Biostats VHM 801 Course
Fall 2019, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to Final Exam, December 2019

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

Subquestion a)

The statistical design is sampling without replacement from a finite population (the batch of cheeses). As the population size ($N = 1000$) is more than 20 times that of the sample size ($n = 20$), we can use a binomial model. If X denotes the number of positive cheeses in the sample, we assume $X \sim \mathrm{B}(n, p)$. Our estimate of p is the observed proportion, $\hat{p} = X/n$. As the observed count is < 15 but $n > 10$, we use the "plus four" method for a confidence interval. The total number of positive cheeses in the population is estimated as $N\hat{p}$, and the corresponding confidence interval is obtained by multiplying the two endpoints of the interval for p by N. In summary,

$$
\begin{aligned}
\hat{p} &= X/n = 4/20 = 0.20, \\
\tilde{p} &= (X+2)/(n+4) = 6/24 = 0.25, \\
95\% \text{ CI for } p &: \quad \tilde{p} \pm z^* \sqrt{\tilde{p}(1-\tilde{p})/(n+4)} \\
&= 0.25 \pm 1.96\sqrt{0.25(1-0.25)/24} = 0.25 \pm 0.173 = (0.077, 0.423), \\
N\hat{p} &= 1000 \cdot 0.20 = 200, \\
95\% \text{ CI for } Np &: \quad N \cdot (0.077, 0.423) = (77, 423).
\end{aligned}
$$

Subquestion b)

The interest is in testing the null hypothesis $\mathrm{H}_0 : p = 0.05$. The most natural alternative hypothesis is one-sided, $\mathrm{H}_a : p > 0.05$, because the interest is in an *increased* presence (prevalence) of Listeria in the cheeses. As $np_0 = 20 \cdot 0.05 = 1 \ll 10$, we should not use the normal approximation for the test. An exact test based on the binomial distribution has $P = P(X_0 \geq 4)$ where $X_0 \sim \mathrm{B}(20, 0.05)$. Our binomial distribution table (from Stevens) gives $P = 0.013 + 0.002 + 0.000 + \ldots = 0.015$; the exact value is $P = 0.016$. Therefore, the test is significant and there is evidence against $p = 0.05$ in favour of an increased prevalence. In other words, the inspectors should indeed worry about an increased prevalence.

It is also possible to base the statistical assessment on the confidence interval from **a)**. As the target value, $p_0 = 0.05$, is outside the interval, we would have had significance at significance level 0.05 for a test with a two-sided alternative. For the one-sided alternative, and because the estimate is in the direction of the alternative, we can therefore conclude that $P < 0.025$.

Subquestion c)

We let X denote the wing length (in mm) of a male finch, and assume $X \sim \mathrm{N}(61.2, 1.8)$. Then

$$P(X > 65) = P\left(\frac{X - 61.2}{1.8} < \frac{65 - 61.2}{1.8}\right) = P(Z > 2.111) = P(Z < -2.111) \approx 0.0174,$$

using Table B of PSLS. The calculation tells us that wing lengths above 65 mm are relatively rare. In order to determine x so that $P(X > x) = 0.01$, we note that in $\mathrm{N}(0, 1)$ we have $P(Z > 2.33) \approx 0.01$ (from the same table), and compute

$$z = \frac{x - 61.2}{1.8} \quad \text{or} \quad x = 61.2 + 1.8z = 61.2 + 1.8 \cdot 2.33 = 65.4.$$

Subquestion d)

From the graphical summary and the stemplot we see that the distribution has a rather peculiar shape. It is definitely not normal (based on the normality test), despite a skewness very close to zero. The kurtosis is negative, hinting at the distribution not having proper tails, but even though this can perhaps be said about the left tail, this characterization seems little helpful. No potential outliers are seen. The most important feature of the distribution is that it has multiple peaks, and while the histogram may not allow us to assess clearly whether this could be due to chance, the stemplot shows a distribution separated into several parts. There is a distinct gap between observations in the range $11.4 - 13.8$ and those above. There may also be a gap between observations in the range $15.2 - 17.0$ and those above 18, with one intermediate observation (17.6). The best description of the distribution is therefore as multimodal (bimodal or trimodal).[1]

A multimodal distribution does not have a single centre, instead there would be centres for each part of the distribution. The overall centre of the distribution essentially reflects the proportions of the multiple parts of the distribution and is of limited interest as a central measure. Although a sample size of 61 observations would allow use of t-distribution procedures for a confidence interval for the mean, it is therefore not meaningful in this situation.

Question 2

Subquestion a)

The data layout is a two-way factorial with yield (X) as the outcome, presence/absence of O_3 and SO_2 as the two factors, and 3 replicates for each of the 4 treatment groups. The chambers were the experimental units (not the individual bean plants within each chamber), and in the completely randomized design these should have been assigned randomly to the 4 treatment groups. One possible randomization procedure would be to randomly permute (reorder) the chamber numbers $1 - 12$.

Subquestion b)

The natural statistical model is a two-way ANOVA model:

$$X_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

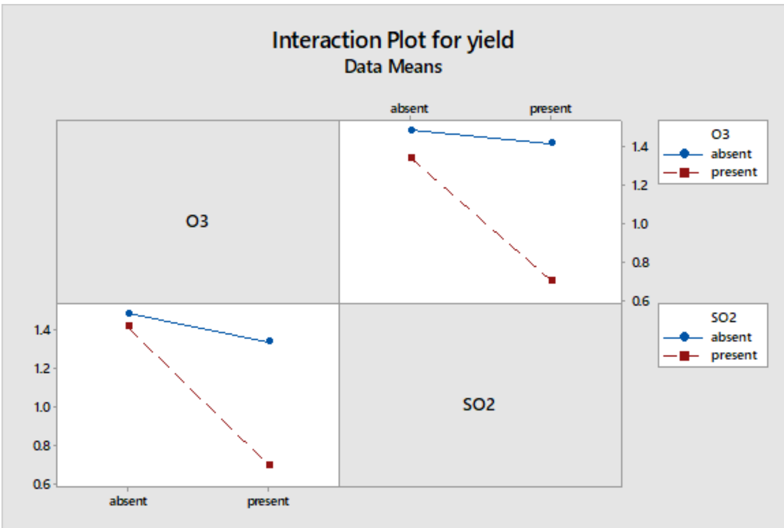
[1] The data are from a 2018 paper in *Nature Communications*; `doi:10.1038/s41467-018-07374-9`.

where $i = 0, 1 \sim$ absence/presence of O_3, $j = 0, 1 \sim$ absence/presence of SO_2, and $k = 1, 2, 3$. The errors ε_{ijk} are assumed i.i.d. and $\sim \mathrm{N}(0, \sigma)$.

Model checking utilizes the provided residual plot panel. The normal plot looks reasonably straight, and definitely fine for such a small dataset, and there are no extreme (beyond ± 2) standardized residuals. The plot against fitted values (of which there are only four distinct values, corresponding to the four treatment groups) shows some difference in the vertical variability and maybe some indication of fanning. This impression however comes mostly from the group with the lowest mean having also the smallest standard deviation. Because the group with the highest mean has the second-lowest standard deviation, there is no clear/strong increase in standard deviation with the mean, and standard variance-stabilizing transformations (such as square-root or log) will not be particularly effective. It is seen that the variance ratio guideline for ANOVAs is violated ($s_{\max}/s_{\min} = 0.2128/0.0557 = 3.8 \gg 1$), but with so few observations per group this may not be critical. It can be suggested to perform tests for equal variances (e.g. Levene's test; it turns out to be clearly non-significant). In summary, the concerns identified about the model assumptions are probably not so serious that they invalidate an ordinary two-way ANOVA analysis, but we may want to exercise caution with comparisons involving the group with the lowest standard deviation.

Subquestion c)

The natural graphical display for a factorial design to show the *effects* (not the raw data) is the interaction plot. It is not included in the Minitab listings, but one version of it can be sketched from the four treatment group means provided.



The two curves are clearly non-parallel, suggesting an interaction effect to be present. Both the graph and the estimates show the yield to be clearly lowest when both air pollutants are presents and otherwise pretty similar. The strength of this pattern will be evaluated by the statistical analysis.

Subquestion d)

The ANOVA table shows a significant ($P = 0.011$) interaction in the effect of the two air pollutants, which gives us evidence that *both* O_3 and SO_2 are of importance for the yield. The tests for the main effects are not needed for this conclusion, and these two tests (both strongly significant) are of less interest in this situation. The four group means indicate, as discussed above, that the presence of both O_3 and SO_2 leads to a decrease in yield, whereas the presence of just one of the pollutants has only a minor effect. The standard error of a group mean is $s/\sqrt{3} = 0.086$, and using $t^* = t_{.975}(8) = 2.306$ the 95% CIs for treatment levels have a margin of error of $2.306 \cdot 0.086 = 0.20$. This shows the lowest CI does not overlap with any of the three others, which in turn have their estimates within all intervals.

This will in itself allow a conclusion to be drawn, but we may further calculate the standard error of a difference of two treatment means as $s\sqrt{2/3} = 0.12$, so that $\text{LSD}_{0.95} = 2.306 \cdot 0.12 = 0.28$. In summary, without correcting for multiple comparisons, the differences between both pollutants present and the other groups are clearly significant, whereas there are no significant differences between the three other groups. Without access to software it is difficult to adjust exactly for the 6 multiple comparisons among the four treatment groups, but from $t_{.9975}(8) = 3.833$ we see that the Bonferroni-adjusted LSD-value would be less than $3.833 \cdot 0.12 = 0.46$, for which the comparisons still have the same significance. Thus, our conclusion supports the interpretation of the interaction plot, that the presence of both pollutants decreases yield, whereas the data do not show any notable effect of the presence of just one of them.

Question 3

Subquestion a)

The gas consumption is clearly a response variable, and the temperature may be either a response or explanatory variable, but it is clear from the description of the study that the interest is in expressing gas consumption as a function of temperature. The two series of records would be considered as independent series (on *this* house), but if the study had involved multiple houses, they would have been paired observations (because on the same house).

The statistical analyses shown are based on linear regression models with temperature (x) and gas consumption (Y) records related by the equation:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 26 \text{ and } 30,$$

where the errors ε_i are assumed independent and $\sim \mathrm{N}(0, \sigma)$. Separate analyses are carried out for the data before and after the insulation. The table below gives estimates of all parameters and confidence intervals for the regression parameters (using $t_{.975}(24) = 2.064$ and $t_{.975}(28) = 2.048$):

Parameter	Statistic	Before insulation	After insulation
intercept (β_0)	estimate 95% CI	6.854 ±0.244 (6.609,7.098)	4.724 ±0.266 (4.458,4.990)
slope (β_1)	estimate 95% CI	-0.393 ±0.040 (-0.434,-0.353)	-0.278 ±0.052 (-0.330,-0.226)
stand. dev. (σ)	estimate	0.281	0.355

In both models, there is strong evidence of the relation between temperature and gas consumption ($P < 0.0005$). The R^2-values are fairly high (0.944 and 0.813 for before and after insulation, respectively), and indicate that the models would be give fairly good predictions (in particular the model before insulation).

Subquestion b)

The residual plots for the analysis of data before insulation look really nice. The normal probability plot is close to straight, and the plot of residuals versus fitted values has no indication of a cone (or fan) shape or lack of fit. The histogram looks perhaps a bit left-skewed, but this could be an artifact caused by the small sample size. The most extreme residuals are only moderately beyond ± 2, and hardly of any concern.

The residuals plots for the data after insulation are not quite so nice. The normal probability plot is not quite straight and has one point clearly off the line to the very left, and also a cluster of points below the line in the center. The plot of residuals versus fitted values is perhaps dominated by one fairly extreme residual; its value seems to be close to -3. The distribution of residuals seems to be left-skewed with most of the points above zero; this could be a result of the single negative residual. It seems worthwhile to examine this point more closely. In the fitted line plot, we identify it as the lowest recorded gas consumption value at the bottom right of the graph. Another value close by might be suspect as well, in particular if those points were recorded in weeks close in time. In summary, this (these) observation(s) should be further inspected, and it would also be of interest to rerun the analysis on the dataset with this (these) observation(s) omitted.

Subquestion c)

The rate of change in gas consumption per degree increase in temperature is the slope. The estimated slopes are given in the table for **a)**, with 95% confidence intervals. The fact that these intervals do not overlap gives evidence at the 5% significance level of a difference in the slopes before and after insulation. A t-test may also be computed based on the estimates and standard errors as follows:

$$t = \frac{-0.393 - (-0.278)}{\sqrt{0.0196^2 + 0.0252^2}} = -3.60.$$

Note that the denominator is the estimated standard error for the difference between the slopes. The degrees of freedom for this statistic is not so easy to determine, but the conservative rule of using the smaller of the df for the two regression analyses gives df$=$24, for which the t-value of -3.60 is highly significant ($P < 2 \cdot 0.001 = 0.002$). There is clear evidence of a difference in slopes before and after insulation; the slope is steeper before insulation.

Subquestion d)

The estimated gas consumption at a temperature of zero degrees is the intercept of the fitted regression; these values are also given in **a)**. The estimate is much higher before than after insulation, and by the non-overlapping confidence intervals the difference is statistically significant at the 5% level (and clearly much stronger than that). A statistical test can be computed in the same way as in **c)**: $t = 12.1$ which is indeed highly significant.

The estimated gas consumption at a temperature of 10 °C can be calculated from the fitted line equations:

$$\begin{aligned} \text{before:} \quad & \hat{y} = 6.854 - 0.3932 \cdot 10 = 2.922, \\ \text{after:} \quad & \hat{y} = 4.724 - 0.2779 \cdot 10 = 1.945. \end{aligned}$$

Without standard errors for these estimates, statistical inference is not possible. Predictions with intervals can be requested from software. Just as for the intercepts, the comparison of these two values should be based on confidence intervals rather than prediction intervals, because we are not considering a new observation but the predicted gas consumption (corresponding to the linear relation). We can get a rough sense of the resulting inference by the following argument. Because the value 10 is roughly as far away from the mean x-value (given in the data table) as 0, the standard errors should not be too different from those of the intercepts. The difference between the two estimated values above far exceed the difference between the margin of errors for the intercept, so it seems plausible that there will also be a significant difference at 10 °C, with the gas consumption being lower after insulation.

Subquestion e)

The temperature x at which the estimated gas consumption is the same before and after insulation can be obtained by solving the equation obtained by putting the estimated linear relations equal to each other:

$$6.854 - 0.393x = 4.724 - 0.278x, \quad \text{or} \quad x = \frac{6.854 - 4.724}{0.393 - 0.278} = 18.5.$$

That is, according to the estimated equations the gas consumption would be lower after insulation at all temperatures up till 18.5 °C. However, one should question whether the equations hold so far beyond the range of temperatures in the study; this would be a case of extreme extrapolation. By inserting 18.5 °C into one of the equations, it is seen that the predicted gas consumption at that temperature is *negative*, clearly indicating that the equations cannot be extrapolated so far. Based on the fact that the gas consumption is estimated to be higher prior to insulation within the entire range of temperatures of the study, and that the equations cannot be extended meaningfully to a point where this is no longer true, we may conclude that the insulation is associated with a reduced gas consumption at all temperatures until gas consumption essentially vanishes.

Subquestion f)

In summary, the insulation led to a reduced gas consumption within the entire temperature range of the study (-1 – 10°C), and our analysis also showed that gas consumption after insulation increased less steeply with drops in temperature than before insulation.