Assignment 2

Name: James La Fontaine

Student Number: 1079860

Tutorial Day and Time: Friday 2:15 PM – 4:15 PM

Tutor's Name: Haoyu Yang

Question 1

1a) $n=9$, $\sigma=0.6$, $\bar{x}=8$

$$Pr\left(-c < \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < c\right) = 0.95 \quad \text{where } c = \Phi^{-1}(0.975)$$

$$Pr\left(-1.96 < \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < 1.96\right) = 0.95$$

$$Pr\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

A 95% CI for μ is $\left(8 - 1.96\frac{0.6}{3},\ 8 + 1.96\frac{0.6}{3}\right)$

$$= (7.608,\ 8.392)$$

1b) $\frac{\text{width}}{2} = c\frac{\sigma}{\sqrt{n}}$

$$\frac{0.2}{2} = 1.96\frac{0.6}{\sqrt{n}}$$

$$n = \left(\frac{1.96 \times 0.6}{0.1}\right)^2 = 138.298 \approx 139 \text{ samples required}$$

1c)

$$Pr\left(-c < \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} < c\right) = 0.95 \qquad \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

$$Pr\left(\bar{X} - c\frac{S}{\sqrt{n}} < \mu < \bar{X} + c\frac{S}{\sqrt{n}}\right) = 0.95$$

$c = 2.306 \qquad S^2 = 0.425 \qquad S = 0.652$

A 95% CI for μ is $\left(8 - 2.306\frac{0.652}{3}, 8 + 2.306\frac{0.652}{3}\right)$

$= (7.499, 8.501)$

This confidence interval is slightly wider than the confidence interval from part (a) due to the sample standard deviation being higher than the assumed σ.

Question 2

2a)

$$n = \frac{c^2 \hat{p}(1-\hat{p})}{\epsilon^2}$$

assume $\hat{p} = 0.8$ as the lowest sample proportion we expect and use it as our estimate to cover the case with maximum uncertainty

$$\Rightarrow n = \frac{1.96^2 \times 0.8 \times 0.2}{0.05^2} \approx 246 \text{ samples required}$$

2b) $n = \frac{1.96^2 \times 0.8 \times 0.2}{0.02^2} \approx 1537$ samples required

Question 3

3a)

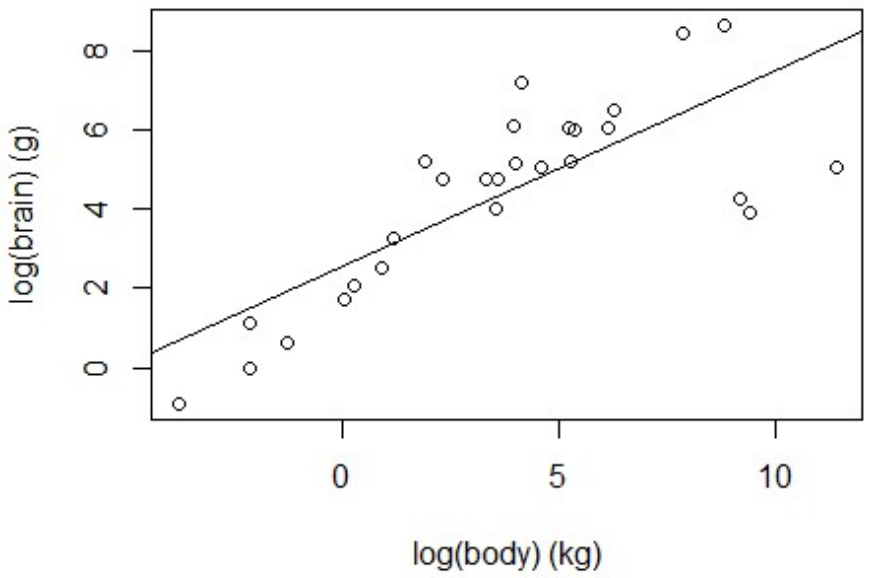
The difference in scale between the raw brain and body measurements is too significant to fit a suitable model and to produce a suitable plot to compare their relationship

3b)

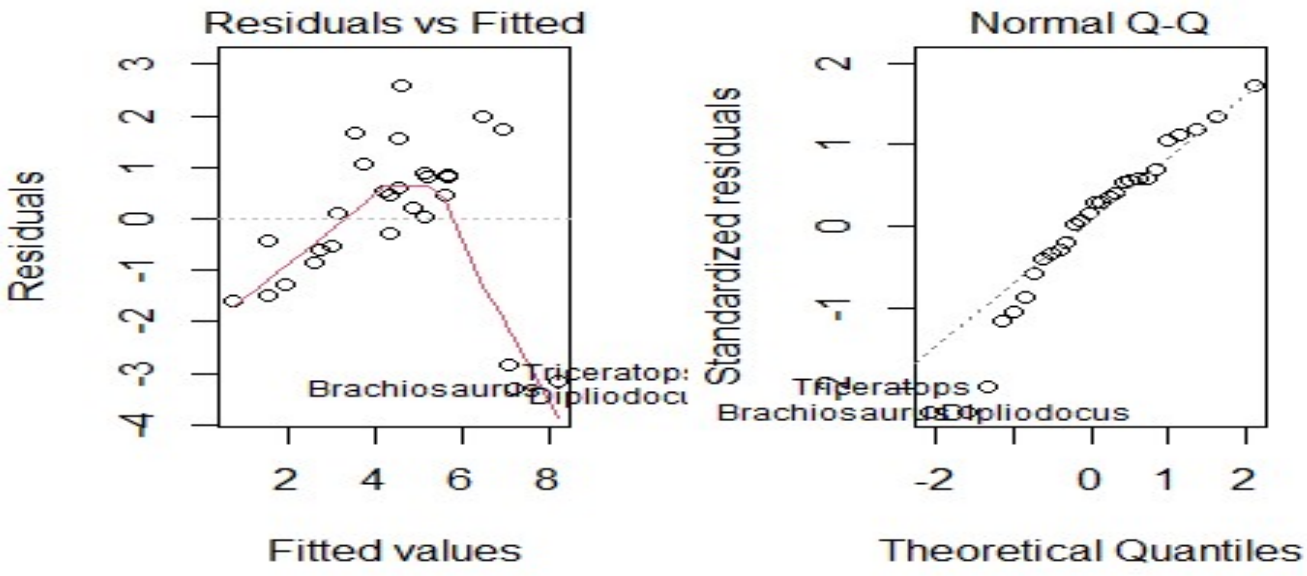
```
data(Animals, package = "MASS")

brain = log(Animals$brain)
body = log(Animals$body)

model1 = lm(brain ~ body, data = log(Animals))

summary(model1)

##
## Call:
## lm(formula = brain ~ body, data = log(Animals))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2890 -0.6763  0.3316  0.8646  2.5835
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.55490    0.41314   6.184 1.53e-06 ***
## body         0.49599    0.07817   6.345 1.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.532 on 26 degrees of freedom
## Multiple R-squared:  0.6076, Adjusted R-squared:  0.5925
## F-statistic: 40.26 on 1 and 26 DF,  p-value: 1.017e-06
```

3c)

```
plot(body, brain, xlab = "log(body) (kg)", ylab = "log(brain) (g)")
abline(model1)
```




```
par(mfrow = c(1, 2))
plot(model1, 1:2)
```



It seems quite plausible to assume that the residuals are normally distributed according to the QQ plot, and the model plot indicates that this linear regression model appears to represent the relationship between an animal's body and brain moderately well. Despite this, the residuals vs fitted values plot produces a systematic 'U' pattern which implies that the assumption of linearity isn't holding. However, this is clearly caused by the outliers in the data.

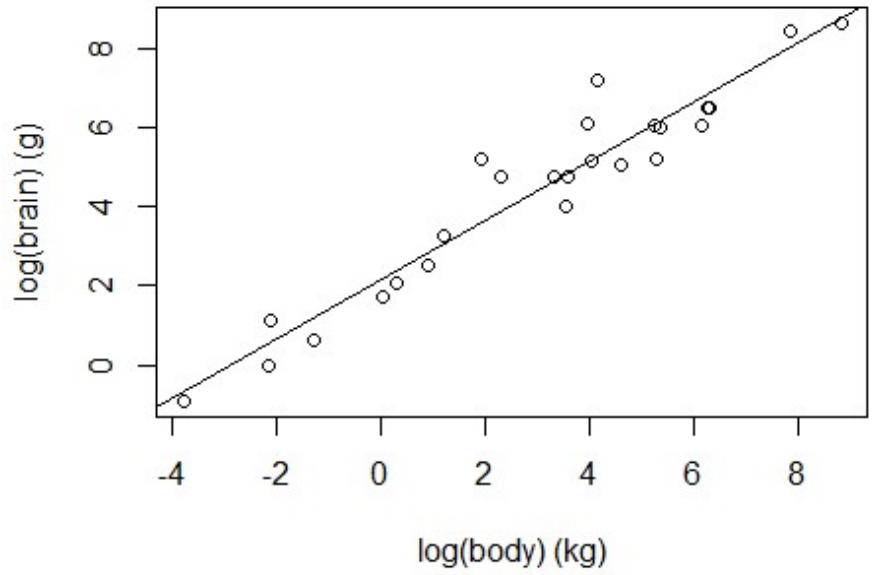
3d)

```
AnimalsNoDinosaurs = Animals[-c(6, 16, 26),]
brain2 = log(AnimalsNoDinosaurs$brain)
body2 = log(AnimalsNoDinosaurs$body)
model2 = lm(brain2 ~ body2, data = log(AnimalsNoDinosaurs))

summary(model2)
```

```
##
## Call:
## lm(formula = brain2 ~ body2, data = log(AnimalsNoDinosaurs))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9125 -0.4752 -0.1557  0.1940  1.9303
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.15041    0.20060   10.72 2.03e-10 ***
## body2        0.75226    0.04572   16.45 3.24e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7258 on 23 degrees of freedom
## Multiple R-squared:  0.9217, Adjusted R-squared:  0.9183
## F-statistic: 270.7 on 1 and 23 DF,  p-value: 3.243e-14
```

3e)

```
plot(body2, brain2, xlab = "log(body) (kg)", ylab = "log(brain) (g)")
abline(model2)
```



3f)

```
newdata = data.frame(body2 = 500)

predict(model2, newdata, interval = "confidence")

##        fit      lwr      upr
## 1 378.2808 331.2781 425.2834
```

A 95% confidence interval for the average brain weight (in grams) of camels is

(331.2781, 425.2834)

Question 4

$n=8 \quad m=12 \quad \bar{x}=8.21 \quad \bar{y}=7.36 \quad S_x=1.610 \quad S_y=0.956$

Let enriched air plant growth = X
normal air- plant growth = Y

$$W=\frac{\bar{X}-\bar{Y}-(\mu_x-\mu_y)}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}} \approx t_{\frac{\left(\frac{S_x^2}{n}+\frac{S_y^2}{m}\right)^2}{\frac{S_x^4}{n^2(n-1)}+\frac{S_y^4}{m^2(m-1)}}}$$

$$\approx t_{10.31}$$

$$Pr\left(-c < \frac{\bar{X}-\bar{Y}-(\mu_x-\mu_y)}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}} < c\right)=0.95$$

$$Pr\left(\bar{X}-\bar{Y}-c\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}} < \mu_x-\mu_y < \bar{X}-\bar{Y}+c\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}\right)$$

$$=0.95$$

A 95% CI for $\mu_x-\mu_y$ is

$$\left(8.21-7.36-2.219\sqrt{\frac{1.61^2}{8}+\frac{0.956^2}{12}},\ 8.21-7.36+2.219\sqrt{\frac{1.61^2}{8}+\frac{0.956^2}{12}}\right)$$

$$=(0.501,\ 1.199)$$

Therefore there is strong evidence that a CO_2-enriched atmosphere increases plant growth.

Question 5

5a)

H_0: p1 = p2

H_1: p1 ≠ p2

5b)

```
x = c(120, 60)
n = c(800, 600)

p1 = prop.test(x, n)

p1

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 7.2105, df = 1, p-value = 0.007248
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.01406774 0.08593226
## sample estimates:
## prop 1 prop 2
##   0.15   0.10

z_statistic = sqrt(p1$statistic)

names(z_statistic) = NULL

cat('z_statistic:', z_statistic, '\n')

## z_statistic: 2.68524

critical_value = qnorm(1-0.05/2)

cat('critical value:', critical_value, '\n')

## critical value: 1.959964
```

The observed value of the Z-statistic is 2.6852. When $\alpha = 0.05$, the rejection region for this test is $|z| > 1.96$. Therefore, we reject H_0 and can conclude that there is evidence suggesting that the rates of babies with low birthweight differ between Africa and the Americas. More specifically, there is evidence that the rate is higher in Africa than the Americas.

5c)

```
p2 = prop.test(x, n, conf.level = 0.99)

p2

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 7.2105, df = 1, p-value = 0.007248
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  0.003235253 0.096764747
## sample estimates:
## prop 1 prop 2
##   0.15   0.10

z_statistic = sqrt(p2$statistic)

names(z_statistic) = NULL

cat('z_statistic:', z_statistic, '\n')

## z_statistic: 2.68524

critical_value = qnorm(1-0.01/2)

cat('critical value:', critical_value, '\n')

## critical value: 2.575829
```

If $\alpha = 0.01$, then the rejection region for the test is $|z| > 2.5758$. Therefore, we still reject H_0 and conclude that there is evidence suggesting that the low birthweight rates differ between the two continents.

5d)

```
conf_int = p1$conf.int

cat('95% confidence interval for difference in rates: (',round(conf_int[1],
4),',', round(conf_int[2], 4),')')

## 95% confidence interval for difference in rates: ( 0.0141 , 0.0859 )
```

Question 6

6a)

```
significance_level = pgeom(4, 0.4, lower.tail = FALSE) + dgeom(4, 0.4)
cat('The probability of committing a Type I Error is', significance_level)
```

The probability of committing a Type I Error is 0.1296

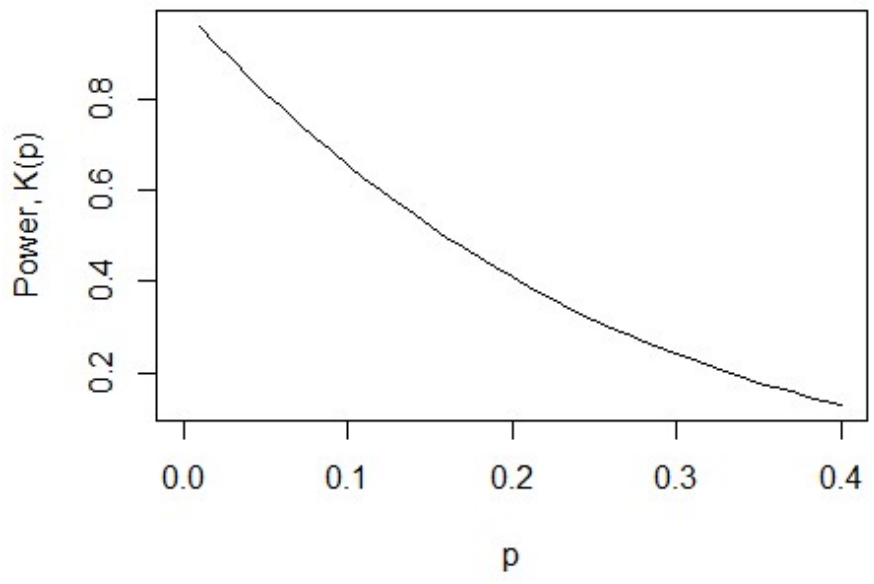
6b)

```
beta = pgeom(4, 0.2) - dgeom(4, 0.2)
cat('The probability of committing a Type II Error is', beta)
```

The probability of committing a Type II Error is 0.5904

6c)

```
K1 = function(p)
    1 - (pgeom(4, p) - dgeom(4, p))
p = seq(0, 0.4, 0.01)
K = K1(p)

plot(p, K, type = 'l', ylab = 'Power, K(p)')
```



6d)

```
cat('significance level for critical value = 4:', pgeom(4, 0.4, lower.tail =
FALSE) + dgeom(4, 0.4), '\n')

## significance level for critical value = 4: 0.1296

cat('significance level for critical value = 5:', pgeom(5, 0.4, lower.tail =
FALSE) + dgeom(5, 0.4), '\n')

## significance level for critical value = 5: 0.07776

cat('significance level for critical value = 6:', pgeom(6, 0.4, lower.tail =
FALSE) + dgeom(6, 0.4), '\n')

## significance level for critical value = 6: 0.046656

cat('significance level for critical value = 7:', pgeom(7, 0.4, lower.tail =
FALSE) + dgeom(7, 0.4), '\n')

## significance level for critical value = 7: 0.0279936
```

Therefore, a test where the null hypothesis is rejected if the observed value of X is greater than or equal to 6 gives an approximate significance level of 0.05. The actual significance level of this test is 0.046656.