Biostats VHM 801 Course
Winter 2018, Atlantic Veterinary College, PEI
Henrik Stryhn

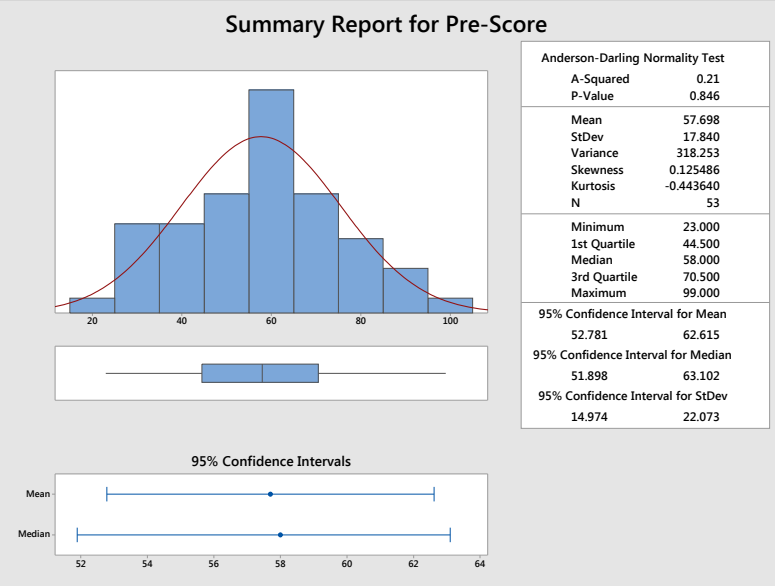
Solution to home assignment II

The solution is more detailed and verbose than required for a 100% mark, and includes additional remarks on outlier tests.

a) Reading skills prior to chess instruction

We denote by $X_1, \ldots, X_{53}$ the DRP-scores prior to chess instruction by the $n = 53$ students. Because the scores are normalized to grade level, it is meaningful to assume these to form a sample from a single population (spanning several grades). It also seems reasonable to assume the scores to be independent; we have no information to the contrary. In other words, we assume the X_i to be i.i.d. (independent and identically distributed) observations. In addition, we assume their distribution to be normal $\mathrm{N}(\mu, \sigma)$ where μ and σ are unknown parameters. This assumption will be critically assessed as part of our statistical analysis. The Graphical Summary (or Summary Report) from Minitab (below) shows descriptive statistics, and we have the parameter estimates:

$$\hat{\mu} = \bar{X} = 57.70; \quad \hat{\sigma} = s = 17.84.$$



The distribution looks unimodel and symmetrical, and is seems pretty close to a normal distribution, possibly with slightly too short tails (not surprisingly, the percentages are bounded by 0% and 100%). Formally, the A-D normality test gives no evidence whatsoever against normality ($P = 0.85 \gg 0.05$). Thus, we have no concerns with using inference based on the t-distribution. The 95% confidence interval (CI) for μ is listed in the display as (52.78,62.62). Because the CI does not include the value 50, we have formal significance at the 5% level for a t-test of $\mathrm{H}_0 : \mu = 50$ against the two-sided alternative $\mathrm{H}_a : \mu \neq 50$. The wording of the question ("lower or higher than 50%") leads to a two-sided H_a. In order to get more complete information about the significance, we compute the t-test statistic,

$$t = (\bar{X} - 50)/(s/\sqrt{n}) = (57.70 - 50)/(17.84/\sqrt{53}) = 3.14, \quad P = 2 \cdot \Pr(t(52) \geq 3.14) = 0.003.$$

The P-value is much lower than 0.05, and we conclude that there is strong evidence against H_0, which therefore must be rejected in favour of the alternative (H_a) that μ differs from 50%. Because the sample mean (57.7) exceeds 50, we have evidence to say that the students who were offered chess instruction have higher reading skill scores than their respective grade average.

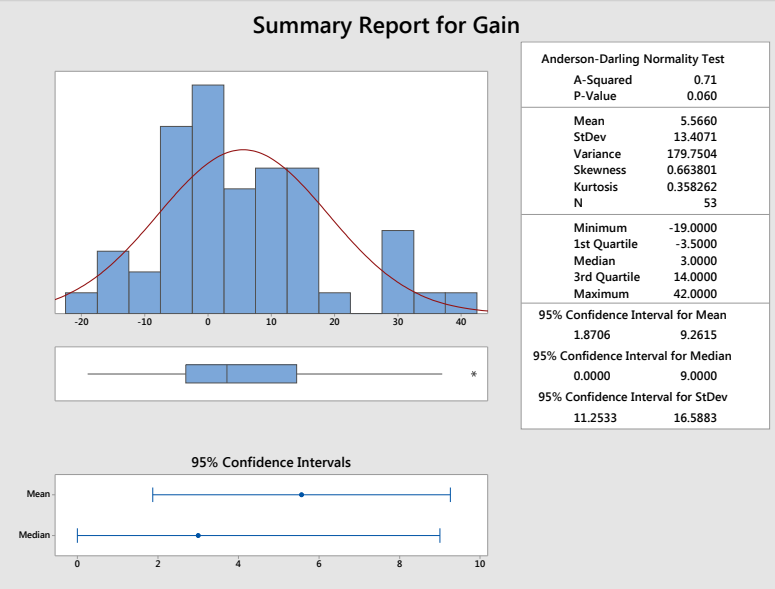
b) Study design and population

The reading scores before and after chess instruction are two paired samples because they were obtained from the same students. We would expect the two reading scores from the same student to be generally similar, where a beneficial effect of the chess instruction would cause the second scores to be (mostly) larger than the first scores.

The students involved in the study were volunteers offered participation after having shown an interest in chess. Therefore the population they might represent are students with an interest in chess and a willingness to commit to the study. We already showed in a) that these students had higher reading skills (as reflected by higher DRP scores) than the general student population, and both the interest in chess and the commitment to the study might further select for students with specific interests and resources. The data do however not offer us any information to assess that. Finally, our results for the effects of chess instruction would obviously require the students to have received similar instruction.

c) Confidence intervals for mean improvement

For each of the $n = 53$ students we work with the gain (improvement) $D_i = Y_i - X_i$, $i = 1, \ldots, n$, that is, the difference between DRP-scores after (Y) and before (X) the chess instruction ("post–pre"). We assume these 53 gains to be i.i.d. from an unspecified distribution with mean μ and standard deviation σ. As will be discussed below, assuming a normal distribution is not as natural here. We use similar descriptive tools as in a), and estimate the parameters by their sample values: $\hat{\mu}_D = \bar{D} = 5.57$ (the estimated mean improvement) and $\hat{\sigma}_D = s_D = 13.41$.



The histogram shows a slight/moderate right-skewness of the distribution, and the skewness is estimated at 0.66. The skewness also shows from the mean being a bit larger than the median, and an

asymmetry in the interquartile range around the median. The normality plot (not shown) is somewhat curved, and the P-value for the A-D normality test ($P = 0.060$) is close to the significance level of 0.05. For these reasons it is not obvious to assume a normal distribution for the gains. We will nevertheless use t-based procedures for (approximate) inference about the mean. A discussion of how valid this approximation might be is deferred to question d).

formula	$t^* = t_{1-\alpha/2}(52)$ $(t_{1-\alpha/2}(50)$	*confidence interval*: $\bar{D} \pm t^* s_D/\sqrt{n}$
95% CI ($\alpha = 0.05$)	2.007 (2.009)	$5.57 \pm 2.007 \cdot 13.41/\sqrt{53} = (1.87, 9.26)$
90% CI ($\alpha = 0.10$)	1.675 (1.676)	$5.57 \pm 1.675 \cdot 13.41/\sqrt{53} = (2.48, 8.65)$

The approximate 95% confidence interval includes the true population mean (μ_D) with (approximate) probability 0.95, the probability referring to the proportion of times within a long series of repeated experiments such intervals would include μ_D. Within the same series of repeated experiments, the 90% confidence intervals would include μ_D at a proportion of 0.90. However, as the study population is a group of non-randomly selected children, it is difficult to imagine how the experiment could be repeated. Another interpretation of the confidence intervals is that they comprise the values for which we have no evidence against μ_D being equal to that value, at the 5% and 10% significance levels, respectively. For example, as 2.0 is in the 95% but not the 90% confidence interval, we have evidence at the 10% level but not at the 5% level that μ_D is not equal to 2. In other words, the P-value for testing $\mathrm{H}_0 : \mu_D = 2$ against $\mathrm{H}_a : \mu_D \neq 2$ is between 0.05 and 0.10.

d) Test for improvement

Our null hypothesis of interest is $\mathrm{H}_0 : \mu_D = 0$, corresponding to no improvement. As the question is worded towards *improvement* of test scores, it is most natural to take a one-sided alternative hypothesis, $\mathrm{H}_a : \mu_D > 0$.

formula	*test*: $t = \bar{D}/(s/\sqrt{n})$, $P = \mathrm{P}(t_{n-1} > t_{\mathrm{obs}})$
results	$t_{\mathrm{obs}} = 5.57/(13.41/\sqrt{53}) = 3.02$, $P = 0.0019$

With a P-value of 0.002 there is clear evidence against the null hypothesis. We therefore reject H_0 and turn instead to the alternative H_a. There is evidence of a mean improvement of the test scores, and as indicated by the confidence intervals the approximate range of the improvement is between 2 and 9 percentage points. This is a mean improvement and does not imply that every child improves.

As noted above, our basic assumption is that the improvements form a sample of independent observations from a common distribution. In part c), we noted the distribution to be somewhat right-skewed, so we did not assume a normal distribution for the improvements. This means that all our inference is approximate only, based on the degree to which the distribution of the sample mean $\bar{D}$ is approximated by a normal distribution. With a minor skewness only and no strong outliers the textbook guidelines for use of z/t procedures (lecture slide 7L−3) are definitely met, but that does not tell us how accurate the approximation is. In any case, our decision to reject H_0 is hardly in question, so we stick with our conclusion of a mean increase in reading skills after the chess instruction.

e) Analysis without observation equal to 42

The boxplot shown in the Graphical Summary indicated 42 as a suspected outlier. Upon inspection of the data it does not look very extreme compared to the second-largest value of 37. Also, none of

the scores for subject no. 7 look unusual. Therefore it is not, based on the data alone, clear that the value 42 should be considered as an outlier.

The reduced dataset has a sample mean of 4.865 and a sample standard deviation of 12.52, a reduction of about 13% and 7% compared to their previous values, respectively. The methods for computing confidence interval and test are unchanged, using these new values and the reduced $n = 52$.

Excluding the value 42	95% CI	$4.865 \pm 2.008 \cdot 12.52/\sqrt{52} = (1.38, 8.35)$
	test	$t_{\text{obs}} = 4.865/(12.52/\sqrt{52}) = 2.80$, $P = 0.0036$

The test for no mean improvement is still clearly significant, even if the P-value has gone up. The confidence interval covers roughly the same range even if has moved somewhat to the left, following the drop in mean improvement scores by 0.7 (from 5.57 to 4.865). None of this is too surprising: when dropping the largest value, the mean improvement goes down and the data point to a smaller mean improvement. However, the impact here can be said to be modest, and the evidence of an improvement is still present. In conclusion, the results are quite robust to inclusion or removal of the largest observation.

f) Outlier assessment of the observation equal to 42

The sample mean and standard deviation of the reduced sample are given above. Let D be a random variable for improvement scores, and assume that $D \sim \text{N}(4.865, 12.52)$. The probability of an observation as large as 42, or larger, is therefore (where $Z \sim N(0, 1)$),

$$\text{P}(D > 42) = \text{P}\left(\frac{D - 4.865}{12.52} > \frac{42 - 4.865}{12.52}\right) = \text{P}(Z > 2.966) = 0.01508 \approx 0.0015,$$

and including the possibility of an extreme observation on the left side of the distribution as well, we get a probability of $2 \cdot 0.0015 = 0.003$. This probability is for a single (pre-determined) student.

Looking at one or more extreme students in a sample (of 53 students) is a different situation. If Y denotes the number of observations among 53 that are at least as extreme as 42, we may use a binomial $B(53, 0.003)$ distribution for Y to compute $\text{P}(Y \geq 1)$, or the rewriting (based on the multiplication rule for independent events)

$$\text{P}(Y \geq 1) = 1 - \text{P}(Y = 0) = 1 - \text{P}(\text{“not extreme”})^{53} = 1 - (1 - 0.003)^{53} = 0.147 \approx 0.15.$$

There is a 15% chance to get at least one observation as extreme as 42 in a sample of 53 observations. This is not a very low probability, so it seems that an observation of 42 (or more extreme) could easily have happened by chance only. Assuming the normal distribution to be valid, it therefore does not seem justified to consider the observed value of 42 as an outlier.

One critical assumption for the calculation is the normal distribution. As the extreme value is in the right tail of the distribution and the most pronounced deviation of the observed data from a normal distribution is a right-skewness, the probability of 0.15 is, if anything, an underestimate of the correct probability. This is because extreme values in the right tail are typically more likely in a right-tailed distribution than in either tail in a symmetrical distribution.

Another important part of the calculation is that the estimated mean and standard deviation are considered equal to true population values. We know very well that these estimates have uncertainty, and that uncertainty is not reflected in the above calculation, which therefore is expected to give too low P-values. When this uncertainty is correctly taken into account, (single) outlier detection under a normal distribution assumption method goes under the name of Grubbs' test, which is one of the options offered in Minitab in the “Basic Statistics – Outlier Test” menu.