E.P.H.E. Évaluation 2 — DEVOIR À LA MAISON – Session Montpellier

UE "Acquisition et traitement statistique de données"- niveau 1

CORRIGÉ

Plein de questions pour commencer…

1. *Si on veut rejeter l'hypothèse nulle pour chacune des différences possibles, on est conduit à faire un…*
   a) *Test bilatéral.*
   b) *Test unilatéral.*
   c) *Test singulier.*
   d) *Test binomial. puisqu'il permet d'étudier les différences entres les valeurs observées et la population*

2. *La covariance sera toujours*
   a) *positive.*
   b) *Plus grande que la variance.*
   c) *Le reflet de la direction de la relation entre les deux variables étudiées.*
   d) *De valeur inférieure à 1.*

3. *Dans l'équation Y=12,6 X + 5,*
   a) *Une différence d'une unité de X conduira à cinq points de différences de la variable dépendante.*
   b) *Y va décroître quand X va croître.*
   c) *La corrélation est obligatoirement significative.*
   d) *Une différence d'une unité de X conduira à 12,6 points de différences de la variable dépendante.*

4. *Dans un exemple hypothétique, la pente d'une droite de régression prédisant Y par X est de – 12. Cela signifie que :*
   a) *Il y a dû y avoir une erreur de calcul.*
   b) *Le coefficient de corrélation des données doit être négatif. Ecart type positif et r = -12*Ecart type*
   c) *L'amplitude du coefficient de corrélation est grande. Je ne comprend pas de qu'est l'amplitude ici.*
   d) *L'interception de la droite à l'ordonnée doit être positive.*

5.Si deux jeux de mesures ont des moyennes semblables mais des variances légèrement différentes, le "t" résultant sera plus proche de :

   a) *1.00*
   b) *3.00*
   c) *0.00*
   d) *C'est impossible à savoir.*

je comprend pas la question mais si c'est le t*obs qui est demandé alors la bonne réponse serai la d puisque nous ne pouvont pas déterminer r²*

5. *Si nous ne pouvons pas rejeter l'hypothèse nulle dans un test t, nous pouvons conclure :*
   a) *Que l'hypothèse nulle est fausse*
   b) *Que l'hypothèse nulle est vraie*
   c) *Que l'hypothèse alternative est fausse*
   d) *Que nous n'avons pas assez d'éléments pour pouvoir rejeter l'hypothèse nulle*

L'hypothèse nulle ne peut pas être fausse ou vrai. On la rejette ou l'accepte faute de « preuve » avec un certain pourcentage d'erreur.

6. *Si 4% de la variabilité de l'espérance de vie est expliquée par la variabilité dans le comportement vis à vis du tabac, alors les valeurs correspondantes de r et r² doivent être respectivement de :*

a) *0,20 et 0,04*
b) *0,04 et 0,16*
c) *0,04 et 0,20*
d) *Il faut plus d'information pour répondre.*

La variabilité de 4% correspond à la variabilité totale. Or pour avoir r^2 il nous faut aussi la variabilité expliquée.

7. *Quand on calcule "s", on divise par n-1 plutôt que par n parce que le résultat est :*
   a) *Plus petit.*
   b) *Plus grand.*
   c) *Une estimation moins biaisée de σ.*
   d) *Plus facile à interpréter*
   e) *Egal à la moyenne de la population.*

n-1 permet de réduir le biais créé par l'utilisation de la moyenne de l'échantillon comme moyenne de population. .

8. *La variance populationnelle est :*
   a) *Une estimation de la variance échantillonnale.*
   b) *Calculée exactement comme la variance échantillonnale.*
   c) *Une estimation biaisée.*
   d) *Le plus souvent une inconnue que l'on cherche à estimer.*

Le calcul de la variance populationnelle est une estimation biaisé puisqu'elle découle d'une observation d'un échantillon et d'une extrapolation et non de toute la population.

9. *Lors d'un stage, un étudiant en psychologie s'interroge sur la manière d'améliorer l'estime de soi d'enfants institutionnalisés. Il interroge au hasard 50 enfants institutionnalisés et leur demande de se coter sur une échelle (métrique), [le chi-carré montre une absence d'ajustement à la normalité].*

Ultérieurement, il pratique avec eux des jeux afin de leur donner confiance en soi et leur demande à nouveau d'évaluer leur estime de soi. L'analyse la plus appropriée pour ce plan est :

   e) *un t de Student pour échantillons appariés*
   f) *un chi-carré d'indépendance*
   g) *un test exact de Fisher*
   h) *un test U de Mann-Withney-Wilcoxon*
   i) *un test W de Wilcoxon*
   j) *une ANOVA simple à groupes indépendants*
   k) *une ANOVA double à groupes indépendants*
   l) *une ANOVA en mesures répétées*
   m) *une ANOVA double mixte (un facteur à groupes indépendants, l'autre en mesures répétées)*

Nous sommes ici dans une comparaison de données de même nature. De plus, on sait que les données ne suivent pas une loi normale, on se retrouve donc à devoir faire un test non paramétrique. Il y a deux échantillons (avant et après les jeux) qui sont appariés (même enfants). On est donc dans le cas où il est préférable de réaliser un test de Wilcoxon pour échantillon appariés.

10. *Pour chacun des exemples suivants, précisez quel test vous utiliseriez ?*
*[Donnez le nom d'un test non-paramétrique ET, **s'il peut être envisagé** (après vérification des prémisses), celui du test paramétrique correspondant.]*

a) *Tester si des malades atteints de la maladie de Parkinson et des malades atteints de la maladie d'Huntington ont des différences dans leurs capacités cognitives (estimées par des résultats de tests notés sur 100). On est en présence de comparaison de données de même nature. Deux échantillons non appariés. On réalise un test de wilcoxon pour échantillon non appariés ou si on obtient une p>0.05 pour le test de normalité de shapiro, et un test d'homoscédasticité, un test t.*

b) *Tester si les individus d'un groupe de patients atteints de la maladie d'Alzheimer sous forme bénigne sont significativement différents de patients gravement atteints quant à leurs résultats de tests de mémoire (pour lesquels nous disposons seulement d'une moyenne de résultats de tests). Il s'agit ici d'un test de comparaison de données à une moyenne. Je ne suis pas sur mais je verrais ici un test t de conformité après avoir fait un test de shapiro ou alors un test de wilcoxon de conformité.*

c) *Tester si un nouveau médicament augmente réellement la capacité de mémoire de patients atteints de la maladie d'Alzheimer en comparant les résultats de tests avant traitement et après deux mois de traitement. On est en présence de comparaison de données de même nature. Les échantillons sont appariés puisque les individus se retrouvent dans les deux échantillons. Pour le test non paramétrique, on peut réaliser un teste de Wilcoxon pour échantillon appariés. Pour le*

test paramétrique et après avec réalisé le test de shapiro, un test t pour échantillon appariés pourra être fait.

d) *Tester si des patients atteints des maladies de Parkinson, d'Huntington et d'Alzheimer ont des fréquences différentes de déficit en mémorisation.* Il s'agit ici de comparer des fréquences et donc des proportions de 3 échantillons. Les échantillons ne sont pas appariés. On peut donc réaliser un test de cochran.

e) *Afin de comparer les revenus des employés d'une entreprise on a relevé les salaires mensuels en milliers de francs d'un échantillon représentatif et normalement distribué de personnes des deux sexes. Vous vous demandez si l'on peut considérer que les salaires sont indépendants du sexe. on désir réaliser ici un test d'indépendance (ou association) sur un échantillon (les employés d'une entreprise). On réalise donc un test de khi².*

f) *Un chercheur a effectué des mesures d'éléments nutritifs en amont et en aval d'une source ponctuelle de pollution sur un cours d'eau. Y-a-t-il des différences entre ses deux séries de mesures ? comparaison de deux échantillons de données appariés. On peut donc réaliser un test de Wilcoxon pour échantillons appariés ou si le test de shapiro montre que les séries suivent une loi normale, un test t.*

11. *Le sous-tableau suivant est un extrait d'un tableau qui présente des mesures de la hauteur (en mm) d'une plante, réalisées dans plusieurs milieux différents [mini = 50 échantillons par milieu]. Un chercheur désire comparer ces données afin de connaître l'effet du milieu sur la taille de cette plante (on admet –sans faire de test de Shapiro - que les données suivent une distribution "Gaussienne").*

Milieu 1	***Milieu 2***	***Milieu 3***	***Milieu 4***	***Milieu 5***
12	*141*	*56*	*87*	*241*
15	*146*	*67*	*105*	*264*
12	*135*	*43*	*79*	*225*
18	*147*	*78*	*123*	*257*
24	*154*	*45*	*114*	*248*
32		*69*		*258*
31				*236*
15				

a. *Quelle analyse permet d'estimer l'effet du milieu sur la hauteur des plantes ? [NE PAS LA FAIRE !]* *une ANOVA.*

b. *Quelles sont les conditions requises pour pouvoir réaliser au mieux cette analyse ?*

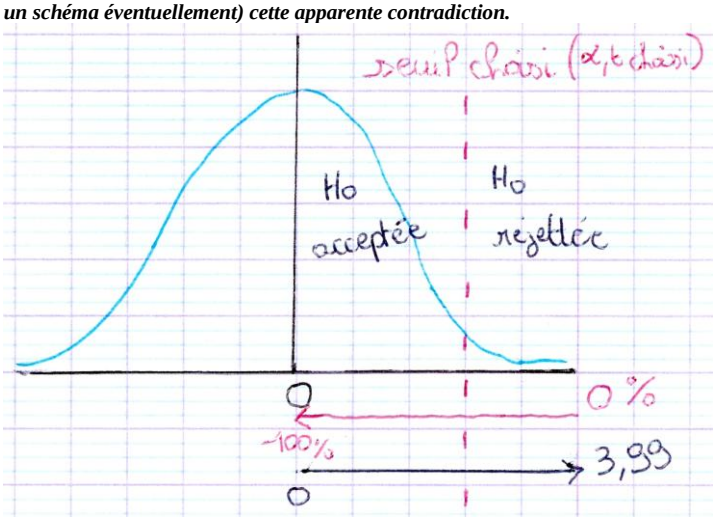
avoir les tailles moyennes des plantes pour chaque milieu.
Avoir éliminer au maximum les facteurs non-contrôlés ou les connaitre pour annihiler leur effets.
Tenir compte des facteurs blocs pour évaluer les résultats.
Essayer d'avoir un plan équilibrés (tous les sous-groupes comportent le même nombre d'individus).

Des questions auxquelles il faut répondre avec précision et brièveté… (ce n'est pas incompatible) pour continuer…

A. ***Après avoir calculé un coefficient de corrélation des rangs de Spearman, un étudiant trouve $r_s = 1,36$. Que peut-il conclure ?*** ***le coefficient de corrélation des rangs de Spearman doit être compris entre -1 et 1. Il y a donc surement eu une erreur de calcul.***

B. ***Après une analyse de variance, si Ho est rejetée, peut-on conclure que toutes les moyennes des populations statistiques comparées sont différentes ?*** Non. H0= les distributions suivent la même loi normale. Rejetter H0 revient à accepter l'hypothèse alternative qui est qu'au moins une distribution à une moyenne qui s'écarte des autres moyennes.

C. ***Dans les tests d'hypothèse :***
Si P(probabilité) statistique > α (seuil choisi), on accepte Ho (hypothèse nulle)
Si P statistique < α, on rejette Ho
Après un test t de Student :
Si la valeur de t observée > t pour le seuil choisi, on rejette Ho
Si la valeur de t observée < t pour le seuil choisi, on accepte Ho
Expliquez simplement (avec un schéma éventuellement) cette apparente contradiction.



Comme le montre le schéma, en rouge se trouve l'échelle sur laquelle placer la P statistique. On voit alors que si P< α, la probabilité se retrouve à droite du seuil, dans la zone où Ho est rejetée. Au contraire, l'échelle correspondant au test t de Student (ici en noir) est inversée par rapport à celle du test d'hypothèse. Pour une valeur observée > au seuil, la valeur se trouve en zone où Ho est rejetée.

D. ***Pour comparer deux jeux de données numériques provenant de populations statistiques indépendantes à distribution Gaussienne, vaut-il mieux utiliser un test t (de Student) de comparaison pour échantillons indépendants ou un test F (de Fisher = ANOVA) ? Justifiez votre réponse.***

Là je ne vois pas du tout. Si vous avez des indices pour me mettre sur la piste. Au début j'aurai dis un test t puisque c'est une comparaison de données de deux échantillons indépendants qui suivent une loi normale. Donc le test de shapiro c'est révélé non significatif et on peux partir sur un test paramétrique. Cependant avant de faire le test t il faudrait faire le test d'homoscédasticité mais l'énoncé n'en parle pas.

E. ***Retrouvez les valeurs manquantes dans ce résumé de résultats d'une ANOVA :***

Source de variation	Somme des carrés	ddl	Moyenne de la somme (CM)	F calculé
Entre groupes	448	4	112	**7**
Erreur (résiduel)	**240**	15	**16**	
TOTAL	688	**19**		

CMr = SCr/(n-k) donc 16= 240/(n-k) soit n-k = 15 donc ddl entre groupe= 19-15= 4

F = CME/CMR donc 7= CME/16 donc CME= 7*16= 112

CME= SCE/k-1 donc 112=SCE/4 donc SCE= 4*112=448

F. ***Les résultats suivants sont obtenus (concernant un test passé par une population d'étudiants) : m = 86,7 avec une variance s^2 de 169.***
Si on tire au hasard un échantillon de 700 étudiants, combien auront une chance d'avoir une note comprise entre 85 et 90 ?

On cherche a déterminer la probabilité d'apparition d'un événement au cours de n=700 expériences identiques et indépendantes. On peut utiliser ici une loi normale réduite ***(je n'en suis pas tout à fait sûr. Un avis sur la question ?)*.**

On pose X= x-m et S l'écart type = Vs^2 = V169 = 13
Soit z1 = X/S = (x-m)/S = (86,7-85)/13 = 0.131
Soit dans la table de Gauss, pour z = 0.131 : 0.0517 soit 0.5- 0.0517 = 0.4483
donc la probabilité qu'on tire au hasard sur 700 tirage une personne dont la note est > à 85 est de 44,88%.
Soit z2 =(90-86.7)/13= 0.254
Soit dans la table de Gauss, pour un z = 0.254 : 0.0987 soit 0.5-0.0987 = 0.4013
donc la probabilité qu'on tire au hasard sur 700 tirage une personne dont la note est > à 90 est de 40,13%%.
Donc la probabilité d'avoir une personne avec une note entre 85 et 90 est de 44,88-40.13 soit 4.75% ***je ne suis pas sûr de cette partie (cette soustraction)***
Donc sur les 700 tirages, 700*0.0475=33.25 auront une chance d'avoir une note comprise entre 85 et 90.

G. Dans un article de revue économique, on trouve le tableau suivant à propos d'une étude cherchant à relier le taux de chômage au niveau d'étude. Cent personnes ont été interrogées.

Niveau d'étude	Chômage	Pas de chômage	Total
Primaire	**43**	**39**	**82**
Secondaire	**27**	**35**	**62**
Supérieur	**10**	**18**	**28**
Total	80	92	**172**

a) *Par quelle méthode statistique pourrait-on étudier ce tableau ? Justifiez votre choix !*

On cherche ici a montrer l'effet de la variable qualitative « étude » sur la variable quantitative chômage. Il s'agis donc ici de tester les différences de variation dans chaque groupe défini par les modalités de la variable explicative s'écartent de manière significative de 0.

On réalise donc un test ANOVA

Ne pourrai t on pas ici réaliser un tableau de contingence ? et faire un test de khi² ?

b) *Quelle(s) critique(s) pourriez-vous faire de ce tableau ?*

Je n'ai pas trouvé de problèmes mathématiques. Suis-je passé au travers de quelque chose ?
Il aurai été bon de noté pour le chômage les périodes. <1 ans, entre 5 et 10 ans …

Deux exercices à faire pour finir…

Tout les calculs ont été réalisés à partir du logiciel R. Je posterai donc dans mes résultats les commandes que j'ai utilisé

Exercice n°1 :

Un laboratoire teste les effets secondaires d'un médicament sur une série de 20 cobayes volontaires. Ce médicament semble altérer la vigilance. On demande donc aux sujets de réagir à un stimulus visuel en appuyant sur un bouton. On note les temps de réaction avant et après prise de médicament. Lors du test sous médicament, il apparaît que certains sujets sont somnolents et il arrive donc qu'ils ne perçoivent pas le stimulus. On décide donc d'introduire dans l'expérience une valeur maximale fixée arbitrairement à 10 pour les sujets qui n'ont pas réagi avant 10 secondes.
Les valeurs sont les suivantes :

Avant : 0.35, 0.31, 0.12, 0.07, 0.17, 0.34, 0.32, 0.25, 0.05, 0.16, 0.18, 0.12, 0.14, 0.27, 0.21, 0.15, 0.13, 0.37, 0.30, 0.24
Après : 10, 0.50, 0.22, 0.13, 10, 0.45, 10, 0.20, 0.45, 0.16, 10, 0.28, 0.12, 0.20, 0.25, 0.18, 0.31, 0.24, 0.22, 10

a) Peut-on dire que le médicament a un effet sur la vigilance ? Justifiez le choix de votre test.

Ho : il n'y pas d'effet sur la vigilance lors de l'utilisation de ce médicament.
Nous sommes ici dans le cas d'un test de comparaison de donnée de même nature. La première étape est donc de déterminer si les échantillons suivent une loi normale.
avant<-c(0.35, 0.31, 0.12, 0.07, 0.17, 0.34, 0.32, 0.25, 0.05, 0.16, 0.18, 0.12, 0.14, 0.27, 0.21, 0.15, 0.13, 0.37, 0.30, 0.24)
apres<- c(10, 0.50, 0.22, 0.13, 10, 0.45, 10, 0.20, 0.45, 0.16, 10, 0.28, 0.12, 0.20, 0.25, 0.18, 0.31, 0.24, 0.22, 10)
shapiro.test(avant) p-value = 0.3772 >0.05 donc le premier échantillon suit une loi normale
shapiro.test (apres) p-value = 1.234e-06< 0.05 donc le second échantillon ne suit pas une loi normale.
On se dirige donc sur un test non-paramétrique. Les deux échantillons étant appariés, nous réalisons le test de Wilcoxon pour échantillons appariés.
wilcox.test(avant,apres,paired=T) p-value = 0.007448
le risque de se tromper en rejetant l'hypothèse H0 est de 0,7%. Les valeurs diffèrent de manière hautement significative. Donc oui, le médicament à un effet sur la vigilance.

b) Aboutirait-on aux mêmes conclusions en excluant tout simplement les sujets sensibles à l'endormissement ?
Justifiez le choix de votre test et commentez les résultats !

Ho : il n'y pas d'effet sur la vigilance lors de l'utilisation de ce médicament.
Nous sommes toujours ici dans le cas d'un test de comparaison de donnée de même nature. La première étape est donc de déterminer si les échantillons suivent une loi normale une fois les sujet sensibles à l'endormissement enlevés.
avant2<-c(0.31, 0.12, 0.07, 0.34, 0.25, 0.05, 0.16, 0.12, 0.14, 0.27, 0.21, 0.15, 0.13, 0.37, 0.30)
apres2<- c(0.50, 0.22, 0.13, 0.45, 0.20, 0.45, 0.16, 0.28, 0.12, 0.20, 0.25, 0.18, 0.31, 0.24, 0.22)
shapiro.test (avant2) p-value = 0.3848 >0.05 donc le premier échantillon suit une loi normale
shapiro.test (apres2) p-value = 0.03815 < 0.05 donc le second échantillon ne suit pas une loi normale.
On se dirige donc sur un test non-paramétrique. Les deux échantillons étant appariés, nous réalisons le test de Wilcoxon pour échantillons appariés.
wilcox.test(avant2,apres2,paired=T) p-value = 0.1319
le risque de se tromper en rejetant l'hypothèse H0 est de 13,19%. On accepte donc Ho. Le médicament n'a alors pas d'effet sur la vigilance.

Commentaires : en enlevant les personnes sensibles à l'endormissement, le résultat change complètement puisqu'on passe d'un rejet hautement significatif à une acceptation de l'hypothèse nulle. L'élimination des valeurs extrêmes peut se justifier dans certain cas mais ici elles représentent ¼ de l'échantillon (5 cas sur les 20). Les éliminés reviendrait donc à fausser le résultat scientifique et ne serait pas pertinent.

Exercice n°2 :

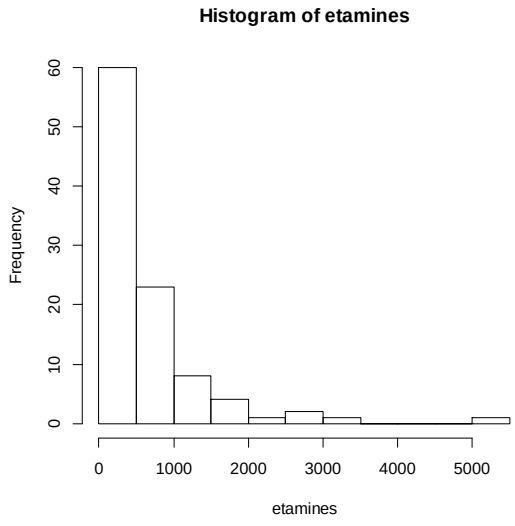
Chez les Angiospermes l'apparition de la fleur, il y a 130 millions d'années, est considérée comme une innovation clé. Le plan d'organisation de la fleur est conservé chez presque tous les Angiospermes. En revanche il existe une incroyable diversité de forme, de couleur et de nombre de pièces au sein de chaque verticille. C'est le cas notamment pour les étamines dont le nombre peut varier de quelques unités (oligandrie) à plusieurs dizaines (icosandrie) voire plusieurs centaines (polyandrie). Exemple pour 100 fleurs dans le tableau suivant :

3	108	203	259	304	422	518	646	865	1483
7	129	205	262	338	423	520	648	888	1567
8	144	215	262	339	425	541	666	995	1591
9	144	216	264	344	437	542	672	1095	1627
17	158	218	271	344	439	569	717	1105	1827
18	173	222	273	351	453	575	732	1262	2110
21	182	224	275	364	461	581	740	1306	2566
23	186	226	276	366	470	603	745	1412	2881
35	187	243	279	403	485	609	803	1469	3277
42	203	253	293	410	499	646	824	1470	5481

Pour "R" : (3, 7, 8, 9, 17, 18, 21, 23, 35, 42, 108, 129, 144, 144, 158, 173, 182, 186, 187, 203, 203, 205, 215, 216, 218, 222, 224, 226, 243, 253, 259, 262, 262, 264, 271, 273, 275, 276, 279, 293, 304, 338, 339, 344, 344, 351, 364, 366, 403, 410, 422, 423, 425, 437, 439, 453, 461, 470, 485, 499, 518, 520, 541, 542, 569, 575, 581, 603, 609, 646, 646, 648, 666, 672, 717, 732, 740, 745, 803, 824, 865, 888, 995, 1095, 1105, 1262, 1306, 1412, 1469, 1470, 1483, 1567, 1591, 1627, 1827, 2110, 2566, 2881, 3277, 5481)

a) En partageant les données en 11 intervalles, tracer l'histogramme de cette série de données. [en raison du grand nombre de données et de modalités, on peut assimiler ces données quantitatives discontinues à des valeurs continues et tracer un histogramme]

```
hist(etamines,br=c(0,500,1000,1500,2000,2500,3000,3500,4000,4500,5000,5500))
```



b) Représentez la même chose avec un diagramme en bâtons (barplot).

Rem. : Faites-le en vous aidant de l'argument plot = F (*Cfr* documents de TD de R)

BON COURAGE !