EXAMEN STATISTIQUE - 1SN

Lundi 15 janvier 2025 (10h-11h30)

Partiel sans document (Une feuille A4 recto-verso autorisée)

Exercice 1 : Estimation (10 points)

On considère n observations $x_1, ..., x_n$ issues d'un vecteur $(X_1, ..., X_n)$ de n variables aléatoires indépendantes de mêmes lois de densités

$$f(x_i;\theta) = \frac{2x_i}{\theta}\exp\left(-\frac{x_i^2}{\theta}\right) I_{\mathbb{R}^+}(x_i),$$

où $I_{\mathbb{R}^+}$ est la fonction indicatrice sur $\mathbb{R}^+$ ($I_{\mathbb{R}^+}(x) = 1$ si $x > 0$ et 0 sinon) et où θ est un paramètre inconnu que l'on cherche à estimer.

1. (1pt) Montrer que l'estimateur du maximum de vraisemblance du paramètre θ est

$$\widehat{\theta}_{\text{MV}} = \frac{1}{n}\sum_{k=1}^{n} X_k^2.$$

On justifiera que la vraisemblance admet un maximum en ce point.
La vraisemblance de l'échantillon s'écrit

$$L(x_1, ..., x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta) \propto \left(\frac{1}{\theta}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{x_i^2}{\theta}\right) I_{\mathbb{R}^+}(x_i),$$

où $\mathbb{I}_{\mathbb{R}^+}$ est la fonction indicatrice sur l'intervalle $\mathbb{R}^+$ et $\propto$ signifie "proportionnel à", la constante de proportionnalité regroupant tous les termes indépendants de θ. On sait qu'il est plus facile de travailler avec la log-vraisemblance définie par

$$\ln L(x_1, ..., x_n;\theta) = -n\ln\theta - \frac{1}{\theta}\sum_{i=1}^{n} x_i^2.$$

Cette log-vraisemblance admet pour dérivée

$$\frac{\partial \ln L(x_1, ..., x_n;\theta)}{\partial\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i^2.$$

Les variations de cette dérivée sont définies par

$$\begin{aligned}\frac{\partial \ln L(x_1, ..., x_n;\theta)}{\partial\theta} \geq 0 \quad &\Leftrightarrow \quad -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i^2 \geq 0 \\ &\Leftrightarrow \quad \theta \leq \frac{1}{n}\sum_{i=1}^{n} x_i^2.\end{aligned}$$

Ceci permet de faire un tableau de variation qui indique que $\theta = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ est le maximum global unique de la vraisemblance, d'où

$$\widehat{\theta}_{\text{MV}} = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

2. (2pt) Déterminer la loi de $Y_k = X_k^2$ et en déduire en utilisant les tables que $E[Y_k] = \theta$ et $\text{var}(Y_k) = \theta^2$.
Le changement de variables $Y_k = X_k^2$ est bijectif de $\mathbb{R}^+$ dans $\mathbb{R}^+$. La densité de Y_k s'obtient avec la formule du changement de variables

$$\begin{aligned}\pi(y_k) =& \frac{2\sqrt{y_i}}{\theta} \exp\left(-\frac{y_i}{\theta}\right) \left|\frac{dx_i}{dy_i}\right| \\ =& \frac{1}{\theta} \exp\left(-\frac{y_i}{\theta}\right) \mathbb{I}_{\mathbb{R}^+}(y_i)\end{aligned}$$

qui est la densité d'une loi gamma $\mathcal{G}\left(1, \frac{1}{\theta}\right)$. On en déduit $E[Y_k] = \theta$ et $\text{var}(Y_k) = \theta^2$.

3. (1pt) L'estimateur $\widehat{\theta}_{\text{MV}}$ est-il un estimateur non biaisé et convergent du paramètre θ ?
Cette question est très classique. La moyenne et la variance de l'estimateur sont

$$E\left[\widehat{\theta}_{\text{MV}}\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \theta.$$

et

$$\text{var}\left[\widehat{\theta}_{\text{MV}}\right] = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}[X_i] = \frac{\theta^2}{n}.$$

Comme $\widehat{\theta}_{\text{MV}}$ est un estimateur non biaisé et que sa variance tend vers 0 lorsque $n \to \infty$, $\widehat{\theta}_{\text{MV}}$ est un estimateur convergent de θ.

4. (2pts) Déterminer la borne de Cramér-Rao d'un estimateur non biaisé de θ. L'estimateur $\widehat{\theta}_{\text{MV}}$ est-il l'estimateur efficace du paramètre θ ?
La dérivée seconde de la log-vraisemblance par rapport à θ est

$$\frac{\partial^2 \ln L(x_1, ..., x_n; \theta)}{\partial \theta^2} = \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n} x_i^2.$$

d'où

$$E\left[-\frac{\partial^2 \ln L(X_1, ..., X_n; \theta)}{\partial \theta^2}\right] = \frac{n}{\theta^2} - \frac{2n}{\theta^3} \times \theta = \frac{n}{\theta^2}.$$

On en déduit que la borne de Cramér-Rao pour un estimateur non-biaisé de θ est

$$\text{BCR} = \frac{-1}{E\left[\frac{\partial^2 \ln L(X_1,...,X_n;\theta)}{\partial \theta^2}\right]} = \frac{\theta^2}{n}.$$

Comme $\text{var}\left[\widehat{\theta}_{\text{MV}}\right] = \text{BCR}$ et que l'estimateur $\widehat{\theta}_{\text{MV}}$ est non biaisé, l'estimateur $\widehat{\theta}_{\text{MV}}$ est l'estimateur efficace du paramètre θ.

5. (1pt) Déterminer $E[X_k]$ et en déduire un estimateur des moments du paramètre θ.
On a

$$E[X_k] = \int_0^{+\infty} \frac{2u^2}{\theta} \exp\left(-\frac{u^2}{\theta}\right) du$$

qui se calcule par exemple avec une intégration par parties. Si on pose $v' = -2\frac{u}{\theta}\exp\left(-\frac{u^2}{\theta}\right)$, on a $v = \exp\left(-\frac{u^2}{\theta}\right)$ et alors $w = -u$, d'où

$$E[X_k] = \int_0^{+\infty} \exp\left(-\frac{u^2}{\theta}\right) du.$$

On reconnait la densité d'une loi normale de variance $2\sigma^2 = \theta$ telle que

$$\int_0^{+\infty} \exp\left(-\frac{u^2}{\theta}\right) du = \frac{\sqrt{2\pi\sigma^2}}{2}.$$

On obtient finalement

$$E[X_k] = \frac{\sqrt{\theta\pi}}{2}.$$

On en déduit

$$\theta = \frac{(2E[X_k])^2}{\pi},$$

et donc un estimateur des moments de θ est

$$\widehat{\theta}_{\text{Mo}} = \frac{4}{\pi}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

6. (3pts) On suppose désormais que le paramètre θ est muni d'une loi a priori de densité

$$f(\theta) = \frac{a}{\theta^2} e^{-\frac{a}{\theta}} I_{\mathbb{R}^+}(\theta) I_{\mathbb{R}^+}(\theta).$$

avec $a > 0$ et où $I_{\mathbb{R}^+}$ est la fonction indicatrice sur $\mathbb{R}^+$ ($I_{\mathbb{R}^+}(\theta) = 1$ si $\theta > 0$ et 0 sinon) Déterminer l'estimateur du maximum a posteriori du paramètre θ noté $\widehat{\theta}_{\text{MAP}}$. Montrer que la loi a posteriori du paramètre θ est une loi inverse gamma dont on déterminera les paramètres. En déduire l'estimateur MMSE de θ noté $\widehat{\theta}_{\text{MMSE}}$.
On a vu que la densité de l'échantillon s'écrit comme suit

$$p(x_1, ..., x_n; \lambda) = \prod_{i=1}^{n} f(x_i; \theta) \propto \left(\frac{1}{\theta}\right)^n \exp\left(-\frac{\sum_{i=1}^n x_i^2}{\theta}\right) I_{\mathbb{R}^+}(x_i).$$

La loi a posteriori de $\theta|X_1, ..., X_n$ vérifie donc

$$p(\theta|x_1, ..., x_n) \propto p(x_1, ..., x_n|\theta) f(\theta) \propto \left(\frac{1}{\theta}\right)^{n+2} \exp\left(-\frac{a + \sum_{i=1}^n x_i^2}{\theta}\right)$$

Cette loi est obtenue en remplaçant n par $n+2$ et $\sum_{i=1}^n x_i^2$ par $a + \sum_{i=1}^n x_i^2$ dans la vraisemblance. On en déduit

$$\widehat{\theta}_{\text{MAP}} = \frac{1}{n+2}\left(\sum_{i=1}^{n} X_i^2 + a\right).$$

Pour déterminer l'estimateur MMSE de θ, il faut calculer la moyenne de la loi a posteriori de $\theta|X_1, ..., X_n$. L'expression de $p(\theta|x_1, ..., x_n)$ donnée ci-dessus est la densité d'une loi inverse gamma $\mathcal{IG}\left(n+1, a + \sum_{i=1}^n X_i^2\right)$ dont la moyenne est dans la table de lois

$$\widehat{\theta}_{\text{MMSE}} = E[\theta|X_1, ..., X_n] = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 + a\right).$$

Exercice 2 : Tests Statistiques (10 **points)**

On considère n observations $x_1, ..., x_n$ issues d'un vecteur $(X_1, ..., X_n)$ de n variables aléatoires indépendantes de mêmes lois de densités

$$f(x_i;\theta) = \frac{2x_i}{\theta} \exp\left(-\frac{x_i^2}{\theta}\right) I_{\mathbb{R}^+}(x_i),$$

où $I_{\mathbb{R}^+}$ est la fonction indicatrice sur $\mathbb{R}^+$ ($I_{\mathbb{R}^+}(x) = 1$ si $x > 0$ et 0 sinon). On désire utiliser les observations $x_1, ..., x_n$ pour déterminer si $\theta = \theta_0 > 0$ ou si $\theta = \theta_1 < \theta_0$. On considère donc le test d'hypothèses

$$H_0 : \theta = \theta_0, \quad H_1 : \theta = \theta_1 \qquad \text{avec } \theta_1 < \theta_0.$$

1. (2pt) Déterminer la statistique T_n du test de Neyman Pearson et la région critique associée (sans chercher pour l'instant à déterminer le seuil associé à cette région noté S_α).
   Le test de Neyman Pearson est défini par

   $$\text{Rejet de } H_0 \text{ si } \frac{L(x_1, ..., x_n; \theta_1)}{L(x_1, ..., x_n; \theta_0)} > S_{1,\alpha}$$

   où $S_{1,\alpha}$ est un seuil dépendant du risque de première espèce α. Mais

   $$\frac{L(x_1, ..., x_n; \theta_1)}{L(x_1, ..., x_n; \theta_0)} > S_{1,\alpha} \Leftrightarrow \ln \frac{L(x_1, ..., x_n; \theta_1)}{L(x_1, ..., x_n; \theta_0)} > S_{2,\alpha}$$
   $$\Leftrightarrow \left(\frac{1}{\theta_0} - \frac{1}{\theta_1}\right) \sum_{i=1}^{n} x_i^2 > S_{3,\alpha}.$$

   Comme $\theta_1 < \theta_0$, on a $\frac{1}{\theta_1} > \frac{1}{\theta_0}$. La règle de décision est donc finalement

   $$\text{Rejet de } H_0 \text{ si } T_n = \sum_{i=1}^{n} X_i^2 < S_\alpha.$$

   La région critique du test est l'ensemble des vecteurs $(x_1, ..., x_n) \in]0, +\infty[^n$ tels que $T_n < S_\alpha$ et la statistique de test est $T_n = \sum_{i=1}^n X_i^2$.

2. (1pt) On admettra que la variable aléatoire $Y_k = X_k^2$ est de moyenne $E[Y_k] = \theta$ et de variance $\text{var}(Y_k) = \theta^2$. En déduire la loi asymptotique de T_n sous les deux hypothèses H_0 et H_1.
   D'après le théorème central limite, on a

   $$\text{Sous } H_0 : T_n \approx \mathcal{N}\left(n\theta_0, n\theta_0^2\right),$$

   et

   $$\text{Sous } H_1 : T_n \approx \mathcal{N}\left(n\theta_1, n\theta_1^2\right),$$

   où $\approx$ signifie "de loi approchée (pour n grand)".

3. (2pts) On note F la fonction de répartition d'une loi du normale $\mathcal{N}(0, 1)$. En utilisant la loi asymptotique trouvée à la question précédente, exprimer le risque de première espèce α en fonction du seuil du test de Neyman Pearson noté S_α, de F, n et θ_0. En déduire la

valeur du seuil S_α en fonction de $F^{-1}(\alpha)$, de n et de θ_0.
Le risque α est défini par

$$\alpha = P[\text{Rejeter } H_0 | H_0 \text{ vraie}] = P\left[T_n < S_\alpha | T_n \sim \mathcal{N}\left(n\theta_0, n\theta_0^2\right)\right],$$

soit

$$\alpha = F\left[\frac{S_\alpha - n\theta_0}{\sqrt{n\theta_0^2}}\right].$$

On en déduit

$$S_\alpha = F^{-1}(\alpha)\sqrt{n\theta_0^2} + n\theta_0.$$

4. (2pts) Déterminer les caractéristiques opérationnelles du récepteur (courbes COR) pour ce test et montrer qu'elles ne dépendent que de $\sqrt{n}$ et de $\frac{\theta_1}{\theta_0}$. Analyser les performances du test en fonction de $\frac{\theta_1}{\theta_0}$ et de n et tracer l'allure des courbes COR pour différentes valeurs de $\frac{\theta_1}{\theta_0}$.
Les courbes COR expriment $\pi = 1 - \beta$ en fonction de α. On sait qu'il suffit d'utiliser l'expression de α et de changer θ_0 en θ_1 pour avoir π, soit

$$\pi = F\left[\frac{S_\alpha - n\theta_1}{\sqrt{n\theta_1^2}}\right],$$

soit en remplaçant S_α par son expression

$$\pi = F\left[\frac{\theta_0}{\theta_1}F^{-1}(\alpha) + \sqrt{n}\left(\frac{\theta_0}{\theta_1} - 1\right)\right].$$

Cette expression de π ne dépend donc que de n et de $\frac{\theta_1}{\theta_0}$. En particulier, quand n augmente ou quand $\frac{\theta_0}{\theta_1}$ augmente (i.e., $\frac{\theta_1}{\theta_0}$ diminue), la performance du test augmente, ce qui est attendu. Les allures approximatives de ces courbes pour différentes valeurs de n et de $\frac{\theta_1}{\theta_0}$ sont représentées ci-dessous :

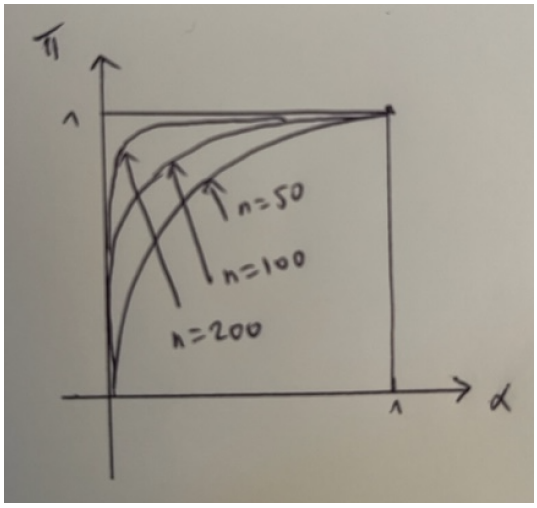

Figure 1: Allure des courbes COR pour différentes valeurs de n.

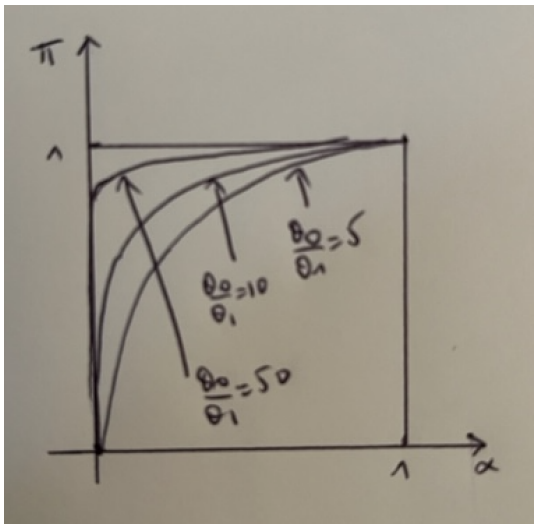

Figure 2: Allure des courbes COR pour différentes valeurs de $\frac{\theta_0}{\theta_1}$.

5. (3pts) On désire vérifier que l'ensemble des observations $(x_1, ..., x_n)$ suit une loi de densité $f(x_i;\theta)$ avec $\theta = 1$ à l'aide d'un test du χ^2. Déterminer la fonction de répartition de cette loi et en déduire que l'intervalle $]0, +\infty[$ est la réunion de trois intervalles équiprobables pour la loi de densité $f(x_i; 1)$ que l'on précisera. On compte le nombre d'observations x_i appartenant à ces trois intervalles et on trouve $K_1 = 13$, $K_2 = 8$ et $K_3 = 9$. Quelle est la valeur de la statistique du test du χ^2 ? Exprimer le seuil de ce test noté S_α en fonction du risque α et de l'inverse de la fonction de répartition d'une loi du χ^2 dont on précisera le nombre de degrés de liberté. On donne $S_{0.05} = 5.991$. Qu'en conclut-on ?
La fonction de répartition associée à la loi de densité $f(x_i;\theta)$ s'écrit

$$F_\theta(x) = \int_0^x f(u;\theta)du = \begin{cases} 0 \text{ si } x \leq 0 \\ \int_0^x \frac{2u}{\theta} \exp\left(-\frac{u^2}{\theta}\right) du = 1 - \exp\left(-\frac{x^2}{\theta}\right) \text{ si } x \geq 0 \end{cases}$$

Pour $\theta = 1$, on obtient donc

$$F_1(x) = \begin{cases} 0 \text{ si } x \leq 0 \\ 1 - \exp\left(-x^2\right) \text{ si } x \geq 0 \end{cases}$$

Le premier intervalle $]0, a[$ est tel que

$$F(a) = \frac{1}{3} \Leftrightarrow 1 - \exp\left(-x^2\right) = \frac{1}{3} \Leftrightarrow x = \sqrt{\ln\left(\frac{3}{2}\right)}$$

Le deuxième intervalle $[a, b]$ est tel que

$$F(b) = \frac{2}{3} \Leftrightarrow 1 - \exp\left(-x^2\right) = \frac{2}{3} \Leftrightarrow x = \sqrt{\ln 3}$$

La statistique du test du χ^2 est

$$\phi = \sum_{i=1}^{3} \frac{(K_i - np_i)^2}{np_i} = \frac{9}{10} + \frac{4}{10} + \frac{1}{10} = \frac{14}{10} = 1.4.$$

On sait que sous l'hypothèse H_0, ϕ suit une loi du χ^2 à $K-1$ degrés de liberté, donc

$$\phi \sim \chi^2_2.$$

Comme $\phi = 1.4 < S_{0.05} = 5.991$, on accepte l'hypothèse H_0 avec le risque $\alpha = 0.05$ donc on en déduit que l'échantillon suit la loi de densité $f(x_i;\theta)$ avec $\theta = 1$ avec le risque $\alpha = 0.05$.

LOIS DE PROBABILITÉ DISCRÈTES

$p_k = P[X = k]$ $\quad$ $p_{1,...,m} = P[X_1 = k_1, ..., X_m = k_m]$

LOI	**Probabilités**	Moyenne	Variance	Fonction Caractéristique
Uniforme	$p_k = \frac{1}{n}$ $k \in \{1, ..., n\}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$\frac{e^{it}(1-e^{itn})}{n(1-e^{it})}$
Bernoulli	$p_1 = P[X = 1] = p$ $p_0 = P[X = 0] = q$ $p \in [0,1] \quad q = 1-p$	p	pq	$pe^{it} + q$
Binomiale $B(n,p)$	$p_k = \binom{n}{k} p^k q^{n-k}$ $p \in [0,1] \quad q = 1-p$ $k \in \{0, 1, ..., n\}, \binom{n}{k} = \frac{n!}{k!(n-k)!}$	np	npq	$(pe^{it} + q)^n$
Binomiale négative	$p_k = \binom{n+k-1}{n-1} p^n q^k$ $p \in [0,1] \quad q = 1-p$ $k \in \mathbb{N}, \binom{n}{k} = \frac{n!}{k!(n-k)!}$	$n\frac{q}{p}$	$n\frac{q}{p^2}$	$\left(\frac{p}{1-qe^{it}}\right)^n$
Multinomiale	$p_{1,...,m} = \frac{n!}{k_1!...k_m!} p_1^{k_1} ... p_m^{k_m}$ $p_j \in [0,1] \quad q_j = 1-p_j$ $k_j \in \{0, 1, \ldots, n\}$ $\sum_{j=1}^m k_j = n \quad \sum_{j=1}^m p_j = 1$	np_j	Variance : np_jq_j Covariance : $-np_jp_k$	$\left(\sum_{j=1}^m p_j e^{it}\right)^n$
Poisson $P(\lambda)$	$p_k = e^{-\lambda}\frac{\lambda^k}{k!}$ $\lambda > 0 \quad k \in \mathbb{N}$	λ	λ	$\exp[\lambda(e^{it} - 1)]$
Géométrique	$p_k = pq^{k-1}$ $p \in [0,1] \quad q = 1-p$ $k \in \mathbb{N}^*$	$\frac{1}{p}$	$\frac{q}{p^2}$	$\frac{pe^{it}}{1-qe^{it}}$

LOIS DE PROBABILITÉ CONTINUES

LOI	**Densité de probabilité**	Moyenne	Variance	Fonction Caractéristique
Uniforme	$f(x) = \frac{1}{b-a}$ $x \in]a,b[$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{itb} - e^{ita}}{it(b-a)}$
Gamma $\mathcal{G}(\nu,\theta)$	$f(x) = \frac{\theta^\nu}{\Gamma(\nu)} e^{-\theta x} x^{\nu-1}$ $\theta > 0,\ \nu > 0$ $x \geq 0$ avec $\Gamma(n+1) = n!\ \forall n \in \mathbb{N}$	$\frac{\nu}{\theta}$	$\frac{\nu}{\theta^2}$	$\frac{1}{\left(1 - i\frac{t}{\theta}\right)^\nu}$
Inverse gamma $\mathcal{IG}(\nu,\theta)$	$f(x) = \frac{\theta^\nu}{\Gamma(\nu)} e^{-\frac{\theta}{x}} \frac{1}{x^{\nu+1}}$ $\theta > 0,\ \nu > 0$ $x \geq 0$ avec $\Gamma(n+1) = n!\ \forall n \in \mathbb{N}$	$\frac{\theta}{\nu-1}$ si $\nu > 1$	$\frac{\theta^2}{(\nu-1)^2(\nu-2)}$ si $\nu > 2$	(*)
Première loi de Laplace	$f(x) = \frac{1}{2} e^{-\lvert x \rvert}, \quad x \in \mathbb{R}$	0	2	$\frac{1}{1+t^2}$
Normale univariée $\mathcal{N}(m,\sigma^2)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$	m	σ^2	$e^{imt - \frac{\sigma^2 t^2}{2}}$
Normale multivariée $\mathcal{N}_p(\boldsymbol{m},\boldsymbol{\Sigma})$	$f(x) = K e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{m})}$ $K = \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}}$ $x \in \mathbb{R}^p$	$\boldsymbol{m}$	$\boldsymbol{\Sigma}$	$e^{i\boldsymbol{u}^T\boldsymbol{m} - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}}$
Khi_2 χ^2_ν $\Gamma\left(\frac{\nu}{2}, \frac{1}{2}\right)$	$f(x) = k e^{-\frac{x}{2}} x^{\frac{\nu}{2}-1}$ $k = \frac{1}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)}$ $\nu \in \mathbb{N}^*,\ x \geq 0$	ν	2ν	$\frac{1}{(1-2it)^{\frac{\nu}{2}}}$
Cauchy $c_{\lambda,\alpha}$	$f(x) = \frac{1}{\pi\lambda\left(1 + \left(\frac{x-\alpha}{\lambda}\right)^2\right)}$ $\lambda > 0,\ \alpha \in \mathbb{R}$	(-)	(-)	$e^{i\alpha t - \lambda\lvert t \rvert}$
Beta $B(a,b)$	$f(x) = k x^{a-1} (1-x)^{b-1}$ $k = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ $a > 0,\ b > 0$ $x \in]0,1[$ avec $\Gamma(n+1) = n!\ \forall n \in \mathbb{N}$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	(*)