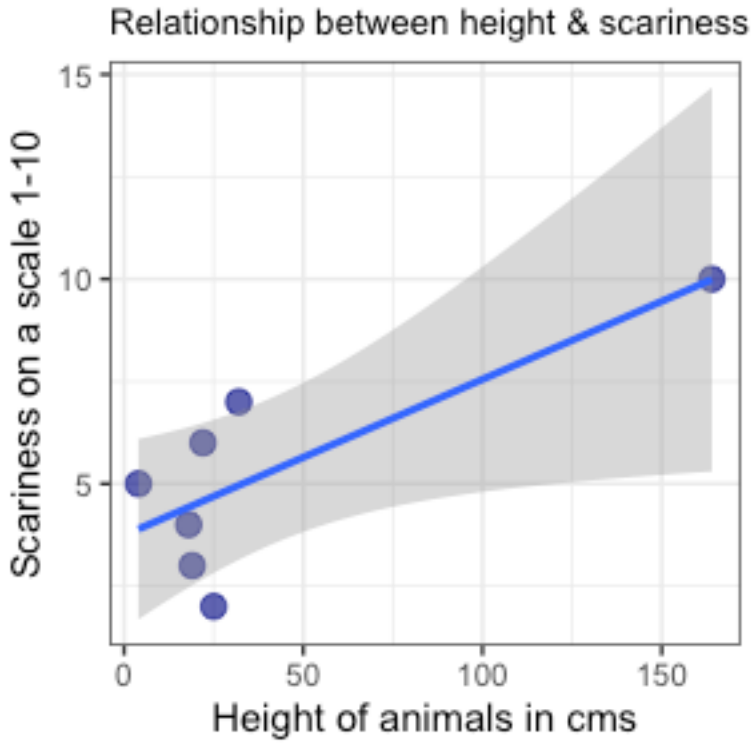
RMHI/ARMP Problem Set 2

Chaitanya Chandrika Raghuvanshi 1117645 [Word Count: 1712]

Q1

```
do %>%
ggplot(mapping=aes(x=height,y=scariness)) +
geom_point(colour="darkblue",alpha=0.7,size=3) +
geom_smooth(method = "lm", se=TRUE) +
theme_bw() +
labs(title = "Relationship between height & scariness",
x="Height of animals in cms", y="Scariness on a scale 1-10")+
theme(plot.title = element_text(size = 10))
```




```
cor.test(do$height,do$scariness, method ="spearman")

##
##  Spearman's rank correlation rho
##
## data:  do$height and do$scariness
## S = 28, p-value = 0.2667
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
## 0.5
```

ANSWER: Having a look at the scatter plot between height and scariness, the relationship between the two doesn't seem linear. Therefore, we should use Spearmans correlation as it doesn't assume linearity unlike Pearson. The correlation coefficient is rho=0.5, S=28, p=.2667. Since the p-value is greater than 0.05, we cannot reject the null hypothesis that there is no correlation between height and scariness. This indicates that the height of the animals does not relate to their level of scariness. [Word count: 81]

Q2

```
dl %>%
ggplot(mapping = aes(x = name,y = hr,fill = type)) +
geom_violin(alpha = 0.4, trim = FALSE, colour = NA)+
geom_boxplot(colour = "black", width = 0.4, position = position_dodge(0.9),
alpha = 0.5) +
theme_bw() +
labs(
title = "Performance on lie detector test",
x = "Name of person",
y = "Heart rate")  +
theme(legend.position = c(.18, .85))+
scale_fill_brewer(palette='Set2')+
scale_colour_brewer(palette='Set2')
```

Q3

```
dl_foxy = dl %>% filter(name == "foxy")
dl_lfb = dl %>% filter(name == "lfb")
t.test(formula = hr~type, data = dl_foxy)

##
##  Welch Two Sample t-test
##
## data:  hr by type
## t = 1.2244, df = 45.691, p-value = 0.2271
## alternative hypothesis: true difference in means between group control and
group test is not equal to 0
## 95 percent confidence interval:
##  -3.479107 14.279107
## sample estimates:
## mean in group control    mean in group test
##                117.68                112.28

t.test(formula = hr~type, data = dl_lfb)

##
##  Welch Two Sample t-test
##
## data:  hr by type
## t = -0.82813, df = 47.991, p-value = 0.4117
## alternative hypothesis: true difference in means between group control and
group test is not equal to 0
## 95 percent confidence interval:
##  -12.203434   5.083434
## sample estimates:
## mean in group control    mean in group test
##                132.68                136.24

t.test(formula = hr~name, data = dl)

##
##  Welch Two Sample t-test
##
## data:  hr by name
## t = -6.3193, df = 97.888, p-value = 7.797e-09
## alternative hypothesis: true difference in means between group foxy and
group lfb is not equal to 0
## 95 percent confidence interval:
##  -25.59747 -13.36253
## sample estimates:
## mean in group foxy  mean in group lfb
##             114.98             134.46
```

ANSWER: We should use Welch t-test as it takes the difference in variance into consideration (as seen in the graph) to check if there is any difference in the means of control and test

groups. The outcome variable is heart rate while the predictor is the type of group (test or control).

For foxy: The difference between the groups was non-significant, t(45.691)=1.224, p=.227, suggesting that the mean heart rates did not significantly differ in the two groups for Foxy. Therefore, the lie detector indicates that Foxy did not lie.

For LFB: The difference between the groups was non-significant, t(47.991)=-0.828, p=.412, suggesting that the mean heart rates did not significantly differ in the two groups for LFB. Therefore, the lie detector indicates that LFB did not lie.

For performance difference: The difference between the groups was significant, t(97.888)=-6.319, p<.001, suggesting that the performances of Foxy and LFB significantly differed. This may indicate that they did not communicate with each other during the test. [Word count: 161]

Q4

Q4a

```
shapiro.test(dd$health[dd$time=="t1"])

##
##  Shapiro-Wilk normality test
##
## data:  dd$health[dd$time == "t1"]
## W = 0.98792, p-value = 0.9673

shapiro.test(dd$health[dd$time=="t2"])

##
##  Shapiro-Wilk normality test
##
## data:  dd$health[dd$time == "t2"]
## W = 0.98239, p-value = 0.8552
```

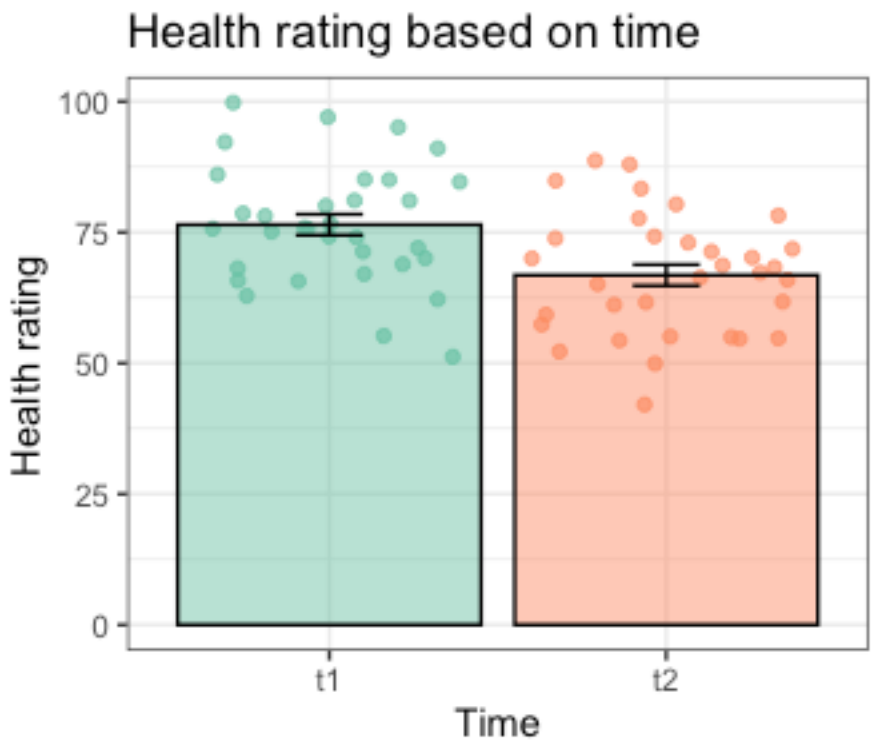
ANSWER: We assume that the population distribution of each time is normal. This assumption can be tested using the Shapiro-Wilk test as the sample size for each group is smaller than 50. For each time t1 and t2, we checked if the the health data is normally distributed. For t1, W=0.988, p=.967 and for t2, W=0.982, p=.855. Since the p-value is greater than 0.05 for both groups, we cannot reject null hypothesis stating that our normality assumption for population distribution of time t1 and t2 upholds. [Word count: 85]

Q4b

```
dd_sum <- dd %>%
group_by(time) %>%
summarise(mean = mean(health),
sd = sd(health),
n = n(),
sderr = sd/sqrt(n)) %>%
```

```
ungroup()

dd_sum %>%
ggplot(mapping=aes(x=time,y=mean,fill=time)) +
geom_jitter(data=dd,mapping=aes(x=time,y=health,colour=time),alpha=0.7,show.l
egend=FALSE) +
geom_col(colour="black",show.legend=FALSE,alpha=0.5) +
geom_errorbar(mapping=aes(ymin = mean-sderr, ymax=mean+sderr),width=0.2) +
theme_bw() +
scale_fill_brewer(palette="Set2") +
scale_colour_brewer(palette="Set2") +
labs(title = "Health rating based on time", x="Time", y="Health rating")
```




```
t.test(formula = health ~ time, data = dd, paired = TRUE)

##
##  Paired t-test
##
## data:  health by time
## t = 9.172, df = 32, p-value = 1.798e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   7.496301 11.776426
## sample estimates:
## mean of the differences
##                9.636364
```

```
cohensD(dd$health[dd$time=="t1"], dd$health[dd$time=="t2"], method='paired')

## [1] 1.596639
```

ANSWER: To test the null hypothesis that there is no difference in average health rating at time t1 and t2, we used a paired-samples t-test as we are comparing the same variable health (outcome variable) at different times (predictor variable), t1 and t2. The graph above shows that t1 has a higher average rating than t2, however, we need to check if the difference is significant using the t-test. The difference in the average health ratings at time t1 and t2, that is, mean of the differences (9.636 units with a 95% confidence interval of [7.496, 11.776]), is statistically significant, t(32)=9.172, p<.001. To measure effect size, we use Cohen's d (d=1.597) and found that the effect size is large, meaning health is significantly affected by time. The results may suggest that people's health has been getting worse in Otherland which could possibly be due to food shortage. [Word count: 146]

Q5

```
dis_table = table(dd$disability[dd$time=="t1"], dd$disability[dd$time=="t2"])
dis_table

##
##         FALSE TRUE
##   FALSE    16    9
##   TRUE      1    7

mcnemar.test(x = dis_table)

##
##  McNemar's Chi-squared test with continuity correction
##
## data:  dis_table
## McNemar's chi-squared = 4.9, df = 1, p-value = 0.02686
```

ANSWER: We use McNemar, chi-squared test as the data is not independent due to each person having two observations for t1 and t2. Since we have a binary outcome of TRUE and FALSE for disability and we measure it twice for t1 and t2, McNemar is the most suitable test. From the table above we can see that 9 people went from non-disabled to disabled, one person went from disabled to non-disabled, 7 remain disabled while 16 remain non-disabled. We see that the null hypothesis, stating that there is no change of people being on or off disability from t1 to t2, is rejected as X2(1)=4.9, p=.027. Therefore, there has been a significant change from being on disability to being off of it or vice-versa, indicating a significant difference between t1 and t2.[Word count: 132]

Q6

ANSWER: It is better to exclude data from time point 1 as for a one way ANOVA we assume that the size groups are independent. If we do not exclude one of the time points, there will be two values for each size, braking our independence assumption. Another thing we could do to solve this issue is we take the average of the health ratings at time t1 and t2 as our numerical variable in our one way ANOVA. Therefore, by taking average of the two, we create a

numerical variable that takes into account both time periods and still maintains our assumptions of independence. [Word count: 104]

Q7

```
health_model = aov(health ~ size, data = ddt2)
shapiro.test(x = health_model$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  health_model$residuals
## W = 0.97546, p-value = 0.6436

leveneTest(health ~ size, data = ddt2)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.1001 0.9051
##       30
```

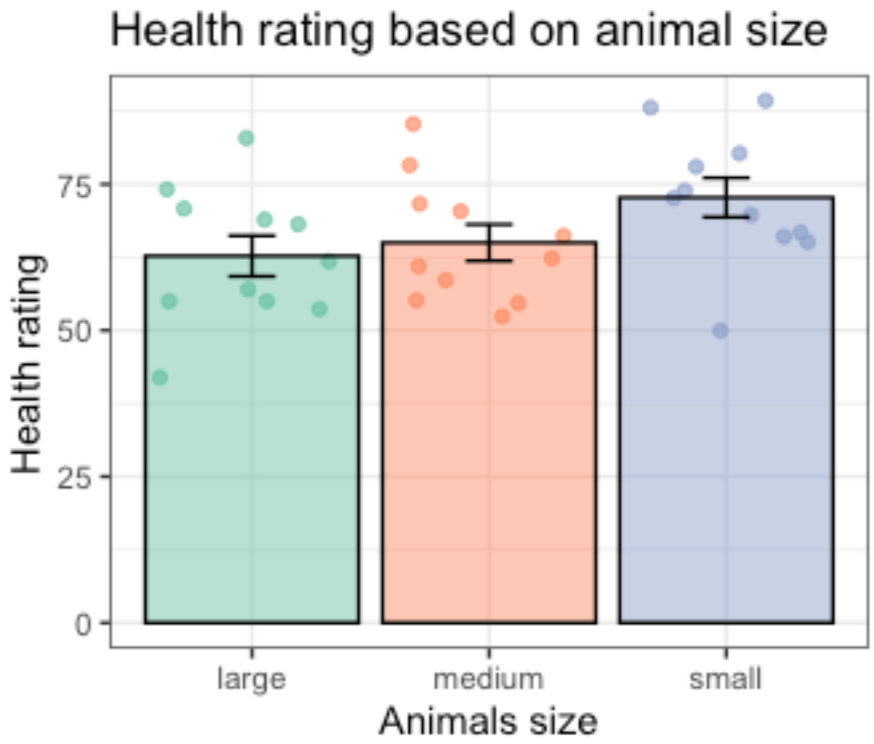
ANSWER: Assumption 1: We assume that residuals are normally distributed and used Shapiro-Wilk test to evaluate the same. The stats, W=0.975, p=.643 tell us that the p-value is not significant and the null hypothesis stating that there is normal distribution of residuals cannot be rejected as p value is > 0.05. This concludes that the residuals are normally distributed and the assumption was not violated. Assumption 2: We assume that the size groups have the same variance and used Levene's test to evaluate the same.The stats, F(2,30)=0.1, p=.905 tell us that the p-value is not significant and the null hypothesis of homogeneity of variance across groups cannot be rejected as p value is > 0.05, suggesting that the different size groups have the same variance and the assumption was not violated. [Word count: 130]

Q8

```
ddt2_sum <- ddt2 %>%
group_by(size) %>%
summarise(mean = mean(health),
sd = sd(health),
n = n(),
sderr = sd/sqrt(n)) %>%
ungroup()

ddt2_sum %>%
ggplot(mapping=aes(x=size,y=mean,fill=size)) +
geom_jitter(data=ddt2,mapping=aes(x=size,y=health,colour=size),alpha=0.7,show
.legend=FALSE) +
geom_col(colour="black",show.legend=FALSE,alpha=0.5) +
geom_errorbar(mapping=aes(ymin = mean-sderr, ymax=mean+sderr),width=0.2) +
theme_bw() +
scale_fill_brewer(palette="Set2") +
scale_colour_brewer(palette="Set2") +
```

```
labs(title = "Health rating based on animal size", x="Animals size",
y="Health rating")
```




```
summary(health_model)

##             Df Sum Sq Mean Sq F value Pr(>F)
## size         2    605   302.3   2.486    0.1
## Residuals   30   3648   121.6

etaSquared(health_model)

##         eta.sq eta.sq.part
## size 0.1421487   0.1421487
```

ANSWER: We used a one way ANOVA test as three groups (small, medium and large) are defined by a single categorical variable named size to predict our numerical value, health. The graph above gives us a description of the health ratings of animals based on their size, showing smaller animals have the highest average health rating followed by medium and then large. We perform the ANOVA test to see if the difference is statistically significant The null hypothesis that the population means for all three size groups are the same cannot be rejected as a one-way ANOVA indicated that there were not significant differences in the health rating by animal size, F(2,30) = 2.486, p = .1, η2 = 0.142. The results mean that the different size groups have no significant difference on the health of a person. The eta squared statistic is used to calculate effect size, interpreting that size explains about 14% of the variation in the health rating for that size.[Word count: 163]

Q9

Q9a

```
reg_model <- lm(health ~ height, data=ddt2)
summary(reg_model)

##
## Call:
## lm(formula = health ~ height, data = ddt2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.1327  -5.9215  -0.2801   7.4251  18.4251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.60675    2.53945  28.198   <2e-16 ***
## height      -0.14741    0.05414  -2.722   0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.52 on 31 degrees of freedom
## Multiple R-squared:  0.193,  Adjusted R-squared:  0.1669
## F-statistic: 7.412 on 1 and 31 DF,  p-value: 0.01054

etaSquared(reg_model)

##           eta.sq eta.sq.part
## height 0.1929598   0.1929598
```

ANSWER: We used a regression model to evaluate whether height is a significant predictor of health rating as both, the predictor variable (height) and the outcome variable (health) are numerical. The regression line is given by health= 71.607 - 0.147xheight. The intercept is 71.607, which is the health rating our model predicts for a person with 0 cms height. The slope for height is -0.147, which means that for each additional centimeter of height, we can expect a decrease of 0.148 in health rating. The null hypothesis stating that height is not a significant predictor for health is rejected as p-value is < 0.05. Using the results, F(1,31)=7.412, p=0.010. we can conclude that height is a significant predictor of health rating. The eta squared statistic is used to calculate effect size, interpreting that height explains about 19% of the variation in the health rating. [Word count: 143]

Q9b

ANSWER: In q8, when using ANOVA we find that the animal size (small, medium and large), has no significant difference on their health rating. However, when we used a regression model in q9 we found the animal height is a significant predictor of their health. The p-value for ANOVA (0.1) is greater than 0.05 while for regression the p-value (0.01) is less than 0.05. Therefore, the categorical variable size was not helpful for our statistical analysis, while the numerical variable height resulted in being a significant predictor of health rating.

Regression provides an in-depth analysis of data as we have the exact height of the animal rather than a less precise categorical measure, size. It acknowledges the existence of random variation that can be found in the data by including residuals, that is, deviation from the regression line.[Word count: 137]

Q10

```
reg_model2 <- lm(health ~ height + income, data=ddt2)
summary(reg_model2)

##
## Call:
## lm(formula = health ~ height + income, data = ddt2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1762  -6.8412   0.1135   7.2637  17.9919
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 56.87169    7.21207   7.886 8.43e-09 ***
## height      -0.11272    0.05363  -2.102   0.0441 *
## income       0.14191    0.06550   2.167   0.0383 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.946 on 30 degrees of freedom
## Multiple R-squared:  0.3022, Adjusted R-squared:  0.2556
## F-statistic: 6.495 on 2 and 30 DF,  p-value: 0.004533

etaSquared(reg_model2)

##           eta.sq eta.sq.part
## height 0.1027656   0.1283593
## income 0.1091970   0.1353055
```

ANSWER: The slope for height is -0.113, which means that for each additional centimeter of height, the height variable leads to a decrease of 0.113 in health rating. While in q9, we expect a decrease of 0.148 in health rating. This is because the plane of best fit in our multiple regression minimises residuals based on not only height but also income, reducing the variance explained by height. The effect of height here is different with respect to model in q9 in terms of its effect size. Height explains about 19% variation in the first model, while it explains only about 10% variation is the second model. This is because the total model minimises the residuals, and some of the variance in the model with only height is taken care of in the second model by variable income, decreasing the variation explained by height in the second model. [Word count: 147]

Q11

Q11a

ANSWER: (i) A (ii) A (iii) B

Q11b

ANSWER: (i) We know that for large value of test statistic χ2, the null hypothesis is doing a bad job in explaining the data and p value is bound to be really small. However, in our scenario p-value is relatively large for a large value of χ2. (ii) If null hypothesis is true, we expect t statistics around 0. Therefore, in our example since we reject null hypothesis in both cases the statistic is likely to be away from zero. Since in A, the test statistic is much closer to 0 and we reject null with a lower p-value than B, A is not possible. (iii) The F-statistic can never be zero as we take a ratio of sum of squares which means we always get a positive value. Therefore B is not possible. [Word count: 132]

Q12

Q12a

ANSWER: (i) B (ii) A (iii) C (iv) D

Q12b

ANSWER: (i) Since there is no x2 term and B has only two axes y and x1 with the plane lying on x2=0, y does not depend on x2 in this graph which is also evident in equation 1 (ii) The interaction between x1 and x2 can be seen in figure B and since the interaction divides the plane in similar parts, x1 and x2 are bound to have similar effect on y. (iii) Since there are two variables x1 and x2 with no interaction, graph C suits the best as there is no interaction in the plane but y value depends on both x1 and x2. (iv) The interaction present in graph D does not equally divide the graph plane, so x1 and x2 are bound to have different slopes. Therefore, in equation 4 we have different slops for x1 and x2 making it the most suitable equation for graph D. [Word count: 151]

Q13

Foxy because she seems nice and smart.