## Solution to home assignment I

The solution presents one way in which the descriptive analysis could be done, but other ways are possible as well. It is more detailed than required for a 100% mark, by including all the variables for the descriptive analysis when only 2 or 3 selected variables were required in different parts of the assignment.

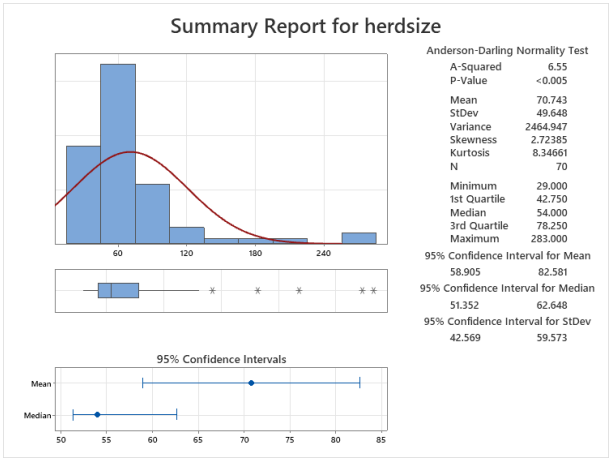
### a. Study and variable types

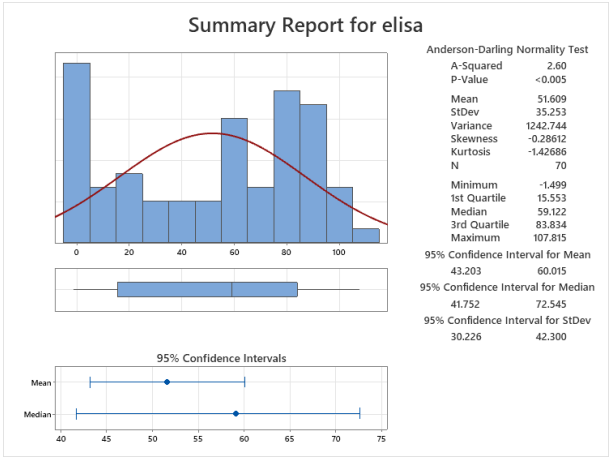
The study is an observational study because the researchers did not impose any treatment on study herds, and the research objectives were achieved by using the data collected from the herds in the study without any attempts to influence the observations.

The variables *herd* and *province* are nominal categorical; *herdsize* is quantitative discrete with 47 different numbers of herd sizes; *elisa*, *milk305*, *fat305*, *prot305* and *parity* are quantitative continuous. Note that the parity for each cow is a discrete variable, but the average parity across all cows in a herd is essentially continuous.

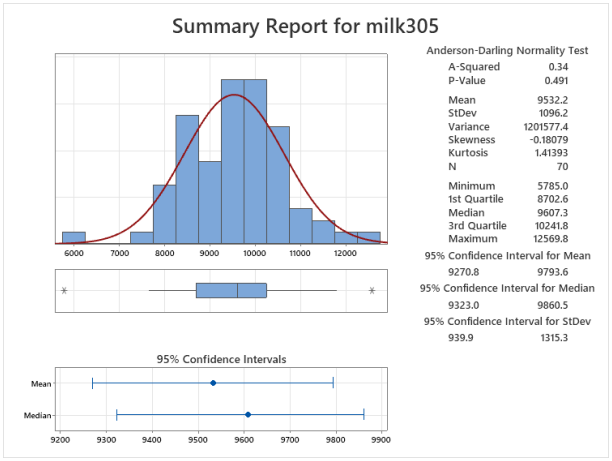
### b. Descriptive statistics

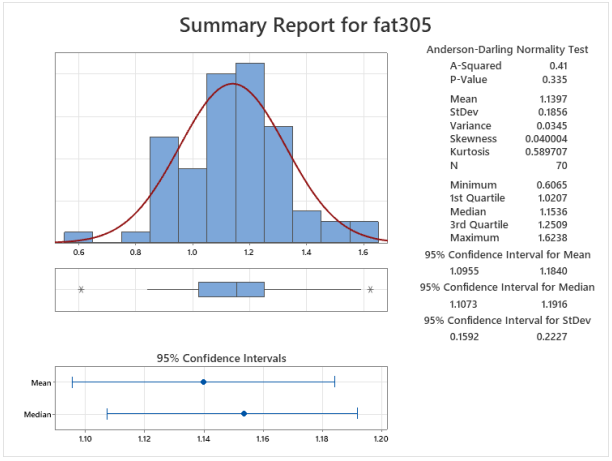
For simplicity the descriptive statistics and graphical displays for the continuous variables will use the Graphical Summary menu in Minitab. With the sample size of 70, the most appropriate graphical representation of the distribution is a histogram, and the display also includes a box-plot and all the commonly used descriptive statistics (plus some we won't consider here, such as the confidence intervals). The default number of bins in the histograms varies and in some cases seems too high, e.g. compared to the $\sqrt{70} \approx 8.4$ guideline. The cut-points for the "suspected outliers" indicated in the box-plot can be computed manually as the $(Q1 - 1.5 \cdot \text{IQR})$ and $(Q3 + 1.5 \cdot \text{IQR})$ on the lower and upper sides of the distribution, respectively. In a normal distribution, the expected number of suspected outliers from this rule in a sample of size 70 is $0.0035 \cdot 70 \approx 0.25$ in both the left and right tails (slide 1L−15). So a single suspected outlier is not implausible even for a variable that is perfectly normal.

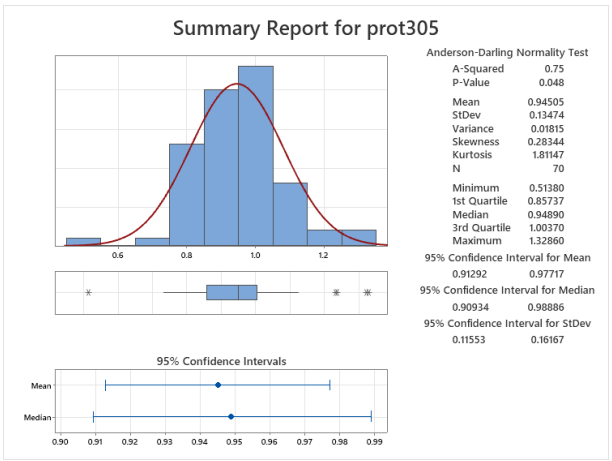


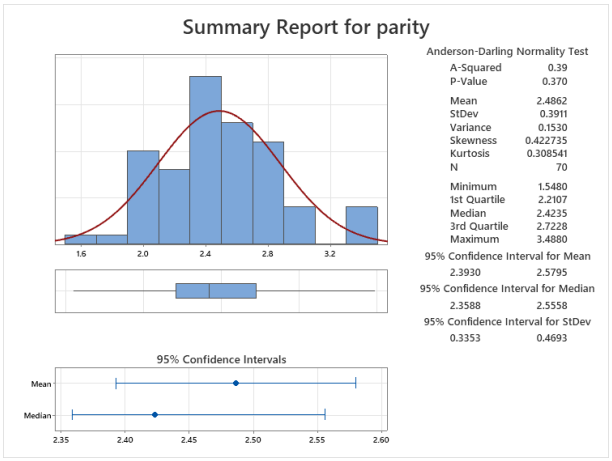




*(continues on the next page)*

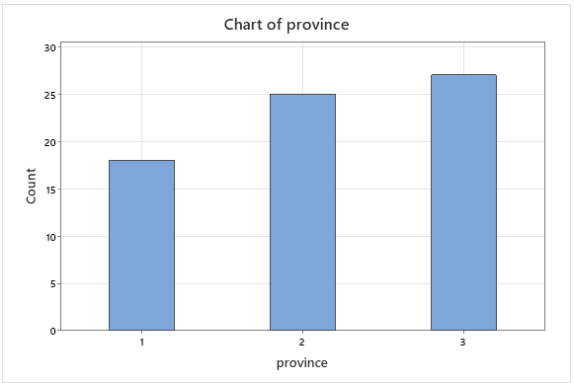



For the categorical variable *province*, descriptive statistics such as the mean and standard deviation and graphical displays such as a histogram (with overlaid normal curve) are less useful (and often quite meaningless). The relevant statistics are the counts or proportions for each category, and graphical representation by a bar graph (right) or a pie chart.




| Province | Frequency | Proportion |
|---|---|---|
| 1 (NB) | 18 | 0.257 |
| 2 (NS) | 25 | 0.357 |
| 3 (PE) | 27 | 0.386 |

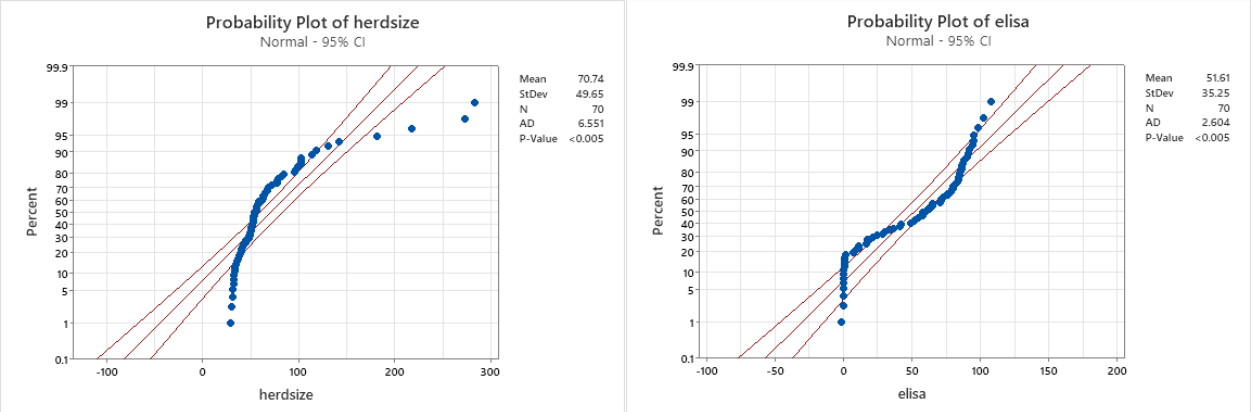
Finally, brief verbal summaries of the distributions based on the computed statistics and graphs:

- *herdsize*: unimodal with a center around 55; strongly right-skewed (skewness = 2.7); an IQR of 35 signals a moderate spread in the center of the distribution, but the right tail is very long; five "suspected outliers" in the right tail, but none of these are apparent errors (just large herds), and there is no reason to suspect they should be real outliers; if very large herds are considered substantially different from the rest of the population, one might consider restricting the population to include only herds up to a certain size, but that will reduce the applicability of study results,

- *elisa*: most naturally described as bimodal with one peak close to zero and another peak around 80; overall slightly left-skewed (skewness= $-0.29$) but this reflects both the peak at zero and a possible left-skewness in the remainder of the distribution; the distribution has narrow tails, reflected by a negative kurtosis (-1.4), with a large spread; no observed "suspected outliers",
- *milk305*: unimodal; centered around 9500 kg; symmetrical distribution with (negative) skewness close to zero and a moderately elevated kurtosis (1.41) indicating heavier tails than a normal distribution, probably mostly due to two "suspected outliers" ($5785 < 6394$ kg in the left tail, and $12570 > 12551$ kg in the right tail, with the suspected outlier cutoffs indicated as well) that both seem as reasonable observations (not errors), but with the lowest value being a bit far from the others, it is not obvious whether it could be considered as a real outlier,
- *fat305*: unimodal; centered around 1.15 kg; symmetrical distribution with skewness very close to zero and also a kurtosis quite close to zero; one "suspected outlier" in each tail, but only the low value could plausibly be a real outlier (similar to the discussion above),
- *prot305*: unimodal; centered around 0.95 kg; fairly symmetrical distribution with a skewness still quite close to zero (0.28), but a quite high kurtosis value (1.8) indicating heavier tails than a normal distribution, presumably caused by five "suspected outliers" (0.51 kg in the left tail, two values around 1.32 kg in the right tail and two lower values around 1.23 kg, just above the cutoff of: $1.0037 + 1.5 \cdot (0.1463) = 1.223$), but all looking reasonable and hardly real outliers (instead reflecting a distribution with heavier tails); note that the low values for *milk305*, *fat305* and *prot305* are all from the same herd, a finding that may be indicative of something strange about that herd and possibly worthwhile to explore further for the investigators,
- *parity*: unimodal; centered around 2.4; fairly symmetrical distribution with only a moderate skewness (0.42) and also a kurtosis quite close to zero; no "suspected outliers".

### c. Normal distribution

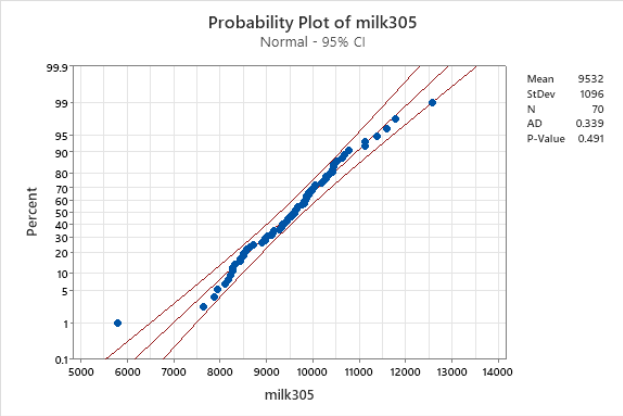
For each of the six quantitative variables, we show a normal probability plot and interpret the results.

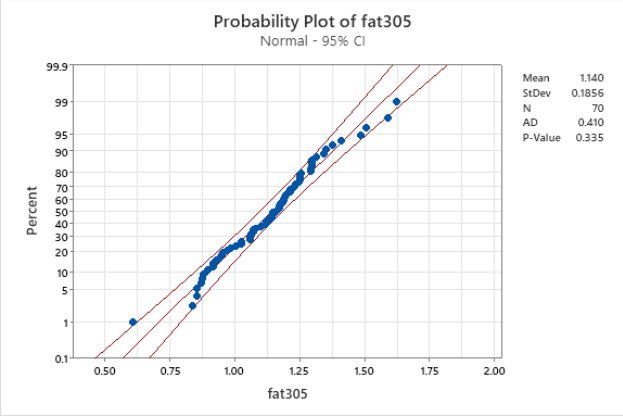



The distribution for *herdsize* is not normal; it was described above as clearly right-skewed, e.g. with its mean (70.7) much greater than the median (54). The AD (Anderson-Darling) normality test gives clear evidence ($P < 0.005$) against normality. Also, the probability plot for *herdsize* is strongly curved with systematic departures from a straight line and many points outside the bands.

The distribution for *elisa* is also not normal; it was described above as bimodal. The AD normality test gives clear evidence ($P < 0.005$) against normality. The probability plot for *elisa* is far from a

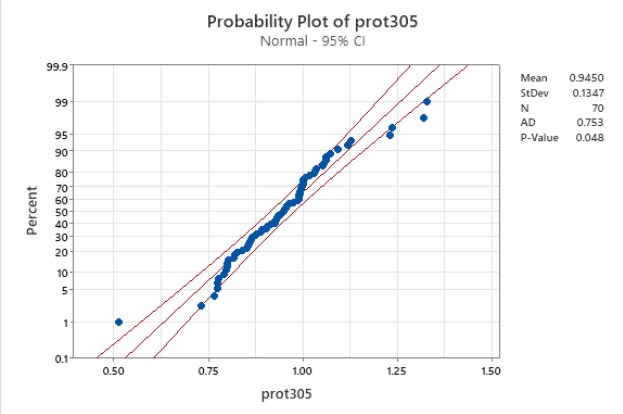
straight line; note how strongly the peak at zero shows at the left side of the graph.

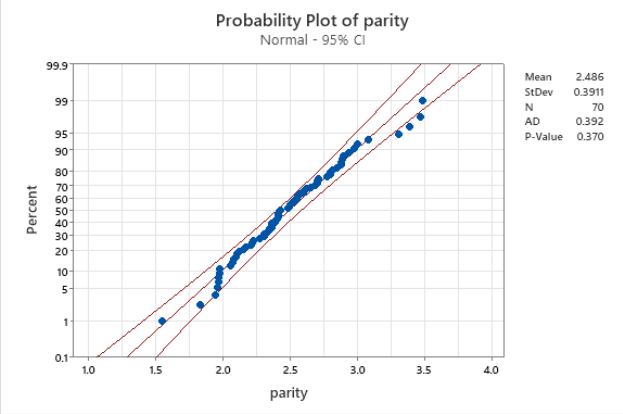



The distribution for *milk305* seems close to normal; the histogram is fairly symmetrical and the overlaid normal curve seems to match the distribution quite well. The normality test gives no evidence whatsoever against normality ($P = 0.49$). The probability plot is pretty straight (no systematic departures from a straight line), with only the one extreme point in the left tail clearly outside the bands.

The distribution for *fat305* also seems close to normal; the histogram is fairly symmetrical and the overlaid normal curve seems to match the distribution quite well. The normality test gives no evidence whatsoever against normality ($P = 0.34$). The probability plot is pretty straight (no systematic departures from a straight line), with a couple of points barely outside the bands.



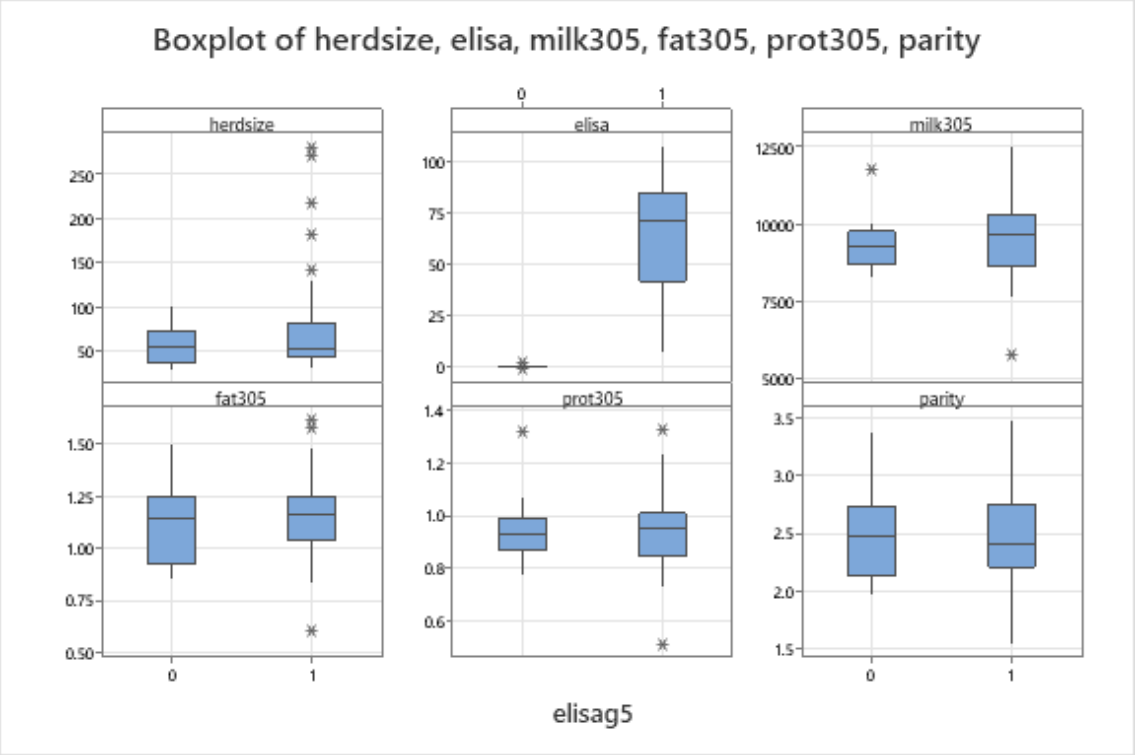

The distribution for *prot305* is symmetrical but shows some evidence against a normal distribution ($P = 0.048$). The normal probability plot seems close to a straight line, except for a few points at both ends. These extreme values are probably the reason why this distribution does not match the normal so well.

The distribution for *parity* seems close to normal; its histogram is fairly symmetrical and the overlaid normal curve matches the distribution well. The normality test gives no evidence whatsoever against normality ($P = 0.37$). The probability plot is pretty straight (no systematic departures from a straight line), with three points just outside the bands at the right tail.

### d. Comparing infected and non-infected herds

In order to dichotomize the ELISA test result at the cut-off value of 5, we can sort the values and manually indicate whether values are below or above the cut-off. In Minitab, we may use the *Data-Recode-To Numeric* menu, select the variable *elisa*, and code the ranges $(-2, 5)$ and $(5, 110)$ as 0 and 1, respectively; these values were chosen based on the *elisa* range from -1.5 to 108. There are 57 out of 70 observations with an ELISA test result greater than 5, corresponding to a proportion of infected herds of 0.814.

The focus in this part is on comparative statistics and graphics. We demonstrate the use of comparative box-plots to show differences in distributions of the other variables between the herd infection status groups. The variable *elisag5* takes the value 1 when *elisa* $> 5$, and 0 otherwise.




The six box-plots show some differences between the infected and non-infected herds. For *herdsize*, although both the medians and boxes are quite similar between the two groups, all five "suspected outliers" with large numbers of lactating cows are in the infected group. For *elisa*, the two groups naturally have very different distributions (by construction); the narrow range for non-infected herds contrasts the wide range for infected herds with an apparent left-skewed distribution. For *milk305*, the medians and box sizes in the infected group are larger than those in the infected group. For *fat305*, the medians are similar between the two groups, and although the displays do not look quite the same, the spread appears quite similar in the two groups. For *prot305*, it seems that the median protein yield in the infected herds is slightly larger than that in the non-infected herd group, but overall the two distributions look quite similar. For *parity*, the non-infected herds may be centered slightly higher than the infected herds and also have a slightly larger spread, but the differences appear small.

*Means of quantitative variables in infected and non-infected herds:*

| Group | *herdsize* | *elisa* | *milk305* | *fat305* | *prot305* | *parity* |
|---|---|---|---|---|---|---|
| *elisa* $\leq 5$ | 58.5 | 0.26 | 9351 | 1.119 | 0.949 | 2.505 |
| *elisa* $> 5$ | 73.5 | 63.3 | 9573 | 1.144 | 0.944 | 2.482 |

For the interpretation of these means, we should really also consider their standard errors, but this is considered beyond the scope of the home assignment. Let it suffice here to say that the differences for *milk305*, *fat305*, *prot305* and *parity* appear small, considering the range of values in these distributions. The huge difference for *elisa* occurred by construction. For *herdsize*, there may be a difference between herd infection status groups, with infected herds having an average larger number of lactating cows.

**e. Selection of herds**

Random selection within predefined groups is called stratified sampling. The random selection may be done for each group separately, or one may use a combined randomization procedure. Let us to begin with consider a single group: category 1 with 23 herds. In Minitab we would then input the herd numbers into a single column, and use the *Calc-Random Data-Sample from Columns* menu to randomly reorder these numbers, preferably after having set a seed (base) to allow us to reconstruct the process if needed.

Next, if we want to randomly select herds in all groups in one process, we can use a similar approach as described in the solution for Supplementary Exercise 3.40. Specifically,

1) number the herds from 1 to 142;

2) generate category labels (23 1's, 15 2's etc.), either manually or using coding to produce patterned data (not straightforward from the menus in Minitab);

3) generate a uniform random number (say between 0 and 1) for each herd, and sort all the data by the random numbers within each category, (i.e., sort on two variables: category first and then random number);

4) select the six first herds within each category in the sorted data.

The 30 selected herds do not accurately represent the entire population of herds despite that all infection categories are included (by construction). Because the total numbers of herds per category are different, the six herds in a large category (e.g., category 4) represent more herds in the population than those in a smaller category (e.g., category 2). So in a sense, the selected herds in the different categories have different weights relative to the population. If one wants to estimate population values (for example means), one needs to account for the unequal weights; different methods exist, but this topic is well beyond the scope of the course.