ÉCOLE UNIVERSITAIRE DE PHYSIQUE ET D'INGÉNIERIE Université Clermont Auvergne	**2024-2025** **M2PFA UP & NP** **IMMAP** **DU Data Scientist**
Statistics homework October 2024	

A propos de l'échantillonnage aléatoire

Il est parfois nécessaire de générer un échantillon de valeurs aléatoires qui suit une distribution expérimentale, qui peut possèder (ou non) une définition analytique explicite. Le but de cet exercice est d'explorer quelques-unes des techniques permettant de produire de tels échantillons.

I - Echantillonner la réciproque de la fonction cumulative.

Suposons que la fonction de répartition (ou fonction cumulative, CDF) de la distribution ciblée, $\mathscr{F}_{\mathrm{X}}$, soit connue, ou tout du moins qu'elle puisse être dérivée numériquement des données expérimentales. Afin de générer un échantillon aléatoire de variables indépendantes X, identiquement distribuée suivant la densité de cumulative $\mathscr{F}_{\mathrm{X}}$, on peut utiliser une transformation de la distribution continue *uniforme* dans l'intervalle $[0,1]$, $\mathrm{U} \sim \mathbb{I}_{[0,1]}$. Cette dernière a la propriété de coïncider avec sa propre réciproque : $\mathbf{P}(\mathrm{U} < p) = \mathscr{F}_{\mathrm{U}}(p) = p = \mathscr{F}_{\mathrm{U}}^{-1}(p)$. La *méthode d'échantillonnage inverse* repose sur le théorème de la réciproque qui stipule que pour toute fonction réciproque $Q_{\mathrm{X}}(p) = \mathscr{F}_{\mathrm{X}}^{-1}(x)$, la variable aléatoire construite comme $\mathrm{X} = Q_{\mathrm{X}}(\mathrm{U})$ admet $\mathscr{F}_{\mathrm{X}}(x)$ comme fonction de répartition.

Dans le cadre de cet exercice, considérons par exemple la fonction *sigmoïde* définie par:

$$\mathscr{F}_{\mathrm{X}}(x;\lambda) = \frac{1}{1+e^{-\lambda x}}.$$

avec pour paramètre $\lambda \in \mathbb{R}^+$. Cette fonction monotone croissante, $\mathscr{F}_{\mathrm{X}} : x \in \mathbb{R} \to \mathscr{F}_{\mathrm{X}}(x,\lambda) \in [0,1]$ constitue une fonction de répartition valide et normalisée pour la variable aléatoire continue X. Il est facile de déterminer la fonction réciproque (ou fonction *quantile*, *percent point fonction*, PPF):

$$\mathscr{F}_{\mathrm{X}}(x,\lambda) = \mathbf{P}(\mathrm{X} < x) = p \implies x = Q_{\mathrm{X}}(p,\lambda) = \mathscr{F}_{\mathrm{X}}^{-1}(p,\lambda) = \frac{1}{\lambda}\ln\left(\frac{p}{1-p}\right).$$

1- Générez un échantillon de $N = 10^5$ probabilités aléatoires p_k, uniformément distribuées dans l'intervalle $[0,1]$.[1] Pour chaque valeur p_k, évaluez la valeur correspondante du quantile de $\mathscr{F}_{\mathrm{X}}$

$$x_k = \mathscr{F}_{\mathrm{X}}^{-1}(p_k;\lambda=1) = \ln(p_k/(1-p_k)).$$

Affichez l'histogramme de densité obtenu pour l'échantillon$\{x_k\}$.

Vous avez maintenant produit un échantillon aléatoires de valeurs distribuées suivant la loi $\mathrm{X} \sim f_{\mathrm{X}}(x;\lambda=1)$, pour la valeur de paramètre $\lambda = 1$.

2- Il est assez facile de dériver explicitement l'expression de la *fonction de densité de probability* (PDF):

$$f_{\mathrm{X}}(x,\lambda) = \frac{\partial \mathscr{F}_{\mathrm{X}}(x;\lambda)}{\partial x} = \frac{\lambda}{4}\,\mathrm{sech}^2\left(\frac{\lambda x}{2}\right),$$

où $\mathrm{sech}(z) = 1/\cosh(z)$ est la fonction *sécante hyperbolique*.

Superposez la courbe de densité, $f_{\mathrm{X}}(x,1)$, sur l'histogramme. Correspond-elle ?

[1] Adaptez N à vos ressources informatiques. Vous pouvez commencer avec une valeur plus faible pour développer votre code, N=10^4 par exemple, puis augmenter progressivement tant que le temps d'exécution reste raisonnable.

3- Afin de valider la distribution générée, estimons le paramètre λ de la PDF théorique en utilisant la méthode du Maximum de Vraisemblance (ML). Définissez la fonction négative du log-vraisemblance en fonction du paramètre λ, évaluée sur l'échantillon $\{x_k\}$:

$$-\ln\mathscr{L}\left(\lambda;\{x_k\}\right) = -\sum_{k=1}^{N}\ln\left[f_{\mathrm{X}}(x_k;\lambda)\right],$$

et trouvez la valeur de l'estimateur $\hat{\lambda}$ qui minimise cette fonction.

4- Evaluez l'incertitude associée, $\hat{\sigma}_\lambda$, en utilisant la méthode graphique, $\Delta\ln\mathscr{L}(\lambda_\pm) = 1/2$. Est-ce que la valeur mesurée sur l'échantillon, $\hat{\lambda} \pm \hat{\sigma}_\lambda$, est compatible avec la valeur théorique attendue $\lambda = 1$?

5- Si oui, félicitations ! Vous savez maintenant générer un échantillon aléatoire à partir de n'importe quelle fonction de répartition.[2] Vous avez cependant travaillé pour rien, sinon la gloire, car cette loi de probabilité bien connue est déjà implémentée dans la librairie *scipy* sous le nom de *loi logistique*, avec toute la machinerie nécessaire à la génération aléatoire d'événements, ou d'évaluation des fonctions PDF, CDF, PPF, ou log-Likelihood ...
Produisez un échantillon alternatif de N variables *logistiques* avec le générateur intégré (scipy.stats.logistic.rvs(size=N)). Comparez cet échantillon avec votre propre échantillonnage, en produisant le *graphe des résidus*, c'est-à-dire l'histogramme de la différence pour chaque *bin* des deux histogrammes, divisées par l'incertitude associée. Vous pourrez faire l'approximation, $\sigma_{n_i} \simeq \sqrt{n_i}$, pour l'incertitude sur le comptage d'événements du *bin* i (attention à la normalisation en densité). Conclure.

6- La loi *logistique* possède un profile symétrique avec des queues plus étalées que la loi Gaussienne-Normale (ce genre de profile est appelé *leptokurtique*). La Valeur attendue et la Variance d'une variable aléatoire *logistique* centrée, $\mathrm{X} \sim f_{\mathrm{X}}(x,\lambda)$, sont:

$$\mathbb{E}\left[\mathrm{X}\right] = 0 \quad \text{et} \quad \mathbb{V}\left[\mathrm{X}\right] = \frac{\pi^2}{3\lambda^2}.$$

Essayons de jouer avec cette Variance pour évaluer le nombre π.
Utilisez l'échantillon $\{x_k\}$ initialement généré avec le paramètre $\lambda = 1$, pour construire un nouvel échantillon $\{z_k\}$, où la variable aléatoire Z est définie comme la valeur centrée réduite de X multipliée par le facteur $1/\sqrt{3}$, i.e.

$$z_k = \frac{x_k - \bar{\mu}_{\mathrm{X}}}{\sqrt{3\bar{\sigma}^2_{\mathrm{X}}}}.$$

Les grandeurs $\bar{\mu}_{\mathrm{X}}$ et $\bar{\sigma}^2_{\mathrm{X}}$, sont respectivement la moyenne et la variance de l'echantillon $\{x_k\}$.[3] Dans la limite des grands N, la variance de l'échantillon tend vers la variance de la distribution, $\bar{\sigma}^2_{\mathrm{X}} \simeq \mathbb{V}\left[\mathrm{X}\right]$. La Variance de la variable aléatoire modifiée Z vaut alors $\mathbb{V}\left[\mathrm{Z}\right] \simeq 1/3$ et sa densité de probabilité est une loi *logistique*, $f_{\mathrm{Z}}(z;\lambda_{\mathrm{Z}})$ de paramètre $\lambda_{\mathrm{Z}} = \pi$.
Répétez les questions I-3 et I-4 pour évaluer l'estimateur $\hat{\lambda}_{\mathrm{Z}}$ qui minimise la fonction négative du log-vraisemblance, $-\ln\mathscr{L}\left(\lambda;\{z_k\}\right)$, évaluée sur l'échantillon modifié $\{z_k\}$, ainsi que l'incertitude associée.[4] Est-ce que la mesure est compatible avec la valeur connue de π ? (et vous pouvez ainsi conclure que les méthodes d'échantillonnage sont généralement très peu efficaces pour déterminer un nombre significatif de décimales de π)

[2] Dans les cas où l'expression analytique de la fonction quantile ne peut pas être dérivée facilement, il est toujours possible d'interpoler la distribution expérimentale pour déterminer les valeurs de x correspondant à n'importe quelle valeur $p = \mathscr{F}_{\mathrm{X}}(x)$ de la fonction.

[3] Vous pouvez rapidement obtenir ces valeurs en important le module python *statistics* qui fournit: $\bar{\mu}_{\mathrm{X}}$ = statistics.mean(X) et $\bar{\sigma}^2_{\mathrm{X}}$ = statistics.variance(X).

[4] L'incertitude sur la détermination de $\bar{\sigma}^2_{\mathrm{X}}$ est négligée ici.

II - La méthode d'échantillonnage accept-reject.

La méthode accept-reject est une technique *Monte-Carlo* qui permet de "sculpter" un échantillon aléatoire d'une densité cible, $g_{\rm X}(x)$, à partir d'une densité de référence, $f_{\rm X}(x)$, qui vérifie le critère

$$\alpha f_{\rm X}(x) \geq g_{\rm X}(x) \;\forall\; x \in supp\,(g_{\rm X})\,,$$

où α est un facteur d'échelle constant qui assure l'inégalité en tout x.
La méthode consiste à

i) générer une probabilité aléatoire p uniformément distribuée dans l'intervalle $[0, 1]$.

ii) générer une valeur aléatoire x qui suit la densité de référence $f_{\rm X}(x)$.

iii) si le rapport des densités $g_{\rm X}(x)/\alpha f_{\rm X}(x) \geq p$ la valeur x est acceptée (rejetée sinon).

iv) répéter cette procédure N fois.

La méthode accept-reject nécessite généralement d'importants échantillons de reference, en particulier pour peupler les régions de faible densité de la distribution cible. Plus les profils entre distribution de référence et cible sont proches, plus le taux de rejection est faible.
Essayons de générer un échantillon aléatoire distribué suivant la loi *Normale*, en utilisant la loi *logistique* plus étalée comme référence.

1- Le maximum de densité vaut $f_{\rm X}(0; 1) = 1/4$ pour la loi *logistique*, et $\mathscr{N}_{\rm X}(0; 0, 1) = 1/\sqrt{2\pi}$ pour la loi *Normale*. Nous pouvons choisir la valeur minimale du facteur d'échelle $\alpha_0 = 4/\sqrt{2\pi}$ telle que la distribution *logistique* "englobe" la loi de densité *Normale* pour tout x.
Tracer les deux functions, $\alpha_0 f_{\rm X}(x, 1)$ et $\mathscr{N}_{\rm X}(x; 0, 1)$, sur un même graphe pour vérifier que l'on a bien $\alpha_0 f_{\rm X}(x) \geq \mathscr{N}_{\rm X}(x) \;\forall\; x$.[5]

2- Pour chacune des valeurs de la distribution *logistique* de l'échantillon $\{x_k\}$, $(k = 1, N)$
(i) évaluez la rapport des densité $r(x_k) = \mathscr{N}_{\rm X}(x_k)/\alpha_0 f_{\rm X}(x_k, 1)$.
(ii) générez une probabilité aléatoire p uniformément distribuée dans l'intervalle $[0, 1]$
(iii) si $r(x_k) \geq p$ gardez x_k dans le sous-échantillon des valeurs acceptées $\{x_m\}_a$, $m = 1, M \leq N$.
Quel est le taux d'acceptance M/N ?

3- Construisez l'histogramme de densité pour le sous-échantillon $\{x_m\}_a$. Vérifiez le profil *Normal* en réalisant l'ajustement d'une loi gaussienne sur le sous-échantillon (avec scipy.stats.norm.fit, par exemple). Donner les valeurs obtenues des paramètres ajustés $\hat{\mu}$ et $\hat{\sigma}$ et tracez la courbe de densité, $\mathscr{N}(x; \hat{\mu}, \hat{\sigma})$, sur l'histogramme.

4- Bien que généralement un choix sous-optimal, la loi *Uniforme* dans un intervalle fini $[a, b]$, englobant tout autre distribution avec le bon facteur d'échelle, peut être utilisée comme distribution de référence pour produire n'importe quelle distribution aléatoire sur un intervalle fini. Déterminez le facteur d'échelle minimal α_0 et produisez un échantillon de valeurs aléatoires sur l'intervalle $[-10, 10]$ pour la loi *logistique* $\mathrm{X} \sim f_{\rm X}(x; 1)$, à partir d'un échantillon de N valeurs aléatoires uniformément distribuées. Quel est le taux d'acceptance M/N ? Essayez de trouver un meilleur choix pour la distribution de référence que la loi uniforme, pour produire un échantillon *logistique*.

[5] La méthode fonctionne identiqueent pour tout facteur d'échelle $\alpha \geq \alpha_0$, mais plus la valeur est grande plus le taux de réjection augmente.