RMHI/ARMP Problem Set 2

Henrik Formoe 1575157 [Word Count: 1498]

Please put your answers here, following the instructions in the assignment description. Put your answers and word count tallies in the locations indicated; if none is indicated that means there is no word count for that question. Remember to knit as you go, and submit the knitted version of this on Canvas.

## Q1

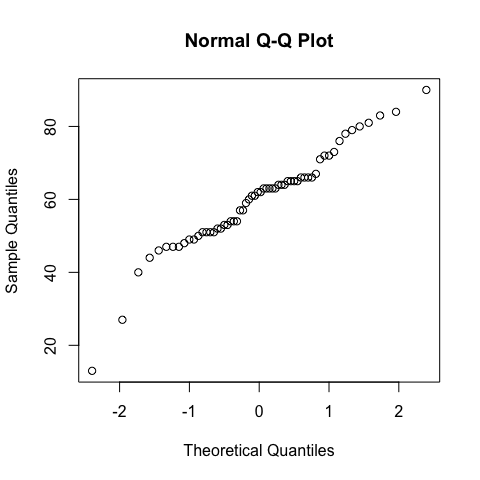
**Q1a**

# any code goes here  
  
ds\_d <- ds %>%  
 group\_by(person) %>%  
 summarise(  
 MeanTime = mean(length),  
 SDtime = sd(length),  
 SEtime = SDtime / sqrt(n())  
) %>%  
 ungroup()  
  
ds\_barplot <- ggplot(data = ds\_d, aes(fill = person)) +  
 geom\_col(data = ds\_d, aes(x = person, y = MeanTime), alpha = 0.3, color = "black", width = 0.7, show.legend = FALSE) +  
 geom\_errorbar(data = ds\_d, aes(x = person, y = MeanTime, ymin = MeanTime - SEtime, ymax = MeanTime + SEtime), width = 0.2) +  
 geom\_jitter(data = ds, aes(x = person, y = length, color = person), position = position\_jitter(width = 0.6, height = 0.6), size = 2, alpha = 0.9, show.legend = FALSE) +  
 scale\_y\_continuous(breaks = seq(0,100, by = 10)) +  
 theme\_bw() +  
 labs(  
 title = "Mean time spent watching per shift:",  
 x = "Person",  
 y = "Time Watching (Seconds)"  
 ) +  
 scale\_fill\_brewer(palette = "Pastel1") +  
 scale\_color\_brewer(palette = "Pastel1")  
  
print(ds\_barplot)



**Q1b**

# any code goes here  
  
qqnorm(ds$length)



shapiro.test(ds$length)

##   
## Shapiro-Wilk normality test  
##   
## data: ds$length  
## W = 0.95669, p-value = 0.03248

shapiro.test(ds$length[ds$person=="BigHippo"])

##   
## Shapiro-Wilk normality test  
##   
## data: ds$length[ds$person == "BigHippo"]  
## W = 0.95357, p-value = 0.4246

shapiro.test(ds$length[ds$person=="QuietHorse"])

##   
## Shapiro-Wilk normality test  
##   
## data: ds$length[ds$person == "QuietHorse"]  
## W = 0.96271, p-value = 0.5992

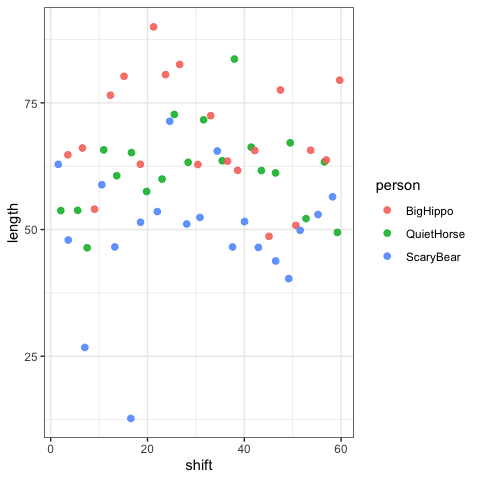
shapiro.test(ds$length[ds$person=="ScaryBear"])

##   
## Shapiro-Wilk normality test  
##   
## data: ds$length[ds$person == "ScaryBear"]  
## W = 0.8947, p-value = 0.03284

leveneTest(length ~ person, data = ds)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 2 0.4662 0.6297  
## 57

jitter1 <- ggplot (ds) +  
 geom\_jitter(aes(x = shift, y = length, color = person), position = position\_jitter(width = 0.6, height = 0.6), size = 2, alpha = 0.9) +  
 theme\_bw()  
   
print(jitter1)



*Note that the order of these assumptions (i.e., what assumption you put in 1 vs 2) does not matter!*

*ASSUMPTION 1: We assume that our data follows a normal distribution. First using qqnorm, and then test it using a shapiro.test to evaluate the assumption of normality. We found overall p = 0.033. Also, we have multiple groups (three persons), and therefore tested each group. BigHippo p = 0.424, QuietHorse p = 0.599, ScaryBear p = 0.033. The null hypothesis is that the data is normally distributed, so this indicates that the data in its whole, and for ScaryBear, is not normally distributed. [Word Count: 81]*

*ASSUMPTION 2: We assume that the variance across all groups is homogenous, and test this using a leveneTest. p = 0.630 shows that there is no significant difference, therefore confirming homogenous variance across groups. [Word Count: 32]*

**Q1c**

# any code goes here  
  
kruskal.test(length ~ person, data = ds)

##   
## Kruskal-Wallis rank sum test  
##   
## data: length by person  
## Kruskal-Wallis chi-squared = 22.47, df = 2, p-value = 1.321e-05

kruskal\_effsize(length ~ person, data = ds)

## # A tibble: 1 × 5  
## .y. n effsize method magnitude  
## \* <chr> <int> <dbl> <chr> <ord>   
## 1 length 60 0.359 eta2[H] large

*ANSWER: Given the results in Q1B, we find the data not to be normally distributed, but the variety across the groups to be homogenous. This means that we will use a Kruskal-Wallis test, because the assumption of normal distribution is violated. We find the chi-squared = 22.47, p-value = 1.321e-05 indiciating a significant difference between the three people and the amount of time spent watching. We find the effect size to be 0.359, suggesting that 35.9 % of the variance in length (seconds) is attributed to the person watching. [Word Count: 88]*

**Q1d**

# any code goes here  
  
watchtime\_person <- aov(length ~ person, data = ds)  
summary(watchtime\_person)

## Df Sum Sq Mean Sq F value Pr(>F)   
## person 2 3789 1894.6 15.9 3.25e-06 \*\*\*  
## Residuals 57 6792 119.2   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

etaSquared(watchtime\_person)

## eta.sq eta.sq.part  
## person 0.3581173 0.3581173

posthocPairwiseT( x = watchtime\_person, p.adjust.methods = "holm")

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: length and person   
##   
## BigHippo QuietHorse  
## QuietHorse 0.0628 -   
## ScaryBear 2.3e-06 0.0011   
##   
## P value adjustment method: holm

*ANSWER: We ran post-hoc tests with holm-correction, and found significant effects with p = 2.3e-06 between BigHippo and ScaryBear, and p = 0.0011 between QuietHorse and ScaryBear. In terms of LFBs question, this shows a significant difference in watch times between SB and BH, and SB and QH. [Word count: 47]*

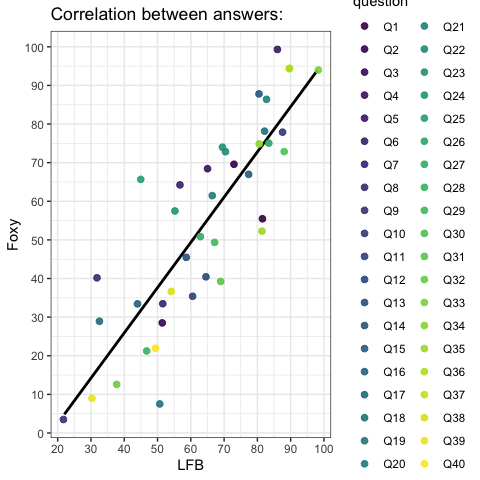
## Q2

**Q2a**

# any code goes here  
  
dl$question <- factor(dl$question, levels = paste0("Q", 1:40), ordered = TRUE)  
  
cor.test(dl$lfb, dl$foxy)

##   
## Pearson's product-moment correlation  
##   
## data: dl$lfb and dl$foxy  
## t = 10.141, df = 38, p-value = 2.31e-12  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.7399745 0.9208905  
## sample estimates:  
## cor   
## 0.8545113

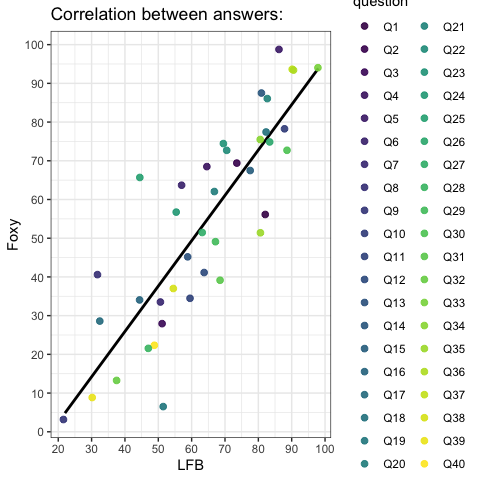
jitter2 <- ggplot (dl) +  
geom\_smooth(method = "lm", aes(x = dl$lfb, y = dl$foxy), se = FALSE, colour = "black", show.legend = FALSE) +  
geom\_jitter(aes(x = dl$lfb, y = dl$foxy, color = question), position = position\_jitter(width = 0.6, height = 0.6), size = 2, alpha = 0.9) +  
 theme\_bw() +  
 scale\_y\_continuous(breaks = seq(0,100, by = 10)) +  
 scale\_x\_continuous(breaks = seq(0,100, by = 10)) +  
 labs(  
 title = "Correlation between answers:",  
 x = "LFB",  
 y = "Foxy"  
 )  
  
   
print(jitter2)



*ANSWER: Our tests show a correlation coefficient of 0.854 with a p-value = 2.31e-12. We used a Pearson’s correlation test, because it assumes linearity between X and Y, hence why we also included the linear regression line which illustrates the positive correlation. In sum, this indicates that LFB and Foxy’s answers overall were very similar. [Word Count: 54]*

**Q2b**

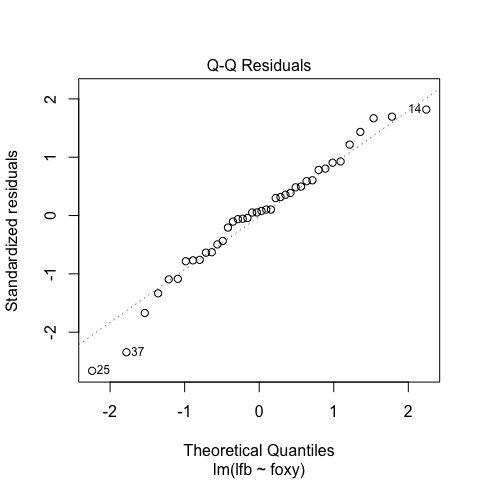
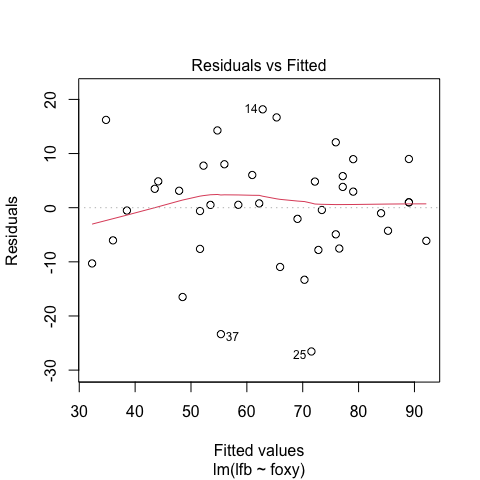
# any code goes here  
  
jitter2 <- ggplot (dl) +  
geom\_smooth(method = "lm", aes(x = dl$lfb, y = dl$foxy), se = FALSE, colour = "black", show.legend = FALSE) +  
geom\_jitter(aes(x = dl$lfb, y = dl$foxy, color = question), position = position\_jitter(width = 0.6, height = 0.6), size = 2, alpha = 0.9) +  
 theme\_bw() +  
 scale\_y\_continuous(breaks = seq(0,100, by = 10)) +  
 scale\_x\_continuous(breaks = seq(0,100, by = 10)) +  
 labs(  
 title = "Correlation between answers:",  
 x = "LFB",  
 y = "Foxy"  
 )  
  
   
print(jitter2)



model1 <- lm(formula = lfb ~ foxy, data = dl)  
summary(model1)

##   
## Call:  
## lm(formula = lfb ~ foxy, data = dl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26.5573 -6.0572 0.6602 5.8847 18.1674   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 30.42665 3.71081 8.199 6.25e-10 \*\*\*  
## foxy 0.62319 0.06145 10.141 2.31e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.13 on 38 degrees of freedom  
## Multiple R-squared: 0.7302, Adjusted R-squared: 0.7231   
## F-statistic: 102.8 on 1 and 38 DF, p-value: 2.31e-12

plot(model1,which=1:2)



*ANSWER: (i) t-statistic and p-value is the same for both the linear regression model and the Pearson correlation test. t= 10.141 and the p-value = 2.31e-12. ; (ii) The correlation coefficient r explains the strength and direction of the relationship between the variables, and how our data points fit along a straight line. It might therefore show a strong/weak and positive/negative relationship (or no relationship = 0). A regression line (Y = b1 X + b0) is the best fitting straight line to our data points, meaning with the least residual values. While r shows strength and direction, the value (b1 X) indicates unit increase which also shows strength and direction, e.g. every time Foxy answers 50, LFB’s answer will be (0.623 x 50 + 30.427 = 61.577).; (iii) It provides us with the adjusted R-squared, and F-statistics. [Word Count: 136]*

**Q2c**

# any code goes here  
  
standardcoefficient1 <- lm(scale(lfb) ~ scale(foxy), data = dl)  
  
lm.beta(standardcoefficient1)

## scale(foxy)   
## 0.8545113

# Also just wanted to try this one.  
  
summary(standardcoefficient1)

##   
## Call:  
## lm(formula = scale(lfb) ~ scale(foxy), data = dl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.38010 -0.31477 0.03431 0.30581 0.94411   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.341e-16 8.320e-02 0.00 1   
## scale(foxy) 8.545e-01 8.426e-02 10.14 2.31e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5262 on 38 degrees of freedom  
## Multiple R-squared: 0.7302, Adjusted R-squared: 0.7231   
## F-statistic: 102.8 on 1 and 38 DF, p-value: 2.31e-12

*ANSWER: The standardised coefficient is 0.854*

**Q2d**

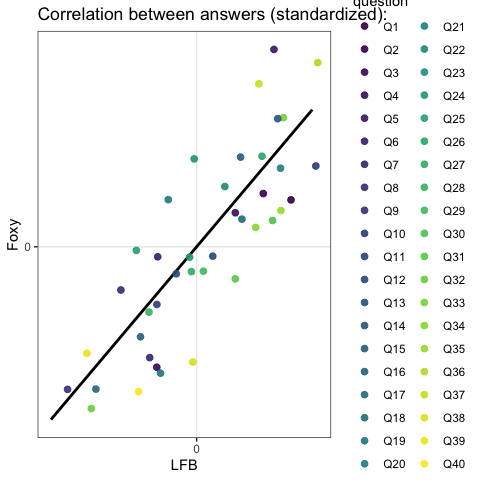
# this code is given  
dl <- dl %>%  
 mutate(lfbZnorm = (lfb - mean(lfb))/sd(lfb),  
 foxyZnorm = (foxy - mean(foxy))/sd(foxy))  
  
# any code goes here  
cor.test(dl$lfbZnorm, dl$foxyZnorm)

##   
## Pearson's product-moment correlation  
##   
## data: dl$lfbZnorm and dl$foxyZnorm  
## t = 10.141, df = 38, p-value = 2.31e-12  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.7399745 0.9208905  
## sample estimates:  
## cor   
## 0.8545113

*ANSWER: This code chunk is converting our raw scores to standardized scores. To be able to compare apple’s to oranges, or LFB’s answers to Foxy’s answers we need standardized scores. The formula for z-scores is: standardized score = (raw scores - population mean) / population SD. For z-scores the mean is by definition zero and SD is one. Z-scores are applied e.g. in WICS-IV, where the mean is one hundred and one SD is 15. The answer is still the same, since the relative values are still the same, but it’s now easier to understand how a value measures compared to the mean, explained by standard deviations. [Word Count: 106]*

**Q2e**

# any code goes here  
  
  
dl$question <- factor(dl$question, levels = paste0("Q", 1:40), ordered = TRUE)  
  
jitter3 <- ggplot (dl) +  
geom\_smooth(method = "lm", aes(x = dl$lfbZnorm, y = dl$foxyZnorm), se = FALSE, colour = "black", show.legend = FALSE) +  
geom\_jitter(aes(x = dl$lfbZnorm, y = dl$foxyZnorm, color = question), position = position\_jitter(width = 0.6, height = 0.6), size = 2, alpha = 0.9) +  
 theme\_bw() +  
 scale\_y\_continuous(breaks = seq(0,100, by = 10)) +  
 scale\_x\_continuous(breaks = seq(0,100, by = 10)) +  
 labs(  
 title = "Correlation between answers (standardized):",  
 x = "LFB",  
 y = "Foxy"  
 )  
  
   
print(jitter3)

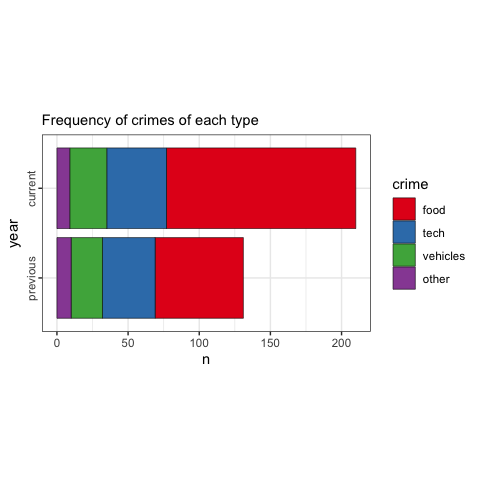


*ANSWER: In the standardized linear regression, the intercept parameters are centered around the mean of each ‘population’, in this case person, and the intercept is at 2.341e-16. Also, the slope is now 0.85, meaning that for every SD increase in Foxy, we expect a 0.85 SD increase in LFB. In other words, the values are displayed according to standard deviations from to the mean, rather than using unit-differences in values of the raw scores. [Word Count: 73]*

## Q3

**Q3a**

# this code is given  
dcPlotTable <- dc %>%  
 group\_by(year,crime) %>%  
 summarise(n = n()) %>%  
 ungroup()  
  
# any code goes here  
  
 dc\_barplot <- ggplot(data = dcPlotTable, aes(x = year, y = n, fill = crime)) +  
 geom\_col(alpha = 1, color = "black", size = 0.2, width = 0.9, show.legend = TRUE) +  
 scale\_y\_continuous(breaks = seq(0,200, by = 50)) +  
 scale\_x\_discrete(guide = guide\_axis (angle=90)) +  
 coord\_flip() +  
 theme\_bw() +  
 theme(aspect.ratio = 3/5) +  
 labs(  
 subtitle = "Frequency of crimes of each type",  
 x = "year",  
 y = "n",  
 fill = "crime"  
 ) +  
 scale\_fill\_brewer(palette = "Set1") +  
 scale\_color\_brewer(palette = "Set1")  
  
print(dc\_barplot)



**Q3b**

# any code goes here  
  
filtered\_dcPlot <- dcPlotTable %>%  
 filter(year == "current") %>%  
 mutate(p = 0.25)  
  
#Extract data for current year  
  
n\_crime <- filtered\_dcPlot$n  
expected\_p <- filtered\_dcPlot$p  
  
#Make data suitable for test  
  
chi\_dcPlot <- chisq.test(x=n\_crime,p=expected\_p)  
cramer\_v\_result1 <- CramerV(table(filtered\_dcPlot$crime, n\_crime))  
#Desctools  
cramer\_v\_result2 <- cramerV(table(filtered\_dcPlot$crime, n\_crime))  
#Rcompanion  
  
print(cramer\_v\_result1)

## [1] 1

print(cramer\_v\_result2)

## Cramer V   
## 1

#Chi-square & CramerV  
  
filtered\_dcPlot <- filtered\_dcPlot %>%  
mutate(st\_dres = chi\_dcPlot$stdres)  
  
#Adding residuals  
  
print(filtered\_dcPlot)

## # A tibble: 4 × 5  
## year crime n p st\_dres  
## <fct> <fct> <int> <dbl> <dbl>  
## 1 current food 133 0.25 12.8   
## 2 current tech 42 0.25 -1.67  
## 3 current vehicles 26 0.25 -4.22  
## 4 current other 9 0.25 -6.93

print(chi\_dcPlot)

##   
## Chi-squared test for given probabilities  
##   
## data: n\_crime  
## X-squared = 174.95, df = 3, p-value < 2.2e-16

*ANSWER: We conducted a chi-square Goodness of Fit to see if the data is how we would expect – if the likelihood is equal for all crime-types. In accordance with LFB’s statement, the chisq (X-squared = 174.95, df = 3, p-value < 2.2e-16) reject the 0-hypothesis, and therefore confirm that some crime-types occur more often than we’d expect, and some a lot less. The residuals show food crime at the top, and “others” occurring less frequently. [Word Count: 74]*

**Q3c**

# any code goes here  
  
dc\_table <- dcPlotTable %>%  
 pivot\_wider(names\_from = "year", values\_from = "n")  
   
print(dc\_table)

## # A tibble: 4 × 3  
## crime previous current  
## <fct> <int> <int>  
## 1 food 62 133  
## 2 tech 37 42  
## 3 vehicles 22 26  
## 4 other 10 9

chisq\_indp <- chisq.test(dc\_table[, -1])  
  
chisq\_indp$observed

## previous current  
## [1,] 62 133  
## [2,] 37 42  
## [3,] 22 26  
## [4,] 10 9

chisq\_indp$expected

## previous current  
## [1,] 74.91202 120.08798  
## [2,] 30.34897 48.65103  
## [3,] 18.43988 29.56012  
## [4,] 7.29912 11.70088

chisq\_indp$residuals

## previous current  
## [1,] -1.4918273 1.1782692  
## [2,] 1.2073041 -0.9535482  
## [3,] 0.8290586 -0.6548039  
## [4,] 0.9997007 -0.7895797

chisq\_indp$stdres

## previous current  
## [1,] -2.905272 2.905272  
## [2,] 1.755136 -1.755136  
## [3,] 1.139715 -1.139715  
## [4,] 1.310952 -1.310952

print(chisq\_indp)

##   
## Pearson's Chi-squared test  
##   
## data: dc\_table[, -1]  
## X-squared = 8.7196, df = 3, p-value = 0.03326

*ANSWER: I conducted a chi-square test of independence (X-squared = 8.719, df = 3, p-value = 0.033) to test if variables are related to each other, looking at thefts committed in the previous and current year. The results show a statistically significant effect indicating a difference in crimes committed in the previous and current year. [Word Count: 54]*

**Q3d**

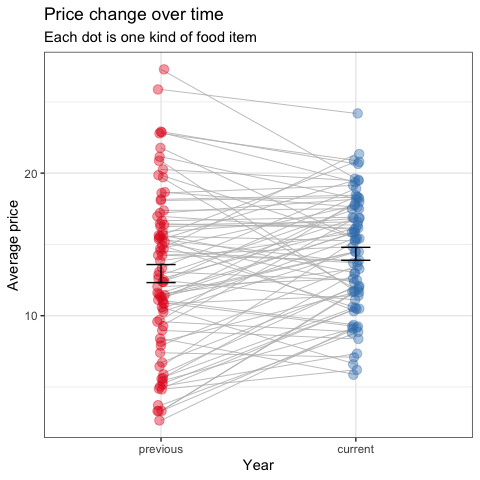
# any code goes here  
  
chisq\_indp$stdres

## previous current  
## [1,] -2.905272 2.905272  
## [2,] 1.755136 -1.755136  
## [3,] 1.139715 -1.139715  
## [4,] 1.310952 -1.310952

*ANSWER: Our adjusted residuals may suggest if crime types were committed less or more frequently than we’d expect, because we may interpret these as indication of how many SD away values in each cell are. Some use +/- 1.96 as an indication of significant results. If so, In the previous year, we see that food-crimes appeared less frequently, and in the current year, food-crimes were comitted more frequently than expected. [Word Count: 69]*

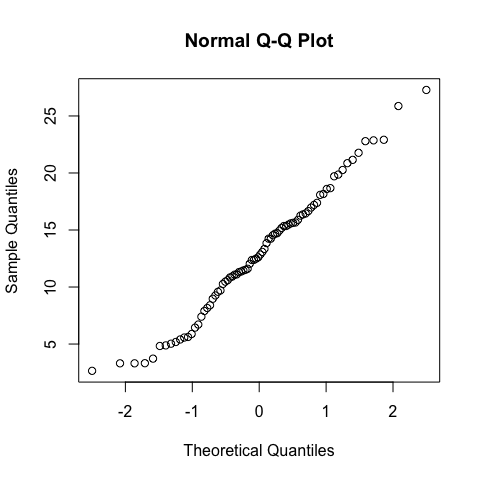
## Q4

# plot the figure   
dfLong <- df %>%  
 pivot\_longer(cols=c(previous,current),names\_to="year",  
 values\_to="price")  
dfLong$year <- factor(dfLong$year,  
 levels=c("previous","current"))  
  
dfSum <- dfLong %>%  
 group\_by(year) %>%  
 summarise(mean=mean(price),  
 sd=sd(price),  
 n=n(),  
 sdErr=sd/sqrt(n)) %>%  
 ungroup()  
  
dfLong %>%  
 ggplot(mapping=aes(y=price,x=year,colour=year)) +  
 geom\_path(aes(group=names),colour="grey",  
 linewidth=0.3) +  
 geom\_jitter(alpha=0.4,size=3,width=0.02,height=0) +  
 geom\_errorbar(data=dfSum,width=0.15,  
 colour="black",  
 mapping=aes(y=mean,  
 ymin = mean-sdErr,  
 ymax = mean+sdErr)) +  
 theme\_bw() +  
 theme(legend.position="none") +  
 scale\_fill\_brewer(palette="Set1") +  
 scale\_colour\_brewer(palette="Set1") +  
 labs(title="Price change over time",  
 subtitle="Each dot is one kind of food item",  
 y="Average price", x="Year")

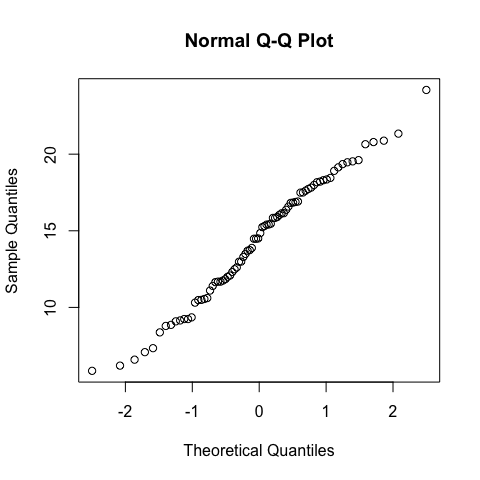


**Q4a**

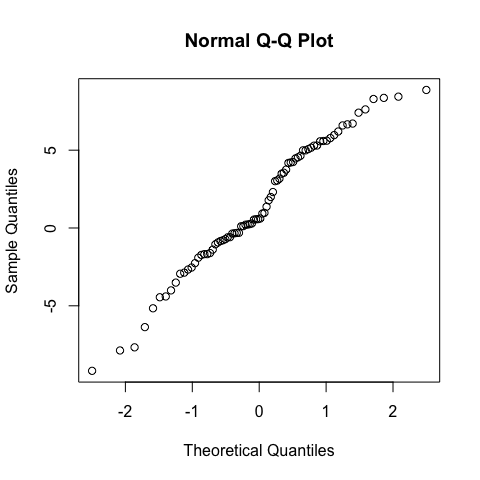
# any code goes here  
  
df <- df %>%  
 mutate(diff = (df$current - df$previous))  
  
  
qqnorm(df$previous)



qqnorm(df$current)



qqnorm(df$diff)



shapiro.test(df$previous)

##   
## Shapiro-Wilk normality test  
##   
## data: df$previous  
## W = 0.98268, p-value = 0.3524

shapiro.test(df$current)

##   
## Shapiro-Wilk normality test  
##   
## data: df$current  
## W = 0.98395, p-value = 0.416

shapiro.test(df$diff)

##   
## Shapiro-Wilk normality test  
##   
## data: df$diff  
## W = 0.97522, p-value = 0.1223

*ANSWER: The main assumption is that our data is normally distributed. We asssesed the datapoints through a qqnorm, before executing the Shapiro-Wilk test to ensure that data is normally distributed for the difference in scores between previous and current year (W = 0.975, p-value = 0.122). Results were not significant indicating that our data indeed is normally distributed. [Word Count: 57]*

**Q4b**

# any code goes here  
  
  
mean(df$previous)

## [1] 12.95425

mean(df$current)

## [1] 14.33988

# I was just interested in looking at the mean of both years too.  
  
t.test(x = df$current, y = df$previous, paired = TRUE)

##   
## Paired t-test  
##   
## data: df$current and df$previous  
## t = 3.0119, df = 79, p-value = 0.003486  
## alternative hypothesis: true mean difference is not equal to 0  
## 95 percent confidence interval:  
## 0.4699315 2.3013185  
## sample estimates:  
## mean difference   
## 1.385625

cd <- cohensD(df$current, df$previous, method = "paired")  
  
print(cd)

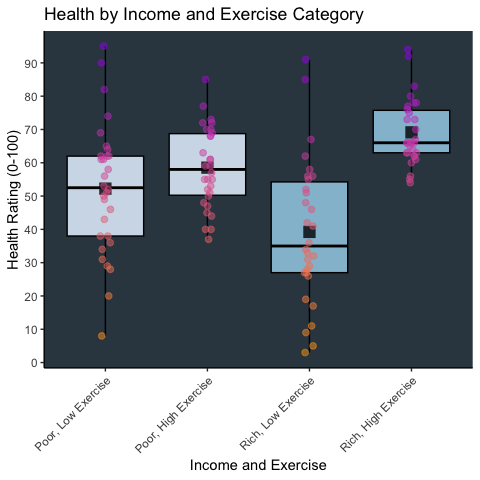
## [1] 0.3367456

*ANSWER: A paired samples t-test was used to analyze Super Size’s question: to compare the two means for the same subject (food items) in this repeated measures design. We found t = 3.012, df = 79, p-value = 0.003 showing a significant difference in the means between the previous and current year, with a mean difference of 1.386. The test also provides a 95 percent confidence interval ranging from 0.470 to 2.301. Further, calculated effect size using Cohen’s D = 0.337. This falls between small effect size ≈ 0.2 and medium effect size ≈ 0.5, indicating an effect, but between small and medium. In light of the research question, this shows a statistically significant increase in food prices (outcome) from the previous year to the current year (predictor). [Word Count: 127]*

## Q5

**Q5a**

# any code goes here  
  
dh\_boxplot <- dh %>%  
 ggplot(data = dh, mapping=aes(x = interaction(inCat, exCat), y = health, fill = exCat)) +  
 geom\_boxplot(alpha = 0.9, color ="black") +  
 stat\_summary(fun = mean, geom = "point", shape = 15, size = 4, color = "black", alpha = 0.8) +  
 geom\_jitter(aes(color = health), size = 2, alpha = 0.5, position = position\_jitterdodge(jitter.width = 0.2, dodge.width = 0.9)) +  
 theme\_classic() +  
 theme(panel.background = element\_rect(fill = "#36454F"),  
 legend.position = "none") +  
 scale\_y\_continuous(breaks = seq(0,100, by = 10)) +  
 scale\_x\_discrete(labels = c("Poor, Low Exercise", "Poor, High Exercise","Rich, Low Exercise", "Rich, High Exercise"), guide = guide\_axis (angle=45)) +  
 scale\_color\_gradient(  
 low = "orange",  
 high = "purple"  
 ) +  
 scale\_fill\_brewer(  
 palette = "Blues") +  
 labs(  
 title="Health by Income and Exercise Category",  
 y="Health Rating (0-100)",  
 x="Income and Exercise")  
  
print(dh\_boxplot)



**Q5b**

# any code goes here  
  
dh\_aov <- aov(health ~ inCat\*exCat, data = dh)  
  
print(dh\_aov)

## Call:  
## aov(formula = health ~ inCat \* exCat, data = dh)  
##   
## Terms:  
## inCat exCat inCat:exCat Residuals  
## Sum of Squares 9828.30 43.20 4177.20 32139.67  
## Deg. of Freedom 1 1 1 116  
##   
## Residual standard error: 16.6453  
## Estimated effects may be unbalanced

summary(dh\_aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## inCat 1 9828 9828 35.473 2.83e-08 \*\*\*  
## exCat 1 43 43 0.156 0.693667   
## inCat:exCat 1 4177 4177 15.077 0.000172 \*\*\*  
## Residuals 116 32140 277   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*ANSWER: A two-way ANOVA was conducted to investigate the income category, exercise category and interaction between them as the predictor variables, and with health as the outcome variable. The ANOVA test shows a statistically significant main effect of income category F(1,116) = 35.47, p < .001, and a significant interaction effect between income and exercise category F(1,116) = 15.08, p < .001. In terms of our research question, this shows that income category does have a significant effect on health, and although exercise does not show a significant effect, it does show a significant interaction between income and exercise category. This can indicate that exercise has different effects depending on income. In Q5A we can see that the health ratings (median and means) between low and high exercise appear to be more extreme in the rich category. As mentioned by Hugo, diet and nutrition is important data we could use to further evaluate these effects. [Word Count: 154]*

## Q6

**Q6a**

# any code goes here  
  
model2 <- lm(health ~ income + exercise, data = dh)  
  
summary(model2)

##   
## Call:  
## lm(formula = health ~ income + exercise, data = dh)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -47.474 -11.639 1.287 11.410 43.464   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 24.593939 8.592920 2.862 0.00499 \*\*   
## income 0.379854 0.084126 4.515 1.52e-05 \*\*\*  
## exercise -0.007382 0.007034 -1.049 0.29613   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.27 on 117 degrees of freedom  
## Multiple R-squared: 0.154, Adjusted R-squared: 0.1396   
## F-statistic: 10.65 on 2 and 117 DF, p-value: 5.627e-05

*ANSWER: A linear multiple regression was conducted to investigate the effects on the outcome variable health by the predictor variables income and exercise. Our model is statistically significant F(2,117) = 10.65, p < .001, with a r-squared of 0.154, indicating that our model accounts for 15.4 % of the variance in health. We found income t(117) = 4.515, p < .001, with a slope of 0.380 suggesting that each increase in income resulted in an 0.380 increase in health. Effects of exercise were not significant t(117) = -1.049, p = .296. In light of the research question, our model suggests a strong relationship between income and health, however the role of exercise is not significant and therefore unclear, and warrants further analysis to determine the role of exercise and health. [Word Count: 129]*

**Q6b**

# any code goes here  
  
model2 <- lm(health ~ income + exercise + income:exercise, data = dh)  
  
summary(model2)

##   
## Call:  
## lm(formula = health ~ income + exercise + income:exercise, data = dh)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -45.147 -8.374 0.236 9.073 47.437   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.215e+02 1.538e+01 7.901 1.75e-12 \*\*\*  
## income -7.240e-01 1.701e-01 -4.257 4.23e-05 \*\*\*  
## exercise -1.952e-01 2.699e-02 -7.232 5.49e-11 \*\*\*  
## income:exercise 2.131e-03 2.989e-04 7.131 9.18e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.3 on 116 degrees of freedom  
## Multiple R-squared: 0.4118, Adjusted R-squared: 0.3966   
## F-statistic: 27.08 on 3 and 116 DF, p-value: 2.399e-13

*ANSWER: (i) The interaction between exercise and income t(116) = 7.131, p < .001, indicates a statistically significant interaction effect between income and exercise on health; (ii) Our model in 6A was statistically significant F(2,117) = 10.65, p < .001, with a r-squared of 0.154. However, our current model F(3,116) = 27.08, p < .001., with a r-square of 0.412, is statically significant and explains 41.2 % of the variance. In sum, this shows that our current model, that includes interaction effects, is able to explain more of the relationship between income and exercise and it’s effect on health, and therefore a better fit to our data.; (iii) All effects in our current model are statistically significant. However, the unstandarized coefficient of income changed to negative in the current model. This indicates that the isolated effect of income on health is negative, and the isolated effect of exercise on health is negative. However, when combined they show a positive effect, therefore there is also an interaction effect highlighting that they should be interpreted together. [Word Count: 172]*

**Q6c**

*ANSWER: In Q5b we found an effect of income and a interaction effect, however in Q6b we found effects from income, exercise and interaction effects. Both analysis lead to the same qualitative conclusion: that there is an effect of income and an interaction effect. But Q6b finds an effect of exercise by itself, and also captures the different effects in more detail as well as providing R-squared helping to explain variance. If the goal is simply to investigate any effect, Q5b is more than sufficient, however if the data is to be used for political decision-making we should go for the method used in Q6b. [Word Count: 104]*

## Q7

**Q7a**

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

**Q7b**

*ANSWER: (i) The chi-square statistic is a measure of our data against a theoretical prediction, where larger values mean a worse fit to the data. A value closer to zero would indicate that our data and the theoretical data is very similar. A negative chi-square value would therefore not make any sense here.; (ii) A t-test determines if the means of two groups are significantly different, by comparing our data to a theoretically predicted mean or another frequency table. In option B the t-value is really big, but at the same time the p-value is not significant. If there was a big effect we would also expect a statstically signficant p-value.; (iii) The small F-value in option A indicates that the variable does not explain a significant amount of variance for the depedent variable, therefore a significant value of p < .001 does not make sense. Like in e.g. Q5b, we would rather expect large F-values to be stastically signficant. [Word Count: 158]*

## Q8

*ANSWER: My favorite character is probably Gladly. He suggested that everyone weighed themselves rather than fighting in Otherland. And also I like the quotes: “I think it’s just that food is very yummy and people like to eat all of the time”, as well as “Well… I am hungry all of the time”.It makes me laugh, and also Gladly’s got a point. Food can be quite yummy!*