Homework 3 – Due 12 am CST, 10 March 2021

The total points on this assignment is 100. For this homework, there is no need to turn in any software code, however, you are required to turn in the results.

1. Let $X_1, X_2, \ldots, X_n$ be independent identically distributed (univariate) random variables from the $N(\mu, \sigma^2)$ distribution. Answer the following questions:

   (a) Provide the conditional distribution of $X_1, X_2, \ldots, X_{n-1}$ given $\bar{X}$. [*10 points*]

   (b) Write a function in R which will generate a random vector $\boldsymbol{X}$ where $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ given $\bar{X}$. (Note that $X_n = n\bar{X} - \sum_{i=1}^{n-1} X_i$.) [*10 points*]

   (c) The file provided `Baby_Giraffe_IAH_0417.tiff` in the usual places provides a digital file in the Tagged Image File Format (TIFF) of a baby giraffe born in the Houston zoo. The R packages `rtiff` or `tiff` can be used to read in this file as in HW 1. Here is the example using the `rtiff` package.

```
library(rtiff)
jiri <- readTiff(fn="Baby_Giraffe_IAH_0417.tiff")
plot(jiri)
```

   Color can be expressed in terms of its primary components namely Red, Green and Blue. Therefore, each pixel has a certain amount of Red, a certain amount of Green and a certain amount of Blue. These are obtained, using the `rtiff` package, at each pixel (in matrix form) by using: `jiri@red`, `jiri@green`, `jiri@blue`. (Make sure that you look at the help on `pixmap-class` to get an idea of what the components of pixmap are). Note also that these values are all between 0 and 1.

   We will first discuss what we are planning to do using the red channel as an example. Suppose that we are assuming that the red color at each pixel (grid-point) is normally distributed with some (pixel-specific) mean μ and common variance $\sigma^2/16 = 0.1^2$. We will try and predict the image at a higher resolution than currently provided. To do so, assume that the observation at each pixel is actually the mean at each of 16 pixels, where each pixel can be subdivided into 4 pixels in both the x- and y- directions. Note that this is a computationally-intensive exercise so you are encouraged to think about an efficient implementation to the programming. So given the mean of this 4×4-block (given by the observed red-channel values), we are to obtain simulated values for each 4×4-block.

   i. Obtain a matrix of the derived (higher resolution) image. [*10 points*]
   ii. Repeat the same for the green and blue channels. [*10 points*]
   iii. All RGB values in R are between 0 and 1. Make sure that the realized values are so by simply truncating values that may be above or below. [*5 points*]
   iv. Create an R object of class `pixmap` that contains these new channels. (*If you do not know what to do here, I would suggest copying the original object into another object and changing the attributes of the object.*) [*10 points*]
   v. Display the higher resolution image. Note that the higher-resolution image will be of lower quality in general because of the information lost in compressing it in the first place. [*10 points*]

   What to turn in? Turn in the derivations and the display and, if you want, the `R code`.

2. The inverse hyperbolic sine (IHS) transformation of Johnson (Biometrika, 1949) has been used by Burbridge, Magee and Robb (JASA, 1988) for the purpose of normalizing transformations for linear regression. The transformation is as follows:

$$g(y;\theta) = \begin{cases} y & \theta = 0 \\ \text{arcsinh}(\theta y)/\theta & \theta \neq 0 \end{cases}$$

which is defined over all θ and y. However, the transformation is symmetric as a function of θ so restricting to $\theta \geq 0$ is enough. Suppose that we have a sample $X_1, X_2, \ldots, X_n$. We want to find θ such that $Y_i = g(X_i, \theta)$ is normally distributed with mean μ_θ and variance σ^2_θ.

(a) Provide the loglikelihood function of $(\theta, \mu_\theta, \sigma^2_\theta)$ given the sample $X_1, X_2, \ldots, X_n$. [*15 points*]

(b) The `danish` object in the `SMPracticals` package in `R` provides data on major insurance claims due to fires in Denmark from 1980 to 1990. The values of the claims have been rescaled for commercial reasons. (There are some outliers.)

  i. For this dataset, provide some description of the untransformed dataset. [*5 points*]

  ii. Then transform the claims data for normality as per the maximum likelihood estimator (limiting yourself to θ in the range [0,4]. Use graphical displays and the Shapiro-Wilks' test to assess whether the transformed data are reasonably normal. [*15 points*]