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

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. All analyses shown used Minitab, but Stata or other programs would give similar figures and results.

a) Study and variable types

The study is *observational*, as opposed to an experiment, because no treatments or other interventions were undertaken. The data may be considered as a simple random sample from the population of heifers in Belgian herds (at that time). Herd information is not included with the data but would typically be available.

Among the four variables (two versions of somatic cell counts are given), breed is a nominal categorical variable (because the values 0−3 are labels and have no numerical meaning or logical order), and the other variables are quantitative. The quantitative variables are all on a continuous scale, except for the discrete (daily) time scale for days in milk. It can be said that the time since calving is truly continuous and only recorded at a discrete scale, but with only 10 possible values the distribution will inevitably have a discrete character.

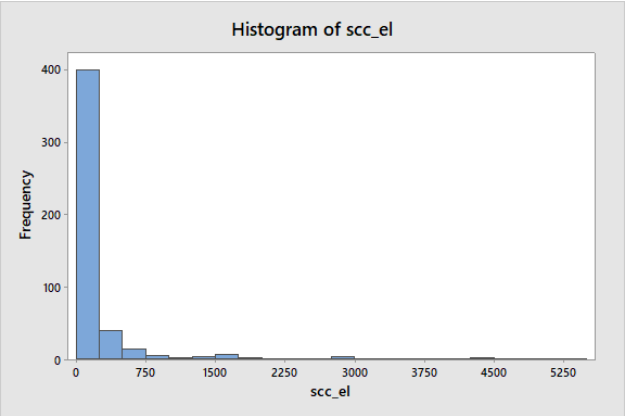
b) Descriptive analysis

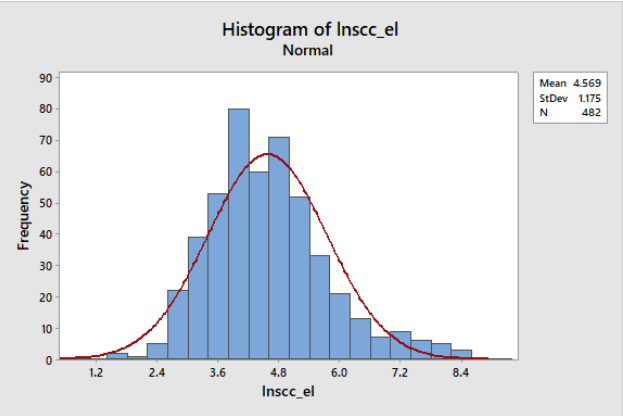
The descriptive analysis consists of a number of graphs/plots and of the most common descriptive statistics for the quantitative variables (see table below). The list of descriptive statistics should as a minimum include the mean, median and standard deviation. With a sample size of several hundred observations, a histogram is the most appropriate figure for the distribution of continuous variables; the number of bins can be adjusted to around $\sqrt{482} = 22$, and the bins should not include impossible values (e.g. < 0 for `scc_el`). The days in milk variable does not have interval values and its distribution should either be shown in a chart with one bar for each value (days between 5 and 14, both included) or a histogram with bins $4.5-5.5,\ 5.5-6.5, \ldots, 13.5-14.5$.

Statistic	`scc_el`	`lnscc_el`	`kg_el`	`dim_el`
mean	237.5	4.57	21.41	9.77
median	85.0	4.44	21.8	10
minimum	5.0	1.61	1.8	5
1st quartile	43.0	3.76	18.0	7
3rd quartile	178.3	5.18	25.0	12
maximum	5016	8.52	37.8	14
standard deviation	543.1	1.18	6.25	2.82
range	5011	6.91	36.0	9
inter-quartile range	135.3	1.42	7.0	5
skewness	5.33	0.75	-0.64	-0.08
kurtosis	33.7	0.76	1.23	-1.24
A-D normality test	98.8	4.08	3.52	9.17
P-value	< 0.005	< 0.005	< 0.005	< 0.005

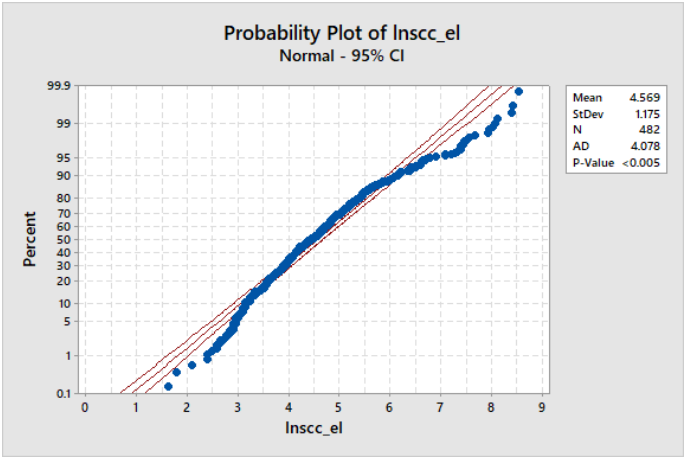
Somatic cell count variables

The histogram and descriptive statistics show that the raw somatic cell count has an extremely right skewed distribution, with the values spanning a very wide range. It is useful to transform the values to get a more symmetrical distribution, and the most commonly used transformation for somatic cell counts is the (natural) log transformation. The distribution of `lnscc_el` is much "nicer" and more symmetrical, although the histogram shows some right-skewness (the skewness is 0.75, and the mean is somewhat above the mean). The center of the distribution is about 4.5, the standard deviation is 1.2, and no obvious outliers are seen. A boxplot (not shown) would indicate several potential ("suspected") outliers mostly in the right tail of the distribution, but recall that the rule behind those is based on a symmetrical distribution, and that the number of potential outliers indicated will also automatically increase with sample size. We should only designate observations that clearly stand out as not belonging with the rest as (true) outliers.





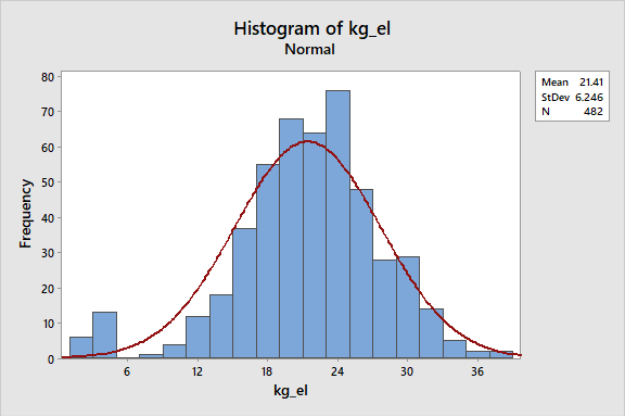
On original scale, and also on logarithmic scale, the A-D normality test gives strong evidence against a normal distribution, and the normal plot (on the latter scale) shows the typical curved shape of a skewed distribution. Therefore assuming a normal distribution does not seem valid on either scale.

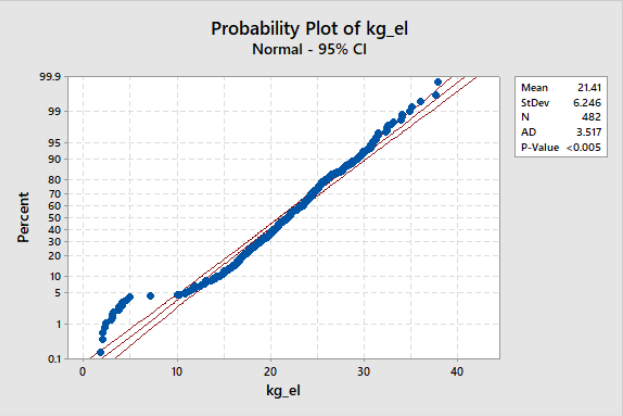


Milk production

The histogram shows that the milk production (in kg) has a roughly symmetrical and bell-shaped distribution, except for a small group of very low values in the range of $1-5$ *kg*. These values can be interpreted as a second mode (which would make the distribution bimodal) or as outliers, in the sense of observations that do not belong with the rest. Focusing for now on the remaining and major

part of the distribution, it seems fairly bell-shaped and centered around 22 *kg* with a spread of about 5 *kg*. The normal plot is reasonably straight except for the deviations in the lower tail; the normality test is clearly significant, but we knew already that the group of low values would contrast with a normal distribution. It could be of interest to explore the distribution without the values in the very low range; we will return to this question in part d) below.

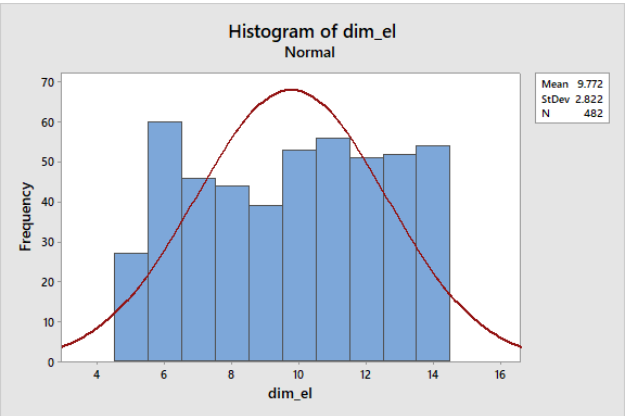


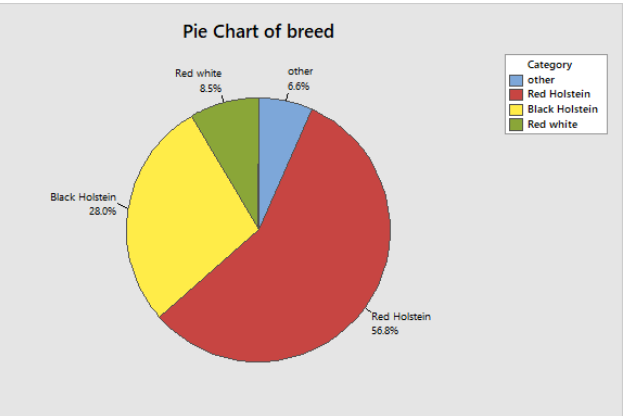


Days in milk and Breed

The chart/histogram for `dim_el` shows that the day of the milk recordings was almost equally distributed on days 5−14 after calving in the dataset. Therefore, the mean of the distribution is close to the midpoint (9.5), and the standard deviation is large. Also, there are no outliers. Note the negative kurtosis to indicate the lack of tails in the distribution. It is obvious from the visual display of the distribution that it is far from bell-shaped. We could also say that by the discreteness of the data, any approximation by a continuous distribution is of little interest (though strictly speaking not invalid in general, because it would depend on what such an approximation was used for).

The distribution on the 4 breed groups is shown in the pie chart; the majority of heifers are Holsteins (either red or black). Because the values associated with the groups are mere labels, it does not make sense to approximate the distribution with any quantitative distribution (continuous or discrete).





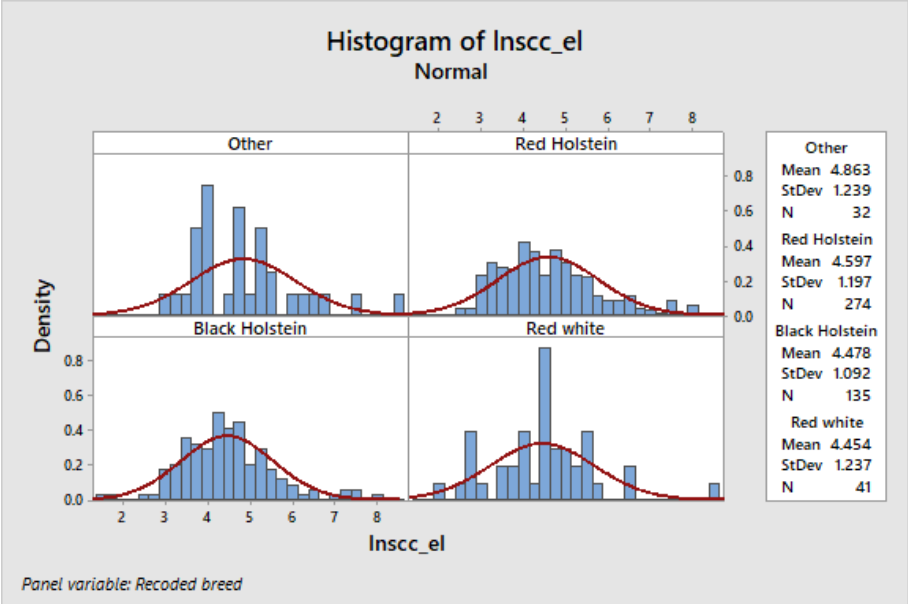
c. Proportion of heifers with somatic cell count above 200

The sample proportion of heifers with `scc_el` > 200 is: $\hat{p} = 109/482 = 0.226 = 22.6\%$; the count of 109 values > 200 is most easily obtained after sorting by `scc_el`. This estimate is valid for any simple random sample no matter the distribution type, because the conditions for a binomial setting are a consequence of the simple random sampling (assuming that sampling from a finite population

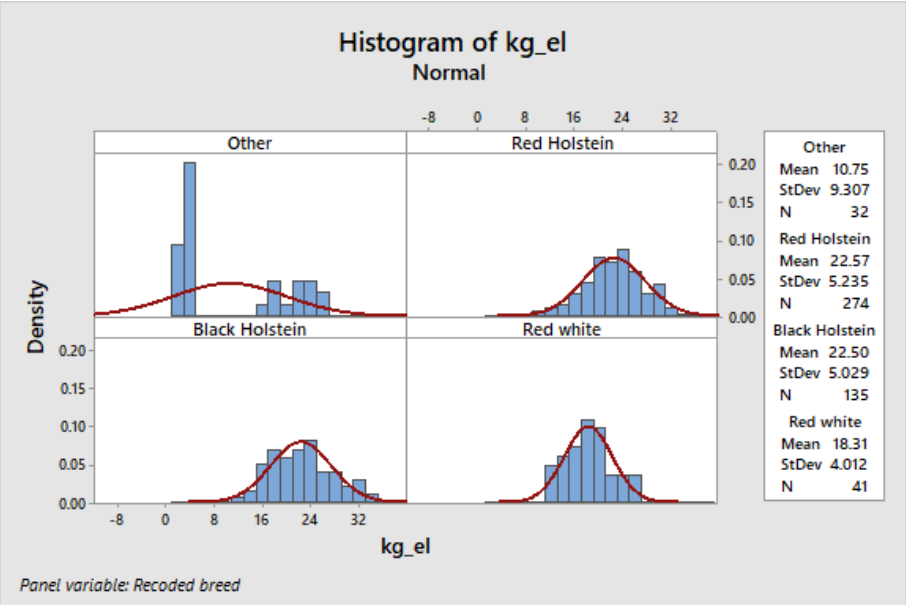
is not a concern). If the data follow a particular theoretical distribution (say a normal distribution) closely, there could be a gain in precision by computing the estimate from the normal distribution; however, if the normal distribution does not approximate the data well, an estimate based on the normal distribution could be seriously biased. The distribution of the somatic cell counts is nowhere near a normal, so it would be a bad choice to use a normal distribution. (The value obtained is $P(Z > (200 - 237.5)/543.1) = P(Z > -0.069) = 0.53$.) The log somatic cell counts are closer to a normal distribution, so that calculation makes more sense: $P(Z > (\ln(200) - 4.5693)/1.1752) = P(Z > 0.620) = 0.268$. This value is closer to the sample proportion but still a fair bit off, which can be attributed to the fact that the normal distribution does not fit that well to the log somatic cell counts. Therefore, our best estimate is the sample proportion, and the standard error for that estimate is $\sqrt{\hat{p}(1-\hat{p})/482} = 0.019$.

d. Comparisons between breeds

We prefer to work with the somatic cell counts on logarithmic scale because the extreme right-skewness of the distribution on original scale makes it difficult to meaningfully describe and visualize the distribution. As the focus here is on a descriptive comparison between the breeds, our main tool will be a display of the distributions with separate histograms (with default binning, though not always optimal) for each breed group (note that it is better to display density than frequency with the very different group sizes). We have no information to further discriminate within the breed groups; in practice, that would presumably be of real interest.



The histograms for log somatic cell counts appear quite similar for the four breed groups, and the same can be said for the sample means and standard deviations. A formal statistical comparison between the four distributions will need to assess the differences relative to what could be expected from random variation. Note that because groups 0 and 3 have the smallest sample sizes, it is not surprising that their distributions look more "rough" than for the two large Holstein groups. With no obvious differences between breeds in the log somatic cell counts, our previous analyses based on combining all breeds seem very reasonable.



A different picture arises for the milk yield. The distribution for breed 0 is clearly bimodal. The previously noted group of low values fall within this breed, and the remaining milk yields for breed 0 seem reasonably similar to those from the other groups. It was subsequently confirmed after consultation with the owner of the dataset, Dr. De Vliegher, that the BWB breed included in group 0 would indeed be expected to have much lower milk yields — it is beef cattle! The distributions for breeds $1-3$ all seem close to normal in shape, and it is not clear how substantial the differences in sample means and standard deviations are, or whether they could be due to chance only. Statistical methods for comparing multiple groups will be covered later in the course (typically referred to as "one-way ANOVA").

As we have already displayed the distributions separately for the breeds, we may in addition want to formally assess normality for each breed group (the A-D normality test is non-significant for breeds $1-3$ with P-values 0.286, 0.056, and 0.207, respectively; not shown) or reduce breed 0 to the heifers with milk yields above 8 *kg*, say, in an attempt to eliminate all beef cattle from this group, and redo the descriptive analysis (a total of 13 observations remain for breed 0, with a mean of 21.5 and standard deviation of 3.23; not shown). We conclude that the main problem with the distribution of milk yield appears to have been the BWB breed included in breed 0, and after that issue has been dealt with the distribution may be reasonably normal but possibly still not quite uniform across breeds.