EXPERIMENTATION ET ANALYSE DES DONNEES – JANVIER 2013

Question 1

Présentez et comparez les tests F de Fisher d'une ANOVA à 1 facteur et d'une régression linéaire: détaillez dans les deux cas les hypothèses du test, ainsi que la manière dont la statistique F est calculée. A l'aide d'un schéma, précisez dans quelles conditions l'on peut rejeter l'hypothèse nulle.

Question 2

Qu'est-ce que le test de Wilcoxon-Mann-Withney et en quoi se différentie-t-il de son équivalent non paramétrique ? Un étudiant a étudié la présence de coliformes fécaux dans 10 stations localisées en aval d'une unité d'épuration avant et après son installation. Il a obtenu les valeurs suivantes (le S1 dénote la Station 1, etc...):

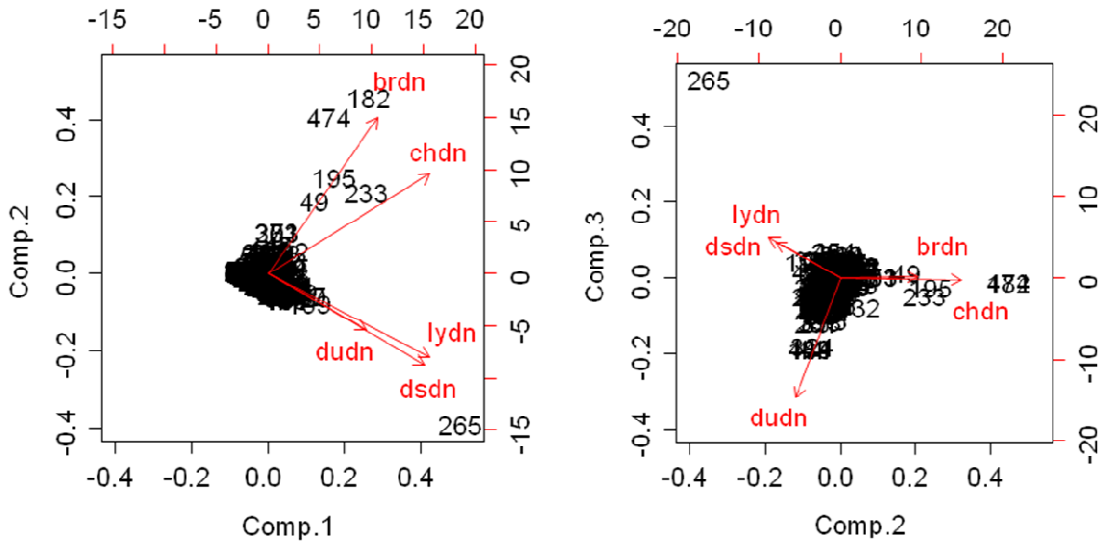
Avant installation:	S1: 123	S2: 899	S3: 67	S4: 98	S5: 44	S6: 133
Après installation:	S1: 10	S2: 349	S3: 82	S4: 59	S5: 44	S6: 92

L'étudiant se demande si l'installation de l'unité d'épuration a eu un effet sur le nombre de coliformes. Faut-il utiliser un test apparié ou non-apparié ? Selon le cas retenu, calculez la valeur de la statistique. Sachant que les valeurs limites pour une statistique W pour $n = 6$ sont pour p = 0.001, p = 0.01, p = 0.05 et p = 0.1 respectivement égales à 28, 27, 24 et 22. Quelles conclusions pouvez-vous tirer de cette analyse ?

Question 3

Un recensement de la volaille a été fait dans un des Etats de l'Inde, et le nombre d'individus de différents types de volaille a été compté dans chaque sous-district (sous division administrative qui correspond à nos communes). Les variables chdn, brdn, dsdn, dudn, lydn désignent les différents types de volaille recensés. On a effectué une analyse en composantes principales et obtenu les résultats suivants:

```
> myPCA = princomp(~ chdn + brdn + dsdn + dudn + lydn, cor=TRUE, data = myD)
> summary(myPCA)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4      Comp.5
Standard deviation     1.6513929 1.1976310 0.8790306 0.24150090 0.086971241
Proportion of Variance 0.5454197 0.2868640 0.1545390 0.01166454 0.001512799
Cumulative Proportion  0.5454197 0.8322837 0.9868227 0.99848720 1.000000000
```



Détaillez les différents éléments fournis en résultats. Que pouvez-vous conclure de cette analyse ? Toutes les variables sont elles indépendants ou certains vous paraissent-elle redondantes ? Si vous deviez faire des recommandations en vue de diminuer le coût du recensement, quelle(s) catégorie(s) de volaille pourrai(en)t ne pas être recensée(s) ? Certaines observations vous semblent-elles suspectes ? Pourquoi ?

Question 4

Un chercheur s'intéresse aux facteurs permettant de prédire la densité de canards (dudnlg; en échelle logarithmique) et la densité de poulets de basse-cour (dsdnlf; en échelle logarithmique) dans chaque sous-district en fonction de la densité de route (rddn), de la densité de population humaine (hpdnlg, échelle logarithmique), et du nombre moyen de récoltes de céréales par an (ncrop). Il a effectué deux régressions linéaires, et a obtenu les résultats suivants:

```
Call:
lm(formula = dsdnlg ~ rddn + hpdnlg + ncrop, data = myD)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.050e-01  1.220e-01   4.141 4.03e-05 ***
rddn        2.151e-07  2.487e-07   0.865    0.388
hpdnlg      5.858e-01  4.502e-02  13.011  < 2e-16 ***
ncrop       5.979e-02  4.237e-02   1.411    0.159
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3101 on 523 degrees of freedom
Multiple R-squared: 0.2598,   Adjusted R-squared: 0.2556
F-statistic: 61.19 on 3 and 523 DF,  p-value: < 2.2e-16


Call:
lm(formula = dudnlg ~ rddn + hpdnlg + ncrop, data = myD)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.423e-01  1.756e-01  -1.949   0.0518 .
rddn        -7.537e-07  3.581e-07  -2.104   0.0358 *
hpdnlg       8.840e-01  6.481e-02  13.639   <2e-16 ***
ncrop        1.309e-01  6.100e-02   2.145   0.0324 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4465 on 523 degrees of freedom
Multiple R-squared: 0.3094,   Adjusted R-squared: 0.3054
F-statistic:  78.1 on 3 and 523 DF,  p-value: < 2.2e-16
```

Que peut-on conclure de ces analyses ? Ces deux modèles sont ils comparables ? Justifiez votre réponse. Quelles analyses supplémentaires pourrait-on effectuer pour s'assurer que ces résultats sont fiables ?