Testing for a difference in distributions - Blood data

Dominika Bakalarz

1 Introduction

This report presents an analysis of two datasets, "home" and "clotting", both concerned with blood testing. The goal of the first part of this report is to determine if the blood pressure measurements taken at home by the patient and the measurements taken in a hospital by a nurse have the same distribution. The goal of the second part of the report is to compare blood clotting times while taking different treatments in order to verify if there is a reduction in clotting time when taking the new treatment.

2 Part 1

There are three variables: Subject is the number of the patient, home is the blood pressure measurement taken by the patient at home, hospital variable is the blood pressure for the same patient measured by a nurse in a hospital. There are 11 individuals.

2.1 Exploratory Data Analysis

As we can see in Figure 1, out of 11 individuals 8 obtained lower blood pressure measurement when measured at home than in a hospital. The remaining three patients obtained a higher blood pressure result at home than in a hospital, but the difference between the two measurements for each patient is minimal, in particular patients 1 and 11 obtained almost exact results in the two different settings. A similar conclusion can be drawn from the boxplots presented in Figure 2. From the first two plots, we can see that the threshold and median of the hospital measurements are higher than for the home measurements. Moreover, the last plot in this figure shows that for more than 75% Subjects the hospital measurement was higher than the one taken at home. Thus, one might suspect that there is an offset in the distributions of home and hospital measurements. This hypothesis will be explored further and will be verified by relevant statistical tests in the next part of the report.

2.2 Testing for a difference in the distributions

Since we have paired data (there are two measurements for each Subject), we are going to use the Wilcoxon Signed Rank Test (SRT). This test is a non-parametric test used to compare two related samples and to assess whether their population mean ranks differ when the sample distributions are not specified.

In general, we would expect blood pressure measurements, taken in the same conditions, to be normally distributed, but the Signed Rank test does not need any assumptions about the underlying distribution.

Let Y be a random variable corresponding to the hospital measurements (that is, we have observations Y_1 =hospital[1], Y_2 =hospital[2], ...Y_{11}=hospital[11]), and Z be a random variable corresponding to the home measurements (with observations $Z_1, Z_2, ...Z_{11}$) Let $X = Y - Z$ and $X_i = Y_i - Z_i$ for $i = 1, ..., 11$.

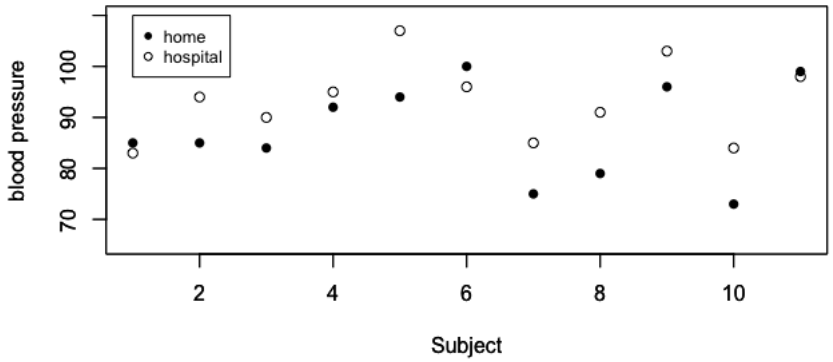

Figure 1: Blood pressure measurements taken 1) at home (full black circles), 2) in a hospital (empty circles), for each of the 11 individuals

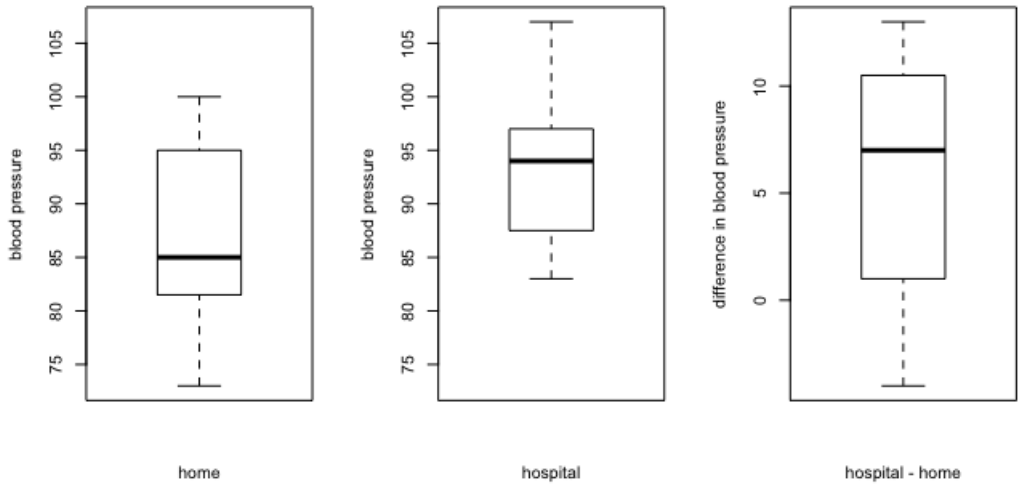

Figure 2: Boxplots for the blood measurements taken at home (left plot), at the hospital (plot in the middle), and the difference between hospital and home measurements (right plot). Note that the scale on the right plot is different.

- If Y and Z have the same distribution then X is a symmetric random variable with mean and median zero.
- If Y has the same distribution as $Z - \Delta$ then X is symmetric about its median $\mu = \Delta$.

We can use SRT to test H0: $\Delta = 0$ versus the alternative hypothesis H1: $\Delta \neq 0$ (or a relevant one-sided alternative).

The test statistic W is defined as:

$$W = \sum_{i=1}^{n} Z_i R_i$$

where n is the number of observations (thus in this part we have n = 11), R_i is the rank of the observation X_i in $X_1, X_2, ..., X_n$ and let $Z_i = \mathbb{1}_{X_i>0}$.

Then, under the null hypothesis, $Z_i \sim$ Bernoulli(1/2) and $R \sim$ Unif($\mathcal{P}_n$). The null for the test statistic W will be simulated by sampling $Z \sim U\{0,1\}^n$ for any fixed permutation of $R = (R_1, R_2, ..., R_n)$.

This method can be easily generalised to test H0: $\mu = \mu_0$ versus H1: $\mu \neq \mu_0$ by adjusting the definitions of Z_i and R_i, as presented in the lectures.

By inspecting sorted data X (differences in measurements) we note that there are no ties and 0 does not occur, thus we can use the built-in function wilcoxon.test() to perform the SRT on X, with H0: $\Delta = 0$ and the two-sided alternative H1: $\Delta \neq 0$. The output of this function is presented in Figure 3 below.

```
        Wilcoxon signed rank test

data:  hospital and home
V = 59, p-value = 0.01855
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
  1.5 10.5
sample estimates:
(pseudo)median
             6
```

Figure 3: SRT result

The obtained p-value is 0.01855, it is very significant (at 5% significance level). Thus there is enough evidence to reject the null hypothesis. Therefore, we conclude that the home and hospital measurements do not come from the same distribution. That is, there is a non-zero offset between the distribution. The estimated value of the offset is 6.

We have answered the questions given in Part 1, but we will perform an additional test to show that 6 is indeed a likely value of the offset. For $i = 1, 2, ..., 11$, let $Z'_i = Z_i + 6$, and let $X'_i = Y - Z'$. We will perform SRT on X' to check if Y and Z' come from the same distribution.

However, X' contains 0, and there are ties in the absolute values of the values of X', so we cannot use the built-in function Wilcoxon.test(). Thus, we will estimate the p-value using the Monte Carlo method. We start with removing 0 from the data and then calculate the observed test statistic W_{obs}=26. Then we simulate W under the null hypothesis 100000 times and check how many of this simulated W values are above 26. We find the estimated p-value to be 0.77. Thus, there is no evidence against the null hypothesis. Therefore, $\Delta = 6$ is a possible offset for this data.

2.3 Interpretation of the results

We have found that values of the measurements taken at the hospital are a little bit higher than the values of the measurements taken at home. A possible explanation is that the individuals could be a bit optimistic about their blood pressure what caused them to report a slightly lower value than what was displayed. Another possible explanation is that the individuals might have found visits at the hospital more stressful than being at home, so their blood pressure in the hospital was indeed slightly higher.

3 Part 2

There are three variables: Subject is the number of the patient, old is the blood clotting time measured for this individual when they were receiving the old treatment, new is the time to clotting for the same individual when they were receiving a new treatment. There are 15 individuals.

3.1 Exploratory Data Analysis

As we can see in Figure 4, 10 out of 15 individuals had shorter blood clotting time when taking the new drug. In particular, for Subjects 1, 4, 8,10, 13, 14 the time shortened significantly. For Subjects 2, 9, 11, 12, 15 the clotting time extended after taking the new drug, but the difference in clotting times for each Subject looks small, in particular, the difference in times for individuals 9, 11, 15 seems to be almost negligible. We can see from the boxplots in Figure 5 that the clotting times after taking the new drug are in general lower and, also, have lower variance. Although, the measurement corresponding to Subject 1 and the new drug might be a potential outlier. It equals 56, what seems to be a suspiciously low value when compared to 281 - the other measurement for Subject 1, and also when compared to the new drug measurements for other Subjects. This point is presented as an outlier in the boxplot in the middle of Figure 5. There is also another point marked as an outlier on this boxplot (with value 416), but since the other measurement for this subject is 391, it is less suspicious. Based on the analysis of the exploratory plots, we suspect that the new drug indeed reduces the blood clotting time, but we are going to verify this hypothesis with appropriate statistical tests.

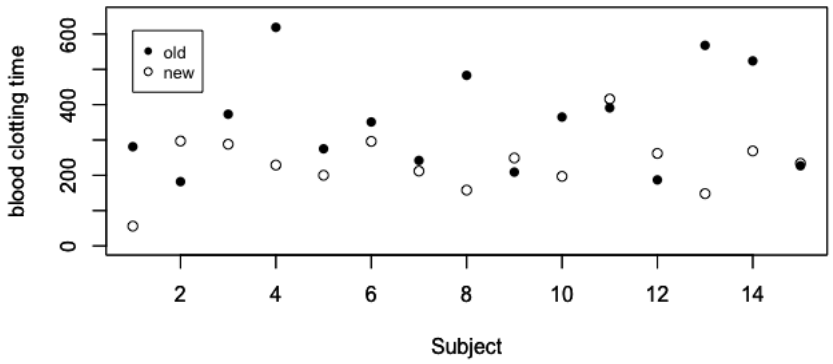

Figure 4: Blood clotting time while using the old drug (full black circles) and the new drug (empty circles). Measurements for Subject 15 are very close. Thus there is a significant overlap between the two circles.

3.2 Testing for a reduction in clotting time

In this question we are working with paired data, so we will again use the Wilcoxon Signed Rank Test (SRT). Again we do not need any assumptions about the underlying distribution for neither the old nor new measurements (although, in general, we would expect this kind of data to be normally distributed).

Let Y be a random variable corresponding to the old drug measurements (that is, we have observations Y_1 =old[1], Y_2 =old[2], ..., Y_{15}=old[11]), and Z be a random variable corresponding to the new

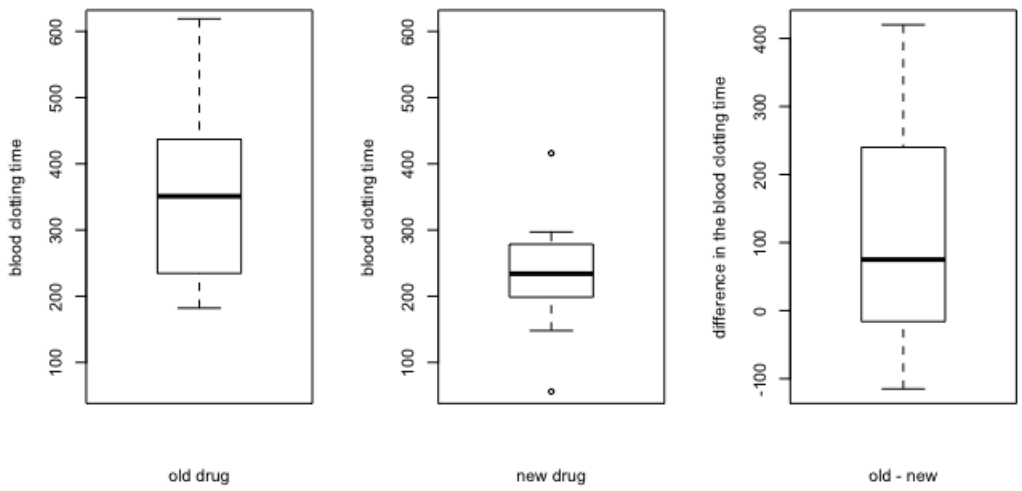

Figure 5: Boxplots for the blood measurements taken while using the old drug (left plot), while using the new drug (plot in the middle), and the difference between the blood clotting times (right plot). Note that the scale on the right plot is different.

drug measurements (with observations $Z_1, Z_2, ...Z_{15}$) Let $X = Y - Z$. We will apply SRT to X, with H0: $\Delta = 0$ versus the one-sided alternative hypothesis H1: $\Delta > 0$. We have decided to choose the one-sided alternative based on the expert knowledge from the question, saying that it is possible that the new drug decreases the time to clotting.

Since there is a tie in the absolute values of the observations X_i, we cannot use the built-in R function, and we will simulate p-value in the same way as in the second test in Part 1. We calculate the observed value of the test statistic W to be 97.5. Then we simulate 100000 times possible values of W under the null hypothesis, and we obtain the estimate of the p-value to be 0.01481. This value is very significant (At 5% significance level). Therefore we have enough evidence to reject the null hypothesis. Thus, we conclude that the new drug causes a reduction in clotting time.

3.3 Confidence Interval for the change

We will calculate the confidence interval for the reduction time using Walsh averages. There are 15 subjects, thus we have 15 * (15+1) /2 = 120 Walsh averages. Using psignrank() function we obtain $\alpha//2$ values for each average, and we choose 24th average as the left tail of the confidence interval because its corresponding $\alpha//2$ value (equal 0.024) is the closest to 0.025. This implies that 96th Walsh average is the right tail of the confidence interval (because 120 - 24 = 96). Thus, we obtain (7.5, 240) to be 95.2% confidence interval for the change in clotting time.

3.4 Interpretation of the results

There is a reduction in clotting time when taking the new drug, with 95.2% confidence level it is between 7.5 and 240.

4 Summary

In the first part, we have found that the measurements do not come from the same distribution, measurements taken at home are a little bit lower. Estimated value of the offset is 6. In the second part, we have found that there is a reduction in the blood clotting time when taking the new drug, and (7.5, 240) is a 95.2% confidence interval for the change. However, in both cases we working with very small datasets. Thus, collecting more data would be a good idea. In particular, it would help to reduce the confidence the interval given in Part 2.

5 Attachments

R code is attached on the next page.