ENSAM - TP

Tran Viet Chi, `chi.tran@math.univ-lille1.fr`, bureau 316 (bâtiment M3).

Exercice 1 (Des hommes et des femmes)
Les données du fichier `salaires.xls` fournissent le salaire, le sexe, la catégorie socio-professionnelle (CSP) et le nombre de jours d'absence pour chaque salarié d'une entreprise.

Partie A Nombre de jours d'absence

Nous nous intéressons au nombre de jours d'absence, X. On note $X_1, \cdots X_N$ les variables aléatoires correspondant aux différents individus de l'échantillon, N étant la taille de l'échantillon.

1. Représenter graphiquement, par sexe, la variable nombre de jours d'absence.

2. Calculer les moyennes et écart-type pour les hommes ($\bar{X}_1$, s_1^2), pour les femmes ($\bar{X}_2$, s_2^2) et pour l'ensemble ($\bar{X}$, s^2). Combien y a-t-il d'hommes (N_1), de femmes (N_2) ? Comparer les moyennes par sexe. Que cela vous inspire-t-il ?

3. (théorique) Montrer la décomposition de la variance empirique suivante :

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2 = \text{Var inter}(X) + \text{Var intra}(X),$$

$$\text{où : Var inter}(X) = \frac{N_1\left(\bar{X}_1 - \bar{X}\right)^2 + N_2\left(\bar{X}_1 - \bar{X}\right)^2}{N} \quad \text{et} \quad \text{Var intra}(X) = \frac{N_1\, s_1^2 + N_2 s_2^2}{N}$$

Quelle est l'interprétation de Var inter(X) ? de Var intra(X) ? Expliquer pourquoi en calculant :

$$R = \frac{\text{Var inter}(X)}{s^2}$$

on obtient la part réelle due au sexe de la différence entre les valeurs observées pour les hommes et pour les femmes.

4. A partir des données précédentes, calculer la variance inter et la variance intra. Conclusion ?

Partie B Données groupées

1. Créer le découpage suivant des salaires :
- on crée une variable auxiliaire *groupe* qui vaut 1 si l'individu gagne moins que 11450, 2 s'il gagne entre 11 450 et 15 650, 3 s'il gagne entre 15 650 et 39 150 et 4 sinon.
- Calculer pour chaque classe l'effectif, la moyenne et la variance des salaires.

2. On suppose qu'on ne dispose maintenant plus que des données résumées (*i.e.* les données en classes, et plus les données individuelles). Retrouver la moyenne et la variance.

3. Supposons qu'on ne dispose plus que du nombre d'employés par classe. Donner une approximation de la moyenne et de la variance.

Exercice 2 (Répartition salariale sur des données groupées, d'après le partiel de 2007-2008)
Dans une entreprise, les salaires sont les suivants :

1. Pour chaque classe $i \in \{1, 2, 3\}$ de salaires (notée $[x_{i-1}, x_i[$ et d'effectif n_i), calculer la fréquence empirique f_i, l'amplitude a_i, le centre $c_i = (x_{i-1} + x_i)/2$, la fréquence empirique cumulée F_i (proportion des salaires inférieurs à x_i), la masse salariale approchée $n_i c_i$ et la masse salariale cumulée approchée

Classe de salaire	Salaires mensuels	Nombre de salariés
1	$[500, 1500[$	50
2	$[1500, 2500[$	125
3	$[2500, 5500[$	25

m_i (approximation de la somme de tous les salaires inférieurs à x_i). On présentera les résultats dans un tableau.

2. Tracer avec **R** l'histogramme de la variable "salaire". Quelle règle faut-il respecter ?

3. Dessiner avec **R** la fonction de répartition. Comme la variable de salaire est quantitative continue, on choisira ici la version continue de la fonction de répartition empirique, obtenue par interpolation linéaire des points (x_i, F_i).

4. Quelle est l'équation de la portion de droite représentant la fonction de répartition empirique sur l'intervalle de salaires $[1500, 2500[$? En déduire la médiane.

5. On s'intéresse à la répartition des salaires sur ces données agrégées.
5.1. Tracer la courbe de Lorenz avec **R** (on pourra utiliser la commande `segments`). Pour des données groupées, cette courbe est obtenue en reliant les points

$$\left(F_i, \frac{\sum_{j\leq i} n_j c_j}{\sum_{j=1}^{3} n_j c_j}\right).$$

5.2. Calculer l'indice de Gini qui est égal à 2 fois l'aire entre la première bissectrice et la courbe. L'indice de Gini est un indicateur qui vaut 1 lorsque seul un individu gagne la totalité de la masse salariale et 0 lorsque tout le monde gagne la même chose. Commenter le résultat obtenu.

Exercice 3 (Logement)
Nous disposons, pour un échantillon de $n = 30$ appartements de 3 pièces mis en location dans un même quartier de Paris :
- du loyer mensuel Y (en francs),
- de la surface X (en m^2).

L'échantillon est noté $(X_i, Y_i)_{i\in[\![1,30]\!]}$.

1. Préciser la population, les variables (donner leur nature).

2. Décrire la distribution des deux variables X et Y (faire une étude avec **R** et résumer les résultats qui vous semblent intéressants).

3. Représenter graphiquement le loyer Y en fonction de la surface X. Premier commentaire ? Un modèle linéaire semble-t-il approprié ?

On s'intéresse au modèle $Y = aX + b + \varepsilon$, où a, b sont des coefficients à déterminer et où ε est un bruit gaussien centré et d'espérance σ^2.

4. Quelle est la variable expliquée, quelle est la variable explicative ?

5. Estimer les coefficients a et b du modèle par la méthode des moindres carrés ordinaires :
- en utilisant la fonction `lm` de **R**,
- en calculant directement les coefficients à l'aide de leur expression théorique.

Quelle est l'interprétation de a ? de b ? Commenter les valeurs numériques obtenues.

6. Superposer au nuage de points dessiné à la question **3** la droite de régression obtenue.

7. Calculer le coefficient de détermination. Commentaire ?

8. Etudier les hypothèses faites sur les résidus :
- Calculer leur moyenne et leur variance empirique. Quelles autres statistiques descriptives ou graphiques vous semble-t-il pertinent de donner ?
- Les résidus suivent-ils une loi normale ? Comment le vérifier ?