Correction TP 3 : Comparaison de populations - Régression Linéaire

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

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.

```
> mean(cadre$Tension)
[1] 146.325
```

Q2.

```
> fumeur = cadre$Tension[cadre$Fumeur == "oui"]
> non_fumeur = cadre$Tension[cadre$Fumeur == "non"]
length(fumeur)
[1] 22
length(non_fumeur)
[1] 18
mean(fumeur)
[1] 151.9091
mean(non_fumeur)
[1] 139.5
```

Q3. On désigne par μ_1 la tension artérielle moyenne chez les fumeurs et μ_2 la tension artérielle moyenne chez les non fumeurs.
H_0 : "$\mu_1 = \mu_2$".
H_1 : "$\mu_1 > \mu_2$".
$\alpha = 5\%$.
Il s'agit d'un test unilateral dont la région critique est de la forme $]c, +\infty[$.

Q4. On supposera que les variances des deux groupes sont égales. En effet on constate que les variances empiriques des deux échantillons sont proches.

```
> sd(fumeur)
[1] 11.81165
> sd(non_fumeur)
[2] 11.88796
```

Le test sur l'égalité des moyennes se fait donc utilisant la variable $T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}}$ (Cf. le cours).

Cette variable suit une loi de student à $n_1 + n_2 - 2$ degrés de liberté. La région critique est de la forme $]t_{\alpha,n_1+n_2-2}, +\infty[$.

Q5.

```
> (mean(fumeur)-mean(non_fumeur))/(sqrt(1/22 + 1/18)*sqrt((21*var(fumeur) +
17*var(non_fumeur))/(38)))
T
[1] 3.296031
t_alpha = -qt(0.05,df=38)
[1] 1.685954
```

On voit que $T > t_{\alpha,38}$ et est bien dans la région critique. Donc on rejette l'hypothèse d'égalité des deux moyennes.

Q6.

```
> age50 = cadre$Tension[cadre$Age <= 50]
> age51 = cadre$Tension[cadre$Age > 50]
>length(age50)
[1] 19
>length(age51)
[1] 21
```

Q6. μ_1 est la moyenne chez les cadres de moins de 50 ans et μ_2 est la moyenne chez les plus de 50 ans.
H_0 : "$\mu_1 = \mu_2$"
et H_1 : "$\mu_1 < \mu_2$".
Le test est unilateral et la région critique est de la forme $]-\infty, c[$. Comme précédemment, on suppose que les variances sont identiques. Donc la variables T suit une loi de student à 38 degrés de liberté. La région critique est $]-\infty, -t_{\alpha,38}[$.

```
  > T = (mean(age50)-mean(age51))/(sqrt(1/19 + 1/21)*sqrt((18*var(age50) +
20*var(age51))/(38)))
> T
[1] 4.921663
```

La valeur T se trouve bien dans la région critique. Donc on rejette l'hypothèse d'égalité des moyenne des deux groupes.

Q7. Parmi les cadres agés de plus de 50 ans et qui sont fumeurs, ceux qui présentent un risque d'hypertension sont ceux qui ont une tension ≥ 140.

```
> risque = cadre$Tension[cadre$Age > 50 & cadre$Fumeur == "oui" & cadre$Tension
> 140] length(risque)
[1] 12
length(risque)/40
[1] 0.3
```

30% des cadres présentent un risque d'hypertension.

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

Q1.

```
> fonds = read.table("fonds-insertion-travail.txt",header=TRUE)
> sum(fonds$NbEmploi)
[1] 3769
```

Q2.

```
> plot(fonds$Montant,fonds$NbEmploi)
```

Q3. La liaison semble effectivement linéaire. Elle est même directe.

Q4.

```
> cov(fonds$Montant,fonds$NbEmploi)/(sd(fonds$Montant)*sd(fonds$NbEmploi))
[1] 0.9577412
```

Le coefficient de correlation est très proche de 1. Donc la correlation entre les deux variables est forte.

Q5.

```
> fm = lm(fonds$NbEmplois ~ fonds$Montant)
> fm
Call :
lm(formula = fonds$NbEmplois   fonds$Montant) Coefficients :
     (Intercept)       fonds$Montant
          -6.424              65.180
> abline(fm)
```

Q6. La droite de régression est $y = a + bx$ avec $a = -6.424$ et $b = 65.180$.
On a $x = \frac{y-a}{b}$.
Pour $y > 0$, $x > 0.986$ million
Pour $y = 20$, $x = 0.405$ million
Pour $y = 100$, $x = 1.632$ million
Pour $y = 200$, $x = 3.667$ millions.
Q7. Pour un montant de 2 millions on aurait 123.963 emplois. Pour un montant de 4 millions on aurait 254.296 emplois.

Q8. L'affirmation du ministère est fausse car le nombre d'emplois augmente en moyenne de 65.180 pour un montant d'un millions supplémentaire.

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

Q1.

```
> solaire = read.table("protection-solaire.txt",header=TRUE)
> mean(solaire$Prix)
[1] 11.005
> mean(solaire$Efficaite)
[1] 67.7
```

Q2.

```
> solaire$Marque[solaire$Efficacite == max(solaire$Efficacite)]
[1] Photoplex Plus
> solaire$Marque[solaire$Prix == max(solaire$Prix)]
[1] Lancome Soleil
> solaire$Efficacite[solaire$Prix == max(solaire$Prix)]
```

Q3.

```
> plot(solaire$Efficacite,solaire$Prix)
```

Vue la dispersion des points, la relation entre le prix et l'efficacité ne semble pas linéaire.

Q4.

```
> fm = lm(solaire$Prix ~ solaire$Efficacite)
> fm
Call :
lm(formula = solaire$Prix   solaire$Efficacite) Coefficients :
     (Intercept)       solaire$Efficacite
          8.81692              0.03232
> abline(fm)

> cov(solaire$Prix,solaire$Efficacite)/(sd(solaire$Efficacite)*sd(solaire$Prix))
[1] 0.1424317
```

La correlation entre le prix et l'efficacité est en effet faible.

Q5. Le coefficient de correlation étant faible, on peut supposer que le prix n'est pas directement lié à l'efficacité. On observe aussi qu'un certain nombre de produits (Photoplex Plus, Copperton UVGuard, Bain de soleil Protection) ont une forte efficacité mais ont un prix relativement bas. Par contre, certains produits ont un prix élevé mais une efficacité basse. On peut donc dire pour que statistiquement l'efficacité des produits ne justifie pas leurs prix.