Biostats VHM 801 Course
Fall 2020, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to home assignment II

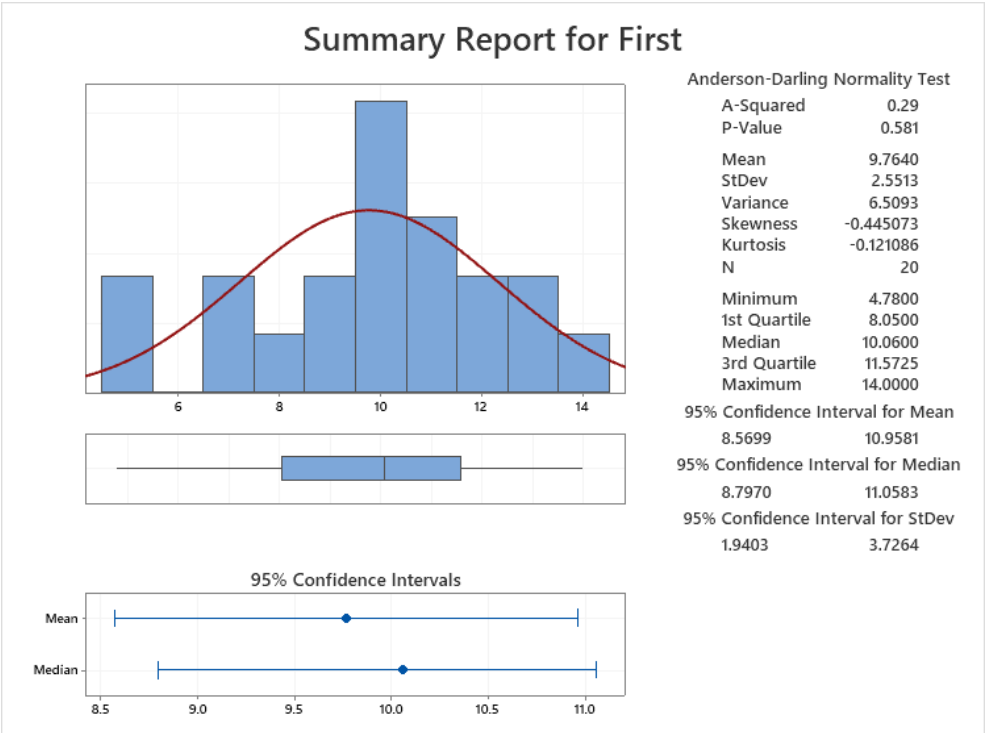
The solution is more detailed than required for a 100% mark, by covering multiple options for several questions. Question 6) was waived in the marking of the assignment.

1. Statistical design

The description of the experiment immediately points to the statistical design as two paired (or matched) samples. Each pair consists of two patients with similar age and disease duration. The purpose of pairing such patients is that it is expected that their reaction to the treatment (medicine or placebo) will be similar, thus reducing the between-patient variation and increasing the power of the study. Within each pair it should be randomly decided which patient will receive the treatment; this can be done using random digits (where odd and even numbers may represent medicine and placebo, respectively) or using random numbers from a uniform distribution (with numbers below and above 0.5 representing the two treatment categories). In addition to the pairs of patients, each patient is being measured twice (before and after the treatment), and these two measurements may also be considered as paired observations. Finally, (single) blinding seems quite feasible for this study, by not informing the patients whether the pill they received was medicine or placebo. If the pills were made to be indistinguishable, their true identity could possibly also be unknown to the personnel tending to the patients (i.e., double blinding).

2. Baseline level of measurements

Because the interest is in describing the measurements before any treatment, we should use the first measures of all patients regardless of their pair and group. So the first step is to combine the values into a sample of 20 observations. A graphical summary of the distribution is shown below:



The histogram is quite noisy with 20 observations (and too many bins), but the overall impression is of a unimodal distribution, centered at around 10 and slightly left-skewed, in part due to two observations at around 5 being somewhat lower than the rest (but not enough to be indicated as suspected outliers). The P-value for the A-D normality test is nowhere near significance, and also the normal probability plot (not shown) confirms the impression of a distribution that is not too far from normal. Throughout, we will develop the statistical inference in a series of steps.

Model. Our inference is based on the first measured locomotion powers (X_i), and we assume that $X_1, \ldots, X_{20}$ are i.i.d. from $\mathrm{N}(\mu, \sigma)$, with both parameters unknown. As usual, i.i.d. stands for independent and identically distributed.

Estimation. In a single sample, estimation is by the sample mean and standard deviation: $\hat{\mu} = \bar{X} = 9.764$ and $\hat{\sigma} = s = 2.551$. If we however, per the instructions of the assignment, assume $\sigma = 2.5$ to be known, we can use z-inference (with $z^* = z_{.95} = 1.645$) to get the intervals:

$$\begin{aligned} 90\%\text{ CI for } \mu &: \ 9.764 \pm 1.645 \cdot 2.5/\sqrt{20} = 9.764 \pm 0.920 = (8.84, 10.68), \\ 90\%\text{ range of } \mathrm{N}(\mu, \sigma) &: \ 9.764 \pm 1.645 \cdot 2.5 = 9.764 \pm 4.113 = (5.65, 13.88). \end{aligned}$$

Alternatively, we could use the estimated $s = 2.551$ and the corresponding $t^* = t_{.95}(19) = 1.729$ to get the intervals:

$$\begin{aligned} 90\%\text{ CI for } \mu &: \ 9.764 \pm 1.729 \cdot 2.551/\sqrt{20} = 9.764 \pm 0.986 = (8.78, 10.75), \\ 90\%\text{ range of } \mathrm{N}(\mu, \sigma) &: \ 9.764 \pm 1.729 \cdot 2.5 = 9.764 \pm 4.411 = (5.35, 14.18). \end{aligned}$$

Model check. We already described the distribution above. With a sample size of 20, t-distribution inference for the population mean should be robust enough to allow the minor deviations from normality. The estimated range is not equally robust, so here the assumed normality is more critical. One alternative could be to use the (empirical) 5% and 95% percentiles of the data, but this procedure has not been discussed in the course, and relies on software and its specific implementation of the calculations. For example, the `centile` command in Stata gives these percentiles as 4.78 and 13.97, respectively.

3. Initial measurements in medicine and placebo groups

As an aid for the solution to this and the following questions, we give a table (on the next page) with values of several variables obtained as differences between the original variables, together with a few key descriptive statistics. Due to the small sample size, graphical descriptive displays are of limited use beyond ensuring that no strongly outlying observations are present. The most obvious graphical displays would be stemplots or dotplots, but these are not included in the solution.

Data. Denote by $X_1, \ldots, X_{10}$ and by $Y_1, \ldots, Y_{10}$ the initial locomotion powers in the knee in the medicine and placebo groups, respectively. Compute, corresponding to the matched pairs design, the differences ("medicine−placebo") $D_i = X_i - Y_i$ in initial locomotion power within each patient pair.

Model. Our inference is based on the differences (D_i), and we assume that $D_1, \ldots, D_{10}$ are i.i.d. from $\mathrm{N}(\mu_D, \sigma_D)$. As usual, i.i.d. stands for independent and identically distributed.

Estimation. In a single sample, estimation is by the sample mean and standard deviation:

$$\begin{aligned} \hat{\mu}_D = \bar{D} = 0.396, \ \hat{\sigma}_D = s_D = 1.381, \\ 95\%\text{ CI for } \mu_D : \ 0.396 \pm 2.262 \cdot 0.437 = 0.396 \pm 0.988 = (-0.59, 1.38), \end{aligned}$$

with the CI computed by the formula: $\bar{D} \pm t^* \, s_D/\sqrt{10}$, using $t^* = t_{.975}(9) = 2.262$ with 9 df.

Difference Pair/Statistic	Medicine 1 − Placebo 1	Medicine 2 − Medicine 1	Placebo 2 − Placebo 1	Diff. M2−M1 − Diff. P2−P1
1	0.06	2.50	2.89	-0.39
2	2.60	-0.39	2.23	-2.62
3	1.93	1.92	4.38	-2.46
4	-1.39	0.30	0.41	-0.11
5	0.61	0.59	4.26	-3.67
6	0.63	0.61	2.27	-1.66
7	-0.01	-0.96	0.00	-0.96
8	0.92	1.61	2.40	-0.79
9	0.67	0.98	2.36	-1.38
10	-2.06	2.53	1.18	1.35
mean	0.396	0.969	2.238	-1.269
std. deviation	1.381	1.173	1.438	1.433
std. error	0.437	0.371	0.455	0.453
skewness	-0.32	-0.14	-0.01	0.11

Model check. With a sample size of only 10 (pairs), it is difficult to assess any problems with the normal distribution. The normal probability plot looks reasonably straight (all points are well within the confidence bounds), and the test for normal distribution is far from significant (not shown). The skewness is quite close to zero. In conclusion, there is no indication of non-normality and certainly no statistical evidence against the normal distribution.

Hypothesis and Test. The statistical hypothesis of interest is H_0: $\mu_D = 0$, corresponding to the same (population) mean in the medicine and placebo groups (that is, $\mu_X = \mu_Y$). There is no reason to prefer any particular alternative direction, so we take a two-sided alternative hypothesis H_a: $\mu_D \neq 0$. Our test statistic is the one-sample t-test:

$$t = \bar{D} / \left(s_D / \sqrt{10} \right) = 0.91, \quad P = 2 \cdot P(t(9) \geq 0.91) = 0.39.$$

The test is clearly non-significant, and there is no indication of a difference in locomotion powers between the two groups prior to treatment. This is what the researchers would hope for and also expect, if the treatments were randomized in each pair, and the major sources of heterogeneity were captured by the pairing.

Added note on analysis with known standard deviation. A known standard deviation of $\sigma = 2.5$ on individual values can be used to compute a (known) standard deviation for the difference (σ_D), if we additionally know the correlation of the values within each pair. Without such information provided, our only option would be to assume the values within each pair as independent, in which case the standard deviation follows from the rules for standard deviations of differences (lecture slide 3L−17) as: $\sigma_D = \sqrt{\sigma^2 + \sigma^2} = \sqrt{2 \cdot 2.5^2} = 3.536$. Using this value, we get $z = 0.35$ and the associated $P = 0.72$. Because the sample standard deviation for the differences is much smaller (due to the actual correlation within each pair being positive), this inference is weaker and not really to be recommended.

4. Improvement in the medicine group

For each patient, measurements are taken before and after treatment administration, and these two measurements should also be considered as matched (within each subject). From the two measurements it is natural to compute the improvement ("second−first") as a single outcome for each patient.

The design for these improvements is a standard matched samples design.

The analysis proceeds in a similar way as in Question 3, only with a different set of differences (the second data column in the table on the previous page). To keep this solution reasonably short, the steps from above will be reviewed in abbreviated form omitting entirely similar statements (such as the model assumption).

Estimates and 95% confidence interval for the improvement in locomotion power in the medicine group:

$$\hat{\mu}_D = \bar{D} = 0.969,\ \hat{\sigma}_D = s_D = 1.173,$$
$$95\%\text{ CI for } \mu_D:\ 0.969 \pm 0.839 = (0.13, 1.81).$$

The confidence interval does not include zero, this being indicative of a significant improvement in power. We may follow up with a statistical test, for which we take the hypotheses H_0: $\mu_D = 0$ and H_a: $\mu_D \neq 0$. The two-sided alternative is most natural if there is interest in detecting any effect of the medicine, both an improvement and a detrimental effect. One could argue that only an improvement is of real interest, and take the one-sided alternative H_a: $\mu_D > 0$. The choice is to some extent subjective, but my preference would be for the two-sided alternative. The t-test takes the form:

$$t = 0.969/\left(1.173/\sqrt{10}\right) = 2.61, \quad P = 2 \cdot P(t(9) \geq 2.61) = 0.028.$$

Using Table C of PSLS, we could conclude from $2.61 > 2.398 = t_{.98}(9)$ that $P < 2 \cdot 0.02 = 0.04$. The P-value against the one-sided alternative H_a: $\mu_D > 0$ equals half of the two-sided P-value. In either case, the test is (weakly) significant, and the analysis therefore provides (weak) evidence of an improvement in locomotion power from before to after treatment. This does not by itself prove that the medicine has a beneficial effect; it only shows that the patients in this group improved between the two measurements.

5. Assessment of treatment effect

The treatment effect becomes apparent by comparison of the medicine and placebo groups. By the pairing of patients, we still have a matched pairs design. As the outcome of interest is the improvement in locomotion power from before treatment (medicine or placebo) to after treatment, we need to form the differences within each pair between these improvements (computed separately for each patient). In effect, the variable of interest is a difference of differences; see the last column in the table on the first page for the values obtained in the 10 pairs.

Once more the analysis is similar to previous questions, and we discuss the results using the same abbreviated form.

Estimates and 95% confidence interval for the difference between improvement in locomotion powers between the medicine and placebo groups:

$$\hat{\mu}_D = \bar{D} = -1.269,\ \hat{\sigma}_D = s_D = 1.433,$$
$$95\%\text{ CI for } \mu_D:\ -1.269 \pm 1.025 = (-2.29, -0.24).$$

The confidence interval does not include zero, this being indicative of a significant difference between the improvements in the two groups. However, as the estimate is negative, and the difference was computed as "medicine−placebo", the result seems to indicate a greater improvement in the placebo group. For a statistical test, we take the hypotheses H_0: $\mu_D = 0$ and the alternative hypothesis H_a as either two-sided or one-sided with similar rationales as above. (Note that it is *not* allowed to adapt

the alternative hypothesis to the direction in the data.) The t-test takes the form, for the two-sided alternative H_a: $\mu_D \neq 0$,

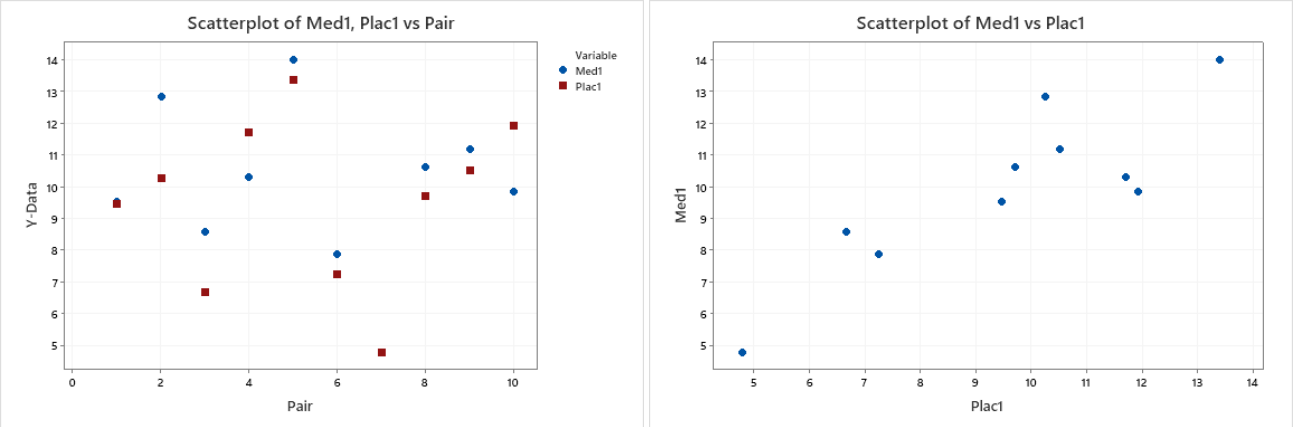
$$t = -1.269/\left(1.433/\sqrt{10}\right) = -2.80, \quad P = 2 \cdot P(t(9) \geq 2.80) = 0.021.$$

The test is (weakly) significant, and therefore provides evidence of a different improvement in locomotion power from before to after treatment in the medicine and placebo groups. The estimates show that the improvement is greater in the placebo group. The P-value against the one-sided alternative H_a: $\mu_D > 0$ is $P = P(t(9) > -2.80) = 0.99$ and thus provides absolutely no evidence against H_0; intuitively, this is because the alternative hypothesis fits even worse to the data than the null hypothesis. With this alternative hypothesis our conclusion would therefore be that there is no evidence of a beneficial treatment effect.

Conclusion. There is some evidence that the medicine works differently than the placebo, but the data indicate the improvement to be less with medicine than with placebo. If we look at the two groups separately, the medicine group shows a weakly significant improvement from before to after mediciation, but the placebo group shows a larger and clearly significant improvement (not shown in this solution, but the estimates are listed in the table). Obviously, it can be recommended to the researchers that they carefully check their protocols and perhaps redo the experiment to verify that the results are correct and reproducible. If so, the medicine is not to be recommended (it should rather be prohibited!). The placebo effect would then show either that a natural improvement occurred between the two measurements on each patient, or that there is some room for improvement of the patients' conditions by non-medical stimulation.

6. Impact/usefulness of pairing

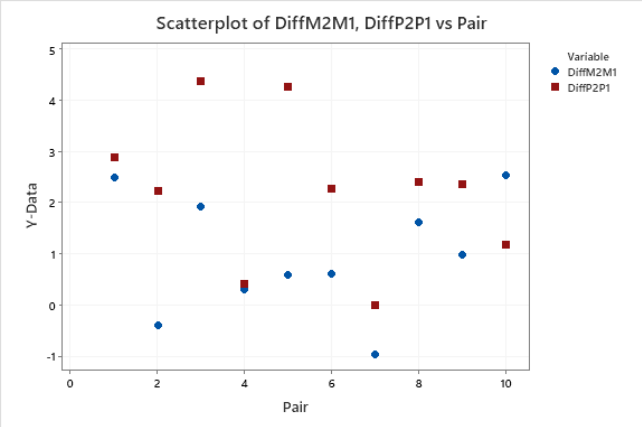
The pairs are blocks, and as such they would be expected to reduce variability in the statistical assessments. Any impact of the pairs would have to show up as less variability within a pair than between pairs. Two graphical displays that can illustrate this are shown below: to the left the two observations within each pair are plotted together (vertically) to allow a visual assessment hereof, and to the right the two values in each pair are plotted against each other. If there was similarity of values within pairs, one would expect the points in the right plot to show a strong positive association, possibly close to the line $y = x$.

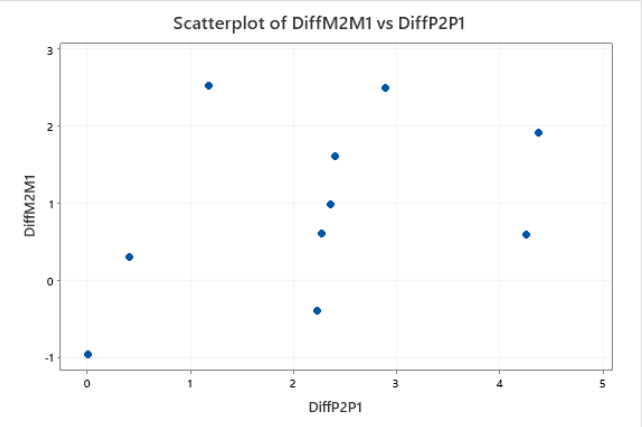


Both plots show values of the first measurement, before any treatment. Because the treatment cannot have had any effect yet, the values should be comparable. Indeed, the left plot seems to show the values within each pair to be mostly rather close, whereas some pairs are quite different, and the

right plot seems to show a positive association. Ways to quantify these impressions are beyond the material covered so far: a one-way ANOVA for the group of pairs in the left plot shows a strongly significant effect of pairs, whereas a correlation coefficient to quantify the association in the right plot is estimated at 0.86 and gives evidence of a positive association.

While these findings are encouraging for the effect of the pairing, they do not relate directly to the purpose of the study. The treatment and placebo effects are reflected in the *differences* between the second and first measurements for each patient. We should therefore redo the assessment for those differences.





The plots are now much less convincing, and the aforementioned patterns are not so obvious any more. Even without the quantification of the patterns available to us (in fact, neither the correlation coefficient nor a (two-way) ANOVA show any significance for the differences), it seems less clear that the pairing has had any noticeable impact on the differences. One interpretation of this finding is that the use of each patient as her/his own control by computing improvements, largely eliminated the dependence of the outcome on the initial state of the patient.

Another idea to assess the impact of the pairs is to compare the results of the paired sample analysis in 5) with the results one would get when assuming two independent samples (for the patient differences within the medicine and placebo groups). The estimated difference, $\hat{\mu}_M - \hat{\mu}_P = \hat{\mu}_D = \bar{D} = \bar{X} - \bar{Y}$ is the same, but the standard errors are not: paired: 0.453; independent: 0.586. Therefore the test statistics are also different: paired: $t = -2.80, P = 0.021$; independent: $t = -2.16, P = 0.045$. These comparisons show that the pairing did yield a moderate increase in precision, but not a dramatic improvement. In summary, the assessment of the usefulness of the pairing is therefore perhaps inconclusive: there was some gain in using the pairs but it did not substantially affect the conclusions of the analysis.