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

Solution to home assignment I

This solution presents one way in which the descriptive analysis could be done, but other ways are possible as well. The solution is more detailed than required for a 100% mark, e.g. by including all the variables when only two selected variables were required for the assignment.

1. Study and variable types

The study is *observational*, as opposed to an experiment, because no treatments or other interventions were undertaken. In the published papers, the herds were described as *purposively selected* (i.e. a convenience sample), but the 30 pigs per herd were randomly selected. The table below lists all variables with a description of their type.

variable	description
farm	categorical, nominal
pig	categorical, nominal
sex	binary/categorical, nominal
dtm	quantitative, continuous (but discretely observed in days), ratio
adg	quantitative, continuous, ratio
mm	quantitative, continuous (but discretely observed, with strange gaps), ratio
ar	categorical, ordinal
lu	categorical, ordinal
epg5	quantitative, discrete (as counts, even if scaled), ratio
worms	quantitative, discrete (as counts), ratio
li	categorical, ordinal

As an elaboration of the description given for *ar*, it can be said that the scores roughly represent a categorization of *mm*, but the scale does not reflect the original scale (for *mm*) in any obvious way, so therefore we should not consider this variable as quantitative.

2. Descriptive analysis

The table below (next page) gives the most useful descriptive statistics for continuous variables; the list of descriptive statistics should as a minimum include the mean, median and standard deviation. There are no missing values so the number of observations is 341 for all variables. Discrete quantitative variables may be described as if they were continuous if their distribution is "well spread-out". For categorical variables it is more useful to give the probabilities (that is, the proportion of pigs) for each possible value, as shown in the table below (next page).

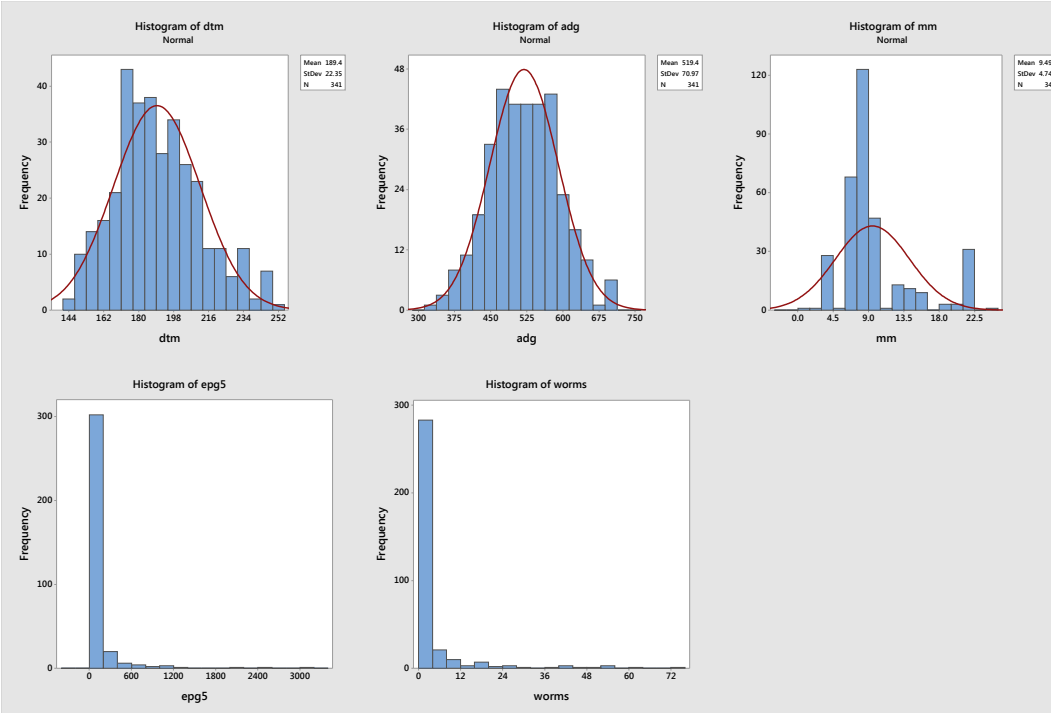
With 341 observations, the preferred graphical display of the continuous distributions is a histogram, which may be overlaid a normal distribution curve to show the agreement with the normal distribution (where of interest). The stemplot gives about the same information but in a more clumsy layout, and a boxplot displays only the descriptive statistics involved and any "suspected outliers".

Descriptive statistics for continuous variables:

statistic	*dtm*	*adg*	*mm*	*epg5*	*worms*
mean	189.4	519.4	9.50	78.6	3.37
minimum	141	317	0	0	0
1st quartile	175	469	6.0	0	0
median	188	519	8.0	0	0
3rd quartile	203	568.5	10.0	0	2.00
maximum	254	707	24.0	3025	72.0
standard deviation	22.3	71.0	4.75	294	9.88
inter-quartile range	28.0	99.5	4.00	0	2.00
skewness	0.43	0.02	1.38	6.48	4.28
kurtosis	-0.07	-0.21	1.15	50.9	19.7

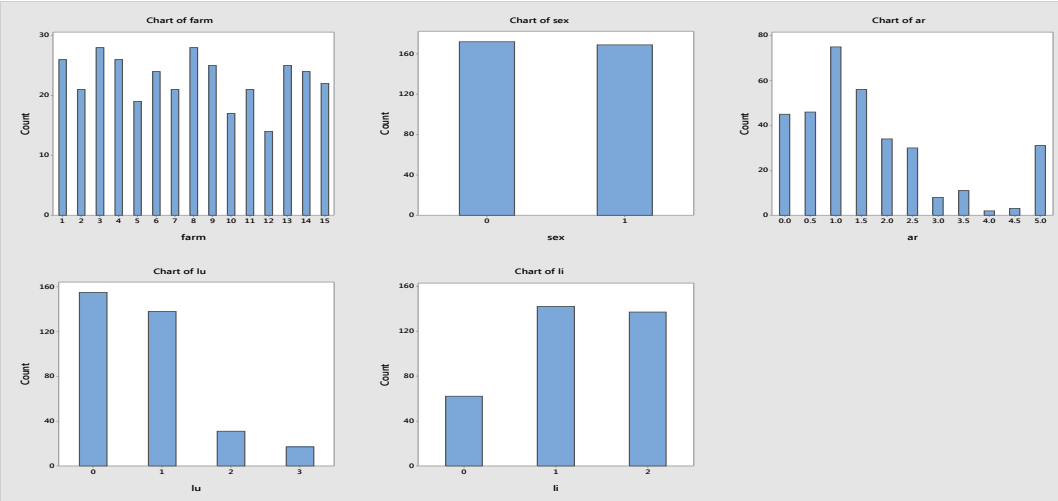
Observed probability distributions for categorical (excl. farm and pig) and discrete variables:

variable	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
farm−1	.076	.062	.082	.076	.056	.070	.062	.082	.073	.050	.062	.041	.073	.070	.065
sex	.504	.496													
ar·2	.132	.135	.220	.164	.100	.088	.023	.032	.006	.009	.091				
lu	.454	.405	.091	.050											
li	.182	.416	.402												



The default number of bins was quite variable, with most bins (24) for *drm* and least bins (15 and 16, respectively) for *epg5* and *worms*. All histograms have been adjusted to have 19 ($\approx \sqrt{341}$) bins. Normal distribution curves were only overlaid for the variables where it seemed meaningful. The preferred graphical display for a categorical variable is a chart with one bar for each value representing

its observed proportion; alternatively, a pie chart could be used as well. Bar charts for the categorical variables are shown below; note that the bars are separated — which distinguishes bar graphs from histograms.



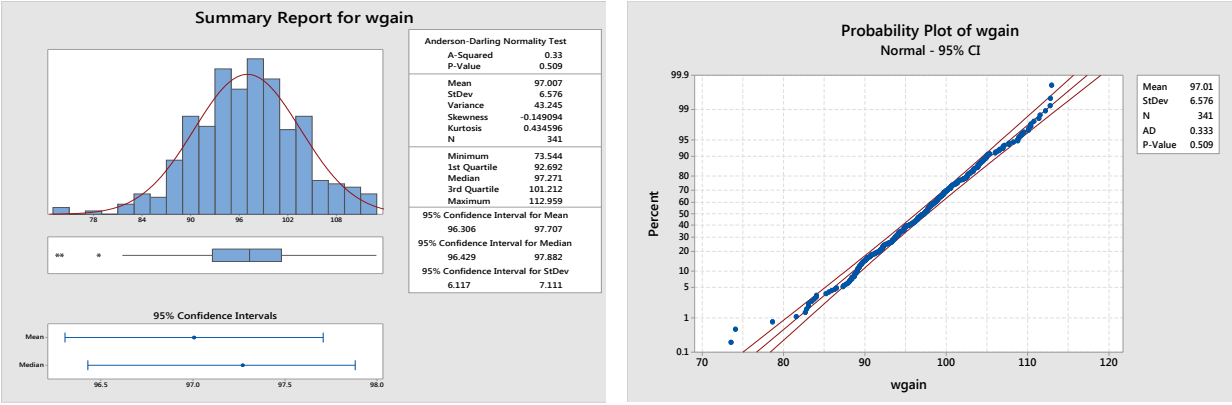
Finally, brief summaries of the distributions based on the computed statistics and graphs:

- *farm*: roughly equally distributed (i.e., equally many pigs per farm), though none of them reached the target sample size of 30,
- *sex*: very close to a 50:50 distribution on the two sexes (female and castrate), with slightly more pigs in the former category,
- *dtm*: unimodal; centered around $185-190$ days with most of the distribution within 20 days hereof; somewhat right-skewed, in particular due to a too short left tail; four suspected outliers indicated in the right tail by the boxplot, but even the largest of these (254) does not seem particularly extreme,
- *adg*: unimodal; centered around 519 g; fairly large spread ($s = 71$) but symmetrical and apparently quite bell-shaped; one suspected outlier in the left tail but nothing to worry about,
- *mm*: strange looking distribution with multiple peaks and gaps, probably resulting from recording at a discrete scale; apparently unimodal even if a second peak in the right tail may exist; centered around $8-10mm$ with a fairly wide spread between minor peaks at $4mm$ and $21mm$; clearly right-skewed; many suspected outliers indicated in the right tail (including the minor peak at $31mm$, but these do not seem as real outliers,
- *ar*: with the ordinal scale, descriptive statistics based on percentiles are somewhat meaningful — e.g., the median equals 1.5, and distribution seems centered at the lower end of the scale with a mode at 1, and a second mode at the upper end (5),
- *lu*: highest (and similar) proportions for categories 0 and 1 ($\leq$ mild), and only about 14% of pigs with higher scores,
- *epg5*: unimodal with both mode and median at 0, therefore centered at 0; strongly right-skewed with a few very large observations, but the main part of the distribution is close to 0 (e.g., also $Q_3 = 0$); for such a right-skewed distribution, the indications of "suspected outliers" in the boxplot are of no use, and all the extreme observations may be fine,
- *worms*: unimodal with both mode and median at 0, therefore centered at 0; strongly right-skewed observations scattered across the range $(0-75)$, but the main part of the distribution is close to 0 (e.g., $Q_3 = 2$); for such a right-skewed distribution, the indications of "suspected outliers" in the boxplot are of no use, and all the extreme observations may be fine,

- *li*: highest (and similar) proportions for categories 1 and 2 ($\geq$ mild), and only about 18% of pigs with a negative score.

3. Total weight gain

We compute the total weight gain (*wgain*) by multiplying *adg* and *dtm*, and convert into *kg* by dividing the resulting value by 1000. Our description of the distribution for *wgain* is based on the graphical summary and normal probability plot shown below.



The distribution is unimodal and centered around 97 *kg*; the spread is quite narrow with a standard deviation of 6.6 *kg*; the distribution appears roughly bell-shaped with a skewness value just a little below zero, and only two values in the left tail seem somewhat off the others and may be real outliers (not too surprisingly, one of these pigs also had the lowest *adg*). The normal probability plot looks straight in the central part of the distribution, but the upper tail is a bit short and the two mentioned potential outliers are too far out in the left tail. The A-D normality test gives $P = 0.509$, and therefore offers no evidence against a normal distribution; according to this test, the distribution could very well be normal. This overall test based on the 341 values does not necessarily mean that there could not be a problem with the two low values; we would need another tool to quantify that, and our assessment would depend on whether there were other reasons to suspect these observations as outlying or they just happened to be the two most extreme ones. For our purpose it suffices to say overall the distribution appears quite normal and that the two values in the left could warrant further exploration.

4. Current standard for average daily gain

An average daily gain of 1.5 *lbs* corresponds to $1.5 \cdot 453.6 = 680.4$ *g*. Among the pigs in our data, only 6 (1.8%) achieved such a high growth rate. Even without carrying out any formal statistical tests (which have not been discussed in the course so far, and were therefore not expected to be used for the answer) it seems clear that current growth rates are higher than typical values in the dataset.

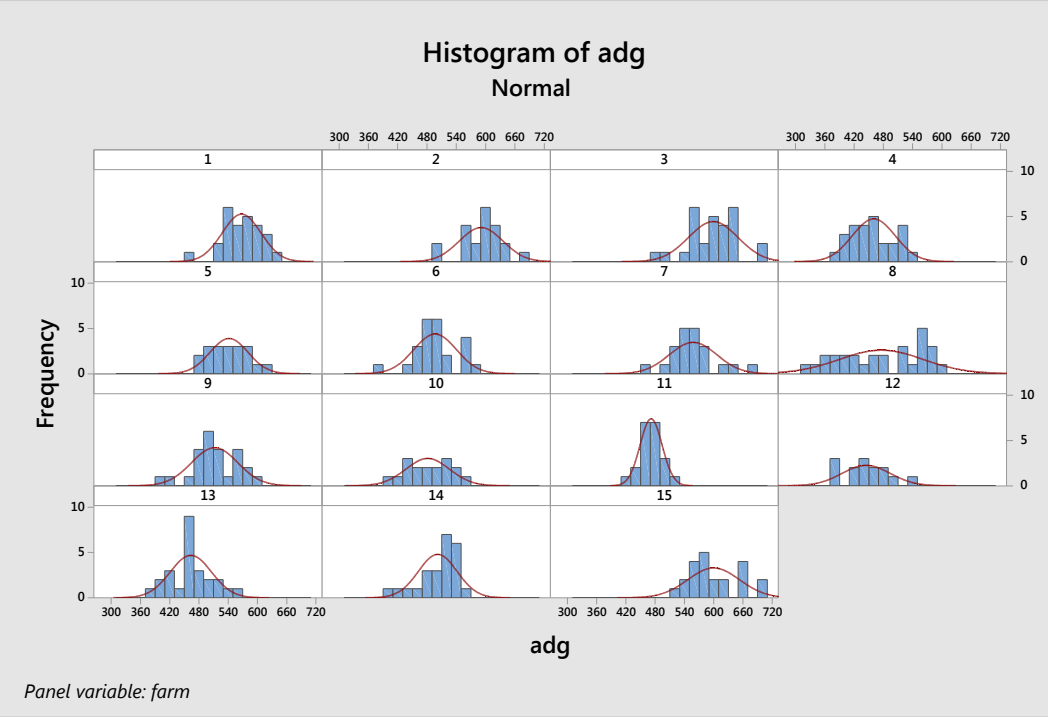
Possible explanations could be related to differences in *i)* measurement or in *ii)* the populations. For *i)*, Our *adg* values were computed from the pigs' entire lifespans, but obviously the growth is not linear, so different values would be obtained for different life stages or stages in the production. It is not clear from the cited reference value in the assignment (nor from the sources it was drawn from) that it was computed in the same way as our *adg*.

For *ii)*, it seems intuitively plausible that the pig populations of today may differ in substantial ways from those on PEI some 30 years ago. Such differences may be due to inherent characteristics of

the animals, such as genetics or breed, or they may be due to differences in production parameters, such as feed. Without detailed knowledge about pig production now (corresponding to the reference value) and then, any further specification of the reasons for the discrepancy may end up as mostly speculative.

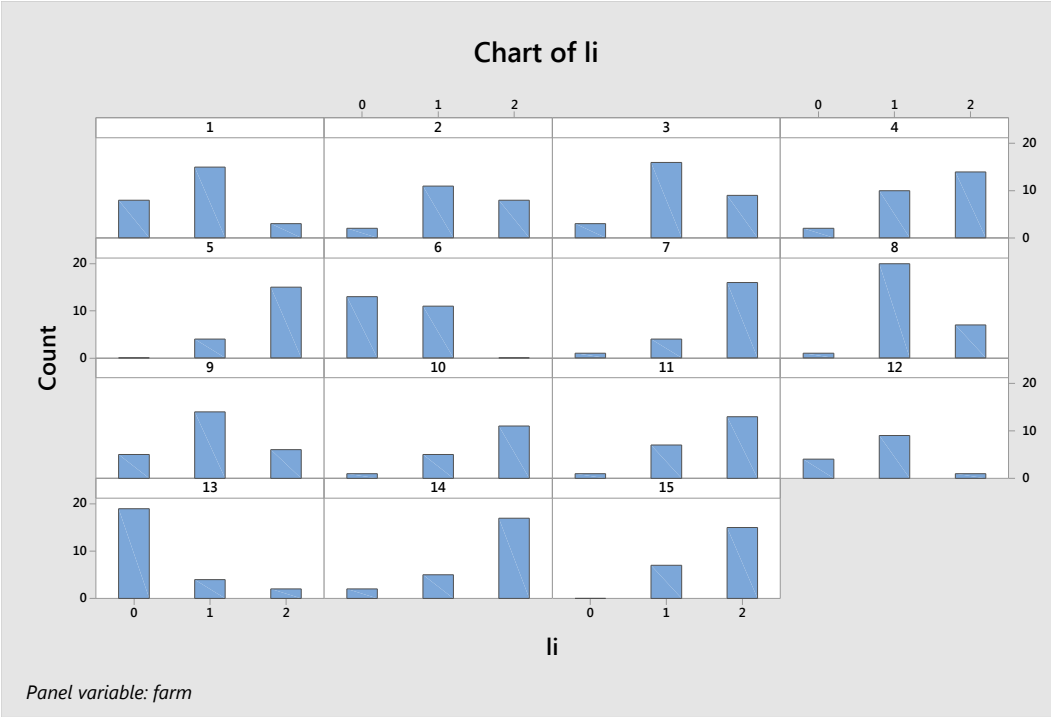
5. Distributions within and across farms

The answer for this question will be limited to the variables *adg* and *li*, representing one continuous and one categorical variable; other variables would be explored similarly. When the data are split onto each farm, the sample size becomes quite low and graphical representations such as histograms become rather noisy. One needs to critically assess whether the actual display is still useful, and in the case of *adg* the distributions depicted still seem reasonable. Considering the population shapes, the most natural descriptive statistics are the mean and standard deviations (for clearly non-normal distributions different measures of center and spread should be used), and these are shown in the table below.



variable/ statistic	farm 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
adg															
mean	568	591	599	460	542	496	558	476	513	481	472	445	464	502	600
std.dev.	40	45	51	44	39	44	49	87	48	45	23	50	43	40	53
li															
$p(0)$	.308	.095	.107	.077	.000	.542	.048	.036	.200	.059	.048	.286	.760	.083	.000
$p(1)$	.577	.524	.571	.385	.211	.458	.191	.714	.560	.294	.333	.643	.160	.208	.318
$p(2)$	.115	.381	.321	.539	.790	.000	.762	.250	.240	.647	.619	.071	.080	.708	.682

For *li*, a bar chart still is the most natural way of representing the distribution, and the proportions for the different *li*-categories were also shown in the table below.



The distributions of weight gains within different farms look reasonably similar (and normal) in shape (considering the low sample size), but there are some fairly large differences in both means (ranging from 445 to 600 *g*) and standard deviations (farms 8 and 11). The distributions of liver scores appear very different across farms, with proportions in the lowest (negative) and highest (severe) categories ranging from zero to well about 0.5 across the farms. Based on these findings it seems fair to say that both growth and parasite infestation varies substantially across farms. Therefore, the overall distributions for these two variables do not represent the distributions within specific farms well.