G. Hunault
Licence Mass

Angers, novembre 2004

Logiciels Statistiques : exercices en Dbase

1. Un remplacement vite fait

La base de données BullQT.dbf est un extrait de la base de données Bull.dbf. Elle contient l'identificateur des individus interrogés et les 6 variables quantitatives pour le dossier BULL. Parmi ces variables QT, le palier de déverminage est exprimé en heures et en minutes dans deux champs séparés. Créer un champ DurPal et utiliser l'instruction REPLACE de Dbase pour remplir ce champ. Discuter ensuite comment calculer la moyenne du palier, avec les deux champs de départ puis avec le champ DurPal.

2. Comparaison de moyennes et de pourcentages

Donner les instructions *Dbase* qui permettent d'ouvrir la base ELF, de calculer la moyenne et la variance de l'age d'abord pour les hommes et ensuite pour les femmes. Au sens de la comparaison de moyennes, y a-t-il une différence à 5 % ?

Effectuer ensuite une comparaison des pourcentages de personnes ayant fait des études supérieures pour les deux modalités de la variable SEXE dans le dossier ELF. On pourra se limiter à une comparaison à 5 %. Effectuer également une comparaison de moyennes pour l'age de hommes ayant fait des études supérieures et des femmes ayant fait des études supérieures dans ce même dossier ELF. On pourra se limiter à une comparaison à 5 %.

Vous n'oublierez pas de conclure par une phrase simple et lisible à chaque fois.

3. Calcul de médiane

Soit X une variable statistique quantitative de taille n et soit x_1, x_2... x_n les différentes valeurs que prend la variable. La médiane m^* de X qui est une mesure dite "de tendance centrale" est la valeur telle que 50 % des valeurs de X sont au-dessus de m^* et 50 % sont au-dessous de m^* pour un nombre impair de valeurs. Pour un nombre pair de valeurs, on utilise la demi-somme des valeurs du milieu de X lorsqu'elles sont rangées par ordre croissant.

Ainsi la médiane de `-1, 2, 3, 4, 15` est `3` et la médiane de `1, 99` est `50`.

Soit U la variable quantitative définie par les $n = 5$ valeurs

```
12 17 14 16 15
```

correspondant à l'étude du poids de matière séche pour un arbre de type pin du Canada, l'unité de mesure étant le gramme.

Donner les expressions Dbase qui calculent les trois quantités $moy(U)$, $ect(U)$, $m^*(U)$ où moy désigne la moyenne, ect l'écart-type et m^* la médiane. Afin de gagner du temps, on pourra utiliser la base de données nommée mstar.dbf ; les variables U, V et W y sont respectivement nommées XU, XV et XW.

Donner ensuite les instructions qui fournissent avec un seul chiffre après la virgule les valeurs numériques associées.

Donner enfin les valeurs numériques correspondantes pour la série théorique V définie par les valeurs `10 11 12 13 14` puis pour la série théorique W définie par les valeurs `10 11 12 11 10 14`.

Pour les plus fort(e)s, écrire un programme median.prg qui demande le nom d'une base de données, d'un champ et qui calcule la médiane de ce champ pour la base de données.

4. Coefficients d'asymétrie et d'aplatissement

Soit X une variable statistique quantitative de taille n et soit x_1, x_2... x_n les différentes valeurs que prend la variable. On se propose ici de calculer les coefficients d'asymétrie et d'aplatissement de la variable (nommés aussi *skewness* et *kurtosis*) à l'aide de *Dbase*. On note $moy(X)$ la moyenne de X et $ect(X)$ son écart-type mathématique exact défini comme la racine de la quantité $moy(X^2) - moy(X)^2$.

On nomme valeur centrée réduite pour la valeur numéro i notée d_i la quantité

$$d_i = \frac{x_i - moy(X)}{ect(X)}$$

Le coefficient d'asymétrie de X, noté $sk(X)$ et le coefficient d'aplatissement X, noté $ku(X)$ sont alors définis par

$$sk(X) = \frac{1}{n}\sum_{i=1}^{n} {d_i}^3 \quad \text{et} \quad ku(X) = \frac{1}{n}\sum_{i=1}^{n} {d_i}^4$$

Soit U la variable quantitative définie par les $n = 5$ valeurs

`12 17 14 16 15`

correspondant à l'étude du poids de matière séche pour un arbre de type pin du Canada, l'unité de mesure étant le gramme.

Donner les expressions Dbase qui calculent $moy(U)$, $ect(U)$, $sk(U)$, $ku(U)$.

Donner ensuite avec un seul chiffre après la virgule les valeurs numériques associées. Donner enfin les valeurs numériques associée pour la série théorique V définie par les valeurs `10 11 12 13 14` et pour la série théorique W définie par les valeurs `10 11 12 11 10`.

Pour les plus fort(e)s, écrire un programme skku.prg qui demande le nom d'une base de données, d'un champ et qui calcule les coefficients d'aplatissement et d'asymétrie de ce champ pour la base de données.

Question annexe : A quoi servent ces coefficients ? Comment s'en sert-on ?

Esquisse de SOLUTION

1. Un remplacement vite fait

Pour créer un champ sous *Dbase*, il faut modifier la structure de la base courante. On réalise cette opération pour la base demandée via les instructions

```
. use BU11QT
. modify structure
```

Il est bien sur possible de taper `modi stru` au lieu de `modify structure` puisque ce sont des abbréviations légales. A la suitre de `modi stru` on se positionne en fin de structure et on définit le nouveau champ (numéro 8) en donnant son nom : `DURPAL`, son type : `N` (pour Numérique) et sa longueur. On peut mettre ici 7 car le nombre d'heures est peu élevé. De façon plus fine, une étude de la base BullQT montre que le plus grand nombre d'heures est 96 et donc que la plus grande valeur possible pour `DURPAL` est 5761 donc 5 chiffres auraient suffi.

Lorsque le champ `DURPAL` est défini, la commande

```
replace all DURPAL with PALH*60+PALM
```

permet de le remplir correctement (par défaut, *Dbase* met 0, comme pour tout numérique).

Si on veut ensuite calculer la moyenne du palier, on peut

- soit calculer la moyenne de `DURPAL` et l'exprimer en heures et minutes,
- soit calculer la moyenne de `PALH` (disons mh), calculer la moyenne de `PALM` (disons mm), et en déduire la moyenne du palier via la formule $60 * mh + mm$ car la moyenne est un opérateur linéaire,
- soit calculer la moyenne de l'expression sur champ `60*PALH+PALM` comme le permet *Dbase*.

Pour formater le résultat, on peut utiliser la fonction `STR` qui convertit un numérique en chaine de caractères avec un nombre au choix de décimales.

Pour exprimer la moyenne en heures et minutes, on peut utiliser la fonction `INT` qui permet d'obtenir la partie entière.

Dans tous les cas de figure, on trouve une moyenne de 5 heures et un peu plus de 5 minutes, ce que montre le "log" de la session *Dbase* suivante :

```
. use bullqt

. modi stru

**********
**********  on utilise les commandes
**********  de l'éditeur pour ajouter
**********  le champ DURPAL
**********

     45 enregs ajouté(s)

** remplissage de DURPAL

. replace all DURPAL with PALH*60+PALM
     45 enregs remplacé(s)

** moyenne : méthode 1

. average durpal
     45 enregs moyenné(s)
durpal
   305

. average durpal*1.0
     45 enregs moyenné(s)
durpal*1.0
     305.2

** moyenne : méthode 2

. average palh*1.000 to mh
     4.978

. average palm*1.000 to mm
     6.556
. ? mh*60+mm
   305
```

```
** moyenne : méthode 3

. average palh*60 + palm to mp
     45 enregs moyenné(s)
palh*60 + palm
   305

** affichage avec deux décimales

. ? str(mp,7,2)
   305.22

** affichage avec deux décimales sans choix de longueur

. ? str(mp,,2)
  305.22

** calcul du nombre d'heures

. store int(mp/60) to nbh
  5

** calcul du nombre de minutes

. store int(mp-60*nbh)
  5
```

2. Comparaison de moyennes et de pourcentages

Il n'y a aucune difficulté à comparer des moyennes, des pourcentages si l'on connait les formules statistiques associées. On pourra les relire à l'adresse

`http://www.info.univ-angers.fr/pub/gh/wstat/formules.ps`

Pour comparer la moyenne d'age des hommes et des femmes dans le dossier `Elf`, nous fournissons une copie de la session Dbase réalisée à l'aide des commandes `SET ALTER TO ...` et `SET ALTER ON` ce qui nous a fourni le fichier `compdb1.txt` :

```
. set alter to compdb1.txt
. set alter on

. use elf
. count to nbh for sexe=0
     35 enregs

. count to nbf for sexe=1
     64 enregs

. sum age/nbh for sexe=0 to mah
     35 enregs sommé(s)
            36.40

. sum age/nbf for sexe=1 to maf
     64 enregs sommé(s)
            35.52

. sum age*age/nbh for sexe=0 to mt2ah
     35 enregs sommé(s)
        1602.17

. sum age*age/nbf for sexe=1 to mt2af
     64 enregs sommé(s)
        1581.27
```

```
. store mt2ah - mah*mah to vah
                 277.2114

. store mt2af - maf*maf to vaf
                 319.9060

. store sqrt(vah) to eah
      16.6497

. store sqrt(vaf) to eaf
      17.8859

. store abs(mah-maf) to dim
        0.88

. store (vah/nbh) + (vaf/nbf) to varp
            12.9189

. store dim/sqrt(varp) to eps
            0.2461

. close alter
```

Au seuil de 5 %, soit la valeur 1.96, on peut donc accepter l'hypothèse que les hommes et les femmes ont la même moyenne d'age.

A l'aide d'instructions similaires pour les deux autres comparaisons, on arrive aux résultats suivants : il y a 35 hommes dans le dossier ELF dont 17 ont fait des études supérieures et 64 femmes dont 22 ont fait des études supérieures.

La comparaison de pourcentage correspondant aux formules citées se résume par le tableau

ia	17	na	35	pa	0.486
ib	22	nb	64	pb	0.344
ii	39	nn	99	p	0.394

On obtient donc une différence réduite d' à peu-près 1.38 et donc au risque de 5 % on peut accepter l'hypothèse que les pourcentages correspondent à une même population.

Pour la comparaison de moyennes, on trouve de même

Variable	nbVal	Moyenne	Variance	Ecart-type	Cdv
A	17	31.235	121.239	11.011	35 %
B	22	32.591	144.514	12.021	37 %

soit une différence réduite : 0.3662 et donc au seuil de 5 % soit 1.96, on peut accepter l'hypothèse d'égalité des moyennes.

3. Calcul de médiane

Pour calculer la moyenne du champ `XU`, la commande `average` de *Dbase* suffit. Pour la variance, nous utilisons la formule "moyenne des carrés moins carré de la moyenne", ce qui demande donc une commande `average` et une commande `store` : l'écart-type se calcule alors simplement à l'aide la racine carré qui se dit `sqrt` en *Dbase*, d'où les instructions :

```
. use mstar

. list stru

Structure de la base de données: D:mstar.dbf
Nombre total d'enregistrements :      5
Date de la dernière mise à jour: 24/04/01
Champ  Nom champ   Type         Dim    Dec
    1  XU          Numerique      3
    2  XV          Numerique      3
    3  XW          Numerique      3
** Total **                      10

. list

  Enreg.    XU  XV  XW
      1     12  10  10
      2     17  11  11
      3     14  12  12
      4     16  13  11
      5     15  14  10

. average xu to moy_xu
      5 enregs moyenné(s)
 xu
 15

. average xu*xu to mct_xu
      5 enregs moyenné(s)
  xu*xu
    222
```

```
. store mct_xu-moy_xu*moy_xu to var_xu
                   3

. store sqrt(var_xu) to ect_xu
        1.72

. disp memo

MOY_XU      pub   N     15    (   14.80000000)
MCT_XU      pub   N    222    (  222.00000000)
VAR_XU      pub   N      3    (    2.96000000)
ECT_XU      pub   N      1.72 (    1.72046505)
```

Ensuite, on trie les valeurs et puisqu'il y en a un nombre impair (n=5) on va à l'aide de l'instruction `goto` au milieu des valeurs pour récupérer la médiane :

```
. use mstar

. sort on xu to mxu
  100% ;  5 enregistrements triés

. use mxu

. list

 Enreg.     XU  XV  XW
      1     12  10  10
      2     14  12  12
      3     15  14  10
      4     16  13  11
      5     17  11  11

. count to n
      5 enregs

. goto (n+1)/2

. store xu to med_xu
  15
```

Au lieu de recommencer les calculs pour les autres valeurs fournies, nous préférons écrire un programme que nous nommons `median.prg`. Voici son contenu :

```
*  median.prg : calcul de médiane

CLEAR
SET talk   off
SET safety off


?  "median.prg "
?  "                         Cpy. (gH) Gilles HUNAULT, 1997"
?  "                         gilles.hunault@univ-angers.fr"
?

** panneau de saisie du nom de la base

STORE dbf() + replicate(" ",20) TO nombase
STORE substr(nombase,1,20) TO nombase
@ 8,03 SAY " Nom de la base ? : "
@ 9,03 SAY " (ou Aide pour explications)"
@ 8,50 GET nombase
READ

IF   substr(upper(nombase+"    "),1,4) = "AIDE"
.OR.  len(trim(ltrim(nombase)))=0
     clear
     ? "calcul de la médiane"
     ?
     ? " Vous devez disposer d'une base de données au sens Dbase"
     ? " (fichier de type .DBF) dont la structure doit être la suivante :"
     ?
     ? "   - un premier champ de type Caractère de dimension 4"
     ? "   - tous les autres champs sont numériques. "
     ?
     ? " le champ que vous voulez utiliser doit être de nature QT."
     ?
     RETURN
ENDIF
```

```
** préparation des noms de fichier

STORE trim(ltrim(nombase)) TO nombase
IF at(".DBF",upper(nombase)) = 0
    STORE nombase+".DBF" TO NOMComplet
ELSE
    STORE nombase        TO NOMComplet
    STORE at(".DBF",upper(nombase)) TO ip
    STORE substr(nombase,1,ip-1) TO nombase
ENDIF

IF .NOT. FILE(NOMComplet)
   ? chr(7)
   ? chr(7)
   ? " Désolé, je ne vois pas ce fichier. Vérifiez son existence avec DIR "
   ? " Ne tapez pas .DBF mais indiquer le chemin d'accés ( PATH ) "
   ?
   ? " Fin anormale d'exécution, code 1 : fichier non trouvé."
   ?
   RETURN
ENDIF

USE   &NOMComplet
STORE Nombase+".med" to Nomsor
DISP  STRU
ACCEPT "Nom de la variable à traiter ? " to nomVar

SET   ALTER  TO &Nomsor
SET   ALTER  ON
SET   SAFETY OFF

** affichage du titre

?
? " Médiane de la variable ", nomVar," base ",Nombase
?

** on trie suivant la variable

sort on &nomvar to bazTmp
```

```
** on compte le nombre de valeurs

USE bazTmp
COUNT to nbval

** s'il y en a un nombre impair, la médiane
** est la valeur du milieu

STORE int(nbVal/2) to milieu
IF .not.  2*milieu=nbVal
   goto 1+milieu
   store &nomVar to median
   ? " il y a un nombre impair de termes "
   ?" la médiane est donc ",median
ELSE
   ** sinon on fait la demi-somme des valeurs du milieu
   goto milieu
   store &nomVar to val1
   skip
   store &nomVar to val2
   store (val1+val2)/2 to median
   ? " il y a un nombre pair de termes "
   ? " la médiane est donc la demi-somme de ",val1," et ",val2
   ?" la médiane est donc ",median
ENDIF
USE

?
? " -- fin de median "
?
close alter
?
? " Résultats dans ",nomsor
?
set talk   on
set safety on
```

4. Coefficients d'asymétrie et d'aplatissement

Si l'on veut effectuer ces calculs directement sous *Dbase* en mode interactif, il faut commencer par calculer la moyenne et l'écart-type avant d'effectuer la sommation à l'aide de la commande `sum` (la double étoile signifie la puissance), soit les instructions :

```
. use mstar

** reprise du calcul de moyenne et écart-type

. average xu to moy_xu
      5 enregs moyenné(s)
 xu
 15

. average xu*xu to mct_xu
      5 enregs moyenné(s)
  xu*xu
    222

. store mct_xu-moy_xu*moy_xu to var_xu
                     3

. store sqrt(var_xu) to ect_xu
         1.72

** calcul de skewness et kurtosis

. sum ( (xu-moy_xu)/ect_xu) )**3/nbval to sk_xu
      5 enregs sommé(s)
               -0.395

. sum ( (xu-moy_xu)/ect_xu) )**4/nbval to sk_xu
      5 enregs sommé(s)
                1.993
```

Pour automatiser ces calculs, il est judicieux d'écrire un programme que nous listons ici :

```
*  skku.prg : calcul de skewness et kurtosis

clear
set talk   off
set safety off


  ?  "skku.prg "
  ?  "                                  Cpy. (gH) Gilles HUNAULT, 1997"
  ?  "                                   gilles.hunault@univ-angers.fr"
  ?

  clear memory

  store dbf() + replicate(" ",20) to nombase
  store substr(nombase,1,20) to nombase
  @ 8,03 say " Nom de la base ? : "
  @ 9,03 say " (ou Aide pour explications)"
  @ 8,50 get nombase
  read

  IF substr(upper(nombase+"    "),1,4) = "AIDE"
 .OR. len(trim(ltrim(nombase)))=0
     clear
     ? "calcul de skewness et kurtosis"
     ? "            "
     ? " Vous devez disposer d'une base de données au sens Dbase"
     ? " (fichier de type .DBF) dont la structure doit être la suivante :"
     ?
     ? "   - un premier champ de type Caractère de dimension 4"
     ? "   - tous les autres champs sont numériques. "
     ?
     ? " le champ que vous voulez utiliser doit être de nature QT."
     ?
     RETURN
  ENDIF

  store trim(ltrim(nombase)) to nombase
```

```
 IF at(".DBF",upper(nombase)) = 0
   store nombase+".DBF" to NOMComplet
 ELSE
   store nombase        to NOMComplet
   store at(".DBF",upper(nombase)) to ip
   store substr(nombase,1,ip-1) to nombase
 ENDIF


IF .NOT. FILE(NOMComplet)
  ? chr(7)
  ? chr(7)
  ? " Désolé, je ne vois pas ce fichier. Vérifiez son existence avec DIR "
  ? " Ne tapez pas .DBF mais indiquer le chemin d'accés ( PATH ) "
  ?
  ? " Fin anormale d'exécution, code 1 : fichier non trouvé."
  ?
  RETURN
ENDIF

 USE   &NOMComplet
 STORE Nombase+".sku" to Nomsor
 DISP  STRU
 accept "Nom de la variable à traiter ? " to nomVar

 SET   ALTER  TO &Nomsor
 SET   ALTER  ON
 SET   SAFETY OFF
 ?
 ? " Calcul de skewness et kurtosis pour la variable ", nomVar," base ",Nombase
 ?

 ** calcul de la moyenne et de l'écart-type

 AVERAGE &nomVar          TO moy
 AVERAGE &nomVar*&nomvar TO mct
 STORE mct-moy*moy        TO var
 STORE SQRT(var)          TO ect

 ** on base la base de données en revue
 ** on calcule au fur et à mesure la valeur centrée réduite
```

```
  STORE 0 TO s_sk
  STORE 0 TO s_ku
  GOTO 1
  DO WHILE .NOT. EOF()
     STORE &nomVar         TO laval
     STORE (laval-moy)/ect TO cr
     STORE s_sk + cr**3    TO s_sk
     STORE s_ku + cr**4    TO s_ku
     SKIP
  ENDDO
  COUNT TO nbVal
  STORE s_sk/nbVal TO sk
  STORE s_ku/nbVal TO ku

  ? " Nombre de valeur " , nbVal
  ? " Moyenne          " , moy
  ? " Ecart-type       " , ect
  ? " Asymétrie        " , sk
  ? " Aplatissement    " , ku
  ?

  disp memo

?
? " -- fin de skku "
?
close alter
?
? " Résultats dans ",nomsor
?
set talk   on
set safety on
```