Assignment 2

EMATM0061: Statistical Computing and Empirical Methods, TB1, 2024

Dr. Rihuan Ke

Introduction

Create an R Markdown for assignment

First, it is recommended that you create a single R Markdown document to include your solutions, with headings created by heading codes such as "## 1.1 (Q1)", "## 3 (Q1)", etc.

It is a good practice to use R Markdown to organise your code and results. You can start with the template called Assignment02_TEMPLATE.Rmd which can be downloaded via Blackboard.

In Section 1, you will need to use R programming to complete the tasks. In section 2 and 3, it is not required to write R code.

You can optionally hand in this assignment by 13:00 Tuesday 1 October. This will help us understand your work but will not count towards your final grade. If you want to hand in the assignment, please submit a PDF file containing your answers (click on the "**Assignment 02**" under the assignment tab at Blackboards to upload the file). There is no requirement on how the PDF file is generated. One example is to choose the output of R-markdown as PDF (which may require LaTex to be installed in your computer). Another example is to choose a html output at R-markdown and convert the html file into a PDF file. If you have multiple PDF files, please combine them into a single PDF file before the submission.

Load packages

Then we need to load two packages, namely Stat2Data and tidyverse, before answering the questions. If they haven't been installed in your computer, please use install.packages() to install them first.

1. Load the tidyverse package:

```
library(tidyverse)
```

2. Load the Stat2Data package and then the dataset Hawks:

```
library(Stat2Data)
data("Hawks")
```

1. Data Wrangling

This part is mainly about data wrangling. Basic concepts of data wrangling can be found in lecture 4.

1.1 Select and filter

(Q1). Use a combination of the **select()** and **filter()** functions to generate a data frame called "hSF" which is a sub-table of the original Hawks data frame, such that

1. Your data frame should include the columns:
    a) "Wing"
    b) "Weight"
    c) "Tail"
2. Your data frame should contain a row for every hawk such that:
    a) They belong to the species of Red-Tailed hawks
    b) They have weight at least 1kg.
3. Use the pipe operator "%>%" to simplify your code.

The data frame should look like this:

```
##   Wing Weight Tail
## 1  412   1090  230
## 2  412   1210  210
## 3  405   1120  238
## 4  393   1010  222
## 5  371   1010  217
```


(Q2) How many variables does the data frame hSF have? What would you say to communicate this information to a Machine Learning practitioner?

How many examples does the data frame hSF have? How many observations? How many cases? 572

1.2 The arrange function

(Q1) Use the arrange() function to sort the hSF data frame created in the previous section so that the rows appear in order of increasing wing span.

Then use the head command to print out the top five rows of your sorted data frame. Your results should look something like this:

```
##    Wing Weight Tail
## 1  37.2   1180  210
## 2 111.0   1340  226
## 3 199.0   1290  222
## 4 241.0   1320  235
## 5 262.0   1020  200
```

1.3 Join and rename functions

The species of Hawks within the data frame "Hawks" have been indicated via a two-letter code (e.g., RT, CH, SS). The correspondence between these codes and the full names is given by the following data frame:

```
##   species_code species_name_full
## 1           CH          Cooper's
## 2           RT        Red-tailed
## 3           SS     Sharp-shinned
```

(Q1). Use **data.frame()** to create a data frame that is called **hawkSpeciesNameCodes** and is the same as the above data frame (i.e., containing the correspondence between codes and the full species names).

(Q2). Use a combination of the functions **left_join()**, the **rename()** and the **select()** functions to create a new data frame called "hawksFullName" which is the same as the "Hawks" data frame except that the Species column contains the full names rather than the two-letter codes.

(Q3). Use a combination of the **head()** and **select()** functions to print out the top seven rows of the columns "Species", "Wing" and "Weight" of the data frame called **"hawksFullName"**. Do this without modifying the data frame you just created. Your result should something like this:

```
##         Species Wing Weight
## 1    Red-tailed  385    920
## 2    Red-tailed  376    930
## 3    Red-tailed  381    990
## 4      Cooper's  265    470
## 5 Sharp-shinned  205    170
## 6    Red-tailed  412   1090
## 7    Red-tailed  370    960
```

Does it matter what type of join function you use here?

In what situations would it make a difference?

1.4 The mutate function

Suppose that the fictitious "Healthy Hawks Society"[1] has proposed a new measure called the "bird BMI" which attempts to measure the mass of a hawk standardized by their wing span. The "bird BMI" is equal to the weight of the hawk (in grams) divided by their wing span (in millimeters) squared. That is,

---

[1] Both the "Healthy Hawks Society" and the concept of "bird BMI" were made up purely for this assignment.

$$\text{Bird-BMI} := 1000 \times \text{Weight}/\text{Wing-pan}^2.$$

(Q1). Use the **mutate()**, **select()** and **arrange()** functions to create a new data frame called "hawksWithBMI" which has the same number of rows as the original Hawks data frame but only two columns - one with their Species and one with their "bird BMI". Also, arrange the rows in descending order of "bird BMI". The top 8 rows of your data frame should look something like this:

```
##   Species  bird_BMI
## 1      RT 852.69973
## 2      RT 108.75741
## 3      RT  32.57493
## 4      RT  22.72688
## 5      CH  22.40818
## 6      RT  19.54932
## 7      CH  15.21998
## 8      RT  14.85927
```

1.5 Summarize and group-by functions

Using the data frame "hawksFullName", from Section 1.3 above, to do the following tasks:

(Q1). In combination with the summarize() and the group_by functions, create a summary table, broken down by Hawk species, which contains the following summary quantities:

1. The number of rows (num_rows);
2. The average wing span in centimeters (mn_wing);
3. The median wing span in centimeters (nd_wing);
4. The trimmed average wing span in centimeters with trim=0.1, i.e., the mean of the numbers after the 10% largest and the 10% smallest values being removed (t_mn_wing);
5. The biggest ratio between wing span and tail length (b_wt_ratio).

Hint: type ?summarize to see a list of useful functions (mean, sum, etc) that can be used to compute the summary quantities. Your final result should look something like this:

```
## # A tibble: 3 × 6
##   Species       num_rows mn_wing md_wing t_mn_wing b_wt_ratio
##   <chr>            <int>   <dbl>   <dbl>     <dbl>      <dbl>
## 1 Cooper's            70    244.     240      243.       1.67
## 2 Red-tailed         577    383.     384      385.       3.16
## 3 Sharp-shinned      261    185.     191      184.       1.67
```

(Q2). Next create a summary table of the following form: Your summary table will show the number of missing values, broken down by species, for the columns Wing, Weight, Culmen, Hallux, Tail, StandardTail, Tarsus, and Crop. You can complete this

task by combining the **select()**, **group_by()**, **summarize()**, **across()**, **everything()**, **sum()** and **is.na()** functions. You should end with a summary table of the following form:

```
## # A tibble: 3 × 9
##   Species        Wing Weight Culmen Hallux  Tail StandardTail Tarsus
Crop
##   <chr>         <int>  <int>  <int>  <int> <int>        <int>  <int>
<int>
## 1 Cooper's          1      0      0      0     0           19     62
21
## 2 Red-tailed        0      5      4      3     0          250    538
254
## 3 Sharp-shinned     0      5      3      3     0           68    233
68
```

Random experiments

vents and sample spaces, and the set

随机试验是一个过程

A random experiment is a procedure (real or imagined) which:

1. has a well-defined set of possible outcomes; 有一组明确的可能结果
2. could (at least in principle) be repeated arbitrarily many times.
   可以 (至少理论上)无限次重复

it random experiments, events and sample spaces

事件(Event) 是一组即一个集合实验可能的结果

An event is a set (i.e. a collection) of possible outcomes of an experiment

ng R codes. If you

ed "Assignment_R

是所有感兴趣的可能结果的集合

e under the "resource

A sample space is the set of all possible outcomes of interest for a random experiment

ples for your

reference.

2.1 Random experiments, events and sample spaces

(Q1) Firstly, write down the definition of a random experiment, event and sample space. This question aims to help you recall the basic concepts before completing the subsequent tasks.

两次点数和为7
6*6
2^36
是的, 尽管不合理

(Q2) Consider a random experiment of rolling a dice twice. Give an example of what is an event in this random experiment. Also, can you write down the sample space as a set? What is the total number of different events in this experiment? Is the empty set considered as an event?

可能的事件数量： 样本空间的子集总数就是**所有可能的事件**的数量。根据集合论，具有 n 个元素的集合有 2^n 个子集。因此，包含36个结果的样本空间，其所有子集的数量是 2^{36} 个事件。**每一个事件都对应一个样本空间的子集**，包括空集、单一结果的集合、多结果的集合，甚至是整个样本空间本身。

2.2 Set theory

Remember that a set is just a collection of objects. All that matters for the identity of a set is the objects it contains. In particular, the elements within the set are unordered, so for example the set {1, 2, 3} is exactly the same as the set {3, 2, 1}. In addition, since sets are just collections of objects, each object can only be either included or excluded and multiplicities do not change the nature of the set. In

particular, the set {1, 2, 2, 2, 3, 3} is exactly the same as the set A = {1, 2, 3}. In general there is no concept of "position" within a set, unlike a vector or matrix.

(Q1) Set operations:

Let the sets A, B, C be defined by A := {1, 2, 3}, B := {2, 4, 6}, C := {4, 5, 6}.

1. What are the unions $A \cup B$ and $A \cup C$?
2. What are the intersections $A \cap B$ and $A \cap C$?
3. What are the complements $A \setminus B$ and $A \setminus C$? 集合B在A中的补集合 集合C在A中的补集合
4. Are A and B disjoint? Are A and C disjoint?
5. Are B and $A \setminus B$ disjoint? 动笔写出来就知道了 尤其是A \ B
6. Write down an arbitrary partition of {1,2,3,4,5,6} consisting of two sets. Also, write down another partition of {1,2,3,4,5,6} consisting of three sets.

(Q2) Complements, subsets and De Morgan's laws

Let Ω be a sample space. Recall that for an event $A \subseteq \Omega$ the complement $A^c := \Omega \setminus A := \{w \in \Omega : w \notin A\}$. Take a pair of events $A \subseteq \Omega$ and $B \subseteq \Omega$.

1. Can you give an expression for $(A^c)^c$ without using the notion of a complement?
2. What is Ω^c?
3. (Subsets) Show that if $A \subseteq B$, then $B^c \subseteq A^c$.
4. (De Morgan's laws) Show that $(A \cap B)^c = A^c \cup B^c$. Let's suppose we have a sequence of events $A_1, A_2, \cdots, A_K \subseteq \Omega$. Can you write out an expression for $(\cap_{k=1}^{K} A_k)^c$?
5. (De Morgan's laws) Show that $(A \cup B)^c = A^c \cap B^c$.
6. Let's suppose we have a sequence of events $A_1, A_2, \cdots, A_K \subseteq \Omega$. Can you write out an expression for $(\cup_{k=1}^{K} A_k)^c$?

(Q3) Cardinality and the set of all subsets:

Suppose that $\Omega = \{w_1, w_2, \cdots, w_K\}$ contains K elements for some natural number K. Here Ω has cardinality K.

Let E be a set of all subsets of Ω, i.e., $E := \{A | A \subset \Omega\}$. Note that here E is a set. Give a formula for the cardinality of E in terms of K.

(Q4) Disjointness and partitions.

Suppose we have a sample space Ω, and events A_1, A_2, A_3, A_4 are subsets of Ω.

1. Can you think of a set which is disjoint from every other set? That is, find a set $A \subseteq \Omega$ such that $A \cap B = \emptyset$ for all $B \subseteq \Omega$.

2. Define events $S_1 := A_1, S_2 = A_2 \setminus A_1, S_3 = A_3 \setminus (A_1 \cup A2), S_4 = A_4 \setminus (A_1 \cup A2 \cup A_3)$. Show that S_1, S_2, S_3, S_4 form a partition of $A_1 \cup A_2 \cup A_3 \cup A_4$.

(Q5) Indicator function.

Suppose we have a sample space Ω, and the event A is a subset of Ω. Let $\mathbf{1}_A$ be the indicator function of A.

1. Write down the indicator function $\mathbf{1}_{A^c}$of A^c (use $\mathbf{1}_A$ in your formula).
2. Can you find a set B whose indicator function is $\mathbf{1}_{A^c} + \mathbf{1}_A$?
3. Recall that $\mathbf{1}_{A\cap B} = \mathbf{1}_A \cdot \mathbf{1}_B$ and $\mathbf{1}_{A\cup B} = \max(\mathbf{1}_A, \mathbf{1}_B) = \mathbf{1}_A + \mathbf{1}_B - \mathbf{1}_A \cdot \mathbf{1}_B$ for any $A \subseteq \Omega$ and $B \subseteq \Omega$. Combining this with the conclusion from Question (Q5) 1, use indicator functions to prove $(A \cap B)^c = A^c \cup B^c$ (De Morgan's laws).

~~**(Q6)** Uncountable infinities (this is an optional extra).~~

This is a challenging optional extra. You may want to return to this question once you have completed all other questions.

Show that the set of numbers $\Omega := [0,1]$ is uncountably infinite.

3. Probability theory

In this section we consider some of the concepts introduced in Lecture 6.

Recall that we have introduced the three key rules of probability. Given a sample space Ω along with a well-behaved collection of events $\mathcal{E}$, a probability $\mathbb{P}$ is a function which assigns a number $\mathbb{P}(A)$ to each event $A \in \mathcal{E}$, and satisfies rules 1, 2, and 3:

: $\mathbb{P}(A) \geq 0$ for any event $A \in \mathcal{E}$

: $\mathbb{P}(\Omega) = 1$ for sample space Ω

: For pairwise disjoint events $A_1, A_2, \cdots$ in $\mathcal{E}$, we have

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

3.1 Rules of probability

(Q1) Construct a probability function based on the Rules of probability

Consider a sample space $\Omega = \{a, b, c\}$ and a set of events $\mathcal{E} = \{A \subseteq \Omega\}$ (i.e., $\mathcal{E}$ consists of all subsets of Ω). Based on the rules of probability, find a probability function $\mathbb{P}: \mathcal{E} \to [0,1]$ that satisfies

$$\mathbb{P}(\{a, b\}) = 0.6 \quad \text{and } \mathbb{P}(\{b, c\}) = 0.5.$$

In your example, you need to define a function called $\mathbb{P}$. The function maps each event in $\mathcal{E}$ to a number. Make sure that your function $\mathbb{P}$ satisfies the three rules, but you don't need to write down the proof (that it satisfies the three rules).

(Q2) Verify that the following probability space satisfies the rules of probability.

Consider a setting in which the sample space $\Omega = \{0,1\}$, and $\mathcal{E} = \{A \subseteq \Omega\} = \big\{\emptyset, \{0\}, \{1\}, \{0,1\}\big\}$. For a fixed $q \in [0,1]$, define a function $\mathbb{P}: \mathcal{E} \to [0,1]$ by

$$\mathbb{P}(\emptyset) = 0,\ \mathbb{P}(\{0\}) = 1 - q,\ \mathbb{P}(\{1\}) = q,\ \mathbb{P}(\{0,1\}) = 1.$$

Show that the probability space $(\Omega, \mathcal{E}, \mathbb{P})$ satisfies the three rules of probability.

3.2 Deriving new properties from the rules of probability

(Q1) Union of a finite sequence of disjoint events.

Recall that in Rule 3, we have

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

for an infinite sequence of pairwise disjoint events $A_1, A_2, \cdots$. Show that for a finite sequence of disjoint events $A_1, A_2, \cdots A_n$, for any integer n bigger than 1, the below equality holds as a consequence of Rule 3:

$$\mathbb{P}(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i)$$

Please note that in left hand side of the equation above we have the union of a finite sequence instead of an infinite sequence.

(Q2) Probability of a complement.

Prove that if Ω is a sample space, $S \subseteq \Omega$ is an event and $S^c := \Omega \setminus S$ is its complement, then we have

$$\mathbb{P}(S^c) = 1 - \mathbb{P}(S).$$

(Q3) The union bound

In Rule 3, for pairwise disjoint events $A_1, A_2, \cdots$, we have

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Recall that in the lecture we have also shown the union bound as a consequence of the rules of probability: for a sequence of events $S_1, S_2, \cdots$, we have $\mathbb{P}(\cup_{i=1}^{\infty} S_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(S_i)$.

Give an example of a probability space and a sequence of sets $S_1, S_2, \cdots$, such that $\mathbb{P}(\cup_{i=1}^{\infty} S_i) \neq \sum_{i=1}^{\infty} \mathbb{P}(S_i)$.

(Q4) Probability of union and intersection of events.

Show that for events $A \subseteq \Omega$ and $B \subseteq \Omega$, we have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$