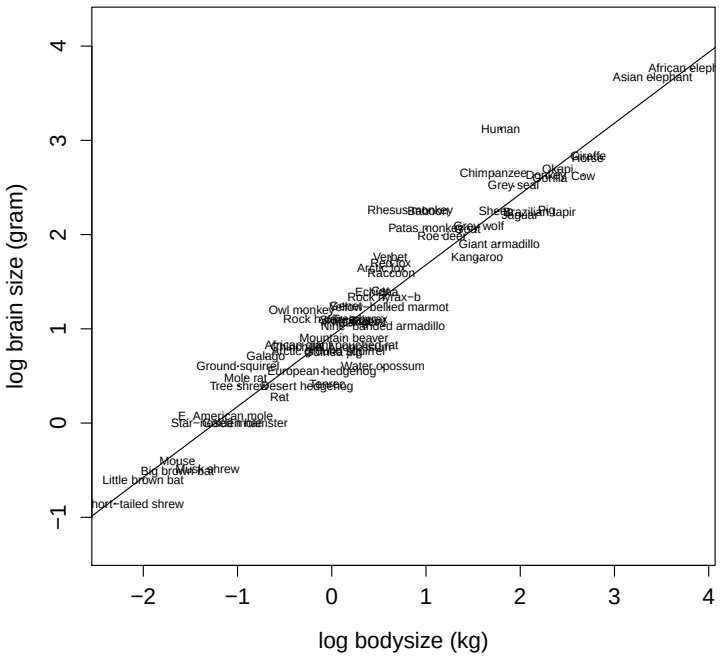
Solution of assignment 1, ST2304

Problem 1 Brain size on average constitutes 0.96% of the total body weight.

It is somewhat hard to see if the relationship between untransformed variables is linear since the distributions of both variables are highly skewed.

The following plot shows the log-transformed variables and the fitted linear regression model $\log_{10}\text{body} = \alpha + \beta \log_{10}\text{brain} + e$ with species names added to the plot



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

$$\text{brain} = 10^{\log_{10}\text{brain}} = 10^{\alpha+\beta\log_{10}\text{brain}} = 10^{\alpha}\text{body}^{\beta} = \alpha'\text{body}^{\beta}. \tag{1}$$

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

This model summary produced by R when fitting this model,

```
> summary(linreg)

Call:
lm(formula = logbrain ~ logbody)

Residuals:
     Min       1Q   Median       3Q      Max
-0.74503 -0.21380 -0.02676  0.18934  0.84613

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.92713    0.04171   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: 0.3015 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
```

shows that the parameter estimates are $\hat{\alpha} = 0.9271$, $\hat{\beta} = 0.7517$ and $\hat{\sigma} = 0.3015$ (the "residual standard error"). Log body size has a significant effect on log brain size (the P value for the test is less $2 \cdot 10^{-16}$. It might be noted that the estimated intercept but not the estimated slope will depend on whether natural or base 10 logarithms are used and whether both the variables are in the same units or not.

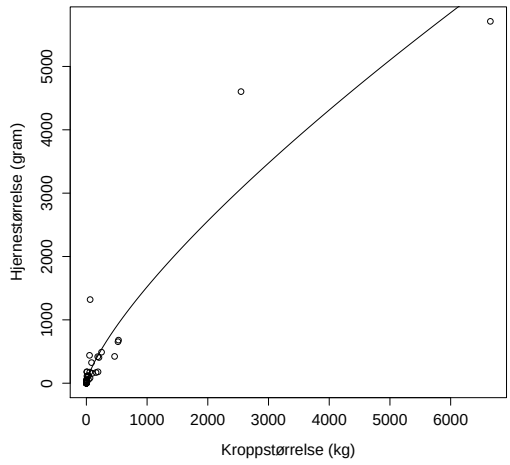
The human species has the largest deviation from the estimated regression line. The expected value of log brain size in humans based on the log body size in humans of 1.79 and the fitted model becomes $0.92 + 0.75 \cdot 1.79 = 2.27$ (in terms of the original non-transformed brain size variable, this corresponds to $10^{2.27} = 186$ grams). According to the regression model log brain size is normally distributed with this expectation and standard deviation equal to 0.3015. From this we find that the probability that log brain size is equal or greater than the observed value of 3.12, $P(\log \text{brain} > 3.12)$ 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.

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

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

which is t-distributed with $n-2 = 62-2 = 60$ degrees of freedom under H_0. Given the observed value of $T = -8.72$ the corresponding P-value for the test, that is, the probability under H_0 that T takes the observed or a more extreme value becomes $2P(T < -8.72) = 2.87 \cdot 10^{-12}$. 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.

Brain size plotted against the body size:



R-code

```
mammals <- read.table("http://www.math.ntnu.no/~jarlet/statmod/mammals.dat",
  header=T)
attach(mammals)
mean((brain/1000)/body)
```

```
logbody <- log(body,10)
logbrain <- log(brain,10)
linreg <- lm(logbrain~logbody)
summary(linreg)
plot(logbody,logbrain,
     xlab="log bodysize (kg)",
     ylab="log brain size (gram)",cex=.1,asp=1)
abline(linreg)
text(logbody,logbrain,species,cex=.5)
# probability of observed or greater log human brain size
pnorm(logbrain[species=="Human"],
      mean=0.9271+0.75169*logbody[species=="Human"],
      sd=0.3015,
      lower.tail=F)
# test of H0: beta=1 vs H1: beta<>1
T <- (0.75169 - 1)/0.02846 # the observed value of the test-statistic
2*pt(T,df=62-2) # the p-value of test

# plot of oringal variables and estimated relationsship between them
plot(body,brain,xlab="Kroppstørrelse (kg)",ylab="Hjernestørrelse (gram)")
curve(10^0.9271*x^0.75169,add=T)
detach(data)
```

Problem 2 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, that all birthdays are on different days become

$$P(\bar{A}) = \frac{\text{Number of outcomes in } \bar{A}}{\text{Number of outcomes in S}} = \frac{365 \cdot 364 \ldots (365 - 23 + 1)}{365^{23}} = \frac{365!/(356 - 23)!}{365^{23}} \qquad (3)$$

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
Warning messages:
1: In factorial(365) : value out of range in 'gammafn'
2: In factorial(356 - 23) : value out of range in 'gammafn'
> factorial(365)
[1] Inf
Warning message:
In factorial(365) : value out of range in 'gammafn'
> 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

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

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

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

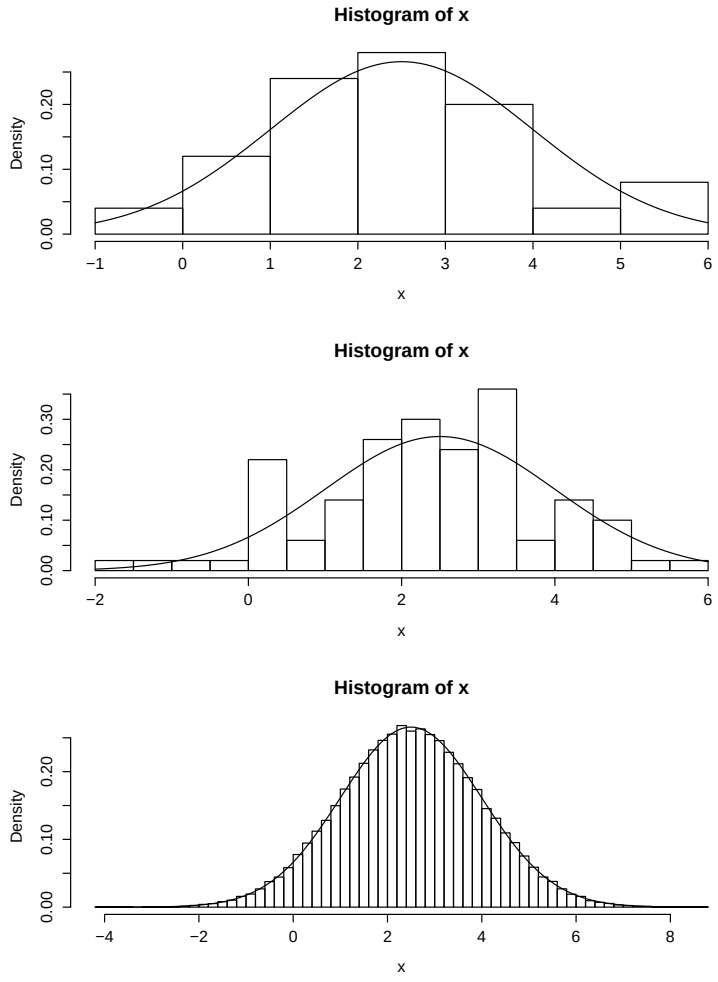
Expression (4) can thus be written as follows in R

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

Hence, the probability of A, $P(A) = 0.51$.

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.

Problem 3 Histogram of respectively 50, 100 og 100000 realisations from a normal distribution with mean 2.5 og standard deviation 1.5 (solid curves).



R-kode:

```
par(mfrow=c(3,1))
x <- rnorm(50,mean=2.5,sd=1.5)
hist(x,freq=F)
curve(dnorm(x,mean=2.5,sd=1.5),add=T)
x <- rnorm(100,mean=2.5,sd=1.5)
hist(x,freq=F,breaks=20)
curve(dnorm(x,mean=2.5,sd=1.5),add=T)
x <- rnorm(100000,mean=2.5,sd=1.5)
hist(x,freq=F,breaks=50)
curve(dnorm(x,mean=2.5,sd=1.5),add=T)
```