

M1-ISL, Initiation à la Recherche. TD. Construire un document computationnel.

Jean-François COUCHOT
`couchot [arobase] femto-st [point] fr`

12 janvier 2025

L'objectif de ce TP est de construire un document computationnel de recherche lié au TD. Voici un guide méthodologique, mais on pourra s'en écarter si on trouve des éléments plus pertinents à expliquer, illustrer.

1 Création du notebook jupyter

Par binôme créer votre un jupyter notebook nommé `nom1_nom2_RR_etendue`. C'est ce notebook qui sera déposé sur moodle.

Permettre que si je réexécute votre code, j'obtienne exactement vos résultats en gérant l'aléa inhérent aux réponses randomisées.

2 Contenu du notebook

Après avoir défini un titre, dans une cellule de texte, résumer le travail qui sera présenté dans le jupyter notebook, son contexte (ce TP en M1 ISL, initiation à la recherche) et présenter le plan. Il serait judicieux de faire une section dans ce notebook par section de ce TP.

Dans tout ce qui suit, on considère une fréquence initiale f de réponses positives à une question embarrassante qui varie entre 0 et 1 et on considère que N est le nombre de personnes qui répondent à ce sondage (qui varie entre 10 et 100000).

2.1 Evaluations reproductibles de l'algorithme de RR de Warner

1. Dans une cellule de texte, donner la démonstration de la valeur de la fréquence \hat{f} estimée de réponses positives lorsque les données sont nettoyées selon le processus de RR.
2. Afficher une figure l'histogramme 3D de la fréquence \hat{f} estimée de réponses positives (en z) lorsque les données sont nettoyées selon le processus de RR. Cet affichage sera la moyenne de 10 expérimentations.
3. Afficher dans une seconde figure l'histogramme 3D de l'écart-type de la fréquence \hat{f} estimée de réponses positives lorsque les données sont nettoyées selon le processus de RR. Cet écart type sera issu des données précédentes.
4. Dans une cellule de texte, donner la démonstration de la valeur théorique de l'écart type de la fréquence \hat{f} .
5. Afficher dans une troisième figure l'histogramme 3D de l'écart-type théorique de la fréquence \hat{f} estimée de réponses positives lorsque les données sont nettoyées selon le processus de RR.
6. Dans une cellule de texte, interpréter et analyser ces deux figures.

2.2 Evaluations reproductibles de variations de RR de Warner : calcul de l'écart-type

Dans ce qui suit, on considère les variables p_1 et p_2 avec le même sens que dans le TD.

1. Dans une cellule de texte, donner la démonstration de la valeur de la fréquence \hat{f}' estimée de réponses positives lorsque les données sont nettoyées selon le processus de RR étendue en suivant les probabilités p_1 et p_2 .

2. Dans une autre cellule de texte, donner la démonstration de la valeur théorique de l'écart-type de la fréquence \hat{f}' .
3. On cherche à représenter le fait que la formule théorique d'écart-type est correcte. Pour cela, on pourra construire une figure contenant un histogramme en 3D où :
 - la fréquence f (entre 0 et 1) de réponses positives à une question embarrassante sera représentées sur l'axe des abscisses ;
 - le nombre N de personnes qui répondent à ce sondage (entre 10 et 100000) sera représentées sur l'axe des ordonnées ; on pourra prendre une échelle logarithmique ;
 - pour chaque paire (p_1, p_2) dans $]0, 1[\times]0, 1[$ (on se limitera d'abord à 10 valeurs pour p_1 et pour p_2 , puis on en prendra 100), on réalisera 5 expérimentations de nettoyage et de calcul de l'estimation \hat{f}' dépendant de p_1, p_2, f et N . L'écart-type sera alors calculé (sur cette série de 5 valeurs) et la valeur absolue de la différence entre celui-ci et la valeur théorique de cet écart-type sera mémorisée. On affichera en z le maximum de cet écart absolu (sur les 10 000 valeurs).
4. Analyser les résultats présentées dans cette figure dans une cellule de texte. On devrait notamment conclure que la formule théorique est "juste".

2.3 Formalisation de l'utilité de cette variation de Warner

Dans ce qui suit, on considère les variables p_1 et p_2 avec le même sens que dans le TD.

1. Dans une cellule de texte, montrer que cette variation est plus utile que l'algorithme original de Warner si et seulement si l'inéquation suivante est établie

$$3p_1^2 - 4f.p_1.(1 - 2p_2).(1 - p_1) - 4p_2.(1 - (1 - p_1)p_2).(1 - p_1) > 0 \quad (1)$$

2. Montrer dans une cellule de texte que pour être aussi respectueux de la vie privée, il est nécessaire que les deux contraintes suivantes soient simultanément établies :

$$p_1 + (1 - p_1)p_2 \leq 3(1 - p_1)p_2 \quad (2)$$

$$p_1 + (1 - p_1)(1 - p_2) \leq 3(1 - p_1)(1 - p_2) \quad (3)$$

3. Donner le code du programme qui permet de trouver des solutions au système composé des équations (1), (2) et (3).
4. Discuter de l'absence de solutions à ce système et donner une conclusion.

3 A rendre

Dans l'espace moodle associé au cours, déposer votre notebook jupiter avant le 18/01/2025 à 18h.