Q1.a

First, I'll input the data into a contingency table.

```
tab <- matrix(c(14, 6, 7, 7, 7, 1), ncol=3, byrow=TRUE) colnames(tab) <- c(‘NStpd’,‘BRqst’,‘StpdGW’)
rownames(tab) <- c(‘UClass’,‘LClass’) tab <- as.table(tab)
```

```
print(tab)
```

```
classdf <- as.data.frame.matrix(tab)
```

```
print(classdf)
```

```
str(classdf)
```

These are our observed frequencies. Now we need a model which predicts the

expected frequencies. The total number of drive-bys by Upper Class (UClass)

folks is 27, and the total number of drive-bys by Lower Class (LClass) folks

is 15. This brings our total sample of drive-bys to 42.

```
rowSums(tab)
```

Out of all the instances of drive-bys, the total amount not stopped (NStpd)

was 21, the total amount stopped and requested a bride (BRqst) was 13, and

the toal amount stopped and given a warning was 8.

```
colSums(tab)
```

Now, we want to know the expected frequency UClass folks that don't get

stopped to compare with our observed frequency of 14. We do this by

multiplying 27/42 x 21/42. This gives us 9/28, or 0.3214. We multiply this

probability by the total number of drive-bys: 0.3214 x 41 = 13.49 (13.5). We

we will do this for all the cells, and this gives us:

Uclass - 13.5 8.36 5.14 | 27

LClass - 7.5 4.64 2.86 | 15

21 13 8 | T: 42

We can note that our marginal totals remain the same. Now we must determine

whether the difference in the expected frequencies and the observed

frequencies are significant. To do this, we apply the Chi-square formula.

```
EO_tab <- matrix(c(14, 13.5, 7, 7.5, 6, 8.36, 7, 4.64, 7, 5.14, 1, 2.86), ncol=2, byrow=TRUE)
colnames(EO_tab) <- c('Obsrvd','Expctd') rownames(EO_tab) <- c('UClass_NStpd','LClass_NStpd',
'UClass_BRqst', 'LClass_BRqst', 'UClass_StpdGW', 'LClass_StpdGW') EO_tab <- as.table(EO_tab)
print(EO_tab)
```

The next operation is to square the difference in the (O)bserved and the

(E)xpected. For example, the difference in the O and the E for Upper-Class

folks not being stopped is .5. We then square this–which gives us 0.25–and

we divide this by E, where E in this instance is 13.5. This gives us 0.01856,

or 0.019. We'd do this for all rows.

Once we have performed the (O-E)2 / E operation on each row, we add the

results to get X^2.

```
chisqr <- 0.019 + 0.033 + 0.666 + 1.200 + 0.673 + 1.209
```

```
print(chisqr)
```

Alternatively, we could use the chisq.test() to perform this much faster:

```
chisq.test(tab)
#Q1.b
```

As we saw when running the chisq.test(tab), the p-value is 0.1502. But how

do we do this longhand?

First we need to prove the chisq.test(tab) output for the degrees of freedom

(df). To do this, we apply the following df = (r-1)(c-1). Let's look back at

our original table:

```
print(tab)
```

We note that there are 2 rows and 3 columns. Now we can apply the DF

procedure:

```
df = (2-1)*(3-1)
print(df)
```

With the X^2

```
p = pchisq(3.8, 2, lower.tail = FALSE)
print(p)
```

The hypothesis for this project would stand: Police are more likely to seek a

bribe from someone from the Upper Class than someone from the Lower Class. The

null hypothesis would then be: Police are more less likely to seek a bribe

from someone from the Upper Class than someone from the Lower Class.

Given that alpha is less than our p-value, we can reject the null hypothesis.

Therefore, it is statistically significant that police are more likely to seek

a bribe from someone from the Upper Class than someone from the Lower Class.

#Q1.c

X2_classdf <- chisq.test(classdf)

resdf <- X2_classdf$residuals

#Q1.d

I'm not sure. I've already spent 10 hours on Q1.

Q2

wb_wp <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv") attach(wb_wp)

summary(wb_wp)

str(wb_wp)

water <- table(wb_wp$reserved, wb_wp$water) barplot(water, main="Water and Women in Politics", xlab="Have Reserved Positions for Women", col=c("darkblue","red"), legend = rownames(water), beside=TRUE)

Q2.a

Null Hypothesis: GPs without reserved seats for women have more repaired or

new drinking water facilities.

Alternative Hypothesis: GPs with reserved seats for women have more repaired

or new drinking water facilities.

Q2.b

w_w <- lm(wb_wp$water\ wb_wp$reserved) w_w

summary(w_w)

summary(w_w)$r.squaredcor(wb_wp$reserved, wb_wp$water)^2

plot(wb_wp$water\ wb_wp$reserved) abline(w_w)

Q2.c

Based on the data, it sppears that the null hypohtesis is false, that is,

GPs with reserved seats for women repair and build new facilties for drinking

water.

Q3

install.packages("ggpubr")

library(ggpubr)

Q3.1

dat<-read.csv("http://stat2.org/datasets/FruitFlies.csv") attach(dat)

summary(dat)

Life_ID <- ggplot(dat, aes(x=ID, y=Longevity, group=Treatment)) + geom_point(aes(shape=Treatment, color=Treatment), size=2)+ scale_shape_manual(values=c(21, 22, 23, 24, 25))+ scale_color_manual(values=c("#1B9E77", "#D95F02", "#7570B3", "#E7298A", "#66A61E"))+ theme(legend.position="top") + geom_smooth(method=lm,se=FALSE,

plot(Life_ID)

Here I plotted Longevity by specimen, keeping the groups visually distinct.

The initial investigation seems to show that two groups, one with 1 pregnant

fruit and the other with no female fruit flies, seem to have the most amount

of days lived.

Q3.2 & 3 & 4

Life_Thorax <- ggplot(dat, aes(x=Thorax, y=Longevity, group=Treatment)) + geom_point(aes(shape=Treatment, color=Treatment), size=2)+ scale_shape_manual(values=c(21, 22, 23, 24, 25))+ scale_color_manual(values=c("#1B9E77", "#D95F02", "#7570B3", "#E7298A", "#66A61E"))+ theme(legend.position="top") + geom_smooth(method=lm,se=FALSE,

plot(Life_Thorax)

Life_Thorax_CC <- cor.test(Thorax, Longevity, method=c("pearson"))

X <- c(Thorax) Y <- c(Longevity)

res <- cor.test(dat*Longevity*, *dat*Thorax, method = "pearson") print(res)

Correlation coefficient for the two variables is 0.6364835. It looks like

there is a linear relationship.

Q.3.5

sample.mean <- mean(dat$Thorax) print(sample.mean)

sample.n <- length(dat*Thorax*)*sample.sd* < −*sd*(*dat*Thorax) sample.se <- sample.sd/sqrt(sample.n) print(sample.se)

alpha = 0.05 degrees.freedom = sample.n - 1 t.score = qt(p=alpha/2, df=degrees.freedom,lower.tail=F) print(t.score)

margin.error <- t.score * sample.se

lower.bound <- sample.mean - margin.error upper.bound <- sample.mean + margin.error print(c(lower.bound,upper.bound))

l.model <- lm(Thorax ~ Longevity)

print(l.model)

confint(l.model, level=0.90)

#Q3.6

thorlifepre <- data.frame(thorax= c(.8))

predict(l.model, newdata = thorlifepre)

I just have no idea what's going on anymore.