GM – Probabilités et Statistique – Corrigés 11

À noter : les **raisonnements/justifications** des étapes de résolution sont également importantes (pas seulement le résultat final).

En salle

Exercice 1 (a) Échantillons indépendants (attribution aux groupes aléatoire), variances égales pour chaque groupe, et soit les mesures sont normalmenet distributés dans chaque group, soit les tailles des échantillons sont suffisamment grandes que les moyennes pourraient être supposées distribuées normalement (TCL).

(b) Tableau ANOVA :

Source	df	SC	CM	F	p
Treatments	$5-1=\mathbf{4}$	$11681.4-4800=\mathbf{6881.4}$	$6881.4/4=\mathbf{1720.4}$	$1720.4/80=\mathbf{21.5}$	4.8×10^{-11}
Error	60	$60\times 80=\mathbf{4800}$	80		
Total	$60+4=\mathbf{64}$	11681.4			

(c) $H:\mu_1=\mu_2=\cdots=\mu_5$
$A:\exists\,\mu_i\neq\mu_j$ (au moins une moyenne diffèrent des autres)
$p_{obs}=4.8\times 10^{-11}<<0.01$, donc on REJETTE l'hypothèse NULLE H. Au moins un groupe a une moyenne qui est (statistiquement) significativement differente que les autres.

(d) La proportion de variation expliquée $=SS_{trts}/SStotal=6881.4/11681.4=0.589\approx 60\%$.

(e) Les tests :

test	t_{obs}
** $H:\mu_C=\mu_V$	$\dfrac{240-225}{\sqrt{80\,(1/25+1/10)}}=4.48$
* $H:\mu_C=\mu_{PG}$	$\dfrac{240-230}{\sqrt{80\,(1/25+1/10)}}=2.99$
** $H:\mu_C=\mu_{DF}$	$\dfrac{240-215}{\sqrt{80\,(1/25+1/10)}}=7.47$
** $H:\mu_C=\mu_{DE}$	$\dfrac{240-200}{\sqrt{80\,(1/25+1/10)}}=11.95$
$H:\mu_V=\mu_{PG}$	$\dfrac{225-230}{\sqrt{80\,(1/10+1/10)}}=-1.25$
$H:\mu_V=\mu_{DF}$	$\dfrac{225-215}{\sqrt{80\,(1/10+1/10)}}=2.50$
** $H:\mu_V=\mu_{DE}$	$\dfrac{225-200}{\sqrt{80\,(1/10+1/10)}}=6.25$
** $H:\mu_{PG}=\mu_{DF}$	$\dfrac{230-215}{\sqrt{80\,(1/10+1/10)}}=3.75$
** $H:\mu_{PG}=\mu_{DE}$	$\dfrac{230-200}{\sqrt{80\,(1/10+1/10)}}=7.50$
** $H:\mu_{DF}=\mu_{DE}$	$\dfrac{215-200}{\sqrt{80\,(1/10+1/10)}}=3.75$

(f) $t_{60,0.995}=2.660$; dans la table : * et ** significatif au niveau nominal (sans ajustment) de 0.01 ($|t_{obs}|>2.660$)

(g) $\alpha' = \alpha/10 = 0.001 \implies t_{60,0.9995} = 3.460$; dans la table : ** significatif au niveau global (ajustment de Bonferroni) de 0.01 ($|t_{obs}| > 3.460$)

Source	df	SC	CM (= SC/df)	F $\left(= \frac{\text{MS(source)}}{\text{MSerror)}}]\right)$	p
Water	(a) # groups − 1 $\implies 2-1=\mathbf{1}$	342.2	(g) 342.2/1 = **342.2**	(k) 342.2/14.25 = **24.01**	0.000365 ***
Sunlight	(b) # groups − 1 $\implies 2-1=\mathbf{1}$	20.2	(h) 20.2/1 = **20.2**	(l) 20.2/14.25 = **1.42**	0.256272
(n) Water × Sunlight (Interaction)	(c) (# groups − 1) × (# groups − 1) $\implies (2-1)\times(2-1)=\mathbf{1}$	132.2	(i) 132.2/1 = **132.2**	(m) 132.2/14.25 = **9.28**	0.010152 *
Error	(d) df total - df group $\implies 15-1-1-1=\mathbf{12}$	(f) 665.6 − 342.2 − 20.2 − 132.2 = **171**	(j) 171/12 = **14.25**		
Total	(e) # obs − 1 $\implies 16-1=\mathbf{15}$	665.6			

Exercice 2 (a)-(n) (see the table)

(o) Soit Y_{ijk} = croissance des plantes (en cm) pour chaque individu k au niveau de l'eau i et au niveau d'ensoleillement j ; μ = constante ; α_i = effet de l'eau ($i = 1, 2$) ; β_j = effet de lumière du soleil ($j = 1, 2$) ; γ_{ij} = effet d'interaction de l'eau et de la lumière du soleil ; ϵ_{ijk} = terme d'erreur pour l'individu k aux niveaux i et j pour l'eau et le soleil, respectivement. Aussi des contraintes : par exemple $\sum_i \alpha_i = \sum_j \beta_j = \sum_{i,j} \gamma_{ij} = 0$, ou $\alpha_2 = \beta_2 = \gamma_{1,2} = \gamma_{2,1} = \gamma_{2,2} = 0$. (Contraintes nécessaires pour se conformer aux df disponibles.)

Structure du modèle : $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$

Suppositions du modèle : $\epsilon_{ijk} \sim$ iid $N(0, \sigma^2)$.

(p) 1. paramètre = α_1 (seulement 1 df pour niveau d'eau)
2. $H : \alpha_1 = 0$
$A : \alpha_1 \neq 0$
3. $F_{obs} = 24.01$
4. Sous H, $F_{obs} \sim F_{1,12} \implies p_{obs} = 0.000365$
5. $p_{obs} = 0.000365 \leq 0.05 = \alpha \implies$ REJETTE l'hypothèse NULLE H ; il y a un effet (très) statistiquement significatif d'eau sur la croissance.

(q) 1. paramètre = β_1 (seulement 1 df pour niveau du soleil)
2. $H : \beta_1 = 0$
$A : \beta_1 \neq 0$
3. $F_{obs} = 1.42$
4. Sous H, $F_{obs} \sim F_{1,12} \implies p_{obs} = 0.256272$
5. $p_{obs} = 0.256272 > 0.05 = \alpha \implies$ NE REJETTE PAS l'hypothèse NULLE H ; l'effet du soleil sur la croissance n'est PAS statistiquement significatif.

(r) 1. paramètres = γ_{ij}, $i, j = 1$
2. $H : \gamma_{11} = 0$ (seulement 1 df interaction – on peut choisire n'importe quel des 4 possibilitités)
$A : \gamma_{11} \neq 0$
3. $F_{obs} = 9.28$
4. Sous H, $F_{obs} \sim F_{1,12} \implies p_{obs} = 0.010152$
5. $p_{obs} = 0.010152 \leq 0.05 = \alpha \implies$ REJETTE l'hypothèse NULLE H ; il existe un effet d'interaction (très) statistiquement significatif entre l'eau et le soleil sur la croissance, ce qui signifie que l'effet du niveau de lumière solaire est différent pour différents niveaux d'eau.

Il existe également des preuves graphiques d'une interaction – les lignes sur le tracé d'interaction ne sont PAS parallèles.

(s) Même si ‘lumière du soleil’ n’est pas *marginalement* significatif, il pourrait le devenir si l ?interaction terme serait supprimé – il se pourrait que l’interaction Le terme capte une partie de la variabilité de la « lumière du soleil ».

En général (mais pas nécessairement toujours), vous devriez partir dans des termes inférieurs lorsque vous incluez un effet d’interaction.

À domicile

Exercice 1 (a) Le nombre d’oiseaux est Total df + 1 = 19 + 1 = 20 ; le designe est *équilibré*, donc le number de mâles = nombre de femelles = 20/2 = 10.

(b) Nombre de traitements hormonaux est df + 1 = 1 + 1 = 2.

(c) Il y a 4 sexe × hormone groupes, donc le numbre d’oiseaux dans chaque groupe est 20/4 = 5.

(d) Il y a **1** df interaction, donc seulement 1 interaction pourrait être estimatée.

(e) (Même que l’Exercice 1(o), avec les substitutions pour les noms des variables)

(f) NON – les traits sont (à peu près) parallèl.

(g) Les 3 NULLE hypothèses possibles sont :

(1) $H : \alpha_1 = 0$

(2) $H : \beta_1 = 0$

(3) $H : \gamma_{11} = 0$

Au niveau de 10%, on REJETTE les hypothèses 1 et 2 ; on NE REJETTE PAS l’hypothèse 3 (pas d’interaction)

(h) Puisque on a REJETTÉ la présence d’un effet d’interaction, on peut réduire le modèle comme la suite :

Structure du modèle : $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$

Suppositions du modèle : $\epsilon_{ijk} \sim$ iid $N(0, \sigma^2)$.

Exercice 2 (a)

Source	df	SC	CM	F	p
Taille	2	26051	13026	199.1	< 2e-16 ***
Filtre	1	1056	1056	16.1	0.00373 ***
Size : Filter	2	804	402	6.1	0.000158 ***
Error	30	1962	65		
Total	35	29874			

(b) Oui, un modéle d’interaction est approprié car la terme d’interaction (Taille :Filtre) est hautement statistiquement significatif.

(c) Le pourcentage de la variabilité totale expliqué par le modèle est : SC modèle / SC Totaux
= (26051 + 1056 + 804)/29874 = 27912/29874 ≈ 93%.

Exercice 3 La non-significativité des effets principaux est trompeuse ici, car la modification du niveau de A a des effets importants sur le résultat pour chaque niveau de B, mais les effets *diffèrent* selon le niveau de B considéré. Par conséquent, nous ne pouvons pas interpréter l’effet de A comme un effet *moyen* ; cela annule les effets, dont nous pouvons constater l’importance.