Biostats VHM 801 Course
Fall 2022, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to Mid-term Exam, November 2022

The solution is much more detailed than required for a 100% score, by its length and by including discussions of several options for some of the questions, in particular for both parts of **b)**.

Question 1

Subquestion a)

We base our descriptive analysis on the graphical summary for time at original scale. Assuming the observations for the 27 air conditioning units to form a simple random sample (or to be i.i.d., independent and identically distributed), the distribution can be described as follows:

- continuous because time since repair can take interval values, even if the time variable is recorded as integers,
- apparently unimodal with a mode towards the lower end of a wide range from 1 to 216 hours,
- clearly right-skewed as seen in the histogram and the boxplot, as well as by the median being less than the mean and the skewness value of 0.85,
- no suspected outliers indicated in the boxplot,
- with the right-skewness the distribution is definitely not normal (also, the A-D test is significant at $P = 0.025$ and offers formal evidence against the normal distribution).

Due to the right-skewness, the mean and standard deviation do not summarize the distribution well. Instead, the 5-number summary is suggested, its values are: $1 - 18 - 63 - 111 - 216$.

Subquestion b)

The two parts *i)* and *ii)* are done separately and by different methods.

Part i)

For a parametric calculation, we will want to assume a normal distribution and compute the probability in that distribution (with its estimated parameters). Effectively, this is a probability calculation, because we do not have the methods to supplement the calculated probability with the uncertainty (i.e., a standard error) coming from estimating the parameters. The only scale among the three scales shown (original, square-root and log-transformed) where a normal distribution assumption seems reasonable is the square-root scale. At that scale, the distribution looks quite symmetrical, and the A-D normality test is totally non-significant ($P = 0.86$). We let X denote the square-root transformed operating time for an airconditioning unit. Using the estimates from the graphical summary, we will work from $X \sim \mathrm{N}(7.9294, 3.8046)$. Our interest is in operating times larger than 100, which corresponds to values on square-root scale larger than $\sqrt{100} = 10$. Then,

$$\begin{aligned} P(X > 10) &= P\left(\frac{X - 7.9294}{3.8046} > \frac{10 - 7.9294}{3.8046}\right) = P(Z > 0.5442) \\ &\approx P(Z > 0.54) = P(Z < -0.54) = 0.2946 \approx 0.295, \end{aligned}$$

using Table B of the PSLS textbook.

Part ii)

Contrary to *i)*, this part involves statistical inference. Without any distributional assumptions we will use the sample proportion of units with operating times above 100 hours. This corresponds to a dichotomous (binary) outcome for each of the 27 air conditioning units, namely whether their operating time exceeds 100 or not. The binomial setting leads to a binomial distribution: $X \sim \mathrm{B}(27, p)$. The observed count of units that lasted longer than 100 hours is $X_{\text{obs}} = 8$, and thus $\hat{p} = X_{\text{obs}}/27 = 8/27 = 0.296$. For confidence intervals on a single proportion we can use the classical (normal approximation) method if the number of events and non-events both exceed 15; this condition is, however, not met with only 8 events. Instead we can use the "plus four" method because the sample size (27) exceeds 10. The plus-four adjusted proportion is $\tilde{p} = (8+2)/(27+4) = 10/31 = 0.3226$. For a 95% CI we will use $z^* = 1.96$, and the calculation goes as follows:

$$\begin{aligned} 95\% \text{ CI:} \quad &= \tilde{p} \pm 1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} = 0.3226 \pm 1.96\sqrt{\frac{0.3226 \cdot (1 - 0.3226)}{31}} \\ &= 0.3226 \pm 0.1645 = (0.158, 0.487). \end{aligned}$$

It is seen that the estimated proportion in the data is quite close to the calculated probability in i), reflecting that the normal distribution approximation is good.

Subquestion c)

We need to employ probability calculations for the specific settings described. For each question, a student can select either the right or wrong answer. For a student that is purely guessing, the probability of getting the answer right is the same for all questions, therefore a binomial setting applies. If X denotes the number of correct answers, we have $X \sim \mathrm{B}(20, p)$, where p is the probability of getting the answer right. For a choice between two answers of which only one is right, $p = 1/2$. The probability of passing the exam is $\mathrm{P}(X \geq 12)$, corresponding to at least 60% correct answers, and using Table 1 from the Stevens textbook we get:

$$\begin{aligned} \mathrm{P}(X \geq 12) \quad &= \quad P(X{=}12) + P(X{=}13) + P(X{=}14) + P(X{=}15) + P(X{=}16) + P(X{=}17) + P(X \geq 18) \\ &= \quad 0.120 + 0.074 + 0.037 + 0.015 + 0.005 + 0.001 + 0 = 0.252 \approx 0.25 \ \text{ or } 25\%. \end{aligned}$$

For a choice between three answers, the probability of getting one answer right is $p = 1/3$. We need to recompute the probability of $X \geq 12$ with $p = 1/3$, but the table only has entries for $p = 0.3$ and $p = 0.4$. We can compute those two probabilities to achieve a range for the probability in question.

$$\begin{aligned} p = 0.3: \quad \mathrm{P}(X \geq 12) = \quad &P(X{=}12) + P(X{=}13) + P(X \geq 14) = 0.004 + 0.001 + 0 = 0.005, \\ p = 0.4: \quad \mathrm{P}(X \geq 12) = \quad &P(X{=}12) + P(X{=}13) + P(X{=}14) + P(X{=}15) + P(X \geq 16) \\ = \quad &0.035 + 0.015 + 0.005 + 0.001 = 0.056. \end{aligned}$$

Thus, the probability of passing the exam is between 0.5% and 5.6%. An exact calculation using statistical software gives the value 0.01297, or 1.3%; as one might expect, this value is closer to the probability for $p{=}0.3$ than for $p{=}0.4$.

Subquestion d)

The situation described involves statistical inference based on the information provided: the class of 25 students had an average score of 0.70 (or 70%, but this solution will use the decimal representation), and we want to compare this result with the country-wide average of 0.73. We will assume the class scores, say $X_1, \ldots, X_{25}$ for the 25 students in the class, to form a simple random sample (or be i.i.d.) from $\mathrm{N}(\mu, \sigma)$, where μ is the class-specific (or instructor-specific) mean. We estimate μ by the sample mean, i.e. $\hat{\mu} = \bar{X} = 0.70$, and wish to carry out a statistical test for the hypothesis $\mathrm{H}_0 : \mu = 0.73$. For the calculations we will need a value for the standard deviation; it could be suggested that the professor obtain the sample standard deviation among the 25 test scores, and proceed along the lines of a 1-sample t-test.

We are also told that test scores across the country have a 95% range of $(0.55, 0.91)$. This is *not* a confidence interval; it is a statement about the distribution of individual test scores. The country-wide average must be estimated much more precisely than such a wide interval. If it had been a confidence interval, we could have inferred from the sample mean (0.70) being easily within the interval, that there was no evidence whatsoever that the specific class differed in its average from the country-wide average. In a normal distribution, the 95% range is computed as $\mu \pm 1.96 \cdot \sigma$ or for simplicity $\mu \pm 2\sigma$. Assuming a normal distribution for scores across the country (we have no way of validating this, but this is a necessary assumption in order to proceed), we can therefore reconstruct the population standard deviation as: $\sigma = (0.73 - 0.55)/2 = 0.09$. Note that the 95% range is symmetrical about its mean which supports the assumed normal distribution. Without access to the sample standard deviation, we can work with this standard deviation, and it seems most natural to take it as a known value, which in turn leads to z-inference. As mentioned above, our null hypothesis is $\mathrm{H}_0 : \mu = 0.73$. The alternative hypothesis is perhaps most naturally taken as two-sided, $\mathrm{H}_a : \mu \neq 0.73$ (a one-sided alternative could be argued as well), because the professor would not know in beforehand whether the students were below or above the national average. Using Table B of PSLS, we get:

$$\begin{aligned}
\hat{\mu} &= \bar{X} = 0.70, \quad \mathrm{SE}(\hat{\mu}) = \sigma/\sqrt{n} = 0.09/\sqrt{25} = 0.018, \\
z &= \frac{\bar{X} - 0.73}{\mathrm{SE}(\hat{\mu})} = -1.667 \approx -1.67, \quad P = 2 \cdot \mathrm{P}(Z < -1.67) = 2 \cdot 0.0475 = 0.095.
\end{aligned}$$

There is no statistical evidence of a different performance of this class than the national average. It would be legitimate to use a t-test instead, assuming the sample standard deviation to be equal to the population value of $\sigma = 0.09$. This approach would lead to a slightly larger P-value, obtained from the t-distribution with $\mathrm{df} = n - 1 = 24$; the critical value for significance at the $\alpha = 0.05$ level is $t_{.975}(24) = 2.064$, from Table C of PSLS. The conclusion remains the same: there is no convincing evidence that the class performed unusually, and the professor does not need to worry (too much).