Biostats VHM 801 Course
Fall 2021, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to Final Exam, December 2021

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 data are from a Danish study carried out in the 1970s.

Subquestion a)

The study was experimental because the calves were assigned to a breeding regime decided by the experimenter. The statistical design is two paired samples, where the twins form the pairs. The pairs may also be interpreted as blocks. Observations from different pairs should (in the absence of information to the contrary) be considered as independent. On the other hand, the values obtained from the two heifers in a pair should be considered as dependent, where the dependence arises from the strong genetic similarity among the two monozygotic twins.

Subquestion b)

Because of the paired samples, statistical inference to compare the two diets must be based on the differences within pairs. The question asks us however to look at the distributions across pairs for the early and normal breeding groups. Assuming the observations for the 23 heifers in each group to form a simple random sample (or to be independent and have identical distribution of their milk yield), the distributions can be described as follows:

- both distributions appear as fairly symmetrical, despite the normal group showing a minor left skewness, and both variables show reasonably straight lines in the normal probability plots and have P-values for the A-D normality test clearly non-significant; in other words, both distributions appear as reasonably normal in shape,
- the mean and the central part of the distribution (i.e., the box) are higher for the normal group, whereas the spread (both in terms of the standard deviation and the IQR) is larger for the early group; both of these findings are in agreement with the stated expectations prior to the study,
- neither distribution shows any suspected outliers.

Subquestion c)

The milk yields for the normally and early bred heifers (3100 and 4500, respectively) do not fall outside the range of normally and early bred heifers among the 23 pairs in our data. Their difference of -1600, however, is quite a bit smaller than any difference value in our data. We can assess whether it would be flagged as a suspected outlier by the rule based on the IQR:

$$\text{lower bound} = Q1 - 1.5 \cdot \text{IQR} = -30 - 1.5 \cdot (1720 - (-30)) = -2655.$$

As seen, it is far from that lower bound and would not be indicated as a suspected outlier. Part of the reason for this is that the central part of the distribution of differences is very wide whereas the

tails are relatively small. The normal probability plot and test for this variable shows that a normal distribution does not describe it (very) well. Without a better understanding of the shape of this distribution, it may be difficult to reach a definitive conclusion about outliers, but it seems fair to say that for the yield to be so much higher in the early bred heifer is unusual, even if not extreme enough to label the pair as an outlier.

Subquestion d)
Because of the paired samples, statistical inference to compare the two diets must be based on the differences within pairs. If we denote by D_i the difference in the i^{th} pair, our statistical model is:

$$D_1, \ldots, D_{23} \text{ are a simple random sample (or i.i.d.) from } N(\mu, \sigma).$$

As already mentioned, the normal distribution assumption may be questionable with a curved pattern for the points in the probability plot (corresponding to a bimodal distribution) and a low (and significant) P-value of 0.035 for the normality test, but we will proceed while keeping in mind that our results may be only approximate. The statistical analysis has the steps:

- *Estimation*: $\hat{\mu} = \bar{D} = 862, \;\; \hat{\sigma} = s = 857,$
- *Test* of H_0: $\mu = 0$ vs. H_a: $\mu > 0$ (based on the stated expectations prior to the study):

  $$t = \bar{D}/(s/\sqrt{23}) = 862/(857/\sqrt{23}) = 4.82, \quad P = 2 \times P(t(22) > 4.82) < 0.001.$$

  (Table C gives the 99.9% percentile in $t(22)$ as 3.792.) The test is strongly significant, so we conclude there is very strong evidence against equal means, and hence in favour of a higher mean yield in the normal breeding group.

The estimated difference was 862, and we should supplement that with a confidence interval (but it was not asked for in the exam).

Subquestion e)
Among the nonparametric methods listed, only the sign test and Wilcoxon signed rank test apply to our one sample of differences. The Mann-Whitney-Wilcoxon test is for two independent samples, and the Kruskal-Wallis test is for two or more independent samples, so they do not apply. A permutation test could be constructed, but this has not been discussed in the course.

Among the four statements shown, only i) and iv) are correct. The outcome of interest is the difference within each pair, and it did show some non-normality. The nonparametric methods provide inference about the median, not the mean. According to the guidelines for how serious violations of the normal distribution are at different sample sizes, the t-test should in fact be robust enough here (with $n > 15$ and no serious skewness or outliers). The distribution of differences is so clearly centered above zero that nonparametric tests should also show clear significance.

Subquestion f)
We assume the milk yield in the normal breeding group to be normally distributed, with the normal distribution parameters equal to the sample estimates: $X \sim \text{N}(4217, 794)$. Then we calculate

$$\text{P}(X > 4000) = \text{P}\left(\frac{X - 4217}{794} > \frac{4000 - 4217}{794}\right) = \text{P}(Z > -0.27) = 0.6064 \approx 0.61,$$

using Table B. The z-score could also be obtained directly: $z = (4000 - 4200)/794 = -0.27$.

For two independent, normally bred heifers, the probability that they both exceed 4000 is obtained by the multiplication rule:

$$\mathrm{P}(X_1 > 4000, X_2 > 4000) = \mathrm{P}(X_1 > 4000) \cdot \mathrm{P}(X_2 > 4000) = 0.61 \cdot 0.61 = 0.37.$$

A similar calculation would not work for two heifers in a pair because of the dependence of their milk yields. Both values need to taken into account simultaneously when computing the probability, essentially a calculation in a two-dimensional normal distribution and outside the scope of the course. For this entire subquestion, another approach would be to estimate the probabilities as proportions in the data, e.g. $14/23 = 0.61$ in the normal breeding group. In view of the reasonably good agreement with a normal distribution, the estimate based on the normal distribution is preferable.

Question 2

The question is based on the EESEE (Electronic Encyclopedia of Statistical Examples & Exercises) story "Surgery in a blanket" included with the textbook.

Subquestion a)

The study is experimental because the research team allocated patients to the two treatment groups. Based on the information provided, there is nothing to link the patients in the two treatment groups, so they should be considered as two independent samples. The study seems to be properly randomized, although no specific detail is given about how the random assignment of patients was carried out. The study is also blinded in several ways: both the patients and the entire team involved with the treatment of the patients at the hospital were unaware of the actual assignment to treatment groups. The blinding helps to avoid biases and is a definite strength of the study design.

Subquestion b)

The days of hospitalization is a quantitative outcome, so the obvious approach is a two-sample t-procedure. Thus we assume the outcomes to be i.i.d. from normal distributions $\mathrm{N}(\mu_n, \sigma_n)$ and $\mathrm{N}(\mu_h, \sigma_h)$ in the normothermic and hypothermic groups, We have no information on which to assess the normality assumption, but with the large and fairly equal sample sizes in the two groups the distributions need to deviate strongly from normal distributions in order to substantially affect the validity of t-distribution inference. In two-sample t-inference without assuming equal variances, it is typically difficult to compute the degrees of freedom, but the Minitab listing gives df = 165. Without that value, we would have to approximate the degrees of freedom by the smallest df for the two samples, i.e. df = 95 (or use inference assuming the variances to be equal). Table C gives $t^* = 2.626$ for df = 100. Then the calculations for the 99% CI for $\mu_n - \mu_h$ take the form,

$$\begin{aligned} \bar{X}_n - \bar{X}_h \pm t^*(80)\sqrt{s_n^2/n_n + s_h^2/n_h} &= 12.1 - 14.4 \pm 2.626 \cdot \sqrt{4.4^2/104 + 6.5^2/96} \\ &= -2.6 \pm 2.08 = (-4.68, -0.52) \approx (-4.7, -0.5). \end{aligned}$$

(The standard error could also have been computed roughly from the 95% CI in the Minitab listing, by dividing its margin of error by 1.984, the t^*-value for a 99% CI.) The confidence interval includes only negative values, thereby constituting evidence, i.e. significance, at the 1% level (i.e., $P < 0.01$) that the hospitalization lengths are shorter in the normothermic group.

Subquestion c)

The statistical model here is two independent binomial distributions, $\mathrm{B}(104, p_n)$ and $\mathrm{B}(96, p_h)$, for the counts of patients with postoperative infections in the normothermic and hypothermic groups,

respectively, Estimates for the proportions are listed in the table as $\hat{p}_n = 0.058$ and $\hat{p}_h = 0.188$. In order to test the hypothesis $H_0 : p_n = p_h$, we can use the ("classical") 2-sample z-test. Equivalently, we can consider the data in a two-way table layout, corresponding to a model I (comparison of independent populations). The Pearson X^2-test is for the same null hypothesis, against the two-sided alternative $H_a : p_n \neq p_h$. The second Minitab listing gives this test as $X^2 = 7.965$, with 1 df and $P = 0.005$. The expected counts of the table are all above 5, so the test meets it condition. We conclude that there is strong evidence that the normothermic group has a lower a proportion of patients with postoperative infections.

Subquestion d)

i) Two independent proportions, similar to **c)** above. The counts are very small so that care is needed with the validity for the test; it will barely meets its condition of all expected counts exceeding 5. The z-test or the X^2-test are computable by hand from the information provided ($z = 2.17,\ P = 0.03$).

ii) The duration of surgery and estimated blood loss during surgery are quantitative variables measured on the same patient. The natural parameter for an association between such two variables is the correlation coefficient (measuring linear association). A statistical analysis for the correlation cannot be done from the information provided. Regression is also possible, but less natural from the wording of the targeted parameter.

iii) The counts of events on the same patients during and after surgery would form a two-way table, for example with rows corresponding to events/non-events during operation and columns corresponding to events/non-events after the operation. Such a table could be formed for each treatment group, and in these tables one could assess whether there was an association between events in the two periods. Thus a two-way table analysis is required, based on a sample from a single population (or model II). The two-way table would need to be constructed from the raw data and entered into software for analysis.

Subquestion e)

1) Incorrect. The margin of errors for the 95% confidence intervals are computed by multiplying the standard errors by the appropriate t^*-values (all close to 2), and no further division by the square-root of the sample size is required.

2) Incorrect. The 95% ranges were computed as (hypothermic group): $3.4 \pm 2 \cdot 1.1 = (1.2, 5.6)$, but they should be based on the standard deviation, not the standard error.

3) Correct. The standard deviations would be incredibly large if computed by multiplying the listed standard errors by the square-root of the sample size, e.g. (hypothermic group): $1.1 \cdot \sqrt{158} = 13.8$. Something must be wrong in the table. One might suspect that the authors listed standard deviations instead of standard errors. (In the paper, the listed P-value for comparison of duration of surgery between the two groups does not match with this interpretation either.)

Question 3

Subquestion a)

The statement suggests to model the association between frequency and inverse height (or between inverse frequency and height). As it must be the waterfall that produces the earth vibrations, and

it would make little sense to see the height of the waterfall as a result of the earth vibrations, the causal relation is from the height to the frequency, and not the converse. Therefore we should take frequency as the response (outcome) variable, and height as the explanatory variable. Among the four analyses provided, the third one matches these considerations, with the model

$$F_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, 10,$$

where $x_i = 1/\text{height}_i$ is the ith waterfall's inverse height, F_i is its frequency, and the errors ε_i are assumed independent and $\sim \text{N}(0, \sigma)$. The predictive power of the model is expressed by $R^2 = 93.7\%$, the proportion of variation explained by the model. The model has (fairly) high predictive ability/power, and the points scatter quite closely to the line. The frequencies for the two smallest waterfalls (with heights less than $10m$) are not fitted very well by the line, and one them produces a fairly large standardized residual of 2.54. This may be of some concern, and one could speculate that the linear relation applies only for smaller values of x, and hence for larger (higher) waterfalls. Note also that the relation between frequency and (untransformed) height is clearly curved and therefore inappropriate for linear regression.

Subquestion b)
Estimates and 95% confidence intervals for the regression parameters (using $t^* = t_{.975,7} = 2.365$):

$$\begin{aligned}
\text{intercept}: \quad & \hat{\beta}_0 = 2.803 \quad 95\%\ \text{CI}: 2.803 \pm 2.365 \cdot 1.89 = 2.803 \pm 4.470 = (-1.67, 7.27), \\
\text{slope}: \quad & \hat{\beta}_1 = 218.1 \quad 95\%\ \text{CI}: 218.1 \pm 2.365 \cdot 21.4 = 218.1 \pm 50.6 = (167.5, 268.7), \\
\text{stand. dev}: \quad & s = 3.863
\end{aligned}$$

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.

Subquestion c)
The test of the association between frequency and inverse height will test the hypothesis $\text{H}_0 : \beta_1 = 0$ against a two-sided alternative. The test statistic is given in the Minitab listing as $t = 10.17$, $P = 0.000$ (or equivalently as $F = 103.43$ with the same P-value), meaning that $P < 0.0005$. We conclude there to be very strong evidence of an association between frequency and inverse height. It is totally out of the question that this could have happened by chance only (under the other model assumptions made).

For the line to pass through the origin, the intercept must be zero. We can test the hypothesis $\text{H}_0 : \beta_0 = 0$ by a t-test, or use the confidence interval for β_0 to assess the hypothesis. The Minitab listing gives $t = 1.48$, $P = 0.182$ for a test of the hypothesis, and the 95% confidence interval clearly includes zero. We conclude that there is no evidence in these data to suggest that the line could not pass through the origin. On the other hand, we do not have x-values very close to zero in the dataset (corresponding to very high waterfalls), so we should not consider this as a "proof" that the line passes through the origin; it only tells that this is in agreement with the actual data.

Subquestion d)
The biggest concern with the linear equation is in the range of x-values above 0.10 (as discussed in **a)**). This is the most obvious range to take an additional observation in. For the estimated equation the observation corresponding to the largest x-value close to 0.20 has a fairly strong influence on the relation, so it is desirable to include an additional observation not too far from this value. For

this solution, we will consider an additional waterfall with a height of $6m$. This corresponds to $x = 1/6 = 0.167$, and the prediction:

$$\hat{F} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0.167 = 39.2.$$

As the observed frequency for a waterfall of this height would be a new observation, its agreement with the model should be based on the prediction interval (and not the confidence interval). We can approximately read off the interval from the graph, as $(28, 50)$. If the new observation falls outside of this range, we have formal evidence at the 5% significance level to say that it is not in agreement with the model fitted from the current data. If the new observation does fall within the interval, we may also consider whether the estimated regression line after inclusion of the new observation differs substantially from the current line. This would be a subjective assessment, and may for example be carried out by plotting the data points with the two fitted lines overlaid.