Robin Boudry, Maarten Rahier, Tom Schipper, Laurens van Paemel. Groep x  
  
**Project Prinstat 2018-2019  
Part 2 : simulation study**

**Data cleaning:** No data were missing, but one observation of a diastolic blood pressure of 66 was considered unlikely and this patient was removed from the dataset. To ensure that all investigated patients were sufficiently exposed to the treatment, only participants with an adherence of at least 80% were selected for analysis.[[1]](#footnote-1)1 This led to a further removal of 29 patients from the dataset.

**1. Power comparison:** Two variables, the diastolic blood pressure measured at the last visit (denoted *dbp\_end*) and the change in blood pressure from baseline (denoted *dbp\_dif)*, were investigated for the first treatment arm and the placebo arm, which consisted of 39 and 44 patients. Although the QQ-plots of the variables revealed deviations from normality, the sample sizes were considered large enough for the Central Limit Theorem to apply and to assume that the sample means are normally distributed. The F-ratio test of the variances lead us to the conclusion that the Welch two-sample T-test was appropriate. The test was applied one-sided, with alternative hypothesis , where is the mean of treatment group and that of the placebo group.

The power of the T-test was estimated via 10 000 Monte Carlo simulations for each of the two variables. The treatment and placebo groups were simulated by random samples from a normal distribution with the mean and standard deviation of the respective group and were then compared via the Welch T-test. The power was determined as the percentage of p-values below 0.05 in a simulation of 10000 tests. The Delta( ) was checked 2fold. **A:** (figure1) as a proportion of the difference in original Delta. **B: (**figure 2) as an absolute difference.

|  |  |
| --- | --- |
| ***A.***  *Figure 1: power of the Welch T-test when applied to the variables* dbp\_dif *(blue) and* dbp\_end *(red) for multiple differences in means(as a percentage of the observed difference in the data) between the two treatment arms.* | ***B.***    *Figure 2: power of the Welch T-test when applied to the variables* dbp\_dif *(blue) and* dbp\_end *(red) for multiple differences in means.* |

Figure **A.** (Power when Delta is expressed as a proportion of the original difference) shows that the variables start at a similar power of 5%, corresponding to the significance level of 0.05, when there is no difference in means and converge to a power of 100% when the difference in means is 75% or more of the observed differences. However, when the means differ 20% - 75% of their observed differences, the test using the variable *dbp\_end* has more power than the test applied to *dbp\_dif*.

Figure **B.** (Power when Delta is expressed as a numerical difference) shows that dbp\_dif has more power in detecting a difference for a certain range of Delta values. We will explain this in the following.

**2. Analytical calculations of the observed Power findings:** To explain the discrepancy in power of the Welch T-test the following power function was used:

Where , and the non-centrality parameter is defined as:

For **A** this function was applied to a difference in means () decreased to 42.5% of the observed difference, as simulating with this led to the largest difference in power. The power is 70% for *dbp\_dif* and 74% for *dbb\_end*.

As , and are fixed, the difference in power can either be due to a different Δ or different values of and for the variables. As a higher increases the power, the variable with the largest and smallest variances should give the highest power. In this case, *dbp\_dif* has smaller values for and , but *dpb\_end* has a larger . The larger here has outweighed the smaller variances, leading to a higher power for *dbp\_end*.

Following the previous statement, we would expect to find a higher power for dbp\_dif when comparing absolute differences (**B**), because of the smaller variance. When using the same method as above mentioned, this is exactly what we find.

**3. Multiple testing**

Next a simulation study was executed to compare treatment 1 and 2. The same T-test as used in question one was applied simultaneously on *dbp\_dif* and *dbp\_end*. Each time the approach that resulted in the smallest p-value was picked. In both separate approaches a significant level of 0.05 was used. Earlier, in the separate cases, this resulted in a type I error of approximately 5% which can be seen as logical.

In the case where each time the approach with the lowest p-value was chosen the type I error increased to approximately 0.077. This is increase is possibly caused by multiple testing problem which occurs when one considers a set of [statistical inferences](https://en.wikipedia.org/wiki/Statistical_inference) simultaneously. By using more than one inference the probability on at least one incorrect rejection of the null hypothesis increases, this probability is called the family-wise error rate.

**4. Bonferroni Correction**

When coming back to the multiple testing problem the question can be asked if it is possible to solve the inflation of the type I error. One possible option is to control the family-wise error rate by using the Bonferroni correction. This method makes use of the Boole’s inequality:

\*m

Where is the significance level and m the number of tests performed simultaneously. When the family wise error must be controlled at , and when in our case two tests must be performed simultaneously, each of these test must be performed at the /2 significance level.

The simulation of the previous question was repeated but this time a significance level of 0.025 was used. Due do this correction the inflation of the type I error was avoided and a type 1 error of approximately 0.05 was found, which corresponds with the initial significance level of 0.05.

It is important to notice that by using the Bonferroni correction it is possible that our test loses power. This happens when the family wise error rate is lower than our specified significance level of 5%.

3. To test whether the difference between the treatment and placebo group is significant, a testing strategy is adopted which considers the treatment’s effect to be significant whenever either the test on the final measure or the test on the difference yields a p-value below 5%. To calculate the Type 1 error rate, H0 is simulated by equalizing the means of both the treatment and placebo group. The final dbp measurement (dbp\_end) and the difference between last measurement during the run-in period (dbp\_dif) are tested simultaneously for 10 000 Monte Carlo simulations.

The Welch T-test is Type 1 error robust, meaning that it controls the Type 1 error rate at 5 %. As a result, the Type 1 error rate for dbp\_end and dbp\_dif separately, converge to 5 % as the number of simulations increases. However, with the adopted testing strategy, the Type 1 error is whenever either measurement under H0 yields a p-value below 0.05, resulting in a Type 1 error rate of 7.2 %.

Due to multiple testing the Type 1 error rate is inflated. The inflated Type 1 error rate can be corrected by applying a Bonferroni correction based on Boole’s inequality.

1. 1 Valgimigli M, Garcia Garcia H, Vrijens B *et al*. Standardized classification and framework for reporting, interpreting, and analysing medication non-adherence in cardiovascular clinical trials: a consensus report from the Non-adherence Academic Research Consortium (NARC). *European Heart Journal* 2018; **00**: 1-16. [↑](#footnote-ref-1)