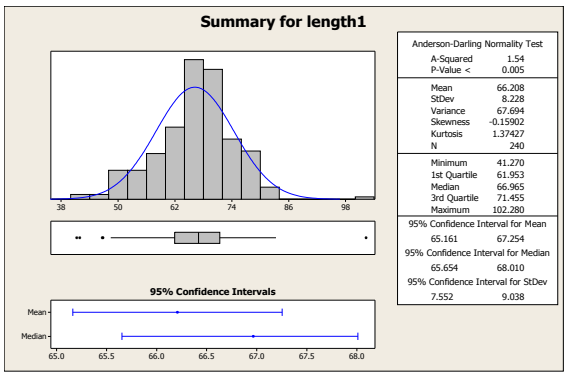
Solution to home assignment I

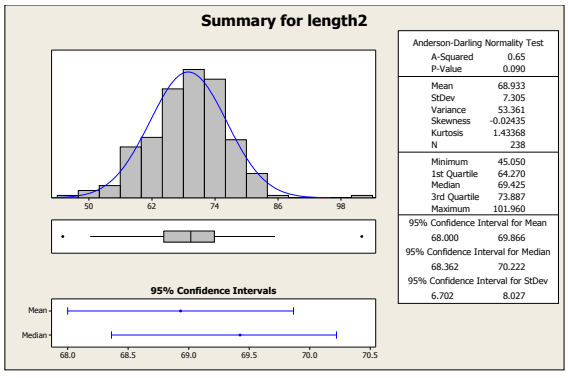
The solution presents one way in which the descriptive analysis could be done, but other ways are possible as well. It is more detailed than required for a 100% mark, by including all the variables for the descriptive analysis when only one continuous parameter was required for the assignment and by a detailed discussion of the questions related to the randomization.

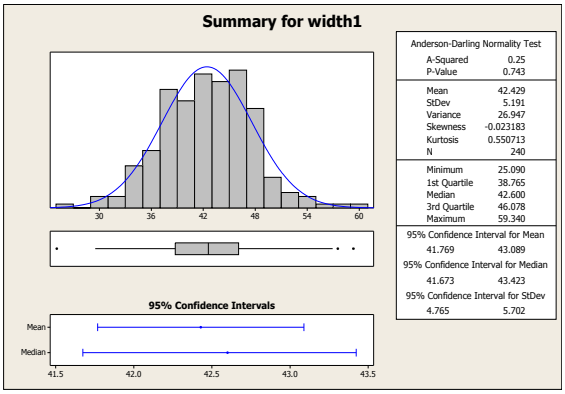
1. Descriptive analysis

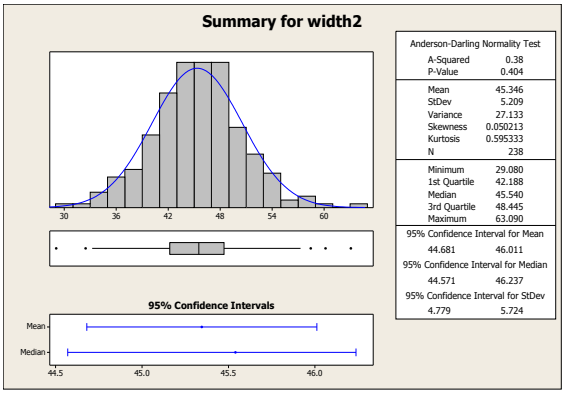
The variables *length*, *width*, *height*, and *weight* are quantitative and continuous (both when measured at the study beginning and end), *cage* is nominal categorical with six categories, and the variables *site* and *source* are both dichotomous (categorical with two categories).

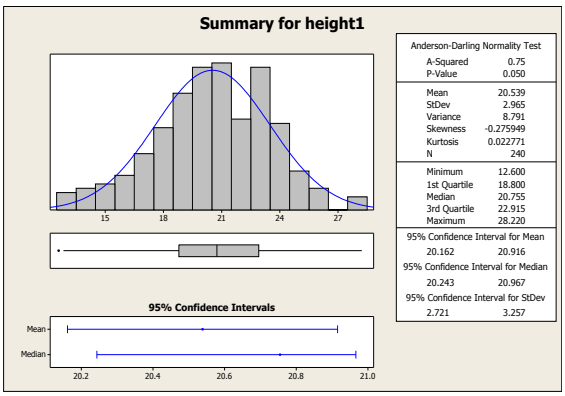
For simplicity the descriptive statistics and graphical display for the continuous variables will use the Graphical Summary menu in Minitab. With the quite large sample size, the most appropriate graphical representation of the distributions is a histogram (stemplots are useful as well), and the display also includes a box-plot and all the commonly used descriptive statistics (plus some we won't consider here, such as the confidence intervals). The histograms for start and end values of the same parameter should preferablybe displayed with the same bins (intervals); the number of bins can be adjusted from the software default (by editing the graph) if the histograms are considered too smooth or too rough. The cut-points for the "suspected outliers" indicated in the box-plot can be computed manually as the $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$ on the lower and upper side of the distribution, respectively. In a normal distribution, we would expect close to one suspected outlier from this rule with a sample of size 240 (calculated as $0.0035 \cdot 240 = 0.84 \approx 1$) in both the left and right tails (slide 1L−15). Some missing values occur among the measurements at the study end; these could be noted and explored.

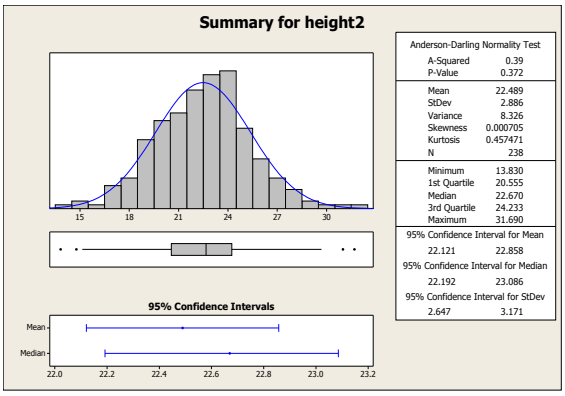


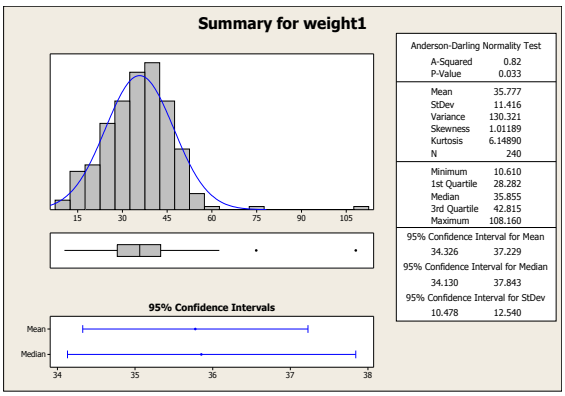


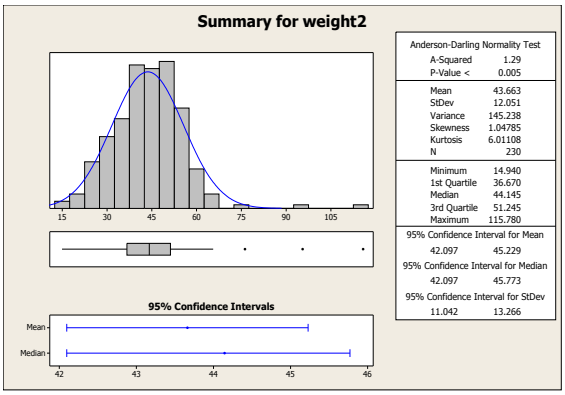




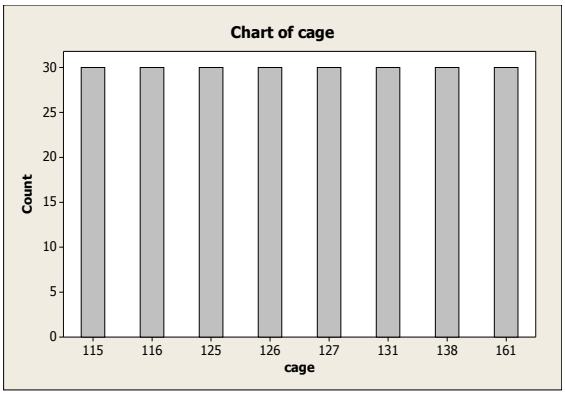


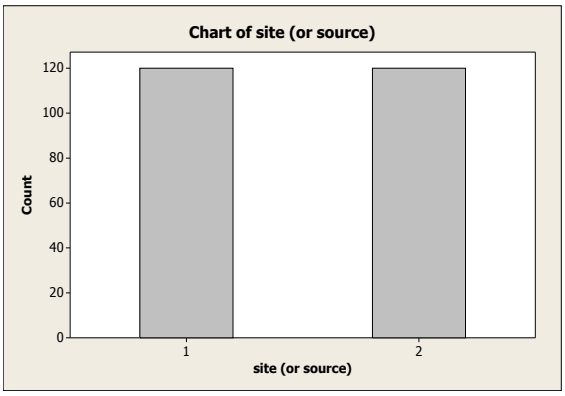






For categorical variables, descriptive statistics such as the mean and standard deviation and graphical displays such as a histogram (with overlaid normal curve) are less useful. The relevant statistics are the counts or proportions for each category, and a bar graph (or pie graph). It turns out all the categorical variables are perfectly uniformly (i.e., equally) distributed on the categories, so we'll just represent the distributions by bar graphs (even though one could argue these to be little informative because of the uniform distributions). The uniform distributions resulted from the experimental design and are not of real interest.





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

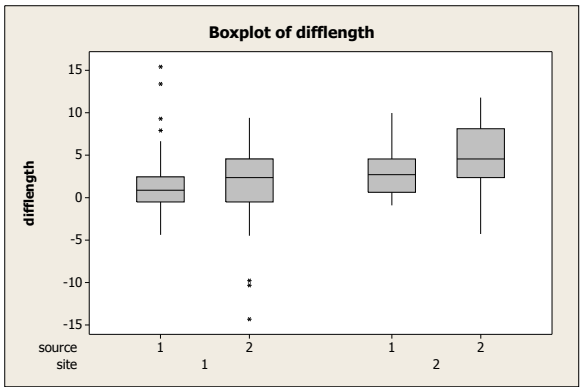
- *length*: unimodal; centered around 66−69 *mm*; fairly symmetrical with a few very extreme "suspected outliers" in both tails of the distributions. These are the main cause of the quite high kurtosis value (clearly > 1), but we prefer to look after the extreme observations directly

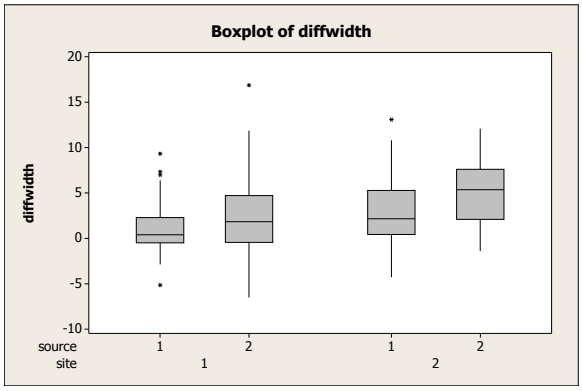
instead of the kurtosis value. The most extreme observations at the beginning and end of the study in both tails are from the same oysters, making it more likely that these were not errors but simply very small or large oysters (which may however not be representative of the population and therefore still be considered as real outliers); central part of the distributions spread fairly evenly around the medians with interquartile ranges of slightly less than 10 mm,

- *width*: unimodal; centered around 42−45 mm; almost perfectly symmetrical with only moderately extreme "suspected outliers" in both tails of the distributions; overall shape close to normal; central part of the distributions spread evenly around the median (16.8−18.0); most extreme observations at the beginning and end correspond to different oysters, but they nevertheless don't appear particularly extreme compared to the rest of the distributions and should probably not be considered as real outliers, one oyster however increased from 25.09 to 41.95 − a suspiciously large increase,
- *height*: unimodal (despite the ruggedness of the histogram for *height1*); centered around 20−22 mm; fairly symmetrical despite some left-skewness for *height1* and a shape not too far from normal; two "suspected outliers" in each tail for *height2* but not very extreme and hardly real outliers,
- *weight*: unimodal; centered around 35 and 44 g, respectively; central part of the distributions slightly left-skewed, but the distributions appear as right-skewed due to several very extreme observations in the right tail; the two largest observations in each distribution correspond to the same oysters and may therefore be valid values, but nevertheless very different from the remainder of the values (and thus perhaps not representative for the population ⇒ outliers).

2. Descriptive analysis for growth in four treatment groups

The four treatment groups are the combinations of the two dichotomous factors oyster source and growth site. Comparisons of the treatments could be based on the statistics and histograms in graphical summaries, as shown above, but we will here use box-plots to graphically represent the five-number summaries (and suspected outliers) of the growth distributions. The growth variables are computed by subtracting initial values from end values, e.g. as *difflength* = *length2* − *length1* (in Minitab, using the `Calc-Calculator` menu).



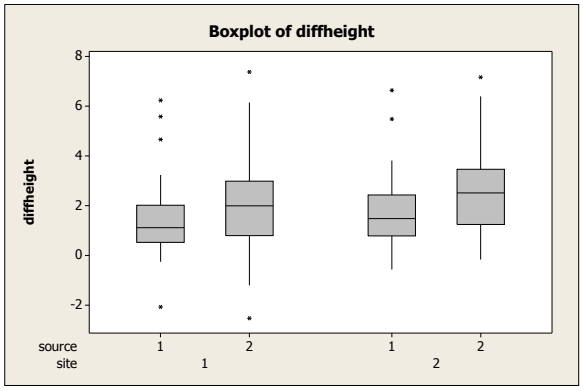


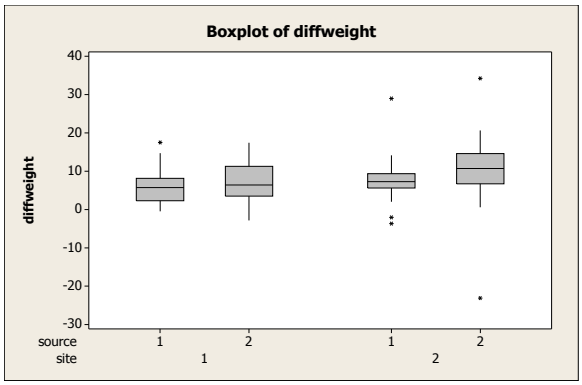
Brief comparative summaries of the distributions of the growth parameters between the four treatment groups (referring also to some of the descriptive statistics in the probability plots for Question 3):

- *length*: Growth was larger at the low salinity site (*site=2*) and for oysters from the low salinity source (*source=2*) when considering the medians and boxes (the middle 50% of the distributions) in the plots. Thus, the best performing treatment group seemed to oysters from a low salinity

source grown at a low salinity site. The variability (as represented by the box size, or IQR) in oyster growth seemed to be larger for the low salinity source within both sites. The negative extreme values seem suspect — how can an oyster become so much smaller?

- *width*: Again when considering the medians and boxes (the middle 50% of the distributions) in the plots, growth was larger at the low salinity site and for oysters from the low salinity source, and the combination of low salinity source and site appeared to perform markedly better than the others. The variability was pretty similar in all groups, considering that the smaller IQR for the high salinity site and source group could be offset by several extreme observations.

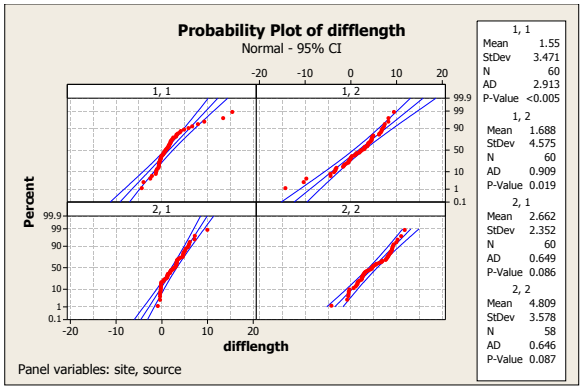


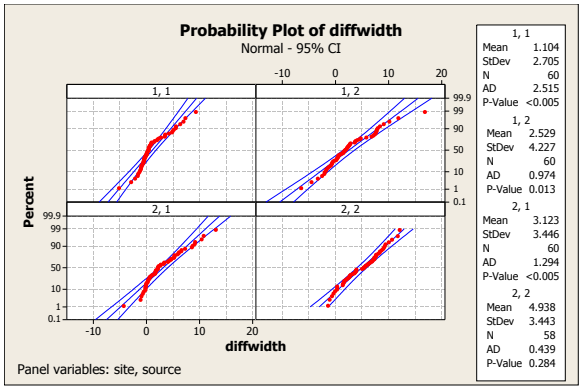


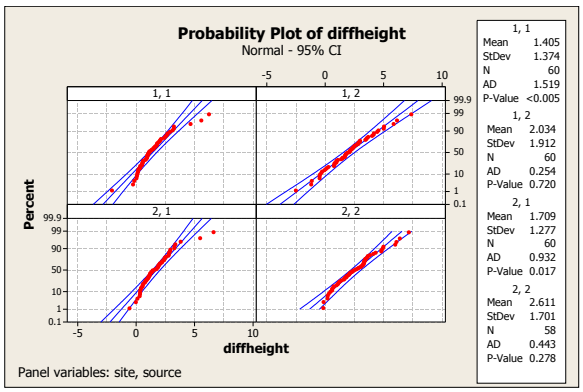
- *height*: The boxplots show less clear differences between sites as for the other growth parameters although the low salinity source and site again seems to outperform all other treatment groups. Contrasting all other parameters, one group at the high salinity site did slightly better than one of the low salinity site groups. The variability seemed larger for this variable relative to the group differences, and there was again a tendency of larger variability among low salinity source oysters.
- *weight*: The boxes of the weight distributions appear visually smaller than for other parameters but this due to scaling caused by very extreme observations at the low salinity site. Also here, the largest growth was found for the low salinity source and site. Assessments of variability will be strongly affected by whether extreme observations are considered as outliers or not; in particular, the loss of more than 20 g for an oyster at site 2 seems suspect.

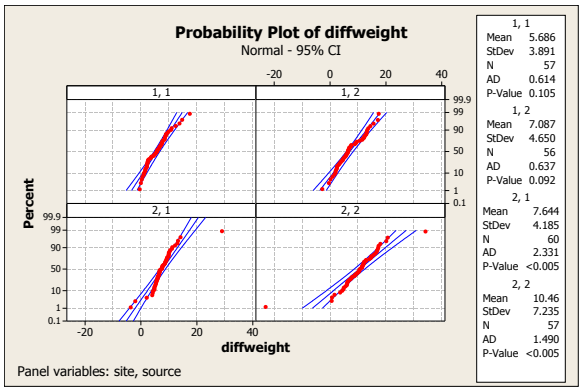
3. Normality of growth variable distributions

The best quantitative ways of assessing normality are probability plots and normality tests; these should be computed separately for each treatment group. Minitab can display the plots in separate windows, in separate panels of the same window (my preference) or overlaid in the same figure. In Minitab, the probability plot includes a P-value for the A-D (Anderson-Darling) test; we interpret low P-values as evidence against the normal distribution at a (not too strict) 0.05 significance level. Generally speaking, it's perfectly possible (and quite common) that the assumption of a normal distribution does not appear equally sensible in all treatment groups. This may be due to randomness in the data, but extreme observations in some groups can also play a role.









For all growth parameters, at least one of the 4 treatment groups showed strong evidence ($P < 0.005$) against a normal distribution. Reasons for the non-normality were either extreme observations (e.g., length and height for the high salinity site and source, as well as weight for the high salinity site) or skewness (e.g., 3 treatment groups for width) or both. Three of the four distributions for the high salinity and site group looked quite close to normal.

4. Randomization

(i) Randomization within or between cages.
If differences in growth are expected between cages, it will be more efficient to compare the two sources within than between cages. This corresponds to considering cages as blocks. Another advantage of having both sources in the same cage is that the design will be less affected if one or several cages are damaged or lost. So even without knowing for sure that growth differences might occur between cages, it is safer to distribute both sources within each cage. All this assumes that the logistical challenges of mixing sources within each cage are manageable; it is definitely easier to have one source per cage.

(ii) Randomization within a cage.
The randomization approach depends on whether it is assumed that all 32 positions in the 4×8 layout offer equally good growth conditions or specific assumptions are made that some positions are potentially advantageous. In the first case, the randomization within a cage corresponds to a completely randomized design, and the task is to randomly distribute 16 oysters from each of the low and high salinity sources onto the 32 positions. This requires labelling of the 32 positions as $1-32$ in some way, and then 16 numbers need to be selected randomly among these (for the low salinity source, for example). Table A can be used for this or statistical software. The figure below shows one such randomization obtained using Minitab where L and H indicate low and high salinity oysters,

respectively. These oysters should be selected randomly from pools of oysters from each of the two sources, and the added numbers 1−16 could correspond to a numbering of oysters within each of the two sources or the order in which these are taken from a pool, see the discussion under (iii) below.

	Column							
Row	1	2	3	4	5	6	7	8
1	L16	H11	H9	L7	L11	H12	L13	L6
2	H2	L8	L14	L5	L1	H8	H10	H13
3	H16	L4	L15	L3	H1	H15	L12	H14
4	L2	L10	H7	H5	H4	H3	H6	L9

The irregularities in the resulting desing reflect the randomness of the procedure. For example, column 6 happened to contain only high salinity oysters.

Other designs will rely on specific assumptions about the possible differences in growth conditions within the cage. For example, one may take rows as blocks (hence randomization is done within each row) or take columns as blocks (randomization within each column). As a slightly different example, we will consider the possibility that oysters in the exterior positions in the cage potentially have different conditions than those in the interior positions. For simplicity, we will form just two blocks, corresponding to exterior and interior positions, as shown in the figure below. Note that the two blocks are not of equal size: there are 20 exterior and 12 interior positions. This is not a concern here because they both hold far more observations than treatments (2). Randomization then follows within each of the two blocks, with one particular randomization shown below which in its first step randomly selects oysters from within each of two sources for exterior and interior positions.

	Column							
Row	1	2	3	4	5	6	7	8
1	L6	H2	L8	L14	H8	H10	H13	H16
2	L4	L16	H11	H9	L7	L11	H12	L15
3	L3	L13	L5	L1	H7	H5	H3	H1
4	H15	L12	H14	L2	L10	H4	H6	L9

(iii) Randomization between sites.
The study sites are allocated to treatments by deciding which oysters (of either low or high salinity source) are deposited at each of the two sites. The logistics of randomly selecting oysters from a pool, e.g. a bag of oysters, and here to allocate them to two sites, is not totally straightforward when oysters cannot be expected to have individual IDs. So randomly reordering numbers (or rows in a worksheet) does not in itself help without an explanation of how the numbers (rows) are linked to the individual oysters. Also, just picking oysters ("randomly") out of the bag does not necessarily correspond to simple random sampling, because the larger oysters might be picked first (or the smaller ones).

When allocating oysters into two groups (corresponding to sites) the effect of such non-random selection can be eliminated by having a randomly determined list of allocations into sites A and B, say, based on the order in which oysters are picked from the bag. That will effectively achieve a complete randomization of the order of sampling, and hence be equivalent to simple random sampling. A commonly used and easier method than simple random sampling is systematic random sampling (Lecture 2, slide 2L−14), which here would correspond to alternating between sites A and B for the oysters picked out of the bag. Systematic random sampling does not achieve a full randomization and is therefore considered slightly inferior; in practice, it should work fine as well.