G. Hunault
Licence MEF

Angers, janvier 2009

Examen de Logiciels Statistiques

1. Analyse de variables QT

On s'intéresse pour l'instant uniquement aux variables quantitatives du dossier HER (Health Exam Results), qui sont des données correspondant à une étude américaine nationale de santé de la fin des années 90. On trouvera ces données et leur descriptif à l'endroit habituel (`k:\stat_ad\`) mais aussi sur le Web à l'adresse

`http://forge.info.univ-angers.fr/~gh/Datasets/datasets.htm`

Effectuez l'étude statistique *univariée* de ces variables quantitatives ; vous ne fournirez comme résultats numériques sur votre copie que ceux demandés dans les questions suivantes :

- A part la colonne d'identification, quelle est la seule variable du dossier qui n'est **pas** quantitative ?

- Quelles sont les trois variables dont le coefficient de variation est **le plus élevé** ?

- Quel tableau chiffré récapitulatif **bien ordonné** de ces trois variables pouvez-vous fournir ?

- Quel commentaire "intelligent" peut-on rédiger pour **chacune** de ces trois variables ?

Effectuez l'étude statistique *bivariée* des variables quantitatives du dossier HER ; vous ne fournirez comme résultats numériques sur votre copie que ceux demandés dans les questions suivantes :

- Quelles sont les deux variables quantitatives les plus corrélées linéairement ? Cette corrélation est-elle significativement différente de zéro ?

- Parmi les deux équations de régression, y en a-t-il une plus intéressante qu'une autre ? Si oui, laquelle ? Peut-on dire qu'il y a une relation de causalité entre ces deux variables ? Si oui, laquelle ?

- Quel commentaire "intelligent" peut-on écrire à propos de cette corrélation ?

- Quelle variable est la plus corrélée à la variable `POIGN` ?

2. Analyse de variables QL

Y a-t-il indépendance entres les modalités de la variable CLASS et celles de la variable AGE dans le dossier TITANIC ? Vous effectuerez le test adéquat et vous détaillerez, si besoin est, les modalités les plus liées. Vous n'oublierez pas de commenter "scientifiquement" les résultats.

3. Un peu de réflexion

Parmi tous les logiciels présentés dans les cours (Excel, R, StatBox, Spss, Statistica, Sas), lesquels savent calculer le coefficient de variation ?

Quels sont les défauts de ce coefficient ? Pourquoi ne peut-on pas s'en servir pour n'importe quelle variable quantitative ?

4. Lecture de fichiers de résultats

Le programme SAS suivant a été exécuté :

```
proc univariate data=her normal ;
     var age taille poids ;
run ;
```

Voici un extrait choisi du listing SAS correspondant :

```
Procédure UNIVARIATE
========================

Variable :  AGE
---------------
N                          80
Moyenne                 34.35
Ecart-type         13.1756392

Tests de normalité

Test                  -Statistique--    -----p Value------
Shapiro-Wilk          W      0.95245    Pr < W      0.0048
Kolmogorov-Smirnov    D      0.10828    Pr > D      0.0207
Cramer-von Mises      W-Sq  0.183197    Pr > W-Sq   0.0086
Anderson-Darling      A-Sq  1.183675    Pr > A-Sq  <0.0050

Variable :  TAILLE
------------------
N                          80
Moyenne                167.04
Ecart-type         9.80388195

Tests de normalité

Test                  -Statistique--    -----p Value------
Shapiro-Wilk          W     0.993139    Pr < W      0.9505
Kolmogorov-Smirnov    D     0.065996    Pr > D     >0.1500
Cramer-von Mises      W-Sq  0.042113    Pr > W-Sq  >0.2500
Anderson-Darling      A-Sq  0.221793    Pr > A-Sq  >0.2500

Variable :  POIDS
-----------------
N                          80
Moyenne                72.2925
Ecart-type         15.8166374

Tests de normalité

Test                  -Statistique--    -----p Value------
Shapiro-Wilk          W     0.981811    Pr < W      0.3134
Kolmogorov-Smirnov    D     0.048732    Pr > D     >0.1500
Cramer-von Mises      W-Sq  0.033239    Pr > W-Sq  >0.2500
Anderson-Darling      A-Sq  0.287177    Pr > A-Sq  >0.2500
```

On effectue des calculs similaires avec Statistica et on obtient les résultats suivants :

```
Classeur : her, Variable: AGE, Distribution : Normale
Chi-Deux = 21.59123, dl = 6 (ajustés) , p = 0.00144
K-S d=.10828, p> .20; Lilliefors p<.05
Shapiro-Wilk W=.95245, p=.00477

Classeur : her, Variable: TAILLE, Distribution : Normale
Chi-Deux = 1.66941, dl = 4 (ajustés) , p = 0.79627
K-S d=.06600, p> .20; Lilliefors p> .20
Shapiro-Wilk W=.99314, p=.95054

Classeur : her, Variable: POIDS, Distribution : Normale
Chi-Deux = 1.12585, dl = 6 (ajustés) , p = 0.98039
K-S d=.04873, p> .20; Lilliefors p> .20
Shapiro-Wilk W=.98181, p=.31339
```

Que peut-on conclure de toutes ces analyses ?

5. Comparaisons de moyennes

Comparez les AGES, POIDS et TAILLES des hommes et des femmes pour le dossier HER. On utilisera uniquement les tests vus en cours et on rédigera "proprement" les conclusions.

6. Choix de logiciel

Si vous deviez installer un logiciel de statistiques pour un entreprise de taille moyenne (de 40 à 100 personnes), quel logiciel choisiriez-vous ? Vous pouvez admettre que l'entreprise dispose d'un budget "conséquent" pour acheter ou louer n'importe quel logiciel de statistiques et de compétences informatiques suffisantes pour installer n'importe quel logiciel téléchargeable. Vous essaierez de justifier votre choix par des considérations pratiques.

ELEMENTS DE SOLUTION

1. Analyse de variables QT

La seule variable du dossier qui n'est pas quantitative est la variable SEXE, variable qualitative nominale. A l'aide des instructions R :

```
source("statgh.r")
her  <- lit.dar("her.dar")
attach(her)

# on extrait la matrice des variables QT seulement

                                ## print( head(her) )
her_qt <- her[,-1]
                                ## print( head(her_qt) )
                                ## print( tail(her_qt) )

# analyse univariée et bivariée

nomcol <- chaineEnListe("age taille poids ttaille
                  pouls sys dia chol  imc jmbg coud poign bras")
unites <- chaineEnListe("ans cm kg cm ppm mmHg mmHg mg
                                              kg/m2 cm cm cm cm")
allQT(her_qt,nomcol,unites)
```

on réalise l'analyse univariée et bivariée des variables QT et on en déduit que les 3 "variables qui varient le plus" (à l'aide du CDV) sont `CHOL`, `AGE` et `POIDS`. Le tableau récapitulatif correspondant est

Nom	Taille	Moyenne	Unite	Ecart-type	Coef. de var.		Min	Max
chol	80	318.050	mg	253.972	79.85	%	2	1252
age	80	34.350	ans	13.093	38.12	%	12	73
poids	80	72.293	kg	15.717	21.74	%	42	116

Voici un commentaire qui ne suppose aucune connaissance biomédicale :

La variable CHOL est assez dispersée ($cdv = 80\,\%$) pour ces 80 individus avec une moyenne de 318 mg et un écart-type de 254 mg alors que les variables AGE et POIDS sont nettement plus concentrées autour de la moyenne avec respectivement $m = 34$ ans, $s = 13$ ans et $m = 72$ kg, $s = 18$ kg.

Les résultats fournis par R montrent que les deux variables les plus corrélées linéairement sont `TTAILLE` et `POIDS` ($r = 0.908$, $p < 10^{-4}$) et que cette corrélation est significativement proche de 1 (ou encore différente de zéro) au seuil de 5 %.

Pour obtenir les deux équations de régression linéaire, on peut écrire en R :

```
analin("Tour de taille",ttaille,"cm","Poids",poids,"kg")
```

et on obtient alors

```
Poids           =   1.09 * Tour de taille  -  23.44
Tour de taille  =   0.76 * Poids           +  33.24
```

La question de savoir si une équation est plus intéressante qu'une autre n'a aucun sens en statistique. De même, la causalité se situe dans un domaine d'expertise qui ne relève pas de compétences statistiques mais plutôt de compétences biomédicales. Un expert du domaine pourrait ainsi dire : « *Il y a de fortes chances qu'aucune de ces deux variables ne soient la cause de l'autre, mais plutôt qu'elles soit toutes les deux dépendants d'une même autre troisième variable "surcharge pondérale", elle-même liée au syndrome métabolique.* »

Si dans la "matrice des corrélations" on lit la ligne pour `POIDS`, on voit que la variable qui lui est le plus corrélée est `COUD` ($r = 0.802$) et à l'aide de liste des corrélations on peut affirmer que cette corrélation est significativement proche de 1 au seuil de 5 % ($p < 10^{-4}$).

2. Analyse de variables QL

A l'aide des instructions R suivantes :

```
titanic <- lit.dar("http://www.info.univ-angers.fr/pub/gh/Datasets/titanic.d
attach(titanic)
mclass <- "équipage 1èreClasse 2èmeClasse 3èmeCatégorie"
mage   <- "enfant adulte"
triCroise("Cabine",CLASS,mclass,"Age",AGE,mage,TRUE)
```

on peut constater que les valeurs observées dans le tri croisé et dans le tableau des valeurs théoriques attendues sous hypothèse d'indépendance sont très différentes :

```
TABLEAU DES DONNEES

          équipage   1èreClasse   2èmeClasse   3èmeCatégorie   Total
enfant           0            6           24              79     109
adulte         885          319          261             627    2092
Total          885          325          285             706    2201

 VALEURS ATTENDUES et MARGES

          équipage   1èreClasse   2èmeClasse   3èmeCatégorie   Total
enfant          44           16           14              35     109
adulte         841          309          271             671    2092
Total          885          325          285             706    2201
```

La différence la plus importante semble être l'absence d'enfants parmi les membres d'équipage, ce qui se comprend aisément. Un test du χ^2 d'indépendance montre qu'il y a effectivement dépendance entre les modalités : $\chi^2 = 118.4133$, $p = 1.694884 \times 10^{-25}$. La liste décroissante des contributions au χ^2 par couple de modalités le montre clairement mais met en évidence que la différence la plus importance est due à un grand nombre d'enfants en cabine de 3ème catégorie et qu'avec la différence précédente (enfants parmi les membres d'équipage) pratiquement tout le χ^2 est "expliqué".

```
  Signe     Valeur    Pct      Ligne     Colonne          Obs    Th
    +       55.465   46.84 %   enfant    3èmeCatégorie    79     35.0
    -       43.828   37.01 %   enfant    équipage          0     43.8
    +        6.924    5.85 %   enfant    2èmeClasse       24     14.1
...
```

3. Un peu de réflexion

Excel ne propose pas de fonction pour calculer le coefficient de variation, pas plus que R. Spss ne le propose pas dans ses menus (mais il existe une fonction `CFVAR` pour les macros). StatBox, Statistica et Sas savent fournir directement le coefficient de variation.

Le défaut du CDV est que si la moyenne m est proche de zéro, ce coefficient varie beaucoup ; si $m = 0$, il ne peut pas être calculé. Il est donc plutôt réservé aux variables QT qui ne changent pas de signe. De plus $X \rightsquigarrow cdv(X)$ n'est pas une fonction additive car $m(X+a) = m(X)+a$ et $s(X+a) = s(X)$, ce qui est peut-être un défaut.

4. Lecture de fichiers de résultats

Les calculs effectués avec SAS et STATISTICA sont des tests de normalité pour les variables AGE, TAILLE et POIDS. Au vu des résultats, il semblerait qu'on puisse considérer que les variables TAILLE et POIDS puissent être considérées comme suivant une distribution normale, mais pas AGE.

5. Comparaisons de moyennes

On utilise les instructions R suivantes :

```
source("statgh.r")
her  <- lit.dar("her.dar") ; attach(her)

ageH <- age[sexe==0]
ageF <- age[sexe==1]
compMoyData("AGE vs SEXE",ageH,ageF)

poiH <- poids[sexe==0]
poiF <- poids[sexe==1]
compMoyData("POIDS vs SEXE",poiH,poiF)

taiH <- taille[sexe==0]
taiF <- taille[sexe==1]
compMoyData("TAILLE vs SEXE",taiH,taiF)
```

On peut alors affirmer au seuil de 5 % qu'il n'y a pas de différence significative entre les moyennes des ages pour les hommes et pour les femmes ($p = 0.445$), qu'il y a une différence significative entre les moyennes des poids ($p = 0.0005$) et aussi entre les moyennes des tailles ($p = 1.130 \times 10^{-11}$).