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

Solution to Final Exam, December 2018

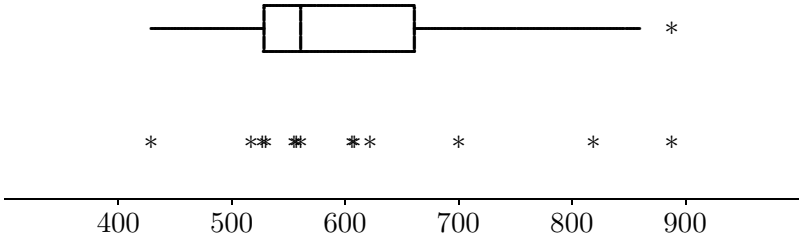
Question 1 was not included in the exam for students who used their midterm mark. The solution is more detailed than expected, by giving additional calculations and detailed interpretations and explanations of all procedures.

Question 1

Subquestion a)

Valid choices for a graphical display are a stemplot, a boxplot or a dotplot, sketched below.

```
4 | 2
5 | 123556
6 | 002
7 | 0
8 | 18
```



From the display and the statistics we can describe the distribution as follows:

- 5-number summary: $429 - 528.5 - 561 - 661 - 888$; in the boxplot, the whiskers extend at most $1.5 \cdot \text{IQR} = 1.5 \cdot (661 - 528.5) = 198.75$ beyond the box,
- some right-skewness visible in the distribution (mean > median, asymmetrical box, suspected outlier in the right tail).
- suspected (mild) outlier in the right tail (888) which does not look particularly extreme relative to the other observations.

In summary, the distribution appears to be somewhat right-skewed and therefore strictly speaking not normal, but the deviations from a normal distribution may not be so strong as to exclude a normal distribution analysis. For the two-sample comparison referred to in the question, one would want to assess the distribution in the other group similarly.

Subquestion b)

For $X \sim \text{N}(3.4, 0.57)$ we calculate,

$$\text{P}(X < 2.5) = \text{P}\left(\frac{X - 3.4}{0.57} < \frac{2.5 - 3.4}{0.57}\right) = \text{P}(Z < -1.58) = 0.057,$$

using a statistical table for $\text{N}(0, 1)$. So 5.7% of babies are born with low birthweight.

The distribution of the average birthweight of three babies from this distribution is normal with the same mean (3.4) and a standard deviation of $0.57/\sqrt{3} = 0.329$. Therefore, by a similar calculation as above,

$$\text{P}(\bar{X} < 2.5) = \text{P}\left(\frac{\bar{X} - 3.4}{0.329} < \frac{2.5 - 3.4}{0.329}\right) = \text{P}(Z < -2.74) = 0.0031.$$

Finally, with each baby born on PEI having a low birthweight probability of 0.05, the probability of none of three babies having low birthweight equals $(1-0.05)^3 = 0.857$. There is about 86% chance that none of three babies born on PEI have low birhtweight. The probability could also have been obtained from the binomial distribution $\text{Bin}(3, 0.95)$.

Subquestion c)

We will assume that the four recordings $X_1, \ldots, X_4$ are i.i.d. from $\text{N}(\mu, \sigma)$ with $\sigma = 0.15$ as stated in the lab manual. It is better to assume σ to be known than estimate it from only 4 observations. In this model, the 95% confidence interval for μ is computed as:

$$95\% \text{ CI for } \mu\text{: } \bar{X} \pm z^* \sigma/\sqrt{n} = 3.80 \pm 1.96 \cdot 0.15/\sqrt{4} = 3.80 \pm 0.15.$$

The 95% CI gives a range within which we are 95% confident the true specific gravity value lies. The 95% confidence means that among repeated sets of (four) measurements, the corresponding CIs will contain the true value with probability 0.95.

We carry out a hypothesis test for $\text{H}_0 : \mu = 3.9$ versus $\text{H}_a : \mu < 3.9$ (as specified in the lab manual), as follows:

$$Z = \frac{\bar{X} - 3.9}{\sigma/\sqrt{n}} = \frac{3.80 - 3.9}{0.15/\sqrt{4}} = -1.33, \quad P = \text{P}(Z < -1.33) = 0.0918.$$

With $P > 0.05$, there is no evidence against the true specific gravity being equal to 3.9, thus the data fail to show convincingly that the true specific gravity is less than 3.9. As the sample mean was less than 3.9, the data may hint at the true value being less than 3.9, but it does not offer evidence for it.

Question 2

Subquestion a)

The first table gives a crosstabulation of the 670 carcasses by two variables, status of the carcass and slaughterhouse. Status is clearly a response variable, and slaughterhouse could be considered either as fixed (explanatory) or random (response) because the submission of samples from the slaughterhouses is indeed random. As the focus of the analysis is on the status and less on the origin of the carcass, it is more natural to consider the slaughterhouse as a fixed variable. The statistical model is therefore of type I — independent binomial distributions $\text{Bin}(n_i, p_i)$, $i = 1, 2, 3$, from the three slaughterhouses. We assume a binomial setting for samples from the same slaughterhouse (same probability of being condemned, and independent outcomes). The hypothesis of interest is whether the proportions of condemnation are the same at the three slaughterhouses, that is, $\text{H}_0 : p_1 = p_2 = p_3$. The computer listing shows that

$$X^2 = 11.27, \ \text{DF} = 2, \ P = 0.004,$$

so there is strong evidence against H_0. The conditions for use of X^2 are easily met, as all expected counts are large. We conclude that some differences exist between the slaughterhouses in their probabilities of condemnation. For further exploration, we may identify cells with the largest deviations between observed and expected counts, and these occur for slaughterhouse I.

However, with only two outcome categories, and because it was specifically requested to compare the proportions of condemned samples, we should estimate the proportions: $\hat{p}_1 = 45/179 = 0.251$, $\hat{p}_2 = 43/318 = 0.135$, $\hat{p}_3 = 27/173 = 0.156$. Apparently slaughterhouse I has a higher rate than the others. If we wanted to further compare the proportions among the slaughterhouses, we would have to do pairwise tests; with the large sample sizes, (classical) z-tests should be fine. This is however a lot of work for an exam, and for a start we could compute the standard errors for these estimates:

$$\text{SE}(\hat{p}_1) = \sqrt{\hat{p}_1(1-\hat{p}_1)/179} = 0.032, \quad \text{SE}(\hat{p}_2) = 0.019, \quad \text{SE}(\hat{p}_3) = 0.028.$$

From these values it is seen that (classical) confidence intervals (for which the margins of errors will be 1.96 times the standard errors) for the proportions will strongly overlap between slaughterhouses II and other (estimates inside the other CI), not overlap between slaughterhouses I and II, and slightly overlap between I and other. Threfore, with non-adjusted comparisons (i.e., no adjustment for doing 3 pairwise comparisons), only the last comparison requires a specific calculation. We will omit the calculation here and return to the question in **b)**.

Subquestion b)

The separate datasets for the two diagnoses (tail bite and other diagnoses) are analyzed by the same statistical model and procedure as in **a)**. The analysis for tail bite carcasses shows a strong significance against H_0 ($X^2 = 13.39,\ P = 0.001$), whereas the analysis for carcasses with other diagnoses shows no significance of H_0 ($X^2 = 2.63,\ P = 0.27$). The non-significance does not prove that there are no differences between slaughterhouses, but the differences are less strong and not sufficient to provide evidence against H_0. We focus therefore on the carcasses with tail bites. The estimated proportions are $\hat{p}_1 = 19/49 = 0.388,\ \hat{p}_2 = 13/95 = 0.137,\ \hat{p}_3 = 4/29 = 0.138$. Also here, slaughterhouse I has the by far largest proportion of condemned carcasses, and the two other groups are so close we don't need to bother about significance. The corresponding standard errors are 0.070, 0.035 and 0.064, respectively; however, we need to be careful with in particular the other group because the condition for a classical 95% CI is clearly violated. Again, CIs for slaughterhouse I and II do not overlap, and we need a (classical) two-sample z-test for comparison of slaughterhouses I and other:

$$Z = \frac{\hat{p}_1 - \hat{p}_3}{\sqrt{\hat{p}(1-\hat{p})[1/49 + 1/29]}} = \frac{0.388 - 0.138}{\sqrt{0.295(1-0.295)[1/49 + 1/29]}} = 2.34$$

where we used $\hat{p} = (19 + 4)/(49 + 29) = 0.295$. We get $P = 2 \cdot \mathrm{P}(Z > 2.34) = 0.019$, giving also evidence for these slaughterhouses to be different; note that the condition for the Z-test is met. In conclusion, we have shown differences to exist between the slaughterhouses in the proportions of condemned carcasses diagnosed as tail bite, and slaughterhouse I has the highest proportion, whereas the other slaughterhouse groups have significantly lower and very similar proportions. In carcasses with other diagnoses however, there were no significant differences between slaughterhouses.

Subquestion c)

(A) The study is an observational study (no treatments were imposed).

(B) One of the computer listings shows a significant difference between the slaughterhouses in the proportions of samples diagnosed as tail bites ($X^2 = 10.4,\ P = 0.006$). The proportions were 0.274, 0.299, 0.168, respectively, and therefore clearly lowest at the other slaughterhouses.

(C) The appropriate procedure is to first examine the two-way tables at each slaughterhouse separately, and if these tables show a consistent pattern we may look at the table combined across slaughterhouses. The slaughterhouse is an obvious lurking variable for the association between status and diagnosis; differences between slaughterhouses have already been seen for both variables. The analysis does indeed show a strong association at slaughterhouse I, and no association at all at the two other slaughterhouses, and the combined table gives a misleading picture. Correlation analysis is not appropriate for categorical variables, and there is really no obvious reason why we would need a new study to estimate the association between diagnosis and status.

Question 3

Subquestion a)

The study[1] is observational in character (measurements were taken, no treatments were imposed), but it is not obvious what population the study may be representative for because the selected

[1] Allison DB, Heshka S, Sepulveda D & Heymsfield SB (1993), Counting calories — caveat emptor, *J. Amer. Med. Assoc.* **270**, 1454–1456.

food products were not chosen randomly from any specified population. The targeted population appears to be that of diet and health food products. The three food origin groups represent different subpopulations, and it seems fair to assume that the samples were independent. Thus, the design may be described as three independent samples from different populations. The design is unbalanced, by its unequal group sizes.

Subquestion b)
The statistical model analysed in the Minitab listing is a one-way analysis of variance model:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, 2, 3;\ j = 1, \ldots, n_i\ (n_1 = 20,\ n_2 = 12,\ n_3 = 8),$$

where Y_{ij} = deviation in caloric content (in percent) for food item j in food origin group i, and the errors ε_{ij} are assumed to be normally distributed $\mathrm{N}(0, \sigma)$.

The dotplots show the three samples to have quite different shapes. The L group is much wider than the others and with a long tail to the right, whereas the other groups are more narrow and approximately symmetrical, except for a possible outlier ("cheese curls") on the lower side for group N. The probability plots show reasonably straight lines, except for group L with only 8 observations. Moreover, the A-D normality tests give no evidence against a normal distribution for any of the groups ($P = 0.27 - 0.56$). The biggest (only) problem is therefore the large difference in variance between the groups. The textbook guideline is clearly violated: $s_L/s_N = 8.0 \gg 2$. The one-way analysis is not acceptable with such large differences in variation between the groups.

The quote from the paper can be criticized on two points. First, there was no evidence against a normal distribution in these data. Second, there is no indication in the study design or the data that the observations may violate the assumption of independence, and any non-parametric procedure (including the one in **c)**) involves the same assumption of independence as the parametric analysis. Therefore, the normality and independence of observations should not have been mentioned in the text.

Subquestion c)
The analysis is Kruskal-Wallis test and therefore a non-parametric one-way ANOVA analysis, in which only the ranks of the observations are used. The analysis assumes independent groups but does not assume normal distributions within groups. It may be done under the additional (Δ-)assumption that the within-group distributions have the same shape. In the present situation, this is clearly not relevant because we described the distributions as very different, in particular they had different variability. Therefore our null and alternative hypotheses should be phrased simply as the three distributions being equal, and not equal, respectively (and *not* in terms of the medians). The test statistic is highly significant ($P < 0.0005$), and the average ranks for the three groups indicate group N to be lower than the other two, which have quite similar average ranks. For further pairwise comparisons, one would need to do Wilcoxon-Mann-Whitney two-sample tests; this analysis requires software, e.g. the Nonparametrics menu in Minitab.

Subquestion d)
The analysis of the transformed data is based on the same statistical model as above (but for the transformed, say Y^*, values). The dotplot still shows somewhat different distribution shapes in the three groups, but the variances are much more similar: our guideline is only mildly violated: $s_L/s_R = 2.3$; this may be acceptable. Only group R seems to have a symmetrical distribution, and group N may be severely left-skewed, in part due to the potential outlier. It is suggested to run the analysis with and without this potential outlier to see if it affects the results. The normal plot for the

residuals has some curvature around the straight line but may be acceptable. The histogram seems to indicate some left-skewness. It is suggested to obtain a P-value for a normality test, either in each group or for the standardized residuals.

The ANOVA table shows a strongly significant F-test for the hypothesis that the group means are the same. Irrespective of the doubts about the compliance with model assumptions, it is clear that strong differences between the groups exist. The means and confidence intervals show that the deviations of caloric content (after transformation) of group N are significantly lower than those of the two other groups. Note that the transformation preserves the order between the groups from the original data. This significance is not adjusted for multiple (3) comparisons but it seems likely to hold also after such an adjustment. For the comparison between groups L and R we compute a t-test:

$$t = \frac{\bar{Y}_L^* - \bar{Y}_R^*}{\sqrt{\text{MSE}((1/8) + (1/12))}} = \frac{0.3536 - 0.1878}{0.1542\sqrt{(1/8) + (1/12)}} = 2.356,$$

which in a t-distribution with df = 37 corresponds to a percentile somewhere between 98% and 99%; therefore $0.02 \leq P \leq 0.04$ for a two-sided alternative hypothesis. We conclude that there is a statistically significant difference between groups L and R. After Bonferroni adjustement (by a factor of 3), this difference would however no longer be significant.

Subquestion e)

We now consider each sample separately, and assume (as before) the values to constitute a SRS (or to be i.i.d.). Both parametric (i.e., based on the normal distribution) and non-parametric analyses are possible. In favour of a non-parametric analysis speaks that the distributions seem to have some non-normal features, both on original and transformed scale. However, the normality tests in **a)** showed no evidence against normal distributions within each sample, so it would be legitimate to use a normal distribution model and test the hypothesis that the mean deviation equals 0. The non-parametric analysis would test that the median deviation equals 0. Without access to the data, only a sign test can be computed. The table below gives values for both tests in all three groups, everywhere against a two-sided alternative.

Statistic	Group N	Group R	Group L
n	20	12	8
$\bar{Y}$	0.13	25.13	81.8
s_Y	10.52	16.07	84.0
$t = \bar{Y}/(s_y/\sqrt{n})$	0.055	5.41	2.75
P-value	$>$0.50	$<$0.001	$<$0.04
$\#Y > 0$	11	11	8
$\#Y \neq 0$	20	12	8
P-value	0.82	0.006	0.008

For the sign test, the P-values were computed as twice the tail probability in binomial distributions with $n = (\#Y \neq 0)$ and $p = 0.5$; these distributions (for $n = 8$ and $n = 20$) are listed in Table 1 of S. The conclusions from the two tests are similar. There is evidence that the means/medians are different from zero in groups R and L, whereas there is nothing to indicate these values to differ systematically from zero for group N. In groups R and L, the means (and medians) are actually greater than zero so the evidence is of population values greater than zero. As the deviation was computed as measured minus stated caloric content, this implies that there is evidence of systematic underreporting of caloric content in food items from these groups.