ASI (L2) : TP9

Objectifs du TP :

Effectuer une analyse textuelle, lire et utiliser des macros *Excel* mises dans des fichiers-texte, lire et utiliser des programmes *Rstat*.

1. Une analyse textuelle

On veut produire les dictionnaires alphabétiques et fréquentiels du texte FOGG.TXT.

Quel(s) logiciel(s) ou quelle(s) page(s) *Web* peut-on utiliser ?

Combien y a-t-il de mots en tout ? et de mots différents ? Quel en est le rapport ?

Quels sont les "vrais" mots les plus fréquents ? Que peut-on en déduire ?

Combien de fois le mot "jour" et le mot "jours" apparaissent-ils ?

On veut maintenant connaitre les environnements du mot "jour" et du mot "jours" avec 3 mots avant et 4 mots après.

Quel(s) logiciel(s) ou quelle(s) page(s) *Web* peut-on utiliser ?

Attention : le texte à analyser ne commence qu'à la ligne 262.

2. Lire et utiliser des macros *Excel*

Après avoir lu le fichier STEMLEAF.TXT essayer de réaliser sous *Excel* le diagramme tige et feuille de la variable `AGE` puis de la variable `FORM_ANBAC` pour le dossier `CETOP2`.

Que pouvez-vous en conclure ?

Après avoir lu le fichier BOXPLOTS.TXT essayer de réaliser sous *Excel* un affichage en boites à moustaches pour les vins puis pour les catégories de vins du dossier `VINS`.

Que voit-on ?

Comment utiliser le fichier-modèle ASGQT.XLT pour réaliser sous *Excel* l'analyse séparée et l'analyse conjointe de toutes les QT d'un dossier, par exemple pour le dossier `ANTAL` ?

3. Lire et utiliser des programmes *Rstat*

Après avoir lu le fichier STATGH.R essayer de réaliser très rapidement l'analyse séparée et l'analyse conjointe de toutes les QT du dossier `ANTAL` puis les catégories de vins dans le dossier `VINS`.

Effectuer ensuite à l'aide des programmes de STATGH.R la comparaison des moyennes d'ages des hommes et des femmes du dossier `ELF`.

Effectuer enfin la comparaison des pourcentages d'hommes et de femmes survivantes dans le dossier `TITANIC`.

Esquisse de SOLUTION

Une analyse textuelle

Si on utilise la page *Web* d'Analexies, on obtient comme résultats :

```
Dans le texte il y a sans doute 6579 mots en tout
                              et 1986 mots differents ;
chaque mot est donc repete 3.31 fois en moyenne.
```

Avec le logiciel `analexies` du disque `K:` on obtient

```
Il y a : 6694 mots  repartis en :
         6675 mots courants  dont 2024 mots courants differents
```

(soit alors une répétition moyenne de 3.298 fois par mots).

La différence tient à la définition de ce qu'est un mot, à la gestion de la ponctuation...

Les "vrais" mots les plus fréquents sont

`fogg`	55
`passepartout`	33
`bombay`	25
`phileas`	24
`heures`	21

On reconnait le nom des héros et la ville des chapitres 9, 10 et 11 ainsi qu'un mot important : *heures*.

Le mot *jour* apparait 3 fois et le mot *jours* 9 fois.

Voici les environnements fournis par le logiciel `analexies` du disque `K:`

```
STATISTIQUES fournies par les dictionnaires
===========================================

   Mot         Occurences
   ---         ----------
   jour           3
   jours          9

Les environnements sont affichés avec 3 mots avants  et 4 mots après.

-----------------------------------------------------------------------

          deux fois par     jour     on faisait de la
       quatre repas par     jour     sans que jamais ni
     se modifier chaque     jour     autrefois on y voyageait
           de bombay ce     jour     là ils célébraient une

       en quatre vingts     jours     non toute cette gymnastique
       en quatre vingts     jours     pourrait bien cacher quelque
           gain de deux     jours     que phileas fogg inscrivit
         bombay à trois     jours     seulement de calcutta le
emploieraient pas trois     jours     à la franchir mais
       en quatre vingts     jours     en attendant et après
      et recomptait les     jours     écoulés maudissait les haltes
          du soleil les     jours     étaient plus courts d
              j ai deux     jours     d avance à sacrifier


-- Fin de recherche d'environnement
```

Lire et utiliser des macros *Excel*

Pour les diagrammes "*Tige et Feuilles*", il n'y a qu'une seule macro dans le fichier STEMLEAF.TXT nommée $StemAndLeaf$. On peut l'exécuter via le menu `Outils / Macro / Macros`. Il faut visiblement deux onglets, l'un nommé *Stem* et l'autre *Data*. Les données doivent être en colonne 1, à partir de la ligne 2 (en ligne 1 il y a sans doute le nom de la variable). Les données doivent être triées par ordre croissant. On trouvera les résultats sur la page suivante.

Pour l'age, on en déduit qu'il à beaucoup de non-réponses (valeur 0) et que la plupart des gens ont de 30 à 40 ans.

Pour le nombre d'années de formation, il y a là aussi beaucoup de non-réponses ; la plupart des gens ont un bac +5.

Pour les diagrammes "*Boites à moustaches*", il y a plusieurs sous-programmes, le premier, nommé $CreerFeuilleBoxPlot$ étant le programme principal. Pour exécuter le programme, il faut sélectionner les données à tracer avant de passer par le menu `Outils / Macro / Macros` (sans les *id* de ligne).

Le programme crée alors un onglet avec les résultats numériques et le graphiques. Il serait plus judicieux de mettre le graphique seul dans une feuille (on le fait "à la main" rapidement avec un clic-droit dans le graphique à l'aide du menu `Emplacement`).

Pour traiter les catégories de vins, il faut transposer les données du dossier `VINS`. Pour ce faire, il faut sélectionner les données, faire `Edition / Copier` puis `Edition / Collage special` et cocher la case Transposé.

Pour utiliser le fichier-modèle `ASGQT.XLT`, il suffit de le recopier sur le disque `D:`, de l'ouvrir en activant les macros et de faire ce que dit l'onglet *Aides*, à savoir :

- insérer les données à partir de la cellule L1C1 dans l'onglet Données ; (en principe la ligne 1 doit contenir le nom des colonnes et la colonne 1 le nom des lignes),
- cliquer sur le bouton "départ". dans l'onglet *Statistiques*,
- lorsque l'affichage est stabilisé, utiliser les boutons 1 à 6 pour trier la colonne correspondante.

Il faut certainement modifier les formats de cellule pour une meilleure lisibilité. Attention au calcul des meilleures corrélation car les bornes 0.6 et 0.9 sont fixes (et sans doute mal gérées...).

Variable AGE

```
  Count   Stem       Leaves  Each digit represents 1 case(s).
     42      0       |00000000000000000000000000000000000000002
      0      1       |
     40      2       |3445555666667777777778888888888899999999
     63      3       |000011111111222222223333344444555555556666666677778888888999999
     70      4       |0000000011111111222222222223333333444444555555556666667777777888999999
     30      5       |000000112222233333334555678899
      1      6       |0
```

Variable FORM_ANBAC

```
  Count   Stem       Leaves  Each digit represents 2 case(s).
     53      0       |00000000000000000000000000
      2      1       |0
     13      2       |000000
      9      3       |00000
     33      4       |0000000000000000
    110      5       |00000000000000000000000000000000000000000000000000000000000000
     15      6       |00000000
      9      7       |0000
      2      8       |0
```

Pays importateurs de vins

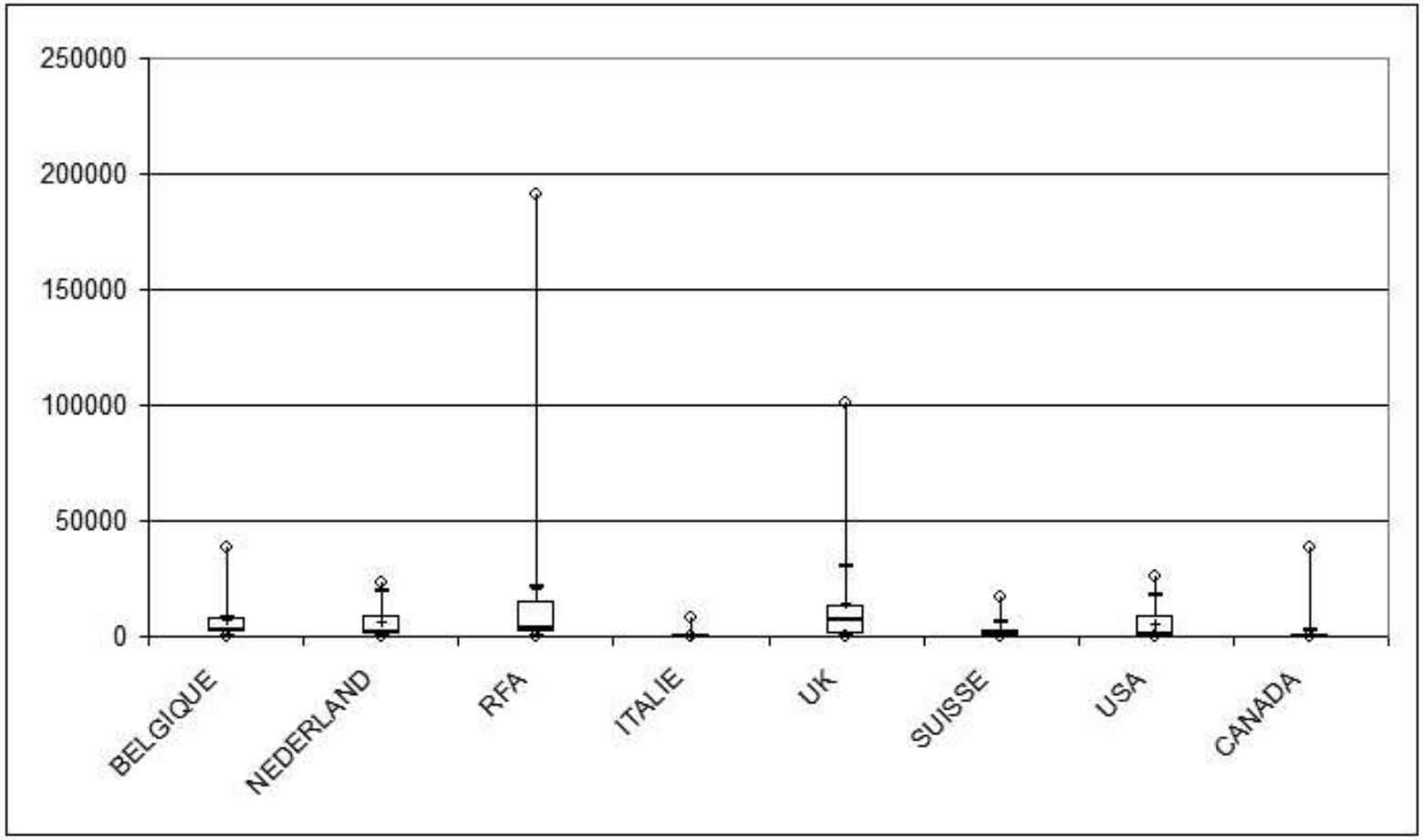
NOM	BELGIQUE	NEDERLAND	RFA	ITALIE	UK	SUISSE	USA	CANADA
q1	1986	707,25	1391,75	1,5	1187	119,75	441,5	71
min	24	74	135	0	284	0	0	0
moust. inf.	24	74	135	0	284	0	0	0
med	2512	1824	3671	27	7398	536	1007	286
moy	7470	6261	20262	1119	14299	2965	5153	2814
moust. sup.	7950	19840	21023	98	30025	6544	17487	2346
max	38747	22806	191140	8037	101108	17327	26192	38503
q3	7729,75	8849,25	15128,75	87,75	13389,75	2713	8793,5	1127
nb atyp. inf.	0	0	0	0	0	0	0	0
nb atyp. sup.	4	2	2	4	2	3	1	2
effectif	18	18	18	18	18	18	18	18

Catégories de vins

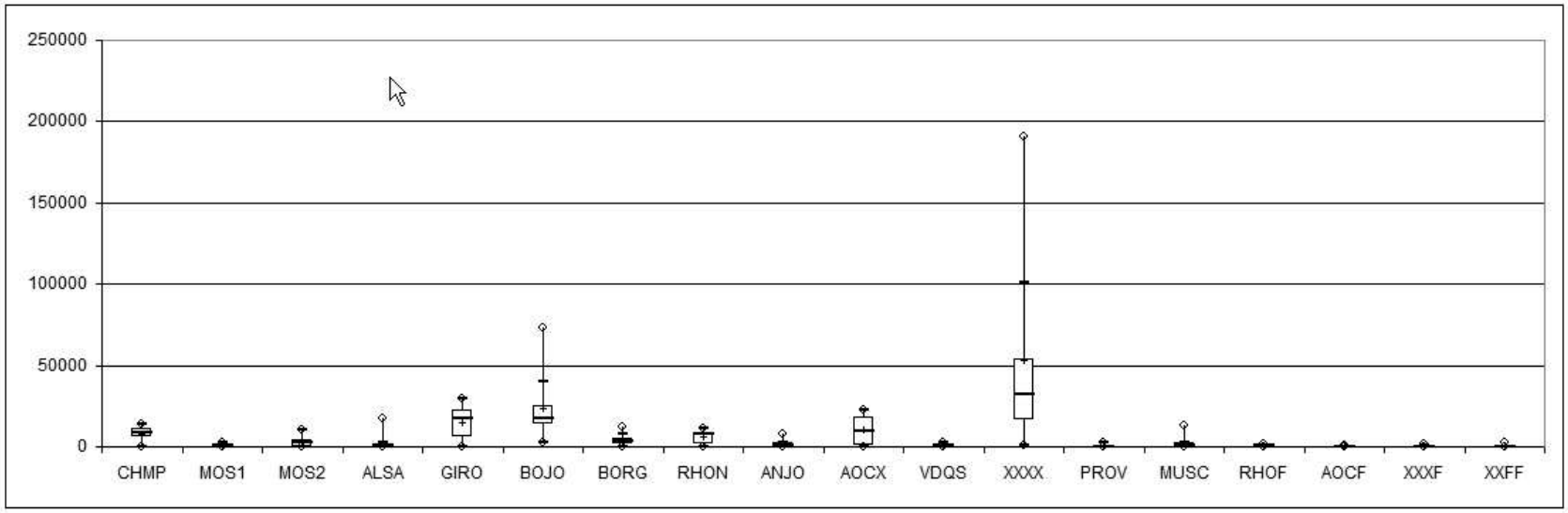
NOM	CHMP	MOS1	MOS2	ALSA	GIRO	BOJO	BORG	RHON	ANJO
q1	6248,25	366	320	366,25	5769,75	13586,25	2051,25	1690,75	140
min	206	30	112	57	56	2346	98	24	0
moust. inf.	206	30	112	57	56	2346	98	24	0
med	8851	792	2240	864	17069	17476	3488	7751	736
moy	8160	974	3332	3005	14922	23716	4388	6137	1752
moust. sup.	13556	2436	10439	2422	30025	39919	7885	11691	2587
max	13556	2436	10439	17183	30025	72977	11791	11691	7582
q3	10934	1414,25	4644,5	2104,75	22383,75	24859,75	5592,25	8763,25	2222,5
nb atyp. inf.	0	0	0	0	0	0	0	0	0
nb atyp. sup.	0	0	0	1	0	1	1	0	1
effectif	8	8	8	8	8	8	8	8	8

NOM	AOCX	VDQS	XXXX	PROV	MUSC	RHOF	AOCF	XXXF	XXFF
q1	1195,25	418,25	16361,25	179,25	494,25	415,25	7,75	0	30
min	50	0	1029	0	0	6	0	0	8
moust. inf.	50	0	1029	0	0	6	0	0	8
med	9998	1023	32348	343	1123	709	81	12	61
moy	10285	1041	52983	746	2591	680	220	275	563
moust. sup.	22806	2258	101108	2514	2908	1009	246	480	208
max	22806	2258	191140	2514	12891	1648	1177	1533	2415
q3	17901	1503,5	54337,25	1206,25	2239	841	181,5	240	582,25
nb atyp. inf.	0	0	0	0	0	0	0	0	0
nb atyp. sup.	0	0	1	0	1	1	1	1	2
effectif	8	8	8	8	8	8	8	8	8

Boxplot de *vins* :



Boxplot de *vins* transposé :



4. Lire et utiliser des programmes *Rstat*

Le fichier STATGH.R permet de réaliser l'analyse de toutes les QT via `anaQT` une fois qu'on a lu le fichier des données. Par exemple, on peut écrire :

```
source("statgh.r")
antaldbf <- read.dbf("antal.dbf")
antalmat <- antaldbf$dbf
nbl      <- (dim(antalmat)[1])
nbc      <- (dim(antalmat)[2])
antal    <- antalmat[1:nbl,2:nbc]

anaQT(antal,colnames(antal))
```

Pour comparer les moyennes d'ages des hommes et des femmes du dossier `ELF` le plus simple est d'utiliser `compMoyData` puisqu'on dispose des données :

```
elf <- read.dbf("elf.dbf")

sexe <- elf$dbf[2]
age  <- elf$dbf[3]

agehom <- age[sexe==0]
agefam <- age[sexe==1]

compMoyData(agehom,agefam)
```

Pour le dossier `TITANIC`, on utilise la fonction `compPourc` :

```
tit <- read.dbf("titanic.dbf")
sexe  <- tit$dbf[4]
surv  <- tit$dbf[5]

nbhom <- sum( sexe==1 )
nbfam <- sum( sexe==0 )

nbhomsurv  <- sum( (sexe==1) & (surv==1) )
nbfamsurv  <- sum( (sexe==0) & (surv==1) )

compPourc(nbhomsurv,nbhom,nbfamsurv,nbfam)
```

Voici le résultat de l'exécution de ces commandes :

```
> ################################################
>      source("statgh.r")

 statgh.r, version 1.34 Mars 2005

> ################################################

>      antaldbf <- read.dbf("antal.dbf")
>      antalmat <- antaldbf$dbf
>      nbl      <- (dim(antalmat)[1])
>      nbc      <- (dim(antalmat)[2])
>      antal    <- antalmat[1:nbl,2:nbc]
>
>      anaQT(antal,colnames(antal))

Données
$DOU0
 [1]  0  0  0  6  6  6  4  0  0  0  0  0  0  4  4  2  2  0
      0  0  0  4  4 12 12 12 12  0  0  0  2  2  4  4  2  2

$DOU1
 [1]  0  2  4  4 10 10 12 14 14  0  4 12 14 16 12 12 14 18
      0  4  4 12 20 14 10 10 10  0  0  0  4  8  8  8  8  6

$DOU2
 [1]  0  4 16 18 12  8  4  6  6  0  6 10 16 10 12 16 10  8
      0  6 20 18 12  7 10 10 10  0  0  0 16  8  8  4  8 12

$DOU3
 [1] 28 28 18 10  8 12 16 16 16 22 22 16  8  8 10 10 14 14
     18 20 14  6  4  7  8  8  8 22 22 34 14 10 14 16 10 10

$DOU4
 [1] 12  6  2  2  4  4  4  4  4 18  8  2  2  2  2  0  0  0
     22 10  2  0  0  0  0  0  0 18 18  6  4  4  6  8 12 10

$DOU5
 [1] 25 13  5  5  9  9  9  9  9 37 17  5  5  5  5  1  1  1
     45 21  5  1  1  1  1  1  1 37 37 13  9  9 13 17 25 21
```

Par cdv décroissant

Nom	Num	Taille	Moyenne	Ecart-type	Coef. de variation	Minimum	Maximum
1	DOU0	36	2.944	3.756	129.4	0.000	12.000
5	DOU4	36	5.444	5.899	109.9	0.000	22.000
6	DOU5	36	11.889	11.799	100.6	1.000	45.000
2	DOU1	36	8.278	5.541	67.9	0.000	20.000
3	DOU2	36	8.639	5.583	65.5	0.000	20.000
4	DOU3	36	14.472	6.784	47.5	4.000	34.000

Par ordre d'entrée

Nom	Num	Taille	Moyenne	Ecart-type	Coef. de variation	Minimum	Maximum
1	DOU0	36	2.944	3.756	129.4	0.000	12.000
2	DOU1	36	8.278	5.541	67.9	0.000	20.000
3	DOU2	36	8.639	5.583	65.5	0.000	20.000
4	DOU3	36	14.472	6.784	47.5	4.000	34.000
5	DOU4	36	5.444	5.899	109.9	0.000	22.000
6	DOU5	36	11.889	11.799	100.6	1.000	45.000

Par moyenne décroissante

Nom	Num	Taille	Moyenne	Ecart-type	Coef. de variation	Minimum	Maximum
4	DOU3	36	14.472	6.784	47.5	4.000	34.000
6	DOU5	36	11.889	11.799	100.6	1.000	45.000
3	DOU2	36	8.639	5.583	65.5	0.000	20.000
2	DOU1	36	8.278	5.541	67.9	0.000	20.000
5	DOU4	36	5.444	5.899	109.9	0.000	22.000
1	DOU0	36	2.944	3.756	129.4	0.000	12.000

Matrice des corrélations

	DOU0	DOU1	DOU2	DOU3	DOU4	DOU5
DOU0	1.000					
DOU1	0.334	1.000				
DOU2	0.223	0.405	1.000			
DOU3	-0.600	-0.705	-0.672	1.000		
DOU4	-0.463	-0.723	-0.693	0.555	1.000	
DOU5	-0.463	-0.723	-0.693	0.555	1.000	1.000

Meilleure corrélation 1 pour DOU5 et DOU4
Formule DOU4 = 0.500 * DOU5 -0.500

Vous pouvez utilisez r.ps pour voir la courbe correspondante

```
 Coefficients de corrélation par ordre décroissant

    1.000  pour          6  et      5  soit  DOU5     DOU4
   -0.723  pour          6  et      2  soit  DOU5     DOU1
   -0.723  pour          5  et      2  soit  DOU4     DOU1
   -0.705  pour          4  et      2  soit  DOU3     DOU1
   -0.693  pour          6  et      3  soit  DOU5     DOU2
   -0.693  pour          5  et      3  soit  DOU4     DOU2
   -0.672  pour          4  et      3  soit  DOU3     DOU2
   -0.600  pour          4  et      1  soit  DOU3     DOU0
    0.555  pour          6  et      4  soit  DOU5     DOU3
    0.555  pour          5  et      4  soit  DOU4     DOU3
   -0.463  pour          5  et      1  soit  DOU4     DOU0
   -0.463  pour          6  et      1  soit  DOU5     DOU0
    0.405  pour          3  et      2  soit  DOU2     DOU1
    0.334  pour          2  et      1  soit  DOU1     DOU0
    0.223  pour          3  et      1  soit  DOU2     DOU0
>
> ################################################
>
>      elf <- read.dbf("elf.dbf")
>
>      sexe <- elf$dbf[2]
>      age  <- elf$dbf[3]
>
>      agehom <- age[sexe==0]
>      agefam <- age[sexe==1]
>
>      compMoyData(agehom,agefam)

 COMPARAISON DE MOYENNES (valeurs fournies)

    Variable    nbVal      Moyenne    Variance     Ecart-type     Cdv
       A          35       36.400      285.365       16.893        46  %
       B          64       35.516      324.984       18.027        51  %

    différence réduite :  0.2431

    au seuil de 5 % soit 1.96, on peut accepter l'hypothèse d'égalité des moyennes.
```

```
> ##############################################
>
>      tit <- read.dbf("titanic.dbf")
>      sexe  <- tit$dbf[4]
>      surv  <- tit$dbf[5]
>
>      nbhom <- sum( sexe==1 )
>      nbfam <- sum( sexe==0 )
>
>      nbhomsurv  <- sum( (sexe==1) & (surv==1) )
>      nbfamsurv  <- sum( (sexe==0) & (surv==1) )
>
>      compPourc(nbhomsurv,nbhom,nbfamsurv,nbfam)

 COMPARAISON DE POURCENTAGES

    population A,    367  individus marqués sur  1731  soit une proportion de  0.212
    population B,    344  individus marqués sur   470  soit une proportion de  0.732
    globalisation,   711  individus marqués sur  2201  soit une proportion de  0.323

    écart-réduit :  21.3746

    au seuil de 5 % soit 1.96, on peut refuser l'hypothèse d'égalité des pourcentages.
```

Remarques sur la transposition des données

Comme nos matrices de données comportent en ligne 1 le nom des colonnes et en colonne 1 le nom des lignes, pour transposer une matrice de données, on ne peut pas se contenter d'utiliser la fonction t car sinon les nouvelles lignes de données sont transformées en caractères. On pourra s'en rendre compte avec les instructions

```
vins     <- read.dbf("vins.dbf")
vinsdata <- vins$dbf
t(vinsdata)
```

Il faut donc procéder plus prudemment, à savoir :

```
vins     <- read.dbf("vins.dbf")
vinsdata <- vins$dbf

anomlig  <- vinsdata[,1]
anomcol  <- colnames(vinsdata)
anblig   <- dim(vinsdata)[1]
anbcol   <- dim(vinsdata)[2]
valnum   <- vinsdata[ ,2:anbcol ]
dim(valnum)

snivdata <- matrix(nrow=anbcol-1,ncol=anblig+1)
snivdata[1:(anbcol-1), 2:(anblig+1)] <- t(valnum)
snivmat <- as.data.frame(snivdata)
snivmat[,1] <- anomcol[2:anbcol]
colnames(snivmat) <- c("VIN",anomlig)
nomcol <- anomlig

nbl      <- (dim(snivmat)[1])
nbc      <- (dim(snivmat)[2])
sniv     <- snivmat[1:nbl,2:nbc]
allQT(sniv,nomcol)
```