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

Solution to home assignment 1

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 four selected variables were required for the assignment, and by discussing answers for both an experimental and observational study. All analyses shown used Minitab 21, but other Minitab versions or other programs, such as Stata, would give similar figures and results.

1. Descriptive analysis

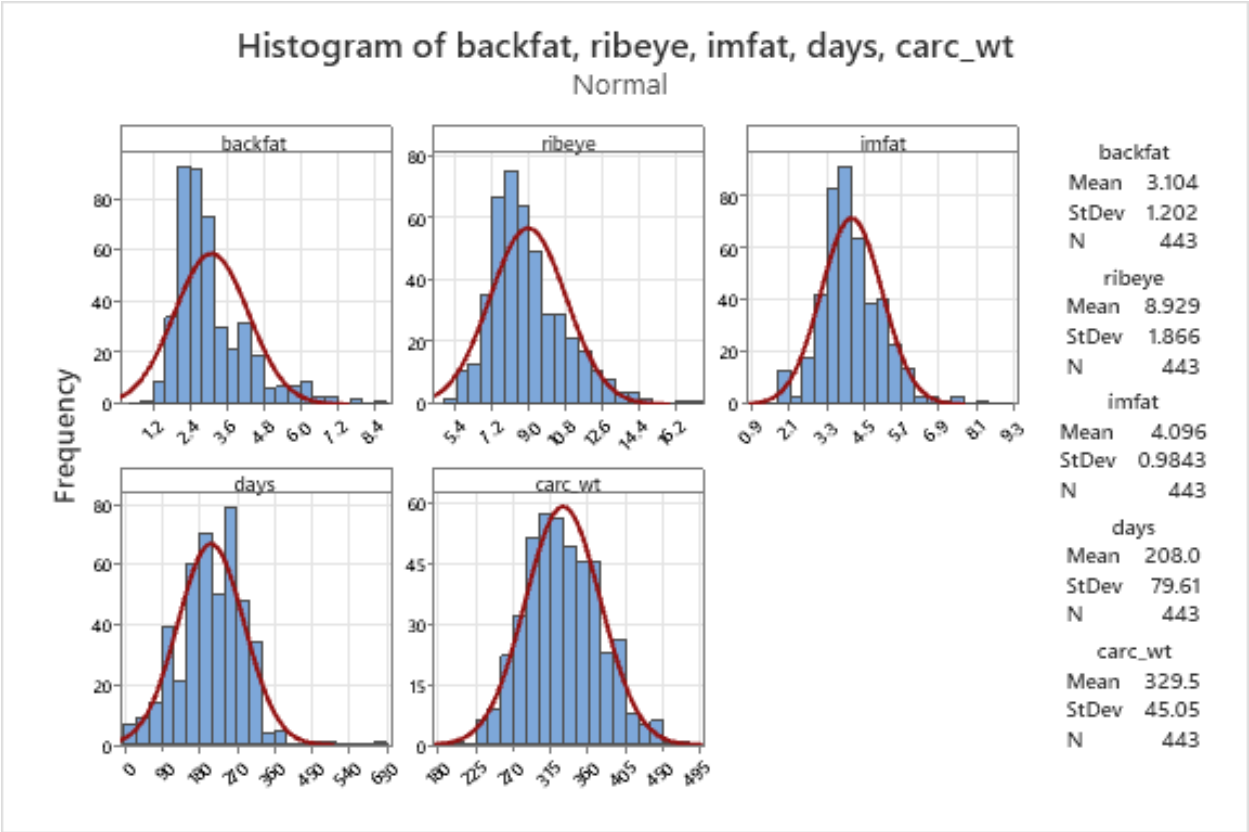
The data may be considered as a simple random sample from the population of beef cattle on PEI entering a feedlot prior to slaughter, with some restrictions corresponding to how the reduced dataset was constructed; these restrictions are not of interest here. The variables *backfat*, *ribeye*, *imfat*, and *carc_wt* are quantitative and measured on a continuous scale, whereas *days* is quantitative and measured on a discrete (integer) scale (although in principle on a continuous scale). The variables *farm*, *grade*, *breed* and *id* are categorical with more than two categories, all nominal except *grade* which is ordinal. The remaining variables (*sex*, *bckgrnd*, and *implant*) are dichotomous (categorical with two categories), where there is no reason to distinguish between nominal and ordinal. The data includes 443 animals with complete records (no missing values).

The table below gives the most useful descriptive statistics for quantitative variables; the list of descriptive statistics should as a minimum include the mean or median for the center, and the standard deviation or interquartile range for the spread, but we include also the five-number summary and the skewness. For categorical variables it is more useful to give the probabilities (that is, the proportion of animals) for each possible value, as shown in the table on the next page (except for the variable *id*, which takes a distinct value for every animal so its distribution is not of real interest).

Descriptive statistics for quantitative variables:

statistic	*backfat*	*ribeye*	*imfat*	*days*	*carc_wt*
mean	3.104	8.929	4.096	208	329.5
minimum	0.900	5.230	1.850	15	209.1
1st quartile	2.250	7.665	3.500	163	297.3
median	2.725	8.610	3.970	214	325.6
3rd quartile	3.610	9.870	4.650	265	360.9
maximum	8.755	17.235	8.360	614	474.5
standard deviation	1.202	1.866	0.984	79.6	45.1
inter-quartile range	1.360	2.205	1.150	102	63.6
skewness	1.47	0.95	0.55	0.12	0.26

With the quite large sample size, the preferred graphical display of the continuous distributions is a histogram, which may be overlaid a normal distribution curve to show the agreement with the normal distribution (where of interest). Note that the boxplot just displays the descriptive statistics involved and any potential outliers rather than the full distribution.



The default number of bins varied between variables, so it was reset at $\sqrt{487} \approx 21$ for all histograms. Note that despite *days* being a discrete variable, due to the large spread of values across distinct days the most appropriate graphical display is a histogram (a bar graph would be too noisy).

Observed probability distributions for categorical (incl. binary) variables:

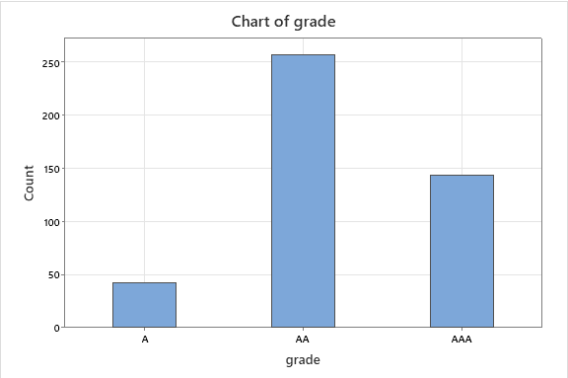
value	*sex*	*bckgrnd*	*implant*
0	.361	.253	.745
1	.639	.747	.255

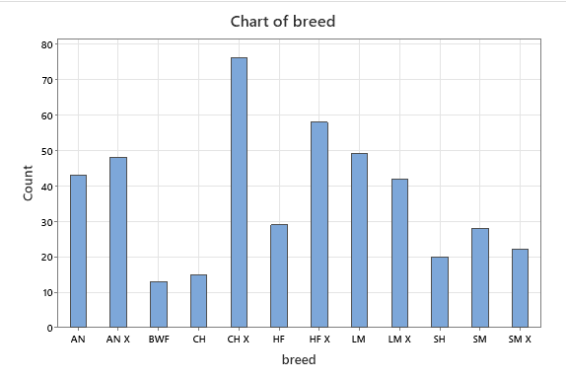
value	*grade*
A	.095
AA	.580
AAA	.325

value	*farm*
1	.126
2	.129
3	.302
4	.183
5	.147
7	.113

value	*breed*
AN	.097
AN X	.108
BWF	.029
CH	.034
CH X	.172
HF	.065
HF X	.131
LM	.111
LM X	.095
SH	.045
SM	.063
SM X	.050

The preferred graphical display for a categorical variable is a chart with one bar for each value representing its probability (or count). To illustrate, the bar graph for the variables *grade* and *breed* are shown (on the next page); note that the bars are separated — which distinguishes the bar graph from a histogram.





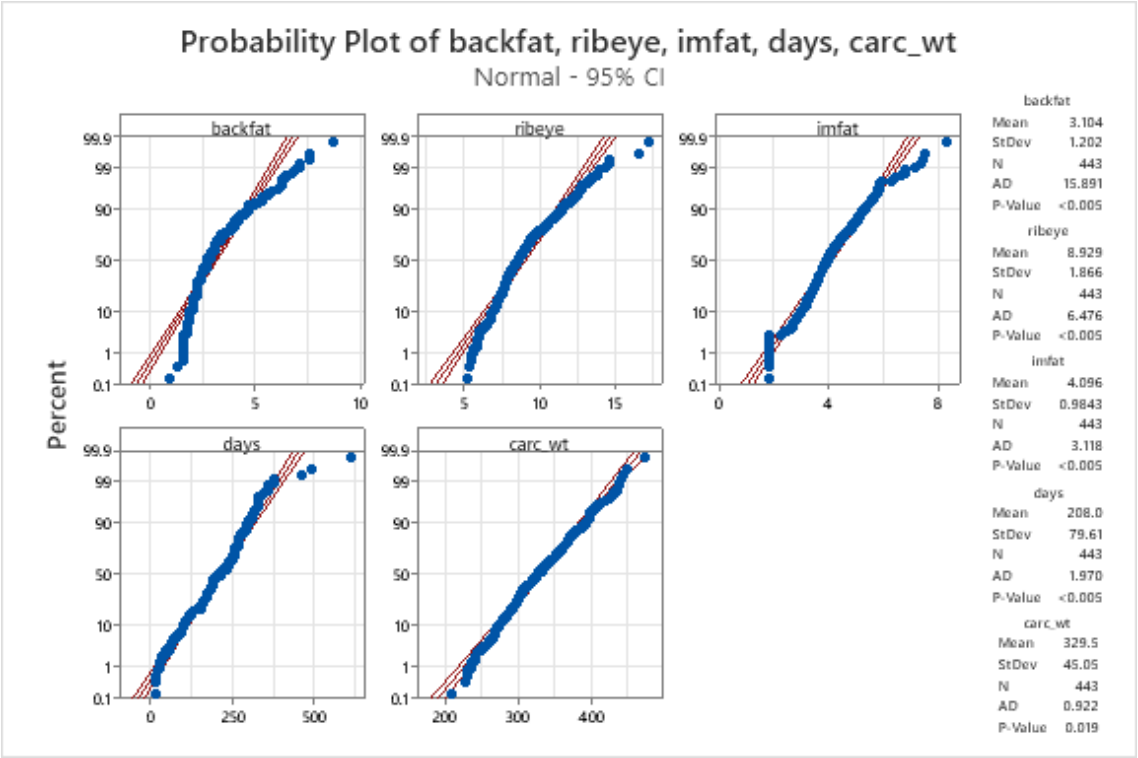
Finally, brief summaries of the distributions based on the computed statistics and graphs:

- *farm*: one farm (no. 3) higher represented than the five other farms, with similar proportions,
- *grade*: the most common grade is AA (more than 50% of carcasses) whereas grade A is quite rare (about 10% of carcasses),
- *breed*: the most common breed is "CH X" (where X presumably means crossbreed) at approx. 17%, and some breeds (e.g. BWF) are quite rare; one might consider combining pure and crossbreeds (as done in the journal article), this would reduce the number of categories and produce a distribution with AN, CH, HF, and LM breeds as the most common,
- *sex*: almost twice as many steers (64%) than heifers,
- *bckgrnd*: about three out of four animals (75%) were backgrounded,
- *implant*: approximately one fourth (25.5%) of the animals had an implant,
- *backfat*: unimodal; centered around 3 mm; strongly right-skewed with a long tail; many potential outliers on the boxplot (beyond 5.65 mm), but only the largest value (8.755) may be strikingly outlying, and in view of the right-skewness such a large value in the right tail may be quite "okay",
- *ribeye*: unimodal; centered a bit above 8.5 cm^2; clearly right-skewed with a pretty long right tail; several potential outliers on the boxplot (beyond 13.18 cm^2), and two of these may seem strikingly outlying. However, closer inspection reveals that all the potential outliers are from farm 3, thus indicating substantial farm differences rather than being isolated extreme values,
- *imfat*: unimodal; centered around 4%; moderately right-skewed with a pretty long tail; several potential outliers on the boxplot (beyond 6.375%), but only one value at 8.36 seems well separated from the rest; the histogram also shows a small peak in the left tail, and this stems from 13 records with a value of 1.85% — although not indicated as potential outliers, this strange pattern in the data should definitely be explored,
- *days*: unimodal (centered around 210 days) or possibly bimodal with modes at approx. 190 and 250 days, in any case with a very wide range ($15 - 614$ days); weakly right-skewed but with several potential outliers in the right tail on the boxplot (beyond 418 days), the largest one (614 days) seems conspicuously higher than the rest and might warrant further investigation,
- *carc_wt*: unimodal; centered around or slightly below 330 kg with a standard deviation of 45 kg; close to symmetrical and bell-shaped; a single potential outlier in the right tail but hardly a real outlier.

2. Assessment of normal distribution

The standard tools to assess whether it is reasonable to assume a variable to be normally distributed are the normal probability plot and a normality test. The figure below shows these plots and gives a P-value for the Anderson-Darling test for each of the 5 quantitative variables.

The plot for *carc_wt* looks reasonably straight, but with $P = 0.019$ the A-D normality test shows some evidence against a normal distribution, probably due to its slight right-skewness and a resulting lack of fit in the lower tail. All other P-values are listed as $P < 0.005$ and thus provide clear evidence against a normal distribution. The plots for *backfat* and *ribeye* show the pattern typical of a right-skewed distribution with both tails clearly off (below) the indicated straight line. Some of these patterns are also seen for *imfat* and *days*. In addition, the plot for *imfat* shows the peak at 1.85 in the left tail, and the plot for *days* shows the two extreme values (at 614 days) in the right tail.



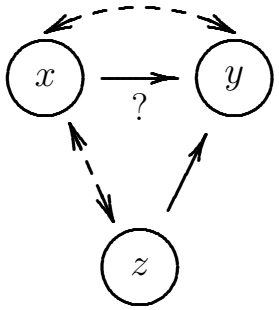
The distribution for *carc_wt* was already not too far from normal, and the square-root and log-transformations further improvement the normality, in probability plots (not included here) and in P-values of 0.25 and 0.41, respectively. For *imfat*, however, no improvements are seen after the transformations, which only aggravate the lower tail problem. After square-root transformation of the three other variables, the A-D normality tests are still clearly significant ($P < 0.005$). The probability plot for *days* on the square-root scale does shows that the lower tail of the distribution not match the normal, reflecting its negative skewness after transformation. Therefore, this variable's distribution cannot be improved by square-root transformation (and log-transformation only makes it worse). The distributions for *backfat* and *ribeye* on square-root scale are still clearly right-skewed, so a further log-transformation is explored as well. On log-scale the right-skewness has further diminished in both distributions, and the probability plot for log-ribeye looks quite straight (not shown) although we still have $P < 0.005$ for both variables. In conclusion, transformation was only beneficial for *carc_wt*, and none of the other variables could be assumed as approximately normal on any of the scales explored, although the distributions for *backfat* and *ribeye* improved substantially towards normality on logarithmic scale.

3. Discussion of study type

There is no indication in the text or the journal article that the study was experimental with respect to the use of implants. This would have required the experimenters to control whether animals were implanted or not, but as it is not mentioned anywhere one would assume that the farmers decided about use of implants. Also, the distribution of implants (e.g. across farms) does not suggest it was controlled. Therefore, the study should be considered as *observational*.

For the purpose of the assignment, let us discuss how randomization could have been implemented in an experimental study comprising 6 (or 8) farms. One important question is whether all animals from a farm would have the same implant status (i.e., whether they all had an implant or they all did not have an implant). Inspection of the data shows that most farms (actually all farms but no. 4) were consistent in their use of implants, with some farms using it and others not using it. Even if it therefore may not be feasible to carry out a study where the use of implants is randomized within each farm, corresponding to a *randomized block design* with farms as blocks, it would most likely be the most efficient design. This is because we would expect between-farm variation in the outcomes, and in a block design the effect of the use of implants would effectively be evaluated within blocks (farms). For the randomization we would then within each farm randomly decide which animals were to get the implant. If we knew the sample size from the farm, we could randomly select half of the animals to get the implant, for example using Minitab. If the sample size was not pre-determined, we could flip a coin to determine the use of implant for each consecutive pair of animals in the farm.

In an observational study, the association between an explanatory variable (x) and an outcome (y) may be confounded by one or more lurking variable(s) (z). Here x = the use of implants and y = some characteristic of the carcasses, such as grade or weight. Several lurking variables can be suggested; a very common confounding variable for studies involving animals in the agricultural food production is farm. As we have not yet discussed methods to quantify relationships in the course, we limit ourselves to a conceptual discussion. Let us specifically discuss y = carcass grade, and z = farm. The diagram below depicts the suggested scenario with a confounding effect of z, and we comment on each of the effects in turn.



- *association* between x and y: the journal article refers to previous research indicating an association between implants and carcass grade; in the present data, we would want to compare the distribution of carcass grades among animals with and without an implant;
- *association* between x and z: we already noted above that the use of implants was highly variable across farms, and this would indicate an association between farm and the use of implants;
- *causal effect* of z on y: it would be entirely possible that the 6 (8) farms in the study performed differently in terms of the resulting carcass grade; in fact, the results presented in the journal article shows estimates from a model with farm effects included.

If all three associations/effects were present in the data, we could say that the association between the use of implants and carcass grade was confounded by farms, or that farms acted as a lurking variable for this association. Other "intuitive" possible lurking variables are sex and breed.