Problem set 1

Each question is worth ten marks.

1. In the Monty Hall problem, a host (H) and a contestant (C) play the following game. (1) H places a prize behind one of three doors. (2) C makes a random guess as to which door hides the prize. (3) H opens one of the unchosen doors that does not hide the prize. (4) C chooses whether to stay with the chosen door, or switch to the other closed door. (5) All remaining doors are opened to show whether C wins the prize.

   What is the probability of winning if C stays with the chosen door in step (4)? What is the probability of winning if C switches to the other closed door? Prove your answers using the basic rules of probability theory.

2. Write an R function `rcnorm( nsamp, mean, sd, nclip )` that returns a vector of `nsamp` independent random numbers that are normally distributed, except that they all fall within `nclip` standard deviations of the mean. The mean is `mean` and the standard deviation of the unclipped distribution is `sd`. Give the mean a default value of 0, the standard deviation a default value of 1, and `nclip` a default value of 2. (Suggestion: start with a sample from `rnorm()`, and keep resampling any values that are outside of the desired range until none are left.)

3. The file `keeling.txt` contains monthly measurements of the atmospheric concentration of CO_2 from March 1958 to present. Download this file from the course github repository. Write an R script that does the following with this data.

   (a) Use `read.table` to read the file as a data frame. Use the optional argument `header=TRUE` to indicate that the file contains a line of headings (`year`, `month`, etc.) that will be used to name the data frame columns. Also use the optional argument `sep=","` to indicate that the columns in the text file are separated by commas. Consult `?read.table` if you need further information about this function.

   (b) Plot the CO_2 concentration for each month from March 1958 to present. Inspect the data in `keeling.txt`, and use one or more of the first four columns to come up with a measure of time for the x-axis. For the CO_2 concentration, use the column `co2_fill`, which is a slightly smoothed version of the measurements.

   (c) Plot the average CO_2 concentration during each year from 1959 to present.

   (d) Plot the average CO_2 concentration by month, averaging over the years from 1959 to present.

   In parts (b), (c), and (d), give the plots titles and axis labels.

Due date: February 25, 2020

Bonus problems

1. Generalize your solution to problem 1 (above) in some interesting way. For example, suppose that instead of three doors there is some arbitrary number n of doors, solve the problem, and see how the advantage of the switching strategy increases or decreases as the number of doors increases.

2. Solve problem 2 (above) using inverse transform sampling. This will allow you to find all the samples from the clipped normal distribution in one pass, without finding and resampling outliers. See the last lines of `randvar.R` from lecture 3 for an example of inverse transform sampling.

3. Improve the solution to problem 3 (above) in some useful way. For example, (1) add error bars showing the standard error of the mean of each data point in parts (c) and (d), (2) add a legend to the plot, (3) make the three plots subpanels of a single figure, and/or (4) save the plots to an .eps or .pdf file.