Cours de Statistique
ENSIIE 1A

Nicolas BRUNEL*

Année 2017-2018

Avant-propos

Il s'agit de notes de cours partielles. Les exemples et calculs seront traités en cours.
Le cours de statistique « MST » est un cours d'introduction à la statistique inférentielle, parfois appelée statistique mathématique. Ce cours présente le cadre mathématique et les fondements de la démarche statistique.
Les UEs de statistiques qui viennent développer et/ou compléter ce cours sont :

- La modélisation statistique (régression, modèle prédictif) : MRR, Modèle de Régression Régularisé (S3)
- L'Analyse de Données Exporatoires (ou analyse de données multivariées, Fouille de Données/Data Mining, apprentissage non-supervisée) : MAD, Analyse de Données (S3)
- Séries Temporelles (lissage, analyse séries chronologiques, prédiction et modèle linéaire) : UE MOST (S4)
- Apprentissage Statistique (apprentissage supervisé, prédiction) : UE MOST (S4)
- Simulation aléatoire et Inférence Bayésienne : Méthodes de Simulation, MESIM (S4)
- Apprentissage Statistique Avancée (modèles d'ensemble, forêts aléatoires, clusterings, SVM) : MAL (S5)
- Apprentissage Profond (Réseaux de Neurones) : MSA (S5)

Le lien entre informatique et statistique a toujours été très fort, et a donné naissance ce qu'on appelle la « Science des Données » (ou Data Science), qui est à l'intersection des statistiques, de l'informatique (programmation, bases de données, calcul intensif, traitement d'images,...) et des mathématiques (au premier chef les probabilités et l'optimisation). Beaucoup de termes ou de sensibilités (plus que « disciplines propres » à part entière) se retrouvent associés tels que l'apprentissage automatique (machine learning), réseaux de neurones (deep learning) et intelligence artificielle....Les TPs et le projet de ce cours seront faits en R (https ://cran.r-project.org/), langage de programmation libre qui est désormais une référence dans la communauté statistique, et dans le monde professionnel pour la manipulation et l'analyse des données, et la modélisation. L'autre langage de programmation dorénavant incontournable (pour l'instant...) en science des données et Python, et ses nombreuses librairies mathématiques, ou de traitement de données.
La lecture de ces notes, et l'assistance aux amphis de cours et TDs-TP, seront avantageusement complétées et étofées par la lecture plus ou moins attentive de de supports de cours disponibles sur le site de la Société Française de Statistique :
https ://www.sfds.asso.fr/fr/enseignement_de_la_statistique/ressources/enseignement_superieur/597-par_thematiques
On pourra aussi regarder le site de Trevor Hastie, et son livre *Computer Age Statistical Inference* qui présente de manière intéressante l'évolution des statistiques et de la dualité modèle/algorithme qui l'a fait fortement évoluer :
http ://web.stanford.edu/~hastie/CASI/order.html
Ses autres livres, en téléchargement libre, sont aussi intéressant (mais plus orienté « apprentissage statistique »).

*nicolas.brunel@ensiie.fr, bureau 108

Table des matières

Première partie
Modélisation statistique

1 Outils probabilistes et modèles statistiques

1.1 Quelques définitions de théorie de la mesure et de probabilités

Définition 1. Une Tribu $\mathcal{T}$ de l'ensemble $\mathcal{X}$ - typiquement R ou R^p, un ensemble dénombrable tel que $\mathbb{N}$ - est un ensemble de sous-ensembles de $\mathcal{X}$ tel que

1. $\mathcal{X} \in \mathcal{T}$
2. Si $A, B \in \mathcal{T}$, alors $A \cup B \in \mathcal{T}$
3. Si $A \in \mathcal{T}$, alors $A^c = \mathcal{X} \backslash A \in \mathcal{T}$
4. Si pour toute suite $(A_n)_{n\geq 0}$ dans $\mathcal{T}$, alors $\cup_{n\geq 0} A_n \in \mathcal{T}$.

Les deux exemples les plus classiques et les plus utiles sont :
— Si $\mathcal{X}$ ensemble fini ou dénombrable, la tribu que l'on utilise presque tout le temps est l'ensemble des sous-ensembles de $\mathcal{X}$, noté $\mathcal{P}(\mathcal{X})$.
— Si $\mathcal{X} = R$, alors on $\mathcal{B}(R)$ (la tribu borélienne) : il s'agit de l'ensemble de sous-ensembles de R obtenus par union et intersection dénombrables des intervalles $]a, b[$ avec $a < b$.
— Si $\mathcal{X} = R^d$, on prend les intersections et unions dénombrables des pavés $\prod_{i=1}^{d}]a_i, b_i[$.
La notion de tribu permet de définir proprement les ensembles pour lesquels on peut définir facilement et « sans surprise » une longueur, une aire, un volume,... et finalement construire une théorie de l'intégration (et des probabilités) possédant de bonnes propriétés : théorème de convergence dominée, théorème de Fubini (pour le calcul des intégrales multiples),...

Définition 2. Une mesure est une fonction définie sur une espace mesurable $(\mathcal{X}, \mathcal{T})$ (i.e définie sur des ensembles) telle que

positivité $\mu(\emptyset) = 0$ et $\mu : \mathcal{T} \longrightarrow R^+$, i.e pour tout ensemble A dans la tribu $\mathcal{T}$, $\mu(A) \geq 0$.

addititivité Si $A \cap B = \emptyset$, $\mu(A \cup B) = \mu(A) + \mu(B)$.

σ-additivité Si la suite $(A_n)_{n\geq 0}$ de $\mathcal{T}$ est tel que $A_i \cap A_j = \emptyset$, alors $\mu(\cup_{n\geq 0} A_n) = \sum_{n\geq 0} \mu(A_n)$.
— Une mesure est dite **σ-finie**, si il existe une suite d'étéments de $\mathcal{T}$ E_n, $n \geq 0$ tel que $\cup_{n\geq 0} E_n = \mathcal{X}$ et $\mu(E_n) < +\infty$.
— Si $\mu(\mathcal{X}) = 1$, alors μ est appelée une **mesure de probabilités** (de manière plus générale, on parle de mesure bornée si $\mu(\mathcal{X}) < +\infty$.

Exemple 1. Il y a essentiellement deux mesures « pratiques » qui sont la mesure de Lebesgue (mesure continue) et la mesure de comptage qui sont en fait déjà bien connues.

— La **mesure de Lebesgue** sur R, notée λ, est telle que pour tout segment $]a, b]$ $(a < b)$, on a $\lambda(]a, b]) = b - a$. A partir de cette mesure, on définit l'intégrale de Lebesgue, qui correspond pour les fonctions continues f (ou assez « régulières », par exemple continue par morceaux) à l'intégrale de Riemann

$$\int_R f(x)\lambda(dx) = \int_{-\infty}^{\infty} f(x)dx = \lim \sum_i f(x_i)\,(x_{i+1} - x_i)$$

On aime bien utiliser la notation différentielle dx. Une fonction est dite intégrable si $\int_R |f|\,(x)dx < \infty$. La mesure λ est ce que l'on appelle une **mesure σ-finie** car $R = \cup_{n\in Z}]n, n+1]$ (les propriétés ne sont pas trop différentes de celles d'une mesure bornée). Dans la foulée, on définit aussi la mesure de Lebesgue sur R^+ $\lambda_{R^+}(A) = \lambda(A \cap R^+)$ et l'intégrale correspondante $\int_{R^+} f(x)dx = \int_{-\infty}^{+\infty} f(x)1_{R^+}(x)dx$. On écrit encore $\lambda_{R^+} = 1_{R^+} \cdot \lambda$.
— La mesure de Lebesgue s'étend à R^d en introduisant la mesure produit (on « tensorise » les mesures)

$$\lambda^{\otimes d}(\prod_{i=1}^{d}]a_i, b_i]) = \prod_{i=1}^{d} (b_i - a_i)$$

La mesure correspondante est parfois simplement notée $\lambda^{\otimes d}(dx) = dx_1 dx_2 \dots dx_d$. De la même façon, à partir de la mesure, l'intégrale de Lebesgue est construite pour tout fonction mesurable (et intégrable)

$$\int_{R^d} h(x)dx = \int h(x_1, \dots, x_d)dx_1 \dots dx_d$$

Et on rappelle que le **théorème de Fubini** (on l'énonce dans le cas $d = 2$ mais c'est vrai en toute généralité) affirme que pour toute fonction à valeurs réelles positives (et mesurable sur R^2) alors

$$\begin{aligned}\int_{R^2} f(x)\lambda^{\otimes d}(dx) &= \int \left(\int f(x_1, x_2)dx_2 \right) dx_1 \\ &= \int \left(\int f(x_1, x_2)dx_1 \right) dx_2\end{aligned}$$

C'est aussi vrai si la fonction n'est plus positive, mais seulement si les marginales $\int f(x_1, x_2)dx_1$, $\int f(x_1, x_2)dx_2$ sont intégrables.

— La mesure discrète la plus simple est la **mesure de Dirac**, qui peut être définie dans R ou dans une ensemble discret $\{x_1, x_2, \dots\}$: la mesure de Dirac en $x \in \mathcal{X}$, notée δ_x est telle que $\delta_x(A) = 1 \iff x \in A$. En particulier, $\delta_x(\{x\}) = 1$, et $\forall y \neq x$, $\delta_x(\{y\}) = 0$. L'intégration relativement à la mesure de Dirac est assez simple

$$\int_{\mathcal{X}} f(y)\delta_x(dy) = f(x).$$

— La **mesure de comptage** est définie à pour un ensemble dénombrable $\mathcal{X}$: $\mu_c = \sum_{x \in \mathcal{X}} \delta_x$. On a alors pour tout ensemble $A \subset \mathcal{X}$,

$$\mu(A) = \sum_{x \in \mathcal{X}} \delta_x(A)$$

C'est à dire le nombre de points x qui sont dans l'ensemble A. L'intégrale par rapport à cette mesure vaut

$$\int_{\mathcal{X}} f(x)\mu_c(dx) = \sum_{x \in \mathcal{X}} f(x)$$

Dans beaucoup d'applications statistiques, les mesures de Lebesgue et de comptage sont les seules dont on a besoin pour décrire les situations rencontrées car les observations sont des valeurs numériques ou vectorielles (dans R^d) ou tout simplement des comptages.

Enfin, une **variable aléatoire** X à valeurs dans $\mathcal{X}$ est une application « mesurable » [1] pour laquelle on connaît seulement les probabilités $P(X \in A)$ pour tout $A \in \mathcal{T}$, $\mathcal{T}$ tribu définie sur $\mathcal{X}$. On appelle aussi $P_X(A)$ la probabilité associée, que l'on appelle la loi de la variable aléatoire X.

Définition 3. Un **modèle statistique** $\mathcal{M}$ sur $\mathcal{X}$ est une famille de lois de probabilités définie sur $\mathcal{X}$ (en fait une tribu $\mathcal{T}$ définie sur $\mathcal{X}$) :

$$\mathcal{M} = \{P_\theta, \theta \in \Theta\}$$

θ est un paramètre inconnu et Θ est l'espace des paramètres. Le plus souvent, l'espace des paramètres est inclus dans R^p ($\Theta \subset R^p$), dans ce cas on parle d'un **modèle paramétrique**. Parfois, le paramètre θ est une fonction, qui peut être vue comme un vecteur de dimension infinie. Dans ces cas-là, on dit que le modèle est non-paramétrique.

Le problème de la statistique est le plus souvent d'aider à prendre une décision à partir d'un échantillon, c'est-à-dire d'une suite d'observations $(x_1, x_2, \dots x_n) \in \mathcal{X}^n$ (un tableau excel, une image ou une suite d'image,...). La modélisation statistique consiste à supposer que ces données

1. Nous allons éviter de faire trop appel aux détails à la théorie de la mesure, mais il s'agit juste de dire qu'il s'agit d'une fonction (variable) pour laquelle on peut définir rigoureusement la mesure image, c'est à dire la probabilité $P(X \in A)$ pour $A \subseteq \mathcal{X}$.

observées sont des réalisations de variables aléatoires dont la **loi inconnue** appartient au modèle statistique $\mathcal{M}$. Le problème pratique revient alors à identifier quelle est la meilleure distribution de probabilité P_θ en accord avec les données, ou à choisir entre P_{θ_0} et P_{θ_1} à partir seulement de l'observation de $(x_1, x_2, \dots x_n)$.

Remarque 1. Cependant, une réflexion sur l'espace et les bonnes mesures se révèle nécessaire lorsque l'on considére des objets un peu plus complexes. Par exemple, si on veut considérer des processus stochastiques (à temps continu) cela revient à définir des mesures de probabilités sur des espaces de fonctions définies sur R. De même, pour des processus spatiaux qui sont indexés par R^2, R^3.

1.2 Densité et mesure dominée

1.2.1 Domination et densité

Dans beaucoup de cas, la mesure de probabilité, qui est une fonction définie sur une tribu, n'est pas très pratique ou maniable, et on lui préfère souvent la densité qui est une fonction directement définie sur l'espace des observations $\mathcal{X}$. Ceci est bien plus maniable pour les calculs et pour l'inférence. Nous revoyons donc ici la notion de densité, et le théorème de Radon-Nikodym qui lui est associé.

Si μ est une mesure sur $\mathcal{X}$ (par exemple λ ou μ_c), il est très facile de définir de nouvelles mesures sur ce même espace grâce au calcul intégral. En effet, pour toute fonction $f : \mathcal{X} \longrightarrow R$, positive et intégrable, on peut définir une mesure de probabilité ν :

$$\forall A \in \mathcal{T},\ \nu(A) = \int_A f(x)\mu(dx)$$

La fonction f est appelée la densité de ν relativement à μ, et on écrit $\nu = f \cdot \mu$.

Si on veut définir une probabilité, il suffit alors de normaliser la densité, i.e de prendre $x \mapsto \frac{f(x)}{\int_{\mathcal{X}} f(x)\mu(dx)}$. La quantité $\int_{\mathcal{X}} f(x)\mu(dx)$ est appelée la constante de normalisation (parfois très compliquée à calculer).

Exemple 2. On prend $h(x) = \exp\left(-x^2\right)$ avec $\mu = \lambda$.

On peut aussi se poser la question inverse : partant de deux mesures μ et ν définies sur l'espace $\mathcal{X}$, est-il possible de trouver une fonction f, telle que l'on puisse écrire $\nu = f \cdot \mu$? C'est à dire : est ce que je peux trouver une fonction $f \geq 0$ définie sur $\mathcal{X}$, telle que pour tout ensemble A, on puisse écrire

$$\nu(A) = \int_A f(x)\mu(dx)$$

La réponse est non en général, et cela est du aux ensembles de **mesure nulle**, i. e tel que $\mu(A) = 0$. En effet, la théorie de l'intégration dit que si A est un ensemble de mesure nulle (on dit **négligeable**), alors pour toute fonction f intégrable, on a $\int_A f(x)\mu(dx) = 0$.

Exemple 3. Pour la mesure de Lebesgue, on $\lambda(\{x\}) = 0$ et pour tout ensemble dénombrable $\lambda(\{x_1, x_2, \dots, x_n\}) = 0$. Dans R^2, $\lambda^{\otimes 2}(\{(t, \cos(t)) \,|t \in [0, 2\pi]\}) = 0$.

Si les mesures μ, ν sont telles que pour tout ensemble A, $\mu(A) = 0 \implies \nu(A) = 0$, alors on dit que la mesure μ domine la mesure ν. On dit aussi que ν est absolument continue relativement à la mesure μ, que l'on note $\nu \ll \mu$.

On voit que si la mesure ν admet une densité relativement à la mesure μ, elle est alors nécessairement dominée par μ. Le théorème de Radon-Nikodym affirme en fait que cette condition de domination est nécessaire et suffisante.

Radon-Nikodym. *Soit μ, ν deux mesures σ-finies définies sur $(\mathcal{X}, \mathcal{T})$, alors*

$$\nu \ll \mu \ \Leftrightarrow \exists\, f\ : \mathcal{X} \to R^+,\ \nu = f \cdot \mu$$

De plus, si f est intégrable par rapport à μ, alors ν est une mesure bornée.

La densité f est parfois appelée la **dérivée de Radon-Nikodym**, notée $\frac{d\mu}{d\nu}$.

Exemple 4. Par définition, la loi normale ou gaussienne $\mathcal{N}(\mu, \sigma^2)$ est la mesure de probabilité ayant pour densité relativement à la mesure de Lebesgue $f(x;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

Nous avons aussi le modèle exponentiel $\mathcal{E}(\lambda)$, tel que la densité par rapport à la mesure de Lebesgue sur R^+ est $f(x,\lambda) = \lambda e^{-\lambda x}$. Par rapport à la mesure de Lebesgue sur R, sa densité est $f(x,\lambda) = \lambda e^{-\lambda x} 1_{[0,+\infty[}(x)$.

Enfin le modèle de Poisson $\mathcal{P}(\lambda)$ de probabilités P_λ de densité $f(x,\lambda) = e^{-\lambda}\frac{\lambda^x}{k!}$ par rapport à la mesure de comptage sur $\mathbb{N}$: $\mu_c = \sum_{k=0}^{\infty} \delta_k$.

Dans les exemples ci-dessus, les modèles $\mathcal{M}$ sont appelés **modèles dominés**, car toutes les mesures de probabilités sont dominées par une même mesure. De plus, toutes les mesures de probabilités se dominent mutuellement. Ce n'est pas toujours le cas, par exemple si on considère le modèle uniforme suivant

$$\mathcal{U} = \left\{ P_\theta = \frac{1}{\theta} 1_{[0,\theta]}(x),\ \theta > 0 \right\}$$

Si $\theta < \theta'$, on a bien $P_\theta \ll P_{\theta'}$, mais l'inverse est bien évidemment faux (cf. $]\theta, \theta']$) : dans ce dernier cas, on ne peut donc pas trouver de fonction h intégrable tel que $P_{\theta'} = h \cdot P_\theta$.

Problème 1. On observe les valeurs $(0.9,\ 0.1,\ 1.4, 0.7)$.

— Quelle est la meilleure loi dans le modèle uniforme $\mathcal{U}$ en accord avec l'échantillon : $\theta = 1$ ou $\theta' = 1.5$?

— Quelle est meilleure loi dans le modèle Gaussien $\mathcal{N}(\mu, 1)$ en accord avec l'échantillon : $\mu = 1$ ou $\mu' = 1.5$?

1.2.2 Changement de mesure et forme générale des densités

Si 2 mesures de probabilités P et Q sont telles que $Q \ll P$, avec $Q = h \cdot P$: soient X v.a.r de loi P et Y v.a.r de loi Q, alors pour toute fonction φ t.q. $\varphi(X)$ est intégrable, on a

$$\begin{aligned} E_Q\left[\varphi(X)\right] &= \int \varphi(x) Q(dx) \\ &= \int \varphi(x) h(x) P(dx) \\ &= E_P\left[(\varphi h)(Y)\right] \end{aligned}$$

Si il existe une mesure μ qui domine $P = p \cdot \mu, Q = q \cdot \mu$ alors

$$\begin{aligned} E_P\left[\varphi(Y)\right] &= \int \varphi(y) p(y) \mu(dy) \\ &= \int \varphi(y) \frac{p(y)}{q(y)} q(y) \mu(dy) \\ &= E_Q\left[\varphi(Y)\frac{p}{q}(Y)\right] \end{aligned}$$

C'est ce que l'on appelle pour le **changement de mesure**. Le ratio $\frac{p}{q}$ est très utile pour l'inférence car il permet de comparer les densités entre elles (lorsque c'est possible) et leur *vraisemblance* au vu d'un échantillon.

Remarque 2. Le changement de mesure est aussi utilisé pour calculer des espérances dont on connaît la densité, mais que l'on ne sait pas forcément simuler. Cette méthode s'appelle l'échantillonnage d'importance (*Importance Sampling*).

En pratique, on rencontre les configurations suivantes :

1. densité continue sur R^p, cf les lois usuelles sur R ou R^+ : loi Gaussienne, Gamma, Pareto, Cauchy,...
2. sur $[a, b]$: loi uniforme, loi Beta,...
3. discret (sur N, $\{0, 1\}$, $\{A, T, G, C\}$,....) : Bernoulli, Géométrique, Binomiale négative, Poisson,...

4. mélange discret + continu (distribution avec des « atomes »). 2 cas standards et fréquents.
   (a) Phénomène de censure : une machine analyse la qualité de l'eau : X_i est la quantité de bactéries mesurée dans un échantillon d'eau. Cependant en dessous d'un certain seul $s > 0$, la machine indique seulement « indétectable ». On suppose que les variables $X_i \geq 0$, $i = 1, \ldots, n$ suivent une loi $\mathcal{E}(\lambda)$, et on veut connaître le paramètre λ (ou $1/\lambda$ quantité moyenne de bactéries). L'échantillon $(x_1, x_2, \ldots, x_n)$ peut se mettre sous la forme (à une permutation près) $(x_1 \ldots x_{n_s}, s, \ldots s)$. La valeur effectivement mesurée est $Y_i = X_i 1_{X_i \geq s} + s 1_{X_i \leq s}$. La loi de Y_i a un atome en s et admet une densité relativement à la mesure $\delta_s + \lambda$.
   (b) Pour une compagnie d'assurance, le montant X_k des primes annuelles versées au client k. Soit N_k le nombre de sinistres déclarés par le client k, et $Y_{j,k}$ la prime versée pour le j -ème sinistre. On suppose $N_k \sim \mathcal{P}(\lambda_k)$ et $Y_{j,k} \sim \exp(\mu_k)$; on a $X_k = \sum_{j=0}^{N_k} Y_{j,k}$. 0 est une atome de la distribution de Y_k ($P(X_k = 0) = P(N_k = 0) = \exp(-\lambda_k)$). C'est un processus de Poisson composé pour lequel on peut calculer la fonction génératrice des moments.

On appelle **support** d'une distribution le plus grand intervalle (ou ensemble) $supp(f)$ de $\mathcal{X}$ telle que la densité $f(x) > 0$ (f densité par rapport à μ). On peut dire que la forme classique d'une densité $f_X(x)$ de v.a. X relativement à une mesure de référence (Lebesgue) $\mu^{ref}(dx)$ est :

$$f_X(x) = \frac{1}{C} \exp\left(h(x)\right) 1_{supp(f_X)}(x)$$

avec $h : X \mapsto R$ fonction telle que $\int_{\mathcal{X}} \exp\left(h(x)\right) 1_{supp(f_X)}(x)$ soit finie, C constante de normalisation, et $supp(f_X)$ le support. Si on a un modèle paramétrique, alors

$$f(x, \theta) = \frac{1}{C(\theta)} \exp\left(h(x, \theta)\right) 1_{A_\theta}(x) \tag{1.1}$$

On voit que le support peut dépendre de θ (et bien sûr la constante de normalisation).

1.3 Différentes représentations de la loi d'une variable aléatoire réelle

Définition 4. La **fonction de répartition** d'une v.a.r X : Pour tout $x \in R$, $F_X(x) = P\left(X \leq x\right)$. Si X admet une densité par rapport à la mesure de Lebesgue alors $F_X(x) = \int_{-\infty}^{x} f_X(x)dx$.

Note 1. C'est une fonction continue à droite, qui a une limite à gauche. Elle est croissante, mais pas toujours strictement.

Définition 5. La **fonction quantile** d'une v.a.r X : Pour tout $\alpha \in [0, 1]$,

$$Q(\alpha) = \inf\left\{x \in R | F_X(x) > \alpha\right\}.$$

On l'appelle fonction de répartition (pseudo) inverse F^{-1} .

Note 2. On a toujours $F\left(Q(\alpha)\right) = \alpha$, mais on peut avoir $Q(F(x)) \neq x$, notamment lorsqu'il existe un segment $]a, b]$ tel que $F(a) = F(b)$, $a < b$.

On peut définir une fonction de répartition multivariée directement : si X_1, X_2 sont deux v.a.r. réelles, alors la fonction de répartition jointe est $F_{X_1,X_2}(x_1, x_2) = P\left(X_1 \leq x_1, X_2 \leq x_2\right)$. Par contre, la définition d'une fonction quantile multivariée est nettement moins évidente, car pour une probabilité α, il existe beaucoup de façon différente de trouver des pavés $]-\infty, a_1] \times]-\infty, a_2]$ telle que $P(X_1 \leq a_1, X_2 \leq a_2) = \alpha$.

Note 3. Lorsque l'on veut faire des calculs de lois, montrer des convergences (en loi),... on peut utiliser les densités (lorsqu'elles existent) et se ramener à du calcul intégral. On peut aussi regarder ce qu'il se passe pour toutes les espérances $E\left[\varphi(X)\right]$, pour des fonctions φ (intégrables pour P_X) et appartenant à une classe de fonctions tests, souvent les fonctions continues bornées.

Mais ce n'est pas toujours très pratique, et on peut considérer en fait une classe plus petite de fonctions tests $x \mapsto \exp(itx)$ ou $x \mapsto \exp(tx)$, pour des valeurs t bien choisies, c'est à dire telle que $E\left[\exp(tX)\right] < \infty$. En faisant varier le paramètre t, nous parcourons l'ensemble des fonctions

tests, et de plus nous tombons sur les transformations de Fourier et de Laplace respectivement. Ces transformations intégrales ont des propriétés analytiques très intéressantes qui permettent de ramener bon nombre de problèmes probabilistes à des problèmes de calcul. Typiquement le calcul de la somme de variables aléatoires indépendantes se ramène au simple produit des fonctions caractéristiques.... Les applications sont très nombreuses, et permettent de traiter des situations très complexes.

Définition 6. Pour X v.a.r, la **fonction caractéristique**

$$\phi_X(t) = E\left[\exp\left(itX\right)\right] = \int_{\mathcal{X}} \exp\left(itx\right) P_X(dx)$$

est définie pour tout $t \in R$. De plus, si X a une densité par rapport à la mesure de Lebesgue alors $\phi_X(t) = \int_{-\infty}^{\infty} \exp\left(itx\right) f_X(x)dx$, c'est la transformation de Fourier . Si X est un vecteur aléatoire dans R^d, la généralisation est directe : pour tout $\boldsymbol{t}$ dans R^d,

$$\phi_X(\boldsymbol{t}) = E\left[\exp\left(i\boldsymbol{t}^\top X\right)\right] = \int_{R^d} \exp\left(i\boldsymbol{t}^\top X\right) f_X(x)dx$$

Définition 7. Pour tout t dans R tel que $E\left[\exp(tX)\right] < \infty$, la **fonction génératrice des moments** est $M_X(t) = E\left[\exp\left(tX\right)\right]$, et on a alors

$$M_X(t) = \sum_{k\geq 0} \frac{t^k E\left[X^k\right]}{k!}$$

1.4 Comportement des variables aléatoires : indicateurs de position et de dispersion, inégalités de concentration

La fonction génératrice des moments $t \mapsto M_X(t)$ (lorsqu'elle existe) montre que les moments $E\left[X^k\right]$ permettent de caractériser complètement une variable. Cependant, les 2 premiers moments jouent un rôle fondamentale :

1. L'espérance $E\left[X\right] = \int_{\mathcal{X}} xP_X(dx)$, plus trivialement appelée la moyenne
2. La variance $V(X) = E\left[\left(X - E\left[X\right]\right)^2\right]$, qui est plus exactement le moment centrée d'ordre 2.

Une variable déterministe est une « variable qui ne varie pas », c'est à dire dont la loi est une distribution de Dirac δ_x , en un certain point x. La moyenne est la meilleure approximation déterministe d'une v.a. X au sens de la perte quadratique.

Proposition 1. *Si $E\left[X\right] < \infty$, alors*

$$E\left[X\right] = \arg\min_{m\in R} E\left[\left(X - m\right)^2\right]$$

La variance mesure la qualité de cette approximation $E\left[\left(X - E\left[X\right]\right)^2\right]$ (toujours au sens de la perte quadratique). Si on change la perte qu'on utilise pour mesurer l'erreur d'approximation, on modifie bien sûr l'approximation. Ainsi si on utilise la perte absolue, on a

Proposition 2. *Si F_X est la fonction de répartition de X, alors*

$$F_X^{-1}(\frac{1}{2}) = \arg\min_{m\in R} E\left[|X - m|\right]$$

La médiane $F_X^{-1}(\frac{1}{2})$ e st la meilleure approximation déterministe au sens de la perte absolue.

Démonstration. Si on X a une densité p/r à λ. $E[|X-m|] = \int_{-\infty}^{m}(x-m)f(x)dx - \int_{m}^{\infty}(x-m)f(x)dx$. Pour chercher le minimum, on dérive par rapport à m. On peut aussi écrire

$$\begin{aligned}
E[|X-m|] &= E[(X-m)1_{X\geq m}] - E[(X-m)1_{m\geq X}] \\
&= \int_0^{+\infty} P(X-m\geq t)dt + \int_0^{+\infty} P(m-X\geq t)\,dt \\
&= \int_0^{+\infty} (P(X\geq m+t) + P(m-t\geq X))\,dt \\
&= \int_0^{+\infty} (1-F(m+t)+F(m-t))\,dt
\end{aligned}$$

De la mêne façon, si on dérive par rapport à m, on trouve la condition :

$$\int_m^{+\infty} f(t)dt = \int_{-\infty}^{m} f(t)dt$$

d'où m est la médiane. □

Les quantiles permettent de localiser très efficacement les plages de valeurs les plus fréquentes d'une v.a.r. : on pensera à la boxplot, et les intervalles interquartiles, etc.... Les quantiles et la médiane sont d'excellents indicateurs de position d'une variable aléatoire, mais aussi permettent de décrire la dispersion.

Cependant la moyenne joue un rôle central, ceci est du au fait qu'une v.a. (ou un échantillon) ont tendance à se concentrer autour de la moyenne. Nous pouvons quantifier ce phénomène d'aggrégation (ou de concentration) grâce à des inégalités assez simples : les inégalités de Markov, et de Bienaimé-Chebychev en particulier.

Inégalité de Markov et Bienaimé-Chebychev.. *Soit X v.a.r et $\varphi : R \longrightarrow R$, fonction croissante telle que $E[\varphi(X)] < +\infty$ alors pour $t \in R$*

$$P(X\geq t) \leq \frac{E[\varphi(X)]}{\varphi(t)}$$

En particulier, si $V(X) < +\infty$

$$P(|X-E[X]|\geq t) \leq \frac{V(X)}{t^2}$$

Ces inégalités permettent de mesurer la probabilité d'une déviation de « t » de X de son espérance $E[X]$. Les amplitudes sont contrôlées par la variance ; plus les déviations t sont importantes, plus elles ont une faible probabilité d'arriver. Si on prend $t = k\sqrt{V(X)} = k\sigma$ (où σ est l'écart-type), on a donc

$$P(|X-E[X]|\geq 2\sigma) \leq \frac{1}{4}$$

pour toute variable admettant une variance. Ces inégalités permettent aussi de montrer que la convergence en moyenne quadratique implique la convergence en probabilités. Cependant ce n'est pas une inégalité très fine (cf. pour une gaussienne par exemple). L'inégalité de Chernoff permet d'avoir de meilleurs bornes en faisant apparaître la fonction génératrice des moments.

Inégalité de Chernoff. *Soit X une v.a.r, dont nous notons la **fonction génératrice des cumulants** $\psi_X(u) = \log E[\exp(uX)] = \log M_X(u)$ (i.e le logarithme de la fonction des moments). On introduit alors la **transformation de Cramér** définie par $\psi_X^*(t) = \inf_{u\geq 0}\{ut - \psi_X(u)\}$ pour tout les valeurs t t.q. $\psi_X^*(t)$ fini. Alors, on a*

$$P(X\geq t) \leq \exp(-\psi_X^*(t))$$

Démonstration. On part de l'inégalité pour tout $t > 0$, on a $1_{R^+}(x-t) \leq \exp(\lambda(x-t))$ pour tout $x \in R$ et pour tout $\lambda \geq 0$ (ou λ t.q $E[\exp(\lambda X)] < \infty$). Donc on en déduit que

$$\begin{aligned} P(X \geq t) &\leq E[\exp(\lambda(X-t))] \\ &\leq e^{-\lambda t} E[\exp(\lambda X)] = e^{-\lambda t + \log(M_X(\lambda))} \\ &\leq e^{-(\lambda t - \log(M_X(\lambda)))} \end{aligned}$$

Comme ceci est vrai pour tout $\lambda \geq 0$, on a donc que c'est vrai en particulier pour λ_t t.q. $\lambda_t = \arg\max_{\lambda \geq} \lambda t - \log(M_X(\lambda))$ et on pose $h_X(t) = \sup_{\lambda \geq 0} \lambda t - \log(M_X(\lambda))$. Pour que $h_X(t)$ soit bien défini, il faut s'assurer que l'optimisation est bien définie et que l'on puisse trouver λ_t. On peut montrer que $\lambda \mapsto \log(M_X(\lambda))$ est une fonction convexe et que la transfo de Cramér est en fait la transformation de Legendre Fenchel d'une application convexe, et que l'on peut bien définir une intervalle pour t afin de garantir la définition. □

On déduit de l'inégalité de Chernoff, les inégalités suivantes :

1. Le cas Gaussien $X \sim \mathcal{N}(0, \sigma^2)$: $M_X(u) = \exp(\frac{u^2\sigma^2}{2})$, $\psi_X(u) = \frac{u^2\sigma^2}{2}$ et $u_t = \frac{t}{\sigma^2}$ et pour tout $t > 0$ $\psi^*(t) = \frac{t^2}{2\sigma^2}$ et donc
$$P(X \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$
2. Le cas Poisson $X \sim \mathcal{P}(\mu)$: $M_X(u) = \exp(-u\mu - \mu)\exp(\mu e^u)$, $\psi_X(u) = \mu(e^u - u - 1)$ et $u_t = \log(1 + \frac{t}{\mu})$ et pour tout $t > 0$ $\psi^*(t) = h(t)$ avec $h(x) = (1+x)\log(1+x) - x$, $x \geq -1$ et donc pour tout $t > 0$
$$P(X \geq t) \leq \exp\left(-\mu h\left(\frac{t}{\mu}\right)\right)$$
On remarque que $h(x) \leq \frac{x^2}{2}$, d'où $P(X \geq t) \leq \exp\left(-\frac{t^2}{2\mu}\right)$.
3. Le cas Bernoulli $P(X = 1) = p$: $M_X(u) = (1-p) + pe^u$ et $\psi_X(u) = \log(1 - p + pe^u) - \lambda p$ et $u_t = \log\frac{(1-p)(p+t)}{p(1-p-t)}$ et pour tout $t \in]p, 1[$, $\psi_X^*(t) = (1-t)\log\frac{1-t}{1-p} + t\log\frac{t}{p}$ et donc
$$P(X \geq t) \leq \exp(-\psi_X^*(t))$$

Un intérêt de l'inégalité de Chernoff est qu'elle est bien adaptée aux sommes de variables aléatoires indépendantes. En effet, si $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, alors on a $\psi_{\bar{X}_n}(u) = n\psi_X(\frac{u}{n})$. Par conséquent, la transformée de Cramér est

$$\begin{aligned} \psi_{\bar{X}_n}^*(t) &= \inf_{u \geq 0}\left\{\lambda t - n\psi_X(\frac{u}{n})\right\} \\ &= n \inf_{\lambda \geq 0}\left\{\frac{u}{n}t - \psi_X(\frac{u}{n})\right\} \\ &= n\psi_X^*(t) \end{aligned}$$

donc

$$P\left(\bar{X}_n \geq t\right) \leq \exp(-n\psi_X^*(t))$$

Et donc pour de nombreuses lois (dites sous-gaussiennes), nous obtenons que

$$P\left(\bar{X}_n \geq t\right) \leq \exp\left(-\frac{nt^2}{2\sigma^2}\right)$$

Cette inégalité est vraie pour tout n (y compris $n = 2, 3$), et ne dépend pas d'une approximation asymptotique obtenue par exemple par le Théorème Central Limite. Par exemple, pour des v.a. de Bernoulli $\mathcal{B}(1, p)$, une inégalité similaire à celle de Chernoff (c'est d'inégalité d'Hoeffding) montre que

$$P\left(\left|\bar{X}_n - p\right| \geq t\right) \leq 2\exp\left(-2nt^2\right)$$

2 Le modèle d'échantillonnage et les modèles statistiques

Nous précisons le mécanisme de génération des observations, et donc la structure des modèles que nous allons considérer.

2.1 L'échantillon et l'hypothèse i.i.d

Une **échantillon indépendant et identiquement distribué (*i.i.d*)** est un $n-$uple de variables aléatoires $(X_1, X_2, \ldots, X_n)$ à valeur dans $\mathcal{X}$ (soit R, $\{0,1\}$,...) qui sont de même loi et indépendantes. En particulier si les X_i admettent une densité $x \mapsto f(x)$ relativement à une mesure de référence μ_{ref}, alors la loi jointe (par rapport à $\mu_{ref}^{\otimes n}$) s'écrit

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$$

L'échantillon i.i.d est un cas restreint de description probabiliste de la loi d'une série d'observations.

Exemple 5. En effet, dans une étude clinique contenant n individus de poids, d'âge,... on s'intéresse à la variable $X_i = \begin{cases} 0 & \text{pas de cancer} \\ 1 & \text{apparition d'un cancer} \end{cases}$, $i = 1, \ldots n$. On peut supposer que les variables $X_i \sim \mathcal{B}(1, p_i)$ sont indépendantes, mais la probabilité p_i dépend de facteur génétique, clinique, et de mode de vie, et l'échantillon n'est pas identiquement distribué.

Exemple 6. On a n sites de prélèvement d'eau de pluie sur l'agglomération d'Evry. On note X_i, $i = 1, \ldots, n$ la quantité d'eau récoltée en chaque site. On suppose que les sites sont suffisamment rapprochés pour que l'on puisse considérer que $X_i \sim \mathcal{N}(\mu, \sigma^2)$ pour tout $i = 1, \ldots, n$. Par contre, deux sites proches ne sont pas indépendants : ils auront tendance à recevoir une quantité de précipitations « similaires », il y a une corrélation non-nulle. On suppose souvent que le vecteur aléatoire $(X_1, \ldots, X_n)$ est un vecteur gaussien $\mathcal{N}(\mu 1_n, \Sigma)$. La matrice de covariance Σ donne la covariance $\Sigma_{ij} = \text{cov}(X_i, X_j))$ entre deux sites. On peut modéliser la dépendance spatiale en introduisant la fonction de covariance suivante $\Sigma_{ij} = \sigma^2 \exp(-\left(\frac{d(i,j)}{h}\right)^2$, où $d(i,j)$ est la distance (euclidienne) entre les sites i et j (on peut vérifier que cette fonction est de type positif et qu'elle définit bien une covariance). Les observations X_i ont bien la même loi, mais ne sont pas indépendantes.

2.2 Des modèles classiques

1. Le modèle gaussien. $\mathcal{N}$
2. Le modèle Bernoulli. $\mathcal{B}$
3. Le modèle Weibull

$$\mathcal{W} = \left\{ P_{a,b} = f(x; a, b) \cdot \lambda \text{ avec } f(x; a, b) = \frac{a}{b} \left(\frac{x}{b}\right)^{a-1} \exp\left(-\left(\frac{x}{b}\right)^{a}\right) 1_{[0,+\infty[}(x);\ a, b > 0 \right\}.$$

4. Le modèle Uniforme sur $[0, \theta]$ $(\mathcal{U})$, loi exponentielle translatée

$$\mathcal{E} = \left\{ f(x, \lambda, a) = \exp\left(-\lambda(x-a)\right) 1_{[a,+\infty[}(x); a, \lambda > 0 \right\}.$$

5. Le modèle de mélange de 2 gaussiennes. On suppose que (X, Y) sont deux v.a. t.q. $X \sim \mathcal{B}(1, p)$ et on la loi de Y conditionnelle est donnée par : $\mathcal{L}(Y|X = 0) = \mathcal{N}(\mu_0, \sigma_0^2)$ et $\mathcal{L}(Y|X = 1) = \mathcal{N}(\mu_1, \sigma_1^2)$. La loi de Y (loi marginale) est un mélange de gaussiennes.

2.3 La vraisemblance d'un échantillon

On suppose que nous avons un échantillon i.i.d $X_1, \ldots, X_n$ (à valeurs dans $\mathcal{X}$) et on suppose que la loi commune appartient à un modèle statistique $\mathcal{M} = \{P_\theta = f(x,\theta) \cdot \mu_{ref},\ \theta \in \Theta\}$ (avec une mesure de référence : Lebesgue, ou mesure comptage). On rappelle que la densité du n-uple est

$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

La densité est une fonction définie sur $\mathcal{X}^n$.

Définition 8. Pour tout $(x_1, x_2, \ldots x_n) \in \mathcal{X}^n$, on définit la **vraisemblance** comme la fonction définie sur Θ à valeurs dans R^+ :

$$\theta \mapsto L(x_1, \ldots, x_n, \theta) = f_{X_1 \ldots X_n}(x_1, \ldots, x_n, \theta)$$

Exemple 7. Si $(x_1, \ldots, x_5) = (-0.14, -0.07, 0.19, -1.94, 2.43)$ est la réalisation d'un échantillon gaussien dont la loi commune est $\mathcal{N}(\mu, 1)$ pour $\mu \in R$ inconnue. Alors la vraisemblance est

$$\begin{aligned} L(x_1, \ldots, x_5, \mu) &= (2\pi)^{-5/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{5} (x_i - \mu)^2 \right) \\ &= (2\pi)^{-5/2}\, e^{-\frac{1}{2}\left((-0.14-\mu)^2 + (-0.07-\mu)^2 + (0.19-\mu)^2 + (-1.94-\mu)^2 + (-2.43-\mu)^2\right)} \end{aligned}$$

La vraisemblance $L(x_1, \ldots, x_n, \theta)$ est une fonction de l'échantillon, c'est donc une fonction aléatoire, dont les propriétés vont nous permettre de faire de l'inférence. De manière générale, on appelle **statistique** $S(X_1, \ldots, X_n)$ toute fonction de l'échantillon avec $S : \mathcal{X}^n \longrightarrow R^m$ (ou d'autres valeurs). On peut définir beaucoup de statistiques différents mais seules quelques unes sont vraiment intéressantes (et ça dépend du modèle utilisé).

On utilise souvent la **log-vraisemblance** $\mathcal{L}(X_1, \ldots, X_n, \theta) = \log L(X_1, \ldots, X_n, \theta)$ qui est bien définie lorsque $X_i \sim P_{\theta'}$ et que $P_\theta \ll P_{\theta'}$. En effet, si $P_{\theta'}$ n'est pas dominé par P_θ, alors on a $\mathcal{L}(X_1, \ldots, X_n, \theta) = -\infty$ avec une probabilité non-nulle. Comme on le verra, ceci est relié à la dérivée de Radon-Nikodym $\frac{dP_{\theta'}}{dP_\theta}$.

La vraisemblance est une des statistiques de référence pour l'inférence, i.e choisir entre plusieurs lois. Si on reprend l'exemple du choix entre deux lois P_{θ_1} et P_{θ_2} (mutuellement dominée), une règle simple de sélection du meilleur paramètre et de prendre celui qui maximise la vraisemblance de l'échantillon (ou la log-vraisemblance). Supposons que l'échantillon $(X_1, \ldots, X_n)$ est de loi commune P_{θ_1}, alors nous allons vérifier que la vraisemblance aura tendance à être plus grande en θ_1 qu'en θ_2. En effet , dire que $\mathcal{L}(x_1, \ldots, x_n, \theta) > \mathcal{L}(x_1, \ldots, x_n, \theta')$, est équivalent à dire que

$$\sum_{i=1}^{n} \log \frac{f(x_i, \theta)}{f(x_i, \theta')} > 0$$

Or en moyenne, nous avons

$$E_{\theta_1}\left[\sum_{i=1}^{n} \log \frac{f(X_i, \theta_1)}{f(X_i, \theta_2)} \right] = n E_{\theta_1}\left[\log \frac{f(X, \theta_1)}{f(X, \theta_2)} \right]$$

et on remarque que $E_{\theta_1}\left[\log \frac{f(X,\theta_1)}{f(X,\theta_2)}\right] = -E_{\theta_1}\left[\log \frac{f(X,\theta_2)}{f(X,\theta_1)}\right] \geq -\log E_{\theta_1}\left[\frac{f(X,\theta_2)}{f(X,\theta_1)}\right]$ par application de l'inégalité de Jensen [2]. Par conséquent,

$$\begin{aligned} E_{\theta_1}\left[\frac{f(X,\theta_2)}{f(X,\theta_1)}\right] &= \int \frac{f(x,\theta_2)}{f(x,\theta_1)} f(x,\theta_1)\mu(dx) \\ &= \int f(x,\theta_2)\mu(dx) = 1 \end{aligned}$$

2. Jensen : Pour toute fonction convexe φ, on a $E[\varphi(X)] \geq \varphi(E[X])$. L'inégalité est stricte si la fonction est strictement convexe. On rappelle donc que la fonction $-\log$ est strictement convexe.

et ainsi $E_{\theta_1}\left[\log \frac{f(X,\theta_1)}{f(X,\theta_2)}\right] \geq -\log(1) = 0$. Donc on a intérêt à sélectionner le paramètre qui maximise la log-vraisemblance, car en moyenne elle est plus élévée pour le bon paramètre que pour les autres. Cette propriété remarquable motive l'introduction de la

Divergence de Kullback-Leibler.

Soit P_θ et $P_{\theta'}$ 2 probabilités telles que $P_\theta \ll P_{\theta'}$, si on note $\frac{dP_\theta}{dP_{\theta'}}$ la dérivée de Radon-Nikodym alors on définit la divergence de KL de $P_{\theta'}$ relativement à P_θ

$$\begin{aligned} D(P_\theta \| P_{\theta'}) &= \int_{\mathcal{X}} \log \frac{dP_\theta}{dP_{\theta'}} P_\theta(dx) \\ &= \int_{\mathcal{X}} \log \frac{f(x,\theta)}{f(x,\theta')} f(x,\theta)\mu(dx) \end{aligned}$$

si les probabilités admettent une densité relativement à une mesure de référence μ t.q. $P_\theta = f_\theta \cdot \mu$ et et $P_{\theta'} = f_{\theta'} \cdot \mu$. C'est une pseudo-distance :

$$\begin{cases} \forall \theta, \theta' \in \Theta, & D(P_\theta \| P_{\theta'}) \geq 0 \\ D(P_\theta \| P_{\theta'}) = 0 & \Leftrightarrow \theta = \theta' \end{cases}$$

Cependant, la divergence n'est pas symétrique et ne vérifie pas l'inégalité triangulaire.

Si on a en tête les inégalités de concentration de type Markov ou Chernoff, la variable

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i,\theta_1)}{f(X_i,\theta_2)},\ i = 1, \dots n$$

ne sera donc pas très loin de $D(P_{\theta_1} \| P_{\theta_2})$: la maximisation de la vraisemblance de l'échantillon est donc une heuristique qui devrait permettre de sélectionner la meilleure loi en accord avec les données. On verra plus tard que la log-vraisemblance contient toute l'information pertinente sur le paramètre inconnu θ.

Exemple 8 (Calcul de la divergence entre deux Gaussiennes et deux Bernoulli.)**.**

Divergence entre Gaussiennes $D\left(\mathcal{N}(\mu_1,\sigma_1^2) \| \mathcal{N}(\mu_2,\sigma_2^2)\right) = \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_1-\mu_2)^2}{\sigma_2^2} - 1 + \log\frac{\sigma_2}{\sigma_1}\right)$

Divergence entre Bernoulli $D\left(\mathcal{B}(1,p_1) \| \mathcal{B}(1,p_2)\right) = p_1 \log\frac{p_1}{p_2} + (1-p_1)\log\frac{1-p_1}{1-p_2}$.

La divergence de Kullback-Leibler est utilisé comme une « distance » entre distributions de probabilités. Bien sûr, on peut définir de nombreuses autres distances telles que la distance en Variation Totale $TV(P,Q) = \sup_A |P(A) - Q(A)|$, ou encore des distances plus « naturelles » telle que les normes L^1 ou L^2 entre les densités $\int |p-q|(x)\mu(dx)$ ou $\int |p-q|^2(x)\mu(dx)$. Le problème de ces distances est qu'elle sont difficiles à calculer à partir d'un échantillon obervée, alors que la divergence de KL s'exprime comme *l'espérance* d'une certaine statistique, on pourra donc la calculer à partir des données.

La vraisemblance n'est pas la seule statistique que l'on peut définir. Une statistique est aussi un moyen de résumer l'information : on passe d'un échantillon de dimension n à une information condensée de dimension réduite :

1. Les moments empiriques $\frac{1}{n}\sum_{i=1}^n X_i^k$ (pour $k = 1 \dots K$)
2. Les quantiles empiriques et statistiques de rang, $M_n = \max(X_1, \dots, X_n)$ et $\min(X_1, \dots, X_n)$.
3. Etc

La liste est longue et n'est limitée que par l'imagination. Calculer une statistique d'un échantillon dans R, par exemple $\bar{X}_n$, revient à passer d'un vecteur de taille n à une simple valeur, i.e un vecteur de taille 1. La question qui se pose est : que perd-on comme information lors de cette réduction ? Cela dépend en fait du modèle que l'on utilise.

Exemple 9. On suppose que $(X_1, \dots, X_n) \sim \mathcal{E}(\lambda)$. Si on considére $m_n = \min(X_1, \dots, X_n)$ et $\bar{X}_n$, on peut vérifier que ces statistiques donnent de l'information sur le paramètre d'intérêt λ. En effet, on a $P(m_n \leq x) = (1 - (1 - F_\lambda(x))^n) = 1 - \exp(-n\lambda x)$, ce qui signifie que $m_n \sim \mathcal{E}(n\lambda)$ et donc $E_\lambda[n \times m_n] = \frac{1}{\lambda}$ et $V(m_n) = \frac{1}{n\lambda^2}$. De manière similaire, on peut vérifier que $E[\bar{X}_n] = \frac{1}{\lambda}$ et $V(\bar{X}_n) = \frac{1}{n\lambda^2}$.

Dans le cas d'un échantillon gaussien $(X_1, \dots, X_n) \sim \mathcal{N}(0, \sigma^2)$, avec σ inconnu, les moments empiriques $M_1 = \frac{1}{n}\sum_{i=1}^n X_i$ ou à $M_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$ n'apportent pas du tout la même information sur la loi de l'échantillon. M_1 est un mauvais résumé (qui ne donne aucune info sur σ), alors que M_2 décrit la variabilité.

Dans l'exemple précédent, on voit que l'on peut construire des statistiques très différentes qui semblent contenir la même information, ou au contraire assez proches qui ne donne pas du tout la même information (voire pas du tout). Dans les cas vues précédemment, on peut essayer d'identifier dans un premier temps les résumés qui sont pertinents et les autres. Une première formalisation possible du concept de « résumé sans perte » est donnée par la notion de statistique exhaustive.

Définition 9. Statistique Exhaustive

Soit $T : (x_1 \dots, x_n) \longrightarrow R^d$ une fonction définie sur $\mathcal{X}^n$, et soit $X_1, \dots X_n$ un échantillon i.i.d dont la loi est dans $\mathcal{M} = \{P_\theta = f_\theta \cdot \mu | \theta \in \Theta \subset R^p\}$. On dit que $T(X_1, \dots X_N)$ est une statistique exhaustive si la loi de l'échantillon $(X_1, \dots, X_n)$ sachant $T(X_1, \dots, X_n)$ ne dépend pas de θ.

Cette définition indique qu'il faut calculer la loi conditionnelle et par exemple la densité conditionnelle $f_{(X_1 \dots X_n)|T(X_1,\dots,X_n)}(x_1, \dots x_n|t)$. A partir de là, on peut déduire une caractérisation assez simple qui dépend de la forme de la vraisemblance.

Lemme de factorisation.. *Soit $X_1, \dots X_n$ un échantillon i.i.d dont la loi est dans*

$$\mathcal{M} = \{P_\theta = f_\theta \cdot \mu | \theta \in \Theta \subset R^p\}.$$

Alors, si il existe 3 fonctions : $T : \mathcal{X}^n \longrightarrow R^d$, $g : R^d \times \Theta \longrightarrow R^d$ et $h : \mathcal{X}^n \longrightarrow R^+$ telles que l'on peut écrire la densité de l'échantillon

$$f(x_1 \dots, x_n, \theta) = g\left(T(x_1 \dots, x_n), \theta\right) h(x_1, \dots, x_n)$$

alors $T(X_1, \dots, X_n)$ est une statistique exhaustive pour le modèle $\mathcal{M}$.

On remarque au passage que la décomposition (T, g, h) n'est pas unique. Il y a une infinité de statistique exhaustive, et par le lemme de factorisation on voir facilement que si T est une statistique exhaustive, alors si S est une statistique t.q on peut écrire $T(X_1, \dots, X_n) = G\left(S(X_1, \dots, X_n)\right)$ (i.e. $T = G(S)$) alors S est aussi exhaustive. On dit que S est une **statistique exhaustive minimale** si pour toute statistique exhaustive, on peut trouver H telle que $S = H(T)$. Une statistique exhaustive minimale est donc le meilleur résumé (le plus compact) d'un échantillon.

Exemple 10. Le modèle gaussien multivarié $\mathcal{M} = \{\mathcal{N}(0, \Sigma), \Sigma \in SDP^+\}$. Une statistique exhaustive est la covariance empirique. On suppose que les X_i sont des vecteurs gaussiens de R^q telle que $E[X_i] = 0$, et le paramètre à estimaer est la matrice de covariance. La densité s'écrit donc $f(x, \Sigma) = (2\pi)^{-q/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}x^\top \Sigma x\right)$. La vraisemblance de l'échantillon est

$$\begin{aligned} L(x_1, \dots, x_n, \Sigma) &= (2\pi)^{-nq/2} \det(\Sigma)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^\top \Sigma^{-1} x_i\right) \\ &= (2\pi)^{-nq/2} \det(\Sigma)^{-n/2} \left(-\frac{1}{2}\sum_{i=1}^n \mathrm{Tr}(x_i^\top \Sigma^{-1} x_i)\right) \end{aligned}$$

où Tr dénote l'opérateur matriciel Trace : Si A matrice carrée dans $R^{q\times q}$, alors $\mathrm{Tr}(A) = \sum_{i=1}^n A_{ii}$. Parmi les propriétés classiques, en plus de la linéarité, on a $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ et $\mathrm{Tr}(A) = \mathrm{Tr}(A^\top)$. Enfin, on rappelle que si $a \in R$, $a = \mathrm{Tr}(a)$. Ce rappel montre que l'on peut écrire

$$\begin{aligned} \sum_{i=1}^n \mathrm{Tr}(x_i^\top \Sigma x_i) &= \sum_{i=1}^n \mathrm{Tr}(\Sigma^{-1} x_i x_i^\top) \\ &= \mathrm{Tr}(\Sigma \sum_{i=1}^n x_i x_i^\top) \end{aligned}$$

Si on note $T(x_1,\dots,x_n)=\frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ qui est matrice définie positive de $R^{q\times q}$, on voit que

$$L(x_1,\dots,x_n,\Sigma)=(2\pi)^{-nq/2}\det(\Sigma)^{-n/2}\exp\left(-\frac{n}{2}\mathrm{Tr}\left(\Sigma^{-1}T\right)\right)$$

Donc $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ est une statistique exhaustive d'après le lemme de factorisation avec $g(t,\Sigma)=\det(\Sigma)^{-n/2}\exp(-\frac{n}{2}\mathrm{Tr}\,(\Sigma T))$ et $h(x_1,\dots x_n)=(2\pi)^{-nq/2}$. On remarque au passage que $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ est la covariance empirique.

On remarque qu'il suffit de regarder les statistiques exhaustives pour comparer et choisir des lois si l'on veut se baser sur le principe du maximum de vraisemblance. En effet, on a vu qu'il suffisait de regarder les rapports de vraisemblance pour calculer une distance entre des lois de probabilités et faire un choix entre deux lois :

$$\begin{aligned}\frac{f(x_1\dots,x_n,\theta_1)}{f(x_1\dots,x_n,\theta_2)}&=\frac{g\left(T(x_1\dots,x_n),\theta_1\right)h(x_1,\dots,x_n)}{g\left(T(x_1\dots,x_n),\theta_2\right)h(x_1,\dots,x_n)}\\&=\frac{g\left(T(x_1\dots,x_n),\theta_1\right)}{g\left(T(x_1\dots,x_n),\theta_2\right)}\end{aligned}$$

Dans ce cas là, on voit que l'échantillon ne rentre en compte dans la comparaison entre les deux lois qu'à travers la valeur de la statistique exhaustive $t=T(x_1,\dots,x_n)$ et le ratio $\frac{g(t,\theta_1)}{g(t,\theta_2)}$.

Exemple 11. On considère le modèle de Weibull

$$\mathcal{W}=\left\{f(x;a,b)=\frac{a}{b}\left(\frac{x}{b}\right)^{a-1}\exp\left(-\left(\frac{x}{b}\right)^a\right)1_{[0,+\infty[}(x);\ a,b>0\right\}$$

On suppose que b est fixé (égale à 1). La vraisemblance est

$$L(x_1,\dots,x_n;a)=\exp\left(-\sum_{i=1}^n x_i^a\right)a^n x_i^{a-1}\prod_{i=1}^n 1_{[0,+\infty[}(x_i)$$

Dans ce cas, si on cherche $g(T,a)=\exp\left(-\sum_{i=1}^n x_i^a\right)a^n x_i^{a-1}$, on voit que l'on ne peut pas faire autrement que prendre $T(x_1,\dots,x_n)=(x_1,\dots,x_n)$, et $g(T,a)=\exp\left(-\sum_{i=1}^n x_i^a\right)a^n x_i^{a-1}$. On peut pas réduire la dimension de l'échantillon facilement, en raison de la fonction puissance.

Définition 10. Statistique Libre en loi

Soit $S:(x_1\dots,x_n)\longrightarrow R^d$ est une statistique libre en loi de X si la loi de $S(X)$ ne dépend pas de θ. On dit aussi statistique pivotale.

Les notions de statistique exhaustive minimale et de statistique libre permettent de séparer l'échantillon $(X_1,\dots,X_n)$ en deux parties : un résumé qui compresse les données et qui en garde toute l'information sur le paramètre θ, et une partie statistique libre qui contient l'aléa de l'échantillon ne dépendant pas du paramètre θ.

Finalement, on peut voir que dans beaucoup de cas, les statistiques exhaustives sont très reliés au paramètre θ inconnu, et dans beaucoup de cas, on a en fait

$$E_\theta\left[T(X_1,\dots X_n)\right]=\theta$$

ou alors $T(X_1,\dots,X_n)$ est « proche » de θ. Nous étudions dans la section suivante les propriétés des estimateurs, et les méthodes de constructions.

Deuxième partie

Estimation ponctuelle et région de confiance

3 Estimation ponctuelle

Soit $\mathcal{M}$ un modèle statistique, on appelle estimateur de θ une statistique $T(X_1, \ldots, X_n)$ t.q θ est approché par T. Intuitivement, il y a deux manières de définir les attentes que l'on peut avoir d'un estimateur :

— Une propriété en moyenne du type $E_\theta [T(X_1, \ldots X_n)] = \theta$, qui permette de garantir que les probabilités $P_\theta (|T(X_1, \ldots X_n) - \theta| \geq \epsilon)$ soient faibles pour ϵ petit.
— Une convergence asymptotique, i.e $T(X_1, \ldots X_n) \longrightarrow \theta$ en un sens à préciser.

Ainsi on parle souvent d'estimateur convergent. Cependant il faut préciser le type de convergence stochastique que l'on considère :

1. la convergence en probabilités : Pour tout $\epsilon > 0$, $P_\theta (|T(X_1, \ldots X_n) - \theta| \geq \epsilon) \longrightarrow 0$ pour $n \to \infty$,
2. la convergence presque-sûre : $P_\theta (\limsup |T(X_1, \ldots X_n) - \theta| \geq \epsilon) = 0$,
3. la convergence en moyenne quadratique : $E_\theta \left[\|T(X_1, \ldots X_n) - \theta\|^2 \right] \to 0$,
4. et enfin, la convergence en loi de $Y_n = T(X_1, \ldots X_n) - \theta$ vers Y tel que pour tout f continue bornée, $E_\theta [f(Y_n)] \longrightarrow E_\theta [f(Y)]$.

En fait, il n'y a pas réellement de définition d'un estimateur. Ce qui est important est de trouver un moyen de départager bonnes ou mauvaises approximations de θ, en se concentrant sur la mesure de la perte ou de l'erreur. Pour cela, on utilise le langage de la théorie de la décision et des fonctions de « perte » qui vont mesurer la qualité de ces approximations. Soit $L : \Theta \times \Theta \longrightarrow R^+$ une fonction de coût. On appelle **Risque** de $T_n = T(X_1, \ldots X_n)$ le coût espéré

$$R(T_n, \theta) = E_\theta [L(T_n, \theta)]$$

Une fois que l'on a défini un tel critère, on peut essayer de déterminer le meilleur estimateur au sens du risque R. Pour faire ceci, on introduit la notion de risque minimax et d'estimateur minimax, i.e

$$\mathcal{R}(T_n) = \sup_{\theta \in \Theta} E_\theta [L(T_n, \theta)]$$

et on recherche le meilleur estimateur qui minimise $\mathcal{R}(T_n)$ (sur l'espace de tous les paramètres possibles). Lorsque $\theta \subset R^p$, le risque le plus couramment utilisé est le risque quadratique, qui s'analyse via la décomposition Biais-Variance :

Décomposition Biais-Variance. *Soit $T_n = T(X_1, \ldots, X_n)$ une statistique de variance finie. Alors on décompose l'Erreur Quadratique Moyenne*

$$\begin{aligned} R(T_n, \theta) &= E_\theta \left[\|T_n - \theta\|^2 \right] \\ &= \|E_\theta [T_n] - \theta\|^2 + Tr(Var_\theta (T_n)) \end{aligned}$$

*La quantité $E_\theta [T_n] - \theta$ est le **biais** de l'estimateur, et le second terme est sa **variance**.*

Exemple 12. Le cadre de la théorie de la décision est très utilisé lorsqu'il s'agit de problème de prédiction. Si on veut prédire Y à l'aide de v.a.r $X_1, \ldots, X_n$, cela veut dire que l'on veut trouver la fonction (ou la statistique) H qui minimise la perte moyenne $E_{(\boldsymbol{X}, Y)} [L (Y, H(X_1, \ldots, X_n))]$.

— Si $Y \in R$ et $L(y, h) = (y - h)^2$, on a alors $H(x_1, \ldots x_n) = E [Y | X_1, \ldots, X_n]$.
— Si $Y \in \{0, 1\}$ et $L(y, y') = \delta_{y \neq y'}$, on a $H(x_1, \ldots, x_n) = \arg\max_{y \in \{0,1\}} P(Y = y | X_1, \ldots, X_n)$.

Un estimateur de θ, noté le plus souvent $\hat{\theta}$ (en omettant souvent la dépendance en l'échantillon), est une statistique $\hat{\theta}(X_1, \ldots, X_n)$ qui est « proche » en un certain sens de θ. Une approche classique est de chercher à minimiser le risque quadratique en deux temps : d'abord on cherche des estimateurs sans biais i.e tels que $E_\theta\left[\hat{\theta}(X_1, \ldots, X_n)\right] = \theta$ (pour tout θ) et dans un second temps on cherche à minimiser la variance. C'est ce qui explique que beaucoup de résultats classiques concernent des estimateurs sans biais et de variance minimale. Cette seconde étape est motivée par l'existence d'une procédure assez systématique pour réduire la variance.

Théorème de Rao-Blackwell.

Soit $\mathcal{M}$ un modèle statistique, $X_1, \ldots, X_n$ un échantillon iid, et $T(X_1, \ldots, X_n)$ une statistique exhaustive. Soit $\hat{\theta}_1$ un estimateur de θ sans biais alors $\hat{\theta}_2 = E_\theta\left[\hat{\theta}_1 | T\right]$ a un risque quadratique plus faible.

Démonstration. On calcule l'espérance du risque quadratique

$$E_\theta\left[\left(\theta - \hat{\theta}_1\right)^2\right] = E_\theta\left[E_\theta\left[\left(\theta - \hat{\theta}_1\right)^2 | T\right]\right]$$

et on remarque que $E_\theta\left[\left(\theta - \hat{\theta}_1\right)^2 | T\right] \geq E_\theta\left[\theta - \hat{\theta}_1 | T\right]^2 = \left(\theta - E_\theta\left[\hat{\theta}_1 | T\right]\right)^2$. Donc $\hat{\theta}_2$ a un risque quadratique plus faible, mais le problème est que ce n'est pas forcément une statistique observable, car son calcul dépend de la connaissance de θ, le paramètre inconnu que l'on souhaite estimer. Cependant, comme on considère une statistique exhaustive T, et que la loi de $X_1, \ldots, X_n | T$ ne dépend donc pas de θ, nous avons en fait $E_\theta\left[\hat{\theta}_1 | T\right] = E\left[\hat{\theta}_1 | T\right]$ (qui ne dépend pas de θ), et donc la statistique améliorée est effectivement calculable. □

3.1 Construction d'estimateur

Un estimateur $\hat{\theta}$ est une statistique construite pour donner une estimation du paramètre d'intérêt θ lorsque l'on suppose que l'échantillon $(X_1, \ldots, X_n)$ provient du modèle $\mathcal{M} = \{P_\theta, \theta \in \Theta\}$. L'objectif de cette section est d'identifier des principes généraux pour construire des estimateurs « automatiquement »et dépasser le caractère « ad hoc » entre-aperçu jusqu'à présent.

Pour clarifier la discussion, on notera **le vrai paramètre** θ^*, c'est-à-dire le paramètre de la loi des observations $X_1, \ldots, X_n$ (loi commune est P_{θ^*}). Le problème de l'estimation est donc de calculer « le meilleur θ » parmi tous les candidats disponibles dans l'ensemble Θ.

3.1.1 La méthode des moments.

Une idée simple est basée sur le fait que le paramètre θ est souvent relié à la moyenne ou la variance de la variable aléatoire X. Par exemple lorsque l'on a un modèle Gaussien $\mathcal{N}(\theta, 1)$ ou $\mathcal{E}(\frac{1}{\theta})$, on a $E_\theta[X] = \theta$. Un estimateur assez naturel est donc $\hat{\theta}(X_1, \ldots X_n) = \bar{X}_n$. Plus généralement, on peut obtenir une relation relativement explicite entre les moments théoriques (lorsqu'ils sont finis) et les paramètres. Si X est une v.a.r, on peut trouver une relation entre θ et $E_\theta\left[X^k\right]$, ce qui revient à dire qu'il existe une fonction $m_k : \Theta \longrightarrow R$ telle que $E_\theta\left[X^k\right] = m_k(\theta)$. A partir de l'échantillon $(x_1, \ldots x_n)$, on calcule des estimateurs empiriques $\frac{1}{n}\sum_{i=1}^n x_i^k = \hat{m}_k$ et on a alors $\hat{m}_k \approx m_k(\theta^*)$ si θ^* est le vrai paramètre. Le principe de la méthode des moments s'écrit :

1. Trouver les expressions explicites pour les fonctions $m_1, \ldots, m_K$ pour $K \geq 1$
2. Calculer les moments empiriques $\hat{m}_k = \frac{1}{n}\sum_{i=1}^n x_i^k$
3. Chercher $\hat{\theta}$ tel que pour tout k, on ait $\hat{m}_k \approx m_k(\hat{\theta})$ pour $k = 1, \ldots, K$ (ce n'est pas forcément possible d'avoir l'égalité exacte $\hat{m}_k = m_k(\hat{\theta})$ pour tous les moments empiriques).

Ainsi la méthode des moments donne un candidat $\hat{\theta}$ tel que $m_k(\hat{\theta}) \approx m_k(\theta^*)$ pour $k = 1, \ldots, K$. Si les fonctions m_k sont assez discriminantes, on espère alors pouvoir affirmer que $\hat{\theta} \approx \theta^*$. Si les fonctions m_k ont de bonnes propriétés, alors cette approche marche très bien et peut donner des estimateurs de bonne qualité. Une première condition est que le nombre de conditions de moments K soit plus grand que $p = \dim(\Theta) =$ le nombre de paramètre à estimer. En effet, le point 3 est

le point critique de la méthode des moments car on doit résoudre le système (nonlinéaire le plus souvent) : trouver θ dans $\Theta \subset R^p$ t.q.

$$\left\{ \begin{array}{lcl} \hat{m}_1 & = & m_1(\theta) \\ \vdots & = & \vdots \\ \hat{m}_K & = & m_K(\theta) \end{array} \right. \tag{3.1}$$

Si $K < p$, alors le système est sous-déterminé, et il y a une infinité de solutions. Si $K = p$, on peut avoir une unique solution. Si $K > p$ alors le système est sur-déterminé. Cependant, on a aucun intérêt à vouloir résoudre exactement le système (3.1) car les termes de gauche ne sont que des approximations de $m_k(\theta^*)$, qui sont perturbées aléatoirement par les fluctuations d'échantillonnage. Même si on a $E_{\theta^*}[\hat{m_k}] = m_k(\theta^*)$, une représentation pertinente est donnée par $\hat{m_k} = m_k(\theta^*) + \epsilon_k$ où ϵ_k est l'écart aléatoire à la moyenne (dont la loi peut être approchée par un Théorème Central Limite). Donc le système s'écrit en fait

$$\left\{ \begin{array}{lcl} m_1(\theta^*) + \epsilon_1 & = & m_1(\theta) \\ \vdots & = & \vdots \\ m_K(\theta^*) + \epsilon_K & = & m_K(\theta) \end{array} \right. \tag{3.2}$$

Pour simplifier, nous allons supposer temporairement que les fonctions m_k sont linéaires, ce qui veut dire que le système 3.1 est linéaire, et peut s'écrire sous la forme

$$\theta^* + \boldsymbol{\epsilon} = A\theta$$

avec A matrice connue de dimension $K \times p$ (le vecteur des moments empiriques est $\hat{\boldsymbol{m}} = \theta^* + \boldsymbol{\epsilon} \in R^K$). Si on a exactement $K = p$ équations, on pourrait essayer d'inverser A, mais à cause de la perturbation aléatoire (fluctuation d'échantillonnage) $\boldsymbol{\epsilon}$, une erreur va se propager sur la solution du système perturbé. Cette erreur est bien connue en Analyse Numérique, et elle est contrôlée par le conditionnement de la matrice A. Pour essayer de minimiser l'impact de la perturbation aléatoire, on essaie d'avoir $K > p$ et de résoudre le système (3.1) non pas classiquement, mais au sens des moindres carrés. En notation vectorielle ($\hat{\boldsymbol{m}} = (\hat{m}_1 \dots \hat{m}_K)$ et $\boldsymbol{m} : \Theta \longrightarrow R^K$), l'estimateur des moments est calculé en résolvant

$$\hat{\theta}^{MM} = \arg\min_{\theta \in \Theta} \|\hat{\boldsymbol{m}} - \boldsymbol{m}(\theta)\|^2 \tag{3.3}$$

Ainsi $\hat{\theta}^{MM}$ est choisi pour satisfaire à peu près toutes les équations simultanément. Pour calculer l'estimateur des moments, lorsque les fonctions m_k sont assez régulières, on utilise alors des algorithmes d'optimisation : par exemple des descentes de gradient et notamment la méthode de Newton, qui est faite pour résoudre les problèmes de type « moindres carrés ».

Exemple 13. Le cas du modèle $\mathcal{E}(\frac{1}{\theta})$. Nous avons $E[X] = \theta$ et $V(X) = \theta^2$ et $E\left[X^2\right] = 2\theta^2$. Nous pouvons définir plusieurs estimateurs des moments selon le nombre de moments que l'on utilise (1 ou 2 par exemple). Soient $\bar{X}_n$ et $\bar{X_n^2}$ les 2 premiers moments empiriques. Alors on peut chercher :

1. $\min_{\theta>0} \left(\bar{X}_n - \theta\right)^2$ d'où $\hat{\theta}_1^{MM} = \bar{X}_n$
2. $\min_{\theta>0} \left(\bar{X}_n - \theta\right)^2 + \left(\bar{X^2}_n - 2\theta^2\right)^2$ et $\hat{\theta}_2^{MM}$ est obtenu par algorithme de Newton. On pourra prendre pour condition initiale de l'algorithme $\theta^{[0]} = \bar{X}_n$.

Exemple 14. Nous considérons à nouveau l'exemple (2) d'un processus de Poisson composé $N \sim \mathcal{P}(\lambda)$ et $Y_j \sim \exp(\mu)$; la variable d'intérêt (l'ensemble des montants des sinistres annuels) est une somme aléatoire de termes aléatoires $X = \sum_{j=0}^{N} Y_j$. Nous n'avons pas de formule explicite pour la densité de cette variable (qui a un atome en 0, et donc la densité s'écrit relativement à la

mesure de référence $\delta_0 + \lambda$). La fonction génératrice des moments est

$$\begin{aligned} M_X(t) &= E\left[E\left[\exp(tX)|N\right]\right] \\ &= \sum_{k=0}^{\infty} P(N=k) E\left[\exp(tY)\right]^k \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} M_Y(t)^k \\ &= e^{-\lambda} \exp\left(\lambda M_Y(t)\right) \\ &= \exp\left(\lambda\left(M_Y(t) - 1\right)\right) \end{aligned}$$

Nous pouvons calculer $M_X^{'}(t) = \lambda M_Y^{'}(t) M_X(t)$ et $M_X^{''}(t) = \lambda M_Y^{''}(t) M_X(t) + \left(\lambda M_Y^{'}(t)\right)^2 M_X(t)$ et en déduire les moments car $E\left[X^k\right] = \frac{dM_X^{(k)}}{dt}(0)$, pour $k \geq 0$. Ainsi, $E\left[X\right] = \lambda M_Y(0)^{'} = \lambda E\left[Y\right]$ et $E\left[X^2\right] = \lambda M_Y^{''}(0) + \left(\lambda E\left[Y\right]\right)^2 = \lambda E\left[Y^2\right] + E\left[X\right]^2$, ce qui donne une variance $V(X) = \lambda E\left[Y^2\right]$. Nous pouvons aussi calculer le moment d'ordre 3 car on a

$$M_X^{(3)}(t) = \left(\lambda M_Y^{(3)}(t) + 3\lambda^2 M_Y^{''}(t) M_Y^{'}(t) + \lambda^3 M_Y^{'3}(t)\right) M_X(t)$$

ce qui donne

$$\begin{aligned} E\left[X^3\right] &= \lambda M_Y^{(3)}(0) + 3\lambda^2 M_Y^{''}(0) M_Y^{'}(0) + \lambda^3 M_Y^{'3}(0) \\ &= \lambda E\left[Y^3\right] + 3\lambda^2 E\left[Y^2\right] E\left[Y\right] + \left(\lambda E\left[Y\right]\right)^3 \end{aligned}$$

Si on suppose que $Y \sim \mathcal{E}(\frac{1}{\mu})$, on peut éventuellement calculer exactement la fonction génératrice correspondante (car $M_Y(t) = \frac{\mu}{\mu - t}$ pour $t < \mu$) mais surtout facilement les moments de X car $E\left[Y\right] = \mu$,$E\left[Y^2\right] = 2\mu^2$, $E\left[Y^3\right] = 6\mu^3$. On peut construire des estimateur des moments pour le vecteur $\theta = (\lambda\ \mu)^\top$:

$$\hat{\theta}_2^{MM} = \arg\min_{\lambda,\mu>0} \left(\bar{X}_n - \lambda\mu\right)^2 + \left(\bar{X^2}_n - \left(2\lambda\mu^2 + \lambda\mu\right)\right)^2$$

ou encore

$$\hat{\theta}_3^{MM} = \arg\min_{\lambda,\mu>0} \left(\bar{X}_n - \lambda\mu\right)^2 + \left(\bar{X^2}_n - \left(2\lambda\mu^2 + \lambda\mu\right)\right)^2 + \left(\bar{X^3}_n - \left(6\lambda\mu^3 + 6\lambda^2\mu^3 + (\lambda\mu)^3\right)\right)^2$$

3.1.2 Le maximum de vraisemblance

Définition 11 (Estimateur du Maximum de vraisemblance). Soit $\mathcal{M} = \{P_\theta = f_\theta \cdot \mu, \theta \in \Theta\}$ un modèle statistique, et $(X_1, \ldots, X_n)$ un échantillon i.i.d. On appelle Estimateur du Maximum de Vraisemblance $\hat{\theta}^{MV}$ tout estimateur tel que $L(X_1, \ldots, X_n; \hat{\theta}^{MV}) = \max_{\theta\in\Theta} L(X_1, \ldots, X_n; \theta)$, autrement dit

$$\begin{aligned} \hat{\theta}^{MV} &\in \arg\max_{\theta\in\Theta} L(X_1, \ldots, X_n; \theta) \\ &\in \arg\max_{\theta\in\Theta} \mathcal{L}_n(X_1, \ldots, X_n; \theta) \end{aligned}$$

où L est la vraisemblance, et $\mathcal{L}$ est la log-vraisemblance.

On utilise la notation $\arg\max$ pour désigner l'ensemble des valeurs qui réalise le maximum global d'une fonction $\theta \mapsto \mathcal{L}(\theta)$. Si cette fonction a un unique maximum global, alors l'ensemble $\arg\max_{\theta\in\Theta} \mathcal{L}(X_1, \ldots, X_n; \theta)$ est un singleton.

On rappelle que $\mathcal{L}_n(X_1, \ldots, X_n; \theta) = \sum_{i=1}^n \log f(X_i, \theta)$ et que $X_i \sim P_{\theta^*}$ avec θ^* le *vrai paramètre*. Il est facile de voir que maximiser la log-vraisemblance est équivalent à minimiser en θ la différence des log-vraisemblances normalisées

$$\min_\theta \frac{1}{n}\mathcal{L}_n(X_1, \ldots, X_n; \theta^*) - \frac{1}{n}\mathcal{L}_n(X_1, \ldots, X_n; \theta)$$

c'est à dire

$$\hat{\theta}^{MV} = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i,\theta^*)}{f(X_i,\theta)}$$

Comme la vraisemblance dépend des observations, le critère

$$\theta \mapsto \frac{1}{n}\mathcal{L}_n(X_1,\ldots,X_n;\theta^*) - \frac{1}{n}\mathcal{L}_n(X_1,\ldots,X_n;\theta)$$

que l'on doit minimiser est donc une fonction en θ qui est aléatoire et dont nous pouvons calculer l'espérance. On obtient donc que pour tout θ

$$E_{\theta^*}\left[\frac{1}{n}\mathcal{L}(X_1,\ldots,X_n;\theta)\right] = \frac{1}{n}\sum_{i=1}^{n} E_{\theta^*}\log f(X_i,\theta)$$

et donc en moyenne, l'estimateur $\hat{\theta}^{MV}$ minimisera $\frac{1}{n}\sum_{i=1}^{n} E_{\theta^*}\log\frac{f(X_i,\theta^*)}{\log f(X_i,\theta)} = D(P_{\theta^*}\|P_\theta)$. En fait, si n est assez grand et si $E_{\theta^*}\log\frac{f(X,\theta^*)}{f(X,\theta)} < \infty$, alors $\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(X_i,\theta^*)}{f(X_i,\theta)}$ converge presque-sûrement vers $E_{\theta^*}\log\frac{f(X,\theta^*)}{f(X,\theta)}$ pour tout θ dans Θ.

Ainsi si n est assez grand, rechercher le maximum de vraisemblance revient à minimiser la divergence de Kullback-Leibler entre la vraie distribution P_{θ^*} et la distribution de l'échantillon. On montrera plus tard, que l'estimateur $\hat{\theta}^{MV}$ est consistent, c'est-à-dire qu'il converge presque-sûrement vers θ^* (qui plus est de manière optimale).

Note 4. Par le lemme de factorisation, on voit que si T est une statistique exhaustive alors

$$\hat{\theta}^{MV} = \arg\max_{\theta\in\Theta}\sum_{i=1}^{n}\log h(T(X_1,\ldots,X_n);\theta)$$

et donc le maximum de vraisemblance ne dépend de l'échantillon $(x_1,\ldots,x_n)$ qu'à travers les valeurs de la statistique T.

Pour calculer le maximum de vraisemblance, on doit donc résoudre un problème d'optimisation, le plus souvent continu (car souvent $\theta \in R^p$) et on utilise les conditions classiques de premier ordre et de second ordre (pour vérifier que l'on a bien un maximum local et global). Le plus souvent la première étape est le calcul des **équations de vraisemblance**, i.e. la recherche des solutions $\hat{\theta}$

$$\nabla_\theta \log L(x_1,\ldots,x_n;\hat{\theta}) = 0 \tag{3.4}$$

Bien sûr, il s'agit d'une condition nécessaire, et il faut vérifier que la matrice Hessienne est bien définie négative

$$\nabla^2_\theta \log L(x_1,\ldots,x_n;\hat{\theta}) < 0$$

pour s'assurer que l'on a bien un maximum local. Enfin, on vérifie que ce maximum local est bien global.

Dans le cas où l'espace Θ est contraint, on doit résoudre un problème d'optimisation sous contrainte, pour lequel il est courant d'utiliser la relaxation lagrangienne. Parfois, la log-vraisemblance peut être une fonction complexe à optimiser, possédant de nombreux optima locaux. Evidemment, plus le modèle est complexe et plus l'optimisation devient l'enjeu prépondérant de l'inférence statistique. A ce moment là, la bonne question à se poser n'est pas de savoir si la log-vraisemblance est régulière, mais plutôt si le problème d'optimisation est convexe, et si on peut trouver un bon algorithme d'optimisation.

On va calculer l'estimateur du maximum de vraisemblance pour quelques modèles classiques.

Exemple 15. EMV de la moyenne et de la variance pour le modèle gaussien $\mathcal{N}(\mu,\sigma^2)$. La log-vraisemblance est

$$\frac{1}{n}\mathcal{L}_n(\mu,\sigma^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

On cherche $\frac{\partial}{\partial \mu}\frac{1}{n}\mathcal{L}_n = 0$ et $\frac{\partial}{\partial \sigma^2}\frac{1}{n}\mathcal{L}_n = 0$. Les équations de vraisemblance sont

$$\begin{cases} \frac{\partial}{\partial \mu}\frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2 & = 0 \\ \frac{1}{2\sigma^2} - \frac{1}{2}\frac{1}{(\sigma^2)^2}\frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2 & = 0 \end{cases}$$

En résolvant le système, on trouve

$$\begin{cases} \hat{\mu} = & \bar{x}_n \\ \widehat{\sigma^2} = & \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x}_n)^2 \end{cases}$$

On trouve donc que l'estimateur du max de vraisemblance sont les moyennes et variances empiriques (on vérifie les conditions du second ordre...). Dans le cas gaussien multivarié, on tombe sur la même chose mais en version vectoriel / matriciel. On pourra partir des calculs effectués dans le cas centré de l'exemple 10.

Exemple 16. EMV dans le cas Bernoulli $\mathcal{B}(1,\theta)$

La vraisemblance s'écrit $L(x_1,\dots,x_n;\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$, et $\frac{1}{n}\mathcal{L}_n = \bar{x}_n \log\theta + (1-\bar{x}_n)\log(1-\theta)$. L'équation de vraisemblance est

$$\frac{1}{\theta}\bar{x}_n - \frac{1}{1-\theta}(1-\bar{x}_n) = 0$$

soit $\frac{\theta}{1-\theta} = \frac{\bar{x}_n}{1-\bar{x}_n}$ et donc $\hat{\theta} = \bar{x}_n$. On vérifie qu'il s'agit bien d'un maximum global en calculant le hessien en $\theta = \hat{x}_n$.

Exemple 17. Modèle de Cauchy $\mathcal{C}(\theta)$. Ce sont des lois de probabilités sur R dont la densité par rapport à λ s'écrit $f(x,\theta) = \frac{1}{\pi(1+(x-\theta)^2}$. La log-vraisemblance normalisée s'écrit

$$\frac{1}{n}\mathcal{L}_n(X_1,\dots,X_n;\theta) = -\log\pi - \frac{1}{n}\sum_{i=1}^n \log\left(1+(X_i-\theta)^2\right)$$

et l'équation de vraisemblance est

$$\frac{1}{n}\mathcal{L}_n'(\theta) = \frac{1}{n}\sum_{i=1}^n \frac{(X_i-\theta)}{1+(X_i-\theta)^2} = 0.$$

Cette équation n'a pas de solution explicite, ni même de solution unique. Pour trouver une solution, on peut avoir recours à des algorithmes d'optimisation numérique du type méthode du gradient. On construit alors une suite (que l'on espère convergente)

$$\theta^{(k+1)} = \theta^{(k)} + \gamma_k \frac{1}{n}\mathcal{L}_n'(\theta^{(k)})$$

La choix de la condition initiale est important, par exemple on peut prendre la médiane de $(x_1,\dots,x_n)$.

3.1.3 Le calcul des estimateurs

Maximum de Vraisemblance et Optimisation La fonction de log-vraisemblance normalisée dans le cas i.i.d est $\frac{1}{n}\mathcal{L}_n(X_1,\dots,X_n;\theta) = \frac{1}{n}\sum_{i=1}^n \log f(X_i,\theta)$. Le vecteur dérivé est $S_n(X_1,\dots,X_n;\theta) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta \log f(X_i,\theta) \in R^p$: il est souvent appelé **vecteur score** normalisé (et on a $S_1(X,\theta) = S(X,\theta) = \nabla_\theta \log f(X,\theta)$ d'où $S_n(X_1,\dots,X_n;\theta) = \frac{1}{n}\sum_{i=1}^n S(X_i,\theta)$).

Enfin, le Hessien de $\frac{1}{n}\mathcal{L}_n$ est souvent utile pour l'optimisation lorsque $\theta \mapsto \log f(x,\theta)$ est deux fois dérivable par rapport à θ. Par définition, le Hessien est égal à

$$H_n(\theta) = H_n(X_1,\dots,X_n;\theta) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta^2 \log f(X_i,\theta).$$

C'est la matrice symétrique dans $R^{p\times p}$ dont les entrées sont égales à $\frac{1}{n}\sum_{k=1}^{n}\frac{\partial^2}{\partial\theta_i\theta_j}\log f(X_k,\theta)$. Et par définition, on note

$$H_1(X,\theta) = H(X,\theta) = \nabla^2_\theta \log f(X,\theta).$$

Pour calculer le maximum de vraisemblance, on a recours à des algorithmes d'optimisation itératifs qui construisent une suite de paramètres $\theta^{(k)}$, $k \geq 0$, partant d'une condition initiale $\theta^{(0)}$:

1. La méthode du gradient.
$$\theta^{(k+1)} = \theta^{(k)} + \gamma_k S_n(\theta^{(k)})$$
2. La méthode de Newton-Raphson ou Fisher's Scoring. Il s'agit de la méthode de Newton pour trouver annuler le vecteur score et résoudre les équation de vraisemblance. Comme on cherche $\hat{\theta}^{MV}$ t.q. $S_n\left(\hat{\theta}^{MV}\right) = 0$, on part du développement de Taylor d'ordre 1 autour de l'itération $\theta^{(k)}$ pour trouver la bonne direction pour se déplacer : $S_n\left(\hat{\theta}^{MV}\right) = S_n\left(\theta^{(k)}\right) + H_n(\theta^{(k)})\left(\hat{\theta}^{MV} - \theta^{(k)}\right)$. L'itération s'écrit alors
$$\theta^{(k+1)} = \theta^{(k)} + \left[H_n(\theta^{(k)})\right]^{-1} S_n(\theta^{(k)})$$
3. Le gradient stochastique. Lorsque l'on a trop de de données, ou que le modèle est trop complexe (temps calcul des dérivées long, maxima multiples, grand espace de paramètres,...), on sélectionne aléatoirement un sous-ensemble de données (1 seule par exemple) et on fait
$$\theta^{(k+1)} = \theta^{(k)} + \gamma_k S_1(X_{\mathbf{i_k}}, \theta^{(k)})$$
ou $\mathbf{i_k}$ est tiré uniformément dans $\{1,\ldots,n\}$, et la suite de pas $(\gamma_k)_{k\geq 0}$ est telle que $\sum_{k=1}^{\infty}\gamma_k = +\infty$ et $\sum_{k=1}^{\infty}\gamma_k^2 < +\infty$.
4. L'algorithme EM.

On a deux propriétés remarquables à propos du score et du Hessien de la log-vraisemblance.

Proposition 3. *Propriétés du Score*

Pour tout $\theta \in \Theta$,

$$E_\theta\left[S\left(X,\theta\right)\right] = 0$$

De plus, la variance du score satisfait

$$V_\theta\left(S(X,\theta)\right) = E_\theta\left[S(X,\theta)S(X,\theta)^\top\right] = -E_\theta\left[H(X,\theta)\right]$$

Cela veut dire qu'en moyenne, le score est nul pour le bon paramètre. Dans ce cas, la variance du score (matrice définie positive) est égale l'opposé de la Hessienne de la log-vraisemblance.

3.1.4 La famille exponentielle (naturelle et générale)

De nombreuses lois couramment utilisées en pratique ont des vraisemblances qui ont une structure mathématique similaire :

Gaussienne

Gamma

Beta

Dirichlet

Bernoulli

Poisson

Wishart

Géométrique

Multinomial avec un nombre d'essais/tirages fixe.

Binomiale Négative avec un nombre d'essais/tirages fixe.

Tous ces modèles statistiques peuvent être rassemblées dans une méta-famille plus grande, la famille exponentielle qui possède des propriétés remarquables et pour lequel l'estimateur du maximum de vraisemblance se calcule plus simplement.

Nous avons vu en effet la forme générale d'une densité par rapport à une mesure de référence μ est $f(x,\theta) = \frac{1}{C(\theta)} \exp\left(h(x,\theta)\right) 1_{A_\theta}(x)$. On s'intéresse au cas où le support A_θ ne dépend pas de θ (et donc toutes les probabilités sont absolument continues l'une par rapport à l'autre), et où la fonction $h(x,\theta) = \langle T(x), B(\theta)\rangle$ a une structure linéaire (bilinéaire plutôt) où $x \mapsto T(x)$ et $\theta \mapsto B(\theta)$ sont des fonctions à valeurs dans R^k et $\langle\cdot,\cdot\rangle$ dénote un produit scalaire dans un espace euclidien bien adapté.

Nous commençons d'abord par définir la

Définition 12 (**Famille Exponentielle Canonique**)**.**

Soit $\mathcal{X}$ l'espace des observations (par exemple R^q ou $\{0,1,\ldots,N\}$) et μ une mesure de référence σ-finie. On appelle famille exponentielle canonique de dimension p engendrée par $\mathbf{T} = (T_1 \ldots T_p)$ et la fonction $h : \mathcal{X} \longrightarrow R^+$ la famille de densités par rapport à μ indexées par $\eta = (\eta_1 \ldots \eta_p)$ t.q

$$f(x,\eta) = \exp\left(\langle \eta, T(x)\rangle - A(\eta)\right) h(x)$$

On a par définition $A(\eta) = \log \int_{\mathcal{X}} \exp\left(\langle \eta, T(x)\rangle\right) h(x)\mu(dx)$, autrement dit $\exp\left(-A(\eta)\right)$ est la constante de normalisation. Et on note $\Xi = \left\{\eta \in R^d \middle| |A(\eta)| < \infty\right\}$.

Exemple 18. Si on prend $\mathcal{X} = R$, $h = 1_{R^+}$, $p = 1$ et $T(x) = x$, on a alors $A(\eta) = \log \int_0^\infty \exp(\eta x)dx = \begin{cases} \log(-1/\eta), & \eta < 0 \\ \infty, & \eta \geq 0 \end{cases}$. Donc nous avons $p(x,\eta) = \exp\left(\eta x - \log(-1/\eta)\right) h(x)$. C'est le modèle exponentiel standard, mais avec la paramétrisation canonique de la loi exponentielle.

Exemple 19. On considère le modèle $\mathcal{X} = R^q$, $\mathcal{N}\left(\eta, I_q\right)$ avec

$$\begin{aligned} f(x,\eta) &= (2\pi)^{-q/2} \exp\left(-\frac{1}{2}(x-\eta)^\top (x-\eta)\right) \\ &= (2\pi)^{-q/2} \exp\left(-\frac{1}{2}xx^\top\right) \exp\left(\eta^\top x - \frac{1}{2}\eta^\top\eta\right) \end{aligned}$$

on a $h(x) = (2\pi)^{-q/2} \exp\left(-\frac{1}{2}xx^\top\right)$, $A(\eta) = \frac{1}{2}\left\|\eta\right\|^2$ et $\langle\cdot,\cdot\rangle$ est le produit scalaire standard de R^q.

On peut aussi considérer le cas du modèle gaussien $\mathcal{N}(0,\eta^{-1})$, i.e le paramètre naturel est l'inverse de la matrice de variance $V_\eta(X) = \Sigma$, voir l'exemple 10. Dans ce cas, on a

$$f(x,\eta) = (2\pi)^{-q/2} \det(\eta)^{1/2} \exp\left(-\frac{1}{2}\mathrm{Tr}\left(\eta^\top T(x)\right)\right)$$

avec $T(x) = xx^\top$, et $\langle A, B\rangle = Tr(A^\top B)$ est le produit scalaire entre matrice carrée. Le **paramètre naturel** $\eta = \Sigma^{-1}$, qui appartient à l'espace des matrices définies positives, est appelé couramment **matrice de précision**.

Une manière de construire (ou générer) des modèles statistiques peut être de prendre des fonctions $x \mapsto T(x)$ éventuellement arbitraire (par exemple $T(x) = \frac{1}{2}x^\top Hx + Bx + c$ ou $T(x) = (x,\, \log(x))$) définie sur l'espace des observations, ainsi qu'une fonction $h(x)$. Il faut alors identifier l'ensemble des valeurs η pour lesquels la densité $f(x,\eta) = \exp\left(\langle \eta, T(x)\rangle - A(\eta)\right) h(x)$.

Cependant, le plus souvent, on essaie plutôt d'identifier si un modèle statistique donné est une famille exponentielle. En effet, dans ce cas-là, un certain nombre de propriétés intéressantes sont satisfaites.

Proposition 4. *Soit $\{f(x,\eta),\, \eta \in \Xi\}$ une famille exponentielle canonique, de dimension p, engendré par $\mathbf{T} = (T_1,\ldots,T_p)$ et la fonction h alors*

1. *L'espace des paramètres naturel Ξ est convexe,*
2. *La fonction $A : \Xi \longrightarrow R$ est convexe,*
3. *Si le modèle est de rang complet (i.e la fonction $(T_1,\ldots,T_p)$ est t.q. les T_i sont linéairement indépendants), alors la fonction A est strictement convexe,*

4. *Pour tout* $i = 1, \dots, p$, $E_\eta [T_i(X)] = \frac{\partial}{\partial \eta_i} A(\eta)$,
5. *Si le modèle est de rang complet, alors la fonction* $\eta \mapsto \nabla_\eta A(\eta)$ *est un difféomorphisme de* $\mathring{\Xi}$ *sur* $\nabla_\eta \mathring{A}(\Xi)$,
6. *Pour tout* i, j, $cov_\eta (T_i(X), T_j(X)) = \frac{\partial^2}{\partial \eta_i \partial \eta_j} A(\eta)$,
7. *La statistique* $\mathbf{T}(X) = (T_1(X), \dots, T_p(X))$ *est une statistique exhaustive minimale.*

De nombreuses propriétés du modèle sont donc déduit de la fonction $\eta \mapsto A(\eta)$. La log-vraisemblance est

$$\frac{1}{n}\mathcal{L}_n(X_1, \dots, X_n; \eta) = \left\langle \eta, \frac{1}{n}\sum_{i=1}^n \mathbf{T}(X_i) \right\rangle - A(\eta) + \sum_{i=1}^n h(X_i) \tag{3.5}$$

et donc l'équation de vraisemblance s'écrit

$$\frac{1}{n}\sum_{i=1}^n \mathbf{T}(X_i) = \nabla_\eta A(\eta). \tag{3.6}$$

Si le modèle est de rang complet, l'application $\eta \mapsto \nabla_\eta A(\eta)$ est un difféomorphisme de $\mathring{\Xi}$ sur $\nabla_\eta \mathring{A}(\Xi)$. On peut vérifier alors que la probabilité que $\frac{1}{n}\sum_{i=1}^n \mathbf{T}(X_i)$ soit sur la frontière de Ξ est nulle. Donc, presque-sûrement, on peut toujours inverser la relation 3.6, et on trouve donc

$$\hat{\eta}^{MV} = \nabla_\eta A^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{T}(X_i)\right)$$

existe et est unique. Enfin, comme la fonction $A(\eta)$ est strictement convexe, on sait que $\eta \mapsto \left\langle \eta, \frac{1}{n}\sum_{i=1}^n \mathbf{T}(X_i)\right\rangle - A(\eta)$ a un unique maximum global. Dans le cas d'une famille exponentielle canonique, on vient donc de voir que le maximum de vraisemblance existe toujours et est unique. Enfin celui-ci vérifie que lorsque $n \longrightarrow \infty$, $\frac{1}{n}\sum_{i=1}^n \mathbf{T}(X_i) \longrightarrow E_{\eta^*}[\mathbf{T}(X)]$, d'après la loi forte des grands nombres (si les $X_i \sim P_{\eta^*}$). Or $E_{\eta^*}[\mathbf{T}(X)] = \nabla_\eta A(\eta^*)$ et on rappelle que $\eta \mapsto \nabla_\eta A^{-1}$ est un difféomorphisme donc que

$$\nabla_\eta A^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{T}(X_i)\right) \longrightarrow \nabla_\eta A^{-1}(E_{\eta^*}[\mathbf{T}(X)])$$

presque-sûrement. On vient donc de prouver que $\hat{\eta}^{MV} \longrightarrow \eta^*$ presque-sûrement (et en probabilités) et donc que l'estimateur du maximum de vraisemblance est asymptotiquement consistent. Enfin on peut montrer que la variance de cet estimateur est $V_{\eta^*}(\hat{\eta}^{MV}) = \nabla^2_\eta A(\eta^*)$ et qu'il est asymptotiquement normal (gaussien). Ceci fait que l'on connaît donc très bien le comportement du maximum de vraisemblance dans ce cas, et que les résultats obtenus sont optimaux (voir la prochaine section 3.2).

Cependant, parmi les modèles cités précédemment, peu sont exprimés directement en fonction du paramètre naturel η. Ils sont plutôt paramétrés via $\theta \in \Theta$ de la manière suivante

$$f(x, \theta) = \exp\left(\langle \varphi(\theta), T(x)\rangle - B(\theta)\right) h(x)$$

et on a en fait une transformation $\varphi : \theta \mapsto \varphi(\theta)$ de Θ dans Ξ. En effet, dans l'exemple de la loi normal vectoriel 19, le paramètre est en fait η^{-1}.

Exemple 20. De même pour le modèle de Bernoulli $\mathcal{B}(1, \theta)$ que l'on peut mettre sous forme canonique. On a

$$\begin{aligned} f(x, \theta) &= \left(\frac{\theta}{1-\theta}\right)^x (1-\theta) \\ &= \exp\left(x \log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right) \end{aligned}$$

On pose alors $\varphi : \theta \mapsto \eta = \log\left(\frac{\theta}{1-\theta}\right)$ et la fonction s'écrit $f(x, \theta) = \exp(x\varphi(\theta) - B(\theta))$. La forme canonique correspondante est $f(x, \eta) = \exp(x\eta - A(\eta))$ avec $A(\eta) = -\log(1-\theta) = \log(1+e^\eta)$. Et on déduit par exemple que $E_\eta[X] = \frac{d}{d\eta}\log(1+e^\eta) = \frac{e^\eta}{1+e^\eta} = \theta$.

3.2 Propriétés des estimateurs

Pour beaucoup de modèles assez simples ou standard (qui sont des familles exponentielles), on vient de voir que le maximum de vraisemblance était un estimateur convergent lorsque n tend vers l'infini (pour le moins il reste proche du vrai paramètre θ^*). Nous avions vu en début de chapitre qu'il importait surtout de pouvoir comparer les estimateurs et de s'intéresser particulièrement au risque quadratique $E_{\theta^*}\left[\left(\hat{\theta}^{MV} - \theta^*\right)^2\right]$. Pour cette raison, nous nous intéressons essentiellement au biais et à la variance de l'estimateur.

3.2.1 Estimation sans biais et Information de Fisher

Définition 13 (**Modèle Régulier**)**.** Une modèle statistique est un modèle régulier si

1. c'est un modèle paramétrique $\mathcal{M} = \{f(x,\theta)\cdot\mu|\theta\in\Theta\}$, dominé par une mesure σ-finie μ. On a $\mathcal{L}(x,\theta) = \log f(x,\theta)$ la log-vraisemblance d'une seule observation.
2. La fonction $\theta \mapsto \mathcal{L}(x,\theta)$ est C^2 sur Θ pour μ-presque tout x dans $\mathcal{X}$.
3. Pour tout θ dans Θ, et il existe un voisinage de θ, pour lequel on peut trouver des fonctions intégrables $h, g > 0$, tel que h domine uniformément le jacobien $\nabla_\theta\mathcal{L}(x,\theta)$ et le hessien $\nabla^2_\theta\mathcal{L}(x,\theta)$ et g domine uniformément (en θ) $f(x,\theta)$.
4. Pour tout $\theta\in\Theta$, la matrice $E_\theta\left[\nabla^2_\theta\mathcal{L}(x,\theta)\right]$ est inversible.

Essentiellement un **modèle régulier** est un modèle pour lequel on peut dériver sous le signe intégral et telle que $E_\theta\left[\nabla^2_\theta\mathcal{L}(x,\theta)\right]$ est inversible.

Définition 14 (**Information de Fisher**)**.** Soit $\mathcal{M}$ un modèle statistique régulier. La matrice

$$I(\theta) = E_\theta\left[\nabla_\theta\mathcal{L}(X,\theta)\nabla_\theta\mathcal{L}(X,\theta)^\top\right] \tag{3.7}$$

est bien définie et ne dépend pas de la mesure de référence μ. La matrice $I(\theta)$ est la matrice d"information de Fisher.

Proposition 5. *Soit $\mathcal{M}$ modèle régulier. La matrice d'information de Fisher admet l'expression alternative suivante*

$$I(\theta) = -E_\theta\left[\nabla^2_\theta\mathcal{L}(X,\theta)\right]$$

Si $(X_1, X_2, \ldots, X_n)$ est un échantillon, alors la matrice d'information de Fisher est

$$\begin{aligned} I_n(\theta) &= E_\theta\left[\nabla_\theta\mathcal{L}_n(X_1,\ldots,X_n;\theta)\nabla_\theta\mathcal{L}_n(X_1,\ldots,X_n;\theta)^\top\right] \\ &= nI_1(\theta) = nI(\theta) \end{aligned}$$

Enfin, on parle de d'estimateur de θ régulier, si $T(X)$ est tel que $T(X)$ est de carré intégrable et si $\theta\mapsto E_\theta[T(X)]$ est différentiable. Le résultat le plus important de cette section est l'inégalité de Cramér-Rao, qui affirme qu'il y a une limite inférieure à la précision que l'on peut espérer avoir pour un estimateur $T(X)$. Cette barrière indépassable est appelée l'inégalité d'information, ou borne de Cramér-Rao

Théorème - Borne de Cramér-Rao. *Soit $\mathcal{M}$ un modèle régulier, et T un estimateur sans biais et régulier de θ, alors*

$$V_\theta\left(T(X)\right) \geq I(\theta)^{-1}$$

Définition 15. Soit $\mathcal{M}$ un modèle statistique régulier. Un estimateur sans biais T de θ est dit **efficace** si sa variance atteint la borne de Cramér-Rao, i.e. si

$$\mathrm{Var}_\theta\left(T(X)\right) = I(\theta)^{-1}$$

La construction d'estimateur sans biais, est une tache assez difficile. En effet, la plupart des estimateurs $T_n(X_1,\ldots,X_n)$ que l'on construit sont tels que le biais $B_n(\theta) = E_\theta\left[T_n(X_1,\ldots,X_n)\right] - \theta$ est lui-même fonction de θ, et qu'il n'est pas possible de débiaiser l'estimateur. Cependant, le biais de l'estimateur peut tendre vers 0, lorsque la taille de l'échantillon croît : si l'estimateur T_n est tel

que $B_n(\theta) \longrightarrow 0$, si $n \longrightarrow \infty$ on dira que c'est un estimateur **asymptotiquement sans biais**. De la même façon, construire une estimateur (sans biais) qui est efficace est délicat, et souvent restreint à des cas de modèles simples pour n fixé. Par contre, si la variance est tel que

$$\mathrm{Var}_\theta\left(T_n(X_1,\ldots,X_n)\right) \longrightarrow \left[nI_1(\theta)\right]^{-1}$$

lorsque $n \longrightarrow \infty$, on dira que l'estimateur T_n est **asymptotiquement efficace**.

3.2.2 Efficacité asymptotique de l'Estimateur du Maximum de Vraisemblance

Définition 16. De manière remarquable, on peut montrer que sous des conditions standards, l'Estimateur du Maximum de Vraisemblance $\hat{\theta}^{MV}$ est asymptotiquement sans biais et efficace, ce qui fait de lui un estimateur de référence car il est optimal au sens de risque quadratique.

Théorème - Consistance de l'Estimateur du Maximum de Vraisemblance. *Soit $\mathcal{M} = \{P_\theta = f(x;\theta)\cdot\mu | \theta \in \Theta\}$ un modèle statistique dominée, et on suppose que l'espace des paramètres $\Theta \subset R^d$ est compact. Supposons que*

1. *$\forall x \in \mathcal{X}$, la fonction $\theta \mapsto \log f(x;\theta)$ est continue (régularité)*
2. *Pour tout $\theta_0 \in \Theta$, $E_{\theta_0}\left[\sup_{\theta\in\Theta}|\log f(X;\theta)|\right] < \infty$ (non-explosion du max)*
3. *$\forall \theta \neq \theta^{'}$, $P_\theta \neq P_{\theta'}$ (identifiabilité du modèle statistique)*

Alors la fonction $\theta \mapsto E_{\theta^*}\left[\log f(X;\theta)\right]$ a un maximum unique au point $\theta = \theta^*$. De plus, si pour tout $\theta^* \in \Theta$, l'estimateur $\hat{\theta}_n^{MV}$ vérifie

$$\lim_{n\to\infty} P_{\theta^*}\left(\mathcal{L}_n(\hat{\theta}_n^{MV}) \geq \mathcal{L}_n(\theta^*)\right) = 1$$

alors $\hat{\theta}_n^{MV}$ converge en probabilités vers θ^* ($P_{\theta^*}\left(\left\|\hat{\theta}_n^{MV} - \theta^*\right\| \geq \epsilon\right) \longrightarrow 0$ pour tout ϵ.

Enfin, on connaît la loi asymptotique de l'estimateur du Maximum de Vraisemblance,

Théorème - Normalité asymptotique du Maximum de Vraisemblance. *Supposons que le modèle $\mathcal{M} = \{P_\theta = f(x;\theta)\cdot\mu | \theta \in \Theta\}$ est un modèle régulier, et soit $S(x,\theta)$ le vecteur score associée. Soit $\hat{\theta}_n^{MV}$ l'estimateur du maximum de vraisemblance, vérifiant*

$$\sqrt{n}\nabla_\theta\left(\frac{1}{n}\mathcal{L}_n(X_1,\ldots,\theta)\right) \overset{P_{\theta^*}-proba}{\longrightarrow} 0$$

alors pour tout θ^* dans Θ, on a

$$\sqrt{n}\left(\hat{\theta}_n^{MV} - \theta^*\right) \overset{P_{\theta^*}-Loi}{\rightsquigarrow} \mathcal{N}\left(0, I_1^{-1}(\theta^*)\right)$$

3.2.3 Rappels sur les convergences stochastiques

4 Régions de confiance

Principe de construction des régions de confiance :

1. Non-asymptotique
2. Asymptotique

Troisième partie
Tests statistiques

5 Théorie des tests

5.1 Règle de décision

Règle de décision.
Le risque 0-1.
Les différentes erreurs.

5.2 Rapport de Vraisemblance

5.2.1 Lemme Neyman-Pearson

5.2.2 Rapport du Maximum de Vraisemblance

5.3 Comparaison d'échantillons

6 Tests d'adéquation

6.1 Kolmogorov-Smirnov

6.2 Test du Khi-2