BIO8940 Advanced stats and Open Science

Work in Progress

Julien Martin

16-02-2021

Table des Matières

Note

Work in progress. New chapters are going to appears regularly meaning that if you download the pdf it might be incomplete by the time we do the practical in class.

Préface

Les exercices de laboratoire que vous retrouverez dans les pages qui suivent sont conçus de manière à vous permettre de développer une expérience pratique en analyse de données à l'aide d'un logiciel (R). R est un logiciel très puissant, mais comme tous les logiciels, il a des limites. En particulier il ne peut réfléchir à votre place, vous dire si l'analyse que vous tentez d'effectuer est appropriée ou sensée, ou interpréter biologiquement les résultats.

Quelques points importants à retenir

- Avant de commencer une analyse statistique, il faut d'abord vous familiariser son fonctionnement. Cela ne veut pas dire que vous devez connaître les outils mathématiques qui la sous-tendent, mais vous devriez au moins comprendre les principes utilisés lors de cette analyse. Avant de faire un exercice de laboratoire, lisez donc la section correspondante dans les notes de cours. Sans cette lecture préalable, il est très probable que les résultats produits par le logiciel, même si l'analyse a été effectuée correctement, seront indéchiffrables.

- Les laboratoires sont conçus pour compléter les cours théoriques et vice versa. À cause des contraintes d'horaires, il se pourrait que le cours et le laboratoire ne soient pas parfaitement synchronisés. N'hésitez donc pas à poser des questions sur le labo en classe ou des questions théoriques au laboratoire.

- Travaillez sur les exercices de laboratoire à votre propre rythme. Certains exercices prennent beaucoup moins de temps que d'autres et il n'est pas nécessaire de compléter un exercice par séance de laboratoire. En fait deux séances de laboratoire sont prévues pour certains des exercices. Même si vous n'êtes pas notés sur les exercices de laboratoire, soyez conscient que ces exercices sont essentiels. Si vous ne les faites pas, il est très peu probable que vous serez capable de compléter les devoirs et le projet de session. Prenez donc ces exercices de laboratoire au sérieux !

- Les 2 premier laboratoires sont conçu pour vous permettre d'acquérir ou de réviser le minimum de connaissances requises pour vous permettre de réaliser les exercices de laboratoires avec R. Il y a presque toujours de multiples façons de faire les choses avec R et vous ne trouverez ici que des méthodes simples. Ceux et celles d'entre vous qui y sont enclins pourront trouver en ligne des instructions plus détaillées et complexes. En particulier, je vous conseille :

  - R pour les débutants http://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf

- An introduction to R http://cran.r-project.org/doc/manuals/R-intro.html
- Si vous préférez des manuels, le site web de CRAN en garde une liste commentée à : http://www.r-project.org/doc/bib/R-books.html
- Une liste impressionnante de très bon livre sur R https://www.bigbookofr.com/
- Finalement, comme aide-mémoire à garder sous la main, je vous recommande R reference card par Tom Short http://cran.r-project.org/doc/contrib/Short-refcard.pdf

Qu'est-ce que R et pourquoi l'utiliser dans ce cours?

R est un logiciel libre et multi-plateforme formant un système statistique et graphique. R est également un langage de programmation spécialisé pour les statistiques.

R a deux très grands avantages pour ce cours, et un inconvénient embêtant initialement mais qui vous forcera à acquérir des excellentes habitudes de travail. Le premier avantage est que vous pouvez tous l'installer sur votre (ou vos) ordinateurs personnel gratuitement. C'est important parce que c'est à l'usage que vous apprendrez et maîtriserez réellement les biostatistiques et cela implique que vous devez avoir un accès facile et illimité à un logiciel statistique. Le deuxième avantage est que R peut tout faire en statistiques. R est conçu pour être extensible et est devenu l'outil de prédilection des statisticiens mondialement. La question n'est plus : " Est-ce que R peut faire ceci? ", mais devient" Comment faire ceci avec R ". Et la recherche internet est votre ami. Aucun autre logiciel n'offre ces deux avantages.

L'inconvénient embêtant initialement est que l'on doit opérer R en tapant des instructions (ou en copiant des sections de code) plutôt qu'en utilisant des menus et en cliquant sur différentes options. Si on ne sait pas quelle commande taper, rien ne se passe. Ce n'est donc pas facile d'utilisation à priori. Cependant, il est possible d'apprendre rapidement à faire certaines des opérations de base (ouvrir un fichier de données, faire un graphique pour examiner ces données, effectuer un test statistique simple). Et une fois que l'on comprend le principe de la chose, on peut assez facilement trouver sur le web des exemples d'analyses ou de graphiques plus complexes et adapter le code à nos propres besoins. C'est ce que vous ferez dans le premier laboratoire pour vous familiariser avec R.

Pourquoi cet inconvénient est-il d'une certaine façon un avantage? Parce que vous allez sauver du temps en fin de compte. Garanti. Croyez-moi, on ne fait jamais une analyse une seule fois. En cours de route, on découvre des erreurs d'entrée de données, ou que l'on doit faire l'analyse séparément pour des sous-groupes, ou on obtient des données supplémentaires, ou on fait une erreur. On doit alors recommencer l'analyse. Avec une interface graphique et des menus, cela implique recommencer à cliquer ici, entre des paramètres dans des boîtes et sélectionner des boutons. Chaque fois avec possibilité d'erreur. Avec une série de commandes écrites, il suffit de corriger ce qui doit l'être puis de copier-coller l'ensemble pour répéter instantanément. Et vous avez la possibilité de parfaitement documenter ce que vous avez fait. C'est comme cela que les professionnels travaillent et offrent une assurance de qualité de leurs résultats.

Installation des logiciels nécessaires

R

Pour installer R sur un nouvel ordinateur, allez au site http://cran.r-project.org/. Vous y trouverez des versions compilées (binaries) ou non (sources) pour votre système d'exploitation de prédilection (Windows, MacOS, Linux).

Note : R a déjà été installé sur les ordinateurs du laboratoire (la version pourrait être un peu plus ancienne, mais cela devrait être sans conséquences).

0.0.1 Text editor or IDE

Tinn-r Atom sublime, emacs, vim

Rstudio

RStudio est un environnement de développement intégré (IDE) créé spécifiquement pour travailler avec R. Sa popularité connaît une progression foudroyante depuis 2014. Il permet de consulter dans une interface conviviale ses fichiers de script, la ligne de commande R, les rubriques d'aide, les graphiques, etc.

RStudio est disponible à l'identique pour les plateformes Windows, OS X et Linux. Pour une utilisation locale sur son poste de travail, on installera la version libre (Open Source) de RStudio Desktop depuis le site https://www.rstudio.com/products/rstudio/download/

Visual Studio Code

Tinn-r

Paquets pour R

- Rmarkdown
- tinytex

Ces 2 paquets devrait être installé automatiquement avec RStudio, mais pas toujours. Je vous recommande donc de les installer manuellement. Pour ce faire, simplement copier-coller le texte suivant dans le terminal R.

```
install.packages(c("rmarkdown", "tinytex"))
```

pandoc

laTex

- tinytex or others

Instructions générales pour les laboratoires

- Apporter une clé USB ou son équivalent à chaque séance de laboratoire pour sauvegarder votre travail.
- Lire l'exercice de laboratoire AVANT la séance, lire le code R correspondant et préparer vos questions sur le code.
- Durant les pré-labs, écouter les instructions et posez vos questions au moment approprié.
- Faites les exercices du manuel de laboratoire à votre rythme, en équipe, puis je vous recommande de commencer (compléter?) le devoir. Profitez de la présence du démonstrateur et du prof…
- Pendant vos analyses, copiez-collez des fragments de sorties de R dans un document (par exemple dans votre traitement de texte favori) et annotez abondamment.
- Ne tapez pas directement vos commandes dans R mais plutôt dans un script. Vous pourrez ainsi refaire le labo instantanément, récupérer des fragments de code, ou plus facilement identifier les erreurs dans vos analyses.
- Créez votre propre librairie de fragments de codes (snippets). Annotez-là abondamment. Vous vous en féliciterez plus tard.

Notes sur le manuel

Vous trouverez dans le manuel des explications sur la théorie, du code R, des explications sur R et des exercises.

Le manuel essaie aussi de mettre en évidence le texte de différentes manières.

Avec des sections à vous de jouer, ui indique un exercise à faire, idéalement sans regarder la solution qui se trouve plus bas.

des avertissements

des avertissements

des points importants

des notes

et des conseils

Resources {-}

Ce document est généré par l'excellente extension bookdown de Yihui Xie. Il est basé sur le précédent manuel de laboratoire *BIO4558 manuel de laboratoire* par Antoine Morin. L'introduction à R est largement reprise de l'excellent manuel de **Julien Barnier** intitulé *Introduction à R et au tidyverse*

Licence

Ce document est mis à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions 4.0 International.

Figure 1: Licence Creative Commons

Partie I

Open Science

Chapitre 1

Introduction to open Science

Chapitre 2

Introduction to Rmarkdown

2.1 Practical

2.1.1 Context

Let's apply what we have learnt in the course on **Writing dynamic and reproducible documents - An introduction to R Markdown**

We will use the awesome `palmerpenguins` dataset , an alternative to Fisher's `iris` dataset, to explore and visualize data.

These data have been collected and shared by Dr. Kristen Gorman and Palmer Station, Antarctica LTER.

The package was built by Drs Allison Horst and Alison Hill, check out the official website.

The package `palmerpenguins` has two datasets.

```
library(palmerpenguins)
```

The dataset `penguins` is a simplified version of the raw data; see `?penguins` for more info:

```
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex
##   <fct>   <fct>           <dbl>         <dbl>            <int>       <int> <fct>
## 1 Adelie  Torge~           39.1          18.7              181        3750 male
## 2 Adelie  Torge~           39.5          17.4              186        3800 fema~
## 3 Adelie  Torge~           40.3          18                195        3250 fema~
## 4 Adelie  Torge~           NA            NA                 NA          NA <NA>
## 5 Adelie  Torge~           36.7          19.3              193        3450 fema~
```

```
## 6 Adelie  Torge~           39.3          20.6             190        3650 male
## # ... with 1 more variable: year <int>
```

The other dataset `penguins_raw` has the raw data; see `?penguins_raw` for more info:

```
head(penguins_raw)
```

```
## # A tibble: 6 x 17
##   studyName `Sample Number` Species Region Island Stage `Individual ID`
##   <chr>               <dbl> <chr>   <chr>  <chr>  <chr> <chr>
## 1 PAL0708                 1 Adelie~ Anvers Torge~ Adul~ N1A1
## 2 PAL0708                 2 Adelie~ Anvers Torge~ Adul~ N1A2
## 3 PAL0708                 3 Adelie~ Anvers Torge~ Adul~ N2A1
## 4 PAL0708                 4 Adelie~ Anvers Torge~ Adul~ N2A2
## 5 PAL0708                 5 Adelie~ Anvers Torge~ Adul~ N3A1
## 6 PAL0708                 6 Adelie~ Anvers Torge~ Adul~ N3A2
## # ... with 10 more variables: `Clutch Completion` <chr>, `Date Egg` <date>,
## #   `Culmen Length (mm)` <dbl>, `Culmen Depth (mm)` <dbl>, `Flipper Length
## #   (mm)` <dbl>, `Body Mass (g)` <dbl>, Sex <chr>, `Delta 15 N (o/oo)` <dbl>,
## #   `Delta 13 C (o/oo)` <dbl>, Comments <chr>
```

For this exercise, we're gonna use the `penguins` dataset.

2.1.2 Questions

1) Install the package `palmerpenguins`.

2) Create a new R Markdown document, name it and save it. Delete everything after line 12. Add a new section title, simple text and text in bold font. Compile ("Knit").

3) Add a chunk in which you load the `palmerpenguins`. The corresponding line of code should be hidden in the output. Load also the `tidyverse` suite of packages. Modify the defaults to suppress all messages.

4) Add another chunk in which you build a table with the 10 first rows of the dataset.

5) In a new section, display how many individuals, penguins species and islands we have in the dataset. This info should appear directly in the text, you need to use inline code . Calculate the mean of the (numeric) traits measured on the penguins.

6) In another section, entitled 'Graphical exploration', build a figure with 3 superimposed histograms, each one corresponding to the body mass of a species.

7) In another section, entitled 'Linear regression', fit a model of bill length as a function of body size (flipper length), body mass and sex. Obtain the output and graphically evaluate the assumptions of the model.

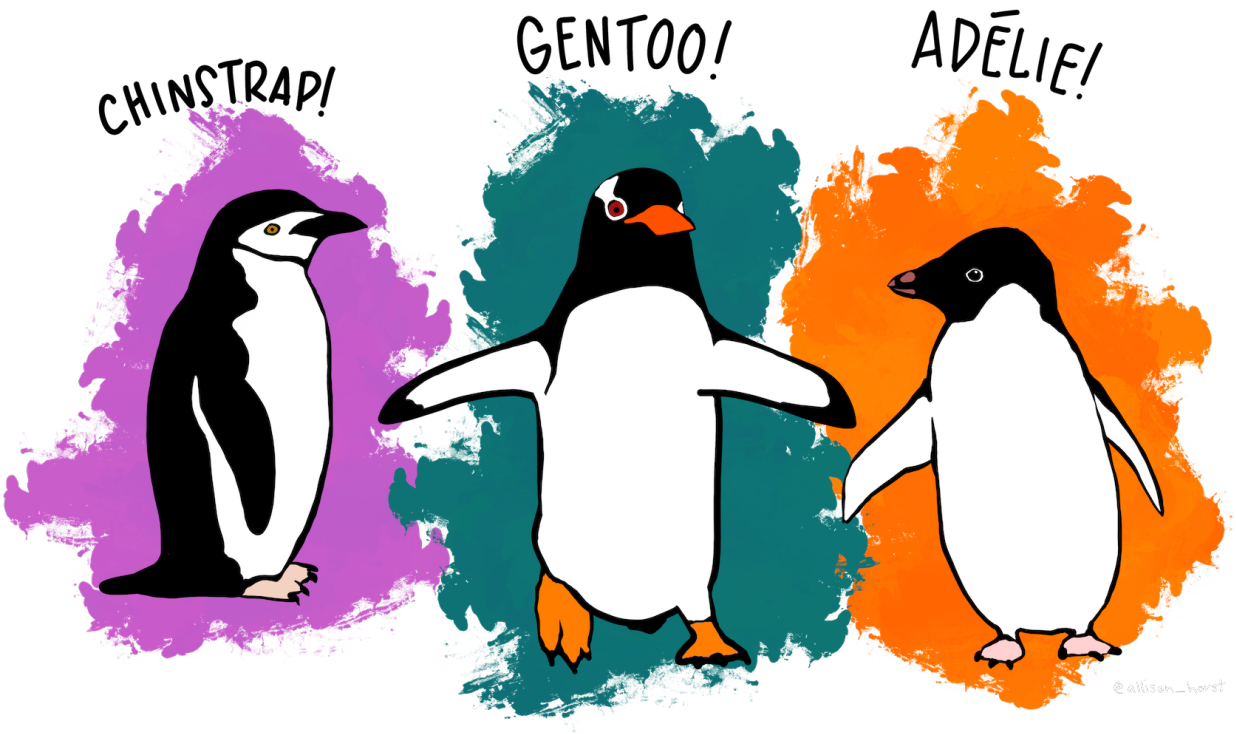
8) Add references manually or using `citr` in `RStudio`.

1. Pick a recent publication from the researcher who shared the data, Dr Kristen Gorman. Import this publication in your favorite references manager (we use Zotero, no hard feeling), and create a bibtex reference that you will add to to the file `mabiblio.bib`.
2. Add `bibliography: mabiblio.bib` at the beginning of your R Markdown document (YAML).
3. Cite the reference iin the text using either typing the reference manually or using `citr`. To use `citr`, instal it first; if everything goes well, you should see it in the pulldown menu `Addins` . Then simply use `Insert citations` in the pull-down menu `Addins`.
4. Compile.

9) Change the default citation format (Chicago style) into the The American Naturalist format. It can be found here https://www.zotero.org/styles. To do soo, add `csl: the-american-naturalist.csl` in the YAML.

10) Build your report in html, pdf and docx format.

2.1.3 Happy coding



Chapitre 3

Introduction to github with R

3.1 Practical

3.1.1 Context

We will configure Rstudio to work with our github account, then create a new project and start using `github`. To have some data I suggest to use the awesome `palmerpenguins` dataset .

3.1.2 Information of the data

These data have been collected and shared by Dr. Kristen Gorman and Palmer Station, Antarctica LTER.

The package was built by Drs Allison Horst and Alison Hill, check out the official website.

The package `palmerpenguins` has two datasets.

```
library(palmerpenguins)
```

The dataset `penguins` is a simplified version of the raw data; see `?penguins` for more info:

```
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex
##   <fct>   <fct>           <dbl>         <dbl>            <int>       <int> <fct>
## 1 Adelie  Torge~           39.1          18.7              181        3750 male
## 2 Adelie  Torge~           39.5          17.4              186        3800 fema~
## 3 Adelie  Torge~           40.3          18                195        3250 fema~
## 4 Adelie  Torge~           NA            NA                 NA          NA <NA>
```

```
## 5 Adelie  Torge~           36.7          19.3               193        3450 fema~
## 6 Adelie  Torge~           39.3          20.6               190        3650 male
## # ... with 1 more variable: year <int>
```

The other dataset `penguins_raw` has the raw data; see `?penguins_raw` for more info:

```
head(penguins_raw)
```

```
## # A tibble: 6 x 17
##   studyName `Sample Number` Species Region Island Stage `Individual ID`
##   <chr>               <dbl> <chr>   <chr>  <chr>  <chr> <chr>
## 1 PAL0708                 1 Adelie~ Anvers Torge~ Adul~ N1A1
## 2 PAL0708                 2 Adelie~ Anvers Torge~ Adul~ N1A2
## 3 PAL0708                 3 Adelie~ Anvers Torge~ Adul~ N2A1
## 4 PAL0708                 4 Adelie~ Anvers Torge~ Adul~ N2A2
## 5 PAL0708                 5 Adelie~ Anvers Torge~ Adul~ N3A1
## 6 PAL0708                 6 Adelie~ Anvers Torge~ Adul~ N3A2
## # ... with 10 more variables: `Clutch Completion` <chr>, `Date Egg` <date>,
## #   `Culmen Length (mm)` <dbl>, `Culmen Depth (mm)` <dbl>, `Flipper Length
## #   (mm)` <dbl>, `Body Mass (g)` <dbl>, Sex <chr>, `Delta 15 N (o/oo)` <dbl>,
## #   `Delta 13 C (o/oo)` <dbl>, Comments <chr>
```

For this exercise, we're gonna use the `penguins` dataset.

3.1.3 Questions

1) Create a github acocunt if not done yet.

2) Configure Rstudio with your github account using the `usethis` package.

3) Store your GITHUB Personal Authorisation Token in your .Renviron file

4) Create a new R Markdown project, and create a new git repository

5) Create a new Rmarkdon document, in your project. Then save the file and stage it.

6) Create a new commit including the new file and push it to github (Check on github that it works).

7) Edit the file. Delete everything after line 12. Add a new section title, simple text and text in bold font. Then knit and compile.

8) Make a new commit (with a meaningful message), and push to github.

9) Create a new branch, and add a new section to the rmarkdown file in this branch. Whatever you want. I would suggest a graph of the data.

10) Creat a commit and push it to the branch.

11) On github, create a pull request to merge the 2 different branches.

12) Check and accep the pull request to merge the 2 branches.

You have successfully used all the essential tools of `git` . You are really to explore and discover its power

3.1.4 Happy git(hub)-ing

Partie II

Statistics

Chapitre 4

Refresher on `glm`

```
m1 <- glm(fish ~ french_captain, data = dads_joke, family = poisson)
```

Exercice 5

On a relevé les notes en maths, anglais et sport d'une classe de 6 élèves et on a stocké ces données dans trois vecteurs :

```
maths <- c(12, 16, 8, 18, 6, 10)
anglais <- c(14, 9, 13, 15, 17, 11)
sport <- c(18, 11, 14, 10, 8, 12)
```

Calculer la moyenne des élèves de la classe en anglais.

Calculer la moyenne générale de chaque élève.

```
## [1] 13.16667
```

```
## [1] 14.66667 12.00000 11.66667 14.33333 10.33333 11.00000
```

Exercice 5

On a relevé les notes en maths, anglais et sport d'une classe de 6 élèves et on a stocké ces données dans trois vecteurs :

```
maths <- c(12, 16, 8, 18, 6, 10)
anglais <- c(14, 9, 13, 15, 17, 11)
sport <- c(18, 11, 14, 10, 8, 12)
```

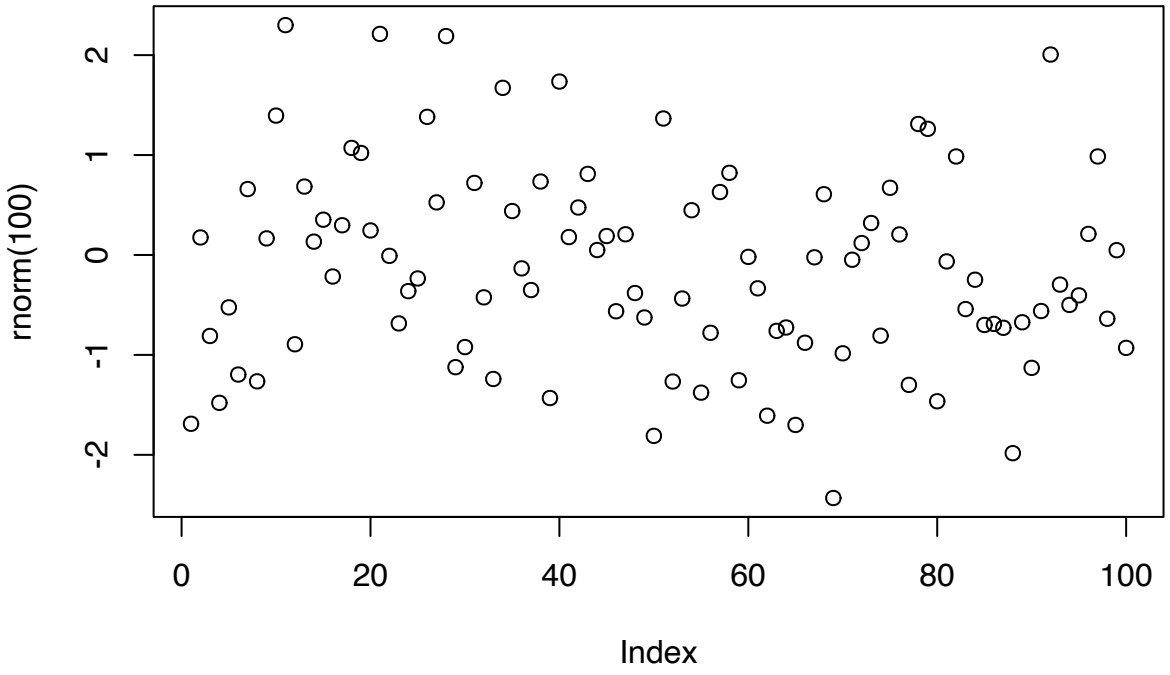
Calculer la moyenne des élèves de la classe en anglais.

Calculer la moyenne générale de chaque élève.

Test pour montrer la solution

```
## [1] 13.16667
```

```
## [1] 14.66667 12.00000 11.66667 14.33333 10.33333 11.00000
```



Chapitre 5

Refresher on `glm`

5.1 Practical

5.1.1 The superb wild unicorns of the Scottish Highlands

Unicorns, a legendary animal and also symbol or Scotland, are frequently described as extremely wild woodland creature but also a symbol of purity and grace. Here is one of most accurate representation of the lengendary animal.

Figure 5.1: The superb unicorn of the Scottish Highlands

Despite their image of purity and grace, unicorns (*Unicornus legendaricus*) are raging fighter when it comes to compete for the best sweets you can find at the bottom of rainbows (unicorn favourite source of food).

We want to know:

- If aggressiveness differs among individuals
- If aggressive behaviour is plastic (change with the environment)
- If aggressive behaviour depends on body condition of focal animal

With respect to plasticity, we will focus on rival size as an 'environment'. Common sense, and animal-contest theory, suggest a small animal would be wise not to escalate an aggressive contest against a larger, stronger rival. However, there are reports in the legendary beasty literature that they get more aggressive as rival size increases. Those reports are based on small sample sizes and uncontrolled field observations by Munro baggers enjoying their whisky after a long day in the hills.

5.1.1.1 Experimental design - what is the structure of the data we have?

Here, we have measured aggression in a population of wild unicorns. We brought some (n=80) individual into the lab, tagged them so they were individually identifiable, then repeatedly observed their aggression when presented with model 'intruders' (animal care committe approved). There were three models; one of average unicorn (calculated as the population mean body length), one that was build to be 1 standard deviation below the population mean, and one that was 1 standard deviation above.

Data were collected on all individuals in two block of lab work. Within each block, each animal was tested 3 times, once against an 'intruder' of each size. The test order in which each male experienced the three instruder sizes was randomised in each block. The body size of all focal individuals was measured at the beginning of each block so we know that too (and have two separate measures per individual).

5.1.1.2 Data

Let's load the data file and make sure we understand what it contains

```
unicorns <- read.csv("data/unicorns_aggression.csv")
```

You can use `summary(unicorns)` to get an overview of the data and/or `head(unicorns)` to see the structure in the first few lines. This data frame has 6 variables:

- Individual **ID**
- Experimental **Block**, denoted for now as a continuous variable with possible values of -0.5 (first block) or +0.5 (second block)
- Individual **body_size**, as measured at the start of each block in kg
- The repeat number for each behavioural test, **assay_rep**
- Opponent size (**opp_size**), in standard deviations from the mean (i.e., -1,0,1)
- **aggression**, our behavioural trait, measured 6 times in total per individual (2 blocks of 3 tests)

5.1.2 Questions

1) Load the libraries (`lme4`)

2) Load the data `unicorns.csv` and look at the summary and data structure

3) Fit a first mixed model with `lmer` that have only individual identity as a random effect and only a population mean.

4) Look at the output of the model (*summary()*)

5) This is a fairly rubbish model so now fit a better model by adding opponent size (*opp_size*) and block (*block*) as fixed effects. Look at the output of the model (*summary()*)

6) Where are the **p-values** ? what I have done wrong. Load `lmerTest` and refit the model using the exact same code. Look at the summary again.

7) Testing for random effects using `ranova()`. You can also do it by hand if you want.

8) Estimate repeatability, either by hand or using the `rpt` package

9) Now what about the effect on an individual body size on its aggression. Do a new model including body size also as a fixed effects. Look at the model output, estimate the probability associated witht eh random effects.

10) Make a few diagnostic plots before you can get to excited by your results (*homoscedasticity*, *Gaussian distribution for residuals*, *linear relation*, *similar within group variance*)

11) Compare repeatability among different models.

5.1.3 Happy mixed-modelling

Figure 5.2: The superb unicorn

Chapitre 6

Refresher on `glm`

```
m1 <- glm(fish ~ french_captain, data= dads_joke, family = poisson)
```

Exercice 5

On a relevé les notes en maths, anglais et sport d'une classe de 6 élèves et on a stocké ces données dans trois vecteurs :

```
maths <- c(12, 16, 8, 18, 6, 10)
anglais <- c(14, 9, 13, 15, 17, 11)
sport <- c(18, 11, 14, 10, 8, 12)
```

Calculer la moyenne des élèves de la classe en anglais.

Calculer la moyenne générale de chaque élève.

```
## [1] 13.16667
```

```
## [1] 14.66667 12.00000 11.66667 14.33333 10.33333 11.00000
```

Exercice 5

On a relevé les notes en maths, anglais et sport d'une classe de 6 élèves et on a stocké ces données dans trois vecteurs :

```
maths <- c(12, 16, 8, 18, 6, 10)
anglais <- c(14, 9, 13, 15, 17, 11)
sport <- c(18, 11, 14, 10, 8, 12)
```

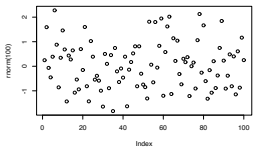
Calculer la moyenne des élèves de la classe en anglais.

Calculer la moyenne générale de chaque élève.

Test pour montrer la solution

```
## [1] 13.16667
```

```
## [1] 14.66667 12.00000 11.66667 14.33333 10.33333 11.00000
```



Partie III

Appendix

R