Assignment 1. ST661 2019

Catherine Hurley

Due on Monday October 14, 6pm

You should complete this assignmnent by filling in the answers into assign1.Rmd (on Moodle). You should knit the file to htlm and upload the html file to Moodle. The upload must be completed by time and date given above.

1. Write code to
   (a) Construct a vector with values 10,11,12,13,14. Do this using :, c and seq. Assign your result to `x`.
   (b) Construct a vector with values 14,10,11,13,12. Call it `y`.
   (c) Extract the third value of `x`.
   (d) Extract the second and fourth values of `y`.
   (e) Write a sentence explaining what the next line of code does.

```
sum(x>y)
```

   (f) Write a sentence explaining what the next line of code does. (Look up the help page for abs.)

```
sum(abs(x-y) == 1)
```

   (g) Write code which implements the following formula.

$$\sum_{1}^{5} x_i^2/y_i$$

   (h) Form a 5 × 2 matrix using the values in `x` and `y`.

2. A 2009 Eurostat report lists the percentage of the population in EU countries living with in a dwelling that (i) is too dark, (ii) has a leaking roof or is damp, (iii) has no bath or shower and (iv) has no indoor flushing toilet.
   Download the data from Moodle and read in the data using

```
house <- read.csv("house.csv",row.names=1)
# or if you are working on the Rstudio server
house < <- read.csv("rstudio_files/ST661/house.csv",row.names=1)
```

   You might need to insert the path to the folder containing the csv file.
   Write R code for following.
   (a) Use head to check the first few rows of the data.
   (b) Extract the row of the data for Ireland.
   (c) Find Spain's value for TooDark.
   (d) For how many countries is the percentage without an indoor toilet 0?
   (e) For how many countries is the percentage without an indoor toilet and without a bath/shower both 0?
   (f) Find the names of the countries where the percentage without an indoor toilet and without a bar/shower are both zero. Your code should give a vector of the country names.
   (g) Find the maximum value of the variable Damp.

3. Download the worldcup.Rdata file from Moodle and type in

```
load("worldcup.Rdata")
```

You might need to insert the path to the folder containing the Rdata file. The data can also be found on the Rstudio server.

The vector ngoals has contains the number of goals scored in every world cup final match from 1990 to 2002. Only goals in the 90 minute regulation time are considered. See next page for more information.

(a) Draw a barplot of the number of goals per match. What is the most frequent number of goals? Find the mean and the median.

(b) The first 52 matches were played in 1990, the next 52 matches were played in 1994, the next 64 matches were played in 1998, and the last 64 matches were played in 2002. Therefore elements 52, 104, 168 and 232 of ngoals give the numbers of matches scored in the four finals. Extract these elements of ngoals into a vector.

(c) Make a vector of length 232 containing the year each match was played. Call your vector year. (Hint: use rep.)

(d) Turn the vector year into a factor. Make a data frame where the first variable is ngoals and the second is year.

(e) Make a vector called matchn containing the number of the match for that world cup. The vector should be 1..52, 1..52, 1..64, 1..64. Add this vector to the data frame.

(f) Extract the rows 52, 104, 168 and 232 of the data frame. Which year's final has the most goals?

(g) Make a boxplot of the number of goals played for each year. What can you say about the 8 goal match?

(h) (Advanced, optional). The vector gtimes contains the cumulative time in minutes from the start of the 1990 world cup in which goals were scored (counting playing time only.) Construct another vector minute containing the minute of each game in which the goal is scored. In what minutes of the matches are goals most likely to occur?

2 THE WORLD CUP DATASET

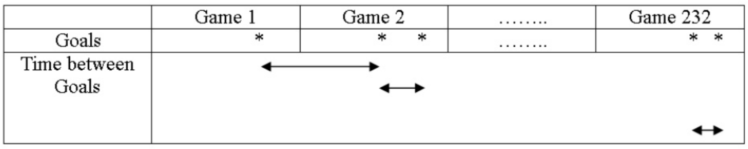
The World Cup tournament pitching the best national teams is the ultimate in the soccer world. It is held every four years. South Korea and Japan co-hosted the latest tournament between 31 May and 30 June 2002. That tournament involved 32 countries, which were initially divided into 8 groups of four. In this first stage, each country played against each of its 3 group peers. The top 2 countries within each group then advanced to a knockout second stage. Ultimately, Brazil beat Germany 2-0 in the 64^{th} and final game.

In 1998, the same tournament format was used in France. The host country was ultimately crowned as the champion. However, when Italy and USA respectively hosted the World Cup in 1990 and 1994, only 24 countries were showcased. In these two gatherings, the respective winners, Germany and Brazil, emerged after 52 games.

Data on all the goal occurrences in these four World Cup tournaments are available at Fifa's World Cup website. Previous tournaments are ignored as the website does not provide information on the sequence of games. As such, they would not help in the illustration of the exponential distribution.

This paper therefore focuses on the 232 games played in the 1990-2002 World Cup tournaments. Only the goals scored in the 90 minutes regulation time are considered. This leaves out goals scored in extra time or in penalty shoot-outs. A regular soccer game consists of two halves scheduled for 45 minutes. However, injury time is often added at the end of each half to compensate for game stoppages arising from player injuries. The extent of injury time in each game is unfortunately not available. For consistency in the analysis, a goal scored in injury time, say at the 92^{nd} minute, is recorded as occurring at the 90^{th} minute. This is because until 1998, goals scored in added times were always recorded at either the 45^{th} or 90^{th} minute. This may have some effect on the fit of the Poisson and exponential distributions to the data.

The first game in Italy90 saw Cameroon scoring a single goal against Argentina at the 67^{th} minute. In the second game, Romania scored 2 goals against the then Soviet Union at the 42^{nd} and 57^{th} minutes. The time to these goals are respectively 65 $(= (90-67)+42)$ and 15 $(= 57-42)$ minutes. Proceeding in the same way, 574 inter-goal times were obtained by the end of the 2002 Final game (game 232). Figure 1 illustrates the computation of the time between goals.



	Game 1	Game 2	……..	Game 232
Goals	*	* *	……..	* *
Time between Goals				