Biostats VHM 801 Course
Fall 2016, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to Final Exam, December 2016

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)

Assuming the observations for the 12 cats to form a simple random sample (or to be independent and have identical distributions of the different outcomes), the distributions can be described as follows:

- *pre* sct-I concentrations:
  - a quantitative variable but with 4 values equal to zero (presumably below a detection limit),
  - one very large outlier in the right tail (4.03 for cat no. 2),
  - strongly right-skewed due to the single large outlier,
  - remaining 7 values fairly equally spread across the range from 0.05 to 0.29,
- *post* sct-I concentrations:
  - a quantitative variable with a total of 6 values equal to zero (two cats in addition to those for *pre*),
  - among the remaining values, the value 0.04 occurs four times,

In summary, the variable *pre* and *change* has a strong outlier (cat no. 2), and both variables have a substantial number of values equal to zero. To assume a normal distribution for any of these variables would not be appropriate.

Subquestion b)

For this question we consider the *change* values, computed as $post - pre$. The distribution of this has similar features as those discussed in **a)**, in particular one large outlier in the right tail (the same cat) and several values equal to zero. Therefore it does not seem reasonable to assume a normal distribution, and we should not use a t-test to assess whether the mean change differs from zero (for completeness: $t = 1.216, \text{df} = 11, P = 0.25$. Instead we will use a sign test. The null hypothesis is H_0 : median $= 0$, and without any indication of a specific alternative hypothesis of interest we should take H_a : median $\neq 0$. Among the 12 values, 8 are non-zero, and we therefore let $X \sim \text{Bin}(8, p)$ denote the number of cats with a positive change. In terms of the probability (p) of a positive change, the hypotheses are $\text{H}_0 : p = 0.5$ and $\text{H}_a : p \neq 0.5$. The observed value is $X = 8$ (all cats have a positive change). Therefore we compute the P-value from the $\text{Bin}(8, 0.5)$ distribution as follows:

$$P = 2 \times \text{P}(X \geq 8) = 2 \times \text{P}(X = 8) = 2 \times 0.5^8 = 2 \times 0.00390625 = 0.0078.$$

The binomial probability could also be obtained as 0.004 from Table C of IPS or Table 1 of Stevens. There is strong evidence against H_0 which must therefore be rejected. We conclude that the median change is not zero, it is greater than zero. There is evidence of a drop in sct-I values from before the treatment to after treatment.

Subquestion c)
Let $X_1, \ldots, X_{12}$ denote the *pre* sct-I concentrations. As above we assume these to form a simple random sample (to be i.i.d.), but we do not assume a normal distribution. Let μ denote the population mean. We calculate an approximate 95% confidence interval (CI) for μ using the t-distribution procedures; the relevant t-percentile is $t_{.975}(11) = 2.201$. From the Minitab listing, we have $\bar{X} = 0.421$ and $s = 1.141$.

$$95\% \text{ CI for } \mu \ : \ \bar{X} \pm t^* s/\sqrt{n} = 0.421 \pm 2.201 \cdot 1.141/\sqrt{12} = 0.421 \pm 0.725 = (-0.304, 1.146).$$

It is seen that the CI extends well below zero, although negative values are impossible. Without a normal distribution of the X_i's, the CI is only approximate, and in this case it is clearly not very good.

The most serious problem with the data is the single very large outlier (cat no. 2). The first thing one would do is to reassess this value in order to confirm its validity. Without this value, the CI will improve substantially. One possible approach to (further) improving the CI is to transform the *pre* sct-I values. A log transformation will need some adjustment because of the values equal to zero, and it will still be hard to achieve an approximate normal distribution after transformation because of the multiple values equal to zero. A non-parametric procedure will yield a confidence interval for the median, but this is not what was asked for; nevertheless, it may still be the most sensible approach.

Subquestion d)
We consider here the velocity values before treatment, and the 12 cats are divided into two groups based on whether their *pre* sct-I value was zero or non-zero. These groups comprise 4 and 8 cats, respectively, and we have no reason to suspect the two samples would not be independent. The listing of the data values show distributions of velocity values (in the two samples) that look fairly regular, except for one extreme outlier in the right tail (for cat no. 6). The options available to us in the Minitab listing are a two-sample t-test and a two-sample non-parametric test (the Mann-Whitney-Wilcoxon rank (MWW) test). Because the outlier could affect the t-test substantially (by violating the normality assumption), the most natural approach is to use the rank test.

We assume the two samples to be independent (i.i.d.) from distributions D_0 and D_1, say, for cats with `sctzero` equal to 0 and 1, respectively. Estimated medians from the two distributions are 1.25 and 1.05, respectively. The outlier will also make it questionable to assume that the two distributions (D_0 and D_1) have the same shape, so we will not phrase our hypotheses in terms of the population medians. Instead we consider,

$$\begin{aligned} \mathrm{H}_0 : &\quad D_0 = D_1 \ \text{ (same distributions)} \\ \mathrm{H}_a : &\quad D_0 \text{ is systematically larger than } D_1. \end{aligned}$$

The text states that cats with sct-I values equal to zero ($\sim D_1$) were expected to have lower velocity values than cats with measureable sct-I values ($\sim D_0$). The MWW test statistic equals $W = 61$, and the Minitab listing gives $P = 0.14$ against a two-sided alternative. With a one-sided alternative H_a and the difference between the distributions in the direction of the H_a, we get the one-sided P by halving this value, i.e. $P = 0.07$. The test is therefore not significant but rather close. We have no evidence to confirm the researchers hypothesis about higher velocity values among cats with non-zero sct-I values, but the data show some indication in that direction.

As additional analysis it would be relevant to ask for confirmation that the outlying value is not an error, and if not it would be of interest to explore the impact of the outlier on the results. Without the outlier, a two-sample t-tests seems perfectly appropriate, and that test may have greater power than the MWW test used here.

Question 2

The paper referred to is Xu, Bates & Schweitzer (1993), *Public Opinion Quarterly* **57**, 232-237.

Subquestion a)

If we let X denote the number of households where an answering machine was picking up the call, the natural model is $X \sim \text{Bin}(1802, p)$ where p is proportion of households with an active answering machine. The statistical design is sampling from a finite population, and if the population under study is large enough (according to our guideline, at least 20 times larger than the sample) a binomial distribution is applicable. We observed $X_{\text{obs}} = 391$, and with the number of positives and negatives both very large, a classical confidence interval can be used.

$$\begin{aligned} \hat{p} &= 391/1802 = 0.217, \\ 95\%\text{ CI} &: 0.217 \pm 1.96\sqrt{0.217(1-0.217)/1802} = 0.217 \pm 0.019 = (0.198, 0.236) \end{aligned}$$

The 95% confidence interval does not include the value stated as the percentage of households with an answering machine attached. (Alternatively, one could compute a z-test for $\text{H}_0 : p = 0.25$; this yields $z = -3.24$ and $P = 0.001$.) Formally, that indicates a disagreement with the present study, so the statement that the proportion was consistent with the other study is not evident. Furthermore, it is not obvious that the parameter estimated was exactly the same. For example, the estimated proportion in the study would not include any households with an answering where the call was picked up by a person instead of the answering machine. The proportion in the study would therefore also most likely be affected by the time of the day these calls were made. So in summary, it may be difficult to attach any specific interpretation to the estimated proportion in the study.

Subquestion b)

The statistical design is independent samples from four populations. In each sample we observe (among other things) the number of surveys completed out of the total number of households contacted — a binomial setting. The statistical model is therefore independent binomial distributions, also termed a model for comparing multiple populations, or model I. In this question we consider three samples and the corresponding binomial distributions $\text{B}(n_i, p_i)$, i=A,B,C. The sample proportions are

$$\hat{p}_A = 48/100 = 0.48, \quad \hat{p}_B = 43/97 = 0.44, \quad \hat{p}_C = 43/94 = 0.46,$$

and the values are seen to be quite close. We test the hypothesis $\text{H}_0 : p_A = p_B = p_C$ against a two-sided alternative hypothesis by a chi-square statistic (from the Minitab listing): $X^2 = 0.272$, df $= 2$, $P = 0.87$. There is no evidence at all against H_0. It seems fair to say that the response rates seem to be the same no matter which type of message is left.

Subquestion c)

As there was no indication of a difference in response rates for the different messages, we pool these samples into a combined sample. The advantage is doing so is that we gain statistical power for the comparison with the control group (no message left). The statistical is two independent binomial distributions: $\text{B}(n_0, p_0)$ and $\text{B}(n_{ABC}, p_{ABC})$, and our interest is in the hypothesis H_0: $p_0 = p_{ABC}$. The alternative hypothesis is most naturally taken as one-sided: $\text{H}_a : p_0 < p_{ABC}$, because it is quite implausible that leaving a message decreases the response rate, and there would be no real interest in investigating that. The estimates are $\hat{p}_0 = 33/100 = 0.33$ and $\hat{p}_{ABC} = 134/291 = 0.46$, in agreement with the alternative hypothesis. The test statistic is

$$z = (\hat{p}_{ABC} - \hat{p}_0)/\sqrt{\hat{p}(1-\hat{p})(1/100 + 1/291)} = 2.2756,$$

where the pooled probability is $\hat{p} = (33 + 134)/(100 + 291) = 0.427$. The z-statistic correponds (approximately) to the 98.85% percentile of N(0,1), so $P = 0.0115$. There is clear evidence against H_0, and we conclude that leaving a message indeed improves the response rate. Note that $z^2 = 5.178$, the chi-square statistic for the two-way table made up of the two samples. The P-value of the chi-square test is twice the one determined here because it uses a two-sided alternative. Note also that the IPS guidelines for use of both the z-statistic and chi- square test are easily satisfied (in a) as well).

Subquestion d)

The most obvious point to critisize in the text of the paper is that it refers to a t-test, which is totally inappropriate in the context of two-way table analyses. It is not a typo because references to t-tests appear repeatedly in the paper. However, the above calculation shows that the t-value is the same as our z-value, and the P-values are also the same. Therefore, the correct method has probably been used, it is just referred to incorrectly. The P-value is that for a two-sided alternative, which seems less natural here, as discussed above. Further, and minor, points of criticism are listed below:

- The text may give the impression that there were substantial differences in the response rates for the three message types, but that these failed to yield statistical significance. This is somewhat misleading because the differences in reponse rates were small and very far from statistical significance.
- One percentage (41.4) is reported with an extra decimal compared to the others, for no obvious reason.
- The analysis of refusals is almost complementary to the analysis of surveys completed, because the two numbers add up (except for the no message group where one household is not accounted for) to the total number of contacts. The only difference between the analyses carried out is in the denominator, presumably corresponding to those households where the researchers were unsuccessful in establishing a contact. It might therefore have been better to analyze the data as a categorical outcome, with the categories: survey completed, survey refused, no contact established.
- To denote the significance level of 0.05 as "normal" is unusual and potentially misleading (it has nothing to do with the normal distribution); it would be better to refer to it as "standard" or "common".

Question 3

Subquestion a)

All the models shown in the Minitab listing are linear regression models. Let x_i and Y_i denote the dose of vitamin K and the concentration of coagulation reagent for rat i, $i = 1, \ldots, 10$. The dose is an explanatory variable, and the concentration is a response variable. Therefore a possible statistical model is a linear regression for Y:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where the errors $\varepsilon_1, \ldots, \varepsilon_{10}$ are i.i.d. and $\sim \mathrm{N}(0, \sigma)$. Results for this model are shown in the first Minitab listing. Below we summarize the different types and their suitability for the data.

1) Y = reagent, x = dose; poor fit, due to clearly non-linear relation,
2) Y = ln(reagent), x = dose; decent fit but still a clearly curved shape, so model not ideal,

3) $Y = \ln(\text{dose})$, $x = \ln(\text{reagent})$; meaningless model because does is not a response variable,

4) $Y = \ln(\text{reagent})$, $x = \ln(\text{dose})$; nice fit with very high R^2 and no apparent violations of model assumptions,

5) $Y = 1/\text{reagent}$, $x = \text{dose}$; decent fit but still a clearly curved shape, so model not ideal.

Based on the above discussion, the fourth model (corresponding to relation (2) in the text) is the best choice of model. Note that R^2-values cannot meaningfully be compared across different observation scales, so we should not base model choice for these data only on R^2-values.

Subquestion b)
The fourth Minitab listing gives parameter estimates and standard errors, and in addition we need $t^* = t_{.95}(8) = 1.86$ for the 90% confidence intervals,

$$\begin{aligned}
\hat{\beta}_1 = -1.761, \quad \text{SE}(\hat{\beta}_1) = 0.068, \quad 90\%\ CI : -1.761 \pm 1.86 \cdot 0.068 = -1.76 \pm 0.13 = (-1.89, -1.63),\\
\hat{\beta}_0 = 6.734, \quad \text{SE}(\hat{\beta}_0) = 0.113, \quad 90\%\ CI : 6.734 \pm 1.86 \cdot 0.113 = 6.73 \pm 0.21 = (6.52, 6.94),\\
\hat{\sigma} = 0.138.
\end{aligned}$$

The relation (2) shows that the parameter b equals the slope in the fitted model; therefore the estimate and CI for b equals those for β_1 above. Relation (2) furthermore shows that the parameter a is related by the intercept β_0 as $\beta_0 = \log(a)$, or $a = \exp(\beta_0) = \text{e}^{\beta_0}$. We therefore get

$$\hat{a} = \exp(\hat{\beta}_0) = \exp(6.734) = 840.5, \quad 90\%\ \text{CI} : (\exp(6.52), \exp(6.94)) = (678.5, 1033).$$

Note that a has an interpretation as the expected concentration for a dose of 1.

Subquestion c)
In our selected model, the hypothesis of no impact of vitamin K dose on blood coagulability corresponds to a slope of zero, that is: $\text{H}_0 : \beta_1 = 0$. The expected relation between vitamin K dose and coagulability is that higher doses facilitate coagulation, and hence a lower concentration of the coagulation reagent is required. Our alternative hypothesis should therefore be $\text{H}_a : \beta_1 < 0$. The estimated slope is indeed negative. The Minitab listing gives the t-test for H_0 as $t = -26.02$ with $P < 0.0005$ for the two-sided alternative hypothesis. Our P-value should be half of this value, but because the two-sided P is already so small, the additional halving has no real impact. We conclude that the association between vitamin K dose and blood coagulability is very strong and in the direction of the expected relation. The very good fit of the model further adds to the impression of a strong relation because the model would offer precise predictions.

Subquestion d)
The new observation of $(6.17, 27.5)$ corresponds to $(1.82, 3.31)$ at the natural log scales. The Minitab print gives a prediction of $3.52854 \approx 3.53$ for a log-dose of 1.82, as well as 95% confidence and prediction intervals. Because our information corresponds to a new observation, the observed value should be compared to the prediction interval of $(3.19, 3.865)$. It is well inside the interval, and we therefore conclude that this new observation is in good agreement with the model.