Assignment 1, ST2304

Problem 1 Read the dataset `mammals.dat` into R using the command

```
mammals <- read.table("https://www.math.ntnu.no/~jarlet/statmod/mammals.dat",
           header=T)
```

This dataset contains the brainsize (in grams) and bodyweight (in kg) of 62 terrestrial mammals. Inspect the dataset by writing `mammals`.

Make the variables contained in the dataframe directly accessible by writing

```
attach(mammals)
```

1. Make a scatterplot (`?plot`) with bodysize along the first axis and brainsize along the second. Does the relationship between brain- and bodysize appear linear?

2. Compute new variables `logbrain` og `logbody` by logtransformation. Use natural logarithms (see `?log`) and make a new scatterplot of log brainsize versus log bodysize. Choose appropriate text labels on each axis (include units in parantheses) using the arguments `xlab` and `ylab` (see `?plot`).

   Based on this plot it might be reasonable to assume that log brainsize depends is normally distributed with expected value depending linearly on log bodysize, that is, the model

   $$\log \text{brain} = \alpha + \beta \log \text{body} + e \tag{1}$$

   where $e \sim N(0, \sigma^2)$. What is the relationship between the original non-transformed variables implied by equation (**??**)?

3. Fit this linear regression model as follows in R

   ```
   linreg <- lm(logbrain~logbody)
   ```

   Inspect the object `linreg`. Also study the more detailed information you get about the fitted model object by writing `summary(linreg)`. What is the estimate of the parameters α, β and σ in equation (**??**)? Is the effect of bodyweight on brainsize statistically significant? Add the estimated regression line to the scatterplot with the command `abline`.

4. Also add small text strings of the name of each species to the plot using the function `text`. Use the argument `cex` to control the fontsize of each text string.

5. Which species has the largest brainsize for its bodyweight, that is, which species deviates the most from the fitted regression line in plot in point 3?

   Based on the fitted model, what is the expected log brain size for this somewhat particular species? Also transform back to the original scale (in grams). From the assumption that each observation is normally distributed around the regression line, compute the probability of a brain size this big or bigger based on the estimated model. You will need to use a statistical table or alternatively, the function `pnorm`.

6. What is the interpretation of the regression coefficient β? In particular, explain why it might be expected that this parameter should take a value of 1. Do a statistical test of this null hypothesis, that is, $H_0 : \beta = 1$ vs. $H_1 : \beta \neq 1$ using the information available from `summary(linreg)`. If your conclusion is that $\beta < 1$, what is your interpretation of this?

7. Finally, go back to the plot of the non-transformed variables and add a curve describing the relationship between (non-transformed) brain- and bodysize (use the function `curve` with the additional argument `add=T`).

Problem 2 According to the central limit theorem, a sum of n independent random variables, if scaled appropriately, approaches a normal distribution in the limit as n goes to infinity. In this exercise we will simulate the distribution of sums of uniformly distributed variables.

10000 realisations of a uniformly distributed random variable U_1 on the interval from 0 to 1 can be simulated as follows

```
u1 <- runif(10000,0,1)
```

or just

```
u1 <- runif(10000)
```

See `?runif`. Make a histogram showing the distribution of the simulated values of U_1 using the `hist` function. Use the optional `breaks` argument if needed.

Using the same method, simulate 10000 realisations of U_2, U_3, U_4 and U_5, all from the same uniform distribution, compute and plot a histogram of the corresponding values of the sum

$$X = U_1 + \cdots + U_n, \tag{2}$$

say, $X = U_1 + U_2$ using the command

```
x <- u1 + u2
hist(x)
```

How does the distribution of X change when you add more terms to the above sum?

For $n = 5$, find the theoretical expected value and standard deviation of X and add a curve showing the appropriate normal density function to the histogram using the command

```
curve(dnorm(x,mean=     ,sd=    ),add=T)
```

Note that you need to use the additional argument `prob=TRUE` when creating the histogram prior to adding the curve to get a single scale on the second axis.

Problem 3 Write an R expression that computes the probability that at least two persons out of 23 have birthdates on the same day of the year. Make the assumption that birthdates are uniformly distributed throughout the year. You may need the to use the `factorial` and `choose` or related functions. Hint: To avoid numerical overflow, you may need to work with the logarithms of the quantities involved whereever possible (carefully read the help pages of the `factorial` and `choose` functions.)