Understanding p-values and the assumptions behind statistical tests

Karim Saied (karim.saied@unil.ch)

22/12/2017

1. Introduction

The current study aims to assess the importance of the assumptions behind statistical tests through their violations under different conditions.

2. Two-sample *t*-test

The two-sample *t*-test allows to assess whether the means of two samples are different. Here different assumptions are tested, namely the normal distribution of data, the sample size and the equality of the variances.

2.1. Normal distribution in both samples

```
set.seed(7); n = 10000; m = 5000; p_values = rep(NA, times=n)

for (i in 1:n){
  p_values[i] = t.test(rnorm(m, mean=0, sd=1), rnorm(m, mean=0, sd=1), var.equal=TRUE)$p.value
}
sum(p_values < 0.05) /length(p_values)
```

```
## [1] 0.0496
```

Let's first consider the ideal siuation where a large sample size (m = 5000), a normal distribution in both samples and an equal variance between them are present. The proportion of p-values < 0.05 is equal to 0.0496, which is close to the expected value, namely 0.05.

2.2. Normal distribution in both samples and low sample size

```
set.seed(7); n = 10000; m = 5; p_values = rep(NA, times=n)

for (i in 1:n){
  p_values[i] = t.test(rnorm(m, mean=0, sd=1),
                       rnorm(m, mean=0, sd=1), var.equal=TRUE)$p.value
}
sum(p_values < 0.05) /length(p_values)
```

```
## [1] 0.0508
```

When the ideal situation described above is met, a low sample size seems not to have an impact.

2.3. Exponential distribution in both samples

```
set.seed(7); n = 10000; m = 5; p_values = rep(NA, times=n)

for (i in 1:n){
  p_values[i] = t.test(rexp(m, rate = 10), rexp(m, rate = 3), var.equal=TRUE)$p.value
}
sum(p_values < 0.05) /length(p_values)
```

```
## [1] 0.2163
```

Using the same sample size as before (low sample size), if both samples display an exponential distribution but with two different rate, one can observe that the proportion of false positives is 21.6%. The simulation was also performed using the same distribution and no rate difference between both samples. The proportion of p-values < 0.05 was 4.3%, which was bellow the expected value of 0.05.

2.4. Poisson distribution in both samples

```
set.seed(7); n = 10000; m = 5000; p_values = rep(NA, times=n)

for (i in 1:n){
  p_values[i] = t.test(rpois(m, lambda=3), rpois(m, lambda=3), var.equal=TRUE)$p.value
}
sum(p_values < 0.05) /length(p_values)
```

```
## [1] 0.0495
```

A Poisson distribution in both samples displaying a large sample size seems not to have an impact on the proportion of false positives.

2.5. Unequal variances between the samples

```
set.seed(7); n = 10000; m = 5000; p_values = rep(NA, times=n)

for (i in 1:n){
  p_values[i] = t.test(rnorm(m, mean=0, sd=50), rnorm(m, mean=0, sd=3), var.equal=TRUE)$p.value
}
sum(p_values < 0.05) /length(p_values)
```

```
## [1] 0.0494
```

In that simulation, using the same large sample size (m = 5000), a normal distribution and an unequal variance between both samples, one can still observe a proportion of p-values close to the expected value (0.05).

2.6. Unequal variances and small sample size

```
set.seed(7); n = 10000; m = 10; p_values = rep(NA, times=n)

for (i in 1:n){
  p_values[i] = t.test(rnorm(m, mean=0, sd=50), rnorm(m, mean=0, sd=3), var.equal=TRUE)$p.value
}
sum(p_values < 0.05) /length(p_values)
```

```
## [1] 0.0615
```

However, using two samples with different variance and a small sample size (m = 10), the proportion of p-values < 0.05 is higher (6.2%) than 0.05.

3. Pearson and Spearman correlations

The Pearson correlation allows to assess the relationship between two variables. In other words, it measures the linear correlation between these variables. Two assumptions are tested, namely the normal data distribution and the absence of outliers.

The Spearman correlation allows to assess the relationship between two variables using the ranks of the variables. In other words, the correlation is calculated between the ranks of the variables instead of their values. It is a non-parametric test.

3.1. Non-normal distribution with Pearson correlation

```
set.seed(7); n = 10000; m = 5000; p_values = rep(NA, times=n)

for (i in 1:n){
  sample1 = rpois(m, lambda=3)
  sample2 = rexp(m, rate=3)
  p_values[i] = cor.test(sample1, sample2)$p.value
}

sum(p_values < 0.05) /length(p_values)
```

```
## [1] 0.0473
```

With two samples displaying a non-normal distribution each, one can observe that the proportion of p-values < 0.05 is 4.7%, which is quite similar to 0.05.

3.2. Presence of outliers with Pearson correlation

```
set.seed(7); n = 10000; m = 5000; p_values = rep(NA, times=n)

for (i in 1:n){
  sample1 = c(rnorm(m, mean=0, sd=1), -7000, 7000)
  sample2 = c(rnorm(m, mean=0, sd=1), -7000, 7000)
  p_values[i] = cor.test(sample1, sample2)$p.value
}

sum(p_values < 0.05) /length(p_values)
```

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

One can observe that the presence of outliers strongly affects the correlation. As a matter of fact, all p-values (100%) are greater than 0.05. It is clear that the Pearson correlation is not able to deal with outliers.

3.3. Presence of outliers with Spearman correlation

```
#spearman
set.seed(7); n = 10000; m = 5000; p_values = rep(NA, times=n)

for (i in 1:n){
  sample1 = c(rnorm(m, mean=0, sd=1), -7000, 7000)
  sample2 = c(rnorm(m, mean=0, sd=1), -7000, 7000)
  p_values[i] = cor.test(sample1, sample2, method="spearman")$p.value
}

sum(p_values < 0.05) /length(p_values)
```

```
## [1] 0.0533
```

Using the Spearman correlation, a non-parametric test, the same conditions as before have been applied and it appears that the proportion of p-values < 0.05 correspond to the expected value (0.05).

5. Conclusion

Regarding the two-sample t-test, if the ideal situation (normal distribution and equal variances) is not met (or as one moves further from it), then the sample size must be large enough to give a reasonable rate of false positives. Otherwise, in general, the test is rather robust toward the violation of its assumptions. A large sample size is not absolutely necessary as long as the sample are normally distributed and their variances are equal. One has to keep in mind that in practice, the ideal situation previously mentioned is never met perfectly.

About the correlation tests, the Pearson correlation is extremely sensitive to outliers. This is why the Spearman correlation, which is a non-parametric test, is more suitable with data containing outliers. According to the results, a non-normal data distribution seems not to be necessary.

Taken together, the results of the current study highlight the fact that the statistical tests assessed, are rather robust as long as several assumption are not violated at the same time.