Chapter 9: Hypothesis Testing

Exercises - Alternative with SIMD

Brian Fogarty

25 August 2018

Contents

EXERCISE I **1**

ANSWERS FOR EXERCISE I **1**

EXERCISE II **1**

ANSWERS FOR EXERCISE II **1**

EXERCISE III **2**

ANSWERS FOR EXERCISE III **2**
Question 3.1 2
Question 3.2 3
Question 3.3 3
Question 3.4 3

EXERCISE IV **4**

ANSWERS FOR EXERCISE IV **4**
Question 4.1 4
Question 4.2 4
Question 4.3 5
Question 4.4 5

EXERCISE I

Write the null hypotheses for the following alternative hypotheses:

1. Individuals who are liberal are more likely to vote for the liberal party/candidate than individuals who are conservative.
2. Cities with high poverty rates are expected to have high murder rates compared to cities with low poverty rates.
3. Countries with free university are expected to have more first time university students than countries without free university.
4. US states with high sales tax are expected to have lower economic growth than US states with low sales tax.
5. Individuals with a high education level are expected to have a good health status compared to individuals with a low education level.

ANSWERS FOR EXERCISE I

1. There is no relationship between individuals' political ideology (or liberal/conservative) and vote choice.
2. There is no relationship between cities' poverty rates and murder rates.
3. There is no relationship between countries' tuition fees and the number of first time university students.
4. There is no relationship between US states' sales tax rate and economic growth.
5. There is no relationship between individuals' education level and health status.

EXERCISE II

Write a statement signifying a statistically significant relationship for each of the previous alternative hypotheses.

ANSWERS FOR EXERCISE II

1. There is a statistically significant relationship between individuals' political ideology (or liberal/conservative) and vote choice. Or, political ideology has a statistically significant effect on individuals' vote choice.
2. There is a statistically significant relationship between cities' poverty rates and murder rates. Or, poverty rates have a statistically significant effect on cities' murder rates.
3. There is a statistically significant relationship between countries' tuition fees and the number of first time university students. Or, whether a country has tuition fees or not has a statistically significant effect on the number of first time university students.
4. There is a statistically significant relationship between US states' sales tax rate and economic growth. Or, sales tax rates have a statistically significant effect on states' economic growth.
5. There is a statistically significant relationship between individuals' education level and health status. Or, education level has a statistically significant effect on individuals' health status.

EXERCISE III

Using `pct_employment_deprived` as the outcome variable and `urban` as the grouping variable from the SIMD data (`simd.csv`), carry out the following **independent samples** tests.

1. Perform a non-directional t-test. Are the groups significantly different?
2. Based on your findings from the previous question, perform an appropriate directional t-test. Is there a significant difference?
3. Perform a non-directional Wilcoxon rank-sum test. How are the significance results different from the t-test using this non-parametric test?
4. Based on your findings from the previous question, perform an appropriate directional Wilcoxon rank-sum test. How are the significance results different from the directional t-test using this non-parametric test?

ANSWERS FOR EXERCISE III

Question 3.1

```
setwd("C:/QSSD/Exercises/Chapter 9 - Exercises/")
getwd()
```

```
[1] "C:/QSSD/Exercises/Chapter 9 - Exercises"
```

```
simd <- read.csv("simd.csv")

library(car)
```

```
Loading required package: carData
```

```
simd$urban <- recode(simd$urban, "0='Rural';1='Urban'",as.factor=TRUE)

t.test(pct_employment_deprived~urban, data=simd)
```

```
    Welch Two Sample t-test

data:  pct_employment_deprived by urban
t = -21.831, df = 6176.8, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.147837 -3.464297
sample estimates:
mean in group Rural mean in group Urban
           8.288453           12.094519
```

Since $p \leq .05$, we conclude that there is a statistically significant difference in the mean employment deprivation percentage between urban and rural Scottish datazones.

Question 3.2

Since rural datazones have a smaller mean than urban datazones, we need to use the option `alternative="less"` in the t-test.

```
t.test(pct_employment_deprived~urban, alternative="less", data=simd)
```

```
    Welch Two Sample t-test

data:  pct_employment_deprived by urban
t = -21.831, df = 6176.8, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -3.519257
sample estimates:
mean in group Rural mean in group Urban
           8.288453           12.094519
```

Since $p \leq .05$, we conclude that rural datazones' mean employment deprivation percentage is statistically significantly smaller than for urban datazones.

Question 3.3

```
wilcox.test(pct_employment_deprived~urban, data=simd)
```

```
    Wilcoxon rank sum test with continuity correction

data:  pct_employment_deprived by urban
W = 3992700, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

We see that `p < 2.2e-16`, which tells us that there is a statistically significant difference between employment deprivation percentages of rural and urban datazones. Therefore, our significance results are the same as they were with the t-test.

Question 3.4

Since we know that rural datazones have a smaller mean than urban datazones, we need to use the option `alternative="less"` in the Wilcoxon rank-sum Test.

```
wilcox.test(pct_employment_deprived~urban, alternative="less", data=simd)
```

```
    Wilcoxon rank sum test with continuity correction

data:  pct_employment_deprived by urban
W = 3992700, p-value < 2.2e-16
alternative hypothesis: true location shift is less than 0
```

We see that `p < 2.2e-16` and thus the rural datazone distribution is significantly less than the urban datazone distribution. Therefore, our significance results are the same as they were with the t-test.

EXERCISE IV

Using `alcohol` and `drugs` from the SIMD data (`simd.csv`), carry out the following **dependent samples** tests.

1. Perform a non-directional t-test. Are the groups significantly different?
2. Based on your findings from the previous question, perform an appropriate directional t-test. Is there a significant difference?
3. Perform a non-directional Wilcoxon Signed Test. How are the significance results different from the t-test using this non-parametric test?
4. Based on your findings from the previous question, perform an appropriate directional Wilcoxon Signed Test. How are the significance results different from the directional t-test using this non-parametric test?

ANSWERS FOR EXERCISE IV

Question 4.1

```
t.test(simd$alcohol,simd$drugs, paired=TRUE)
```

```
    Paired t-test

data:  simd$alcohol and simd$drugs
t = 3.2978, df = 6973, p-value = 0.0009792
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.032719 7.991084
sample estimates:
mean of the differences
               5.011901
```

Since $p \leq .05$, we conclude that there is a statistically significant difference between the means of the standardised ratios of hospital stays related to alcohol misuse and drug misuse.

Question 4.2

Since the differences of means value is positive, it implies that the mean of the alcohol standardised ratio is greater than the mean of the drug standardised ratio.

```
t.test(simd$alcohol,simd$drugs, alternative="greater", paired=TRUE)
```

```
    Paired t-test

data:  simd$alcohol and simd$drugs
t = 3.2978, df = 6973, p-value = 0.0004896
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 2.511794      Inf
sample estimates:
mean of the differences
               5.011901
```

Since $p \leq .05$, we conclude that the mean of the alcohol standardised ratio is significantly greater than the mean of the drug standardised ratio.

Question 4.3

```
wilcox.test(simd$alcohol,simd$drugs, paired=TRUE)
```

```
    Wilcoxon signed rank test with continuity correction

data:  simd$alcohol and simd$drugs
V = 14747000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Based on the p-value, we conclude that the distributions for the alcohol standardised ratio and the drug standardised ratio across all Scottish datazones are statistically significantly different. Therefore, our significance results are the same as they were with the t-test.

Question 4.4

We need to use the option `alternative="greater"`.

```
wilcox.test(simd$alcohol,simd$drugs, alternative="greater", paired=TRUE)
```

```
    Wilcoxon signed rank test with continuity correction

data:  simd$alcohol and simd$drugs
V = 14747000, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0
```

We see that `p < 2.2e-16` and thus the alcohol distribution is significantly greater than the drug distribution. Therefore, our significance results are the same as they were with the t-test.