

Assignment 2

● Graded

Student

James La Fontaine

Total Points

18.5 / 20 pts

Question 1

Question 1

6 / 6 pts

1.1 1(a) 2 / 2 pts

Assuming $\sigma = 0.6$ based on previous experience, calculate a 95% confidence interval for μ .

✓ + 2 pts Perfect: CI is (7.608,8.392) with working

+ 1 pt Reasonable attempt but error (e.g. using a T distribution) or correct answer but no working

+ 0 pts Incorrect/no attempt

1.2 1(b) 2 / 2 pts

Still assuming $\sigma = 0.6$, suppose we want a 95% confidence interval (for μ) that has width at most 0.2. How many experiments do we need to run?

✓ + 2 pts Perfect: the sample size needs to be at least 139, with working

+ 1 pt Reasonable attempt but incorrect (e.g. stating final answer is n=138.3 or rounding down to 138 or got 11.76 because forgot to square or got 35 because used 0.2), or answer correct but no working

+ 0 pts Incorrect/no attempt

1.3 1(c) 2 / 2 pts

If σ is unknown, calculate a 95% confidence interval for μ . Comment on how this compares with the confidence interval from part (a).

✓ + 2 pts Perfect (7.5,8.5) using $t=2.306$ or 2.31 AND comment 'The width of the CI is a little wider than the CI from part (a).'

+ 1 pt Reasonable attempt but no comment or some mistake e.g. using $c=1.96$ OR answer correct but no working

+ 0 pts No attempt/incorrect

Question 2

Question 3

6.5 / 8 pts

2.1 3(a)

0.5 / 1 pt

- + 1 pt Good comment e.g. 'brain and body weight do not have a linear correlation/relationship/don't fit on trend line', 'data are highly skewed' or 'they are bounded below by zero', minuscule correlation, or fit a model on raw data and concluded R-squared is small, p-value for body is insignificant, etc.

The raw data are unsuitable for fitting this model. Why?

- ✓ + 0.5 pts Comment shows some understanding but could be improved e.g. 'there's no obvious relationship between body and brain weight', 'data not evenly spread' or similar

- + 0 pts Incorrect or no attempt (e.g. simply stating 'data has a wide range/variance', 'the distribution is a bad fit, errors would over' or giving summary stats is not a valid attempt)

2.2 3(b)

1.5 / 1.5 pts

- ✓ + 0.5 pts Transforming data to log scale e.g. `LogAnimals <- log(Animals)`. It's ok to do this within the `lm` function.

- ✓ + 0.5 pts Fitting a model with `log(brain)` or `brain` as the response (award even if didn't log data)
`m1 <- lm(brain ~ body, data = LogAnimals)`

- ✓ + 0.5 pts Applying the summary function and returning output (award even if model wrong e.g. flipped x & y and/or no logs unapplied)

```
summary(m1)

##
## Call:
## lm(formula = brain ~ body, data = LogAnimals)
##
## 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
```

- + 0 pts Incorrect or no attempt

2.3 3(c)

1.5 / 1.5 pts

- ✓ + 0.5 pts At least **TWO** diagnostic plots (a plot of x against y counts). If fit the wrong model, but still presenting two plots, award mark.

Comment

- ✓ + 1 pt Three points are different to the rest: Triceratops, Brachiosaurus and Diplodocus (or 'the dinosaurs'). Note some students have made this comment in part d, that's ok too.

+ 0.5 pts Reasonable comment (e.g. analysing diagnostics) but didn't state which points are unusual (i.e. dinosaurs) AND don't have them labelled on the graph. It's **not** enough to just state that points 6, 16 and 26 are different.

+ 0 pts Comment missing or totally irrelevant

+ 0 pts Incorrect or no attempt

2.4 3(d)

1 / 1 pt

- ✓ + 0.5 pts Omitting three outlier points

- ✓ + 0.5 pts Refitting model on updated data & providing summary output for new model (numbers need to match below to get mark)

```
LogAnimals2 <- LogAnimals[-c(6, 16, 26), ] # omit dinosaurs
m2 <- lm(brain ~ body, data = LogAnimals2)
summary(m2)

##
## Call:
## lm(formula = brain ~ body, data = LogAnimals2)
##
## 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 ***
## body        0.75226    0.04572   16.45 3.24e-14 ***
## ---
```

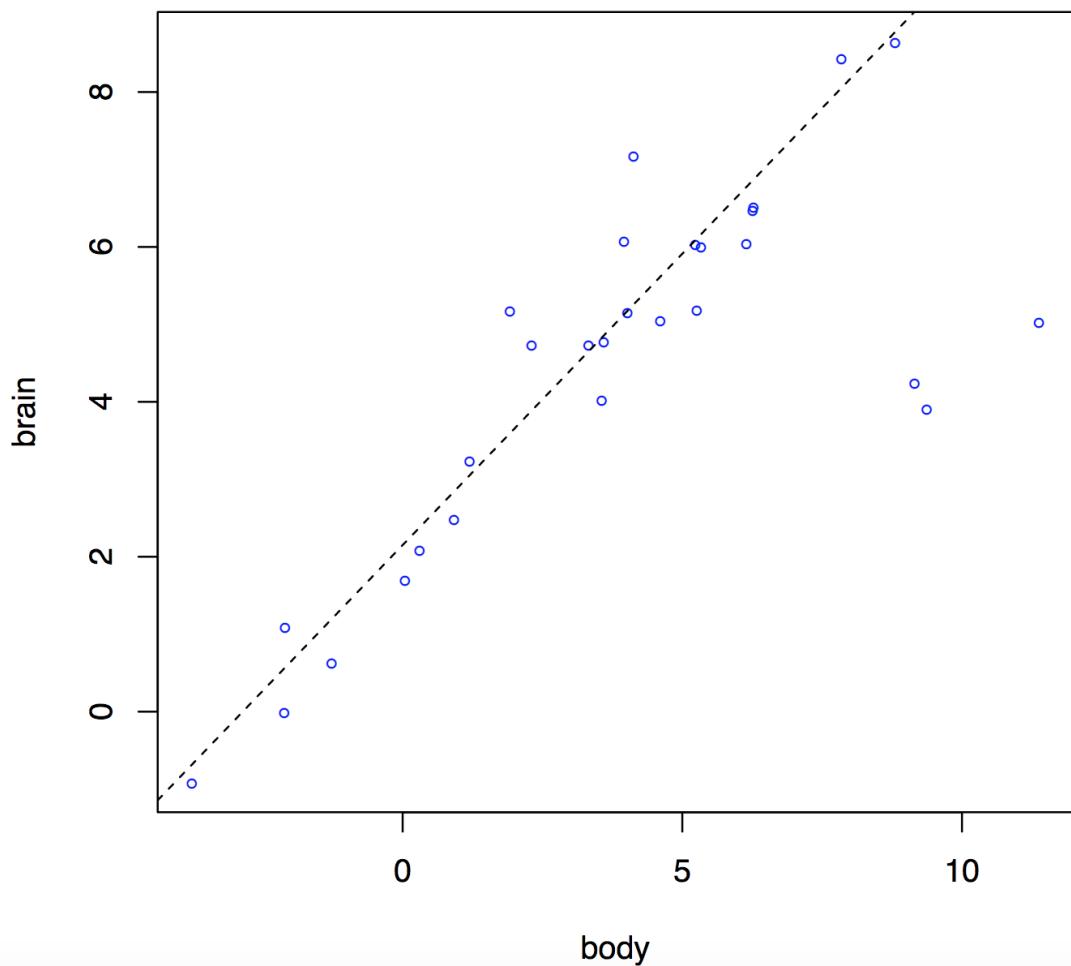
+ 0 pts Incorrect or no response

2.5 3(e)

1 / 1 pt

Show a plot of the data together with the new regression line

✓ + 1 pt Correct graph (ok to either include or exclude dinosaurs, labels not required)



+ 0.5 pts Correct graph but swapped x & y

+ 0 pts Incorrect or no attempt

2.6 3(f)

1 / 2 pts

Give a 95% confidence interval for the average brain weight of camels that weigh on average 500 kg.

+ 2 pts Correct

```
newdata = data.frame(body = log(500))
camelConfLog <- predict(m2, newdata, interval = "confidence")
camelConfLog

##          fit      lwr      upr
## 1 6.825418 6.399984 7.250851

exp(camelConfLog[-1]) # transform to usual scale (g)

## [1] 601.8355 1409.3038
```

- ✓ + 1 pt Reasonable attempt but some mistake e.g. prediction instead of confidence interval, didn't back-transform to grams, didn't apply log to 500, used wrong model, answer not returned OR simply stating correct answer with no code/working given

+ 0 pts Incorrect or no attempt

Question 3

Question 5

5 / 5 pts

3.1 5(a)

1 / 1 pt

State hypotheses

✓ + 1 pt $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$ (ok to have != or words 'does not equal to' instead of \neq)

+ 0.5 pts Correct but hypotheses wrong way around OR used one-sided alternative (making both of these mistakes=0 marks).

+ 0 pts Incorrect or no attempt. Any answer involving μ , \hat{p} or \bar{X} receives 0.

3.2 5(b)

2 / 2 pts

Carry out a test that has significance level $\alpha = 0.05$. What is your conclusion?

✓ + 2 pts Multiple possible correct answers (must include correct conclusion i.e. reject null or 'rates differ'): Prop.test(): p-value=0.007248<5% correct=FALSE: p-value=0.005676<5% using test stat: Z=2.6852 or 2.7659>1.96 There may be other correct methods. If student had a one-tailed hypothesis for Q5a, award full marks for either a (correct) one or two-tailed approach.

+ 1 pt Some sensible attempt but missing some element e.g. missing relevant code/working, no conclusion, wrong alpha used, etc. OR used CI to do the hypothesis test.

+ 0 pts Incorrect/no attempt

3.3 5(c)

1 / 1 pt

✓ + 1 pt Still reject H_0 or 'rates differ'

+ 0 pts Incorrect or no attempt

3.4 5(d)

1 / 1 pt

Give a 95% confidence interval for the difference in rates.

✓ + 1 pt prop.test: (0.01407, 0.08593)
w/o cont correct: (0.01553, 0.08447)
There are other correct methods

+ 0.5 pts Small mistake in decimals e.g. stating CI is (0.014, 0.86) or (0.14, 0.086)

+ 0 pts Incorrect (e.g. two separate CIs) or no attempt

Question 4

Neatness, matching and attempts

1 / 1 pt

4.1 Neatness, matching and attempts

1 / 1 pt

✓ + 1 pt Q2, Q4 and Q6 ALL have some attempt

+ 0 pts One or more of these questions not attempted: 2, 4, 6

- 0.5 pts Questions weren't matched (i.e. 'Student did not select pages for questions')

- 1 pt Incredibly messy (e.g. blank pages, illegible writing, code and graphs handdrawn/handwritten, etc.)

No questions assigned to the following page.

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

Questions assigned to the following page: [1.1](#) and [1.2](#)

Question 1

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

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

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

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

A 95% CI for μ is $\left(8 - 1.96 \frac{0.6}{\sqrt{9}}, 8 + 1.96 \frac{0.6}{\sqrt{9}}\right)$

$$= (7.608, 8.392)$$

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

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

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

Question assigned to the following page: [1.3](#)

1c)

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

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

$$c = 2.306 \quad s^2 = 0.425 \quad s = 0.652$$

$$\text{A } 95\% \text{ CI for } \mu \text{ is } \left(8 - 2.306 \frac{0.652}{3}, 8 + 2.306 \frac{0.652}{3}\right) \\ = (7.499, 8.501)$$

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 .

No questions assigned to the following page.

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

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

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

Questions assigned to the following page: [2.1](#) and [2.2](#)

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))

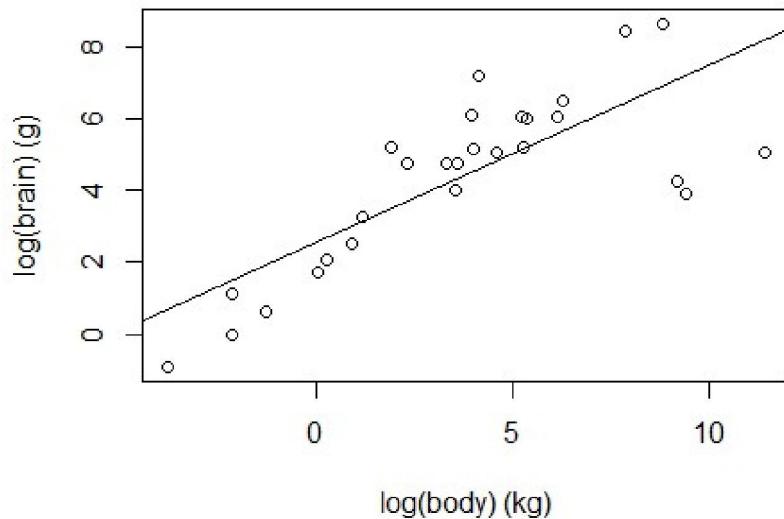
summary(model1)

##
## 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
```

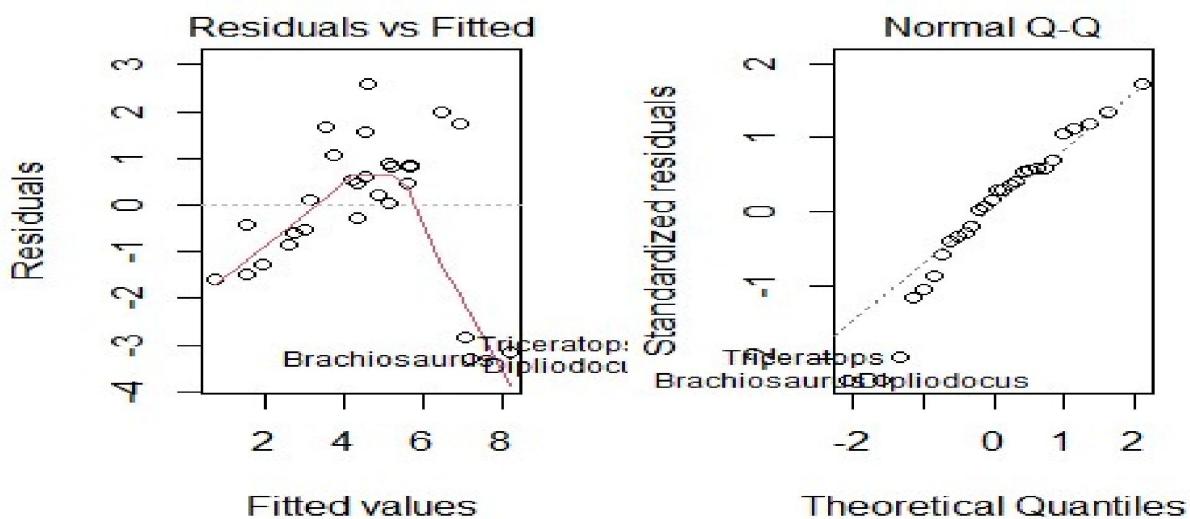
Question assigned to the following page: [2.3](#)

3c)

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



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



Questions assigned to the following page: [2.3](#) and [2.4](#)

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))

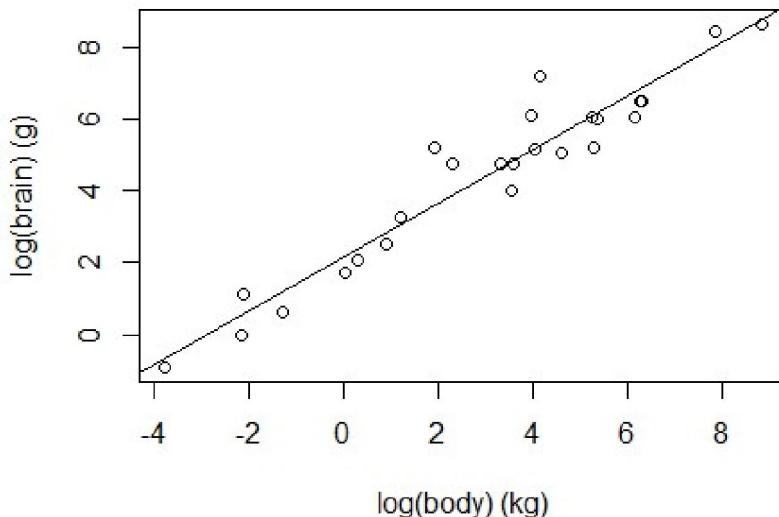
summary(model2)

##
## 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
```

Questions assigned to the following page: [2.5](#) and [2.6](#)

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)

No questions assigned to the following page.

Question 4

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

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

$$W = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \approx t \frac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{m}\right)^{1/2}}{\sqrt{\frac{S_x^4}{n^2(n-1)} + \frac{S_y^4}{m^2(m-1)}}}$$

$$\approx t_{10.31}$$

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

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

A 95% CI for $\mu_X - \mu_Y$ is

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

$$= (0.501, 1.199)$$

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

Questions assigned to the following page: [3.1](#) and [3.2](#)

Question 5

5a)

$H_0: p_1 = p_2$

$H_1: p_1 \neq p_2$

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.

Questions assigned to the following page: [3.3](#) and [3.4](#)

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 assigned to the following page: [4.1](#)

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)

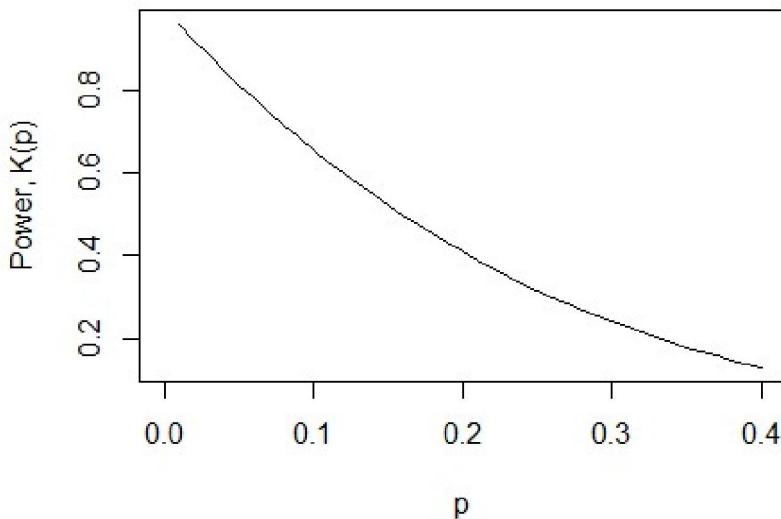
```
beta = pgeom(4, 0.2) - dgeom(4, 0.2)
cat('The probability of committing a Type II Error is', beta)
```

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)')
```



No questions assigned to the following page.

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.