T.D. de Probabilités et Statistiques

Licence Informatique

Gilles HUNAULT, 1998

T.D. 7 : Corrélation de rangs, Compararaisons...

1. Corrélation des rangs de Kendall
2. Algorithmes de comparaison de moyennes et pourcentages
3. Comparaison de moyennes : durées de tri
4. Coefficient de corrélation des rangs
5. Nombre d'appels sur une "hotline"
6. Calculs sur la loi de Mendel
7. χ^2 d'indépendance pour une vente de livres

1. Corrélation des rangs de Spearman et Kendall

Enoncé

Calculer le coefficient de corrélation des rangs de Spearman et de Kendall pour les valeurs A et B correspondants à des rangs de préférence pour 6 types de petits gâteaux pour le gouter

Numero du gateau	Rang A	Rang B
1	2	2
2	5	4
3	6	1
4	3	5
5	1	3
6	4	6

On rappelle que si ε_i est la différence entre les rangs A_i et B_i, ρ_S vaut $1-6.(\Sigma\ \varepsilon_i^2)/n.(n^2-1)$ où n est le nombre d'objets à hiérarchiser. On donnera les algorithmes de calcul des deux coefficients supposant n donné, les rangs de A et B étant mis dans des tableaux tels que $A[i]$ et $B[i]$ correspondent à A_i et B_i.

On rappelle aussi que si r_i désigne le nombre d'inversions ($B_j > A_i$) entre A et B après la valeur i lorsque A est trié par ordre croissant, ρ_K vaut $-1 + 4.(\Sigma\ r_i)/n.(n-1)$.

Solution

Les ε_i valent ici respectivement 0,1,5,2,2,2 donc la somme de leur carré est 38 pour n=6, d'où ρ_S=1-6.38/6.35 soit -0,086.

L'algorithme du calcul est simplement

```
sdc ← 0
pour i de1a n
     eps ← A[i]-B[i]
     sdc ← sdc + epseps
finpour
denom ← n(n n-1)
rhos  ← 1 -6sdc/denom
```

Si on trie A par ordre croissant et B en conséquence, les rangs deviennent

```
Rang A   1  2  3  4  5  6
Rang B   3  2  5  6  4  1
```

et les nombres r_i sont respectivement 3,3,1,0,0 de somme 7. ρ_K vaut donc $-1 + 4.7/6.5$ soit -0,067. On remarquera que dans les deux calculs de corrélation des rangs ρ est négatif ce qui semble indiquer plutot un désaccord entre A et B... Un algorithme possible de calcul avec tri des valeurs de A et B est `/* kendall : on commence par le calcul des inversions */`

```
   r = 0
   pour j de 1 a  n-1
      x = b.j
      pour k de j+1 a n
         si b.k > x alors
            r = r + 1
         finsi
      finpour sur k
   finpour sur j
   rk = (4*r/(n*(n-1))) -1
   écrire " Kendall rk = " rk

 /* spearman : calcul direct */

   s = 0
   pour j de 1 a n
      x = b.j
      y = a.j
      s = s + (x-y)*(x-y)
   finpour
   d = n*(n*n-1)
   écrire " n*(n*n-1)  " d
   rs = 1 - 6*s/d
   écrire "Spearman s = " s " et rs = "  rs
```

2. Algorithmes de comparaisons

Enoncé

Donner l'algorithme du calcul de comparaison de pourcentages et du calcul de comparaison de moyennes pour des variances connues. On ne supposera pas la fonction racine existante.

Solution

```
/* compourc.rex : comparaison de pourcentages */


   ia ← ...
   na ← ...
   ib ← ...
   nb ← ...

   pa ← ia/na
   pb ← ib/nb
   p  ←(ia+ib)/(na+nb)
   df ← pa-pb
   si df <0 alors
      df ← -1*df
   finsi
   q ← p*(1-p)*(1/na+1/nb)
   r ← racine(q)

   écrire "ia " ia " na " na " pa " pa
   écrire "ib " ib " nb " nb " pb " pb
   écrire "ii " ia+ib " nn " na+nb " p " p
   écrire "df " df " r2 " q " r " r " r*r " r*r
   écrire "eps" df/r

   écrire " résultat du test : " df*racine(10)/r
```

```
/* sous-programme  */

   racine:
    arg x
    si x> 1 alors rr ← 1
            else rr ← 0.9
    do i ← 1 to 10
       rr ← (x/(2*rr)) + rr/2
    fin
   renvoyer rr

/* compmoy.rex : comparaison de moyennes */

   n ← ...

   a.0 ← ...
   a.1 ← ...
   ...
   a.n ← ...

   na ← 0
   nb ← 0
   sa ← 0
   sb ← 0
   ca ← 0
   cb ← 0
   do i ← 0 to n
      x.i ← i
      na ← na + a.i
      nb ← nb + b.i
      sa ← sa + x.i*a.i
      sb ← sb + x.i*b.i
      ca ← ca + x.i*x.i*a.i
      cb ← cb + x.i*x.i*b.i
   fin
```

```
ma ← sa/na
mb ← sb/nb
va ← (ca/na) - (ma*ma)
vb ← (cb/nb) - (mb*mb)
ea ← racine(va)
eb ← racine(vb)
df ← ma - mb
si df < 0 alors df ← -df
d ← (va/na)+(vb/nb)
r ← racine(d)
delta ← df/r

écrire " A  na = " na " ma = " ma " va = " va " ea = " ea
écrire " B  nb = " nb " mb = " mb " vb = " vb " eb = " eb
écrire " |ma-mb| " df " r = racine( " d " ) = " r
écrire " delta = " delta
```

3. Comparaison de moyennes entre ordinateurs

Enoncé

On dispose de deux ordinateurs (A) et (B). Des simulations de taille de fichiers pour des calculs de tris en mémoire ont amenés aux valeurs suivantes pour les fichiers temporaires

taille (Meg)	ordi. A	ordi. B
3	6 fois	8 fois
4	2	3
8	5	11
11	9	7

Effectuez une comparaison de moyennes.

Solution

Tous calculs effectués, on trouve $n_a = 22$, $m_a = 7.5$, $v_a = 11.7045455$ et $n_b = 29$, $m_b = 6.93103448$, $v_b = 9.5814507$ d'où $|m_a - m_b| = 0.56896552$, $r = \sqrt{0.862419647} = 0.928665519$, $\delta = 0.612670018$. Ce nombre est inférieur à 1.96 donc on peut conclure que les deux ordinateurs se comportent de la même façon sur cet exemple.

4. Coefficient de corrélation des rangs

Enoncé

Montrer que le coefficient de *Spearman* est en fait le coefficient usuel de corrélation linéaire. Donner les plages de variation de ρ_K et ρ_S.

Solution

Soient $R = r_i$ et $S = s_i$ les rangs pour n objets de 1 à n. Alors $m(R) = m(S) = (n+1)/2$ et $\sigma(R) = \sigma(S) = \sqrt{(n^2-1)/12}$ car R et S correspondent à la variable uniforme discrète $\mathcal{UD}(n)$. Posons $\varepsilon_i = r_i - s_i$. Comme $\Sigma\ (r_i - s_i)^2 = \Sigma\ r_i^2 + \Sigma\ s_i^2 - 2\Sigma\ r_i s_i$, on en déduit que $\Sigma\ \varepsilon_i^2 = 2n(n+1)(2n+1)/6 - 2\Sigma\ r_i s_i$. D'autre part, $\rho = cov/\sigma_X\sigma_Y = \Sigma\ r_i s_i/n - m_X m_Y$ donc $\Sigma\ r_i s_i = n\rho\sigma_X\sigma_Y - nm_X m_Y$. Reportant cette valeur dans l'équation précédente, on obtient $\Sigma\ \varepsilon_i^2 = n(n+1)(2n+1)/3 - 2\ (\ \rho n(n^2-1)/12 - n(n+1)^2/4\ )$. Développant ces calculs, avec $n(n+1)/12$ en facteur, on a $\Sigma\ \varepsilon_i^2 = n(n+1)/12\,[\ 4(2n+1) - 6(n+1)] - n\rho(n^2-1)/6\ ]$, soit finalement $\Sigma\ \varepsilon_i^2 = n(n^2-1)/6\ (\ 1-\rho\ )$. On peut donc écrire le ρ normal comme $1 - 6\Sigma\ \varepsilon_i^2/n(n^2-1)$ ce qui justement la définition de ρ_s.CQFD.

Si on a les mêmes rangs, alors a_i=b_i donc $\varepsilon_i = 0$ donc $\rho_S = 1$. Ou encore : $S = R$ donc $\rho(R,S) = 1$. Par contre, si les choix sont inverses : $S = n+1-R$ donc $\rho(R,S) = -1$, la démonstration directe avec $b_i = n+1-a_i$ étant plus longue...

Si on a les mêmes rangs, alors a_i=b_i donc il y a $n-1$ inversions puis $n-2$ puis... jusqu'à 1. Alors S vaut n(n-1)/2 d'où 4/D vaut 2 et donc $\rho_K = 1$. Par contre si les choix sont inverses il n'y a aucune inversion donc $\rho_K = 1$.

5. Nombre d'appels sur une "hotline"

Enoncé

On compte le nombre d'appels obtenus sur une "ligne chaude" pour l'aide en ligne d'un nouveau logiciel de gestion. On obtient les valeurs suivantes

nb jours	2	3	4	5	6	7	8	9
nb appels	03	03	05	38	39	75	26	01

Lors de la dernière mise à jour du logiciel, cette même ligne avait enregistré 115 appels en tout soit une moyenne de 6 jours avec un écart-type de 1.2083. Peut-on comparer l'utilisation de la ligne ?

Solution

Bien sûr qu'on peut comparer ! σ=1.2083 donne $V = \sigma^2 = 1.46$, et pour cette année, $\Sigma\ n_i = 190$, $m_a = 6.32$, $v_a = 1.587$ donc $\delta = 0.32/\sqrt{0.0211} = 2.20$. Au seuil de 5 %, la différence est donc significative.

6. Calculs sur la loi de Mendel

Enoncé

Le brillant généticien Mendel avait introduit un schéma de répartition 9/3/3/1 pour décrire la couleur dominante et la couleur récessive de certaines espèces florales. Des élèves de Deug Biologie à l'Université d'Angers décident de vérifier cela sur un exemple. Pour cela, ils prennent des fleurs de balsamine avec un croisement blanc/pourpre. Ils obtiennent les valeurs suivantes

couleur	pourpre	rose	blanc/lavande	blanc
nb. fleurs	1790	547	548	213

le schéma de Mendel est-il vérifié ?

D'autres élèves penchent plutôt pour une équirépartition. Leurs données sont

couleur	pourpre	rose	blanc/lavande et blanc
nb. fleurs	1065	326	457

Peut-on leur donner raison ? ou doit-on encore prendre le schéma de Mendel ?

Solution

Le total général est 3098 et la répartition théorique de Mendel donnerait comme effectifs 1743, 581, 581 et 194 (ou 193). Le calcul du χ^2 donne 7 alors que le χ^2 théorique maximal autorisé à 5 % est 7.81. Le schéma est donc vérifié.

Pour le deuxième jeu de valeur, le total général est 1829 et la répartition théorique de Mendel donnerait comme effectifs 1029, 343 et 457 (343+114). Le calcul du χ^2 donne 2.89 alors que le χ^2 théorique maximal autorisé à 5 % est 5.99 Le schéma est donc encore vérifié. Par contre l'équirépartition donnerait 610 comme effectif moyen d'où un χ^2 de 520 qui est inacceptable.

7. χ^2 d'indépendance pour une vente de livres

Enoncé

Une enquête récente a fourni les données suivantes relatives au croisement lieu de vente (Grand Magasin sauf Fnac, Fnac, Vente par correspondance, Librairie Générale) et famille de livres (notée par souci de confidentialité famille A, B, C, D, E et F) :

	GM	Fnac	Vpc	Lib.G
A	1	0	0	0
B	11	4	3	6
C	11	1	7	0
D	6	4	11	4
E	7	12	11	10
F	1	1	1	3

Y a-t-il un lien entre lieu de vente et famille de livre ?

La vente de tels livres est-elle équirépartie par lieu ?

Solution

L'utilisation du programme Awk vu en cours donne comme résultat :

```
6 lignes et 4 colonnes
   1     0     0     0  +     1
  11     4     3     6  +    24
  11     1     7     0  +    19
   6     4    11     4  +    25
   7    12    11    10  +    40
   1     1     1     3  +     6
+++++++++++++++++++++++++++++++++
  37    22    33    23  +   115

   0.322     0.191     0.287     0.200   +     1.000
   7.722     4.591     6.887     4.800   +    24.000
   6.113     3.635     5.452     3.800   +    19.000
   8.043     4.783     7.174     5.000   +    25.000
  12.870     7.652    11.478     8.000   +    40.000
   1.930     1.148     1.722     1.200   +     6.000
++++++++++++++++++++++++++++++++++++++++++++++++++++
  37.000    22.000    33.000    23.000   +     0.000

 chi-deux 28.151
```

Le chi-deux maximal théorique pour $(l-1)*(c-1)$=5*3=15 degrés de libertés vaut 25. Notre χ^2 lui est supérieur donc on peut dire qu'il n'y a pas indépendance, ou qu'il y a liaison, la plus forte contribution étant 3.91, fournie par la ligne 3, colonne 1.

Puisque le total général est 115, s'il y avait équirépartition de la vente des livres par lieu, chaque lieu aurait 28,75 livres. Conservant cette valeur pour les TH, le χ^2 d'ajustement est alors 5.73 alors que le χ^2 théorique maximal est 7.81 pour 3 ddl. Une déformation "raisonnable" des valeurs pour avoir des nombres théoriques entiers pourrait être de prendre comme répartition 30 28 29 28 ce qui fournit un χ^2 de 4.63 ; là encore, on peut accepter l'hypothèse d'équirépartition.