UNIVERSITÉ DE NANTES

Licence 3 Mathématiques

Statistique Inférentielle

Paul Rochet

Contents

1 Modèle d'échantillonage **2**

2 Inférence dans un modèle paramétrique **3**
2.1 Méthode des moments . . . 3
2.2 Maximum de vraisemblance . . . 6

3 Estimation non-paramétrique **9**
3.1 Fonction de répartition . . . 9
3.2 Quantiles . . . 12

4 Intervalles de confiance **14**
4.1 Le modèle Gaussien . . . 15
4.2 Intervalles de confiance asymptotiques . . . 17

5 Tests **18**
5.1 Définitions générales . . . 18
5.2 Tests sur un paramètre de position . . . 20
5.2.1 Test du signe . . . 20
5.2.2 Test de Student sur la moyenne . . . 21
5.3 Tests d'adéquation . . . 22
5.3.1 Test du χ^2 . . . 22
5.3.2 Test de Kolmogorov-Smirnov . . . 24

1 Modèle d'échantillonage

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé et $\mathcal{X}$ un sous-ensemble de $\mathbb{R}$. On observe une réalisation $(x_1, ..., x_n) \in \mathcal{X}^n$ d'un vecteur aléatoire $(X_1, ..., X_n)$. On suppose les composantes $X_1, ..., X_n$ indépendantes et de même loi μ inconnue. On dira que le vecteur $(X_1, ..., X_n)$ est un échantillon de variables indépendantes et identiquement distribuées, abrégé en iid.

Remarque 1.1. *On distingue l'échantillon $(X_1, ..., X_n)$ qui est aléatoire, et les données $(x_1, ..., x_n) = (X_1(\omega), ..., X_n(\omega))$ qui sont la réalisation d'un événement ω. Lors d'une étude statistique, on observe seulement l'image de ω par les fonctions $X_1, ..., X_n$.*

Contrairement aux statistiques descriptives qui consistent essentiellement à résumer l'information contenue dans des données à l'aide d'indicateurs (moyenne, médiane, variance etc...) ou de représentations graphiques (diagrammes, boxplots, histogrammes etc...), la statistique inférentielle cherche à fournir un modèle probabiliste qui explique les données observées. Les observations sont donc traitées comme des réalisations de variables aléatoires de loi inconnue, dont on cherche à retrouver certaines caractéristiques (ex: espérance, probabilité d'un événement, densité etc...).

La statistique inférentielle permet de répondre à différentes questions:

1. Le choix d'un modèle permettant de décrire le phénomène aléatoire
2. L'estimation de la loi des observations
3. La construction d'intervalles de confiances
4. La vérification d'hypothèses par des tests

Plus le modèle est précis, plus il sera facile d'évaluer la loi des observations. Par exemple, si on dispose d'un échantillon $X_1, ..., X_n$ issu d'une loi normale $\mathcal{N}(m, \sigma^2)$, il suffit d'estimer la moyenne et la variance des observations pour connaître la loi. En revanche, on prend le risque de considérer un modèle qui n'est pas exact. Si les observations ne sont pas Gaussiennes, on ne retrouvera jamais leur loi en se limitant à un modèle Gaussien.

Définition 1.2. *Un modèle $\mathcal{M}$ est un ensemble de lois possibles pour les observations. Le modèle est paramétrique s'il est indexé par un ensemble de dimension finie: $\mathcal{M} = \{\mu_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$. Dans le cas contraire, le modèle est non-paramétrique.*

Les familles de lois usuelles sont des exemples de modèles paramétriques (ex: modèle Gaussien $\mathcal{M} = \{\mathcal{N}(m, \sigma^2), m \in \mathbb{R}, \sigma^2 > 0\}$, modèle de Poisson $\mathcal{M} = \{\mathcal{P}(\lambda), \lambda > 0\}$, modèle exponentiel $\{\mathcal{E}(\lambda), \lambda > 0\}$ etc...). Les modèles non-paramétriques sont moins spécifiques et sont utilisés dans les situations où on dispose de peu d'information sur la loi μ des observations. Par exemple, μ est une loi continue (densité), μ admet un moment d'ordre deux, μ est une loi symétrique par rapport à zéro ou encore μ est une loi de probabilité (aucune information) sont des exemples de modélisation non-paramétriques.

Lorsqu'on choisit un modèle paramétrique, on prend le risque que la vraie loi des données n'appartienne pas au modèle. Un modèle non-paramétrique permet donc une évaluation parfois moins précise mais plus robuste de cette loi.

Définition 1.3. *Une statistique $T_n = T_n(X_1, ..., X_n)$ est une fonction (mesurable) connue de l'échantillon.*

Une statistique T_n est une variable aléatoire dont on observe une réalisation $t_n = T_n(x_1, ..., x_n)$.

Définition 1.4. *Un paramètre $\theta \in \Theta$ est une quantité inconnue qui dépend de la loi des observations.*

Formellement, un paramètre est l'image de μ par une fonctionnelle connue $\theta(.)$ (par exemple, l'espérance ou la médiane sous la loi μ). L'ensemble de définition du paramètre θ, noté Θ est donc l'image du modèle $\mathcal{M}$ par la fonctionnelle $\theta(.)$: $\Theta = \theta(\mathcal{M})$. Pour simplifier les notations, la dépendance en μ est souvent omise et θ désigne simplement une valeur de l'ensemble Θ, ou même la valeur associée à la vraie loi $\theta = \theta(\mu)$.

Une statistique utilisée pour estimer un paramètre θ est appelée *estimateur* et souvent notée $\hat{\theta}_n$. Un estimateur $\hat{\theta}_n$ d'un paramètre θ est dit *consistant* s'il converge en probabilité vers θ quand la taille de l'échantillon n tend vers l'infini, et ce quelque soit la vraie loi des observations. Si la convergence est presque sûre, alors $\hat{\theta}_n$ est dit fortement consistant.

S'il existe une suite réelles $(a_n)_{n\in\mathbb{N}}$ qui tend vers l'infini telle que la suite $a_n(\hat{\theta}_n - \theta)$ est bornée en probabilité, alors on dit que $\hat{\theta}_n$ est a_n-consistant. La vitesse de convergence d'un estimateur $\hat{\theta}_n$ est une suite $(a_n)_{n\in\mathbb{N}}$ qui vérifie

$$a_n\big(\hat{\theta}_n - \theta\big) = O_p(1) \quad \text{et} \quad a_n\big(\hat{\theta}_n - \theta\big) \neq o_p(1).$$

Remarque 1.5. *Si un estimateur $\hat{\theta}_n$ est tel que $a_n\big(\hat{\theta}_n - \theta\big)$ converge en loi vers une limite non dégénérée (penser notamment au TCL avec la moyenne empirique qui converge vers l'espérance pour $a_n = \sqrt{n}$), alors $\hat{\theta}_n$ est a_n-consistant. La réciproque n'est évidemment pas vraie, une suite bornée en probabilité ne converge pas forcément en loi.*

Le biais d'un estimateur $\hat{\theta}_n$ de θ est la quantité $b(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$ qui n'est définie que si l'espérance de T est définie, c'est-à-dire si $\mathbb{E}|\hat{\theta}_n| < \infty$. Si le biais est nul, on dira que T est un estimateur sans biais.

Remarque 1.6. *Il ne faut pas oublier que l'espérance $\mathbb{E}(\hat{\theta}_n)$ est calculée sous la loi μ des observations, qui est inconnue. Il est en fait sous-entendu que l'estimateur T de θ est sans biais si*

$$\forall \mu \in \mathcal{M} \ , \ \mathbb{E}_\mu(\hat{\theta}_n) = \theta(\mu).$$

Cette propriété ne dépend pas de la vraie valeur de μ et peut donc être connue par le statisticien en pratique.

L'erreur quadratique moyenne (EQM) d'un estimateur $\hat{\theta}_n$ de θ est la quantité

$$\mathrm{EQM}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n - \theta)^2 \in [0, +\infty].$$

On dit que $\hat{\theta}_n$ converge dans $\mathbb{L}^2$ si son EQM tend vers zéro (pour tout $\mu \in \mathcal{M}$).

Lemme 1.7 (Décomposition biais-variance)**.** *Pour tout estimateur $\hat{\theta}_n$ qui admet un moment d'ordre 2,*

$$\mathrm{EQM}(\hat{\theta}_n) = b(\hat{\theta}_n)^2 + \mathrm{var}(\hat{\theta}_n).$$

La formule de décomposition biais-variance permet de montrer immédiatement qu'un estimateur converge dans $\mathbb{L}^2$ si et seulement si son biais et sa variance tendent simultanément vers zéro. Comme la convergence dans $\mathbb{L}^2$ implique la convergence en probabilité, un estimateur qui converge en moyenne quadratique est consistant.

2 Inférence dans un modèle paramétrique

2.1 Méthode des moments

On rappelle que les moments d'une variable aléatoire réelle X sont donnés par

$$\mathsf{m}_p := \mathbb{E}(X^p) \ , \ p = 1, 2, ...$$

Le moment d'ordre p n'est défini que si la quantité $\mathbb{E}\big(|X|^p\big)$ est finie. On peut bien sûr généraliser cette définition à des moments p non-entiers.

Proposition 2.1. *Si X admet un moment d'ordre $p > 0$, alors X admet également un moment d'ordre q pour tout $q \in [0, p[$.*

Preuve. Soit $q \in [0, p[$, la fonction $\phi : x \mapsto x^{p/q}$ est convexe sur $[0, +\infty[$ puisque $p/q > 1$. Par l'inégalité de Jensen,

$$\mathbb{E}\big(|X|^p\big) = \mathbb{E}\Big(\big(|X|^q\big)^{p/q}\Big) \geq \Big(\mathbb{E}\big(|X|^q\big)\Big)^{p/q}$$

d'où le résultat. □

Les moments sont faciles à estimer à partir d'un échantillon iid $X_1, ..., X_n$. En effet, par la loi forte des grands nombres, on sait que si m_p existe alors

$$\widehat{\mathsf{m}}_p := \frac{1}{n}\sum_{i=1}^{n} X_i^p$$

converge presque sûrement vers m_p quand $n \to \infty$. Si de plus, X_1 admet un moment d'ordre $2p$, alors le TCL est également valable et on a

$$\sqrt{n}\big(\widehat{\mathsf{m}}_p - \mathsf{m}_p\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}\big(0, \mathsf{m}_{2p} - \mathsf{m}_p^2\big).$$

Remarque 2.2. *L'estimateur $\widehat{\mathsf{m}}_p$ correspond au moment d'ordre p d'une variable aléatoire de fonction de répartition F_n (la fonction de répartition empirique). On l'appelle le moment empirique d'ordre p.*

Supposons maintenant que la loi des observations appartienne à un modèle paramétrique $\mathcal{M} = \{\mu_\theta\ ,\ \theta \in \Theta\}$ pour lequel chacun des paramètres $\theta = (\theta_1, ..., \theta_k)$ s'exprime en fonction des moments m_p. Les estimateurs des moments $\widehat{\mathsf{m}}_p$ permettent donc d'estimer naturellement les paramètres du modèle par "injection", c'est-à-dire en remplaçant les moments théoriques m_p par leurs versions empiriques $\widehat{\mathsf{m}}_p$ dans l'expression de $\theta = \theta(\mathsf{m}_1, \mathsf{m}_2, ...)$.

Exemples

1. Soit $X_1, ..., X_n$ iid dans le modèle Gaussien $\mathcal{N}(\theta, \sigma^2)$. Comme θ est le moment d'ordre 1: $\mathsf{m}_1 = \mathbb{E}(X_i)$, un estimateur des moments de θ est

$$\widehat{\theta} = \widehat{\mathsf{m}}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

   De plus, on remarque que $\sigma^2 = \mathsf{m}_2 - \mathsf{m}_1^2$. On déduit un estimateur de σ^2 par la méthode des moments:

$$\widehat{\sigma}^2 = \widehat{\mathsf{m}}_2 - \widehat{\mathsf{m}}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

   par la loi forte des grands nombres (et continuité de la fonction $(x, y) \mapsto x - y^2$ pour le second).

2. Soit un échantillon iid $X_1, ..., X_n$ de loi de Poisson de paramètre $\lambda > 0$ inconnu. On sait que l'espérance de la loi de Poisson est égale au paramètre, on peut donc construire un estimateur de λ par

$$\hat{\lambda} = \widehat{\mathsf{m}}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Poussons le vice plus loin en remarquant que λ est également la variance dans le modèle de Poisson: $\lambda = \mathsf{m}_2 - \mathsf{m}_1^2$. On déduit un second estimateur de λ:

$$\widetilde{\lambda} = \widehat{\mathsf{m}}_2 - \widehat{\mathsf{m}}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

Ces deux estimateurs sont consistants.

3. On rappelle que l'espérance et la variance d'une variable aléatoire X de loi gamma $\Gamma(a,b)$ sont données par

$$\mathbb{E}(X) = \frac{a}{b} \quad \text{et} \quad \mathrm{var}(X) = \frac{a}{b^2}.$$

Soit $X_1, ..., X_n$ iid de loi gamma $\Gamma(a,b)$ avec $a, b > 0$. Pour estimer les paramètres a, b inconnus du modèle, on les exprime en fonction des moments en résolvant un système à (ici) deux équations

$$\begin{cases} \mathsf{m}_1 = \dfrac{a}{b} \\ \mathsf{m}_2 - \mathsf{m}_1^2 = \dfrac{a}{b^2} \end{cases} \iff \begin{cases} a = \dfrac{\mathsf{m}_1^2}{\mathsf{m}_2 - \mathsf{m}_1^2} \\ b = \dfrac{\mathsf{m}_1}{\mathsf{m}_2 - \mathsf{m}_1^2} \end{cases}$$

On déduit des estimateurs $\widehat{a}, \widehat{b}$ par la méthode des moments en remplaçant les moments théoriques $\mathsf{m}_1, \mathsf{m}_2$ par les versions empiriques (on remarquera que $\widehat{\mathsf{m}}_2 - \widehat{\mathsf{m}}_1^2$ est la variance empirique et donc strictement positive presque sûrement à partir de $n \geq 2$).

Comme on a pu le constater dans les exemples, il n'y a pas qu'une façon de construire des estimateurs par la méthode des moments, puisque l'expression d'un paramètre θ en fonction de moments n'est en général pas unique. En pratique, on choisira généralement l'expression issue des premiers moments par simplicité, mais aussi parce que les moments empiriques d'ordre petit ont une variance plus faible. Pour estimer k paramètres qui ne sont pas liés, il faut au minimum k moments et résoudre un système d'équations.

Proposition 2.3. *On suppose l'existence du moment d'ordre k. Soit un paramètre $\theta = g(\mathsf{m}_1, ..., \mathsf{m}_k)$ avec $g : \mathbb{R}^k \to \mathbb{R}$ une fonction continue. L'estimateur par la méthode des moments*

$$\widehat{\theta} := g\left(\widehat{\mathsf{m}}_1, ..., \widehat{\mathsf{m}}_k\right)$$

converge presque sûrement vers $\theta = g(\mathsf{m}_1, ..., \mathsf{m}_k)$.

Preuve. Par la LFGN, $\hat{\mathsf{m}}_j$ converge presque sûrement vers m_j pour tout $j = 1, ..., k$. Comme une intersection finie d'événements de probabilité 1 est de probabilité 1, le vecteur $(\mathsf{m}_1, ..., \mathsf{m}_k)$ converge lui-aussi presque sûrement. Par continuité de g (la continuité au point $(\mathsf{m}_1, ..., \mathsf{m}_k)$ suffit), $\widehat{\theta} \xrightarrow{ps} \theta$. $\square$

Remarque 2.4. *Alors que les moments empiriques sont bien des estimateurs sans biais des moments théoriques, rien n'indique que l'estimateur des moments $\widehat{\theta} = g\left(\hat{\mathsf{m}}_1, ..., \hat{\mathsf{m}}_k\right)$ reste sans biais dès que la fonction g n'est pas linéaire. C'est le cas par exemple de la variance empirique*

$$\widehat{\sigma}^2 = \widehat{\mathsf{m}}_2 - \widehat{\mathsf{m}}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2,$$

qui n'est pas sans biais (mais l'est asymptotiquement).

Sous des conditions de régularité supplémentaires sur g, on peut étudier plus en détail le comportement asymptotique d'un estimateur de la forme $\widehat{\theta} := g\left(\widehat{\mathsf{m}}_1, ..., \widehat{\mathsf{m}}_k\right)$. On utilise pour cela la Δ-méthode qui est basé sur un développement limité de g au point $(\mathsf{m}_1, ..., \mathsf{m}_k)$. Par souci de simplicité, on se limitera ici au cas $k = 1$.

Proposition 2.5 (Δ-méthode)**.** *On suppose l'existence du moment d'ordre* 2 *et on note* $\sigma^2 = \mathsf{m}_2 - \mathsf{m}_1^2$ *la variance. Soit* $\theta = g(\mathsf{m}_1)$ *pour une fonction* $g : \mathbb{R} \to \mathbb{R}$ *de classe* $\mathcal{C}^1$ *dans un voisinage de* m_1*. L'estimateur* $\widehat{\theta} := g\big(\widehat{\mathsf{m}}_1\big)$ *vérifie*

$$\sqrt{n}\big(\widehat{\theta} - \theta\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}\big(0, g'(\mathsf{m}_1)^2\sigma^2\big).$$

Preuve. Par le TCL, $\sqrt{n}(\widehat{\mathsf{m}}_1 - \mathsf{m}_1)$ converge en loi vers une $\mathcal{N}(0, \sigma^2)$. On applique la formule de Taylor-Lagrange à l'ordre 1,

$$\widehat{\theta} = g\big(\widehat{\mathsf{m}}_1\big) = \theta + (\widehat{\mathsf{m}}_1 - \mathsf{m}_1)g'(\xi_n),$$

pour ξ_n compris entre m_1 et $\widehat{\mathsf{m}}_1$ (on admettra que ξ_n est mesurable et donc bien une variable aléatoire). On a donc

$$\sqrt{n}\big(\widehat{\theta} - \theta\big) = \sqrt{n}(\widehat{\mathsf{m}}_1 - \mathsf{m}_1)g'(\xi_n).$$

Comme $\widehat{\mathsf{m}}_1$ est consistant, $g'(\xi_n) \xrightarrow{\mathbb{P}} g'(\mathsf{m}_1)$ par continuité de g' en m_1. On conclut par le lemme de Slutsky. $\square$

2.2 Maximum de vraisemblance

Dans un problème d'estimation paramétrique, une notion très utilisée est le *maximum de vraisemblance.* L'idée est de choisir comme estimateur la valeur du paramètre qui rend la plus probable l'observation de l'échantillon $x_1, ..., x_n$.

Un modèle paramétrique $\{\mu_\theta : \theta \in \Theta\}$ est dit *identifié* si l'application $\theta \mapsto \mu_\theta$ est injective sur Θ. Autrement dit, un modèle est identifié si chaque valeur du paramètre est associée à une unique loi du modèle.

Définition 2.6. *Soit* $\mathcal{M} = \{\mu_\theta, \theta \in \Theta\}$ *un modèle paramétrique identifié et* $X = (X_1, ..., X_n)$ *le vecteur des observations.*

1. *Si pour tout* $\theta \in \Theta$*,* μ_θ *est une loi discrète, la vraisemblance du modèle est la fonction*

$$\theta \mapsto V(\theta, X) = \prod_{i=1}^{n} \mu_\theta(\{X_i\}),\ \theta \in \Theta.$$

2. *Si pour tout* $\theta \in \Theta$*,* μ_θ *est une loi continue de densité* f_θ*, la vraisemblance du modèle est la fonction*

$$\theta \mapsto V(\theta, X) = \prod_{i=1}^{n} f_\theta(X_i),\ \theta \in \Theta.$$

Un estimateur du maximum de vraisemblance est une variable aléatoire qui maximise $\theta \mapsto V(\theta, X)$.

Exemples

1. Modèle binomial: on observe le résultat de n lancés de dés $X = (X_1, ..., X_n) \in \{0, 1\}^n$. Les X_i sont des réalisations indépendantes d'une variable de Bernoulli de paramètre p_0 inconnu. La probabilité d'observer l'échantillon est de

$$\prod_{i=1}^{n} p_0^{X_i}(1 - p_0)^{1-X_i} = p_0^S(1 - p_0)^{n-S},$$

où $S = \sum_{i=1}^n X_i$. La vraie valeur p_0 étant inconnue, on peut écrire cette probabilité comme une fonction de p:

$$V(p, X) = p^S (1-p)^{n-S}$$

qui est la probabilité d'observer l'échantillon $X = (X_1, ..., X_n)$ si les observations suivent la loi de Bernoulli de paramètre p (qui n'est autre que la vraisemblance du modèle). L'estimateur du maximum de vraisemblance est la valeur $\hat{p}$ qui maximise la probabilité $V(p, X)$,

$$\hat{p} = \arg\max_{p \in [0,1]} V(p, X).$$

On note $v(p, X) = \log V(p, X)$ la log-vraisemblance, maximiser $V(., X)$ revient à maximiser $v(., X)$. La fonction $p \mapsto v(p, X)$ est strictement concave sur $[0, 1]$, le maximiseur est donc l'unique solution de la condition du premier ordre

$$\frac{\partial}{\partial \theta} v(\hat{p}, X) = \frac{S}{\hat{p}} - \frac{n-S}{1-\hat{p}} = 0 \Longleftrightarrow \hat{p} = \frac{S}{n}.$$

Dans ce modèle l'estimateur du maximum de vraisemblance est donc défini comme la variable aléatoire $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$.

2. Modèle de Poisson: la vraisemblance pour un modèle de Poisson $\mathcal{M} = \{\mathcal{P}(\lambda), \lambda > 0\}$ est donnée par

$$V(\lambda, X) = e^{-n\lambda} \frac{\lambda^S}{\prod_{i=1}^n X_i!}.$$

Maximiser $V(\lambda, X)$ revient à maximiser $\lambda \mapsto \log(\lambda) S - n\lambda$. Le maximiseur est

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

3. Loi uniforme: on considère le modèle $\mathcal{M} = \{\mathcal{U}(0, \theta), \theta > 0\}$, qui sont des lois de densités

$$f_\theta(x) = \begin{cases} 1/\theta & \text{si } x \in [0, \theta] \\ 0 & \text{sinon.} \end{cases}$$

La vraisemblance du modèle est donnée par

$$V(\theta, X) = \begin{cases} \theta^{-n} & \text{si } \max\{X_1, ..., X_n\} \leq \theta \\ 0 & \text{sinon.} \end{cases}$$

La fonction $\theta \mapsto \theta^{-n}$ est décroissante sur $[0, +\infty[$, on déduit donc que la vraisemblance est maximale pour $\hat{\theta} = \max\{X_1, ..., X_n\}$.

4. Loi normale: ...

Le maximum de vraisemblance est stable par transformation des observations et du paramètre.

Proposition 2.7. *Soit un modèle paramétrique $\mathcal{M} = \{\mu_\theta : \theta \in \Theta\}$ et $\hat{\theta}$ un estimateur du maximum de vraisemblance de θ.*

i) Soit $g : \mathcal{X} \to \mathcal{Y}$ un $\mathcal{C}^1$ diffeomorphisme, $\hat{\theta}$ est un estimateur du maximum de vraisemblance pour les observations $Y_i = g(X_i)$.

ii) Soit $g : \Theta \to \mathrm{H}$ une fonction injective, $\hat{\eta} = g(\hat{\theta})$ est un estimateur du maximum de vraisemblance de $\eta = g(\theta)$.

Preuve. Supposons pour simplifier que X_i est de densité f_θ avec $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}$. La densité de Y_i est alors donnée par $h_\theta = f_\theta \circ g^{-1}/|g' \circ g^{-1}|$ par la formule du changement de variable. On montre alors facilement que

$$V(\theta, Y) = \prod_{i=1}^{n} h_\theta(Y_i) = \frac{1}{\prod_{i=1}^{n} |g'(X_i)|} V(\theta, X),$$

ce qui implique $i)$. La preuve se généralise facilement à des espaces $\mathcal{X}, \mathcal{Y}$ quelconques. Pour $ii)$, la vraisemblance de l'échantillon en fonction de η est donnée par

$$\eta \mapsto V(g^{-1}(\eta), X) \ , \ \eta \in H.$$

Il suffit de remarquer que

$$V(g^{-1}(\hat{\eta}), X) = V(\hat{\theta}, X) = \sup_{\theta \in \Theta} V(\theta, X) = \sup_{\eta \in \mathrm{H}} V(g^{-1}(\eta), X),$$

par l'injectivité de g. □

Une transformation du paramètre peut rendre un modèle paramétrique non identifié si la transformation g n'est pas injective. C'est évidemment à éviter, mais dans ce cas, on peut définir la vraisemblance par

$$\eta \mapsto \sup_{\theta : g(\theta) = \eta} V(\theta, X).$$

Sous des conditions (raisonnables mais techniques) de régularité sur le modèle, on peut montrer l'estimateur du maximum de vraisemblance existe, est unique et consistant. On donne les étapes du raisonnement (on suppose la loi des observations continue mais le raisonnement marche aussi dans le cas discret) . Soit $\ell_\theta(x) = \log(f_\theta(x))$, la log-vraisemblance s'écrit

$$v(\theta, X) = \sum_{i=1}^{n} \ell_\theta(X_i).$$

Par ailleurs, on note θ_0 la vraie valeur du paramètre et $g(\theta) = \int \ell_\theta(x) f_{\theta_0}(x) dx$.

- *Condition 1:* La famille de fonctions $\{\ell_\theta, \theta \in \Theta\}$ vérifie la loi des grands nombres uniformément en θ (c'est une classe de Glivenko-Cantelli sous f_{θ_0}) au sens où:

$$\sup_{\theta \in \Theta} \Big| \frac{1}{n} \sum_{i=1}^{n} \ell_\theta(X_i) - g(\theta) \Big| \xrightarrow[n \to \infty]{ps} 0.$$

Pour garantir la consistance de $\hat{\theta}_{MV}$, on a maintenant besoin que le modèle soit fortement identifié dans le sens où ℓ_θ ne peut pas se rapprocher arbitrairement de ℓ_{θ_0} sans que $\theta \to \theta_0$.

- *Condition 2:* Soit X de densité f_{θ_0}, pour toute suite $(\theta_k)_{k \in \mathbb{N}}$ dans Θ,

$$\ell_{\theta_k}(X) \xrightarrow[k \to \infty]{ps} \ell_{\theta_0}(X) \implies \theta_k \xrightarrow[k \to \infty]{} \theta_0.$$

Cette condition est indispensable si on veut pouvoir retrouver θ_0 à partir des $\ell_\theta(X_i), \theta \in \Theta$. Par l'inégalité de Jensen, elle implique en particulier que la seule façon de maximiser la fonction g est de converger vers $\theta = \theta_0$. En combinant les conditions 1 et 2, $v(., X)$ converge uniformément (presque sûrement) vers une fonction g maximale seulement en θ_0, ce qui entraîne la convergence presque sûre de $\hat{\theta}_{MV}$ vers θ_0.
Pour obtenir le comportement asymptotique à l'ordre 1 de $\hat{\theta}_{MV}$, on écrit le développement de Taylor-Lagrange de la dérivée de la log-vraisemblance.

- *Condition 3:* $\theta \mapsto \ell_\theta$ est de classe $\mathcal{C}^2$ pour (presque) tout x.

En posant $s_\theta(x) = \partial \ell_\theta(x)/\partial\theta$ et $h_\theta(x) = \partial^2 \ell_\theta(x)/\partial\theta^2$, on a

$$\frac{1}{n}\sum_{i=1}^n s_{\hat\theta_{MV}}(X_i) = \frac{1}{n}\sum_{i=1}^n s_{\theta_0}(X_i) + \big(\hat\theta_{MV} - \theta_0\big)\frac{1}{n}\sum_{i=1}^n h_{\tilde\theta}(X_i)$$

avec $\tilde\theta$ compris entre θ_0 et $\hat\theta_{MV}$. Sous réserve que la condition du premier ordre $\sum_{i=1}^n s_{\hat\theta_{MV}}(X_i) = 0$ soit vérifiée, on trouve

$$\sqrt{n}\big(\hat\theta_{MV} - \theta_0\big) = -\frac{\sqrt{n}}{n}\sum_{i=1}^n s_{\theta_0}(X_i) \times \frac{1}{\frac{1}{n}\sum_{i=1}^n h_{\tilde\theta}(X_i)}.$$

On peut alors conclure à condition que $\frac{1}{n}\sum_{i=1}^n h_{\tilde\theta}(X_i)$ converge bien vers la limite attendue.

- *Condition 4:*$\{h_\theta, \theta \in \Theta\}$ est une classe de Glivenko-Cantelli sous f_{θ_0}.

Soit $\mathcal{I}(\theta_0) = -\int h_{\theta_0}(x) f_{\theta_0}(x)dx$, on montre sous des conditions de dérivabilité sous le signe intégral:

$$\frac{\sqrt{n}}{n}\sum_{i=1}^n s_{\theta_0}(X_i) \xrightarrow[n\to\infty]{loi} \mathcal{N}\Big(0, \mathcal{I}(\theta_0)\Big) \quad \text{et} \quad \frac{1}{n}\sum_{i=1}^n h_{\tilde\theta}(X_i) \xrightarrow[n\to\infty]{ps} -\mathcal{I}(\theta_0).$$

On conclut par le lemme de Slutsky:

$$\sqrt{n}\big(\hat\theta_{MV} - \theta_0\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}\Big(0, \mathcal{I}(\theta_0)^{-1}\Big).$$

3 Estimation non-paramétrique

On se place maintenant dans le cadre non-paramétrique, où peu (voire aucune) information sur la loi des observations n'est disponible. On définit donc implicitement comme modèle $\mathcal{M}$ l'ensemble des lois sur $\mathbb{R}$.

3.1 Fonction de répartition

On rappelle que la fonction de répartition associée à une loi μ est donnée par

$$F(x) = \mu\big(]-\infty, x]\big) = \mathbb{P}(X \leq x),\ x \in \mathbb{R}.$$

La fonction de répartition caractérise la loi, c'est-à-dire que deux lois μ et ν sont égales si, et seulement si, leurs fonctions de répartition sont égales. On rappelle qu'une fonction de répartition est une fonction croissante et telle que $\lim_{x\to-\infty} F(x) = 0$ et $\lim_{x\to+\infty} F(x) = 1$. De plus, F est continue à droite et limitée à gauche (càdlàg) avec

$$\forall x \in \mathbb{R},\ \lim_{y\to x^+} F(y) = F(x) \quad \text{et} \quad \lim_{y\to x^-} F(y) = F(x^-) = \mathbb{P}(X < x).$$

On a donc en particulier $\mathbb{P}(X = x) = F(x) - F(x^-)$. Par ailleurs, lorsque μ admet une densité f continue, la fonction de répartition vaut

$$F(x) = \int_{-\infty}^x f(y)dy,\ x \in \mathbb{R}.$$

Dans ce cas F est de classe $\mathcal{C}^1$ de dérivée $F' = f$ par le théorème fondamental de l'analyse.

Pour estimer la fonction de répartition en un point x, on utilise la proportion des observations inférieures à x,

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\},\ x \in \mathbb{R},$$

où $\mathbb{1}$ est la fonction indicatrice. La fonction F_n est appelée *fonction de répartition empirique.*

Proposition 3.1. *Pour tout $\omega \in \Omega$, l'estimation $F_n : \mathbb{R} \to [0,1]$ est une fonction de répartition.*

Preuve. Soit $\omega \in \Omega$, on pose $(x_1, ..., x_n) = (X_1(\omega), ..., x_n(\omega))$. La fonction de répartition empirique associée est

$$\big(F_n(x)\big)(\omega) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i \leq x\},\ x \in \mathbb{R}.$$

Montrons qu'elle vérifie les propriétés d'une fonction de répartition. Pour tout $x \geq \max\{x_1, ..., x_n\}$, $F_n(x) = 1$. De même, pour tout $x < \min\{x_1, ..., x_n\}$, $F_n(x) = 0$. On a donc clairement

$$\lim_{x\to-\infty} F_n(x) = 0 \quad \text{et} \quad \lim_{x\to+\infty} F_n(x) = 1.$$

Pour tout i, la fonction $x \mapsto \mathbb{1}\{x_i \leq x\}$ est croissante, continue sur $\mathbb{R} \setminus \{x_i\}$ et càdlàg en x_i. On déduit que F_n est croissante et càdlàg comme somme finie de fonctions croissantes càdlàg. On a précisemment en x_i,

$$\lim_{x\to x_i^+} F_n(x) = \frac{\#\{j : x_j \leq x_i\}}{n} = F_n(x_i)$$

et F_n admet une limite à gauche

$$\lim_{x\to x_i^-} F_n(x) = \frac{\#\{j : x_j < x_i\}}{n} = F_n(x_i) - \frac{\#\{j : x_j = x_i\}}{n}. \qquad \square$$

Remarque 3.2. *L'ensemble $\{j : x_j = x_i\}$ admet au moins un élément: x_i. La fonction de répartition empirique F_n n'est donc jamais continue en x_i puisque la limite à gauche est strictement inférieure à la limite à droite.*

Lorsque tous les x_i sont différents (ce qui arrive presque sûrement lorsqu'ils sont de loi continue), la fonction de répartition empirique F_n obtenue pour une réalisation ω est la fonction de répartition de la loi uniforme discrète sur $\{x_1, ..., x_n\}$, parfois appelée loi empirique. On approche donc en quelque sorte la loi des observations par une loi discrète sur l'ensemble des réalisations. En particulier, $F_n(x)$ ne peut prendre que $n+1$ valeurs différentes: $0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1$.

Proposition 3.3. *Pour tout $x \in \mathbb{R}$, $F_n(x)$ est un estimateur sans biais de $F(x)$ qui converge presque sûrement vers $F(x)$ quand $n \to \infty$.*

Preuve. Soit $Y_i = \mathbb{1}\{X_i \leq x\}$, les variables aléatoires $Y_1, ..., Y_n$ sont indépendantes de loi de Bernoulli $\mathcal{B}(F(x))$. On a donc en particulier, $\mathbb{E}(Y_i) = F(x)$ et $\mathrm{var}(Y_i) = F(x)(1 - F(x)) < \infty$. On déduit

$$\mathbb{E}(F_n(x)) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} Y_i\Big) = \mathbb{E}\left(Y_1\right) = F(x).$$

Par la loi forte des grands nombres, $F_n(x)$ converge presque sûrement vers $F(x)$. $\square$

La loi forte des grands nombres permet d'établir facilement la convergence presque sûre de la fonction de répartition empirique vers la vraie fonction de répartition en tout point. Le théorème de Glivenko-Cantelli donne un résultat plus fort, à savoir, la convergence uniforme presque sûre de la fonction de répartition empirique.

Théorème 3.4 (Glivenko-Cantelli). *Soit $X_1, .., X_n$ un échantillon iid de fonction de répartition continue F. La fonction de répartition empirique F_n converge presque sûrement uniformément vers F quand $n \to \infty$:*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{p.s.} 0.$$

Pour la preuve, on aura besoin du lemme suivant.

Lemme 3.5. *Si X est une variable aléatoire réelle de fonction de répartition F, alors pour tout $x \in \mathbb{R}$,*

$$\mathbb{1}\{X \leq x\} \overset{ps}{=} \mathbb{1}\{F(X) \leq F(x)\}.$$

En particulier, si F est continue, alors $U := F(X)$ est de loi uniforme sur $]0, 1[$.

Preuve du lemme. Comme F est croissante, on sait que

$$\{F(X) \leq F(x)\} = \{X \leq x\} \cup \{F(X) \leq F(x), x < X\}.$$

Soit $u = \sup\{y : F(y) \leq F(x)\}$, on distingue deux cas. Si $F(u) = F(x)$ alors

$$\mathbb{P}(F(X) \leq F(x), x < X) = \mathbb{P}(F(X) \leq F(x), x < X \leq u) \leq F(u) - F(x) = 0.$$

Sinon, soit u_n une suite croissante qui tend vers u et telle que $F(u_n) = F(x)$ pour tout $n \in \mathbb{N}$. On a

$$\begin{aligned}\mathbb{P}(F(X) \leq F(x), x < X) &= \lim_{n\to\infty} \mathbb{P}(F(X) \leq F(x), x < X \leq u_n)\\ &\leq \lim_{n\to\infty} F(u_n) - F(x) = 0,\end{aligned}$$

où la première égalité découle de l'axiome des probabilité $\lim_{n\to\infty} \mathbb{P}(\cup_{k=1}^n A_k) = \mathbb{P}(\cup_{n=1}^\infty A_n)$. On déduit $\mathbb{1}\{X \leq x\} \overset{ps}{=} \mathbb{1}\{F(X) \leq F(x)\}$. Si F est continue, du fait que $\lim_{x\to-\infty} F(x) = 0$ et $\lim_{x\to\infty} F(x) = 1$, il existe, pour tout $u \in]0, 1[$, un (ou plusieurs) $x \in \mathbb{R}$ tel que $F(x) = u$. Par conséquent, l'égalité $\mathbb{P}(F(X) \leq u) = u$ est vérifiée pour tout $u \in]0, 1[$, ce qui prouve que $F(X)$ est de loi uniforme sur $]0, 1[$. □

La preuve du théorème de Glivenko-Cantelli utilise également le théorème suivant, dû à Polyà (parfois appelé théorème de Dini) que l'on admettra.

Théorème 3.6. *Une suite F_n de fonctions croissantes définies sur un compact $[a, b] \subset \mathbb{R}$ qui converge simplement vers une fonction F continue sur $[a, b]$, converge uniformément vers F.*

Preuve du théorème de Glivenko-Cantelli. D'après le lemme 3.5, on a

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \sup_{x\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i \leq x\} - F(x) \right| \overset{ps}{=} \sup_{x\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{F(X_i) \leq F(x)\} - F(x) \right|.$$

Par hypothèse, on sait que $F(x)$ peut prendre toutes les valeurs de l'intervalle $]0, 1[$. En posant $u = F(x)$ et $U_i = F(X_i)$ qui suit une loi uniforme sur $[0, 1]$, on a donc

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \overset{ps}{=} \sup_{u\in[0,1]} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{U_i \leq u\} - u \right|,$$

les valeurs en $u = 0$ et $u = 1$ étant trivialement 0. Il suffit de montrer le résultat pour un échantillon $U_1, ..., U_n$ de variables aléatoires uniformes sur $[0, 1]$. On suppose donc sans perte de généralité que

$$F_n(u) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{U_i \leq u\} \quad \text{et} \quad F(u) = u,\ u \in [0, 1].$$

Soit $q \in \mathbb{Q} \cap]0,1[$, on sait par la proposition 3.3 que $\mathbb{P}(F_n(q) \to q) = 1$. Comme une union dénombrable d'ensembles de probabilité nulle est de probabilité nulle, on déduit que l'événement $\Omega_0 := \{\forall q \in \mathbb{Q} \cap]0,1[, F_n(q) \to F(q)\}$ est de probabilité 1. Soit $u \in]0,1[$ et $(u_k)_{k\in\mathbb{N}}$ et $(v_k)_{k\in\mathbb{N}}$ deux suites rationnelles adjacentes de limite u telles que $u_k \leq u \leq v_k$, $\forall k \in \mathbb{N}$. La croissance de F_n entraîne

$$\forall k \in \mathbb{N},\ F_n(u_k) \leq F_n(u) \leq F_n(v_k).$$

En faisant tendre n vers l'infini, on déduit que sur Ω_0,

$$F(u_k) \leq \liminf_{n\to\infty} F_n(u) \leq \limsup_{n\to\infty} F_n(u) \leq F(v_k).$$

Comme le résultat est vrai pour tout k, il reste vrai en faisant tendre k vers l'infini. On sait de plus que $F_n(0) = 0$ et $F_n(1) = 1$ presque sûrement. On obtient $\Omega_0 = \{\forall u \in [0,1], F_n(u) \to u\}$. Finalement, d'après le théorème 3.6, F_n converge uniformément vers F sur Ω_0. □

Remarque 3.7. *Le théorème de Glivenko-Cantelli reste vrai lorsque F n'est pas continue mais la preuve précédente n'est plus valable dans ce cas.*

On peut montrer que la loi de $K_n := \sqrt{n}\ \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ ne dépend pas de F lorsque celle-ci est continue. Cette loi dépend de n mais converge vers la loi de Kolmogorov quand $n \to \infty$. Ce résultat qui est une conséquence du théorème de Donsker, est notamment utile pour la construction du test de Kolmogorov-Smirnov.

3.2 Quantiles

Définition 3.8. *Soit $\alpha \in]0,1[$, on appelle quantile d'ordre α d'une variable aléatoire X, tout réel q_α vérifiant*

$$\mathbb{P}(X \leq q_\alpha) \geq \alpha \quad \textit{et} \quad \mathbb{P}(X \geq q_\alpha) \geq 1 - \alpha.$$

Si X admet une densité f, alors on sait que $]0,1[\subseteq F(\mathbb{R})$, ce qui implique qu'il existe pour tout $\alpha \in]0,1[$, un quantile q_α tel que

$$\mathbb{P}(X \leq q_\alpha) = \alpha \quad \text{et} \quad \mathbb{P}(X \geq q_\alpha) = 1 - \alpha.$$

Si de plus F est bijective, le quantile d'ordre α est unique pour tout $\alpha \in]0,1[$.

Remarque 3.9. *La médiane est le quantile d'ordre* 0.5. *Les premier et troisième quartiles sont les quantiles d'ordre* 0.25 *et* 0.75 *respectivement.*

Définition 3.10. *La fonction quantile F^- est l'inverse généralisée de F définie par*

$$F^-(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\},\ u \in]0,1[.$$

La fonction quantile F^- est toujours définie, même lorsque F n'est pas continue ou est constante sur un intervalle. Le graphe de F^- est le symétrique par rapport à la première bissectrice du graphe de F pour lequel une partie constante devient un point de discontinuité.

Proposition 3.11. *$F^-(\alpha)$ est le plus petit quantile d'ordre α. Si de plus, F est bijective d'un intervalle $]a,b[$ (avec $a, b \in \mathbb{R} \cup \{-\infty, +\infty\}$) dans $]0,1[$, la fonction quantile est la fonction réciproque F^{-1} de F.*

Lemme 3.12. *Soit U une variable aléatoire de loi uniforme sur $]0,1[$, $F^-(U)$ est une variable aléatoire de fonction de répartition F.*

Preuve. Soit $(x_n)_{n\in\mathbb{N}}$ une suite décroissante de $\{x \in \mathbb{R} : F(x) \geq u\}$ qui converge vers sa borne inférieure $F^-(u)$. Par construction, $F(x_n) \geq u$ et par la continuité à droite de F,

$$F \circ F^-(u) = \lim_{n\to\infty} F(x_n) \geq u.$$

Ainsi, pour tout $x \in \mathbb{R}$,

$$\mathbb{P}(F^-(U) \leq x) \leq \mathbb{P}\big(F \circ F^-(U) \leq F(x)\big) \leq \mathbb{P}\big(U \leq F(x)\big) = F(x).$$

Soit $v > u$, le fait que F est croissante entraîne $\{x \in \mathbb{R} : F(x) \geq v\} \subseteq \{x \in \mathbb{R} : F(x) \geq u\}$ et

$$F^-(v) = \inf\{x \in \mathbb{R} : F(x) \geq v\} \geq \inf\{x \in \mathbb{R} : F(x) \geq u\} = F^-(u),$$

F^- est donc croissante. Comme $F^- \circ F(x) = \inf\{y \in \mathbb{R} : F(y) \geq F(x)\} \leq x$, on conclut

$$F(x) = \mathbb{P}\big(U \leq F(x)\big) \leq \mathbb{P}\big(F^-(U) \leq F^- \circ F(x)\big) \leq \mathbb{P}\big(F^-(U) \leq x\big). \qquad \square$$

Remarque 3.13. *Contrairement au lemme 3.5, ce résultat ne nécessite pas d'hypothèse de continuité et est valable pour toute fonction de répartition F.*

Pour l'estimation des quantiles à partir d'un échantillon iid $X_1, ..., X_n$, on définit la fonction de quantile empirique

$$F_n^-(u) = \inf\{x \in \mathbb{R} : F_n(x) \geq u\}.$$

C'est la fonction de quantile associée à la fonction de répartition empirique.

Définition 3.14. *Le quantile empirique d'ordre α est la variable aléatoire $\hat{q}_\alpha = F_n^-(\alpha)$.*

Le quantile empirique d'ordre α peut se définir également à partir des statistiques d'ordre de l'échantillon $X_1, ..., X_n$.

Définition 3.15. *Les statistiques d'ordre $X_{(1)}, ..., X_{(n)}$ sont les valeurs ordonnées dans l'ordre croissant de l'échantillon $X_1, ..., X_n$.*

Les statistiques d'ordre $X_{(k)}$ pour $k = 1, ..., n$ sont des variables aléatoires qui vérifient par construction $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$. Dans la suite, on notera n_α le plus petit entier supérieur ou égal à $n\alpha$.

Proposition 3.16. *Pour tout $\alpha \in]0, 1[$, $\hat{q}_\alpha = X_{(n_\alpha)}$.*

Preuve. F_n est constante par morceaux et vérifie $F_n(x) = 0$ si $x < X_{(1)}$, $F_n(x) = 1$ si $x \geq X_{(n)}$ et

$$F_n(x) = \frac{k}{n}\ ,\ x \in [X_{(k)}, X_{(k+1)}[\ ,\ k = 1, ..., n-1,$$

avec potentiellement $[X_{(k)}, X_{(k+1)}[= \emptyset$ si $X_{(k)} = X_{(k+1)}$. On a en particulier $F_n(X_{(n_\alpha)}) \geq n_\alpha/n \geq \alpha$ et $\forall x < X_{(n_\alpha)}$, $F_n(x) \leq (n_\alpha - 1)/n < \alpha$. Ainsi,

$$F_n^-(\alpha) = \inf\{x \in \mathbb{R} : F_n(x) \geq \alpha\} = X_{(n_\alpha)}. \qquad \square$$

Le quantile empirique est un quantile de l'ensemble $\{x_1, ..., x_n\}$. Par exemple, la médiane empirique $F_n^-(0.5)$ vaut $X_{(\frac{n}{2}+1)}$ lorsque n est pair et $X_{(\frac{n+1}{2})}$ lorsque n est impair.

Remarque 3.17. *Le quantile empirique $\hat{q}_\alpha = F_n^-(\alpha)$ converge presque sûrement vers $q_\alpha = F^-(\alpha)$ quand $n \to \infty$. La preuve (admise) est une conséquence du théorème de Glivenko-Cantelli.*

Théorème 3.18. *Supposons $X_1, ..., X_n$ iid de densité f, la k-ième statistique d'ordre $X_{(k)}$ a pour densité*

$$f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}\big(1-F(x)\big)^{n-k} f(x),$$

pour tout x point de continuité de f.

Preuve. Soit $h > 0$, on a pour tout $x \in \mathbb{R}$

$$\begin{aligned}
\mathbb{P}(x < X_{(k)} \leq x+h) &= \mathbb{P}\Big(\#\{i : X_i \leq x\} \leq k-1,\ \#\{i : X_i \leq x+h\} \geq k\Big) \\
&= \sum_{j=0}^{k-1} \sum_{\ell=k-j}^{n-j} \mathbb{P}\Big(\#\{i : X_i \leq x\} = j,\ \#\{i : x < X_i \leq x+h\} = \ell\Big) \\
&= \sum_{j=0}^{k-1} \sum_{\ell=k-j}^{n-j} \binom{n}{j}\binom{n-j}{\ell} F(x)^j \big(F(x+h)-F(x)\big)^{\ell}\big(1-F(x)\big)^{n-j-\ell}.
\end{aligned}$$

On déduit

$$f_{(k)}(x) = \lim_{h\to 0} \frac{1}{h}\ \mathbb{P}(x < X_{(k)} \leq x+h) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}\big(1-F(x)\big)^{n-k} f(x),$$

les termes pour $\ell \geq 2$ étant négligeables quand $h \to 0$. □

Le quantile empirique $\hat{q}_\alpha = X_{(n_\alpha)} = F_n^-(\alpha)$ est un estimateur du quantile $q_\alpha = F^-(\alpha)$ qui est consistant si F est bijective sur un voisinage de q_α. Dans ce cas, on a même la convergence en loi de la statistique renormalisée $\sqrt{n}(\hat{q}_\alpha - q_\alpha)$ vers une loi normale, comme énoncé dans la proposition suivante que l'on admettra.

Proposition 3.19. *Si F est de classe $\mathcal{C}^1$ dans un voisinage de q_α et $F'(q_\alpha) = f(q_\alpha) > 0$, alors*

$$\sqrt{n}(\hat{q}_\alpha - q_\alpha) \xrightarrow[n\to\infty]{loi} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f(q_\alpha)^2}\right).$$

4 Intervalles de confiance

Jusqu'à présent, nous avons vu des méthodes d'estimation qui ont pour objectif, étnat donné un paramètre inconnu, de proposer une valeur en fonction des observations. Plutôt que de se limiter à une valeur, un point de vue différent consiste à proposer un ensemble de valeurs probables.

Définition 4.1. *On appelle intervalle de confiance pour θ de niveau $1-\alpha \in [0,1]$, un intervalle aléatoire $I_\alpha = I_\alpha(X)$ fonction des observations, tel que*

$$\mathbb{P}(I_\alpha \ni \theta) = 1-\alpha.$$

Idéalement, l'égalité $\mathbb{P}(I_\alpha \ni \theta) = 1-\alpha$ doit être vraie sous toutes les lois μ du modèle. Cette condition est d'autant plus difficile à vérifier que le modèle est grand.

Le statisticien choisit à l'avance la valeur de $\alpha \in [0,1]$ qui correspond à la marge d'erreur qu'il s'autorise pour donner un encadrement du paramètre. Plus la marge d'erreur est faible, plus l'intervalle est grand.

Plus généralement, tout sous-ensemble $R_\alpha = R_\alpha(X) \subset \Theta$ tel que $\mathbb{P}(R_\alpha \ni \theta) = 1-\alpha$ est appelée région de confiance de niveau $1-\alpha$. Il n'existe pas vraiment de méthode générale pour construire un intervalle (ou une région) de confiance.

Définition 4.2. *Une fonction $g : \mathbb{R}^n \times \Theta \to \mathbb{R}$ est pivotale si la loi de $g(X, \theta)$ ne dépend pas de la loi des observations (ne dépend pas de θ dans un modèle paramétrique $\mathcal{M} = \{\mu_\theta : \theta \in \Theta\}$).*

Exhiber une fonction pivotale suffit à construire une région de confiance.

Proposition 4.3. *Si g est pivotale, alors pour tout ensemble $G_\alpha \subset \mathbb{R}$ tel que $\mathbb{P}\big(g(X, \theta) \in G_\alpha\big) = 1 - \alpha$, la région*

$$R_\alpha(X) := \{\theta \in \Theta : g(X, \theta) \in G_\alpha\}$$

est une région de confiance de niveau $1 - \alpha$.

Preuve. Il suffit de remarquer que $\mathbb{P}\big(\theta \in R_\alpha(X)\big) = \mathbb{P}\big(g(X, \theta) \in G_\alpha\big) = 1 - \alpha$. □

Remarque 4.4. *On peut construire un ensemble G_α vérifiant $\mathbb{P}\big(g(X, \theta) \in G_\alpha\big) = 1 - \alpha$ précisément parce que g est pivotale est donc que la loi de $g(X, \theta)$ est connue et ne dépend pas de θ.*

Prenons un exemple simple: soit $X_1, ..., X_n$ iid dans le modèle exponentiel $\mathcal{M} = \{\mathcal{E}(\theta) : \theta > 0\}$. On sait que $\theta X_i \sim \mathcal{E}(1)$. En sommant sur tout l'échantillon, on obtient une loi $\Gamma(n, 1)$. La fonction

$$g(X, \theta) := \theta \sum_{i=1}^{n} X_i$$

est donc pivotale puisque la loi de $g(X, \theta)$ ne dépend pas de θ (et donc de la loi des observations). Soit $q_\cdot$ le quantile de la loi $\Gamma(n, 1)$, on sait que pour tout $t \in]0, \alpha[$,

$$\mathbb{P}\Big(q_t \leq \theta \sum_{i=1}^{n} X_i \leq q_{1-\alpha+t}\Big) = 1 - \alpha.$$

On déduit l'intervalle de confiance $I_\alpha = [q_t / \sum_{i=1}^n X_i, q_{1-\alpha+t} / \sum_{i=1}^n X_i]$. En faisant varier t, on peut alors chercher un intervalle de confiance de longueur minimale, pour obtenir un encadrement le plus précis possible de θ au niveau α fixé.

4.1 Le modèle Gaussien

On suppose ici que les observations sont iid de loi normale $\mathcal{N}(\mathsf{m}, \sigma^2)$. On veut construire un intervalle de confiance sur la moyenne m et/ou la variance σ^2.

1. *Moyenne avec variance connue.* On suppose que les observations $X_1, ..., X_n$ sont distribuées suivant une loi normale de variance σ^2 connue et de moyenne m inconnue. Dans ce modèle on sait que

$$\hat{\mathsf{m}} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathcal{N}\Big(\mathsf{m}, \frac{\sigma^2}{n}\Big).$$

La loi de l'estimateur est connue en fonction de m. En particulier, on peut se ramener à une variable aléatoire normale centrée réduite par

$$g(\mathsf{m}, X) = \sqrt{n}\ \frac{\hat{\mathsf{m}} - \mathsf{m}}{\sigma} \sim \mathcal{N}(0, 1).$$

La fonction $g(\mathsf{m}, X) = \sqrt{n}(\hat{\mathsf{m}} - \mathsf{m})/\sigma$ est donc pivotale puisque sa loi ne dépend pas de m. On insiste sur le fait que l'on n'observe pas une réalisation de $g(\mathsf{m}, X)$ puisque c'est une fonction du paramètre inconnu m. La symétrie de la densité Gaussienne nous permet d'écrire

$$\mathbb{P}(-q_{1-\frac{\alpha}{2}} \leq g(\mathsf{m}, X) \leq q_{1-\frac{\alpha}{2}}) = 1 - \alpha,$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite. On cherche maintenant à isoler m dans l'expression de $g(\mathsf{m}, X)$ de manière à en obtenir un encadrement. On écrit

$$\mathbb{P}\Big(-q_{1-\frac{\alpha}{2}} \leq \sqrt{n}\,\frac{\hat{\mathsf{m}}-\mathsf{m}}{\sigma} \leq q_{1-\frac{\alpha}{2}}\Big) = \mathbb{P}\Big(\hat{\mathsf{m}} - \frac{q_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}} \leq \mathsf{m} \leq \hat{\mathsf{m}} + \frac{q_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}}\Big) = 1-\alpha.$$

On remarque que les bornes de l'intervalles sont fonctions des observations et donc connues. On obtient ainsi un intervalle de confiance pour m avec une marge d'erreur de α qui correspond à la probabilité que la vraie valeur du paramètre ne soit pas dans l'intervalle.

2. *Moyenne avec variance inconnue.* Ici, on ne peut pas utiliser la statistique précédente car elle dépend de σ^2. On remplace donc la variance par l'estimateur sans biais de $\hat{\sigma}^2$ donné par

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mathsf{m}})^2.$$

On utilise alors les propriétés suivantes, que l'on admettra

- $(n-1)\hat{\sigma}^2/\sigma^2$ suit une loi du Khi-2 à $n-1$ degrés de liberté, notée $\chi^2(n-1)$
- $\hat{\sigma}^2$ et $\hat{\mathsf{m}}$ sont indépendantes.

On déduit que la fonction

$$g(\mathsf{m}, X) = \sqrt{n}\,\frac{\hat{\mathsf{m}}-\mathsf{m}}{\hat{\sigma}} = \frac{\sqrt{n}(\hat{\mathsf{m}}-\mathsf{m})/\sigma}{\hat{\sigma}/\sigma} \overset{loi}{=} \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2(n-1)/(n-1)}} \Big\} \textit{ indépendantes}$$

est pivotale. Sa loi est appelée loi de Student à $n-1$ degrés de liberté et notée $\mathcal{T}(n-1)$. On peut alors construire un intervalle de confiance de la même façon que pour le cas σ^2 connu remplaçant les quantiles de la loi normale standard par ceux de la loi de Student à $n-1$ degrés de liberté.

3. *Variance avec moyenne connue.* On sait $(X_i - \mathsf{m})/\sigma$ est de loi $\mathcal{N}(0,1)$. Si m est connu, la fonction

$$g(\sigma^2, X) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mathsf{m})^2$$

est donc pivotale, de loi $\chi^2(n)$. Soit $k_{\alpha,n}$ le quantile d'ordre α de la loi du Khi deux à n degrés de liberté, qui vérifie $\mathbb{P}(\chi^2(n) \geq k_{\alpha,n}) = 1-\alpha$. On a

$$\mathbb{P}(k_{\frac{\alpha}{2},n} \leq g(\sigma^2, X) \leq k_{1-\frac{\alpha}{2},n}) = 1-\alpha \Longleftrightarrow \mathbb{P}\left(\frac{\sum_{i=1}^{n}(X_i-\mathsf{m})^2}{k_{1-\frac{\alpha}{2},n}} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n}(X_i-\mathsf{m})^2}{k_{\frac{\alpha}{2},n}}\right) = 1-\alpha.$$

4. *Variance avec moyenne inconnue.* On a vu que $\hat{\sigma}^2$ vérifie que $(n-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-1)$. On obtient donc un intervalle de confiance en utilisant les quantiles de la loi $\chi^2(n-1)$ (et non $\chi^2(n)$ comme dans l'exemple précédent).

4.2 Intervalles de confiance asymptotiques

Dans beaucoup de modèles, il est difficile d'exhiber une fonction pivotale. On se contente alors de fonction *asymptotiquement pivotale*, c'est à dire dont la loi quand $n \to \infty$ ne dépend plus de la loi des observations. Dans un modèle paramétrique $\mathcal{M} = \{\mu_\theta : \theta \in \Theta\}$, on construit facilement une fonction asymptotiquement pivotale à partir d'un estimateur asymptotiquement Gaussien de θ. On note q_α le quantile d'ordre α de la loi $\mathcal{N}(0,1)$.

Proposition 4.5. *Soit θ un paramètre et $\hat{\theta}$ un estimateur asymptotiquement Gaussien de θ avec*

$$\sqrt{n}\big(\hat{\theta}-\theta\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}\big(0,\sigma^2\big),$$

avec $\sigma^2<\infty$. Si $\hat{\sigma}^2$ est un estimateur consistant de σ^2, alors pour tout $\alpha\in]0,1[$, pour tout $t\in]0,\alpha[$

$$I_\alpha=\Big[\hat{\theta}+q_t\frac{\hat{\sigma}}{\sqrt{n}}\ ,\ \hat{\theta}+q_{1-\alpha+t}\frac{\hat{\sigma}}{\sqrt{n}}\Big]$$

est un intervalle de confiance pour θ de niveau $1-\alpha$ asymptotiquement.

Preuve. Par le lemme de Stutsky, $S=\sqrt{n}\big(\hat{\theta}-\theta\big)/\hat{\sigma}\xrightarrow{loi}\mathcal{N}(0,1)$ quand $n\to\infty$. On déduit

$$\forall\alpha\in]0,1[\ ,\ \forall t\in]0,\alpha[\ ,\ \mathbb{P}\big(\theta\in I_\alpha\big)=\mathbb{P}\Big(q_t\le\frac{\sqrt{n}\big(\hat{\theta}-\theta\big)}{\hat{\sigma}}\le q_{1-\alpha+t}\Big)\underset{n\to\infty}{\longrightarrow}1-\alpha$$

par continuité de la fonction de répartition de la loi normale. □

Dans un modèle paramétrique paramétré par θ, la variance asymptotique σ^2 est fonction de θ. Si cette fonction est continue, un estimateur consistant est donné par $\hat{\sigma}^2=\sigma(\hat{\theta})^2$. Par la proposition précédente, (en prenant $t=\alpha/2$ par exemple), on obtient l'intervalle de confiance

$$I_\alpha=\Big[\hat{\theta}-q_{1-\frac{\alpha}{2}}\frac{\sigma(\hat{\theta})}{\sqrt{n}}\ ,\ \hat{\theta}+q_{1-\frac{\alpha}{2}}\frac{\sigma(\hat{\theta})}{\sqrt{n}}\Big],$$

de niveau α asymptotiquement. Une autre région de confiance asymptotique, en général plus précise car n'utilisant pas l'approximation de $\sigma(\theta)$ par $\sigma(\hat{\theta})$, est donnée par

$$R_\alpha=\Big\{\theta:-q_{1-\frac{\alpha}{2}}\le\sqrt{n}\frac{\hat{\theta}-\theta}{\sigma(\theta)}\le q_{1-\frac{\alpha}{2}}\Big\}.$$

Cette région de confiance n'est pas forcément un intervalle si $\sigma(.)$ n'est pas monotone.

Exemples

1. Soit $\theta=\mathbb{E}(X)$ dans le modèle non-paramétrique des lois qui admettent un moment d'ordre deux fini. Soit $\hat{\theta}=\frac{1}{n}\sum_{i=1}^n X_i$, on sait par le TCL que $\sqrt{n}(\hat{\theta}-\theta)$ converge en loi vers une $\mathcal{N}(0,\sigma^2)$ où $\sigma^2=\mathrm{var}(X)$. Soit $\hat{\sigma}^2$ la variance empirique, on construit un intervalle de confiance asymptotique de θ:
$$I_\alpha=\Big[\hat{\theta}-q_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{n}}\ ,\ \hat{\theta}+q_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{n}}\Big].$$

2. On cherche un intervalle de confiance sur le paramètre p d'un modèle de Bernoulli. On sait que $\sqrt{n}(\hat{p}-p)$ est asymptotiquement Gaussien de moyenne 0 et variance $p(1-p)$. On a donc par le lemme de Slutsky,
$$\sqrt{n}\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}}\xrightarrow[n\to\infty]{loi}\mathcal{N}(0,1).$$
On obtient l'intervalle de confiance asymptotique de niveau α:
$$\Big[\hat{p}-q_{1-\frac{\alpha}{2}}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\ ,\ \hat{p}+q_{1-\frac{\alpha}{2}}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\Big].$$
On obtient une région de confiance différente (mais toujours de niveau α asymptotiquement) en considérant la statistique $\sqrt{n}(\hat{p}-p)/\sqrt{p(1-p)}$ (exercice).

3. Modèle de Poisson. Si les observations suivent une loi de Poisson de paramètre λ, alors moyenne et variance sont égales et $\sqrt{n}(\hat{\lambda}-\lambda)$ converge en loi vers une $\mathcal{N}(0,\lambda)$, pour $\hat{\lambda}=\frac{1}{n}\sum_{i=1}^n X_i$.

5 Tests

En statistique, choisir un modèle revient à fixer des conditions sur la loi des observations. Lorsque l'on souhaite vérifier si ces conditions sont raisonnables, on peut utiliser un *test statistique*. Formellement, un test est une variable aléatoire construite à partir des observations qui prend la valeur 1 si on rejette l'hypothèse, ou 0 si on ne rejette pas l'hypothèse. La décision de rejeter ou non l'hypothèse est toujours sensible à l'aléa des observations, ce qui nous oblige à fixer à l'avance une marge d'erreur α qui correspond à la probabilité de rejeter l'hypothèse alors qu'elle est vraie. Pour tester une hypothèse $\mathcal{H}_0$, on construit une variable aléatoire dont la loi est connue si $\mathcal{H}_0$ est vraie. Si la valeur observée de cette variable est "aberrante" par rapport à sa loi sous $\mathcal{H}_0$, on décidera de rejeter l'hypothèse.

5.1 Définitions générales

On observe un échantillon iid $X = X_1, ..., X_n$ de loi μ dont on veut tester une ou plusieurs propriétés (par exemple, tester si μ est une loi normale, μ est centrée etc...). De manière générale, on s'intéresse à tester une hypothèse de la forme $\mathcal{H}_0 : \mu \in \mathcal{M}_0$ pour $\mathcal{M}_0$ un certain sous-ensemble du modèle initial $\mathcal{M}$. L'hypothèse $\mathcal{H}_0$ est appelée hypothèse nulle. L'ensemble des possibilités si $\mathcal{H}_0$ n'est pas vérifiée est l'hypothèse alternative, notée $\mathcal{H}_1$. On a donc $\mathcal{H}_1 : \mu \in \mathcal{M}_1 = \mathcal{M} \setminus \mathcal{M}_0$.

Définition 5.1. *Un test de niveau α d'une hypothèse $\mathcal{H}_0$ est une statistique $\Phi_\alpha(X)$ qui suit une loi de Bernoulli de paramètre α sous l'hypothèse $\mathcal{H}_0$.*

Il est indispensable que la variable aléatoire Φ_α soit une fonction connue des observations, c'est-à-dire que l'on peut observer une réalisation de Φ_α. C'est précisément la valeur observée de Φ_α qui va nous permettre de juger de la crédibilité de l'hypothèse $\mathcal{H}_0$. Pour une valeur observée $\Phi_\alpha(X) = 1$, on rejette $\mathcal{H}_0$ au niveau α. Sinon, on ne rejette pas $\mathcal{H}_0$.

Définition 5.2. *La région critique R_α est l'ensemble des réalisations qui conduisent à rejeter $\mathcal{H}_0$ au niveau α. On a donc la relation*

$$\Phi_\alpha = \mathbb{1}\{X \in R_\alpha\}.$$

En pratique, les régions critiques $R_\alpha, \alpha \in]0,1[$ forment une famille de sous-ensembles croissants de $\mathbb{R}^n$ (ou plus généralement $\mathcal{X}^n$ l'espace des observations) au sens où

$$\forall \alpha' > \alpha : R_{\alpha'} \supset R_\alpha.$$

Autrement dit, rejeter l'hypothèse $\mathcal{H}_0$ au niveau α implique de la rejeter également pour tous les niveaux supérieurs. On dispose donc d'un continuum de tests $\Phi_\alpha, \alpha \in]0,1[$ qui donnent des événements $\{\Phi_\alpha = 1\}, \alpha \in]0,1[$ croissants.

Dans beaucoup de cas, la région critique peut s'exprimer simplement à partir d'une statistique $T = T(X) \in \mathbb{R}$, appelée *statistique de test*, sous la forme

$$R_\alpha = \{x \in \mathcal{X}^n : T(x) > s_\alpha\},$$

où s_α est un seuil adapté. Dans ce cas, on rejette $\mathcal{H}_0$ si T dépasse le seuil s_α.

Un test est dit *exact* si la probabilité de rejeter $\mathcal{H}_0$ lorsqu'elle est vraie (l'erreur de première espèce) est égale au niveau choisi α. Si cette condition est trop difficile (ou impossible) à vérifier, on peut se contenter d'un test asymptotiquement exact, dont l'erreur de première espèce tend vers α quand $n \to \infty$.

Définition 5.3. *On appelle* p-value *du test le plus petit niveau α pour lequel on rejette $\mathcal{H}_0$ au vu de la réalisation de Φ_α:*

$$\text{p-value} = \inf\{\alpha > 0 : \Phi_\alpha(X) = 1\}.$$

On rejettera donc l'hypothèse $\mathcal{H}_0$ si la p-value est supérieure au niveau du test. L'écart entre la p-value et α quantifie la marge avec laquelle on accepte $\mathcal{H}_0$ dans le cas où elle est plus grande que α et la marge avec laquelle on rejette $\mathcal{H}_0$ dans le cas contraire.

Proposition 5.4. *Pour un test exact, la p-value est une variable aléatoire de loi uniforme sur* $[0,1]$ *sous* $\mathcal{H}_0$.

Preuve. On suppose que $\mathcal{H}_0$ est vraie. Soit $u \in]0,1[$,

$$\mathbb{P}(\text{p-value} \leq u) = \mathbb{P}\Big(\inf\{\alpha > 0 : \Phi_\alpha(X) = 1\} \leq u\Big) = \mathbb{P}\Big(\bigcup_{0<\alpha\leq u} \{\Phi_\alpha(X) = 1\}\Big).$$

Comme les événements $\{\Phi_\alpha(X) = 1\}, \alpha \in [0,1]$ sont croissants et de probabilité α, on déduit que $\mathbb{P}(\text{p-value} \leq u) = \lim_{\alpha\to u} \alpha = u$. □

Exemple On veut tester l'hypothèse $\mathsf{m} = \mathsf{m}_0$ dans le modèle Gaussien où les observations $X_1, ..., X_n$ sont de loi $\mathcal{N}(\mathsf{m}, 1)$. On sait que, si $\mathcal{H}_0$ est vraie

$$T = \sqrt{n}\,\big(\hat{\mathsf{m}} - \mathsf{m}_0\big) \sim \mathcal{N}(0,1).$$

Supposons dans un premier temps que m peut prendre toutes les valeurs réelles, l'hypothèse alternative peut donc s'écrire $\mathcal{H}_1 : \mathsf{m} \neq \mathsf{m}_0$. Une valeur "raisonnable" de T (càd conforme à l'hypothèse $\mathcal{H}_0$) sera proche de zéro, le cas contraire conduira à penser que $\mathsf{m} \neq \mathsf{m}_0$. Une procédure naturelle de test consiste donc à rejeter $\mathcal{H}_0$ si $|T|$ est trop grand. Le seuil est ensuite calculé en fonction du niveau fixé, de manière à avoir une probabilité de rejeter à tort égale à α. On construit donc le test bilatéral:

$$\Phi_\alpha = \mathbb{1}\{|T| > q_{1-\frac{\alpha}{2}}\} = \mathbb{1}\Big\{\hat{\mathsf{m}} - \frac{q_{1-\frac{\alpha}{2}}}{\sqrt{n}} \leq \mathsf{m}_0 \leq \hat{\mathsf{m}} + \frac{q_{1-\frac{\alpha}{2}}}{\sqrt{n}}\Big\}.$$

On rejette plus facilement $\mathcal{H}_0$ lorsque le niveau α est élevé. En faisant décroître α, le seuil à partir duquel on ne rejette plus l'hypothèse nulle nous donne la p-value du test. On obtient facilement ici

$$\text{p-value} = 2\big(1 - F(|T|)\big),$$

où F est la fonction de répartition d'une $\mathcal{N}(0,1)$. Supposons maintenant que m ne peut pas prendre de valeur inférieure à m_0. L'hypothèse alternative s'écrit alors $\mathcal{H}_1 : \mathsf{m} > \mathsf{m}_0$. Les valeurs de T conformes à l'hypothèse nulle sont les valeurs petites. En particulier une valeur négative de T nous conduira à penser que $\mathsf{m} = \mathsf{m}_0$ puisque $\mathsf{m} < \mathsf{m}_0$ n'est pas possible. On décide donc de rejeter $\mathcal{H}_0$ si T est grand. Le seuil donnant un niveau de α est $q_{1-\alpha}$, qui correspond au test unilatéral:

$$\Phi_\alpha = \mathbb{1}\{T > q_{1-\alpha}\}.$$

La p-value vaut ici $1 - F(T)$.

Définition 5.5. *La puissance du test, notée* $1 - \beta(.)$, *est la probabilité de rejeter* $\mathcal{H}_0$ *sous une loi* ν *vérifiant l'hypothèse alternative*

$$1 - \beta(\nu) = \mathbb{E}_\nu\big(\Phi_\alpha(X_1, ..., X_n)\big)\ ,\ \nu \in \mathcal{M}_1.$$

La puissance est une fonction de la loi ν, définie sur l'ensemble $\mathcal{M}_1$. La quantité $\beta(\nu)$ est la probabilité de ne pas rejeter $\mathcal{H}_0$ si l'échantillon $X_1, ..., X_n$ est issu de la loi ν. On dira qu'un test est consistant si la puissance tend vers 1 quand $n \to \infty$, pour tout $\nu \in \mathcal{M}_1$.

Il y a deux types d'erreurs possibles lors d'un test statistique:

- Rejeter $\mathcal{H}_0$ alors qu'elle est vraie constitue l'erreur de première espèce, de probabilité égale au niveau α du test. Elle est fixée à l'avance par le statisticien.
- Ne pas rejeter $\mathcal{H}_0$ lorsque l'échantillon est issu d'une loi $\nu \in \mathcal{M}_1$ constitue l'erreur de deuxième espèce, de probabilité $\beta(\nu)$. Un test sera d'autant meilleur (plus puissant) que l'erreur de deuxième espèce est faible, à un niveau α fixé, pour les différentes valeur de $\nu \in \mathcal{M}_1$. La construction d'un test puissant dépend donc fortement de l'hypothèse alternative $\mathcal{H}_1$.

5.2 Tests sur un paramètre de position

5.2.1 Test du signe

Le test sur la médiane, ou test du signe, est un des plus vieux tests non-paramétriques existant. Il permet de tester si la médiane de X, $\mathrm{med}(X) = F^{-}(1/2)$, est égale à une certaine valeur m_0. Comme la médiane n'est pas forcément unique, l'hypothèse nulle que l'on considère ici est

$$\mathcal{H}_0 : F(\mathsf{m}_0) = 1/2,$$

ce qui revient à tester, dans le cas où F est continue, si m_0 est un quantile d'ordre $1/2$ de X. On utilise pour ce test la statistique

$$M_n = \sum_{i=1}^{n} \mathbb{1}\{X_i > \mathsf{m}_0\} - \sum_{i=1}^{n} \mathbb{1}\{X_i < \mathsf{m}_0\} = \#\{i : X_i > \mathsf{m}_0\} - \#\{i : X_i < \mathsf{m}_0\}.$$

Comme pour le test sur la moyenne, on peut utiliser un test bilatéral ou unilatéral. Soit q_α le quantile d'ordre $\alpha \in]0,1[$ de la loi normale centrée réduite, on construit les tests

$$\begin{aligned}
&\Phi_\alpha = \mathbb{1}\{|M_n| > \sqrt{n} q_{1-\frac{\alpha}{2}}\} && \text{pour tester } \mathcal{H}_0 \text{ contre } \mathcal{H}_1 : F(\mathsf{m}_0) \neq 1/2 \\
&\Phi_\alpha = \mathbb{1}\{M_n > \sqrt{n} q_{1-\alpha}\} && \text{pour tester } \mathcal{H}_0 \text{ contre } \mathcal{H}_1 : F(\mathsf{m}_0) < 1/2 \\
&\Phi_\alpha = \mathbb{1}\{M_n < \sqrt{n} q_{\alpha}\} && \text{pour tester } \mathcal{H}_0 \text{ contre } \mathcal{H}_1 : F(\mathsf{m}_0) > 1/2
\end{aligned}$$

Théorème 5.6. *Si F est continue en m_0, ces trois tests sont de niveau α asymptotiquement et sont consistants.*

Preuve. Comme F est continue en m_0, $\mathbb{P}(X_i = \mathsf{m}_0) = 0$ et $M_n = 2\sum_{i=1}^{n} \mathbb{1}\{X_i > \mathsf{m}_0\} - n$ presque sûrement. Or, par le théorème central limite,

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i > \mathsf{m}_0\} - \mathbb{P}(X > \mathsf{m}_0)\Big) \xrightarrow[n\to\infty]{loi} \mathcal{N}\big(0, 1/4\big).$$

Sous $\mathcal{H}_0$, $\mathbb{P}(X > \mathsf{m}_0) = 1/2$ et $M_n/\sqrt{n}$ converge vers une loi normale standard. Sous l'hypothèse alternative, il est facile de voir en distingant les cas, que $M_n/\sqrt{n}$ diverge et que le test est consistant. $\square$

Remarque 5.7. *La statistique $\sum_{i=1}^{n} \mathbb{1}\{X_i > \mathsf{m}_0\}$ suit une loi binomiale $\mathcal{B}(n, 0.5)$ sous $\mathcal{H}_0$, on peut donc utiliser les quantiles de la loi binomiale pour construire un test plus précis pour des petites valeurs de n.*

5.2.2 Test de Student sur la moyenne

On s'intéresse ici à tester une hypothèse de la forme $\mathcal{H}_0 : \mathbb{E}(X) = \mathsf{m}_0$ où m_0 est un réel connu. Le modèle considéré est l'ensemble des lois non constantes qui admettent un moment d'ordre deux

$$\mathcal{M} = \{\nu : 0 < \mathrm{var}_\nu(X) < \infty\}.$$

L'ensemble des lois vérifiant l'hypothèse nulle est $\mathcal{M}_0 = \{\nu : \mathbb{E}_\nu(X) = \mathsf{m}_0,\ 0 < \mathrm{var}_\nu(X) < \infty\}$ et l'hypothèse alternative est simplement $\mathcal{H}_1 : \mathbb{E}(X) \neq \mathsf{m}_0,\ 0 < \mathrm{var}(X) < \infty$.

On définit la moyenne empirique $\overline{X}_n$ et l'écart-type empirique $\hat{\sigma}$ par

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Le test non-paramétrique sur la moyenne est basé sur le théorème central limite, et plus précisément, sur le comportement de la statistique

$$S_n := \sqrt{n}\ \frac{\overline{X}_n - \mathsf{m}_0}{\hat{\sigma}},$$

appelée *statistique de Student*. Si $\mathcal{H}_0$ est vraie, S_n converge vers une loi normale standard quand n tend vers l'infini. Soit q_α le quantile d'ordre $\alpha \in]0,1[$ de la loi normale standard, le test est défini par

$$\Phi_\alpha(X_1, ..., X_n) = \mathbb{1}\left\{|S_n| > q_{1-\frac{\alpha}{2}}\right\}.$$

On choisit donc de rejeter $\mathcal{H}_0 : \mathbb{E}(X) = \mathsf{m}_0$ si la statistique $|S_n|$ prend une valeur trop grande pour une loi normale standard.

Remarque 5.8. *Le nom de la statistique provient du fait que S_n suit, sous $\mathcal{H}_0$, une loi de Student à $n-1$ degrés de liberté dans le cas où les données X_i sont Gaussiennes.*

Théorème 5.9. *Le test $\Phi_\alpha = \mathbb{1}\{|S_n| > q_{1-\frac{\alpha}{2}}\}$ est asymptotiquement de niveau α,*

$$\forall \nu \in \mathcal{M}_0,\ \lim_{n\to\infty} \mathbb{E}_\nu(\Phi_\alpha) = \alpha.$$

De plus, $\forall \nu \in \mathcal{M}_1,\ \lim_{n\to\infty} \mathbb{E}_\nu(\Phi_\alpha) = 1$ (le test est consistant).

Preuve. On sait que $\hat{\sigma}^2$ converge en probabilité (et même presque sûrement) vers $\sigma^2 = \mathrm{var}(X) > 0$. Par le théorème central limite, on sait que

$$Y_n := \sqrt{n}\ \frac{\overline{X}_n - \mathbb{E}(X)}{\sigma} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1).$$

D'après le théorème de Slutsky, on a donc

$$Y_n \times \frac{\sigma}{\hat{\sigma}} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1).$$

Sous $\mathcal{H}_0$, $\mathbb{E}(X) = \mathsf{m}_0$ et $S_n = Y_n \times \frac{\sigma}{\hat{\sigma}}$ converge donc en loi vers une $\mathcal{N}(0,1)$. Comme la fonction de répartition de la loi normale est continue, la convergence en loi implique la convergence simple des fonctions de répartition et

$$\lim_{n\to\infty} \mathbb{P}\big(|S_n| > q_{1-\frac{\alpha}{2}}\big) = \mathbb{P}\big(|\mathcal{N}(0,1)| > q_{1-\frac{\alpha}{2}}\big) = \alpha.$$

Sous $\mathcal{H}_1$, posons $\mathsf{m}_1 = \mathbb{E}(X) \neq \mathsf{m}_0$. On a

$$S_n = \sqrt{n}\ \frac{\overline{X}_n - \mathsf{m}_0}{\hat{\sigma}} = \sqrt{n}\ \frac{\overline{X}_n - \mathsf{m}_1}{\hat{\sigma}} + \sqrt{n}\ \frac{\mathsf{m}_1 - \mathsf{m}_0}{\hat{\sigma}} \underset{n\to\infty}{\approx} \mathcal{N}(0,1) + \sqrt{n}\ \frac{\mathsf{m}_1 - \mathsf{m}_0}{\sigma}$$

et il est clair que $\mathbb{P}\big(|S_n| > q_{1-\frac{\alpha}{2}}\big) \to 1$. □

Le test de l'hypothèse $\mathcal{H}_0 : \mathbb{E}(X) = \mathsf{m}_0$ contre $\mathcal{H}_1 : \mathbb{E}(X) \neq \mathsf{m}_0$ est un test bilatéral. Si on sait à l'avance que l'espérance est forcément supérieure ou égale à m_0 (resp. inférieure ou égale), il est plus efficace d'utiliser un test unilatéral avec comme alternative $\mathcal{H}_1 : \mathbb{E}(X) > \mathsf{m}_0$ (resp. $\mathcal{H}_1 : \mathbb{E}(X) < \mathsf{m}_0$). Le test unilatéral est alors défini par $\Phi_\alpha = \mathbb{1}\{S_n > q_{1-\alpha}\}$ (resp. $\Phi_\alpha = \mathbb{1}\{S_n < q_\alpha\}$).

5.3 Tests d'adéquation

Un test d'adéquation est un test sur une valeur particulière de la loi des observations. Etant donné une loi μ_0 connue, on veut tester si l'échantillon $X_1, ..., X_n$ est issu de cette loi.

5.3.1 Test du χ^2

Le test d'adéquation du χ^2 est utilisé à l'origine pour tester si un échantillon est issu d'une loi discrète μ_0. L'hypothèse nulle s'écrit donc $\mathcal{H}_0 : \mu = \mu_0$, pour μ_0 est une loi discrète connue à support fini $\{t_1, ..., t_k\}$. On se place dans le modèle non-paramétrique, l'hypothèse alternative est $\mathcal{H}_1 : \mu \neq \mu_0$. En notant $\pi_j = \mu_0(\{t_j\})$ et $p_j = \mathbb{P}(X = t_j)$ pour $j = 1, ..., k$, l'hypothèse nulle peut s'écrire de façon équivalente $\mathcal{H}_0 : p_j = \pi_j,\ j = 1, ..., k$. On remarque également que l'hypothèse alternative peut s'écrire $\mathcal{H}_1 : \exists j,\ p_j \neq \pi_j$.

On estime naturellement les probabilités inconnues p_j par les fréquences d'apparition des valeurs t_j dans l'échantillon

$$\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i = t_j\}\ ,\ j = 1, ..., k.$$

On définit alors la statistique de test

$$Q_n := n\sum_{j=1}^{k} \frac{(\hat{p}_j - \pi_j)^2}{\pi_j} = n\Big(\sum_{j=1}^{k} \frac{\hat{p}_j^2}{\pi_j} - 1\Big).$$

Proposition 5.10. *Sous $\mathcal{H}_0 : \mu = \mu_0$, Q_n converge en loi vers un $\chi^2(k-1)$ quand $n \to \infty$. Par ailleurs, Q_n tend vers l'infini en probabilité sous $\mathcal{H}_1 : \mu \neq \mu_0$.*

Preuve. On pose $\hat{p} = (\hat{p}_1, ..., \hat{p}_k)^\top$, $\pi = (\pi_1, ..., \pi_k)^\top$ et

$$Y_i = \begin{bmatrix} \mathbb{1}\{X_i = t_1\} \\ \vdots \\ \mathbb{1}\{X_i = t_k\} \end{bmatrix}\ ,\ i = 1, ..., n.$$

Par le théorème central limite vectoriel, $\sqrt{n}\big(\hat{p} - \pi\big) = \sqrt{n}\ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \pi) \xrightarrow[n\to\infty]{loi} \mathcal{N}(0, \Sigma)$ où

$$\Sigma = \mathrm{var}(Y) = \begin{bmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \dots & -\pi_1\pi_k \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & \dots & -\pi_2\pi_k \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_1\pi_k & -\pi_2\pi_k & \dots & \pi_k(1-\pi_k) \end{bmatrix}.$$

Soit $\mathsf{D} = \mathrm{Diag}(\sqrt{\pi})$, la matrice diagonale de diagonale $\sqrt{\pi} = (\sqrt{\pi_1}, ..., \sqrt{\pi_k})^\top$. On déduit du résultat précédent que

$$\hat{q} := \sqrt{n}\ \mathsf{D}^{-1}\big(\hat{p} - \pi\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}(0, \Gamma),$$

où $\Gamma = \mathsf{D}^{-1}\Sigma\mathsf{D}^{-1} = \mathrm{I} - \sqrt{\pi}\sqrt{\pi}^\top$ (I désigne l'identité de $\mathbb{R}^k$). On remarque dans un premier temps que $\hat{q}$ est orthogonal à $\sqrt{\pi}$. Soit $\{\mathsf{v}_1, ..., \mathsf{v}_{k-1}\}$ une collection de vecteurs orthonormés de $\mathbb{R}^k$, orthogonaux à $\sqrt{\pi}$ et $\mathsf{V} = (\mathsf{v}_1, ..., \mathsf{v}_{k-1}) \in \mathbb{R}^{k\times(k-1)}$. Comme la norme Euclidienne est invariante par changement de base othonormée, on déduit

$$Q_n = \|\hat{q}\|^2 = (\hat{q}^\top\sqrt{\pi})^2 + \|\mathsf{V}^\top\hat{q}\|^2 = \|\mathsf{V}^\top\hat{q}\|^2.$$

En remarquant que $\mathsf{V}^\top \Gamma \mathsf{V} = \mathrm{I}$ (l'identité de $\mathbb{R}^{k-1}$), on déduit

$$\mathsf{V}^\top \hat{q} = \sqrt{n}\ \mathsf{V}^\top \mathsf{D}^{-1}\big(\hat{p} - \pi\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}(0, \mathrm{I}),$$

d'où le résultat $Q_n \xrightarrow[n\to\infty]{loi} \chi^2(k-1)$. Par ailleurs, on voit facilement que $Q_n \xrightarrow[n\to\infty]{\mathbb{P}} +\infty$ sous $\mathcal{H}_1$. □

Ce résultat permet de construire le test d'adéquation du χ^2 qui s'écrit

$$\Phi_\alpha = \mathbb{1}\{Q_n > q_{1-\alpha}\},$$

où $q_{1-\alpha}$ désigne le quantile d'ordre $1-\alpha$ de la loi du $\chi^2(k-1)$. Par la continuité de la fonction de répartition d'un χ^2, on déduit facilement de la proposition précédente que le test est asymptotiquement de niveau α et qu'il est consistant.

L'idée du test d'adéquation du χ^2 peut se généraliser à tout type de loi. Supposons maintenant que les observations sont distribuées suivant une loi μ quelconque sur $\mathbb{R}$ et que l'on veut tester une valeur particulière $\mathcal{H}_0 : \mu = \mu_0$. On procède de la façon suivante:

- On choisit une partition $\{A_j\}_{j=1,\dots,k}$ de $\mathbb{R}$ telle que $\pi_j := \mu_0(A_j) > 0$ pour tout $j = 1, .., k$.
- On calcule les proportions d'observations pour chaque ensemble A_j,

$$\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \in A_j\},\ j = 1, ..., k.$$

- D'après la proposition 5.10, on sait que

$$Q_n = n\sum_{j=1}^{k} \frac{(\hat{p}_j - \pi_j)^2}{\pi_j}$$

  suit asymptotiquement un $\chi^2(k-1)$ si $\mu(A_j) = \pi_j$ pour tout $j = 1, .., k$.

Remarque 5.11. *Le test $\Phi_\alpha = \mathbb{1}\{Q_n > q_{1-\alpha}\}$ est donc asymptotiquement de niveau α pour tester l'hypothèse $\mathcal{H}_0 : \mu = \mu_0$. En revanche, il n'est plus consistant puisqu'on peut construire une loi $\mu_1 \neq \mu_0$ telle que $\mu_1(A_j) = \pi_j$ pour tout $j = 1, .., k$. La puissance du test sous μ_1 convergera donc vers $1-\alpha$, ce qui n'est pas satisfaisant. Une solution à ce problème est d'augmenter le nombre de classes $k = k_n$ de la partition avec n, suffisamment lentement pour pouvoir approcher la loi de la statistique Q_n par un $\chi^2(k_n)$.*

5.3.2 Test de Kolmogorov-Smirnov

Comme le test d'adéquation du χ^2, le test de Kolmogorov-Smirnov permet de tester une valaur particulière de la loi via l'hypothèse nulle $\mathcal{H}_0 : \mu = \mu_0$. En revanche, il n'est valable que pour des lois μ_0 continues. L'idée du test repose sur le théorème suivant.

Théorème 5.12 (Donsker)**.** *Soit $X_1, ..., X_n$ un échantillon iid de fonction de répartition F continue et F_n la fonction de répartition empirique. Alors,*

$$K_n = \sqrt{n}\ \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{loi} K,$$

où K est une variable aléatoire de loi de Kolmogorov de fonction de répartition

$$F_K(x) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}\exp(-2j^2x^2),\ \ , x \geq 0.$$

La fonction de répartition de la loi de Kolmogorov est continue est strictement croissante sur $]0, +\infty[$, il existe donc un unique $q_{1-\alpha}$ tel que $F_K(q_{1-\alpha}) = 1 - \alpha$, ce qui permet de construire un test asymptotiquement de niveau α par $\Phi_\alpha = \mathbb{1}\{K_n > q_{1-\alpha}\}$. On peut montrer également que ce test est consistant pour toute fonction de répartition $F_1 \neq F$.