Biostats VHM 801 Course
Fall 2019, Atlantic Veterinary College, PEI
Henrik Stryhn

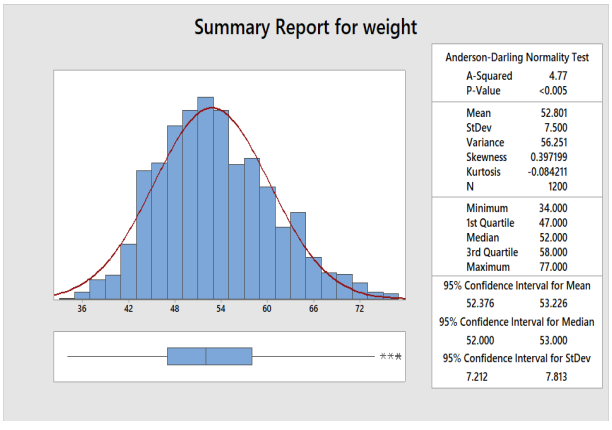
Solution to home assignment I

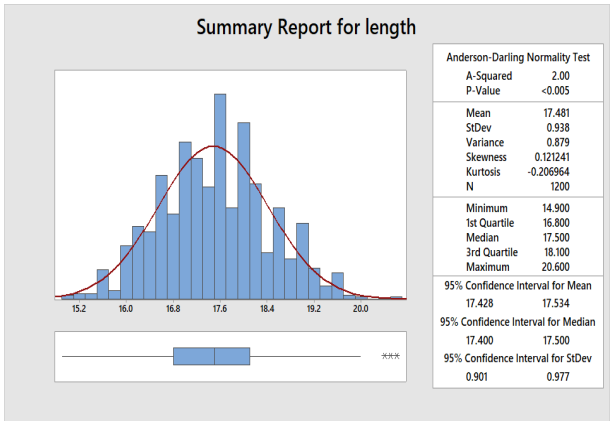
The solution presents one way in which the descriptive analysis could be done, but other ways are possible as well. It is much more detailed than required for a 100% mark, by including all the variables for the descriptive analysis when only 3 selected variables were required for the assignment, by including multiple ways to carry out the randomization in Question 4 and by giving detailed answers to both its subquestions. All analyses shown used Minitab 18, but Stata or other programs would give similar figures and results.

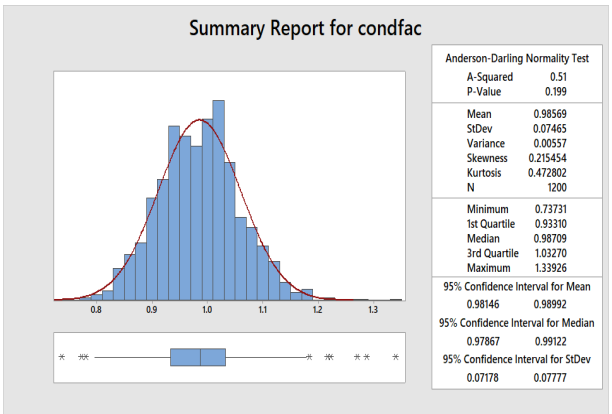
1. Descriptive analysis

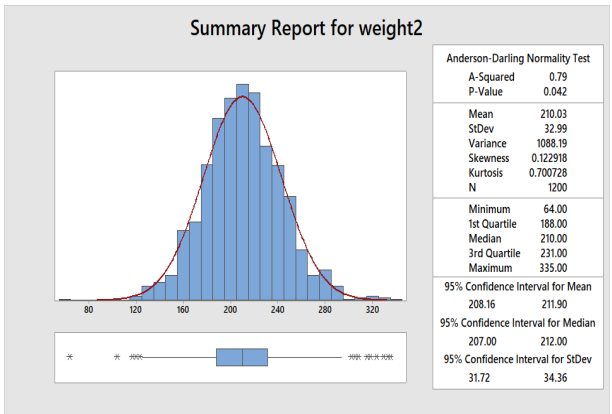
The variables *weight*, *length*, *condfac*, and *weight2* are quantitative and continuous, *vaccine* and *sex3* are categorical and nominal with six and three categories, respectively, and the remaining variables, *sex*, *operc*, *jaw*, and *prec*, are all dichotomous (categorical with two categories).

For simplicity the descriptive statistics and graphical display for the continuous variables will use the Graphical Summary menu in Minitab. With the 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). The cut-points for the "suspected outliers" indicated in the box-plot can be computed manually as $(Q_1 - 1.5 \times IQR)$ and $(Q_3 + 1.5 \times IQR)$ on the lower and upper side of the distribution, respectively. In a normal distribution, the expected number of suspected outliers from this rule in a sample of size 1200 is $0.35\% \cdot 1200 = 4.2$ in both the left and right tails (slide 1L–18).





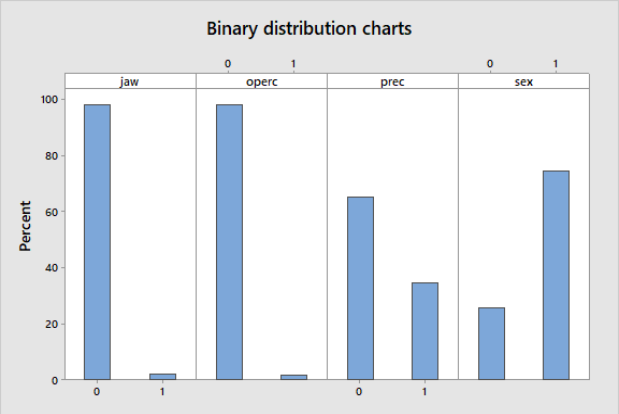


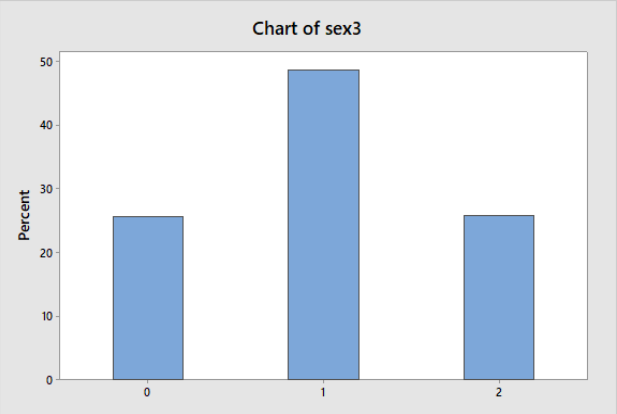


For categorical variables, descriptive statistics such as the mean and standard deviation and graphical displays such as a histogram (with overlaid normal curve) are less useful. The relevant statistics are the counts or proportions for each category, and a bar graph. The variable *vaccine* has a perfectly uniform distribution on the six vaccine groups (200 fish, or 16.67% in each group), so this variable is not included below (as its distribution is already described). The variable *sex3* is really a composite of *sex* and *prec*, whereby the two categories for males from *prec* are contrasted with the females (that have missing values for *prec*). It is valid to compute proportions for the resulting three categories, but it is perhaps more natural to consider the proportions for *sex* and *prec* separately.

Observed probability distributions for binary variables (counts and proportions in parenthesis):

value	*sex*	*operc*	*jaw*	*prec*	*sex3*
0	307 (.256)	1178 (.982)	1176 (.980)	584 (.654)	307 (0.256)
1	893 (.744)	22 (.018)	24 (.020)	309 (.346)	584 (0.487)
2					309 (0.258)





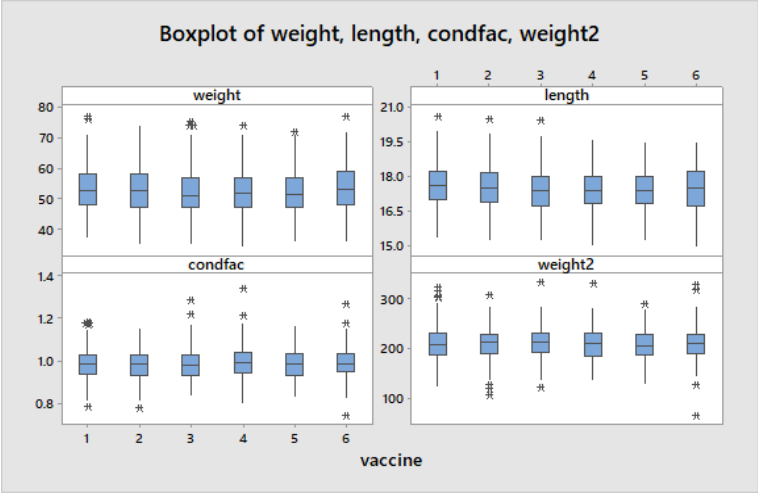
Finally, brief summaries of the distributions based on the computed statistics and graphs:

- *weight*: unimodal; centered around 52 *g*; slightly right-skewed (skewness = 0.40) with "suspected outliers" only in the longer, right tail; central part of the distribution spread quite evenly around the median ($47 - 58$); the overlaid normal curve approximates the histogram quite well, but AD-test gives very strong evidence against a normal distribution ($P < 0.005$); a total of 4 "suspected outliers" (> 74.5), and none of these seem suspiciously large,
- *length*: unimodal; centered around 17.5 *cm*; central part of the distribution spread quite evenly around the median ($16.8 - 18.1$); roughly symmetrical and bell-shaped; rugged shape of the histogram, due to some values far more (e.g. 17.0 (88 obs.), 18.0 (97)) or less (e.g. 17.1 (25), 18.1 (23)) frequently represented than expected in a smooth distribution — this probably represents measurement error or round-off; due to the ruggedness, clear evidence against a normal distribution; only 3 high (> 20.05) "suspected outliers", and none of these seem as real outliers,
- *condition factor*: unimodal; centered around 0.986; fairly symmetrical central part of distribution; a minor right-skewness and some extreme observations in both tails (3 low (< 0.784) and 6 high (> 1.182) "suspected outliers"), perhaps the largest of these (fish 777) looks a bit suspect; no evidence against a normal distribution (by the AD-test),
- *weight at transfer*: unimodal; centered around 210 *g*; fairly symmetrical and narrow central part of distribution ($188 - 231$); long tails on both sides (presumably the reason for the elevated kurtosis (0.70)) and a total of 14 "suspected outliers" ($5 < 123.5$; $9 > 295.5$), of which perhaps the lowest value looks suspect; some indication by the AD-test against a normal distribution,

- *operc* and *jaw*: both distributions have around 98% of observations at 0, corresponding to normal fish, and 2% of fish with deformities,
- *prec* and *sex*: 74% of the fish are males, and of these 35% show early maturation (*prec*=1).

2. Descriptive analysis, including groups defined by categorical variables

The focus in this part is on comparative statistics and graphics. We demonstrate the use of comparative box-plots to show the difference between vaccine groups, and only give means for the two categories of the binary variables (these could be represented graphically as well).



The four box-plots show no obvious differences between the vaccine groups; the medians and boxes are very similar across groups, and the differences between asterisks seem mostly of random nature. For the three measurements at vaccination, we would not have expected to find any differences between vaccine groups because the fish were randomly allocated to them at that time.

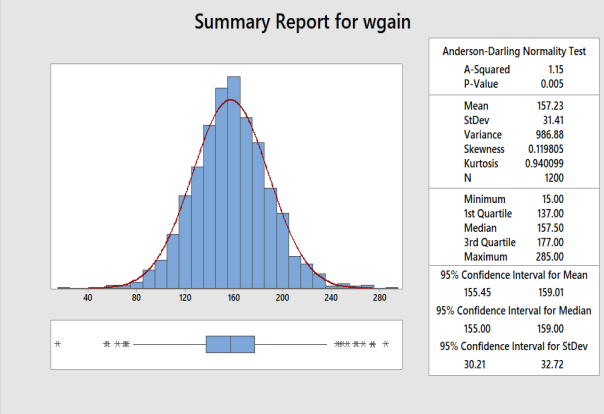
Means of continuous variables across categories:

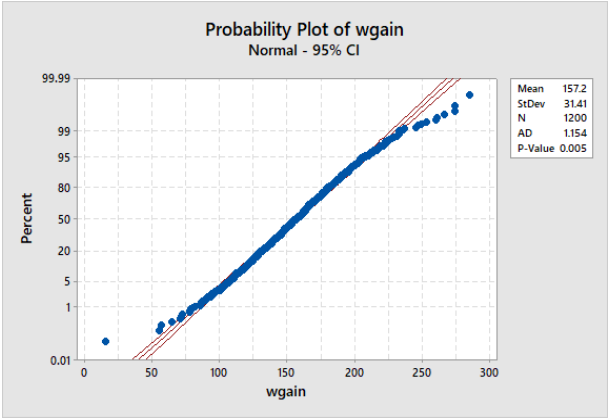
variable	category	*weight*	*length*	*condfac*	*weight2*
operc	0	52.84	17.49	0.9857	210.4
	1	50.91	17.26	0.9867	188.1
jaw	0	52.79	17.48	0.9851	210.3
	1	53.17	17.33	1.0148	196.6
prec	0	52.64	17.45	0.9886	208.4
	1	51.65	17.34	0.9885	213.7
sex	0	54.26	17.69	0.9774	209.4
	1	52.30	17.41	0.9886	210.2

For the interpretation of these means, we should really also consider their standard errors, but this is considered beyond the scope of the home assignment. Let it suffice here to say that the groups of fish with the first two deformities (*operc* and *jaw*) are very small, so that mean differences should be interpreted cautiously. For *weight*, there seems to be a difference between sexes, with females being heaviest and precocious males being lightest (and the non-precocious males inbetween). The same pattern is seen for *length*, which also shows smaller sizes of fish with operculum and jaw deformities; given the small sample size, it is not clear whether these are substantial findings or not. The findings for *condfac* are opposite, which given its computation from weight and length is perhaps not too surprising. Finally, it seems that there may be effects of all variables except *sex* on the weight at transfer; e.g., operculum and jaw deformities may be associated with smaller weights at transfer.

3. Descriptive analysis and normal distribution for the weight gain

We compute the weight gain as $wgain = weight2 - weight$. For our descriptive analysis, we use again the Graphical Summary, and supplement with a normal probability plot.





The histogram looks symmetrical, the median and mean are close, and the box of the box-plot shows perfect symmetry. However, the kurtosis is large (0.94), and the box-plot shows a large number of "suspected outliers"; we count $7 + 10 = 17$ observations less than 77 g or greater than 237 g. It is also seen that the overlaid normal curve does not quite reach the peak of the histogram. The probability plot shows a straight line in the central part of the distribution, but the tails fall clearly off the line. The bounds in the probability plot are very narrow, due to the large sample size, but even without these it is clear that the tails are too heavy (or large) for a normal distribution. It is no surprise that the AD (Anderson-Darling) normality test gives clear significance against a normal distribution. Although some of the smallest and the largest values of *wgain* are somewhat off the rest, it is not clear that these would not be proper values from the distribution. There is, however, one very small value (fish 878 with a weight gain of only 15 g, mostly due to having the lowest weight (64 g) at transfer), and it certainly looks suspect. Generally speaking, when as in this case the tails are too heavy for a normal distribution, it is not easy to assess whether observations are truly outlying or not, and we should always be cautious in decisions to declare some observations as "true outliers".

4. Randomization

For a fish population size of $N = 2100$, we would expect to have $2100/6 = 350$ fish per vaccine group. In this solution, four valid methods will be described for the randomization. Because the actual sample size, given as 2000, will not be known before the sampling is completed, any randomization scheme can obviously not be based on that number. As mentioned in the assignment, it is unreasonable to assume that the order in which the fish are sampled will be totally random. For example, the stronger fish may be able to resist capture longer and therefore be sampled at the end. For this reason, even if the total number of fish sampled was known in advance, it would not be valid to simply allocate the first one sixth of the fish sampled to the first vaccine group, and so on.

i) Pre-randomization based on tagging prior to vaccination or on sampling order
If it is assumed that the 2100 fish are tagged prior to vaccination, we can randomly assign vaccines to the 2100 fish numbers, and then at vaccination identify the fish and its vaccine group by the tag. Alternatively, if the fish were not pre-tagged, we could randomly assign vaccine groups $1 - 6$ to the sampling order from 1 to 2100: this would require a large table giving the vaccine groups for each fish number sampled (e.g., fish no. $1 \rightarrow$ vaccine 4, no. $2 \rightarrow 3$, no. $3 \rightarrow 6$, no. $4 \rightarrow 3$, and so on). For a total loss of $n = 100$ fish from the population, the count (X) of lost

fish from vaccine group 1 (say) would approximately follow a binomial distribution $\text{Bin}(100, 1/6)$; this is a sampling from a finite population, where $n < N/20$. Therefore, $\text{E}(X) = 100 \times (1/6) = 16.7$ and $\text{sd}(X) = \sqrt{100 \times (1/6) \times (5/6)} = 3.7$. For the number ($Y$) of fish from vaccine group 1 included in the study, i.e. $Y = 350 - X$, we therefore have $\text{E}(Y) = 350 - \text{E}(X) = 333.3$ and $\text{sd}(Y) = \text{sd}(X) = 3.7$. The possible range for X is all values $0, \ldots, 100$, and therefore the possible range for Y is all numbers $250, \ldots, 350$. The calculated mean and standard deviation show that the likely range is much smaller.

ii) Random allocation of each sampled fish
If each fish sampled is independently, randomly allocated to a vaccine group (similar to rolling a die for each fish), the number of fish (Y) in vaccine group 1 would be binomially distributed $B(n, 1/6)$, where n now stands for the total number of fish included in the study. For $n = 2000$, we get $\text{E}(Y) = 2000 \times (1/6) = 333.3$ and $\text{sd}(Y) = \sqrt{2000 \times (1/6) \times (5/6)} = 16.7$. It is seen that this method produces a much more variable count in each vaccine group than method *i)*. The range of Y is all numbers $0, \ldots, 2000$ (although most of these counts would be highly unlikely).

iii) Random allocation within suitable blocks
One drawback of method *ii)* is that it relies on the long-term averaging to achieve an approximately equal count for each of the vaccine groups. By blocking the random allocation in groups of e.g. 30 fish (i.e., 5 fish per vaccine), equal counts are obtained within each complete block. The procedure would randomly allocate fish sampled as no. $1, \ldots, 30$ to vaccine groups according to a pre-defined random allocation of these numbers to the 6 vaccine groups, then fish sampled as no. $31, \ldots, 60$ would be randomly allocated according to another pre-defined random allocation, and so on. Blocking could also eliminate some of the effects of the fish not being sampled in totally random order, if the effects correspond to a trend over time. The block size (30 in the example) should be a multiple of 6, but may be chosen as larger or smaller depending on the logistics.

A total sample size of 2000 fish corresponds to 66 complete blocks ($66 \times 30 = 1980$) and 20 fish from one incomplete block. The count (X) for vaccine group 1 among the fish from the incomplete block cannot be assumed to follow a binomial distribution $B(20, 1/6)$ because 20 out of the 30 fish in the block were sampled (so this is sampling from a finite population). It is intuitively obvious that the mean equals $\text{E}(X) = 20 \times (1/6)$, and it can be proven by the addition formula for means applied to the 0/1-variables $S_1, \ldots, S_{20}$ indicating whether each of the sampled fish were allocated to vaccine group 1. Therefore, the mean of the total count ($Y = 66 \times 5 + X$) for group 1 equals $\text{E}(Y) = 66 \times 5 + 20/6 = 333.3$. Its standard deviation is more difficult to compute; using formula for the hypergeometric distribution (which are not covered by the course), it can be shown to be equal to 0.98. However, the range can be established as the numbers $330, \ldots, 335$, because the possible range for X is $0, \ldots, 5$.

iv) Randomization by looping through groups 1–6
Systematic random sampling is a sampling method by which every k^{th} member of the population is sampled, after the first member is selected randomly (slide 2L–12). By choosing the block size of method *iii)* equal to 6, and using an allocation within a block starting from a randomly selected start number (e.g., if the number 3 is chosen, the sequence of vaccines is 3, 4, 5, 6, 1, 2), gives a randomization scheme similar to systematic random sampling. With a randomly selected start number, the mean and range can be determined in a similar way as for method *iii)*: $\text{E}(Y) = 333 + 2 \times (1/6) = 333.3$, and the possible values are 333 and 334 (only). Without a random component (e.g. if the looping simply starts from 1), such a method should not be called a randomization (instead it is a treatment allocation), and may be criticized as inadequate.

It is seen that all four methods have the same mean count of observations per vaccination group, and the differences are in their variability and patterns. Generally, systematic random sampling would be sensitive to the occurrence of certain patterns in the sampled fish, e.g. by different people catching the fish. For this reason, it is sometimes referred to as quasi-randomized. If such patterns do not exist, the method is attractive by its simplicity and the narrow range it gives for the counts in each treatment group. The only method with an equally narrow range would be a blocking scheme with block size 6, where the allocation within each block is truly random, but that seems hard to manage logistically. Method *ii)* gives a good safeguard against unexpected patterns in the allocation, but at the cost of quite variable group sizes. Method *i)* is the obvious choice if the fish are pre-tagged, and this was the method actually used for the study.