TD8 : perception de la science (correction)

```
> require(ade4)
> source("fonctions.R")
```

En 1993, l'*International Social Survey Programme* (ISSP) a interrogé 365 personnes sur leur rapport à la science. On s'intéresse à 4 affirmations particulières :

— `a` : « nous croyons trop souvent à la science, pas assez aux sentiments et à la foi » ;
— `b` : « au total, la science fait plus de mal que de bien » ;
— `c` : « tous les changements que les humains font à la nature, vont probablement empirer les choses » ;
— `d` : « la science moderne résoudra nos problèmes environnementaux sans trop changer notre mode de vie ».

Pour chacune, les réponses possibles sont : tout à fait d'accord (5), d'accord (4), sans avis (3), pas d'accord (2), pas du tout d'accord (1).

```
> science=read.csv("science-simple.csv", colClasses="factor")
> science1=science[,c(1:4)]
> burt1=acm.burt(science1, science1)
> acm1=dudi.acm(science1,scannf=F,nf=4)
> acm1=dudi.fixsigns(acm1, sign.co=c(1,-1, 1, -1))
> poids1=as.data.frame(acm1$cw)
> colnames(poids1)="poids"
> inert1=inertia.dudi(acm1, col=T)
> burt2=burt1
> burt2[1,8:9]=burt2[8:9,1]=NA
> burt2[2,6]=burt2[6,2]=NA
> burt2[3,6:7]=burt2[6:7,3]=NA
> burt2[5,7:8]=burt2[7:8,5]=NA
> burt2[6,6]=burt2[7,7]=burt2[8,8]=burt2[9,9]=NA
> conting=burt1[1:5,11:15]
> stats=stats.conting(conting)
```

1 Un premier regard

Les données sont résumées dans le tableau de Burt ci-dessous :

```
> burt2
```

	a.1	a.2	a.3	a.4	a.5	b.1	b.2	b.3	b.4	b.5	c.1	c.2	c.3	c.4	c.5	d.1	d.2	d.3	d.4	d.5
a.1	14	0	0	0	0	7	5	NA	NA	0	1	3	4	5	1	3	4	6	1	0
a.2	0	80	0	0	0	NA	30	13	15	2	3	20	23	27	7	14	26	12	24	4
a.3	0	0	94	0	0	NA	NA	23	24	2	1	15	35	32	11	10	25	36	21	2
a.4	0	0	0	133	0	15	32	37	31	18	4	14	22	74	19	17	27	36	43	10
a.5	0	0	0	0	44	5	NA	NA	11	11	2	3	8	16	15	9	11	9	11	4
b.1	7	NA	NA	15	5	NA	0	0	0	0	8	9	15	20	3	13	13	8	17	4
b.2	5	30	NA	32	NA	0	NA	0	0	0	3	28	30	41	8	15	37	32	26	0
b.3	NA	13	23	37	NA	0	0	NA	0	0	0	9	24	40	12	8	24	28	21	4
b.4	NA	15	24	31	11	0	0	0	NA	0	0	7	20	39	16	9	14	25	27	7
b.5	0	2	2	18	11	0	0	0	0	33	0	2	3	14	14	8	5	6	9	5
c.1	1	3	1	4	2	8	3	0	0	0	11	0	0	0	0	2	2	1	6	0
c.2	3	20	15	14	3	9	28	9	7	2	0	55	0	0	0	8	20	13	14	0
c.3	4	23	35	22	8	15	30	24	20	3	0	0	92	0	0	5	17	47	22	1
c.4	5	27	32	74	16	20	41	40	39	14	0	0	0	154	0	21	39	32	52	10
c.5	1	7	11	19	15	3	8	12	16	14	0	0	0	0	53	17	15	6	6	9
d.1	3	14	10	17	9	13	15	8	9	8	2	8	5	21	17	53	0	0	0	0
d.2	4	26	25	27	11	13	37	24	14	5	2	20	17	39	15	0	93	0	0	0
d.3	6	12	36	36	9	8	32	28	25	6	1	13	47	32	6	0	0	99	0	0
d.4	1	24	21	43	11	17	26	21	27	9	6	14	22	52	6	0	0	0	100	0
d.5	0	4	2	10	4	4	0	4	7	5	0	0	1	10	9	0	0	0	0	20

Question 1 *Des données sont manquantes dans le tableau ci-dessous (*`NA`). En utilisant les propriétés du tableau, retrouvez-les.

On commence par calculer les effectifs totaux pour la question `b`. On peut le faire en additionnant les colonnes de la sous matrice (`b`, `c`). Par exemple, (b.1, `b.1`)$= 8+9+15+20+3 = 55$. On peut de même calculer les autres données et on obtient pour la diagonale

```
> diag(as.matrix(burt1))[6:9]
```

```
b.1 b.2 b.3 b.4
 55 110  85  82
```

Ensuite, on calcule

— (a.1, b.4)= $82 - 11 - 31 - 24 - 15 = 1$,
— (a.1, b.3)= $14 - 7 - 5 - 1 = 1$,
— (a.5, b.3)= $85 - 37 - 23 - 13 - 1 = 11$
— (a.5, b.2)= $44 - 5 - 11 - 11 - 11 = 6$
— (a.2, b.1)= $80 - 30 - 13 - 15 - 2 = 20$
— (a.3, b.1)= $55 - 5 - 15 - 20 - 7 = 8$
— (a.3, b.2)= $110 - 6 - 32 - 30 - 5 = 37$

Les autres valeurs sont obtenues par symétrie par rapport à la diagonale.

Le tableau finalement obtenu est

```
> burt1
```

	a.1	a.2	a.3	a.4	a.5	b.1	b.2	b.3	b.4	b.5	c.1	c.2	c.3	c.4	c.5	d.1	d.2	d.3	d.4	d.5
a.1	14	0	0	0	0	7	5	1	1	0	1	3	4	5	1	3	4	6	1	0
a.2	0	80	0	0	0	20	30	13	15	2	3	20	23	27	7	14	26	12	24	4
a.3	0	0	94	0	0	8	37	23	24	2	1	15	35	32	11	10	25	36	21	2
a.4	0	0	0	133	0	15	32	37	31	18	4	14	22	74	19	17	27	36	43	10
a.5	0	0	0	0	44	5	6	11	11	11	2	3	8	16	15	9	11	9	11	4
b.1	7	20	8	15	5	55	0	0	0	0	8	9	15	20	3	13	13	8	17	4
b.2	5	30	37	32	6	0	110	0	0	0	3	28	30	41	8	15	37	32	26	0
b.3	1	13	23	37	11	0	0	85	0	0	0	9	24	40	12	8	24	28	21	4
b.4	1	15	24	31	11	0	0	0	82	0	0	7	20	39	16	9	14	25	27	7
b.5	0	2	2	18	11	0	0	0	0	33	0	2	3	14	14	8	5	6	9	5
c.1	1	3	1	4	2	8	3	0	0	0	11	0	0	0	0	2	2	1	6	0
c.2	3	20	15	14	3	9	28	9	7	2	0	55	0	0	0	8	20	13	14	0
c.3	4	23	35	22	8	15	30	24	20	3	0	0	92	0	0	5	17	47	22	1
c.4	5	27	32	74	16	20	41	40	39	14	0	0	0	154	0	21	39	32	52	10
c.5	1	7	11	19	15	3	8	12	16	14	0	0	0	0	53	17	15	6	6	9
d.1	3	14	10	17	9	13	15	8	9	8	2	8	5	21	17	53	0	0	0	0
d.2	4	26	25	27	11	13	37	24	14	5	2	20	17	39	15	0	93	0	0	0
d.3	6	12	36	36	9	8	32	28	25	6	1	13	47	32	6	0	0	99	0	0
d.4	1	24	21	43	11	17	26	21	27	9	6	14	22	52	6	0	0	0	100	0
d.5	0	4	2	10	4	4	0	4	7	5	0	0	1	10	9	0	0	0	0	20

Question 2 *On s'intéresse au lien entre les réponses aux affirmations* `a` *et* `c` *du sondage. Donnez le tableau de contingence de leurs modalités. Le χ^2 correspondant est 46.12. En utilisant la table donnée à la fin de ce sujet, que peut-on dire de la dépendance entre les réponses à ces deux questions ?*

Le tableau de contingence de ces deux questions est le sous tableau suivant du tableau de Burt :

```
> conting
```

	c.1	c.2	c.3	c.4	c.5
a.1	1	3	4	5	1
a.2	3	20	23	27	7
a.3	1	15	35	32	11
a.4	4	14	22	74	19
a.5	2	3	8	16	15

Le χ^2 du tableau doit être interprété comme ayant $(5-1)(5-1) = 16$ degrés de liberté. D'après la table du χ^2 fournie à la fin, cela correspond à des valeurs critiques de $26,296$ à 5% et $32,000$ à 1%. L'interprétation est la suivante : si les variables sont indépendantes, alors, par exemple, $P(\chi^2_{16} > 32) = 0,01$. Si la valeur mesurée du χ^2 est plus grande que cette valeur critique, alors on peu affirmer que l'hypothèse H_0 d'indépendance est fausse.

Dans notre cas à 1%, la valeur du χ^2 est nettement au dessus de la valeur critique. On peut en déduire que H_0 est fausse la dépendance entre les variables `a` et `c` est très nette.

Si on demande à R de calculer la p-value, on obtient une toute petite valeur : 9.3e-05.

2 Analyse des correspondances multiples

On procède à l'analyse des correspondances multiples des données ci-dessus. Les 10 premières valeurs propres sont données ci-dessous, suivies dans l'ordre pour les 4 premières colonnes par : les coordonnées des catégories, leur poids, leur contribution aux axes et leur qualité de représentation par les sous espaces (ces 3 derniers en %).

```
> round(acm1$eig[1:10], 2)

 [1] 0.45 0.39 0.33 0.32 0.27 0.26 0.26 0.24 0.23 0.22
```

```
> round(acm1$co, 2)
```

	Comp1	Comp2	Comp3	Comp4
a.1	0.90	-1.39	0.38	-1.91
a.2	0.46	-0.80	0.18	0.40
a.3	0.58	0.65	0.40	-0.19
a.4	-0.38	0.19	-0.65	0.27
a.5	-1.21	-0.07	0.68	-0.53
b.1	0.25	-1.54	-0.47	-0.96
b.2	0.69	-0.09	0.47	0.50
b.3	-0.04	0.61	-0.20	0.11
b.4	-0.30	0.52	-0.32	-0.11
b.5	-1.87	-0.01	0.52	-0.05
c.1	0.31	-2.90	-1.46	-1.81
c.2	0.77	-0.60	0.75	0.89
c.3	0.70	0.57	0.13	-0.88
c.4	-0.21	0.08	-0.63	0.40
c.5	-1.47	0.00	1.14	-0.20
d.1	-0.62	-0.83	0.73	-0.26
d.2	0.25	-0.24	0.55	0.73
d.3	0.50	0.84	0.09	-0.75
d.4	-0.01	-0.19	-0.97	0.26
d.5	-1.96	0.09	-0.08	-0.25

```
> round(poids1*100,1)
```

	poids
a.1	1.0
a.2	5.5
a.3	6.4
a.4	9.1
a.5	3.0
b.1	3.8
b.2	7.5
b.3	5.8
b.4	5.6
b.5	2.3
c.1	0.8
c.2	3.8
c.3	6.3
c.4	10.5
c.5	3.6
d.1	3.6
d.2	6.4
d.3	6.8
d.4	6.8
d.5	1.4

```
> round(inert1$col.abs,1)
```

	Axis1	Axis2	Axis3	Axis4
a.1	1.7	4.8	0.4	11.0
a.2	2.6	9.0	0.5	2.7
a.3	4.9	7.1	3.1	0.7
a.4	3.0	0.8	11.6	2.1
a.5	9.9	0.0	4.2	2.7
b.1	0.5	22.9	2.5	10.9
b.2	7.9	0.2	5.0	5.9
b.3	0.0	5.6	0.7	0.2
b.4	1.1	3.9	1.7	0.2
b.5	17.7	0.0	1.8	0.0
c.1	0.2	16.3	4.8	7.7
c.2	5.0	3.5	6.3	9.5
c.3	7.0	5.3	0.3	15.3
c.4	1.0	0.2	12.5	5.4
c.5	17.6	0.0	14.0	0.5
d.1	3.1	6.4	5.7	0.7
d.2	0.9	0.9	5.7	10.6
d.3	3.8	12.2	0.2	12.2
d.4	0.0	0.6	19.2	1.4
d.5	11.8	0.0	0.0	0.3

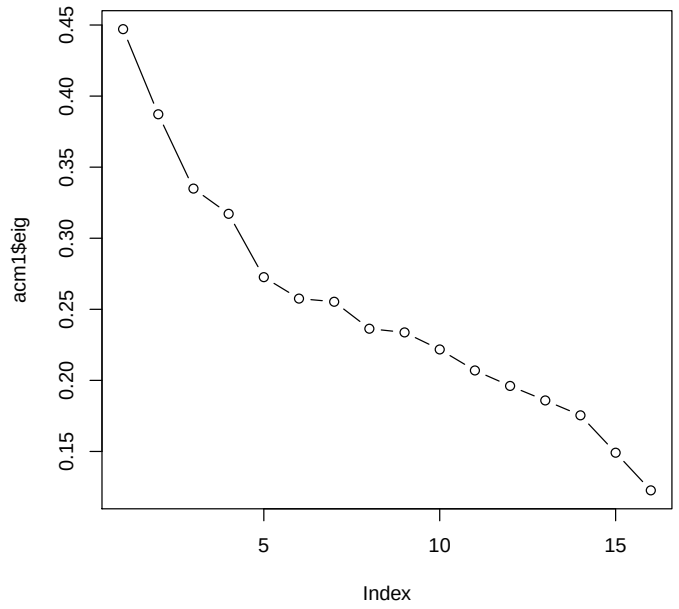
```
> round(inert1$col.cum[,-(acm1$nf+1)],1
```

	Axis1	Axis1:2	Axis1:3	Axis1:4
a.1	3.2	10.9	11.5	26.0
a.2	6.0	23.8	24.7	29.1
a.3	11.8	26.6	32.2	33.4
a.4	8.5	10.5	35.0	39.1
a.5	20.2	20.2	26.6	30.4
b.1	1.1	43.0	46.8	63.2
b.2	20.3	20.6	30.2	40.9
b.3	0.0	11.4	12.6	12.9
b.4	2.6	10.4	13.4	13.8
b.5	34.8	34.8	37.5	37.5
c.1	0.3	26.4	33.0	43.1
c.2	10.6	17.0	26.9	41.1
c.3	16.7	27.8	28.3	54.2
c.4	3.2	3.7	32.6	44.4
c.5	36.9	36.9	58.8	59.5
d.1	6.6	18.3	27.2	28.3
d.2	2.2	4.1	14.3	32.3
d.3	9.4	35.4	35.7	56.9
d.4	0.0	1.3	36.7	39.1
d.5	22.3	22.3	22.4	22.7

Question 3 *Qu'est-ce que la répartition des valeurs propres nous dit sur les variables d'origine et sur la qualité de l'analyse ? Combien de valeurs propres doit-on retenir a priori ? Si on n'en retient que 2, quelle sera la part d'inertie expliquée ?*

Pour faciliter la discussion, on donne la représentation des valeurs propres (qui n'est pas obligatoire) :

```
> plot(acm1$eig, type="b")
```



On voit ici que la décroissance des valeurs propres est très lente. Cela veut dire que la dépendance entre les variables d'origine est plutôt faible. L'analyse sera donc assez mauvaise et donnera peu d'informations.

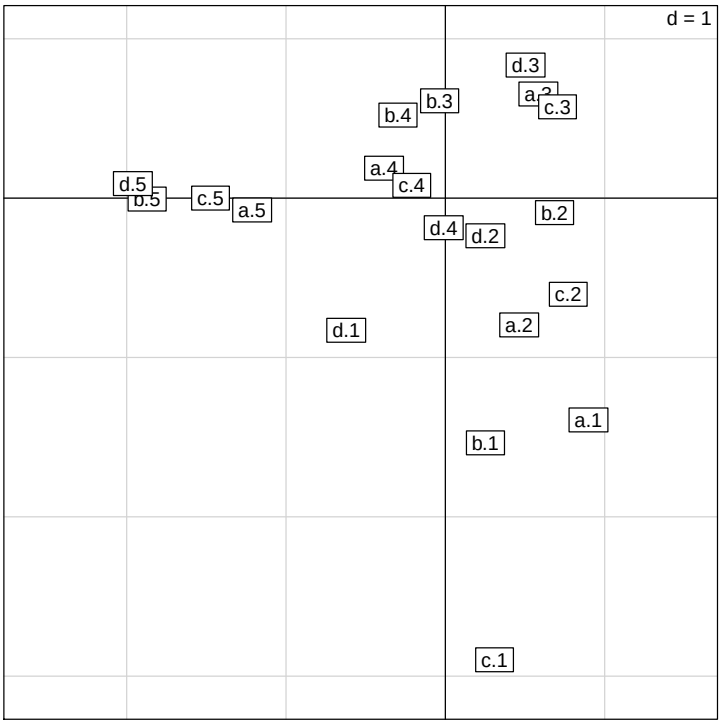
Le nombre total de valeurs propres est $20 - 4 = 16$. La somme vaut $16/4 = 4$ et les axes intéressants correspondent à une valeur propre supérieure à $1/4 = 0,25$. Selon ce critère il faudrait retenir les 7 premiers axes, ce qui est bien sûr trop.

Si on ne retient que 2 axes, l'inertie expliquée est $0,84$, soit 21% de l'inertie totale. C'est très faible.

Question 4 *Quelles sont les catégories qui déterminent les deux premiers axes principaux ? (on détaillera les critères et on cherchera à être précis dans la réponse).*

On fournit pour référence une représentation des deux premiers axes qui n'est pas dans le sujet.

```
> s.label(acm1$co)
```



Si on travaille à partir des coordonnées, on peut prendre un coefficient multiplicateur de 3, et on obtient des valeurs limite pour les coordonnées de $\sqrt{3 \times 0,45} = 1,16$ sur le premier axe et $\sqrt{3 \times 0,39} = 1,08$ sur le second. On obtient :

— axe 1 : en négatif `d.5` $(-1,96)$, `b.5` $(-1,87)$, `c.5` $(-1,47)$, `a.5` $(-1,21)$; en positif, rien ;

— axe 2 : en négatif `c.1` (-2.90), `b.1` $(-1,54)$, `a.1` $(-1,39)$; en positif, rien.

Si on prend un coefficient multiplicateur de 2, le résultat est le même. On peut être tenté d'ajouter d'autres termes (`d.1` sur l'axe 2, par exemple), mais c'est à mon sens une erreur, vu le « trou » qu'il y a entre les valeurs.

Question 5 *Comment peut-on interpréter les axes à partir de ces données ? Que peut on dire de la question* `d` *?*

Le premier axe met en avant les personnes qui sont tout à fait d'accord avec les affirmations et qui se méfient donc de la science ; il n'y a pas vraiment d'opposition, les autres sont plutôt dans un paquet homogène. Il y a un « piège » avec la question `d`, pour laquelle une réponse 5 correspond à des personnes plutôt pro-science. En rappelant que nous avons ici des personnes *a priori* très anti-science, on peut supposer qu'elles se sont laissé entraîner au lieu de lire le contenu des questions. Normalement, on devrait avoir `d.1` ici.

Le second axe met en avant du coté négatif des personnes qui ont confiance en la science (sauf la question `d`). Là encore, elles sont opposées à des personnes ayant une opposition plus mesurée. Par contre, ces personnes ayant une meilleure idée de la science ne se sont pas trompées sur la question `d` et ne mettent pas `d.1` en avant.

En regardant les réponses à la question `d`, on peut supposer qu'un certain nombre de personnes (surtout « anti-science ») n'ont pas remarqué que la question a un sens positif. La répartition entre `a` et d n'est en effet pas aussi régulière qu'entre `a` et `c`, par exemple.

Même si ces résultats sont difficiles à interpréter, on peut noter qu'ils ne mettent en avant que des personnes ayant des opinions très tranchées, et que les personnes qui doutent sont plus comparables entre elles. C'était prévisible, puisque ces questions ont des effectifs faibles, ce qui conduit mécaniquement à une grande contribution à l'inertie totale.

Question 6 *Quels sont les catégories les mieux représentées dans le premier plan principal ? Commentez et expliquez ce que l'on observe.*

On peut lire directement les qualités de représentation dans la seconde colonne. Les catégories les mieux représentées sont `b.1` (43%), `c.5` $(36,9\%)$ et `d.3` $(35,4\%)$. On peut remarquer que toutes ces variables sont mal représentées (qualité $< 50\%$). Ce n'est pas étonnant, vu que la qualité de l'analyse est mauvaise.

3 Variables supplémentaires

```
> science.suppl=science[,-c(1:4)]
> suppl1=acm.suppl(acm1,science.suppl)
> suppl1$test[4,2]=NA
```

On ajoute à l'analyse de nouvelles variables quantitatives :

— sexe : homme (`sex.h`) ou femme (`sex.f`) ;

— âge : `age.16-24` ans, `age.25-34`, `age.35-44`, `age.45-54`, `age.55-64`, et enfin `age.65--` (65 ans et plus) ;

— niveau d'éducation : école primaire (`edu.pri`), secondaire (`edu.sec`) ou éducation supérieur (`edu.sup`).

On donne ci-dessous les coordonnées de ces catégories sur les 2 premiers axes principaux, leur effectif et les valeurs test correspondantes.

```
> round(suppl1$li[,1:2], 4)
            Axis1   Axis2
sex.f     -0.1058 -0.0623
sex.h      0.1227  0.0722
age.16-24  0.0653 -0.1584
age.25-34  0.0626 -0.0070
age.35-44 -0.1227  0.0876
age.45-54  0.1855 -0.0219
age.55-64  0.1198 -0.0059
age.65--  -0.3070  0.0469
edu.pri   -0.0968  0.0467
edu.sec    0.0956  0.0245
edu.sup    0.1034 -0.3006
```

```
> effectif=suppl1$eff
> as.data.frame(effectif)
          effectif
sex.f          196
sex.h          169
age.16-24       42
age.25-34       80
age.35-44       72
age.45-54       62
age.55-64       54
age.65--        55
edu.pri        183
edu.sec        142
edu.sup         40
```

```
> round(suppl1$test[,1:2], 2)
          Axis1 Axis2
sex.f     -2.17 -1.28
sex.h      2.17  1.28
age.16-24  0.45 -1.09
age.25-34  0.63    NA
age.35-44 -1.16  0.83
age.45-54  1.60 -0.19
age.55-64  0.95 -0.05
age.65--  -2.47  0.38
edu.pri   -1.85  0.89
edu.sec    1.46  0.37
edu.sup    0.69 -2.01
```

Question 7 *La valeur test pour* `age.25-34` *est manquante sur l'axe 2. Calculez-là.*

Pour calculer la valeur manquante, on utilise la formule

$$VT = -0,0070 \times \sqrt{80}\sqrt{\frac{365-1}{365-80}} = -0,07$$

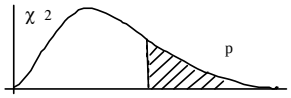
Question 8 *Quelles sont les catégories qui sont liées aux deux premiers axes ? On justifiera les propriétés utilisées. Quelles conclusions peut-on en tirer ?*

Les valeurs-test données ici permettent de savoir quelles sont les catégories liées aux axes. Une catégorie est liée à un axe si
- c'est une catégorie supplémentaire, c'est-à-dire non utilisée dans l'analyse
- son effectif est assez grand (mettons 30)
- sa valeur-test sur l'axe est supérieure à 2 ou 3 en valeur absolue

Les catégories liées aux axes sont ici
- axe 1 : en négatif, `sex.f` $(-2,17)$ et `age.65--` $(-2,47)$, en positif `sex.h` $(2,17)$; les femmes et/ou les personnes de plus de 65 ans sont plus sujets que la moyenne à rejeter fermement la science ;
- axe 2 : en négatif `edu.sup` $(-2,01)$, en positif, rien ; les personnes ayant une éducation supérieure ont plus tendance que la moyenne à être très confiants en la science.

Il faut ajouter que la majorité des gens sont plus mitigés. Enfin, la présence de `sex.h` sur l'axe 1 n'est pas une surprise, puisque ce n'est que le pendant de `sex.f` (c'est toujours pareil avec les variables à 2 modalités). Il dit tout de même que les hommes ont moins tendance que les femmes à être fermement anti-science.

TABLE DU CHI-DEUX : $\chi^2(n)$



n p	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
1	0,0158	0,0642	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635
2	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210
3	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,341
4	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277
5	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086
6	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812
7	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475
8	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090
9	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666
10	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209
11	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725
12	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217
13	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,472	27,688
14	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141
15	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578
16	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000
17	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409
18	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	32,346	34,805
19	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	33,687	36,191
20	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	35,020	37,566
21	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932
22	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289
23	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	38,968	41,638
24	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	40,270	42,980
25	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314
26	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	42,856	45,642
27	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	44,140	46,963
28	18,939	21,588	23,647	27,336	31,391	34,027	37,916	41,337	45,419	48,278
29	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	46,693	49,588
30	20,599	23,364	25,508	29,336	33,530	36,250	40,256	43,773	47,962	50,892

Pour n > 30, on peut admettre que $\sqrt{2\chi^2} - \sqrt{2n-1} \approx N(0,1)$