Biostats VHM 801 Course
Fall 2022, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to Final Exam, December 2022

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. The requirements for a 100% score were lowered slightly for some parts of the questions.

Question 1

Subquestion a)

It is a retrospective observational study. The basic statistical design is two paired samples, consisting of the observations on Friday 6th and Friday 13th. This structure is in particular clear for the Drivers data. The pairs arise because the observations are taken on days that are close in time, thus e.g. in the same season of the year. One could expect differences across the year related to weather and season-specific activities that people might be engaged in. Therefore, the two observations one week apart would be expected to be more similar than on any two days where one is a Friday 6th and the other is a Friday 13th. For the Shoppers data, the counts have been totalled across dates, but instead the data has paired samples for each store. The Accidents data are only described (here) in terms of totals, for which the paired nature of the underlying data is more or less gone.

As stated in the published paper, the population studied broadly corresponded to people living in the study area. The statistical inference may be valid for a wider population (e.g. British, urban, with similar socio-economic and ethnic composition).

Subquestion b)

The nine counts of shoppers on the two Fridays should be considered as two paired samples (as discussed above). The counts are quantitative observations, so the statistical analysis most naturally focuses on the difference between the counts in each pair. If we let $D_1, \ldots, D_9$ denote the counts of shoppers on the 13th minus the corresponding counts on the 6th, our statistical model is that the D_i constitute a simple random sample and are assumed i.i.d. (independent and identically distributed). A parametric analysis would assume a normal distribution $\mathrm{N}(\mu_D, \sigma_D)$ for the differences (see below for discussion of this assumption). We estimate the parameters by the sample statistics, and test the hypothesis $\mathrm{H}_0 : \mu_D = 0$ (corresponding to no mean difference between the counts of shoppers on the 6th and 13th) against a two-sided alternative by a t-test:

$$\hat{\mu}_D = \bar{D} = 232.4,\ \hat{\sigma}_D = s_D = 198.2,\ t = \frac{\bar{D}}{s_D/\sqrt{9}} = 3.52,\ P = 2 \cdot \mathrm{P}(t_8 > 3.52) < 2 \cdot 0.005 = 0.01.$$

The t-test statistic is significant at the 1% significance level, and thus provides clear evidence against the null hypothesis. We conclude that the number of shoppers is larger on the 13th; the estimated difference is 232 persons. Compared to the total number of shoppers (on the 13th), this is a actually a small relative difference: $232/25085 = 0.93\%$. Also the 95% CI for the mean total count difference is quite narrow.

It is difficult to assess whether the normal distribution is appropriate from a sample of only 9 observations. The graphical summary for the variable `diff13min6` shows a unimodal and roughly symmetrical distribution, and the A-D normality showes absolutely no evidence against a normal

distribution ($P = 0.88$). At least it could be said that the distribution at hand does not show any indications of non-normality. A normal distribution could also be motivated from the central limit theorem, because the counts (and differences) are summed up across a number of dates; such a summation would generally tend to create a more normal distribution. In reaction to the statement of the paper, it is a cautious approach to use a non-parametric procedure when the data do not allow assessment of normality with adequate power to detect deviations, but the statement could be read to imply that the data showed indications of non-normality, when in fact the choice was probably motivated by this generally cautious approach. Therefore, the statement is perhaps not well worded.

Subquestion c)

The statistical model is again that the differences $D_1, \ldots, D_5$ constitute a sample of i.i.d. quantitative observations. The non-parametric test that can be computed manually is the sign test, whereas the alternative Wilcoxon signed rank test requires computer analysis. The null hypothesis is H_0 : median$(D) = 0$, corresponding to the equal probabilities of the 6th value being the larger value, and of the 13th value being the larger. The analysis is based on the count (X) of the number of dates at which the 6th value is larger (one could also consider the count for when the 13th value is larger). We assume $X \sim \mathrm{B}(5, p)$ and will test the hypothesis $\mathrm{H}_0 : p = 0.5$ against the alternative $\mathrm{H}_a : p \neq 0.5$. The two-sided alternative seems most natural here as there is no particular focus on one alternative. The statistical analysis has the following steps:

$$\hat{p} = 0/5 = 0, \; P = 2 \cdot \mathrm{P}(X \leq 0) = 2 \cdot 0.03125 = 0.0625,$$

where the P-value is computed directly from the $\mathrm{B}(5, 0.5)$ distribution (in Table 1 of Stevens), or as $P(X = 0) = 0.5^5 = 0.03125$. As $P > 0.05$ there is no formal evidence against the null hypothesis, so we cannot reject that the two Friday counts are equally likely to be the largest. As P is close to 0.05, we could say that there is an indication that traffic is higher on the 6th than the 13th. The paper actually reported a P-value less than 0.05, which could be obtained by using a one-sided alternative (here, $\mathrm{H}_a : p < 0.5$). It is not clear whether a one-sided alternative hypothesis is warranted.

Subquestion d)

If a total of $n = 110$ accidents happened on the two Fridays, the proportion of these that happened on the 13th would follow a binomial distribution $\mathrm{B}(n, p)$. The confidence interval can be computed using the classical, normal approximation (because the sample easily includes more than 15 events and 15 non-events):

$$\hat{p} = 65/110 = 0.591, \; 95\% \text{ CI: } \hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/110} = 0.591 \pm 0.092 = (0.499, 0.683).$$

The confidence interval barely includes $p = 0.5$ (corresponding to same probabilities on the two days), so there is no formal evidence (although it is very close) that the probability of transport accidents happening on the 13th is larger than accidents happening on the 6th. We could say that the data show some indication of larger risk of transport accidents on the 13th than on the 6th. As above, using a one-sided alternative hypothesis could bring the P-value below 0.05 (as was done in the paper).

Subquestion e)

If working under the assumption that accidents involving animals are equally likely on the two Fridays, the count of such accidents that occured on a Friday 13th should follow a binomial distribution, say $X \sim \mathrm{B}(4, 0.5)$. The statistical table for the binomial distribution then gives:

$$P(X > 2) = P(X = 3) + P(X = 4) + 0.250 + 0.063 = 0.313.$$

The probabilities can also be computed manually, e.g. $P(X = 4) = 0.5^4 = 0.0625$.

Question 2

Before answering the question, we reflect on the experimental study design. The experimental unit is a plant, the spraying treatments (including control) is the treatment factor and controlled, and the presence or absence of lesions is the response (outcome). The variable `plants` holds the total number of plants with the same response for each treatment.

Subquestion a)

The first analysis (top left) is a simple linear regression with plants as the outcome (y-variable) and treatments as the predictor (x-variable). Just focusing on the x-variable, it is completely meaningless to predict with a linear equation involving treatment numbers because these are mere labels with no real meaning. Treatment is a categorical variable and cannot be used for prediction in regression. The second analysis (top right) is a one-way ANOVA with treatments as groups, so the treatments are dealt with appropriately. The outcome, however, is not meaningful because the two counts (plants with and without lesions) are represented as replicates within each treatment, while they actually mean completely different things (i.e., lesions and non-lesions); this issue also exists with the regression analysis. One would want to look at how many plants had lesions out of the total or compare the number of plants with and without lesions, but the one-way ANOVA does not do that. Finally, the third analysis is a two-way table analysis with a X^2-test. This analysis is for two categorical variables and counts of observations in each of the combined categories, and that is indeed what we have. So this is a meaningful analysis, to be further discussed in **b)**.

Subquestion b)

Denote by N_{ij} the number of plants subjected to treatment j that were classified into lesion group i, where $i =$ no, yes and $j = 1, \ldots, 7$. The outcome (or response) in the study is the presence or absence of lesions at the plants. The treatments were imposed by the experimentator and is therefore an explanatory variable. Within each treatment group, a certain number of plants were inspected − these correspond to a binomial setting. Thus we assume a binomial distribution in each column: $N_{1j} \sim \text{Bin}(N_j, p_j)$, where p_j is the probability of lesions and $N_j = N_{0j} + N_{1j}$ is the total number of plants in the jth treatment group. The binomial settings involve the primary assumptions:

- same probability p_j of lesions for all plants in the group,
- independence of lesions between different plants.

In addition, we assume the outcomes for the different treatments to be independent. Overall this is model I: independent binomials over the columns (or model for comparing several populations).

The hypothesis of interest is H_0: $p_1 = \ldots = p_7$, same probability of lesions for the 7 treatments. We test the hypothesis against an unspecified alternative (H_a: some differences between p_j's) using a Pearson chi-square statistic:

$$X^2 = 15.466, \quad \text{df} = 6, \quad P = 0.017.$$

The test is significant at the 5% level and close to significant at the 1% level. Thus, there is a moderate significance against H_0, and we can be pretty sure that some differences between treatments exist. We also note that all expected values in the cells are above 5, so that we can safely use the test. To quantify the treatment effects, we estimate the probability of lesions within each group: $\hat{p}_j = N_{1j}/N_j$, see the table below in which the treatments are ordered by estimated proportions of plants lesions.

Statistic	Formula	Treatment j						
		1	4	5	7	3	6	2
plants with lesions	N_{1j}	17	11	15	29	11	22	7
group size	$N_j = N_{0j} + N_{1j}$	20	13	24	54	22	48	17
proportion of lesions	$\hat{p}_j = N_{1j}/N_j$	0.850	0.846	0.625	0.537	0.500	0.458	0.412
standard error	$\sqrt{\hat{p}_j(1-\hat{p}_j)/N_j}$	0.080	0.100	0.099	0.068	0.107	0.072	0.119

The table shows that the control group (treatment 1) does indeed do worst but that treatment 4 is only marginally better. Treatments 6 and 2 show the lowest proportions of plants with lesions.

Subquestion c)

The problem of comparing treatment levels after an ANOVA table has been extensively discussed in the course (and is explained in the textbook), but we have not discussed treatment comparisons after a chi-square test much. The simplest method (which should *not* be used after the ANOVA, but is valid for the chi-square test) is to carry out separate analyses for each comparison. For example, to compare treatments 1 and 2, restrict attention to these two groups and carry out a 2-sample z-test for proportions (or any other method valid to compare two proportions, such as inference based on separate CIs in 2 out of 3 scenarios). To achieve an overall significance level of at least 5%, we can use the Bonferroni method and compare against an adjusted significance level for each comparison of $5\%/6 = 0.83\% = 0.083$; there are in total 6 comparisons against the control. The standard errors in the table give an impression of which groups might turn out to be significant; classical (plus four would be better, but we want to keep it simple) 95% confidence intervals would have a margin of error of $1.96 \times \mathrm{SE}(\hat{p}_j)$. Only the intervals for treatments 6 and 2 are non-overlapping with the one for treatment group 1. We therefore start out with the z-test for treatment 2, and proceed downwards:

- treatment 2: $z = (\hat{p}_1 - \hat{p}_2)/\sqrt{p_0(1-p_0)(1/20+1/17)} = 2.78$ (using the pooled proportion $p_0 = (17+7)/(20+17) = 0.649$), $P = 2 \times P(z > 2.78) = 2 \times 0.0027 = 0.0054$,
- treatment 6: $z = (\hat{p}_1 - \hat{p}_6)/\sqrt{p_0(1-p_0)(1/20+1/48)} = 2.98$ (using $p_0 = (17+22)/(20+48) = 0.574$), $P = 2 \times P(z > 2.98) = 2 \times 0.0014 = 0.0028$,
- treatment 3: $z = (\hat{p}_1 - \hat{p}_3)/\sqrt{p_0(1-p_0)(1/20+1/22)} = 2.40$ (using $p_0 = (17+11)/(20+22) = 0.667$), $P = 2 \times P(z > 2.40) = 2 \times 0.0082 = 0.016$,
- treatment 7: $z = (\hat{p}_1 - \hat{p}_7)/\sqrt{p_0(1-p_0)(1/20+1/54)} = 2.06$ (using $p_0 = (17+29)/(20+54) = 0.622$), $P = 6 \times 2 \times P(z > 2.47) = 2 \times 0.0068 = 0.014$.

Note that all z-tests meet their conditions of expected counts of plants with and without treatment being above 5 ($p_0 N_j$ and $(1-p_0)N_j$, respectively, in both the treatment and control groups). As the comparisons with both treatment 3 and 7 turned out to be non-significant at the Bonferroni-adjusted significance level, it is easy to guess that the same happens for the comparisons with the two remaining treatments where the difference in proportions is much smaller. We conclude, there is evidence at a simultaneous 5% error level that treatments 6 and 2 perform better than the control. Treatments 3, 7 and 5 show some improvement relative to control but not enough to qualify for statistical significance (at an overall 5% significance level), whereas treatment 4 performs more or less the same as the control.

Question 3

Before answering the question, we note that the two variables `f` and `w` can both be considered as response variables with random variation; neither of them are controlled, although one might suspect that the species of birds were selected to represent a broad range of bird sizes.

Subquestion a)

In the plots on logarithmic scale, the relation between the two variables is positive, strong and approximately linear (the points are close to a straight line). On original scale, the linear relation is less obvious, but otherwise the message is the same. The Pearson correlation coefficient r expresses the strength and direction of linear association between the variables; at $r = 0.99$, it is positive and close to its maximum value of 1. Independence corresponds to $r = 0$, and the 95% CI shows that 0 is a totally unreasonable value for the population correlation ρ; in other words, there must be very strong evidence against $\rho = 0$. The main difference between the plots on the two scales is that the points on logarithmic scale are more regularly spaced on both axes, whereas the points on original scale form a cluster close to zero ($\sim$ small birds) with some larger values and two very large values. As already mentioned, it could also be said that the relation looks less linear on original scale.

Subquestion b)

All four models are simple linear regression models, but they differ in their choice of response (y) and predictor (x) variables. The first two models have both variables `w` and `f` on original scale, but their roles as response and predictor have been switched between the models. The statement suggests to predict `f` from `w`, therefore we should have `f` as the response and `w` as the predictor. In the last two models, both variables are on log-transformed scale, and again with the roles of y and x switched. In choosing the preferred scale, we should look at where the model assumptions may be met to a reasonable degree. The residual plots on original scale show a cluster of points at the lower end, and the remaining points scatter more and more around the horizontal line at 0 as we move along the x-axis ("cone" or "fan" shape plot). This is clearly indicative of unequal variability about the line. Even without the fitted line overlaid on the scatterplots of `f` and `w`, it seems plausible that the points are very close to the line around 0 and scatter more about the line for larger values of x. This would be a clear violation of the assumption of equal error variances, and effectively rules out the models on original scale. Note that R^2-values on different scales are not comparable, so R^2 cannot be used to select the best model.

The last model in the listing, with `lnf` as the response and `lnw` as the predictor, is therefore the best choice among those shown, and the residual plots looks reasonably okay (the low value of y for the lowest value of x may be a concern, but the standardized residual is only slightly below -2). The model can be written as,

$$\texttt{lnf}_i = \beta_0 + \beta_1 \texttt{lnw}_i + \varepsilon_i, \quad i = 1, \ldots, 14,$$

where the errors ε_i are assumed independent and $\sim \mathrm{N}(0, \sigma)$.

Subquestion c)

Estimates and 95% confidence intervals for the regression parameters (using $t^* = t_{.975}(12) = 2.179$):

$$\begin{aligned}
\text{intercept}: &\quad \hat{\beta}_0 = -1.577, \quad 95\%\ \text{CI}: -1.577 \pm 2.179 \cdot 0.219 = -1.577 \pm 0.477 = (-2.054, -1.100),\\
\text{slope}: &\quad \hat{\beta}_1 = 0.9742, \quad 95\%\ \text{CI}: 0.9742 \pm 2.179 \cdot 0.0377 = 0.9742 \pm 0.0821 = (0.8921, 1.0563),\\
\text{stand. dev}: &\quad s = 0.3247
\end{aligned}$$

The intercept and slope are the two parameters determining the linear relation between frequency and inverse height, whereas the standard deviation is of the vertical deviations about the line. The predictive power of the model is expressed by $R^2 = r^2 = 98.24\%$, the proportion of variation explained by the model. The model has high predictive ability, and the points scatter quite closely to the line. There is very strong evidence against $\beta_1 = 0$, both by $t = 25.86$ and $F = 668.7$. In part **a)**, we already noted the high value for r, the strong significance (the tests are the same) and the points being close to the line.

Subquestion d)

We can use the estimated equation to predict `lnf` for a given value of `lnw`. For `w` $= 2.7$ we get `lnw` $= 0.993$, and

$$\widehat{\texttt{lnf}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0.993 = -1.577 + 0.9742 \cdot 0.993 = -0.61,$$

which in turn corresponds to $\hat{\texttt{f}} = \exp(-0.61) = 0.54$. This value is substantially lower than the actual weight (1.0). To statistically assess whether this species matches the estimated relationship, we would need a prediction interval for our prediction on logarithmic scale, and we would then base our inference on whether the observed value on log-scale ($\ln(1.0) = 0$) falls inside the prediction interval. If yes, the species is in agreement with out estimated model. We note that the prediction is for a value of `lnw` outside the range in the data, i.e. an extrapolation. Extrapolations are a potential concern because we do not know whether the linear relation will hold.

Subquestion e)

The equation $\texttt{lnf} = \ln(k) + \texttt{lnw}$ corresponds to taking the slope $\beta_1 = 1$ in the regression model above (and shifting notation for the intercept: $\beta_0 = \ln(k)$, or $k = \exp(\beta_0)$). Therefore, we can test the validity of the estimation approach by testing $\mathrm{H}_0 : \beta_1 = 1$ against a two-sided alternative $\mathrm{H}_a : \beta_1 \neq 1$. The 95% CI for β_1 from **c)** includes 1 and therefore tells us that there is no significance against H_0, i.e. $P > 0.05$. We can also compute the t-test: $t = (\hat{\beta}_1 - 1)/\mathrm{SE}(\hat{\beta}_1) = (0.9742 - 1)/0.0377 = -0.68$, which is far from significance in the $t(14)$-distribution. We conclude that the estimated regression line seems to agree with the proportionality relation (from the text). Further,

$$\begin{aligned} \hat{k} &= \exp(\beta_0) = \exp(-1.577) = 0.207 \approx 0.21, \\ 95\%\ \mathrm{CI} &: (\exp(-2.054), \exp(-1.100)) = (0.128, 0.333) \approx (0.13, 0.33). \end{aligned}$$

Therefore our model estimates a proportionality factor of 21% with a 95% ranging from 13% to 33%. We could actually improve on this estimation if we re-estimated the intercept in a model with the slope restricted to 1, but this calculation cannot be done from the information provided.