INSA – Génie Mathématique **Année 2017/2018**

GM4 – Statistique

Polycopié d'exercices

Antoine Godichon-Baggioni

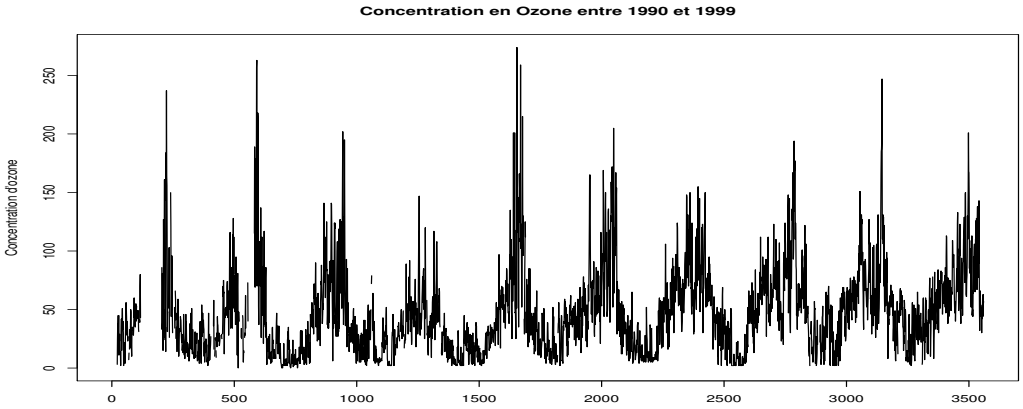
Statistique – Feuille de TD 1

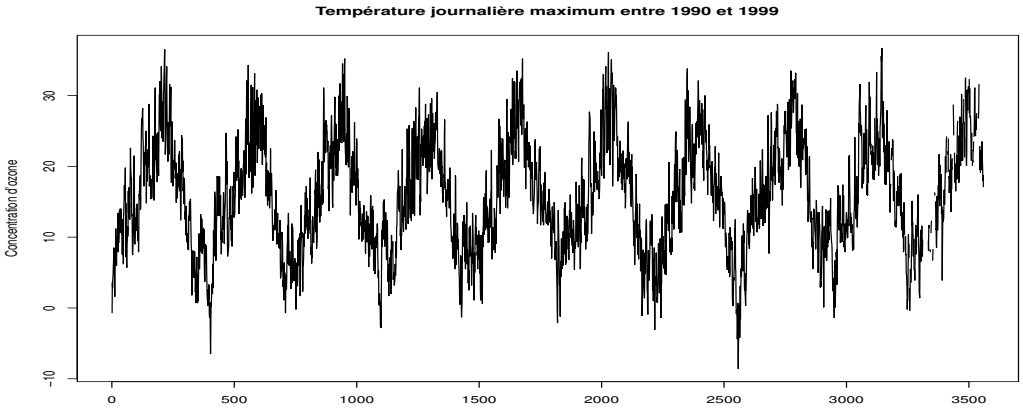
Séries Chronologiques

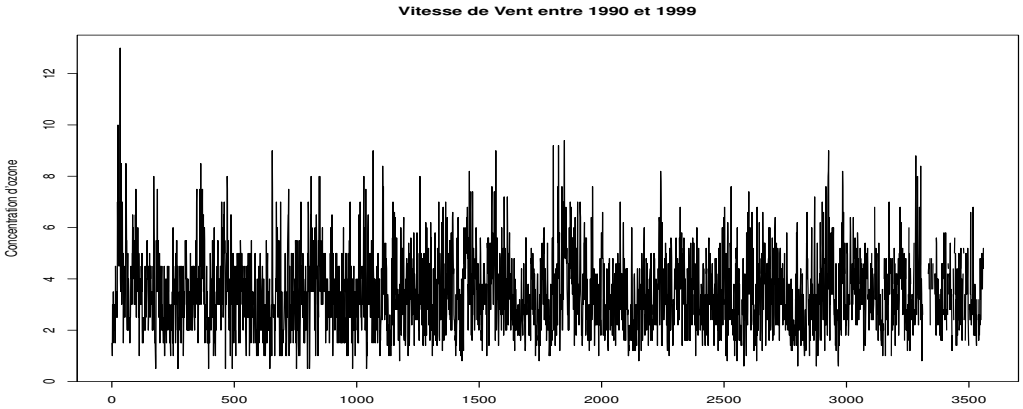
Exercice E.1 (Description d'une chronique d'ozone) On dispose de données réelles de pollution provenant d'une station de mesures d'AirParif, l'organisme chargé de la surveillance et de la prévision de la qualité de l'air en Ile de France. Ces données sont des mesures journalières sur la période 1990-1999 et concernent :

- le maximum de la température du jour (en $^{o}\mathtt{C}$)
- la vitesse moyenne de vent entre 14h et 18h (en $\mathtt{m/s}$)
- le maximum de la concentration en ozone du jour entre 14h et 18h (en $\mu\mathtt{g/m}^3$)

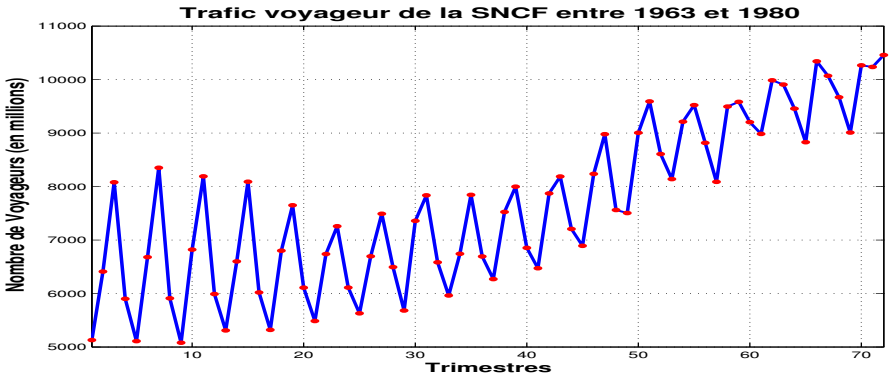
Commenter les graphiques suivants.







Exercice E.2 (Trafic SNCF) Le graphique ci-dessous présente l'évolution trimestrielle du trafic voyageur de la SNCF en deuxième classe, entre 1963 et 1980. Ces données, exprimées en millions de voyageurs kilomètres, sont extraites du livre de Gourieroux et Montfort.



Décrire la chronique.

Exercice E.3 (Coefficients de la Droite des Moindres Carrés) On considère la série statistique double $(t_i, y_i)_{1 \leq i \leq n}$. On souhaite ajuster au nuage de points associé à cette série, une droite des moindres carrés d'équation $y = a\,t\, +\, b$.

1. Montrer que les coefficients a et b de la DMC ont pour expression :

$$a \;=\; \frac{\sum_{i=1}^{n} \left(t_i - \overline{t}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{n} \left(t_i - \overline{t}\right)^2} \quad \text{et} \quad b = \overline{y} - a\,\overline{t}$$

où $\overline{t}$ et $\overline{y}$ désignent respectivement les moyennes empiriques des séries (t_i) et (y_i). On rappelle que l'on cherche a et b de telle sorte que la quantité

$$f(a,b) \;=\; \sum_{i=1}^{n}\Big(y_i - a\,t_i - b\Big)^2$$

soit minimale.

2. Démontrer la décomposition en sommes de carrés suivante :

$$\sum_{i=1}^{n}(y_i-\overline{y})^2 \;=\; \sum_{i=1}^{n}\Big(a\,t_i+b-\overline{y}\Big)^2 \;+\; \sum_{i=1}^{n}\Big(y_i-a\,t_i-b\Big)^2$$

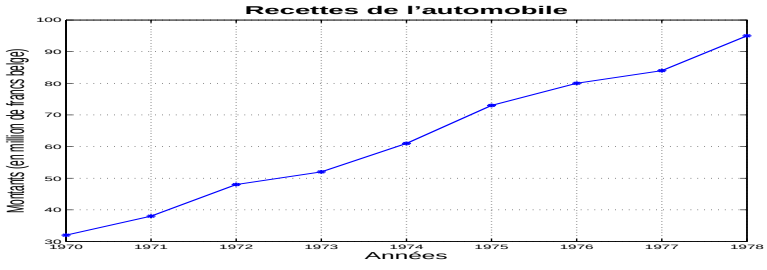
et en déduire que

$$\sum_{i=1}^{n}\Big(y_i-a\,t_i-b\Big)^2 \;=\; \sum_{i=1}^{n}(y_i-\overline{y})^2 \;-\; a^2\sum_{i=1}^{n}\Big(t_i-\overline{t}\Big)^2$$

Exercice E.4 (Recettes de l'industrie automobile) On considère la série des montants rapportés par l'industrie automobile au Trésor public belge entre 1970 et 1978 (en millions de francs belges) :

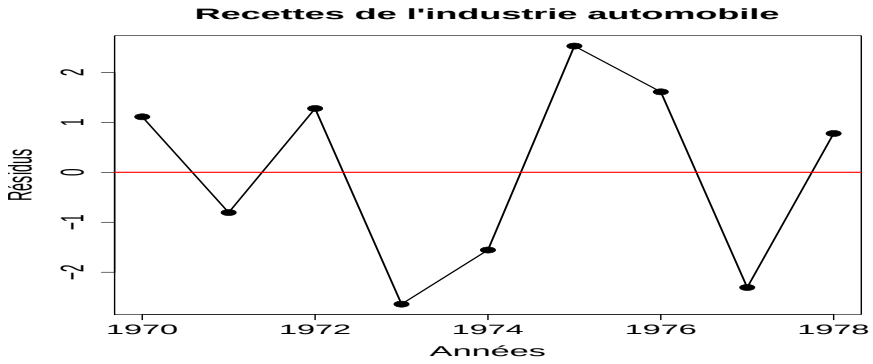
t_i	1970	1971	1972	1973	1974	1975	1976	1977	1978
y_i	32	38	48	52	61	73	80	84	95

1. Voici la représentation graphique de la série. Commenter.



2. On se propose d'ajuster à cette série une tendance linéaire de la forme $f(t) = a\,t+b$. Déterminer a et b par la méthode des moindres carrés.

3. Représenter la tendance obtenue sur le graphique précédent.
4. Calculer le coefficient de corrélation linéaire empirique que l'on notera r.
5. Le graphe des résidus est le suivant.



   Commenter ce graphique.
6. Calculer la somme des carrés des résidus, ainsi que le coefficient de détermination ou pourcentage de variance expliquée.
7. Commenter les résultats obtenus.

Pour vous faciliter les calculs, on vous donne les valeurs suivantes :

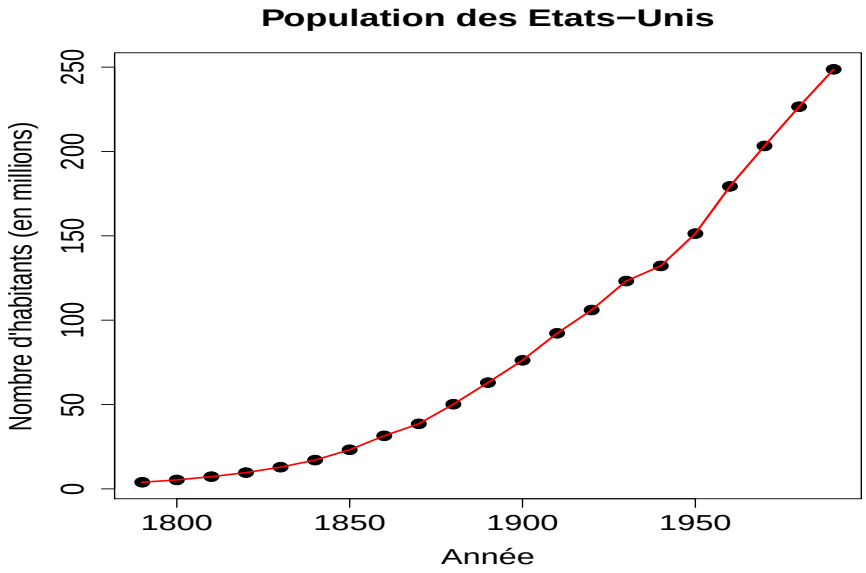
$\sum t_i$	$\sum y_i$	$\sum \left(t_i - \overline{t}\right)^2$	$\sum \left(t_i - \overline{t}\right)\left(y_i - \overline{y}\right)$	$\sum \left(y_i - \overline{y}\right)^2$
17 766	563	60	475	3 788.22

Exercice E.5 (Méthode des 2 points - Recette automobile) Reprendre les données de l'exercice 4 sur les recettes de l'automobile et déterminer l'équation de la droite d'ajustement obtenue par la méthode des deux points, en prenant les points médians des sous-séries que vous aurez choisies. Représenter cette droite sur le graphique. Calculer la somme des carrés des résidus associés. Comparer avec celle obtenue avec l'ajustement par la droite des moindres carrés. Commenter.

Exercice E.6 (Etude de la population américaine) On souhaite étudier l'évolution de la population américaine entre 1790 et 1990. On dispose des données suivantes :

Années	Nombre d'habitants (en millions)
1790	3.929214
1800	5.308483
1810	7.239881
1820	9.638453
1830	12.860702
1840	17.063353
1850	23.191876
1860	31.443321
1870	38.558371
1880	50.189209
1890	62.979766
1900	76.212168
1910	92.228496
1920	106.021537
1930	123.202624
1940	132.164569
1950	151.325798
1960	179.323175
1970	203.302031
1980	226.545805
1990	248.709873

Compte-tenu de la représentation graphique de cette série, on souhaite modéliser sa tendance à l'aide d'un polynôme de degré d, degré qu'on souhaitera le plus petit possible.



1. Décrire la chronique.

2. Déterminer les polynômes des moindres carrés de degré $d = 0, 1, 2$ et 3. On reportera dans le tableau ci-dessous les valeurs des différents coefficients où (P_d) désigne le polynôme de degré d.

Polynôme	a_0	a_1	a_2	a_3
P_0				
P_1				
P_2				
P_3				

Tableau 1 – *Estimations des différents polynômes*

3. Calculer les différentes parts de variance expliquée. On pourra synthétiser les différents résultats obtenus dans le Tableau ci-dessous

Modèle	SCR	Part de variance
(P_0)		
(P_1)		
(P_2)		
(P_3)		

Tableau 2 – *Somme des Carrés des Résidus et Part de Variance*

Quelle valeur de d suggérez-vous ?

4. Proposer alors une prévision de la population américaine en 2000. Vérifier, en cherchant sur Internet par exemple, si cette prévision est correcte.

Pour tous les calculs, on pourra utiliser les formules du cours (...!!!) combinées avec Matlab ou bien un tableur.
Si vous décidez d'utiliser Excel, ouvrir le fichier *PopUSA.xls* joint à la feuille de TD. Puis, si ce n'est pas déjà fait, activer dans l'onglet ***Outils*** à la rubrique ***Macros complémentaires***, l'***Utilitaire d'Analyse***.
Ensuite, dans l'onglet ***Outils***, lancez l'***Utilitaire d'Analyse*** et choisir ***Régression linéaire***. Il suffit alors de lancer la macro avec comme variable Y la colonne "nombre d'habitants" et comme variable X soit la colonne des "Années", soit les colonnes "Années" et "Années^2", soit les colonnes "Années", "Années^2" et "Années^3".
Il suffit alors de lire les résultats. Pour déterminer les informations utiles, commencer par faire à la main le calcul des coefficients de la droite des moindres carrés (polynôme de degré 1) ainsi que la part de variance expliquée, puis lancer la macro avec comme variable Y la colonne "nombre d'habitants" et comme variable X la colonne des "Années".

Statistique – Feuille de TD 2

Séries Chronologiques (*Suite*)

Exercice E.7 Soit $n \geq 2$ et $Y_1, Y_2, \ldots, Y_n$ des variables aléatoires indépendantes, de même loi $\mathcal{N}(\mu, \sigma)$ où σ désigne l'écart-type. On se propose d'étudier les propriétés des estimateurs respectifs des paramètres supposés inconnus μ et σ^2.

1. Rappeler l'expression de l'estimateur noté $\widehat{\mu}$ du paramètre μ.
2. Supposons le paramètre μ connu. Donner l'expression de l'estimateur du paramètre σ^2.
3. Supposons maintenant μ inconnu. Donner l'expression de l'estimateur noté S^2 du paramètre σ^2.
4. Donner, en justifiant, la loi de $\widehat{\mu}$. Quelles propriétés sur $\widehat{\mu}$ déduit-on de ce résultat ?
5. Démontrer que

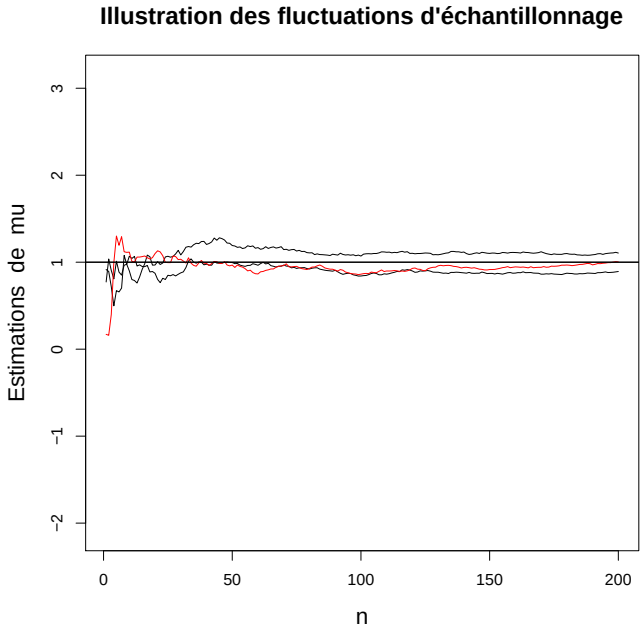
$$\sum_{i=1}^{n} (Y_i - \mu)^2 \;=\; \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 \;+\; n\left(\overline{Y} - \mu\right)^2$$

6. Donner en justifiant la loi de $\displaystyle\sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2$.
7. Déduire du résultat de la question 4, la loi de $\displaystyle n\left(\frac{\overline{Y} - \mu}{\sigma}\right)^2$.
8. Sachant que $\dfrac{\sqrt{n}(\overline{Y} - \mu)}{S}$ suit une loi de Student à $(n-1)$ degrés de liberté, notée T_{n-1}, construire un intervalle de confiance bilatéral pour le paramètre μ au niveau de confiance $1-\alpha$. On pourra noter $t_{n-1,\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la Student à $(n-1)$ `ddl`, ie.

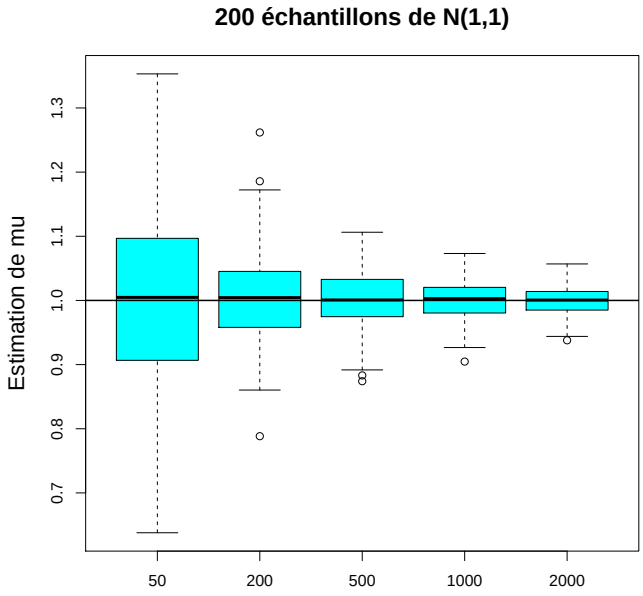
$$\mathbb{P}\left[|T_{n-1}| \leq t_{n-1,\alpha/2}\right] = 1-\alpha \iff \mathbb{P}\left[T_{n-1} \leq t_{n-1,\alpha/2}\right] = 1-\alpha/2$$

9. On décide d'étudier par simulation le comportement de l'estimateur du paramètre μ pour des tailles d'échantillon petites et modérées. Pour cela, on simule 200 échantillons de tailles $n = 50, 200, 500, 1000$ et 2000 d'une loi $\mathcal{N}(1, 1)$.

   On présente dans la figure ci-dessous quelques trajectoires (valeurs successives de l'estimation) obtenues avec des échantillons de taille $n = 200$. Commenter.



On présente dans la figure ci-dessous les boîtes à moustaches des 200 estimations pour diffféntes valeurs de n. Commenter.



Enfin, pour chacun des 200 échantillons et pour chaque valeur de n, on a construit l'intervalle de confiance à 95% du paramètre μ et on a noté l'appartenance ou non de

μ à cet intervalle. Le tableau ci-dessous présente les résultats obtenus. Commenter.

n	Nombre d'IC contenant μ (/200)	Pourcentage	Quantile $t_{n-1,0.975}$
50	192	0.96	2.0096
200	184	0.92	1.9720
500	185	0.92	1.9648
1000	186	0.93	1.9623
2000	188	0.94	1.9611

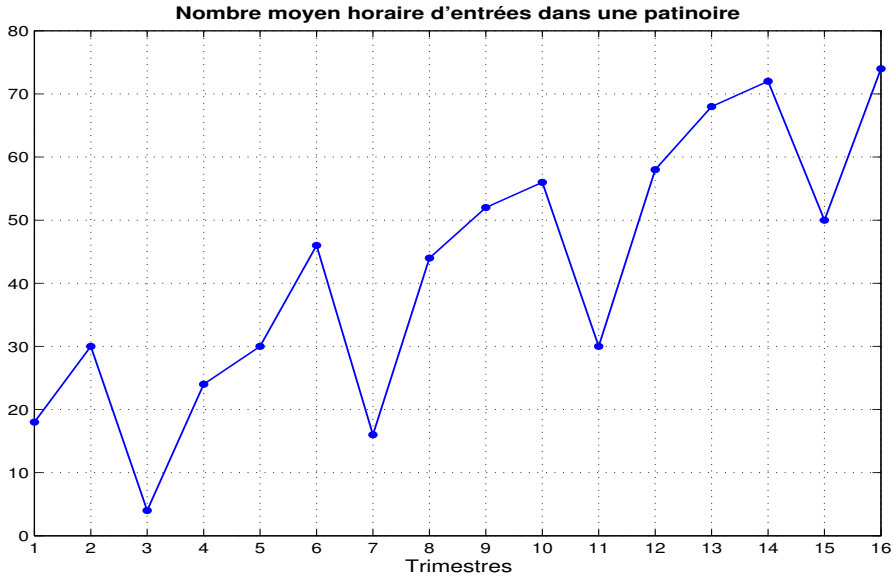
Rappel. *Pour $n \geq 1$, soient $Z_1, \ldots, Z_n$ n variables aléatoires indépendantes et de même loi $\mathcal{N}(0,1)$. Alors, $S_n = \sum_{k=1}^{n} Z_k^2$ suit une loi du khi-deux à n degrés de liberté (ddl) et on note :* $S_n \sim \chi^2(n)$.

Corollaire. *Si X suit une loi $\chi^2(n)$, alors $\mathbb{E}[X] = n$ et $\mathbb{V}\mathrm{ar}(X) = 2n$.*

Exercice E.8 (Fichier *Patinoire.xls*) On considère la série trimestrielle du nombre horaire moyen d'entrées dans une patinoire de 1993 à 1996.

Années	Trimestres			
	1	2	3	4
1993	18	30	4	24
1994	30	46	16	44
1995	52	56	30	58
1996	68	72	50	74

Voici la représentation graphique de la série. Pour simplifier, on remplacera la date du premier trimestre 1993 par 1, la date du second trimestre 1993 par 2, etc ...



1. Commenter ce graphique.

2. Calculer les séries MM(2), MMC(2), MM(3), MM(4) et MMC(4). On complètera le tableau ci-dessous.

(t_i)	(y_i)	$\left(\frac{t_i+t_{i+1}}{2}\right)$	$MM(2)$	$MMC(2)$	$MM(3)$	$MM(4)$	$MMC(4)$
1	18						
		1.5					
2	30						
		2.5					
3	4						
4	24						
5	30						
6	46						
7	16						
8	44						
9	52						
10	56						
11	30						
12	58						
13	68						
14	72						
15	50						
16	74						

3. Représenter sur le même graphique que la série initiale les séries MMC(2), MM(3) et MMC(4). Commenter.

Statistique – Feuille de TD 3

Séries Chronologiques (*Suite et Fin*) – Modèle Linéaire Gaussien simple (*Début*)

Exercice E.9 L'objet de cet exercice est de proposer une modélisation de la série des ventes trimestrielles d'un grand magasin parisien. La série $(y_i)_{1\leq i\leq 16}$ des Ventes (en Milliers d'Euros) est donnée pour la période 1995-1998 dans le tableau ci-dessous et la représentation graphique se trouve en Figure 1.

t_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
y_i	662	742	683	842	717	792	742	875	767	805	767	917	790	843	820	972

1. Justifier le choix d'un modèle additif pour modéliser cette série chronologique. Décrire les composantes en présence.

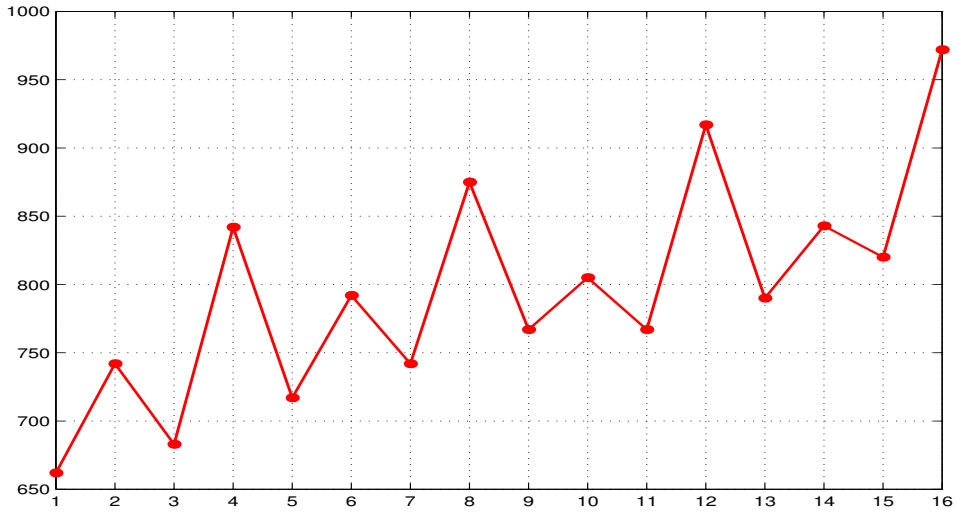


FIGURE 1 – Ventes trimestrielles (en Milliers d'Euros)

2. Dans une première étape, on se propose d'estimer la tendance par une moyenne mobile centrée.
   (a) Quel ordre p doit-on utiliser ?
   (b) Quel est l'intérêt d'utiliser une moyenne mobile centrée ?
   (c) Compléter le tableau suivant :

t_i	1	2	3	4	5	6	7	8
MMC(p)				752.25	765.875	777.375	787.75	795.625
t_i	9	10	11	12	13	14	15	16
MMC(p)		808.75	816.875	824.5	835.875			

3. L'estimation des coefficients saisonniers a donné $\widehat{s}_1 = -41.6$, $\widehat{s}_2 = 2.6$ et $\widehat{s}_3 = -49.4$. En utilisant une propriété des coefficients saisonniers, déduire la valeur de $\widehat{s}_4$.
4. Calculer la série corrigée des variations saisonnières (CVS_i) aux instants $t_i = 13, 14, 15, 16$.
5. On décide d'améliorer l'estimation de la tendance en ajustant une droite à la série corrigée des variations saisonnières. Donner l'équation de cette droite obtenue par la méthode des moindres carrés. On donne les valeurs numériques suivantes :

   $\overline{t} = 8.5, \quad \overline{\mathrm{CVS}} = 796, \quad \mathrm{Cov}(t\,;\,\mathrm{CVS}) = 217.54, \quad \mathrm{Var}(t) = 21.25, \quad \mathrm{Var}(\mathrm{CVS}) = 2\,320.63$

6. Le graphe des résidus est donné en Figure 2. Expliquer comment est obtenue la série des résidus $(\widehat{e}_i)_{1\leq i\leq 16}$ et commenter le graphique.

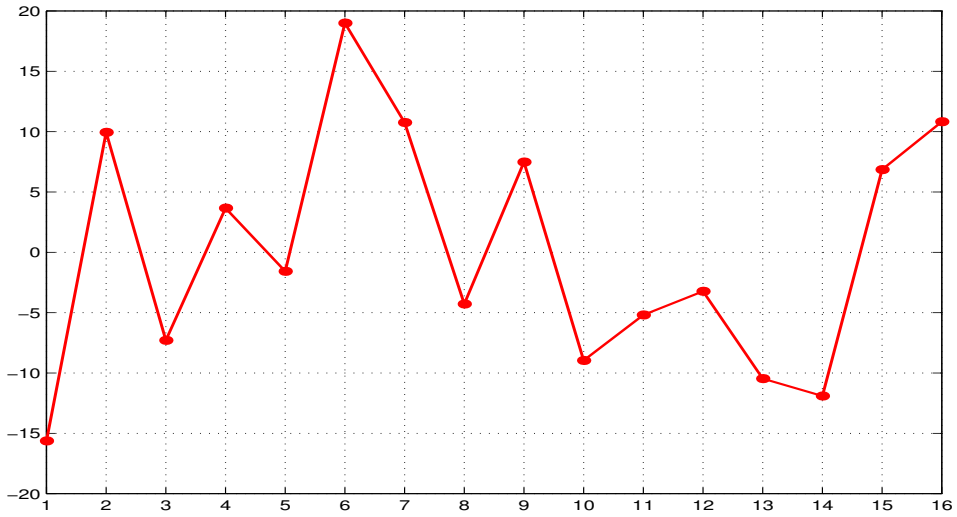


FIGURE 2 – Résidus (en Milliers d'Euros)

7. Effectuer une prévision des 4 trimestres de l'année 1999. Serait-il réaliste d'utiliser ce modèle pour prévoir les valeurs des trimestres de l'année 2005 ?

Exercice E.10 Le tableau ci-dessous donne les résultats du traitement d'une série trimestrielle notée $(y_i)_{1\leq i\leq 20}$. On propose d'utiliser un modèle multiplicatif. La tendance est estimée par une droite d'équation $y = 20.5\,t + 261$.

t_i	y_i	$\widehat{f_i}$	$1+\widehat{s_i}$	CVS_i
1	170	281,5	0,66	
2	300	302		300
3	610	322,5	1,91	319
4	120	343	0,43	279
5	250	363,5	0,66	379
6	410	384		410
7	790	404,5		414
8	190	425	0,43	442
9	290	445,5	0,66	
10	460	466		460
11	890	486,5	1,91	466
12	250	507	0,43	581
13	450	527,5		682
14	550	548		
15	1100	568,5	1,91	576
16	270	589		628
17	320	609,5	0,66	485
18	600	630		600
19	1260	650,5	1,91	
20	280	671	0,43	651

où

- $(\widehat{f_i})$ désigne la tendance estimée par la droite d'équation : $y = 20.5t + 261$;
- $(\widehat{s_i})$ désigne la suite des coefficients saisonniers centrés estimés ;
- (CVS_i) désigne la série corrigée des variations saisonnières.

1. Rappeler l'équation du modèle multiplicatif en précisant les hypothèses sur les différentes composantes.
2. Calculer la valeur du coefficient saisonnier $\widehat{s_2}$.
3. Rappeler la définition de la série corrigée des variations saisonnières (CVS_i) et calculer la valeur CVS_1.
4. Compléter les valeurs manquantes du tableau.
5. Calculer la valeur $\widehat{y_{20}}$ prédite par le modèle.

Exercice E.11 *Préliminaires pour la démonstration du Théorème 1 du cours sur le MLG.*

On reprend les notations du cours sur le MLG simple. On dispose de n observations bidimensionnelles $(x_i, y_i)_{1\leq i\leq n}$, où les données $(x_i)_{1\leq i\leq n}$ ne sont pas des réalisations de

variables aléatoires et les données $(y_i)_{1\leq i\leq n}$ sont les réalisations de n variables aléatoires $(Y_i)_{1\leq i\leq n}$ liées aux données $(x_i)_{1\leq i\leq n}$ par la relation :

$$\forall\, i \in \Big\{1, 2, \cdots, n\Big\},\quad Y_i \;=\; \alpha\, x_i \;+\; \beta \;+\; \varepsilon_i \tag{1}$$

où $\alpha, \beta \in \mathbb{R}$ et $\varepsilon_1, \cdots, \varepsilon_n$ i.i.d. de loi $\mathcal{N}(0\,, \sigma)$ où σ désigne l'écart-type.

1. Montrer que $\overline{Y} \;\sim\; \mathcal{N}\left(\alpha\overline{x} + \beta\,;\, \dfrac{\sigma}{\sqrt{n}}\right)$.
2. Montrer que $\displaystyle\sum_{i=1}^{n} (x_i - \overline{x}) \;=\; 0$, ie. la somme des termes d'une série centrée vaut 0.
3. Montrer que A peut encore s'écrire sous la forme

$$A \;=\; \frac{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})\, Y_i}{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

4. Etablir la loi de A.
5. Montrer que A et $\overline{Y}$ sont non corrélés, c'est à dire que $\mathrm{Cov}(A\,,\, \overline{Y}) = 0$.
6. Ecrire sous forme vectorielle le modèle (1).

Rappels. Dans le cas gaussien, la non corrélation est équivalente à l'indépendance.

Rappels sur la Covariance

Définition 1. *Soit (X, Y) un couple de variables aléatoires. On appelle* **covariance** *de (X, Y), notée $\mathrm{Cov}(X, Y)$, le nombre réel, s'il existe, donné par*

$$\mathrm{Cov}(X, Y) \;=\; \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(Y - \mathbb{E}\left[Y\right])\right]$$

Remarque. On peut noter que cette définition est parfaitement symétrique par rapport aux deux coordonnées X et Y. On a ainsi $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$. De plus, on peut généraliser le théorème de Kœnig :

$$\mathrm{Cov}(X, Y) \;=\; \mathbb{E}\left[XY\right] \;-\; \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$$

Par conséquent, si X et Y sont deux variables aléatoires indépendantes alors $\mathrm{Cov}(X, Y) = 0$. Notons que la réciproque est généralement fausse
Enfin, on a bien évidemment $\mathrm{Cov}(X, X) = \mathbb{V}\mathrm{ar}(X)$.

Propriété. L'opérateur covariance étant linéaire par rapport à chacune de ses coordonnées, si X_1, X_2, Y_1, Y_2 sont quatre variables aléatoires et $\alpha_1, \alpha_2, \beta_1, \beta_2$ quatre nombres réels alors

$$\begin{aligned}\mathrm{Cov}(\alpha_1 X_1 + \alpha_2 X_2, \beta_1 Y_1 + \beta_2 Y_2) &= \alpha_1\beta_1\mathrm{Cov}(X_1, Y_1) + \alpha_1\beta_2\mathrm{Cov}(X_1, Y_2) \\ &\quad + \alpha_2\beta_1\mathrm{Cov}(X_2, Y_1) + \alpha_2\beta_2\mathrm{Cov}(X_2, Y_2)\end{aligned}$$

Corollaire. *Soit (X, Y) un couple de variables aléatoires et soient α, β deux nombres réels. Alors,*

$$\mathbb{V}\mathrm{ar}(\alpha X + \beta Y) = \alpha^2\mathbb{V}\mathrm{ar}(X) + \beta^2\mathbb{V}\mathrm{ar}(Y) + 2\alpha\beta\,\mathrm{Cov}(X, Y)$$

Statistique – Feuille de TD 4

Modèle Linéaire Gaussien Simple (*Suite*)

Exercice E.12 *Démonstration du Théorème 1 du cours sur le MLG simple.*

En suivant les différentes étapes décrites dans la preuve du Théorème 1 (*cf. cours*), rédiger la démonstration de ce Théorème.

Exercice E.13 *Démonstration du Théorème 2 du cours sur le MLG simple.*

Nous allons démontrer dans cet exercice quelques éléments de la preuve du Théorème 2,

1. Etablir que
$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2 \;=\; \frac{1}{\sigma^2}\sum_{i=1}^{n}\Big(Y_i - Ax_i - B\Big)^2 \;+\; \frac{1}{\sigma^2}\sum_{i=1}^{n}\Big((A-\alpha)x_i + (B-\beta)\Big)^2$$
2. Montrer que $\displaystyle\frac{1}{\sigma^2}\sum_{j=1}^{n}\varepsilon_j^2 \sim \chi^2_n$.
3. Montrer que
$$\sum_{i=1}^{n}\Big((A-\alpha)x_i + (B-\beta)\Big)^2 \;=\; (A-\alpha)^2\sum_{j=1}^{n}(x_i-\overline{x})^2 \;+\; n\left(\overline{Y} - \alpha\overline{x} - \beta\right)^2$$
4. En déduire, en utilisant un résultat sur le chi-deux (voir ci-après), que
$$\frac{1}{\sigma^2}\sum_{j=1}^{n}\Big((A-\alpha)x_j + (B-\beta)\Big)^2 \sim \chi^2_2$$

Quelques rappels sur le khi-deux

Théorème. *Pour $n \geq 1$, soient $Z_1, \ldots, Z_n$ n variables aléatoires indépendantes et de même loi $\mathcal{N}(0,1)$. Alors, $S_n = \displaystyle\sum_{k=1}^{n} Z_k^2$ suit une loi du khi-deux à n degrés de liberté (ddl) et on note : $S_n \;\sim\; \chi^2(n)$.*

Corollaire. *Si X suit une loi $\chi^2(n)$, alors $\mathbb{E}[X] = n$ et $\mathbb{V}\mathrm{ar}(X) = 2n$.*

Corollaire. *Soit X et Y deux variables aléatoires indépendantes de loi respectives $\chi^2(n)$ et $\chi^2(m)$. Alors,*

$$Z = X + Y \sim \chi^2(n+m)$$

Conséquence. *Soit X, Y et Z trois variables aléatoires positives telles que $Z = X + Y$. Si $Z \sim \chi^2(n)$ et $X \sim \chi^2(p)$, alors $Y \sim \chi^2(n-p)$ et on a l'indépendance entre X et Y.*

Construction d'une Student et d'une Fisher

Soient U et V deux variables aléatoires indépendantes.

- *si $U \sim \mathcal{N}(0\,,\,1)$ et $V \sim \chi^2_n$ alors $\dfrac{U}{\sqrt{V\ /\ n}} \sim T_n$ (Student à n degrés de liberté)*
- *si $U \sim \chi^2_p$ et $V \sim \chi^2_q$ alors $\left(\dfrac{U\ /\ p}{V\ /\ q}\right) \sim F(p\,;\,q)$ (Fisher à p et q degrés de liberté)*

Exercice E.14 Démontrer que

$$\frac{a\sqrt{n\,d_x^2}}{s} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

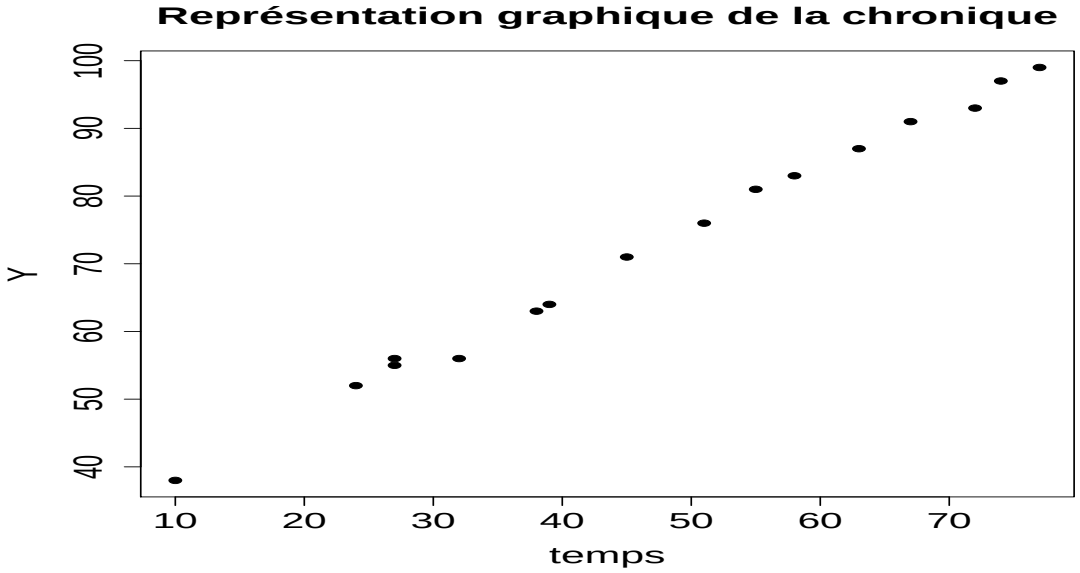
On se rappellera que

$$\sum_{i=1}^{n}\Big(y_i - \overline{y}\Big)^2 = \sum_{i=1}^{n}\Big(a\,x_i + b - \overline{y}\Big)^2 + \sum_{i=1}^{n}\Big(y_i - a\,x_i - b\Big)^2$$

Exercice E.15 (Fichier *carbone.xls*) On présume, que dans un acier donné, la résistance à la traction, notée R, est liée linéairement et de manière significative à la teneur en carbone, notée C. Pour confirmer cette hypothèse, on fait une série de mesures sur des échantillons d'acier. La teneur en carbone est exprimée en $^o/_{ooo}$ et la résistance à la traction de l'acier en kg/mm^2. Les mesures obtenues ont été consignées dans le tableau suivant.

	Echantillons d'acier															
C	72	55	63	38	10	45	77	67	58	74	51	39	27	27	24	32
R	93	81	87	63	38	71	99	91	83	97	76	64	55	56	52	56

1. Justifier la modélisation des données par un modèle linéaire gaussien (pour construire votre argumentation, vous pouvez utiliser le graphique donné ci-dessous). Introduire alors ce modèle. On notera respectivement $(x_i)_{1\leq i\leq 16}$ les différentes teneurs en carbone et $(y_i)_{1\leq i\leq 16}$ les différentes valeurs de la résistance à la traction de l'acier.



2. Donner des intervalles de confiance pour tous les paramètres du modèle, au niveau de confiance 95%.
3. Tester au risque 5% le caractère significatif de la liaison linéaire entre la teneur en carbone d'un acier et sa résistance à la traction.
4. Construire une prévision de la résistance à la traction d'un acier dont la teneur en carbone serait de 70 $‰$. Donner l'intervalle de prévision associé, au niveau de confiance 95%.

Pour les différentes applications numériques, on vous fournit les informations suivantes :

$\sum x_i$	$\sum y_i$	$\sum (x_i - \overline{x})^2$	$\sum (x_i - \overline{x})(y_i - \overline{y})$	$\sum (y_i - \overline{y})^2$
759	1162	6 179.9375	5 599.625	5 095.75

Statistique – Feuille de TD 5

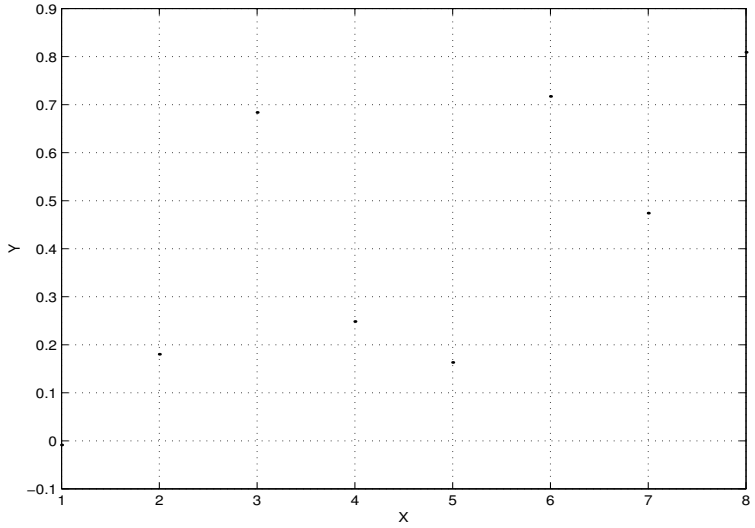
Modèle Linéaire Gaussien Simple (*Suite et Fin*)

Exercice E.16 (Fichier *experim.xls*) Un expérimentateur étudie l'effet d'un caractère X sur un caractère Y. Pour cela il effectue 8 mesures et consigne les résultats obtenus dans le tableau suivant :

X	1	2	3	4	5	6	7	8
Y	-0.0087	0.1802	0.6838	0.2484	0.1632	0.7173	0.4741	0.8090

On note respectivement $(x_i)_{1\leq i\leq 8}$ et $(y_i)_{1\leq i\leq 8}$ les différentes valeurs de X et de Y.

1. Justifier la modélisation des données par un modèle linéaire gaussien (pour construire votre argumentation, vous pouvez utiliser le graphique donné ci-dessous). Introduire alors ce modèle.



2. Tester, au risque 5%, le caractère significatif de l'influence du caractère X sur le caractère Y.

On fournit les informations suivantes.

$\sum x_i$	$\sum y_i$	$\sum (x_i - \overline{x})^2$	$\sum (x_i - \overline{x})(y_i - \overline{y})$	$\sum (y_i - \overline{y})^2$
36	3.2672	42	3.6044	0.6479

Exercice E.17 Traiter, à l'aide du module *Régression Linéaire* du logiciel *Excel*, les questions 2 de l'exercice E.1 de la feuille de TD 5 et de l'exercice E.4 de la feuille de TD 4. On sortira notamment les tables d'analyse de variance.

Exercice E.18 *(DS1 2006-2007)* On dispose de n ($n \geq 2$) observations bidimensionnelles $(x_1, y_1), \ldots, (x_n, y_n)$ qui sont telles que

$$\sum_{i=1}^{n} x_i \;=\; 0 \quad \text{et} \quad \sum_{i=1}^{n} y_i \;=\; 0$$

Les données $x_1, \ldots, x_n$ ne sont pas des réalisations de variables aléatoires, mais sont imposées par la nature des choses. En revanche, les données $y_1, \ldots, y_n$ sont supposées être des réalisations de n variables aléatoires $Y_1, \ldots, Y_n$, qui sont liées aux données $x_1, \ldots, x_n$ de la manière suivante :

$$\forall\; i \in \Big\{1, 2, \ldots, n\Big\}\;,\quad Y_i \;=\; \alpha\, x_i \;+\; \varepsilon_i$$

où $\alpha \in \mathbb{R}$ et où $\varepsilon_1, \ldots, \varepsilon_n$ sont n variables aléatoires indépendantes de même loi $\mathcal{N}(0\,,\,\sigma)$; α et σ sont les paramètres inconnus du modèle. Notons que par construction, les variables aléatoires $Y_1, \ldots, Y_n$ sont indépendantes.

Remarque. Le paramètre β (du MLG simple) n'apparaît pas dans cette modélisation car les données sont centrées.

1. Montrer que l'estimateur des moindres carrés du paramètre α est A défini par

$$A \;=\; \Big(\sum_{i=1}^{n} x_i\, Y_i\Big) \Big/ \Big(\sum_{i=1}^{n} x_i^2\Big)$$

   On rappelle que cet estimateur rend minimale la quantité

$$f(A) \;=\; \sum_{i=1}^{n} \Big(Y_i \;-\; A\, x_i\Big)^2$$

2. Donner la loi de A. Quelles propriétés tire-t-on de ce résultat ?
3. Après avoir montré que

$$\sum_{i=1}^{n} (Y_i - A\, x_i)^2 \;=\; \sum_{i=1}^{n} (Y_i - \alpha\, x_i)^2 \;-\; \sum_{i=1}^{n} (A\, x_i - \alpha\, x_i)^2$$

   et déduit ensuite que

$$\sum_{i=1}^{n} (Y_i - A\, x_i)^2 \;=\; \sum_{i=1}^{n} \varepsilon_i^2 \;-\; (A - \alpha)^2 \sum_{i=1}^{n} x_i^2$$

montrer que $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(Y_i - A\,x_i)^2$ est un estimateur sans biais de σ^2.

4. Sachant que $\dfrac{(n-1)\,S^2}{\sigma^2} \sim \chi^2_{n-1}$ et que A et S^2 sont indépendants, donner la loi des variables aléatoires $\dfrac{(A-\alpha)\sqrt{\sum_{i=1}^{n} x_i^2}}{S}$ et $\dfrac{(A-\alpha)^2\,\sum_{i=1}^{n} x_i^2}{S^2}$.
5. Construire un intervalle de confiance au niveau de confiance $(1-\delta)$ pour le paramètre α.
6. Soit α_0 un réel donné. En utilisant un des résultats de la question 4, proposer un test au risque δ de l'hypothèse nulle $H_0 : \ll\alpha = \alpha_0\gg$ contre l'hypothèse alternative $H_1 : \ll\alpha \neq \alpha_0\gg$.
7. On s'intéresse à une donnée x_0 pour laquelle on n'a pas observé de y_0. Préciser les hypothèses à faire sur la variable aléatoire Y_0 dont y_0 serait une réalisation.
8. Proposer une prévision de y_0. On la notera $\widehat{y}_0$.
   De quelle variable aléatoire, notée $\widehat{Y}_0$, $\widehat{y}_0$ est-elle une réalisation ?
   Donner, en justifiant, la loi de $\widehat{Y}_0$.
9. Construire un intervalle de confiance, au niveau de confiance $(1-\delta)$, du paramètre $\mathbb{E}(Y_0)$.

Statistique – Feuille de TD 6

Modèle Linéaire Gaussien Simple & ANOVA 1

Exercice E.19 (ANOVA 1) Soit $p \geq 2$ un entier et soient $n_1, n_2, \cdots, n_p$ des entiers supérieurs à 2. Soient $\Big((Y_{ij})_{1\leq j\leq n_i}\Big)_{1\leq i\leq p}$ des variables aléatoires globalement indépendantes, avec pour tout $1 \leq i \leq p$ et tout $1 \leq j \leq n_i$,

$$Y_{ij} \ \sim \ \mathcal{N}\left(\mu_i \, , \, \sigma^2\right).$$

Les paramètres $(\mu_i)_{1\leq i\leq p}$ et σ^2 sont inconnus. On introduit les notations suivantes (notations qui seront utilisées dans le cours d'ANOVA 1) :

$$\begin{aligned} SCM &= \sum_{i=1}^{p} n_i \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \\ SCR(M_p) &= \sum_{i=1}^{p}\sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \\ SCR(M_1) &= \sum_{i=1}^{p}\sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 \end{aligned}$$

où $\overline{Y}_{i\bullet} = \dfrac{1}{n_i}\displaystyle\sum_{j=1}^{n_i} Y_{ij}$ et $\overline{Y}_{\bullet\bullet} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{p} n_i\, \overline{Y}_{i\bullet}$ avec $n = \displaystyle\sum_{i=1}^{p} n_i$.

1. Démontrer la décomposition en somme de carrés suivante :

$$SCR(M_1) \ = \ SCM \ + \ SCR(M_p)$$

2. Pour tout $1 \leq i \leq p$, rappeler la loi de $\dfrac{1}{\sigma^2}\displaystyle\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$. Donner alors, en justifiant, la loi de $SCR(M_p)/\sigma^2$ et en déduire un estimateur sans biais de σ^2.
3. Dans cette question on suppose que les (μ_i) sont tous égaux et égaux à μ. Donner alors la loi de $SCR(M_1)/\sigma^2$ et en déduire celle de SCM/σ^2.
4. Rappeler la loi de $\overline{Y}_{i\bullet}$ et donner celle de $\overline{Y}_{\bullet\bullet}$. En déduire alors $\mathbb{E}\left[\left(\overline{Y}_{i\bullet} - \mu_i\right)^2\right]$ et $\mathbb{E}\left[\left(\overline{Y}_{\bullet\bullet} - \mu\right)^2\right]$ où l'on a noté $\mu = \dfrac{1}{n}\displaystyle\sum_{i=1}^{p} n_i\, \mu_i$.

   Dans ce qui suit, on se propose de montrer que

$$\mathbb{E}\left[\frac{SCM}{p-1}\right] \ = \ \sigma^2 \ + \ \frac{1}{p-1}\sum_{i=1}^{p} n_i \ (\mu_i - \mu)^2 \qquad (2)$$

5. Montrer que l'on a

$$SCM = \sum_{i=1}^{p} n_i \left(\overline{Y}_{i\bullet} - \mu_i\right)^2 + \sum_{i=1}^{p} n_i \left(\mu_i - \mu\right)^2 + \sum_{i=1}^{p} n_i \left(\overline{Y}_{\bullet\bullet} - \mu\right)^2 + 2\sum_{i=1}^{p} n_i \left(\overline{Y}_{i\bullet} - \mu_i\right)\left(\mu_i - \mu\right) - 2\sum_{i=1}^{p} n_i \left(\overline{Y}_{i\bullet} - \mu_i\right)\left(\overline{Y}_{\bullet\bullet} - \mu\right)$$

6. En déduire, en utilisant notamment les résultats de la question 3, que

$$\mathbb{E}\left[SCM\right] = (p+1)\sigma^2 + \sum_{i=1}^{p} n_i \left(\mu_i - \mu\right)^2 - 2\sum_{i=1}^{p} n_i \,\mathbb{E}\left[\left(\overline{Y}_{i\bullet} - \mu_i\right)\left(\overline{Y}_{\bullet\bullet} - \mu\right)\right]$$

7. Après avoir montré que $\overline{Y}_{\bullet\bullet} - \mu$ pouvait s'écrire sous la forme $\dfrac{1}{n}\displaystyle\sum_{\ell=1}^{p} n_\ell\Big(\overline{Y}_{\ell\bullet} - \mu_\ell\Big)$, montrer que pour tout $1 \leq i \leq p$,

$$\mathbb{E}\left[\left(\overline{Y}_{i\bullet} - \mu_i\right)\left(\overline{Y}_{\bullet\bullet} - \mu\right)\right] = \frac{n_i}{n}\,\mathbb{E}\left[\left(\overline{Y}_{i\bullet} - \mu_i\right)^2\right] = \frac{\sigma^2}{n}.$$

   En déduire alors l'égalité (2).
8. A quelle condition $SCM/(p-1)$ est-il un estimateur sans biais de σ^2 ?

Exercice E.20 (DS1 2007-2008) On considère 3 portées de porcelets dont les poids à la naissance sont indiqués dans le tableau ci-dessous. On note y_{ij} le poids du $j^{\text{ème}}$ porcelet de la portée i.

Portée	1	2	3
y_{ij}	3.2 3.3 3.2 2.9 3.3 2.5 2.6 2.8	2.6 2.6 2.9 2.0 2.0 2.1	3.1 2.9 3.1 2.5
n_i	8	6	4
$\overline{y}_{i\bullet}$	2.975	2.3667	2.9
$\left(\overline{y}_{i\bullet} - \overline{y}\right)^2$	0.0482	0.1512	0.0209

On donne de plus $\sum_{i=1}^{3}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}\right)^2 = 3.0644$.

On supposera que pour tout $i = 1, 2, 3$ et tout $j = 1, \cdots, n_i$, la donnée y_{ij} est une réalisation d'une variable aléatoire Y_{ij} satisfaisant

$$Y_{ij} \;=\; \mu_i \;+\; \varepsilon_{ij}$$

où $\mu_i \geq 0$ et (ε_{ij}) est une suite de variables aléatoires gaussiennes, centrées, de même variance σ^2 et indépendantes.

1. Donner des estimations des paramètres inconnus mis en jeu dans le modèle.
2. Compléter le tableau d'analyse de variance suivant

Source des variations	Somme des carrés	Degrés de liberté	Moyenne des carrés	Statistique F	p-value
Facteur *Portée*	.....	.....	.....	.....	
Résiduelle	.....	.....	.....		
Totale	.....	.....			

3. Dites au risque 5% et 1%, si le poids moyen des porcelets peut être considéré comme le même d'une portée à l'autre.
4. Les portées 1 et 3 ont le même père : y-a-t'il, au risque 5% une différence entre les deux portées ?

Statistique – Feuille de TD 7

Analyse de Variance à un facteur & MLG Multiple

Exercice E.21 Dans un lycée, plusieurs élèves mesurent la glycémie sanguine de 5 d'entre eux. Il est raisonnable d'admettre que les erreurs de mesures sont indépendantes et qu'elles ont un écart-type σ identique, ceci pour toutes les mesures effectuées.
Le tableau ci-dessous rassemble les mesures obtenues

	Elève 1	Elève 2	Elève 3	Elève 4	Elève 5
y_{ij}	3.50 4.27 4.57 4.37 4.47 3.43	4.42 4.37 3.87 5.08 4.73 4.48 5.12 4.18	6.13 6.08 5.22 5.28 5.52 5.58 5.67	5.83 5.38 5.12 4.73 5.98 5.57 5.32	5.82 5.92 6.30 6.13 5.20 5.70 4.98 4.52 6.05
$\overline{y}_{i\bullet}$	4.1017	4.5313	5.6400	5.4186	5.6244
n_i	6	8	7	7	9
$\left(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}\right)^2$	1.0069	0.3293	0.2861	0.0982	0.2697

où y_{ij} désigne la $j^{\text{ème}}$ mesure de l'élève i. On donne de plus, $\sum_{i=1}^{5}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{\bullet\bullet})^2 = 21.0313$. On suppose que pour tout $i = 1, \ldots, 5$ et tout $j = 1, \ldots, n_i$, la donnée y_{ij} est une réalisation d'une variable aléatoire Y_{ij} satisfaisant

$$Y_{ij} \;=\; \mu_i \;+\; \varepsilon_{ij}$$

où $\mu_i \geq 0$ et (ε_{ij}) est une suite de variables aléatoires indépendantes de même loi $\mathcal{N}(0, \sigma^2)$.

1. Donner des estimations des paramètres inconnus du modèle.
2. Compléter le tableau d'ANOVA 1 suivant

Source des variations	Somme des carrés	Degrés de liberté	Moyenne des carrés	Statistique F	p-value
Facteur *Eleve*	.....	.....	.....	.....	4,45 E-07
Résiduelle	.....	.....	.....		
Totale	.....	.....			

3. Tester, au risque 5%, puis 1%, l'existence d'un effet "élève" sur le taux de glycémie moyen.
4. Il se trouve que les élèves 1 et 2 ont le même père : y-a-t'il, au risque 5% une différence du taux de glycémie entre les deux élèves ?

Exercice E.22 On se place dans $\mathbb{R}^n$ muni de la norme euclidienne. Soient $\mathcal{V}$ et $\mathcal{W}$ deux sous-espaces vectoriels de $\mathbb{R}^n$ tels que $\mathcal{V} \subset \mathcal{W}$. On note respectivement $P_{\mathcal{V}}$ et $P_{\mathcal{W}}$ les projections orthogonales sur $\mathcal{V}$ et $\mathcal{W}$. Soit $Y \in \mathbb{R}^n$. Démontrer que

$$\|Y - P_{\mathcal{V}}(Y)\|^2 \;=\; \|Y - P_{\mathcal{W}}(Y)\|^2 \;+\; \|P_{\mathcal{W}}(Y) - P_{\mathcal{V}}(Y)\|^2$$

Exercice E.23 On reprend les notations du cours. En utilisant l'exercice précédent, démontrer que

$$\left\|Y - \mathbf{1}_n\overline{Y}\right\|^2 \;=\; \|XT - \mathbf{1}_n\overline{Y}\|^2 \;+\; \|Y - XT\|^2$$

et

$$\|Y - X^{(q)}T^{(q)}\|^2 \;=\; \|XT - X^{(q)}T^{(q)}\|^2 \;+\; \|Y - XT\|^2$$

Exercice E.24 Rédiger la démonstration du théorème 1 du cours sur le MLG multiple.

Statistique – Feuille de TD 8

Modèle Linéaire Gaussien Multiple

Exercice E.25 (Preuve du Théorème 1) On reprend les notations du cours. On considère le modèle linéaire Gaussien multiple défini par

$$Y \;=\; X\theta \;+\; \varepsilon$$

où ε est un vecteur gaussien de $\mathbb{R}^n$ de loi $\mathcal{N}(0\,,\,\sigma^2 I_n)$ et X est supposée de rang $(p+1)$.

Démontrer le théorème suivant.

Théorème 1 Sous les hypothèses du modèle linéaire Gaussien multiple, on a

- $T = (X'X)^{-1}X'Y \;\sim\; \mathcal{N}(\theta\,,\,\sigma^2(X'X)^{-1})$
- $\dfrac{(n-p-1)S^2}{\sigma^2} = \dfrac{\|Y-XT\|^2}{\sigma^2} \;\sim\; \chi^2_{(n-p-1)}$
- T et S^2 sont indépendants.

Exercice E.26 (DS Septembre 2007)

On souhaite étudier la variation du taux d'hémoglobine dans le sang au cours d'une opération chirurgicale en fonction de la durée de l'opération et du volume de sang perdu pendant l'opération. On dispose des résultats suivants où y_i représente la valeur observée en pourcentage de la variation du taux d'hémoglobine, x_i est la durée de l'opération en heures décimales et x'_i est le volume en litres de sang perdu.

y_i	-1.70	-4.61	-5.82	-1.17	-4.23	-3.31	+0.42	-2.98
x_i	1.75	1.33	1.43	1.86	1.81	1.66	1.60	2.00
x'_i	0.52	0.59	0.61	0.50	0.54	0.49	0.27	0.47

Dans ce qui suit, on supposera que pour tout i, y_i est une réalisation d'une variable aléatoire Y_i de loi $\mathcal{N}(\beta + \alpha x_i + \alpha' x'_i\,,\sigma^2)$. Pour la suite, on notera $(M2)$ ce modèle, $(M1)$ le modèle dans lequel n'intervient que la durée de l'opération et $(M1')$ le modèle dans lequel n'intervient que le volume de sang perdu.

1. Calculer les estimations b, a, a' et s^2 des paramètres inconnus β, α, α' et σ^2.
2. Tester, au risque 5%, l'hypothèse selon laquelle la variation du taux d'hémoglobine ne dépend ni de la durée de l'opération ni du volume de sang perdu.

3. Dans le modèle $(M2)$, tester, au risque 5%, l'hypothèse selon laquelle la variation du taux d'hémoglobine ne dépend pas de la durée de l'opération.
4. Construire une prévision de la variation du taux d'hémoglobine dans le sang d'un patient qui subit une opération d'une durée $x_0 = 1.25$ et dont le volume en litres de sang perdu est $x'_0 = 0.52$. Donner un intervalle de prévision au niveau 95%.

On fournit les informations suivantes :

$$\begin{pmatrix} n & \sum x_i & \sum x'_i \\ \sum x_i & \sum x_i^2 & \sum x_i\, x'_i \\ \sum x'_i & \sum x_i\, x'_i & \sum x_i'^2 \end{pmatrix}^{-1} = \begin{pmatrix} 15.3185 & -6.0699 & -10.0172 \\ -6.0699 & 3.0898 & 1.7624 \\ -10.0172 & 1.7624 & 14.1480 \end{pmatrix}$$

et

$$\sum_{i=1}^{8} y_i = -23.4, \qquad \sum_{i=1}^{8} (y_i - \overline{y})^2 = 28.8442,$$
$$\sum_{i=1}^{8} x_i\, y_i = -38.044, \qquad \sum_{i=1}^{8} (x'_i - \overline{x}')^2 = 0.0761,$$
$$\sum_{i=1}^{8} x'_i\, y_i = -12.9324, \qquad \sum_{i=1}^{8} (x'_i - \overline{x}')(y_i - \overline{y}) = -1.2617.$$
$$\sum_{i=1}^{8} (y_i - a\, x_i - a'\, x'_i - b)^2 = 6.9950$$

Exercice E.27 (DS2 2006-2007) On considère la série chronologique $(t_i, y_i)_{1 \leq i \leq 20}$ suivante :

t_i	1	2	3	4	5	6	7	8	9	10
y_i	5.7	7.2	7.7	2.9	5.7	7.0	6.0	10.4	10.2	8.0
t_i	11	12	13	14	15	16	17	18	19	20
y_i	12.7	14.0	15.8	12.7	21.3	17.2	25.0	23.2	28.9	32.9

Dans ce qui suit, on supposera que les données $(y_i)_{1 \leq i \leq 20}$ sont les réalisations de variables aléatoires $(Y_i)_{1 \leq i \leq 20}$ liées aux instants $(t_i)_{1 \leq i \leq 20}$ par la relation :

$$(M_3) \qquad Y_i = P_3(t_i) + \varepsilon_i \quad \text{avec} \quad (\varepsilon_i) \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

et où P_3 est le polynôme à coefficients réels de degré 3 défini par

$$P_3(t) = \alpha_0 + \alpha_1\, t + \alpha_2\, t^2 + \alpha_3\, t^3$$

On notera par (M_3) ce modèle. L'objet de cette étude est de voir si un polynôme de degré inférieur ne suffirait pas à modéliser la tendance de la série chronologique.

1. Ecrire ce modèle sous la forme vectorielle $Y = X\theta + \varepsilon$, en précisant le contenu et la dimension de chacune des variables.
2. Déterminer les estimations notées a_0, a_1, a_2, a_3 et s^2 des paramètres inconnus α_0, α_1, α_2, α_3 et σ^2. On explicitera les quantités intervenant dans les formules qui permettent le calcul de ces estimations.

   Pour le calcul des estimations, vous pouvez utiliser la sortie *Excel* ad-hoc (Tableau 3 ou 4) se trouvant en fin d'exercice, obtenue avec l'outil *Régression linéaire* de l'*Utilitaire d'analyse*.
3. On notera par (M_0), (M_1) et (M_2) les sous-modèles du modèle (M_3) respectivement associés aux polynômes

$$\begin{aligned} P_0(t) &= \alpha_0 \\ P_1(t) &= \alpha_0 + \alpha_1 t \\ P_2(t) &= \alpha_0 + \alpha_1 t + \alpha_2 t^2 \end{aligned}$$

   Déterminer pour chaque sous-modèle, les estimations des paramètres du polynôme associé. Comme dans la question 1, on explicitera les formules qui permettent d'obtenir les différentes estimations. On pourra synthétiser les différents résultats obtenus dans le Tableau ci-dessous

Polynôme	a_0	a_1	a_2	a_3
P_0				
P_1				
P_2				
P_3				

Tableau 3 – *Estimations des différents polynômes*

4. Calculer les sommes des carrés des résidus des différents modèles. On pourra synthétiser les différents résultats obtenus dans le Tableau ci-dessous

Modèle	SCR	ddl
(M_0)		
(M_1)		
(M_2)		
(M_3)		

Tableau 4 – *Somme des Carrés des Résidus*

5. Déterminer à l'aide d'un ou plusieurs tests statistiques le degré du polynôme qui suffit pour modéliser la tendance de la série chronologique. On commencera par expliquer la démarche utilisée. On réalisera les tests au niveau 5%.

6. Construire une prévision de la série à l'instant $t = 21$. Donner un intervalle de prévision au niveau de confiance 95%.

Pour vous aider, on fournit informations suivantes.

$$\sum_{i=1}^{20}(t_i-\overline{t})^2 = 665 \qquad \sum_{i=1}^{20} t_i = 210 \qquad \sum_{i=1}^{20} t_i^4 = 722\,666 \qquad \sum_{i=1}^{20} y_i = 274.5$$

$$\sum_{i=1}^{20}(y_i-\overline{y})^2 = 1\,376.82 \qquad \sum_{i=1}^{20} t_i^2 = 2\,870 \qquad \sum_{i=1}^{20} t_i^5 = 12\,333\,300 \qquad \sum_{i=1}^{20} t_i\,y_i = 3\,757.6$$

$$\sum_{i=1}^{20}(t_i-\overline{t})(y_i-\overline{y}) = 875.35 \qquad \sum_{i=1}^{20} t_i^3 = 44\,100 \qquad \sum_{i=1}^{20} t_i^6 = 216\,455\,810 \qquad \sum_{i=1}^{20} t_i^2\,y_i = 59\,373$$

$$\sum_{i=1}^{20} t_i^3\,y_i = 998\,048.2$$

- $\begin{pmatrix} 20 & 210 \\ 210 & 2870 \end{pmatrix}^{-1} = \begin{pmatrix} 0.21579 & -0.015789 \\ -0.015789 & 0.0015038 \end{pmatrix}$
- $\begin{pmatrix} 20 & 210 & 2870 \\ 210 & 2870 & 44100 \\ 2870 & 44100 & 722666 \end{pmatrix}^{-1} = \begin{pmatrix} 0.55351 & -0.10789 & 0.00439 \\ -0.10789 & 0.0266 & -0.001196 \\ 0.00439 & -0.001196 & 0.00005696 \end{pmatrix}$

ainsi que les deux sorties *Excel* suivantes :

RAPPORT DÉTAILLÉ

Coefficient de détermination multiple	0,9709
Observations	20

ANALYSE DE VARIANCE

	ddl	Somme des carrés	Moyenne des carrés	F	Valeur critique de F
Régression	2	1298,0379	649,019	140,053	2,748E-11
Résidus	17	78,7796	4,6341		
Total	19	1376,8175			

	Coefficients	Erreur-type	Statistique de Student	Probabilité	Limite inf pour seuil de conf. à 95%	Limite sup pour seuil de conf. à 95%
Constante	6,9208	1,6016	4,3213	0,0005	3,5418	10,2998
t	-0,5974	0,3512	-1,7009	0,1072	-1,3385	0,1436
t^2	0,0911	0,0162	5,6092	3,1286E-05	0,0569	0,1254

Tableau 3 – Sortie *Excel* n°1 obtenue avec l'outil *Régression linéaire* de l'*Utilitaire d'analyse*

RAPPORT DÉTAILLÉ

Coefficient de détermination multiple	0,9713
Observations	20

ANALYSE DE VARIANCE

	ddl	Somme des carrés	Moyenne des carrés	F	Valeur critique de F
Régression	3	1298,9725	432,991	88,995	3,398E-10
Résidus	16	77,8450	4,865		
Total	19	1376,8175			

	Coefficients	Erreur-type	Statistique de Student	Probabilité	Limite inf pour seuil de conf. à 95%	Limite sup pour seuil de conf. à 95%
Constante	6,1476	2,409405599	2,551490219	0,02134	1,0399	11,2553
t	-0,2029	0,969455592	-0,209294103	0,83686	-2,2581	1,8523
t^2	0,0453	0,105910675	0,427610069	0,67464	-0,1792	0,26982
t^3	0,001455	0,00332045	0,438291795	0,66704	-0,005584	0,008494

Tableau 4 – Sortie *Excel* n$^{\circ}2$ obtenue avec l'outil *Régression linéaire* de l'*Utilitaire d'analyse*

Statistique – Feuille de TD 9

Modèle Linéaire Gaussien Multiple & ANOVA 2

Exercice E.28 On souhaite expliquer le taux d'oxyde de carbone (CO) contenu dans les cigarettes à l'aide de 3 variables explicatives :

- le taux de goudron (TAR) en g ;
- le taux de nicotine (NICOTINE) en mg ;
- le poids de la cigarette (WEIGHT) en g.

On dispose de 24 mesures (voir fichier *cigarette.xls*). A l'aide d'Excel et de l'outil d'analyse *Régression linéaire* que l'on mettra en œuvre plusieurs fois, déterminer, en utilisant la méthode de régression descendante, l'ensemble des variables explicatives qui a une influence significative sur le taux de CO. On fera tous les tests au niveau 5%.

Exercice E.29 On souhaite expliquer la consommation des voitures (en $\ell/100km$) à l'aide de 4 variables explicatives :

- le Prix (en Francs) ;
- la Cylindrée (en cm^3) ;
- la Puissance (en kW) ;
- le Poids (en kg).

On dispose de 31 mesures (voir fichier *consommation.xls*). A l'aide d'Excel et de l'outil d'analyse *Régression linéaire* que l'on mettra en œuvre plusieurs fois, déterminer, en utilisant la méthode de régression ascendante, l'ensemble des variables explicatives qui a une influence significative sur la consommation des voitures. On fera tous les tests au niveau 5%.

Exercice E.30 Reprendre les données de l'exemple 1 de l'introduction du cours d'Analyse de la variance à 2 facteurs.

1. Tester la présence d'un effet traitement *niveau de la fertilisation / rotation*
2. Tester l'hypothèse selon laquelle il n'y a pas d'interaction.
3. Tester l'effet du facteur *niveau de la fertilisation*.
4. Tester l'effet du facteur *rotation*.

Tous les tests se feront au risque 5% et 1%. Pour vous aider, on fournit les informations

suivantes.

Source de variation	Somme de carrés	Degrés de liberté (ddl)
Fertilisation	$S_A^2 = 145.71$	1
Rotation	$S_B^2 = 1\,180.48$	2
Interaction	$S_{AB}^2 = 333.63$	2
Résiduelle	$SCR(M_{2\times 3}) = 3\,052.47$	54
Totale	4 712.29	59

Exercice E.31 On injecte de l'insuline à 24 lapins en leur donnant des doses notées A_1, A_2 et A_3, préparées suivant deux protocoles différents notés B_1 et B_2. La réduction de sucre dans leur sang a été mesurée et elle a donné les résultats suivants :

Réduction	B_1	B_2	Moyenne
A_1	17 21 49 54 $\overline{y}_{11\bullet} = 35.25$	33 37 40 16 $\overline{y}_{12\bullet} = 31.5$	$\overline{y}_{1\bullet\bullet} = 33.375$
A_2	64 48 34 63 $\overline{y}_{21\bullet} = 52.25$	41 64 34 64 $\overline{y}_{22\bullet} = 50.75$	$\overline{y}_{2\bullet\bullet} = 51.5$
A_3	62 72 61 91 $\overline{y}_{31\bullet} = 71.5$	56 62 57 72 $\overline{y}_{32\bullet} = 61.75$	$\overline{y}_{3\bullet\bullet} = 66.625$
Moyenne	$\overline{y}_{\bullet 1\bullet} = 53$	$\overline{y}_{\bullet 2\bullet} = 48$	$\overline{y}_{\bullet\bullet\bullet} = 50.5$

On désigne par y_{ijk} la réduction de sucre dans le sang des lapins. On suppose que les données $\left(y_{ijk}\right)$ sont les réalisations de variables aléatoires $\left(Y_{ijk}\right)$ satisfaisant pour tout $i = 1, 2, 3$, tout $j = 1, 2$ et tout $k = 1, 2, 3, 4$,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

où (ε_{ijk}) est une suite de variables aléatoires indépendantes de loi $\mathcal{N}(0, \sigma^2)$..

1. Décrire le plan d'expérience.
2. Donner des estimations de tous les paramètres inconnus du modèle.
3. Dresser le tableau d'analyse de variance.
4. Tester l'hypothèse selon laquelle il n'y a pas d'interaction.
5. Tester l'hypothèse selon laquelle la dose d'insuline n'a pas d'effet significatif sur la réduction de sucre.
6. Tester l'hypothèse selon laquelle le protocole de préparation n'a pas d'effet significatif sur la réduction de sucre.

On fournit les informations suivantes : $\overline{y}^2_{\bullet\bullet\bullet} = 2\,550.25$, $\sum_{i=1}^{3} \overline{y}^2_{i\bullet\bullet} = 8\,205.03125$, $\sum_{j=1}^{2} \overline{y}^2_{\bullet j\bullet} = 5\,113$, $\sum_{i=1}^{3}\sum_{j=1}^{2} \overline{y}^2_{ij\bullet} = 16\,465.75$ et $\sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{4} y^2_{ijk} = 69\,358$.

Exercice E.32 (GM4 - DS2 - Janvier 2007) Dans une usine de production de pièces détachées pour l'automobile, trois équipes (matin, soir et nuit) se relaient sur une même chaîne de montage. Elles y occupent quatre postes de travail A, B, C et D. On relève sur 3 mois, le nombre total de pièces défectueuses sur la production d'un mois, que l'on ventile par équipe et par poste de travail. On dispose donc de 3 observations pour chaque combinaison *Equipe* - *Poste*. Les résultats sont les suivants :

	A ($j=1$)			B ($j=2$)			C ($j=3$)			D ($j=4$)		
équipe du matin ($i=1$)	27	24	26	15	12	13	24	23	25	5	5	8
équipe du soir ($i=2$)	21	27	25	13	17	17	25	27	30	13	7	5
équipe de nuit ($i=3$)	29	28	25	25	23	23	28	34	27	7	10	13

Tableau 5 – *Nombre de pièces défectueuses par poste et par équipe.*

On pourra noter y_{ijk} le nombre de pièces défectueuses relevées le $k^{\text{ème}}$ mois par l'équipe i travaillant sur le poste j.

On souhaite interpréter ce tableau à l'aide d'une analyse de la variance en étudiant l'effet des deux facteurs *Equipe* et *Poste de travail* sur le nombre de pièces défectueuses produites dans cette chaîne de montage.

Indication. Vous pouvez répondre aux questions de la **partie II** (sauf la question **2.d**) sans avoir traité la **partie I**.
Tous les tests se feront au niveau 5%.
Les différents tableaux à compléter se trouvent à la fin de la feuille de TD.

Partie I. Dans un premier temps, on analyse séparément l'effet des deux facteurs *Equipe* et *Poste de travail* sur la production à travers le nombre de pièces défectueuses produites avec un modèle d'analyse de la variance à un facteur. Vous pouvez répondre à la question **2** sans avoir traité la question **1**.

1. **On analyse tout d'abord l'influence de l'équipe sur la production.**
   (a) Compléter le Tableau de données 6 (voir dernière feuille du sujet).
   (b) Décrire avec soin le modèle envisagé avec toutes ses hypothèses pour étudier l'effet du facteur *Equipe* sur la production de pièces défectueuses. On précisera quels sont les paramètres du modèle et leurs contraintes, et on donnera la dimension du modèle.
   (c) En vous servant du Tableau 6 et du Tableau 7 d'analyse de la variance que vous aurez préalablement complété, donner une estimation des paramètres de ce modèle (paramètres d'espérance et de variance résiduelle). On donnera directement l'expression des estimations de ces paramètres, puis leurs valeurs sur les données.
   (d) En analysant la table de l'ANOVA 1 (Tableau 7), peut-on conclure à l'existence d'une différence qualitative entre les performances globales des équipes ?
   On commencera par préciser les hypothèses H_0 et H_1 en termes de valeurs sur les paramètres ainsi qu'en termes de comparaison de modèles (on donnera alors la dimension de chacun des deux modèles que l'on compare), puis on donnera la statistique de test, sa loi sous H_0 et la région de rejet.

2. **On étudie maintenant l'influence du poste de travail sur la production.**
   (a) Compléter le Tableau de données 8.
   (b) Décrire avec soin le modèle envisagé avec toutes ses hypothèses pour étudier l'effet du facteur *Poste* sur la production de pièces défectueuses. On précisera quels sont les paramètres du modèle et leurs contraintes, et on donnera la dimension du modèle.
   (c) En vous servant du Tableau 8 et du Tableau 9 d'analyse de la variance que vous aurez préalablement complété, donner une estimation des paramètres de ce modèle (paramètres d'espérance et de variance résiduelle). On donnera directement l'expression des estimations de ces paramètres, puis leurs valeurs sur les données.
   (d) En analysant la table de l'ANOVA 1 (Tableau 9), peut-on conclure que les postes de travail présentent des difficultés de montage inégales ?
   On commencera par préciser les hypothèses H_0 et H_1 en termes de valeurs sur les paramètres ainsi qu'en termes de comparaison de modèles (on donnera alors la dimension de chacun des deux modèles que l'on compare), puis on donnera la statistique de test, sa loi sous H_0 et la région de rejet.

Partie II. On désire à présent tenir compte de l'effet conjoint des deux facteurs dans la modélisation pour analyser les données du Tableau 5.

1. Décrire le plan d'expérience associé aux données du Tableau 5. S'agit-il d'un plan d'expérience orthogonal ?

2. Dans toute cette question, on se place dans un modèle **complet** d'analyse de la variance à deux facteurs pour étudier les données du Tableau 5.
   (a) Expliquez pourquoi il est pertinent d'envisager de modéliser les données par un tel modèle. Décrire avec soin ce modèle avec toutes ses hypothèses. On précisera quels sont les paramètres du modèle, les contraintes d'identifiabilité et on donnera la dimension du modèle.

   A l'aide du Tableau 10, que l'on aura préalablement complété,
   (b) Tester l'effet conjoint de l'équipe et du poste sur la production.
   (c) Tester l'effet de l'interaction sur la production.
   (d) Tester l'effet de l'équipe sur la production.
   (e) Tester l'effet du poste sur la production.

   Pour chacun des tests, vous indiquerez quelles sont les hypothèses testées en termes de paramètres et en termes de comparaison de modèles. Vous donnerez la dimension de chaque sous-modèle du modèle complet envisagé. Vous donnerez également la statistique de test, sa loi sous H_0 et la région de rejet.

   Hors barême. Cette question est facultative et ne pourra rapporter que des points supplémentaires.
3. On se place maintenant dans le modèle d'analyse de la variance à deux facteurs sans interaction.
   (a) Vous semble t-il pertinent d'étudier ce modèle d'après les résultats de la question **1** ?
   (b) Compléter le Tableau 11 d'analyse de la variance dans ce modèle sans interaction à partir de celui établi dans le modèle complet (Tableau 10).
   (c) A l'aide du tableau 11, tester si l'équipe, puis le poste ont un effet significatif sur le nombre de pièces défectueuses.
   (d) Comparer les résultats obtenus avec ceux obtenus dans les questions **1.d** et **2.d** de la **Partie I**. Vous semble t-il plus pertinent de faire une analyse de la variance de chaque facteur séparément comme dans la **Partie I.** ou bien une analyse de la variance à deux facteurs comme dans la **Partie II** ?

Tableaux relatifs à l'Exercice 32

Equipe	Matin	Soir	Nuit	Global
nombre d'observations				
nombre moyen de pièces défectueuses				

Tableau 6 – *Nombre de pièces défectueuses par équipe*

Source	ddl	Sommes des Carrés	Carrés Moyens	Z
Modèle (Equipe)	2	184,72	92,36	.....
Résidu	33	2291,83	69,45	
Total	35	2476,56		

Tableau 7 – *Table de l'ANOVA 1 du nombre de pièces défectueuses en fonction de l'équipe*

Poste	A	B	C	D	Global
nombre d'observations					
nombre moyen de pièces défectueuses					

Tableau 8 – *Nombre de pièces défectueuses par poste*

Source	ddl	Sommes des Carrés	Carrés Moyens	Z
Modèle (poste)	3	2061,89	687,3	.....
Résidu	.....	.....	.....	
Total	35	2476,56		

Tableau 9 – *Table de l'ANOVA 1 du nombre de pièces défectueuses en fonction du poste*

Source de variation	ddl	Somme des Carrés	Carrés Moyens	Z
Modèle(Traitement)	11	2324,56	211,32	.....
Résiduelle	24	152	6,33	
Total	35	2476,56		

Source de variation	ddl	Sommes des Carrés	Carrés Moyens	Z
équipe	2	184,72	92,36	.....
poste	.....	.....	.....	.....
équipe * poste	6	77,94	12.99	.....
Résiduelle	.....	.....	.....	
Total	35	2476,56		

Tableau 10 – *Tables de l'ANOVA à deux facteurs dans le modèle complet avec interaction.*

Source de variation	ddl	Sommes des Carrés	Carrés Moyens	z
équipe	.....	.....	.....	.....
poste	.....	.....	.....	.....
Résiduelle	.....	.....	.....	
Total	.....	.....		

Tableau 11 – *Table de l'ANOVA à deux facteurs dans le modèle sans interaction.*