TISD M1 Pro - Rattrapage du 5 mai 2009

Exercice 1 (Activité féminine, 4 pts)
Nous disposons d'un échantillon de femmes pour lesquelles nous connaissons la décision d'activité (`ACT=1` si la personne est active, `ACT=0` sinon) et le statut conjugal (`COHAB=1` si la femme vit en couple et `COHAB=2` sinon). On donne ci-dessous le tableau de contingence associé :

	COHAB=1	COHAB=2
ACT=0	25 676	13 563
ACT=1	43 575	17 145

TAB. 1 – *Tableau de contingence donnant les effectifs pour le croisement des variables ACT et COHAB*

1 (0.5 pt). Préciser la taille de la population et le type des variables étudiées.

2 (1). Calculer les fréquences marginales pour chaque variable.

3 (2.5). Faire un test du χ^2 d'indépendance de ces deux variables. On donne (en langage R) :

```
qchisq(0.95,1)=3.841459        qchisq(0.95,3)=7.814728
qchisq(0.95,2)=5.991465        qchisq(0.95,4)=9.487729
```

Rappeler le comportement de la statistique de test sous les hypothèses nulle et alternative. Qu'en déduisez-vous ?

Exercice 2 (Répartition salariale sur des données groupées, 6 pts)
Dans une entreprise, les salaires (considérés comme une variable quantitative continue) sont les suivants :

Classe de salaire	Salaires mensuels	Nombre de salariés
1	$[500, 2000[$	30
2	$[2000, 3500[$	150
3	$[3500, 5500[$	20

1 (1). Tracer l'histogramme de la variable "salaire".

2 (0.5). Tracer la fonction de répartition de la variable salaire (on rappelle que la variable de salaire est quantitative continue).

3 (1). A partir du tableau des données groupées, calculer une approximation du salaire moyen.

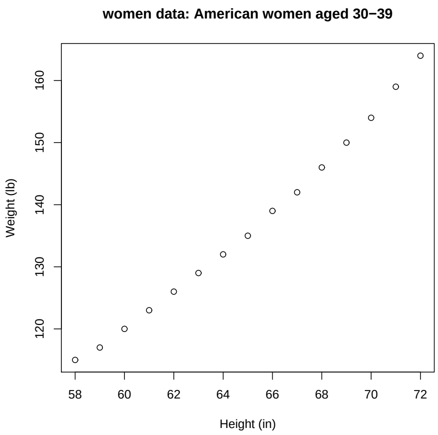
4 (1.5). Calculer la médiane de la distribution des salaires. Comparer au résultat de la question **3**.

5. On s'intéresse à la répartition des salaires sur ces données agrégées.
5.1 (0.5). Tracer la courbe de Lorenz. On rappelle que cette courbe est obtenue à partir des points

$$\left(F_i, \frac{\sum_{j\leq i} n_j c_j}{\sum_{j=1}^{3} n_j c_j}\right), \quad i \in [\![1,3]\!],$$

où F_i est la fréquence cumulée, n_i l'effectif et c_i le centre de la classe $i \in [\![1,3]\!]$.
5.2 (1). Calculer l'indice de Gini.
5.3 (0.5). Commenter.

Exercice 3 (Relation poids-taille, 10 pts)
On étudie la relation poids (`weight` en anglais, en inches) - taille (`height`, en livres) pour un échantillon de 15 Américaines âgées de 30 à 39 ans. On obtient le graphique suivant :



1 (0.5 pt). Que vous inspire ce graphique ?

2 (1.5). Sous **R**, la table s'appelle `women`. On donne :

```
mean(women$height)=65
mean(women$weight)=136.7333             mean(women$weight^2)=18920.2
mean(women$height^2)=4243.667           mean(women$height*women$weight)=8952.067
```

Calculer la moyenne et la variance des variables `taille` et `poids`, ainsi que la covariance des variables `taille` et `poids`.

3 (1.5). On obtient sous **R** :

```
var(women$height)= 20
var(women$weight)=240.2095              cov(women$height,women$weight)=69
```

D'où viennent les différences ? Ré-obtenir ces valeurs à partir des calculs de la question 2.

4 (1.5). A partir des résultats des questions précédentes, calculer les estimateurs MCO de la régression suivante :

$$\text{poids} = a \times \text{taille} + b + \varepsilon$$

5 (1). On donne le résultat suivant, obtenu sous **R** :

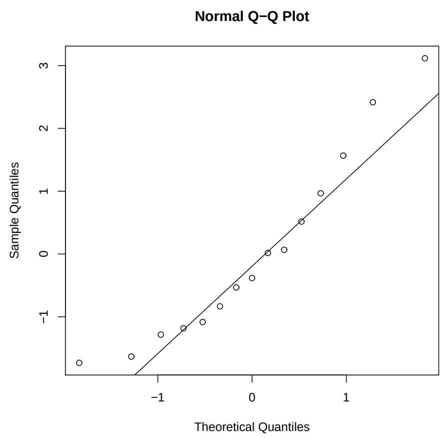
```
> reg<-lm(women$weight ~ women$height)
> summary(reg)

Call:
lm(formula = women$weight ~ women$height)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7333 -1.1333 -0.3833  0.7417  3.1167

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -87.51667    5.93694  -14.74 1.71e-09 ***
women$height   3.45000    0.09114   37.85 1.09e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.525 on 13 degrees of freedom
Multiple R-Squared: 0.991,      Adjusted R-squared: 0.9903
F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

Que pensez-vous de la qualité de la régression (donnez 3 éléments pour justifier votre réponse).

6 (1). Quelles sont les hypothèses pour pouvoir lire les tests de Student et de Fisher donnés à la question précédente ? Pour vérifier cela, on tape sous **R** le code suivant :

```
eps<-reg$residuals
qqnorm(eps)
qqline(eps)
```

ce qui nous fournit la figure ci-contre :



Qu'en pensez-vous ?

7 (1). Nous supposons les résidus gaussiens. Quelle est la loi de $\widehat{a}$ conditionnellement aux variables explicatives ? (on pourra noter x_i et y_i la taille et le poids respectivement de l'individu i)

8 (0.5). Quelle sont les hypothèses nulle et alternative du test de Student calculé sur la ligne `women$height   3.45000    0.09114   37.85 1.09e-14 ***` dans la question **5** ?

9 (1.5). A partir de la question **7**, établir l'expression littérale de la statistique du test de Student de la question **8**.