Biostats VHM 801 Course
Fall 2023, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to Final Exam, December 2023

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. Some parts of Questions 1 and 3 were waived in the marking.

Question 1

The data are from the VER2 textbook (Dohoo *et al.* (2009), *Veterinary Epidemiological Research*, 2nd edition) used in the epidemiology courses at AVC. The study is published in Sanchez *et al.* (2002), Evaluation of the repeatability of a crude adult indirect *Ostertagia ostertagi* ELISA and methods of expressing test results, *Veterinary Parasitology* **109**, 75−90.

Subquestion a)

It seems reasonable to assume the milk samples for the 40 cows to form a simple random sample (or to be i.i.d.). The distributions can be described as follows:

- both distributions appear as unimodal and somewhat left-skewed, more strongly so for the averaged OD values, and this left-skewness together with the fairly large sample size probably also explains the deviations from normality detected ($P = 0.012$) by the Anderson-Darling test for the averaged OD values and indicated ($P = 0.112$) for the individual OD values,
- the mean, median and the central part of the distribution (e.g., the box formed by Q1 and Q3) are quite similar for the two variables,
- the standard deviation, the IQR and the range are a bit smaller for the averaged OD values, which therefore overall can be said to be slightly less variable,
- neither distribution shows any suspected outliers (by the $1.5 \times$ IQR rule).

Subquestion b)

Two options for estimating the proportion of cows with OD values above 0.45 are calculation from a normal distribution and direct estimation from the sample. A normal distribution calculation should not be used for the averaged OD values with the stronger skewness and formal evidence against a normal distribution. Direct estimation from the sample requires us to determine how many of the 40 values are less than 0.45. The dotplot shows (as far as we can determine from the resolution) that both samples have 10 values below (and hence 30 values above) 0.45; therefore, the sample proportions are $30/40 = 0.75$. The normal distribution calculation for the individual OD values goes as follows, using the sample mean and standard deviation,

$$\mathrm{P}(X > 0.45) = \mathrm{P}\left(\frac{X - 0.6534}{0.2541} > \frac{0.45 - 0.6534}{0.2541}\right) = \mathrm{P}(Z > -0.80) = 0.7881 \approx 0.79.$$

Subquestion c)

The center of a distribution can be represented by the mean or median, and the Minitab displays include 95% confidence intervals (CIs) for the mean and median of each distribution. Both distributions show evidence against a normal distribution, which may affect the validity of the CI for the mean.

On the contrary, the CI for the median does not rely on distributional assumptions. According to the guidelines for the use of t-distribution procedures, the large sample size and the relatively normal shape of the distribution for the raw cv (coefficient of variation) should make the CI for the mean okay to use, but the quite strong outliers in the distribution for the adjusted cv makes that interval more questionable, and one should consider the median instead. The median also more meaningfully reflects the center in a strongly right-skewed distribution.

Neither distribution has its center and its corresponding CI anywhere close to 5%, so the cv cannot be labelled as "(very) good". The adjusted cv has its median close to 10% and the corresponding CI includes 10% as well as values above 10%. We are therefore *not* confident that the center is less than 10%, corresponding to a good or acceptable cv level, but it is around that level. The raw cv has its center well above 10%, and both the mean and median are close to 15% with their CIs including 15% as well as larger values. We can therefore not rule out that the center of the raw cv distribution is above 15%, and hence "critically high", and the raw cv is certainly not "good or acceptable".

Subquestion d)

We saw in **c)** that the center of the raw cv distribution is higher than for the adjusted cv distribution; this agrees with the expectation that adjusting the OD values could lead to lower variability among repeated samples. The adjusted cv distribution has however a very long right tail with two suspected outliers, which are both larger in value than all values from the raw cv distribution. This contributes to a larger standard deviation in the adjusted cv distribution, which is most likely not desirable. So although the adjusted cv values appear lower overall, there is also an increased variability and apparent susceptibility to extreme values for the adjusted values.

For statistical inference to compare the two distributions, we must be aware that they are computed from the same 40 samples. Hence the values cannot be assumed as independent and should instead be treated as two paired samples. We would need the differences or at least the individual pairs to carry out statistical inference to compare the distributions. The differences between the means and medians can be computed from the distributions, but we cannot attach any standard errors or confidence intervals to those differences from the information available.

Subquestion e)

Only the directly estimated proportion of values above 0.45 allows easy calculation of a confidence interval; it is more difficult to incorporate the uncertainty of the estimated parameters in the normal distribution into the calculation. The direct estimation of the proportion corresponds to a binomial model $\text{Bin}(40, p)$. The classical (normal approximation) confidence interval does not apply here because there are only 10 "negative" samples when at least 15 is required. The plus-four adjusted proportion is $\tilde{p} = (30 + 2)/(40 + 4) = 32/44 = 0.7273$. For a 95% CI we will use $z^* = 1.96$, and the calculation goes as follows:

$$\begin{aligned} 95\%\text{ CI:} &= \tilde{p} \pm 1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} = 0.7273 \pm 1.96\sqrt{\frac{0.7273 \cdot (1-0.7273)}{44}} \\ &= 0.7273 \pm 0.1360 = (0.591, 0.863). \end{aligned}$$

For part ii), we note that if the six test results for the same sample were independent of each other, then the average would have a smaller standard deviation according to the formula,

$$\text{sd}(\bar{X}) = \sigma/\sqrt{n} \approx s/\sqrt{6} = 0.2541/\sqrt{6} = 0.10.$$

In fact, the standard deviation for the averaged values was only a bit smaller than for the individual values, at 0.2390 compared to 0.2541. The central limit theorem also tells us that the distribution of the average should be closer to normal than of the individual values; our analysis in **a)** indicated

the opposite to be the case. Therefore, the data do *not* suggest that the six repeated test values are independent of each other. Intuitively, the values would also be expected to be dependent.

Question 2

The question is based on the paper Sherif *et al.* (2004), Detection of cotinine in neonate meconium as a marker for nicotine exposure in utero, *Eastern Mediterranean Health Journal* **10**, 96−105. Exercise 27.45 of PSLS 3e is based on the same data.

Subquestion a)

The study is observational because no treatments were imposed. The study design is three independent samples.

Subquestion b)

The statistical model used in the paper is a one-way ANOVA (a two-way ANOVA would make no sense because the mothers are just coded 1−10 in each group but there are 30 distinct mothers):

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

where Y_{ij} is the cotinine concentration for mother j in smoker group i, $i = 1, 2, 3$ and $j = 1, \ldots, 10$. The errors (ε_{ij}) are assumed independent and distributed $\sim \mathrm{N}(0, \sigma)$. The test for equal means across smoking groups, $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3$, is computed at $F = 10.45$ which in $F \sim F(2, 27)$ corresponds to $P < 0.001$ (because $F_{.999}(2, 27) = 9.02$ in Table E). The Minitab print gives $P = 0.000$, meaning that $P < 0.0005$. The statement $P = 0.01$ in the paper is wrong and misleading by understating the significance.

Only the (unadjusted) pairwise comparison between active smokers and non-smokers can be established as significant from the Minitab display because the confidence intervals do not overlap. For the two other comparisons we compute $\mathrm{LSD} = t^* s_p \sqrt{2/n} = 2.052 \cdot 89.42\sqrt{2/10} = 82.06$, using $t^* = t_{.975}(27) = 2.052$. It is seen that the difference between the means of the active smoker and passive smoker groups exceeds this value, whereas the corresponding difference between passive smokers and non-smokers does not. Therefore, the stated significances in the paper are correct for non-adjusted pairwise comparisons. This is considered a satisfactory answer. Bonferroni adjustement for multiple comparisons involves adjustment for $3 \cdot 2/2 = 3$ pairs. The calculation would require Minitab access to either use a different $t^* = t_{.992}(27)$ or the Comparisons menu with the model being fitted as a General Linear Model. The pairwise t-tests can also be calculated manually:

$$1 \text{ vs. } 2: \; t = \frac{367.2 - 263.4}{89.42\sqrt{2/10}} = 2.596, \quad 1 \text{ vs. } 3: \; t = \frac{367.2 - 185}{89.42\sqrt{2/10}} = 4.556.$$

As $t_{.99}(27) = 2.473$, the unadjusted P-value for the first comparison has $P < 0.01$ and therefore the Bonferroni-adjusted P-value satisfies $P < 3 \cdot 0.01 = 0.03$. The second comparison is strongly significant as is quite obvious from the means and confidence intervals. We conclude that a Bonferroni-adjustment for pairwise comparisons does not change the conclusion, and this method could also have been used in the paper. Note that the comparison 2 vs. 3 was already non-significant before any adjustment.

Subquestion c)

Three types of invalidation of model assumptions can be detected from the information provided:

- unequal variances: the ratio between largest and smallest group standard deviations ($143.7/24.23 = 5.9$) clearly exceeds the textbook guideline of 2. Implications: the overall F-test should remain significant because the difference between active smokers and non-smokers is very large (relative to the variation). Assuming same variances may however affect pairwise comparisons between groups. For example, the comparison between active and passive smokers may not really be significant (the pairwise t-test based on these two groups only yields: $t = (367.2 - 263.4)/\sqrt{143.7^2/10 + 52.5^2/10} = 2.15$; possibly significant, depending on the degrees of freedom, but in any case less strongly), and the comparison between passive smokers and non-smokers may be significant (pairwise t-test based on these two groups only: $t = (263.4 - 185)/\sqrt{52.5^2/10 + 24.23^2/10} = 4.3$ (very clearly significant);

- outliers: the residual plots shows one fairly large outlier to the right: observation 700 in the active smoker group (a suspected outlier based on the IQR rule: $436 + 1.5(436 - 264) = 694$, but only barely so). Implications: analyzing without this observation will decrease both the mean and spread for active smokers, but probably not enough to meet the equal variances requirement, and the impact for statistical significance may be fairly limited (strongest on pairwise comparison active vs. passive smokers);

- distribution shape: even without the outlier, some right-skewness may remain in the active smokers group (based on Q1 and Q3), and there is some left-skewness in the passive smokers group. So the data may not approximate normal distributions well within the groups. Implications: difficult to assess but most likely only minor impact on significances.

Subquestion d)

Two approaches can be suggested for analysis of the data: a non-parametric analysis using Kruskal-Wallis test and transformation of the outcome.

The rank-based Kruskal-Wallis test does not assume normal distributions for the within-group distributions and does not assume equal variances, and hence avoids the problems with model assumptions discussed above. Pairwise comparisons after the (significant) Kruskal-Wallis test must be done by the Mann-Whitney-Wilcoxon two-sample test for comparing two independent samples. A Bonferroni correction adjustment for multiple testing is possible, using the approach referred to in Exercise 15.47 of IPS 7e. Because the three distributions are very different in shape (for example by their different standard deviations), the assumption about equal shapes should *not* be made, and the conclusions can therefore not be phrased in terms of medians.

Transformation to achieve approximately equal variances and approximately normal within-group distributions may be possible because the standard deviation is incresing with the mean (across the 3 groups). A log-transformation is suggested but other transformations may be considered if this does not work (additional analysis will show that the log-transformation is ineffective; an inverse transformation ($1/Y$) achieves approximate variance homogeneity but problems with extreme residuals remain, now in the left tail). Assuming that a transformation can be found that offers reasonable compliance with model assumptions on transformed scale, the analysis for statistical significance proceeds in a similar way as in **b)** with the overall F-test and pairwise comparisons between groups (possibly adjusted for multiple comparisons).

Question 3

The data for the reductions in systolic and diastolic blood pressures are from Siddiqui *et al.* (2020), Reserpine substantially lowers blood pressure in patients with refractory hypertension: A proof-of-concept study, *American Journal of Hypertension* **33**, 741−747.

Subquestion a)
The data are measurements before and after treatment on the same subject, and therefore a paired samples design. Denote by X_i and Y_i the diastolic blood pressure for patient i, $i = 1, \ldots, 6$, before and after treatment, respectively. Let further $D_i = X_i - Y_i$, the drop in blood pressure. The statistical model (Model I) is that the differences $D_1, \ldots, D_6$ are i.i.d. and $\sim \mathrm{N}(\mu, \sigma)$. The parameter μ is the mean drop in blood pressure, and $\hat{\mu} = 22.0$ with $\mathrm{SE}(\hat{\mu}) = 6.44$ (also computable as: $15.73/\sqrt{6} = 6.44$). To assess whether the data give evidence of a treatment effect we test the null hypothesis $\mathrm{H}_0 : \mu = 0$. It is most natural to use a one-sided alternative $\mathrm{H}_a : \mu > 0$ because the focus on lowering the blood pressure. The test statistic is $t = \hat{\mu}/\mathrm{SE}(\hat{\mu}) = 3.42$ which corresponds to $0.005 < P < 0.01$ in a $t(5)$-distribution. There is clear evidence ($P < 0.01$) against H_0, suggesting that the treatment *is* effective. A 95%-confidence interval for the blood pressure reduction is: $22.0 \pm 2.571 \cdot 6.44 = 22.0 \pm 16.6 = (5.4, 38.6)$.

Subquestion b)
The Pearson correlation coefficient quantifies the linear association between two continuous variables. The treatment effect is represented by the reduction (or drop) in blood pressure. The table of correlations between all variables includes the value $r = 0.916$ for the correlation between the reductions in systolic and diastolic blood pressures. For statistical inference we could request a confidence interval from the Correlations menu in Minitab; this interval is not computable from the information provided. We can however compute a t-test for $\mathrm{H}_0 : \rho = 0$ against a one-or two-sided alternative (either can be argued in this context). We calculate,

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.916\sqrt{\frac{6-2}{1-0.916^2}} = 4.57,$$

which corresponds to $0.01 < P < 0.02$ against a two-sided alternative in a $t(4)$-distribution. Because $t_{.995}(4) = 4.604$, we must have $P \approx 0.01$. We conclude that despite the small sample size there is evidence of a positive correlation in the systolic and diastolic blood pressure reductions.

Subquestion c)
The two models are linear regressions for the post-treatment value and drop:

$$\begin{aligned} \text{(II)}: \quad Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i, \\ \text{(III)}: \quad D_i &= \gamma_0 + \gamma_1 X_i + \epsilon_i, \end{aligned}$$

where in both models the errors are assumed i.i.d. and $\sim \mathrm{N}(0, \sigma)$. As the pre-treatment diastolic blood pressure never takes the value zero, the intercepts (β_0 and γ_0) have no intuitive interpretations on their own. The slopes (β_1 and γ_1) are the "effects" of the pre-treatment blood pressure on the post-treatment blood pressure and the drop in blood pressure, respectively. The σ is the standard deviation about the line. The listings give the following estimates:

$$\begin{aligned} \text{(II)}: \quad & \hat{\beta}_0 = 50.62, \ \hat{\beta}_1 = 0.2738, \ \hat{\sigma} = 9.735, \\ \text{(III)}: \quad & \hat{\gamma}_0 = -50.62, \ \hat{\gamma}_1 = 0.7262, \ \hat{\sigma} = 9.735. \end{aligned}$$

For the 95% confidence intervals for the regression parameters, we use $t^* = 2.776$ from the $t(4)$-distribution. Because the models are directly related (to be explained further below), only the confidence intervals for Model II are included here:

$$\begin{aligned} \beta_0 &: \ 50.62 \pm 2.776 \cdot 24.4 = 50.62 \pm 67.73 = (-17.1, 118.4), \\ \beta_1 &: \ 0.2738 \pm 2.776 \cdot 0.240 = 0.2738 \pm 0.6662 = (0.39, 0.94). \end{aligned}$$

From $D_i = X_i - Y_i$ we see that model III can be written: $X_i - Y_i = \gamma_0 + \gamma_1 X_i + \epsilon_i$, or by rearrangement of the equation: $Y_i = -\gamma_0 + (1-\gamma_1)X_i - \epsilon_i$. Therefore, the regression parameters can be identified as: $\beta_0 = -\gamma_0$ and $\beta_1 = 1 - \gamma_1$. The errors are merely switched in sign ($\varepsilon_i = -\epsilon_i$), and therefore the standard deviation about the line is the same for the two models. The fit of the two models is the same, and the fitted line plots do not indicate any systematic model violations.

In model II, the hypothesis $\mathrm{H}_0 : \beta_1 = 0$ corresponds to no association (correlation) between the pre- and post-treatment values. It may be a bit surprising that the data provide no evidence of such an association ($P = 0.318$), but the estimated slope is not that large and it is a small dataset. In model III, the hypothesis $\mathrm{H}_0 : \gamma_1 = 0$ corresponds to no association between the drop in blood pressure and the pre-treatment value. With a numerically larger value of the slope, we get a weak significance ($P = 0.039$) against this hypothesis. It does seem plausible that the drop in blood pressure could depend on the pre-treatment value, and the positive slope means that higher pre-treatment values are assosiated with larger reductions.

Subquestion d)

The wording of the question gives some freedom to the interpretation of the 95% interval. It is for a new patient and would therefore most naturally be considered a prediction interval, but it could also be interpreted as a confidence interval for the expected value of the patient. The table below gives both of these intervals for the three models.

	Model I	Model II	Model III
estimate	88.0	80.8	80.8
confidence interval	(71.4,104.6)	(67.8,93.6)	(67.8,93.6)
prediction interval	(44.2,131.8)	(50.8,110.7)	(50.8,110.7)

The estimate and intervals for Model II are given directly in the computer listing. The estimate and confidence interval for Model I was obtained by subtracting the estimate and interval from **a)** from 110. The intervals for Model III were obtained similarly by subtracting the intervals of the listing from 110. The prediction interval for model I was obtained from the following formula (which is not in the course): $\hat{\mu} \pm t^* s\sqrt{1 + (1/6)}$; subsequently, this interval was subtracted from 110. Three observations are made. One is that estimate of model I is quite a bit higher than the others. Two is that the estimates and intervals for models II and III are identical; this is because the two models give essentially the same fit to the data. Three is that the intervals from model I are wider than those of model II/III.

Subquestion e)

Model I allows us to directly estimate the treatment effect and assess its significance. Model II is for the post-treatment blood pressure and does not allow us to compute treatment effects. Model III is for the drop in blood pressure but the model is more general than model I because it allows the treatment effect to depend on the pre-treatment level. For $\gamma_1 = 0$, model III becomes model I. However, we previously noted some evidence against that hypothesis. It must therefore be concluded that model III is more accurate than model I. This is the reason for the wider intervals of model I in c). Model III is more useful than model II because it allows assessment of the treatment effect (for any given pre-treatment blood pressure). The fact that R^2 is (much) higher for model III than model II is not important in itself because R^2-values cannot be compared between different outcomes anyway. As the errors are the same, it is the change in outcome that causes the different R^2-values.