Andrew Downie
0786342

Question 2

Question a) Is there a relationship between tea temperature and area of residence?

Here is the given table describing the relationship between temperature and area of residence:

Area of residence	Temperature				Total
	< 60	60-64	65-69	>= 70	
Urban	4200	3646	1566	537	9949
Rural	14758	15260	6490	2125	38633
Total	18958	18906	8056	2662	48582

Formulate an appropriate null and alternative hypotheses:

H_0: tea temperature and area of residence are independent. (The representation of each tea temperature is the same whether the area of residence is urban or rural)

H_A: tea temperature and area of residence are not independent. (The representation of each tea temperature is different, depending on whether the area of residence is urban or rural)

Hypothesis test:

If H_0 were true, we would expect the number of people in each category to be:

$$\text{Expected category / cell value} = \frac{(\text{row total x column total})}{\text{grand total}}$$

We can fill in a table with expected values like so:

Area of residence	Temperature				Total
	< 60	60-64	65-69	>= 70	
Urban	3882.367	3871.718	1649.77	545.1451	9949
Rural	15075.63	15034.28	6406.23	2116.855	38633
Total	18958	18906	8056	2662	48582

Now that we have the expected values and the given values in two tables, we can compare them to calculate **the chi-square test statistic:**

$$\chi 2 = \frac{(4200 - 3882.367)^2}{3882.367} + \frac{(14758 - 15075.63)^2}{15075.63} + \frac{(3646 - 3871.718)^2}{3871.718}$$
$$+ \frac{(15260 - 15034.28)^2}{15034.28} + \frac{(1566 - 1649.77)^2}{1649.77} + \frac{(6490 - 6406.23)^2}{6406.23}$$
$$+ \frac{(537 - 545.1451)^2}{545.1451} + \frac{(2125 - 2116.855)^2}{2116.855}$$

$$\chi 2 = 54.729$$

P-value:

Using software, we can discover: p-value < 0.00001

Output from R:

```
        Pearson's Chi-squared test

data:  teaData
X-squared = 54.729, df = 3, p-value = 7.843e-12
```

Conclusion:

With a p-value so small, there is evidence against the null hypothesis, and therefore evidence for the alternative hypothesis, regardless of any reasonable alpha value. In the context of this problem, this means that tea temperature and area of residence are not independent. In other words knowing the area of residence for a data set will give hints about the expected distribution of tea temperature consumed for that data set, and vice versa.

Question b) Would the conclusion from question a hold if the columns 65-69 and >= 70 were combined?

To figure out if collapsing the columns labeled 65-69 and >= 70 we could simply add the last two columns together resulting a table that looks like this:

Area of residence	Temperature			Total
	< 60	60-64	>= 65	
Urban	4200	3646	2103	9949
Rural	14758	15260	8615	38633
Total	18958	18906	10718	48582

```
        Pearson's Chi-squared test

data:  tea_stripped_2b
X-squared = 54.068, df = 2, p-value = 1.817e-12
```

By combing the last two columns the resulting chi squared test statistic and matching p-value are indeed different, but not different enough to change the conclusion. Both before and after combining these two columns the evidence against the null hypothesis is overwhelmingly strong at any reasonable alpha level.

This result is expected, as even though the >= 70 column had much smaller values than the rest of the table, the >= 70 columns smallest value was still bigger than 500, which is both a big and significant sample.