Noiret L,
A. Rogozan

TP
Statistiques Descriptives
Analyse multidimensionnelle de Données Qualitatives

Architecture des Systèmes d'Information
3ème année

La semaine dernière, nous avons vu comment mesurer la liaison entre deux variables quantitatives graphiquement (nuage de points) et à l'aide d'un indicateur (corrélation linéaire). Cette semaine, nous allons voir les outils permettant de d'explorer l'association entre variables qualitatives :

1. le tableau de contingence
2. les mesures d'association (Chi2, coefficient Phi pour deux variables binaires, coefficient de Tschuprow, coefficient de Cramer) qui se calculent à partir du tableau de contingence.

1 Les données

Les données que nous allons analyser dans ce TP sont extraites de https://github.com/caesar0301/awesome-public-datasets et concerne les caractéristiques des passagers du Titanic. La description des données originales est disponible sur cette page web : https://www.kaggle.com/c/titanic/data.
Vous pouvez récupérez ces données en téléchargeant le fichier `titanic.csv` sur Moodle.

Libellé	Signification
Survived	0 : Décédé, 1 : survivant
Pclasse	Indique la classe du passager : 1, 2 ou 3ème classe
Sex	Indique si le passager est un homme ou une femme
Age	Age du passager
SibSp	Frères/sœurs à bord : 0 : aucun, 1 : Un, 2 : plus de un
Parch	Parents à bord : 0 : aucun, 1 : Un, 2 : plus de un
Fare	Prix du billet
Embarked	Port d'embarquement : 67 : Cherbourg, 81 : Queenstown, 83 : Southampton

TABLE 1 – Données du fichier titanic-quali.csv

1. Importez les données dans une table.
2. On va commencer par créer une nouvelle variable $Sex01$ dont la valeur est 1 si le passager est une femme, 0 sinon.
   $titanic.Sex01 = strcmp(titanic.Sex, 'female'))$.
3. Quel est le type de chaque variable ?
4. Pour chaque variable qualitative, indiquer le nombre de champs et le nombre de donnés manquantes (*summary*).
5. Choisissez une représentation graphique adaptée pour les variables $SipSp$, $Embarked$ et $Fare$.
   Aide : transformer la variable $Embarked$ en variable catégorielle (voir TP précédent).

2 Tableaux de contingence

2.1 La survie selon le genre

1. Calculez le tableau de contingence sur la survie *Survived vs* le genre des passagers (*Sex01*). Profitez de la qualité binaire des données pour calculez le tableau de contingence manuellement.
   Vérifier ensuite vos calculs en utilisant une fonction de Matlab (*crosstab*).

2. Analysez ce que vous observez et proposez une hypothèse.

2.2 La survie selon la classe et le port d'embarquement

1. Calculez le tableau de contingence sur la survie *vs* la classe des passagers.
2. Calculez le tableau de contingence sur la survie *vs* le port d'embarquement.
3. Formulez des hypothèses sur vos observations.

3 Dépendance de variables

1. Implémentez (dans des fonctions) le calcul du χ^2 et des mesures issues du χ^2 :
   — Φ^2,
   — Coefficient de Tschuprow,
   — Coefficient de Cramer.
   Comment s'interprête la valeur de chaque coefficient ?
2. Pour chaque mesure , calculez un tableau présentant l'interdépendance pour l'ensemble des variables utilisées.
3. Que pensez-vous de vos hypothèses précédentes ?

4 Profils

1. Implémentez le calcul des profils-colonnes pour une paire de descripteurs de votre choix.
2. Affichez ces profils-colonnes sous forme de graphiques (voir figure 1). Mettez en relation ce que vous observez et les hypothèses que vous avez faites.

5 Tableaux de Burt

1. Calculez le tableau de Burt pour l'ensemble des variables et leurs modalités.

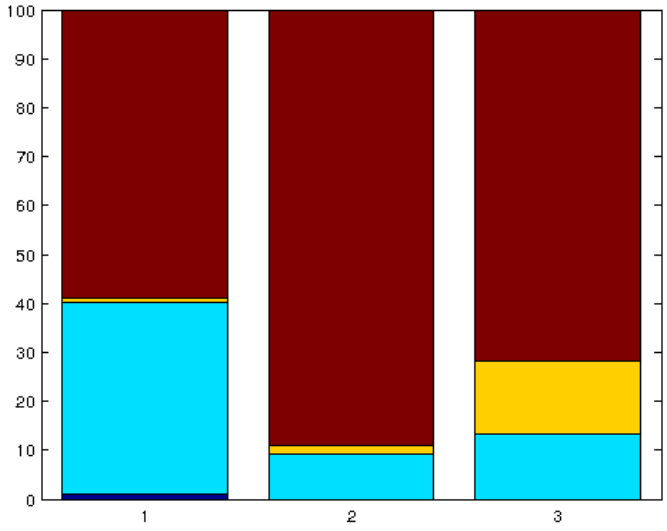

FIGURE 1 – Profils-colonnes relatifs à la table de contingence de la classe des passagers vs au port d'embarquement.

Bonus pour les plus rapides

1. Tracer l'histogramme de la variable *Age* en choisissant vous même le découpage en classes d'âges. Prenez des classes d'amplitudes égales, puis d'amplitude variables. Commentez ce que vous voyez.
2. Explorer les relations entre les variables quantitatives.