# Assignment 3

*STAT221*

*To be submitted on Learn by 3pm on Tuesday 6 October 2020*

Where you are asked to do something in R you should include ALL of the code used to produce the result in your assignment submission so it can be reproduced and checked. But do not include any code which is not required (i.e. be concise, just as you would be expected to do in writing essays or any other maths/stats work). Your code should include comments at the main steps to explain what it is doing. Part of the assessment in this assignment is for well written R code. Use the examples in the lectures and labs as a guide to showing clear and concise programs. When asked to "explain" or discuss a particular result you will be expected to write one or two sentences, e.g. where relevant to explain "why" a result occurred or "how" to do something.

Graphs are expected to have relevant axis labels and titles, but at this stage legends (for when there are multiple things displayed on one graph) are not expected.

The assignments covers two topics:

1. density estimation; and
2. smoothing.

## Q1. Bin Width Choice for Histogram Density Estimator

a. Generate a sample of 1,000 independent observations from a gamma distribution with scale $\theta = 1$ and shape $k = 5$, using the `rgamma()` function.
   Create a sample density histogram and overlay the true gamma density function. Use the default settings in the `hist()` function for estimating the bin width (essentially the default is Sturge's Rule). Does the bin width obtained using Sturge's Rule look right?

b. On a single plot, include three subgraphs to show the estimated density histogram using the three bin width rules:

- Sturge's Rule
- Scott's Normal Reference Rule and
- Freedman-Diaconis Rule.
  On each plot overlay the true density function. In your opinion which bin width estimator looks the most reasonable?

c. Repeat part (a) and (b) for 1,000 simulated datapoints from a standard Cauchy distribution. Comment as to which of the above rules appears to provide the most reasonable estimate of the bin width?

## Q2. Christchurch Pollution Data

On Learn there is a Christchurch pollution dataset from ECan. The dataset is stored in a comma-separated variable format (hence the .csv extension on the filename).

Be careful when downloading the datafile that your browser does not change the file format or filename.

The following code changes the working directory of R to your P: drive in a folder called STAT221, where it assumes you have saved this datafile. If you have saved the datafile in a different directory you will have to change the first line. The data is imported into R using the `read.csv()` function. The code then shows you the variables names in the dataset and the daily average PM$_{10}$ pollution concentrations in $\mu g/m^3$.

```
setwd("P:/STAT221")
chch <- read.csv("chchpollution.csv")
# column names
names(chch)
# summary of data.frame content
str(chch)
# daily average particulate matter concentration
chch$PM10
```

a. When considering their policy on residential solid fuel burner usage, ECan focus on the daily average PM$_{10}$ concentrations during the months May-September, as these are when the high concentrations predominantly caused by residential heating occur.
   Create a histogram of the May-September daily average PM$_{10}$ concentrations. Use a bin width of 1 $\mu g/m^3$.

b. Plot a frequency polygon density estimate of the same data. Include appropriate axis labels and a title.

c. Plot an average shifted histogram density estimate of the same data. Include appropriate axis labels and a title.

d. Plot the empirical cumulative distribution function of this data. To do this use the function `ecdf`, and include appropriate axis labels and a title.

e. Use the empirical cumulative distribution function to estimate the probability of a randomly chosen day in May-September having an exceedance of the government target level of 50 $\mu g/m^3$. In doing so, you are assuming that the historical concentrations are representative of future behaviour.

## Q3. Smoothing

a. Plot the data contained in the file Q3a_data.txt, and fit a range of smoothers to the data, varying the function arguments to try to get a good fit. Plot four smoothers on the plot of the data, using different line types or colours, that indicate the range of smoothers that you considered. Plot your final choice for the best smoother on a separate plot. Discuss the fit of the smoothers you tried, and give an explanation as to how you made your final choice of smoother.

b. Do the same as in part a, but this time using the data contained in the file Q3b_data.txt. As in question 3a, produce two plots: one showing the range of smoothers you considered, and a second plot showing your final choice of which you think is the 'best' smoother. Discuss the fit of the smoothers you tried, and give an explanation as to how you made your final choice of smoother.

c. Generate your own dataset of 50 equally spaced points from the model

$$y_i = g(x_i) + e_i$$

   where the errors follow an iid Normal distribution with mean zero and standard deviation of 0.5, and where the true relationship is

$$g(x) = sin(x) + sin(2x).$$

   Generate the points between $x$ values of 0 and $2\pi$, and use your university ID number as the seed. Plot your simulated data and add a curve that shows the true relationship. Fit different smoothers to this data and find the smoother that you think fits best, add it to your plot above, and explain why you chose this smoother.

d. Plot the data contained in the file Q3d.csv. This file contains a binary response variable, *chd* that indicates whether or not a person has coronary heart disease, and a predictor variable, *age*, that gives the persons age in years. Add a loess smoother to your plot and then explain what the smoother tells you about this data.