Chapitre 3

Échantillonnage et estimation

3.1 Introduction

La **théorie de l'échantillonnage** étudie les liens entre une population et des échantillons de cette population. À partir d'informations relatives à la loi d'une variable X pour une population donnée, on en déduit le comportement d'échantillons aléatoires simples relatifs à cette variable.
Dans la pratique c'est le problème inverse qui se pose. En général on ne connaît pas la loi de X, on ne connaît pas tous ses paramètres et on souhaite obtenir des informations à partir de l'observation d'un échantillon. Ce problème fait partie de la **théorie de l'estimation**.

Souvent on s'intéresse à la valeur d'un paramètre bien précis de la loi de X, espérance, variance, proportion. Ce paramètre noté θ est appelé **paramètre d'intérêt**, c'est un nombre dont la valeur est inconnue. On cherche à évaluer ce nombre à partir de l'observation d'un échantillon. À partir des données de l'observation d'un échantillon, on détermine une valeur numérique $\hat{\theta}$ qu'on appelle **estimation ponctuelle du paramètre d'intérêt**.

On peut aussi définir un **intervalle de confiance** c'est-à-dire déterminer un intervalle $[\theta_1; \theta_2]$ qui contient la vraie valeur du paramètre θ inconnu avec une grande probabilité fixée à priori.

Exemple 3.1.1 On veut estimer l'espérance mathématique de la variable X, "note des étudiants à l'examen", vérifiant $X \rightsquigarrow \mathcal{N}(m, \sigma)$. On prélève 50 copies dans la population, on les corrige, on obtient 50 notes x_1, x_2,...,x_{50} et on détermine la moyenne de cet échantillon $\overline{x} = \dfrac{x_1 + x_2 + \ldots + x_{50}}{50}$, on obtient 9,1. Intuitivement on peut estimer m par 9,1. On dit que $9,1$ est une estimation ponctuelle de m. On remarque que si on avait pris un autre échantillon, l'estimation serait différente. On pourrait aussi conclure que la moyenne m appartiendrait à l'intervalle $[8,4; 9,8]$ avec une probabilité de 0,9 par exemple. L'intervalle $[8,4; 9,8]$ est alors un intervalle de confiance au risque d'erreur 0,1.

3.2 Estimation ponctuelle

3.2.1 Introduction

L'ensemble des hypothèses relatives au problème d'estimation de paramètre est appelé **modèle statistique**. Celui-ci comprend :

— des hypothèses relatives à la loi de la variable X, par exemple $X \rightsquigarrow \mathcal{N}(m, \sigma)$, m et σ étant inconnus, ou X suit une loi inconnue. Le paramètre θ doit être défini, par exemple $\theta = E(X)$, $\theta = \sigma(X)$, $\theta = p$. On écrira $X \rightsquigarrow l(x, \theta)$ où x est la réalisation de X.

— La méthode de construction de l'échantillon doit être précisée, échantillon aléatoire simple par exemple. On n'utilisera dans ce cours que des échantillons aléatoires simples.

Rappel sur le choix d'un échantillon : Les échantillons étudiés sont tous aléatoires, le hasard intervient dans le choix de leurs éléments. Cependant deux procédures sont possibles pour construire un échantillon aléatoire :

— échantillon non exhaustif : pour construire un échantillon de taille n, on procède par n tirages au hasard avec remise (remise de l'individu dans la population après chaque tirage),

— échantillon exhaustif : pour construire un échantillon de taille n, on procède par n tirages au hasard sans remise ou par le tirage simultané de n individus.

Si la population est très grande, on peut considérer un échantillon exhaustif comme non exhaustif.

Rappel sur les échantillons aléatoire simples : On considère l'exemple suivant.

Exemple 3.2.1 Considérons un économiste chargé de réaliser un étude de marché pour une entreprise qui souhaite lancer une nouvelle marque de fromage. Il commence par analyser la consommation de fromage en France. Il doit réaliser un sondage et demander aux personnes interrogées combien de fois elles ont consommé de fromage la semaine dernière. La consommation de fromage est extrêmement variable et incertaine. Certaines n'en mangent jamais, d'autres en mangent plusieurs fois par jour. On a donc un grand nombre de réalisations possibles. À chacune de ces réalisations potentielles est associée une probabilité, la consommation hebdomadaire de fromage est donc une variable aléatoire. Notons X la quantité consommée et plus précisément le nombre de fois par semaine qu'un individu mange du fromage. Cette variable X a une distribution de probabilité, une loi qu'on note $l(x)$. L'espérance et la variance de X sont deux paramètres de cette loi. $X \rightsquigarrow l(x, m, \sigma)$ avec $m = E(X)$ et $\sigma = \sigma(X)$. À priori, la loi de X, m et σ sont inconnus. Considérons un prélévement au hasard de n individus avec remise dans la population. Observer les quantités consommées de fromage pour ces n individus revient à observer la réalisation de la variable X pour ces n individus, choisis au hasard, indépendamment les uns des autres et avec remise. Les consommations de ces n individus peuvent être considérées comme n variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et de même loi que X c'est-à-dire $l(x, m, \sigma)$.

Les n variables aléatoires indépendantes $X_1, X_2, \ldots, X_n$ constituent un échantillon aléatoire simple de la variable X si et seulement si

$$E(X_1) = E(X_2) = \ldots = E(X_n) = E(X) = m,$$
$$\sigma(X_1) = \sigma(X_2) = \ldots = \sigma(X_n) = \sigma(X) = \sigma.$$

Une fois ces n personnes interrogées, on dispose de n valeurs numériques des quantités consommées. On appelle ces n valeurs numériques observations ou encore réalisations. Ce sont des nombres réels qu'on notera $x_1, x_2, \ldots, x_n$.

On considère donc un modèle statistique $X \rightsquigarrow l(x, \theta)$ et un échantillon aléatoire simple $X_1, X_2, \ldots, X_n$. On recherche une statistique fonction des variables $X_1, X_2, \ldots, X_n$ susceptible de fournir la meilleure estimation possible du paramètre d'intérêt. Cette statistique est appelée **estimateur**.

Population	**Échantillon aléatoire simple**	**Observations**
$X \rightsquigarrow l(x, m)$	$X_1, X_2, \ldots, X_n$ estimateur $\overline{X}$ (par exemple)	$x_1, x_2, \ldots, x_n$ $\overline{x}$ estimation ponctuelle de m

Remarque 3.2.1 Dans le cas de la variable "note", on pourrait prendre comme estimation de m :

$$\frac{x_1 + x_{50}}{2}, \frac{x_1 + x_3 + x_5 + \ldots + x_{49}}{25}, \frac{x_2 + x_4 + \ldots + x_{50}}{25}, \ldots.$$

Dès lors le problème est celui du choix d'un estimateur. Comment va t-on décider quelle statistique utiliser en fonction du paramètre θ recherché ?

3.2.2 Estimateur sans biais

Définition 3.2.1 *Soit $X \rightsquigarrow l(x,\theta)$ un modèle statistique et soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire simple de X. On appelle **estimateur sans biais** du paramètre θ toute statistique $T = T(X_1, X_2, \ldots, X_n)$ telle que $E(T) = \theta$.*

Définition 3.2.2 *Si $E(T) \neq \theta$, T est **biaisé** et le biais vaut $E(T-\theta) = E(T) - \theta$.*

Considérons différentes statistiques ainsi que des tirages non exhaustifs (les tirages ont lieu avec remise) :

1. Prenons l'exemple de la statistique moyenne échantillon.
   Supposons que nous nous intéressons par exemple à l'espérance de la consommation hebdomadaire de fromage. On constitue un échantillon aléatoire simple en tirant au hasard n personnes de la population. Un enquêteur les interroge et obtient les réalisations numériques $x_1, x_2, \ldots, x_n$ des n variables aléatoires $X_1, X_2, \ldots, X_n$. La variable aléatoire "consommation moyenne" est $\overline{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$ et la réalisation de la variable $\overline{X}$ est $\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$.
   On remarquera que la consommation moyenne $\overline{x}$ de l'échantillon varie en fonction de l'échantillon, c'est-à-dire que pour des échantillons différents, on obtient des moyennes d'échantillons différentes.

   Définition 3.2.3 *La variable*

   $$\boxed{\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}}$$

   est appelée variable moyenne échantillon.

   Si l'on considère par exemple 20 échantillons de taille n, on obtient la moyenne $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_{20}$ de chacun de ces échantillons. On peut s'attendre à ce que ces 20 valeurs soient proches de l'espérance m de la consommation hebdomadaire.
   — Espérance de la variable moyenne : soit $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire simple relatif à la variable X. Pour $i = 1, 2, \ldots, n$ on a $E(X_i) = E(X) = m$. Donc $E(\overline{X}) = E\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{nm}{n} = m$.
   — Variance de la variable moyenne : pour i variant de 1 à n, $V(X_i) = V(X)$, les variables $X_1, X_2, \ldots, X_n$ sont indépendantes donc $V(\overline{X}) = V\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n^2} V\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n V(X_i) = \frac{nV(X)}{n^2} = \frac{V(X)}{n} \Leftrightarrow \sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$.

   Proposition 3.2.1 *$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$ est un estimateur sans biais de $\theta = m$ car $E(\overline{X}) = m$.*

2. Prenons l'exemple de la statistique variance échantillon.
   Reprenons l'exemple sur la consommation de fromage. La variabilité des comportements individuels de la consommation est mesurée par l'écart-type σ de la consommation X. On considère un n-échantillon aléatoire simple $X_1, X_2, \ldots, X_n$ de X et la statistique

   $$\boxed{\Sigma^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\overline{X})^2}$$

   La réalisation de cette variable aléatoire Σ^2 est la variance de l'échantillon, notée

$$\sigma'^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\overline{x})^2.$$

Déterminons l'espérance de la variable variance : on a $E(\Sigma^2) = E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right) - E\left(\overline{X}^2\right)$. Donc, $E(\Sigma^2) = \frac{1}{n}\sum_{i=1}^{n}E(X_i^2) - E\left(\overline{X}^2\right) = \frac{1}{n}\sum_{i=1}^{n}\left(V(X_i) + E(X_i)^2\right) - (V(\overline{X}) + E(\overline{X})^2)$. En utilisant les formules précédentes relatives à la variable moyenne échantillon, on trouve $E(\Sigma^2) = \frac{1}{n}\sum_{i=1}^{n}\left(V(X) + E(X)^2\right) - \left(\frac{V(X)}{n} + E(X)^2\right) = V(X) - \frac{V(X)}{n} = \left(1 - \frac{1}{n}\right)\sigma^2$.

Proposition 3.2.2 $\Sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$ *est un estimateur biaisé de* $\theta = \sigma^2$ *car* $E(\Sigma^2) = \left(1 - \frac{1}{n}\right)\sigma^2$, *le biais est alors* $E(\Sigma^2) - \sigma^2 = -\frac{\sigma^2}{n} < 0$

On remarque que l'espérance de Σ^2 ne donne pas une image parfaite de σ^2, variance de X dans la population. Elle est systématiquement plus petite que σ^2 et lorsque n tend vers $+\infty$, $E(\Sigma^2)$ tend vers σ^2. Pour remédier à cet inconvénient de la la statistique Σ^2, on introduit la statistique S^2.

3. Prenons l'exemple de la statistique S^2.
Soit $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire simple de la variable X. On définit la variable

$$\boxed{S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

La réalisation de S^2 est notée

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

On remarque alors que $n\Sigma^2 = (n-1)S^2 = \sum_{i=1}^{n}X_i^2 - n(\overline{X})^2$ et $n\sigma'^2 = (n-1)s^2 = \sum_{i=1}^{n}x_i^2 - n(\overline{x})^2$.

Déterminons l'espérance de la variable S^2 : $E(S^2) = E\left(\frac{n}{n-1}\Sigma^2\right) = \frac{n}{n-1}E(\Sigma^2) = \frac{n}{n-1}\left(1 - \frac{1}{n}\right)\sigma^2 = \sigma^2$.

Proposition 3.2.3 *On déduit du résultat sur l'espérance de la variable* S^2 *que* $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ *est un estimateur sans biais de* $\theta = \sigma^2$ *car* $E(S^2) = \sigma^2$.

4. Prenons l'exemple de la statistique fréquence F.
On considère une population où une certaine proportion, un certain pourcentage d'individus ont une caractéristique donnée. Dans toute la population, avoir ou non une caractéristique donnée est une épreuve de Bernoulli. Par exemple, à l'issue d'une chaîne de fabrication, un article est défectueux avec la probabilité p, non défectueux avec la probabilité $1 - p = q$. Pour chaque article fabriqué, on peut définir la variable aléatoire X : lorsque l'article est défectueux, X prend la valeur 1 et $p(\{X = 1\}) = p$, lorsque l'article n'est pas défectueux, X prend la valeur 0 et $p(\{X = 0\}) = q$. X suit une loi de Bernoulli de paramètre p. Considérons un n-échantillon aléatoire simple de cette variable X soit $X_1, X_2, \ldots, X_n$ de réalisation $x_1, x_2, \ldots, x_n$. Ces n variables aléatoires indépendantes suivent toutes la même loi, celle de X, c'est-à-dire $\mathcal{B}(p)$. Leur somme $Y = X_1 + X_2 + \ldots + X_n$ suit une loi binomiale de paramètres n et p, $Y \rightsquigarrow \mathcal{B}(n,p)$ et $p(\{Y = k\}) = C_n^k p^k (1-p)^{n-k}$ pour k variant de 0 à

n.
On définit ensuite la variable fréquence

$$\boxed{F = \frac{Y}{n} = \frac{X_1 + X_2 + \dots X_n}{n}}$$

Cette variable fréquence est une variable aléatoire dont l'univers image est $\left\{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\right\}$ dont on connaît la distribution de probabilité : $p\left(\left\{F = \frac{k}{n}\right\}\right) = p(\{Y = k\}) = C_n^k p^k q^{n-k}$.
La réalisation de cette variable fréquence est $f = \frac{x_1 + x_2 + \dots + x_n}{n}$, c'est-à-dire la fréquence de l'échantillon ou encore fréquence (ou pourcentage) des articles défectueux dans l'échantillon.

— Espérance de la variable fréquence F : on sait que $F = \frac{Y}{n}$ et $E(Y) = np$ donc $E(F) = \frac{1}{n}E(Y) = \frac{np}{n} = p$.

— Variance de la variable F : on sait que $F = \frac{Y}{n}$ et $V(Y) = npq$ donc $V(F) = \frac{1}{n^2}V(Y) = \frac{npq}{n^2} = \frac{pq}{n}$ et $\sigma(F) = \sqrt{\frac{pq}{n}}$.

On déduit du résultat sur l'espérance de la variable F que

Proposition 3.2.4 *$F = \frac{X_1 + X_2 + \dots + X_n}{n}$ est un estimateur sans biais de $\theta = p$ car $E(F) = p$.*

Dans le cas de tirages exhaustifs (les tirages ont lieu sans remise), si l'on désigne par N la taille de la population et par n la taille de l'échantillon, on obtient les résultats suivants, en faisant intervenir les facteurs d'exhaustivité :

1. $E(\overline{X}) = m$
2. $V(\overline{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$
3. $E(\Sigma^2) = \sigma^2\left(1 - \frac{1}{n}\right)\frac{N}{N-1}$
4. $E(S^2) = \sigma^2\frac{N}{N-1}$
5. $E(F) = p$
6. $V(F) = \frac{pq}{n}\frac{N-n}{N-1}$

Jusqu'ici, nous avons toujours parlé d'échantillons formés avec remise, situation où après avoir été choisie, une unité statistique de la population est remise avec les autre unités de la population, faisant en sorte qu'elle pourrait être choisie à nouveau pour faire partie de l'échantillon. Par le fait même, le choix se faisait toujours à partir d'un ensemble identique, donc les variables X_k étaient toutes identiquement distribuées et indépendantes. Dans notre situation, au fur et à mesure que les unités statistiques sont choisies, cela crée un "trou" dans la population et celle-ci est modifiée. Les variables $X_1, X_2, \dots, X_n$ perdent leur indépendance, elles ne sont plus des répliques de X : le résultat d'un processus influence le résultat du processus suivant, il y a dépendance. Sans remise, les échantillons seront choisis comme ce qui a été décrit pour la loi hypergéométrique. On remarque d'ailleurs que le facteur de correction $\frac{N-n}{N-1}$ qui apparaît dans la variance est le même facteur que celui qui était apparu dans cette loi.

3.3 Estimations pas intervalle de confiance

3.3.1 Préliminaires

Dans le cadre de l'estimation ponctuelle, on associe un nombre, une estimation à un paramètre dont la valeur est inconnue. La précision de cette estimation peut être déterminée en calculant un **intervalle de confiance** pour ce paramètre, c'est-à-dire un intervalle contenant la valeur inconnue du paramètre avec une grande probabilité donnée.

Définition 3.3.1 *Soit un modèle statistique $X \rightsquigarrow l(x,\theta)$ et soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire simple relatif à la variable X. On dit que $[C_1; C_2]$ est un **intervalle de confiance**, de niveau $1-\alpha$, du paramètre θ si on a*

$$p(\{C_1 \leq \theta \leq C_2\}) = 1 - \alpha.$$

Les bornes de l'intervalle C_1 et C_2 sont les statistiques basées sur l'échantillon aléatoire. À priori C_1 et C_2 sont des variables aléatoires, une fois les réalisations de l'échantillon obtenues, on dispose des valeurs numériques $x_1, x_2, \ldots, x_n$. On remplace C_1 et C_2 par leurs réalisations et on obtient les bornes de l'intervalle recherché. Cet intervalle est une réalisation de l'intervalle de confiance $[C_1; C_2]$.

Remarque 3.3.1 Généralement, on prend des intervalles à risque symétrique, c'est-à-dire tels que

$$p(\{\theta < C_1\}) = p(\{\theta > C_2\}) = \frac{\alpha}{2}$$

3.3.2 Intervalle de confiance pour une proportion

Dans une population donnée de grande taille, la proportion d'individus p ayant une caractéristique donnée $\mathcal{C}$ est inconnue. On désire déterminer, à partir d'un tirage d'un échantillon non exhaustif de taille n de la population, un intervalle de confiance au risque α de p.
Le tirage de cet échantillon peut être modélisé par un n-échantillon au hasard tiré d'une variable aléatoire F qui suit une loi de Bernoulli de paramètre p. Soient donc $X \rightsquigarrow \mathcal{B}(p)$ une loi de Bernoulli de paramètre p et $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire simple. La fréquence $F = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ est un bon estimateur (estimateur sans biais, convergent et efficace) du paramètre p, où chacune des variables aléatoires X_i suit une loi de Bernoulli. La fréquence est un estimateur asymptotiquement normal et on utilise l'approximation $F \rightsquigarrow \mathcal{N}\left(p, \sqrt{\dfrac{pq}{n}}\right)$ pour $n \geq 30$, $np \geq 5$ et $nq \geq 5$. Ces conditions seront appelées les **conditions de normalité**.

Les tables statistiques (cf Annexe A) fournissent les valeurs Z_α telles que $p(\{-Z_\alpha < Z < Z_\alpha\}) = 1 - \alpha$ avec $Z \rightsquigarrow \mathcal{N}(0,1)$. On applique cette relation à la variable $Z = \dfrac{F-p}{\sqrt{\frac{pq}{n}}}$ qui suit une loi normale $\mathcal{N}(0,1)$. On obtient

$$\boxed{p\left(\left\{-Z_\alpha < \frac{F-p}{\sqrt{\frac{pq}{n}}} < Z_\alpha\right\}\right) = 1 - \alpha}$$

Remarquons que $-Z_\alpha < \dfrac{F-p}{\sqrt{\frac{pq}{n}}} < Z_\alpha \Leftrightarrow -Z_\alpha < \dfrac{p-F}{\sqrt{\frac{pq}{n}}} < Z_\alpha \Leftrightarrow F - Z_\alpha\sqrt{\dfrac{pq}{n}} < p < F + Z_\alpha\sqrt{\dfrac{pq}{n}}$. On obtient un intervalle de confiance de p au niveau de confiance $1-\alpha$ soit $\left[F - Z_\alpha\sqrt{\dfrac{pq}{n}}; F + Z_\alpha\sqrt{\dfrac{pq}{n}}\right]$.
Pour un risque $\alpha = 5\%$, on trouve $Z_\alpha = 1,96$ et l'intervalle de confiance est :

$$\left[F - 1,96\sqrt{\frac{pq}{n}}; F + 1,96\sqrt{\frac{pq}{n}}\right],$$

pour un risque $\alpha = 1\%$, on trouve $Z_\alpha = 2,58$ et l'intervalle de confiance est :

$$\left[F - 2,58\sqrt{\frac{pq}{n}}; F + 2,58\sqrt{\frac{pq}{n}}\right],$$

Cet intervalle pose un problème pratique important, on peut affirmer que la proportion p appartient à cet intervalle avec une probabilité de $1 - \alpha$ mais les bornes de cet intervalle dépendent de p, la proportion inconnue. Deux possibilités sont utilisées :

1. on remplace p et q par leurs estimations ponctuelles f et $1 - f$. La réalisation de l'intervalle de confiance est alors $\left[f - Z_\alpha\sqrt{\frac{f(1-f)}{n}}; f + Z_\alpha\sqrt{\frac{f(1-f)}{n}}\right]$.

   Exemple 3.3.1 *En vue d'un contrôle de qualité on observe la fabrication d'un objet par une machine durant une période donnée. On décide de tirer un échantillon non exhaustif de taille* $n = 1000$ *dans la fabrication. On constate que 60 d'entre eux sont défectueux. Déterminer au risque de* 5% *un intervalle de confiance de la proportion d'objets défectueux durant la période donnée.*

   Les données sont $n = 1000$ et $f = \dfrac{60}{1000} = 0,06$. La taille de l'échantillon est grande ($n > 30$). Le risque de 5% conduit à $Z_\alpha = 1,96$. L'intervalle de confiance numérique est donc, au risque de 5% :

   $$\left[0,06 - 1,96\sqrt{\frac{0,06(1-0,06)}{1000}}; 0,06 + 1,96\sqrt{\frac{0,06(1-0,06)}{1000}}\right] = [0,045; 0,075].$$

   La proportion p d'objets défectueux fabriqués par la machine est, au risque de 5%, telle que $4,5\% \leq p \leq 7,5\%$.

2. Deuxième méthode : $pq = p(1-p) = -p^2 + p = f(p)$. Alors $f'(p) = -2p + 1$ et on en déduit que f est croissante sur $[0; \frac{1}{2}[$ et décroissante sur $]\frac{1}{2}; 1]$. Dans le cas où p est voisin de $\frac{1}{2}$, on remplace pq par sa valeur maximale $\frac{1}{4}$. La réalisation de l'intervalle de confiance est alors $\left[f - \dfrac{Z_\alpha}{2\sqrt{n}}; f + \dfrac{Z_\alpha}{2\sqrt{n}}\right]$.

   Cette méthode, qui permet un calcul rapide, donne un intervalle de confiance de grande amplitude car la valeur $\dfrac{1}{4}$ du produit $p(1-p)$ est surestimée.

   Exemple 3.3.2 *On a besoin d'estimer rapidement la proportion p d'accidents du travail dans une entreprise de construction. On a constaté sur un échantillon de* 200 *jours ouvrables qu'il y a eu 18 accidents. Déterminer, au risque de* 5%*, un intervalle de confiance de la proportion d'accidents.*

   Les données sont $n = 200$ et $f = \dfrac{18}{200} = 0,09$. Pour un calcul rapide, l'intervalle de confiance numérique est donc, au risque de 5% :

   $$\left[0,09 - 1,96\frac{1}{2\sqrt{200}}; 0,09 + 1,96\frac{1}{2\sqrt{200}}\right] = [0,02; 0,159].$$

   La proportion d'accidents est au risque de 5% telle que $2\% \leq p \leq 15,9\%$. On se rappellera que cette méthode augmente l'amplitude de l'intervalle de confiance. Le calcul fait avec la première méthode donnerait une proportion d'accident p telle que $5\% \leq p \leq 12,99\%$.

Remarque 3.3.2

- On a utilisé l'approximation normale déduite du théorème central limite pour établir l'intervalle de confiance. Il est donc nécessaire que $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$. Dans la pratique, p est inconnue, on vérifie ces conditions sur f donc $n \geq 30$, $nf \geq 5$ et $n(1-f) \geq 5$.
- La longueur de l'intervalle de confiance est $L(\alpha, n) = 2Z_\alpha\sqrt{\dfrac{f(1-f)}{n}}$.

- La précision de l'estimation obtenue est $\frac{1}{2}L(\alpha,n) = Z_\alpha\sqrt{\frac{f(1-f)}{n}}$.
- Z_α étant une fonction décroissante de α (risque pris par le statisticien), lorsque $1-\alpha$ augmente, α diminue, Z_α augmente, la longueur de l'intervalle augmente.
- Lorsqu'on a choisi la valeur de α, on peut imaginer de déterminer la taille de l'échantillon nécessaire pour atteindre une précision donnée l soit $Z_\alpha\sqrt{\frac{f(1-f)}{n}} < l$. On obtient $n > Z_\alpha^2\frac{f(1-f)}{l^2}$.

Exemple 3.3.3

1. On réalise un sondage en vue de prévoir le résultat de l'élection présidentielle. On effectue un tirage aléatoire simple de 750 électeurs. Parmi eux, 324 déclarent qu'ils ont l'intention de voter pour le candidat A tandis que 426 électeurs affirment qu'ils vont voter pour le candidat B. Donner un intervalle de confiance au niveau 95% pour la proportion d'électeurs qui vont voter pour le candidat A.

   On définit la variable X de la manière suivante :
   — si un électeur quelconque vote pour le candidat A, $X = 1$ et $p(\{X = 1\}) = p$,
   — s'il vote pour le candidat B, $X = 0$ et $p(\{X = 0\}) = 1 - p$.
   X suit une loi de Bernoulli de paramètre p. $X_1, X_2, \ldots, X_{750}$ est un échantillon aléatoire simple. Une estimation ponctuelle de p est $f = \frac{324}{750} = 0,432$. L'intervalle de confiance est donné par

$$\left[f - Z_\alpha\sqrt{\frac{f(1-f)}{n}}; f + Z_\alpha\sqrt{\frac{f(1-f)}{n}}\right].$$

   Les conditions de normalité sont vérifiées car $n = 750 \geq 30$, $nf = 750 \times \frac{324}{750} = 324 \geq 5$, $n(1-f) = 426 \geq 5$. On obtient dans la table $Z_{0,05} = 1,960$. L'intervalle numérique est donné par :

$$\left[0,432 - 1,96\sqrt{\frac{0,432 \times 0,568}{750}}; 0,432 + 1,96\sqrt{\frac{0,432 \times 0,568}{750}}\right] = [0,39656; 0,47746].$$

   On peut affirmer que $p(\{0,39656 < p < 0,47746\}) = 0,95$ et $p(\{p < 0,39656\}) = p(\{p > 0,46746\}) = 0,025$.
   Pour $1-\alpha = 0,99$ donc un risque de 1% on a $Z_\alpha = 2,575829$, l'intervalle de confiance est $[0,3854; 0,4786]$.

   Remarque 3.3.3 Cet exemple fait apparaître le peu d'intérêt que présente souvent les sondages tels qu'on les donne dans la presse c'est-à-dire sans donner l'intervalle de confiance ni le niveau de confiance.

2. On souhaite estimer avec une précision de 2% au niveau de confiance $1 - \alpha = 90\%$ le pourcentage de sujets non immunisés après une vaccination. Sur combien de sujets l'observation doit-elle porter sachant que le pourcentage observé de personnes non immunisées est
   (a) $f = 0,20$
   (b) $0,2 < f < 0,3$

   (a) Supposons les conditions de normalité $n \geq 30$, $nf \geq 5$, $n(1-f) \geq 5$ vérifiées avec $\alpha = 0,10$, $f = 0,20$ et $Z_{0,10} = 1,645$. Il faut que

$$Z_\alpha\sqrt{\frac{f(1-f)}{n}} \leq 0,02 \Leftrightarrow n \geq \frac{Z_\alpha^2 f(1-f)}{(0,02)^2} = \frac{(1,645)^2 \times 0,2 \times 0,8}{0,0004} = 1083.$$

   (b) Étudions les variations de $f(1-f)$. Soit $g(x) = x(1-x) = -x^2 + x$, $g'(x) = -2x + 1$. On en déduit alors que g est croissante sur $[0,2; 0,3]$ (avec $g(0,2) = 0,2 \times 0,8$ et $g(0,3) = 0,3 \times 0,7$). Ainsi, $0,16 < f(1-f) < 0,21$ et par conséquent,

$$Z_\alpha\sqrt{\frac{f(1-f)}{n}} < Z_\alpha\sqrt{\frac{0,21}{n}} \leq 0,02 \Leftrightarrow n \geq \frac{(1,645)^2 \times 0,21}{0,0004} \simeq 1421.$$

3.3.3 Intervalle de confiance pour l'espérance

1. La variance σ^2 est supposée connue.
La variable aléatoire parente X suit une loi de probabilité de paramètre $m = E(X)$ inconnu et de variance σ^2 connue. Soit $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire simple de X. On sait alors qu'un bon estimateur ponctuel de m est $\overline{X}$ (estimateur sans biais, convergent et efficace) et que

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \rightsquigarrow \mathcal{N}(m, \frac{\sigma}{\sqrt{n}}) \text{ et } Z = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow \mathcal{N}(0,1).$$

Les tables fournissent la valeur Z_α, pour α donné, telle que $p(\{-Z_\alpha < Z < Z_\alpha\}) = 1 - \alpha$. Or

$$-Z_\alpha < Z < Z_\alpha \Leftrightarrow -Z_\alpha < \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} < Z_\alpha$$
$$\Leftrightarrow -Z_\alpha \frac{\sigma}{\sqrt{n}} < \overline{X} - m < Z_\alpha \frac{\sigma}{\sqrt{n}} \Leftrightarrow -Z_\alpha \frac{\sigma}{\sqrt{n}} < m - \overline{X} < Z_\alpha \frac{\sigma}{\sqrt{n}}$$

On obtient ainsi $p(\{\overline{X} - Z_\alpha \frac{\sigma}{\sqrt{n}} < m < \overline{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}\}) = 1 - \alpha$ c'est-à-dire un intervalle de confiance de m au niveau de confiance $1 - \alpha$ soit $\left[\overline{X} - Z_\alpha \frac{\sigma}{\sqrt{n}}; \overline{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right]$. Dans la pratique, on dispose d'un échantillon non exhaustif tiré au hasard de la population. Cet échantillon fournit une réalisation de $\overline{X}$ par le calcul de la moyenne $\overline{x}$. Ainsi l'échantillon donne une réalisation de l'intervalle de confiance au risque α qui est $\left[\overline{x} - Z_\alpha \frac{\sigma}{\sqrt{n}}; \overline{x} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right]$.

Exemple 3.3.4 Une machine M fabrique des engrenages en grande série. Des études antérieures permettent de dire que les mesures des diamètres forment une population normale d'écart-type $\sigma = 0,042$ cm. On extrait un échantillon non exhaustif de la fabrication journalière de taille $n = 200$ engrenages. La moyenne des diamètres sur cet échantillon est $\overline{x} = 0,824$ cm. Donner au seuil de confiance 95% un intervalle de confiance de la moyenne m des diamètres des engrenages.
Considérons D la variable aléatoire égale au diamètre des engrenages. L'énoncé dit que $D \rightsquigarrow \mathcal{N}(m, \sigma = 0,042)$. Soit $D_1, D_2, \ldots, D_{200}$ un 200-échantillon au hasard de D. Les $n = 200$ variables aléatoires D_i suivent la même loi $\mathcal{N}(m, \sigma = 0,042)$ que D. Soit m le diamètre moyen inconnu des engrenages. On considère alors l'estimateur sans biais et convergent $\overline{D} = \frac{1}{200}\sum_{i=1}^{200} D_i$ de m. Une réalisation de $\overline{D}$ est $\overline{d} = 0,824$. On sait que l'intervalle de confiance au rsique α est $\left[\overline{D} - Z_\alpha \frac{\sigma}{\sqrt{n}}; \overline{D} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right]$. Pour un risque de 5% on a $Z_\alpha = 1,96$. Ainsi, l'intervalle de confiance est $\left[\overline{D} - 1,96\frac{0,042}{\sqrt{200}}; \overline{D} + 1,96\frac{0,042}{\sqrt{200}}\right]$. L'échantillon fournit une réalisation de cet intervalle de confiance à savoir $\left[0,824 - 1,96\frac{0,042}{\sqrt{200}}; 0,824 + 1,96\frac{0,042}{\sqrt{200}}\right]$ soit $[0,818; 0,830]$.

2. La variance est inconnue.
Dans la pratique, si l'espérance $m = E(X)$ est inconnue, a fortiori, la variance $\sigma^2 = E[(X - m)^2]$ est également inconnue. Or nous venons de voir que l'intervalle de confiance de m tel qu'il vient d'être défini dépend de σ. Il est alors tentant de remplacer σ par son estimation ponctuelle s fournie par l'estimateur S^2. Ce nombre n'est autre que l'écart-type calculé sur l'échantillon de taille n avec $n - 1$ degrés de liberté (ddl). Dans ces conditions, on utlise le procédé dit de **Studentisation** qui consiste à remplacer la variable centrée réduite $Z = \frac{\overline{X} - E(\overline{X})}{\sigma(\overline{X})}$ par la variable $T = \frac{\overline{X} - E(\overline{X})}{\frac{S}{\sqrt{n}}}$ qui suit une loi de Student à $n - 1$ ddl. La table de Student nous permet de déterminer $t_{n-1,\alpha}$ tel que pour $n - 1$ ddl on ait

$$p(-t_{n-1,\alpha} \leq T \leq t_{n-1,\alpha}) = 1 - \alpha.$$

On obtiendra alors l'intervalle de confiance au risque α :

$$\left[\overline{X} - t_{n-1,\alpha}\frac{S}{\sqrt{n}}; \overline{X} + t_{n-1,\alpha}\frac{S}{\sqrt{n}}\right]$$

dont une réalisation sur l'échantillon est $\left[\overline{x} - t_{n-1,\alpha}\frac{s}{\sqrt{n}}; \overline{x} + t_{n-1,\alpha}\frac{s}{\sqrt{n}}\right]$.

Exemple 3.3.5 *Dans l'atmosphère, le taux de gaz nocif, pour un volume donné, suit une loi normale d'espérance et de variance inconnues. On effectue n prélèvements conduisant aux valeurs numériques* $x_1, x_2, \ldots, x_n$.

(a) Sur un échantillon de taille $n = 10$*, on observe* $\overline{x} = 50$ *et* $s^2 = 100$.
Quel est l'intervalle de confiance à 5% du taux moyen m de gaz dans l'atmosphère ?

(b) Quel serait cet intervalle si la variance σ^2 *du taux de gaz nocif était connue et valait exactement 100 ?*

(a) Considérons X la variable aléatoire égale au taux de gaz nocif dans l'atmosphère. L'énoncé dit que $X \rightsquigarrow \mathcal{N}(m, \sigma)$ avec m et σ^2 inconnues (m représente le taux moyen de gaz nocif dans l'atmosphère). Soit $X_1, X_2, \ldots, X_n$ un n-échantillon au hasard de X. Les n variables aléatoires X_i suivent la même loi $\mathcal{N}(m, \sigma)$ que X. Les valeurs observées $x_1, x_2, \ldots, x_n$ sont une réalisation du n-échantillon de X. On considère alors l'estimateur sans biais et convergent $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ de m.
Une réalisation de $\overline{X}$ sur l'échantillon est $\overline{x} = 50$. L'intervalle de confiance de m au risque α est l'intervalle aléatoire $\left[\overline{X} - Z_\alpha\frac{\sigma}{\sqrt{n}}; \overline{X} + Z_\alpha\frac{\sigma}{\sqrt{n}}\right]$ où Z_α est déterminé par $p(-Z\alpha < T < Z_\alpha) = 1 - \alpha$ avec $T \rightsquigarrow \mathcal{N}(0, 1)$.
Or σ est inconnu, on le remplace donc par son estimation s. L'intervalle de confiance sur l'échantillon devient : $\left[\overline{X} - t_{n-1,\alpha}\frac{S}{\sqrt{n}}; \overline{X} + t_{n-1,\alpha}\frac{S}{\sqrt{n}}\right]$ où $t_{n-1,\alpha}$ est déterminé par $p(-t_{n-1,\alpha} < T < t_{n-1,\alpha}) = 1 - \alpha$ avec T qui suit une loi de Student à $n - 1$ ddl.
Pour $\alpha = 5\%$ et $n - 1 = 9$, on obtient dans la table $t_{9;0,05} = 2,262$. L'intervalle de confiance recherché est donc : $\left[\overline{X} - 2,262\frac{S}{\sqrt{n}}; \overline{X} + 2,262\frac{S}{\sqrt{n}}\right]$. Une réalisation de cet intervalle de confiance sur l'échantillon est : $\left[\overline{x} - 2,262\frac{s}{\sqrt{n}}; \overline{x} + 2,262\frac{s}{\sqrt{n}}\right]$ soit $\left[50 - 2,262\frac{10}{\sqrt{10}}; 50 + 2,262\frac{10}{\sqrt{10}}\right] = [42,84; 57,15]$ numériquement, qui est donc l'intervalle de confiance, au risque 5% du taux moyen du gaz nocif dans l'atmosphère.

(b) Si la variance est connue et égale à 100, on utilise la table de la loi normale pour déterminer $Z_{0,05} = 1,96$. L'intervalle de confiance est alors $\left[\overline{X} - 1,96\frac{\sigma}{\sqrt{n}}; \overline{X} + 1,96\frac{\sigma}{\sqrt{n}}\right]$. Et une réalisation sur l'échantillon est : $\left[50 - 1,96\frac{10}{\sqrt{10}}; 50 + 1,96\frac{10}{\sqrt{10}}\right] = [43,80; 56,20]$.
L'utilisation de la loi normale donne un intervalle de confiance d'amplitude plus petit que celui obtenu à l'aide de la loi de Student. Ce résultat est cohérent car pour l'utilisation de la loi normale, on a supposé que l'écart-type σ était connu. On a donc une meilleure connaissance de la loi de X que dans le cas où σ est inconnu.

Remarque 3.3.4 Si la taille de l'échantillon est "grande" ($n > 30$), on peut utiliser la loi normale à la place de la loi de Student. C'est pour cette raison qu'on trouve dans la littérature l'expression : "la loi de Student est la loi des petits échantillons".

3.3.4 Intervalle de confiance pour la variance

Théorème 3.3.1 *Soit X une variable aléatoire telle que $X \rightsquigarrow \mathcal{N}(m,\sigma)$ et $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire de X. On utilise $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ comme estimateur sans biais et convergent de σ^2. Alors, la variable aléatoire $\dfrac{n-1}{\sigma^2}S^2$ suit une loi de χ^2 à $n-1$ ddl. On note*

$$\frac{n-1}{\sigma^2}S^2 \rightsquigarrow \chi^2.$$

Un intervalle de confiance au risque α est de la forme $[a;b]$ où a et b sont deux variables aléatoires construites à partir d'un n-échantillon au hasard de X telles que $p(a \leq \sigma^2 \leq b) = 1-\alpha$ or

$$a \leq \sigma^2 \leq b \Leftrightarrow \frac{1}{b} \leq \frac{1}{\sigma}^2 \leq \frac{1}{a} \Leftrightarrow \frac{(n-1)S^2}{b} \leq \frac{(n-1)S^2}{\sigma^2} \leq \frac{(n-1)S^2}{a}.$$

Posons $U = \dfrac{(n-1)S^2}{\sigma^2}$, $t_b = \dfrac{(n-1)S^2}{b}$ et $t_a = \dfrac{(n-1)S^2}{a}$. Alors,

$$p(a \leq \sigma^2 \leq b) = 1-\alpha \Leftrightarrow p(t_b \leq U \leq t_a) = 1-\alpha \Leftrightarrow \left\{p(U < t_b) = \frac{\alpha}{2} \text{ et } p(U > t_a) = \frac{\alpha}{2}\right\}.$$

Comme U suit une loi de χ^2, on détermine les valeurs de t_a et t_b à l'aide d'une table du χ^2 à $n-1$ degrés de liberté. Comme $b = \dfrac{(n-1)S^2}{t_b}$ et $a = \dfrac{(n-1)S^2}{t_a}$, l'intervalle de confiance cherché est alors : $\left[\dfrac{(n-1)S^2}{t_a}; \dfrac{(n-1)S^2}{t_b}\right]$. Une réalisation de cet intervalle de confiance sur l'échantillon est : $\left[\dfrac{(n-1)s^2}{t_a}; \dfrac{(n-1)s^2}{t_b}\right]$ où s est l'estimation ponctuelle de la variance de la population avec $n-1$ ddl.

Exemple 3.3.6 *Une variable aléatoire X est distribuée selon une loi normale de paramètres m et σ inconnus. On dispose d'un n-échantillon associé à X de taille $n = 16$. Sur cet échantillon on observe $\sum_{i=1}^{16}(x_i - \overline{x})^2 = 1500$. Déterminer un intervalle de confiance de la variance au seuil de confiance de 95%.*
La moyenne et la variance de la population sont inconnues. L'intervalle de confiance au seuil 95% de σ^2 est donné par $\left[\dfrac{(n-1)S^2}{t_a}; \dfrac{(n-1)S^2}{t_b}\right]$ et sa réalisation sur l'échantillon est donnée numériquement par $\left[\dfrac{(16-1)s^2}{t_a}; \dfrac{(16-1)s^2}{t_b}\right]$. On a $s^2 = \dfrac{1}{15} \times 1500 = 100$. Pour calculer t_a et t_b, on utilise la variable aléatoire U qui suit une loi du χ^2 à $16-1 = 15$ ddl. D'après les résultats précédents on a $p(U > t_b) = 0,975$ et $p(U < t_a) = 0,025$. La table donne $t_b = 6,26$ et $t_a = 27,5$. On obtient alors l'intervalle de confiance cherché $\left[\dfrac{1500}{27,5}; \dfrac{1500}{6,26}\right] = [54,54; 239,60]$.

3.4 Exercices

Exercice 18 Une entreprise fabrique des sacs en plastique biodégradables pour les enseignes de distribution. Elle s'intéresse au poids maximal que ces sacs peuvent supporter sans se déchirer. On suppose ici que le poids maximal que ces sacs peuvent supporter suit une loi normale d'espérance mathématique 58 (kg) et d'écart-type 3 (kg).

1. Sur 200 sacs reçus, une grande enseigne de distribution constate un poids moyen de 57,7 kg.
   (a) Donner un intervalle de confiance bilatéral de la moyenne des poids sur un échantillon de taille 200, au seuil de 1%.

(b) Quelle est votre conclusion sur le poids moyen constaté ?

2. Donner le poids moyen dépassé dans 97% des cas, sur un échantillon de taille 200.

Exercice 19 Les résultats d'un enquête effectuée sur une population de 1500 salariés d'une entreprise ont montré que dans 65% des cas, les individus avaient au moins un crédit en cours. Trouver la probabilité pour que 2 échantillons de 200 personnes chacun indiquent plus de 10 points d'écart entre les proportions de personnes ayant au moins un crédit en cours.

Exercice 20 Afin de mieux gérer les demandes de crédits de ses clients, un directeur d'agence bancaire réalise une étude relative à la durée de traitement des dossiers, supposée suivre une distribution normale. Un échantillon non exhaustif de 30 dossiers a donné :

Durée (mn)	0-10	10-20	20-30	30-40	40-50	50-60
Effectif	3	6	10	7	3	1

1. Calculer la moyenne et l'écart-type des durées de traitement des dossiers de cet échantillon.
2. En déduire les estimations ponctuelles de la moyenne m et de l'écart-type σ de la population des dossiers.
3. Donner une estimation de m par l'intervalle de confiance au seuil de risque 5%.

Exercice 21 La société G@E a mis au point un logiciel de gestion destiné essentiellement aux PME. Après une enquête dans la région Aquitaine-Limousin-Poitou-Charentes (ALPC), auprès de 100 entreprises déjà équipées d'un matériel informatique (PC) apte à recevoir ce logiciel, la société G@E décide de fixer le prix de vente à 200€. Elle espère diffuser son produit auprès de 68% des PME de la région (cette valeur constituera la proportion de ventes sur l'échantillon). On peut admettre que les 100 PME interrogées constituent un échantillon représentatif des 13250 PME formant le marché potentiel (en 2015).

1. Déterminer l'intervalle de confiance de la proportion p des entreprises intéressées par le logiciel, au seuil de risque 1%.
2. Quelle aurait dû être la taille de l'échantillon pour que l'amplmitude de l'intervalle de confiance soit de 20 points (erreur de $0,1$) ?

Exercice 22 Des observations sur une longue période de la fabrication d'un certain type de boulons ont montré que la résistance à la rupture suit une loi normale dont l'écart-type est $\sigma = 34,5$. Lors d'un contrôle de fabrication, on tire un échantillon de 8 éléments dans la population des boulons fabriqués qui est de très grand effectif. On trouve pour moyenne de résistance à la rupture dans cet échantillon $\overline{x}_e = 225$.

1. Pourquoi à votre avis le tirage des boulons se fait-il sans remise ? Pourquoi peut-on le considérer malgré tout comme un tirage avec remise ? (utiliser le facteur d'exhaustivité $\dfrac{N-n}{N-1}$)
2. Déterminer un estimation de la moyenne des résistances à la rupture des boulons de la fabrication. Prouver que la variable aléatoire $\overline{X}$ suit la loi normale $\mathcal{N}(225, 12.198)$ et donner l'intervalle de confiance, pour cette moyenne m, au seuil de 95%. Qu'en est-il au seuil de 99%.

Exercice 23 Lors d'un contrôle de qualité sur une population d'appareils électro-ménagers, au cours d'un mois de fabrication, on prélève d'une manière non exhaustive un échantillon de 1000 appareils. Après un test de conformité, on constate que 60 appareils ont un défaut.

1. Justifier que la variable « proportion d'appareils électro-ménagers défectueux » suit une loi binomiale de paramètres que l'on précisera. Les conditions de normalité sont-elles vérifiées et si oui, que permettent-elles d'affirmer ?

2. Donner un intervalle de confiance du pourcentage d'appareils défectueux, au risque de 5%.

Exercice 24 Soient les notes de mathématiques d'un étudiant :

4, 5, 8, 10, 12, 13

1. Calculer la moyenne et l'écart-type de la population des notes.
2. Former tous les échantillons exhaustifs possibles de taille 2
3. Calculer l'espérance et l'écart-type de la distribution d'échantillonnage des moyennes.

Exercice 25 On considère une population U de $N = 5$ individus pour lesquels on connaît les valeurs de la variable Y : $y_1 = 3$, $y_2 = 1$, $y_3 = 0$, $y_4 = 1$, $y_5 = 5$. On choisit un 3-échantillon aléatoire simple S dans cette population.

1. Donner les valeurs de la moyenne, de la médiane et de la variance de la variable Y dans la population. Lister tous les échantillons possibles de taille $n = 3$. Quelle est la probabilité de sélection de chaque échantillon ?
2. Pour un échantillon donné, on estime la moyenne (respectivement la médiane) de la population. Calculer les valeurs de ces estimateurs pour chaque échantillon et en déduire que l'estimateur de la moyenne est sans biais alors que l'estimateur de la médiane est biaisé.
3. Pour chaque échantillon, calculer l'estimateur S^2_{YS} de S^2_{YU} et en déduire que cet estimateur est sans biais (on rappelle que $S^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$).

Exercice 26

1. En utilisant la table de nombres aléatoires données ci-après, engendrer un échantillon non exhaustif de taille $n = 10$, de la variable normale X de moyenne $\mu = 5$ et d'écart-type $\sigma = 2$.
2. Vérifier la normalité de l'échantillon obtenu à l'aide d'un graphique gausso-arithmétique.
3. Calculer une estimation sans biais de la moyenne de X donnée par cet échantillon.
4. Calculer une estimation sans biais de la variance de X, donnée par cet échantillon.
5. En déduire l'intervalle de confiance à 95% de la moyenne.

EXTRAITS D'UNE TABLE DE NOMBRES AU HASARD

(Kendall et Babington Smith, table tirée de Christian Labrousse, Statistique, Tome2, Dunod, Paris, 1962)

02 22 85 19 48 74 55 24 89 69 15 53 00 20 88 48 95 08
85 76 34 51 40 44 62 93 65 99 72 64 09 34 01 13 09 74
00 88 96 79 38 24 77 00 70 91 47 43 43 82 71 67 49 90
64 29 81 85 50 47 36 50 91 19 09 15 98 75 60 58 33 15
94 03 80 04 21 49 54 91 77 85 00 45 68 23 12 94 23 44
42 28 52 73 06 41 37 47 47 31 52 99 89 82 22 81 86 55
09 27 52 72 49 11 30 93 33 29 54 17 54 48 47 42 04 79
54 68 64 07 85 32 05 96 54 79 57 43 96 97 30 72 12 19
25 04 92 29 71 11 64 10 42 23 23 67 01 19 20 58 35 93
28 58 32 91 95 28 42 36 98 59 66 32 15 51 46 63 57 10
64 35 04 62 24 87 44 85 45 68 41 66 19 17 13 09 63 37
61 05 55 88 25 01 15 77 12 90 69 34 36 93 52 39 36 23

98 93 18 93 86 98 99 04 75 28 30 05 12 09 57 35 90 15
61 89 35 47 16 32 20 16 78 52 82 37 26 33 67 42 11 93
94 40 82 18 06 61 54 67 03 66 76 82 90 31 71 90 39 27
54 38 58 65 27 70 93 57 59 00 63 56 18 79 85 52 21 03
63 70 89 23 76 46 97 70 00 62 15 35 97 42 47 54 60 60
61 58 65 62 81 29 69 71 95 53 53 69 20 95 66 60 50 70
51 68 98 15 05 64 43 32 74 07 44 63 52 38 67 59 56 69
59 25 41 48 64 79 62 26 87 86 94 30 43 54 26 98 61 38
85 00 02 24 67 85 88 10 34 01 54 53 23 77 33 11 19 68
01 46 87 56 19 19 19 43 70 25 24 29 48 22 44 81 35 40
42 41 25 10 87 27 77 28 05 90 73 03 95 46 88 82 25 02
03 57 14 03 17 80 47 85 94 49 89 55 10 37 19 50 20 37
18 95 93 40 45 43 04 56 17 03 34 54 83 91 69 02 90 72