UE BIOSTATISTIQUES et BIOINFORMATIQUE
Contrôle Terminal - Session 1
Jeudi 11 janvier 2018
Durée 90 minutes

Instructions

Ce formulaire sera analysé par lecture optique, toute intervention manuelle rendue nécessaire par le non-respect des règles ci-dessous sera susceptible d'être sanctionnée par un retrait de points.

- Pour sélectionner une case, remplissez-la intégralement au stylo à bille en **noir** : □ → ■.
- Ne pas utiliser de crayon à papier.
- Pour corriger effacez la case avec du correcteur blanc.
- N'inscrivez rien dans l'en-tête ou dans les marges des pages.
- Il n'y a qu'une réponse juste pour chaque question.
- Une réponse incorrecte à une question entraine une perte de points égale à $\frac{1}{K-1}$, avec K le nombre de réponses proposées à la question.

Identité

Renseignez les champs ci-dessous et codez votre numéro d'étudiant ci-contre.

Nom et Prénom :

.....

Numéro d'étudiant :

.....

□0	□0	□0	□0	□0	□0	□0	□0
□1	□1	□1	□1	□1	□1	□1	□1
□2	□2	□2	□2	□2	□2	□2	□2
□3	□3	□3	□3	□3	□3	□3	□3
□4	□4	□4	□4	□4	□4	□4	□4
□5	□5	□5	□5	□5	□5	□5	□5
□6	□6	□6	□6	□6	□6	□6	□6
□7	□7	□7	□7	□7	□7	□7	□7
□8	□8	□8	□8	□8	□8	□8	□8
□9	□9	□9	□9	□9	□9	□9	□9

Le sujet comporte 4 parties qui peuvent être traitées indépendemment. Il est cependant conseillé de lire attentivement les parties 1, 2 et 3 avant d'aborder la partie 4. Tous les tests et intervalles de confiance seront réalisés avec un risque $\alpha = 0.05$.

Des chercheurs s'intéressent aux différences de concentration d'une hormone H entre 3 élevages de poulets de Bresse (espèce *Gallus gallus domesticus*) différant par le type d'alimentation qu'ils reçoivent.

1 Construction des échantillons

Dans chacun des 3 élevages, les chercheurs prennent au hasard 20 poulets adultes (constituant un échantillon aléatoire simple de l'élevage correspondant). Pour chacun d'entre eux, ils effectuent une analyse préliminaire recherchant la présence d'un virus V étant connu pour compromettre le dosage de l'hormone. Les résultats de cette analyse préliminaire sont résumés dans le tableau ci-dessous :

	Elevage 1	Elevage 2	Elevage 3
Test Négatif	14	19	6
Test Positif	6	11	14

Question 1 Quel test permet de tester H_0 : « la proportion de poulets infectés par le virus V est la même dans les trois élevages » ?

- ☐ Une ANOVA à un facteur contrôlé
- ☐ Un test du χ^2 d'ajustement à une loi binomiale
- ☐ Un test du χ^2 d'ajustement à une loi normale
- ☐ Un test du χ^2 d'indépendance/homogénéité

Question 2 Que vaut la valeur observée de la statistique du test ?

☐ 11.01 ☐ 8.15 ☐ 6.54 ☐ 7.12

Question 3 Que vaut la valeur seuil de la statistique de ce test ?

☐ 3.84 ☐ 1.96 ☐ 6.12 ☐ 5.99

Question 4 Que pouvez-vous en conclure ?

- ☐ H_0 est rejetée avec un risque d'erreur de première espèce inconnu
- ☐ H_0 est rejetée avec un risque d'erreur de deuxième espèce inconnu
- ☐ H_0 est rejetée avec un risque d'erreur de première espèce de 5%
- ☐ H_0 est rejetée avec un risque d'erreur de deuxième espèce de 5%

Question 5 Quelle conclusion biologique vous semble la plus pertinente ?
Les données montrent

- ☐ une différence significative dans la proportion d'individus d'infectés par le virus V entre les 3 échantillons
- ☐ une différence significative de proportion d'individus infectés par le virus V entre les 3 élevages
- ☐ que la proportion d'individus infectés par le virus V est significativement plus grande dans le troisième élevage
- ☐ que la proportion d'individus infectés par le virus V est la même dans les 3 élevages

2 Comparaison de la concentration hormonale moyenne entre les 3 élevages

Pour chaque poulet testé négatif au virus V, les chercheurs procèdent alors à un dosage de la concentration hormonale (variable réponse $Y = log(concentrationdeH)$). Les résultats sont consignés dans le tableau ci-dessous (les concentrations en hormone H sont exprimées en nmol/L) :

	Elevage 1	Elevage 2	Elevage 3
n	14	9	6
$\sum y_i$	6.80	13.51	4.25
$\sum y_i^2$	5.94	20.10	3.92

Nous considérerons ici que les conditions d'application des tests sont respectées.

Question 6 A partir de ces données, donnez une estimation de la moyenne et de la variance de Y dans le premier élevage :

- ☐ $\mu_1 = 0.486$ et $\sigma_1^2 = 0.203$
- ☐ $\hat{\mu_1} = 0.486$ et $\hat{\sigma_1^2} = 0.203$
- ☐ $\bar{y_1}$=0.486 et $S_1^2 = 0.203$
- ☐ $\bar{y_1}$=0.486 et $S_1^2 = 0.188$

Question 7 Dans les questions suivantes, nous allons réaliser une ANOVA pour tester H_0 : « la concentration hormonale moyenne est la même dans les trois élevages ». Laquelle de ces propositions formalise mathématiquement l'hypothèse nulle H_0 correspondant à ce test :

- ☐ $\mu_i = \mu_j$ pour tous les élevages i et j
- ☐ $\mu_i \neq \mu_j$ pour au moins deux élevages i et j

- ☐ $\bar{y_i} = \bar{y_j}$ pour tous les élevages i et j
- ☐ Trois hypothèses sont testées : $H0_1 : \mu_1 = \mu_2$; $H0_2 : \mu_2 = \mu_3$, $H0_3 : \mu_1 = \mu_3$
- ☐ $\bar{y_i} \neq \bar{y_j}$ pour au moins deux élevages i et j

A partir des informations présentées dans le tableau 2, on calcule les sommes des carrés des écarts suivantes :

— $SCE_{tot} = 9.16$

— $SCE_{res} = 3.37$

Question 8 Que vaut la statistique F observée pour ce test :

☐ 0.63 ☐ 22.37 ☐ 14.52 ☐ 8.85

Question 9 Que vaut la valeur seuil de F associée à ce test :

☐ 3.37 ☐ 5.53 ☐ 1.96 ☐ 3.84

Question 10 Que pouvez-vous en conclure ?

- ☐ H_0 est rejetée avec un risque d'erreur de deuxième espèce inconnu
- ☐ Les trois hypothèses nulles sont rejetées avec un risque d'erreur de première espèce de 5%
- ☐ H_0 est rejetée avec un risque d'erreur de première espèce inconnu
- ☐ H_0 est rejetée avec un risque d'erreur de deuxième espèce de 5%
- ☐ H_0 est rejetée avec un risque d'erreur de première espèce de 5%

Question 11 Quelle conclusion biologique vous semble la plus pertinente ?

- ☐ Les trois concentrations hormonales moyennes différent toutes entre elles
- ☐ Les concentrations hormonales moyennes ne sont pas significativement différentes entre les trois élevages
- ☐ Les données montrent une différence significative de concentration hormonale moyenne entre les trois échantillons
- ☐ Les données montrent une différence significative de concentration hormonale moyenne entre les trois élevages
- ☐ Les concentrations hormonales moyennes sont les mêmes dans les trois élevages

3 Concentration hormonale en phase de croissance

Les chercheurs souhaitent désormais évaluer si la concentration de l'hormone H (variable réponse $Y = log$(concentration de H), avec H en nmol/L) varie en fonction de l'âge du poulet (X, en mois) durant sa phase de croissance juvénile. L'âge de chaque poussin est connu d'après la bague qui lui a été posée à sa naissance, et les individus atteignent la maturité sexuelle autour de 5 mois. A cette fin, ils prélèvent au hasard 30 juvéniles dans le troisième élevage et recherchent la présence du virus. Ils obtiennent 22 poulets négatifs chez qui ils réalisent un dosage de l'hormone H. Les résultats sont donnés dans la figure et le tableau ci-dessous :

$\sum x_i$	53.03
$\sum y_i$	20.00
$\sum x_i^2$	151.01
$\sum y_i^2$	24.41
$\sum x_i y_i^2$	49.29

Question 12 A partir de ces données, calculez le coefficient de corrélation linéaire entre X et Y :

☐ 1.90 ☐ 0.09 ☐ 0.84 ☐ 0.90

La droite de régression de Y par X a pour équation $Y = aX + b$.

Question 13 Donnez une estimation de la pente et de l'ordonnée à l'origine de cette droite de régression :

☐ $a = 1.2$ et $b = -3.12$ ☐ $\hat{a} = 0.047$ et $\hat{b} = 0.80$
☐ $a = 0.047$ et $b = 0.80$ ☐ $\hat{a} = 1.2$ et $\hat{b} = -3.12$

Question 14 Vous semblerait-il pertinent de considérer également la régression de X par Y :

☐ Non car l'âge est une variable indépendante, cela n'a pas de sens de l'expliquer par Y
☐ Oui car l'âge est également une variable aléatoire ici
☐ Non car l'âge est une variable contrôlée par l'expérimentateur, cela n'a pas de sens de l'expliquer par Y

Dans les questions qui suivent, nous testerons la significativité de la liaison entre X et Y par le test du coefficient de corrélation. Nous réaliserons un test bilatéral car nous n'avons qu'une idée imprécise du rôle que pourrait jouer cette hormone durant la phase de croissance.

Question 15 Que sont H_0 et H_1 ?

☐ H_0 : « X et Y sont linéairement indépendantes » et « H_1 : X et Y sont linéairement liées »
☐ H_0 : « X et Y sont linéairement liées » et H_1 :« X et Y sont linéairement indépendantes »
☐ H_0 : « X et Y ne sont pas significativement liées » et « H_1 : X et Y sont significativement liées »
☐ H_0 : « X et Y sont significativement liées » et « H_1 : X et Y ne sont pas significativement liées »

Question 16 Que vaut la statistique du test ?

☐ -0.21 ☐ 1.27 ☐ 0.40 ☐ -1.41

Question 17 Que vaut la valeur seuil de la statistique du test ?

☐ 2.086 ☐ 3.841 ☐ 2.845 ☐ 1.960

Question 18 Que pouvez-vous en conclure ?

☐ H_0 est rejetée avec un risque d'erreur de deuxième espèce inconnu
☐ H_0 est acceptée avec un risque d'erreur de deuxième espèce inconnu
☐ H_0 est rejetée avec un risque d'erreur de première espèce de 5%
☐ H_0 est acceptée avec un risque d'erreur de première espèce de 5%

Question 19 Quelle conclusion biologique vous semble la plus pertinente ?

☐ Les données mettent en évidence un lien significatif entre la concentration hormonale et l'âge
☐ L'âge n'affecte pas la concentration hormonale
☐ Les données ne permettent pas de mettre en évidence une variation de la concentration hormonale avec l'âge selon un modèle linéaire
☐ Les données montrent que l'hormone est indépendante de l'âge

Question 20 Aurait-ce été une bonne idée de réaliser le même test en ne considérant qu'un seul individu mesuré à intervalles de temps réguliers (par exemple toutes les 2 semaines entre 0 et 5 mois) en vérifiant à chaque fois qu'il n'a pas contracté le virus ?

☐ Non car les données n'auraient pas été indépendantes
☐ Non, cela ne pose pas de problème au niveau du test mais l'échantillon aurait été trop petit
☐ Oui, cela ne pose pas de problème au niveau du test et aurait permis de réduire la variabilité de Y liée aux différences individuelles

4 Robustesse à la non représentativité potentielle de notre échantillon

On soupçonne dans les jeux de données analysés précédemment (parties 2 et 3) que les individus échantillonnés ne soient pas représentatifs de leurs populations respectives, c'est-à-dire de leur élevage. Ainsi, sont notamment exclus des analyses les individus détectés comme étant porteurs du virus V.

Question 21 Pourquoi y a-t-il une incertitude sur la condition de représentativité des échantillons ?

- ☐ Parce que les individus prélevés au départ (avant de tester s'ils sont infectés) ne constituent pas un échantillon représentatif de chaque élevage
- ☐ Parce que la proportion d'infectés n'est probablement pas la même dans tous les élevages
- ☐ Parce que tous les individus issus d'un même élevage pourraient ne pas avoir la même probabilité d'être infectés

La question qui se pose donc naturellement est celle de la robustesse des tests réalisés (ANOVA dans la partie 2 et corrélation dans la partie 3) au non-respect de cette condition de représentativité.

Question 22 Que signifierait ici le fait que les tests réalisés soient robustes au non-respect de la condition de représentativité ?

Que, malgré le fait que la procédure d'échantillonnage ne conduise pas à constituer un échantillon représentatif,

- ☐ la probabilité de rejeter H_0 à tort sera différente de α
- ☐ la probabilité de rejeter H_0 à tort sera bien égale à α
- ☐ la probabilité d'accepter H_0 à tort sera bien égale à β
- ☐ la probabilité d'accepter H_0 à tort sera bien égale à α

Le but est ici de faire une proposition intuitive sur la robustesse des tests vus dans les parties 2 et 3. A cette fin, nous envisageons la situation où en réalité, le virus touche préférentiellement les individus ayant une concentration hormonale H plus basse que la moyenne.

Question 23 Concernant l'analyse faite dans la partie 2, supposons donc qu'il n'y ait en réalité aucune différence de concentration hormonale entre les trois élevages considérés. Dans ce cas, vous semblerait-il possible que le fait de ne garder que les individus non infectés dans nos échantillons conduise à augmenter le risque de mettre en évidence un effet « élevage » qui n'existerait pas ?

- ☐ Non, car le fait de ne prendre que les individus infectés ne peut pas biaiser l'estimation de la moyenne
- ☐ Oui, et ce même si la proportion d'individus infectés était la même dans les trois élevages
- ☐ Non, car le fait de ne prendre que les individus infectés peut biaiser l'estimation de la moyenne mais de manière identique dans les trois populations, ne créant pas artificiellement d'écart systématique entre les moyennes estimées
- ☐ Oui, mais uniquement s'il y avait une différence de proportion d'individus infectés entre les trois élevages car il se pourrait que le risque d'être infecté soit lié à la concentration hormonale

Question 24 Concernant la partie 3, supposons qu'il n'y ait en réalité aucun lien entre la concentration hormonale H et l'âge durant la phase de croissance juvénile. Supposons également que les individus, une fois infectés par le virus, conservent celui-ci toute leur vie. Dans ce cas, vous semblerait-il possible que le fait de ne considérer que les individus non infectés dans notre échantillon conduise à augmenter le risque de mettre en évidence un effet « âge » qui n'existerait pas ?

- ☐ Oui, car les individus plus âgés pourraient être plus fréquemment infectés par le virus
- ☐ Non, car notre échantillon est représentatif de la population des individus non infectés et, si il n'y a pas de lien entre Y et X, il n'y en aura pas non plus au niveau de cette sous-population

- Fin du sujet -