Biostats VHM 801 Course
Fall 2021, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to Mid-term Exam, October 2021

The solution is more detailed than required for a 100% score, by including discussions of several options for some of the questions, in particular for both parts of **e)**. The data are from Baldi B & Moore DS, *The Practice of Statistics in the Life Sciences*, 2nd ed., with a few minor modifications, and the study was conducted at the City University of New York.

Question 1

Subquestion a)

The study is experimental because the students were randomly divided into the two groups receiving the different subliminal messages; thus, the treatments (i.e. the messages) were imposed by the experimenter). The randomization would correspond to a completely randomized design. Furthermore, the design has characteristics of both 2 independent samples and 2 paired samples. For every student, the scores at the beginning and end of the summer course are paired samples. The two groups of students (receiving different messages) are two independent samples. The three variables involving the test scores (`before`, `after`, `diff`) are quantitative although the test scores only appear to take integer values (and may therefore be considered as discrete). The `group` variable is (nominal) categorical, or binary/dichotomous as there are only two groups.

Subquestion b)

We base our descriptive analysis for the `before` variable on the dotplot and the descriptive statistics of the first table. The discrete nature of the variable is obvious in the dotplot, which is dominated by the peak at 18, making the distribution unimodal. The distribution appears to have a longer tail to the right and thus to be somewhat right-skewed, which is also indicated by a skewness of 0.32, the mean being slightly larger than the median, and the box formed by the quartiles being wider to the right than to the left of the median. The dotplot could suggest a potential outlier to the left, and we may assess whether it is a suspected outlier by the rule of the boxplot. The size of the box is $\text{IQR} = \text{Q3} - \text{Q1} = 4.25$, and the lowest value (15) is nowhere near at being $1.5 \times \text{IQR}$ less than Q1 (18). So it is not a suspected outlier by this rule. Nor are there any suspected outliers in the right tail. The biggest problem with fitting a normal distribution to these data is presumably the discreteness, because the right-skewness is not that strong. A normality test could tell us whether there is formal evidence against a normal distribution; it was however not included among the statistics shown.

Subquestion c)

The focus in this question is on the variable `diff`, holding the improvements in test score for each of the 18 students. The Minitab output does not include any graphical summaries for `diff`. We must therefore rely on the (limited) descriptive statistics to get an impression of the data. The values in both groups seem to spread out fairly well with no obvious outliers, and there is no strong skewness.

The natural statistical model is therefore that we have two independent samples of normally distributed observations. If we denote the difference scores in the positive group by $X_1, \ldots, X_{10}$ and the difference scores in the negative group by $Y_1, \ldots, Y_8$, the statistical model could be stated as:

$$X_1, \ldots, X_{10} \sim \text{N}(\mu_1, \sigma_1) \quad \text{and} \quad Y_1, \ldots, Y_8 \sim \text{N}(\mu_2, \sigma_2),$$

and all variables are also assumed independent. The second Minitab listing gives both estimates for the mean parameters and a 95% confidence interval based on this model:

$$\hat{\mu}_1 - \hat{\mu}_2 = 3.18, \quad 95\% \text{ CI for } (\mu_1 - \mu_2) : (-0.26, 6.61).$$

We are also asked to carry out a statistical test to assess the significance of the difference in test scores between the two groups. The wording of the question invites a one-sided alternative ("a greater improvement"). The same Minitab listing also gives a t-test statistic.

$$\mathrm{H}_0 : \mu_1 = \mu_2, \quad \mathrm{H}_a : \mu_1 > \mu_2,$$
$$t_{\text{obs}} = 1.98, \ \text{df} = 14.$$

As we want to test against a one-sided alternative, we need to compute the P-value manually. The critical value at a 5% significance level is 1.761 (from Table C of PSLS), hence the test is significant at $P < 0.05$. The P-value in the listing for the two-sided alternative is twice that of the value for a one-sided alternative, hence $P = 0.067/2 = 0.034$.

We conclude that there is some (weak) evidence against H_0 and hence of a greater improvement in the positive message group. The estimated difference in improvements was 3.18 (test score units). The 95% CI included zero but this was because a test with a two-sided alternative would be just non-significant. We are 95% confident that the true difference in improvements is roughly in the range between 0 and 7 test score units.

Subquestion d)

Knowing from **c)** that there might be some beneficial effect of the positive messages, we should preferably assess the impact of the course in the neutral message group. We therefore consider the $Y_1, \ldots, Y_8$ scores in the neutral group, and use the same model as above — thus a single sample from a normal distribution. The estimated mean improvement is $\hat{\mu}_2 = 8.13$. For a "measure of its precision" there are two obvious choices: the standard error (SE) of the mean, and a confidence interval. The standard error of the mean gives the standard deviation in the distribution of the mean. The confidence interval converts this information into an interval estimate of the true population mean (μ_2). The calculations are (with $t^* = t_{.975}(n_2 - 1) = t_{.975}(7) = 2.365$ from Table C)

$$\begin{aligned} \text{SE} &= s_2/\sqrt{n_2} = 3.48/\sqrt{8} = 1.23, \\ 95\% \text{ CI for } \mu_2 &: \hat{\mu}_2 \pm t^* s_2/\sqrt{n_2} = 8.13 \pm 2.91 = (5.22, 11.04). \end{aligned}$$

Both the estimate and the confidence interval are far from 0, the value corresponding to no improvement from before to after the summer course. The data demonstrate quite clearly, intuitively with strong statistical significance despite that we have not done a formal test, an improvement in the math skills, as measured by the two tests.

Subquestion e), part i)

All improvements in test score in the neutral group were positive. This corresponds to observing $X_{\text{obs}} = 8$ out of $n_2 = 8$ positives in a binomial trial, with $X \sim B(8, p)$. For this to have happened "by chance alone" would mean that $p = 0.5$, corresponding to equal chances of a positive and negative improvement. We can substantiate the claim by testing the hypothesis $\mathrm{H}_0 : p = 0.5$ against the alternative $\mathrm{H}_a : p > 0.5$ (it is difficult to imagine that the students would perform worse after the course). With the small sample size, we can only use the exact test in $B(8, 0.5)$ which for the observed $X_{\text{obs}} = 8$ gives $P = P(X \geq X_{\text{obs}}) = P(X = 8) = 0.5^8 = 0.0039$ (Table 1 of S lists this value as 0.004). Therefore we conclude that the instructor was indeed right to say that this would not have happened by chance alone; the fact that all 8 students in the neutral group improved their score demonstrated a beneficial effect of the course. Note that the procedure we used really corresponds to the sign test.

Subquestion e), part ii)

For a comparison of the initial test scores in the two groups we have again a two-sample situation. One possibility is to assume two normal distributions in the same way as for part **c)**. The Minitab listing does not provide a t-test for testing the two `before` means being equal, but calculation is possible from the descriptive statistics provided. We could also use the `2-sample t` menu in Minitab. One detail to consider is whether the alternative hypothesis would be one- or two-sided: there is no reason to expect a difference in any particular direction, so two-sided is the obvious choice. Another detail is whether to use the test with a pooled standard deviation: as the two standard deviations are fairly close, one would not expect major differences between the two versions of the two-sample t-test. If the calculation is to be done using software anyway, one may as well choose the non-pooled t-test so as to not make any assumptions about the standard deviations.

Before proceeding with this approach we should consider the model assumptions (these were listed in **c)** so they are not repeated here). Contrasting the overall distribution we described in **b)**, we should look at the two groups separately. It is difficult to assess normality from the information provided. There is no indication of extreme values. One possible concern is that the distribution of initial test scores in the positive group has half of its values equal to 18, and this value is at the left end of the distribution. This is perhaps a bit too discrete (and skewed) for a normal distribution (and normality tests are indeed significant when carried out by statistical software, but that information is not available in the listing). It is not obvious whether this problem with one of the two distributions will invalidate the t-distribution based inference because of the robustness of the t-procedures (as discussed in Session 6).

One alternative is to use a non-parametric test to compare the two groups: this would be a Wilcoxon-Mann-Whitney two-sample rank test. This test is available within the `Nonparametrics` menu in Minitab. The null hypothesis tested would be that the two distributions (in the positive and neutral groups) of initial test scores are equal, against a non-specific alternative that they differ in some way.