T.D. de Probabilités et Statistiques

Licence Informatique

Gilles HUNAULT, 1998

T.D. 5 : Statistiques Descriptives

1. Types de variables (QT,QL,QX...) en Statistiques Descriptives
2. Calculs concrets de QT : temps de transit
3. Calculs concrets de QL : bande passante
4. Valeurs de a et b pour $|\rho|$ =1
5. Programme de test si d est une distance

1. Types de variables (QT,QL,QX...)

Enoncé

Dans le cours, on utilise principalement les variables QT (quantitatives) et QL (qualitatives). Peut-il y avoir d'autres types de variables ? On pourra par exemple imaginer que les variables correspondent à des questions pour un questionnaire de type enquête, ou que les variables sont des mesures issues de capteurs...

Solution

Il y a bien sûr de nombreux autres types de variable. Citons

les **QM**	ou *variables multi-réponses*,
les **QP**	ou *variables pourcentages*,
les **QS**	ou *variables (anti-)scores*,
les **QF**	ou *variables floues*,
les **QH**	ou *variables hiérarchiques*,
les **QE**	ou *variables d'énonciation.*
les **QX**	ou *variables textuelles.*

Pour la plupart de ces variables, les calculs se ramènent à des calculs de comptages comme pour les **QL** ou des sommation pondérées comme pour les **QT**. On laisse au lecteur le soin de vérifier ces dernières affirmations sur des exemples concrets et d'essayer de trouver quel affichage "intelligent" doit alors être utilisé.

2. Calculs concrets de QT : temps de transit

Enoncé

Soit T la variable quantitative "*temps de transit*" exprimée en minutes dont les valeurs sont, dans l'ordre, $[97, 12, 192, 25, 48]$. Soit maintenant D la variable quantitative "*durée de transport*" exprimée en heures et dont les valeurs sont, dans l'ordre $[16, 2, 32, 4, 8]$. Effectuez l'analyse séparée puis conjointe des ces deux variables. On présentera les résultats suivant un ordre "intelligent". Calculer aussi le coefficient de corrélation et, si besoin est, les coefficients a et b de la relation linéaire correspondante à savoir T$=a.$D$+b$.

Solution

Tous calculs numériques effectués, on obtient les résultats suivants (une fois D convertie en minutes, soit les valeurs $[960, 120, 1920, 240, 480]$) :

Variable	Moyenne	Ecart-type	σ/m
D	744.0	654.6	87 %
T	74.8	65.4	87 %

Le coefficient de corrélation liénaire ρ vaut ici 1 ce qui indique une liaison linéaire. Utilisant les formules $a = \rho.\sigma_Y/\sigma_X$ et $b = m_y - a.m_X$ pour la liaison $Y = a.X + b$ avec Y =D et X =R, on obtient D = 10,01.T-4,75. L'ordre d'affichage est ici indifférent puisqu'on a le même coefficient de variation.

3. Calculs concrets de QL : bande passante

Enoncé

Soient $[10, 12, 17, 12, 10, 17, 10]$ les valeurs (codées) de la variable qualitative B "*Bande Passante*" où 10 correspond à la gamme FM, 12 à la gamme UHF et 17 à la gamme VHF. On considère également la variable R "*Type de Radio*" dont les valeurs codées sont $[1, 1, 1, 2, 2, 2, 2]$. Le code 1 signifie "Radio de qualité moyenne" et le code 2 "Radio de qualité supérieure". Effectuez l'analyse séparée puis conjointe des ces deux variables. On présentera les résultats suivant un ordre "intelligent". Peut-on parler de liaison entre B et R ?

Solution

Le comptage à la main des différents croisements possibles donne le tri croisé suivant (effectifs absolus) :

	FM	UHF	VHF	Total
R q.moyenne	1	1	1	3
R q.superieure	2	1	1	4
Total	3	2	2	7

Puisque 1/7 est à peu-près 14.3 %, on peut donc présenter les variables suivant le classement suivant

Variable	Mode	%	Autres modalites		
R	q. moyenne	57	q. superieure (43 %)		
B	bande FM	43	UHF	(29 %)	VHF (29 %)

Compte-tenu des faibles effectifs (7 valeurs en tout), il est difficile de tirer des conclusions.

4. Valeurs de a et b pour $|\rho|$ =1

Enoncé

Dans le cours, on affirme que si $|\rho(X,Y)| = 1$ alors X et Y sont liés par la relation linéaire $Y = a.X + b$ avec $a = \rho(X,Y).\sigma(Y)/\sigma(X) = (m(X).m(Y) - m(XY))/d$ et $b = m(Y) - a.m(X)$
$= (m(X).m(XY) - m(Y)mt(X,2))/d$ où $d = m(X)^2 - mt(X,2)$. Démontrez ces formules.

Solution

Si $Y = a.X + b$ cela signifie que pour tout i, $Y_i = a.X_i + b$. Appliquant l'opérateur moyenne à ces relations, c'est à dire en sommant et en divisant par n, on obtient l'équation E : $moy(Y) = a.moy(X) + b$. De même, si $Y_i = a.X_i + b$ pour tout i, en multipliant par x_i puis en sommant et en divisant par n, on obtient l'équation F : $moy(XY) = a.moy(X^2) + b.moy(X)$. X et Y étant donnés, (E,F) est un système de deux équations linéaires en les deux inconnuées a et b dont la résolution donne les valeurs indiquées avec d. Pour revenir à $a = \rho(X,Y).\sigma(Y)/\sigma(X)$, il suffit de remarquer que $d = V(X) = \sigma(X)^2$ et que $m(X).m(Y) - m(XY) = cov(X,Y)$ et puisque $\rho(X,Y) = cov(X,Y)/\sigma(Y)/\sigma(X)$, le tour est joué!

5. Programme de test si d est une distance

Enoncé

Soit $X = (X[i,j])$ est un tableau de données où les n individus i sont en ligne et les variables j sont en colonne, on note $d(j_1,j_2)$ la valeur d'une fonction définie sur JxJ où J représente l'ensemble des colonnes. Par exemple $d(j_1,j_2) = \rho((j_1,j_2)$. On suppose que les $d(j_1,j_2)$ sont données par un un tableau nommé tabdist tel que tabdist[j,k]=$d(j,k$. Ecrire un programme Awk qui teste si tabdist peut être considéré comme une matrice triangulaire inféreure de distances.

Solution

```
# tstdist.awk : teste si le fichier peut être
#               considéré comme une matrice triangulaire
#               inféreure de distances

BEGIN { fs = "tstdist.sor"
        print "\n print " Etude de " ARGV[1]\n"
} # fin du  Begin

# lecture : on transfère les données dans le tableau tabdist

(FNR>1) {
  j = FNR-1 ;
  nomelt[j] = $1
  for (i=2;i<=NF;i++) {     k = i-1
      tabdist[j,k] = $i ; tabdist[k,j] = $i
      if ($i<0) {
         if (nbneg==0) {
            nbneg = 1 ; ilig = j ; jcol = k
         } # fin si nbneg = 0
      } # fin si $i <0
  } # fin pour i
} # fin de lecture après la ligne 1
```

```
END{

  if (nbneg>0) {
     print
     print "L'élément en ligne " ilig " colonne " jcol " est négatif"
     print "puisqu'il vaut " tabdist[ilig,jcol] " ; votre fichier " FILENAME
     print "ne peut donc PAS être considéré comme le fichier d'une matrice "
     print "triangulaire inférieure."
     print
    exit
  } # fin de si

  nbelt = FNR - 1

  print " test de positivité OK "

  # 1. Vérification de d(x,x)=0, soit 0 sur la diagonale

  ok1 = 1
  i   = 1
  while (i<=nbelt) {
    # print i " : " tabdist[i,i] " - "
    if (tabdist[i,i]!=0) {
    print
    print "L'élément sur la diagonale en position " i " est non nul"
    print "puisqu'il vaut " tabdist[i,i] " ; votre fichier " FILENAME
    print "ne peut donc PAS être considéré comme le fichier d'une matrice "
    print "triangulaire inférieure de distances."
    print

       i   = nbelt + 1
       ok1 = 0
    } # fin si
    i++
  } # fin tant que sur i
```

```
  if (ok1==1) {
     print " test de diagonale  OK "
     # 2. Vérification de l'inégalité triangulaire
     ok2 = 1 ;
     i   = 1
      while (i<=nbelt) {
        j = 1
        while (j<=nbelt) {
          j++ ;
          k = 1
          while (k<=nbelt) {
            ## test de l'inégalité proprement dite
            x = tabdist[i,j] ; y = tabdist[i,k] ; z = tabdist[k,j]
            if (x>y+z) {
                ok2 = 0
print
print "    Les éléments x = " x " ligne " i " colonne " j
print "                 y = " y " ligne " i " colonne " k
print "                 z = " z " ligne " k " colonne " j
print "    ne vérifient pas l'inégalité triangulaire puisque "
print "    " x " > " y+z " = " y " + " z " soit x > y + z "
print "    c'est à dire d(i,j) > d(i,k) + d(k,j) "
print "    pour i = " i", j = " j " et k " k"."
print " "
print "    Votre fichier " FILENAME
print "    ne peut donc PAS être considéré comme le fichier d'une matrice "
print "    triangulaire inférieure de distances."
print
               i = nbelt+1 ;
               j = nbelt+1 ;
               k = nbelt+1
            } # fin de si sur (x>y+z)
            k++
          } # fin pour k
        } # fin pour j
        i++
      } # fin pour i
  } # fin si ok1==1
```

```
  if (ok2==1) {
print " test triangulaire  OK "
print " "
print " D’après les test effectués, il semble bien que "
print "    votre fichier " FILENAME
print "    PEUT être considéré comme le fichier d’une matrice "
print "    triangulaire inférieure de distances."
print
  } # fin de si

  print " "
  print "  -- fin de tstdist "
  print " "

} # fin du END
```