PO91Q: Fundamentals in Quantitative Research Methods

Week 5 - Solutions

Dr Florian Reiche

F.Reiche@warwick.ac.uk

1 | Data Exploration

Before starting, we need to load libraries and install packages if not already installed. In these exercises we will be using the `tidyverse` package.

1. Set your working directory, place the data set in it, and load it into R.
2. Create a new RScript for this case study and annotate it as you go through the exercises presented here.
3. Load the `tidyverse` package.

1.1 Descriptive Statistics

1. Produce descriptive statistics for all three numerical variables.

```
summary(simd$alc16)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   48.15   66.83   99.03   98.71  116.07  205.29
```

```
summary(simd$mortality16)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   79.04   89.49   94.86   96.72  103.17  126.68
```

```
summary(simd$mortality20)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   79.09   89.82   94.26   95.58  100.79  113.59       1
```

1.2 Visualisation

Let's visualise the distribution of the variable `alc16`.

```
ggplot(simd, aes(x=alc16)) +
  geom_density(aes(y=..density..)) +
  theme_classic() +
  scale_x_continuous(name="Alcohol-Related Hospital Admissions (2011-2014)")+
  ylab('Density') +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=13))
```

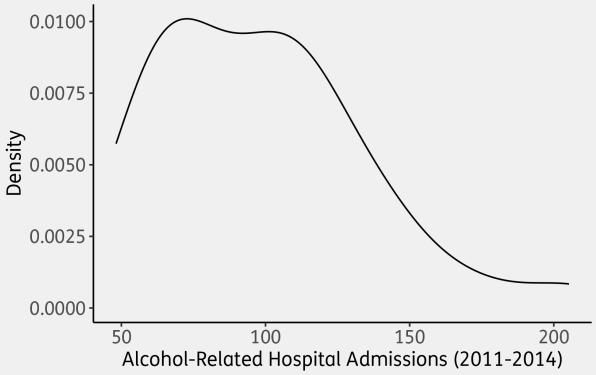

Figure 1: Distribution of Alcohol-Related Hospital Admissions (2011-2014)

1. Reproduce Figure 1.
2. What does the distribution tell us about alcohol-related admissions to hospital?
   - Not perfectly normally distributed, with a positive skew.
3. How does the shape of the distribution in Figure 1 relate to the descriptive statistics calculated in Section 1.1?
   - In a positvely skewed distribution the median is larger than the mean which is the case here. The maximum is also well above the third quartile.
4. What would happen to the shape of the distribution if the median was smaller than the mean?
   - If they were identical, then this would be a normal distribution. If the median was smaller than the mean then we would be dealing with a negatively skewed distribution.

1.3 Hypothesis

We are interested in how alcohol-related admissions to hospital have affected mortality rates in Scotland. The following scatter plot uses the variables `alc16` and `mortality20`.

```
ggplot(simd, aes(alc16, mortality20)) +
  geom_point() +
  xlab('Alcohol-Related Hospital Admissions (2011-2014)') +
  ylab('Standardised Mortality Ratio (2014-2018)') +
  theme_classic() +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=13))
```

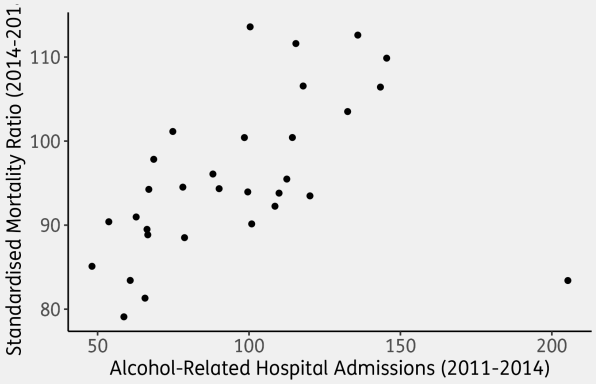

Figure 2: Alcohol-Related Hospital Admissions and Mortality

1. Reproduce Figure 2.
2. Based on this scatter plot, formulate the alternative and the null hypotheses:

H_0: The higher the rate of alcohol-related admissions to hospital between 2011 and 2014, the higher the standardised mortality ratio in 2014-2018

H_A: The rate of alcohol-related admissions to hospital between 2011 and 2014 and the standardised mortality ratio in 2014-2018 are unrelated.

1.4 Sampling

The data frame `simd` which we have been using so far represents the population. Let us now draw a random sample of 15 councils as follows:

```
set.seed(6)
sample <- sample_n(simd, 15)
```

1. Explain the purpose of the `set.seed` function.
   - It creates a pseudo-random number.

1.5 Inferential Statistics

Let us now see if the mortality ratio has changed between the two waves of 2016 and 2020. This is the worked example from the lecture, but I am repeating it here deliberately, so that you can carry out the example yourself.

1. As a first step, create a new variable measuring the difference between `mortality20` and `mortality16`. Make sure that increases are positive and decreases negative.

```
sample$diff <- with(sample, mortality20-mortality16)
```

2. What is the sample mean of the differences in mortality rates, variable `diff`?

```
summary(sample$diff)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.9276 -0.1155  1.5762  0.9063  2.2114  2.8747
```

3. The sample size of 15 is small. Will it be appropriate to conduct a t-test? Why? Why not?
   - Yes, as the population distribution is likely to be normal.
4. Find out whether the difference in mortality rates is significantly different from zero.

```
t.test(sample$diff, mu=0,
       data=sample)
##
##  One Sample t-test
##
## data:  sample$diff
## t = 1.9689, df = 14, p-value = 0.06909
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.08096415  1.89353389
## sample estimates:
## mean of x
## 0.9062849
```

   - There is no statistically significant difference in mortality ratios between the two waves.

5. Draw a graph which depicts the direction of the alternative hypothesis and the p-value. Try not to look at the lecture slides.

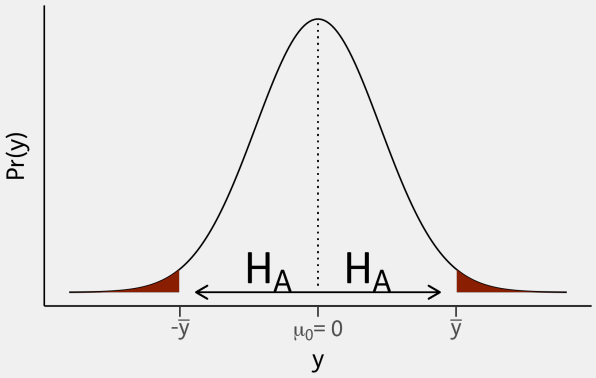

Figure 3: Two-Sided Significance Test

6. Suppose the Scottish Government claims that mortality rates have decreased. Test this claim.

```
t.test(sample$diff, mu=0,
       data=sample,
       alternative = "less")
##
##  One Sample t-test
##
## data:  sample$diff
## t = 1.9689, df = 14, p-value = 0.9655
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##      -Inf 1.717019
## sample estimates:
## mean of x
## 0.9062849
```

   - Absolutely not!

7. Again, draw a graph which depicts the direction of the alternative hypothesis and the p-value. Try not to look at the lecture slides.

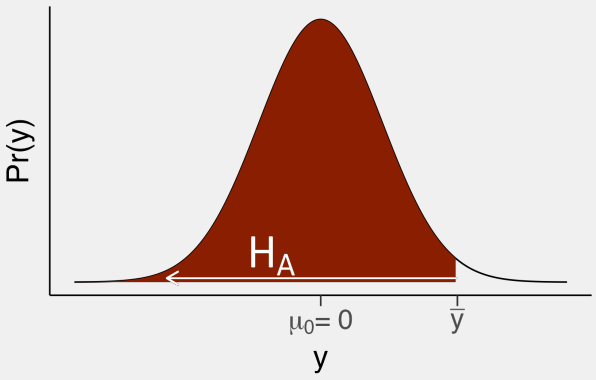

Figure 4: Left-Sided Significance Test

8. Drawing on the results from Exercises 5 and 7, reason about the p-value you would obtain if you tested the hypothesis that mortality rates have increased between the waves of 2016 and 2020.

- Using Figure 4, the p-value for a right-sided test is indicated by the remaining white area. This area must be half of the blue area in Figure 3. $\frac{0.06909}{2} = 0.03454$. You can confirm this with:

```
t.test(sample$diff, mu=0,
       data=sample,
       alternative = "greater")
##
##  One Sample t-test
##
## data:  sample$diff
## t = 1.9689, df = 14, p-value = 0.03454
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  0.09555077        Inf
## sample estimates:
## mean of x
## 0.9062849
```

- This is significant. Mortality rates have indeed increased.

1.6 Causality

1. Identify the elements of symmetry and asymmetry in the setup of this case study.
   - Symmetry: Alcohol abuse leads to health issues and possibly death. It's not really a theory, but medical reasoning.
   - Asymmetry: I have taken the mortality from a later wave than the alcohol-related admissions to hospital. So, rverse causality is not possible, but bear in mind that a time-lag might be insufficient to justify asymmetry.
2. Consider again Figure 5 from the lecture. Which aspects of establishing causality has the case study addressed? What is missing?
   - Practically everything is still missing, bar the theory and historical context, and perhaps asymmetry. We have not touched anything else in this graph, yet.

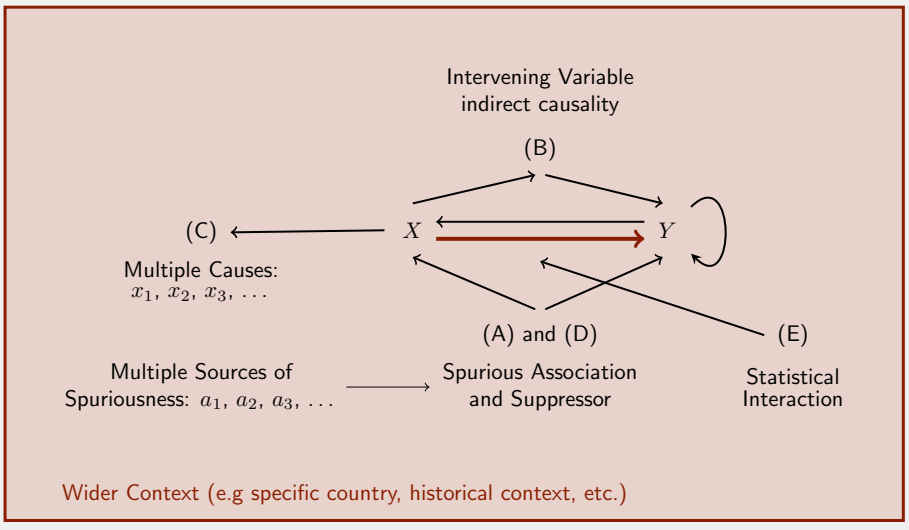

Figure 5: Causality Framework