

Assginment 2 Report, DATA 400

Instructor: Dr. Spectrum Han

Yohen Thounaojam, 56112204

Introduction

This report is a part of Assignment 2 of DATA 400, 2021W2 at UBCO. It is composed of three parts (1 report each). Please refer to the Table of Contents for more information.

Page	Content
2	Ex 3.5
7	Ex 4.1
9	Ex 4.3

Other Information

- All R code used in this report can be found in the .Rmd file in the submission. Moreover, for the relevant portions of the report, the code has been ‘*echoed*’ back into this PDF.

Ex 3.5 from Chapter 3: EFT Data

Page 66

Introduction

The data shown below in *Table 1.1* is from an experiment conducted on a sample of 24 primary school children. Each child completed the *Embedded Figures Test* (EFT), which measures ‘field dependence’, i.e., the extent to which a person can abstract the logical structure of a problem from its context.

##	group	time	EFT
## 10	row	739	139
## 11	row	430	74
## 12	row	410	31
## 13	corner	342	48
## 14	corner	222	23
## 15	corner	219	9

Table 1.1 6/24 rows of EFT Data. **time** in seconds.

After each child completed the EFT, they were allocated to one of two experimental groups. The experiment asked the children to construct a 3x3 pattern from nine colored blocks, taken from the Wechsler Intelligence Scale for Children.

In this report, we will first perform some simple Exploratory Data Analysis to get an overall idea of the data. Then, we will consider certain methods to come to conclusions of the data and try to answer the questions we are aiming for.

Later in the report, all the results of the analysis will be presented to come to a conclusion.

Objective

The two experimental groups were: *Corner* and *Row*. They differed in the instructions they were given for the task.

1. The ‘row’ group were told to start with a row of three blocks.
2. The ‘corner’ group were to start with a corner of three blocks.

The goal of the experiment was to answer the following questions:

1. Did the different instructions produce any change in the average time to complete the pattern?
2. Was this time affected by field dependence?

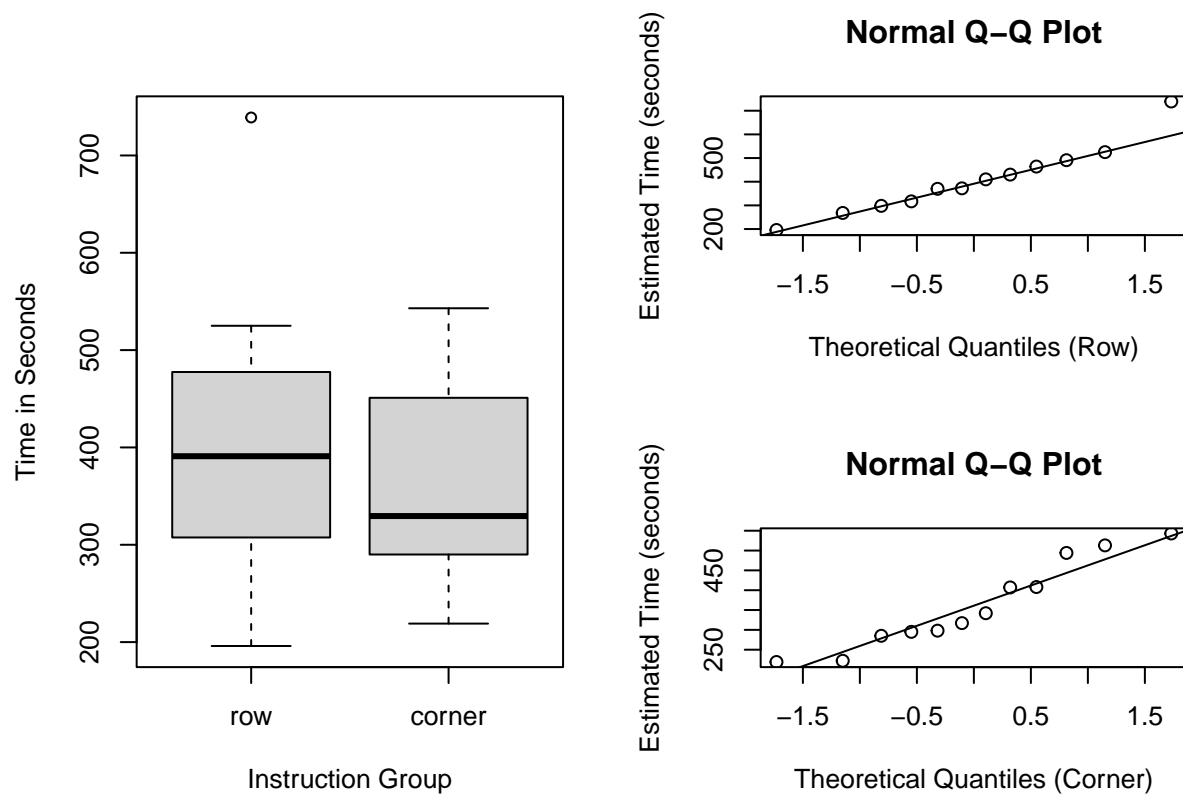
Next, let us look at some Exploratory Data Analysis.

Exploratory Data Analysis

```
##      group      time      EFT
## row   :12  Min.   :196.0  Min.   : 7.00
## corner:12  1st Qu.:297.2  1st Qu.: 31.00
##        Median :371.0  Median : 49.00
##        Mean   :384.3  Mean   : 55.46
##        3rd Qu.:470.8  3rd Qu.: 66.00
##        Max.   :739.0  Max.   :139.00
```

Table 1.2 Summary of data in EFT dataset.

From **Table 1.2** we see that there are two groups of interest, *row* and *corner*. Now that we have an overall idea of the dataset, we can look at some basic Box and Q-Q plots below.



From this box plot above, we can infer the median values of the two groups of interest. Also, we see from the *Normal Q-Q plots* that the *Row* values follow a *Normal Distribution* but not the *Corner* values.

Method

To explain the methods of analysis used in this section, let us recall the questions we want answered. 1. Did the different instructions produce any change in the average time to complete the pattern?

- To answer this question, we will first need to calculate the average time of completion for each *instruction group* and then later compare it.

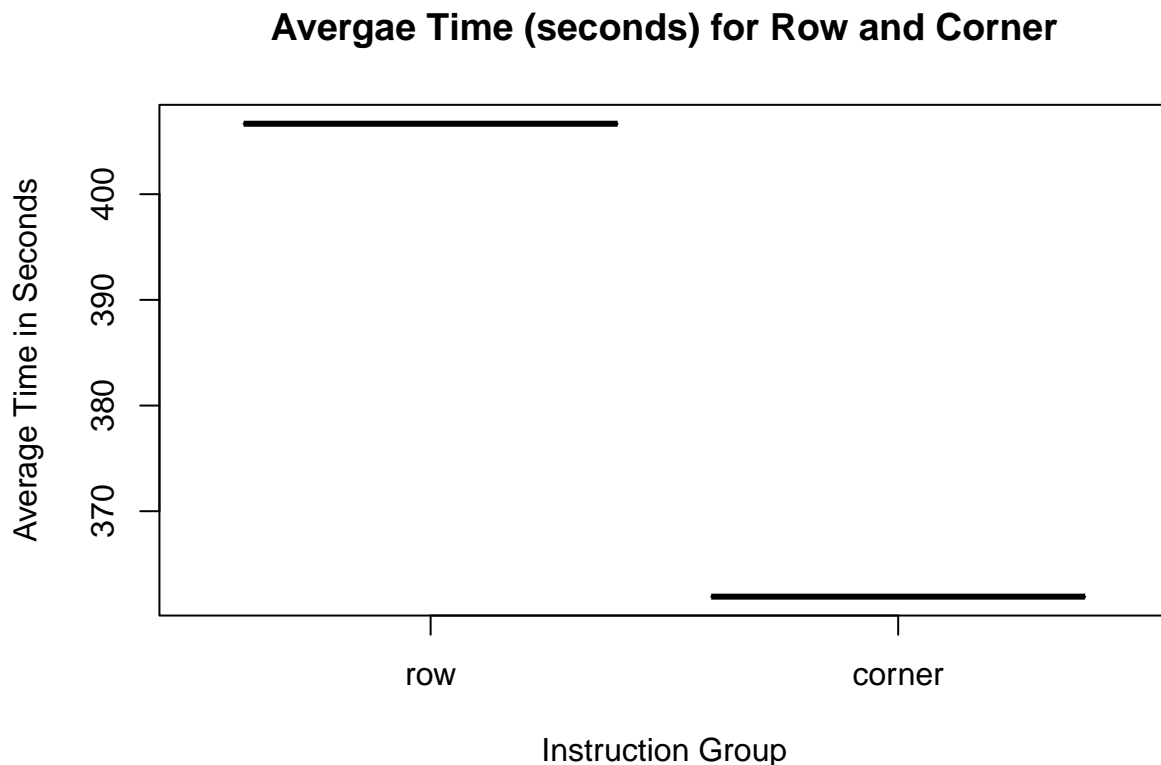
```
EFT_means <- aggregate(time ~ group, EFT, mean)
```

2. Was this time affected by field dependence?

- To answer this question, we can compute the **Pearson's Product-Moment Correlation Coefficient** to see if there is any significant correlation between the two variables, *time* and *EFT*.

Results - Did the different instructions produce any change in the average time to complete the pattern?

In the plot below, we will use the *mean* values we calculated and plot them to answer Question 1.



Here, it is clear that the instruction type changed the result ever so slightly, by ~50 seconds. Instruction Group of *row* has a **higher average completion time**.

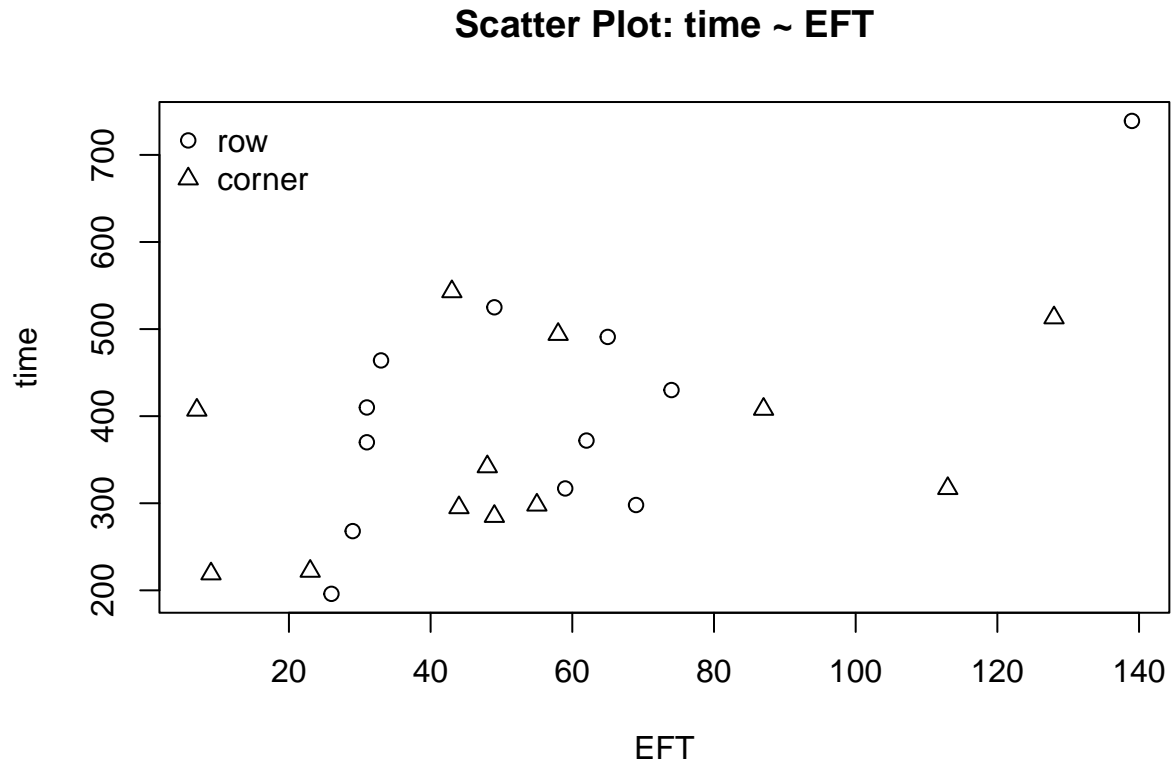
Results - Was this time affected by field dependence?

Next, let us compute the Pearson's product-moment correlation coefficient.

```
##  
## Pearson's product-moment correlation  
##  
## data: time and EFT  
## t = 3.0706, df = 22, p-value = 0.005596  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.1852645 0.7790021  
## sample estimates:  
## cor  
## 0.5477268
```

A correlation coefficient of **0.547** implies only a very slightly **moderate/medium correlation**. This is not significant enough to say the **EFT** affected the **average completion time**.

This can further be proven by the plot below.



Discussions and Conclusion

The conclusions of this analyses are short and can be answered in just two lines:

1. Yes, different instruction groups produce a change in average time. The **row group had a gigher average completion time than the corner group**.
2. No, after computing the Pearson's product-moment correlation coefficient, we can conclude that **Completion Time is not affected by Field Dependence or EFT**.

Recommendations

1. The experiments were conducted on only 24 children which is not a large enough dataset to come to any conclusions. Depending on the method of data sampling, there could have been many biases introduced to the results of this experiment.

Ex 4.1 from Chapter 4: orallesions data

Page 66

Introduction

The data shown below in *Table 2.1* gives the distribution of the oral lesion site found in house-house surveys in three geographic regions of rural India.

##		region		
##	site of lesion	Kerala	Gujarat	Andhra
##	Buccal mucosa	8	1	8
##	Commissure	0	1	0
##	Gingiva	0	1	0
##	Hard palate	0	1	0
##	Soft palate	0	1	0
##	Tongue	0	1	0
##	Floor of mouth	1	0	1
##	Alveolar ridge	1	0	1

Table 2.1 orallesions dataset.

Objective

As laid out in the exercise, this report aims to find:

1. The p -value from Fisher's Test and the corresponding p -value from Fisher's test.
2. Corresponding p -value from applying the usual chi-square test of the data.

Exploratory Data Analysis and Preparation

Since the aim of the exercise is clear and short, we do not need to perform EDA.

* *Please see the next page for results.*

Results

Fisher's Test

```
fisher.test(orallesions)

##
## Fisher's Exact Test for Count Data
##
## data:  orallesions
## p-value = 0.01904
## alternative hypothesis: two.sided
```

Chi-Square Test

```
chisq.test(orallesions)

##
## Pearson's Chi-squared test
##
## data:  orallesions
## X-squared = 20.698, df = 14, p-value = 0.1096
```

Discussions and Conclusion

- p-value from Fisher's Exact Test = 0.01904
- p-value from Chi-Square Test = 0.1096

We see that the p-value we computed using the Chi-Square Test is significant while the p-value from the Fisher's Exact Test is not. Looking back at the difference between the two tests, we might be able to come to some conclusions.

We know that the Chi-Squared test is used to compare the distribution of a categorical variable in a sample or a group with the distribution in another. It also relies on *approximation* while Fisher's Test is *exact*.

So we must be careful about which test we use depending on the *sample size* and the *accuracy* we want.

In conclusion, since the dataset in question is arguably small, we must look at the p-value obtained from Fisher's and conclude that it is not significant.

Ex 4.3 from Chapter 4

Page 66

Introduction

In this exercise, we will generate two groups with measurements following a normal distribution but having different means. Later compare the p-values we compute with two tests, the Wilcoxon Mann-Whitney rank sum test and a permutation test (using `independence_test`).

Objective

The goal of this exercise is to compare the p-values of the mentioned test over multiple replications of the experiment, a 1000 or so.

Next, we find out where the differences come from.

Method

First, we show a sample generation of the data below using **rnorm**, meaning it will have a normal distribution.

```
library("coin")

## Loading required package: survival

sample_data <- data.frame(
  score = rnorm(20),
  sex = as.factor(c(rep("Girl", 10), rep("Boy", 10)))
)

summary(sample_data)
```

```
##      score      sex
## Min.   :-1.9185  Boy :10
## 1st Qu.: -0.1843  Girl:10
## Median :  0.3619
## Mean    :  0.4173
## 3rd Qu.:  0.8644
## Max.    :  2.0713
```

Next, let us check the means:

```
aggregate(score ~ sex, sample_data, mean)

##      sex      score
## 1  Boy  0.5903585
## 2 Girl  0.2441886
```

Results

Below is the method in which we will perform the tests:

```
wilcox.test(sample_data$score ~ sample_data$sex)

##
## Wilcoxon rank sum exact test
##
## data: sample_data$score by sample_data$sex
## W = 58, p-value = 0.5787
## alternative hypothesis: true location shift is not equal to 0

independence_test(sample_data$score ~ sample_data$sex, distribution = exact())

##
## Exact General Independence Test
##
## data: sample_data$score by sample_data$sex (Boy, Girl)
## Z = 0.74795, p-value = 0.4695
## alternative hypothesis: two.sided
```

Next, let us look at replication of the above over 1000 times.

```
library("coin")

# Create empty list to append p_values of 1000 replications
w_values <- rep(NA, 1)
i_values <- rep(NA, 1)

# Loop over 1000 times
for (x in 1:1000) {

  #Create sample data
  sample_data <- data.frame(
    score = rnorm(20),
    sex = as.factor(c(rep("Girl", 10), rep("Boy", 10)))
  )

  # Extract p_values from each test
  w_pValue <- wilcox.test(sample_data$score ~ sample_data$sex)$p.value
  i_pValue <- pvalue(independence_test(sample_data$score ~ sample_data$sex
    , distribution = exact()))

  # Append the p_values to the above lists
  w_values <- c(w_values, w_pValue)
  i_values <- c(i_values, i_pValue)
}
```

We will perform some data cleaning.

```
# Remove NA value which we added to create empty list
w_values <- w_values[!is.na(w_values)]
i_values <- i_values[!is.na(i_values)]

# Create DF of two lists
p_values <- data.frame(
  wilcox = w_values,
  independence = i_values
)
```

Below is the summary of the data obtained from the experiment.

```
##      wilcox      independence
## Min.   :0.0007253 Min.   :0.001472
## 1st Qu.:0.2798610 1st Qu.:0.270205
## Median :0.5288489 Median :0.499080
## Mean   :0.5258009 Mean   :0.503697
## 3rd Qu.:0.7959363 3rd Qu.:0.751264
## Max.   :1.0000000 Max.   :0.998853
```

To look further into it and get a general idea, let us look at a boxplot of the two tests' p-values.



Discussions and Conclusion

As shown in the box plot and the summary of the data collected when repeated over 1000 times, the mean p-values are different.

The differences come from the fact that **Wilcoxon rank-sum test is not a test for differences in the mean while the permutation test is a permutation test for differences in the mean**, and we recall from the Methods section, each sample from the 1000 repetitions have difference mean values.

Acknowledgements

Data Source: HSAUR3 | *Link: Documentation*

References

Question from the A Handbook of Statistical Analyses Using R | *Link*