Solution to home assignment II

The solution is more detailed than required for a 100% mark, by including detailed calculations, discussions and interpretations, including multiple ways of answering the different questions.

a) Statistical comparison with population mean value

We consider a single sample (or i.i.d. observations) from a population, in this case the 2013 population of dairy farms in the Maritime provinces of Canada. The interest is in comparing the population mean with the overall Canadian mean from 2012. The question states that we should consider the 2012 mean as a value with negligible uncertainty. Because we do not know the population standard deviation, our inference will be based on a t-distribution.

Denote by $X_1, \ldots, X_n$, with $n = 70$, the average 305-day milk productions (in kg) for the herds, and assume these to be i.i.d. from $\mathrm{N}(\mu, \sigma)$. The first home assignment included a descriptive analysis of this variable, showing that looks reasonably normal, despite a left-tail suspected outlier; therefore, t-distribution inference is well justified. We estimate the two unknown parameters by their sample values: $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = s_X$. Then we test the hypothesis $\mathrm{H}_0 : \mu = 9979$ by a t-test. We use a two-sided alternative hypothesis, $\mathrm{H}_a : \mu \neq 9979$, because there is nothing in the wording of the research question to suggest otherwise. The calculations are:

$$\hat{\mu} = \bar{X} = 9532.2, \quad \hat{\sigma} = s_X = 1096.2,$$
$$t = \frac{\bar{X} - 9979}{s_X/\sqrt{n}} = \frac{9532.2 - 9979}{1096.2/\sqrt{70}} = -3.410, \quad P = 2 \cdot \mathrm{P}(t(69) > 3.410) = 0.0011 \text{ (software)}.$$

Using a statistical table, from the value $t_{.999}(60) = 3.232$ we would conclude that $P < 0.002$. We reject the null hypothesis; thus, the data give clear evidence that the 2013 milk production in the population represented by the sample is different from, and indeed less than (as seen by the sample mean), the 2012 Canadian milk production. The 95% CI for μ (using $t^* = 1.995$ from software) is: $9532.2 \pm 261.4 \approx (9271, 9794)$, well below the population value.

If the deviation from the population value was due to bias in the sampling procedure (as indicated in the Home assignment I, the herds were selected by a more complex procedure than simple random sampling), there could be a real concern that the data would not be representative of the 2013 population in the Maritime provinces. However, the main reason for the discrepancy is most likely that dairy herds in the Maritime provinces are different from those across the country in general; in fact, data at the Canadian Dairy Information Centre indicate that the 2012 milk production was lower in the Maritime provinces. A more appropriate comparison value would therefore have been a suitably computed 2012 average milk production for the Maritime provinces, but we will not pursue this further.

b) Statistical comparison of herds with different BLV-infection status

The classification of herds as BLV-infected by the `herd_blv` variable separates the 70 herds into 61 BLV-infected and 9 BLV-noninfected herds. The herds in the two groups are now considered as independent samples from distinct populations, and we assume the milk productions in the two samples to come from separate normal distributions.

In notation, the values $X_1, \ldots, X_{n_1}$, with $n_1 = 9$, in the BLV non-infected group follow $\mathrm{N}(\mu_1, \sigma_1)$, and the values $Y_1, \ldots, Y_{n_2}$ in the BLV-infected group, with $n_2 = 61$, follow $\mathrm{N}(\mu_2, \sigma_2)$. A descriptive analysis for the two samples shows the distributions as reasonably normal, despite some skewness among the non-infected herds (due to one somewhat extreme herd, but it is a very small sample) and two suspected outliers among the infected herds (both A-D tests are nonsignificant at $P = 0.245$ and $P = 0.446$, respectively); again, t-distribution inference seems justified. We formulate the null hypothesis of equal means, $\mathrm{H}_0 : \mu_1 = \mu_2$, and it seems most natural to use a two-sided alternative based on the inconsistent information in the literature referred to in the question. It would also be valid to use the one-sided alternative $\mathrm{H}_a : \mu_1 > \mu_2$ based on the biological hypothesis. We now calculate,

$$\hat{\mu}_1 = \bar{X} = 9375.6, \quad \hat{\mu}_2 = \bar{Y} = 9555.3, \quad \hat{\sigma}_1 = s_X = 1098.5 \quad \hat{\sigma}_2 = s_Y = 1103.0,$$
$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{(s_X^2/n_1) + (s_Y^2/n_2)}} = \frac{9375.6 - 9555.3}{\sqrt{(1098.5^2/9) + (1103.0^2/61)}} = -0.458, \quad P = 0.66 \text{ (Minitab)}.$$

Minitab computed the P-value as: $P = 2 \cdot \mathrm{P}(t(10) > 0.458)$. Different software algorithms produce minor differences in the approximate degrees of freedom for the t-statistic, without affecting the conclusion. The P-value for the one-sided alternative $\mathrm{H}_a : \mu_1 > \mu_2$ equals: $P = \mathrm{P}(t(10) > -0.458) = 0.67$ (Minitab). Note that the P-value is greater than 0.5 because the observed means are in the opposite direction of the alternative hypothesis. Note also that we follow the PSLS recommendation against the t-test based on a common variance assumption $\sigma_1^2 = \sigma_2^2$ ($t = -0.457$, $\texttt{df} = 68$, $P = 0.65$). Regardless of the chosen H_a, there is no evidence whatsoever against the null hypothesis, and hence the present data give no indication of differences in milk production in BLV-infected and non-infected herds. It could be said that the sample size is small and unbalanced between the two samples. Even so, with the sample means being very close and also showing the opposite trend, there is no support at all for the biological hypothesis, and we have no reason to hypothesize that an effect might be seen in a larger sample.

c) Statistical comparison of cows with different BLV-infection status

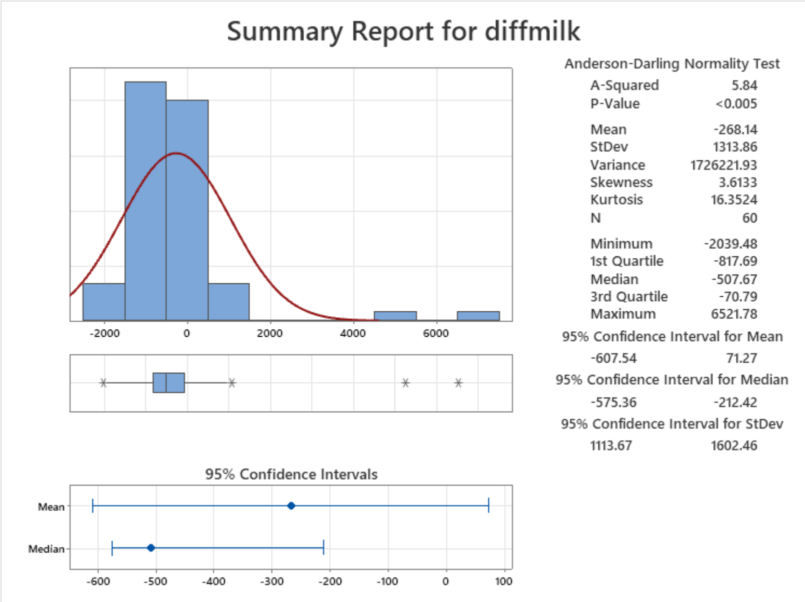
The average milk production values for BLV-infected and non-infected cows within the same herd are paired (or dependent) because the cows in the two groups shared the herd environment. Some herds have higher average milk production than others, for other reasons than their BLV-infection status, and therefore milk production values for cows in the same herd cannot be assumed independent. With our two values from each herd, the design can be considered as two paired samples. We could also say that the herds form blocks. Not all herds have both groups of cows represented, and these herds therefore do not contribute to the comparison between the two groups in a two paired samples design. More advanced methods (beyond VHM 801) exist that would allow inclusion of all the herds, but these also require additional assumptions. The statistical analysis will be based on the differences, say non-infected minus infected, for the herds with both groups represented.

Let $D_1, \ldots, D_n$ ($n = 60$) denote the differences in milk production (`milk305_blv0-milk305_blv1`) within each herd. Almost all (9 out of 10) of the 10 herds not included had a herd BLV-status of 0, meaning that there were no BLV-infected cows in the herd; in one herd (with BLV-status of 1), all the cows were BLV-infected. The natural model is a single sample (of differences) assumed to follow $\mathrm{N}(\mu_D, \sigma_D)$, and the focus of our interest will be the mean difference μ_D and the hypothesis $\mathrm{H}_0 : \mu_D = 0$, again assuming a two-sided $\mathrm{H}_a : \mu_D \neq 0$. Proceding in a similar way as for **a)**, we get

$$\hat{\mu}_D = \bar{D} = -268.1, \quad \hat{\sigma}_D = s_D = 1313.9,$$
$$t = \frac{\bar{D}}{s_D/\sqrt{n}} = -1.581, \quad P = 2 \cdot \mathrm{P}(t(59) > 1.581) = 0.119 \text{ (software)}.$$

The P-value is not significant, but not so far off. However, the mean difference is in the opposite direction of the biological hypothesis (and the P-value for the one-sided $\mathrm{H}_a : \mu_D > 0$ is $P = 0.94$).

Before proceeding to conclusions we should assess the model assumptions; a graphical summary for the differences (`diffmilk`) is shown below.



It is seen that the distribution includes two extreme (suspected) outliers; these can be identified as the values 6521.8 and 5262.7 *kg* for herds 15 and 22, respectively. These values correspond to situations where the average milk production in the infected group was more than 5000 *kg* lower than in the non-infected group; that seems as a very large difference (compared to the other differences). The A-D normality test is clearly significant and gives strong evidence against the normal distribution.

By comparing with each herd's total average milk production (`milk305`), it is also seen that this average is much closer to the average for the non-infected group, and in fact the infected cow averages in these two herds seem very low when compared to other herds. This suggests that there may only have been very few infected cows in these herds, making the average very uncertain and perhaps also unrepresentative of the average milk production among BLV-infected cows. Altogether, we have some indication that these herds are real outliers.

At this point, our options for further analysis are:

i) stick to original analysis, arguing that the robustness of t-procedures would make it approximately valid: in view of the extreme outliers, this does not seem tenable (despite $n = 60$);

ii) use a non-parametric procedure in order to reduce sensitivity of the analysis to the outliers: this would be fine, but it was not really an option at the time the home assignment was done;

iii) reanalyze without the two outliers: this is useful, at least to explore the impact these outliers have on the results obtained above.

For option *iii*), we obtain a much nicer distribution that looks approximately normal without the two outliers, and the significance test of $\text{H}_0 : \mu_D = 0$ against $\text{H}_a : \mu_D \neq 0$ gives $t = -5.82$ and $P \ll 0.001$, with a 95% CI for μ_D of $(-646, -315)$. Therefore our conclusion above did actually depend strongly on the outliers, and without them there would strong significance against equal milk productions in BLV-infected and non-infected cows, with the former group having the *largest* milk production. A nonparametric analysis of data for all 63 herds by a sign test (the Wilcoxon signed rank test is less natural here due to its assumed symmetry) leads to the same finding.

In conclusion, the analyses in **b**) and **c)** both failed to demonstrate a lower milk production in BLV-infected than non-infected herds and cows. Actually the average milk productions were higher in BLV-infected than non-infected herds and cows. The finding for herds was completely non-significant and should therefore not lead us to speculate about a real effect. The finding for cows was very strong when two herds with suspect looking average milk productions for infected cows were removed or a non-parametric test was used. Further exploration is needed to decide how much weight we can attribute to the finding, in particular it would be of interest to know how many cows these extreme averages represented. One could also suggest to explore options for analysis of the data for all cows.

d) Sensitivity and specificity of BLV-classification

The BLV infection status (`herd_blv`) effectively splits the 70 herds into two groups corresponding to infected and non-infected herds. For the solution we will adopt terminology from diagnostic testing, and refer to this classification as the true (or gold standard) status. Our "test" classification is the designation of herds with a bulk-tank ELISA above 5 as infected, and non-infected otherwise. If we denote the number of samples from infected and non-infected herds by n_1 and n_2 (where $n_1+n_2 = 70$), and furthermore the number of herds that "tested" BLV-infected (*positive*) among the truly infected herds by X_1 and the number of herds that "tested" non-infected (*negative*) among the non-infected herds by X_2, our statistical models are:

$$X_1 \sim \text{Bin}(n_1, p_1), \quad \text{and} \quad X_2 \sim \text{Bin}(n_2, p_2),$$

where p_1 is the sensitivity (Se) and p_2 is the specificity (Sp) of the test. It seems reasonable to assume binomial settings for the two samples, even if we have more detailed information about the herds. In a practical application of the "test", one would typically not have such extensive information about the cows in the herd. Estimates and 95% confidence intervals are given in the table below. In both cases, the conditions for using the classical confidence intervals are violated, so we use the "plus four" and exact intervals instead.

Statistic	Formula	Se (`herd_blv`=1)	Sp (`herd_blv`=0)
# obs.	n	61	9
count	X	57	9
estimate	$\hat{p} = X/n$	0.934	1
"+4 estimate"	$\tilde{p} = (X+2)/(n+4)$	0.908	0.846
95% CI	$\tilde{p} \pm 1.96\sqrt{\tilde{p}(1-\tilde{p})/(n+4)}$	$(0.837, 0.978)$	$(0.650, 1)$
95% CI	"exact" (Minitab)	$(0.841, 0.977)$	$(0.684, 1)$

For the sensitivity, the condition for use of the "plus four" interval is easily met ($n_1 \gg 10$), and this interval is a sensible choice; the exact interval is however seen to be very similar. The ability of the "test" to detect a BLV-infected herd is good (estimated at 0.93, and with 95% confidence above about 0.84). To get a more narrow confidence interval, one simply needs a larger number of infected herds.

For the specificity, the condition for use of the "plus four" interval is not met, and an exact interval is preferable. The interval above is the one provided by Minitab ("Adjusted Blaker method"), whereas the Clopper-Pearson confidence interval offered by the `binom.test` function in R and by default in Stata equals (0.664,1). (In Stata, this is, perhaps confusingly, labeled as a one-sided 97.5% CI, but it is the appropriate two-sided 95% CI.) For the purpose of the assignment, any of these answers are considered acceptable. It is seen that despite the perfect classification of the 9 non-infected herds the confidence interval is very wide, and the confidence obtained is still quite vague or uncertain. There is definitely a need for a larger sample of non-infected herds to evaluate the "test".