Problem Set 2

Applied Stats/Quant Methods 1

Jack Merriman

Question 1

(a)

I begin by creating a matrix with the observations from the corruption data

```
corruption <- matrix(c(14, 6, 7, 7, 7, 1), ncol=3, byrow=TRUE)
colnames(corruption) <- c("notStopped", "bribeRequested", "stoppedWarned")
rownames(corruption) <- c("upperClass", "lowerClass")
```

Then I create an empty matrix with the same dimensions and for each cell divide the row totals by the column totals, before multiplying by the grand total, to assign expected values for the original data set to the new matrix

```
expCorruption <- matrix(, ncol = 3, nrow = 2)
for (i in 1:nrow(corruption)){
  expCorruption[i, 1] <- (sum(corruption[i,])*sum(corruption[,1]))/sum(
    corruption)
  expCorruption[i, 2] <- (sum(corruption[i,])*sum(corruption[,2]))/sum(
    corruption)
  expCorruption[i, 3] <- (sum(corruption[i,])*sum(corruption[,3]))/sum(
    corruption)
}
```

Then I make use of vectorised operations to apply $\frac{Row\ total}{Grand\ total} \times Column\ total$ to every cell and find the squared residuals

```
chiCorruption <- ((corruption-expCorruption)^2)/expCorruption
chiCorruption
```

Then by summing the resulting matrix I find the χ^2 test statistic which is 3.791168

```
chi <- sum(chiCorruption)
chi
[1] 3.791168
#I can then check my workings by using the chisq.test() function
chisq.test(corruption)
[1] X-squared = 3.7912, df = 2, p-value = 0.1502
#the chi squared value matches the value I calculated by hand
```

(b)

The p-value for the chi^2 test statistic can be calculated using the `pchisq()` function, where the degrees of freedom are $df = (rows - 1)(columns - 1) = (2 - 1)(3 - 1) = 2$

```
pchisq(chi, df = 2, lower.tail = FALSE)
[1] 0.1502306
```

(c)

Using the same method as I used in *(a)* I assign the standardised residuals to each cell of an empty vector using the $\frac{f_{observed} - f_{expected}}{standard\ error}$ formula this time.

```
resCorruption <- matrix(, ncol = 3, nrow = 2)
for (i in 1:nrow(corruption)){
  resCorruption[i, 1] <- (corruption[i,1]-expCorruption[i,1])/(
    sqrt(expCorruption[i,1]*(1-(sum(corruption[,1])/sum(corruption)))*
           (1-(sum(corruption[i,])/sum(corruption)))))
  resCorruption[i, 2] <- (corruption[i,2]-expCorruption[i,2])/(
    sqrt(expCorruption[i,2]*(1-(sum(corruption[,2])/sum(corruption)))*
           (1-(sum(corruption[i,])/sum(corruption)))))
  resCorruption[i, 3] <- (corruption[i,3]-expCorruption[i,3])/(
    sqrt(expCorruption[i,3]*(1-(sum(corruption[,3])/sum(corruption)))*
           (1-(sum(corruption[i,])/sum(corruption)))))
}
```

This outputs the following values:

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.322	-1.642	1.523
Lower class	-0.322	1.642	-1.523

(d)

None of the absolute values of our standardised residuals are greater than 3, so that means none of the observations are outliers.

Question 2

(a)

The null hypothesis is that there is no observable linear relationship between the number of new or repaired drinking-facilities in villages and the presence of a policy mandating a female council lead, notated as:
$H_0 : \rho_{y \sim x} = 0$
and the converse alternative hypothesis as:
$H_A : \rho_{y \sim x} \neq 0$
Where x is whether or not the policy is in place, and y is the number of new or repaired drinking facilities.

(b)

I run a bivariate regression using the `lm()` function and assign it to a variable so I can create a confidence interval with `confint()`

```
policyData <- read.csv(url("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv"))
#specify the variables being examined
polModel <- lm(formula = water ~ reserved, data = policyData)
confint(polModel)
[1]              2.5 %    97.5 %
(Intercept) 10.240240 19.23640
reserved     1.485608 17.01924
```

We can see that the confidence intervals for the correlation coefficient are: $1.49 \leq \rho \leq 17.02$
As the 0 falls outside of this 95% confidence interval, we reject the null hypothesis.

(c)

We find the coefficient with `summary()`:

```
summary(polModel)
[1] Residuals:
    Min       1Q    Median     3Q     Max
 -23.991 -14.738  -7.865   2.262 316.009
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.738      2.286   6.446 4.22e-10 ***
reserved       9.252      3.948   2.344   0.0197 *
```

Our sample coefficient is 9.252, as our x variable is a binary variable with only two possible values (0 and 1), we can see that on average, villages with the policy mandating a female council leader have **on average** 9.252 more new or repaired drinking facilities than those who do not (the intercept).