Biostats VHM 801 Course
Winter 2026, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to Final Exam, April 2026

Question 1 was not included in the exam for students who used their midterm mark. The solution is more detailed than expected, by giving additional calculations and detailed interpretations and explanations of all procedures.

Question 1

The question is based on the article, de Fornel *et al.* (2007), Effects of radiotherapy on pituitary corticotroph macrotumors in dogs: A retrospective study of 12 cases, *Canadian Veterinary Journal* **48**, 481–486. Only 11 of the dogs returned for follow-up measurements. The recent article referred to is, Pham *al.* (2026), Defining variations in the size of normal pituitary glands and pituitary macrotumors based on canine skull morphology, *Veterinary Radiology & Ultrasound* **67**, e70135.

Subquestion a)

Tumor heights before and after the radiotherapy constitute paired samples on each dog. We have no information about the selection of the dogs, but it seems reasonable to assume that they represent a carefully defined study population (defined by certain inclusion criteria, one of which must be the presence of tumors). We would definitely assume the measurements on different dogs to be independent. In reply to the question, no it is not valid to use a two-sample t-test because it applies to two independent samples, and (as explained) our data constitute two paired samples.

Subquestion b)

If $D_1, \ldots, D_{11}$ denotes the differences (before minus after) in tumor heights for the 11 dogs, we assume that the D_i's constitute a simple random sample (or are i.i.d.) from $\mathrm{N}(\mu_D, \sigma_D)$. We estimate the mean by $\hat{\mu}_D = 14.1 - 8.8 = 5.3$ mm, and the standard deviation by $\hat{\sigma}_D = s_D = 5$ mm. Then

$$95\%\text{ CI for } \mu_D : \hat{\mu}_D \pm t^* s_D/\sqrt{n} = 5.3 \pm 2.228 \cdot 5/\sqrt{11} = 5.3 \pm 3.36 = (1.94, 8.66)$$

where $t^* = t_{.975}(10) = 2.228$. Next we test $\mathrm{H}_0 : \mu_D = 0$ against the one-sided alternative $\mathrm{H}_a : \mu_D > 0$ (because we were asked for evidence of a *reduction* in tumor size) by a t-test,

$$t = \hat{\mu}_D/(s_D/\sqrt{n} = 5.3/(5/\sqrt{11}) = 3.516$$

which corresponds to $P < 0.005$ (because $t_{.995}(10) = 3.169$, Table C of PSLS). We conclude that there is clear evidence of a reduction in tumor height, the likely magnitude of which is indicated by the CI.

Subquestion c)

If tumor heights (say X) follow a $\mathrm{N}(\mu, \sigma)$ distribution, and we use the estimates before the radiotherapy: $\hat{\mu} = 14.1$ and $\hat{\sigma} = 3.6$, then the calculation of the proportion of tumors of height less than 10 goes as follows,

$$\mathrm{P}(X < 10) = \mathrm{P}\left(\frac{X - 14.1}{3.6} > \frac{10 - 14.1}{3.6}\right) = \mathrm{P}(Z < -1.1389) \approx= \mathrm{P}(Z < -1.14) = 0.1271 \approx 0.127.$$

The calculated percentage of 12.7% is substantially lower than the quoted 21.3% from the 2026 paper, most likely reflecting differences in the populations of dogs, the two (retrospective) samples of dogs could be seen as representative for.

Subquestion d)

We have two non-parametric tests to assess change between two paired samples, both based on the differences being assumed a simple random sample (i.i.d. observations). They both test the hypothesis H_0 : median$(D) = 0$ against one- or two-sided alternatives. The sign test is based on the signs of the differences only, whereas the Wilcoxon signed rank test is based on ranks. Without access to the data we can only use the sign test. The information provided does not allow us to determine the sign of all the differences, but we do know that 10 out of 11 were positive. The sign test is based on a binomial distribution for the count (say X) of the number of positive differences.

If the 11^{th} dog showed no change, that dog will not be used for the sign test, and we therefore assume $X \sim \mathrm{Bin}(10, p)$, where p is the proportion of dogs for which the radiotherapy causes the tumor to shrink. We will test $\mathrm{H}_0 : p = 0.5$ against (in this case because of the focus on a reduction in tumor height) the one-sided alternative $\mathrm{H}_a : p > 0.5$. The binomial distribution table gives the P-value as: $P = \mathrm{P}(X = 10) = 0.001$, and it could also be computed directly as $P = 0.5^{10} = 0.00098$. The result is strong evidence against H_0 and therefore in favour of a reduction in tumor size.

The outcome for the 11^{th} that would give weakest evidence against H_0 is an increase in its tumor height. This would mean we still observe $X = 10$ but now from a binomial distribution $(11, p)$. This binomial distribution is not in our table, but it can be calculated exactly as

$$\begin{aligned} P &= \mathrm{P}(X \geq 10) = \mathrm{P}(X = 10) + \mathrm{P}(X = 11) = \binom{11}{10}(\tfrac{1}{2})^{10}(1 - \tfrac{1}{2})^1 + \binom{11}{11}(\tfrac{1}{2})^{11} \\ &= \binom{11}{1}(\tfrac{1}{2})^{11} + \binom{11}{0}(\tfrac{1}{2})^{11} = (11 + 1)(\tfrac{1}{2})^{11} = 0.0059. \end{aligned}$$

We could get a conservative estimate for this probability by computing it instead in $\mathrm{Bin}(10, 0.5)$, for which the binomial table gives: $P \leq \mathrm{P}(X \geq 9) = \mathrm{P}(X = 9) + \mathrm{P}(X = 10) = 0.0098 + 0.0010 = 0.011$, where again $X \sim B(10, 0.5)$. In both cases, there is still clear evidence to reject H_0 and hence in favour of a reduction in tumor height from before to after the radiotherapy.

Subquestion e)

Without access to the data it is impossible to say whether the author's use of a "Wilcoxon test" was correct or not. As two Wilcoxon rank tests exist (the Wilcoxon signed rank test, applicable to a single sample, and the two-sample Wilcoxon rank sum test), the statement of the method used is equivocal. Also, no justification for the use of a Wilcoxon test is given. The authors stated their P-value as $P < 0.02$; first, it is preferable to state an exact P-value instead of an upper bound, and second it seems that the P-value may be understated as the sign test gave a considerably lower P-value. As a last point of criticism, it is inconsistent to state means and standard deviations for the two groups in combination with a non-parametric analysis; it would have been better to state medians and inter-quartile ranges. Descriptive statistics for the differences could also have been included, as these relate more directly to the analysis.

Question 2

Subquestion a)

Let X denote the number of respondents answering that drinking and driving is a very serious or extremely serious problem among the $n = 1201$ respondents of the survey. The natural model is a binomial distribution, $X \sim \mathrm{Bin}(n, p)$. The statistical design is (simple random) sampling of a binary outcome (because focusing on the combined categories $5-6$) from a finite population, where the population is much larger (> 20 times) than the sample size. Alternatively, one may argue an

approximate binomial setting. We estimate the population proportion p by the sample proportion, $\hat{p} = 0.88$. With the large sample size (specifically, number of "positives" and "negatives" $\gg 15$), the normal approximation procedure is totally adequate for inference about p:

$$95\%\ \text{CI}: \quad p: \quad \hat{p} \pm z_{.975}\sqrt{\hat{p}(1-\hat{p})/n} = 0.880 \pm 1.96\sqrt{0.88 \cdot 0.12/1201} = 0.880 \pm 0.018 = (0.862, 0.898),$$
$$99\%\ \text{CI}: \quad p: \quad \hat{p} \pm z_{.995}\sqrt{\hat{p}(1-\hat{p})/n} = 0.880 \pm 2.576\sqrt{0.88 \cdot 0.12/1201} = 0.880 \pm 0.024 = (0.856, 0.904).$$

Subquestion b)

The statement means that 95% (19 times out of 20) confidence intervals have margin of errors of at most 2.9%. If the survey was based on a simple random sample, the model would have been the binomial model used for **a)**. The calculation of the margin of error uses the same formula as above: $1.96\sqrt{p(1-p)/n}$. Here $n = 1201$, and p would be replaced by the sample proportion for the question of interest. Different questions and answer categories would have different sample proportions and therefore different margin of errors. The message of the statement is that all the margin of errors would be at most 0.029. The largest margin of error is obtained for $p=0.5$, leading to $1.96\sqrt{0.5 \cdot 0.5/1201} = 0.028$. The difference is either due to round-off (using the factor 2 instead of 1.96) or to a different sampling scheme for which the margin of error would then be calculated by a more complex formula.

The formula for the required sample size to achieve a desired margin of error (m) is (lecture slide 12−9):

$$n \geq p(1-p)(z^*/m)^2.$$

With $m = 0.029$, and using again $p = 0.5$ for the worst-case situation and $z^* = 1.96$, we get: $n \geq 1141.97 \approx 1142$ (with $p = 0.88$: $n \geq 482.3 \approx 483$). It requires at least 1142 respondents to achieve a margin of error of at least 0.029 irrespective of p. As noted above, this calculation is based on simple random sampling and the simple binomial model.

For an answer based on the printed Minitab sample size menus, one should choose the menu for a Proportion (Binomial) parameter, insert the value 0.5 for the worst case scenario, and insert the value 0.029 for the margin of error of confidence intervals. The confidence level and type of confidence interval in the lower menu should stay at their default values. The value returned is 1174, close to the value from the classical CI and reflecting that there is no major difference between classical and exact confidence intervals with such a large sample size.

Subquestion c)

The number of respondents who considered drinking and driving very or extremely serious must have been approximately $1201 \cdot 0.88 = 1056.88 \approx 1057$. Accordingly, $1201 - 1057 = 144$ responses fell in categories $1-4$ of the question. Among the 1057 respondents 7.6% reported at least one instance of drinking and driving, corresponding to $1057 \cdot 0.076 = 80.33 \approx 80$ responses. Conversely, the 21.2% among the 144 respondents corresponds to $144 \cdot 0.212 = 30.5 \approx 31$ responses. The total number of respondents reporting at least one instance of drinking and driving must therefore have been approximately $80 + 31 = 111$, and the proportion is therefore approximately $111/1201 = 0.092$, or 9.2%.

The reported overall value of 7.7% seems suspect because it is too close to the 7.6% in the group who considered drinking and driving a very or extremely serious problem. The overall value would always be a weighted average of the averages within each group (7.6% and 21.2%), but for the overall average to be so close to one of the group averages would require this group to be vastly bigger than the other group. As seen in the the calculation above, a group consisting of 88% is not nearly enough to get an overall proportion of 7.7%.

Subquestion d)

The statistical model is two independent binomial distributions (or binary samples, or data for proportions). The two groups correspond to whether the respondents considered drinking and driving a very or extremely serious problem, or not. Among the Minitab listings, the first two correspond to one-sample models (for each of the two groups), the third corresponds to a two-sample continuous data model, whereas the fourth corresponds exactly to the two-sample binomial model. The test for $\mathrm{H}_0 : p_1 = p_2$ against a two-sided alternative $\mathrm{H}_a : p_1 \neq p_2$ (where p_1 and p_2 are the population proportions of drivers admitting to drinking and driving in the two groups) is reported as: $z = 2.54$ and $P = 0.011$.

Before using the test, we should check that the conditions for its application are satisfied. The easiest way to check the conditions is from its equivalence with the Pearson X^2-test (specifically, $X^2 = z^2$), when the data are set up in a 2×2-table (the last Minitab listing). The rule for the Pearson chi-square test is that all cells of the table have expected values less than 5, which in this case is not satisfied. Therefore, the z or Pearson chi-square test is not entirely appropriate to use. A Fisher's exact test would give a definitive P-value. However, even based on the information provided, it seems likely that some significance between the proportions remains. Therefore, we may conclude that the data contain some indication (and possibly statistical evidence at the 5% significance level; Fisher's exact test will confirm this) that the proportions of drivers admitting to drinking and driving are different in the two groups, i.e. higher for drivers considering drinking and driving less serious. Intuitively, such a difference would perhaps not seem too surprising.

Question 3

Subquestion a)

The study is clearly an experiment because the experimenter decided (randomly) how long the leaves for subjected to the light. As the 15 leaves were randomly assigned to the 3 groups, it is a completely randomized design. The experimental unit is the leaf, and the outcome measured is the leaflet angle. If we denote by Y_{ij} the leaflet angle of leaf j in treatment group i, where $i = 1, 2, 3$ ($\sim$ 30, 45, 60 minutes) and $j = 1, \ldots, 5$, the natural statistical model (and the one assumed in the Minitab listing) is that of a one-way ANOVA: $Y_{ij} = \mu_i + \varepsilon_{ij}$, with $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma)$. The hypothesis of equal means in the 3 delay groups, $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3$, is tested by the ANOVA F-statistic: $F = 6.56$, $P = 0.012$. We conclude that the F-test is significant, and we can reject H_0 and state that some differences between the delay groups exist.

The Minitab listing only includes the group means (among the Descriptive statistics). We can compute the margin of error for 95% confidence intervals as: $t^* s_p/\sqrt{5} = 2.179 \cdot 7.6245/\sqrt{5} = 7.43$, where the t^*-values was from a t-distribution with df = DFE = 12. If we were to carry out pairwise comparisons without adjustment for multiple comparisons, this value shows us that confidence intervals for delay groups 30 and 60 would not overlap, and are hence significantly different, and that confidence intervals for delay groups 30 and 45 would contain the estimate for the other group, and are hence not significantly different. We are unable to conclude about statistical significance between 45 and 60 minute delays based on the confidence intervals. We could compute an LSD-value for pairwise comparisons as follows,

$$\mathrm{LSD}_{.95} = t^*(12) s_p \sqrt{2/n} = 2.179 \cdot 7.6245 \sqrt{2/5} = 10.5.$$

This shows that the unadjusted comparison between 45 and 60 minute delays is also significant. We might want to supplement the analysis with an adjustment for multiple comparisons, and in Minitab both the One-way ANOVA menu and the Comparisons menu after a General linear model can be used to request different types of multiple comparison adjustments.

Subquestion b)
For each of the statements we indicate whether it is correct or false, with a brief explanation. Alternative interpretations of some of the statements were accepted as well, if accompanied with a suitable explanation. A bonus was given if both the statements that (erroneously) were listed for (b9) were assessed correctly. Also, statement (b7) was waived.

(b1) False; in the one-way ANOVA, the groups are considered as three categories, so it makes no sense to compute a correlation.

(b2) False; the one-way ANOVA is superior to multiple 2-sample t-tests; note that pairwise comparisons following the ANOVA F-test differ from regular two-sample t-tests by using the pooled (error) standard deviation and its degrees of freedom.

(b3) False; it is virtually impossible to assess the normal distribution based on the information provided (see also (b4)).

(b4) Correct; it is difficult to detect any deviations from a normal distribution based on only 5 observations per group. Interestingly, the A-D normality test gives $P = 0.027$ for the delay 60 group, whereas other normality tests (including Shapiro-Wilk) are non-significant.

(b5) Correct, in the sense that the boxplot did not display any suspected outliers.

(b6) Correct, because $s_{\max}/s_{\min} = 11.61/1.67 = 6.95 \gg 2$.

(b7) False; trying out a log-transformation can be a good idea, but it will work well only if the standard deviations (roughly) increase with the means, but that is clearly not the case here (based on the estimates).

(b8) Correct.

(b9) False; if the assumptions of a normal distribution model/analysis are met, this will usually offer a better analysis, in particular for a small dataset.

(b9) False; with a total of 15 observations the dataset is not too small for a non-parametric analysis, in fact the Kruskal-Wallis test shows a significant difference between groups.

(b10) True; a sensitivity analysis to explore the impact of making different assumptions on the results is a general method to confirm the validity of the conclusions.

Subquestion c)
The Minitab display is for a linear regression with delay as a quantitative explanatory variable. This model can be written as,

$$Y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij}$$

where the delays are $(x_1, x_2, x_3) = (30, 45, 60)$ and the errors are assumed $\sim \mathrm{N}(0, \sigma)$. The estimated intercept and slope are $\hat{\beta}_0 = 157.7$ and $\hat{\beta}_1 = -0.5733$. The estimated residual standard deviation is $s = 7.44277$. The display does not include an ANOVA table or standard errors for the regression coefficients to compute a t-test ($t = \hat{\beta}_1/\mathrm{SE}(\hat{\beta}_1)$) for the slope. There are ways to work around the missing information, but if none of these come to mind it would be sensible to request Minitab output to include either of these results.

The simplest workaround comes from the fact that the t-tests for the correlation and slope are identical, and the former can directly be computed from $r = -\sqrt{R^2} = -\sqrt{0.507} = -0.712$, as follows

$$t = r\sqrt{(n-2)/(1-r^2)} = 0.712\sqrt{(15-2)/(1-0.507)} = -3.66,$$

with $P < 0.0025$ from $t(13)$ when testing $H_0 : \beta_1 = 0$ against $H_a : \beta_1 < 0$. A two-sided alternative would give $P < 0.005$. The conclusion is the same, the data show evidence of a (linear) decline.

It is also possible to construct the ANOVA table for the linear regression model by combining the information for Total from the 1-way ANOVA (this row is always the same) and the estimated standard deviation ($= \sqrt{\text{MSE}}$) or the $R^2 = (1 - \text{SSE/SST})$.

Source	DF	SS	MS	F	P
Regression	1	739.6	739.6	13.35	< 0.01
Error	13	720.1	55.395		
Total	14	1459.7			

We computed SSE $= 7.44277^2 \cdot 13$, $F = 739.6/55.395$, and for the $F(1, 13)$-distribution the relevant percentiles from Table E of IPS are $F_{.99}(1, 12) = 9.33$ and $F_{.999}(1, 12) = 18.64$. We conclude that the regression model is clearly significant; that is, the hypothesis $\beta_1 = 0$ is clearly rejected (against a two-sided alternative).

Subquestion d)

First, the predicted values from the regression model are obtained by inserting 30, 45 and 60 minutes into the estimated regression equation:

$$\hat{y}_{30} = 157.7 - 0.5733 \cdot 30 = 140.5;\ \hat{y}_{45} = 157.7 - 0.5733 \cdot 45 = 131.9;\ \hat{y}_{60} = 123.3.$$

Next, the margin of error for the 95% confidence intervals from the one-way ANOVA was already computed in **a)**, the value is 7.43. It is clear that all the predicted values are very well within the confidence intervals; the predicted values are all quite close to the group means. This means that the linear regression model does not give a substantially poorer fit to the data than the one-way ANOVA model.

Other ways we could have reached this conclusion are by comparing the residual standard deviations of the two models (7.625 vs. 7.443) or the R^2-value of the two models (52.2% vs. 50.7%). For both statistics, the values are quite close. We can therefore conclude that the fits of the two models to the data are quite similar. The linear regression model has the advantages of including one parameter less and of enabling predictions at other delays than those included in the study. On these grounds one might prefer the linear regression model, if the assumed linear association was biologically meaningful.