Biostats VHM 801 Course
Fall 2024, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to Final Exam, December 2024

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 minor parts of the questions were waived in the marking.

Question 1

The data are from Supplementary Exercise 7.134 of IPS7e.

Subquestion a)

The statistical design is independent samples from two populations: male college students and female industrial workers. The outcome is the number of pins inserted in a fixed time interval. Although this is a count variable, one might expect the distribution to spread out over such a wide range that it can be approximated well by a continuous distribution. We will assume the scores to follow normal distributions; further discussion is deferred to **b)**. The statistical model is therefore that the scores X_i, $i = 1, \ldots 750$, in the students group are i.i.d. and $\sim \mathrm{N}(\mu_1, \sigma_1)$, and that the scores Y_i, $i = 1, \ldots 412$, in the workers group are i.i.d. and $\sim \mathrm{N}(\mu_2, \sigma_2)$. Furthermore, the data for students and workers are assumed independent. The parameter estimates are given in the question. The hypothesis of interest is $\mathrm{H}_0 : \mu_1 = \mu_2$, and with an explicitly stated expectation that the workers would outperform the students the most natural alternative is one-sided, $\mathrm{H}_a : \mu_1 < \mu_2$. We can test the hypothesis by a t-test. The estimated standard deviations are fairly close, but the Minitab listings are for the test without assuming $\sigma_1 = \sigma_2$, so there is little point in making that assumption. The first Minitab listing matches the data, but has a two-sided alternative; therefore,

$$t = -8.95, \quad \mathrm{df} = 893, \quad P < 0.0005/2 = 0.00025.$$

The P-value for a one-sided H_a is half of the P-value for the two-sided H_a, when the direction in the test matches the alternative; it is the case here (because the mean is lower for students than workers). We conclude that the scores of male students and female workers are different, with the female workers performing on the average 2.2 units higher than the male students.

Subquestion b)

The five-number summary gives the minimum, first quartile (Q1), median, third quartile (Q3) and maximum. The mean and median are fairly close (37.32 vs. 37). The distances of Q1 and Q3 from the median are the same (3.5). Both of these features indicate a fairly symmetrical distribution. The distance from the median to the smallest and largest values are quite different though (14 vs. 9). The simple rule to detect potential outliers based on the quartiles, gives that values more than $1.5 \cdot \mathrm{IQR} = 1.5 \cdot (\mathrm{Q3} - \mathrm{Q1}) = 10.5$ apart from Q1 and Q3 in the lower and upper tails, respectively, may be considered as potential outliers. It is seen that the lower tail has one observation which is exactly borderline for being a potential outlier. Given the large sample size, this could easily have happened by chance. Further exploration of the distribution of scores e.g. using a normal probability plot, or a histogram, stemplot or dotplot would aid in determining any substantial outliers. In any case, the borderline outlier is hardly very strong. At large sample sizes, the t-procedures are fairly robust to minor skewness and moderate outliers. Therefore, the information about the distribution does not raise any flags about the validity of the analysis in **a)**. However, one would need to check the distribution of scores in the student group as well.

Subquestion c)
With the assumed normal distribution for the scores, a 95% range for the distribution is obtained from the 68-95-99.7 rule. The interval $\mu \pm 2\sigma$ spans approximately 95% of the distribution. Inserting our estimates for μ and σ, we get the interval: $37.32 \pm 2 \cdot 3.83 = 37.32 \pm 7.66 = (29.66, 44.30)$. If the assumption of a normal distribution was not tenable, one could either assume another distribution in order to determine a 95% range, or one could use the observed 2.5% and 97.5% quartiles in the data to determine the range. Both of these approaches would require access to the full set of values.

Subquestion d)
Based on the five-number summary for the scores at the 15th minute of the test, we can make the following observations/comparisons:

- the location of the distribution has shifted upwards: the median is 7 units higher, and also the other quartiles and the extremes are higher,
- the spread of the distribution seems to be larger as well; the inter-quartile range (IQR) was 7 in the 1st minute but is 10 in the 15th minute, and also the distances to the minimum and maximum have increased,
- the distribution still appears roughly symmetrical; the median is still exactly between Q1 and Q3, and the distances to the extremes are not identical but still faily similar,
- the distribution for 15th minute scores does not have any potential outliers; in fact, the extremes are well within $1.5 \cdot \text{IQR}$ at both tails.

In summary, the scores at the 15th minute are higher and more dispersed, but still appear to be roughly symmetrical.

The scores at the 1st and 15th minute for the workers) constitute paired samples. Therefore one would need to compute the differences $(D_i, i = 1, \ldots, 412)$ in order to assess the statistical significance of the difference in the distributions just described. If a normal distribution for the differences is tenable, one could test the hypothesis of no difference between the distributions by $\text{H}_0 : \mu_D = 0$ by a t-test.

Subquestion e)
With the notation used above ($X \sim$ students, $Y \sim$ workers), $p = \text{P}(X < Y) = \text{P}(Y - X > 0)$; therefore, we need to determine the distribution of $(Y - X)$. We have no other choice but to use the same normal distributions as above. Because X are Y are independent, the distribution of $(\bar{Y} - \bar{X})$ is normal, with

$$\begin{aligned} \text{E}(Y-X) &= \mu_2 - \mu_1, \\ \text{Var}(Y-X) &= \text{Var}(Y) + \text{Var}(X) = \sigma_1^2 + \sigma_2^2, \\ \text{sd}(Y-X) &= \sqrt{\text{Var}(Y-X)}. \end{aligned}$$

Inserting the estimates, yields $(Y - X) \sim \text{N}(2.2, 5.77)$. Therefore $p = \text{P}(Z > -2.2/5.77 = -0.38) = \text{P}(Z < 0.38) = 0.648 \approx 0.65$, using Table B of PSLS. Hence option *ii)* is the correct choice. This conclusion could have been reached without the exact calculation. Because workers had the higher mean, we would need to have $p > 0.5$. On the other hand, the standard deviations in the two distributions are both substantially larger than the mean difference, so even if the uncertainty came from one distribution only, p could not become very large. For example, $-2.2/3.83 = -0.57$, and the left tail probability beyond that value is still less than 0.25.

Question 2

The question is based on the paper Kontiokari *et al.* (2001), Randomised trial of cranberry-lingonberry juice and lactobacillus Gg drink for the prevention of urinary tract infections in women, *British Medical Journal* **322**, 1571–1573.

Subquestion a)

In each treatment group, the data on the 50 women correspond to a binomial setting. Therefore we assume the count of women with recurring UTI, X, to be binomially distributed: $X \sim \mathrm{B}(50, p)$. For confidence intervals on a single proportion we can use the classical (normal approximation) method if the number of cases and non-cases both exceed 15; for these data, this is the case only in the control group. Instead we can use the "plus four" method if the sample size exceeds 10; this condition is met for the cranberry group. Using these methods, we get,

$$\text{cranberry group:} \quad = \quad \tilde{p} \pm 1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} = \frac{10}{54} \pm 1.96\sqrt{\frac{(10/54)(44/54)}{54}} = 0.186 \pm 0.104 = (0.082, 0.290)$$

$$\text{control group:} \quad = \quad \hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{18}{50} \pm 1.96\sqrt{\frac{(18/50)(32/50)}{50}} = 0.360 \pm 0.133 = (0.227, 0.493).$$

The "plus four" CI is also valid in the latter case and equals $(0.242, 0.499)$. (For reference, the (invalid) classical CI for the cranberry group equals $(0.058, 0.262)$.) It is seen that the two confidence intervals overlap, but that none of the estimates are included in the other confidence interval. Therefore we cannot from the confidence intervals alone determine whether a test comparing the proportions in the two groups will be statistically significant. We would need to compute a two-sample test or a confidence interval for the difference between the two proportions.

Subquestion b)

The statistical model for the full dataset would assume independent binomial distributions for the three groups with separate proportions (probabilities) in each group. In terms of data laid out in a two-way table of counts this would be a Model I for comparing independent populations. The third Minitab listing shows the correctly constructed two-way table. The Pearson chi-square statistic is $X^2 = 7.29$ with $P = 0.026$ in a χ^2-distribution with 2 df. The test is significant, and we therefore reject the null hypothesis H_0 of equal proportions of UTI recurrence in the three groups. The estimated proportions in the three groups are: cranberry group: $\hat{p}_1 = 8/50 = 0.16$; lactobacillus group: $\hat{p}_2 = 19/49 = 0.39$; control group: $\hat{p}_3 = 18/50 = 0.36$. Based on the estimated proportions it appears that the cranberry group has lower UTI recurrence than the two other groups, which are very similar.

Subquestion c)

The correct answers are:

i) (c); the condition for the chi-square approximation is in terms of the expected cell counts under the null hypothesis.

ii) (b); the (cranberry, Yes) cell has a chi-square contribution of of $(8-15.10)^2/15.10 = 3.34$, and the next largest chi-square contribution is from the (cranberry, No) cell at $(42-34.9)^2/34.9 = 1.44$, and hence substantially smaller.

iii) (b); note that the null and alternative hypotheses should not be worded in terms of (in)dependence in the present design where group membership is a controlled (explanatory) variable.

iv) (a); with 3 groups, the number of comparisons is $3 \cdot (3-1)/2 = 3$, and the Bonferroni method applies much more generally than to ANOVA: basically to any setting with multiple tests.

Subquestion d)

The Minitab listing does not give us information about a chi-square test involving only the cranberry and control groups, and we therefore have to compute either the chi-square test or the equivalent two-sample z-test manually. The statistical model is two independent proportions, as already stated. For the two-sample z-test we need to compute the pooled probability between from the two samples: $\hat{p} = (8 + 18)/(50 + 50) = 0.26$. Then

$$z = \frac{\hat{p}_1 - \hat{p}_3}{\sqrt{\hat{p}(1-\hat{p})(1/50 + 1/50)}} = \frac{0.16 - 0.36}{\sqrt{0.26 \cdot 0.74 \cdot (2/50)}} = -2.28.$$

Using Table B of PSLS, the P-value is $P = 2 \cdot 0.0113 = 0.023$. There is a significant difference between the proportions of UTI recurrence in the cranberry and control groups. A Bonferroni adjustment for 3 comparisons would make this P-value non-significant, but we should not adjust for multiple comparisons if this was the primary comparison for the study. The lactobacillus group had a proportion of UTI recurrence very similar to the control group, and we can hardly imagine any significant difference between these groups. If we wanted to also compare the lactobacillus group to the cranberry group, the z-statistic is going to become more extreme because the proportions are a bit further apart. So without any adjustments for multiple testing, the conclusion is simply that the cranberry group is significantly lower than the other two, which in turn are not significantly different.

Subquestion e)

The shown sample size calculation based on power corresponds to the stated difference of clinical importance, but the required sample size for 80% power is 62, not 70. If the focus is entirely on reducing the recurrence in UTIs, it could be suggested to use a one-sided alternative hypothesis. That will further reduce the sample size. So it is fair to say that the stated 70 women per group seems to be an overestimate.

The actual study had only 50 women per group, so it did not reach the calculated sample size. The study gave a significant result for the comparison of the cranberry group to control, but it would still be fair to ask for an explanation of why the stated sample size was not met. In the article, the authors explained that the cranberry juice supplier stopped producing the juice.

Question 3

The data are from Denmark and were most likely obtained in an experiment conducted in the pharmaceutical industry.

Subquestion a)

The experiment is a completely randomized two-factor (2×3) design with replication. The experimental unit is the rat, and the two factors are insulin type (A, B) and dose (2.25, 4.00, 5.75), together comprising 6 treatments. Randomization should take place by randomly assigning the 24 rats to the 6 treatment groups with four rats per group.

Subquestion b)

The natural statistical model, and also the model behind the analyses shown in the listings, is a two-way ANOVA model,

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

where $i = \text{A,B} \sim$ insulin type, $j = 1, 2, 3 \sim$ dose group, and $k = 1, 2, 3, 4$. The errors ε_{ijk} are assumed i.i.d. and $\sim \text{N}(0, \sigma)$.

The main model assumptions are the normal distribution and the equal standard deviation across all 6 treatment groups. We can assess normality by the residual plots. The normal plot looks reasonably straight, and there does not appear to be any extreme residuals; the residuals are not standardized so we cannot compare to the N(0,1) scale. In fact, the distribution seems to have a bit short tails (in the histogram indicated by a lack of bell-shape), something that is often seen for data with a limited range (the percent reductions are presumably between 0 and 100%). However, with the relatively small dataset, this may still be within what is acceptable. It could be suggested to store the standardized residuals, and compute both descriptive statistics and a normality test for them. The P-value for the normality test should not be interpreted too rigidly, though (it is indeed non-significant). The plot of residuals against fitted values does not show any indication of systematically varying spread across the treatments (such as a "cone"/"fan" shape). The 6 standard deviations in the table vary a bit more than recommended by the ratio rule ($s_{\max}/s_{\min} = 18.95/7.32 = 2.59 > 2$), but with only four replicates per group this does not seem alarming at all. It could be suggested to do a formal variance test across the 6 groups (P-values for variance tests are indeed non-significant). In summary, no serious concerns with the model assumptions can be seen.

Subquestion c)

A general effect of insulin type and dose-specific effects of insulin type correspond to the main effect of insulin type and the interaction between insulin type and dose, respectively. The ANOVA table shows both of these effects to be clearly non-significant. As a reflection of the non-significant interaction, the interaction plot shows two almost parallel curves, also indicating the absence of an interaction. The main effect plot shows a slightly larger mean reduction for type A than B; the data table gives these values as $\bar{X}_{\text{A}\cdot\cdot} = 53.08$ and $\bar{X}_{\text{B}\cdot\cdot} = 48.00$. In summary, there is no evidence whatsoever that the two insulin types would have different effects on the reduction of glucose levels. For a 95% CI for the difference between the two means, we compute the standard error as

$$\text{SE}(\bar{X}_{\text{A}\cdot\cdot} - \bar{X}_{\text{B}\cdot\cdot}) = \sqrt{\text{MSE}}\sqrt{(1/12) + (1/12)} = 13.92\sqrt{2/12} = 5.683,$$

and hence the 95% CI becomes: $53.08 - 48.00 \pm t^* \cdot 5.683 = 5.08 \pm 11.94$. In the calculation we used $t^* = t_{.975}(18) = 2.101$. We already knew from the non-significance of the main effect for insulin type that the CI would include 0, but the CI gives a range of plausible values for the difference between the two insulin types.

Subquestion d)

By the non-significant interaction already discussed above, we consider only the main effect of dose for which the F-test in the ANOVA table is strongly significant ($F = 11.45, P = 0.001$). Hence, some differences in the effects of the doses do exist. The main effect plot shows that glucose reduction increases with increasing doses (not too surprisingly), and the display of 95% CIs in the Minitab listing shows non-overlapping intervals for doses 2.25 and 5.75, corresponding to a significant difference at the 5% significance level. In order to assess the significance of pairwise comparisons with the middle dose, we could compute an LSD-value,

$$\text{LSD}_{.95} = t^*\sqrt{\text{MSE}}\sqrt{(1/8) + (1/8)} = 14.62,$$

where we again used $t^* = 2.101$, and the denominator of 8 corresponds to each dose mean being over 8 observations. It is seen that all means are separated by more than the LSD-value, hence they should all be considered significantly different at a significance level of 5% for each comparison. An

adjustment for multiple comparisons is possible and seems likely to eliminate at least the significant difference between the two highest doses (because their mean difference is very close to the unadjusted LSD-value). We conclude that significant dose effects exist, and that higher doses are associated with stronger glucose reductions. The main effect plot appears fairly linear although not totally straight.

Subquestion e)
The Minitab listing corresponds to a simple linear regression model,

$$\text{reduc}_i = \beta_0 + \beta_1 \,\text{dose}_i + \varepsilon_i, \quad i = 1, \ldots, 24,$$

where the errors (ε_i) are i.i.d. and $\sim \text{N}(0, \sigma)$. The estimated slope is $\hat{\beta}_1 = 9.5$ and strongly significant (different from zero), by the F-value and its P-value in the ANOVA-table. The estimated intercept is $\hat{\beta}_0 = 12.54$, but a dose of zero is far outside the actual data, so an interpretation of the intercept is probably not of interest. The estimated standard deviation about the line is $s = 13.0$. The fitted line plot does not show any obvious model deviations, but one should also look at the residuals. The scatterplot is a bit deceiving by showing multiple observations with the same value as a single point. It is possible to conduct a formal lack-of-fit test for the linearity assumption, but such tests are beyond the scope of the course (discussed e.g. in VHM 802); in our case, it is far from significant.

For a dose of 5, the predicted value is $\hat{Y} = 12.54 + 9.5 \cdot 5 = 60.04 \approx 60.0$, which can also be read off the plot. The relevant interval for a rat with this dose, is the prediction interval (PI). Rough values read off the plot are 32 for the lower bound and 88 for the upper bound, hence a very wide interval $(32, 88)$. This reflects that the predictive ability of the model is at most moderate ($R^2 = 54\%$). We are 95% confident that the reduction in serum glucose levels experienced by a rat given a dose of 5 units would be within this interval.