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

Solution to Final Exam, April 2018

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)

Some of the descriptive statistics indicate right-skewness, while others point towards a quite symmetrical distribution:

- the mean (0.657) is only slightly larger than the median 0.653 ($\sim$ symmetry),
- the distance from the median to the first quartile (0.029) is about the same as (actually a bit larger than) the distance to the third quartile (0.025) ($\sim$ symmetry),
- the distance from the median to the minimum (0.064) is much smaller than the distance to the maximum (0.164) ($\sim$ right-skewness),
- the skewness parameter is 1.43 (that is, >0 and pretty large, indicating right-skewness).

The kurtosis of the distribution cannot be interpreted meaningfully with such a large skewness. Finally, we compute IQR $= 0.678 - 0.624 = 0.054$, and $1.5 \cdot \text{IQR} = 0.081$. This shows that there is at least one potential outlier at the upper end of the distribution ($0.678 + 0.081 = 0.759 < 0.817$), and it is substantially beyond the cut-off so it appears to be a strong (suspected) outlier. There are no potential outliers at the lower end of the distribution. In summary, the data seem to be inconsistent with a normal distribution by its right-skewness, which in turn seems to be due to one (or several) extreme values in the right tail. Suggested further analyses are a normal plot and a normality test. (Normality tests give $P < 0.005$.)

Subquestion b)

Our statistical model is that the 50 pill weights, $Y_1, \ldots, Y_{50}$, form a sample of independent observations from the same distribution (i.i.d.) with mean μ and standard deviation σ. As discussed above, assuming a normal distribution does not seem reasonable. Therefore, our analysis is going to be approximate, utilizing that the sample mean, $\bar{Y}$, has an approximate normal distribution even if the individual weights are non-normal. The guidelines for use of procedures based on the t-distribution (slide 7L–3) are somewhat inconclusive in this case, because even with $n = 50 > 40$ one needs to be careful with strong outliers in the distribution. The parameter estimates are: $\hat{\mu} = \bar{Y} = 0.657$ and $\hat{\sigma} = s = 0.045$. The 90% confidence interval is given by:

$$\mu : \bar{Y} \pm t^* s/\sqrt{n} = 0.657 \pm 1.684 \cdot 0.045/\sqrt{50} = 0.657 \pm 0.011.$$

For this calculation, we used $t^* = t_{.95}(49) \approx t_{.95}(40) = 1.684$ from Table C of PSLS.

Subquestion c)

We want to test the hypothesis H_0: $\mu = 0.65$ vs. the two-sided alternative H_a: $\mu \neq 0.65$:

$$\begin{aligned} t &= (\bar{Y} - 0.65)/(s/\sqrt{50}) = (0.657 - 0.65)/(0.045/\sqrt{50}) = 1.10, \\ P &= 2 \times P(t(49) > 1.10) > 2 \times 0.1 = 0.2, \quad \text{and clearly non-significant.} \end{aligned}$$

The assessment of the P-value comes from $t_{.90}(40) = 1.303$, again from Table C of PSLS. There is no statistical evidence that the production would be off its target. Considering that the test outcome is far from significant, it seems reasonable to assume the conclusion to be valid despite the non-normality of the distribution.

Subquestion d)

The box will contain more than 100 pills if the total weight of 100 pills does not exceed 66 grams. The requested probability is for at least 100, or more than 99, pills in the box, which would happen if the total weight of 99 pills does not exceed 66 grams. Denote by S_n the total weight of n pills. We have $\mathrm{E}S_n = n\mu$ and $\mathrm{sd}S_n = \sqrt{n}\sigma$, which we for $n = 99$ estimate by $99 \cdot 0.657 = 65.043$ and $\sqrt{99} \cdot 0.045 = 0.448$, respectively. The probability is therefore (approximately),

$$\mathrm{P}(S_{99} < 66) = \mathrm{P}(Z < (66 - 99\mu)/\sqrt{99}\sigma) \approx \mathrm{P}(Z < (66 - 65.043)/0.448) = \mathrm{P}(Z < 2.136) \approx 0.984,$$

from Table B in PSLS.

Question 2

Denote by X_{ij} the blood coagulation time for animal j in fodder group i, where i = C,V1,V2 and $j = 1, \ldots, n_i$ ($n_1 = 6, n_2 = 3, n_3 = 3$). Note that the animal numbering 1–12 is not used in this notation.

Subquestion a)

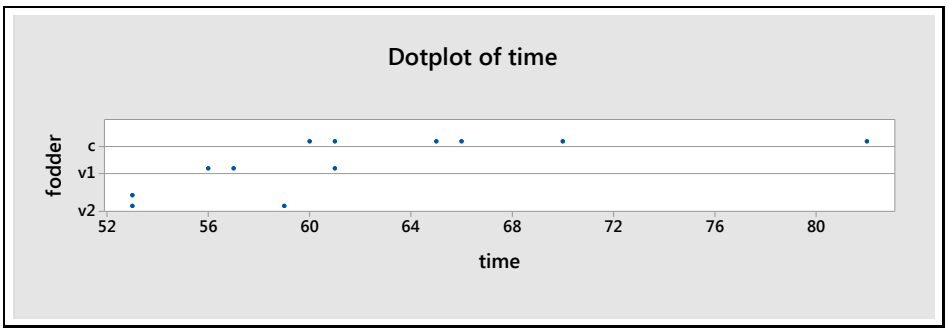
The statistical design is three independent samples. The experimental unit is the animal, and the fodder group is the single factor and also designates the treatment. The design is unbalanced, with unequal replication in the three fodder groups. The design could also, less precisely, be described as a one-way ANOVA layout.

Subquestion b)

The statistical model suggested by the design is a one-way ANOVA model:

$$X_{ij} = \mu_i + \varepsilon_{ij}, \quad \text{where } \varepsilon_{ij} \sim \mathrm{N}(0, \sigma).$$

The model assumes the coagulation times within each fodder group to be independent and normally distributed $\mathrm{N}(\mu_i, \sigma)$, with the same standard deviation (σ) in the 3 fodder groups. In order to visually assess the model assumptions, it might be helpful to sketch a dotplot for the observations.



The dotplot, as well as an inspection of the data, show that animal no. 10 in the C group has a higher value than the other animals in that group. The list of residuals and standardised residuals (from the GLM analysis) shows this animal to have the clearly largest residual and a standardised residual of 2.54. This value is well beyond the 95% range (-2,2) for the N(0,1) distribution; it corresponds to the 99.45% percentile (Table B of PSLS). It is also seen from the Minitab output for the oneway ANOVA command that the standard deviation in group C is quite a bit larger than the standard deviations in the other groups, violating the largest/smallest ≤ 2 guideline. Considering the small group sizes this may not be critical for the analysis but it nevertheless illustrates the impact of animal no. 10 (if animal no. 10 is removed, the standard devation within the C group drops to about the half). Given these two indications that the value for animal no. 10 does not truly represent the population (for fodder C) taken together with the fact that animal no. 10 *was* sick (a biological reason for its unrepresentativeness), we decide to analyse the experiment without animal no. 10. If its value had been in reasonable agreement with the other values in its group, it would be justified to keep it in, making the assumption that the sickness did not affect the blood coagulation, but such an assumption seems suspect here. On the other hand, if it was not known that the animal was sick, it would have been more natural to keep it in the dataset and try (also) a nonparametric analysis.

Subquestion c)

The one-way ANOVA analysis proceeds by estimating the parameters of the model ($\hat{\mu}_i$'s are the group means, and $\hat{\sigma}$ is the pooled standard devation s_p) and by setting up the ANOVA table. Both of these are listed in the (second) one-way ANOVA print. The F-test statistic for testing homogeneity between groups (that the mean blood coagulation times are the same for the 3 fodder groups) has a value of 7.12, with (2,8) degrees of freedom, and a corresponding P-value of 0.017. We conclude that there is evidence (statistical significance) against the group means being equal, that is, some differences exist. The group means show the control group to have the largest blood coagulation times; this is explored further in **d)** and **e)**. There are no signs of problems with the model assumptions for the reduced dataset. A non-parametric analysis (Kruskal-Wallis test) is also possible, but offers no advantages and complicates both the group comparisons and the quantification of the effects.

Subquestion d)

We want to conduct a pairwise comparison between the two vitamin supplementations. This can be done in several ways. We can compute confidence intervals for the two group means, or for their difference; in both cases, we will need $t^* = t_{.975}(8) = 2.306$ (Table C of PSLS). The 95% CIs for the means of groups V1 and V2 will have a margin of error of $t^*\,\mathrm{SE}(\bar{X}_{i\cdot}) = 2.306 \times 2.07 = 4.77$, where the SE was taken from the Minitab listing. With this margin of error it is clear that the CIs for groups V1 and V2 will not only overlap, but also have the estimate (mean) for the other group inside the interval. This implies there is no significant difference (at the 5% significance level) between the two groups.

We could also do a CI for the difference $\mu_{V1}-\mu_{V2}$ or equivalently compute an LSD value to compare the two groups, as follows:

$$\mathrm{LSD}_{0.95} = t^*\, s_p \sqrt{(1/3)+(1/3)} = 2.306 \times 3.59166\sqrt{2/3} = 6.76.$$

The difference between the group means is only is $\bar{X}_{V1} - \bar{X}_{V2} = 58 - 55 = 3.0$, hence smaller than the LSD-value. This also shows that there is no significant difference between the two groups. Finally, we could convert the last calculation into a t-test for $\mathrm{H}_0 : \mu_{V1} = \mu_{V2}$,

$$t = (\bar{X}_{V1} - \bar{X}_{V2})/(s_p\sqrt{2/3}) = 3.0/(3.59166\sqrt{2/3}) = 1.02,$$

which is clearly non-significant in a $t(8)$-distribution; Table C of PSLS gives $t_{.85}(8) = 1.108$, and hence $P = 2 \times P(t(8) > 1.02) > 0.30$. There is no evidence in the data of a different effect of the two vitamin products.

Subquestion e)

For this question it is helpful and reasonable to work under the assumption that the two vitamin products have the same effect, following our analysis in **d)**. For this we compute a combined estimate for vitamin supplementation as $\hat{\mu}_V = \frac{1}{2}(\bar{X}_{V1} + \bar{X}_{V2}) = 56.5$. Because the question asks for a quantification of the effect (relative to control) with an interval, we should do a 95% for $\mu_C - \mu_V$. The estimated difference is $\hat{\mu}_C - \hat{\mu}_V = 64.4 - 56.5 = 7.90$. The (reduced) C group has 5 observations and the combined vitamin group have 6 observations, so the CI becomes:

$$\hat{\mu}_C - \hat{\mu}_V \pm t* s_p \sqrt{(1/5) + (1/6)} = 7.90 \pm 2.306 \times 3.59166 \times 0.6055 = 7.90 \pm 5.02.$$

Because the CI does not include zero, a formal significance test for $\mathrm{H}_0 : \mu_C = \mu_V$ will be non-significant, or $P < 0.05$; this is a satisfactory answer. We can also compute the t-statistic as

$$t = (\bar{X}_C - \bar{X}_V)/(s_p \sqrt{(1/5) + (1/6)}) = 7.90/(3.59166 \times 0.6055) = 3.63, \quad P < 2 \times 0.005 = 0.01,$$

where we used $t_{.995}(8) = 3.355$ from Table C of PSLS. There is clear evidence that the vitamin supplementation has an effect, and the estimates show that it decreases blood coagulation time.

Question 3

Subquestion a)

For the purpose of predicting values of the new method as a function of the established (Gerber) method, we use a linear regression model with Gerber (X) and enzymatic (Y) measurements related by the equation:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, 44,$$

where the errors ε_i are assumed independent and $\sim \mathrm{N}(0, \sigma)$. The Minitab listing gives the parameter estimates:

$$\hat{\beta}_0 = 0.0872, \quad \hat{\beta}_1 = 0.9693, \quad \hat{\sigma} = 0.0795.$$

The Minitab listing also shows an R^2-value of 99.6%, reflecting a very strong relation between the two variables, and indicating that predictions by the model will be very precise. For a measured value by the Gerber method of 4.0, the predicted value for the enzymatic method is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 4.0 = 3.96.$$

The residual plots for the estimated model look good. There is one quite low standardized residual (not too far from -3), and one might want to check whether this observation could be an error. If that is not the case, one would probably keep it in the model as the unexplained variation is very small even with this observation included.

Subquestion b)

The Minitab listing also gives standard errors for the regression parameters. Using $t^* = t_{.975}(42) \approx t_{.975}(40) = 2.021$ we have the following 95% confidence intervals:

$$\begin{aligned} \beta_0 : &\quad 0.0872 \pm 2.021 \cdot 0.0289 = (0.029, 0.146), \\ \beta_1 : &\quad 0.9693 \pm 2.021 \cdot 0.00929 = (0.951, 0.988). \end{aligned}$$

The identity relation ($y = x = 0 + 1 \cdot x$) corresponds to $\beta_0 = 0$ and $\beta_1 = 1$. It is seen that none of these values fall within the corresponding confidence intervals, providing evidence at the 5% significance level against the identity relation. One could also compute t-statistics; the Minitab listing already gives $t = 3.02$ and $P = 0.004$ for the intercept. A test of $\mathrm{H}_0 : \beta_1 = 1$ is computed as follows:

$$t = (\hat{\beta}_1 - 1)/\mathrm{SE}(\hat{\beta}_1) = -3.30,$$

which in a t-distribution with df=42 corresponds to a left tail area close to 0.001, therefore $P \approx 0.002$. The two tests show that there is in fact strong evidence against the identity relation. The estimated slope is a bit smaller than 1, and the estimated intercept is a bit greater than zero. With a positive intercept and a slope less than 1, the enzymatic method is estimated to give larger values than the Gerber method in the lower range of milk fat values, and it will give smaller values than the Gerber method in the higher range.

Subquestion c)

We define the differences as $D_i = Y_i - X_i$, and assume the $n = 44$ differences $D_1, \ldots, D_{44}$ to be i.i.d. and $\sim \mathrm{N}(\mu_D, \sigma_D)$. In this model, the parameter μ_D represents the systematic difference between measurements by the two methods, the bias. The parameters of the model are estimated by the sample values: $\hat{\mu}_D = 0.0005$ and $s_D = 0.0882$. We compute a test for $\mathrm{H}_0 : \mu_D = 0$ against a two-sided alternative as follows,

$$t = \bar{D}/(s_D/\sqrt{n}) = 0.0005/(0.0882/\sqrt{44}) = 0.04,$$

which is obviously totally non-significant in a t-distribution with df=43. There is no evidence of any overall bias between the two methods.

Subquestion d)

For the assumed normal distribution $\mathrm{N}(\mu_D, \sigma_D)$, according to the 68-95-99.7 rule the 95% range is the interval $(\mu_D - 1.96\sigma_D, \mu_D + 1.96\sigma_D)$. Inserting the estimates yields the interval

$$(0.0005 - 1.96 \cdot 0.0882, 0.0005 + 1.96 \cdot 0.0882) = (-0.1724, 0.1734).$$

For this calculation one could also be use $\mu_D = 0$, in consequence of the test just carried out. Also, the value 1.96 may be rounded off to 2 in the formula with minimal loss of information. In the literature, the interval (with estimated μ_D) is referred to as the Bland-Altman limits of agreement.

Next, for an observed fat content of $X = 4.0$ by the Gerber method, we would estimate the enzymatic value as $\hat{Y} = X + \hat{\mu}_D = 4.0 + 0.0005 = 4.0005$. The value is essentially the same as by the Gerber method because the estimated bias between the two methods was so close to zero. The interval combining the expected range for differences with the Gerber value becomes $4.0005 + (-0.1724, 0.1734) = (3.8281, 4.1739) \approx (3.83, 4.17)$. From the regression model of **a)** we would have to request a *prediction interval* for a new observation of 4.0 to get a comparable interval. This interval cannot (with meaningful calculation effort) by computed by hand; software gives the interval as $(3.80, 4.13)$.

Subquestion e)

The differences not being constant across the range of measurements could for example mean that differences were mostly positive at one end of the scale and mostly negative at another end of the scale. In the data, the scale is represented by the values of either X or Y, or perhaps more objectively by the average $(X + Y)/2$; using the average avoids choosing one of the variables in favour of the other one, and is potentially more precise.

A descriptive analysis could look at the 44 differences to determine whether any trend is seen across the range of the measurements. The Minitab listing is sorted in ascending order of X (Gerber), but the order would be almost the same by sorting on Y or the average. By browsing through the listing, for low values of X the differences seem to be mostly positive, and for high values of X the differences seem to be mostly negative. It is difficult to assess whether this pattern could be generated by random variation alone.

A quantitative analysis could utilize the information provided about the correlation between the differences and the other variables. For example, the correlation between the difference and the average is $r = -0.427$. The negative value confirms our impression from the descriptive analysis. To assess whether the data provide evidence of a non-zero correlation, we can use the t-test for correlations:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = -0.427\sqrt{\frac{44-2}{1-(-0.427)^2}} = -3.06,$$

which in a t-distribution with df=42 corresponds to $P < 0.005$ (using $t_{.9975}(40) = 2.971$ from Table C of PSLS). Thus, there is strong evidence of a negative correlation between the differences and the average. Replacing the average by X or Y in the calculation would change only little because the corresponding correlations with the differences are similar. We conclude that the simple model of **c)** is inappropriate for these data because the differences are not constant. An alternative, quantitative analysis is based on fitting a linear regression model of the differences on the averages (or X or Y). The significance test for the differences to be constant across the range is the same (because testing correlation and slope equal to zero give identical tests), but the regression approach leads to modified limits of agreement. Details can e.g. be found in the paper, Bland JM & Altman DG (1999), Measuring agreement in method comparison studies, *Statist. Meth. Med. Res.* **8**, 135−60.