GC – Probabilités et Statistique – Séries 10a + 11a + 11b

En salle

Exercice 1 On a relevé, pendant 25 années, le rendement de blé y (en tonnes par hectare) en fonction de la quantité de pluie annuelle x_1 (en centimètres) et de la température moyenne annuelle x_2 (en degrés Fahrenheit).

Les estimation des paramètres : $\hat{\beta}_0 = 32.051 \quad \hat{\beta}_1 = 0.480 \quad \hat{\beta}_2 = -0.555$

Les valeurs diagonales de la matrice $\mathbf{Z} = (\mathbf{X}'\mathbf{X})^{-1}$:

$$\text{diag } (\mathbf{X}'\mathbf{X})^{-1} = (z_{11},\, z_{22},\, z_{33}) = (3.437,\, 0.0014,\, 0.0021)$$

Pour le modèle complet Ω :

source	df	SC	CM	F	p-valeur
régression	2	184.2	?	?	0.0020
erreur	22	243.3	?		
total	?	?			

(a) Écrire le modèle linéaire statistique complet (Ω).
(b) Quelle est l'expression matricielle de l'estimateur des moindres carrés ordinaires du vecteur $\boldsymbol{\beta}$ des paramètres ?
(c) Trouver $Var(\hat{\boldsymbol{\beta}})$.
(d) Compléter le tableau d'ANOVA.
(e) Trouver R^2 et R^2-ajusté.
(f) Tester $H : \beta_1 = \beta_2 = 0$ au niveau $\alpha = 5\%$ (veiller à spécifier A).
(g) Calculer un intervalle de confiance à 95% pour le paramètre β_2.

Exercice 2 Une chercheuse s'intéresse aux mérites relatifs des régimes par rapport aux médicaments hypocholestérolémiants. Pour chacun des 65 sujets qui ont commencé l'étude avec un cholestérol élevé, elle enregistre le niveau total de cholestérol sanguin (en mg par décilitre) après 6 mois de participation à l'étude. Les patients sont répartis en 5 groupes : un groupe de contrôle (C) qui reçoit un placebo, un groupe de régime végétarien (V), un groupe de régime pauvre en graisses (PG), un groupe à faible dose des médicaments (DF) et un groupe avec une dose élevée de médicaments (DE). Les résumés des données :

Groupe	Moyenne	Écart-type	n_{groupe}
C	240	1.22	25
V	225	1.18	10
PG	230	1.10	10
DF	215	1.02	10
DE	200	1.11	10

Tableau d'ANOVA :

Source	df	SC	CM	F	p
Traitements					4.8×10^{-11}
Erreur	60		80		
Total		11681.4			

(a) Quelles suppositions doivent être remplies pour obtenir une p-valeur valide en utilisant ANOVA ?
(b) Si vous aviez toutes les données, quels examens graphique et numérique feriez-vous pour vérifier ces suppositions ?
(c) Compléter le tableau.

(d) Sur la base de ces données, existe-il des preuves qu'au moins l'une des moyennes des groupes diffère(nt) ? Justifier votre réponse en effectuant un test d'hypothèses approprié. Énoncer les hypothèses nulle et alternative. Quelle est votre conclusion, si $\alpha = 0.01$? Interpréter le résultat.

(e) Quel pourcentage de la variabilité est expliqué par la différence entre les traitements ?

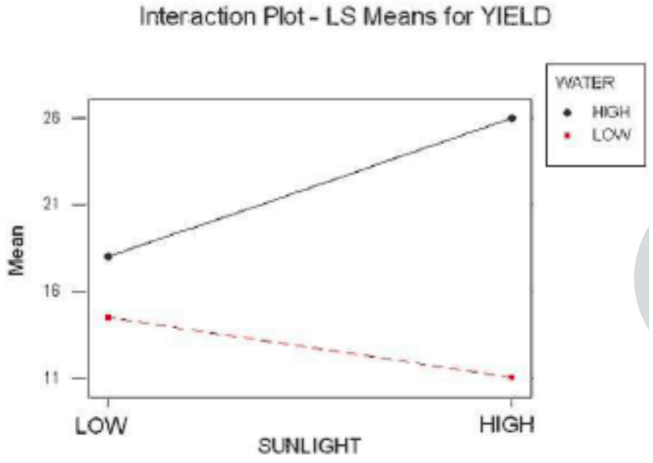
(f) Calculer les statistiques de test t_{obs} pour les 10 tests de comparaison (chaque paire).

(g) Quelles paires de moyennes sont différentes au niveau (nominal - c'est-à-dire sans ajustement) de signification $\alpha = 0{,}01$?

(h) Selon la méthode Bonferroni, trouver le seuil pour chaque test si vous voulez un seuil global de $\alpha = 0{,}01$. Quelles paires de moyennes sont différentes au niveau de signification (global) $\alpha = 0{,}01$?

Exercice 3 La quantité de soleil et l'arrosage influent-ils sur la croissance des géraniums ? La croissance des plantes de 16 plantes est mesurée en centimètres. Chaque combinaison de lumière du soleil et d'eau (élevé/faible) comporte 4 plantes ; p. ex., la combinaison d'eau à haut niveau avec un ensoleillement élevé comprend 4 plantes d'une longueur de 21 à 30 cm. Le tableau d'ANOVA (partiellement complet) se trouve ci-dessous, ainsi qu'un graphique d'interaction.

Source	df	SC	CM	F	$P(> F)$
Water	(a)	342.2	(g)	(k)	0.000365 ***
Sunlight	(b)	20.2	(h)	(l)	0.256272
(n)	(c)	132.2	(i)	(m)	0.010152 *
Error	(d)	(f)	(j)		
Total	(e)	665.6			



(a)-(n) Complèter le tableau.

(o) Écrire le modèle **complete** (théorique), y compris toutes les suppositions du modèle.

(p) La quantité (niveau) d ?arrosage affecte-t-elle la croissance des géraniums en pot ? Réaliser le test d'hypothèse pertinent en veillant à rédiger chacune des 5 étapes, et *interpréter vos résultats* (utiliser $\alpha = 0,05$).

(q) La quantité (niveau) de lumière solaire affecte-t-elle la croissance des géraniums en pot ? Réaliser le test d'hypothèse pertinent en veillant à rédiger chacune des 5 étapes, et *interpréter vos résultats* (utiliser $\alpha = 0,05$).

(r) L'effet du niveau d'ensoleillement dépend-il du niveau d'arrosage ? Réaliser le test d'hypothèse pertinent en veillant à rédiger chacune des 5 étapes, et *interpréter vos résultats* (utiliser $\alpha = 0,05$).

(s) Est-il judicieux de supprimer la variable « sunlight » du modèle ? **Expliquer.**

À domicile

Exercice 1 Un jeu de données mettent en évidence l'oxydation de l'ammoniaque pour produire de l'acide nitrique. La perte de la matière (y = 'stackloss') doit être expliquée par le volume de l'air (x_1 = 'Air Flow'), la température de l'eau refroidissante (x_2 = 'Water Temp'), et la concentration de l'acide (x_2 = 'Acid Conc.').

Considérer les deux sorties du logiciel statistique `R` représentant les estimation pour les deux modèles suivants :

(A) $\Omega : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$,
(B) $\omega : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$

(a) Commenter tous les chiffres et donner des interprétations détaillées. Pour les 'p-values', donner les hypothèses H et A du test et la conclusion.
(b) Combien des observations y a-t-il ?
(c) Pour le modèle Ω, vérifier le calcul de l'estimation de σ, l'écart-type de l'erreur dans le modèle.
(d) Construire un intervalle de confiance à 95% pour β_2 dans le modèle Ω.
(e) Effectuer le test H : ω est le vrai modèle, en utilisant le test F au niveau de 5% ; en tirer des conclusions.

Exercice 2 Des résultats ont été recueillis pour quatre groupes indépendants de dix sujets chacun. Le tableau d'ANOVA :

Source	df	SC	CM	F	p
(h)	(a)	(d)	116	(g)	1.8×10^{-5}
Erreur	(b)	360	(f)		
Total	(c)	(e)			

(a)-(g) Compléter le tableau.
(h) Tester l'hypothèse que les moyennes des groupes sont égales (vous pouvez supposer que les suppositions de l'ANOVA sont vérifiées). Utiliser $\alpha = 0.05$.
(i) Sous H, il y a combien de degrés de liberté de la statistique de test ?

Exercice 3 Le nombre de grammes de graisse par portion pour trois variétés de pizza différentes de plusieurs fabricants est mesuré et un tableau d'ANOVA partielle est fournie ci-dessous :

Tableau d'ANOVA :

Source	df	SC	CM	F	p
		23.0			0.53
Erreur					
Total	20	339.9			

(a) Supposons que les suppositions d'ANOVA pour qu'un test soit valide sont satisfaites. Compléter le tableau d'ANOVA.
(b) Au niveau de signification $\alpha = 0{,}05$, y a-t-il une différence significative dans les contenus des matières graisses moyennes pour les trois variétés de pizza ?
(c) Est-il logique de faire des comparaisons post-hoc (c.-à-d., après le test global) par paire ? Expliquer.

```
Coefficients:
            Estimate Std. t value Pr(>|t|)
                     Error
(Intercept) -39.9    11.9  -3.4   0.0038
Air Flow      0.7     0.1   5.3   5.8e-05
Water Temp    1.3     0.4   3.5   0.0026
Acid Conc.   -0.2     0.2  -0.97  0.3440
---
Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared: 0.9136,     Adjusted R-squared: 0.8983
F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09


Analysis of Variance Table

Response: stack.loss
           Df  Sum Sq Mean Sq F value  Pr(>F)
stack.x     3 1890.4  630.1    59.5    3.01633e-09
Residuals  17  178.8   10.5
```

```
Coefficients:
            Estimate Std. t value Pr(>|t|)
                     Error
(Intercept) -50.4   5.1    -9.8   1.22e-08
Air Flow      0.7   0.1     5.3   4.90e-05
Water Temp    1.3   0.4     3.5   0.0024
---
Residual standard error: 3.239 on 18 degrees of freedom
Multiple R-squared: 0.9088,     Adjusted R-squared: 0.8986
F-statistic: 89.64 on 2 and 18 DF,  p-value: 4.382e-10


Analysis of Variance Table

Response: stack.loss
           Df  Sum Sq Mean Sq F value  Pr(>F)
stack.x     2 1880.4   940.2  89.6     4.38154e-10
Residuals  18  188.8   10.5
```

Exercice 4 Vingt-deux patients subissant un pontage aortocoronaire ont été randomisés pour l'un des trois groupes de ventilation :

(i) un mélange de 50% protoxyde d'azote et de 50% d'oxygène, en continu pendant 24 heures

(ii) un mélange de 50% protoxyde d'azote et de 50% d'oxygène, seulement pendant l'opération

(iii) pas de protoxyde d'azote, mais 35-50% d'oxygène, en continu pendant 24 heures.

La question d'intérêt est de savoir si le folate moyen des globules rouges est différent pour les trois méthodes de ventilation. Les données sont analysées par ANOVA.

(a) Quelles suppositions doivent être satisfaites pour obtenir une p-valeur valide en utilisant ANOVA ?

(b) Si vous aviez toutes les données, quels examens graphiques et numériques feriez-vous pour évaluer ces suppositions ?

(c) Quelle est l'hypothèse nulle pour le test d'ANOVA ?

(d) Utiliser le tableau ANOVA ci-dessous (obtenu à partir du logiciel R) afin de déterminer si l'hypothèse nulle est rejetée à un niveau de 5%. Interpréter le résultat.

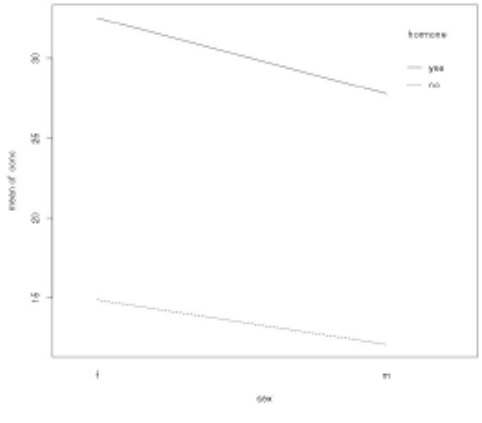
(e) Expliquer pourquoi effectuant un seul test conjoint (ANOVA) est préférable à effectuant plusieurs tests par paires.

```
> redcell.aov<-aov(Folate~Group)
> summary(redcell.aov)
            Df Sum Sq Mean Sq F value  Pr(>F)
Group        2  15516    7758  3.7113 0.04359 *
Residuals   19  39716    2090
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

Exercice 5 Nous voulons tester si la concentration plasmatique de calcium (mg/100 ml) des oiseaux mâles et femelles est affectés par le traitement hormonal. Les sorties de R de l'expérience sont données ci-dessous. Vous pouvez supposer que le désign est équilibrée.

```
             Df Sum Sq Mean Sq F value   Pr(>F)
sex           1   70.3    70.3   3.733   0.0713 .
hormone       1 1386.1  1386.1  73.585 2.22e-07 ***
sex:hormone   1    4.9     4.9   0.260   0.6170
Residuals    16  301.4    18.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
interaction.plot(sex,hormone,conc)
```



(a) Combien d'oiseaux y a-t-il dans l'étude ? Combien de mâles ? Combien de femelles ?

(b) Combien de traitements hormonaux existe-t-il ?

(c) Combien d'oiseaux y a-t-il dans chaque groupe d'hormone $\times$ sexe $\times$?

(d) Combien de termes d'interaction peuvent être estimés ? Comment vous le savez ?

(e) Écrire le modèle **complète** (théorique), y compris toutes les suppositions du modèle.

(f) Existe-t-il une preuve graphique d'un effet d'interaction ? **Expliquer.**

(g) Écrire les 3 tests d'hypothèses possibles qui peuvent être effectués dans cette expérience. Quelles hypothèses NULLES sont REJETÉES au niveau $\alpha = 10\%$?

(h) En utilisant les résultats de la partie (g), écrire le modèle final (théorique).