Licence L2 Mathématique - Informatique — Probabilités - Statistiques
Année universitaire 2006-2007

TP 3 : Comparaison de populations - Régression Linéaire

Exercice 1 : Comparaison de la tension artérielle chez des cadres d'une entreprise

Une entreprise œuvrant dans le domaine des télécommunications a sélectionné 40 individus occupant des postes de cadre dans l'entreprise pour vérifier leur tension artérielle systolique. On a également mesuré d'autres caractères comme l'age, l'indice corporelle, si l'individu est fumeur ou non et le niveau de responsabilité dans l'entreprise.

Dans cette étude, on va comparer la tension artérielle des individus selon qu'ils soient fumeurs ou non et selon leur age.

Téléchagez le fichier `cadre-entreprise.txt` à l'adresse `http://epoc.isima.fr/~diarrassouba/enseignements.htm` et lisez les données qu'il contient. La lecture des données se fait avec la fonction `read.table()`. Cette fonction permet de lire un fichier contenant plusieurs colonnes de données (comme c'est le cas ici). Elle génère un objet contenant plusieurs vecteurs dont les valeurs sont les données des colonnes du fichier. Chaque vecteur de l'objet est accessible par un nom. Par défaut, chaque colonne est nommé `x$V1 x$V2`,..., mais vous pouvez précisez vous-même le nom que vous souhaitez en spécifiant (par exemple) l'argument `header=TRUE` à la fonction `read.table()`. Les noms attribués aux différents vecteurs sont alors les noms qui figurent en entête de chaque colonne dans le fichier.

Tapez les commandes suivantes et observez le résultat.

```
cadre = read.table("cadre-entreprise.txt",header=TRUE)
cadre
cadre$No
cadre$Tension
cadre$Age
cadre$Fumeur
summary(cadre) //Obtenir un résumé sur chaque caractère mesuré
```

Q1. Calculez la tension artérielle moyenne des personnes sélectionnées.

Q2. Regroupez les tensions artérielles des individus en deux groupes : "Fumeur" et "Non-fumeur". Pour cela, tapez les commandes suivantes :

```
fumeur = cadre$Tension[cadre$Fumeur == "oui"]
non_fumeur = cadre$Tension[cadre$Fumeur == "non"]
```

Combien y a-t-il de fumeurs et de non-fumeurs parmi ces 40 individus ? Calculez les tensions artérielles moyennes pour ces deux groupes.

Q3. On veut tester au seuil $\alpha = 5\%$ les hypothèses H_0 : "la tension artérielle moyenne est la même pour les deux groupes d'individus" et H_1 : "La tension artérielle moyenne des fumeurs est supérieure à celle des non-fumeurs". Quelle est la forme de la région critique ?

Q4. Quelle loi suit la variable $T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_2}{n_1} + \frac{\sigma_2}{n_2}}}$ lorsque $n_1 < 30$ et $n_2 < 30$ avec σ_1 et σ_2 inconnus ?

Q5. Calculez la valeur $t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{var(X_1)}{n_1} + \frac{var(X_2)}{n_2}}}$ où X_1 et X_2 représente respectivement les tensions artériels des fumeurs et celles des non fumeurs. Comparez la valeur t au quantile adéquate et concluez sur le test d'hypothèse.

Q6. On va maintenant comparez les tensions artérielles des individus en fonction de leur age. Regroupez les individus en deux groupes : ceux agés de 50 ans et moins, et ceux de plus de 50 ans. Déterminez les effectifs respectifs de chaque groupe.

Q6. Réalisez un test d'hypothèse au seuil de $\alpha = 5$ % en considérant les hypothèses H_0 : "La tension artérielle moyenne est identique quelque soit la catégorie d'age" et H_1 : "La tension artérielle moyenne est supérieure pour les cadres agés de plus de 50 ans". Vous détaillerez les calculs en donnant forme de la région critique, la loi suivie par la variable T, la valeur de t, le quantile auquel vous la comparez et la conculsion du test.

Q7. D'après ces tests, on est sûr à 95% que le fait d'être fumeur et/ou être agé de plus de 50 ans augmentent la tension artérielle chez les cadres de l'entreprise. Sachant qu'une tension artérielle est considérée comme normale lorsque la tension systolique est ≤ 140, déterminez la proportion de cadre dans cet échantillon qui sont agés de plus de 50 ans, sont fumeurs et présentent un risque d'hypertension artétrielle.

Exercice 2 : Analyse par régression d'un fonds d'aide à la réinsertion du travail

Dans le cadre d'une aide à la réinsertion sociale de personnes en difficulté, le gouvernement du Canada a créé un "Fonds de lutte contre la pauvreté par la réinsertion du travail". Ce fonds permet de créer des emplois dans différentes régions et offre à ses bénéficiaires des salaires suffisament décents. Le ministère en charge de ce fonds souhaite déterminer si un lien statistique peut exister entre le montant accordé à chaque région et le nombre d'emplois qui y sont créés. On utilisera pour cela une analyse par régression linéaire simple.

Téléchargez le fichier `fonds-insertion-travail.txt`. Ce fichier donne pour chaque région le montant (en millions de $) accordé par le fonds et le nombre d'emplois qui ont été créés. Lisez les données (fonction `read.table()` en précisant `header = TRUE`) et stockez le résultat dans un objet que vous nommerez `fonds`.

Q1. Calculez le nombre total d'emplois créés ainsi que le montant total accordé par le fonds.

Dans la suite, on tentera d'exprimer le nombre d'emplois (E) en fonction du montant accordé (M) par une relation linéaire.

Q2. Tracez le diagramme de dispersion correspondant en utilisant la fonction `plot()` et en précisant comme vecteur en abscisse le vecteur `fonds$Montant` et en ordonnée le vecteur `fonds$NbEmplois`.

Q3. La liaison entre les deux variables semble-t-elle linéaire ? Quel est le sens de cette liaison ?

Q4. Calculez le coefficient de corrélation entre ces deux variables avec la formule vue en cours

$$\rho = \frac{cov(fonds\$Montant, fonds\$NbEmplois)}{sd(fonds\$Montant) * sd(fonds\$NbEmplois)}.$$

Pensez-vous que la corrélation entre ces deux variables est faible ou forte.

Q5. Calculez les coefficients a et b de la droite de régression en utilisant la fonction `lm()` et tracez la droite de régression (fonction `abline()`) (consultez l'annexe donnée plus bas sur "Comment calculer une droite de régression linéaire simple").

Q6. En vous servant de l'équation de la droite de régression, donnez les montants qu'il faudrait accorder à une région pour obtenir en moyenne un nombre d'emplois > 0, égal à 20, ègal à 100, égal à 200.

Q7. D'après cette relation statistique, quel serait le nombre moyen d'emplois créés pour un montant accordé de 2 millions ? Pour un montant de 4 millions ?

Q8. Le ministère en charge du fonds mentionne qu'une augmentation de 1 million du montant accordé à une région permettrait d'augmenter en moyenne de 100 le nombre d'emplois créés. Etes-vous d'accord avec cette affirmation ? Pourquoi ?

Exercice 3 : Etude des produits de protection solaire : corrélation Efficatité-Prix

Dans cet extercice, on se propose d'étudier une éventuelle corrélation entre l'efficacité de certains produits de protection solaire (protection contre les UVA) et leur prix. L'efficacité considérée ici est le pourcentage de rayons UVA qui sont bloqués par un produit donné. On tentera d'exprimer de façon linéaire le prix en fonction de cette efficacité.

Le fichier `protection-solaire.txt` donne pour 20 produits choisis au hasard leur prix et leur efficacité. Téléchargez-le et lisez les données qu'il contient (fonction `read.table()`). Dans la suite, on représentera le prix et l'efficacité des produits par deux variables aléatoires P et E.

Q1. Calculez le prix et l'efficacité moyens pour tous les produits.

Q2. Donnez le nom et le prix du produit le plus efficace. Donnez le nom du produit le plus cher. Quelle est son efficacité ?

Q3. Tracez le diagramme de dispersion de la relation "Prix en fonction de l'Efficacité". Peut-on dire, au vue du graphique, que les variables P et E sont liées par une relation linéaire ?

Q4. Calculez les coefficients de la droite de régression (fonction `lm()`) et tracez-là (fonction `abline()`). Calculez le coefficient de corrélation entre le prix et l'efficacité. Concluez sur le degré de corrélation de ces deux variables.

Q5. Au vue de ces statistiques, pensez-vous que pour mieux se protéger du soleil il vaut mieux acheter un produit cher. Vous justifierez votre réponse en interprétant le coefficient de corrélation ainsi que le diagramme de dispersion.

Annexe : Comment calculer une droite de régression linéaire simple

A partir de deux échantillons de données x et y obtenues par des tirages de deux variables aléatoires X et Y, on peut calculer avec "R" la droite de régression linéaire simple $Y = a + bX$ en utilisant la fonction `lm()` (**L**inear **M**odel). Cette fonction sert à définir la relation que l'on souhaite établir entre les variables X et Y. Par exemple, la commande `lm(y~x)` définit une

relation linéaire simple de Y en fonction de X. La fonction `lm()` calcule alors à partir des échantillons x et y, les coefficients a et b de la droite de régression. Le résultat renvoyé est un objet contenant deux valeurs : la première (désignée par `Intercept`) est la valeur a (ordonnée à l'origine), la deuxième valeur correspond au coefficient b (pente de la droite). Pour tracer la droite de régression, on utilise la fonction `abline()`. Si `fm` est l'objet renvoyé par la fonction `lm()`, alors les commandes suivantes permettent de tracez le diagramme de dispersion ainsi que la droite de régression :

```
plot(x,y)
fm = lm(y~x)
fm
abline(fm)
```

Remarque : Dû au fonctionnement de la fonction `abline()`, il est recommandé de tracer d'abord le diagramme de dispersion avec la fonction `plot()` avant de tracer la droite de régression.