Introduction à R pour la recherche biomédicale
Exercices

Le langage R

1. Créer un vecteur contenant 15 fois le terme "toto" et 10 fois le terme "titi". Convertir ce vecteur en facteur, avec "titi" comme premier niveau.
2. Créer une matrice $\mathcal{M}$ telle que $\mathcal{M} = \begin{pmatrix} 6 & 7 & 2 \\ 1 & 5 & 9 \\ 8 & 3 & 4 \end{pmatrix}$. Vérifier que toutes les sommes par ligne et par colonnes sont égales. (Pour plus d'informations, `http://bit.ly/sxYUHk`.)
3. Simuler 20 lancers d'une pièce équilibrée ($p = 0.5$). Répéter l'expérience avec k lancers, `k=seq(10, 10000, by=100)`. Calculer à chaque fois la proportion de piles.
4. Créer un jeu de données fictifs composées de deux colonnes d'aléas gaussiens (de moyenne 10 et 12) et d'un facteur équilibré sur le nombre de mesures. Exporter le au format `csv`.
5. À partir des mesures de tailles issus des élèves de deux classes,

   ```
   d <- data.frame(height=rnorm(40, 170, 10),
                   class=sample(LETTERS[1:2], 40, rep=T))
   d$height[sample(1:40,1)] <- 220
   ```

   indiquer si la personne la plus grande provient de la classe A ou B.
6. Créer un facteur traitement, `tx`, de longeur 60, dont les niveaux sont équilibrés (30 `std` et 30 `new`) et organisés selon le motif suivant :

   ```
   std std std new new new ... std std std new new new
   ```

   Remplacer le label `std` par `old`. Faire en sorte que la séquence de labels débute par `new` et non `old` (i.e., intervertir les deux labels). Enfin, mélanger aléatoirement les labels.
7. Charger le fichier `lungcancer.txt`. Celui-ci contient des problèmes de codage (format des variables, valeurs manquantes, etc.). Inspecter les données (e.g., avec `summary()`) et recoder en conséquence.
8. Charger le fichier `adl.sav` (ne pas oublier d'utiliser la library `foreign`), vérifier la présence éventuelle de données manquantes Indiquer le nombre de cas par groupe de traitement (`group`).
9. Charger le fichier `anorectic.sav`. Indiquer le nombre de cas, le nombre de données complètes par sujet*, le nombre total de sujets AN (`Anorexia Nervosa`) et leur score moyen `mood`. * Attention à bien recoder le facteur `time` en conséquence.
10. Charger le fichier `birthwt` du package `MASS`. Indiquer quelles sont les variables assimilables à des facteurs et la distribution de leur niveaux selon le status du poids à la naissance (`low`). Calculer l'âge médian des mères des bébés dont le poids est inférieur à 2,5 kg à la naissance.

Analyse exploratoire

11. Charger les données `cereal.csv`. Afficher l'ensemble des distributions uni- et bivariées.
12. À partir du fichier `anorectic.sav`, produire un résumé numérique des scores moyens (variables 1 à 16), avec leurs écarts-types, *à la baseline* (`time=1`, t_1).
13. Avec les données `anorectic.sav`, toujours à t_1, calculer l'ensemble des coefficients de corrélation de Spearman des scores cités ci-dessus, et indiquer quels sont ceux * qui sont > 0.3 en valeurs absolues. * On peut extraire la partie inférieure d'une matrice de corrélation avec `lower.tri`.
14. Représenter graphiquement l'évolution des 16 scores moyens au cours du temps, à l'aide d'un diagramme de dispersion de type `xyplot`.
15. Avec le fichier de données `adl.sav`, effectuer un tri à plat de toutes les variables binaires (`diabetic` à `psd`) par groupe de traitement. Utiliser un `dotplot` pour résumer les scores moyens sur les variables `travel` à `housekpg`, toujours par groupe de traitement.
16. Avec les données `cereal.csv`, transformer `company` en un facteur à trois modalités : `g mills`, `kellogs`, et `other`. Utiliser un diagramme parallèle (`parallel`) pour afficher toutes les variables numériques en mettant en évidence les trois principales compagnies (voir l'argument `group=`). Discutez les résultats.
17. Le graphique réalisé à l'exercice 14 rend difficile les comparaisons entre certains items et la légende est peu lisible. Proposer une solution alternative reposant sur un `dotplot` avec les labels ordonnés par scores croissants.
18. Résumer la distribution des âges dans le fichier `adl.sav`, par groupe de traitement. Afficher les fonctions de répartition des âges par groupe de traitement, dans le même graphique. Faire de même avec un graphique en quantile (QQ plot, en considérant une loi normale).
19. Charger les données `anorexia` disponible dans le package `MASS`. Représenter graphiquement le poids des patientes, en kg, avant et après traitement, selon le groupe de traitement. Résumer numériquement la distribution des différences `Postwt-Prewt`, pour chaque groupe de traitement.
20. Toujours avec les données `MASS::anorexia`, afficher un diagramme de dispersion de l'ensemble des données (`Postwt` *vs.* `Prewt`), en mettant en évidence les différents groupes de traitement (couleur et/ou symbole) et en superposant une courbe de type `lowess`. Comparer avec des droites de régressions spécifiques à chaque groupe. Quelle conclusion en tirer ?

Mesures et tests d'association

21. À partir des données `sleep`, comparer et discuter les résultats d'un test t pour échantillons appariés *vs.* ceux d'un test pour échantillons indépendants.
22. Avec les mêmes données, calculer la moyenne `di` des scores de différences. Recalculer cette moyenne des différences 499 fois en rééchantillonnant (avec remise) les 10 individus. Afficher la distribution des différences ainsi obtenues. Entre quelles valeurs se situent 95 % des résultats ? * Attention à bien rééchantillonner les individus et pas le tableau entier. Il est probablement plus facile de réarranger les données sous forme d'un tableau avec 10 lignes et 2 colonnes.
23. Avec les données `anorectic.sav`, comparer les scores `mood` à t_1 et t_4 chez les patientes `Anorexia Nervosa` à l'aide d'un test de Wilcoxon. Ce type de test est-il justifié ?
24. Avec les mêmes données, à t_1, tester la corrélation (Spearman) entre `weight`, `binge`, `vomit` et `purge`. Les résultats restent-ils significatifs lorsque l'on multiplie les p-valeurs par 6 ? Faire de même à t_4.
25. En utilisant les données `MASS::anorexia`, comparer les poids des patientes des trois groupes à l'inclusion. Peut-on mettre en évidence une différence significative ?
26. Charger les données `PAINT.DAT` et afficher une matrice de dispersion de l'ensemble des variables. La corrélation entre `LYMPHO` et `WBC` est-elle significativement différente de zéro ? Cette statistique de test est-elle influencée par les "points les plus extérieurs" (ceux définissant l'enveloppe) de la distribution jointe de `LYMPHO` et `WBC` (utiliser `chull`) ? Comment pourrait-on tester si celle-ci est supérieure à 0.5 ? Quelle est la corrélation partielle entre ces deux mêmes variables lorsque l'on ajuste sur `HAEMO` (utiliser `pcor()` dans le package `ppcor`) ?
27. Charger les données `blood.txt` (cf. fichier `blood.r` dans le répertoire `pub/` sur le site). Aggréger les données sur le sexe et réaliser un test du χ^2 pour tester l'indépendence entre le groupe sanguin de la mère et du père. Que suggère le résultat du test ?
28. Considérons le tableau de données suivant :

	AA	GA	GG
BMI < 25	30	246	380
BMI > 25	30	130	184

    Il s'agit de la distribution des génotypes observés sur deux groupes d'individus constitués selon leur indice de masse corporelle. Ces données de génotypage proviennent du SNP (polymorphisme de séquence) `rs1042717`. Existe-t-il une relation entre ces deux variables ? En considérant le génotype comme une variable ordinale, où l'on code la fréquence de l'allèle mineur G, calculer la statistique $M^2 = n \cdot r^2$, où r^2 est le coefficient de corrélation de Bravais-Pearson et n le nombre total d'observations. Reporter la p-valeur associée au quantile observé pour une loi $\chi^2(1)$ (voir `pchisq`).
29. Avec les données `adl.sav`, comparer la durée moyenne de séjour (`los`) entre les deux groupes de traitement. Permuter aléatoirement les labels du facteur de groupe et recalculer cette différence de moyennes, 999 fois. Quelle est la proportion de cas où la différence de moyennes sous permutation est au moins aussi extrême que la différence observée sur l'échantillon original ?
30. Soit la fonction ci-dessous, qui génère deux échantillons aléatoires tirés d'une loi normale (d'écart-type `sd`) en introduisant une différence de moyennes `dm` et renvoit la p-valeur associée à un test de Student.

```
sim.data <- function(n, alloc.ratio=1, dm=.2, sd=c(1,1),
                     verbose=FALSE, ...) {
  n1 <- n2 <- n
  if (alloc.ratio != 1) n2 <- n1*alloc.ratio
  x <- c(rnorm(n1, 0, sd[1]), rnorm(n2, 0+dm, sd[2]))
  grp <- factor(rep(c(1:2), c(n1,n2)))
  out <- t.test(x ~ grp, ...)
  if (verbose) print(out)
  invisible(out$p.value)
}
```

    Un exemple d'utilisation serait : `sapply(10:20, sim.data)` pour avoir les p-valeurs d'un test t lorsque `n` varie entre 10 et 20 ; ou bien la même chose en fixant la différence de moyenne à zéro, `sapply(10:20, sim.data, dm=0)`. Calculer la fréquence de rejet de l'hypothèse nulle d'absence de différence entre les deux groupes, en variant les différents paramètres (essentiellement, `n` et `dm`). Quelles conclusions peut-on tirer de ces simulations ?

Modèle linéaire et applications

31. Avec les données `birthwt` du package `MASS`, effectuer une régression linéaire en considérant le poids des bébés (`bwt`) comme réponse et le poids des mères (`lwt`) comme prédicteur, en considérant les unités de mesure d'origine. Faire de même en considérant les poids des bébés et des mères en *kg*, puis en centrant les poids des mères sur leur moyenne. Les résultats des tests statistiques sur les coefficients changent-ils ? Quelle est l'interprétation des coefficients de régression dans chacun des trois cas ?
32. Décrire la distribution des résidus du modèle précédent. Identifier les observations susceptibles d'influencer les paramètres estimés à partir du modèle de régression, s'il en existe. Dans ce dernier cas, ré-estimer les paramètres du modèle sans ces observations.
33. Avec les mêmes données, régresser l'ethnicité des mères sur le poids des bébés (en *g*), sans oublier de recoder la variable `race` en facteur. Commenter les résultats. Refaire l'analyse après avoir modifié le codage des contrastes comme suit :

```
options(contrasts=c("contr.sum", "contr.poly"))
```

    (Pour comprendre comment R code les contrastes, comparer `contr.sum(3)` à `contr.treatment(3)`, ces derniers étant utilisés par défaut.)
    Quelle est l'interprétation des coefficients de régression dans les deux cas ?
34. Toujours avec les données `birthwt` et le poids des bébés en *g* comme variable réponse, effectuer une ANOVA à deux facteurs, `race` et `smoke`, incluant l'interaction `race:smoke`. L'interaction est-elle significative ? Si non, décrire les effets

principaux.

35. Les conditions d'application du modèle précédent sont-elles vérifiées ? Examiner la distribution des résidus.
36. Importer les données `hersdata.txt`.[1] Estimer les coefficients de régression du modèle suivant :
    `glucose ~ exercice + age + drinkany + BMI`,
    pour les patients non diabétiques (`diabetes==0`). Comparer avec un modèle n'incluant que la variable `exercice`. Effectuer une analyse des résidus du modèle initial.
37. À partir du modèle de régression multiple, prédire le taux de glucose (*mg/dL*) pour un patient ne consommant pas d'alcool, faisant de l'exercice, ayant un BMI de 22 et d'âge médian. Comparer avec la prédiction faite pour un patient ayant les mêmes caractéristiques mais faisant de l'exercice (`exercise==1`). Représenter graphiquement les prédictions concernant le taux de glucose en fonction de `age` et `BMI` (surface de réponse).
38. Importer les données `CYSTIC.DAT`.[2] Décrire graphiquement l'ensemble des données. Considérant le modèle
    `PEmax ~ Weight + BMP + FEV + RV`,
    estimer les coefficients de régression, identifier les observations exerçant un effet levier s'il en existe, et discuter la qualité d'ajustement du modèle.
39. Ajouter `Age` et `Sex` comme prédicteurs dans le modèle précédent. Cela change-t-il l'interprétation des résultats. Si oui, pourquoi ?
40. Avec les données `babies.txt`*, effectuer une ANOVA sur mesures répétées en considérant le logarithme de la force d'inspiration (`inspirat`) comme variable dépendante et la pression maximale (`maxp`) comme variable indépendante. Comparer avec une approche par modèle de régression avec effet aléatoire. * Attention aux valeurs manquantes !

[1] Contexte de l'étude : Un taux de glucose > 125 *mg/dL* signe le diabète, alors qu'un taux situé entre 100 et 125 *mg/dL* est un facteur de risque. On s'intéresse à l'effet protecteur de l'exercice physique chez les sujets à risque (voir le "codebook" pour une description plus complète des données, `http://bit.ly/KY5ReI`).

[2] Contexte de l'étude : 25 patients (`sex=0` pour les hommes) avec fibrose kystique pour lesquels on dispose d'une mesure de malnutrition (`PEmax`) et d'un ensemble de variables explicatives (données anthropométriques et fonction pulmonaire).