Modélisation statistique - TD7
Tests d'hypothèses et Anova

Dans ce TD, on s'intéresse aux pourboires enregistrés par un serveur dans un restaurant américain au début des années 1990. Le restaurant était dans un centre commercial. Il y avait une zone fumeurs et une zone non-fumeurs. Les données indiquent le montant total de l'addition, le pourboire, le sexe de la personne qui a payé et donné le pourboire, la zone du restaurant (fumeur et non-fumeur), le jour de la semaine et le moment de la journée (journée ou soirée). Il y a 3 colonnes intrinsèquement numériques de données : TOTBILL et TIP sont exprimées en dollars (des années 90) alors que SIZE donne le nombre de convives. SEX=0 indique un homme, SEX=1 une femme. SMOKER=0 est mis pour un non-fumeur et SMOKER=1 pour un fumeur. DAY=3 correspond à mardi, DAY=4 à mercredi, etc. TIME=0 signifie en journée, TIME=1 signifie en soirée. IDEN sert bien sûr à identifier un repas.

1 Chargement des données

1. Charger la table de données `Tips.csv`.
2. Que représente chaque colonne ? Vérifiez que la classe de chaque colonne est en adéquation avec le type de variable (qualitative ou quantitative).
3. Changer la classe des colonnes lorsque c'est nécessaire. Utiliser les fonctions `as.numeric()`, `as.factor()`, `as.character()`. Vérifier la procédure.

2 Relation entre le montant de la facture et le montant du pourboire

1. On cherche à définir la relation entre le montant de la facture et le montant du pourboire. Quels outils de statistiques descriptives connaissez-vous pour étudier le lien entre deux variables qualitatives ? Appliquer ces outils avec **R**.
2. Que concluez-vous quant à la relation entre le montant de la facture et le montant du pourboire ?
3. Dans la suite, on cherchera à savoir si les clients ont des comportements différents, selon qu'ils sont un homme ou une femme, fumeur ou non-fumeur, et selon le jour de la semaine. Proposer une nouvelle variable, plus pertinente que le montant du pourboire pour comparer ces comportements. Créer cette nouvelle variable dans la table avec l'instruction `tab$NouvelleVariable = ...`

3 Comparaison des comportements des clients

3.1 En fonction du moment de la journée

1. Un employé affirme que les clients sont plus généreux en soirée qu'en journée. Etablir un test d'hypothèse permettant de répondre à cette question. Sur quel paramètre porte ce test ? Donner les hypothèses nulle et alternative, la statistique de test ainsi que sa loi sous $\mathcal{H}_0$. Quel test allez-vous utiliser ? Que doit-on savoir avant de réaliser ce test ?
2. Le test de Fisher de comparaison des variances suppose que les variables soient normales. A l'aide d'un histogramme, vérifier que cette hypothèse est acceptable pour les deux échantillons. Procéder ensuite au test de Fisher. Que concluez-vous ?
3. Procéder au test de comparaison des moyennes et interpréter la sortie **R**.

3.2 En fonction du sexe du client

1. Reprendre les questions précédentes dans le but de tester l'égalité des pourboires selon le sexe du client.

3.3 En fonction de zone Fumeur/Non Fumeur

1. Reprendre les questions précédentes dans le but de tester l'affirmation d'un second salarié qui affirme que les clients fumeurs sont plus généreux que les nons fumeurs.

3.4 En fonction du jour de la semaine

Dans cette partie, on va mettre en oeuvre une analyse de la variance pour tester l'égalité des pourboires selon le jour de la semaine.

1. Vérifier que la classe de la colonne `DAY` de votre table de donnée `tab` est bien `factor`.
2. Réaliser une analyse de la variance à un facteur : le jour de la semaine. Rappeler le modèle.
3. L'instruction `anova` donne la table d'analyse de la variance. Rappeler quel test est réalisé et à quoi coorespond chaque ligne et colonne. Interpréter la table et conclure.
4. Vérifier les hypothèses du modèle ainsi que la présence de points abberants ou influents avec `plot`.
5. Si vous deviez poursuivre l'étude, que feriez-vous ?