5: Pairwise Association

1

The director of Gainesville Sun newspaper is studying the relationship between the type of community in which a subscriber resides and the section of the newspaper he or she reads first. For a sample of readers, she collected the sample information in the following table:

	National News	Sports	Comics
City	170	124	90
Suburb	120	130	100
Rural	130	90	88

a

At the 95% confidence level, can we conclude that there is a relationship between the type of community where the person resides and the section of the paper read first, i.e. are they dependent?

✓ **Answer** ⌄

After conducting a chi-squared test on the data, we get a result of:

	National News	Sports	Comics
City	1.49678055	0.06059503	1.51276038
Suburb	3.1483254	1.80782229	0.46958868
Rural	0.27605267	1.34198298	0.41323804

$\chi^2 = 10.527146061659542$
$p = 0.032425107389563265$

With a null hypothesis that there is independence between the variables and our hypothesis that there isn't indepencence between our variables, by our χ^2 test probability of 0.032425107389563265 and an $\alpha = 0.05$, we reject our null hypothesis that there is independence and thus conclude that these two variables are dependent.

b

What is the largest chi-square value? How would you interpret this?

✓ **Answer**

The largest χ^2 value was 3.1483254, which means that this particular value was highly unexpected — where our expectation was that there is no dependency between the two variables. This means that this variable in particular does not fit the trends of both variables that are intersecting here.

c

Is the number of people that read national news in rural areas greater or less than expected?

✓ **Answer**

With our expected values as:

	National News	Sports	Comics
City	154.77927063	126.77159309	102.44913628
Suburb	141.07485605	115.54702495	93.378119
Rural	124.14587332	101.68138196	82.17274472

The amount of people that read national news first in rural areas being 130 was larger than the expected number of people that would, at 124.14587332 people.

d

Is the number of people that read comics in cities greater or less than expected?

✓ **Answer**

The amount of people that read comics first in cities being 90 was less than the expected number of people that would, at 102.44913628 people.

2

The Federal Correction Agency is investigating whether a male released from a prison make a different adjustment to civilian life if he returns to his hometown or if he goes elsewhere to live? To put it another way, is there a relationship between adjustment to civilian life and place of residence after release from prison?

The counts are given in the following contingency table:

	Outstanding	Good	Fair	Unsatisfactory
Hometown	27	35	33	25
Not Hometown	13	15	27	25

At 99% confidence can we conclude that adjustment to civilian life and residence after release are dependent?

✓ **Answer**

With χ^2 values of

	Outstanding	Good	Fair	Unsatisfactory
Hometown	0.375	0.83333333	0.25	0.83333333
Not Hometown	0.5625	1.25	0.375	1.25

We get
$\chi^2 = 5.729166666666667$
$p = 0.12555660994179765$

With a null hypothesis that there is independence between the variables and our hypothesis that there isn't indepencence between our variables, by our χ^2 test probability of 0.12555660994179765 and an $\alpha = 0.01$, we fail to reject our null hypothesis that there is independence between the variables.

3

In this exercise, our aim is to quantify the associations between continuous variables and assess the statistical significance of these associations. For this purpose, we will use the two datasets that are provided with the assignment:

- The file `p3a.csv` contains a matrix of size 2400×2 which has 2400 samples and 2 variables.
- The file `p3b.csv` contains a matrix of size 110×2 which has 110 samples and 2 variables.

a

For the two variables provided in `p3a.csv`, assess the association between them by computing Pearson correlation r_a and computing a p-value p_a for the null hypothesis of no

association. Select a significance level α and reject the null-hypothesis if the p-value is less than α. Explain, in complete sentences, your findings: Is there a statistically significant association (at α level) between the provided variables? What is the magnitude and the direction of the association?

✓ **Answer**

$r_a = 0.38087503578373005$
$p_a = 1.0409455130062538 \times 10^{-83}$

At an alpha level of $\alpha = 0.01$, we can statistically reject our null hypothesis that there is independence between the two variables, as our $p_a = 1.04 \times 10^{-83}$, making it statistically significant.

However, our $0.5 > r_a > 0$, so it has a relatively weak association, but it is a positive association. In a linear regression we also obtain a slope of 0.5675 ± 0.0281, which supports our r_a value as well.

b

Repeat part (a) for the variable pair provided in `p3b.csv` and compute Pearson correlation r_b and p-value p_b. Compare the Pearson correlations r_a and r_b as well as the p-values p_a and p_b. Explain your findings: Which variable pair (in part a or b) has a stronger association according to the comparison of the correlations? Which variable pair has a stronger association according to the comparison of the p-values?

Next, draw scatter plots (variable 1 vs. variable 2) to visualize the data for both part (a) and part (b). Which variable pair (in part a or b) has a stronger association do you think according to the scatter plots? Does your conclusion agree with the comparisons of the p-values and correlation coefficients? If not, explain why would this happen.

✓ **Answer**

$r_b = 0.9312196333264214$
$p_b = 3.73732100843862 \times 10^{-49}$

At an alpha level of $\alpha = 0.01$, we can statistically reject our null hypothesis that there is independence between the two variables, as our $p_b = 3.74 \times 10^{-49}$, making it statistically significant.

However, our $r_b > 0.9$, so it has a strong positive linear association. In a linear regression we also obtain a slope of 0.6259 ± 0.0236, which supports our r_b value as well.

Our p_b value is significantly larger than our p_a value, although still significant to the point of disproving the null hypothesis of independence. This p-value has no significance other than either disproving the null or being inconclusive, so these p-values are effectively the same, as they show association of some kind.

Our r_b value, however, is significantly higher than our r_a value. This is because our data is much more clustered and linearly related than our first dataset. This agrees with the visual characteristics of the data, as the first one looks like a shotgun blast, but the second one actually looks linearly related.

The plots do agree with the changes in both our r and p-values.