CentraleSupélec

Rapport confidentiel
CEI UX-KEY

Estimation des incertitudes pour les métriques d'ergonomie web (UX Design)

Etudiants :
Mohamed EL MEJDANI
Nawel SAKLY

Encadrants :
Frédéric PENNERATH
Loïc CUERONI

6 mai 2024

Table des matières

1 Introduction

Dans le cadre de notre sujet, on s'intéresse à l'analyse du comportement des utilisateurs de sites internet. La collecte des données pour un échantillon d'utilisateurs a permis d'associer à chaque page différentes métriques statistiques traduisant autant de qualités ergonomiques attendues (intuitivité, fluidité, etc). Chaque métrique a sa fonction de score associée qui associe à chaque observation une valeur entre 0 et 1.

L'estimation des scores des métriques des pages web est confrontée à deux formes d'incertitudes :

— L'incertitude aléatoire (data uncertainty) : c'est l'incertitude intrinsèque au processus de génération des données et qui est dûe au caractère aléatoire du phénomène étudié. Dans le cadre de ce sujet, cette incertitude est reliée au fait que le comportement des utilisateurs est aléatoire ce qui rend difficile la prédiction exacte des métriques.

— L'incertitude épistémique (model uncertainty) : c'est l'incertitude intrinsèque au processus d'apprentissage et qui est dûe au manque d'observations / données. Le nombre de données qu'on a sur les utilisateurs est limité , ce qui engendre ce type d'incertitude.

Ces notions d'incertitudes vont être plus développées dans ce qui suit.

Dans ce projet, notre objectif est de tenir compte de l'incertitude épistémique associée à l'estimation des scores. Pour aborder cette incertitude, on adopte une approche basée sur des intervalles de confiance. En modélisant le score et en obtenant l'intervalle probable de sa variation, nous sommes ainsi en mesure de quantifier l'incertitude et de déterminer le nombre d'échantillons requis pour son estimation. Vu que les scores sont des fonctions des métriques, on commence par modéliser les métriques avant de procéder à la modélisation du score.

2 Etude des données

On prend la métrique d'intuitivité (intuitiveness) comme exemple dans ce qui suit et on étudie la distribution des données obtenues sur cette métrique. On rappelle la définition de la métrique d'intuitivité :

Définition 1. *L'*intuitivité *est le temps écoulé entre la fin de chargement de la page et la première interaction d'action du visiteur. Cette métrique quantifie la compréhension au premier regard de l'utilisateur sur la page visitée (ou LOM).*

La distribution de l'intuitivité est tracée pour plusieurs LOMs dans la figure 1. On peut assimiler visuellement la distribution des données à la distribution Gamma (voir figure 1). La distribution Gamma est généralement paramétrée à l'aide des deux paramètres : le *paramètre de forme* α et le *paramètre d'intensité* β. Une variable aléatoire X suit la loi $\Gamma(\alpha, \beta)$ si sa fonction de densité de probabilité peut se mettre sous la forme :

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}\beta^{\alpha}e^{-\beta x}}{\Gamma(\alpha)}$$

pour x>0 ; où

$$\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}dt$$

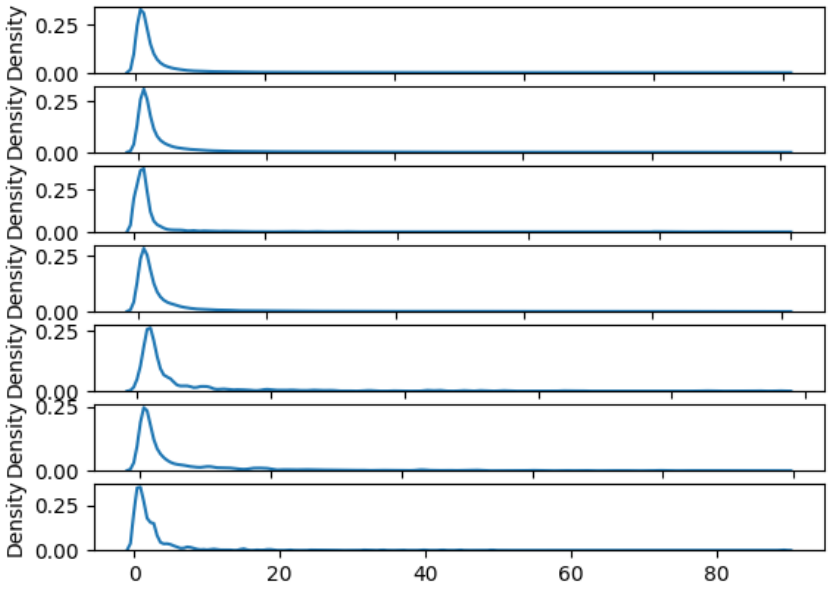


FIGURE 1 – Courbes de la distribution pour les LOMs du fichier endsomethingthou.csv

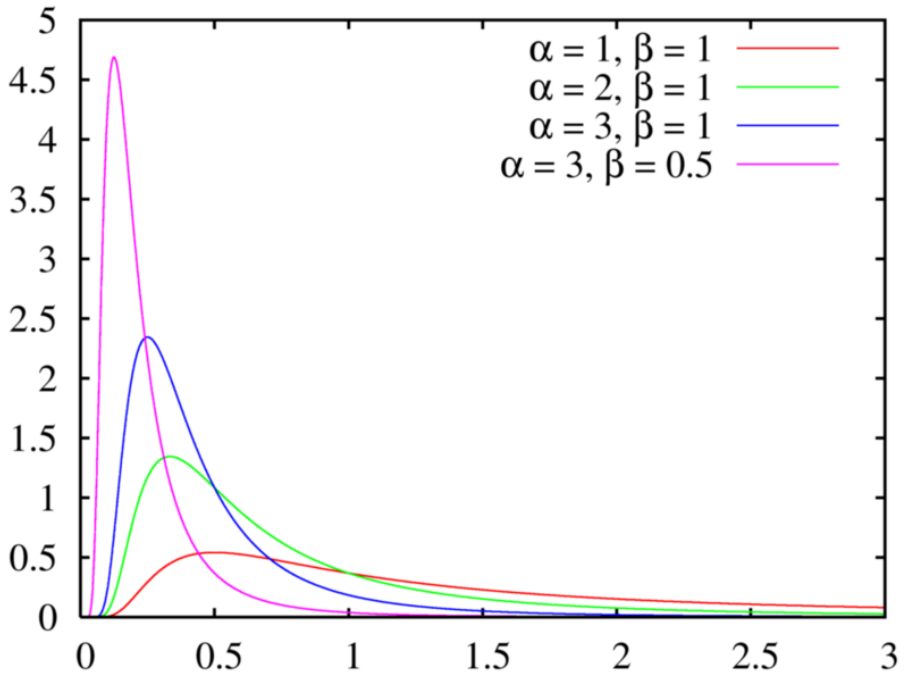


FIGURE 2 – La distribution Gamma (source : www.statisticshowto.com/gamma-distribution/)

En tenant compte de la distribution des données, on peut maintenant les modéliser.

Remarque : Dans ce qui suit, on étudie la métrique d'intuitivité qui suit la distribution Gamma comme exemple, mais tout le processus est généralisable à d'autres métriques et d'autres distributions.

3 Modélisation des métriques

Les approches classiques en statistique (estimateurs ponctuels tels que MLE, MAP, etc) ne permettent pas de modéliser l'incertitude épistémique induite par le nombre limité d'échantillons. Pour ce faire il est nécessaire de recourrir à un paradigme statistique plus puissant qui est l'*inférence bayésienne*.

3.1 Inférence bayésienne

L'apprentissage bayésien tient compte de l'incertitude épistémique en modélisant les paramètres comme des variables aléatoires. La distribution de ces variables paramètres est mise à jour à chaque fois que de nouvelles données nous parviennent.

Le cadre de l'inférence bayésienne est basé sur la formule de Bayes pour faire la mise à jour de la distribution des paramètres θ du modèle lors de l'observation de données $\mathcal{D}$:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta).p(\theta)}{p(\mathcal{D})}$$

où :

- $\mathcal{D} = \{t_1, t_2, ..., t_N\}$ où t_i est l'observation. Dans notre exemple, ce sont les valeurs d'intuitivité collectées des utilisateurs.
- $p(\mathcal{D} \mid \theta)$: *fonction de vraisemblance* (likelihood) du modèle. Cette fonction définit la distribution théorique de la donnée $\mathcal{D}$ observable, conditionnellement aux paramètres θ. Cette fonction découle d'un choix de modélisation, i.e. une distribution gamma dans le cas de la métrique d'intuitivité.
- $p(\theta)$: *distribution a priori* représentant l'état de croyance sur les distributions des paramètres avant observation.
- $p(\theta \mid \mathcal{D})$: *distribution a posteriori*, i.e. après la nouvelle observation, résultat de l'inférence.
- $p(\mathcal{D})$: constante de normalisation (Model evidence)

En général la distribution a priori est une distribution paramétrique $p(\theta) = g_\kappa(\theta)$ dont les paramètres κ sont appelés *hyper-paramètres* afin de les distinguer des paramètres θ du modèle.

Cette formule peut être appliquée de manière itérative pour permettre une inférence "online" et mettre à jour les distributions de manière incrémentale, au fur et à mesure qu'on reçoit des nouvelles données. Cet aspect online de l'inférence bayésienne est intéressant dans le cadre de ce projet vu qu'on a intérêt à pouvoir améliorer notre modèle à chaque fois qu'on reçoit de nouvelles données utilisateurs et en temps réel. On peut donc incorporer incrémentalement ces nouvelles observations dans le modèle. Dans ce contexte, les nouvelles données sont intégrées au modèle existant pour mettre à jour les distributions de probabilité des paramètres du modèle.

3.1.1 Choix du prior

Dans notre cas, notre vraisemblance suit la loi Gamma comme démontré dans la section 2 et donc $\theta = (\alpha, \beta)$. Il est important de remarquer que l'utilisation itérative de la formule de Bayes peut facilement engendrer des calculs compliqués. Ceci dépend du choix du prior, c'est là où l'avantage de choisir un *prior conjugué* (conjugate prior) à la vraisemblance apparaît.

Définition 2. *Une* distribution a priori *est dite conjuguée relativement à la fonction de vraisemblance considérée si son application dans la formule de Bayes produit une distribution a posteriori qui appartient à la même famille de distribution paramétrique que l'apriori.*

Dans ce cas, la formule d'inférence se simplifie et se résume à une mise à jour de la valeur des hyper-paramètres κ selon une certaine formule.

Dans notre cas spécifique, on cherche donc une distribution a priori conjuguée à la distribution Gamma. Une formule pour un tel prior a été développée par Miller(1980) :

$$p(\theta) \propto \frac{p^{\alpha-1}e^{-\beta q}}{\Gamma(\alpha)^r \beta^{-\alpha s}}$$

Les formules de mise à jour des hyper-paramètres p, q, r, s sont donc :

$$\begin{cases} p & \leftarrow p \times \prod_{i=1}^{n} x_i \\ q & \leftarrow q + \sum_{i=1}^{n} x_i \\ r & \leftarrow r + n \\ s & \leftarrow s + n \end{cases}$$

Démonstration. En choisissant ce prior, on trouve que, après une observation x, le postérior trouvé a la distribution suivante :

$$p(\theta \,|\, x) = \frac{\beta^{\alpha}\, x^{\alpha-1}\, e^{-\beta x}}{\Gamma(\alpha)} \times \frac{p^{\alpha-1}\, e^{-\beta q}}{\Gamma(\alpha)^r\, \beta^{-\alpha s}} = \frac{(p \times x)^{\alpha-1}\, e^{-\beta\,(q+x)}}{\Gamma(\alpha)^{r+1}\, \beta^{-\alpha\,(s+1)}}$$

□

3.2 Echantillonnage

Pour pouvoir implémenter cette approche, on a besoin de pouvoir échantillonner de la distribution considérée pour le prior.
Mais notre distribution choisie n'est pas facile à échantillonner analytiquement. Donc on a besoin d'utiliser un algorithme d'échantillonnage approximatif. On choisit l'algorithme connu de Metropolis Hastings :

Algorithme de Metropolis-Hastings

Entrée : État initial $x^{(0)}$, nombre d'itérations T.

1. Initialiser $x^{(0)}$ de manière arbitraire.
2. Pour $t = 1, 2, \ldots, T$:
   (a) Tirer une proposition y à partir d'une distribution de proposition $Q(x^{(t-1)}, \cdot)$.
   (b) Calculer la probabilité d'acceptation :

$$\alpha(x^{(t-1)}, y) = \min\left\{1, \frac{p(y)Q(y, x^{(t-1)})}{p(x^{(t-1)})Q(x^{(t-1)}, y)}\right\}.$$

   (c) Tirer u à partir d'une distribution uniforme sur $[0, 1]$.
   (d) Si $u < \alpha(x^{(t-1)}, y)$:

— Accepter la proposition : $x^{(t)} = y$.

(e) Sinon :

— Rejeter la proposition : $x^{(t)} = x^{(t-1)}$.

Cet algorithme est une méthode de simulation qui permet de tirer des échantillons iid [1] d'une distribution cible p donnée.
L'idée principale de l'algorithme est de construire un modèle de marche aléatoire sous la forme d'une chaîne de Markov telle que sa distribution stationnaire, atteinte au bout d'un nombre infini de sauts, corresponde à la distribution cible p. À chaque déplacement de la marche aléatoire, un nouvel échantillon est obtenu à partir du précédent auquel on applique un "saut" selon une certaine distribution Q, puis accepté ou rejeté en fonction d'une règle d'acceptation probabiliste basée sur la probabilité de transition entre l'état actuel et le nouvel état. Cette règle d'acceptation est définie de telle sorte à garantir que la distribution asymptotique de la marche aléatoire corresponde à p.

3.3 Implémentation et expérimentation sur Python

Pour implémenter la partie d'échantillonnage, on a besoin de choisir la distribution de proposition Q de l'algorithme de Metropolis-Hasting. On choisit une distribution normale centrée de dimension 2 du fait de sa simplicité et de son caractère symétrique autour du zéro (déplacement nul).
Afin d'éviter des problèmes de calcul (les valeurs deviennent vite très grandes ou très petites, ce qui entraine des erreurs d'overflow/underflow), il est nécessaire de remplacer certaines expressions dans l'algorithme par leurs logarithmes (distribution jointe ainsi que le $log(p)$).
Les résultats qu'on a obtenu à partir de l'inférence bayésienne sont obtenus dans la figure 3.

Cette figure montre la convergence progressive des distributions de α et β au fur et à mesure qu'on ajoute des données à notre modèle. On commence avec des courbes réparties et on finit par des courbes presque superposées, une fois qu'on a pris en compte toutes les données.

1. Indépendants et Identiquement Distribués

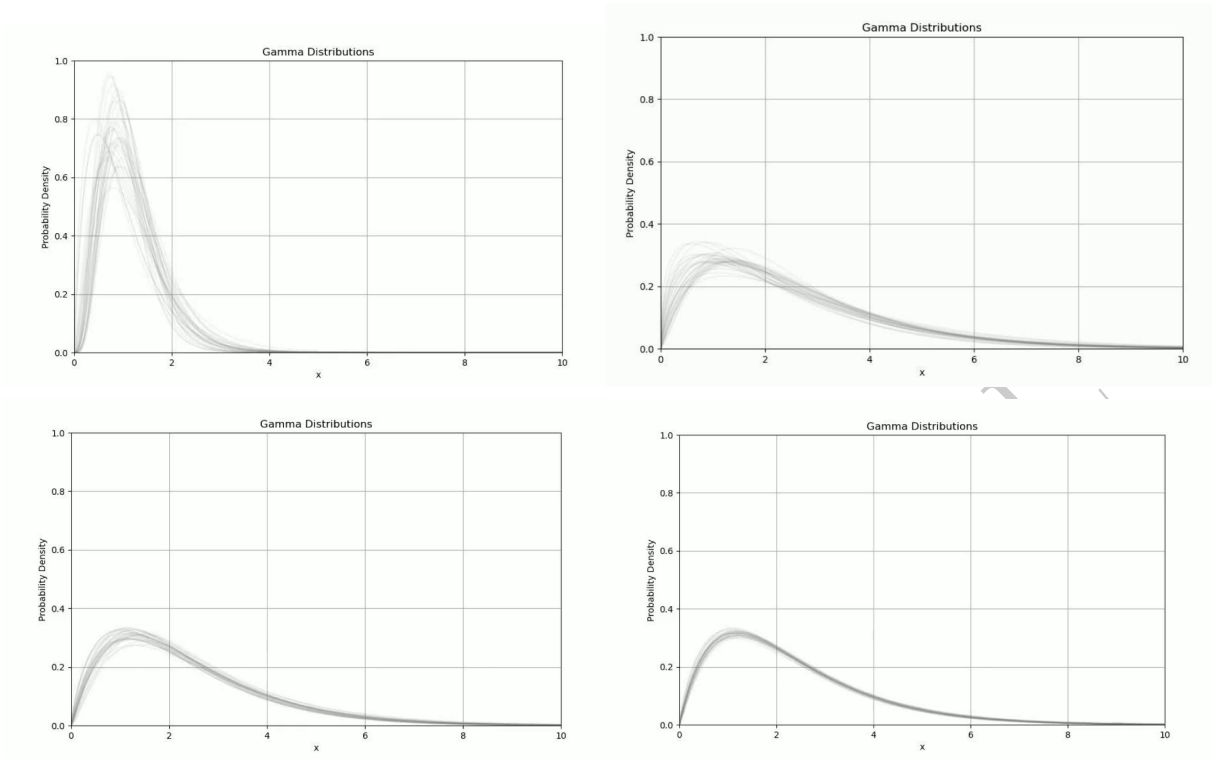


FIGURE 3 – Résultats de convergence de l'inférence bayésienne

4 Intervalles de confiance

4.1 Aspects théoriques

Maintenant qu'on a pu modéliser les données, on peut passer aux intervalles de confiance.

Définition 3. *:L'*intervalle de confiance *est une plage de valeurs susceptibles d'inclure la valeur du paramètre pris en compte avec un certain degré de confiance.*

On a besoin de calculer les intervalles de confiance sur les scores des métriques. Pour faire ça, on tient compte uniquement de la variabilité sur $\theta = (\alpha, \beta)$. C'est donc l'incertitude épistémique due au manque de données.
Le score doit donc être vu comme fonction de θ et non de la métrique T, selon la relation :

$$S = f(T_{median}|T \sim Gamma(\theta)) = F_S^{-1}(0.5|\theta)$$

L'intervalle de confiance est centré sur l'espérance de S lorsqu'on marginalise sur θ.
Pour pouvoir faire ça, il faut trouver la fonction de répartition de la fonction de score.
Pour faire ça, on utilise les résultats de l'inférence bayésienne faite sur la métrique correspondante pour générer les courbes de la fonction de répartition du score F_S.
On utilise donc la relation suivante, compte tenu que le score est une fonction décroissante de l'intuitivité :

$$F_S(s) = P(S \leq s) = P(f(T) \leq s) = 1 - P(T \leq f^{-1}(s)) = 1 - F_T(f^{-1}(s))$$

— f étant la fonction de score continue.

— S et T étant les variables aléatoires modélisant respectivement le score et la métrique.
— F_S et F_T étant les fonctions de répartition correspondantes à S et à T respectivement.
— s étant une valeur réelle.

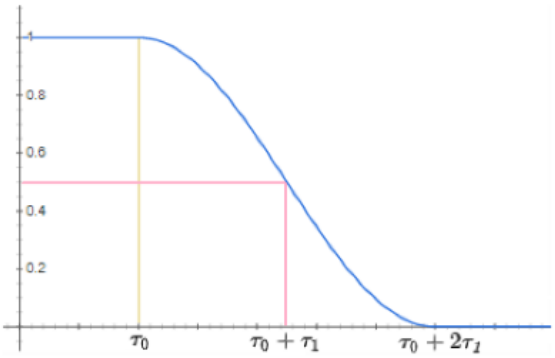
On considère dans notre exemple la fonction de score correspondante à la métrique d'intuitivité :

Formule

- $Score_{t1} = \mathrm{f}(t_{1med})$
  - si $t_{1med} \leq \tau_0, \mathrm{f}(t_{1med}) = 1$
  - si $\tau_0 < t_{1med} \leq \tau_0 + 2\tau_1$,

$$\mathrm{f}(t_{1med}) = \frac{1}{2}\left(1 + \cos\left(\frac{\pi}{2\tau_1}(t_{1med} - \tau_0)\right)\right)$$

  - si $t_{1med} > \tau_0 + 2\tau_1, \mathrm{f}(t_{1med}) = 0$



4.2 Implémentation et expérimentation sur Python

4.2.1 Modélisation du score

Pour faire ça, on échantillonne des valeurs (α, β) du modèle à chaque itération de l'inférence bayésienne.Sachant que la métrique suit la distribution $\Gamma(\alpha, \beta)$, on suit les étapes suivantes, pour chaque couple (α, β) échantillonné :

1. On prépare la liste des valeurs de score , uniformément répartis entre 0 et 1
2. Pour chaque valeur du score, on trouve la valeur de la métrique T correspondante en utilisant la fonction inverse de la fonction du score. Pour cela, on suppose que la fonction du score est une fonction continue, et que la formule du score est valable partout et pas seulement sur t_{median}.
3. Avec les valeurs de la métrique, on calcule la fonction de répartition correspondante aux valeurs de T, en utilisant la fonction de répartition de la loi Gamma, ayant pour paramètres les valeurs échantillonnées de (α, β).
4. Maintenant, en associant la liste des valeurs du score avec la dernière liste obtenue, on obtient la courbe de répartition de la fonction de score.

Pour pouvoir appliquer la procédure mentionnée, on suppose que si t est entre τ_0 et $\tau_1 + \tau_0$, l'expression de la fonction du score est applicable directement à t (pas t_{median}).
On obtient, après convergence, la fonction de répartition obtenue sur la figure 4.

La figure 4 montre que une concentration prévue des scores en 0 et en 1.

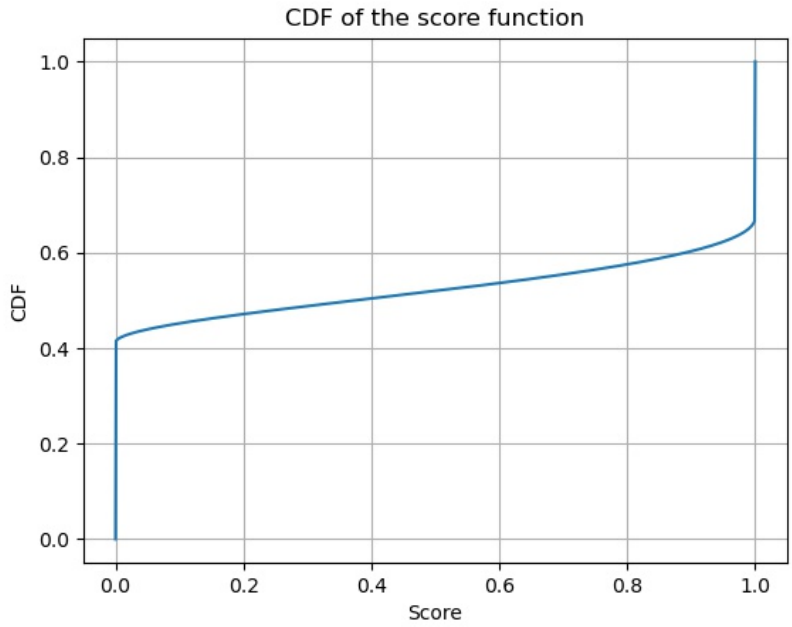


FIGURE 4 – Fonction de répartition du score après convergence

4.2.2 Intervalle de confiance

On prend donc N échantillons $\theta_i = (\alpha_i, \beta_i)$. Pour chaque échantillon, on déduit le score

$$s_i = F_S^{-1}(0.5|\theta_i)$$

Enfin, on peut calculer s_{moy} comme suit :

$$s_{moy} \simeq \frac{1}{N}\sum_{i=1}^{N} s_i$$

On peut donc obtenir l'intervalle de confiance $[s_{min}, s_{max}]$ en triant par ordre croissant les s_i dans un tableau s_{table} :

$$\begin{cases} s_{min} = s_{table}[i_{esp} - \delta_i + 1] \\ s_{max} = s_{table}[i_{esp} + \delta_i] \end{cases}$$

où i_{esp} et δ_i vérifient :

$$\begin{cases} s_{table}[i_{esp}] < s_{moy} \leq s_{table}[i_{esp} + 1] \\ \delta_i = \frac{(1-\alpha)}{2N} \end{cases}$$

On trace l'évolution de l'intervalle de confiance à $\alpha = 0.1$ du score en fonction du nombre de données prises en compte lors de l'inférence bayésienne et on obtient la figure 5.
On remarque que même avec un nombre important d'échantillons pris en compte, la valeur moyenne du score continue à évoluer et on n'arrive pas à bien estimer sa valeur probable. Si on trace la bande virtuelle de l'intervalle de confiance final horizontalement sur toute

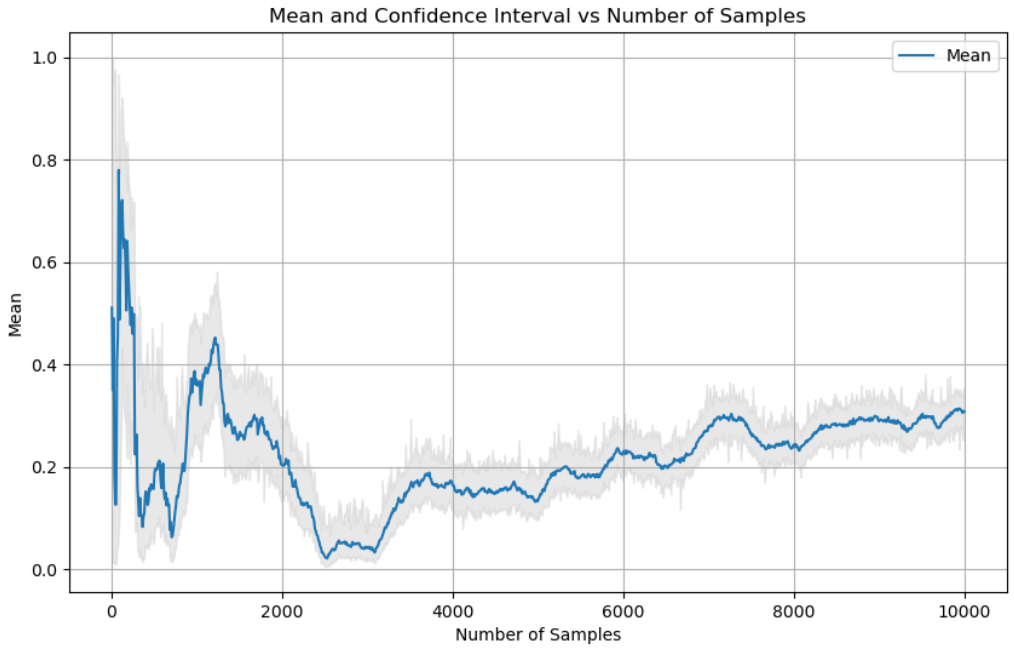


FIGURE 5 – Evolution de l'intervalle de confiance en ajoutant des données à l'inférence bayésienne

la figure, les autres valeurs sont théoriquement censés être dans cette bande avec une probabilité de $0,9$. Mais, c'est évident de la figure 5 que ce n'est pas le cas.
Ceci peut être dû au fait que lors du temps de collecte des échantillons, le site a eu plusieurs changements, ce qui a fait varier la valeur moyenne autant. La distribution des données n'est donc pas stationnaire.
Pour tester cette hypothèse, on essaie de forcer la stationnarité en permutant les données au début. On obtient donc la figure 6.

Les résultats obtenus cette fois sont beaucoup mieux et on voit bien qu'à partir d'environ 6000 échantillons, on a déjà la valeur moyenne probable et l'intervalle de confiance du score qui reste presque constant après.
En pratique, cette procédure de permutation ne peut pas être utilisée vu qu'on receuille les données au fur et à mesure dans le temps, contrairement à notre cas où on a déjà toutes les données à la fois.
Quelques solutions sont possibles pour remédier à ce problème de stationnarité :

- on peut par example utiliser des fenêtres glissantes les sliding windows pour détecter les changements des populations
- on peut attacher des poids aux données en fonction de leur ancienneté pour forcer le modèle à oublier les données les plus anciennes.

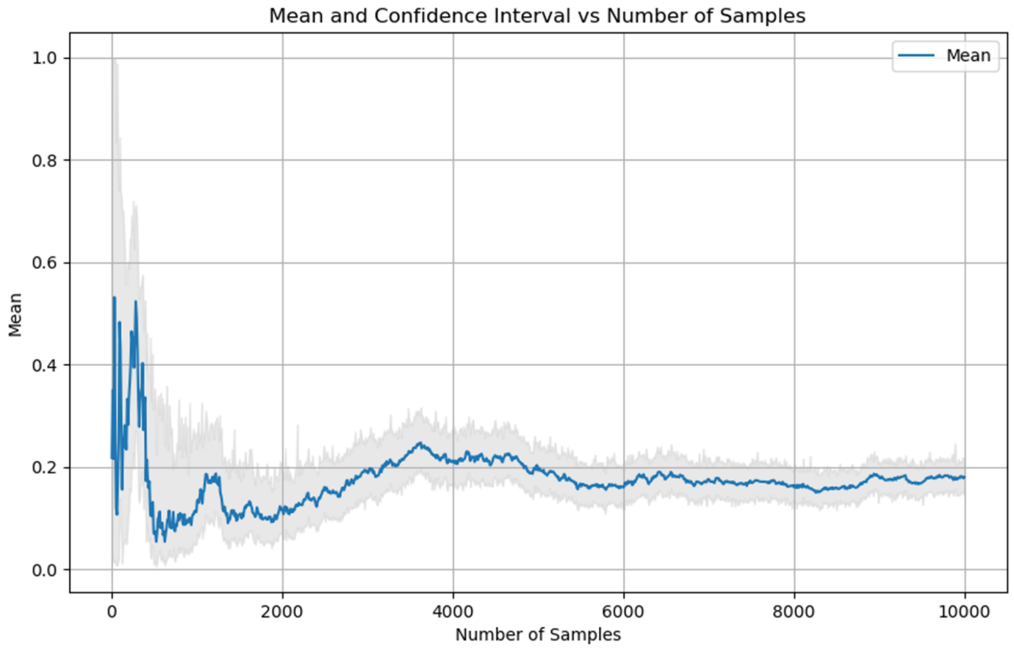


FIGURE 6 – Evolution de l'intervalle de confiance en ajoutant des données à l'inférence bayésienne après permutation des données

5 Score de vraisemblance

5.1 Introduction théoriques

La vraisemblance d'un ensemble de données sous un modèle statistique est une mesure de la probabilité que le modèle génère les données observées. Pour un modèle donné, évaluer la vraisemblance implique souvent de calculer des intégrales de haute dimension, ce qui peut être complexe et coûteux en termes de calcul. L'estimateur de la moyenne harmonique offre une alternative en utilisant une approximation basée sur des échantillons tirés du modèle.

Description mathématique Considérons un modèle gamma paramétré par la densité $f(\alpha, \beta)$. Pour un ensemble de données, la vraisemblance logaritmique serait la somme des logarithmes des densités de chaque observation. Cependant, l'estimateur de la moyenne harmonique utilise des échantillons du modèle pour approximer cette somme.

5.2 Introduction à l'estimateur de la moyenne harmonique

Dans le contexte des méthodes bayésiennes, l'évaluation précise de la vraisemblance peut devenir un défi, surtout lorsque les modèles deviennent complexes ou que les données sont de grande dimension. L'estimateur de la moyenne harmonique offre une solution alternative pour estimer l'intégrale de la vraisemblance sur la distribution postérieure des paramètres, en évitant les difficultés liées aux calculs directs.

5.3 Principe de l'estimateur de la moyenne harmonique

L'estimateur de la moyenne harmonique dans le cadre bayésien est utilisé pour approximer l'intégrale de la vraisemblance inverse sur la distribution postérieure. Le principe repose sur l'idée que l'inverse de la moyenne de l'inverse de la vraisemblance peut être une approximation utile de l'intégrale nécessaire pour le calcul bayésien. Cette méthode est particulièrement utile lorsque la vraisemblance est difficile à intégrer directement à cause de sa forme ou de sa dépendance envers les paramètres de haute dimension.

5.4 Formulation mathématique

Mathématiquement, l'estimation de la vraisemblance $p(x)$ pour des données x est donnée par :

$$p(x) = \int p(x|\theta)p(\theta)d\theta$$

où $p(x|\theta)$ est la vraisemblance des données étant donné les paramètres θ, et $p(\theta)$ est la distribution a priori des paramètres.
L'estimateur de la moyenne harmonique est calculé comme :

$$\frac{1}{\hat{p}(x)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{p(x|\theta_i)}$$

où θ_i sont des échantillons tirés de la distribution postérieure $p(\theta|x)$. Cette formule montre que l'estimateur est l'inverse de la moyenne des inverses des vraisemblances calculées pour chaque échantillon de la distribution postérieure.

5.5 Avantages et limitations

Avantages :
— **Praticabilité** : Utile pour des modèles où le calcul direct est infeasible.
— **Simplicité d'implémentation** : Relativement facile à mettre en œuvre avec des échantillonnages de Monte Carlo.

Limitations :
— **Sensibilité aux valeurs extrêmes** : Peut être très influencé par les échantillons où la vraisemblance est extrêmement basse.
— **Biais** : Tendance à sous-estimer la vraisemblance, surtout avec un nombre insuffisant d'échantillons.

5.6 Implémentation sur Python

5.6.1 Échantillonnage des paramètres

La classe Likelihood implémente cette approche en générant des échantillons des paramètres α et β du modèle gamma. Ces échantillons sont utilisés pour calculer la vraisemblance logaritmique de chaque échantillon par rapport aux données.

5.6.2 Calcul de la vraisemblance

Le calcul de la vraisemblance est ajusté pour la stabilité numérique en soustrayant le maximum des vraisemblances logaritmiques calculées (s_max) de chaque log-probabilité avant d'appliquer l'exponentielle. Cela évite les problèmes de sous-débordement numérique lors de l'exponentiation de très petits nombres.

5.6.3 Normalisation et moyenne

La vraisemblance finale est obtenue en prenant la moyenne logarithmique des probabilités ajustées et en la re-normalisant par le nombre d'échantillons pour obtenir une estimation de la vraisemblance par observation.

5.7 Résultats

5.7.1 Données générées

Pour évaluer l'utilité de cet estimateur de la moyenne harmonique, nous avons réalisé plusieurs expériences. Nous avons généré des données aléatoires suivant une distribution gamma, ainsi que d'autres ensembles de données suivant différentes distributions, telles qu'une loi uniforme. Nos résultats montrent que, systématiquement, l'estimateur fournit une log-vraisemblance significativement meilleure pour les données suivant la distribution gamma. Cette observation confirme l'efficacité de l'estimateur à discriminer entre différentes distributions en fonction de leur adéquation avec le modèle gamma.

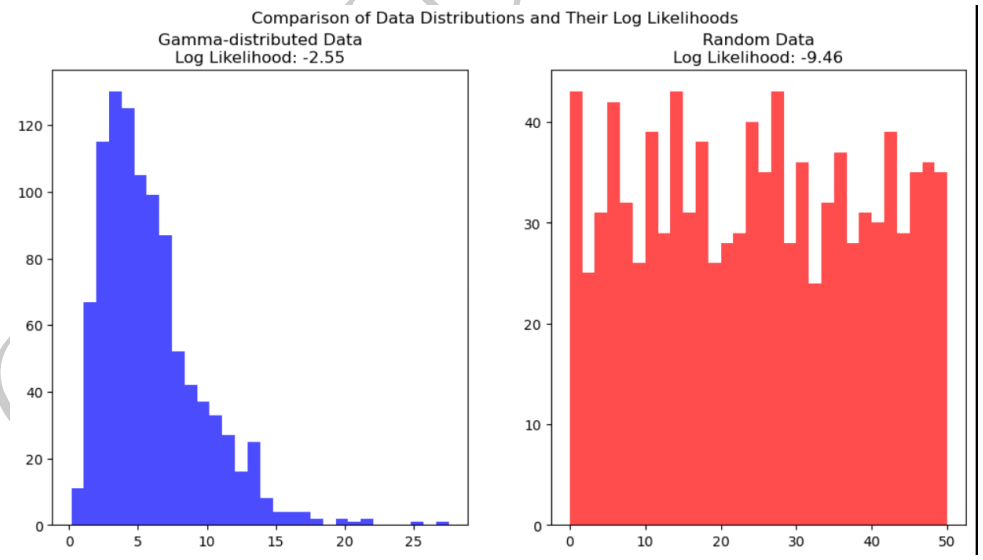


FIGURE 7 – Comparaison de la likelihood d'une donnée qui suit la loi gamma et d'une autre donnée aléatoire

5.7.2 Données réelles

Pour tester l'efficacité de l'estimateur de la moyenne harmonique sur des données réelles, nous avons appliqué notre algorithme de statistique bayésienne pour identifier la

distribution gamma qui correspond le mieux à chaque jeu de données. Une fois cette distribution optimale identifiée, nous avons utilisé l'estimateur de log-vraisemblance pour évaluer la conformité des données à cette distribution. Nous avons observé que les données qui présentaient initialement des caractéristiques similaires à celles d'une distribution gamma affichaient une excellente log-vraisemblance. Cette constatation confirme la pertinence de l'estimateur de la moyenne harmonique pour valider la qualité de l'ajustement d'une distribution gamma aux données réelles, en particulier lorsque les données semblent suivre cette forme de distribution

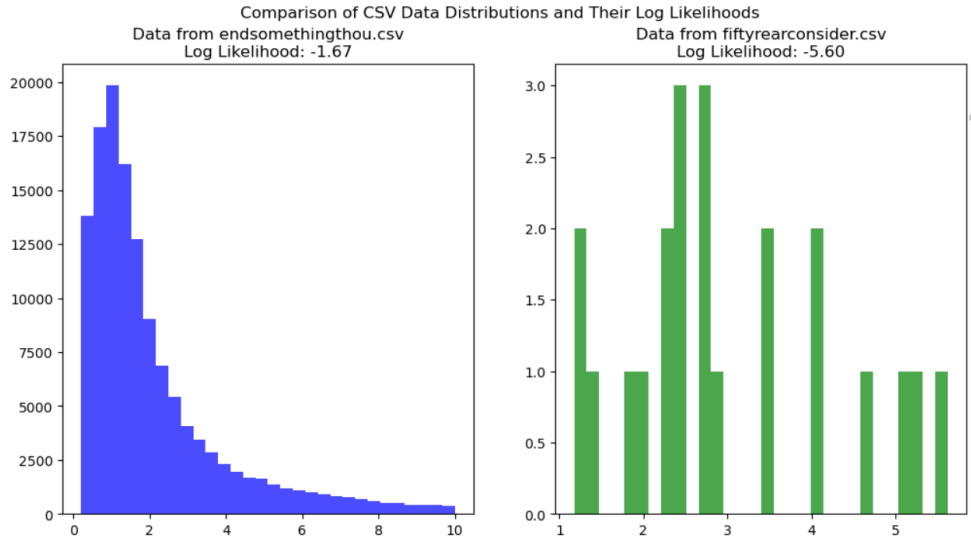


FIGURE 8 – Comparaison de la likelihood d'une donnée qui suit la loi gamma et d'une autre donnée atypique

6 Conclusion

Dans ce projet, l'inférence bayésienne a été utilisée pour modéliser l'incertitude épistémique concernant les métriques et, par conséquent, les scores. Cette approche nous a permis de traiter les paramètres des distributions des métriques comme des variables aléatoires, augmentant ainsi la flexibilité et l'adaptabilité de notre modèle. En ajustant les intervalles de confiance et en intégrant progressivement de nouvelles données utilisateurs, nous avons optimisé notre modèle. Cela nous a aidés à déterminer le nombre d'échantillons nécessaire pour estimer les scores des pages web et à établir un intervalle probable de variation. L'évaluation du score de vraisemblance a confirmé la compatibilité de notre modèle avec les données observées.

Il reste encore quelques pistes à explorer à l'issue de ce projet :

— Une piste intéressante, abordée dans la section 4.2.2, est celle de la segmentation des données pour tenir compte des variations des sites, ce qui pourrait résoudre les problèmes de non-stationnarité lors du calcul des intervalles de confiance.

— Une autre direction concerne le choix du modèle des données, évoqué en section 2. L'analyse de diverses métriques pour plusieurs LOMs montre que les distributions ne sont pas toujours clairement identifiables. Parfois, il existe des distributions

complexes qui nécessitent l'utilisation de modèles de mélange, comme le modèle de mélange gaussien (GMM).

— Concernant le calcul du score de vraisemblance, il serait bénéfique de tester l'algorithme sur d'autres bases de données, y compris celles qui suivent la loi gamma et d'autres données atypiques, afin de déterminer le seuil de vraisemblance approprié et de s'assurer de la robustesse du code.

Remarque : Ce rapport est accompagné du code complèt avec les implémentations cités.