Solution of assignment 1, ST2304

August 1, 2016

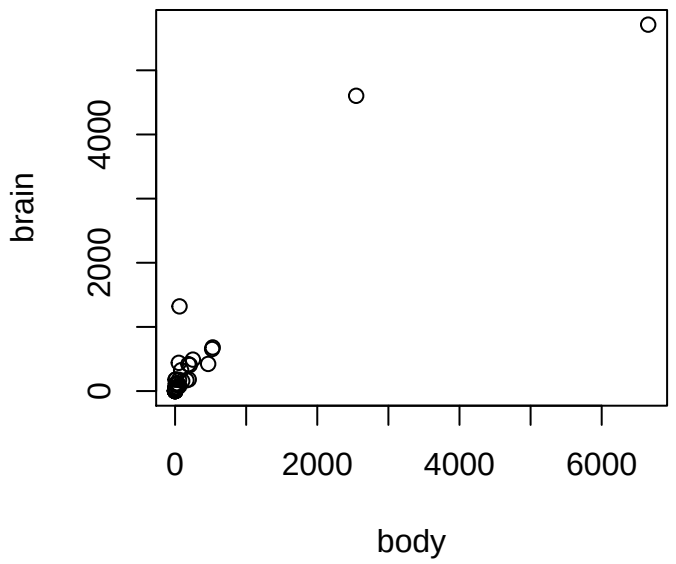
Problem 1

Brain size on average constitutes 0.96% of the total body weight.

```
> # Load
> mammals <- read.table("https://www.math.ntnu.no/~jarlet/statmod/mammals.dat",
+                       header=T)
>
> # Attach to search path. I.e. make variables available by name.
> attach(mammals)


> round(mean((brain/1000)/body)*100, 2)
[1] 0.96
```

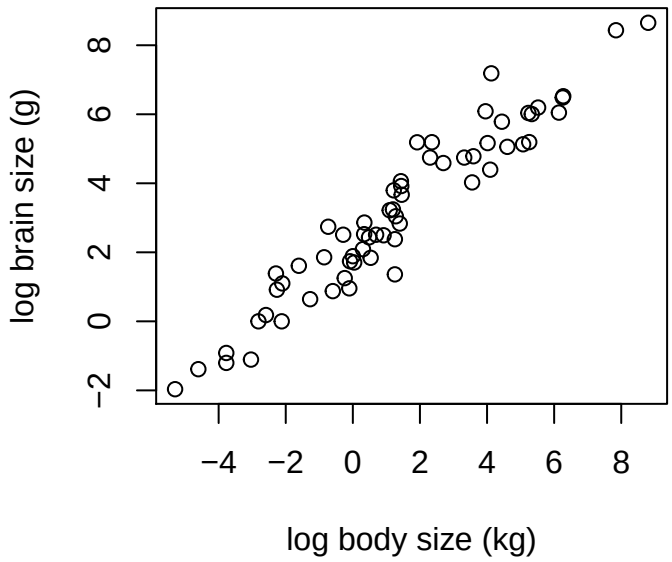
1. A simple scatterplot visualizes the relationship between the two variables. It is somewhat hard to see if the relationship between untransformed variables is linear since the distributions of both variables are highly skewed.

```
> plot(x=body, y=brain)
```



2. The variables were log-transformes using the function `log()`. Natural logarithms are the default option (see `?log()`). In the following scatterplot, log-transformed variables are shown and appropriate axis labels have been added.

```
> logbody <- log(body)
> logbrain <- log(brain)
>
> plot(x=logbody, y=logbrain,
+      xlab="log body size (kg)",
+      ylab="log brain size (g)")
```



In terms of the original untransformed variables the relationship between brain and body size becomes

$$\text{brain} = \exp^{\alpha+\beta \log \text{brain}} = \exp^{\alpha}\text{body}^{\beta} = \alpha'\text{body}^{\beta}. \tag{1}$$

where exp denotes the exponential function.

3. We fit the linear regression model $\log \text{body} = \alpha + \beta \log \text{brain} + e$ and inspect the estimated parameter values.

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

Call:
lm(formula = logbrain ~ logbody)

Coefficients:
(Intercept)      logbody
     2.1348       0.7517
```

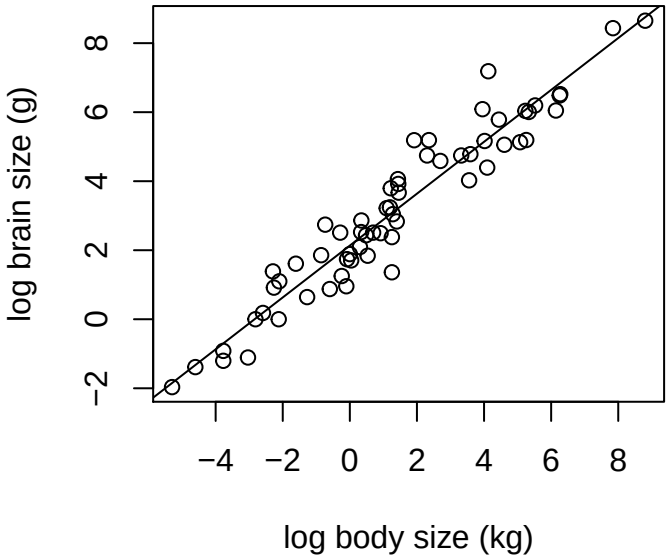
The model summary for this model produced by R provides us with additional details and tests of significance.

```
> summary(linreg)

Call:
lm(formula = logbrain ~ logbody)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71550 -0.49228 -0.06162  0.43597  1.94829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13479    0.09604   22.23   <2e-16 ***
logbody      0.75169    0.02846   26.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-squared:  0.9208,Adjusted R-squared:  0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

We see that the parameter estimates are $\hat{\alpha} = 2.13$, $\hat{\beta} = 0.75$ and $\hat{\sigma} = 0.69$ (the "residual standard error"). Log body size has a significant effect on log brain size (the P value for the test is much less than 0.05, a commonly chosen level of significance). It might be noted that the estimated slope is independent on whether natural or base 10 logarithms are used and whether both variables are in the same units or not. However, the estimated intercept is dependent on these decisions.

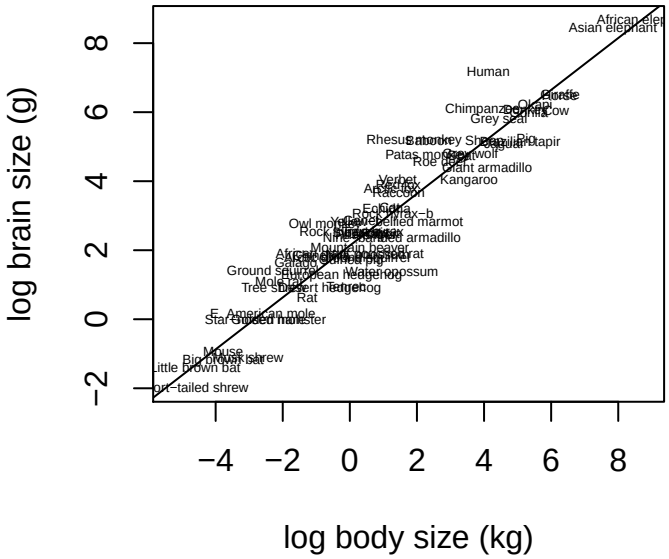
Now we can add the estimated regression line to the scatterplot of brain size on body size.

```
> plot(x=logbody, y=logbrain,
+      xlab="log body size (kg)",
+      ylab="log brain size (g)")
>
> abline(linreg)
```



4. To allow one to see which species that are located where in the graph one may add species names as labels in the graph. By also adding `type="n"` in the R-code we avoid plotting the species names on top of the default bullets in the graph.

```
> plot(x=logbody, y=logbrain,
+      xlab="log body size (kg)",
+      ylab="log brain size (g)",
+      type="n")
>
> abline(linreg)
> text(logbody,logbrain,species,cex=.4)
```



5. The human species has the largest deviation from the estimated regression line.

```
> # Human observed
> obs_human_brain <- logbrain[species == "Human"]
> obs_human_brain
[1] 7.185387
> obs_human_body <- logbody[species == "Human"]
> obs_human_body
[1] 4.127134
> # Human predicted
> pred_human_brain <- coef(linreg)[1] +
+   coef(linreg)[2] * logbody[species == "Human"]
> pred_human_brain
(Intercept)
   5.237098
> # Original scale
> exp(pred_human_brain)
(Intercept)
   188.1233
```

The log brain size of humans are 7.19. The expected value of log brain size in humans based on the log body size in humans of 4.13 and the fitted model becomes $\hat{\alpha} + \hat{\beta} \times \log$ body size $= 2.13 + 0.75 \times 4.13 = 5.24$. On the original non-transformed scale, the predicted brain size is $exp^{5.24} = 188$ grams, while the observed brain size is $exp^{7.19} = 1320$ grams.

```
> # Probability of brain larger than observed
> prob_human_brain <- pnorm(q=obs_human_brain,
+                           mean=pred_human_brain,
+                           sd=summary(linreg)$sigma,
+                           lower.tail=FALSE)
> # P-value
> prob_human_brain
[1] 0.002506931
> # As percent
> round(prob_human_brain * 100, 2)
[1] 0.25
```

According to the regression model log brain size is normally distributed with the expectation equal to the predicted value and standard deviation equal to $\hat{\sigma} = 0.69$. From this we find that the probability that log brain size is equal or greater than the observed value of 7.19, $P(\log \text{brain} > 7.19)$ is 0.25%. That is, very small. Some authors, e.g. Geoffrey Miller have suggested that large brain size in Humans evolved as a result of runaway selection.

6. According to equation 1, for $\beta = 1$ brain size is directly proportional to body size, whereas for $\beta < 1$, larger species tend to have disproportionately smaller brain sizes.

```
> # Sample size
> n <- nrow(mammals)
> n
[1] 62
```

A test of $H_0 : \beta = 1$ vs $H_1 : \beta \neq 1$ can be based on the test statistic

$$T = \frac{\hat{\beta} - \beta_0}{\widehat{SE(\hat{\beta})}} = \frac{\hat{\beta} - 1}{\widehat{SE(\hat{\beta})}} \tag{2}$$

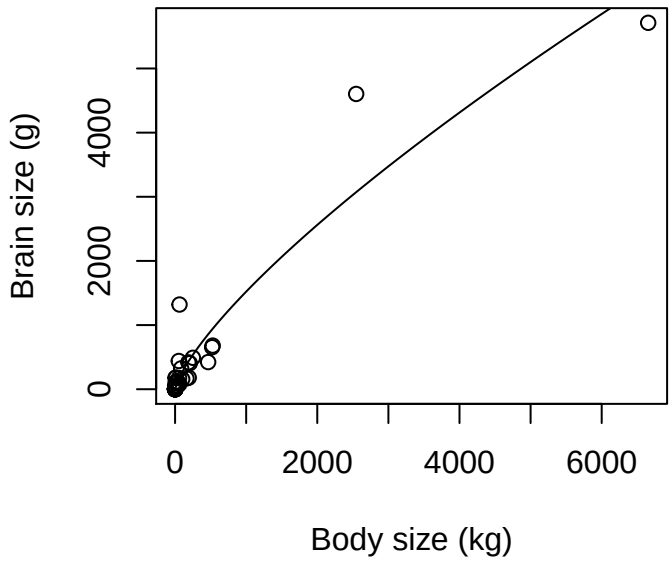
which is t-distributed with $n - 2 = 62 - 2 = 60$ degrees of freedom under H_0. We can extract the necessary estimates from the summary table, then calculate the t- and P-value for the test.

```
> # Extract estimates from summary table
> beta <- coef(summary(linreg))[2, 1]
> beta_se <- coef(summary(linreg))[2, 2]
>
> # T-test
> tval <- (beta-1)/beta_se
> tval
[1] -8.723929
> pval <- 2*pt(tval, df = n-2)
> pval
[1] 2.884369e-12
```

Given the observed t-value of $T = -8.72$ the probability under H_0 that T takes the observed or a more extreme value (the P-value for the test) becomes $2 \times P(T < -8.72) < 0.0001$. If we chooce a level of significance $\alpha = 0.05$ we can thus reject the null hypothesis that brain size is directly proportional to body size in favour of H_1. The estimated $\hat{\beta} = 0.75$ indicates that mammals with large body size have disproportionally smaller brains. Curiously, metabolic rate has the same allometric relationship to body size as brain size.

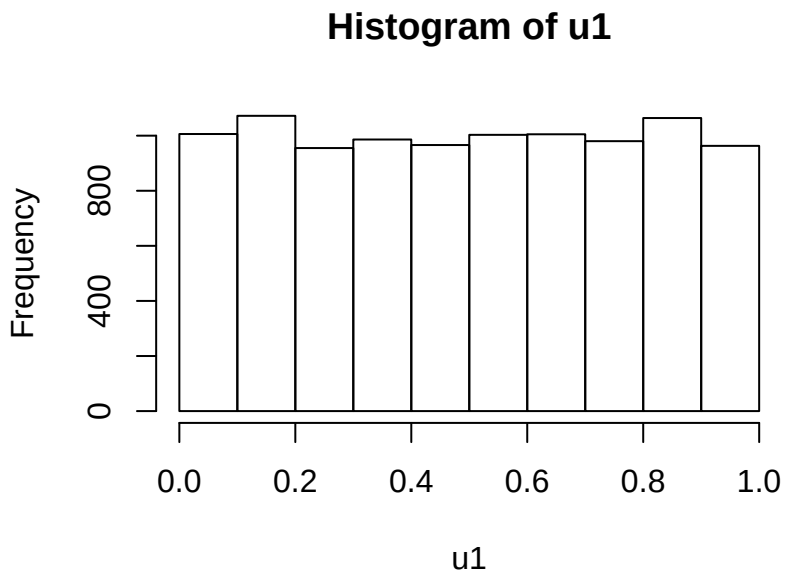
7. The relationship between brain and body size is shown graphically below on the original scale. The curve corresponding to equation 1 has been added to the graph.

```
> # Create graph
> plot(x=body, y=brain,
+      xlab="Body size (kg)",
+      ylab="Brain size (g)")
>
> # Coefficients
> coef(linreg)[1]
(Intercept)
   2.134789
> coef(linreg)[2]
  logbody
0.7516859
> # Add curve to the graph
> curve(expr = exp(2.134789)*x^0.7516859, add=TRUE)
```



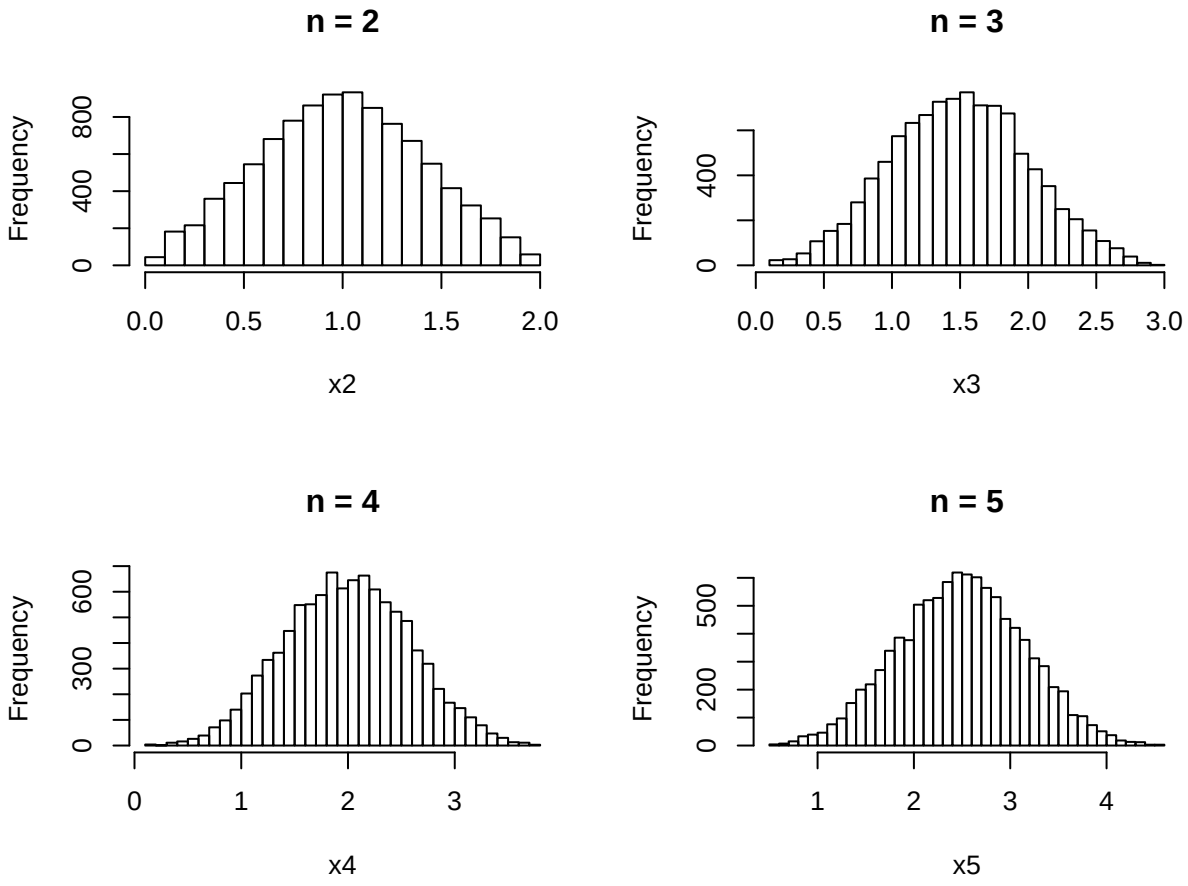
Problem 2
We made 10'000 realisations of the uniformly distributed random variables $U_1, U_2, ..., U_5$. A histogram U_1 is shown.

```
> # Simulations
> u1 <- runif(10000)
> u2 <- runif(10000)
> u3 <- runif(10000)
> u4 <- runif(10000)
> u5 <- runif(10000)
>
> # Histogram
> hist(u1, breaks = 10)
```



Histograms of $X_n = U_1 + ... + U_n$ for $n = (2, .., 5)$ is shown below. `par(mfrow=c(2,2))` alines the graphs in two columns and rows.

```
> x2 <- u1 + u2
> x3 <- x2 + u3
> x4 <- x3 + u4
> x5 <- x4 + u5
>
> par(mfrow=c(2,2))
> hist(x2, breaks = 20, main = "n = 2")
> hist(x3, breaks = 30, main = "n = 3")
> hist(x4, breaks = 40, main = "n = 4")
> hist(x5, breaks = 50, main = "n = 5")
```



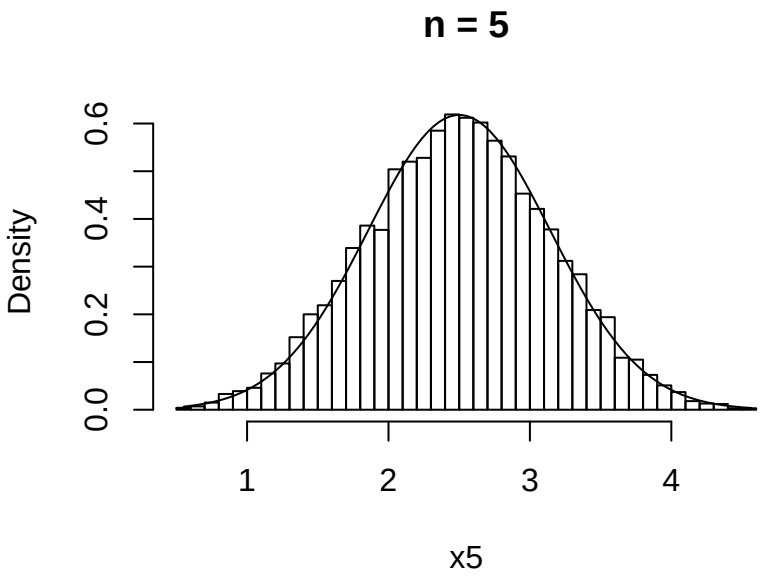
We see that as we increase the number of terms in the sum the distribution approaches a normal distribution.

```
> # Expectation and variance of uniform distribution
> eu <- (1/2)*(0+1)
> vu <- (1/12)*(1-0)^2
> eu
[1] 0.5
> vu
[1] 0.08333333
> # Expectation and variance of the
> # sum of 5 independent variables from
> # the same distribution
> ex <- 5*eu
> vx <- 5*vu
> ex
[1] 2.5
> vx
[1] 0.4166667
```

The uniform distribution has expected value and variance given by $E(U) = \mu = \frac{1}{2}(a+b) = 0.5$ and $V(U) = \sigma^2 = \frac{1}{12}(b-a)^2 = 0.08$, where a = minimum value and b = maximum value of the distribution. From the central limit theorem we know that the sum (X) of $n = 5$ independent random variables, approaches a normal distribution with $E(X) = n\mu = 5 \times 0.5 = 2.5$ and $V(X) = n\sigma^2 = 5 \times 0.08 = 0.42$.

Below is the histrogram with probability densities of the sum of five uniformly distributed variables with the theoretically expected normal distribution overlain.

```
> hist(x5, breaks = 50, main = "n = 5", freq = FALSE)
> curve(dnorm(x, mean = 5*eu, sd = sqrt(5*vu)),
+       from = min(x5), to = max(x5), add = TRUE)
```



Problem 3
Let A denote the event that at least two persons have birthdays on the same day. Based on a combinatorical argument, the probability of the complement of this ($\bar{A}$), that all birthdays are on different days become

$$\begin{aligned} P(\bar{A}) &= \frac{\text{Number of outcomes in } \bar{A}}{\text{Number of outcomes in S}} \\ &= \frac{365 \times 364 \times \cdots \times (365 - 23 + 1)}{365^{23}} = \frac{365!/(356 - 23)!}{365^{23}} \end{aligned} \tag{3}$$

where S denote all possible combinations of birthdays. This exercise is really about how we can do numerical computations involving very small numbers (e.g. probabilites) or large numbers (e.g. in combinatorics). If we try to evalutate the above expression in R we get

```
> factorial(365)/factorial(356-23)/356^23
[1] NaN
> factorial(365)
[1] Inf
> Inf/Inf
[1] NaN
```

that is "not a number". This error arise because 365! is larger than the largest double precision decimal number R can handle,

```
> .Machine$double.xmax
[1] 1.797693e+308
```

so the most sensible thing R can do is to handle the numerator and denominater as infinite represented by `Inf` in R. However, there is no way R can know the value of `Inf/Inf`, thus we get `NaN`.

The way around this problem is to work with logarithms of the quantities appearing in the above fraction by rewriting (3) to the following form

$$\frac{365!/(356-23)!}{365^{23}} = \exp(\ln \frac{365!/(365-23)!}{365^{23}}) = \exp(\ln 365! - \ln 342! - 23 \ln 365) \qquad (4)$$

If we study the help page of `factorial` we see that `lfactorial` computes $\ln x!$, for example,

```
> lfactorial(365)
[1] 1792.332
```

Expression (4) can thus be written as follows in R

```
> exp(lfactorial(365)-lfactorial(342)-23*log(365))
[1] 0.4927028
```

Hence, the probability of A, $P(A) = 1 - P(\hat{A}) = 1 - 0.493 = 0.507$.

Many functions in R optionally computes logarithmic values, in some cases by by specifying an optional `log=TRUE` argument, e.g. `pnorm` and `dnorm`. This is sometimes needed to avoid numerical underflow, for example, in computations of a log likelihood.