Problem Set 2

Luke Duggan, 16316834

October 15, 2021

Question One

(a) Pearson's χ^2 test statistic is $\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$; where we sum over all "bins" i in our contingency table, O_i is the observed quantity in bin i, and E_i is the "expected quantity" in bin i (it is also possible to write this as a nested sum over all rows i and columns j, but not necessary for our purposes). We have the O_i's; but what are the E_i's?

Suppose the class of the driver and the response of the police are independent. Then, to take an example, the probability that an upper class driver is not stopped should just be the product of the probabilities that (A) a driver is upper class and (B) that they are not stopped (since, from elementary probability theory, independence of random variables implies that the joint probability is the product of the marginal probabilities). This gives the probability that an event occurs: the expected number of occurrences is the sample size times this probability (formally, the event that an observation falls into a bin or doesn't fall into that bin is a Bernoulli random variable, with the quantity discussed above the probability p that the event occurs; then, the expected value of the associated binomial random variable with n such Bernoulli trials is, from elementary probability theory, np).

Thus, $E_{1,1}$, the expected number of upper class drivers not stopped, is:

$$E_{1,1} = \frac{27}{42} * \frac{21}{42} * 42 = 13.5$$

The expected values for the other bins are calculated similarly, and are:

$$E_{2,1} = \frac{15}{42} * \frac{21}{42} * 42 = 7.5$$
$$E_{1,2} = \frac{27}{42} * \frac{13}{42} * 42 = approx.8.36$$
$$E_{2,2} = \frac{15}{42} * \frac{13}{42} * 42 = approx.4.64$$
$$E_{1,3} = \frac{27}{42} * \frac{8}{42} * 42 = approx.5.14$$
$$E_{2,3} = \frac{15}{42} * \frac{8}{42} * 42 = approx.2.86$$

This supplies the expected values E_i. To calculate Pearson's test statistic, for each i we subtract the expected value from the observed value in that bin, square the result, and divide it by the expected value. We then sum over all the i, i.e., over every bin in the table. Doing this, we obtain a value of the test statistic equal to approximately 3.79. These calculations are replicated in the attached R script file.

(b) To find the p-value, we first calculate the degrees of freedom, which in this case is 2. Then, we can use R's inbuilt p-value function (see the attached R script). We find that the p-value is approximately equal to 0.15.

The p-value is the probability of the test statistic taking the observed value under the assumption that the null hypothesis is true. We find that this probability is around 15%. Since this is greater than the significance level 10%, we cannot at that significance level reject the null that class and the response of the police are independent.

(c) The residual is just the difference between the observed and expected value. The *standardized* residual is this quantity divided by the square root of the expected value. Note that the standardized residual for bin i is the square root of the term for i in Pearson's test statistic. These residual values are as follows (all values are rounded to two decimal places):

$$\begin{aligned}
E_{1,1} &= \tfrac{(14-13.5)}{\sqrt{13.5}} = .14 \\
E_{2,1} &= \tfrac{(7-7.5)}{\sqrt{7.5}} = -.18 \\
E_{1,2} &= \tfrac{(6-8.36)}{\sqrt{8.36}} = -.82 \\
E_{2,2} &= \tfrac{(7-4.64)}{\sqrt{4.64}} = 1.09 \\
E_{1,3} &= \tfrac{(7-5.14)}{\sqrt{5.14}} = .82 \\
E_{2,3} &= \tfrac{(1-2.86)}{\sqrt{2.86}} = -1.1
\end{aligned}$$

Or, in tabular form:

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	.14	-.82	.82
Lower class	-.18	1.09	-1.1

(d) The standardized residuals can be used to compare the different categories in terms of the magnitude of the difference from their expected values. A large standardized residual indicates a great divergence from the expected value; a small residual indicates

that the observed value is close to the expected value. Similarly, a positive standardized residual indicates the observed value has overshot the expected value; a negative residual indicates that it is below the expected value.

More importantly: each term in Pearson's test statistic is asymptotically chi-square distributed. Since a chi-square random variable is the square of a standard normal random variable, it stands to reason that the standardized residuals, which are the square roots of the terms in Pearson's test statistic, should be standardly normally distributed. Thus, the sum of the standardized residuals can also be used as a test statistic, one which is asymptotically standardly normally distributed.

Question Two

(a) We could form a linear model where the number of new / repaired water facilities is the outcome variable and the binary variable of whether a given council has implemented the reservation policy is the explanatory variable.

Then, if we want a two-tailed alternative hypothesis, the null should be that the slope coefficient is zero, that is, there is no linear effect of the reservation policy on the number of water facilities. The alternative hypothesis is that the reservation policy has some linear effect on the implementation of a policy desired by female voters: although this effect could be positive or negative.

(As a side note, however, it seem that this would not be the best way to test the hypothesis that women tend to support policies that female voters want. It seems we should instead have a one-tailed alternative hypothesis: the null should be that the slope coefficient is nonpositive, that is, that the reservation policy is either unrelated to or actually detrimental to the implementation of this policy that female voters want; and the alternate hypothesis should be that the coefficient is strictly positive).

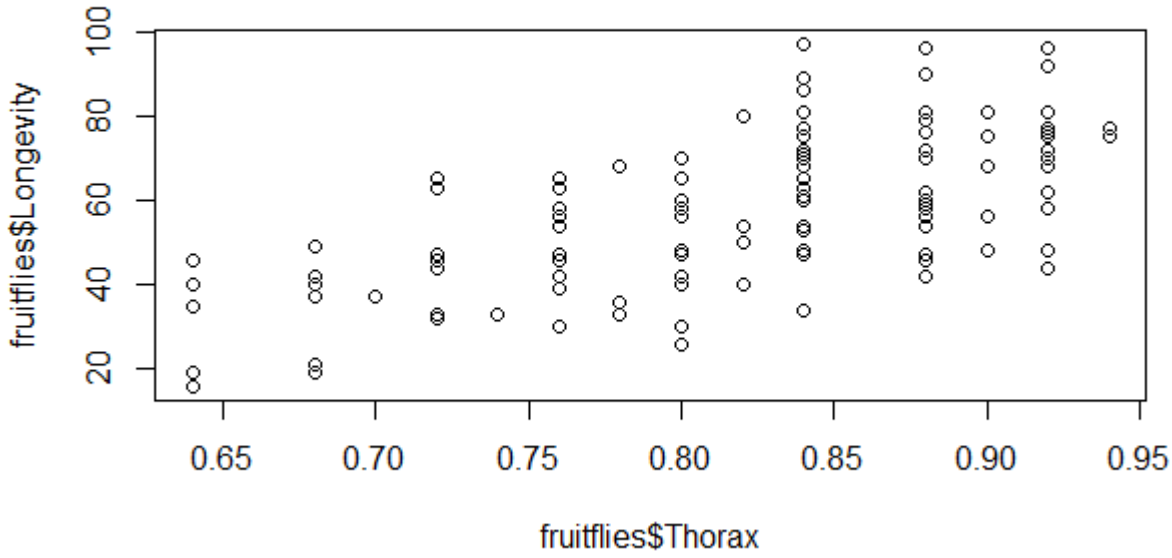
(b) This is answered in the attached R script file. The estimated parameters are calculated with R's built-in functions and manually.

(c) We see that the estimated slope coefficient is about 9.25. The interpretation of this coefficient is as follows. Since the "reserved" variable is a binary variable, it can only take the values 0 and 1. If it takes the value 1, this is associated with an increase of 9.25 in the outcome variable. In English, if a council has the reservation policy, this is associated with about 9.25 more new / repaired water facilities in the council. R also tells us that this point estimate has an associated p-value of just under .02, that is, under the null that the slope coefficient is just 0, there is only a 2% probability of the observed result. Thus, among conventional significance levels, the point estimate is significant at the 5% level.

Question Three

(a) This is answered in the attached R script file.

(b) 2. We below plot lifespan vs thorax.



Graphically, there appears to be at least a moderate positive linear relationship. We can verify this analytically by computing Pearson's correlation coefficient for the two variables. This is done in the attached R script file (the built-in function is used, since it is not hard to calculate the coefficient by hand: the sample covariance of the two variables divided by the product of their sample standard deviations). The result is around 0.64, indeed indicating a moderately strong positive linear relationship.

(c) The regression is performed in the attached R script file (since we've already exhibited a manual calculation of the OLS estimates once, we just used the built-in function this time). We obtain an estimated slope parameter of around 144.33. As with all linear models, this is the "partial effect" of the explanatory variable on the outcome variable: a unit increase in the explanatory variable is associated with so large an increase in the outcome variable. In this case, it would seem an extra millimeter in the length of a fruit fly's thorax is associated with a longer lifespan of around 144 days.

(d) To test for a significant linear relationship, we perform a t-test on the null hypothesis that the slope parameter is zero, against the two-tailed alternative hypothesis that it is nonzero. If we fail to reject the null at some level of significance, then there is no statistically significant linear relationship between the variables at that level.

In the attached R script file, this t-test is performed using results provided by R's lm() call and also manually (to an extent).

Extracting the results of our regression call in R, we see that both point estimates are highly statistically significant, at the 0.001 significance level. Both have test statistic values well in excess of the rejection threshold at this level of significance; what comes to the same thing, under the null that the parameter is zero, both values of the t statistic are associated with very small p-values. Attending specifically to the estimate of the slope parameter, we can say with very great confidence that there is a linear relationship between thorax length and the lifespan of a fruit fly.

(e) Using either R's built in function for computing confidence intervals or manually constructing the interval ourselves, we see in the attached script file that the 90% confidence interval for our point estimate of the slope parameter is, roughly, [118.2, 170.5].

(f) To predict the lifespan of an individual fly with a thorax of 0.8mm, it suffices to plug 0.8 into the fitted model / regression line created from our point estimates. This yields a fitted value of around 54.41 days.

We can go further, however, and obtain a prediction interval. This is done in the attached script file. As we see there, the fitted values agree, but we also see that at the 95% confidence level, the fly could live for as little as 27 days and as many as 81. As we will now see, the wide range of this estimate has to do with individual variability, which will be "averaged out" when considering the confidence interval of the mean.

Computing the 95% confidence interval for the average lifespan of fruitflies with a thorax of 0.8mm, we observe that the fitted value agrees with before, but that the endpoints of the interval are (rounding up) around 52 and 57. Thus, while there may be large amounts of variability in the lifespans of fruit flies with thoraxes 0.8mm in length, at the 95% significance level, the variability in the *average* is nowhere near as large.

(g) Unfortunately, I could not get my code to run for this question. I have included it in the R script file to indicate what I was trying to do.