Homework-6: Pairwise Association

Problem 1

The New York Times director is investigating the relationship between where a subscriber lives and which sections of the newspaper they read first. Use the collected sample information to answer the following questions:

	Politics	Editorial	Sports
City	170	124	90
Suburb	120	130	100
Rural	130	90	88

a. At the 95% confidence level, is it reasonable to conclude that the section read first and the community where a reader resides are dependent?

```r
politics <- c(170,120,130)
editorial <- c(124,130,90)
sports <- c(90,100,88)
NewYorkTimeData <- data.frame(politics, editorial, sports)

result <- chisq.test(NewYorkTimeData)
p_value <- result$p.value
print(p_value)
```

```
[1] 0.03242511
```

```r
alpha <- 0.05
```

Since the p-value of 0.03242511 is smaller than the alpha value of 0.05, we reject the null hypothesis. We can not be 95% sure that the section read first and the community where a reader resides are dependent.

b. What is the largest chi-square value? What is the meaning of this value?

```
total <- sum(NewYorkTimeData)
row_totals <- rowSums(NewYorkTimeData)
col_totals <- colSums(NewYorkTimeData)
expected <- outer(row_totals, col_totals) / total

# Calculate chi-squared contributions for each cell
chi_squared_contributions <- (NewYorkTimeData - expected)^2 / expected

# Total chi-square statistic
total_chi_square <- sum(chi_squared_contributions)

# Print chi-squared contributions for each cell
print(chi_squared_contributions)
```

```
   politics  editorial    sports
1 1.4967806 0.06059503 1.5127604
2 3.1483254 1.80782229 0.4695887
3 0.2760527 1.34198298 0.4132380
```

The largest chi-squared value is a someone who resides in a city and reads the sports column first. This chi-squared value represents the overall measure of the association between the sections read first in the New York Times and the communities where readers reside. A larger chi-square value indicates a stronger association between the variables.

c. Which values are greater than expected (e.g city-sports group)?

```
expected <- chisq.test(NewYorkTimeData)$expected
greater_than_expected <- NewYorkTimeData > expected
print(greater_than_expected)
```

```
     politics editorial sports
[1,]     TRUE     FALSE  FALSE
[2,]    FALSE      TRUE   TRUE
[3,]     TRUE     FALSE   TRUE
```

City-politics, rural-politics, suburb-editorial, suburb-sports, rural-sports

d. Which values are smaller than expected (e.g city-sports group)?

```r
expected <- chisq.test(NewYorkTimeData)$expected
less_than_expected <- NewYorkTimeData < expected
print(less_than_expected)
```

```
     politics editorial sports
[1,]    FALSE      TRUE   TRUE
[2,]     TRUE     FALSE  FALSE
[3,]    FALSE      TRUE  FALSE
```

City-editorial, suburb-politics, rural-editorial, city-sports

Problem 2

Investigate the relationship between where a prisoner resides after the release and how they adjust to civilian life using the sample information given:

	Excellent	Good	Satisfactory	Poor
Big city	27	35	33	25
Rural area	13	15	27	25

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

```r
observed <- matrix(c(27, 35, 33, 25, 13, 15, 27, 25), nrow = 2, byrow = TRUE)
result <- chisq.test(observed)

p_value <- result$p.value
print(p_value)
```

```
[1] 0.1255566
```

```r
alpha <- 0.01
```

Since the p-value of 0.1255566 is larger than alpha 0.01, we fail to reject the null hypothesis. We can be 99% confident that the adjustment to civilian life for a prisoner and residence for the prisoner after release are dependent.

Problem 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 x 2 which has 2400 samples and 2 variables.

→ The file p3b.csv contains a matrix of size 110 x 2 which has 110 samples and 2 variables.

Part (a)

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

```
datap3a <- read.csv("p3a.csv")
datap3b <- read.csv("p3b.csv")


correlation_result <- cor.test(datap3a$X.0.64901, datap3a$X2.8005)

r_a <- correlation_result$estimate
p_a <- correlation_result$p.value

# Print the results
print(paste("Pearson correlation coefficient (ra):", r_a))
```

```
[1] "Pearson correlation coefficient (ra): 0.381681173968854"
```

```
print(paste("P-value (pa) for the null hypothesis of no association:", p_a))
```

```
[1] "P-value (pa) for the null hypothesis of no association: 4.74510896514213e-84"
```

Select a significance level and reject the null-hypothesis if the p-value is less than .

```
correlation_result <- cor.test(datap3a$X.0.64901, datap3a$X2.8005)
p_value <- correlation_result$p.value
print(p_value)
```

```
[1] 4.745109e-84
```

> I picked alpha to be 0.05 (for a 95% confidence rate).Since the p-value of 0.03242511 is smaller than the alpha value of 0.05, we reject the null hypothesis.

Explain, in complete sentences, your findings:

1. Is there a statistically significant association (at level) between the provided variables?

2. What is the magnitude and the direction of the association?

Part (b)

Repeat part (a) for the variable pair provided in p3b.csv and compute Pearson correlation rb and p-value pb. Compare the Pearson correlations ra and rb as well as the p-values pa and pb.

```r
datap3b <- read.csv("p3b.csv")


correlation_result <- cor.test(datap3b$X.0.343, datap3b$X.0.72006)

r_a <- correlation_result$estimate
p_a <- correlation_result$p.value

# Print the results
print(paste("Pearson correlation coefficient (ra):", r_a))
```

```
[1] "Pearson correlation coefficient (ra): 0.932898471193837"
```

```r
print(paste("P-value (pa) for the null hypothesis of no association:", p_a))
```

```
[1] "P-value (pa) for the null hypothesis of no association: 2.87193936452566e-49"
```

Explain, in complete sentences, your findings:

1. Which variable pair (in part a or b) has a stronger association according to the comparison of the correlations?

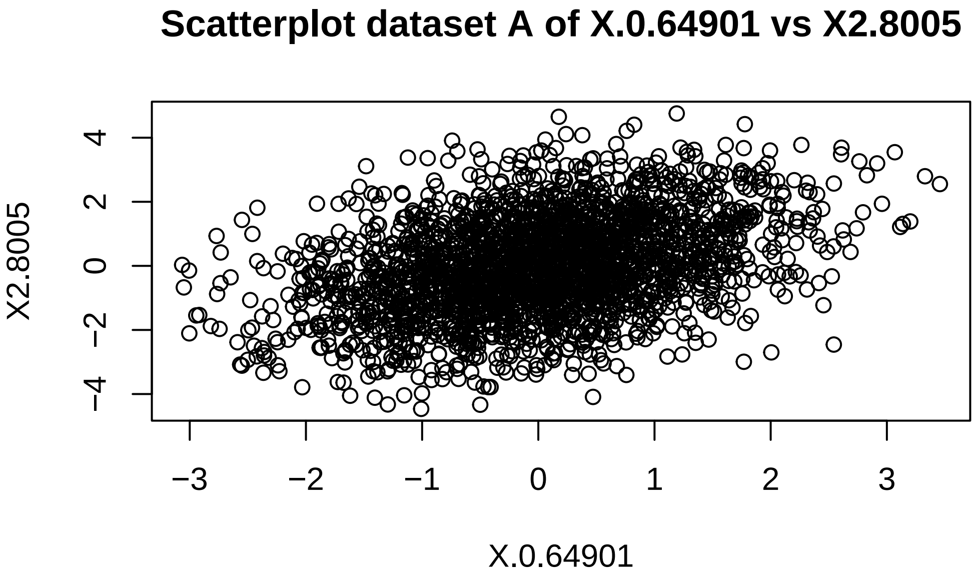
After comparing the absolute values of the correlation coefficents, I would say that b has a stronger association because 0.932898471193837 is MUCH bigger than 0.381681173968854.

2. Which variable pair has a stronger association according to the comparison of the p-values?

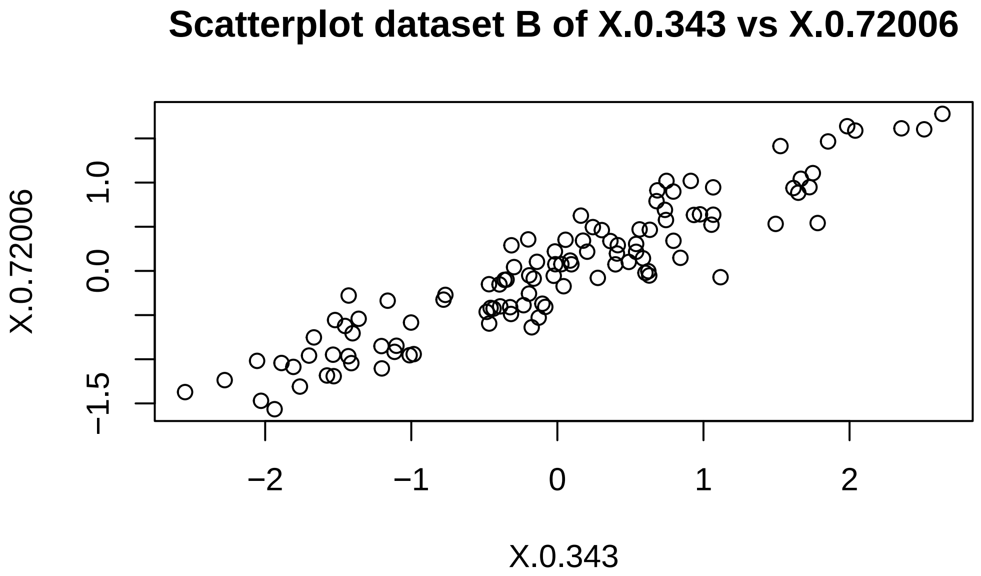
   The p-value for a is 4.745109e-84 and the p-value for b is 4.745109e-84 which means that the variable pair have the same strength of association.

3. Draw scatter plots (variable 1 vs. variable 2) to visualize the data for both part (a) and part (b).

```
plot(datap3a$X.0.64901, datap3a$X2.8005,
     xlab = "X.0.64901", ylab = "X2.8005",
     main = "Scatterplot dataset A of X.0.64901 vs X2.8005")
```



```
plot(datap3b$X.0.343, datap3b$X.0.72006,
     xlab = "X.0.343", ylab = "X.0.72006",
     main = "Scatterplot dataset B of X.0.343 vs X.0.72006")
```



4. Which variable pair (in part a or b) has a stronger association do you think according to the scatter plots?

   According to the scatterplots, I think that b has a stronger association since it is more linear and less dense in one area.

5. Does your conclusion agree with the comparisons of the p-values and correlation coefficients? If not, explain why this would happen.

   My conclusion agrees with the comparisons made with the p-values and the correlation coefficents.