Biostats VHM 801 Course
Fall 2001, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to assignment I

The solution is more detailed than expected to obtain a 100% mark. All analyses shown were done using Minitab (version 12), but other programs may be used as well.

1. Descriptive analysis

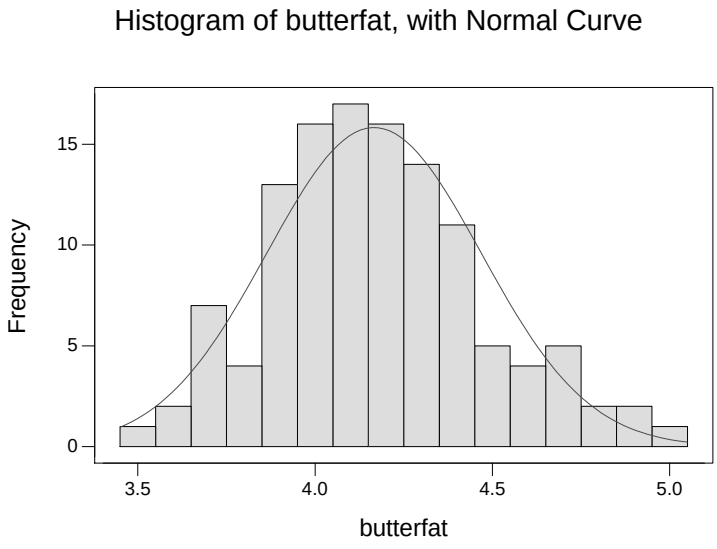
For the entire assignment our model assumption for the 120 butterfat values is that they constitute a sample from a suitable population (of cows). More formally, we assume them independent and identically distributed (i.i.d.).
The descriptive analysis consists of a number of graphs/plots and of the most common descriptive statistics, which are listed in the table below. The list of descriptive statistics should as a minimum include the mean, median and standard deviation.

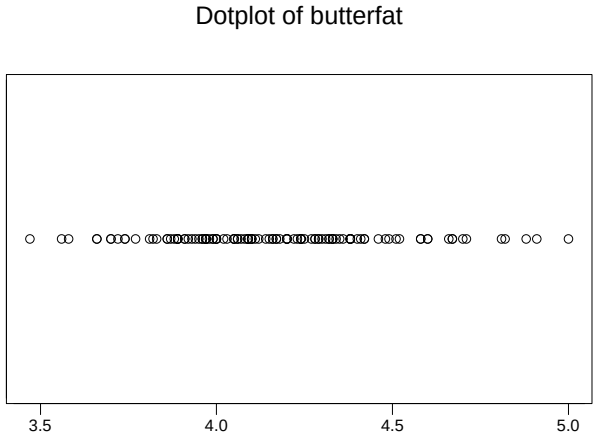
statistic	value	statistic	value	statistic	value
mean	4.166	standard deviation	0.302	skewness	0.373
median	4.145	1st quartile	3.963	kurtosis	0.056
minimum	3.47	3rd quartile	4.348	range	1.53
maximum	5.00	inter-quartile range	0.385		

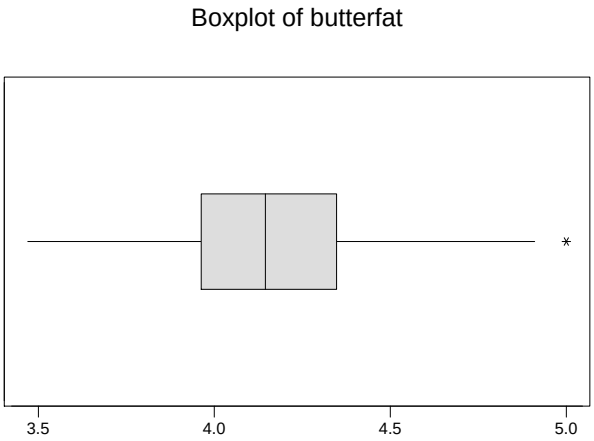
The graphical displays should as a minimum include a histogram or a stemplot. With more than 100 observations a histogram will typically show the distribution shape quite well.

```
Stem-and-leaf of butterfat
N  = 120
Leaf Unit = 0.010

    1   34 7
    3   35 68
    5   36 66
   11   37 002447
   22   38 12366789999
   38   39 1123456677777899
   54   40 0002355566789999
  (12)  41 000124566778
   54   42 0000234444578899
   38   43 012233345688888
   23   44 0122689
   16   45 1288
   12   46 00677
    7   47 01
    5   48 128
    2   49 1
    1   50 0
```





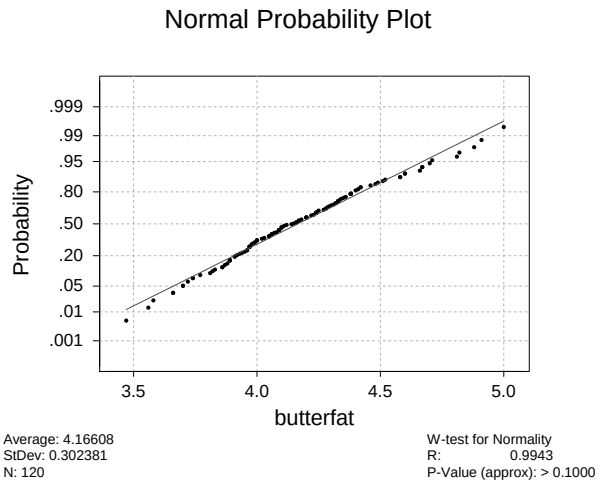


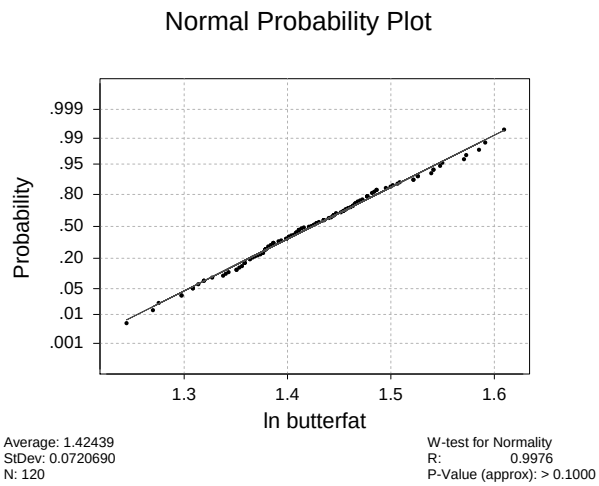
Summary of descriptive statistics:

- the center is around 4.15, the mean is only slightly larger than the median (indicating a roughly symmetric or slightly right skewed distribution), and the distribution is unimodal with a mode approximately at the center,
- the standard deviation is around 0.3 and the interquartile range is about 0.40, the two interquartiles are at approximately the same distance from the mean, but slightly larger to the right (also indicative for a roughly symmetric or slightly right skewed distribution),
- the skewness and kurtosis are positive but only moderate in value (both are zero for a normal distribution),
- the boxplot indicates a possible outlier to the right, but the dotplot shows that the largest value is by no means clearly outlying, just a bit larger than the second-largest.

2. Normal distribution assumptions

The validity of a normal distribution assumption is usually checked graphically by a histogram with overlaid normal density curve and/or by a normal probability plot. Minitab offers two versions of the probability plot, both applicable. In addition, the assumption of a normal distribution may be tested by a hypothesis test (but this has not been covered in the course so far). The probability plots below are for "raw" the butterfat values (left) and for the log-transformed (natural log) butterfat values (right). Both plots show the points quite nicely around the fitted line (corresponding to a normal distribution with parameters estimated by sample mean and standard deviation). Clearly the right plot is more convincing, but also the plot for raw butterfat values seems okay. The P-values for the "W-test" of normality indicate that a normal distribution fits both samples quite well — there is no convincing evidence against a normal distribution. The histogram with overlaid density curve from part 1 also indicates a normal distribution to be reasonable for the "raw" butterfat values. Therefore, unless other reasons point to a preference of log-transformation, we conclude that the improvement of the normal distribution assumption is of no practical importance, and we prefer to work with the untransformed values.





3. Probability calculations

Without assuming any particular distribution for the sample, the natural way to calculate (estimate) the probability of values above 4.50 in the population is to compute the observed proportion af values above 4.50 in the sample. Among the 120 values 104 are below 4.50, and accordingly 16 are above 4.50. Our estimate is therefore $16/120 = 13.3\%$. This procedure correspond to using the data as a binomial setting for the event that each butterfat value is below or above 4.50.

Assuming a normal distribution $\mathrm{N}(\mu, \sigma)$ for the data $(X_1, \ldots, X_{120})$, we estimate the parameters by

$$\hat{\mu} = \bar{X} = 4.1661, \quad \text{and} \quad \hat{\sigma} = s = 0.3024.$$

Calculation of the probability of values above 4.50 from $X \sim \mathrm{N}(4.1661, 0.3024)$ goes as follows,

$$\mathrm{P}(X > 4.50) = \mathrm{P}\left(\frac{X - 4.1661}{0.3024} > \frac{4.50 - 4.1661}{0.3024}\right) = \mathrm{P}(Z > 1.1042) = \mathrm{P}(Z < -1.1042) = 0.135,$$

using Minitab (to calculate the probability without standardization is okay as well). The two estimates probabilities are very close, reflecting that with more than 100 observations the estimates are rather precise, and that the normal distribution fits the data well.