Richard Littauer
2539658

Examination for Statistics in Linguistics

1.

```
## First we should attach and examine the data
> library(languageR)
> attach(lexdec)
> str(lexdec)
```

A)

```
## Here we can see, and deduce, the scale of measurement:
Correct : Nominal, with two levels.
Length : Interval
Class : Nominal
Family Size : Ratio
```

B)

```
## Length does modify correctness significantly, as seen using a two sample t-test
## I chose a t-test because it seemed like the best way to model a DV when the IV was
continuous
> t.test(Length~Correct)

    Welch Two Sample t-test

data:  Length by Correct
t = -2.4254, df = 69.593, p-value = 0.01789
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.01428434 -0.09885145
sample estimates:
  mean in group correct mean in group incorrect
               5.889586                6.446154
## Here is the standard deviation and variation for Length:
> sd(Length)
[1] 1.857397
> var(Length)
[1] 3.449924

## Class is also a significant influence
## Here, a two-sample test would not work, but a one sample test can be used for two
predictors that are both nominal
> t.test(xtabs(~Correct+Class))

    One Sample t-test

data:  xtabs(~Correct + Class)
t = 1.8553, df = 3, p-value = 0.1606
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -296.6951 1126.1951
sample estimates:
mean of x
   414.75
## Here is the xtab() for that
> xtabs(~Correct+Class)
           Class
Correct     animal plant
  correct      884   710
```

```
  incorrect     40    25

## FamilySize is another indication
## I used a Two-sample t-test as it can be used to measure a dependent variable when the
independent variable is continuous, as mentioend above
> t.test(FamilySize~Correct)

    Welch Two Sample t-test

data:  FamilySize by Correct
t = 3.1321, df = 70.855, p-value = 0.002523
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1116560 0.5029038
sample estimates:
  mean in group correct mean in group incorrect
              0.7148457               0.4075658
## Some more readouts for this
> sd(FamilySize)
[1] 0.8709066
> var(FamilySize)
[1] 0.7584784


c)
## Provided with this file are the three graphs.
> barplot(xtabs(~Correct+Length), xlab="Length", ylab="Amount correct")
# For 1c_1, Correct and Length, I went with a barplot, as this clearly shows the proportion
correct or incorrect overall. It doesn't show this very clearly in comparison for different
lengths, however, although this can be ascertained by comparing the differences visually
with a small amount of concentration.

> barplot(xtabs(~Correct+Class), horiz = FALSE, xlab = "Class", ylab = "Correctness")
# For 1c_2, Correct and Class, I went with a barplot, as well, as this clearly shows the
nature of the xtab table.

> plot(FamilySize~Correct, xlab="Correctness", ylab="Family Size with Outliers")
# For 1c_3, Family Size and Correctness, I went with a simple boxplot, as this shows the
data fairly well while showing outliers and deviation, which are helpful in understanding
it.

2)
a)
## It is a good idea to look at the the structure first, to make sure that we are dealing
with the same data formats with similar scales of measurement. After checking this using
str(), it is useful to look at the nature of the three Nominal predictors we'll be looking
at, using xtabs() again, or using subsets.

> str(lexdec)
> xtabs(~Sex+Correct+NativeLanguage)

, , NativeLanguage = English

   Correct
Sex correct incorrect
  F     533        20
  M     387         8

, , NativeLanguage = Other

   Correct
```

```
Sex correct incorrect
  F     526        27
  M     148        10

##Since none of these raise a flag (for instance, no values in one column, or the wrong sort
of scale or data), we can go ahead with the linear model.

b)
## The Null hypothesis is that neither the sex or native language of the participant, nor
whether they were correct or not, should influence the reaction times. The alternative
hypothesis is that one, two, or all of these factors do influence reation times - meaning,
for intance, that females may react faster, or males faster. All alternative hypotheses must
be ruled out for the null hypothesis to be validated.

c)

> calculate <- lm(RT~Correct)
> calculate

Call:
lm(formula = RT ~ Correct)

Coefficients:
     (Intercept)  Correctincorrect
          6.3839            0.0307

> anova(calculate)
Analysis of Variance Table

Response: RT
            Df Sum Sq  Mean Sq F value Pr(>F)
Correct      1  0.059 0.058875  1.0094 0.3152
Residuals 1657 96.647 0.058326
> calculate <- lm(RT~Sex)
> calculate

Call:
lm(formula = RT ~ Sex)

Coefficients:
(Intercept)         SexM
    6.37658      0.02554

> anova(calculate)
Analysis of Variance Table

Response: RT
            Df Sum Sq  Mean Sq F value  Pr(>F)
Sex          1  0.241 0.240532  4.1317 0.04225 *
Residuals 1657 96.465 0.058217
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> calculate <- lm(RT~NativeLanguage)
> calculate

Call:
lm(formula = RT ~ NativeLanguage)

Coefficients:
        (Intercept)  NativeLanguageOther
             6.3183               0.1558
```

```
> anova(calculate)
Analysis of Variance Table

Response: RT
                 Df Sum Sq Mean Sq F value    Pr(>F)
NativeLanguage    1  9.865  9.8647  188.23 < 2.2e-16 ***
Residuals      1657 86.841  0.0524
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

## Here we can see that correctness did not influence the reaction times of the
participants, while their sex was mildly significant, and the Native Language of the
participant was very statisitically significant.

d)
## A Type I error here would be to say that Correctness was a significant influence for
reaction times, as it discards the null hypothesis although there is no justification for
this. A Type II error here would be to say that the Sex and Native Language of the
participants did not influence the reaction times - this would mean keeping the null
hypothesis although the alternative hypothesis have been shown to be more probably, and the
null hypothesis does not stand.
```