T.D. de Probabilités et Statistiques

Licence Informatique

Gilles HUNAULT, 1998

T.D. 6 : Statistiques Descriptives

1. Asg de QT : dossier VINS
2. Asg de QL : dossier ELF88QL
3. Calculs concrets de χ^2 : pièces de fonderie
4. Approximations $\mathcal{B}$ et $\mathcal{P}$: filtrage de substrats
5. Distance ou pas ?

1. Analyse Statistique du dossier VINS

Enoncé

La Direction Générale des Impots publie régulièrement au Journal Officiel une Statistique Mensuelle des Vins. Le J.O. du 4 novembre 1987 fournit en particulier le tableau de données suivant où sont croisées des catégories de vins avec des pays exportateurs. L'unité commune est l'hectolitre. Les sigles se veulent explicites ; ainsi BOJO signifie Beaujolais, ANJO est mis pour Anjou...

ID	BELG	NEDE	RFA	ITAL	UK	SUIS	USA	CANA
CHMP	7069	3786	12578	8037	13556	9664	10386	206
MOS1	2436	586	2006	30	1217	471	997	51
MOS2	3066	290	10439	1413	7214	112	3788	330
ALSA	2422	1999	17183	57	1127	600	408	241
GIRO	22986	22183	21023	56	30025	6544	13114	3447
BOJO	17465	19840	72977	2364	39919	17327	17487	2346
BORG	3784	2339	4828	98	7885	3191	11791	1188
RHON	7950	10537	7552	24	8172	11691	1369	1798
ANJO	2587	600	2101	0	7582	143	872	131
AOCX	17200	22806	15979	50	20004	1279	4016	944
VDQS	1976	1029	1346	0	2258	212	1017	487
XXXX	38747	19151	191140	7992	101108	1029	26192	38503
PROV	1375	1150	2514	0	284	401	9	236
MUSC	2016	2908	1529	0	12891	18	716	653
RHOF	785	1648	1009	6	775	643	542	35
AOCF	160	246	135	8	1177	26	7	0
XXXF	24	1533	160	0	480	0	0	0
XXFF	2415	74	208	8	1705	12	36	47

Analyser ces variables quantitatives (analyse conjointe et séparée). On présentera les résultats comme convenu.

Nous fournissons également les sommes suivantes :

BELG	NEDE	RFA	ITAL	UK	SUIS	USA	CANA
134463	112705	364707	20143	257379	53363	92747	50643

UK*RFA	CANA*RFA	UK*BELG	UK*UK
23597981730	7650378991	5922865383	13719495029
RFA*RFA	CANA*CANA	ITAL*ITAL	
43220226841	1506353705	136070627	

Solution

Nous donnons ici l'affichage "intelligent" des résultats tels que nous le donne notre programme de statistiques AsgQT.Prg (écrit sous Dbase). Nous encourageons le lecteur (et la lectrice !) à vérifier "à la main" ces résultats à partir des sommes fournies.

```
ETUDE de la base  L:VINS

Caractéristiques de la base

Il y a 18  lignes et 9 colonnes (au sens de Dbase).

Analyse par variable

Tri par ordre décroissant de moyenne

 Champ  Nom du   Moyenne  Ecart-Type    Cdv     Min     Max
        champ       m          s        s/m
   1    NUM       non numérique ; son type est :  N
   4    RFA      20261.50 44616.09     220      135   191140
   6    UK       14298.83 23616.47     165      284   101108
   2    BELG     7470.167 9993.916     134       24    38747
   3    NEDE     6261.389 8227.603     131       74    22806
   8    USA      5152.611 7336.798     142        0    26192
   7    SUIS     2964.611 4882.127     165        0    17327
   9    CANA     2813.500 8704.627     309        0    38503
   5    ITAL     1119.056 2511.413     224        0     8037
```

Tri par ordre décroissant de coefficient de variation

Champ	Nom du champ	Moyenne m	Ecart-Type s	Cdv s/m	Min	Max
9	CANA	2813.500	8704.627	309	0	38503
5	ITAL	1119.056	2511.413	224	0	8037
4	RFA	20261.50	44616.09	220	135	191140
6	UK	14298.83	23616.47	165	284	101108
7	SUIS	2964.611	4882.127	165	0	17327
8	USA	5152.611	7336.798	142	0	26192
2	BELG	7470.167	9993.916	134	24	38747
3	NEDE	6261.389	8227.603	131	74	22806

Analyse par couple de variables

Matrice des corrélations

	BELG	NEDE	RFA	ITAL	UK	SUIS	USA
BELG	1.0000						
NEDE	0.8702	1.0000					
RFA	0.8692	0.5818	1.0000				
ITAL	0.5856	0.2895	0.6998	1.0000			
UK	0.9416	0.6997	0.9693	0.6906	1.0000		
SUIS	0.3353	0.5177	0.1984	0.3098	0.2462	1.0000	
USA	0.8699	0.6799	0.8477	0.7172	0.8935	0.4681	1.0000
CANA	0.8143	0.4582	0.9476	0.6585	0.9256	-0.0246	0.7469

Liste des meilleures corrélations linéaires
par ordre décroissant de rho.

On prend comme borne de corrélation sûre la valeur 0,9
et comme borne de corrélation possible la valeur 0,6.

*	0.969	UK	RFA
*	0.948	CANA	RFA
*	0.942	UK	BELG
*	0.926	CANA	UK
	0.894	USA	UK
	0.870	NEDE	BELG

```
0.870 USA   BELG
0.869 RFA   BELG
0.848 USA   RFA
0.814 CANA  BELG
0.747 CANA  USA
0.717 USA   ITAL
0.700 ITAL  RFA
0.700 UK    NEDE
0.691 UK    ITAL
0.680 USA   NEDE
0.659 CANA  ITAL
```

```
Valeurs des formules linéaires pour les corrélation sûres.

 Corrélation    0.969  :  UK    =  0.513  *  RFA   +  3903.587
 Corrélation    0.969  :  RFA   =  1.831  *  UK    -  5921.349
 Corrélation    0.948  :  CANA  =  0.185  *  RFA   -   932.386
 Corrélation    0.948  :  RFA   =  4.857  *  CANA  +  6596.411
 Corrélation    0.942  :  UK    =  2.225  *  BELG  -  2322.591
 Corrélation    0.942  :  BELG  =  0.398  *  UK    +  1772.722
 Corrélation    0.926  :  CANA  =  0.341  *  UK    -  2064.845
 Corrélation    0.926  :  UK    =  2.511  *  CANA  +  7233.245
```

2. Analyse Statistique du dossier ELF

Enoncé

Dans le cadre d'une enquête linguistique sur la féminisation des noms de métiers le Ministère des Droits de la Femme a établi un questionnaire comprenant un signalétique de 7 variables et 26 questions. Nous reproduisons ici le codage des 4 variables qualitatives du signalétique.

```
Sexe            : 0 Homme        1 Femme
Niveau d'étude  : 0 Non réponse  1 Primaire ( niveau CEP )
  2 Secondaire ( niveau BEPC ) 3 Secondaire ( niveau BAC ) 4 Supérieur
Régionalité     : 0 NR, 1 faible, 2 moyenne 3 forte 4 très forte
Usage de la langue : 0 NR, 1 peu fréquent, 2 commun, 3 très particulier
```

Les résultats informatiques sont

SEXE	0	35 (35.35 %)	1	64 (64.65 %)
ETUD	0	3 (3.03 %)	1	6 (6.06 %)
	2	30 (30.30 %)	3	21 (21.21 %)
	4	39 (39.39 %)		
REGI	0	2 (2.02 %)	1	35 (35.35 %)
	2	14 (14.14 %)	3	5 (5.05 %)
	4	43 (43.43 %)		
USAG	0	66 (66.67 %)	1	18 (18.18 %)
	2	13 (13.13 %)	3	2 (2.02 %)

Présentez-les comme convenu et essayez de les commenter.

Solution

USAG	NR	66	(66.67 %)	Peu	18	(18.18 %)
	Commun	13	(13.13 %)	Très	2	(2.02 %)
SEXE	Femme	64	(64.65 %)	Homme	35	(35.35 %)
REGI	Très	43	(43.43 %)	Faible	35	(35.35 %)
	Moyenne	14	(14.14 %)	Forte	5	(5.05 %)
	NR	2	(2.02 %)			
ETUD	Sup	39	(39.39 %)	Bepc	30	(30.30 %)
	Bac	21	(21.21 %)	Primaire	6	(6.06 %)
	NR	3	(3.03 %)			

Commentaires

La plupart (67 %) des gens dans cette population d'une centaine de personnes n'ont pas répondu à la question sur l'usage de la langue. Il y a eu une majorité de femmes interrogées puisqu'on a en gros une répartition 1/3 d'hommes et 2/3 de femmes. La régionalité se décline nettement (43 %) en forte régionalité et un peu moins nettement en faible régionalité (35 %). On trouve beaucoup (40 %) de gens ayant fait des études supérieures et pas mal de gens de niveau bepc (30 %) ou bac (21 %).

3. Calculs concrets de χ^2 : pièces de fonderie

Enoncé

Le conditionnement de pièces de fonderie a conduit à ventiler la population totale (180 pièces) en 6 lots contenant respectivement 29, 41, 31, 29, 18, 32 pièces. Le conditionnement théorique aurait abouti pour la même population aux valeurs 30, 40, 30, 30, 20, 30.

Le calcul du χ^2 est-il possible ? Si oui, combien trouve-t-on ?

Discuter alors la significativité du test sous-jacent.

Solution

Puisque :

- le total des effectifs observés et des effectifs théoriques est le même (il vaut 180 dans les deux cas),
- chaque effectif est positif et supérieur à 5
- le total des effectifs est supérieur à 50,

le calcul du χ^2 est possible.

Le détail du calcul des différences pondérées et de leur somme est le suivant

```
  obs      th   d=obs-th       d*d/th    chideux
29.00   30.00    -1.00      0.0333333    0.03
41.00   40.00     1.00      0.0250000    0.06
31.00   30.00     1.00      0.0333333    0.09
29.00   30.00    -1.00      0.0333333    0.12
18.00   20.00    -2.00      0.2000000    0.33
32.00   30.00     2.00      0.1333333    0.46
```

Le χ^2 est donc 0.46 ; pour 6 classes, le χ^2 théorique avec 6-1=5 degrés de liberté au seuil de 5 % est 11.10 ; au seuil de 5 % on peut donc accepter l'hypothèse que les deux distributions suivent le même modèle.

4. Approximations $\mathcal{B}$ et $\mathcal{P}$

Enoncé

Soit x_i le nombre de fois où on doit filtrer un substrat organique avant d'être sur de sa pureté. Compte-tenu des techniques modernes de filtration, il est très peu probable que ce nombre dépasse 5 fois et on admettra donc que la valeur "5 fois" représente en fait l'évènement "5 fois ou plus". On fournit dans le tableau suivant le nombre n_i de substrats ayant été filtré x_i fois.

x_i	0	1	2	3	4	5
n_i	70	60	20	20	8	70

- Donner le total, la moyenne et la variance du nombre de filtrations.
- Effectuer une approximation des effectifs n_i par la loi binomiale.
- Effectuer une approximation des effectifs n_i par la loi de Poisson.
- Conclure en comparant ces approximations.

Vous n'oublierez pas de détailler les calculs, de justifier les effectifs théoriques réels et arrondis, de fournir le χ^2 utilisé, le nombre de degrés de liberté *etc.*

Solution

Commençons par remarquer que la distribution des n_i est en "creux" (forme parabolique positive). L'approximation par des lois unimodales en "plein" (forme parabolique négative) que sont les lois binomiale et de Poisson sera donc certainement inacceptable.

La somme des n_i vaut 248.

Leur moyenne pondérée est $m = 2{,}185$,

La variance estimée 4,103 soit un écart-type de 2,026.

L'application du modèle binomial $\mathcal{B}(n, p)$ doit donc de faire avec les valeurs $n = 5$ et $p = m/n = 0{,}437$.

Le détail de la construction des probabilités dans le modèle binomial $\mathcal{B}(5; 0, 437)$ est

xi	c(n,k)	p^k	(1-p)^k	probabilite
0	1	1.000000	0.056516	0.0565
1	5	0.437097	0.100400	0.2194
2	10	0.191054	0.178362	0.3408
3	10	0.083509	0.316860	0.2646
4	5	0.036501	0.562903	0.1027
5	1	0.015955	1.000000	0.0160

ce qui aboutit aux effectifs théoriques suivants

Ni	Xi	Pi	Obs	ThExact	ThArrondi
70	0	0.057	70	14.016	14
60	1	0.219	60	54.417	54
20	2	0.341	20	84.510	85
20	3	0.265	20	65.622	66
8	4	0.103	8	25.478	25
70	5	0.016	70	3.957	4

Il n'est alors pas possible de calculer directement le χ^2 car un effectif est trop faible (inférieur à 5). Venant regrouper les deux dernières classes, on dispose alors de valeurs observées et théoriques suivantes

obs	70	60	20	20	78
th	14	54	85	66	29

Ces deux séries de valeurs sont "manifestement très différentes" et c'est ce que montre le calcul du χ^2 qui vaut 389,23.

i	obs	th	obs-th	tc	cumul
1	70	14	56	224.00000	224.000
2	60	54	6	0.66667	224.667
3	20	85	-65	49.70588	274.373
4	20	66	-46	32.06061	306.433
5	78	29	49	82.79310	389.226

Or pour 4 degrés de liberté, au seuil de 5 %, le χ^2 théorique maximal autorisé lu dans la table est 9,49. On doit donc refuser l'hypothèse que nos données suivent une loi binomiale.

L'approximation par la loi de Poisson commence par utiliser la valeur 2.185 de la moyenne pondérée de nos données comme paramètre λ de la loi de Poisson. Si on calcule les 6 premières valeurs de cette loi, on obtient les valeurs

xi	ni	pi
0	70	0,11242
1	60	0,24570
2	20	0,26849
3	20	0,19559
4	08	0,10687
5	70	0,04671

La somme de ces probabilités vaut 0,97578 et on doit donc rajouter 1-0,97578=0,02422 à p_5. Après avoir multiplié les probabilités par 248 et après avoir arrondi les valeurs obtenues, on a la mauvaise surprise d'obtenir un total d'effectifs théoriques égal à 250.

xi	ni	pi*	thir	thi*
0	70	0,11242	27,88016	28
1	60	0,2457	60,93360	61
2	20	0,26849	66,58552	67
3	20	0,19559	48,50632	49
4	08	0,10687	26,50376	27
5	70	0,07093	17,59064	18
somme	250	1	248	250

On retranche alors 1 aux effectifs de probabilités les plus proches de $x,5$ et on peut ensuite calculer le χ^2. Les deux effectifs incriminés sont ici 48,50632 et 26,50376. On leur associe alors respectivement les valeurs 48 et 16.

Finalement, les effectifs à comparer et le calcul du χ^2 est résumé dans le tableau suivant.

xi	obs	thi*	thi	diff	tc	cumul
0	70	28	28	42	63,00000	63,00000000
1	60	61	60	-1	0,01639	63,01639344
2	20	67	67	20	32,97015	95,98654270
3	20	49	48	20	16,33333	112,3198760
4	08	27	26	8	12,46154	124,7814145
5	70	18	18	70	150,22222	275
somme	248	250	248		275	

On obtient 275 ce qui est beaucoup trop élevé pour qu'on puisse accepter l'hypothèse que nos données suivent une loi de Poisson.

5. Distance ou pas ?

Enoncé

L'appplication $\rho\ :\ (X,Y)\ \rightarrow\ \rho(X,Y)$ est une application qui à un couple de variables quantitatives associe un coefficient. Est-ce une distance ? Idem pour son carré ρ^2.

Soient a, b et c trois vecteurs de même longueur dont tous les éléments sont positifs ou nuls. Si la somme des éléments de c est non nulle, on appelle χ^2 entre a et b en base c la quantité $d_c(a,b)$ définie par $\Sigma(a_i-b_i)^2/ci$. La fonction $(a,b)\ \rightarrow\ d_b(a,b)$ est-elle une distance ?

Soient p et q deux variables qualitatives binaires c'est à dire qu'elles ne prennent que les valeurs 0 et 1. On appelle coefficient de ressemblance de Russel-Rao entre p et q la quantité $c_1(p,q)=1_p1_q/(1_p1_q+1_p0_q+0_p1_q)$ où x_py_q désigne le nombre de fois où on trouve simultanément x pour p et y pour q. De même, le coefficient de Sokal-Michener entre p et q est $c_2(p,q)=(1_p1_q+0_p0_q)/n$ où n est la longueur commune des deux variables. Les fonctions $(p,q)\ \rightarrow\ c_1(p,q)$, $(p,q)\ \rightarrow\ 1-c_1(p,q)$ et $(p,q)\ \rightarrow\ 1-c_2(p,q)$ sont-elles des distances ?

Solution

Une distance est une fonction positive. Or $\rho(X, -X)$ vaut -1. Donc ρ ne peut être une distance. Une distance vérifie $d(X, X) = 0$ donc puisque $\rho^2(X, X) = 1$, ρ^2 ne peut pas être une distance. Par contre $1 - \rho^2$ est un candidat intéressant.

La fonction $(a, b) \rightarrow d_b(a, b)$ n'est pas une distance car ce n'est pas une fonction symétrique, puisque pour des vecteurs de longueur 1 $(a-b)^2/b$ est différent de $(a-b)^2/a$. Par contre, on laisse au lecteur le soin de démontrer que pour tout vecteur c donné, $(a, b) \rightarrow d_c(a, b)$ est une distance.

La fonction $(p, q) \rightarrow c_1(p, q)$ n'est pas une distance car $c_1(p, p) = 1$. La fonction $(p, q) \rightarrow 1 - c_1(p, q)$ n'est pas définie partout : par exemple si p ne contient que des 1, $c_1(p, p)$ vaut 0/0. Sinon, elle est symétrique, nulle sur la diagonale. Pour l'inégalité triangulaire, c'est une autre histoire.

$(p, q) \rightarrow 1 - c_2(p, q)$ est une distance car si $c_2(x, y) = (1_x 0_y + 0_x 1_y)/n$ alors $c_2(x, x) = (1_x 0_x + 0_x 1_x)/n = 0/n = 0$ et $c_2(y, x) = (1_y 0_x + 0_y 1_x)/n = (1_x 0_y + 0_x 1_y)/n = c_2(x, y)$. De plus, si $x = 1, y = 0$ alors $x = 1, z = 0$ vaut 1 et $y = 0, z = 1$ vaut 1. De même, si $x = 0, y = 1$ alors $x = 0, z = 1$ vaut 1 et $y = 1, z = 0$ vaut 1 donc $c_2(x, y) = (1_x 0_y + 0_x 1_y)/n$ est inférieur ou égal à $(1_x 0_z + 0_y 1_z)/n + (1_z 0_x + 0_z 1_y)/n$ qu'on peut écrire en $(1_x 0_z + 1_z 0_x)/n + (0_y 1_z + 0_z 1_y)/n$ soit encore $c_2(x, z) + c_2(y, z)$, CQFD.