Biostats VHM 801 Course
Fall 2021, Atlantic Veterinary College, PEI
Henrik Stryhn

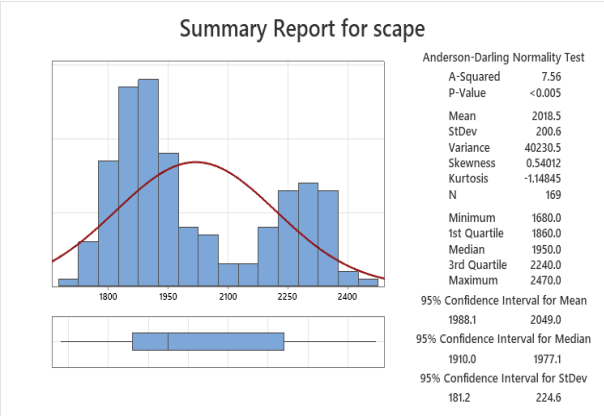
Solution to home assignment 1

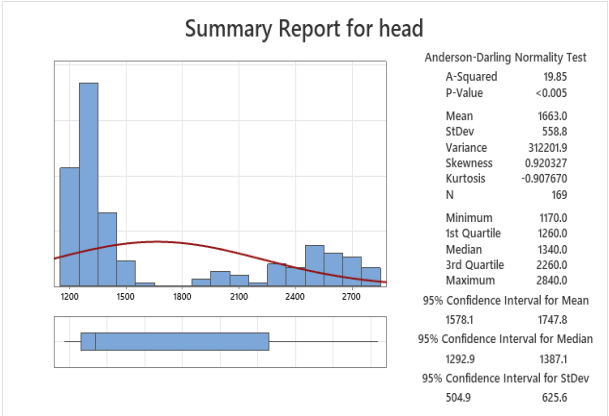
The solution presents one way in which the descriptive analysis could be done, but other ways are possible as well. It is more detailed than required for a 100% mark, among other things by including both the quantitative variables for the descriptive analysis when only one selected variable was required for the assignment. All analyses shown used Minitab 20, but other Minitab versions or other programs, such as Stata, would give similar figures and results.

1. Descriptive analysis for entire sample

The variables *scape* and *head* are quantitative and continuous, even if apparently recorded in gaps of 10 units (micron). The variable *worker* is nominal categorical with only two categories, so it could also be labelled as dichotomous or binary. According to the text, the data should be considered as a (simple random) sample from the population of ants.

For simplicity, the descriptive statistics and graphical display for the continuous variables will use the Graphical Summary menu in Minitab. With the fairly large sample size, 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 of less interest here; the graphs for the confidence intervals have been omitted below).

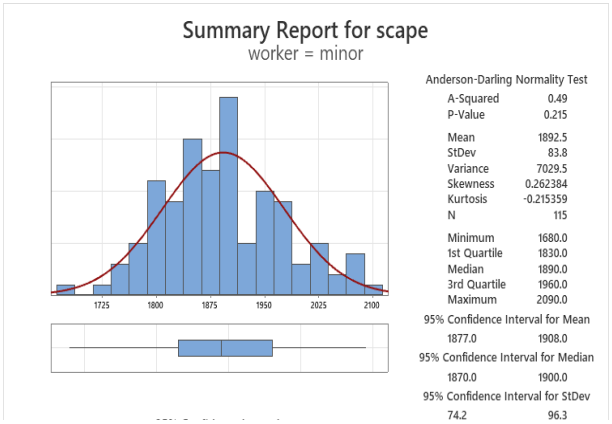


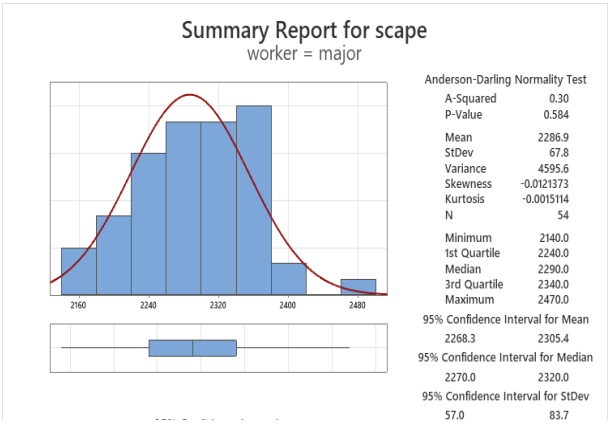


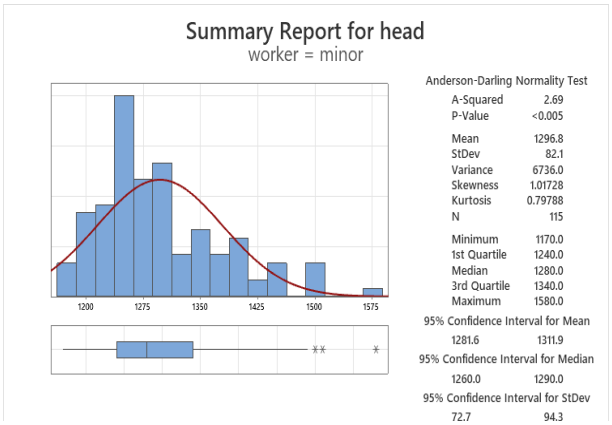
Both distributions appear of a strange shape that at a closer look is seen to be strongly indicative of a bimodal distribution. The bimodality is most pronounced for *head* where the two parts seem completely separated in the histogram. It is not clear from the histogram for *scape* whether a complete separation is also possible for this variable; one would need to look at the individual data points to see that. Because of the bimodality, the usual descriptors for distributions related to center and spread are of limited interest. For example, it is not informative to say that the mean for *head* equals 1663, when there are no observations around that value! Also the median and the skewness are of limited meaning for the combined distribution, because they mostly reflect the proportions of values from the two parts of the distribution. It is difficult to see from the graphs whether outliers could exist; the rule for suspected outliers from the boxplot is pretty much useless in this case where the boxes are so large; indeed, no suspected outliers are shown. Finally, from the discussion it should be quite obvious that neither distribution is anywhere close to normal. The very low P-values for the normality test confirm this, as would the corresponding normal probability/quantile plots (not included here, because the conclusion is so obvious without them).

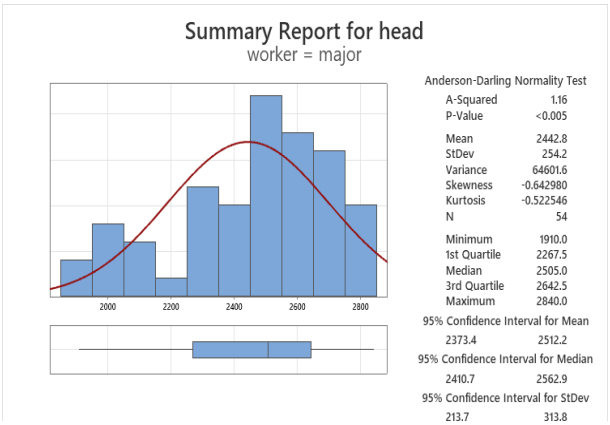
2. Separate descriptive analyses for minor and major workers

When redoing the descriptive analysis for the two types of workers separately, we see that those types correspond exactly to the two modes earlier identified. This is not a coincidence — the journal article describes how the two types were defined based on the morphological features (including measurements of scape length and head width). We include in our analysis for each of the two variables (and worker types) once more the Graphical Summary and add as well the normal plot. For obscure reasons, Minitab produces rather strange looking plots in the Graphical Summary when plots for the two types of workers are requested simultaneously; the plots below were constructed by first splitting the data into separate worksheets.





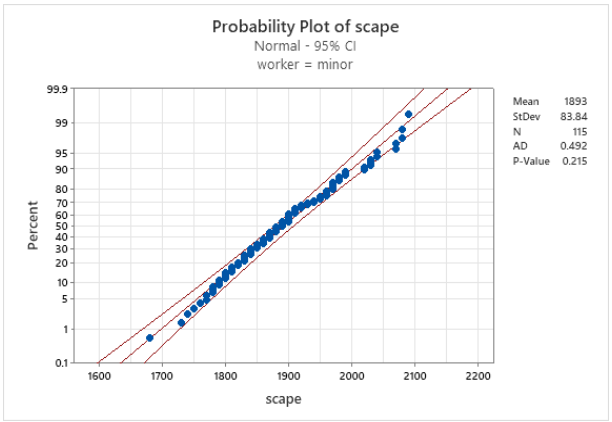


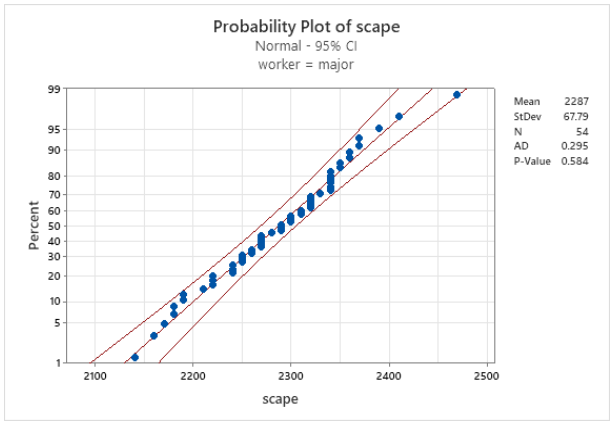


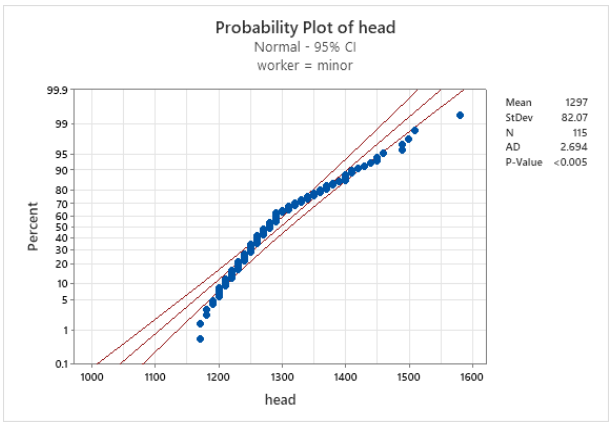
The two variables are reviewed separately, also including interpretations of the normal probability plots shown on the next page. For both body dimensions, the "major workers" are larger in size than the "minor workers".

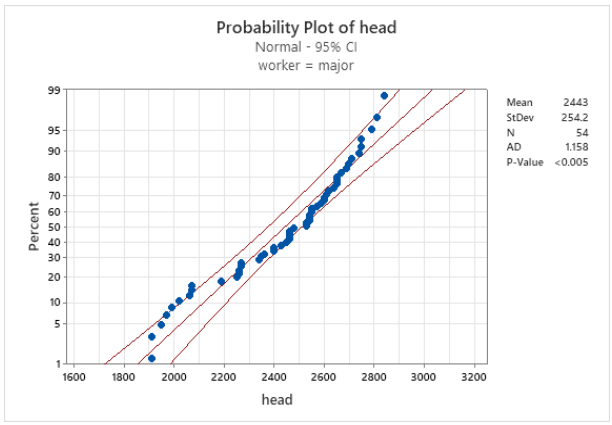
- *scape*: both distributions are unimodal and roughly normal in shape; centered a bit below 1900 and 2300 for minor and major workers, respectively; somewhat larger spread for minor workers, both in terms of the standard deviation and the IQR; both distributions fairly symmetrical (as would be implied by them being roughly normal); neither distribution has suspected outliers indicated in the boxplot; and normality tests are clearly non-significant ($P = 0.22$ and $P = 0.58$ with the AD-test), supporting them to be roughly normal. In summary, the distributions are of similar shape, but the one for major workers is shifted upwards and has a somewhat smaller spread relative to minor workers.
- *head*: the distributions are unimodal (with the possibility of a minor mode to the left for major workers, but it's not clear when looking at individual observations, e.g. in a dotplot, not shown) but of clearly different shapes; centered around 1300 and 2500, respectively; much smaller

spread for minor workers (e.g., the IQR values are 100 and 375); right-skewed and left-skewed (skewness values of 1.02 and -0.64, respectively); three suspected outliers in the right tail for minor workers (of which one looks a bit suspicious), but none for major workers; clear evidence against a normal distribution ($P < 0.005$) for both distributions (notice the different shapes in the normal probability plots). In summary, the distributions look quite dissimilar, but appear as clearly separated.





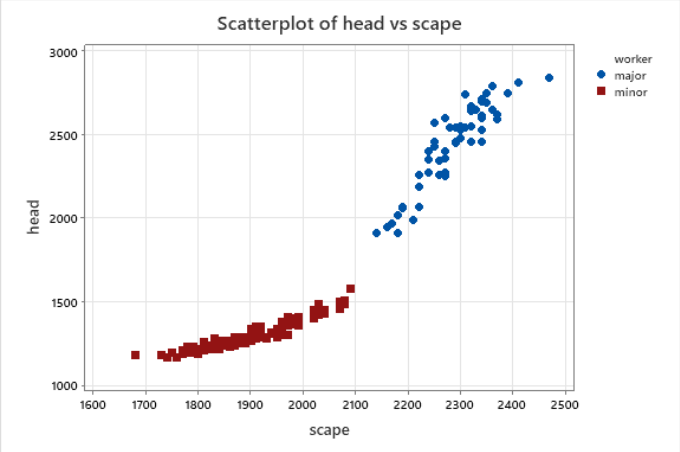




3. Relationship between scape length and head width

The relationship between two quantitative variables is best shown with a scatterplot. The two groups of workers may be shown in different plots or overlaid in the same plot with different plotting symbols and/or colours. We choose to put head width on the y-axis and scape length on the x-axis (as in the article), but there is nothing in the description of the study to suggest it cannot be done the other way around (see plot on next page).

It is evident from the plot that the two variables are related (as one would expect from body measurements of the same animals). The relationships appear both as positive, and maybe even roughly linear (with some noise, in particular for major workers). Visually the relationships appear quite distinct, and one would certainly not be able to fit one straight line through all the points.



4. Scape length probabilities

The scape length distributions were found to be reasonably normal, so for the calculations here we assume normality with the estimated parameters ($X \sim$ minor workers, $Y \sim$ major workers):

$$X \sim \mathrm{N}(1892.5, 83.8), \quad \text{and} \quad Y \sim \mathrm{N}(2286.9, 67.8).$$

The first two calculations below can be done using the standard normal table (e.g. Table B of PSLS) after calculating the z-score, as shown below; it is also perfectly fine to do the calculations directly in the respective distributions (e.g. with Minitab's Probability Distribution Plot menu, as shown below).

a) Calculated directly, the z-score equals $z = (2100 - 2286.9)/67.8 = -2.7566$, or we can write

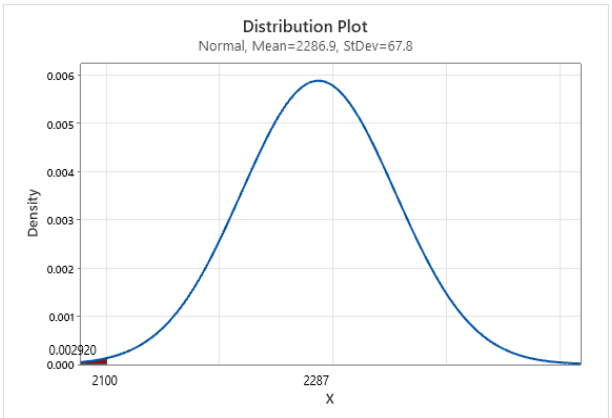
$$\mathrm{P}(Y < 2100) = \mathrm{P}\left(\frac{Y - 2286.9}{67.8} < \frac{2100 - 2286.9}{67.8}\right) = \mathrm{P}(Z < -2.7566) = 0.0029,$$

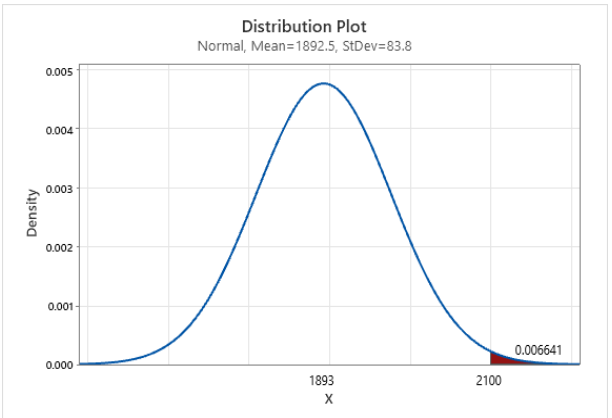
from Table B, or 0.002920 with a few added decimals from software.

b) Calculated directly, the z-score equals $z = (2100 - 1892.5)/83.8 = 2.2761$, or we can write

$$\mathrm{P}(X > 2100) = \mathrm{P}\left(\frac{X - 1892.5}{83.8} > \frac{2100 - 1892.5}{83.8}\right) = \mathrm{P}(Z > 2.4761) = \mathrm{P}(Z < -2.4761) = 0.0066,$$

from Table B, or 0.006641 with a few added decimals from software. The distribution plots for a) and b) are shown below.





c) The calculations above show that for $x = 2100$, the probability for minor workers is larger. When decreasing x, the probability for major workers will decrease and the probability for minor workers will increase, therefore the two probabilities will become more different. This means we need to increase x, and an approximate solution can be found by trial and error (which is considered an acceptable approach).

We can also solve by setting up an equation on z-scale, using the formulas from above with 2100 replaced by x and noting that the probability in b) is most easily calculated by switching from "$>$" to "$<$" (as indicated). The calculations go as follows,

$$\begin{aligned}
\mathrm{P}(Y < x) &= \mathrm{P}(X > x), \quad \text{or} \\
\mathrm{P}\left(Z < \frac{x - 2286.9}{67.8}\right) &= \mathrm{P}\left(Z > \frac{x - 1892.5}{83.8}\right) = \mathrm{P}\left(Z < -\frac{x - 1892.5}{83.8}\right), \quad \text{or} \\
\frac{x - 2286.9}{67.8} &= -\frac{x - 1892.5}{83.8}, \quad \text{or} \\
83.8(x - 2286.9) &= -67.8(x - 1892.5), \quad \text{or} \\
(83.8 + 67.8)x &= 83.8 \cdot 2286.9 + 67.8 \cdot 1892.5 = 130682.2, \quad \text{or} \\
x &= 130682.2/(83.8 + 67.8) = 2110.51.
\end{aligned}$$

With this value for x, the two probabilities both equal 0.004640.

The above method determined the cut-point between the two distributions by making the (mis)classification probabilities equal. Another idea is to choose x to make the two density curves equal. The resulting equations become a bit cumbersome to solve (a quadratic equation in x), but one can approximate the solution by calculating the density curves on a suitable grid of values (e.g. using Minitab's Calc−Probability Distribution menu), leading to a solution slightly below 2107.5. Several other statistical procedures can be applied to this problem as well, such as discriminant analysis and logistic classification. The methods differ in the assumptions made about the distributions and will therefore not give exactly the same answers.