SCHOOL OF MATHEMATICS AND STATISTICS

STAT 432 **Assignment 4** **Due: 5pm Thu 15 Oct 2020**

Bootstrap Methods and MCMC

This assignment is worth 10%

Please include your R code, with comments, in the main pages. Using RMarkdown is recommended. e-Submission to **louise.mcmillan@vuw.ac.nz**

1. The NZ government tracks the prices of various commonly bo"ught household items, in order to estimate the Consumer Price Index (CPI). The dataset `food_prices_kg2019.csv` contains the prices for January 2019, for all items that are listed in units of 1kg.

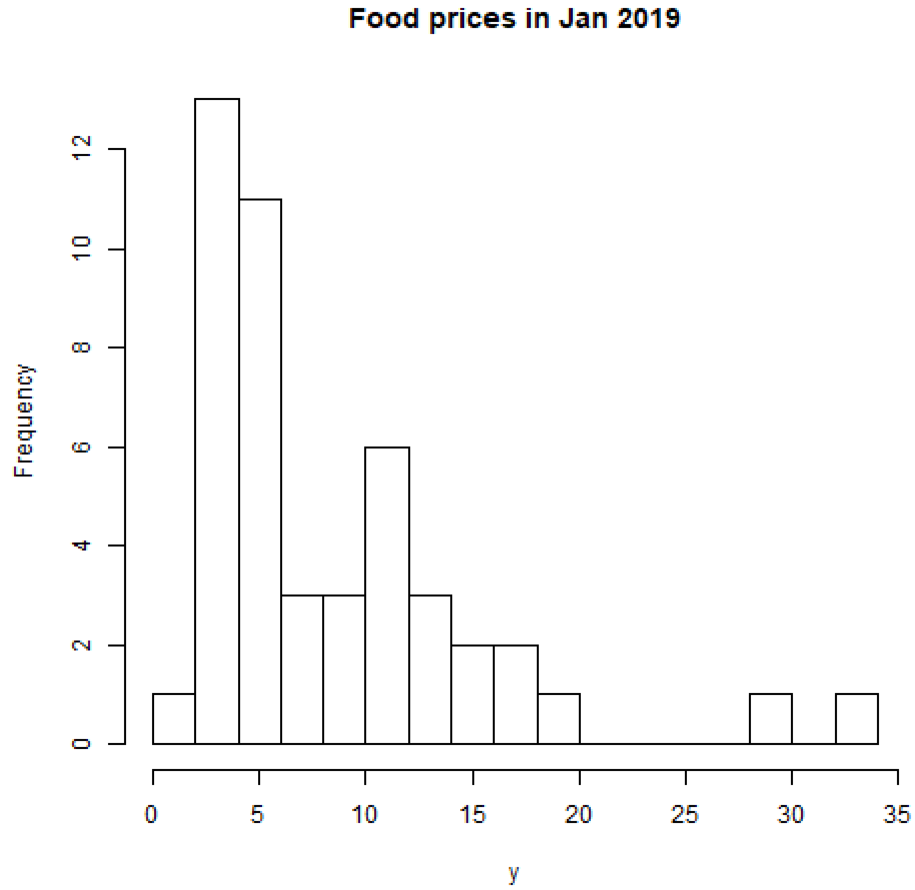


Figure 1: Food prices for items in kg units, from January 2019

(a) Give an estimate of the Interquartile Range θ using the data set (give all your answers here and in the questions below to 3 decimal places).

(b) Give a bootstrap estimate of the standard error of the estimate $\widehat{\theta}$. Use this to construct a standard 95% bootstrap confidence interval.

(c) Give Efron's percentile 95% confidence interval for θ.

(d) Give Hall's percentile 95% confidence interval θ.

(e) Give an estimate of the **bias** in the estimate of θ. Is the size of this bias of concern?

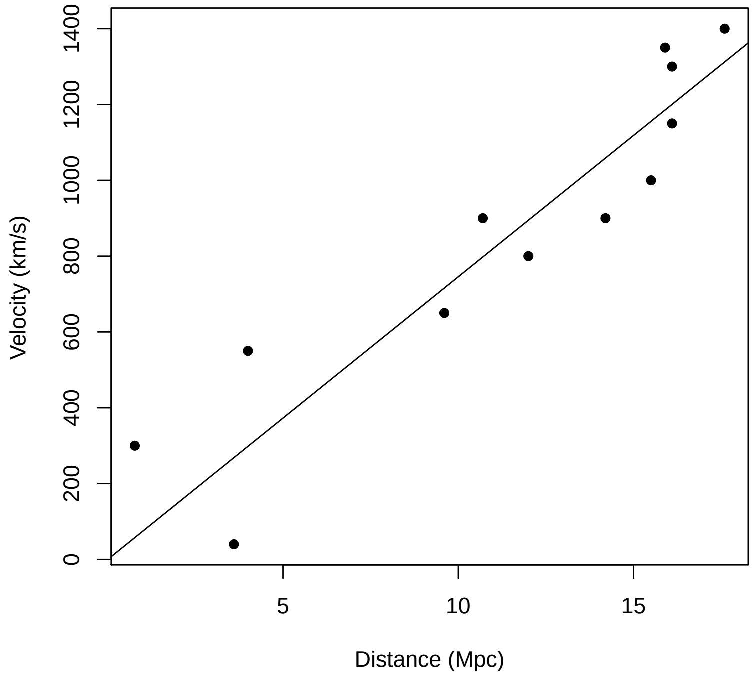
(f) Making a simple bias correction to Efron's percentile interval, quote this new interval and use it to test whether the IQR could be below 4NZD. Explain your reasoning clearly and state the significance level of the test.

2. The expansion rate of the Universe can be characterised by the Hubble Constant β. If a galaxy is at a distance x_i from earth, its apparent recessional velocity due to the expansion is y_i, then $y_i \simeq \beta x_i$.

A dataset of galaxy distances and their recessional velocities is given in the table below, and plotted in the figure, together with its fitted regression line $\widehat{y} = \widehat{\beta} x$.

Group	Galaxy	Distance, x (Mpc)	Velocity, y (km/s)
Local Grp	N0224	0.77	300
M81	N3031	3.6	40
Cen A	N5128	4.0	550
N1023	N1023	9.6	650
Leo I	N3379	10.7	900
N7331	N7331	12.0	800
UMa	N3928	14.2	900
Coma I	N4278	15.5	1000
Coma II	N4494	15.9	1350
Virgo	N4486	16.1	1150
Dorado	N1549	16.1	1300
Fornax	N1399	17.6	1400

The data are in the file `galaxies.csv`.



In the simple proportional model

$$y_i = \beta x_i + \varepsilon_i \qquad \varepsilon_i \overset{\text{ind}}{\sim} N(0, \sigma^2 x_i)$$

maximum likelihood estimators for β and σ are:

$$\begin{aligned} \widehat{\beta} &= T_1(\mathbf{x},\mathbf{y}) = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} \\ \widehat{\sigma}^2 &= T_2(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^n \frac{1}{x_i}(y_i - \widehat{\beta} x_i)^2 \end{aligned}$$

(a) Briefly explain how a **parametric** bootstrap estimate of the standard error of $\widehat{\beta}$ can be calculated.

(b) Create an estimate of the Hubble Constant β with a standard parametric bootstrap confidence interval,

See if you can do this by writing your own code, but also try using the `boot` package in R. Give your code in both cases.

(c) Write down the likelihood and log likelihood of the model, and verify the expressions for the estimates of β and σ^2 given above.

(d) Maximise the log likelihood using the `optim()` function in R to find estimates of β and σ^2 with their standard errors. Compare these to your bootstrap estimates and standard errors.

3. This question is on Markov chain Monte Carlo methods.

(a) Markov chains used for Markov chain Monte Carlo methods need to be irreducible, aperiodic and positive recurrent. Choose 2 of these properties and explain, in 1 or 2 sentences, why they are necessary for MCMC.

(b) Dirichlet distributions are increasingly widely used in statistics. (Additional info, not required for the assignment: if you are familiar with Bayesian statistics, the Dirichlet distribution is the conjugate prior of the Multinomial distribution in the same way that the Beta distribution is the conjugate prior of the Binomial distribution. Also Dirichlet distributions are related to Dirichlet process priors used in Bayesian nonparametrics.)

Dirichlet distributions can be thought of as multivariate forms of Beta distributions. A value x generated from a d-dimensional Dirichlet distribution is a vector of d values between 0 and 1 that sum to 1. A d-dimensional Dirichlet distribution is parametrized by a vector of d values $\boldsymbol{\alpha} > 0$. Often the distribution is simplified by having all the values in $\boldsymbol{\alpha}$ being equal to a single value α.

Write R code to generate values from the 3-dimensional Dirichlet distribution with parameter $\boldsymbol{\alpha} = (1,1,1)$ using the Metropolis-Hastings method. Use `set.seed()` at the start of your code to fix the values every time you compile your assignment.

You can use the function `ddirichlet` from the `mixtools` package to calculate the target density rather than coding it from scratch.

Test these two proposal distributions: (1) Uniform up to 0.5 either side of the current point; (2) Uniform up to 0.1 either side of the current point. Note that both of these proposal distributions are symmetric distributions, which may simplify your code.

Since the values you generate at a given timestep should add up to 1, that means that those 3 values are not all independent. So in a given timestep, generate the first 2 values, which will be independent, and calculate the 3rd value at each time step as 1 minus the sum of the other 2 values. But when you calculate the ratio that is part of the acceptance probability, you will need to use all 3 values to calculate the probability densities.

For each proposal distribution, simply generate the values from it without worrying about whether they are within the range [0,1]; the `ddirichlet` function will return 0 as the density value for points that are outside the accepted range. Also do not worry about the warnings you may get from the `ddirichlet` function like "`log(x) produces NAs`".

For each proposal distribution:

i. Run your chain for 1000 steps of burn-in and 4000 steps after burn-in. Start at the point (0.1,0.1,0.8).

ii. Plot the trace plot of the burn-in values, and the trace plot of the last 1000 values of the run, for at least 1 of the 3 dimensions.

iii. Calculate the acceptance ratio for the last 1000 points in the chain.

iv. Comment in 1 sentence on whether the burn-in period is long enough. Comment in 1 sentence whether the overall chain appears to be mixing well. (You can comment on the other 2 dimensions if you want, but they should be similar to the first dimension.)

(c) Comment in 1 sentence on which of the two proposal distributions you would choose. For that set of results, plot a histogram of the last 1000 values, for the 1st dimension.

(d) Run the following code as an alternative way to generate 1000 points from the Dirichlet distribution (using the fact that the Dirichlet distribution is related to the gamma distribution). Plot a histogram of the first dimension of those 1000 points. Comment in 1 sentence on whether your chosen MCMC chain is a good approximation to the Dirichlet distribution generated using this other method.

```
rDirichlet <- function(ndraw, alpha){
    gamdraw <- rgamma(ndraw, shape=alpha, rate=1)
    gamdraw / sum(gamdraw)
}
xdir.gen <- do.call(rbind, lapply(1:1000, function(i) rDirichlet(3,1)))
```