University of Manitoba
Department of Statistics

STAT 4600: Computational Statistics

Assignment 3

Due: Tuesday, February 6

Question 1: Sampling Distributions using Monte Carlo simulation

Consider the following simple model for data that can be corrupted by the presence of outliers. For $i = 1, 2, \ldots$ we assume

$$X_i | Z_i = 0 \sim N(0, 1),$$

and

$$X_i | Z_i = 1 \sim N(8, 1),$$

with $Z_i \sim$ Bernoulli(0.05). Intuitively, each time a value of X_i is to be obtained, we first obtain Z_i, which is an indicator variable associated with X_i being an outlier ($Z_i = 1$, with probability 0.05) or not ($Z_i = 0$, with probability 0.95). Outliers are characterized by taking a value close to 8, while most observations take values closer to zero. Finally, with the exception of the above dependence of X_i on Z_i for each value of i, it is assumed that all random variables are independent.

(A) We want to visually compare the sampling distributions of the sample mean $\bar{X}$ and the sample median M computed from samples of size 10 associated with the above population model.

For this, first generate 1000 samples of size 10. For each sample, obtain the sample mean $\bar{X}_j$ and median M_j ($j = 1, 2, \ldots, 1000$). Then, make a histogram for each set of values. From these two histograms, does it appear that the sample mean and sample median have approximately normal sampling distributions?

(B) Do the same as above with samples of size 25.

(C) Repeat with samples of size 100.

(D) Do you see an emerging pattern?

Question 2: Back to English Premiership Soccer

For this second look at the soccer data, we will be using the dates of the games to look at teams that perform especially well (or bad) on some specific dates. First, start by constructing the `soccer` data frame by using the following command

```
soccer <- read.csv('EPL_1617.csv', header=T, stringsAsFactors = FALSE)
```

Note that the dates in the first column of the data frame `soccer` are character strings and that games are not ordered chronologically within the data frame.

(A) Construct a function named `dated.summary` that returns a table like the one produced in Assignment 1, but only accounting for games that took place between two dates, `first.date` and `second.date`, to be provided by the user as arguments. These dates can be expected to be character strings in the format `'yyyy-mm-dd'`. The function should also use the `game.results` for the whole season (you will be passing on the `soccer` data) as an argument.

Your function should return an appropriate message if no games took place between the given dates or if the two dates aren't ordered properly.

Finally, use your function to see if any team performed exceptionally well or poorly in

- the first seven weeks of the season,
- the last eight weeks of the season.

Note: Using the function you created in Assignment 1 (if it was as generic as required) should allow you to do this in just a few lines!

(B) Construct an R function named `team.progression` that produces, given the `game.results`, a table showing how the points total of each team progressed over the season after each game, as per Figure 1. Specifically, the output of the function should be a matrix with appropriate team names as row names for improved readability.

For instance, looking at the first row of this table, we see that AFC Bournemouth lost their first two games, then tied their third and won their fourth, for a running total of 0, 0, 1 and 4 points over their first four games of the season. Arsenal, on the other hand, lost their first game, tied their second and won their next two, for a running total of 0, 1, 4 and 7 points over their first four games.

Again, it is desired that the function be fully adaptive to the provided results, i.e., that no information outside of the file that is read be necessary to produce the wanted table.

(C) From the table produced in (B), make a (nice!) graph that can be used to assess Chelsea's domination over the season.

The specifics of what the graph should look like are up to you, but you should keep in mind that a lot of this is about how to display the information clearly and in a way that is insightful. One approach would be to compare Chelsea's performance to that of a few carefully selected teams. Write a few comments to explain what conclusion you draw from your graph.

Finally, do the same to assess Sunderland's performance over the season.

Figure 1: Points progression matrix that should be produced for Question 2

	1	2	3	4	5	6	7	8	9	10	...	38
AFC Bournemouth	0	0	1	4	4	7	8	11	12	12	...	46
Arsenal	0	1	4	7	10	13	16	19	20	23	...	75
Burnley	0	3	3	4	4	7	7	7	10	11	...	40
Chelsea	3	6	9	10	10	10	13	16	19	22	...	93
Crystal Palace	0	0	1	4	7	10	11	11	11	11	...	41
Everton	1	4	7	10	13	13	14	15	15	18	...	61
Hull City	3	6	6	7	7	7	7	7	7	7	...	34
Leicester City	0	1	4	4	7	7	8	8	11	12	...	44
Liverpool	3	3	4	7	10	13	16	17	20	23	...	76
Manchester City	3	6	9	12	15	18	18	19	20	23	...	78
Manchester United	3	6	9	9	9	12	13	14	14	15	...	69
Middlesbrough	1	4	5	5	5	5	6	6	7	10	...	28
Southampton	1	1	2	2	5	8	9	12	13	13	...	46
Stoke City	1	1	1	1	1	2	3	6	9	12	...	44
Sunderland	0	0	1	1	1	1	2	2	2	2	...	24
Swansea City	3	3	3	4	4	4	4	4	5	5	...	41
Tottenham Hotspur	1	4	5	8	11	14	17	18	19	20	...	86
Watford	1	1	1	4	7	7	8	11	12	15	...	40
West Bromwich Albion	3	3	4	4	7	8	9	10	10	10	...	45
West Ham United	0	3	3	3	3	3	4	7	10	10	...	45