Sécurité Appliquée-TP 3
Protection de la vie privée-PVP

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

10 décembre 2020

On considère ici à nouveau le jeu de données Adults comme au TP2 avec la même sélection d'attributs que dans ce TP.

Dans cette partie, on considère que l'on veut représenter un histogramme des statuts conjugaux qui respecte la vie privée. Un budget ϵ est arbitrairement fixé (par exemple à 0.7).

On considère qu'on dispose d'un générateur de bruit `g = truncatedTwoSidedGeoDist(p,c,mx)` qui suit une loi géométrique positive et négative de paramètre de succès $1-p$ et tronquée sur l'intervalle $[-c, mx-c]$.

Histogramme des statuts maritaux

Dans cet exercice, on considère que l'on veut représenter un histogramme qui respecte la vie privée des statuts maritaux qui sont « Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse ».

1. Récupérer les effectifs bruts correspondant à chacune des valeurs données ci-dessus.
2. On souhaite retourner le nombre de personnes de la catégorie « Married-civ-spouse », tout en respectant la (0.7,0)-DP. Quelle est la sensibilité d'une telle requête ? Implanter un premier code qui réalise cela.
3. Supposons maintenant que l'on souhaite combiner cette requête avec celles pour produire l'histogramme complet des 7 valeurs possibles. Pour avoir un epsilon global égal à 0.7, quelle doit être la valeur d'ϵ à chaque fois ? Implanter le code.
4. Implanter la fonction qui produit l'erreur entre l'histogramme retourné à l'utilisateur et l'histogramme brut. L'appliquer à la question précédente.
5. On considère maintenant une requête SQL qui retourne déjà un 7-uplet des effectifs par catégorie. Quelle est la sensibilité d'une telle requête ? Proposer alors un second algorithme de publication d'histogramme qui se base sur ce 7-uplet de réponse. Évaluer l'erreur commise et la comparer avec celle trouvée à la section précédente.