L'estimation paramétrique

Comme nous l'avons vu plusieurs fois depuis le début de l'année, on modélise souvent le caractère que l'on étudie par une variable aléatoire X afin de rendre compte de la variabilité qu'il peut exister entre les individus. A partir des connaissances que l'on a du caractère et de la population étudiés, on fait souvent l'hypothèse que la loi de probabilité associée à la variable X est d'une certaine forme. Les modélisations paramétriques consistent à faire l'hypothèse que la loi de X (ou une caractéristique de celle-ci) appartient à une famille particulière (dite paramétrique) dont les éléments ne se différencient qu'au travers de quelques paramètres. En d'autres termes, on suppose que la loi de X est connue, à quelques paramètres près. Nous allons proposer des méthodes permettant de donner une valeur approchée de ces paramètres à partir des données recueillies sur un échantillon.

0.1 Quelques exemples de modèles paramétriques

L'objectif de ce paragraphe est de vous présenter de manière assez succinte des exemples de modèles paramétriques. On évoquera seulement les familles paramétriques de lois de probabilité dans ce cours. Sachez cependant que la notion de modèle paramétrique peut avoir un sens plus large en statistique ; on dit par exemple que le modèle de régression linéaire que nous avons vu à la fin de la partie sur les statistiques descriptives est un modèle paramétrique car il ne dépend que des paramètres a et b, même si les lois des variables ne sont pas supposées appartenir à des familles paramétriques.

Lois binomiales $Bin(n,p)$

La loi de Bernoulli (Bin(1,p)) sert à modéliser un caractère de type "succès ou échec". Le paramètre $0 \leq p \leq 1$ représente la probabilité de réussir. Par exemple, on peut modéliser le le fait d'obtenir pile en lançant une pièce par une de Bernoulli de paramètre $p = 0.5$ (en supposant que la pièce n'est pas truquée). La loi binomiale $Bin(n,p)$ est une généralisation de la loi de Bernoulli qui sert à modéliser le nombre de succès que l'on obtient sur n essais. Par exemple, le nombre de fois que l'on obtient pile sur n lancés. Elle est définie par

$$\forall 0 \leq k \leq n,\ \mathbb{P}(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k},$$

son espérance et sa variance sont données par :

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

Exercice : Représenter la loi binomiale Bin(10,0.5) avec R en tapant

$$plot(function(x)\ dbinom(x, 10, 0.5), 0, 10, type =' h')$$

Lois géométriques $G(p)$ **:**

La loi géométrique $G(p)$ sert à modéliser le nombre d'essais infructueux nécessaires avant d'obtenir un premier succès. Le paramètre $0 \leq p \leq 1$ représente ici aussi la probabilité d'un succès lors d'un

essai. La loi géométrique peut par exemple être utilisée pour modéliser le nombre de lancés où la pièce est tombée sur face avant d'obtenir pour la première fois pile. Elle est définie par :

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

son espérance et sa variance sont données par

$$\mathbb{E}[X] = \frac{1-p}{p},\ Var(X) = \frac{1-p}{p^2}.$$

Exercice : Représenter la loi géométrique $G(0.5)$ avec R en tapant

$$plot(function(x)\ dgeom(x, 0.5), 0, 20, type =' h')$$

Lois de Poisson $P(\lambda)$:
La loi de Poisson sert notamment à modéliser des événements rares ou le nombre d'événements survenant dans un intervalle de temps donné, Elle peut par exemple être utilisée dans les télécommunications pour modéliser le nombre de communications échangées dans une période donnée. La paramêtre $\lambda > 0$ correspond à la valeur moyenne du caractère étudié (par exemple le nombre moyen de télécommunications). La loi $P(\lambda)$ est définie par

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

son espérance et sa variance son données par

$$\mathbb{E}[X] = \lambda,\ Var(X) = \lambda.$$

Exercice : Représenter la loi de poisson $P(3)$ avec R en tapant

$$plot(function(x)\ dpois(x, 3), 0, 20, type =' h')$$

Lois uniformes continues $U(a, b)$:
La loi uniforme sert à modéliser des événements qui ont la même probabilité de survenir sur toutes le parties de mêmes longueurs de l'intervalle [a,b] (on suppose bien sûr que $a \leq b$). Elle est notamment utilisée lorsque l'on manque d'information comme nous l'avons vu dans la construction de l'histogramme à partir des données classées. Elle a pour densité

$$f(x) = \frac{1}{b-a}1_{[a,b]}(x),$$

son espérance et sa variance sont données par

$$\mathbb{E}[X] = \frac{a+b}{2},\ Var(X) = \frac{(b-a)^2}{12}.$$

Exercice : Représenter la densité de la loi uniforme $U(0, 1)$ avec R en tapant

$$plot(function(x)\ dunif(x, 0, 1), -1, 2)$$

Lois exponentielles $\mathcal{E}(\lambda)$:
La loi exponentielle sert à modéliser des durées de vie. Le paramètre $\lambda > 0$ est l'inverse de la durée de vie moyenne. Elle est notamment utilisée en radioactivité. Chaque atome radioactif possède une durée de vie qui suit une loi exponentielle. Le paramètre λ s'appelle alors la constante de désintégration. La demi-vie, temps au bout duquel la population initiale a diminué de moitié, correspond à la médiane $\frac{\ln(2)}{\lambda}$ de la loi exponentielle $\mathcal{E}(\lambda)$. Elle a pour densité

$$f(x) = \lambda e^{-\lambda x} 1_{[0;+\infty[}(x),$$

son espérance et sa variance sont données par

$$\mathbb{E}|X] = \frac{1}{\lambda},\ Var(X) = \frac{1}{\lambda^2}.$$

Exercice : Représenter la densité de la loi exponentielle $\mathcal{E}(3)$ avec R en tapant

$$plot(function(x)\ dexp(x, 3), 0, 20)$$

Lois normales $\mathcal{N}(m, \sigma)$:
Les lois normales sont très couramment utilisées en statistique, notamment à cause du théorème de la limite centrale. On peut par exemple les utiliser pour modéliser la répartition de caractéristiques biométriques au sein d'une population ou bien, comme nous l'avons vu, les valeurs prises par le ratio de W. BEAVER parmi des entreprises. Les lois normales sont adaptées pour modéliser des distributions en forme de cloche, symétriques autour d'une valeur moyenne m et dont la dispersion autour de cette moyenne est contrôlée par l'écart-type $\sigma > 0$. La densité s'écrit

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}},$$

son espérance et sa variance sont données par

$$\mathbb{E}[X] = m,\ Var(X) = \sigma^2.$$

Exercice : Représenter la densité de la loi normale $\mathcal{N}(0, 1)$ avec R en tapant

$$plot(function(x)\ dnorm(x, 0, 1), -10, 10)$$

Lois de $\chi^2(k)$:
C'est la loi d'une somme de carrés de k variables de loi $\mathcal{N}(0, 1)$. Le paramètre k est un entier naturel que l'on appelle degré de liberté. Elle apparaît notamment lorsque l'on suppose que X est de loi normale et que l'on regarde la loi de S^2. Elle a pour espérance k et pour variance $2k$
Exercice : Représenter la densité de la loi du $\chi^2(5)$ avec R en tapant

$$plot(function(x)\ dchisq(x, 5), 0, 20)$$

Lois de Student $S(k)$:
Considérons deux variables aléatoires indépendantes U et V de lois respectives $\mathcal{N}(0, 1)$ et $\chi^2(k)$.

La loi de Student est la loi d'un quotient $Z = \frac{U}{\sqrt{\frac{V}{k}}}$. Supposons que l'on considère des variables aléatoires $(X_1, \ldots, X_n)$ indépendantes et de loi $\mathcal{N}(m, \sigma)$, avec σ inconnu. La variable

$$T = \frac{(\overline{X} - m)}{\sqrt{\frac{S}{n}}},$$

suit une loi de Student à $n-1$ degrés de liberté.
Exercice : Représenter la densité de la loi de Student $S(5)$ avec R en tapant

$$plot(function(x)\ dt(x, 5), -20, 20)$$

Lois de Fisher $F(d_1, d_2)$: Considérons deux variables indépendantes U et V de lois respectives $\chi^2(d_1)$ et $\chi^2(d_2)$. La loi de Fisher $F(d_1, d_2)$ est la loi du quotient

$$\frac{U/d_1}{V/d_2}.$$

Comme nous l'avons déjà vu la loi de Fisher apparaît notamment lorsque l'on fait une analyse de variance puisqu'elle correspond à la loi de la statistique de Fisher F lorsque le facteur n'a pas d'effet.
Exercice : Représenter la densité de la loi de Fisher $F(5, 10)$ avec R en tapant

$$plot(function(x)\ df(x, 5, 10), 0, 20)$$

Lois Gamma $\Gamma(a, s)$: Les lois Gamma sont des généralisations des lois exponentielles et du χ^2. La loi d'une somme de N variables exponentielles $\mathcal{E}(\lambda)$ indépendantes est une loi Gamma $\Gamma(N, \frac{1}{\lambda})$. Ces lois permettent de modéliser un grand nombre de distributions des variables à valeurs positives. Les paramètres a et s de cette loi sont strictement positifs et jouent des rôles différents. La paramètre a est un paramètre de forme tandis que le paramètre s est un paramètre d'échelle. Son espérance et sa variance sont donnés par

$$\mathbb{E}[X] = as,\ Var(X) = as^2.$$

Exercice : Représenter la densité de la loi Gamma $\Gamma(1, 1)$ avec R en tapant

$$plot(function(x)\ dgamma(x, 5, 1), 0, 20)$$

puis essayer de tracer $\Gamma(1, 0.1)$ et $\Gamma(5, 1)$.

0.2 But de l'estimation.

On suppose maintenant que l'on a choisi de modéliser le caractère que l'on souhaite étudier par une variable dont la loi de probabilité n'est pas totalement connue mais est supposée appartenir à une famille paramétrique (comme celles évoquées dans la section précédente). On dispose de données $(x_1, \ldots, x_n)$ qui sont les observations des variables aléatoires $(X_1, \ldots, X_n)$ modélisant le caractère étudié pour les n individus de l'échantillon. On s'intéresse donc naturellement maintenant à la manière dont on peut tirer profit des informations recueillies pour donner une valeur vraisemblable (estimation) des paramètres inconnus de la loi.

Exemple : On étudie par exemple la durée du trajet "domicile fac ". Soit D ce caractère. On dispose d'un échantillon de taille $n = 80$.

temps de trajet en mn	$[0,10]$	$]10,20]$	$]20,30]$	$]30,40]$	$]40,50]$	$]50,60]$
effectif	32	30	10	6	1	1
pourcentage	0.40	0.375	0.125	0.075	0.0125	0.0125

En regardant l'histogramme, on pense à une loi exponentielle. Le paramètre noté $\lambda > 0$ de cette loi exponentielle est inconnu et on doit "l'estimer".

On cherche pour λ une valeur vraisemblable compte tenu des valeurs observées $d_1 = D_1(\omega), d_2 = D_2(\omega), \cdots d_n = D_n(\omega)$.

On sait par exemple que plus λ est grand, plus D prend des petites valeurs.

De façon plus précise, $E(D) = \frac{1}{\lambda}$. Ainsi, d'aprés la loi des grands nombres, on a $\frac{D_1+\cdots+D_n}{n} \to \frac{1}{\lambda}$ si n est grand. Par conséquent, l'inverse de la moyenne empirique $\overline{d}$ approche λ (car la fonction inverse est continue en $\lambda > 0$). Et donc dans notre exemple on est tenté d'utiliser comme estimateur $\frac{n}{D_1+\cdots+D_n}$ et d'approcher la valeur inconnue de (on dit souvent "estimer") λ par $\frac{n}{d_1+\cdots+d_n} = \frac{1}{14,6}$. (que l'on appelle estimation de λ).

Nous donnons dans la section suivante une définition plus générale de la notion d'estimateur ponctuel.

0.3 L'estimateur ponctuel.

La définition d'un estimateur ponctuel est très vague :

Définition 1 *Si $(X_1, X_2, \cdots, X_n)$ est un échantillon de X dont la loi dépend d'un paramètre θ, on appelle estimateur ponctuel de θ toute fonction $f(X_1, \cdots, X_n)$. On appelle estimation toute valeur $f(x_1, \cdots, x_n)$ que prend l'estimateur une fois l'expérience réalisée .*

Remarque : Un estimateur est une variable aléatoire (car il dépend des observations qui sont elles mêmes des variables aléatoires) alors qu'une estimation est une valeur numérique.

Dans cette définition le paramètre θ peut être un vecteur de paramètres réels. C'est par exemple le cas pour les lois normales pour lesquelles $\theta = (m, \sigma)$. D'autre part, cette définition ne requiert pas que l'estimateur (souvent noté $\hat{\theta}$) soit "proche" ni même qu'il soit lié au paramètre θ. Cela est un peu étonnant car notre objectif est de l'utiliser pour donner une valeur vraisemblable de θ.

D'autre part, on voit bien que l'on peut définir différents estimateurs pour un même paramètre inconnu θ. On a besoin de critères permettant de comparer la qualité de ces différents estimateurs. En pratique on voudrait que l'estimateur soit "proche" de θ, mais on peut donner plusieurs sens au mot "proche" car l'estimateur est une variable aléatoire. On définit plus précisément ci-dessous différentes qualités des estimateurs qui motivent leur utilisation pour estimer de manière pertinente le paramètre inconnu θ.

A. La convergence : si on interrogeait toute la population, l'estimateur devrait valoir θ. Ainsi, si on a un échantillon suffisamment grand, on voudrait que l'estimateur soit proche de θ.

Définition 2 *On dit que l'estimateur est convergent ssi*

$$f(X_1, ..., X_n) \to \theta, \ \textit{quand } n \to \infty.$$

Remarque : il existe différentes notions de convergence ($\mathbb{L}^p$, presque sûre, en probabilité en loi, ...) pour des variables aléatoires mais nous ne les évoquerons pas en détail ici.

Premiers exemples :
La moyenne empirique est d'aprés la loi des grands nombres un estimateur convergent de la moyenne. En effet, sous certaines hypothèses sur la loi de X,

$$\frac{X_1 + X_2 \cdots + X_n}{n} \to \mathbb{E}[X] \text{ qd } n \to +\infty.$$

Il en est de même pour la variance empirique.

B. Pas de biais : on voudrait qu'il n'y ait pas d'erreur systématique, c'est à dire qu'en moyenne (i.e. espérance), l'estimateur soit égal à θ.

Définition 3 *On dit que l'estimateur est sans biais ssi*

$$\mathbb{E}[f(X_1, \cdots, X_n)] = \theta.$$

Dans le cas contraire on dit que l'estimateur est biaisé. Le biais est la quantité

$$B_n = \mathbb{E}[f(X_1, \cdots, X_n)] - \theta.$$

Premiers exemples :
La moyenne empirique est un estimateur sans biais de l'espérance. En effet, grace à la linéarité de l'espérance on a .

$$\mathbb{E}[\frac{X_1 + \cdots X_n}{n}] = \frac{n * (\mathbb{E}[X])}{n} = \mathbb{E}[X].$$

Par contre, si on veut estimer $\sigma^2 = \mathbb{E}\left[(X - E(X))^2\right]$ la variance d'une variable X, on peut proposer $S_n^2 = \frac{(X_1 - \overline{X}_n)^2 + \cdots + (X_n - \overline{X}_n)^2}{n}$ la variance empirique. C'est un estimateur convergent d'après la loi des grands nombres, mais on peut montrer (dans le cas Gaussien par exemple) que ce n'est pas un estimateur sans biais. En effet,

$$\mathbb{E}\left[(X_1 - \overline{X}_n)^2 + \cdots + (X_n - \overline{X}_n)^2\right] = (n-1)\sigma^2$$

(ce résultat repose sur des résultats théoriques sur les lois Gaussiennes que vous n'avez sans doute pas vu, cependant on peut le vérifier pour $n = 1$). On peut montrer que pour avoir un estimateur sans biais de la variance, il faut diviser par $n - 1$ et non pas par n.

Donc dans l'avenir on prendra plutôt

$$S_{n-1}^2 = \frac{(X_1 - \overline{X}_n)^2 + \cdots + (X_n - \overline{X}_n)^2}{n-1}$$

comme estimateur de la variance (notamment dans un échantillon Gaussien).
Remarque : Quand n est grand, les deux estimateurs sont équivalents ! S_{n-1}^2 est aussi un estimateur convergent.
Remarque : Le biais de S^2 tend vers 0 lorsque n tend vers l'infini. On dit que c'est un estimteur asymptotiquement sans biais.

C. Efficacité d'un estimateur : Il est légitime de demander en plus des qualités précédentes que l'estimateur ait de bonnes qualités en terme de précision. Cette précision se mesure au travers du moment d'ordre 2 suivant

$$\mathbb{E}[(f(X_1, \ldots, X_n) - \theta)^2].$$

Dans le cas d'un estimateur sans biais l'expression ci-dessus n'est autre que la variance de l'estimateur. Par conséquent, pour comparer la précision de deux estimateurs sans biais, on peut comparer leur variance.

Si on dispose de plusieurs estimateurs sans biais et convergents, on choisit celui de variance minimale.

La précision d'un estimateur est cependant limitée. On peut montrer, sous certaines conditions, que pour chaque valeur θ du paramètre et pour un biais b donné, le moment d'ordre 2 centré sur θ d'un estimateur de biais b ne peut pas être inférieur à une certaine valeur appelée borne de Rao-Cramer (qui se calcule à partir de la densité de la loi des Xi, de θ et de b). Dans le cas des estimateurs sans biais, cette borne minimale est appelée variance minimale.

Définition 4 *Un estimateur sans biais de variance minimum est dit* **efficace.**

Un tel estimateur n'existe pas toujours. On appelle efficacité d'un estimateur sans biais le rapport de la variance minimale à la variance de l'estimateur.

Premiers exemples :

Pour la moyenne empirique, l'erreur quadratique est $\frac{Var(X)}{n}$.

Lorsque X suit une loi normale, il est aussi pertinent de considérer la médiane $\tilde{X}$ comme estimateur de m. On pourrait montrer que c'est un estimateur sans biais dont la variance (lorsque n est assez grand) est supérieure à celle de la moyenne empirique. C'est donc un estimateur moins précis que la moyenne.

De manière plus générale, on peut montrer que la moyenne empirique est un estimateur efficace de m lorsque X est de loi normale.

Supposons à nouveau que X suit une loi normale. L'erreur quadratique de S_{n-1} est $\frac{2(\sigma^2)^2}{n-1}$. On peut montrer que dans ce cas la variance minimale est de $\frac{2(\sigma^2)^2}{n}$. Par conséquent S_{n-1} n'est pas un estimateur efficace de σ^2. Cependant son efficacité tend vers 1 lorsque n tend vers l'infini. On dit que c'est un estimateur asymptotiquement efficace.

Supposons enfin que l'on cherche à estimer la proportion p d'individus correspondant à un critère donné au sein d'une population. On introduit des varabales de Bernoulli $X_1, \cdots, X_n$ qui valent 1 si l'individu correspondant de l'échantillon correspond au critère et 0 sinon. Ces variables sont indépendantes (echantillonnage aléatoire simple) et de loi Bernoulli(p). Il semble assez logique de considérer comme estimateur la proportion d'individus de l'échantillon correspondant au critère c'est à dire

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

La somme des X_i suit une loi Binomiale(n,p) d'espérance np donc $\hat{p}$ est un estimateur sans biais. On peut montrer que c'est un estimateur efficace.

Remarque : des résultats généraux (que nous n'avons pas le loisir d'aborder ici) donnent des critères précis permettant d'obtenir des estimateurs efficaces pour une grande variété de lois de

probabilités (familles exponentielles).

Les estimateurs proposés dans les cas précédents sont, pour la plupart, assez intuitifs.

Problème : comment obtenir, en général, des estimateurs possédant les qualités énoncées précédemment ?
Nous allons maintenant présenter deux méthodes générales permettant d'obtenir des estimateurs

0.4 Méthode des moments :

Une première approche, appelée méthode des moments consiste à tirer profit de la loi des grands nombres qui nous dit (sous certaines hypothèses, voir TP sur l'echantillonnage) que les moments empiriques d'ordre k

$$\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

convergent vers les moments théorique

$$\mu_k = \mathbb{E}[X^k].$$

La méthode des moments consiste à exprimer les p premiers moments μ_j de la loi de X en fonction des p paramètres $\theta = (\theta_1, \theta_2, \cdots, \theta_p)$ inconnus. Cela amène à un système de p équations à p inconnues :

$$\begin{aligned}
&\mu_1 = h_1(\theta_1, \ldots, \theta_p)\\
&\ldots\\
&\mu_p = h_p(\theta_1, \ldots, \theta_p)
\end{aligned}$$

que l'on cherche à résoudre afin d'obtenir l'expression de chaque paramètre θ_j en fonction des moments $\mu_1, \ldots, \mu_p$. On obtient ensuite des estimateurs en remplaçant les moments théoriques μ_l par les moments empiriques $\overline{X^l}$ dans les expressions obtenues pour les θ_j.

Remarque : il peut arriver que l'on ne considère pas (ou pas seulement) les p premiers moments afin d'obtenir un système d'équations permettant d'aboutir à une unique solution. Nous n'en parlerons pas en détails dans ce cours car en général les premiers moments suffisent.

Premiers exemples :
La méthode que nous avons utilisée au début de ce TP pour la loi exponentielle est un exemple de l'utilisation de la méthode des moments. On a remarqué que $\mu_1 = \frac{1}{\lambda}$ puis déduit que $\lambda = \frac{1}{\mu_1}$ et enfin remplacé μ_1 par $\overline{X}$ pour aboutir à l'estimateur $\hat{\lambda} = \frac{1}{\overline{X}}$.
Supposons que l'on étudie un échantillon $(X_1, \ldots, X_n)$ constitué de variables aléatoires indépendantes de même loi $\mathcal{N}(m, \sigma)$. Les premiers moments s'expriment en fonction de m et σ de la manière suivante :

$$\begin{aligned}
\mu_1 &= m\\
\mu_2 &= \sigma^2 + m^2.
\end{aligned}$$

On obtient facilement en résolvant ce système les identités suivantes

$$\begin{aligned}
m &= \mu_1\\
\sigma &= \sqrt{\mu_2 - \mu_1^2},
\end{aligned}$$

ce qui nous amène à considérer les estimateurs

$$\hat{m} = \overline{X},\ \hat{\sigma} = \sqrt{\overline{X^2} - \overline{X}^2}.$$

On peut remarquer que dans ce cas on retrouve les estimateurs $\overline{X}$ et $\sqrt{S^2}$ que nous avions déjà considéré.

0.5 Maximum de vraisemblance :

La méthode des moments décrite dans la section précédente est assez intuitive et permet de proposer un certain nombre d'estimateurs convergents. Cependant, ces estimateurs n'ont pas toujours d'aussi bonnes propriétés que ceux que l'on obtient en considérant les estimateurs obtenus par la méthode du maximum de vraisemblance que nous allons voir maintenant.

La **méthode du maximum de vraisemblance** permet d'aboutir dans de nombreux cas à des estimateurs efficaces. Son principe consiste à choisir comme estimation du paramètre θ, la valeur la plus vraisemblable, c'est-à-dire celle qui à la plus forte probabilité de provoquer les valeurs observées dans l'échantillon.

La loi d'une variable X est caractérisée par la probabilité des valeurs possibles (cas discret) ou par sa densité (cas continu). De la même façon, si X suit une loi de paramètre θ, la loi du n-échantillon (composé de variables indépendantes) est caractériséeée par $\Pi_{i=1\cdots n} P_\theta(s_i)$ (variables discrêtes) ou $\Pi_{i=1\cdots n} f_\theta(s_i)$(variables continues), pour toutes les valeurs $(s_1, ...s_n)$ prises par les variables $X_1, \ldots, X_n$. Dans les deux cas, ce produit est une fonction des valeurs s_i et du paramètre θ, que l'on notera $L(s_1, s_2, \cdots s_n; \theta)$.

Définition 5 *On appelle estimateur du maximum de vraisemblance de θ une valeur (de t) qui maximise la fonction de vraisemblance*

$$t \mapsto L(X_1, \ldots, X_n; t).$$

L'estimation de θ par le maximum de vraisemblance revient donc à chercher le paramètre avec lequel $L(x_1, \ldots, x_n, t)$ est maximale, c'est à dire "la valeur du paramètre avec laquelle on avait le plus de chance d'obtenir ce qu'on a obtenu".

Remarque : pour des raisons de commodité de calcul, on utilise souvent la fonction de logvraisemblance

$$LL : t \mapsto \ln(L(X_1, \ldots, X_n, t))$$

qui est le logarithme népérien de la fonction de vraisemblance, elles sont maximales en même temps (car le logarithme népérien est strictement croissant sur $\mathbb{R}_*^+$).

Reprenons l'exemple du trajet domicile fac. Si T suit une loi exponentielle de paramè tre θ, la "probabilité" infinitésimale (ou densité) d'avoir obtenu les résultats $x_1, \cdots x_n$ est

$$L(\vec{x}, \theta) = \theta e^{-\theta x_1} \times \theta e^{-\theta x_2} \cdots \times \theta e^{-\theta x_n} = \theta^n e^{-\sum x_i}.$$

La log-vraisemblance est donc

$$n * \ln(\theta) - \theta \sum_{i=1}^{n} x_i$$

Pour maximiser la log-vraisemblance qui est concave, on dérive par rapport à θ et on annule la dérivée ce qui donne $\frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0$ et donc

$$\hat{\theta} = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}.$$

Sous des hypothèses très générales on peut montrer que l'estimateur du maximum de vraisemblance est efficace ou asymptotiquement efficace et que sa distribution d'échantillonnage est asymptotiquement normale.

EXERCICES :

A.Premiers exercices

1. Soit $(X_1, \ldots, X_n)$ un échantillon constitué de variables indépendantes et de loi $Bin(10, p)$. Quel est l'estimateur de p par la méthode des moments. Quel est l'estimateur du maximum de vraisemblance de p ?
2. Soit $(X_1, \ldots, X_n)$ un échantillon constitué de variables indépendantes et de loi $\mathcal{N}(m, \sigma)$. Quel est l'estimateur de $\theta = (m, \sigma)$ par la méthode de moments. Quel est l'estimateur du maximum de vraisemblance de θ ?
3. Soit $(X_1, \ldots, X_n)$ un échantillon constitué de variables indépendantes et de loi $\mathcal{G}(p)$. Quel est l'estimateur de p par la méthode des moments. Quel est l'estimateur du maximum de vraisemblance de p ?
4. Soit $(X_1, \ldots, X_n)$ un échantillon constitué de variables indépendantes et de loi $U(0, \theta)$. Quel est l'estimateur de θ par la méthode des moments. Quel est l'estimateur du maximum de vraisemblance de θ ? Construire un estimateur sans bias à partir de ce

B.Exercices des annales

1. On peut modéliser la hauteur des éonomies mensuelles (exprimées en milliers de maravedis) d'un ménagedans tel pays lointain, par une variable aléatoire de densité

   $$f(x) = 2kxexp(-kx^2) \text{ pour } x > 0 \text{ et } 0 \text{ sinon.}$$

   Donner l'estimateur de k par la méthode du maximum de vraisemblance.
2. Dans 10 corps d'armée Prussiens, pendant une période de 20 ans allant de 1875 à 1894, le nombre de morts par corps d'armée et par an du à la ruade d'un cheval est donnée par le tableau suivant :

Morts	0	1	2	3	4
Effectif	109	65	22	3	1

   – Quelle est la population étudiée ? Quelle est la taille de l'échantillon considéré ?
   – Démontrer que l'estimateur du maximum de vraisemblance du paramètre λ d'une loi de Poisson est égal à la moyenne arithmétique.
   – Calculer sa valeur à partir des données.
3. Le revenu mensuel X en francs constants des ménages d'une population de revenu minimum θ peut être modélisée par une variable aléatoire de densité

   $$f(x, \theta) = \frac{3\theta^3}{x^4} \text{ si } x \geq \theta \text{ et } 0 \text{ sinon.}$$

   On considèr un échantillon $(X_1, \ldots, X_n)$ de cette loi et on veut estimer le pramètre θ.

- Donner la fonction de vraisemblance associée dans les deux cas suivants $\theta > \min(X_1, \ldots, X_n)$ et $\theta \leq \min(X_1, \ldots, X_n)$.
- En déduire que l'estimateur du maximum de vraisemblance est $\hat{\theta} = \min(X_1, \ldots, X_n)$.
- Montrer que $min(X_1, \ldots, X_n)$ admet la densité

$$g(x, \theta, n) = \frac{3n\theta^{3n}}{x^{3n+1}} \text{ si } x \geq \theta \text{ et } 0 \text{ sinon.}$$

- Montrer que $\mathbb{E}[min(X_1, \ldots, X_n)] = \frac{3n}{3n-1}\theta$ et en déduire un estimateur sans biais de θ.

C. Revenir aux exercices de modélisation sur les données discrêtes et continues pour comprendre ce que l'on a fait.

0.6 L'estimation par intervalle de confiance.

L'estimation ponctuelle d'un paramètre, c'est-à-dire la connaissance de la seule valeur estimée de ce paramètre, n'a d'intérêt que si l'on a une idée de la précision avec laquelle il a été estimé.

La plupart du temps, on complète cette estimation en donnant une fourchette

$$[a(x_1, \ldots, x_n); b(x_1, \ldots, x_n)]$$

appelée intervalle de confiance. Elle correspond à la valeur prise sur l'échantillon par l'intervalle aléatoire $[a(X_1, \ldots, X_n); b(X_1, \ldots, X_n)]$ que l'on construit de telle manière qu'il y ait une grande probabilité que la vraie valeur du paramètre se trouve à l'intérieur.

0.6.1 Rappels sur la notion de quantile

Les notions de fonction de répartition et de fonction quantile associées à la loi d'une variable aléatoire X sont réciproques l'une de l'autre. Elles permettent de répondre à deux problématiques complémentaires :

Qu'elle est la probabilité que X soit inférieure ou égale à une valeur seuil donnée ?
On appelle fonction de répartition associée à la loi la variable aléatoire X la fonction $F_X : s \mapsto \mathbb{P}(X \leq s)$. La valeur de la fonction de répartition au point $s = 5$ correspond donc à la probabilité que la variable X soit inférieure ou égale à 5.
Rappel : on obtient la fonction de répartition avec R en utilisant le préfixe "p" suivi du nom de la loi.

Quelle valeur seuil faut-il considérer pour que X lui soit inférieure ou égale avec (au moins) une probabilité donnée ?
Remarquons tout d'abord que la fonction de répartition F_X n'est pas toujours continue et strictement croissante. Par conséquent il arrive que la solution, en s, de $F_X(s) = t$ n'existe pas ou ne soit pas unique. En d'autres termes, il se peut que pour certaines valeurs de t il n'exise pas de valeur seuil s telle que la probabilité que X lui soit inférieure ou égale soit exactement égale à t. Il se peut également que l'on ait plusieurs valeurs seuil possibles.
La fonction quantile associée la loi de X correspond à la fonction inverse généralisée de la fonction de répartition. On la définit de manière générale par

$$Q_X(t) = \inf\{s, F_X(s) \geq t\}.$$

On appelle donc quantile d'ordre t (noté $Q_X(t)$) la plus petite valeur seuil s telle que la probabilité que $X \leq s$ soit au moins t. Remarquons cependant que lorsque la variable X est de loi continue et si la densité est strictement positive, alors F_X est strictement croissante et l'on peut définir $Q_X(t)$ comme la solution en s de $F_X(s) = t$. Le quantile $Q_X(t)$ correspond alors exactement à la seule valeur seuil pour laquelle on a exactement une probabilité t que X lui soit inférieure ou égale.
Rappel : on obtient la fonction quantile avec R en utilisant le préfixe "q" suivi du nom de la loi.

La probabilité que X soit entre deux valeurs a et b (avec $a \leq b$) s'obtient de la manière suivante :

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a).$$

Cela est toujours vrai pour les variables de loi continues si on considère des inégalités strictes ou larges car $P(X = a) = P(X = b) = 0$. Attention, ce n'est pas le cas lorsque X est une variable discrète car dans ce cas $P(X < s) < F_X(s)$.

On se concentre pour ce paragraphe sur le cas de variables dont la loi est continue. On peut visualiser la valeur de la fonction de répartition au point s comme l'aire comprise sous le graphe de la densité jusqu'à la valeur s (c'est à dire sur l'intervalle $]-\infty; s]$). Chercher le quantile $Q_X(t)$, consiste à trouver la plus petite valeur s telle que l'aire comprise sous la courbe soit égale à t. Enfin, la probabilité que X se trouve dans l'intervalle $[a; b]$ est l'aire comprise sous la courbe de la densité entre les points a et b.

Les fonctions de répartition de certaines lois de probabilité n'ont pas de forme explicite. Leurs valeurs sont données dans des tables. On peut se servir de ces tables pour trouver la valeur approchée des quantiles. Vous trouverez par exemple les tables de la loi normale centrée réduite $\mathcal{N}(0, 1)$ ainsi que des lois du χ^2 et des lois de Student. Elles sont présentées de manières un peu différentes.

La table de la loi $\mathcal{N}(0, 1)$ dont vous disposez donne les valeurs de la fonction de répartion $F_X(t)$ pour différentes valeurs de t. La première colonne de la table correspond aux premières décimales de t tanis que la première ligne correspond aux centièmes. On trouve par exemple la valeur de $F_X(0.25)$ dans la case correspondant à la ligne correspondant à "0.2" (dans la première colonne) et dans la colonne correspondant à "0.05"(dans la première ligne), c'est à dire la valeur 0.5987. *La table ne donne que les valeurs de la fonction de répartition pour des valeurs de t positives. Comme la densité de la loi $\mathcal{N}(0, 1)$ est symétrique par rapport à 0 on a $F_X(-t) = 1 - F_X(t)$, ce qui nous permet d'obtenir les valeurs $F_x(t)$ pour $t < 0$. Par exemple, $F_X(-0.25) = 1 - F_X(0.25) = 1 - 0.5987 = 0.4013$. Enfin, lorsque la variable Z est de loi $\mathcal{N}(m, \sigma)$, la variable $\frac{Z-m}{\sigma}$ est de loi $\mathcal{N}(0, 1)$. Par conséquent,*

$$F_Z(t) = \mathbb{P}(Z \leq t) = \mathbb{P}(\frac{Z - m}{\sigma} \leq \frac{t - m}{\sigma}) = F_X(\frac{t - m}{\sigma}),$$

où X est une variable aléatoire de loi normale centrée réduite $\mathcal{N}(0, 1)$. Supposons par exemple que Z suit une loi $\mathcal{N}(1, 2)$. La probabilité que Z soit inférieur ou égal à 2 est

$$F_Z(2) = F_{\frac{Z-1}{2}}(\frac{2 - 1}{2}) = F_{\frac{Z-1}{2}}(0.5) = 0.6915.$$

La table de la loi de Student donne seulement les valeurs de t telles que $P(|X| > t) = P$ pour différentes valeurs de P et du degré liberté ν. Nous verrons l'intérêt de donner simplement ce type de valeurs dans ce qui suit.

Enfin, la table du χ^2 donne quant à elle les valeurs de t telles que $P(X \leq t) = P$ (c'est à dire de $Q_X(P)$) pour différentes valeurs de P et du degré de liberté ddl.

0.6.2 Construction d'intervalles de confiance

Exemple introductif : *Un sondage effectué à quelques jours du second tour d'une élection présidentielle sur un échantillon de 100 personnes donne 47% d'intentions de vote pour Mr Dupont et 53% pour Madame Durand. On supposera que l'échantillonnage est aléatoire et simple. Notons p la probabilité qu'un électeur vote Madame Durand. Les observations recueillies sur notre échantillon nous amènent à estimer p par 53%. Toutefois, afin de pouvoir conclure quelque chose de ce sondage, il est nécessaire d'avoir plus d'informations sur la marge d'erreur de notre estimation. En d'autres termes, quel risque court-on en concluant que p est supérieur à 50% ? Pour répondre à ces questions, on donne en plus un intervalle de confiance $[a(x_1, \ldots, x_n); b(x_1, \ldots, x_n)]$. Il s'agit de la valeur observée (calculée à partir de nos observations) d'un intervalle aléatoire $[a(X_1, \ldots, X_n); b(X_1, \ldots, X_n)]$ construit à partir de notre échantillon $(X_1, \ldots, X_n)$ de telle sorte que l'on ait une grande probabilité (appelée niveau de confiance) que p y appartienne.*

Démarche générale :

On fixe un nombre α compris entre 0 et 1, généralement proche de 0 et appelé **risque.** *Le nombre $1-\alpha$, généralement proche de 1 est appelé* **niveau de confiance.**

Etant donné un échantillon $(X_1, X_2, \cdots, X_n)$ de variables aléatoires, on cherche un intervalle aléatoire $I(X_1, \cdots, X_n) = [a(X_1, \cdots, X_n), b(X_1, \cdots, X_n)]$, tel que α étant fixé, on ait :

$$P\left(\theta \in I(X_1, \cdots, X_n)\right) \geq 1 - \alpha. \tag{1}$$

Une observation $I(x_1, x_2, \cdots x_n)$ de $I(X_1, \cdots, X_n)$ est appelée intervalle de confiance de θ au niveau $1-\alpha$.

On appelle $1-\alpha$ le niveau de confiance de l'intervalle de confiance I, et on dit que θ est dans l'intervalle I avec un niveau de confiance $1-\alpha$.

Remarque : Lorsque l'on considère des risques plus faibles, la largeur de l'intervalle de confiance est plus grande. Afin d'avoir moins de risques de se tromper en concluant que $\theta \in I(X_1, \cdots, X_n)$, on considère un intervalle plus large et on perd donc en précsion.

Remarque : Il existe en général une infinité d'intervalles de confiance vérifiant (1) (même lorsque la probabilité vaut $1-\alpha$). Afin d'avoir un maximum de précision, on choisit en général les intervalles d'amplitude la plus courte. Par conséquent on privilégira notamment les intervalles tels que $P\left(\theta \in I(X_1, \cdots, X_n)\right) = 1-\alpha$ s'ils existent (voir remarque précédente). D'autre part, parmi les intervalles vérifiant cette dernière égalité, on considèrera plutôt certains types d'intervalles en fonction de la nature de la densité de la loi que l'on considère (voir exemples suivants).

Remarque : La largeur des intervalles de confiance décroît lorsque la taille de l'échantillon augmente (voir les exemples suivants).

A. Estimation d'une moyenne par intervalle de confiance :

On construit un intervalle de confiance pour la moyenne en utilisant l'estimateur de la moyenne $\overline{X}$.

A1. Cas d'un échantillon de grande taille : *on peut utiliser le théorème de la limite centrale, qui dit que $\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}}$ suit approximativement la loi $N(0,1)$ quand n est grand ($n > 30$). De plus, quand n est grand on peut approximer σ^2 par la variance empirique S_{n-1}.*
La densité de la loi normale $\mathcal{N}(0,1)$ est symétrique et concentrée autour de sa moyenne 0. Par conséquent, l'intervalle de confiance de largeur minimale est symétrique par rapport à 0. On recherche donc une constante δ_α positive telle que

$$\mathbb{P}(\frac{\overline{X}-m}{\frac{S_{n-1}}{\sqrt{n}}} \in [-\delta_\alpha; \delta_\alpha]) \geq 1-\alpha.$$

Cela revient à chercher δ_α telle que

$$P\left(\frac{|\overline{X}-m|}{\frac{S_{n-1}}{\sqrt{n}}} > \delta_\alpha\right) \leq 1-\alpha.$$

Comme la densité de la loi gaussienne est symétrique par rapport à 0, pour toute variable aléatoire Z de loi $\mathcal{N}(0,1)$ et $\delta > 0$ on a $\phi(-\delta) = \mathbb{P}(Z \leq -\delta) = \mathbb{P}(Z \geq \delta)$. Par conséquent la valeur de δ_α

correspond à la valeur t telle que $\mathbb{P}(Z \leq t) = 1 - \frac{\alpha}{2}$ et δ_α est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$ que l'on note $\phi^{-1}(1 - \frac{\alpha}{2})$. On peut se servir de la table pour obtenir les valeurs du quantile et on en déduit finalement que

$$m \in [\overline{X} - \frac{S_{n-1}}{\sqrt{n}}\phi^{-1}(1 - \frac{\alpha}{2}), \overline{X} + \frac{S_{n-1}}{\sqrt{n}}\phi^{-1}(1 - \frac{\alpha}{2})]$$

avec une sécurité de $1 - \alpha$.

On donnera donc comme intervalle de confiance de niveau de risque $\alpha = 0.05$ pour m l'intervalle

$$[\overline{x} - \frac{s_{n-1}}{\sqrt{n}} \times 1.96, \overline{x} + \frac{s_{n-1}}{\sqrt{n}} \times 1.96].$$

On voit clairement que la largeur de l'intervalle de confiance décroît lorsque la taille de l'échantillon augmente.

Exercice [LAIT] : *La production laitière individuelle annuelle de vaches laitières peut être considérée, dans certaines conditions de race et de région, comme une variable aléatoire, d'écart-type pratiquement égal à 1000 litres et d'espérance variant d'une race à l'autre et d'une région à l'autre, aux environs de 4000 à 5000 litres (Vann Snick et Kint, 1963). Supposons que pour une race donnée dans une région donnée, on ait mesuré la production de 35 bêtes, choisies au hasard et indépendamment, et que la moyenne observée dans ces conditions soit égale à 4790 litres. Quel est l'intervalle de confiance de la production laitière individuelle annuelle moyenne de la population considérée, au niveau de confiance 99% ? Que peut-on faire l'écart-type est inconnu mais que l'écart-type observé sur les 35 vaches est de 1053 litres ?*

A2. Cas d'un échantillon de petite taille : *on ne peut plus utiliser le théorème de la limite centrale et donc on ne peut pas faire grand chose. Le seul cas résolvable est celui où on suppose que le modèle est Gaussien. Dans ce cas, si l'écart-type est supposé connu, on centre et réduit et on utilise la loi $N(0,1)$ comme dans l'exemple précédent.*

Par contre si on ne connait pas σ^2, on l'estime par l'estimateur S^2_{n-1} et on centre et on réduit. On peut montrer que dans ce cas la variable centrée réduite suit la loi de student de paramètre $n-1$. La densité de la loi de student est elle aussi symétrique et concentrée autour de 0. En suivant les mêmes étapes que dans le paragraphe précédent, on obtient les intervalles de confiance suivants où les quantiles $t_{1-\frac{\alpha}{2}}$ de la loi de Student(n-1) remplacent ceux de la loi $\mathcal{N}(0,1)$.

$$[\overline{x} - \frac{s_{n-1}}{\sqrt{n}} \times t_{1-\frac{\alpha}{2}}, \overline{x} + \frac{s_{n-1}}{\sqrt{n}} \times t_{1-\frac{\alpha}{2}}].$$

Remarque : On lit facilement la valeur de $t_{1-\frac{\alpha}{2}}$ sur la table de la loi de student sur la ligne corespondant au degré de liberté $n-1$ et dans la colonne correspondant à $P = \alpha$.

Exercice : *On souhaite connaître la durée de vie d'un composant dans des conditions extrêmes d'utilisation (température élevée, humidité, ...). Pour cela on fait fonctionner 3 appareils jusqu'à la*

panne dans un environnement expérimental et on observe les durées de vies suivantes : $D_1 = 45$, $D_2 = 48$ et $D_3 = 49.5$. Sachant que la durée de vie de ces appareils suit une loi normale, donner un intervalle de confiance avec une sécurité de 95% pour la durée de vie moyenne.

Exercice : *Dans une étude concernant le rôle de la température corporelle sur l'activité diurne d'insectes coléoptères de la famille des ténébrionidés, Kenagy et Stevenson (1982) ont mesuré le poids et la température interne du corps d'individus actifs de plusieurs espèces. Chez eleodes obscura, la température moyenne des 9 individus étudiés s'élève à 23,5°C avec un écart-type de 4,5. Quel est l'intervalle de confiance de la température moyenne du corps en activité de cette espèce au niveau de confiance 0.95, sachant que la distribution des températures corporelles d'individus d'espèces différentes de la même famille suit une loi normale ?*

A3. Estimation d'une proportion p : *Pour fixer les idées, revenons à notre exemple introductif et considérons la proportion p d'individus favorables à Mme Durand. Le nombre N de personnes favorables dans l'échantillon considéré, suit une loi binomiale $B(n, p)$. L'estimateur ponctuel de p est $F = \frac{N}{n}$.*

Si n est grand (n>30) et $nf(1-f) > 12$, on peut approximer la loi binomiale par la loi normale $N(np, np(1-p))$. Ainsi, $\frac{F-p}{\sqrt{p(1-p)/n}}$ peut-être approximé par une loi $N(0,1)$ et on a donc :

$$P\left(\frac{|F-p|}{\sqrt{p(1-p)/n}} > \delta\right) \approx 2\Phi(\delta) - 1.$$

On choisit donc $\delta_\alpha = \Phi^{-1}(1-\frac{\alpha}{2})$ et l'on a

$$\mathbb{P}(p \in [F - \frac{\sqrt{p(1-p)}}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2}), F + \frac{\sqrt{p(1-p)}}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2})]) \approx 1-\alpha.$$

Cependant, les bornes de l'intervalle dépendent de p qui est inconnue.

Une première approche consiste à trouver analytiquement les valeurs de p telles que $p \in [F - \frac{\sqrt{p(1-p)}}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2}), F + \frac{\sqrt{p(1-p)}}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2})]$ en étudiant des inégalités faisant intervenir des polynômes d'ordre 2 en p. Cette approche est cependant assez fastidueuse. On utilisera plutôt l'une des deux méthodes "approchées" suivantes qui ont l'avantage d'être plus simples.

On peut tout d'abord donner un intervalle un peu plus large en utilisant que $p(1-p) \leq 1/4$:

$$p \in [f - \frac{1/2}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2}), f + \frac{1/2}{\sqrt{n}}\phi^{-1}(1-\frac{\alpha}{2})]$$

avec une sécurité de $1-\alpha$. Si on veut être plus précis, on remplace $p(1-p)$ par $f(1-f)$.

Lorsque l'on ne peut pas utiliser une approximation de la loi binomiale par une loi normale, il est possible d'envisager dautres approches mais nous n'en parlerons pas ici.

Revenons à notre exemple : Nous avons $n = 100 > 30$ et $nf(1-f) = 100 \times 0.53 \times 0.47 = 24.91 > 12$. On peut donc appliquer l'approximation de la loi binomiale par la loi gaussienne, ce

qui nous donne pour $\alpha = 0.05$ *comme intervalle de confiance*

$$\begin{aligned} & [f - \frac{\sqrt{f(1-f)}}{\sqrt{n}} \times 1.96, f + \frac{\sqrt{f(1-f)}}{\sqrt{n}} \times 1.96] \\ = \quad & [0.53 - \frac{\sqrt{0.53 \times 0.47}}{\sqrt{100}} \times 1.96, 0.53 + \frac{\sqrt{0.53 \times 0.47}}{\sqrt{100}} \times 1.96] \\ \approx \quad & [0.4322; 0.6278] \end{aligned}$$

On ne peut donc pas vraiment conclure que Madame Durand a plus de 50% d'intentions de vote dans la population toute entière avec un risque de 5%.

Exercice : *Retrouver ces valeurs avec le logiciel R.*

Exercice : *On examine un échantillon de 100 patients présentant un syndrome cutané S. On constate que 25% d'entre eux sont atteints de la maladie M. Donner un intervalle de confiance avec sécurité de 99%pour la proportion des personnes ayant la maladie M parmi celles qui présentent le syndrome S. Utiliser le logiciel R pour en donner la valeur.*

B. Estimation d'une variance par intervalle de confiance pour un échantillon Gaussien :

1. *Estimation de* σ^2 *avec m connue.*
   On utilise que $T = \sum_{i=1}^{n} \frac{(X_i - m)^2}{\sigma^2}$ *est une somme de n variables indépendantes de loi* $N(0,1)$ *au carré. Or une telle somme suit une loi* $\Gamma(\frac{n}{2}, \frac{1}{2})$*, appelée aussi loi de Khi-deux à n degré de liberté.*
   Puis dans la table de Khi-deux on cherche u et v tels que $P(u \leq \chi^2(n) \leq v) = 1 - \alpha$*, on choisit par exemple u et v tels que* $P(\chi^2(n) \leq u) = P(\chi^2(n) \geq v) = \frac{\alpha}{2}$
   Par exemple, $n = 30, \quad 1 - \alpha = 0.95$ *et* $u = 16.79$ *et* $v = 50.89$*. D'où l'intervalle de confiance pour* σ^2 *:* $[\frac{\sum(x_i - m)^2}{v}, \frac{\sum(x_i - m)^2}{u}]$.

   Quand le degré de liberté $\nu > 30$*, on ne peut plus utiliser la table du khi-deux et on utilise l'approximation suivante :*

   Le quantile $\chi^2_{\nu,\alpha}$ *ayant la proba* α *d'être dépassé peut-être approximé quand* $\nu > 30$ *par* $\frac{\left(t_\alpha + \sqrt{2 * \nu - 1}\right)^2}{2}$*, où* t_α *est la valeur ayant la proba* α *d'être dépassée par la loi* $N(0,1)$*. Ainsi*

$$u = \frac{\left(t_{\frac{\alpha}{2}} + \sqrt{2 * \nu - 1}\right)^2}{2}$$

$$v = \frac{\left(t_{1-\frac{\alpha}{2}} + \sqrt{2 * \nu - 1}\right)^2}{2}$$

2. *Estimation de* σ^2 *quand m est inconnue.*
   On procède exactement de la même façon en remplaçant m par son estimateur $\overline{X}$ *et* $\frac{\sum(X_i - \overline{X})^2}{\sigma^2}$ *suit une loi de Khi-deux à* $n - 1$ *degrés de liberté.*

Exercice [LAIT(suite)] : *Supposons que l'on modélise la production de lait par une loi normale. On observe sur 15 vaches un écart-type de 1048 litres. Donner un intervalle de confiance de niveau de confiance* 95% *pour l'écart-type* σ *en supposant que la moyenne* m *est inconnue. Supposons maintenant qu'une autre étdude a montré que la moyenne de production de lait vaut* 4750 *litres. Quel est alors l'intervale de confiance de* σ *? Mêmes questions lorsque l'échantillon est constitué de 35 vaches.*

C. Détermination du nombre d'observations nécessaires pour atteindre une certaine précision.

L'amplitude d'un intervalle de confiance est une fonction décroissante de α *et de* n*.*

Par conséquent, pour augmenter la précision de l'estimation c'est-à-dire réduire l'amplitude de l'intervalle de confiance, il faut :

- *soit augmenter le risque* α *(ou ce qui revient au même, diminuer le niveau de confiance* $1-\alpha$*)*
- *soit augmenter la taille de l'échantillon.*

Comme on souhaite en général ne pas augmenter le risque, on peut souhaiter déterminer le nombre d'observations nécessaires pour atteindre une précision donnée pour un risque α *fixé.*

Revenons à notre exemple concernant les élections. On a vu qu'avec un échantillon de taille 100, l'intervalle de confiance de risque $\alpha = 0.05$ *a une précision de* 0.0978 *et donc une amplitude de* 0.1956 *autour de la valeur* 0.53*. Il contient des valeurs plus petites que* 50% *donc on ne peut conclure que Mme Durand a plus d'électeurs en sa faveur dans la population totale avec un risque de* 5%*. On pourrait voir que pour que avoir une précision de* 3% *de notre intervalle de confiance et qu'il ne contienne pas de valeur plus faible que* 0.5 *il nous faut prendre un risque d'environ* 0.548 *ce qui est tout de même assez élevé. On préfère plutôt considérer un échantillon de taille plus grande qui nous permettrait d'avoir une précision de* 3% *autour de la valeur* 0.53 *pour le même risque* $\alpha = 0.05$*. On voit par exemple que pour* $n = 1064$*, on a une précision supérieure à* 3%*. On voit cependant clairement que la précision de notre intervalle de confiance dépend de la valeur de* p *et donc de celle de* f*. Pour avoir une précision de* 3% *indépedamment des valeurs de* p *et* f*, on doit prendre* n *tel que* $\frac{\sqrt{1/4}}{\sqrt{n}} * \Phi^{-1}(1-\frac{\alpha}{2}) \leq 0.03$ *c'est à dire tel que* $n \geq (\frac{1/4}{0.03^2} * 1.96^2 \approx 1067.111$*.*

Exercice : retrouver cela avec le logiciel R et par les calculs.

Exercice [LAIT(suite)] : *Supposons que l'on modélise par une loi normale la production de lait. Quelle taille d'échantillon faudrait-il pour avoir un intervalle de confiance (de niveau de confiance 95%) d'amplitude inférieure à 400 litres en supposant que l'écart-type est de 1000 litres ?*

D. Exercices supplémentaires

1. *On a mesuré la consommation d'oxygène par minute rapporté au poids sur 6 champions cyclistes et sur 7 champions de natation. Les résultats sont répertoriés dans le tableau suivant :*

Cyclistes	*73*	*71*	*69*	*72*	*74*	*70*	
Nageurs	*64*	*69*	*73*	*68*	*69*	*67*	*66*

   - *On fait l'hypothèse que la consommation d'oxygène suit une loi normale d'écart-type* $\sigma = 2.3$*. Donner un intervalle de confiance de la moyenne pour chaque catégorie de champion avec un niveau de confiance de 0.95. Trouver sa valeur avec des commandes R. Mêmes questions si* σ *n'est pas connu.*

- *On ne fait maintenant aucune hypothèse sur l'écart-type. Donner un intervalle de confiance de niveau de confiance 0.95 de la variance pour chaque catégorie de champion.*
- *Que pensez-vous de l'hypothèse initiale sur l'écart-type ?*

2. *L'office des Lacs Poldaves (O.L.P.) veut estime le nombre N de brochets qui vivent dans la région des grands lacs. Pour cela, on marque puis rejette à l'eau 1000 brochets adultes. Puis on offre une récompense aux pécheurs pour toute bague de marquage renvoyée à l'O.L.P. dans les deux mois qui suivent (cette période est jugée assez brève pour négliger les effets de natalité ett de mortalité).*
   - *Sachant que sur les 10000 péchés dans les deux mois , 50 sont marqués, donner une estimation ponctuelle de la proportion p de brochets marqués qui vivent dans la région des grands lacs.*
   - *Donner un intervalle de confiance à 95% de cette proportion p.*
   - *En déduire une fourchette pour le nombre N.*
3. *Aux portes de votre ville on parle d'ouvrir un hypermarché. Pour en apprécier l'opportunité, un buerau d'études interroge un échantillon représentatif de 400 personnes.*
   - *A la question "êtes-vous favorables à la création d'un supermarché géant dans votre ville ?" 340 personnes ont répondu positivement. Donner une estimation ponctuelle, puis par intervalle de confiance (avec une sécurité de 95%) de la proportion de personnes favorables.*
   - *L'organisme chargé de l'étude recherche, entre autres, s'il est judicieux de prévoir une cafétéria dans le supermarché. Dans cette perspective, l'une des questions porte sur la fréquence hebdomadaire des repas pris en dehors du domicile. Il ressort de l'enquête que le nombre moyen de repas est 1.4 avec un écart-type estimé de 3. Donner une estimation par intervalle de confiance avec une sécurité de 80% du nombre moyen de repas pris en dehors du domicile. Donner également un intervalle de confiance le l'écart-type avec une sécurité de 90%.*
4. *On a mesuré la quantité d'alcool total (mesurée en gr/l) contenue dans 10 cidres doux du marché. On suppose que la quantité d'alcool des cidres suit une distribution normale de moyenne m et d'ecart-type σ. On obtient les valeurs suivantes :*

$$5.42; 5.55; 5.61; 5.91; 5.93; 6.15; 6.20; 6.79; 7.07; 7.37.$$

   - *Déterminer un intervalle de confiance de la moyenne m avec un niveau de confiance $\alpha = 95\%$:*
     a) si l'on suppose que $\sigma = 0.6g/l$,
     b) si σ est inconnu.

   - *Déterminer l'intervalle de confiance de la mesure de σ^2 avec un niveau de confiance de 95% :*
     a) si l'on suppose que $m = 6.2g/l$,
     b) si m est inconnu.