STA238 Tutorial 5

Luis Ledesma

2023-03-01

1 Announcements

- You can upload your work on Crowdmark from the end of the tutorial session to 5pm Friday of that week.
- All questions must be solved using RStudio.

2 Recall: Last tutorial

Last tutorial: based on a random sample, we conducted hypothesis tests, and we looked at the difference between a one-tailed and a two-tailed test. In addition, we looked at rejection regions and how these relate to the p-value.

Main takeaways:

1. For hypothesis testing, one must state what is the distribution of the test statistic assuming that the null hypothesis is true.
2. The rejection region for a one-tailed and a two-tailed test will be different (but, the significance level and overall area of the rejection region will be the same).
3. The hypothesis test comparing the difference between two means of two samples relies on the fact that the samples are independent.

3 Tutorial activity

We want to:

1. Conduct a paired t -test hypothesis test comparing the means before and after a specific event (for instance) for each subject.
2. Conduct a paired t-test with 0.10 significance.
3. Understand and check the assumptions for a paired t-test.

3.1 Conducting a paired t-test

If we want to conduct a paired t-test to determine whether the mean for population 2 μ_2 is larger than the mean for population 1 μ_1, then we would set up our hypothesis test as:

$$H_0 : \mu_1 - \mu_2 = \mu_d = 0 \quad H_a = \mu_1 - \mu_2 = \mu_d < 0$$

The sign of the inequality is important (we want to test whether μ_2 is larger than μ_1, if μ_d is lower than 0, then this holds).

3.1.1 Carrying out the hypothesis test and computing the test statistic

If we were to compute the differences between both populations, one would get the following vector of differences:

$$-5, -2, -5, -1, -6, 0, -7, -4, -3, -4$$

The test statistic under the null hypothesis is given by:

$$t = \frac{\bar{x}_d}{\frac{s_d}{\sqrt{n_d}}} \sim t_{10-1}$$

A t-distribution with 9 degrees of freedom. One computes:

$$\bar{x}_d = \frac{1}{n_d}\sum d_i = \frac{-37}{10} = -3.7$$

$$s_d^2 = \frac{1}{n_d - 1}\left(\sum d_i^2 - \frac{(\sum d_i)^2}{n_d}\right) = \frac{1}{9}\left(181 - \frac{(-37)^2}{10}\right) = 4.9$$

And $s_d = 2.2134$. Combining the above, the computed test statistic is:

$$t_c = \frac{\bar{x}_d}{\frac{s_d}{\sqrt{n_d}}} = \frac{-3.7}{\frac{2.2134}{\sqrt{10}}} = -5.29$$

Recall that we are doing a one-tailed paired t-test. The rejection region will then be:

```
qt(0.1,df=9,lower.tail=TRUE)
```

```
## [1] -1.383029
```

Thus, $RR = \{t_c < -1.382\}$. Then, our computed t-statistic is in the RR, so we reject H_0 at 0.1 significance.

3.1.2 Computing the confidence interval

Recall that we want the 90% confidence interval for μ_d:

$$CI = \bar{x}_d \pm t_{0.05,9}\frac{s_d}{\sqrt{n_d}} = -3.7 \pm 1.283 = (-4.983, -2.417)$$

Remark: We use the 0.05-th quantile for the t-distribution as our test statistic under the null hypothesis follows a t-distribution.

Then, the interpretation will be that we are 90% confident that μ_d lies in this interval. If we were to expand on this definition, if we were to repeat this experiment multiple times, 90% of the computed confidence intervals would capture the difference between the population means μ_d (or, the true value).

3.1.3 Assumptions for paired t-tests

This hypothesis test relies on the following assumptions:

1. **Sample was randomly collected.**
2. **Differences between the paired observations come from a normal distribution.**

3.2 Conducting another paired t-test with R-Studio

Now, we will do a similar analysis by using R-Studio:

```
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("readxl")

database <- read_excel("TUT5Q2Data.xlsx")
head(database)
```

```
## # A tibble: 6 x 3
##   Session Before After
##     <dbl>  <dbl> <dbl>
## 1       1     52    59
## 2       2     42    54
## 3       3     46    55
## 4       4     42    51
## 5       5     43    42
## 6       6     30    43
```

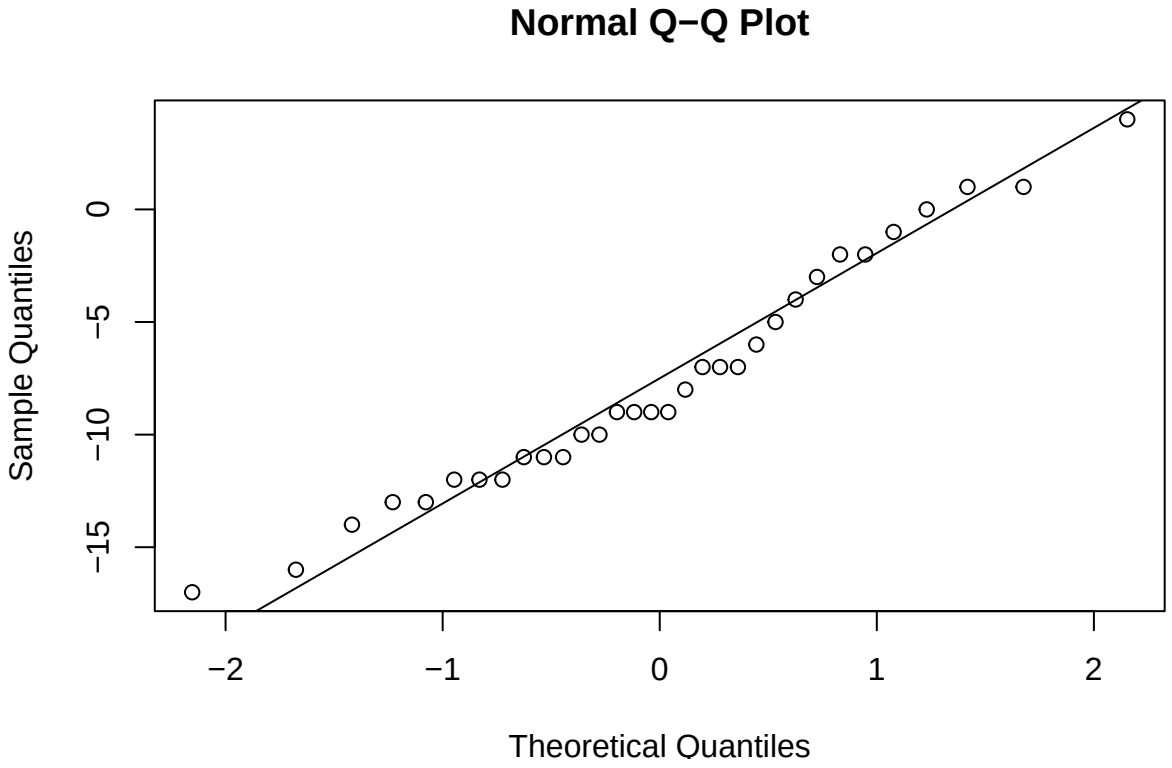
The data should be analyzed as paired differences since we are interested in whether the health status of a patient improves after handling a museum object, and each subject corresponds to a pair with before and after measurements.

The differences between the measurements will be:

```
database$Difference <- database$Before-database$After
diffmeasurements <- database$Difference
```

To check the assumptions for normality, we will look at a QQ-plot of the difference between the measurements:

```
qqnorm(diffmeasurements)
qqline(diffmeasurements)
```



The points lie near the straight line, so we may assume that these come from a normal distribution, satisfying the assumptions for the paired t-test.

To compute the mean and standard deviation of the differences:

```
mean(diffmeasurements)
```

```
## [1] -7.625
```

```
sd(diffmeasurements)
```

```
## [1] 5.271653
```

Now, the 90% confidence interval for the true mean difference will be:

```
mean(diffmeasurements)-qt(0.95,df=31)*(sd(diffmeasurements)/sqrt(32))
```

```
## [1] -9.205063
```

```
mean(diffmeasurements)+qt(0.95,df=31)*(sd(diffmeasurements)/sqrt(32))
```

```
## [1] -6.044937
```

Thus, the 90% confidence interval will be $(-9.2051, -6.0449)$. To interpret this confidence interval, we are 90% confident that the true value of the mean difference is in this interval. Since the lower and upper bounds of the confidence interval are negative, **it seems that the health status of the patients improved after handling the museum object, as there is evidence at** 0.1 **significance that the population mean after is greater than the one before.**