Covariance. Coefficient de corrélation

Exercice 1

Soit la série suivante de dix associations de valeurs x_i et y_i :

x_i	12	10	17	18	13	10	9	14	17	12
y_i	42	38	50	49	43	40	37	42	48	40

Représenter graphiquement le nuage de points correspondant.
Calculer la covariance et le coefficient de corrélation des deux séries.
Réponses : $Cov(x_i, y_i) = 12.92$ *;* $\rho = 0.9665$. *Forte corrélation entre les deux séries.*

Exercice 2

Quinze élèves, désignés par les lettres de A à O ont été classés une première fois par une épreuve de français, une seconde fois par une épreuve de mathématiques. Calculer le coefficient décrivant la corrélation entre ces deux classements.

Elèves	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Fran.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Math.	9	3	1	11	2	5	8	13	4	10	7	14	15	6	12

Réponses : $Cov(F, M) = 9.47$ *;* $\rho = 0.51$. *Remarquez qu'il s'agit ici d'un coefficient de corrélation des rangs. On pourra consulter le paragraphe* Corrélation des rangs de Spearman *d'un ouvrage de statistiques. Le coefficient de corrélation peut également être obtenu à l'aide de la formule :*

$$\rho = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}$$

dans laquelle N est le nombre de sujets et d_i est la différence entre le rang obtenu sur la première variable et celui obtenu sur la seconde.

Exercice 3

Dans une étude psychométrique, on considère la variable QI, notée X et la variable *Combinatoire*, notée Y pour 20 individus.

	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
x_i	99	122	108	125	108	113	94	85	112	125
y_i	3.9	5.0	5.3	8.3	5.5	6.6	5.5	2.2	5.3	5.3

	i11	i12	i13	i14	i15	i16	i17	i18	i19	i20
x_i	108	91	91	109	125	94	120	112	106	91
y_i	4.6	3.7	4.1	2.7	6.8	2.7	5.4	6.2	2.5	2.4

Aide pour les calculs : $\sum x_i = 2138$, $\sum y_i = 94$, $\sum x_i^2 = 231666$, $\sum y_i^2 = 494.36$, $\sum x_i y_i = 10336.4$.

1) Représenter graphiquement le nuage de points (x_i, y_i).
2) Calculer la moyenne et la variance de chacune des deux séries (x_i) et (y_i) , puis la covariance et le coefficient de corrélation des deux séries.
Réponses : $\overline{x} = 106.9$ *;* $\overline{y} = 4.7$ *;* $\sigma^2(x) = 155.69$ *;* $\sigma^2(y) = 2.628$ *;* $Cov(x, y) = 14.39$ *;* $\rho = 0.7114$.

Corrélation. Droites de régression

Exercice 4 *Données Budget*
Il s'agit d'un extrait d'une enquête (ONU 1967) sur les budgets-temps (temps passé dans différentes activités au cours de la journée).
Les colonnes comprennent 3 variables numériques, le temps passé en : Profession (PROF), Transport (TRAN) et loisirs (LOIS). Les temps sont notés en centièmes d'heures. Le code suivant est utilisé pour identifier les lignes :
H : hommes, F : femmes, A : actifs, N : non actifs, M : mariés, C : célibataires,
U : USA, W : pays de l'ouest, E : Est sauf Yougoslavie, Y : Yougoslavie.

Budget	PROF	TRAN	LOIS
HAU	610	140	315
FAU	475	90	305
FNU	10	0	430
HMU	615	140	305
FMU	179	29	373
HCU	585	115	385
FCU	482	94	336
HAW	653	100	330
FAW	511	70	262
FNW	20	7	368
HMW	656	97	321
FMW	168	22	311
HCW	643	105	388
FCW	429	34	392
HAY	650	140	365

Budget	PROF	TRAN	LOIS
FAY	560	105	235
FNY	10	10	380
HMY	650	145	358
FMY	260	52	295
HCY	615	125	475
FCY	433	89	408
HAE	650	142	334
FAE	578	106	228
FNE	24	8	398
HME	652	133	310
FME	436	79	231
HCE	627	148	463
FCE	434	86	380
Moy	451	86	346
Ety	223	47	63

1) Représenter le nuage de points correspondant aux variables PROF et TRAN, puis celui correspondant aux variables PROF et LOIS.
2) Calculer la covariance et le coefficient de corrélation pour le couple de variables (PROF, TRAN), puis pour le couple (PROF, LOIS). Dans chacun des deux cas, commenter le coefficient de corrélation obtenu.
3) Déterminer l'équation de la droite de régression de TRAN selon les valeurs de PROF. Quelle est la part de la variance de TRAN qui est "expliquée" par PROF ?
Réponses : 2) Cov(PROF,TRAN)=9805.12 , r(PROF,TRAN)=0.93 ; Cov(PROF,LOIS)=-2651.87 , r(PROF,LOIS)=-0.19. Seule la corrélation entre PROF et TRAN est importante. L'équation de la droite de régression est : TRAN=0.1977 PROF - 3.15. La part de la variance de TRAN "expliquée par" PROF est de $\frac{Var(\widehat{TRAN})}{Var(TRAN)} = r^2 = 0.87$.

Exercice 5
On mène une étude sur les variations circadiennes de la charge mentale induite par une tâche simple et répétitive. (*circadien signifie "sur un cycle de 24 heures"*).
On considère un échantillon homogène de sujets et on relève, à différents moments de la journée :
- la vitesse d'exécution d'une tâche répétitive simple (nombre d'appuis sur un bouton par minute)
- l'indice de charge mentale induite (mesuré à partir du temps de réaction à un stimulus auditif simple).

On obtient les résultats suivants (moyennes obtenues sur l'ensemble des sujets observés).

Moment	8	10	12	14	16	18	20	22	24
Vitesse	64,54	66,61	71,01	70,10	70,08	68,42	66,63	64,12	63,12
Indice	1,117	1,130	1,171	1,140	1,141	1,129	1,107	1,072	1,052

1) Construire un nuage de points en plaçant en abscisse la variable "moment de la journée", en ordonnée la vitesse, et en choisissant judicieusement les unités.
D'après ce graphique :
- semble-t-il exister une relation entre le moment de la journée et la vitesse ?
- serait-il pertinent de calculer un coefficient de corrélation linéaire pour évaluer l'intensité de cette relation ?
2) Mêmes questions pour les variables vitesse et indice de charge mentale.
3) Calculer la covariance et le coefficient de corrélation linéaire entre les variables vitesse (x_i) et indice de charge mentale (y_i). La corrélation est-elle significative au seuil de 1% ? Quelle est la part de la variance des y_i qui est "expliquée" par celle des x_i ?
Vu la faible amplitude des variations de l'indice, on aura soin de garder un nombre suffisant de décimales dans les calculs intermédiaires. On utilisera par ailleurs les résultats intermédiaires suivants :
$\sum x_i = 604,63$; $\sum y_i = 10,059$; $\sum x_i^2 = 40686,3023$, $\sum y_i^2 = 11,253289$; $\sum x_i y_i = 676,53294$
4) Déterminer une équation de la droite de régression de l'indice de charge mentale en fonction de la vitesse. Construire cette droite sur le graphique précédent.
Réponses : 1 et 2) Il semble exister une relation entre le moment de la journée et la vitesse, mais cette relation n'est pas linéaire, et ne peut donc pas être étudiée à l'aide d'un coefficient de corrélation. En revanche, il semble exister une relation linéaire entre la vitesse et l'indice de charge mentale.
3) $Cov(x_i, y_i) = \frac{676.53294}{9} - \frac{604.63}{9} \times \frac{10.059}{9} = 0.084$. $Var(x_i) = \frac{40686.3023}{9} - (\frac{604.63}{9})^2 = 7.39$. σ_x*=2.72. De même,* $\sigma_y = 0.0344$ *et finalement,* $\rho = \frac{0.084}{2.72 \times 0.0344} = 0.899$. *Il existe donc une forte corrélation positive entre ces deux variables.*
4) Equation de la droite de régression : y=0.0114 x + 0.354

Exercice 6
Une batterie de tests a été soumise à 42 élèves de classes de Cinquième. Cette batterie comprenait notamment un test numérique (variable N) et un test verbal (variable V). Les scores obtenus sont indiqués dans le tableau ci dessous :

Suj.	AG	BAK	BAR	BED	BEN	BO	CA	CE	CH	CO	CU	DAL	DAS	DE
N	12	6	15	12	12	10	9	8	10	15	13	8	9	15
V	16	15	23	16	13	17	14	13	19	18	14	13	20	26

Suj.	DJ	FL	GAB	GAG	GAU	GR	HA	HE	KI	LAB	LAE	LI	LOM	LOU
N	10	8	12	15	10	11	6	11	6	13	6	5	15	7
V	13	15	18	20	19	15	14	25	10	20	12	9	23	15

Suj.	MA	MO	NOR	NOU	PA	PI	RE	RI	RO	SA	SC	TH	VIE	VIT
N	6	10	5	11	13	11	12	15	11	13	15	6	2	15
V	8	17	10	19	8	14	17	18	21	21	15	19	10	21

1) Représenter graphiquement le nuage de points (N_i, V_i).
2) Calculer la moyenne et la variance de chacune des deux séries (N_i) et (V_i) , puis la covariance et le coefficient de corrélation des deux séries. Commenter le coefficient de corrélation obtenu.
Aide pour les calculs : $\sum N_i = 434$, $\sum V_i = 683$, $\sum N_i^2 = 4974$, $\sum V_i^2 = 11915$, $\sum N_i V_i = 7445$.
3) Déterminer une équation de la droite de régression de V selon les valeurs de N. Tracer cette droite sur le graphique précédent.
Réponses : 2) $\overline{N} = 10.33$ *;* $\overline{V} = 16.26$ *;* $\sigma^2(N) = 11.651$ *;* $\sigma^2 V = 19.241$ *;* $Cov(N, V) = 9.22$ *;* $\rho = 0.62$.
3) Equation de la droite de régression : $V = 0.79N + 8.099$.

Exercice 7
Dans une expérience de perception, on étudie l'évaluation des longueurs de figures géométriques. Le sujet est invité à évaluer les longueurs des figures, en s'aidant d'une figure de référence dont il connaît la longueur (9 cm). On note X la variable *longueur des figures* et Y la variable *évaluation des longueurs*.

Partie I
Dans la *condition 1*, les figures sont 11 *bâtonnets*. Les données recueillies pour un sujet sont les suivantes :

x_i	2.5	4.6	6.3	7.6	8.5	9.0	9.5	10.4	11.7	13.4	15.5
y_i	2.8	4.4	6.2	7.8	8.2	9.0	9.6	10.6	12.0	13.6	15.2

Aide pour les calculs : $\sum y_i = 99.4$; $\sum y_i^2 = 1039.24$.

1) Représenter sur un graphique le nuage de points (x_i, y_i) et la droite d'équation $y = x$. Comment peut-on qualifier la qualité de l'ajustement ainsi obtenu ?
2) Calculer les écarts $e_i = y_i - x_i$, puis la moyenne des carrés de ces écarts. Comparer la moyenne des carrés des écarts à la variance de Y.
3) Dans cette condition expérimentale, on fait l'hypothèse selon laquelle les évaluations des longueurs sont *égales* aux longueurs des figures représentées. Ce modèle constitue-t-il une explication satisfaisante des observations ?

Partie II
Dans la *condition 2*, les figures sont des *cercles* de périmètres égaux aux longueurs des bâtonnets de la condition 1. On note U la variable *périmètre de la figure* et V la variable *évaluation du périmètre*. Les données recueillies sont alors les suivantes :

u_i	2.5	4.6	6.3	7.6	8.5	9.0	9.5	10.4	11.7	13.4	15.5
v_i	1.8	3.6	5.8	7.2	8.4	9.0	9.8	11.0	13.2	16.1	21.0

Aide pour les calculs :
$\sum u_i = 99.0$; $\sum v_i = 106.9$; $\sum u_i^2 = 1033.22$; $\sum v_i^2 = 1344.73$; $\sum u_i v_i = 1167.90$.

1) Représenter sur un nouveau graphique le nuage de points (u_i, v_i) et la droite d'équation $v = u$.
Cette droite constitue-t-elle un ajustement satisfaisant des données observées ?
2) Calculer la covariance et le coefficient de corrélation linéaire des variables U et V.
Déterminer une équation de la droite de régression de V par rapport à U et tracer cette droite sur le graphique.
3) Quelle est la part de la variance de la variable V qui est "expliquée" par ce modèle ?

Ce modèle constitue-t-il une explication satisfaisante des observations ?
Réponses : I-2) $\sum e_i^2 = 0.54$ *;* $Var(Y) = 12.82$*. Le modèle proposé semble représenter correctement le comportement des sujets.*
II-2) $Cov(u_i, v_i) = 18.71$ *;* $\rho = 0.98$*. Equation de la droite de régression :* $V = 1.45U - 3.31$*.*
II-3) La part de la variance expliquée est 97.3%.

Exercice 8
On dit que dans une famille, les aînés ont tendance à être plus indépendants que leurs cadets. Un chercheur élabore une échelle d'indépendance en 25 points et procède à l'évaluation de 20 aînés et du frère ou de la sœur qui suit directement chacun des aînés. Imaginons qu'il obtienne les résultats suivants :

Paire	Aîné	Cadet
1	8	9
2	13	15
3	8	10
4	5	7
5	12	10
6	15	13
7	5	8
8	15	12
9	16	13
10	5	9

Paire	Aîné	Cadet
11	17	13
12	12	8
13	2	7
14	13	8
15	19	14
16	18	12
17	14	8
18	17	11
19	18	12
20	20	10

1) Quels sont les individus statistiques et les variables étudiés ? Calculer la moyenne des scores des aînés et celle des scores des cadets et formuler une conclusion descriptive.
2) Un collaborateur du premier chercheur suggère que la différence observée sur une paire dépend essentiellement du score de l'aîné.
a) Pour chaque paire, on appelle x_i le score de l'aîné et y_i la différence (algébrique) des scores entre l'aîné et le cadet.
Calculer le coefficient de corrélation entre les deux séries étudiées.
On pourra utiliser les résultats intermédiaires suivants : $\sum x_i = 252$ *;* $\sum y_i = 43$ *;* $\sum x_i^2 = 3722$ *;* $\sum y_i^2 = 415$ *;* $\sum x_i y_i = 920$*.*
b) Quel commentaire peut-on faire sur le coefficient de corrélation obtenu ?
c) Déterminer une équation de la droite de régression des (y_i) par rapport aux (x_i).
d) Représenter sur un graphique le nuage de points (x_i, y_i) et la droite déterminée au c).
3) Suggérer d'autres facteurs de variation que l'on aurait pu prendre en compte dans une telle étude.
Réponses : 1) $T_+ = 164$; $T_- = 46$*. En utilisant la table de Wilcoxon, on trouve, au seuil de 5%,* $T_{m,c} = 60$*. L'hypothèse "les ainés sont plus indépendants que leurs cadets" est donc confirmée. On peut aussi utiliser l'approximation par une loi normale. On obtient* $Z_{obs} = 2.18$ *et la conclusion est la même.*
2) $Cov(x_i, y_i) = 18.91$; $r_{obs} = 0.9002$*. Pour ddl* $= 18$*, un seuil de 5%, et un test bilatéral, on a :* $r_{crit} = 0.44$*. La corrélation est donc significative.*
La droite de régression a pour équation : $Y = 0.69X - 6.56$*.*
3) L'étude ne tient pas compte de facteurs de variation tels que le sexe, la différence d'âge, la composition de la famille, etc.

Exercice 9

Wagner, Compas et Howell (1988) ont étudié la relation entre le stress et la santé mentale chez des universitaires de première année. A l'aide d'une échelle qu'ils ont mise au point, ils mesurent le stress perçu par le sujet dans son cadre social et son environnement. Ils ont également demandé aux étudiants de remplir la liste de contrôle d'Hopkins qui évalue la présence ou l'absence de 57 symptômes psychologiques.
Pour dix des sujets étudiés, ils ont obtenu les résultats suivants :

Sujet	Stress (x_i)	Symptômes (y_i)
1	30	99
2	27	94
3	9	80
4	20	70
5	3	61
6	16	86
7	5	62
8	10	81
9	23	74
10	34	121

1) Construire un nuage de points en plaçant en abscisse la variable "stress" en ordonnée la variable "symptômes" et en choisissant judicieusement les unités.
Commenter ce graphique.
2) a) Calculer la covariance et le coefficient de corrélation linéaire entre les deux variables.
N.B. On pourra utiliser les résultats suivants : $\sum x_i^2 = 4185 \quad \sum y_i^2 = 71576$; $\sum x_i y_i = 16123$.
b) Commenter le coefficient de corrélation obtenu.
c) Déterminer une équation de la droite de régression des (y_i) par rapport aux (x_i).
d) Représenter sur un graphique le nuage de points (x_i, y_i) et la droite déterminée au c).
Réponses : 2) $Cov(x_i, y_i) = 146.74$; $r = 0.82$. *Le coefficient trouvé montre une liaison entre les variables. La droite de régression a pour équation :* $Y = 1.39X + 58.20$.

Test du χ^2

Exercice 10

Un psychologue fait l'hypothèse que certaines difficultés du langage écrit chez l'enfant sont en relation avec des facteurs dits *instrumentaux*, notamment la latéralisation. Sur un échantillon, ce praticien recueille les données suivantes :

latéralisation	difficultés du langage écrit	
	oui	non
droitier	12	25
ambidextre	14	7
gaucher	17	13

Examiner ces résultats et formuler avec précision une hypothèse statistique. L'éprouver à l'aide du test statistique approprié et indiquer si, au seuil $\alpha = .05$, on peut conclure que la latéralisation a une incidence sur les difficultés du langage écrit.

Réponse : Effectifs théoriques, contributions au χ^2 :

	oui	*non*
d.	*18.08*	*18.92*
a.	*10.26*	*10.74*
g.	*14.66*	*15.34*

	oui	*non*
d.	*2.04*	*1.95*
a.	*1.36*	*1.30*
g.	*0.37*	*0.35*

$\chi^2_{obs} = 7.39$; Pour ddl $= 2$ et $\alpha = 0.05$, $\chi^2_c = 5.99$. Il existe donc une relation entre les deux variables.

Exercice 11

Pour 650 enfants en consultation psychiatrique, on a observé d'une part le rang dans la fratrie : *rang 1, rang 2, rang 3* et d'autre part le diagnostic porté sur l'enfant : réaction de type soit dépressif (d), soit anxieux (a), soit schizophrénique (s). Les observations sont résumées dans le tableau suivant :

	Diagnostic			
Rang	d	a	s	Total
rang 1	98	70	57	225
rang 2	108	75	61	244
rang 3	61	68	52	181
Total	267	213	170	650

1) Quelles sont les variables statistiques étudiées. Quelle est la nature de chacune d'elles ? Quelle est la nature (tableau protocole, tableau d'effectifs, tableau de contingence) du tableau ci-dessus ?

2) On se propose d'étudier à l'aide d'un test du χ^2 s'il existe un lien entre le type de maladie et le rang dans la fratrie.

a) Calculer le tableau des effectifs théoriques correspondant au tableau ci-dessus.

b) La valeur χ^2_{obs} observée sur l'échantillon étudié est obtenue comme somme des contributions des différentes cases du tableau.

Détailler le calcul de la contribution de la première case.

c) Le calcul complet donne : $\chi^2_{obs} = 5.75$.

Déterminer le nombre de degrés de liberté. Utiliser la table du χ^2 pour déterminer au seuil de 5% si la liaison est significative.

Réponses : 1) Variables : rang (ordinale) et diagnostic (nominale). Tableau de contingence.

2) a)

	d	*a*	*s*
r1	*92.4*	*73.7*	*58.8*
r2	*100.2*	*80*	*63.8*
r3	*74.3*	*59.3*	*47.3*

2) b) Contribution première case : $\frac{(98-92.4)^2}{92.4} = 0.34$.
2) c) $ddl = (l-1)(c-1) = 4$. *Au seuil de 5%,* $\chi^2_c = 9.49$. *L'hypothèse d'indépendance des deux variables ne peut pas être rejetée.*

Exercice 12
Lors d'une enquête, on a interrogé 150 personnes prises au hasard sur leurs connaissances en langues étrangères. Les résultats obtenus sont les suivants :

	Hommes	Femmes
Anglais	37	24
Allemand	9	19
Espagnol	16	15
Aucune	20	10

Les connaissances en langues étrangères dépendent-elles du sexe dans la population dont est issu l'échantillon étudié ? On répondra à cette question en effectuant un test au seuil de 5%.
Réponses : Le tableau proposé est un tableau de contingence. On va donc procéder à un test du χ^2. *Effectifs théoriques :*

	Hommes	*Femmes*
Anglais	*33.35*	*27.65*
Allemand	*15.31*	*12.69*
Espagnol	*16.95*	*14.05*
Aucune	*16.40*	*13.60*

$\chi^2_{obs} = \frac{(37-33.35)^2}{33.35} + \ldots + \frac{(10-13.60)^2}{13.60} = 8.48$
Pour $\alpha = 5\%$ *et* $ddl = (2-1)(4-1) = 3$, *on a :* $\chi^2_c = 7.815$. *La différence entre les sexes est donc significative.*

Exercice 13
Dans le cadre d'une enquête sur le SIDA réalisée en Allemagne durant l'été 1990 (A. Hahn, W.H. Eirmbter et R. Jacob), on a interrogé 2089 personnes. Le questionnaire comportait notamment l'item suivant : *Le sida est un péril omniprésent. Indiquez si vous êtes : d'accord, indécis, pas d'accord.* Le croisement de la réponse du sujet avec son âge donne le tableau de contingence suivant :

Classe d'âge	d'accord	indécis	pas d'accord	Total
18 à < 30	43	116	365	524
30 à < 40	36	116	273	425
40 à < 50	32	95	217	344
50 à < 60	38	114	167	319
60 et plus	67	160	250	477
Total	216	601	1272	2089

1) Réaliser un test du χ^2 permettant de répondre à la question suivante : "Les réponses des sujets dépendent-elles de leur âge ? "
2) Comparer de même à l'aide d'un test du χ^2 les réponses des deux dernières classes

d'âge, puis les réponses des moins de 30 ans à celles des 60 ans et plus.
Réponses : 1) Le tableau des effectifs théoriques est donné par :

Classe d'âge	*d'accord*	*indécis*	*pas d'accord*
18 à < 30	*54.2*	*150.8*	*319.0*
30 à < 40	*43.9*	*122.3*	*258.8*
40 à < 50	*35.6*	*99.0*	*209.4*
50 à < 60	*33.0*	*91.8*	*194.2*
60 et plus	*49.3*	*137.2*	*290.5*

Celui des contributions au χ^2 *est donné par :*

Classe d'âge	*d'accord*	*indécis*	*pas d'accord*
18 à < 30	*2.31*	*8.01*	*6.61*
30 à < 40	*1.44*	*0.32*	*0.78*
40 à < 50	*0.36*	*0.16*	*0.27*
50 à < 60	*0.76*	*5.58*	*3.82*
60 et plus	*6.33*	*3.78*	*5.63*

On obtient $\chi^2_{obs} = 46$. *Or pour un seuil de 1% et ddl* $= (5-1)(3-1) = 8$, *on a :* $\chi^2_c = 20.09$. *La réponse du sujet est donc dépendante de son âge.*
2) En ne considérant que les deux dernières classes d'âge, on obtient le tableau d'effectifs théoriques suivant :

Classe d'âge	*d'accord*	*indécis*	*pas d'accord*
50 à < 60	*42.1*	*109.8*	*167.1*
60 et plus	*62.9*	*164.2*	*249.9*

On obtient alors $\chi^2_{obs} = 0.92$. *Or, pour* $\alpha = 5\%$ *et ddl* $= 2$, $\chi^2_c = 5.99$. *Les réponses des sujets sont cette fois indépendantes de leur appartenance à l'une ou l'autre classe d'âge. Dans le dernier cas,* $\chi^2_{obs} = 31.62$, *ce qui est significatif d'une dépendance.*

Exercice 14

Extrait de "Registres mis en jeu par la notion de fonction". I. Guzman-Retamal - Annales de Didactique et de Sciences Cognitives. N.2 (1989) - IREM Strasbourg

Dans une classe expérimentale de 20 élèves et une classe témoin de 24 élèves, on a utilisé deux pédagogies différentes pour enseigner un chapitre du cours de mathématiques. On fait passer aux deux classes un test de connaissances commun, composé d'un certain nombre d'items, et on compare la réussite des deux classes item par item. Commentez et exploitez les trois tableaux suivants :

Item	4.1	6.2	3.2
Expérimentale	.90	.65	0
Témoin	.92	.62	.04

Item	4.2	5.2	10.2
Expérimentale	.65	.85	.50
Témoin	.25	.33	.12

Item	1.2	1.3	8.1
Expérimentale	.25	.25	.20
Témoin	.71	.67	.75

N.B. La correction de Yates pourra être utilisée lors d'éventuels calculs de χ^2.
Réponses : Avant de procéder à un éventuel test du χ^2, *il faut, pour chaque item, reconstituer le tableau d'effectifs observés. Par exemple, on obtient pour l'item 6.2 :*

	Succès	*Echec*	*Total*
Expérimentale	*13*	*7*	*20*
Témoin	*15*	*9*	*24*
Total	*28*	*16*	*44*

On obtient ainsi pour cet item $\chi^2_{obs} = 0.00022$ *(sans correction de Yates). Pour les items 4.1, 6.2 et 3.2, il n'y a pas de différence significative entre les deux groupes (mais seul l'item 6.2 permet de calculer un* χ^2*). Pour les autres items, le test du* χ^2 *montre une différence significative entre les deux groupes. Notez que la méthode utilisée par les auteurs de l'article, qui consiste à tester chaque item indépendamment des autres, est très critiquable.*

Exercice 15

Un psychosociologue cherche à étudier les attitudes de trois groupes sociaux professionnellement contrastés concernant l'éducation des enfants. Il fait ainsi l'hypothèse que les groupes auront des réponses différenciées pour un ensemble de principes et de méthodes à propos de l'éducation des enfants. Tous les items de son questionnaire utilisent une échelle ordinale sémantique permettant d'évaluer le degré d'adhésion des sujets à certaines propositions. L'échelle de mesure est constituée de cinq catégories ordonnées de réponses : *absolument opposé, opposé, indifférent, d'accord, absolument d'accord*. Les distributions obtenues auprès des trois groupes à propos d'un item sont données dans le tableau ci-dessous :

	abs. opposé	opposé	indifférent	d'accord	abs.d'accord
Groupe 1	8	10	12	11	11
Groupe 2	12	11	18	25	17
Groupe 3	21	20	10	13	11

1) Quelles sont les variables étudiées ? Quelle est la nature de chacune d'elles ? Quelle est la nature du tableau proposé ?

2) Former le tableau obtenu en regroupant les modalités "absolument opposé" et "opposé" d'une part, les modalités "d'accord" et "absolument d'accord" d'autre part.

3) Tester la validité de l'hypothèse du chercheur en procédant à un test du χ^2 au seuil de 5 % sur le tableau obtenu à la question 2.

Réponses

1) Les variables étudiées sont d'une part le groupe social auquel appartient le sujet, d'autre part, la réponse du sujet à une question relative à l'éducation des enfants. La variable "groupe" est nominale, la variable "réponse" est ordinale. Le tableau proposé est un tableau de contingence.

2) Le tableau obtenu est alors :

	oppos.	*indifférent*	*accord*
Groupe 1	*18*	*12*	*22*
Groupe 2	*23*	*18*	*42*
Groupe 3	*41*	*10*	*24*

3) Les tableaux des effectifs théoriques et des contributions au χ^2 *sont donnés par :*

	oppos.	*indifférent*	*accord*
Groupe 1	*20.3*	*9.9*	*21.8*
Groupe 2	*32.4*	*15.8*	*34.8*
Groupe 3	*29.3*	*14.3*	*31.4*

	oppos.	*indifférent*	*accord*
Groupe 1	*0.26*	*0.44*	*0.002*
Groupe 2	*2.73*	*0.30*	*1.50*
Groupe 3	*4.68*	*1.28*	*1.75*

On obtient alors $\chi^2_{obs} = 12.97$. *Ici, ddl* $= 4$. *Au seuil de 5%, on a :* $\chi^2_{crit} = 9.488$. *On accepte donc l'hypothèse d'un lien entre le groupe social et la réponse donnée.*

Exercice 16

Une étude par enquête a été réalisée en 1990 sur la représentation de l'arbre d'ornement. Cette étude repose sur l'hypothèse générale d'un lien entre la représentation de l'arbre et l'horizon temporel des sujets. Quatre terrains d'enquête ont été sélectionnés :

1) Paris XIIè (quartier pauvre en espaces verts)
2) Paris XIVè (quartier riche en espaces verts)
3) Evry (quartier pavillonnaire)
4) Evry (quartier d'immeubles)

Dans l'item 2.1 du questionnaire, les sujets interrogés devaient donner leur opinion (accord/désaccord) sur l'affirmation :

L'arbre est le lien entre le passé et l'avenir.

Les tableaux de contingence obtenus en croisant les réponses des sujets d'une part avec leur lieu d'habitation, d'autre part avec leur sexe sont donnés ci-dessous.

	Accord	Désaccord
Paris XII	99	26
Paris XIV	105	17
Evry - Pavillons	43	18
Evry - Immeubles	50	8

TAB. 1 – Croisement réponse/lieu d'habitation

	Accord	Désaccord
Hommes	139	42
Femmes	158	27

TAB. 2 – Croisement réponse/sexe du sujet

1) Effectuer des tests du χ^2 pour déterminer au seuil de 5% si :
a) Les réponses dépendent du lieu d'habitation.
b) Les réponses varient selon le sexe.

Réponses :

1) a) Les tableaux d'effectifs théoriques et de contributions au χ^2 *donnent :*

	Accord	*Désaccord*
Paris XII	*101.43*	*23.57*
Paris XIV	*99*	*23*
Evry - Pavillons	*49.50*	*11.50*
Evry - Immeubles	*47.07*	*10.93*

	Accord	*Désaccord*
Paris XII	*0.058*	*0.251*
Paris XIV	*0.304*	*1.565*
Evry - Pavillons	*0.854*	*3.674*
Evry - Immeubles	*0.183*	*0.787*

On obtient $\chi^2_{obs} = 7.74$. *Ici, ddl* $= 3$ *et donc, au seuil de 5%,* $\chi^2_{crit} = 7.81$. *On accepte donc l'hypothèse* H_0 *d'indépendance entre les variables.*

b) Dans ce cas, les effectifs théoriques et les contributions au χ^2 *sont donnés par :*

	Accord	*Désaccord*
Hommes	*146.88*	*34.12*
Femmes	*150.12*	*34.88*

	Accord	*Désaccord*
Hommes	*0.42*	*1.82*
Femmes	*0.41*	*1.78*

On obtient $\chi^2_{obs} = 4,43$. Ici, ddl $= 1$ et donc, au seuil de 5%, $\chi^2_{crit} = 3.84$. On accepte l'hypothèse H_1 de dépendance entre les variables : les réponses varient selon le sexe.

Exercice 17

Une vaste enquête sur le thème "les jeunes et la culture de l'écran" a été réalisée d'avril à juin 1997 auprès d'un échantillon de 1417 jeunes de 6 à 17 ans par J. Jouët et D. Pasquier. En particulier, pour 1087 d'entre eux, les auteurs ont évalué :
– la durée d'écoute de la télévision un jour de week-end ;
– la fréquence de lecture de livres.
Le tableau obtenu en croisant les résultats obtenus pour ces deux items est le suivant :

	une demi-heure à une heure	deux heures	trois heures et plus
Gros lecteurs	126	157	172
Petits lecteurs	72	111	162
Non lecteurs	49	80	158

1) Quelles sont les variables étudiées ? Quelle est la nature de chacune d'elles ? Quelle est la nature du tableau proposé ?
2) On veut montrer qu'il existe un lien entre la durée d'écoute de la télévision et la fréquence de lecture de livres, en utilisant un test du χ^2 au seuil de 1%.
a) Former le tableau des effectifs théoriques correspondant aux effectifs observés ci-dessus.
b) Calculer les contributions au χ^2 de chacun des couples de modalités, et la valeur du χ^2 observé.
c) Déterminer le nombre de degrés de liberté, la valeur critique du χ^2 au seuil de 1%, et conclure.
3) On veut étudier plus finement le lien existant entre les deux variables étudiées.
a) Calculer la fréquence f_{jk} d'apparition de chaque couple (j, k) de modalités des variables (en pourcentage de l'effectif total).
b) Calculer de même les fréquences marginales (f_j) et (f_k) pour chaque variable prise isolément.
c) Etant donné la modalité j de la première variable et la modalité k de la seconde, on appelle *taux de liaison entre j et k*, la quantité :

$$t_{jk} = \frac{f_{jk} - f_j f_k}{f_j f_k}$$

Calculer les taux de liaison correspondant aux neuf cases du tableau ci-dessus.
d) On dit que deux modalités s'attirent si le taux de liaison correspondant est positif, qu'elles se repoussent s'il est négatif. Dans l'exemple proposé, quelles sont les modalités qui s'attirent ? Quelles sont celles qui se repoussent ?
Quelle interprétation peut-on donner de ces résultats ?
Réponses :
1) a) Les deux variables étudiées sont la durée d'écoute de la télévision, regroupée en trois modalités et la fréquence de lecture de livres, elle aussi regroupée en 3 modalités. Ces deux variables, telles qu'elles sont codées, sont ordinales. Le tableau proposé est un tableau de contingence.
2) Le tableau des effectifs théoriques et celui des contributions au χ^2 sont donnés par :

	une demi-heure à une heure	*deux heures*	*trois heures et plus*
Gros lecteurs	*103.4*	*145.7*	*205.9*
Petits lecteurs	*78.4*	*110.4*	*156.2*
Non lecteurs	*65.2*	*91.9*	*129.9*

	une demi-heure à une heure	*deux heures*	*trois heures et plus*
Gros lecteurs	*4.94*	*0.88*	*5.58*
Petits lecteurs	*0.52*	*0.00*	*0.22*
Non lecteurs	*4.02*	*1.54*	*6.08*

On trouve $\chi^2_{obs} = 23.78$. *Ici, ddl* $= (3-1)(3-1) = 4$. *Au seuil de 1%, on a donc* $\chi^2_{crit} = 13.277$. *Le résultat est significatif d'une liaison entre les deux variables étudiées.*
3) Les fréquences des couples de modalités et des marges du tableau sont données par :

f_{jk}	*.5 à 1*	*2*	*3 et +*	f_j
GL	*11.6%*	*14.4%*	*15.8%*	*41.9%*
PL	*6.6%*	*10.2%*	*14.9%*	*31.7%*
NL	*4.5%*	*7.4%*	*14.5%*	*26.4%*
f_k	*22.7%*	*32.0%*	*45.3%*	*100%*

Dans le calcul des t_{jk}, *on veillera à ne pas oublier qu'il s'agit de pourcentages, c'est-à-dire de valeurs entre 0 et 1.*

t_{jk}	*.5 à 1*	*2*	*3 et +*
GL	*0.21*	*0.078*	*-0.16*
PL	*-0.08*	*0.005*	*0.03*
NL	*-0.24*	*-0.12*	*0.21*

On constate une attirance entre les modalités "gros lecteur" et "faible durée d'écoute" d'une part, entre les modalités "non lecteur" et "forte durée d'écoute".
A l'inverse, on constate une répulsion entre "gros lecteur" et "forte durée d'écoute" de même qu'entre "non lecteur" et "faible durée d'écoute".
L'interprétation évidente est : plus on écoute la télévision et moins on a le temps de lire. Cette interprétation est cependant trop simpliste, car la même enquête montrait des résultats assez différents des précédents en ce qui concerne la lecture de magazines.

Test d'adéquation à une loi théorique

Exercice 18 Dans l'expérience complète de Mendel, deux attributs des pois étaient considérés : la couleur (*jaune* ou *verte*, et la forme (*ronde* ou *ridée*), avec la triple hypothèse suivante : *rond* dominant et *ridé* récessif ; *jaune* dominant et *vert* récessif ; indépendance entre couleur et forme. Le modèle mendélien prédit alors qu'à la deuxième génération, on obtiendra en moyenne la distribution suivante :

Types de pois	*rond jaune*	*rond vert*	*ridé jaune*	*ridé vert*
Fréquences théoriques	$\frac{9}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$

On réalise l'expérience de Mendel et on obtient les résultats expérimentaux suivants :

Types de pois	*rond jaune*	*rond vert*	*ridé jaune*	*ridé vert*
Effectifs	325	116	101	30

Comparer à l'aide de la distance du χ^2 et d'un test du χ^2 l'adéquation entre la distribution théorique et la distribution observée (N.B. Le seul paramètre théorique évalué à partir des observations est ici l'effectif total. Le nombre de degrés de liberté est donc égal à $4-1=3$.)

Réponse : Distribution théorique pour un effectif de 572 individus statistiques :

RoJ	*RoV*	*RiJ*	*RiV*
322	*107*	*107*	*36*

$\chi^2_{obs} = \frac{(325-322)^2}{322} + \ldots + \frac{(30-36)^2}{36} = 2.12$. *Or, pour un seuil de 5% et ddl* $= 3$, *on a* $\chi^2_c = 7.81$. *On accepte donc l'hypothèse selon laquelle l'échantillon observé est issu d'une population distribuée selon la loi théorique indiquée.*

Exercice 19

C. Guimelli (1988) a réalisé une étude sur la représentation de la chasse auprès de 128 chasseurs faisant partie de sociétés de chasse du Gard et de l'Hérault. Les données ci-dessous permettent de comparer la répartition selon la catégorie socio-professionnelle dans la population des chasseurs au niveau national et dans l'échantillon régional de cette étude, échantillon tout venant constitué sans contrôle a priori de la CSP.

C.S.P.	Répartition nationale	Echantillon étudié
Agriculteurs	29.3%	35
Ouvriers et employés	32.6%	30
Commerçants et artisans	8.9%	18
Cadres moyens	7.5%	8
Professions libérales	3.3%	6
Cadres supérieurs et industriels	2.7%	3
Retraités	11.7%	20
Etudiants	3.1%	8
Divers	0.9%	–
Total	100%	128

Au vu des données fournies, peut-on affirmer que les membres des sociétés de chasse de l'Hérault et du Gard se répartissent par CSP de la même manière que la population nationale des chasseurs ? Répondre à cette question en procédant à un test d'adéquation à une loi théorique après regroupement des modalités d'effectifs trop faibles.

Réponse : effectifs théoriques et contributions au χ^2 *:*

CSP	*A*	*O.E.*	*C.A.*	*C.M.*	*PL/CS*	*R.*	*E/D*	*Tot.*
Théo.	*37.5*	*41.7*	*11.4*	*9.6*	*7.7*	*15.0*	*5.1*	*128*
Ctr	*0.17*	*3.28*	*3.82*	*0.27*	*0.22*	*1.67*	*1.65*	*11.07*

On a : $\chi^2_{obs} = 11.07$. *Ici, ddl=7, et pour un seuil de 5%,* $chi^2_{crit} = 14.067$. *On n'a donc pas mis en évidence de différence dans la répartition en CSP entre les deux populations.*

Exercice 20

1) Dans une étude devenue classique (1939), deux chercheurs ont montré à des enfants noirs une poupée noire et une poupée blanche en leur demandant de choisir celle avec laquelle ils voudraient jouer. Sur 252 enfants, 169 ont choisi la poupée blanche tandis que 83 préféraient la poupée noire.

Est-il possible que le choix de l'une ou l'autre poupée soit fait au hasard ? Pour tester cette hypothèse, procéder à un test d'ajustement de l'échantillon des 252 observations par rapport à la loi théorique :

	Poupée blanche	Poupée noire
Fréquence	50%	50%

2) Un deuxième groupe de recherche a reproduit l'expérience précédente en 1970. Les études n'étaient pas exactement identiques, mais les résultats se sont avérés intéressants : sur 89 enfants noirs, 28 ont choisi la poupée blanche et 61 ont préféré la poupée noire. Etudier de même si le choix de la poupée se fait au hasard.

3) Une troisième équipe de chercheurs réunit les résultats précédents dans un tableau de contingence et procède à un test du χ^2 sur ce tableau :

	Exp. 1	Exp. 2
Poupée noire	83	61
Poupée blanche	169	28

Formuler avec précision l'hypothèse nulle et l'hypothèse alternative correspondant à ce test. Réaliser le test et conclure.

Réponses : Pour les questions 1 et 2 le test d'ajustement conduit aux tableaux de calcul suivants :

	P. Bl.	*P. N.*	*Tot.*
Théo.	*126*	*126*	*252*
Obs.	*169*	*83*	*252*
Ctr.	*14.7*	*14.7*	*29.4*

	P. Bl.	*P. N.*	*Tot.*
Théo.	*44.5*	*44.5*	*89*
Obs.	*28*	*61*	*89*
Ctr.	*6.12*	*6.12*	*12.2*

Dans les deux cas, ddl $= 1$*, et donc, au seuil de 1%,* $\chi^2_{crit} = 6.635$*. Dans les deux cas, le choix ne se fait donc pas au hasard.*

3) Les variables mises en jeu sont ici d'une part le choix de l'enfant poupée noire ou poupée blanche) et d'autre part l'expérience considérée (expérience de 1939, expérience de 1970).

L'hypothèse H_0 *est l'indépendance des variables. Autrement dit, les choix des enfants sont les mêmes en 1939 et en 1970.*

L'hypothèse H_1 *postule au contraire un changement dans le comportement des enfants entre les deux expériences.*

Le tableau d'effectifs théoriques et celui des contributions au χ^2 *sont donnés par :*

	Exp. 1	*Exp.2*
P.N.	*106.41*	*37.58*
P.Bl.	*145.58*	*51.42*

	Exp. 1	*Exp.2*
P.N.	*5.15*	*14.59*
P.Bl.	*3.76*	*10.66*

On obtient $\chi^2 obs = 34.17$*. Pour ddl* $= 1$ *et un seuil de* 1%*, on a* $\chi^2_{crit} = 6.63$*. L'hypothèse* H_1 *est donc retenue.*