EXAMEN STATISTIQUE 1MF2E

Lundi 24 mars 2025

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

Exercice 1 : Test Statistique (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

$$f(x_i; a) = \begin{cases} (a-1)x_i^{-a} & \text{si } x_i > 1 \\ 0 \text{ sinon} \end{cases}$$

On désire utiliser les observations $x_1, ..., x_n$ pour déterminer si $a = a_0 > 1$ ou si $a = a_1 > a_0$. On considère donc le test d'hypothèses

$$H_0 : a = a_0, \quad H_1 : a = a_1 \qquad \text{avec } a_1 > a_0 > 1.$$

1. (2pt) Déterminer la statistique T_n du test de Neyman Pearson et la région critique associée. En admettant que $E[\ln X_i] = \frac{1}{a-1}$, vérifier que la région critique du test est en accord avec cette espérance mathématique.
   Le test de Neyman Pearson est défini par

   $$\text{Rejet de } H_0 \text{ si } \frac{L(x_1, ..., x_n; a_1)}{L(x_1, ..., x_n; a_0)} > S_{1,\alpha}$$

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

   $$\frac{L(x_1, ..., x_n; a_1)}{L(x_1, ..., x_n; a_0)} > S_{1,\alpha} \Leftrightarrow \ln \frac{L(x_1, ..., x_n; a_1)}{L(x_1, ..., x_n; a_0)} > S_{2,\alpha}$$
   $$\Leftrightarrow (a_0 - a_1) \sum_{i=1}^{n} \ln(x_i) > S_{3,\alpha}.$$

   Comme $a_1 > a_0$, on a la règle de décision

   $$\text{Rejet de } H_0 \text{ si } T_n = \sum_{i=1}^{n} \ln(x_i) < S_\alpha.$$

   La région critique du test est l'ensemble des vecteurs $(x_1, ..., x_n) \in]1, +\infty]^n$ tels que $T_n < S_\alpha$ et la statistique de test est $T_n = \sum_{i=1}^{n} \ln(x_i)$.
   On rejette donc l'hypothèse H_0 si la moyenne des variables aléatoires $\ln(X_i)$ est petite, ce qui est en accord avec $a_1 < a_0$.

2. (1pt) En admettant que $Y_i = \ln X_i$ suit une loi gamma de paramètres 1 et $a-1$, i.e., $Y_i \sim \mathcal{G}(1, a-1)$, déterminer la loi approchée de T_n sous les deux hypothèses H_0 et H_1 provenant du théorème central limite.
   La moyenne et la variance de Y_i sont données dans la table

   $$E[Y_i] = \frac{1}{a-1} \text{ et } \text{var}[Y_i] = \frac{1}{(a-1)^2}.$$

   Donc d'après le théorème central limite, on a

   $$\frac{T_n - \frac{n}{a-1}}{\sqrt{\frac{n}{(a-1)^2}}} \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, 1).$$

On en déduit

$$\text{Sous } H_0 : T_n \approx \mathcal{N}\left(\frac{n}{a_0-1}, \frac{n}{(a_0-1)^2}\right),$$

$$\text{Sous } H_1 : T_n \approx \mathcal{N}\left(\frac{n}{a_1-1}, \frac{n}{(a_1-1)^2}\right),$$

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

3. (2pts) On note G la fonction de répartition d'une loi du normale $\mathcal{N}(0,1)$. Exprimer le risque de première espèce α en fonction du seuil du test de Neyman Pearson noté S_α, de G, n et de a_0. En déduire la valeur du seuil S_α en fonction de $G^{-1}(\alpha)$ et de n et a_0.
Le risque α est défini par

$$\alpha = P[\text{Rejeter } H_0 | H_0 \text{ vraie}] = P\left[T_n < S_\alpha | T_n \sim \mathcal{N}\left(\frac{n}{a_0-1}, \frac{n}{(a_0-1)^2}\right)\right],$$

soit

$$\alpha = G\left[\frac{S_\alpha - \frac{n}{a_0-1}}{\sqrt{\frac{n}{(a_0-1)^2}}}\right] = G\left(\frac{a_0-1}{\sqrt{n}} S_\alpha - \sqrt{n}\right).$$

d'où

$$S_\alpha = \frac{\sqrt{n}}{a_0-1}\left[G^{-1}(\alpha) + \sqrt{n}\right]$$

4. (2pts) Déterminer les caractéristiques opérationnelles du récepteur (courbes COR) pour ce test et montrer qu'elles ne dépendent que de n et de $r(a_0,a_1) = \frac{a_1-1}{a_0-1}$ (et évidemment des fonctions G et G^{-1}). Analyser les performances du test en fonction de $r(a_0,a_1)$ et tracer l'allure approximative des courbes COR pour différentes valeurs de a_1 lorsque $a_0 = 2$.
Le risque β est défini par

$$\beta = P[\text{Rejeter } H_1 | H_1 \text{ vraie}] = P\left[T_n \geq S_\alpha | T_n \sim \mathcal{N}\left(\frac{n}{a_1-1}, \frac{n}{(a_1-1)^2}\right)\right],$$

soit (en remplaçant a_0 par a_1 dans l'expression de α

$$\pi = 1 - \beta = G\left(\frac{a_1-1}{\sqrt{n}} S_\alpha - \sqrt{n}\right).$$

Les courbes COR sont donc définie par

$$\pi = G\left[\frac{a_1-1}{a_0-1}\left(G^{-1}(\alpha) + \sqrt{n}\right) - \sqrt{n}\right],$$

soit

$$\pi = G\left[r(a_0,a_1)\left(G^{-1}(\alpha) + \sqrt{n}\right) - \sqrt{n}\right].$$

L'allure approximative de ces courbes pour différentes valeurs de a_1 est représentée ci-dessous :

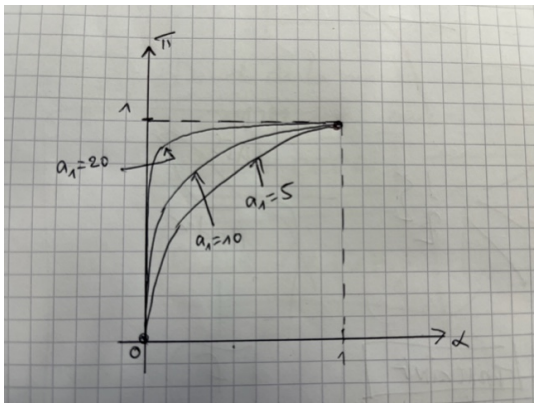

Figure 1: Allure des courbes COR pour différentes valeurs de a_1.

5. (3pts) On désire vérifier que les observations suivent la loi de densité $f(x_i;a)$ avec $a=2$ à l'aide d'un test de Kolmogorov.

   - Déterminer la fonction de répartition de la loi de densité $f(x_i;a)$ notée F lorsque $a=2$.
   - On observe l'échantillon de taille $n=4$ suivant : $x_1=2$, $x_2=3$, $x_3=4$ et $x_4=\frac{3}{2}$. Le tableau suivant résume les quantités nécessaires pour effectuer le test

$x_{(i)}$	2	2.5	3	4
$F\left(x_{(i)}\right)$	0.5000	0.6000	0.6667	0.7500
$\hat{F}\left(x_{(i)}^-\right)$	0	0.25	0.50	0.75
$\hat{F}\left(x_{(i)}^+\right)$	0.25	0.50	0.75	1
$E_i^- = \lvert F\left(x_{(i)}\right) - \hat{F}\left(x_{(i)}^-\right)\rvert$	0.5000	0.3500	0.1667	0.0000
$E_i^+ = \lvert F\left(x_{(i)}\right) - \hat{F}\left(x_{(i)}^+\right)\rvert$	0.2500	0.1000	0.0833	0.2500

   où $(x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)})$ est l'échantillon ordonné.
   Expliquer ce que représentent $\hat{F}\left(x_{(i)}^-\right)$ et $\hat{F}\left(x_{(i)}^+\right)$.
   - Rappeler la région critique du test de Kolmogorov. Pour $\alpha=0.01$ et $n=4$, on a $S_{0.01}=0.7342$. Que peut-on en conclure ?

Correction

- La fonction de répartition de $f(x_i;2)$ s'écrit

$$F(x)=\int_1^x f(u;2)du=\begin{cases} 0 \text{ si } x\leq 1 \\ \int_1^x \frac{1}{u^2}du = 1-\frac{1}{x} \text{ si } x>1 \end{cases}$$

- $\hat{F}\left(x_{(i)}^-\right)$ et $\hat{F}\left(x_{(i)}^+\right)$ représentent les valeurs à gauche et à droite de la fonction en escaliers $\hat{F}$ au point $x_{(i)}$.
- Le test de Kolmogorov rejette l'hypothèse H_0 si

$$D_n=\max_{i\in\{1,\dots,n\}} \max\{E_i^+, E_i^-\} > S_\alpha.$$

  Dans notre cas $D_n=0.5000 < S_{0.01}=0.7342$ et donc on accepte l'hypothèse H_0 (les observation suivent la loi de densité $f(x_i;a)$ avec $a=2$) avec le risque $\alpha=0.01$

Exercice 2 : 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 même loi de densité

$$f(x_i; a) = \begin{cases} (a-1)x_i^{-a} & \text{si } x_i > 1 \\ 0 \text{ sinon} \end{cases}$$

avec $a > 1$.

1. (1pts) Montrer que l'estimateur du vraisemblance du paramètre a noté $\widehat{a}_{\text{MV}}$ est défini par

$$\widehat{a}_{\text{MV}} = 1 + \frac{n}{\sum_{i=1}^n \ln X_i}.$$

La vraisemblance de l'échantillon s'écrit

$$L(x_1, ..., x_n; a) = \prod_{i=1}^n f(x_i; a) = (a-1)^n \prod_{i=1}^n \frac{1}{x_i^a} I_{]1,+\infty[}(x_i),$$

où $\mathbb{I}_{]1,+\infty[}$ est la fonction indicatrice sur l'intervalle $]1, +\infty[$. On sait qu'il est plus facile de travailler avec la log-vraisemblance définie par

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

Cette log-vraisemblance admet pour dérivée

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

Les variations de cette dérivée sont définies par

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

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

$$\widehat{a}_{\text{MV}} = 1 + \frac{n}{\sum_{i=1}^n \ln(X_i)}.$$

2. (2pts) On rappelle que si Y est une variable aléatoire réelle de loi gamma de paramètres ν et θ, notée $Y \sim \mathcal{G}(\nu, \theta)$, alors $Z = \frac{1}{Y}$ suit une loi inverse-gamma de paramètres ν et θ, i.e., $Z \sim \mathcal{IG}(\nu, \theta)$. Montrer que la variable aléatoire $Y_i = \ln X_i$ suit une loi gamma dont on déterminera les paramètres. Quelle est la loi de $Y = \sum_{i=1}^n \ln X_i$? En déduire que $\widehat{a}_{\text{MV}} = 1 + nZ$ où Z est une variable aléatoire de loi inverse gamma $\mathcal{IG}(n, a-1)$.

La variable aléatoire Y_i est à valeurs dans $\mathbb{R}^+$. En faisant un changement de variables (et en faisant attention de ne pas oublier le Jacobien ;-)), on obtient la densité de Y_i

$$g(y_i; a) = \begin{cases} (a-1)e^{-(a-1)y_i} \text{ si } y_i > 0 \\ 0 \quad \text{sinon} \end{cases}$$

On reconnaît une loi gamma $\mathcal{G}(1, a-1)$ dont la fonction caractéristique (donnée dans la table) est

$$\phi_{Y_i}(t) = E[e^{iY_i t}] = \frac{1}{1 - i\frac{t}{a-1}}.$$

La fonction caractéristique de Y est le produit des fonctions caractéristiques des variables Y_i car ces variables sont indépendantes. On a donc

$$\phi_Y(t) = E[e^{iYt}] = \frac{1}{\left(1 - i\frac{t}{a-1}\right)^n}$$

qui est la fonction caractéristique d'une loi gamma $\mathcal{G}(n, a-1)$, donc $Y \sim \mathcal{G}(n, a-1)$. D'après l'indication $Z = \frac{1}{Y} = \frac{1}{\sum_{i=1}^n \ln(X_i)}$ suit une loi inverse gamma $\mathcal{G}(n, a-1)$, d'où $\widehat{a}_{\text{MV}} = 1 + nZ$, où Z suit une loi inverse gamma $\mathcal{IG}(n, a-1)$.

3. (3pts) Déterminer le bias de l'estimateur $\widehat{a}_{\text{MV}}$ et en déduire un estimateur sans biais du paramètre a noté $\widetilde{a}_{\text{MV}}$. Déterminer la variance de $\widetilde{a}_{\text{MV}}$ et en déduire que cet estimateur est convergent.
La loi gamma $\mathcal{G}(n, a-1)$ est de moyenne $\frac{a-1}{n-1}$ et de variance $\frac{(a-1)^2}{(n-1)^2(n-2)}$. Donc

$$E[\widehat{a}_{\text{MV}}] = 1 + \frac{n(a-1)}{n-1}.$$

On en déduit un estimateur non biaisé de a

$$\widetilde{a}_{\text{MV}} = \frac{(n-1)\widehat{a}_{\text{MV}} + 1}{n}.$$

Cet estimateur est de variance

$$\text{var}[\widetilde{a}_{\text{MV}}] = \frac{(n-1)^2}{n^2}\text{var}[\widehat{a}_{\text{MV}}] = \frac{(n-1)^2}{n^2} \times n^2 \text{var}[Z] = \frac{(a-1)^2}{n-2}.$$

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

4. (1pt) Déterminer la borne de Cramer-Rao pour un estimateur non biaisé du paramètre a. L'estimateur $\widetilde{a}_{\text{MV}}$ est-il l'estimateur efficace du paramètre a ?
La dérivée seconde de la log-vraisemblance est

$$\frac{\partial^2 \ln L(x_1, ..., x_n; a)}{\partial a^2} = \frac{-n}{(a-1)^2}.$$

d'où

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

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

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

Comme var $[\widetilde{a}_{\text{MV}}] > \frac{(a-1)^2}{n}$, $\widetilde{a}_{\text{MV}}$ n'est pas l'estimateur efficace du paramètre a. On notera que cet estimateur est asymptotiquement efficace.

5. (3pts) On suppose désormais que le paramètre a est muni d'une loi a priori de densité

$$f(a) = \begin{cases} e^{1-a} \text{ si } a > 1 \\ 0 \text{ sinon} \end{cases}$$

Déterminer l'estimateur du maximum a posteriori du paramètre a noté $\widehat{a}_{\text{MAP}}$. À l'aide de la densité de la variable $W = a|x_1, ..., x_n$ (a sachant $x_1, ..., x_n$), connue à une constante multiplicative près, montrer que la loi de $U = W - 1$ est une loi gamma dont on déterminera les paramètres (on remarquera que la densité de W est non nulle sur l'intervalle $]1, \infty[$). En déduire l'estimateur MMSE de a noté $\widehat{a}_{\text{MMSE}}$.
La loi a posteriori du paramètre a vérifie

$$\begin{aligned} p(a|x_1, ..., x_n) &\propto p(x_1, ..., x_n|a) f(a) \\ &\propto (a-1)^n \exp\left\{-a\left[1 + \sum_{i=1}^{n} \ln(x_i)\right]\right\} \mathcal{I}_{]1,+\infty[}(a), \end{aligned}$$

qui a la même forme que la log-vraisemblance si on remplace $\sum_{i=1}^{n} \ln(x_i)$ par $1 + \sum_{i=1}^{n} \ln(x_i)$. L'estimateur du maximum a posteriori du paramètre a s'obtient donc à partir de l'estimateur du maximum de vraisemblance de a en faisant cette même transformation, soit

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

Si on note $W = a|x_1, ..., x_n$ et qu'on fait un changement de variables $U = -1 + W$, on obtient la densité de U:

$$\begin{aligned} p(u|x_1, ..., x_n) &\propto u^n \exp\left\{-(u+1)\left[1 + \sum_{i=1}^{n} \ln(x_i)\right]\right\} \mathcal{I}_{]0,+\infty[}(u) \\ &\propto u^n \exp\left\{-u\left[1 + \sum_{i=1}^{n} \ln(x_i)\right]\right\} \mathcal{I}_{]0,+\infty[}(u). \end{aligned}$$

qui est une loi gamma $\mathcal{G}(n+1, 1 + \sum_{i=1}^{n} \ln(x_i))$. L'estimateur MMSE est donc

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

LOIS DE PROBABILITÉ CONTINUES

$\mathbf{m}$: moyenne σ^2 : variance **F. C. :** fonction caractéristique

LOI	**Densité de probabilité**	$\mathbf{m}$	σ^2	**F. C.**
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 $\mathcal{N}_p(\boldsymbol{m},\boldsymbol{\Sigma})$	$f(x) = Ke^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m})}$ $K = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}}$ $\boldsymbol{x} = (x_1, ..., x_p)^T \in \mathbb{R}^p$	$\boldsymbol{m}$	$\boldsymbol{\Sigma}$	$e^{i\boldsymbol{u}^T\boldsymbol{m} - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}}$
Khi$_2$ χ^2_ν $\Gamma\left(\frac{\nu}{2}, \frac{1}{2}\right)$	$f(x) = ke^{-\frac{x}{2}} x^{\frac{\nu}{2}-1}$ $k = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}$ $\nu \in \mathbb{N}^*,\ x \geq 0$	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 $B(a,b)$	$f(x) = kx^{a-1}(1-x)^{b-1}$ $k = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ $a > 0,\ b > 0$ $x \in]0,1[$ avec $\Gamma(n+1) = n!\ \forall n \in \mathbb{N}$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	(*)

LOIS DE PROBABILITÉ DISCRÈTES

$\mathbf{m}$: moyenne σ^2 : variance **F. C. :** fonction caractéristique

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

LOI	**Probabilités**	$\mathbf{m}$	σ^2	**F. C.**
Uniforme	$p_k = \frac{1}{n}$ $k \in \{1, ..., n\}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$\frac{e^{it}\left(1-e^{itn}\right)}{n\left(1-e^{it}\right)}$
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 $B(n, p)$	$p_k = C_n^k p^k q^{n-k}$ $p \in [0, 1] \quad q = 1 - p$ $k \in \{0, 1, ..., n\}$	np	npq	$\left(pe^{it} + q\right)^n$
Binomiale négative	$p_k = C_{n+k-1}^{n-1} p^n q^k$ $p \in [0, 1] \quad q = 1 - p$ $k \in \mathbb{N}$	$n\frac{q}{p}$	$n\frac{q}{p^2}$	$\left(\frac{p}{1-qe^{it}}\right)^n$
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	$\exp\left[\lambda\left(e^{it} - 1\right)\right]$
Géométrique	$p_k = pq^{k-1}$ $p \in [0, 1] \quad q = 1 - p$ $k \in \mathbb{N}^*$	$\frac{1}{p}$	$\frac{q}{p^2}$	$\frac{pe^{it}}{1-qe^{it}}$