Report using two-sample t-test as a statistical test

Maria Kunevich

2022-12-01

Two-sample *t-test*

The two-sample is a statistical hypothesis testing technique in which two independent samples are compared to determine if the means of two populations are statistically different.

The assumption for the *t-test* is that the data follow the normal distribution. However, this assumption can be waived if the sample size is large enough due to the Central limit theorem. For the independent samples *t-test*, when each group is larger than 15, the data can be skewed and the test results will still be valid.

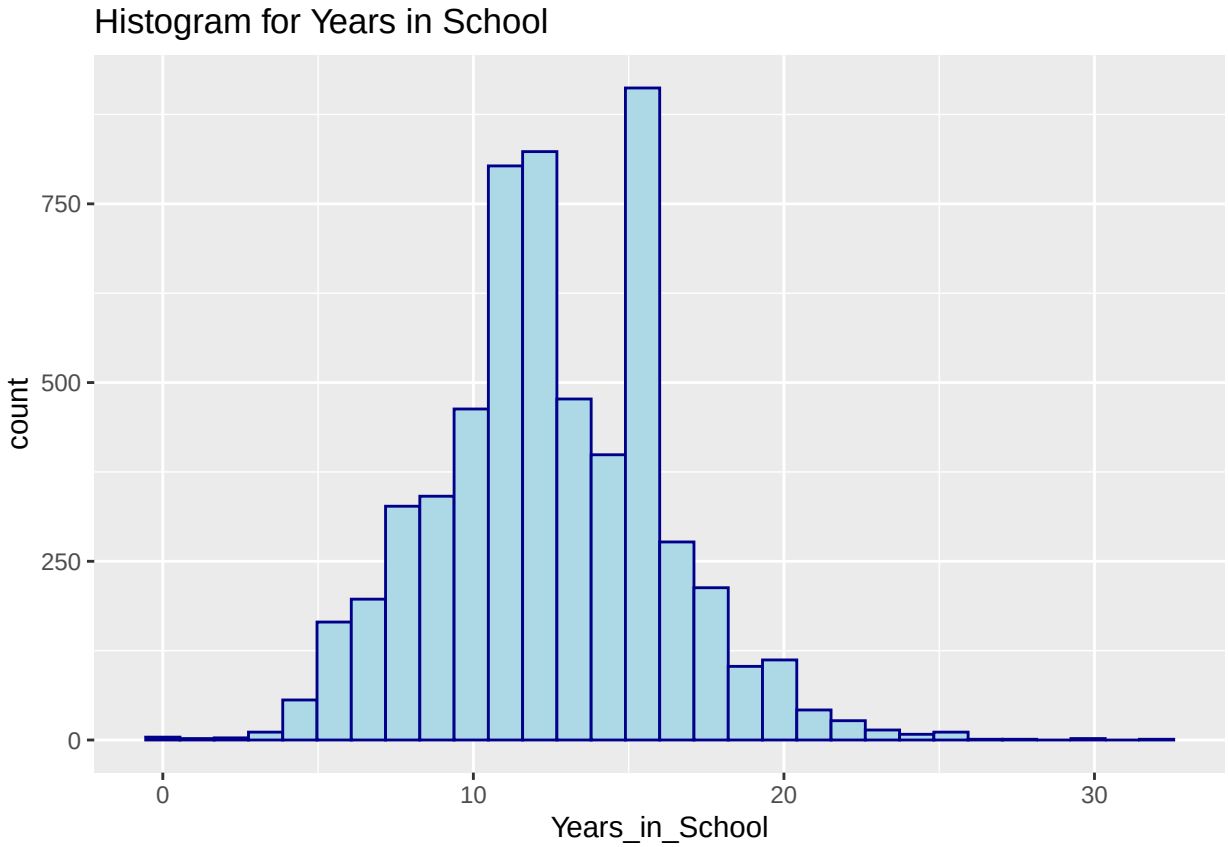
In `Life in Estonia` dataset I will look at an independent binary variable: *Gender* in comparison to the same continuous, dependent variable: *Years in school*.

Research Question:

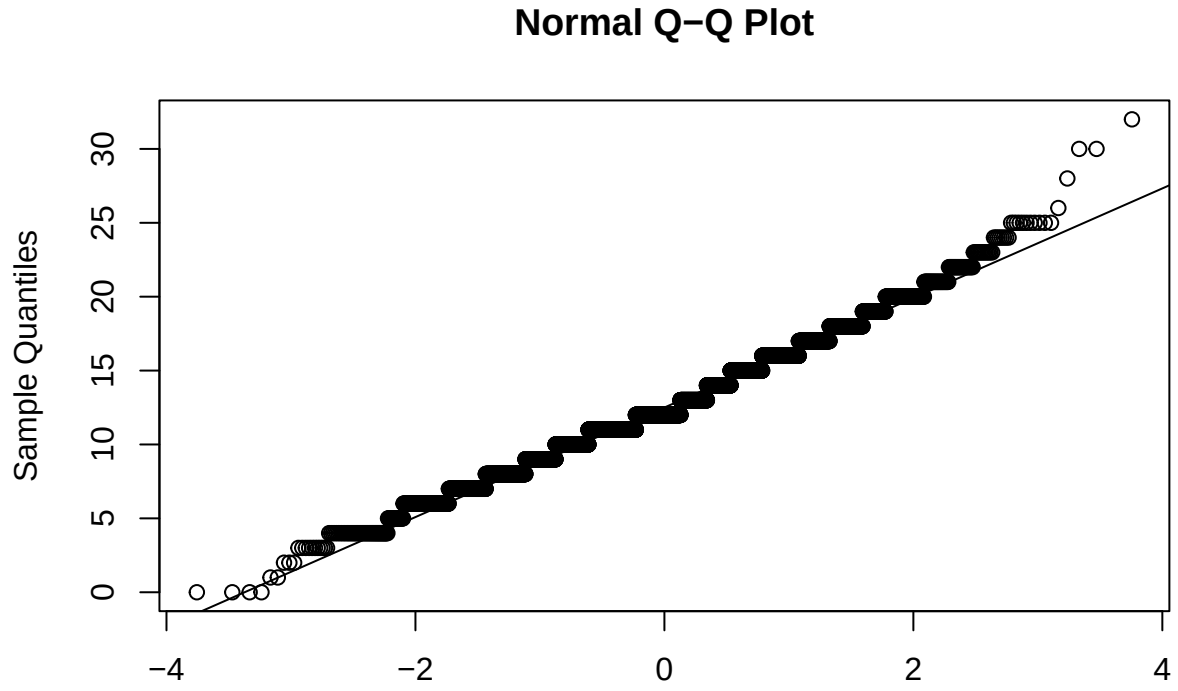
Is there any difference between male and female Estonian residents in terms of time spent studying?

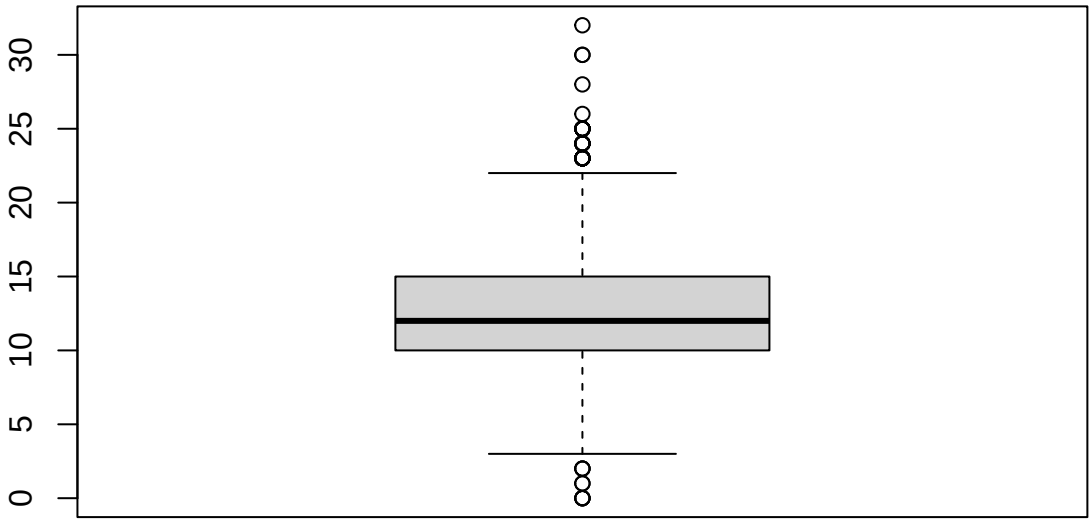
Let's first explore both variables and derive some descriptive statistics from the data:

First, I created a separate dataset with two variables: `gender` and `years in school` and performed some data wrangling. For the variable `years in school` there are 41 missing values, the median is 12 years at school and the standard deviation is 3.68. The histogram for `years in school` displays data that are approximately normally distributed, but there seems to be some abnormal number of counts for the data point `15` years in school.



At the same time, both qq-plot and boxplot for `years in school` variable reveal that there are some extreme data points at both ends, i.e. there are people who didn't spent any time at school or spent only a couple of years (four people with `0` years at school, two people with `1` years at school, three people with `2` years at school) and there are people who spent more than 30 years in school (two people with `30` years at school and one person with `32` years at school).

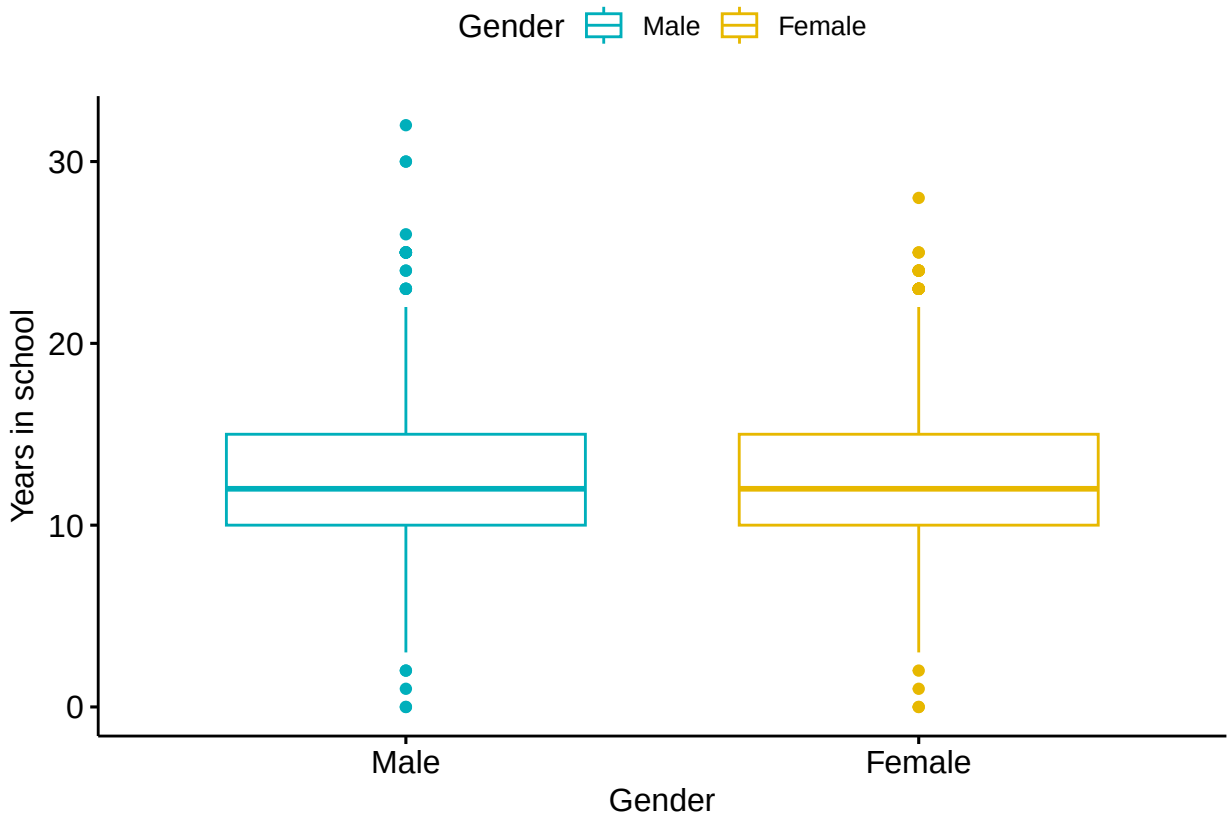




The table below presents the means and standard deviations for `Years in School` for both groups: `male` and `female`:

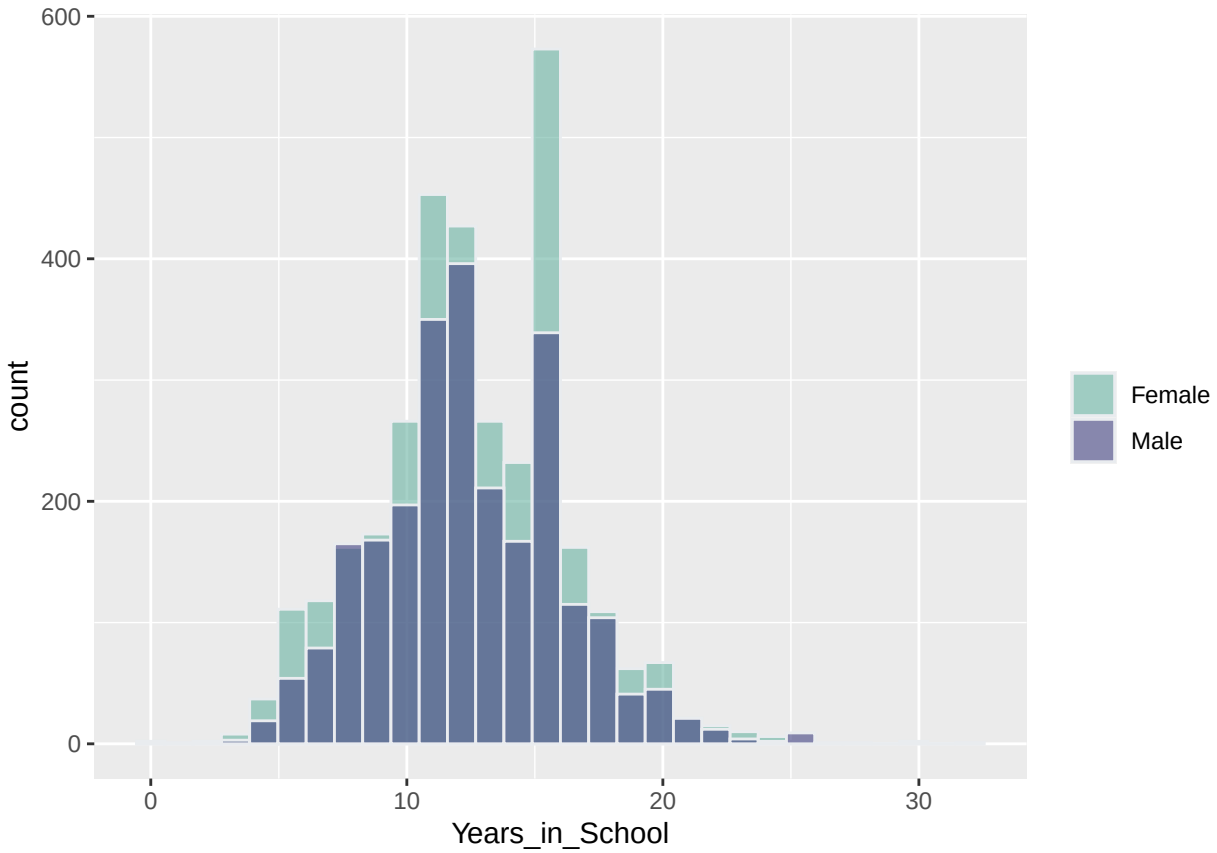
```
## # A tibble: 2 x 4
##   Gender     N  Mean    SD
##   <chr>  <int> <dbl> <dbl>
## 1 Female  3308  12.6  3.69
## 2 Male    2528  12.5  3.67
```

The mean number of years at school for men is 0.06 years lower than the mean for women. The standard deviations are similar so the groups are equally spread out. The box plot for *Years in school* by *Gender* reveals similar information, but also highlights the number of outliers for both groups. The outliers for both groups are presented in the boxplot:



To run the independent t-test I should first check the assumptions:

- Assumption 1: Are the two samples independents? Yes, since the samples from men and women are not related.
- Assumption 2: Are the data from each of the two groups follow a normal distribution? To check this assumption, I will produce histograms of the dependent variable by the independent. Both histograms are approximately normally distributed so the assumption has been met.



- Assumption 3. Do the two populations have the same variances? I will use F-test to test for homogeneity in variances. This can be performed with the function `var.test()` as follow:

```
##
##  F test to compare two variances
##
## data:  Years_in_School by Gender
## F = 1.0101, num df = 3284, denom df = 2509, p-value = 0.7907
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9383072 1.0869153
## sample estimates:
## ratio of variances
##           1.010066
```

The p-value of F-test is p = 0.7907. It's greater than the significance level alpha = 0.05. In conclusion, there is no significant difference between the variances of the two sets of data. Since the standard deviations are similar, so the assumption of equal variances has been met. Therefore, we can use the classic t-test which assumes equality of the two variances.

The null hypothesis: the means for the two populations are equal.

Alternative hypothesis: the means for the two populations are not equal.

```
##
##  Two Sample t-test
##
## data:  Years_in_School by Gender
## t = 0.62589, df = 5793, p-value = 0.5314
```

```
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -0.1302690  0.2524651
## sample estimates:
## mean in group Female   mean in group Male
##             12.56986             12.50876
```

In the result above:
- t is the t-test statistic value (t = -0.62589),
- df is the degrees of freedom (df= 5793),
- *p-value* is the significance level of the t-test (p-value = 0.5314),
- *conf.int* is the confidence interval of the mean at 95% (conf.int = [-0.2524651, 0.1302690]),
- *sample estimates* is the mean value of the sample (mean = 12.50876, 12.56986).

Conclusion:

The p-value of the test is 0.53, which is more than the significance level alpha = 0.05. We can conclude that men's average years in school is not significantly different from women's average years in school.