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 3. 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 intégrable relativement à chaque argument.

— 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 » 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 4. 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 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 5. 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 6. 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 7. Pour la mesure de Lebesgue, on $\lambda(\{x\}) = 0$ et pour tout ensemble dénombrable $\lambda(\{x_1, x_2, \ldots, x_n\}) = 0$. 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 8. 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 9. 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 10. 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 $Y_{j,k} \sim \exp(\lambda_k)$; on a $X_k = \sum_{j=0}^{N_k} Y_{j,k}$. 0 est une atome de la distribution de Y_k.

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

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

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

Définition 13. 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 14. 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 15. 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 16. 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 17. 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 18. *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 19. *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. $E\left[|X - m|\right] = \int_{-\infty}^{m}(x - m)f(x)dx - \int_{m}^{\infty}(x - m)f(x)dx$. Pour chercher le minimum, on dérive par rapport à m. On peut aussi écrire

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

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*

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

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

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

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.

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 20. 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 distribuée.

Exemple 21. 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 $X_1,\dots,X_n$ (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

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

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

Définition 22. Pour tout $(x_1,x_2,\dots x_n)\in\mathcal{X}^n$, on définit la **vraisemblance** comme la fonction définie sur Θ à valeurs dans R^+ :

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

Exemple 23. Si $(x_1,\dots,x_5) = (-0.14,-0.07,0.19,-1.94,2.43)$ 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

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

La vraisemblance $L(x_1,\dots,x_n,\theta)$ 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** $S(X_1,\dots,X_n)$ 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** $\mathcal{L}(X_1,\dots,X_n,\theta) = \log L(X_1,\dots,X_n,\theta)$ 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 $\mathcal{L}(X_1,\dots,X_n,\theta) = -\infty$ 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 $(X_1,\dots,X_n)$ 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 $\mathcal{L}(x_1,\dots,x_n,\theta) > \mathcal{L}(x_1,\dots,x_n,\theta')$, 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

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

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 [1]. Par conséquent,

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

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

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 un 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.

1. 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.

Exemple 25. On suppose que $(X_1, \dots, X_n) \sim \mathcal{E}(\lambda)$. Si on considére $m_n = \min(X_1, \dots, X_n)$ et $\bar{X}_n$, intuitivement ses informations ne convoit pas la même quantité d'information sur le paramètre d'intérêt λ. Est ce

$P(m_n \leq x) = (1 - (1 - F_\lambda(x))^n) = 1 - \exp(-n\lambda x)$.

Une formalisation possible de la concept de « résumé sans perte » est donnée par la notion de statistique exhaustive.

Définition 26. 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 27. 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

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

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

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

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

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

Donc $\frac{1}{n} \sum_{i=1}^n x_i x_i^\top$ est une statistique exhaustive d'après le lemme de factorisation avec $g(t, \Sigma) = \det(\Sigma)^{-n/2} \exp(-\frac{n}{2} \mathrm{Tr}\left(\Sigma T\right))$ 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 :

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

et donc de calculer la valeur d'une stat. exhaustive t sur l'échantillon et de calculer $\frac{g(t,\theta_1)}{g(t,\theta_2)}$.

Exemple 28. 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

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

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 $T(x_1,\ldots,x_n) = (x_1,\ldots,x_n)$, 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 29. Statistique Libre en loi

Soit $S : (x_1\ldots,x_n) \longrightarrow R^d$ 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 $(X_1,\ldots,X_n)$ 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

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

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