M1-ISL
Option Initiation à la Recherche. TD1

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

16 janvier 2022

1 Algorithme de réponse randomisée

On considère la méthode de réponse randomisée [1] qui permet de répondre de manière "anonyme" à une question sur des données sensibles déjà vue en CM.

Exercice 1.1 (Code implantant le nettoyage de la réponse). *Donner le code (3 lignes tout au plus) de la fonction* `reponse_randomisee(rep_o)` *qui prend en paramètre une réponse originale booléenne et qui retourne une réponse du même type selon l'algorithme de réponse randomisée original (avec des pièces de monnaie) donné dans le cours.*

Exercice 1.2 (Code implantant l'estimateur). *Donner le code (3 lignes tout au plus) de la fonction* `estimation_nbre_oui_initiaux(reps)` *qui prend en paramètre la série de réponses nettoyées et retourne une estimation du nombre de* OUI *qui étaient présents initialement.*

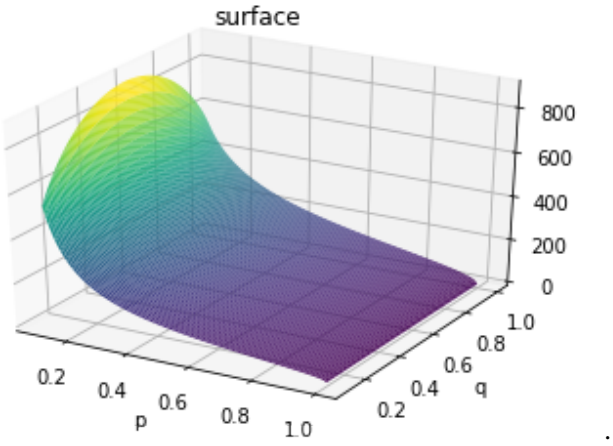
Exercice 1.3 (Evaluation numérique statistique du mécanisme).

1. *Commencer par donner le code de la fonction* `eval_rep_randomisee(nb_exp,data)` *qui prend en paramètres le nombre* `nb_exp` *et la séquence de valeurs binaires originales* `data`. `nb_exp` *fois, le jeu de données* `data` *est nettoyé selon l'algorithme de réponse randomisée puis le nombre de* OUI *qui étaient présents initialement est estimé. Cette estimation est accumulée dans une liste qui est retournée.*
2. *Donner le code qui calcule la moyenne, la variance/l'écart-type de cette liste.*
3. *On considère le jeu de données aldult. La donnée sensible est le fait qu'une personne soit vendeuse (Occupation = Sales). Évaluer la démarche de réponse randomisée sur cette question en répétant 100 fois l'expérimentation.*

Exercice 1.4 (Extension de l'algorithme).

1. *Définir le nouvel estimateur $\hat{f}'$ de la réponse randomisée qui repose sur les probabilités p de dire la vérité et q de sortir un* OUI *si le tirage ne contraint pas de dire la vérité.*
2. *Modifier le code précédent pour qu'il prenne en compte cette extension.*
3. *On considère que $p \in [0.1, 0.2, \ldots, 1.0]$ et $q \in [0.1, 0.2, \ldots, 1.0]$. Pour chaque paire (p, q), estimer pratiquement l'écart-type de $\hat{f}'$ avec 5 évaluations.*
4. *Illustrer ceci à l'aide d'un histogramme en 3D, comme donné à la figure 1.4. On pourra s'inspirer de la référence* [2]
5. *Quels semblent être les bons candidats (p, q) pour minimiser l'écart-type ?*

1. Warner, S. L. (1965). Randomized response : A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60(309), 63-69.
2. https://jakevdp.github.io/PythonDataScienceHandbook/04.12-three-dimensional-plotting.html



2 Aspects théoriques

Exercice 2.1 (Premières preuves).

1. *Montrer que l'estimateur est non biaisé. Pour cela, montrer que* $E[\hat{(}f)] = f$.
2. *Estimer théoriquement la variance de l'estimateur :*
   (a) *On rappelle que la variance est définie comme la moyenne des carrés des écarts à la moyenne. Montrer que pour une VAR* X, $V[X] = E[X^2] - E[X^2]$.
   (b) *On considère un utilisateur dont la réponse originale à la question est* OUI. *Soit* X *la VAR qui vaut 1 si sa réponse randomisée est* OUI *et 0 sinon. Évaluer* $E[X]$, $E[X^2]$. *En déduire que* $V[X] = \delta_1(1-\delta_1)$ *avec* $\delta_1 = p + (1-p)q$.
   (c) *De manière similaire, on considère un utilisateur dont la réponse originale à la question est* NON. *Soit* Y *la VAR qui vaut 1 si sa réponse randomisée est* OUI *et 0 sinon. Montrer que* $V[Y] = \delta_0(1-\delta_0)$ *avec* $\delta_0 = (1-p)q$.
   (d) *Soit* R *VAR qui compte le nombre de réponse randomisées égales à* OUI. *Montrer que* $V[R] = f.V[X] + (N-f).V[Y]$.
   (e) *Montrer alors que* $V[\hat{f}] = \frac{V[R]}{p^2}$.

Exercice 2.2 (Contredire la pratique en intégrant la PVP).

1. *Dans ce contexte d'extension, quels sont les rapports de probabilités maximaux suivants ?*

$$\frac{\Pr[\mathcal{M}(x_1) = y]}{\Pr[\mathcal{M}(x_2) = y]} \text{ pour } x_1, x_2, r \in \{\text{OUI}, \text{NON}\}$$

2. *On souhaite garantir que ce mécanisme ne dévoilera pas plus d'information que celui de réponse randomisée initial. Montrer que ceci nécessite que les deux contraintes suivantes soient simultanément établies :*

$$\frac{p + (1-p)q}{(1-p)q} \leq 3 \qquad \frac{p + (1-p)(1-q)}{(1-p)(1-q)} \leq 3$$

3. *Montrer qu'il est nécessaire que* p *soit inférieur à* $1/2$ *pour que ce système ait une solution.*
4. *Intuitivement, discuter de la variance de l'estimateur* $\hat{f}'$ *en fonction de* p.
5. *Conclure sur l'intérêt de cette extension.*