Study Title ____________________ Date ____________

Exercice 1. Soit $X_1, \ldots, X_n$ un n-échantillon d'espérance μ et de variance finie $\sigma^2 > 0$. On pose

$$S_n = \sum_{k=1}^{n} X_k \quad \text{et} \quad Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

1) Montrer que pour tous $a \leqslant b$,

$$\lim_{n \to +\infty} \mathbb{P}(a \leqslant Z_n \leqslant b) = \frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-\frac{s^2}{2}}\, \mathrm{d}s.$$

2) On suppose maintenant que $X \sim \mathcal{P}(1)$. Montrer que $\mathbb{P}(Z_n \leqslant 0) = \mathbb{P}(S_n \leqslant n)$, puis en déduire un équivalent lorsque $n \to +\infty$ de

$$P_n = \sum_{k=0}^{n} \frac{n^k}{k!}.$$

<u>Rappel</u>: indépandances 2 à 2 : (X_i, X_j) $X_i \perp\!\!\!\perp X_j$
indépandances mutuellement : $(X_i, X_j, \ldots) \perp\!\!\!\perp (X_k, X_p, \ldots)$ } indépandances mutuel plus forte

1) Soit $a \leq b$, on sait que d'après le TCL, on a que :

$Z_n \xrightarrow[n \to +\infty]{\mathcal{L}} Z$ où $Z \sim \mathcal{N}(0,1)$

Par définition, on a que : $\forall t \in \mathbb{R}$, $\mathbb{P}(Z_n \leq t) \xrightarrow[n \to +\infty]{} \mathbb{P}(Z \leq t)$

Or $\mathbb{P}(a \leq Z_n \leq b) = \mathbb{P}(Z_n \leq b) - \mathbb{P}(Z_n \leq a) + \mathbb{P}(Z_n = a)$.

$n \longrightarrow +\infty$

$= \mathbb{P}(Z \leq b) - \mathbb{P}(Z \leq a) + \underbrace{\mathbb{P}(Z = a)}_{0}$ car $Z \sim \mathcal{N}(0,1)$.

Pourquoi $\mathbb{P}(Z_n = a) \longrightarrow \mathbb{P}(Z = a)$? :

$X_n \xrightarrow[n \to +\infty]{\mathcal{L}} X$ $\forall A \in \mathcal{B}(\mathbb{R})$ tel que ∂A est de mesure nulle alors $\mathbb{P}(X_n \in A) \xrightarrow[n \to +\infty]{} \mathbb{P}(X \in A)$

Donc $\mathbb{P}(a \leq Z_n \leq b) \xrightarrow[n \to +\infty]{} \mathbb{P}(a \leq Z \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-s^2/2}\, ds$

2) $X \sim \mathcal{P}(1)$ $X(\Omega) = \mathbb{N}$ (càd que c'est une loi discrète).

$\forall k \in \mathbb{N}$ $\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$

En particulier $\mu = 1$ et $\sigma = 1$ on a : $Z_n = \frac{S_n - n}{\sqrt{n}}$ or $Z_n \leq 0$ d'où $\frac{S_n - n}{\sqrt{n}} \leq 0 \iff S_n \leq n$.

On a $\mathbb{P}(S_n \leq n) = \sum_{k=1}^{n} \mathbb{P}(S_n = k)$ (les $X_i \in \mathbb{N}$) or $S_n \sim \mathcal{P}(n)$, en effet, $\forall t \in \mathbb{R}$:

$\varphi(t) = \mathbb{E}\left[e^{itS_n}\right] = \mathbb{E}\left[\prod_{k=1}^{n} e^{itX_k}\right] \underset{\text{par indépandance}}{=} \prod_{k=1}^{n} \mathbb{E}\left[e^{itX_k}\right] = \varphi_X(t)^n$ → fonction caractéristique

Or $\varphi_{\mathcal{P}(\lambda)}(t) = e^{\lambda(e^{it} - 1)}$. Donc $\varphi_X(t)^n = \left[e^{1(e^{it} - 1)}\right]^n = e^{n(e^{it} - 1)} = \varphi_{\mathcal{P}(n)}(t)$.

Donc $\mathbb{P}(S_n \leq n) = \sum_{k=1}^{n} \mathbb{P}(S_n = k) = \sum_{k=1}^{n} \frac{n^k}{k!} e^{-n} = e^{-n} P_n$

Et $\mathbb{P}(S_n \leq n) \xrightarrow[n \to +\infty]{} \mathbb{P}(Z_n \leq 0) = \frac{1}{2}$ (avec $Z \sim \mathcal{N}(0,1)$ → symétrique et centré en 0)

Pour conclure $e^{-n} P_n \xrightarrow[n \to +\infty]{} \frac{1}{2} \iff P_n \underset{+\infty}{\sim} \frac{1}{2} e^n$

Exercice 2. On considère deux suites de v.a.r. telles que

$$U_n \xrightarrow{\mathcal{L}} U \sim \mathcal{N}(0,1) \quad \text{et} \quad V_n \xrightarrow{\mathbb{P}} 2 \ \ (\text{avec } V_n > 0).$$

1) Étudier la limite de $-2\,U_n$, $U_n + V_n$, $\mathbb{P}(U_n > 0)$, $U_n\,V_n^2$, $\frac{U_n^2}{\sqrt{V_n}}$ et $\mathbb{P}(V_n < 2)$.

2) A-t-on la certitude que $U_n \underset{\text{non}}{\xrightarrow{\mathbb{P}}} U$? Que $V_n \underset{\text{oui}}{\xrightarrow{\mathcal{L}}} 2$? car en loi $\overset{\not\Rightarrow}{\Leftarrow}$ en proba

1) $-2U_n \xrightarrow{L} -2U \sim \mathcal{N}(0,4)$ d'après Slutsky

En effet si on définit $\tilde{V}_n = -2$ p.s $\forall n \in \mathbb{N}$ on a $\tilde{V}_n \xrightarrow[n\to+\infty]{\mathbb{P}} -2$ donc $\tilde{V}_n U_n \xrightarrow[n\to+\infty]{L} -2U$

* $U_n + V_n \xrightarrow[n\to+\infty]{L} U_n + 2 \sim \mathcal{N}(2,1)$.
* $\mathbb{P}(U_n > 0) \xrightarrow[n\to+\infty]{L} \mathbb{P}(U > 0) = \frac{1}{2}$
* $U_n V_n^2 \xrightarrow[n\to+\infty]{L} 4U \sim \mathcal{N}(0,16)$
* $\frac{U_n^2}{\sqrt{V_n}} \xrightarrow[n\to+\infty]{L} \frac{U^2}{\sqrt{2}}$ → on retrouve Slutsky par composition

* $\mathbb{P}(V_n < 2) \xrightarrow[n\to+\infty]{} \mathbb{P}(2<2) = 0$ car $V_n \xrightarrow[n\to+\infty]{\mathbb{P}} 2 \Rightarrow V_n \xrightarrow[n\to+\infty]{L} 2$

Par définition : $V_n \xrightarrow[n\to+\infty]{\mathbb{P}} 2$ ssi $\forall \varepsilon > 0$ $\mathbb{P}(|V_n - 2| \geq \varepsilon) \xrightarrow[n\to+\infty]{} 0$

Exercice 3. Soient $\bar{x}$ la moyenne empirique et s^2 la variance empirique d'une série d'observations $x_1, \ldots, x_n$. Montrer que les transformations

$$x_i \mapsto x_i - \bar{x} \quad \text{et} \quad x_i \mapsto \frac{x_i}{\sqrt{s^2}}$$

permettent de centrer et de réduire les données, respectivement. Proposer d'autres transformations de données utiles.

Exercice 4. À l'aide de la table de la $\mathcal{N}(0,1)$ ou du logiciel R, déterminer les probabilités suivantes :

1) $\mathbb{P}(\mathcal{N}(0,1) > 0.1)$ $1 - \mathbb{P}(\mathcal{N}(0,1) \leq 1) = 1 - 0{,}5398$
2) $\mathbb{P}(\mathcal{N}(0,1) \leqslant 1.96) = 0{,}9750$
3) $\mathbb{P}(\mathcal{N}(1,4) < -0.5) = \mathbb{P}(\mathcal{N}(1,4) \leq -0{,}5) - \underbrace{\mathbb{P}(\mathcal{N}(1,4) = -0{,}5)}_{=0}$
4) $\mathbb{P}(\mathcal{N}(1,4) > 0) = 1 - \mathbb{P}(\mathcal{N}(1,4) \leq 0)$
5) $\mathbb{P}(\mathcal{N}(1,4) = 1) = 0$
6) $\mathbb{P}(|\mathcal{N}(1,4)| \leqslant 4.92)$

$\mathbb{P}(\mathcal{N}(1,4) < -0{,}5) = \mathbb{P}(2 \times \mathcal{N}(0,1) + 1 < -0{,}5) = \mathbb{P}(\mathcal{N}(0,1) < -0{,}75) = \mathbb{P}(\underbrace{-\mathcal{N}(0,1)}_{=\mathcal{N}(0,1)} < 0{,}75)$

$\mathbb{P}(|\mathcal{N}(1,4)| \leq 4{,}92) = \mathbb{P}(-4{,}92 \leq 2 \times \mathcal{N}(0,1) + 1 \leq 4{,}92) = \mathbb{P}\left(\frac{-5{,}92}{2} \leq \mathcal{N}(0,1) \leq \frac{3{,}92}{2}\right)$

$= \mathbb{P}(\mathcal{N}(0,1) \leq 3{,}92) - \mathbb{P}(\mathcal{N}(0,1) \leq -5{,}92)$

Study Title ______________________ Date __________

Exercice 5. On souhaite sonder une population de taille N sur un caractère d'intérêt (par exemple, le niveau de satisfaction au sein d'une entreprise). Si toute la population était sondée, on obtiendrait comme moyenne et variance corrigée x_i constantes

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

En tout, n personnes seront sondées. L'idée est de construire une expérience avec $n \ll N$ dont la moyenne des résultats est représentative de m.

1) Dans un premier temps, un sondage aléatoire simple et sans remise est proposé : on tire au sort simultanément les n individus pour les sonder. On note $S \subset \{1, \dots, N\}$ l'ensemble des individus choisis.

a) Montrer que la moyenne m_S issue du sondage satisfait

$$\mathbb{E}[m_S] = m.$$

b) Montrer que $\mathbb{V}(\mathbb{1}_{\{i\in S\}}) = f(1-f)$ et que, lorsque $i \neq j$, $\mathbb{C}\mathrm{ov}(\mathbb{1}_{\{i\in S\}}, \mathbb{1}_{\{j\in S\}}) = f(\frac{n-1}{N-1} - f)$, où $f = \frac{n}{N}$ est le taux de sondage. En déduire que

$$\mathbb{V}(m_S) = (1-f)\,\frac{s^2}{n}.$$

$S \subset \{1, \dots, N\}$ sous-ensemble aléatoire choisi uniformément

1) a) $m_S = \dfrac{\sum_{i\in S} x_i}{\# S} = \dfrac{1}{n}\sum_{i\in S} x_i = \dfrac{1}{n}\sum_{i=1}^{N} \mathbb{1}_{\{i\in S\}} x_i$

$$\mathbb{E}[m_S] = \frac{1}{n}\sum_{i=1}^{N} \underbrace{\mathbb{E}\left[\mathbb{1}_{\{i\in S\}}\right]}_{=\mathbb{P}(i\in S) = \frac{n}{N}} x_i = \frac{1}{n}\sum_{i=1}^{N}\frac{n}{N}x_i = \frac{1}{N}\sum_{i=1}^{N} x_i = m$$

b) $\mathbb{1}_{\{i\in S\}} \longrightarrow$ ça vaut 0 ou 1 donc $\sim \mathcal{B}\left(\frac{n}{N}\right)$

$$\mathbb{V}(\mathbb{1}_{\{i\in S\}}) = \mathbb{E}\left[\mathbb{1}_{\{i\in S\}}^2\right] - \mathbb{E}\left[\mathbb{1}_{\{i\in S\}}\right]^2 = \frac{n}{N} - \frac{n^2}{N^2} = f(1-f)$$

$i\neq j$ $\quad \mathrm{Cov}(\mathbb{1}_{\{i\in S\}}, \mathbb{1}_{\{j\in S\}}) = \mathbb{E}\left[\mathbb{1}_{\{i\in S\}} \times \mathbb{1}_{\{j\in S\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{i\in S\}}\right]\cdot\mathbb{E}\left[\mathbb{1}_{\{j\in S\}}\right] = \mathbb{P}(i\in S, j\in S) - f^2$

$$= \mathbb{P}(i\in S)\,\mathbb{P}(j\in S \mid i\in S) - f^2 = f\,\frac{n-1}{N-1} - f^2 = f\left(\frac{n-1}{N-1} - f\right)$$

Sans perte de généralité, on peut supposer que $m = 0$, en considérant $\tilde{x}_i = x_i - m$.
Donc on veut mq $\mathrm{Var}(m_S) = (1-f)\frac{s^2}{n}$ avec $s^2 = \frac{1}{N-1}\sum_{i=1}^{N} x_i^2$

$$\mathrm{Var}(m_S) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{N}\mathbb{1}_{\{i\in S\}}x_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{N}\mathbb{1}_{\{i\in S\}}x_i\right) = \frac{1}{n^2}\mathrm{Cov}\left(\sum_{i=1}^{N}\mathbb{1}_{\{i\in S\}}x_i, \sum_{j=1}^{N}\mathbb{1}_{\{j\in S\}}x_j\right)$$

$$= \frac{1}{n^2}\sum_{1\leq i,j\leq n}\mathrm{Cov}\left((\mathbb{1}_{\{i\in S\}}, \mathbb{1}_{\{j\in S\}})\, x_i x_j\right) = \frac{1}{n^2}\left[f(1-f)\sum_{i=1}^{N}x_i^2 + f\left(\frac{n-1}{N-1} - f\right)\sum_{i\neq j}x_i x_j\right]$$

or $f\left(\frac{n-1}{N-1} - f\right) = \frac{n}{N}\left(\frac{n-1}{N-1} - \frac{n}{N}\right) = \frac{n}{N}\left(\frac{n-N}{(N-1)N}\right) = \frac{-n}{N(N-1)}\left(\frac{N-n}{N}\right) = \frac{-f}{(N-1)}(1-f)$

et $\sum_{i\neq j}x_i x_j = \sum_{i=1}^{N}x_i\sum_{j\neq i}x_j = \sum_{i=1}^{N}x_i(N\underbrace{m}_{=0} - x_i) = -\sum_{i=1}^{N}x_i^2$ et $m = \frac{1}{N}\sum_{i=1}^{N}x_i = 0 \iff \sum_{i=1}^{N}x_i = 0$.

Donc $\mathrm{Var}(m_S) = \frac{1}{n^2}\left[f(1-f)\times\sum_{i=1}^{N}x_i^2 + \frac{f}{N-1}(1-f)\sum_{i=1}^{N}x_i^2\right] = \frac{1-f}{N}\left(\sum_{i=1}^{N}x_i^2\right)\left(\frac{f}{n} + \frac{f}{n(N-1)}\right)$

$$= (1-f)\times\frac{1}{n}\left(\sum_{i=1}^{N}x_i^2\right)\frac{f}{n}\times\left(1 + \frac{1}{N-1}\right) = (1-f)\times\frac{1}{n}\left(\sum_{i=1}^{N}x_i\right)\frac{1}{\cancel{N}}\times\frac{\cancel{N}}{N-1} = (1-f)\frac{s^2}{n}$$

$\longrightarrow$ chercher méthode de sondage stratifié proportionnel pour plus d'explication.

Study Title ______ Date ______

2) Dans un second temps, on stratifie la population en c catégories. On sonde indépendamment n_k personnes dans la strate k de taille N_k. On a donc $N_1 + \ldots + N_c = N$ et $n_1 + \ldots + n_c = n$.

a) Expliquer, sans faire de calcul, la raison pour laquelle on a directement

$$\mathbb{E}[m_{S_k}] = m_k \quad \text{et} \quad \mathbb{V}(m_{S_k}) = (1 - f_k)\,\frac{s_k^2}{n_k}$$

en reprenant les notations de la section précédente.

b) On considère la moyenne pondérée

$$m_S^* = \frac{1}{N}\sum_{k=1}^{c} N_k\, m_{S_k}.$$

Montrer que

$$\mathbb{E}[m_S^*] = m \quad \text{et que} \quad \mathbb{V}(m_S^*) = \frac{1}{N^2}\sum_{k=1}^{c} N_k^2\,(1 - f_k)\,\frac{s_k^2}{n_k}.$$

a) Car là on fait dans le cas général (1 strate individuel)

$$m_{S_k} = \frac{\sum_{i\in S_k} x_i}{\# S_k} = \frac{1}{n_k}\sum_{i=1}^{N_k} x_i \mathbb{1}_{\{i\in S_k\}}$$

si $m=0$ $\quad s^2 = \frac{1}{N-1}\sum_{i=1}^{N} x_i^2 \qquad m_s = \frac{1}{n}\sum_{i=1}^{N} x_i \mathbb{1}_{\{i\in s\}} \qquad m_k = \frac{1}{N_k}\sum_{i\in \mathcal{P}_k} x_i \quad$ et $\quad f_k = \frac{n_k}{N_k}$

Ici on peut juste considérer qu'on applique la même méthode pour un échantillon de taille N_k indépendamment des autres échantillons $\mathcal{P}_{\tilde{k}}$, $\tilde{k} \neq k$

b) $\mathbb{E}[m_s^*] = \frac{1}{N}\sum_{k=1}^{c} N_k\, \mathbb{E}[m_{S_k}] = \frac{1}{N}\sum_{k=1}^{c} N_k\, m_k = \frac{1}{N}\sum_{k=1}^{c} N_k \frac{1}{N_k}\sum_{i\in\mathcal{P}_k} x_i = \frac{1}{N}\sum_{k=1}^{c}\sum_{i\in S_k} x_i = \frac{1}{N}\sum_{i=1}^{N} x_i = m$

$\mathbb{V}(m_s^*) = \frac{1}{N^2}\sum_{k=1}^{c} N_k^2\, \mathbb{V}(m_{S_k}) = \frac{1}{N^2}\sum_{k=1}^{c} N_k^2 (1-f_k)\frac{s_k^2}{n_k} \qquad$ (m_{S_k} $\perp\!\!\!\perp$ donc cov nul.)

c) Dans le cas d'un sondage stratifié proportionnel (c'est-à-dire que, dans chaque strate, $f_k = f$), montrer que $m_S^* = m_S$ et que

$$\mathbb{V}(m_S^*) = \frac{1-f}{n}\sum_{k=1}^{c}\frac{N_k}{N}\,s_k^2.$$

d) On peut montrer que, lorsque le design est adapté (effectifs suffisants dans les strates et strates homogènes),

$$\mathbb{V}(m_S^*) \leqslant \mathbb{V}(m_S).$$

Commenter cette inégalité puis en déduire le principe du sondage stratifié optimal. Discuter sur la mise en pratique (ne pas oublier que les s_k^2 sont inconnus...)

c) $\mathbb{V}(m_s^*) = \frac{1}{N^2}\sum_{k=1}^{c} N_k^2 (1-f_k)\frac{s_k^2}{n_k} = (1-f)\sum_{k=1}^{c}\left(\frac{N_k}{N}\right)^2 \frac{s_k^2}{n_k} = (1-f)\sum_{k=1}^{c} \frac{\cancel{n_k}}{n}\times\frac{N_k}{N}\frac{s_k^2}{\cancel{n_k}}$

$= \frac{1-f}{n}\sum_{k=1}^{c}\frac{N_k}{N} s_k^2 \qquad\qquad f_k = f \iff \frac{n}{N} = \frac{n_k}{N_k} \iff \frac{N_k}{N} = \frac{n_k}{n}$

$m_s^* = \frac{1}{N}\sum_{k=1}^{c} N_k\, m_{S_k} = \frac{1}{N}\sum_{k=1}^{c} N_k \times \frac{1}{n_k}\sum_{i=1}^{N}\mathbb{1}_{\{i\in S_k\}} x_i \underset{f_k=f}{=} \frac{1}{N}\sum_{k=1}^{c}\frac{N}{n}\sum_{i=1}^{N}\mathbb{1}_{\{i\in S_k\}} x_i = \frac{1}{n}\sum_{k=1}^{c}\sum_{i=1}^{N}\mathbb{1}_{\{i\in S_k\}} x_i$

$= \frac{1}{n}\sum_{i=1}^{N}\mathbb{1}_{\{i\in s\}} x_i = m_s$

d) Pour améliorer notre estimation, on peut essayer de chercher un estimateur avec la variance la plus faible possible. Ici on remarque sur l'événement $\{\forall k \in \{1,\ldots,c\}\,;\, s_k \leq s\}$ on aurait $\frac{1-f}{n}\sum_{k=1}^{c}\frac{N_k}{N}s_k^2 \leq \frac{1-f}{n}s^2 \underbrace{\sum_{k=1}^{c}\frac{N_k}{N}}_{=1} = \frac{1-f}{n}s^2$

Study Title ______ Date ______

Exercice 6. Les statistiques vérifiant les conditions du Thm. 2.1 (la loi de $(X_1, \ldots, X_n) \mid T_n$ ne dépend pas de θ) sont qualifiées d'*exhaustives* pour le paramètre θ. On va montrer que, dans le cadre d'une expérience de Bernoulli, la moyenne empirique $\bar{X}_n$ est exhaustive pour le paramètre p, l'équilibre d'une pièce. On a

effectivement l'intuition que toute l'information sur l'équilibre de la pièce est contenue dans le nombre moyen de 'pile' (ou de 'face') obtenu, mais montrons-le. Établir que

$$\forall \underline{k} = (k_1, \ldots, k_n) \in \{0,1\}^n, \quad \mathbb{P}(X_1 = k_1, \ldots, X_n = k_n \mid \bar{X}_n) = \frac{1}{\binom{n}{n\bar{X}_n}}$$

Proposer une interprétation combinatoire de ce résultat.

Soit $(X_i)_{1 \leq i \leq n}$ un n-échantillon. On considère $X_i \sim \mathcal{B}(p)$ $\quad \theta = p \in [0,1]$

et $T_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. On veut mq $\forall \underline{k} \in \{0,1\}^n$ et $\forall \bar{x} \in \{\frac{k}{n} ; k \in [\![0,n]\!]\}$:

$$\mathbb{P}(X_1 = k_1, \ldots, X_n = k_n \mid \overline{X}_n = \bar{x}) = \begin{cases} \frac{1}{\binom{n}{n\bar{x}}} & \text{si } \square \\ 0 & \text{sinon} \end{cases}$$

Donc $\mathbb{P}(X_1 = k_1, \ldots, X_n = k_n \mid \overline{X}_n = \bar{x}) = \dfrac{\mathbb{P}(\{X_1 = k_1, \ldots, X_n = k_n\} \cap \{\overline{X}_n = \bar{x}\})}{\mathbb{P}(\overline{X}_n = \bar{x})}$

or $\{X_1 = k_1, \ldots, X_n = k_n\} \cap \{\overline{X}_n = \bar{x}\} \neq \emptyset$ si $\boxed{\bar{x} = \frac{1}{n}\sum_{i=1}^n k_i}$

Soit $k_i \in \{0,1\}^n$ et $\bar{x} = \frac{1}{n}\sum_{i=1}^n k_i$ d'où :

$\{X_1 = k_1, \ldots, X_n = k_n\} \subset \{\overline{X}_n = \bar{x}\} \Rightarrow \mathbb{P}(X_1 = k_1, \ldots, X_n = k_n \mid \overline{X}_n = \bar{x}) = \dfrac{\mathbb{P}(X_1 = k_1, \ldots, X_n = k_n)}{\mathbb{P}(\overline{X}_n = \bar{x})} = A$

$A = \dfrac{\mathbb{P}(X_1 = k_1) \times \ldots \times \mathbb{P}(X_n = k_n)}{\mathbb{P}(\overline{X}_n = \bar{x})} = \dfrac{p^{n\bar{x}}(1-p)^{1-n\bar{x}}}{\mathbb{P}(\overline{X}_n = \bar{x})}$ car $\mathbb{P}(X_i = k_i) = p^{k_i}(1-p)^{1-k_i}$

or $\mathbb{P}(\overline{X}_n = \bar{x}) = \mathbb{P}(n\overline{X}_n = n\bar{x}) = \binom{n}{n\bar{x}} p^{n\bar{x}} (1-p)^{1-n\bar{x}}$ car $n\overline{X}_n = \sum_{i=1}^n X_i \sim \mathcal{B}(n,p)$

Donc $\mathbb{P}(X_1 = k_1, \ldots, X_n = k_n \mid \overline{X}_n = \bar{x}) = 1 \div \binom{n}{n\bar{x}}$

On aurait pu remarquer si $\overline{X}_n = \bar{x}$, alors on doit choisir $n\bar{x}$ variables X_i égales à 1.
Sachant qu'on se place dans un cadre d'équiprobabilité.
Donc $\mathbb{P}(X_1 = k_1, \ldots, X_n = k_n \mid \overline{X}_n = \bar{x}) = \dfrac{1}{\#\{k \in \{0,1\}^n, \sum_{i=1}^n k_i = n\bar{x}\}} = \dfrac{1}{\binom{n}{n\bar{x}}}$

Exercice 7. On propose dans cet exercice un exemple de mise en pratique de la célèbre méthode statistique dite de la *capture/recapture*, pour évaluer la taille d'une population. On fait l'hypothèse qu'il n'y a ni naissances ni morts durant l'expérience, et l'on cherche à évaluer le nombre $N > 0$ de loups dans une réserve. La méthode repose sur le protocole suivant :

— On capture $N_c > 0$ loups, on les marque puis on les relâche.

— On recapture $n > 0$ loups, un par un. Pour chacun d'entre eux, on vérifie s'il est marqué ou non, puis on le relâche.

On admet que les conditions expérimentales sont suffisamment réfléchies pour que les captures/recaptures soient indépendantes et que chaque loup ait la même probabilité d'être (re)capturé. Les résultats asymptotiques portent sur n. tirage avec remise.

1) On appelle S_n le nombre de loups déjà marqués parmi les n loups recapturés. Quelle est la loi de S_n ? En déduire que $\frac{S_n}{nN_c}$ est un estimateur sans biais et consistant de $\frac{1}{N}$.

2) Justifier que $\frac{nN_c}{S_n}$ n'est pas un estimateur pertinent de N (sauf si $N_c = N$, cas critique que l'on exclura de l'étude).

1) Le nombre de loups marqués et comptés issu du tirages indépandants et avec même probabilité i.e on peut voir S_n comme la somme de v.a.r de Bernouilli de paramètre $\frac{N_c}{N}$ et indépandants : $\mathbb{P}(S_n = k) = \binom{n}{k}\left(\frac{N_c}{N}\right)^k\left(1-\frac{N_c}{N}\right)^{n-k}$ $k \in \{0,\dots,n\}$

• Sans biais : $\mathbb{E}\left[\frac{S_n}{nN_c}\right] = \frac{\mathbb{E}[S_n]}{nN_c} = \frac{n\frac{N_c}{N}}{nN_c} = \frac{1}{N}$

• Consistance : si on note b_i le i-ème tirage pour la variable S_n ; càd :

$S_n = \sum_{i=1}^n b_i$ avec $b_i \sim \mathcal{B}\left(\frac{N_c}{N}\right)$ et $(b_i)_{1\le i\le n}$ $\perp\!\!\!\perp$

Donc d'après la LGN $\frac{S_n}{n} \xrightarrow[n\to+\infty]{p.s} \mathbb{E}[b_1] = \frac{N_c}{N}$ Donc $\frac{S_n}{nN_c} \xrightarrow[n\to+\infty]{p.s} \frac{1}{N}$

2) S_n peut prendre la valeur 0 (surtout au début pour n petit).

3) Pour pallier le problème de la question précédente, on pose

$$T_n = \frac{N_c\,(n+1)}{S_n+1}.$$

a) Montrer que T_n est un estimateur consistant de N.

b) Montrer que

$$\forall\, N > 0, \quad \mathbb{E}[T_n] = N\left[1-\left(1-\frac{N_c}{N}\right)^{n+1}\right].$$

c) Déduire que T_n est un estimateur biaisé mais asymptotiquement sans biais de N.

a) $T_n = \frac{N_c\left(\frac{n+1}{n}\right)}{\boxed{\frac{S_n}{n}+\frac{1}{n}}} \xrightarrow[n\to+\infty]{p.s} \frac{N_c}{\left(\frac{N_c}{N}\right)} = N$ (preuve: 2x Slutsky et 1x continuité de la fonction $\frac{1}{x}$)

Slutsky $\xrightarrow[n\to+\infty]{p.s} \frac{N_c}{N}+0$

b) Soit $N>0$, $\mathbb{E}[T_n] = \mathbb{E}\left[\frac{N_c(n+1)}{S_n+1}\right] = N_c\cancel{(n+1)} \cdot \sum_{k=0}^{n} \frac{1}{k+1}\binom{n}{k}\left(\frac{N_c}{N}\right)^k\left(1-\frac{N_c}{N}\right)^{n-k} = N_c\sum_{k=0}^{n}\binom{n+1}{k+1}\left(\frac{N_c}{N}\right)^k\left(1-\frac{N_c}{N}\right)^{n-k}$

or $\forall k \in \{1,\dots,n\}$ $\frac{1}{k+1}\binom{n}{k} = \frac{1}{k+1}\frac{n!}{k!(n-k)!} = \frac{(n+1)!}{(k+1)!(n+1-(k+1))!} \times \frac{\cancel{1}}{n+1} = \frac{1}{n+1}\binom{n+1}{k+1}$

$\mathbb{E}[T_n] = N\sum_{k=0}^{n}\binom{n+1}{k+1}\left(\frac{N_c}{N}\right)^{k+1}\left(1-\frac{N_c}{N}\right)^{n+1-(k+1)} = N\underbrace{\sum_{k=1}^{n+1}\binom{n+1}{k}\left(\frac{N_c}{N}\right)^{k}\left(1-\frac{N_c}{N}\right)^{n+1-k}}_{\text{binôme de Newton } (a+b)^n} = N\left(\underbrace{1}_{\text{de 0 à } n+1} - \underbrace{\left(1-\frac{N_c}{N}\right)^{n+1}}_{k=0}\right)$

c) T_n est biaisé $\mathbb{E}[T_n] - N = -N\left(1-\frac{N_c}{N}\right)^{n+1} \neq 0$ mais $\left|1-\frac{N_c}{N}\right| < 1$ d'où $\mathbb{E}[T_n]-N \xrightarrow[n\to+\infty]{} 0$.

Exercice 8. Soient $X_1, \ldots, X_n$ un n-échantillon (avec $n \geqslant 2$) dont la loi parente est paramétrée par θ, et $\widehat{\theta}_n$ un estimateur arbitraire de θ. Notons

$$\forall i \in \{1, \ldots, n\}, \qquad \widehat{\theta}_{n-1}^{(i)} = \widehat{\theta}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

Il s'agit de l'estimation partielle de θ sur l'échantillon privé de sa i-ème variable. Pour fixer les idées, si par exemple $\widehat{\theta}_n = \bar{X}_n$, alors on a

$$\forall i \in \{1, \ldots, n\}, \qquad \widehat{\theta}_{n-1}^{(i)} = \frac{1}{n-1} \sum_{\substack{k=1 \\ k \neq i}}^{n} X_k.$$

À partir de $\widehat{\theta}_n$, on construit un estimateur *jackknife* selon la formule

$$\widetilde{\theta}_n = n\,\widehat{\theta}_n - \frac{n-1}{n} \sum_{i=1}^{n} \widehat{\theta}_{n-1}^{(i)}.$$

L'appellation *jackknife* vient de la traduction anglaise de *couteau suisse*, en référence à ses nombreuses applications possibles. Dans cet exercice, on étudie sa propriété de réduction de biais.

1) Supposons que le biais initial s'exprime sous la forme

$$B_\theta(\widehat{\theta}_n) = \mathbb{E}[\widehat{\theta}_n] - \theta = \sum_{j=1}^{+\infty} \frac{a_j}{n^j} \quad \text{et donc que} \quad B_\theta(\widehat{\theta}_{n-1}^{(i)}) = \sum_{j=1}^{+\infty} \frac{a_j}{(n-1)^j}$$

où $(a_j)_{j \geqslant 1}$ est une suite de réels indépendants de n, éventuellement nuls à partir d'un certain rang.

a) Montrer que

$$B_\theta(\widetilde{\theta}_n) = n\, B_\theta(\widehat{\theta}_n) - \frac{n-1}{n} \sum_{i=1}^{n} B_\theta(\widehat{\theta}_{n-1}^{(i)}).$$

b) En déduire qu'il existe une suite $(b_{n,j})_{j \geqslant 2}$ telle que

$$B_\theta(\widetilde{\theta}_n) = \sum_{j=2}^{+\infty} \frac{b_{n,j}}{[n(n-1)]^{j-1}}$$

où pour tout $j \geqslant 2$, $b_{n,j} = o(n^{j-1})$ lorsque $n \to +\infty$, et $b_{n,j} = 0$ si $a_j = 0$.

c) Constater la réduction du biais dans les cas usuels, c'est-à-dire lorsque seul a_1 est non nul, et lorsque seuls a_1 et a_2 sont non nuls.

a) $B_\theta(\tilde{\theta}_n) = \mathbb{E}[\tilde{\theta}_n] - \theta = \mathbb{E}[n\hat{\theta}_n] - \left(\frac{n-1}{n}\right) \mathbb{E}\left[\sum_{i=1}^{n} \hat{\theta}_{n-1}^{(i)}\right] - \theta = n\left(\mathbb{E}[\hat{\theta}_n] - \theta\right) + (n-1)\theta\frac{n}{n} - \left(\frac{n-1}{n}\right)\sum_{i=1}^{n} \mathbb{E}[\hat{\theta}_{n-1}^{(i)}]$

$= n\, B_\theta(\hat{\theta}_n) - \left(\frac{n-1}{n}\right)\sum_{i=1}^{n}\left(\mathbb{E}[\hat{\theta}_{n-1}^{(i)}] - \theta\right) = n\, B_\theta(\hat{\theta}_n) - \frac{n-1}{n}\sum_{i=1}^{n} B_\theta(\hat{\theta}_{n-1}^{(i)})$

b) $B_\theta(\tilde{\theta}_n) = n\sum_{j=1}^{+\infty}\frac{a_j}{n^j} - \frac{n-1}{n}\sum_{i=1}^{n}\sum_{j=1}^{+\infty}\frac{a_j}{(n-1)^j} = \sum_{j=1}^{+\infty}\frac{a_j}{n^{j-1}} - \sum_{j=1}^{+\infty}\frac{a_j}{(n-1)^{j-1}} = \sum_{j=2}^{+\infty} a_j \cdot \frac{(n-1)^{j-1} - n^{j-1}}{(n(n-1))^{j-1}}$

d'où $b_{n,j} = a_j\left((n-1)^{j-1} - n^{j-1}\right)$ et $\frac{b_{n,j}}{n^{j-1}} = a_j\left(\left(\frac{n-1}{n}\right)^{j-1} - 1\right) \xrightarrow[n\to+\infty]{} 0$ donc $b_{n,j} = o(n^{j-1})$

c) * si $a_1 \neq 0$ et $a_j = 0$ $\forall j \geqslant 2$ $\quad B_\theta(\tilde{\theta}_n) = 0$

* a_1 et a_2 non nuls mais $a_j = 0$ $\forall j > 2$

$B_\theta(\tilde{\theta}_n) = a_2\left(\frac{n-1-n}{n(n-1)}\right) = \frac{-a_2}{n(n-1)} = O(n^{-2})$ → plus rapide.

$B_\theta(\hat{\theta}_n) = \frac{a_1}{n} + \frac{a_2}{n^2} = O(n^{-1})$

Study Title ____________________ Date __________

2) Application : on souhaite estimer l'espérance $\mathbb{E}[X] = \mu$ et la variance $\mathbb{V}(X) = \sigma^2$ de l'échantillon.

a) On choisit naturellement $\widehat{\mu}_n = \bar{X}_n$. Calculer son estimation jackknife $\widetilde{\mu}_n$. Y a-t-il réduction de biais ?

b) On estime σ^2 par la variance empirique $\widehat{\sigma}_n^2 = S_n^2$. Comme la construction de son estimation jackknife $\widetilde{\sigma}_n^2$ est assez calculatoire, on admettra sans chercher à le démontrer que

$$\forall\, i \in \{1, \dots, n\}, \qquad n\, \widehat{\sigma}_n^2 - (n-1)\, \widehat{\sigma}_{n-1}^{2\,(i)} = \frac{n}{n-1}\, (X_i - \bar{X}_n)^2$$

où l'on a noté $\widehat{\sigma}_{n-1}^{2\,(i)}$ la variance empirique partielle. En déduire l'expression de $\widetilde{\sigma}_n^2$. L'expression obtenue est-elle en accord avec la réduction de biais prévue par la question 1.c) ?

a) $\hat{\mu}_n = \frac{1}{n} \sum_{j=1}^{n} X_j$ avec $\hat{\mu}_{n-1}^{(i)} = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} X_j$

$$\Rightarrow \tilde{\mu}_n = n \hat{\mu}_n - \frac{n-1}{n} \sum_{i=1}^{n} \hat{\mu}_{n-1}^{(i)} = \sum_{j=1}^{n} X_j - \frac{1}{n} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} X_j = \sum_{j=1}^{n} X_j - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_j - X_i)$$

$$= \cancel{\sum_{j=1}^{n} X_j} - \cancel{\frac{1}{n}\, n \sum_{j=1}^{n} X_j} + \frac{1}{n} \sum_{i=1}^{n} X_i = \hat{\mu}_n \longrightarrow \text{sans biais.}$$

b) $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2 = \hat{\sigma}_n^2$

$$n \hat{\sigma}_n^2 - (n-1)\, \hat{\sigma}_{n-1}^{(i)} = \frac{n}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

$$\tilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(n \hat{\sigma}_n^2 - (n-1)\, \hat{\sigma}_{n-1}^{(i)} \right) = n \hat{\sigma}_n^2 - \frac{n-1}{n} \sum_{i=1}^{n} \hat{\sigma}_{n-1}^{2(i)} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \underset{\overset{\shortparallel}{\hat{\mu}_n}}{\bar{X}_n})^2 = \sigma_n^2$$

Study Title ______________________ Date ____________

Exercice 10. On considère un modèle de Cauchy $X \sim \mathcal{C}(\theta)$, c'est-à-dire que la v.a.r. parente du n-échantillon $X_1, \ldots, X_n$ suit une loi de Cauchy dont la densité est donnée par

$$\forall x \in \mathbb{R}, \quad f_X(x\,;\theta) = \frac{1}{\pi\,(1+(x-\theta)^2)}$$

où $\theta \in \Theta = \mathbb{R}$ est un paramètre de position. On cherche à estimer θ.

1) Justifier que la méthode des moments n'est pas applicable dans sa version standard.

2) On considère la fonction h définie par

$$h(x) = \mathbb{1}_{\{x>0\}} - \mathbb{1}_{\{x \leqslant 0\}}.$$

Montrer que

$$\int_{\mathbb{R}} h(x)\, f_X(x\,;\theta)\,\mathrm{d}x = \frac{2}{\pi} \arctan\theta.$$

3) Appliquer la méthode des moments généralisée (cf. la remarque p. 14) pour proposer un estimateur $\widehat{\theta}_n$ de θ. Étudier sa consistance forte.

1) On remarque que $|x|\, f_x(x;\theta) \underset{+\infty}{\sim} \frac{1}{\pi |x|}$ d'où le fait que $\mathbb{E}[|X|]$, et donc $\mathbb{E}[X]$ n'existe pas (d'après le critère de Riemann, ce n'est pas intégrable)

2) $\mathbb{E}[h(X)] \underset{\text{transfert}}{=} \int_{\mathbb{R}} \left(\mathbb{1}_{\{x>0\}} - \mathbb{1}_{\{x\leq 0\}}\right) f_x(x;\theta)dx = \int_0^{+\infty} f_x(x;\theta)dx - \int_{-\infty}^0 f_x(x;\theta)\,dx$

$= \frac{1}{\pi}\int_0^{+\infty} \frac{dx}{1+(x-\theta)^2} - \frac{1}{\pi}\int_{-\infty}^0 \frac{dx}{1+(x-\theta)^2} \underset{u=x-\theta}{=} \frac{1}{\pi}\left[\int_{-\theta}^{+\infty} \frac{du}{1+u^2} - \int_{-\infty}^{-\theta} \frac{du}{1+u^2}\right]$

$= \frac{1}{\pi}\left(\left[\arctan u\right]_{-\theta}^{+\infty} - \left[\arctan u\right]_{-\infty}^{-\theta}\right) = \frac{1}{\pi}\left(\frac{\pi}{2} + \arctan(\theta) + \arctan\theta - \frac{\pi}{2}\right) = \frac{2}{\pi}\arctan(\theta)$

3) Par la méthode généralisée :

$\frac{1}{n}\sum_{k=1}^{n} h(X_k) \xrightarrow[n\to+\infty]{\mathbb{P}} \mathbb{E}[h(X)]$ d'où $\hat{\theta}_n = \tan\left(\frac{\pi}{2n}\sum_{k=1}^{n} h(X_k)\right) \xrightarrow[n\to+\infty]{\mathbb{P}} \theta$

(d'après le CMT car $x \mapsto \tan(x)$ est continue)

Study Title ______ Date ______

Exercice 9. Dans une étude remontant à 1965, C. R. Rao, un célèbre statisticien indien, s'intéresse à la classification génétique d'un ensemble de représentants d'une même espèce animale en 4 groupes. Son modèle statistique s'écrit

$$\mathbb{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3, N_4 = n_4) = K(\underline{n}) \left(\frac{2+\pi}{4}\right)^{n_1} \left(\frac{1-\pi}{4}\right)^{n_2+n_3} \left(\frac{\pi}{4}\right)^{n_4}$$

où $\underline{n} = (n_1, n_2, n_3, n_4) \in \mathbb{N}^4$ est tel que $n_1+n_2+n_3+n_4 = n$, $\pi \in \Theta =]0,1[$ est un paramètre et $K(\underline{n}) > 0$ est une constante de normalisation indépendante de π. Selon ce modèle, N_i est une variable aléatoire comptant le nombre d'individus classés dans la catégorie $i \in \{1, \ldots, 4\}$.

1) On note $\ell_{\text{class}}(\underline{n}\,;\pi) = \mathbb{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3, N_4 = n_4)$ la vraisemblance de la classification des n animaux en (n_1, n_2, n_3, n_4), et $\ell\ell_{\text{class}}(\underline{n}\,;\pi)$ la log-vraisemblance associée. Montrer que

$$\frac{\partial}{\partial \pi}\, \ell\ell_{\text{class}}(\underline{n}\,;\pi) = \frac{-n\,\pi^2 + (n_1 - 2\,(n_2+n_3) - n_4)\,\pi + 2\,n_4}{\pi(2+\pi)(1-\pi)}.$$

2) L'expérience de Rao est composée de $n = 197$ animaux répartis dans les 4 groupes selon les effectifs $\underline{n} = (125, 18, 20, 34)$. Aider Rao à ajuster numériquement son paramètre π sur ses données.

1) $\ell\ell_{class}(\underline{n};\pi) = \ln\left(\ell_{class}(\underline{n};\pi)\right) = \ln\left(K(\underline{n})\right) + n_1 \ln\left(\frac{2+\pi}{4}\right) + (n_2+n_3)\ln\left(\frac{1-\pi}{4}\right) + n_4 \ln\left(\frac{\pi}{4}\right)$

$= \ln\left(K(\underline{n})\right) \quad n_1 \ln(2+\pi) + (n_2+n_3)\ln(1-\pi) + n_4 \ln \pi - \underline{n} \ln 4$

$\frac{\partial}{\partial\pi}\ell\ell_{class}(\underline{n};\pi) = \frac{n_1}{2+\pi} - \frac{n_2+n_3}{1-\pi} + \frac{n_4}{\pi} = \frac{\pi n_1(1-\pi) + (n_2+n_3)(2+\pi)\pi + (2+\pi)(1-\pi)n_4}{(2+\pi)(1-\pi)\pi}$ $_{//A}$

$A = -n_1\pi^2 + \pi n_1 - 2n_2\pi - 2n_3\pi - \cancel{\pi^2 n_2} - \cancel{\pi^2 n_3} + 2n_4 \underbrace{- 2\pi n_4 + \pi n_4}_{-\pi n_4} - \cancel{\pi^2 n_4}$

$= -(\underbrace{n_1+n_2+n_3+n_4}_{n})\pi^2 + (n_1 - 2(n_2+n_3) - n_4)\pi + 2n_4$

Maximum de vraisemblance = $\{f_\theta\,;\theta \in I\}$ une famille de densités. On suppose que $X \sim f_{\theta_0}$. On a un n-échantillon $X_1,\ldots,X_n$. La densité de $(X_1,\ldots,X_n)$ est donnée par $g_{\theta_0}(x_1,\ldots,x_n) = \prod_{i=0}^{n} f_{\theta_0}(x_i)$

On espère donc que $g_\theta(X_1,\ldots,X_n)$ soit maximum pour $\theta = \theta_0$

On cherche π_0 tel que : $\frac{\partial}{\partial\pi}\ell\ell_{class}(\underline{n};\pi_0) = 0 \iff -\underline{n}\pi^2 + (n_1 - 2(n_2+n_3) - n_4)\pi + 2n_4 = 0$.

$\iff \hat{\pi}_n = \frac{-b_n + \sqrt{b_n^2 - 4a_n c_n}}{2a_n}$ avec $a_n = -n$; $b_n = n_1 - 2(n_2+n_3) - n_4$ et $c_n = 2n_4$

2) Application numérique : $a_n = -197$ $\quad b_n = 15$ $\quad c_n = 68$ $\qquad \hat{\pi}_n \approx 60\%$

Exercice 11. On se place dans le cadre du modèle $X \sim \mathcal{G}(p)$ pour $p \in \Theta =]0,1[$, et l'on va chercher à améliorer une estimation naïve de p par la méthode de Rao-Blackwell.

1) Montrer que $\bar{X}_n$ est un estimateur sans biais de $\frac{1}{p}$.

2) En déduire que l'estimateur des moments $\widehat{p}_n$ de p est fortement consistant mais biaisé. Pour le biais, il est suffisant de montrer que la propriété est fausse pour $n = 1$. On rappelle que

$$\forall |z| < 1, \qquad \ln(1-z) = -\sum_{k=1}^{+\infty} \frac{z^k}{k}.$$

3) Montrer que l'estimateur $\widetilde{p}_n = \mathbb{1}_{\{X_1=1\}}$ est un estimateur sans biais de p. Pourquoi peut-on qualifier cette estimation de naïve ?

4) Vérifier que S_n est exhaustive pour p, puis montrer que l'amélioré de Rao-Blackwell de $\widetilde{p}_n$ par rapport à S_n est donné par

$$\widetilde{p}_n^* = \frac{1-\frac{1}{n}}{\bar{X}_n - \frac{1}{n}}.$$

5) Étudier le biais et la consistance de $\widetilde{p}_n^*$.

6) Un fabricant souhaite évaluer la fiabilité des machines qu'il met en vente, et pour cela il récolte auprès de ses clients le jour où est survenue la première panne de chacune des $n = 20$ machines qu'il a vendues.

succès : tombé en panne

Machine	1	2	3	4	5	6	7	8	9	10	
Fonctionnement normal	21 j	24 j	44 j	2 j	23 j	27 j	31 j	11 j	20 j	25 j	+1
tombé en panne :	22^e	25^e	45^e	3^e	24^e	...					

Machine	11	12	13	14	15	16	17	18	19	20	
Fonctionnement normal	34 j	28 j	30 j	32 j	31 j	17 j	24 j	18 j	23 j	22 j	+1

Proposer une estimation fiable de la probabilité qu'une de ses machines tombe en panne dès le premier jour de mise en service ainsi que du temps de fonctionnement moyen avant la première panne.

$\mathbb{P}(X=k) = p(1-p)^{k-1}$ avec $k \in \mathbb{N}^*$

1) $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{\cancel{n}} \cdot \cancel{n}\frac{1}{p} = \frac{1}{p}$ (car X_i i.i.d $\forall i$ et $X_i \sim \mathcal{G}(p)$.)

Donc $\bar{X}_n$ est sans biais

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to+\infty]{\mathbb{P}} \frac{1}{p}$ d'après la loi des grands nombres. Donc $\bar{X}_n$ est consistant

2) On a que $\frac{1}{\bar{X}_n}$ est un estimateur consistant de p car $x \mapsto \frac{1}{x}$ est continue sur $\mathbb{R}_+$

$|1-p|<1.$

Soit $n=1$ $\quad \mathbb{E}\left[\frac{1}{X_1}\right] = \mathbb{E}\left[\frac{1}{X}\right] = \sum_{k=1}^{+\infty}\frac{1}{k}\mathbb{P}(X=k) = \sum_{k=1}^{+\infty}\frac{1}{k}p(1-p)^{k-1} = \frac{p}{1-p}\sum_{k=1}^{+\infty}\frac{(1-p)^k}{k} = \frac{p}{1-p}\left[-\ln(1-(1-p))\right]$

$= -\frac{p\ln p}{1-p} \neq p$ car on aura jamais $\ln p = 1-p$.

3) $\mathbb{E}[\widetilde{p}_n] = \mathbb{P}(X_1 = 1) = p \rightarrow$ sans biais

Ça ne prend pas en compte la taille de l'échantillon, $\widetilde{p}_n$ renvoi 0 ou 1 à chaque fois

4) $S_n = \sum_{k=1}^{n} X_k$; on veut montrer que la loi $(X_1, \dots, X_n | S_n)$ est indépendante de p.

Soient $\bar{k} = (k_1, \dots, k_n) \in (\mathbb{N}^*)^n$ et $s \in \mathbb{N}^*$, $\mathbb{P}((X_1,\dots,X_n) = \bar{k} | S_n = s) = \frac{\mathbb{P}((X_1,\dots,X_n)=\bar{k},\, S_n = s)}{\mathbb{P}(S_n = s)} = 0$ si $\sum_{i=1}^{n} k_i \neq s$

On suppose que $s = \sum_{i=1}^{n} k_i$

$A = \mathbb{P}(X_1 = k_1, \dots, X_n = k_n, \underbrace{S_n = s}_{\text{inclus}}) = \mathbb{P}(X_1 = k_1, \dots, X_n = x_n) = \prod_{i=1}^{n}\mathbb{P}(X_i = k_i) = \prod_{i=1}^{n} p(1-p)^{k_i - 1} = p(1-p)^{s-n}$

$\mathbb{P}_n(S_n = n) = ?$ le nb d'essais nécessaires pour avoir n succès ?

$\mathbb{E}\left[\mathbb{1}_{\{X_n = 1\}} | S_n = s\right] = \mathbb{P}(X_1 = 1 | S_n = s) = \frac{\mathbb{P}(X_1 = 1, S_n = s)}{\mathbb{P}(S_n = s)} \qquad s \in \mathbb{N} \quad s \geq n :$

$\{X_1 = 1\} \cap \{X_1 + \dots + X_n = s\} = \{X_1 = 1\} \cap \{\underbrace{X_2 + \dots + X_n}_{T_{n-1}} = s-1\} \rightarrow$ sont $\perp\!\!\!\perp$ maintenant

$S_n \sim \mathcal{P}(n,p)$ loi de pascal, voir formulaire

$\forall k \in \{n, n+1, \dots\}$ $\mathbb{P}(S_n = k) = \binom{k-1}{k-n} p^n (1-p)^{k-n}$

$T_{n-1} \sim \mathcal{P}(n-1, p)$ $\forall k \in \{n-1, n, \dots\}$ $\mathbb{P}(T_{n-1} = k) = \binom{k-1}{k-n+1} p^{n-1} (1-p)^{k-n+1}$

Donc $$\tilde{p}_n^*(s) = \frac{\mathbb{P}(X_1 = 1)\,\mathbb{P}(T_{n-1} = s-1)}{\mathbb{P}(S_n = s)} = \frac{p\binom{s-2}{s-n} p^{n-1}(1-p)^{s-n}}{\binom{s-1}{s-n} p^n (1-p)^{s-n}} = \frac{\frac{(s-2)!}{(s-n)!(n-2)!}}{\frac{(s-1)!}{(s-n)!(n-1)!}} = \frac{n-1}{s-1}$$

$s-2 < s-n$ donc $\binom{k}{n}$ $k<n$

$$= \frac{1 - \frac{1}{n}}{\frac{s}{n} - \frac{1}{n}} \quad \forall s \geq n$$

5) $\tilde{p}_n^*$ est sans biais par Rao-Blackwell.

$$\tilde{p}_n^* = \frac{1 - \frac{1}{n}}{\overline{X}_n - \frac{1}{n}} \xrightarrow{p.s} p$$ fortement consistant.

per CMT

$\overline{X}_n \xrightarrow[n \to +\infty]{p.s} \frac{1}{p}$ par LFGN.

6) $\hat{\mathbb{P}}(X = 1) = \tilde{p}_n^*$ $\hat{\mathbb{E}}[X] = \overline{X}_n$

Exercice 12. On veut estimer la moyenne d'une variable gaussienne par une approche bayésienne. Pour simplifier, on suppose que la variance σ^2 est connue. Déterminer la loi *a posteriori* ainsi que l'estimateur de la moyenne *a posteriori* de X sur la base d'un n-échantillon $X_1, \ldots, X_n$ dans le modèle hiérarchique $X \sim \mathcal{N}(\mu, \sigma^2)$ et $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$. Commenter le résultat obtenu et comparer avec l'approche fréquentiste.

On veut calculer:

$$\pi(\mu|\underline{X}) \propto l_x(\underline{X}|\mu)\pi(\mu) = \prod_{i=1}^n f_x(x_i;\mu,\sigma^2) = \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \times \frac{1}{\sqrt{2\pi\tau_0^2}} \times e^{-\frac{(\mu-\mu_0)^2}{2\tau_0^2}}$$

$$\propto \exp\left[-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\tau_0^2}\right] = \exp\left[-\frac{1}{2}\sum_{i=1}^n\left(\frac{(x_i-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{n\tau_0^2}\right)\right]$$

Si on regarde dans l'exponentielle: on regarde a, b, c tel que:

$$\frac{(x_i-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{n\tau_0^2} = \frac{(\mu-a)^2}{b} + c$$

$$= \frac{n\tau_0^2(x_i^2 - 2x_i\mu + \mu^2) + \sigma^2(\mu^2 - 2\mu\mu_0 + \mu_0^2)}{\sigma^2 n\tau_0^2}$$

$$= \frac{\mu^2(\sigma^2 + n\tau_0^2) - 2(x_i n\tau_0^2 + \mu_0\sigma^2)\mu + n\tau_0^2 x_i^2 + \sigma^2\mu_0^2}{\sigma^2 n\tau_0^2}$$

$$= \frac{n\tau_0^2+\sigma^2}{\sigma^2 n\tau_0^2}\left(\mu - \left(\frac{x_i n\tau_0^2 + \mu_0\sigma^2}{n\tau_0^2+\sigma^2}\right)\right)^2 + c \qquad c \perp\!\!\!\perp \mu$$

On a alors : $\pi(\mu|\underline{X}) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^n \frac{n\tau_0^2+\sigma^2}{\sigma^2 n\tau_0^2}\left(\mu - \frac{x_i n\tau_0^2+\mu_0\sigma^2}{n\tau_0^2+\sigma^2}\right)^2\right]$

Pour maximiser, on peut utiliser le log:

$$\log \pi(\mu|\underline{X}) \propto \sum_{i=1}^n \frac{n\tau_0^2+\sigma^2}{\sigma^2 n\tau_0^2}\left(\mu - \left(\frac{x_i n\tau_0^2+\mu_0\sigma^2}{n\tau_0^2+\sigma^2}\right)\right)^2 \propto \sum_{i=1}^n\left(\mu - \left(\frac{x_i n\tau_0^2+\mu_0\sigma^2}{n\tau_0^2+\sigma^2}\right)\right)^2$$

$$\frac{\partial \log}{\partial\mu}\pi(\mu|\underline{X}) \propto \sum_{i=1}^n\left(\mu - \left(\frac{x_i n\tau_0^2+\mu_0\sigma^2}{n\tau_0^2+\sigma^2}\right)\right) = n\mu - \frac{n\tau_0^2\sum_{i=1}^n x_i + n\mu_0\sigma^2}{\sigma^2+n\tau_0^2}$$

Donc $\frac{\partial\log}{\partial\mu}\pi(\mu|\underline{X}) = 0 \iff \hat{\mu}_n = \frac{\tau_0^2\sum_{i=1}^n x_i + \mu_0\sigma^2}{n(\sigma^2+n\tau_0^2)} = \frac{\tau_0^2\bar{X}_n}{\sigma^2+n\tau_0^2} + \frac{\mu_0\sigma^2}{n(\sigma^2+n\tau_0^2)}$

Study Title ______________ Date ______

Exercice 13. On veut estimer le paramètre $\lambda > 0$ d'un modèle de Poisson à l'aide d'une approche bayésienne. Pour cela, on considère le modèle $X \mid \lambda \sim \mathcal{P}(\lambda)$ avec un *a priori* exponentiel $\lambda \sim \mathcal{E}(1)$. On note $\underline{X} = (X_1, \ldots, X_n)$ le vecteur des observations.

1) En reprenant les notations du cours, montrer que la loi *a posteriori* est caractérisée par

$$\pi(\lambda \mid \underline{X}) \propto \mathrm{e}^{-(n+1)\lambda} \lambda^{n\bar{X}_n} \mathbb{1}_{\{\lambda \geqslant 0\}}$$

et reconnaître une loi usuelle de paramètres a_n et b_n à identifier.

2) En admettant que la loi en question est de mode $(a_n - 1)/b_n$, en déduire l'estimateur bayésien du maximum *a posteriori* de λ.

3) Rappeler l'expression de l'unique EMV de λ vu en cours (sans le redémontrer). Comparer et commenter (en 1 ou 2 lignes) l'expression de ces estimateurs.

1. $\pi(\lambda \mid \underline{X}) \propto \ell(\underline{X} \mid \lambda)\, \pi(\lambda) = \left(\prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{X_i}}{X_i!} \right) \times 1 \times e^{-1 \times \lambda} \mathbb{1}_{\{\lambda \geq 0\}}$

$\propto e^{-n\lambda} \lambda^{n\overline{X}_n} e^{-\lambda} \mathbb{1}_{\{\lambda \geq 0\}} = e^{-(n+1)\lambda} \lambda^{n\overline{X}_n} \mathbb{1}_{\{\lambda \geq 0\}}$

On reconnaît la loi Gamma : $\lambda \mid \underline{X} \sim \Gamma(n\overline{X}_n + 1, n+1)$ avec $a_n = n\overline{X}_n + 1$
$b_n = n+1$.

2) mode : $\frac{a_n - 1}{b_n}$ $\hat{\lambda}_n = \frac{n\overline{X}_n + 1 - 1}{n+1} = \frac{n\overline{X}_n}{n+1} = \frac{n}{n+1}\overline{X}_n$

3) $\hat{\lambda}_n = \overline{X}_n$ pour l'estimateur bayésien, quand $n \geq 1$ alors les 2 estimateurs sont positifs.

Exercice 14. On considère un n-échantillon $X_1, \dots, X_n$ dont la v.a.r. parente X est de loi de Pareto de paramètre λ où $\lambda > 1$ est à estimer. Sa densité est donnée par

$$\forall x \geqslant 1, \qquad f_X(x\,;\lambda) = \frac{\lambda}{x^{\lambda+1}}.$$

1) Déterminer l'EMV $\widehat{\lambda}_n$ de λ.

2) Par la suite, on va supposer que $n > 2$. On s'intéresse maintenant au comportement de

$$L_n = \sum_{i=1}^{n} \ln(X_i).$$

   a) En admettant que la somme de n v.a.r. i.i.d. de loi $\mathcal{E}(\lambda)$ est de loi $\Gamma(n, \lambda)$, montrer que $L_n \sim \Gamma(n, \lambda)$.

   b) Montrer que

$$\mathbb{E}\left[\frac{1}{L_n}\right] = \frac{\lambda}{n-1}.$$

   c) En déduire que l'EMV est un estimateur biaisé de λ mais qu'il est asymptotiquement sans biais. Proposer alors un estimateur non biaisé $\widehat{\lambda}_n^*$ de λ.

   d) Montrer que

$$\mathbb{E}\left[\frac{1}{L_n^2}\right] = \frac{\lambda^2}{(n-1)(n-2)}.$$

3) On s'intéresse maintenant à la variance de l'estimateur $\widehat{\lambda}_n^*$.

   a) Montrer que

$$\mathbb{V}(\widehat{\lambda}_n^*) = \frac{\lambda^2}{n-2}.$$

   b) Étudier l'efficacité de l'estimateur $\widehat{\lambda}_n^*$.

4) On cherche enfin à construire le test le plus puissant de $\mathcal{H}_0$: "$\lambda = \lambda_0$" contre $\mathcal{H}_1$: "$\lambda < \lambda_0$" pour une valeur-test $\lambda_0 > 1$.

   a) Montrer que pour tout $k > 0$,

$$\Big\{R_X(\underline{X}\,;\lambda_0, \lambda_1) > k\Big\} = \Big\{n \ln \frac{\lambda_1}{\lambda_0} + (\lambda_0 - \lambda_1)\, L_n > \ln k\Big\}.$$

   b) Construire un test de $\mathcal{H}_0$: "$\lambda = \lambda_0$" contre $\mathcal{H}_1$: "$\lambda = \lambda_1$" lorsque $\lambda_1 < \lambda_0$ qui soit le plus puissant au niveau α.

   c) Le test est-il UPP$_\alpha$ pour l'alternative $\mathcal{H}_1$: "$\lambda < \lambda_0$" ?

1) $\forall x \geqslant 1$, $\forall \lambda$ $\quad \ell_X(\underline{X}\,;\lambda) = \dfrac{\lambda^n}{\left(\prod_{i=1}^n X_i\right)^{\lambda+1}} \Rightarrow \ell\ell_X(\underline{X}\,;\lambda) = n \ln \lambda - \left(\sum_{i=1}^n \ln X_i\right).(\lambda+1)$ $\quad \mathcal{L}_X$ dérivable sur Θ

$\Rightarrow \dfrac{\partial \ell\ell_X}{\partial \lambda}(\underline{X}\,;\lambda) = \dfrac{n}{\lambda} - \sum_{i=1}^n \ln(X_i) = \dfrac{n}{\lambda} - L_n = 0 \iff \lambda_n^\triangledown = \dfrac{n}{L_n}$

$\Rightarrow \dfrac{\partial \ell\ell_X}{\partial \lambda}(\underline{X}\,;\lambda) = -\dfrac{n}{\lambda^2} < 0$ $\quad$ donc $\hat{\lambda}_n = \lambda_n^\triangledown = \dfrac{n}{L_n}$

2) a) Soit $Y = \ln X$, $\quad Y(\Omega) = \mathbb{R}^+$

$\forall t \geqslant 0 \quad \mathbb{P}(Y \leq t) = \mathbb{P}(X \leq e^t) = \displaystyle\int_1^{e^t \geqslant 1} \frac{\lambda}{x^{\lambda+1}}\,dx = \lambda \left[\frac{x^{-\lambda}}{-\lambda}\right]_1^{e^t} = \lambda \left[\frac{(e^t)^{-\lambda}}{-\lambda} - \frac{1}{-\lambda}\right] = 1 - e^{-\lambda t}$

$\Rightarrow f_Y(t\,;\lambda) = (1 - e^{-\lambda t})' = \lambda e^{-\lambda t}$ $(t \geqslant 0)$, d'où $Y \sim \mathcal{E}(\lambda)$ et $L_n \sim \Gamma(n, \lambda)$.

$\mathbb{E}\left[\dfrac{1}{L_n}\right] = \displaystyle\int_{\mathbb{R}^+} \frac{1}{x} \cdot \frac{\lambda^n x^{n-1} e^{-\lambda x}}{\Gamma(n)}\,dx = \frac{\lambda^n}{(n-1)!} \int_0^{+\infty} x^{n-2} e^{-\lambda x}\,dx \underset{s=\lambda x}{=} \frac{\lambda^n}{(n-1)!} \int_0^{+\infty} \left(\frac{s}{\lambda}\right)^{n-2} e^{-s} \frac{ds}{\lambda} = \frac{\lambda (n-2)!}{(n-1)!} = \frac{\lambda}{n-1}$

Donc $\mathbb{E}[\hat{\lambda}_n] = n\, \mathbb{E}\left[\dfrac{1}{L_n}\right] = \underbrace{\dfrac{n}{n-1}}_{\xrightarrow[n\to+\infty]{} 1} \lambda \neq \lambda$ $\quad$ mais $\lim\limits_{n \to +\infty} \mathbb{E}[\hat{\lambda}_n] = \lambda$. $\quad \Rightarrow \quad \lambda_n^*$ est sans biais

d) $\mathbb{E}\left[\frac{1}{L_n^2}\right] = \dots = \frac{\lambda^n}{(n-1)!}\int_0^{+\infty}\left(\frac{s}{\lambda}\right)^{n-3} e^{-s}\frac{ds}{\lambda} = \frac{\lambda^2 \cdot (n-3)!}{(n-1)!} = \frac{\lambda^2}{(n-1)(n-2)}$

3) a) D'où $Var\left[\hat{\lambda}_n^*\right] = (n-1)^2\,\mathbb{E}\left[\frac{1}{L_n^2}\right] - \lambda^2 = (n-1)^2 \times \frac{\lambda^2}{(n-1)(n-2)} - \lambda^2 = \frac{n-1}{n-2}\lambda^2 - \lambda^2 = \lambda^2\left[\frac{n-1}{n-2} - 1\right] = \frac{\lambda^2}{n-2}$

$\underset{(X\geq 1)}{I_X(\lambda)} = -\mathbb{E}\left[\frac{\partial^2}{\partial\lambda^2}\ln f_X(X;\lambda)\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\lambda^2}\left(\ln\lambda - (\lambda+1)\ln x\right)\right] = \frac{1}{\lambda^2} \Rightarrow I_n(\lambda) = \frac{n}{\lambda^2} \neq Var^{-1}(\lambda_n^*)$, donc

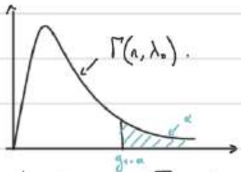
$\hat{\lambda}_n^*$ n'est pas efficace mais assymptotiquement efficace car $Var(\hat{\lambda}_n^*) \cdot I_n(\lambda) = \frac{n}{n-2} \xrightarrow[n\to+\infty]{} 1$

4) a) $\mathcal{H}_0$: "$\lambda = \lambda_0$" vs $\mathcal{H}_1$: "$\lambda = \lambda_1$" avec $\lambda_1 < \lambda_0$

Soit $R = R_x(\underline{X};\lambda_0,\lambda_1) = \frac{\ell_x(\underline{X};\lambda_1)}{\ell_x(\underline{X};\lambda_0)} = \left(\frac{\lambda_1}{\lambda_0}\right)^n\left(\prod_{i=1}^n X_i\right)^{\lambda_0-\lambda_1}$

$R > k \underset{\ln \text{ continue et croissante}}{\Longleftrightarrow} n\ln\left(\frac{\lambda_1}{\lambda_0}\right) + (\lambda_0-\lambda_1)L_n > \ln k$

b) $\alpha = \underset{\text{sous } \mathcal{H}_0}{\mathbb{P}_0}(R>k) = \mathbb{P}_0\left(L_n \underset{\lambda_0>\lambda_1}{>} \frac{1}{\lambda_0-\lambda_1}\left(\ln k - n\ln\frac{\lambda_1}{\lambda_0}\right)\right)$



où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\Gamma(n,\lambda_0)$.

Sous $\mathcal{H}_0$: $L_n \sim \Gamma(n,\lambda_0)$.

Par Neymann-Pearson, le test le plus puissant de $\mathcal{H}_0$ contre $\mathcal{H}_1$ s'écrit : $T = \mathbb{1}_{\{L_n > q_{1-\alpha}\}}$

c) Oui, conclusion valable uniformément en λ_1, à condition que $\lambda_1 < \lambda_0$

Study Title ______________________ Date __________

Exercice 15. Soit le modèle $X \sim \mathcal{N}(\mu, \sigma^2)$.

1) Montrer, en utilisant l'information de Fisher, que

$$\sqrt{n}\begin{pmatrix} \bar{X}_n - \mu \\ S_n^2 - \sigma^2 \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Gamma) \quad \text{avec} \quad \Gamma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}.$$

2) En déduire un test asymptotique de $\mathcal{H}_0$: "$\mu = 0$" contre $\mathcal{H}_1$: "$\mu \neq 0$". Construire de même un test asymptotique de $\mathcal{H}_0$: "$\sigma^2 = 1$" contre $\mathcal{H}_1$: "$\sigma^2 \neq 1$".

1) Le modèle $X \sim \mathcal{N}(\mu, \sigma^2)$ est régulier, on a vu que l'unique EMV vaut : $(\bar{X}_n, S_n^2)$

D'après la proposition 2.10. (p. 29) :

$$\sqrt{n}\begin{pmatrix} \bar{X}_n - \mu \\ S_n^2 - \sigma^2 \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, I_X^{-1}(\mu, \sigma^2)\right)$$

$$I_X(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \quad \text{d'où} \quad I_X(\mu, \sigma^2)^{-1} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}$$

↳ p. 28

$$\Rightarrow \begin{cases} \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2) & \Rightarrow \sqrt{n}\,\dfrac{\bar{X}_n - \mu}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \quad (1) \\ \sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 2\sigma^4) & \Rightarrow \sqrt{n}\,\dfrac{S_n^2 - \sigma^2}{\sigma^2\sqrt{2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \quad (2) \end{cases}$$

⚠ $\cancel{\Leftrightarrow}$

(1) $\overset{\text{(Slutsky)}}{\Rightarrow}$ $\sqrt{n}\,\dfrac{\bar{X}_n - \mu}{\sqrt{S_n^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

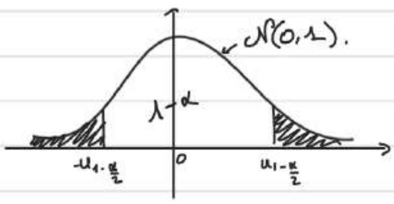
↳ on peut aussi utiliser S_n^{*2} car $n \to +\infty$ donc on s'en fiche, c'est des tests asymptotiques *

(2) $\overset{\text{(Slutsky)}}{\Rightarrow}$ $\sqrt{n}\,\dfrac{S_n^2 - \sigma^2}{S_n^2\sqrt{2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$

↳ idem que *

2)

<u>Sous H_0</u> : (1) $\Longrightarrow \sqrt{n}\,\dfrac{\bar{X}_n}{\sqrt{S_n^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$



H_0 : "$\mu = 0$" vs H_1 : "$\mu \neq 0$"

$$T = \mathbb{1}\left\{\left|\sqrt{n}\,\frac{\bar{X}_n}{\sqrt{S_n^2}}\right| > u_{1-\frac{\alpha}{2}}\right\}$$

<u>Sous H_0</u> : "$\sigma^2 = 1$" (2) $\Rightarrow \sqrt{n}\,\dfrac{S_n^2 - 1}{\sqrt{2}\,S_n^2} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$

$$T = \mathbb{1}\left\{\left|\sqrt{n}\,\frac{S_n^2 - 1}{\sqrt{2}\,S_n^2}\right| > u_{1-\frac{\alpha}{2}}\right\}$$

Exercice 16. On a vu à de nombreuses reprises que la moyenne empirique $\bar{X}_n$ permettait l'estimation de p dans le modèle $X \sim \mathcal{B}(p)$. Montrer, en utilisant l'inégalité de Bienaymé-Tchebychev, que

$$\forall \varepsilon > 0, \quad \mathbb{P}(|\bar{X}_n - p| < \varepsilon) \geqslant 1 - \frac{1}{4\,n\,\varepsilon^2}.$$

En déduire un intervalle de confiance exact de sécurité $1-\alpha$ pour le paramètre p, avec un risque $\alpha = 1/(4\,n\,\varepsilon^2)$.

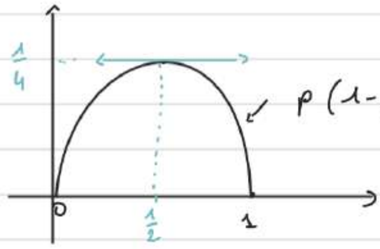
D'après l'inégalité de Bienaymé-Tchebychev : $\varepsilon > 0$

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \geq \varepsilon) \leq \frac{Var(Z)}{\varepsilon^2}$$

Soit $\varepsilon > 0$

$$\mathbb{P}(|\bar{X}_n - p| < \varepsilon) = 1 - \mathbb{P}(|\bar{X}_n - p| \geq \varepsilon) \leq 1 - \frac{p(1-p)}{n\,\varepsilon^2}$$

car $Z = \bar{X}_n$ d'où $\mathbb{E}[Z] = p$ et $Var(Z) = \frac{p(1-p)}{n}$



$p(1-p) \leq \frac{1}{2}\left(1 - \frac{1}{2}\right) \; \forall p.$ d'où $\mathbb{P}(|\bar{X}_n - p| < \varepsilon) \leq 1 - \frac{1}{4n\varepsilon^2}$

Avec $\alpha = \frac{1}{4n\varepsilon^2}$, on doit choisir $\varepsilon = \frac{1}{2\sqrt{n\alpha}}$ car $\varepsilon > 0$.

$$\mathbb{P}(|\bar{X}_n - p| < \varepsilon) = \mathbb{P}(-\varepsilon + \bar{X}_n < p < \varepsilon + \bar{X}_n) \geq 1 - \alpha$$

Donc $IC_{1-\alpha}(p) = \left] -\frac{1}{2\sqrt{n\alpha}} + \bar{X}_n \; ; \; \frac{1}{2\sqrt{n\alpha}} + \bar{X}_n \right[$ (ex $\alpha = 5\%$ ou $\alpha = 1\%$)

Exercice 17. Soit $Z_1, \ldots, Z_n$ un n-échantillon de loi $Z \sim \mathcal{N}(0,1)$ et

$$\forall\, 1 \leqslant i \leqslant n, \qquad X_i = \mu + \sigma\, Z_i.$$

Il est donc clair que $X_1, \ldots, X_n$ est un n-échantillon de loi $X = \mu + \sigma\, Z \sim \mathcal{N}(\mu, \sigma^2)$.

1) On appelle $\bar{Z}_n$ et $T_n^{*\,2}$ la moyenne empirique et la variance empirique corrigée de $Z_1, \ldots, Z_n$. Utiliser le théorème de Cochran pour établir l'indépendance entre $\bar{Z}_n$ et $T_n^{*\,2}$ ainsi que les distributions

$$(n-1)\, T_n^{*\,2} \sim \chi^2(n-1) \qquad \text{et} \qquad \sqrt{n}\, \frac{\bar{Z}_n}{\sqrt{T_n^{*\,2}}} \sim t(n-1).$$

2) On considère maintenant l'échantillon $X_1, \ldots, X_n$.

a) Exprimer $\bar{X}_n$ en fonction de $\bar{Z}_n$ et $S_n^{*\,2}$ en fonction de $T_n^{*\,2}$. En déduire que $\bar{X}_n$ et $S_n^{*\,2}$ sont elles aussi indépendantes.

b) En déduire que

$$\frac{(n-1)\, S_n^{*\,2}}{\sigma^2} \sim \chi^2(n-1) \qquad \text{et que} \qquad \sqrt{n}\, \frac{\bar{X}_n - \mu}{\sqrt{S_n^{*\,2}}} \sim t(n-1).$$

c) Retrouver les intervalles de confiance de sécurité $1-\alpha$ pour les paramètres μ et σ^2 donnés en cours à partir de ces distributions.

d) Construire les mêmes tests qu'à la question 2 de l'exercice ~~précédent~~ 15. Sont-ils plus ou moins précis ? Dans quel contexte peut-on raisonnablement considérer qu'ils sont équivalents ?

1) $F = \mathrm{Vect}\{U\}$ avec $U = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$ d'où $F = \{\alpha U,\ \alpha \in \mathbb{R}\}$

Alors $d = \dim F = 1$.

$P_F = U(U^T U)^{-1} U^T = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (n)^{-1} (1 \ldots 1) = \frac{1}{n} \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix}$

Donc $P_F \underline{Z} = \frac{1}{n} \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} \bar{Z}_n \\ \vdots \\ \bar{Z}_n \end{pmatrix} \Rightarrow \|P_F \underline{Z}\|^2 = n(\bar{Z}_n)^2$

On a $F^\perp = \{V \in \mathbb{R}^n,\ U^T V = 0\}$ $\quad \dim F = n-1$

$P_{F^\perp} = I_n - P_F = \begin{pmatrix} 1-\frac{1}{n} & & -\frac{1}{n} \\ & \ddots & \\ -\frac{1}{n} & & 1-\frac{1}{n} \end{pmatrix}$ Donc $P_{F^\perp} \underline{Z} = \begin{pmatrix} Z_1 - \bar{Z}_n \\ \vdots \\ Z_n - \bar{Z}_n \end{pmatrix}$

$\Rightarrow \|P_{F^\perp} \underline{Z}\|^2 = \sum_{i=1}^n (Z_i - \bar{Z}_n)^2 = (n-1) \cdot T_n^{*2}$

Par Cochran : • $n(\bar{Z}_n)^2 \perp\!\!\!\perp (n-1) T_n^{*2} \Rightarrow \bar{Z}_n \perp\!\!\!\perp T_n^{*2}$

• Et $n(\bar{Z}_n)^2 \sim \chi^2(1)$ et $(n-1) T_n^{*2} \sim \chi^2(n-1)$

Remarque : $n(\bar{Z}_n)^2 = (\sqrt{n}\, \bar{Z}_n)^2 = \left(\sqrt{n}\, \frac{\bar{Z}_n - 0}{1}\right)^2 \sim (\mathcal{N}(0,1))^2 \sim \chi^2(1)$.

sans Cochran.

voir p.41.

ce n'est pas le TCL.

Voir Formulaire : loi de Student $\frac{N}{\sqrt{Z_n/n}} \sim t(n)$ où $N \sim \mathcal{N}(0,1)$, $Z_n \sim \chi^2(n)$ et $N \perp\!\!\!\perp Z_n$

Study Title ______ Date ______

D'où $\sqrt{n}\,\dfrac{\bar{Z}_n}{\sqrt{T_n^{*2}}} = \dfrac{\sqrt{n}\,\bar{Z}_n}{\sqrt{\dfrac{(n-1)T_n^{*2}}{n-1}}} \sim t(n-1)$

$\sqrt{n}\,\bar{Z}_n \sim \mathcal{N}(0,1)$ voir plus haut

$\dfrac{(n-1)T_n^{*2}}{n-1} \sim \dfrac{\chi^2(n-1)}{n-1}$

$\perp\!\!\!\perp$

2) a) $\forall\, 1 \leq i \leq n, \quad X_i = \mu + \sigma Z_i$

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n}(\mu + \sigma Z_i) = \mu + \sigma \bar{Z}_n$

$S_n^{*2} = \dfrac{\|\underline{X} - \bar{X}_n \mathbb{U}\|^2}{n-1} = \dfrac{\|\mu\mathbb{U} + \sigma\underline{Z} - (\mu + \sigma\bar{Z}_n)\mathbb{U}\|^2}{n-1} = \dfrac{\|\sigma(\underline{Z} - \bar{Z}_n\mathbb{U})\|^2}{n-1} = \sigma^2 T_n^{*2}$

b) $\dfrac{(n-1)S_n^{*2}}{\sigma^2} = \dfrac{(n-1)\sigma^2 T_n^{*2}}{\sigma^2} = (n-1)T_n^{*2} \sim \chi^2(n-1).$

$\sqrt{n}\,\dfrac{\bar{X}_n - \mu}{\sqrt{S_n^{*2}}} = \sqrt{n}\,\dfrac{\cancel{\mu} + \cancel{\sigma}\bar{Z}_n - \cancel{\mu}}{\sqrt{\cancel{\sigma^2}T_n^{*2}}} = \sqrt{n}\,\dfrac{\bar{Z}_n}{\sqrt{T_n^{*2}}} \sim t(n-1).$

d) $T = \mathbb{1}\left\{\left|\sqrt{n}\,\dfrac{\bar{X}_n}{\sqrt{S_n^{*2}}}\right| > t_{1-\frac{\alpha}{2}}\right\}$ ("$\mu_0 = 0$").

$t_{1-\frac{\alpha}{2}}$: $t(n-1)$.

$T = 1 - \mathbb{1}\left\{z_{\frac{\alpha}{2}} \leq (n-1)S_n^{*2} \leq z_{1-\frac{\alpha}{2}}\right\}$

$z_{\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}$: $\chi^2(n-1)$

Ces tests sont plus précis car c'est $\forall n$ tandis que pour l'exercice 15 c'est asymptotiques
IQs sont équivalents si n grand.

Study Title ____________ Date ______

Exercice 18. Dans une population biologique diploïde, on observe l'expression d'un gène pour lequel l'allèle A est dominant et l'allèle a est récessif (ce qui signifie que (A, A), (A, a) et (a, A) expriment un caractère C_A tandis que (a, a) ne l'exprime pas). Les lois de Mendel garantissent que, en croisant les individus à partir d'une population initiale équitablement composée de (A, A) et de (a, a), on aboutit à un équilibre où C_A s'exprime dans 75 % des cas. Dans notre population de $n = 200$ individus, on observe C_A à 165 reprises.

1) Peut-on considérer que le gène de notre population satisfait l'équilibre de Mendel, au risque de 5% ? Au risque de 1% ? (Indications : $u_{0.975} \approx 1.96$ et $u_{0.995} \approx 2.58$).

2) Le caractère C_A s'exprime-t-il trop fréquemment pour satisfaire l'équilibre, au risque de 5% ? Au risque de 1% ? (Indications : $u_{0.95} \approx 1.64$ et $u_{0.99} \approx 2.33$).

3) Exprimer la p-valeur de ces tests à l'aide de la fonction de répartition de la loi $\mathcal{N}(0, 1)$.

1) $\underline{X} = (C_A^{(1)}, \dots, C_A^{(200)}) \Rightarrow h(\underline{X}) = \sum_{i=1}^{200} C_A^{(i)} = 165$

On vérifie si $C_A \sim \mathcal{B}(\frac{3}{4})$ $\quad \mathcal{H}_0 :$ " $p = \frac{3}{4}$ " $\quad \mathcal{H}_1 :$ " $p \neq \frac{3}{4}$ "

$\overline{X}_n = \frac{165}{200} = 82{,}5\%$, $Z = \sqrt{200}\, \frac{\overline{X}_n - 75\%}{\sqrt{\frac{3}{4} \times \frac{1}{4}}} = \sqrt{200} \times \frac{4 \times 7{,}5\%}{\sqrt{3}} \approx 2{,}45$

Si $\mathcal{H}_0$ est vraie : $Z = \sqrt{n}\, \frac{\overline{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$. (TCL).

$\underline{\alpha = 5\%}$ Si $Z \in [\pm u_{0{,}975}]$: non rejet de $\mathcal{H}_0$

or $Z \notin [-1{,}96 , 1{,}96]$ donc rejet de $\mathcal{H}_0$ au profit de $\mathcal{H}_1$

$\underline{\alpha = 1\%}$ Si $Z \in [\pm u_{0{,}995}]$: non rejet de $\mathcal{H}_0$

or $Z \in [-2.58 ; 2.58]$ donc non rejet de $\mathcal{H}_0$.

2) Si on regarde $\overline{X}_n = 0{,}825 > 0{,}75$. $\quad \mathcal{H}_0 :$ " $p = \frac{3}{4}$ " $\quad \mathcal{H}_1 :$ " $p > \frac{3}{4}$ "

$T = \mathbb{1}_{\{Z > u_{1-\alpha}\}}$ avec $\alpha = 5\%$ ou $\alpha = 1\%$

Or $T = 1$ dans les deux cas donc on rejette $\mathcal{H}_0$ dans les deux cas.

3) $p\text{-val} = 2(1 - F_0(Z)) \approx 1{,}24\%$ $\quad F_0$ fonction de répartition de la loi gaussien

$p\text{-val} = 1 - F_0(Z) \approx 0{,}62\%$

Study Title ______________________ Date ____________

Exercice 19. On s'intéresse au revenu mensuel moyen d'un panel de $n = 800$ individus selon la répartition hommes/femmes. On observe les effectifs suivants [1] où R désigne le revenu en euros.

	C_1 $R < 1200$	C_2 $1200 \leqslant R < 1800$	C_3 $1800 \leqslant R < 2500$	C_4 $R \geqslant 2500$	
Hommes	20	135	145	80	380
Femmes	75	280	50	15	420
	95	415	195	95	800

On donne également quelques quantiles (approximatifs) d'ordre β de la loi du khi-deux à m degrés de liberté.

$z_\beta(m)$	$\beta = 0.01$	$\beta = 0.025$	$\beta = 0.05$	$\beta = 0.1$	$\beta = 0.9$	$\beta = 0.95$	$\beta = 0.975$	$\beta = 0.99$
$m = 3$	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.3
$m = 4$	0.297	0.484	0.711	1.06	7.78	9.49	11.1	13.3
$m = 6$	0.872	1.24	1.64	2.20	10.6	12.6	14.4	16.8
$m = 8$	1.65	2.18	2.73	3.49	13.4	15.5	17.5	20.1

1) On s'intéresse dans un premier temps à l'existence d'un lien significatif entre le sexe d'un individu et son revenu moyen. La statistique du test du khi-deux d'indépendance vaut sur ces données $D_n^2 \approx 171.7$. Effectuer le test au niveau de risque de 1%. Quelle est votre conclusion ? Δ_n ?

2) Dans un second temps, au vu des effectifs totaux en colonnes, on propose le modèle suivant : si l'on choisit aléatoirement un individu de la population, la probabilité qu'il appartienne à la classe C_i vaut p_i avec

$$p_1 = \frac{1}{8}, \quad p_2 = \frac{1}{2}, \quad p_3 = \frac{1}{4} \quad \text{et} \quad p_4 = \frac{1}{8}.$$

On effectue le test du khi-deux d'adéquation avec le logiciel R. Les sorties sont indiquées ci-dessous.

```
> Eff = c(95, 415, 195, 95)
> chisq.test(Eff, p = c(1/8, 1/2, 1/4, 1/8))

        Chi-squared test for given probabilities

data:  Eff
X-squared = 1.1875, df = 3, p-value = [?]

> pchisq(1.1875, c(3, 4, 6, 8))
[1] 0.243996137 0.119847658 0.022502377 0.003236124
```

$\rightarrow D^2$?

a) Retrouver par le calcul la sortie `X-squared = 1.1875`.

b) Compléter la valeur manquante [?]. Quelle est votre conclusion au niveau de risque 5% ?

1) $D_n^2 \xrightarrow{\mathcal{L}} \chi^2((r-1)(s-1)) = \chi^2(3)$ $\quad r = 2$ et $s = 4$.

or $\beta = 0,99$ pour un risque de 1% et $n = 3$

or $T = \mathbb{1}_{\{D_n^2 > 0,99\}} = 1$ donc on rejette $\mathcal{H}_0$: " $X \perp\!\!\!\perp Y$ "
(et on se doute que p-val. très petit).

2) a) D^2 p44. test d'adéquation à recalculer.

b) p-val = $1 - 0,24 = 0,76 \gg 0.05$ On ne peut donc rejeter $\mathcal{H}_0$

Study Title ____________________ Date __________

Exercice 20. Les populations gaussiennes offrent de nombreuses facilités de traitement statistique. Pour deux n-échantillons $X_1, \dots, X_n$ et $Y_1, \dots, Y_n$ supposés gaussiens et indépendants, on cherche à savoir si $\mu_X = \mu_Y$ et si $\sigma_X^2 = \sigma_Y^2$.

1) On s'intéresse dans un premier temps à un test d'égalité des variances. Montrer que

$$\frac{\sigma_Y^2 \, S_{n,X}^{*\,2}}{\sigma_X^2 \, S_{n,Y}^{*\,2}} \sim F(n-1, n-1).$$

$I = \left[\frac{S_y^2}{S_x^2} f_{\frac{\alpha}{2}} ; \frac{S_y^2}{S_x^2} f_{1-\frac{\alpha}{2}} \right]$

En déduire un IC pour $\frac{\sigma_Y^2}{\sigma_X^2}$ puis un test de $\mathcal{H}_0$: "$\sigma_X^2 = \sigma_Y^2$" contre $\mathcal{H}_1$: "$\sigma_X^2 \neq \sigma_Y^2$".

2) Dans un second temps, en supposant les variances égales, on veut tester l'égalité des espérances. On pose $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. On montre, par un calcul fastidieux et le théorème de Cochran, que

$$\frac{(\bar{X}_n - \bar{Y}_n) - (\mu_X - \mu_Y)}{H_n} \sim t(2(n-1)) \quad \text{où} \quad H_n = \sqrt{\frac{S_{n,X}^{*\,2} + S_{n,Y}^{*\,2}}{n}}.$$

$J = \left[(\bar{X} - \bar{Y}) \pm H t_{1-\frac{\alpha}{2}} \right]$

En déduire un IC pour $\mu_X - \mu_Y$ puis un test de $\mathcal{H}_0$: "$\mu_X = \mu_Y$" contre $\mathcal{H}_1$: "$\mu_X \neq \mu_Y$".

3) Déduire de ce qui précède une méthodologie statistique pour décider si deux populations gaussiennes indépendantes et de même taille peuvent être considérées comme homogènes (dans le sens issues d'un même modèle). Test d'égalité des variances $T = \mathbb{1}_{\{1 \notin I\}}$, Test d'égalité des espérances $T = \mathbb{1}_{\{0 \notin J\}}$

4) Une équipe d'ornithologues souhaite évaluer l'envergure moyenne d'une certaine espèce d'oiseaux peuplant une réserve. Leur hypothèse de départ est que, en vertu du grand nombre d'individus, il paraît judicieux de considérer que l'envergure de chaque oiseau est issue une variable aléatoire gaussienne. La méthodologie est la suivante :
— Un premier échantillon de $n = 100$ individus est capturé et relâché avec un marquage. On note $X_1, \dots, X_n$ les mesures effectuées, supposées suivre la loi $\mathcal{N}(\mu_X, \sigma_X^2)$.
— Un second échantillon de $n = 100$ individus non marqués, afin de garantir l'indépendance entre les échantillons, est capturé. On note $Y_1, \dots, Y_n$ les mesures effectuées, supposées suivre la loi $\mathcal{N}(\mu_Y, \sigma_Y^2)$.

Les mesures donnent

$H_n = \sqrt{\frac{4{,}006}{100}}$ $\quad I = [1{,}104 \times 0{,}67 , 1{,}104 \times 1{,}49]$

$$\frac{S_{n,Y}^{*\,2}}{S_{n,X}^{*\,2}} \approx 1.104, \qquad S_{n,X}^{*\,2} + S_{n,Y}^{*\,2} \approx 4.006 \quad \text{et} \quad \bar{X}_n - \bar{Y}_n \approx 0.322.$$

$J = \left[0{,}322 \pm \sqrt{\frac{4{,}006}{100}} \times 1{,}97 \right]$

Tester s'il est raisonnable, au risque de 5%, de considérer que tous les oiseaux sont bien issus d'une seule population gaussienne. (Indications : $f_{0.025}(99, 99) \approx 0.67$, $f_{0.975}(99, 99) \approx 1.49$ et $t_{0.975}(198) \approx 1.97$).

1\. Par définition, on dit que $Z \sim F(n_1, n_2)$ si $Z = \frac{S_1 / d_1}{S_2 / d_2}$

où $S_i \sim \chi^2(d_i)$ indépendants.

D'après la proposition 3.1 on a que $\frac{(n-1) S_n^{*2}}{\sigma^2} \sim \chi^2(n-1)$

Donc $\begin{cases} \frac{(n-1) S_{n,x}^{*2}}{\sigma_x^2} \sim \chi^2(n-1) \\ \frac{(n-1) S_{n,y}^{*}}{\sigma_y^2} \sim \chi^2(n-1) \end{cases}$ $\perp\!\!\!\perp$ car les échantillon sont $\perp\!\!\!\perp$

$\Rightarrow \frac{\sigma_y^2 S_{n,x}^{*2}}{\sigma_x^2 S_{n,y}^{*2}} \sim F((n-1), (n-1))$.

Sous H_0 : $\sigma_x = \sigma_y$ $\Rightarrow$ $\frac{S_{n,x}^{*2}}{S_{n,y}^{*2}} \sim F((n-1), (n-1))$

On peut donc construire un test bilatéral avec

$T = 1 - \mathbb{1}_{\{f_{\frac{\alpha}{2}} < \frac{S_{n,X}^{*2}}{S_{n,Y}^{*2}} < f_{1-\frac{\alpha}{2}}\}}$ où f_β est le quantile de seuil β pour $F((n-1, n-1))$

2) De nouveau sous H_0 on a $\frac{\bar{X}_n - \bar{Y}_n}{H_n} \sim t(2(n-1))$

Sachant que $t(2(n-1))$ est symétrique, on peut construire un test bilatéral de risque α avec

$$S = 1 - \mathbb{1}_{\{-t_{1-\frac{\alpha}{2}} < \frac{\bar{X}_n - \bar{Y}_n}{H_n} < t_{1-\frac{\alpha}{2}}\}}$$

3) On peut vérifier dans un premier temps que les variances sont égales. Si c'est le cas, on peut alors vérifier l'égalité des moyennes.

Dans les cas contraires H_0 : " $\sigma_X = \sigma_Y$ et $\mu_X = \mu_Y$ "

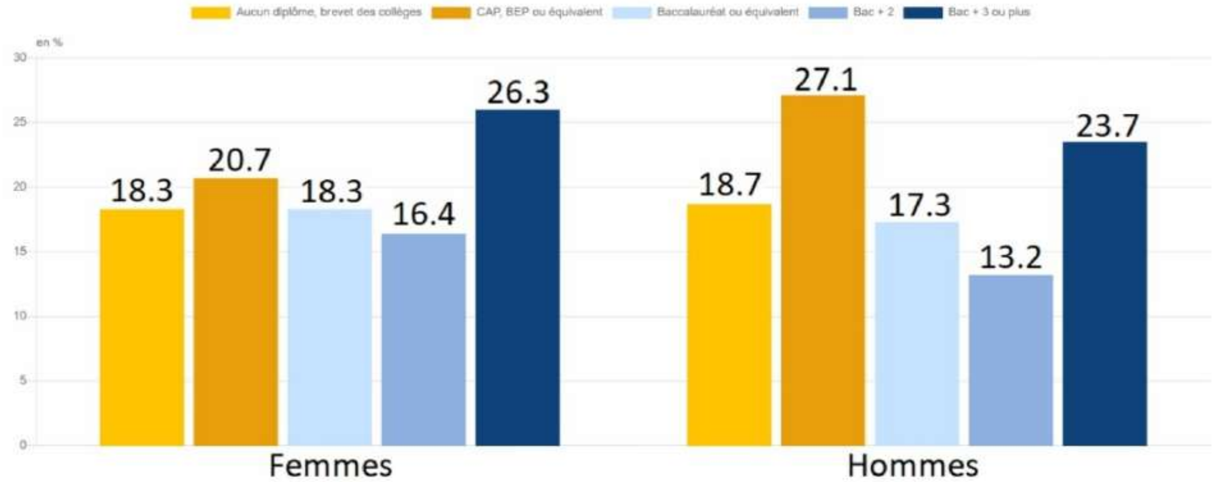
$$R = \mathbb{1}_{\{T=1 \text{ ou } S=1\}} = 1 - \mathbb{1}_{\{T=0 \text{ et } S=0\}}$$

4) $I = [0{,}74 ; 1.64]$ $\quad J = [-0{,}07 ; 0{,}71]$

$1 \in I$ donc on ne rejette pas l'égalité des variances
$0 \in J$ donc on ne rejette pas l'égalité des espérances.

Il semblerait qu'on ait affaire à 2 populations homogènes.

Exercice 21. On peut trouver sur le site de l'INSEE le schéma suivant, relatif aux chiffres correspondant au diplôme le plus élevé selon le sexe en 2020 en France (mesurés en % de la population). Les ordonnées approximatives ont été ajoutées sur les bâtons pour une meilleure lisibilité.



Diriez-vous que ces chiffres traduisent une égalité hommes/femmes face aux diplômes obtenus, ou au contraire que ces chiffres dénotent un déséquilibre hommes/femmes ? Vous décrirez précisément chaque étape de votre démarche, en adoptant un niveau de risque de 5% dans votre conclusion. Les 2 premiers termes du calcul de la statistique de test sont suffisants, ensuite on admettra pour gagner du temps qu'elle vaut ≈ 1.37.

Quelques quantiles : $z_{0.025}(4) \approx 0.48$, $z_{0.05}(4) \approx 0.71$, $z_{0.95}(4) \approx 9.49$, $z_{0.975}(4) \approx 11.1$, $z_{0.025}(5) \approx 0.83$, $z_{0.05}(5) \approx 1.15$, $z_{0.95}(5) \approx 11.1$, $z_{0.975}(5) \approx 12.8$, $z_{0.025}(8) \approx 2.18$, $z_{0.05}(8) \approx 2.73$, $z_{0.95}(8) \approx 15.5$, $z_{0.975}(8) \approx 17.5$, $z_{0.025}(10) \approx 3.25$, $z_{0.05}(10) \approx 3.94$, $z_{0.95}(10) \approx 18.3$, $z_{0.975}(10) \approx 20.5$.

$z_{1-\alpha}$
$z_{1-0,05}$

	Aucun diplôme, Brevet des collèges	CAP, BEP ou équivalent	BAC ou Équivalent	BAC+2	> BAC+3	
Femmes	18,3	20,7	18.3	16.4	26,3	100
Hommes	18,7	27,1	17,3	13.2	23.7	100
	37	47.8	35.6	29.6	50	200

On va mener un test d'indépendance par la méthode de Khi-deux
H_0 : " $X \perp\!\!\!\perp Y$ " vs $H_1 = \overline{H_0}$

$$D^2 = \frac{\left(18,3 - \frac{37 \times 100}{200}\right)^2}{\frac{37 \times 100}{200}} + \frac{\left(27,1 - \frac{47,8 \times 100}{200}\right)^2}{\frac{47,8 \times 100}{200}} + \dots \approx 1,37.$$

Le test s'écrit $T = \mathbb{1}_{\{D^2 > z_{0,95}\}} = \mathbb{1}_{\{D^2 > 9,49\}} = 0$.

quantile de $\chi^2((2-1)\times(5-1))$

Ces données ne semblent pas remettre en question l'indépendance entre hommes / femmes et niveau d'études (risque 5%)

Study Title ______________________ Date ____________

Exercice 22. Pour modéliser une intention de vote au second tour d'une élection entre deux candidats A et B, on pose $X = 1$ pour le candidat A et $X = 0$ pour le candidat B. Ainsi $X \sim \mathcal{B}(p)$ et $p = \mathbb{P}(X = 1)$ est la probabilité qu'un votant choisisse le candidat A. Le sondage est réalisé sur un panel indépendant de n personnes dont les intentions sont notées $X_1, \ldots, X_n$.

1) Un institut de sondage crédite le candidat A de 52 % d'intentions de vote sur la base d'un panel de $n = 100$ participants. Justifier que ce chiffre ne donne pas un avantage significatif au candidat A sur son concurrent, au risque de 5%.

2) À partir de combien de personnes interrogées dans le panel le score de 52 % d'intentions de vote donne-t-il un avantage significatif au candidat A, toujours au risque de 5% ?

Quelques quantiles : $u_{0.95} \approx 1.64$, $u_{0.975} \approx 1.96$, $u_{0.99} \approx 2.33$, $u_{0.995} \approx 2.58$.

Exo 1. 1) $X \sim \mathcal{B}(p)$ et donc $\mathbb{E}[X] = p$ et $\mathrm{Var}(X) = p(1-p)$.

C'est un n-échantillon, donc par le TCL,

$$\sqrt{n}\,\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Mais par la LFGN, $\bar{X}_n \xrightarrow{p.s.} p$ et par le CMT, $\sqrt{\bar{X}_n(1-\bar{X}_n)} \xrightarrow{p.s.} \sqrt{p(1-p)}$. Il reste à appliquer Slutsky,

$$\sqrt{n}\,\frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} = \underbrace{\sqrt{n}\,\frac{\bar{X}_n - p}{\sqrt{p(1-p)}}}_{\xrightarrow{\mathcal{L}} \mathcal{N}(0,1)} \times \underbrace{\frac{\sqrt{p(1-p)}}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}}_{\xrightarrow{p.s.} 1} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

On en déduit immédiatement $ICA_{1-\alpha}(p) = \left[\bar{X}_n \pm \frac{\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n}}\, u_{1-\frac{\alpha}{2}}\right]$.

2) Ici on a (au moins) 2 façons de faire. On peut tester :

- H_0 : "$p = \frac{1}{2}$" contre H_1 : "$p \neq \frac{1}{2}$". (1)

  aucun candidat avantagé — A ou B est avantagé

  H_0 hypothèse nulle

- H_0 : "$p = \frac{1}{2}$" contre H_1 : "$p > \frac{1}{2}$" (2) plus logique

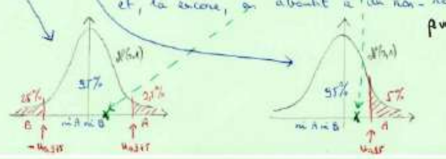
  A est avantagé

Comme le sondage donne 52% à A, la seconde approche paraît mieux adaptée mais les 2 raisonnements étaient acceptés.

Test (1) : On trouve $ICA_{0,95}(p) \simeq [0,422 \,;\, 0,618]$ et $0,5 \in ICA_{0,95}(p)$ d'où un non-rejet de H_0 (au risque 5%). 52% est dans la marge d'erreur, A n'est pas avantagé. (On regarde si $\sqrt{n}\,\frac{\bar{X}_n - \frac{1}{2}}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}$ tombe dans $[\pm u_{0,975} \simeq 1,96]$).

Test (2). Cette fois le test est unilatéral (à droite), on compare donc la statistique de test $\sqrt{n}\,\frac{\bar{X}_n - \frac{1}{2}}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \simeq 0,400$ avec le quantile $u_{0,95} \simeq 1,64$ et, là encore, on aboutit à un non-rejet de H_0.

p-value $= \mathbb{P}(\mathcal{N}(0,1) > 0,400)$
$= 1 - F_{\mathcal{N}(0,1)}(0,4)$
$= 0,34 > \alpha = 0,05$
on ne rejette pas H_0.



3) Là encore je propose les 2 solutions.

Test (1) : on cherche le plus petit entier n tel que $\bar{X}_n - \frac{\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n}}\, u_{0,975} > 0,5$ (borne gauche de l'ICA). Alors, $0,52 - \frac{\sqrt{0,52 \times 0,48}}{\sqrt{n}}\, 1,96 > 0,5$

$$\Leftrightarrow n > \frac{0,52 \times 0,48 \times (1,96)^2}{(0,02)^2} \simeq 2397,2, \text{ d'où } \underline{n = 2398}.$$

Il faudrait donc au moins 2398 sondés pour que 52 % avantage statistiquement A (au risque 5 %), avec le protocole (1).

Test (2) : on cherche le plus petit entier n tel que $\sqrt{n}\,\frac{\bar{X}_n - \frac{1}{2}}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} > u_{0,95}$ (borne gauche de la zone de rejet). Alors, $\sqrt{n} \times \frac{0,52 - \frac{1}{2}}{\sqrt{0,52 \times 0,48}} > 1,64$

$$\Leftrightarrow n > \frac{(1,64)^2 \times 0,52 \times 0,48}{(0,02)^2} \simeq 1678,3, \text{ d'où } \underline{n = 1679}.$$

Il faudrait donc au moins 1679 sondés pour que 52% avantage statistiquement A (au risque 5 %), avec le protocole (2).

Exercice 23. On exécute ci-dessous des lignes de script R qui simulent et testent des données. Interpréter et commenter les sorties.

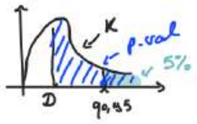
$\mathcal{H}_0 \to \underline{X} \sim \mathcal{N}(\mu, \sigma^2)$.

```
> n = 1000
> X = rnorm(n, 1, 2)
> stats::ks.test(X, "pnorm", 1, 2)

        One-sample Kolmogorov-Smirnov test

data:  X
D = 0.021049, p-value = 0.7674
alternative hypothesis: two-sided
```

espérance ; écart-type ; n-échantillon, $(X_1, \dots, X_n)$ iid $\sim \mathcal{N}(1,4)$

librairie ; fct de répartition de la loi normale

Le test d'adéquation



donc non rejet de $\mathcal{H}_0$ (à 5%)

```
> stats::shapiro.test(X)

        Shapiro-Wilk normality test

data:  X
W = 0.9987, p-value = 0.6868
```

Le test de normalité (c'est gaussien, oui/non).

68% >> 5% → on ne rejette pas $\mathcal{H}_0$

```
> Y = 2*X+1
> stats::ks.test(X, Y)
```

$Y \sim \mathcal{N}(3,16)$ $\quad X \sim \mathcal{N}(1,4)$.

```
        Two-sample Kolmogorov-Smirnov test

data:  X and Y
D = 0.362, p-value < 2.2e-16
alternative hypothesis: two-sided
```

est-ce que ces 2 échantillons ont la même loi ? $F_X = F_Y$? $X \sim Y$?

≈ 0 ⇒ rejet certain de $\mathcal{H}_0$

```
> Fum = c(21, 35, 11, 17)
> Tot = c(100, 150, 90, 100)
> stats::prop.test(Fum, Tot)

        4-sample test for equality of proportions without continuity correction

data:  Fum out of Tot
X-squared = 5.0158, df = 3, p-value = 0.1706
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3    prop 4
0.2100000 0.2333333 0.1222222 0.1700000
```

→ Fumeur

→ tailles des 4 groupes

1er groupe : $\frac{21}{100}$ personnes qui fument

Le test de proportion.

$\mathcal{H}_0$: " $p_1 = p_2 = p_3 = p_4$ "

$\mathcal{H}_1$: contraire de $\mathcal{H}_0$.

→ non rejet de $\mathcal{H}_0$ au risque de 5%

Remarque : p-value inattendue due à n pas suffisamment grand (refaire avec ×10. p-val $\approx 10^{-11}$ → rejet certain de $\mathcal{H}_0$).

$p_1 \neq p_2 \neq p_3 \neq p_4$

```
> L = c(11, 10, 10, 10, 11, 12, 11, 11, 10, 9, 11, 8, 7, 8, 11)
> T = c(12, 13, 15, 9, 17, 15, 15, 16, 20, 11, 12, 10, 16, 15, 13)
> stats::wilcox.test(L, T, paired = TRUE, alternative = "less")

        Wilcoxon signed rank test with continuity correction

data:  L and T
V = 2, p-value = 0.0005332
alternative hypothesis: true location shift is less than 0
```

couple

$L < T$ (lièvre court plus vite que la tortue)

< 0,01 → rejet de $\mathcal{H}_0$: " L = T " au risque de 1%.

```
> X1 = rnorm(n)
> X2 = rnorm(n/2)
> X3 = rnorm(n/4)
> Grp = as.factor(c(rep(1, n), rep(2, n/2), rep(3, n/4)))
> car::leveneTest(c(X1, X2, X3), Grp)

        Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    2  0.5099 0.6007
      1747
```

$\sim \mathcal{N}(0,1)$; $\sim \mathcal{N}(0,1)$; $\sim \mathcal{N}(0,1)$ } de tailles ≠.

rep$(2, \frac{n}{2})$: $\underbrace{2 \dots 2}_{\frac{n}{2}}$

$Var(X_1) = Var(X_2) = Var(X_3)$?

non rejet de $\mathcal{H}_0$

```
> E = rnorm(n)
```

~ iid

```
> X = rep(0, n)
> Y = rep(0, n)
> for (i in 2:n){
        X[i] = E[i-1] + E[i]
  }

> stats::cor.test(E[1:(n-1)], E[2:n], method="pearson")
```

test de corrélation de pearson.

```
        Pearson's product-moment correlation

data:  E[1:(n - 1)] and E[2:n]
t = -1.1544, df = 997, p-value = 0.2486
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.09833808  0.02554519
sample estimates:
        cor
-0.03653681
```

→ non rejet de H_0 (à 5%) : $E[1:(n-1)] = F_1$ et $F_2 = E[2:n]$ "non correlées" (ou H_0 : $Corr(F_1, F_2) = 0$)

```
> stats::cor.test(X[1:(n-1)], X[2:n], method="pearson")

        Pearson's product-moment correlation

data:  X[1:(n - 1)] and X[2:n]
t = 16.855, df = 997, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4211779 0.5178012
sample estimates:
      cor
0.4709006
```

→ rejet certain de H_0 → Les X sont correlées

```
> stats::Box.test(E, lag=5)
```

"bruit blanc"

"décorelées"

→ on tient compte de 5 valeurs "en m̄ temps"

```
        Box-Pierce test

data:  E
X-squared = 4.8897, df = 5, p-value = 0.4295
```

non rejet de H_0.

Exercice 24. On considère la fonction de répartition empirique d'un n-échantillon $X_1, \ldots, X_n$ donnée par

$$\forall x \in \mathbb{R}, \qquad \widehat{F}_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{X_k \leqslant x\}}.$$

On veut montrer qu'elle possède de très bonnes propriétés d'estimation fonctionnelle pour la vraie répartition F_X de l'échantillon. Cela justifie en particulier le succès du test de Kolmogorov-Smirnov.

1) Montrer que, ponctuellement (c'est-à-dire, à x fixé), l'estimation est sans biais, fortement consistante et asymptotiquement normale.

2) Le *théorème de Glivenko-Cantelli* stipule que

$$\|\widehat{F}_n - F_X\|_\infty \xrightarrow{\text{p.s.}} 0.$$

Pour simplifier, on se place dans le cas continu pour démontrer ce résultat. Soit la discrétisation

$$-\infty = x_0 < x_1 < \ldots < x_{d-1} < x_d = +\infty$$

où les abscisses sont choisies de sorte que $F_X(x_i) - F_X(x_{i-1}) = \frac{1}{d}$ pour $i = 1, \ldots, d$. Pour tout $x \in \mathbb{R}$, il existe donc $j \in \{1, \ldots, d\}$ tel que $x \in [x_{j-1}, x_j]$.

a) Montrer que

$$\widehat{F}_n(x) - F(x) \leqslant \widehat{F}_n(x_j) - F(x_j) + \frac{1}{d}.$$

b) Montrer de même que

$$\widehat{F}_n(x) - F(x) \geqslant \widehat{F}_n(x_{j-1}) - F(x_{j-1}) - \frac{1}{d}.$$

c) En déduire que

$$\|\widehat{F}_n - F_X\|_\infty \leqslant \max_{i=1,\ldots,d} |\widehat{F}_n(x_i) - F(x_i)| + \frac{1}{d}$$

puis conclure.

```
X=c(18.7,15.9,11.9,10.9,23.0,16.6,17.4,17.9,10.8,13.5)
```

Study Title Date

Exercice 25. Le but de cet exercice est de bien saisir le test de Kolmogorov-Smirnov, probablement un des plus utilisés en statistique. Un générateur de nombres aléatoires nous fournit le résultat suivant lorsqu'on lui demande 10 variables gaussiennes $\mathcal{N}(16, 4)$.

18.7	15.9	11.9	10.9	23.0	16.6	17.4	17.9	10.8	13.5

Effectuer un test de Kolmogorov-Smirnov sur le jeu de données au risque de 5 % en utilisant R pour obtenir les valeurs théoriques, et critiquer le résultat. On utilisera comme quantile $\approx 0.410\sqrt{10}$.

```
#H0 = X suit la loi normale(16,4)
#ecdf = fonction de répartition empirique
#ecdf(X)
#plot(ecdf(X))
#curve(pnorm(x,16,2),col="red",lty=2,add=T)
#F=pnorm(X,16,2)
#points(X,F,col="red")
X=sort(X)
F=sort(F)
lap=(1:10)/10
lav=(0:9)/10
d=max(abs(F-lap),abs(F-lav))
racine_n_fois_delta=sqrt(10)*d
quantile=0.410*sqrt(10)
#on a pas sqrt(n )*delta > quantile donc T=0
#on ne rejette pas H0, au risuqe de 5%, les données ont pu
être générées selon la loi normale (16,4)

# mais attention !!! test asymptotique avec n=10 !!!


#manière plus immédiate :
#stats::ks.test(X,"pnorm",16,2)
```

Study Title ______________________ Date __________

Exercice 26. Une expérience classique en psychologie consiste à montrer à un panel des photos de visages, afin que chaque personne désigne un coupable. Alors que, sans connaissance préalable, le panel devrait logiquement sélectionner ses coupables de manière uniforme, l'expérience tend à montrer que l'aspect des visages influence fortement le choix. On suppose qu'il y a V visages présentés à un panel de n personnes et que l'un d'entre eux semble agressif. On appelle S_n le nombre de fois où le visage agressif a été désigné et π la probabilité qu'il a d'être sélectionné par chaque membre du panel, de sorte que $S_n \sim \mathcal{B}(n, \pi)$. On notera π_0 la valeur de π dans le cas idéal où le panel est équitable.

1) Soit $k \in \{0, \dots, n\}$. Déterminer $p_0(k) = \mathbb{P}_{\pi_0}(S_n = k)$ en fonction de k, n et V, sous l'hypothèse que le panel est parfaitement équitable (c'est-à-dire qu'il n'est pas influencé par des *a priori* visuels).

2) Supposons maintenant que π varie de manière uniforme dans $[0, 1]$, pour tenir compte de toutes les imperfections possibles du panel, et posons

$$\forall\, k \in \{0, \dots, n\}, \qquad p(k) = \mathbb{P}(S_n = k) = \int_0^1 \mathbb{P}_\pi(S_n = k)\, \mathrm{d}\pi.$$

Montrer que $p(k) = \frac{1}{n+1}$. On rappelle que la fonction bêta d'Euler est définie par

$$\beta(a, b) = \int_0^1 s^{a-1}\,(1-s)^{b-1}\,\mathrm{d}s$$

et qu'elle est liée à la fonction Gamma par la relation

$$\beta(a, b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}.$$

3) Le panel est constitué de $n = 40$ personnes et $V = 6$ visages sont présentés. Le visage agressif est choisi $k = 14$ fois. Donner l'expression exacte de $p_0(14)$ et de $p(14)$. Diriez-vous que le panel a été équitable ou qu'il a été influencé par des *a priori* ? (On admet que $p_0(14) \approx 0.00259$).

4) Reformuler ce problème sous l'angle de vue bayésien.

1. $p_0(k) = \binom{n}{k} \frac{1}{V^k}\left(1 - \frac{1}{V}\right)^{n-k}$

2. $p(k) = \int_0^1 \mathbb{P}_\pi(S_n = k)\, d\pi = \binom{n}{k} \int_0^1 \pi^k (1-\pi)^{n-k}\, d\pi = \binom{n}{k} \cdot \beta(k+1, n-k+1) = \binom{n}{k} \cdot \frac{\Gamma(k+1) \cdot \Gamma(n-k+1)}{\Gamma(n-k+1+k+1)}$

$= \frac{n!}{k!\,(n-k)!} \; \frac{k!\,(n-k)!}{(n+1)!} = \frac{n!}{(n+1)!} = \frac{1}{n+1}$

3. $p_0(14) = \binom{40}{14} \frac{1}{6^{14}} \left(\frac{5}{6}\right)^{26} \approx 0{,}00259.$

$p(14) = \frac{1}{41} \approx 0{,}02439 > p_0(14).$ $\rightarrow$ Il semble que le panel n'était pas équitable.

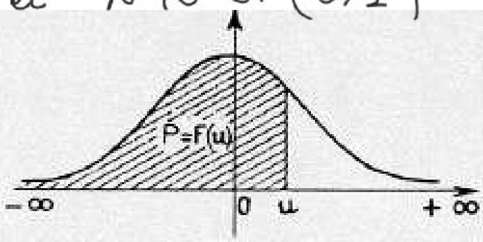
4. $S_n \sim B(n, \pi)$ et $\pi \sim \mathcal{U}_{[0,1]}$ a priori

$p(\pi | S_n) \propto p(S_n | \pi)\, p(\pi)$ a posteriori

$p(\pi | S_n = k) = \frac{\mathbb{P}_\pi(S_n = k)\, \mathbb{1}_{[0,1]}(\pi)}{\int_0^1 \mathbb{P}_\pi(S_n = k)\, d\pi}$

Fonction de répartition de la loi N(0, 1)

$\mathbb{P}(N \leq t) \quad t \geqslant 0$ et $N \sim \mathcal{N}(0,1)$



$u = x + y$

y \ x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
	$\mathbb{P}(N \leq x+y)$									
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7290	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Table pour les grandes valeurs de u

u	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
F(u)	0,99865	0,99904	0,99931	0,99952	0,99966	0,99976	0,999841	0,999928	0,999968	0,999997

Study Title ______________________ Date __________