Corrélation. Droites de régression

Exercice 1 *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, la corrélation est-elle significative ?
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 significative. 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 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 : $Cov(F, M) = 9.47$; $\rho = 0.51$. La corrélation est à peine significative à 5%. 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 égelement 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

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 differents 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 4 *Données Tailles*
Le tableau ci-dessous donne la taille de 10 garçons (variable Z) ainsi que la taille de leurs parents (le père X et la mère Y).

	X	Y	Z
i1	160.0	161	165.0
i2	165.0	155	162.5
i3	170.0	155	165.0
i4	172.5	165	175.0
i5	175.0	170	180.0
i6	180.0	166	177.5
i7	185.0	167	180.0
i8	187.5	172	190.0
i9	190.0	175	195.0
i10	195.0	168	187.5

On cherche s'il existe une relation entre la taille du fils et celle de ses parents et, si oui, quelle est la part respective de la mère et du père. Pour cela, on procède à la régression de Z sur X et Y. On donne les résultats intermédiaires suivants :
$\sum X_i = 1780$; $\sum Y_i = 1654$; $\sum Z_i = 1777,5$
$\sum X_i^2 = 318012.5$, $\sum Y_i^2 = 273974$; $\sum Z_i = 317068.75$
$\sum X_i Y_i = 294932.5$; $\sum X_i Z_i = 317437.5$; $\sum Y_i Z_i = 294632.5$
$Var(X) = 117.25$; $Var(Y) = 40.24$; $Var(Z) = 111.81$
$Cov(X,Y) = 52.05$; $Cov(X,Z) = 104.25$; $Cov(Y,Z) = 63.40$.
1) Quels sont les coefficients de corrélation des variables prises deux à deux ?
2) On utilise un logiciel de traitement statistique pour déterminer l'équation du plan de régression de Z par rapport à X et Y. On obtient :
$Z = 0.4455X + 0.9993Y - 66.83$.
Calculer les valeurs estimées de Z pour chacun des 10 individus statistiques (variable $\hat{Z}$).
On donne par ailleurs : $\sum \hat{Z}_i = 317060.06$ et $\sum Z_i \hat{Z}_i = 317054.34$.
3) Déterminer le coefficient de corrélation multiple.
4) Quelle est la proportion de variance prise en compte par la régression ?
5) Les coefficients de corrélation partielle sont donnés par : $R_{xz,y} = 0.91$; $R_{yz,x} = 0.95$
Quel est, de la taille du père et de celle de la mère, le meilleur prédicteur de la taille du fils ?
6) Prédire la taille d'un garçon, sachant que son père mesure 188cm et sa mère 171cm.
Réponses : N.B. Calculs exécutés à l'aide d'un logiciel de traitement statistique.
1) Les coefficients de corrélation des variables prises deux à deux sont donnés par : r(X,Y)=0.76 ; r(X,Z)=0.91 ; r(Y,Z)=0.95.
2) Les valeurs estimées de Z sont données par :

i1	*i2*	*i3*	*i4*	*i5*	*i6*	*i7*	*i8*	*i9*	*i10*
165.34	*161.57*	*163.8*	*174.90*	*181.01*	*179.24*	*182.47*	*188.58*	*192.69*	*187.92*

3) Coefficient de corrélation multiple : R=0.991.

4) D'où $R^2 = 0.98$. Le modèle explique 98% de la variance observée de la variable Z. Cette proportion est très élevée, mais il s'agit de données fictives....
5) Le coefficient de corrélation partielle le plus élevé est celui liant taille de la mère et taille du fils.
6) Taille du fils si X=188 et Y=171 : Z = 188.

Exercice 5 *Données Evalcour*
L'assocation des étudiants d'une grande université (américaine) a publié une évaluation de plus de cent cours enseignés durant le semestre précédent. Les étudiants de chaque cours avaient rempli un questionnaire d'évaluation portant sur différents aspects du cours; l'évaluation se faisait sur une échelle en cinq points (1=très mauvais, 5=excellent). Les données figurant dans les deux tableaux 1 et 2 pages 4 et 5 sont les données réelles. Elles représentent les scores moyens enregistrés sur 6 variables pour un échantillon de 50 cours. Ces variables étaient :

Qual-Glob	Pédagogie	Examen	Connaissan	Résultat	Inscriptio
3.4	3.8	3.8	4.5	3.5	21
2.9	2.8	3.2	3.8	3.2	50
2.6	2.2	1.9	3.9	2.8	800
3.8	3.5	3.5	4.1	3.3	221
3	3.2	2.8	3.5	3.2	7
2.5	2.7	3.8	4.2	3.2	108
3.9	4.1	3.8	4.5	3.6	54
4.3	4.2	4.1	4.7	4	99
3.8	3.7	3.6	4.1	3	51
3.4	3.7	3.6	4.1	3.1	47
2.8	3.3	3.5	3.9	3	73
2.9	3.3	3.3	3.9	3.3	25
4.1	4.1	3.6	4	3.2	37
2.7	3.1	3.8	4.1	3.4	83
3.9	2.9	3.8	4.5	3.7	70
4.1	4.5	4.2	4.5	3.8	16
4.2	4.3	4.1	4.5	3.8	14
3.1	3.7	4	4.5	3.7	12
4.1	4.2	4.3	4.7	4.2	20
3.6	4	4.2	4	3.8	18
4.3	3.7	4	4.5	3.3	260
4	4	4.1	4.6	3.2	100
2.1	2.9	2.7	3.7	3.1	118
3.8	4	4.4	4.1	3.9	35
2.7	3.3	4.4	3.6	4.3	32

Tableau 1: Première partie des données

1. la qualité globale des exposés (Qual-Glob)
2. les aptitudes pédagogiques du professeur (Pédagogie)

Qual-Glob	Pédagogie	Examen	Connaissan	Résultat	Inscriptio
4.4	4.4	4.3	4.4	2.9	25
3.1	3.4	3.6	3.3	3.2	55
3.6	3.8	4.1	3.8	3.5	28
3.9	3.7	4.2	4.2	3.3	28
2.9	3.1	3.6	3.8	3.2	27
3.7	3.8	4.4	4	4.1	25
2.8	3.2	3.4	3.1	3.5	50
3.3	3.5	3.2	4.4	3.6	76
3.7	3.8	3.7	4.3	3.7	28
4.2	4.4	4.3	5	3.3	85
2.9	3.7	4.1	4.2	3.6	75
3.9	4	3.7	4.5	3.5	90
3.5	3.4	4	4.5	3.4	94
3.8	3.2	3.6	4.7	3	65
4	3.8	4	4.3	3.4	100
3.1	3.7	3.7	4	3.7	105
4.2	4.3	4.2	4.2	3.8	70
3	3.4	4.2	3.8	3.7	49
4.8	4	4.1	4.9	3.7	64
3	3.1	3.2	3.7	3.3	700
4.4	4.5	4.5	4.6	4	27
4.4	4.8	4.3	4.3	3.6	15
3.4	3.4	3.6	3.5	3.3	40
4	4.2	4	4.4	4.1	18
3.5	3.4	3.9	4.4	3.3	90

Tableau 2: Seconde partie des données

3. la qualité des tests et examens (Examen)
4. la connaissance de la matière dont témoigne le professeur, telle qu'elle est perçue par les étudiants (Connaissan)
5. les résultats auxquels s'attendent les étudiants pour ce cours (Résultat, de très bon à insuffisant)
6. le nombre d'inscriptions à ce cours (Inscriptio)

Les résultats de statistiques descriptives concernant les variables précédentes sont donnés dans le tableau 3.
Les coefficients de corrélation des variables prises deux à deux sont donnés dans le tableau 4.
Les coefficients de l'équation de régression multiple de la première variable en fonction des cinq autres sont donnés par le tableau 5.
Ecrire l'équation de régression correspondante, et la vérifier sur l'extrait donné dans le tableau 6.

	Qual-Glob	Pedagogie	Examen	Connaissan	Resultat	Inscriptio
Effectif	50	50	50	50	50	50
Moyenne	3.55	3.664	3.808	4.176	3.486	88.0
Variance	0.376429	0.283167	0.2432	0.166351	0.123269	21042.2
Ecart-type	0.613538	0.532135	0.493153	0.407862	0.351097	145.059

Tableau 3: Statistiques descriptives sur les données

	Qual-Glob	Pedagogie	Examen	Connaissan	Resultat	Inscriptio
Qual-Glob	1	0.8039	0.5956	0.6818	0.3008	-0.2396
Pedagogie	0.8039	1	0.7197	0.5263	0.4691	-0.4511
Examen	0.5956	0.7197	1	0.4515	0.6100	-0.5581
Connaissan	0.6818	0.5263	0.4515	1	0.2242	-0.1279
Resultat	0.3008	0.4691	0.6100	0.2242	1	-0.3371
Inscriptio	-0.2396	-0.4511	-0.5581	-0.1279	-0.3371	1

Tableau 4: Coefficients de corrélation

Paramètre	Estimation
CONSTANTE	-1.19483
Inscriptio	0.000525491
Examen	0.131981
Resultat	-0.184308
Connaissan	0.488984
Pedagogie	0.763237

Tableau 5: Coefficients de l'équation de de régression

Ligne	Observé	Ajusté	Résidu
1	3.4	3.77339	-0.373387
2	2.9	2.6592	0.240795
3	2.6	2.54643	0.0535724
4	3.8	3.45119	0.348812
5	3.0	2.74242	0.257584

Tableau 6: Comparaison des valeurs observées et des valeurs ajustées

Régresseur	Coef
Pedagogie	0,6544
Examen	0,1213
Connaissan	0,4751
Resultat	-0,1656
Inscriptio	0,1990

Tableau 7: Coefficients de corrélation partielle

Enfin, le dernier tableau (tableau 7) donne les coefficients de corrélation partiels entre la variable Qual-Glob et les prédicteurs.
La valeur du coefficient de corrélation multiple vérifie : $R^2 = 0.755$.

Exercice 6
Aux élections européennes de juin 1984, les votes pour la liste du Front National ont été très variables dans l'espace et leur comparaison avec d'autres variables fait apparaître un certain nombre de relations. Les variables choisies dans le tableau 8 sont les suivantes :

- LPEN : pourcentage de voix de la liste FN ;
- ETRA : pourcentage d'étrangers dans la population en 1982 ;
- DELI : Nombre pondéré de délinquance pour 100 habitants en 1980 ;
- CRCH : Taux mensuel moyen d'évolution du chômage entre le 31.08.81 et le 30.04.83 ;
- TXCH : Pourcentage de chômeurs dans la population active au 30.09.83 ;
- URBA : Pourcentage de population urbaine en 1982.

Les individus statistiques sont ici les régions de France Métropolitaine (ILEF=Ile de France, CHAM=Champagne-Ardennes, etc.). L'échelle régionale n'est certainement pas la meilleure pour une telle étude et les conclusions valent pour les agrégats spatiaux et non des personnes.[1]
Le tableau 9 donne les valeurs des coefficients de corrélation des variables prises deux à deux. On voit que les votes pour l'extrême droite sont fortement corrélés à trois variables : taux d'urbanisation (URBA), taux de délinquance (DELI) et taux d'étrangers (ETRA). D'autre part, ces trois variables sont fortement corrélées entre elles, elles sont donc partiellement redondantes.
Une première régression multiple est réalisée en utilisant les 5 variables explicatives. Le coefficient de corrélation multiple vaut $R = 0.934$ et le coefficient de détermination, $R^2 = 0.872$
Les coefficients de corrélation partielle entre la variable LEPEN et chacune des autres variables sont alors ceux indiqués dans le tableau 10.
On peut tester la significativité de ces coefficients de corrélation. Le nombre de degrés de liberté à prendre en compte est $21 - 6 = 15$. Au seuil de 5%, $r_{crit} = 0.4821$.

[1] D'après *Initiation aux méthodes statistiques en Géographie*, Groupe Chadule, Masson Ed., 1994

REG	LEPEN	ETRA	DELI	CRCH	TXCH	URBA
ILEF	14.5	13.3	6	0.23	7.1	93.6
CHAM	10.7	5.4	4	0.07	9.5	62.4
PICA	10.8	4.6	4	0.22	9.7	60.7
HNOR	8.9	3.3	4	0.01	11	69.1
CENT	9.3	5.1	3	0.51	7.8	62.9
BNOR	7.6	1.7	4	0.38	9.8	53.4
BOUR	10.1	5.4	3	0.72	8.6	57.9
NORD	9.1	4.8	4	0.21	11.8	86.4
LORR	12.4	8	4	0.51	9.2	72.4
ALSA	12.5	8.1	3	1.25	7.4	73.2
FCOM	12	7.4	4	0.19	8.2	58.8
PAYS	6.8	1.4	3	0.58	9.6	60.1
BRET	6.8	0.7	3	0.84	9.4	55.6
POIT	6.7	1.7	3	0.48	10	50.5
AQUI	8.3	4.6	4	0.85	9.5	64.6
MIDI	8.1	4.8	3	0.54	8.5	59.3
LIMO	4.8	2.7	3	0.57	6.9	50.9
RHON	12.9	9.1	4	0.57	7.5	76.9
AUVE	7.4	4.6	2	0.85	8.3	58.2
LANG	13.2	6.5	4	1.44	11.4	70.7
PROV	19	8.2	6	1.13	10.5	89.6

Tableau 8: Voix du FN aux élections européennes de 1984

	LEPEN	ETRA	DELI	CRCH	TXCH	URBA
LEPEN	1	0.81	0.76	0.25	0.05	0.77
ETRA		1	0.62	0.08	-0.35	0.76
DELI			1	-0.14	0.19	0.75
CRCH				1	0.00	0.08
TXCH					1	0.13
URBA						1

Tableau 9: Corrélations entre les variables

	ETRA	DELI	CRCH	TXCH	URBA
LPEN	0.6910802	0.52652694	0.5494687	0.45550366	-0.2161422

Tableau 10: Corrélation partielle entre LEPEN et les 5 variables

On retire alors la variable qui a le plus faible coefficient de corrélation partielle, c'est-à-dire URBA et on réalise une régression multiple de la variable LPEN par rapport aux quatre variables explicatives restantes.
On trouve alors : $R = 0.930$, $R^2 = 0.865$ et les nouveaux coefficients de corrélation partielle indiqués dans le tableau 11. Notons que R ne change pratiquement pas : la variable

URBA n'apporte pas d'information supplémentaire par rapport aux quatre variables restantes.

	ETRA	DELI	CRCH	TXCH
LEPEN	0.73354021	0.49258622	0.5282723	0.4116398

Tableau 11: Corrélation partielle entre LEPEN et 4 variables

Pour tester la significativité de ces coefficients, on prend ici $ddl = 16$ et donc $r_{crit} = 0.4683$.
Retirons de même la variable qui a le plus faible coefficient de corrélation partielle, c'est-à-dire TXCH.
On trouve alors : $R = 0.915$, $R^2 = 0.838$ et les coefficients de corrélation partielle du tableau 12.

	ETRA	DELI	CRCH
LEPEN	0.6829056	0.6825633	0.5516189

Tableau 12: Corrélation partielle entre LEPEN et 3 variables

A ce stade, $r_{crit} = 0.4683$, et tous les coefficients sont significatifs. On peut donc dire que les votes pour l'extrême-droite aux élections européennes de 1984, à l'échelle régionale, ont varié en fonction de trois circonstances : le taux d'étrangers, le taux de délinquance, et à un degré moindre, l'évolution du chômage.
La régression n'a pas été faite dans un but de prévision, et l'équation de régression n'a qu'un intérêt limité :

$$\text{LPEN} = 0.52\ \text{ETRA} + 1.69\ \text{DELI} + 2.37\ \text{CRCH} - 0.4$$

Comparaison de deux variances

Exercice 7
Deux méthodes de dosage de l'azote ont été répétées, à partir d'un même échantillon, 25 fois avec la méthode A, 30 fois avec la méthode B. Les résultats sont rassemblés dans les tableaux ci-dessous.

Méthode A	
x_i (en g)	n_i
37	1
39	2
40	2
41	4
42	7
43	4
44	2
46	2
47	1
Total	25

Méthode B	
x_i (en g)	n_i
39	2
40	1
41	6
42	9
43	8
44	3
45	1
Total	30

1) Tester l'hypothèse: "les valeurs moyennes obtenues par les deux méthodes sont égales". (Autrement dit, les méthodes sont-elles exactes?)
2) Comparer les variances des échantillons traités avec les deux méthodes. (Autrement dit, les deux méthodes ont-elles la même précision?)
Réponses : 1) Les paramètres de statistiques descriptives sont donnés par :

	Méthode A	*Méthode B*
Moyenne	*42.08*	*42.10*
Variance	*4.95*	*1.89*
Variance corrigée	*5.16*	*1.96*

Le test de comparaison des deux moyennes (groupes indépendants) conduit à : $t_{obs} = -.04$*, évidemment non significatif aux seuils traditionnels. On ne peut donc pas refuser l'hypothèse* H_0 *d'égalité des moyennes.*
2) La statistique de test suit une loi de Fisher à $ddl_1 = 24$ *et* $ddl_2 = 29$ *degrés de liberté. On obtient :* $F_{obs} = 2.63$*. Au seuil de 1% unilatéral, on a* $F_{crit} = 2.49$*. On conclut donc à une différence des variances.*

Exercice 8

Au cours de certaines expériences, on est amené à mesurer le *temps de réaction* (TR) des sujets. C'est le temps qui s'écoule entre la présentation d'un stimulus (par exemple, une lampe qui s'allume devant le sujet) et la réaction que ce stimulus doit déclencher (par exemple, presser un bouton).
Première expérience. — Le tableau 1 fournit les TR d'une personne qui a réagi 20 fois à l'allumage d'une lampe rouge. On constate que ces 20 TR ne sont pas égaux. Ces variations d'un moment à l'autre sont imprévisibles à partir des informations dont on dispose dans l'expérience.
Deuxième expérience. — Le sujet voit maintenant s'allumer devant lui une lampe qui peut être rouge, verte ou jaune. il doit réagir si la lampe est rouge, mais ne doit pas réagir dans les deux autres cas. Le tableau 1 fournit 20 TR mesurés dans ces conditions. On observe de nouveau des variations imprévisibles d'un moment à l'autre.
Troisième expérience. — Les conditions sont les mêmes que dans la première expérience (une seule lampe) avec une seule différence : au lieu d'être rouge, la lampe donnant le signal de la réaction est verte. La troisième ligne du tableau donne les résultats. Les temps sont de nouveau différents entre eux.

Numéro d'ordre des 20 présentations	1	2	3	4	5	6	7	8	9	10
1ère expérence	20	15	18	25	17	32	18	17	19	23
2è expérience	32	40	33	37	35	29	42	62	50	39
3è expérience	16	18	19	18	15	18	17	32	23	19

Numéro d'ordre des 20 présentations	11	12	13	14	15	16	17	18	19	20
1ère expérience	19	21	15	22	17	17	21	19	17	23
2è expérience	45	47	52	37	38	39	40	41	42	39
3è expérience	23	20	18	25	15	15	17	23	17	19

La dispersion des TR est-elle la même dans chacune des trois conditions expérimentales? Pour répondre à cette question, comparer deux à deux les variances des trois séries de données.

Réponses : Les variances des trois séries sont données par :

	Variance	*Variance corrigée*
1ère expérience	*14.89*	*15.67*
2è expérience	*53.85*	*56.68*
3è expérience	*16.23*	*17,08*

Pour $ddl_1 = 19$ *et* $ddl_2 = 19$ *et un seuil de 5%, on a :* $F_{crit} = 3.00$. *Ici,* $F_{2,1,obs} = 3.61$, $F_{2,3,obs} = 3.31$, $F_{3,1,obs} = 1.09$. *Pour les expériences 1 et 3, l'hypothèse nulle (même variance) peut être retenue. En revanche, l'expérience 2 conduit à une variance différente de celles des deux autres.*

Exercice 9 *Dossier "pedago"*

Lors d'une expérience pédagogique, on s'intéresse à l'effet comparé de deux pédagogies des mathématiques chez deux groupes de 10 sujets:

- pédagogie traditionnelle (p_1)
- pédagogie moderne (p_2)

On note la performance à une épreuve de combinatoire.

p_1 traditionnelle	
s1	5.0
s2	4.0
s3	1.5
s4	6.0
s5	3.0
s6	3.5
s7	3.0
s8	2.5
s9	1.5
s10	2.5

p_2 moderne	
s11	4.0
s12	5.5
s13	4.5
s14	6.5
s15	4.5
s16	5.5
s17	1.0
s18	2.0
s19	4.5
s20	4.5

1) Vérifier que les paramètres des deux échantillons sont donnés par:

	p_1	p_2
Moyenne	3.250	4.250
Ecart-type	1.365	1.553
Variance	1.863	2.413
Ecart-type corrigé	1.439	1.637
Variance corrigée	2.069	2.681

2) Avant d'appliquer un test de comparaison de moyennes, on veut s'assurer que l'on peut supposer les variances égales dans les populations parentes. Procéder à un test de comparaison de variances permettant de s'en assurer.

Réponses : 2) On obtient $F_{obs} = 1.30$. *Or, pour* $ddl_1 = 9$, $ddl_2 = 9$ *et un seuil de 5%, on lit dans la table :* $F_{crit} = 3.18$. *L'hypothèse* H_0 *(égalité des variances) est donc retenue.*

Exercice 10

1) Pour $ddl_1 = 2, ddl_2 = 4$, la densité f de la loi de Fisher-Snedecor est donnée, pour $x \geq 0$ par:

$$f(x) = \frac{8}{(2+x)^3}$$

Construire point par point la courbe de la fonction f.

2) Pour $ddl_1 = 4, ddl_2 = 4$, la densité g de la loi de Fisher-Snedecor est donnée pour $x \geq 0$ par:

$$g(x) = \frac{6x}{(1+x)^4}$$

Construire point par point la courbe de la fonction g.