de la variance.

On montre que sous H_0, T_{n+m-2} suit une loi de Student à $n+m-2$ degrés de liberté. La règle de décision est également donnée à partir de l'hypothèse alternative :

– Pour $H_1 : m_X \neq m_Y$: on décide H_0 si :

$$\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \times \frac{|\overline{x}_n - \overline{y}_m|}{\hat{s}} \leq t,$$

et t est donné en fonction du risque de première espèce :

$$\alpha = P\left(\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \times \frac{\left|\overline{X}_n - \overline{Y}_m\right|}{\hat{S}} \geq t | H_0\right).$$

On en déduit que t est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $n+m-2$ degrés de liberté.

– Pour $H_1 : m_X > m_Y$: on décide H_0 si :

$$\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \times \frac{\overline{x}_n - \overline{y}_m}{\hat{s}} \leq t,$$

et t est donné en fonction du risque de première espèce :

$$\alpha = P\left(\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \times \frac{\overline{X}_n - \overline{Y}_m}{\hat{S}} \geq t | H_0 \right).$$

On en déduit que t est le quantile d'ordre $1-\alpha$ de la loi de Student à $n+m-2$ degrés de liberté.

(h) Test de comparaison de variances, moyennes connues : test de Fisher

ATTENTION : Test possible uniquement si les variables suivent des lois normales.

On observe des réalisations de deux échantillons $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ issus respectivement de lois normales de moyenne (connue) m_X et variance (inconnue) σ_X^2 et de moyenne (connue) m_Y et variance (inconnue) σ_Y^2.

Pour ce type de test, l'hypothèse nulle est $\sigma_X^2 = \sigma_Y^2$ et la statistique utilisée est :

$$F_{n,m} = \frac{\Sigma_X^2}{\Sigma_Y^2},$$

où :

$$\begin{aligned} \Sigma_X^2 &= \frac{1}{n}\sum_{j=1}^{n}(X_j - m_X)^2. \\ \Sigma_Y^2 &= \frac{1}{m}\sum_{j=1}^{m}(Y_j - m_Y)^2. \end{aligned}$$

Sous H_0, $F_{n,m}$ suit une loi de Fisher à n et m degrés de liberté.

Pour $H_1 : \sigma_X^2 \neq \sigma_Y^2$, la règle de décision est : on décide H_0 si :

$$a \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq b,$$

où a et b sont déterminés à partir du risque de première espèce :

$$\alpha = 1 - P\left(a \leq \frac{\Sigma_X^2}{\Sigma_Y^2} \leq b | H_0 \right).$$

a est donc le quantile d'ordre $\frac{\alpha}{2}$ et b le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Fisher à n et m degrés de liberté.

(i) Test de comparaison de variances, moyennes inconnues : test de Fisher
ATTENTION : Test possible uniquement si les variables suivent des lois normales.
On observe des réalisations de deux échantillons $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ issus respectivement de lois normales de moyenne (inconnue) m_X et variance (inconnue) σ_X^2 et de moyenne (inconnue) m_Y et variance (inconnue) σ_Y^2.
Pour ce type de test, l'hypothèse nulle est $\sigma_X^2 = \sigma_Y^2$ et la statistique utilisée est :

$$F_{n-1,m-1} = \frac{\hat{S}_X^2}{\hat{S}_Y^2},$$

où :

$$\begin{aligned} \hat{S}_X^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2. \\ \hat{S}_Y^2 &= \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2. \end{aligned}$$

Sous H_0, $F_{n-1,m-1}$ suit une loi de Fisher à $n-1$ et $m-1$ degrés de liberté.
Pour $H_1 : \sigma_X^2 \neq \sigma_Y^2$, la règle de décision est : on décide H_0 si :

$$a \leq \frac{\hat{s}_X^2}{\hat{s}_Y^2} \leq b,$$

où a et b sont déterminés à partir du risque de première espèce :

$$\alpha = 1 - P\left(a \leq \frac{\hat{S}_X^2}{\hat{S}_Y^2} \leq b | H_0\right).$$

a est donc le quantile d'ordre $\frac{\alpha}{2}$ et b le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Fisher à $n-1$ et $m-1$ degrés de liberté.

5.1.6 Test sur la valeur d'un paramètre quelquonque

Dans de paragraphe, les données ne sont plus nécessairement issues d'une loi normale et ou le paramètre n'est plus nécessairement la moyenne ou la variance.
On observe un échantillon, réalisation d'un échantillon iid $(X_1, \ldots, X_n)$ d'une loi de paramètre θ.
Pour l'hypothèse nulle $H_0 : \theta = \theta_0$, la statistique est :

$$\sqrt{I_{X_1,\ldots,X_n}(T_n)} \times (T_n - \theta_0),$$

où T_n est l'estimateur du maximum de vraisemblance de θ et $\theta \to I_{X_1,\ldots,X_n}(\theta)$ est la fonction d'information de Fisher pour l'échantillon.
Sous H_0 et pour $n \geq 30$, $\sqrt{I_{X_1,\ldots,X_n}(T_n)} \times (T_n - \theta_0)$ suit une loi normale centrée-réduite.
De la même façon, la règle de décision dépend de l'hypothèse alternative H_1 :

– Pour $H_1 : \theta \neq \theta_0$, on décide H_0 si :

$$\sqrt{I_{X_1,\ldots,X_n}(t_n)} \times |t_n - \theta_0| \leq z,$$

où t_n est l'estimation observée (réalisation de T_n). z est déterminé à partir du risque de première espèce :

$$\alpha = P\left(\sqrt{I_{X_1,\ldots,X_n}(T_n)} \times |T_n - \theta_0| \geq z | H_0\right).$$

Ainsi z est le quantile d'ordre $1 - \dfrac{\alpha}{2}$ de la loi normale centrée-réduite.

– Pour $H_1 : \theta > \theta_0$, on décide H_0 si :

$$\sqrt{I_{X_1,\ldots,X_n}(t_n)} \times (t_n - \theta_0) \leq z,$$

z est déterminé à partir du risque de première espèce :

$$\alpha = P\left(\sqrt{I_{X_1,\ldots,X_n}(T_n)} \times (T_n - \theta_0) \geq z | H_0\right).$$

Ainsi z est le quantile d'ordre $1 - \alpha$ de la loi normale centrée-réduie.

5.2 Exercices sur test de la valeur d'un paramètre

5.2.1 Exercice 1

1. (a) Risque de première espèce : probabilité de décider H_1 alors que c'est H_0 qui est vraie. Dans le contexte de l'exercice, probabilité de ne pas indemniser les riverains alors que l'on aurait du les indemniser.
   Risque de seconde espèce : probabilité de décider H_0 alors que c'est H_1 qui est vraie. Dans le contexte de l'exercice, probabilité d'indemniser les riverains alors que l'on n'aurait pas du.

   (b) Le test est du type :

   – $H_0 : m = m_0$.
   – $H_1 : m = m_1$.
   avec $m_0 = 80$ et $m_1 = 78$, sous les deux hypothèses, les lois sont **normales de même variance**(**TRES IMPORTANT :** la règle de décision ne serait pas du type ci-dessous si les lois n'étaient pas normales ou avaient des variances différentes), donc la stratégie de décision est : On décide H_0 si :

   $$\sqrt{n} \times \frac{\overline{x}_n - 80}{7} > z,$$

   où z est un nombre réel que l'on va déterminer.
   Le risque de première espèce est $\alpha = 0.05$, on cherche donc z tel que :

   $$0.05 = P\left(\sqrt{n} \times \frac{\overline{X}_n - 80}{7} \leq z | H_0\right).$$

Sous H_0, $Z = \sqrt{n} \times \dfrac{\overline{X}_n - 80}{7}$ suit une loi normale centrée-réduite, ainsi z est le quantile d'ordre 0.05 de la loi normale centrée-réduite, d'où $z_{0.05} = -z_{0.95} = -1.65$. On décide alors H_0 si :

$$\sqrt{n} \times \frac{\overline{x}_n - 80}{7} > -1.65.$$

Pour $n = 100$, $\overline{x}_n = 79.1$, on a $\sqrt{n} \times \dfrac{\overline{x}_n - 80}{7} = -1.29$, ainsi on décide H_0, on rejette donc H_1. Pour $\overline{x}_n = 79.1$, on décidera d'indemniser les riverains. On décide H_0 entraîne que l'on rejete H_1. On n'accepte pas pour autant H_0 car n'est pas contraire de H_1.

(c) Le risque de deuxième espèce est donné par :

$$\begin{aligned} \beta &= P\left(10 \times \frac{\overline{X}_n - 80}{7} > -1.65 | H_1\right) \\ &= P(\overline{X}_n > 78.845). \end{aligned}$$

Sous H_1, $\overline{X}_n$ suit une loi normale de moyenne $m_1 = 78$ et d'écart-type $\dfrac{\sigma}{\sqrt{n}} = \dfrac{7}{10} = 0.7$, ainsi :

$$\begin{aligned} \beta &= P(Z > \frac{78.845 - 78}{0.7}) \\ &= 1 - P(Z \leq \frac{78.845 - 78}{0.7}) \\ &= 1 - P(Z \leq 1.21), \end{aligned}$$

où Z suit une loi normale centrée-réduite. On en déduit que $\beta = 1 - 0.8869 = 0.1131$. $\beta > \alpha$, on a donc privilégié l'hypothèse H_0.

2. (a) Le risque de première espèce est alors la probabilité d'indemniser les riverains alors que l'on n'aurait pas du.
Le risque de seconde espèce est la probabilité de ne pas indemniser les riverains alors que l'on aurait du.

(b) Dans ce cas, $m_0 < m_1$ et donc, on décide H_0 si :

$$\sqrt{n} \times \frac{\overline{x}_n - 78}{7} < z,$$

où z est un nombre réel que l'on va déterminer.
Le risque de première espèce est $\alpha = 0.05$, on cherche donc z tel que :

$$0.05 = P\left(\sqrt{n} \times \frac{\overline{X}_n - 78}{7} \geq z | H_0\right).$$

Sous H_0, $Z = \sqrt{n} \times \dfrac{\overline{X}_n - 78}{7}$ suit une loi normale centrée-réduite, ainsi z est le quantile d'ordre 0.95 de la loi normale centrée-réduite,

d'où $z_{0.95} = 1.65$. On décide alors H_0 si :

$$\sqrt{n} \times \frac{\overline{x}_n - 78}{7} < 1.65.$$

Pour $n = 100$, $\overline{x}_n = 79.1$, on a $\sqrt{n} \times \frac{\overline{x}_n - 78}{7} = 1.57 < 1.65$, ainsi on décide H_0, on rejette donc H_1.

(c) Le risque de deuxième espèce est donné par :

$$\begin{aligned} \beta &= P\left(10 \times \frac{\overline{X}_n - 78}{7} \leq 1.65 | H_1\right) \\ &= P(\overline{X}_n \leq 79.155 | H_1). \end{aligned}$$

Sous H_1, $\overline{X}_n$ suit une loi normale de moyenne $m_1 = 80$ et d'écart-type $\frac{\sigma}{\sqrt{n}} = \frac{7}{10} = 0.7$, ainsi :

$$\begin{aligned} \beta &= P(Z \leq \frac{79.155 - 80}{0.7}) \\ &= 1 - P(Z \leq \frac{80 - 79.155}{0.7}) \\ &= 1 - P(Z \leq 1.21), \end{aligned}$$

où Z suit une loi normale centrée-réduite. On en déduit que $\beta = 1 - 0.8869 = 0.1131$. $\beta > \alpha$, on a donc privilégié l'hypothèse H_0.

3. **IMPORTANT :** On remarque que pour la même valeur observée, tantôt on décide d'indemniser les riverains et tantôt on décide de ne pas les indemniser. Cela est dû au fait que l'on a imposer le risque de première espèce qui dépend bien sûr de l'ordre des hypothèses. Le fait d'imposer le risque de première espèce fait que l'on va de toute façon privilégier une hypothèse (dans ce cas H_0) par rapport à l'autre.

4. Afin de ne privilégier aucune hypothèse, on choisit :

$$c = \frac{80 + 78}{2} = 79.$$

Dans le cas où le test est :
- $H_0 : m = 80$.
- $H_1 : m = 78$.

On décide H_0 si $\overline{x}_n > 79$. Dans ce cas, on décide H_0 et donc rejette H_1 car $\overline{x}_n = 79.1 > 79$.
Les risques de première espèce et de seconde espèce sont égaux et peuvent être calculés via :

$$\alpha = \beta = P\left(\overline{X}_n \leq 79 | H_0\right).$$

Sous H_0, $\overline{X}_n$ suit une loi normale de moyenne $m_0 = 80$ et d'écart-type $\frac{\sigma}{\sqrt{n}} = \frac{7}{10} = 0.7$, d'où :

$$\begin{aligned}\alpha = \beta &= P(Z \leq \frac{79-80}{0.7}) \\ &= 1 - P(Z \leq \frac{80-79}{0.7}) \\ &= 1 - P(Z \leq 1.43) \\ &= 1 - 0.9236 \\ &= 0.0764.\end{aligned}$$

Dans le cas où le test est :
– $H_0 : m = 78$.
– $H_1 : m = 80$.
On décide H_0 si $\overline{x}_n < 79$. Dans ce cas, on décide H_1 et donc rejette H_0 car $\overline{x}_n = 79.1 > 79$.
Les risques de première et seconde espèce sont exactement les mêmes que précédemment.

5.2.2 Exercice 2

1. Trivial : $m_0 = 23.65$.
2. Pour un tel type de test, la règle de décision est de la forme : on décide H_0 (cad $m = 23.65$ versus $m \neq 23.65$) si :

$$\sqrt{n} \times \frac{|\overline{x}_n - 23.65|}{0.02} \leq z.$$

Le risque de première espèce étant $\alpha = 0.1$, on doit déterminer z tel que :

$$0.1 = P\left(\sqrt{n} \times \frac{|\overline{X}_n - 23.65|}{0.02} > z | H_0\right).$$

Sous H_0, $\sqrt{n} \times \frac{|\overline{x}_n - 23.65|}{0.02}$ suit une loi normale centrée-réduite, ainsi z est le quantile d'ordre 0.95 de la loi centrée-réduite, soit $z_{0.95} = 1.65$. On décide donc H_0 si

$$\sqrt{n} \times \frac{|\overline{x}_n - 23.65|}{0.02} \leq 1.65.$$

Pour $n = 10$ et $\overline{x}_n = 23.661$, on a : $\sqrt{n} \times \frac{|\overline{x}_n - 23.65|}{0.02} = 1.74 > 1.65$, on décide H_1, on rejette donc H_0 ainsi on accepte son contraire qui est H_1.
Calculons le risque de seconde espèce pour $m_1 = 23.64$, c'est à dire :

$$\begin{aligned}\beta(23.64) &= P(\sqrt{10} \times \frac{|\overline{x}_n - 23.65|}{0.02} \leq 1.65 | m = 23.64). \\ &= P(23.639 \leq \overline{X}_n \leq 23.660 | m = 23.64).\end{aligned}$$

Sous $m = 23.64$, $\overline{X}_n$ suit loi normale de moyenne 23.64 et d'écart-type $\frac{0.02}{\sqrt{10}}$, ainsi :

$$\begin{aligned}
\beta(23.64) &= P(\sqrt{10} \times \frac{23.639 - 23.64}{0.02} \leq Z \leq \sqrt{10} \times \frac{23.660 - 23.64}{0.02}) \\
&= P(-0.16 \leq Z \leq 3.16) \\
&= \phi(3.16) - (1 - \phi(0.16)) \\
&= 0.99921 - (1 - 0.5636) \\
&= 0.56281.
\end{aligned}$$

Ainsi la puissance du test pour $m_1 = 23.64$ est $1 - 0.56281 = 0.43719$.
Calculons le risque de seconde espèce pour $m_1 = 23.66$, c'est à dire :

$$\begin{aligned}
\beta(23.66) &= P(\sqrt{10} \times \frac{|\overline{x}_n - 23.65|}{0.02} \leq 1.65 | m = 23.66). \\
&= P(23.639 \leq \overline{X}_n \leq 23.660 | m = 23.66).
\end{aligned}$$

Sous $m = 23.66$, $\overline{X}_n$ suit loi normale de moyenne 23.66 et d'écart-type $\frac{0.02}{\sqrt{10}}$, ainsi :

$$\begin{aligned}
\beta(23.66) &= P(\sqrt{10} \times \frac{23.639 - 23.66}{0.02} \leq Z \leq \sqrt{10} \times \frac{23.660 - 23.66}{0.02}) \\
&= P(-3.32 \leq Z \leq 0) \\
&= 0.5 - (1 - \phi(3.32)) \\
&= 0.5 - (1 - 0.99955) \\
&= 0.49955.
\end{aligned}$$

Ainsi la puissance du test pour $m_1 = 23.66$ est $1 - 0.49955 = 0.50045$.

3. Le risque de première espèce est donné par :

$$\begin{aligned}
\alpha &= P(\text{décider } H_1 | H_0) \\
&= P(\overline{X}_n \notin [23.6409, 23.6624] | m = 23.65) \\
&= 1 - P(\overline{X}_n \in [23.6409, 23.6624] | m = 23.65).
\end{aligned}$$

Sous H_0, $\overline{X}_n$ a pour moyenne 23.65 et écart-type $\frac{0.02}{\sqrt{10}}$, donc :

$$\begin{aligned}
\alpha &= 1 - P(\sqrt{10} \times \frac{23.6409 - 23.65}{0.02} \leq Z \leq \sqrt{10} \times \frac{23.6624 - 23.65}{0.02}) \\
&= 1 - P(-1.44 \leq Z \leq 1.96) \\
&= 1 - (\phi(1.96) - 1 + \phi(1.44)) \\
&= 1 - 0.975 + 1 - 0.9251 \\
&= 0.0999.
\end{aligned}$$

Calculons le risque de seconde espèce pour $m_1 = 23.64$, c'est à dire :

$$\begin{aligned}\beta(23.64) &= P(23.6409 \leq \overline{X}_n \leq 23.6624 | m = 23.64).\\ &= P(0.14 \leq Z \leq 3.54)\\ &= \phi(3.54) - \phi(0.14)\\ &= 0.9998 - 0.5557 = 0.4441.\end{aligned}$$

Ainsi, la puissance du test pour $m_1 = 23.64$ est $1 - 0.4441 = 0.5559$.
Calculons le risque de seconde espèce pour $m_1 = 23.66$, c'est à dire :

$$\begin{aligned}\beta(23.64) &= P(23.6409 \leq \overline{X}_n \leq 23.6624 | m = 23.66).\\ &= P(-3.02 \leq Z \leq 0.38)\\ &= \phi(0.38) - 1 + \phi(3.02)\\ &= 0.648 - 1 + 0.99874 = 0.64674.\end{aligned}$$

Ainsi, la puissance du test pour $m_1 = 23.66$ est $1 - 0.64674 = 0.35326$.
Les puissances étant différentes pour différentes valeurs de m_1, ainsi le test n'est pas UPP.

4. **ATTENTION !!!!!!!!!!! :** on ne teste pas cette fois-ci la valeur de la moyenne mais celle de l'écart-type. L'écart-type étant la racine carrée de la variance, ainsi tester la valeur de l'écart-type est équivalent à tester la valeur de la variance.
Ainsi le test :
– $H_0 : \sigma = 0.02$
– $H_1 : \sigma > 0.02$
est équivalent au test :
– $H_0 : \sigma^2 = 0.0004$
– $H_1 : \sigma^2 > 0.0004$.
ATTENTION : Pour les tests concernant les variances, on ne fait plus une différence mais un rapport dans la règle de décision.
La stratégie de décision de ce test est : on décide l'hypothèse H_0 si :

$$\frac{\hat{s}_n^2}{0.0004} \leq c,$$

où $\hat{s}_n^2$ est l'estimation sans biais de la variance (c'est le carré de l'écart-type corrigé, **attention :** l'écart-type corrigé est la racine carrée de la variance sans biais, cependant ce n'est pas l'écart-type sans biais). La constante c est calculée à partir du risque de première espèce :

$$\begin{aligned}0.05 &= P(\text{décider } H_1 | H_0)\\ &= P\left(\frac{\hat{S}_n^2}{0.0004} > c | H_0\right).\end{aligned}$$

Sous H_0, $R_{n-1} = \dfrac{n-1}{0.0004}\hat{S}_n^2$ suit une loi du Chi-deux à $n-1$ degrés de liberté. Comme $n = 10$, on cherche c tel que :

$$0.05 = P(R_9 > 9c).$$

$9c$ est donc le quantile d'ordre 0.95 de la loi du Chi-deux à 9 degrés de liberté. Soit $9c = 16.919$ et donc $c = 1.88$. Ainsi, on décide H_0 si :

$$\frac{\hat{s}_n^2}{0.0004} \leq 1.88.$$

Pour $\hat{s}_n = 0.033$, on a : $\hat{s}_n^2 = 0.001089$ donc $\frac{\hat{s}_n^2}{0.0004} = 2.7225 \geq 1.88$, ainsi on rejette H_0. On n'accepte pas pour autant H_1 car H_0 et H_1 ne sont pas contraires.

5.2.3 Exercice 3

Pour ce type de test, la règle de décision est de la forme : on décide H_0 si :

$$\sqrt{n} \times \frac{|\overline{x}_n - 12|}{3.4} \leq z.$$

Le risque de première espèce étant égal à 0.05, on cherche z tel que :

$$0.05 = P\left(\sqrt{n} \times \frac{|\overline{X}_n - 12|}{3.4} > z | H_0\right).$$

Sous H_0, $\sqrt{n} \times \dfrac{|\overline{X}_n - 12|}{3.4}$ suit une loi normale centrée-réduite, ainsi z est le quantile d'ordre 0.975 de la loi normale centrée-réduite, d'où $z = 1.96$. On décide donc H_0 si :

$$\sqrt{n} \times \frac{|\overline{x}_n - 12|}{3.4} \leq 1.96.$$

Pour $n = 25$ et $\overline{x}_n = 10.8$, on a $\sqrt{n} \times \dfrac{|\overline{x}_n - 12|}{3.4} \simeq 1.76$ donc on rejette H_1 ainsi on accepte son contraire qui est H_0.

5.2.4 Exercice 4

Bien lire l'énoncé : Peut-on affirmer que le résultat moyen est inférieur à 20 signifie : peut-on rejetter son contraire. Les notes ne pouvant dépasser 20, son contraire est $m = 20$. Le test est donc :

- $H_0 : m = 20$.
- $H_1 : m < 20$.

Nous ne connaissons pas l'écart-type et on ne connaît pas la forme de la loi. Du fait que l'on ne connaisse pas l'écart-type, la règle de décision est de la forme : on décide H_0 si :

$$\sqrt{n} \times \frac{\overline{x}_n - 20}{\hat{s}_n} \geq c,$$

où $\hat{s}_n$ est l'écart-type corrigé, racine carrée de la variance sans biais (**attention : ce n'est pas l'écart-type sans biais**). Le risque étant $\alpha = 0.01$, on cherche donc c tel que :

$$0.01 = P\left(\sqrt{n} \times \frac{\overline{X}_n - 20}{\hat{S}_n} \leq c | H_0\right).$$

Attention : Les notes ne suivant pas nécessairement une loi normale, on ne peut pas affirmer que $\sqrt{n} \times \dfrac{\overline{X}_n - 20}{\hat{S}_n}$ suit une loi de Student sous H_0.

Cependant, comme $n \geq 30$, on peut considérer que sous H_0, $Z = \sqrt{n} \times \dfrac{\overline{X}_n - 20}{\hat{S}_n}$ suit une loi normale centrée-réduite. c est donc le quantile d'ordre 0.01 de la loi normale centrée-réduite, ainsi $c = -2.33$. On décide donc H_0 si :

$$\sqrt{n} \times \frac{\overline{x}_n - 20}{\hat{s}_n} \geq -2.33.$$

Pour $n = 160$, $\overline{x}_n = 19.3$ et $\hat{s}_n = 5.8$, on a : $\sqrt{n} \times \frac{\overline{x}_n - 20}{\hat{s}_n} = -1.52 \geq -2.33$ donc on décide H_0, on rejette donc H_1 et on accepte donc son contraire qui est H_0 car les notes ne peuvent pas être plus grandes que 20. On ne peut alors pas affirmer, avec un risque 0.01, que la moyenne est inférieure à 20.

5.2.5 Exercice 5

Soit X la variable aléatoire suivant la loi de Bernouilli $\mathcal{B}(1, p)$ valant 1 si le patient est guéri et 0 sinon. p (paramètre inconnu) correspond à la proportion de patients guéris. Le test considéré dans cet exercice est :

- $H_0 : p = 0.75$.
- $H_1 : p > 0.75$.

Soit $(x_1, \ldots, x_n)$ un échantillon observé, réalisation d'un échantillon iid $(X_1, \ldots, X_n)$ de la variable aléatoire X. Pour $n \geq 30$, la règle de décision est de la forme : on décide H_0 si :

$$\sqrt{I_{X_1,\ldots,X_n}(\hat{p}_{MV,n}(x_1, \ldots, x_n))} \times (\hat{p}_{MV,n}(x_1, \ldots, x_n) - 0.75) \leq c,$$

où $I_{X_1,\ldots,X_n}$ est l'information de Fisher et $\hat{p}_{MV,n}(x_1, \ldots, x_n)$ est l'estimateur du maximum de vraisemblance de p pour l'échantillon $(x_1, \ldots, x_n)$. En reprenant le résultat de l'exercice 2 sur les intervalles de confiance, on montre que $\hat{p}_{MV,n}(x_1, \ldots, x_n) = \overline{x}_n$ et que $I_{X_1,\ldots,X_n}(p) = \frac{n}{p(1-p)}$ et donc $I_{X_1,\ldots,X_n}(\hat{p}_{MV,n}(x_1, \ldots, x_n)) = \frac{n}{\overline{x}_n(1-\overline{x}_n)}$. On en déduit que la règle de décision est de la forme : on décide H_0 si :

$$\sqrt{n} \times \frac{\overline{x}_n - 0.75}{\sqrt{\overline{x}_n(1 - \overline{x}_n)}} \leq c.$$

Le risque étant $\alpha = 0.05$, on cherche c tel que :

$$0.05 = P\left(\sqrt{n} \times \frac{\overline{X}_n - 0.75}{\sqrt{\overline{X}_n(1 - \overline{X}_n)}} \geq c | H_0\right).$$

Comme $n \geq 30$, sous H_0, $Z = \sqrt{n} \times \dfrac{\overline{X}_n - 0.75}{\sqrt{\overline{X}_n(1 - \overline{X}_n)}}$ suit la loi normale centrée-réduite, c est donc le quantile d'ordre 0.95 de la loi normale centrée-réduite,

ainsi $c = 1.65$. On décide donc H_0 si :

$$\sqrt{n} \times \frac{\overline{x}_n - 0.75}{\sqrt{\overline{x}_n(1-\overline{x}_n)}} \leq 1.65.$$

Pour $n = 300$, $\overline{x}_n = \frac{243}{300} = 0.81$, on a : $\sqrt{n} \times \frac{\overline{x}_n - 0.75}{\sqrt{\overline{x}_n(1-\overline{x}_n)}} = 2.65 > 1.65$ donc on décide H_1 et par conséquent on rejette H_0. On n'accepte pas pour autant H_1 car n'est pas le contraire de H_0.

5.2.6 Exercice 6

Bien lire l'énoncé : Peut-on accepter que la proportion soit inférieure à 0.5 signifie peut-on rejetter son contraire qui est proportion supérieure ou égale à 0.5. Rejetter le fait que la proportion soit supérieure ou égale à 0.5 signifie "rejetter qu'elle soit égale à 0.5" et "rejetter qu'elle soit supérieure (strictement) à 0.5". Commençons par le test :

- $H_0 : p = 0.5$.
- $H_1 : p > 0.5$,

afin de rejetter le fait que la proportion soit supérieure à 0.5. De manière similaire à l'exercice 5, la stratégie de décision est de la forme :

$$\sqrt{n} \times \frac{\overline{x}_n - 0.5}{\sqrt{\overline{x}_n(1-\overline{x}_n)}} \leq c.$$

c est déterminé à partir du risque de première espèce $\alpha = 0.05$ soit :

$$0.05 = P\left(\sqrt{n} \times \frac{\overline{X}_n - 0.5}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} > c | H_0\right).$$

Sous H_0 et comme $n = 225 \geq 30$, $\sqrt{n} \times \frac{\overline{X}_n - 0.5}{\sqrt{\overline{X}_n(1-\overline{X}_n)}}$ suit une loi normale centrée-réduite. Ainsi c est le quantile d'ordre 0.95 , d'où $c = 1.65$. On décide donc H_0 si :

$$\sqrt{n} \times \frac{\overline{x}_n - 0.5}{\sqrt{\overline{x}_n(1-\overline{x}_n)}} \leq 1.65.$$

Pour $n = 225$, $\overline{x}_n = \frac{75}{225} = 0.33333$, on a : $\sqrt{n} \times \frac{\overline{x}_n - 0.5}{\sqrt{\overline{x}_n(1-\overline{x}_n)}} = -5.3$. Ainsi, on décide H_0 et donc on rejette H_1 (on n'accepte pas pour autant H_0).
Effectuons maintenant le test :

- $H_0 : p = 0.5$.
- $H_1 : p < 0.5$,

afin de rejetter le fait que la proportion soit égale à 0.5.
La stratégie de décision est de la forme : on décide H_0 si :

$$\sqrt{n} \times \frac{\overline{x}_n - 0.5}{\sqrt{\overline{x}_n(1-\overline{x}_n)}} > c.$$

On cherche c tel que :

$$0.05 = P\left(\sqrt{n} \times \frac{\overline{X}_n - 0.5}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \leq c | H_0\right).$$

On en déduit que $c = -1.65$ quantile d'ordre 0.05 de la loi normale centrée-réduite. On décide donc H_0 si :

$$\sqrt{n} \times \frac{\overline{x}_n - 0.5}{\sqrt{\overline{x}_n(1-\overline{x}_n)}} > -1.65.$$

Comme $-5.3 < -1.65$, on décide H_1 et donc on rejette H_0.
Avec un risque de 0.05, nous avons rejetter le fait que la proportion soit supérieure ou égale à 0.5, ainsi nous acceptons le fait qu'elle soit inférieure (strictement) à 0.5.

5.2.7 Exercice 7

Exercice difficile si on se débrouille mal en lecture d'énoncé.
Premier point : Combien de télespectateurs ont changé d'opinion. Réponse : 500 et c'est la taille n de l'échantillon considéré.
Notons p la proportion de téléspectateurs ayant changé d'opinion en faveur du candidat A. Etre plus favorable à A signifie que $p > 0.5$. Ainsi accepter l'hypothèse $p > 0.5$ signifie rejetter l'hypothèse $p \leq 0.5$. Commençons par le test :

- $p = 0.5$.
- $p < 0.5$,

afin de rejetter $p < 0.5$. En utilisant la même démarche que dans l'exercice 6, on décide H_0 si :

$$\sqrt{n} \times \frac{\overline{x}_n - 0.5}{\sqrt{\overline{x}_n(1-\overline{x}_n)}} > -1.65.$$

Pour $n = 500$, $\overline{x}_n = \frac{350}{500} = 0.7$, on a $\sqrt{n} \times \frac{\overline{x}_n - 0.5}{\sqrt{\overline{x}_n(1-\overline{x}_n)}} = 9.76$. Ainsi, on rejette H_1.
On effectue ensuite le test :

- $p = 0.5$.
- $p > 0.5$,

afin de rejetter l'hypothèse $p = 0.5$. De même, la règle de décision est : on décide H_0 si :

$$\sqrt{n} \times \frac{\overline{x}_n - 0.5}{\sqrt{\overline{x}_n(1-\overline{x}_n)}} < 1.65.$$

Comme $9.76 > 1.65$, ainsi on rejette H_0. En conclusion, on peut affirmer avec un risque de 0.05 que la prestation a été plus favorable à A qu'à B.

5.3 Exercices sur comparaison de moyenne et variance

5.3.1 Exercice 1

Il y a une erreur dans l'énoncé, on a $\sum_{i=1}^{7}(x_i - \overline{x}_n)^2 = 1619.43$.

1. (a) Estimateur sans biais de m_X : $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

   Estimateur sans biais de m_Y : $\overline{Y}_m = \frac{1}{m}\sum_{i=1}^{m} Y_i$.

   Estimateur sans biais de σ_X^2 : $\hat{S}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$.

   Estimateur sans biais de σ_Y^2 : $\hat{S}_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y}_n)^2$.

   Les estimations correspondantes sont : $\overline{x}_n = 162.28$, $\hat{s}_X^2 = 269.9$, $\overline{y}_n = 175.33$ et $\hat{s}_Y^2 = 257.89$.

   (b) Le test :
   - $H_0 : \sigma_X^2 = \sigma_Y^2$.
   - $H_1 : \sigma_X^2 \neq \sigma_Y^2$

   est un test de comparaison de variance. La statistique utilisée est :

   $$\frac{\hat{S}_X^2}{\hat{S}_Y^2},$$

   qui suit une loi de Fisher $\mathcal{F}(6,5)$ sous H_0.
   D'après H_1, la règle de décision est de la forme : on décide H_0 si :

   $$a \leq \frac{\hat{s}_X^2}{\hat{s}_Y^2} \leq b,$$

   où a et b sont déterminés à partir du risque de première espèce $\alpha = 0.02$, c'est à dire :

   $$0.02 = 1 - P\left(a \leq \frac{\hat{S}_X^2}{\hat{S}_Y^2} \leq b | H_0\right).$$

   a et b sont respectivement les quantiles d'ordre 0.01 et 0.99 de $\mathcal{F}(6,5)$, soit $b = 10.672$, a est aussi l'inverse du quantile d'ordre 0.99 de la loi de Fisher $\mathcal{F}(5,6)$, d'où $a = \frac{1}{8.746} = 0.11$. Ainsi, on décidera H_0 si :

   $$0.11 \leq \frac{\hat{s}_X^2}{\hat{s}_Y^2} \leq 10.672.$$

   On a $\frac{\hat{s}_X^2}{\hat{s}_Y^2} = 1.04$, donc on décide H_0, ainsi on rejette H_1 donc on accepte son contraire qui est H_0.

(c) Dans ce test, nous ne connaissons pas les écart-types. D'après le résultat précédent, on peut considérer que les deux écart-types sont égaux. Ainsi pour $H_0 : m_X = m_Y$, la statistique utilisée est :

$$T_{n+m-2} = \frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \times \frac{\overline{X}_n - \overline{Y}_m}{\hat{S}},$$

où $\hat{S}^2$ est l'intra-variance empirique donnée par :

$$\hat{S}^2 = \frac{(n-1)\hat{S}_X^2 + (m-1)\hat{S}_Y^2}{n+m-2},$$

où $\hat{S}_X^2$ et $\hat{S}_Y^2$ sont respectivement les estimateurs sans biais de la variance à partir des échantillons $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$.
Sous H_0, $T_{n+m-2} = T_{11}$ suit une loi de Student à 11 degrés de liberté.
Pour $H_1 : m_X < m_Y$, la règle de décision est :

$$\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \times \frac{\overline{x}_n - \overline{y}_m}{\hat{s}} \geq t,$$

où t est déterminé à partir du risque de première espèce :

$$0.05 = P\left(T_{n+m-2} \leq t | H_0\right),$$

ainsi t est le quantile d'ordre 0.05 de la loi de Student à 11 degrés de liberté, il se lit dans la colonne $0.1 = 0.05 \times 2$. Ainsi $t = -1.796$.
Pour un risque 0.05, on décidera H_0 si :

$$\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \times \frac{\overline{x}_n - \overline{y}_m}{\hat{s}} \leq -1.796. \tag{59}$$

On a $\hat{s} = 264.44$ et $\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}} \times \frac{\overline{x}_n - \overline{y}_m}{\hat{s}} = -1.44$, donc on décidera H_1 et rejettera H_0.
Lorsque le risque est $\alpha = 0.1$, la valeur de t est le quantile d'ordre 0.1 de la loi de Student, il se lit dans la colonne 0.2 et vaut donc $t = -1.363$. Cette fois-ci on décidera H_0 et donc rejettera H_1.

2. (a) Sous H_0, $\overline{X}_n - \overline{Y}_m$ suit une loi normale de moyenne 0 et de variance $\left(\frac{1}{7}+\frac{1}{6}\right) \times 17^2 \simeq 89.45$.

   (b) Connaissant les écart-types, pour $H_0 : m_X = m_Y$, la statistique utilisée est :

$$Z = \frac{\overline{X}_n - \overline{Y}_m}{17 \times \sqrt{\frac{1}{n}+\frac{1}{m}}},$$

qui suit une loi normale centrée-réduite sous H_0.
Pour $H_1 : m_X < m_Y$, la règle de décision est : on décide H_0 si :

$$\frac{\overline{x}_n - \overline{y}_m}{17 \times \sqrt{\frac{1}{n}+\frac{1}{m}}} \geq z,$$

où z est déterminé par le risque de première espèce :

$$\alpha = P\left(\frac{\overline{X}_n - \overline{Y}_m}{17 \times \sqrt{\frac{1}{n} + \frac{1}{m}}} \leq z | H_0\right).$$

z est le quantile d'ordre α de la loi normale centrée-réduite.
Pour $\alpha = 0.05$, on trouve $z = -1.65$ et donc on décide H_0 si :

$$\frac{\overline{x}_n - \overline{y}_m}{17 \times \sqrt{\frac{1}{n} + \frac{1}{m}}} \geq -1.65.$$

On a : $\dfrac{\overline{x}_n - \overline{y}_m}{17 \times \sqrt{\frac{1}{n} + \frac{1}{m}}} = -1.38$, on décide donc H_0. Ainsi, on rejette H_1.
Pour $\alpha = 0.1$, $z = -2.33$, dans ce cas on rejette H_0.

5.3.2 Exercice 3

ATTENTION : on ne sait pas que la variable est normale, on ne peut pas utiliser un test de Student.
On utilise alors le TCL. Soit m_X le poids moyen des étudiants faisant du sport et m_Y celui des étudiants ne faisant pas de sport. Soient σ_X et σ_Y les écart-types correspondants. On notera $\overline{X}_n$, $\overline{Y}_m$, $\hat{S}_X$ et $\hat{S}_Y$ les estimateurs sans biais correspondants. Pour $H_0 : m_X = m_Y$, on utilisera la statistique :

$$Z = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\hat{S}_X^2}{n} + \frac{\hat{S}_Y^2}{m}}},$$

qui suit une loi normale centrée-réduite sous H_0 du fait que n et m soient plus grands que 30.
Bien lire l'énoncé : Accepter l'hypothèse que $m_X > m_Y$ signifie rejetter son contraire, c'est à dire $m_X < m_Y$ ou $m_X = m_Y$.
Commençons par le test d'hypothèse alternative $H_1 : m_X < m_Y$ afin de rejetter H_1.
Pour une telle hypothèse, on décide H_0 si :

$$\frac{\overline{x}_n - \overline{y}_m}{\sqrt{\frac{\hat{s}_X^2}{n} + \frac{\hat{s}_Y^2}{m}}} \geq z,$$

où z donné par :

$$0.05 = P\left(\frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\hat{S}_X^2}{n} + \frac{\hat{S}_Y^2}{m}}} \leq z | H_0\right).$$

z est donc le quantile d'ordre 0.05 de la loi normale centrée-réduite, soit $z = -1.65$. On décide H_0 si :

$$\frac{\overline{x}_n - \overline{y}_m}{\sqrt{\frac{\hat{s}_X^2}{n} + \frac{\hat{s}_Y^2}{m}}} \geq -1.65.$$

On a : $\dfrac{\overline{x}_n - \overline{y}_m}{\sqrt{\frac{\hat{s}_X^2}{n} + \frac{\hat{s}_Y^2}{m}}} = 1.32$, ainsi on décide H_0 et donc on rejette $m_X < m_Y$, on accepte alors $m_X \geq m_Y$.
Effectuons maintenant le test d'hypothèse $H_1 : m_X > m_Y$. Pour ce test, la règle de décision est :

$$\frac{\overline{x}_n - \overline{y}_m}{\sqrt{\frac{\hat{s}_X^2}{n} + \frac{\hat{s}_Y^2}{m}}} \leq z,$$

et on montre que z est le quantile d'ordre 0.95 de la loi normale centrée-réduite. Ainsi $z = 1.65$, on décide donc H_0 et on rejette $m_X > m_Y$. Ainsi, on accepte seulement que $m_X = m_Y$.

5.3.3 Exercice 4

1. Sans savoir que les variables suivent des lois normales, on ne peut pas faire de test de comparaison de variance.
2. Comme les tailles des deux échantillons est plus grande que 30, on peut faire un test de comparaison de moyenne. Pour $H_0 : m_1 = m_2$, la statistique utilisée est :

   $$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\hat{S}_1^2}{n} + \frac{\hat{S}_2^2}{m}}},$$

   qui suit une loi normale centrée-réduite sous H_0.
3. On cherche à accepter l'hypothèse que $m_1 < m_2$ et donc à rejetter l'hypothèse que $m_1 \geq m_2$. Commençons le test avec $H_1 : m_1 > m_2$ pour lequel la règle de décision est : on décide H_0 si :

   $$\frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n} + \frac{\hat{s}_2^2}{m}}} \leq z.$$

   z est déterminé à partir de :

   $$0.05 = P(Z \geq z | H_0),$$

   ainsi $z = 1.65$. On décide H_0 si :

   $$\frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n} + \frac{\hat{s}_2^2}{m}}} \leq 1.65.$$

Comme $\overline{x}_1 < \overline{x}_2$, on a $\dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n} + \frac{\hat{s}_2^2}{m}}} < 0 \leq 1.65$, donc on décide H_0 et on rejette $m_1 > m_2$. Ainsi on accepte $m_1 \leq m_2$.
Effectuons maintenant le test avec $H_1 : m_1 < m_2$. La règle de décision est alors : on décide H_0 si :

$$\frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n} + \frac{\hat{s}_2^2}{m}}} \geq z.$$

On trouve alors $z = -1.65$. Comme $\dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n} + \frac{\hat{s}_2^2}{m}}} = -1.58 \geq -1.64$ on décide donc H_0 et on rejette $m_1 < m_2$. Des deux tests, on accepte que $m_1 = m_2$.

5.3.4 Exercice 5

Pour ce type de test, la statistique utilisée est :

$$T_{n+m-2} = \frac{1}{\sqrt{\frac{1}{n} + \frac{1}{m}}} \times \frac{\overline{X}_1 - \overline{X}_2}{\hat{S}},$$

qui suit une loi de Student à $n+m-2 = 15$ degrés de liberté sous H_0. Accepter que les scores moyens sont moins bons en milieu rural qu'en milieu urbain c'est rejetter $m_1 \geq m_2$. Commençons par le test $H_1 : m_1 > m_2$, dans ce cas la règle de décision est :

$$\frac{1}{\sqrt{\frac{1}{n} + \frac{1}{m}}} \times \frac{\overline{x}_1 - \overline{x}_2}{\hat{s}} \leq t.$$

On montre de même que pour un risque $\alpha = 0.05$, t est le quantile d'ordre 0.95 de la loi de Student à 15 degrés de liberté. D'où $t = 1.753$. Comme $\overline{x}_1 < \overline{x}_2$, on décide H_0, on rejette donc $m_1 > m_2$, on accepte donc $m_1 \leq m_2$.
Pour le test $H_1 : m_1 < m_2$, la règle de décision est :

$$\frac{1}{\sqrt{\frac{1}{n} + \frac{1}{m}}} \times \frac{\overline{x}_1 - \overline{x}_2}{\hat{s}} \geq t.$$

On montre que $t = -1.753$, on a $\dfrac{1}{\sqrt{\frac{1}{n} + \frac{1}{m}}} \times \dfrac{\overline{x}_1 - \overline{x}_2}{\hat{s}} = -2.15$, on rejette H_0 et donc on accepte $m_1 < m_2$.