Université Claude Bernard Lyon 1

L3APAS (Semestre 2) Statistiques
Session 1 13 Mai 2011
Durée : 2 h.

Examen CCF2 de Statistiques

Document autorisés : tout type (papier ou numérique), voire ordinateur personnel

IMPORTANT : Il est conseillé de d'abord traiter les exercices 1 à 3, assez classiques, en finissant par l'exercice 4, un peu moins classique.

AVERTISSEMENT
L'ensemble des fichiers de données nécessaires pour cet examen ('`M1IGAPASA10data.txt`', '`pise.txt`' et '`pps006.txt`') est normalement disponible à la fois

- en ligne sur `http://utbmjb.chez-alice.fr/UFRSTAPS/index.html` à la rubrique habituelle (voir 'examen', en bas de la page) ;
- en cas de problème internet, sur le réseau de l'université Lyon I : il faut aller sur :
  - 'Poste de travail',
  - puis sur le répertoire '`P:`' (appelé aussi '`\\teraetu\Enseignants`'),
  - puis '`jerome.bastien`',
  - enfin sur '`L3APAS\examen\CCF2`'.

Exercice 1.
On étudie le fichier de données '`M1IGAPASA10data.txt`'.

(1) Analyser la variable '`main.ecriture`'.

(2) Analyser la variable '`taille`'.

Exercice 2.
La tour de Pise est une merveille architecturale mais les ingénieurs sont inquiets de son inclinaison. C'est pourquoi ils réalisent régulièrement des études de son évolution. Le fichier '`pise.txt`' contient pour 13 années consécutives une mesure cette inclinaison (en dixième de millimètres au delà de 2,90 m pour être très précis par rapport à un point qui devrait être à la vertical).
Étudier le croisement de la variable '`Annee`' et de la variable '`Inclinaison`'.
Les ingénieurs ont-ils de bonnes raisons d'être inquiets ?

Exercice 3.
Les données étudiées correspondent aux notes de gymnastes au championnat de France UFOLEP, le 1er et 2 juin 2002 à Elancourt. La gymnastique est une discipline où l'on retrouve (chez les féminines) quatre agrès : le sol, le saut de cheval, la poutre et les barres asymétriques. La compétition regroupe deux fédérations : l'UFOLEP (Union Française des Œuvres Laïques d'Éducation Physique) et la FFG (Fédération Française de Gymnastique). Certaines filles sont en double appartenance, c'est-à-dire qu'elles sont affiliées aux deux fédérations : UFOLEP et FFG. Elles ont le droit de faire les deux compétitions (UFOLEP et FFG). D'autres sont en simple appartenance, en l'occurrence ici elles sont affiliées à l'UFOLEP. Les gymnastes sont regroupées par âge et par niveaux (catégorie) et les rencontres se font par catégories. L'ordonnancement des niveaux est le suivant (pour une compétition nationale) : niveau 3, promotion, honneur, espoir et excellence. Les gymnastes d'une même catégorie sont notées par les mêmes juges dans un esprit d'équité et de justice. À chaque agrès, il y a au moins trois juges et la note médiane est conservée. Le classement est obtenu en faisant la somme des notes des quatre agrès.
On étudie le fichier de données '`pps006.txt`'.
Les différentes variables sont

- '`poutre`' : la note obtenue à la poutre ;
- '`barres`' : la note obtenue aux barres ;
- '`saut`' : la note obtenue au saut ;
- '`sol`' : la note obtenue au sol ;
- '`total`' : la somme totale (somme des quatre précédentes)
- '`cate`' : variable qualitative correspondant à l'appartenance :
  - '`s`' : simple appartenance ;
  - '`d`' : double appartenance.
- '`niv`' : variable qualitative correspondant à la catégorie
  - '`espo`' : catégorie espoir ;
  - '`niv3`' : catégorie niveau 3.

(1) Analyser la variable '`total`'.

(2) Analyser la variable '`cate`'.

(3) À partir de là, on peut se poser plusieurs questions concernant les résultats obtenus et la discipline elle-même :

(a) Y'a-t-il une différence entre les filles qui sont en simple appartenance et les filles qui sont en double appartenance : Étudier le croisement de la variable '`total`' et de la variable '`cate`'.

(b) Y'a-t-il un lien entre la note finale et le niveau : Étudier le croisement de la variable '`total`' et de la variable '`niv`'.

(c) Y'a-t-il des liens entre les différents agrès :

(i) On pourra tout d'abord calculer la moyenne et l'écart type de notes obtenues pour chacun des agrès ;

(ii) On pourra ensuite étudier successivement la relation entre la note totale et la note obtenue au premier agrès, puis au second, puis au troisième, puis au quatrième et conclure.

Exercice 4.

Attention, cet exercice est à part ; il met en évidence votre débrouillardise sous R, utilise peu de notions théoriques acquises en cours et met l'accent sur vos capacités d'observation. C'est donc un bonus.

Vous trouverez sur le site deux fichiers spéciaux : '`texte1.Rdata`' et '`aexclure.Rdata`', qui sont des fichiers de donnés sous un format spécifique à R. Vous ne pouvez donc pas les ouvrir avec un éditeur de texte, mais en procédant comme suit :

– Téléchargez-les comme d'habitude.
– Ouvrez-les de la façon suivante : taper

```
load("texte1.Rdata")
load("aexclure.Rdata")
```

– Vous devez alors obtenir un data.frame qui s'appelle '`d`' et un tableau de caractères qui s'appelle '`aexclure`'. Vérifiez-le en tapant

```
ls()
aexclure
d
```

Le data.frame '`d`' contient 11152 caractères, issu d'un texte en français.

(1) Analyser la variable '`lettres`' : on s'intéresse donc aux nombres et aux pourcentages des signes présents dans un texte. Vous pourrez, pour vous affranchir des chiffres, des caractères majuscules non accentués et de différents caractères de ponctuation, taper les choses suivantes :

```
u <- table(d$lettres, exclude = aexclure)
u
sort(u/sum(u) * 100)
```

(2) D'autres textes en littérature française, allemande, italienne, espagnole ou anglaise ont été analysés de la même façon, grâce à R. *Vous n'avez pas à faire ces analyses, puisqu'elles ont été faites pour vous.* Ces textes sont numérotés 1 à 9. Les résultats figurent dans les tableaux 1 à 3 pages 5–6. Dans ces tableaux, seul le pourcentage des 15 lettres apparaissant les plus fréquemment est donné.

Tous les tableaux sont donnés à partir de la page 5 de l'énoncé.

(a) On donne, dans le tableau 4 page 7, les pourcentages d'apparition des lettres dans un corpus (ensemble d'ouvrages) français contenant un grand nombre de lettres.

Les pourcentages mesurés dépendent de la nature du texte en question. On observe beaucoup de variation sur les lettres les moins fréquents, (la fréquence du 'œ' varie entre 0,002 % et 0,09 % pour trois textes pris au hasard), elle est également sensible même pour les lettres les plus fréquentes (l'ordre de fréquence des lettres a, s, i, t et n, qui sont les plus fréquentes à part e, fluctue d'un texte à l'autre, mais reste autour de celle donnée dans le tableau 4).

En vous attachant aux lettres les plus fréquentes et en vous fondant sur vos observations, essayer de trouver parmi les tableaux 1 à 3 pages 5–6, lesquels semblent correspondre à des ouvrages en français.

Un indice pour le texte 7 :

Un polar raconta la façon dont un constituant primordial du français disparut au cours d'un fait qui prit d'abord l'air d'un larcin ; plus tard, on crut qu'il s'agissait d'un kidnapping. Son papa (signifiant plutôt moral) fut fort connu ; il participait à l'Oulipo.

Un indice pour le texte 9 :

Ce thème permet de prendre le sens de ce célèbre texte de Perec (les Revenentes) en cette terre extrême de l'ex-CEE (pensez "Éternel été" ; c'est bête de chercher chez les Grecs).

(b) En français, donnez des exemples concrets de contexte dans lesquels les fréquences des lettres et des symboles utilisés peuvent être beaucoup modifiés par rapport à un texte "standard".

(c) Les textes 1 à 9 constituent tout ou partie des œuvres suivantes (dans le désordre) :
- Macbeth
- la Divine Comédie
- Andromaque
- Die Leiden des jungen Werther
- Les Misérables
- la Disparition
- Britannicus
- un pastiche des "Revenentes"
- une traduction des "Revenentes" en espagnol

Sauriez-vous rendre à chaque texte son auteur (vous avez le droit à trois jockers) ?

Corrigé

Un corrigé sera disponible sur `http://utbmjb.chez-alice.fr/UFRSTAPS/index.html`

Ensemble des tableaux

lettre texte 1	pourcent. texte 1	lettre texte 2	pourcent. texte 2	lettre texte 3	pourcent. texte 3
e	16.65	e	16.93	e	16.43
a	9.19	s	9.57	s	9.88
t	8.50	u	8.64	u	8.59
i	8.39	r	8.57	r	8.29
s	8.25	i	7.65	i	7.87
n	7.66	n	7.55	n	7.66
u	7.07	o	7.29	a	7.18
r	7.02	a	6.96	o	7.18
l	6.44	t	6.67	t	6.52
o	5.69	l	5.33	l	4.91
d	3.82	m	3.52	m	3.50
c	3.16	d	3.22	d	3.40
m	3.16	p	2.90	c	3.11
p	2.83	c	2.78	p	3.03
é	2.17	v	2.41	v	2.45

TABLE 1. Les pourcentages d'apparition des lettres pour les textes 1 à 3

lettre texte 4	pourcent. texte 4	lettre texte 5	pourcent. texte 5	lettre texte 6	pourcent. texte 6
e	12.78	e	14.11	e	19.21
a	11.48	t	10.11	n	12.26
i	10.57	o	9.15	i	9.97
o	9.92	a	8.55	r	7.28
r	8.23	h	7.85	s	6.49
n	7.44	n	7.56	h	6.43
c	6.56	s	7.18	t	6.29
l	6.25	r	6.92	a	5.57
s	5.82	i	6.45	d	5.56
t	5.68	d	4.91	u	4.42
d	3.73	l	4.78	c	4.13
u	3.53	u	3.94	l	3.93
m	3.02	c	3.13	m	3.31
p	2.60	m	2.92	g	3.03
v	2.38	y	2.43	o	2.12

TABLE 2. Les pourcentages d'apparition des lettres pour les textes 4 à 6

lettre texte 7	pourcent. texte 7	lettre texte 8	pourcent. texte 8	lettre texte 9	pourcent. texte 9
i	13.43	e	33.54	e	41.07
a	12.63	t	10.49	r	10.19
o	10.88	s	9.39	n	8.07
n	9.60	n	7.32	s	7.87
u	9.13	l	6.83	t	6.56
t	8.60	r	6.46	d	5.15
s	7.66	c	5.00	l	5.04
r	6.58	é	4.63	c	2.52
l	5.84	m	3.29	é	2.32
d	3.36	d	3.05	q	2.32
p	3.36	è	2.81	m	2.22
m	2.42	p	2.32	v	2.12
c	2.35	x	1.95	p	1.82
v	2.22	v	1.71	¡	1.41
f	1.95	h	1.22	E	1.31

TABLE 3. Les pourcentages d'apparition des lettres pour les textes 7 à 9

lettre	pourcentage
e	14.715
s	7.948
a	7.636
i	7.529
t	7.244
n	7.095
r	6.553
u	6.311
l	5.456
o	5.378
d	3.669
c	3.260
p	3.021
m	2.968
é	1.904
v	1.628
q	1.362
f	1.066
b	0.901
g	0.866
h	0.737
j	0.545
à	0.486
x	0.387
y	0.308
è	0.271
ê	0.225
z	0.136
w	0.114
ç	0.085
ù	0.058
k	0.049
î	0.045
œ	0.018
ï	0.006
ë	0.000

TABLE 4. Les pourcentages d'apparition dans un corpus (1 533 629 lettres) en français