Comments on DA#1

EDA requires detective work to understand the characteristics and features of a set of data. It is important to keep an open mind and not jump to conclusions that would preclude consideration of alternatives. EDA is not definitive. It is a process of discovery. EDA may result in insights, and also in questions for further discussion with investigators. The abalone data analysis assignment presents these aspects.

The original investigators were unable to predict the age of abalones based on physical measurements. In this sense their study was a failure. This EDA assignment had the specific objective of ascertaining plausible reasons for this outcome based on the data that were collected. Study results would necessarily pertain solely to the immediate locations sampled and surrounding environs. This does not mean another study designed and conducted differently with additional variables would have the same outcome. The scientific process involves learning from failure. That may be the case here.

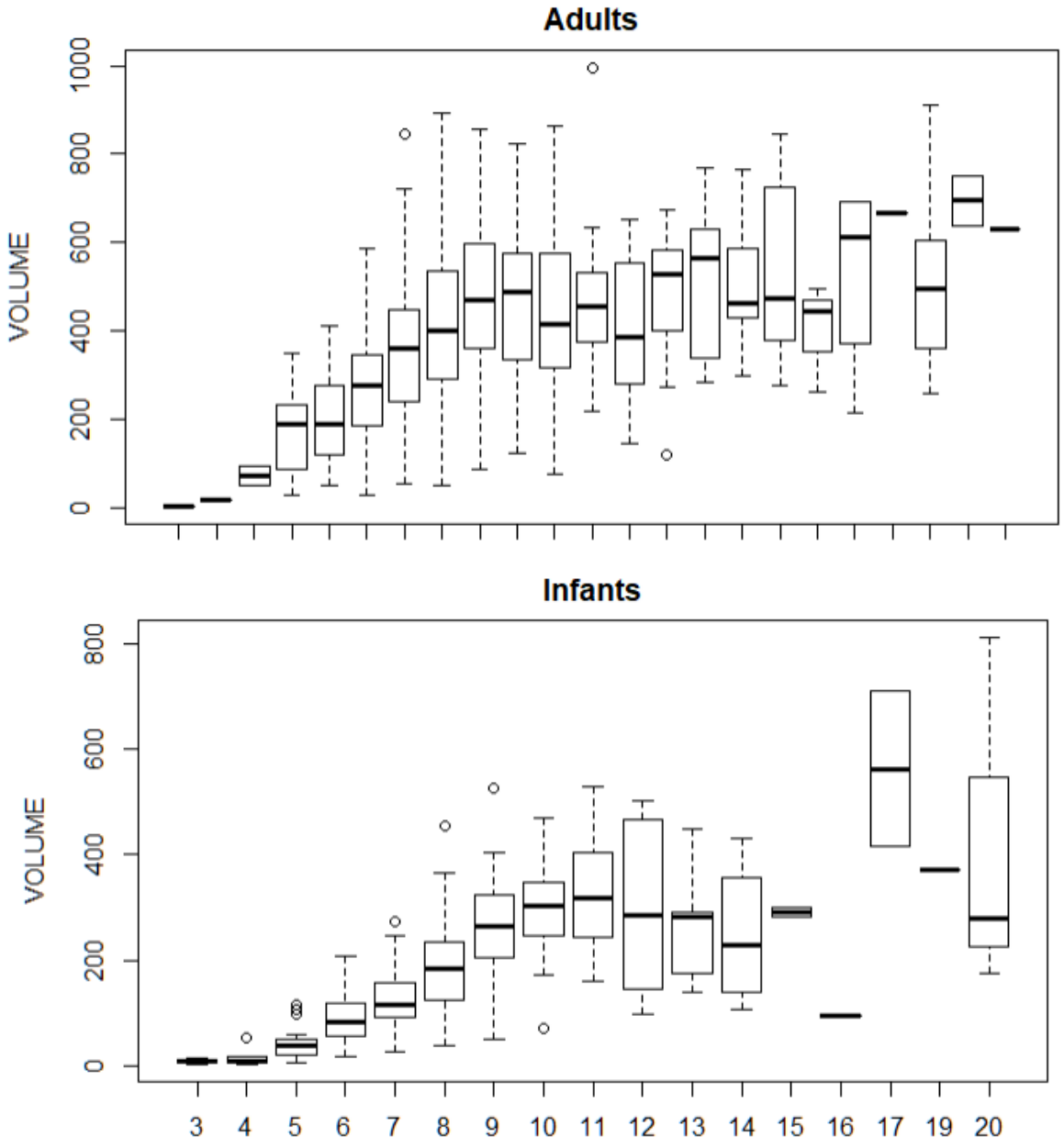
1) The first part of the EDA assignment suggests the data do not constitute a random sample. Examining the data reveals the count of infants in A1 to be less than the count of infants in A2. To sustain an adult population there needs to be a sufficient number of infants in A1 to do this. One might reasonably expect the counts to be reversed. There are different possible explanations. Possibly the abalone population is damaged in some way, or the sampling procedure was in areas more heavily populated with older abalones or the sampling was selective focusing on larger, older abalones leaving the smaller ones behind. This is the type of observation that should be discussed with the investigators if they were available. Regardless, it is possible to examine the relationship between age and physical dimensions among those abalones harvested recognizing we are dealing with this specific set of data which may not be indicative of other locations.

2) Why are there “infants” in A4 and A5? The notion that all abalones mature sexually by 6 years may not pertain to the population studied. Delayed sexual differentiation is a possibility. Determining sexual differentiation at different ages or at different times of the year could also play into this. This is not to say the investigators were sloppy. Identifications could be indeterminate in some cases due to the state of the art.

These data indicates “infants” are on average smaller than adults particularly in A4 and A5. This becomes apparent when variables such as VOLUME or WHOLE are compared between the infants and the adults as a function of RINGS. Note the boxplots below.

This observation is deserving of additional study. There may be something fundamental taking place with these abalones from a physiological perspective. Such a discovery may be more important than figuring out why age prediction is difficult.

3) The rule of thumb regarding maturation after reaching 10 rings is one of slowing growth rather than sexual development. The boxplots below tell the same story about growth whether Adults (Males and Females combined) or Infants are considered. The same type of thing is present for WHOLE and SHUCK. The limited number of abalones with larger ring counts leads to the variability observed in the boxplots.



4) What about the defined age classes? These are a way to classify abalones so that generational differences can be studied without the necessity of resorting to individual ring counts. Conclusions reached based on age classes can always be verified to see if they hold up when ring count is considered. Looking at the abalones in terms of age classes amounts to viewing the data as a cohort study of different abalone generations in the sense that older abalones will have been exposed to more risks, and differing circumstances, than younger abalones. Having worked on marine environmental studies, I can attest to the fact that the marine environment is far from uniform. The variability we see is consistent with this perspective.

5) Regarding the failure of the original study, from a statistical perspective, it boils down to not being able to determine a useful descriptive relationship between physical dimensions and ring count (or age classes). Two things are at work here. First, as abalones mature their growth slows down so that changes in physical dimensions do not correspond closely to increases in ring count. Second, apparently there is considerable variability in growth rates resulting in abalones of differing sizes having the same or similar ring counts. We might question the precision of counting rings; and, this should be discussed with the investigators. However, at the moment not having evidence to the contrary, I do not view errors in ring counting as the probable reason for the rather large observed variability in physical dimensions.

We could attempt to develop descriptive relationships between physical dimensions and age (ring count) for each sex. Even if outliers were thrown out, and we looked at each sex individually, we would find substantial unexplained variability in physical dimensions within each age class and no useful descriptive relationship for precise age prediction across these segments. If such a relationship does not exist for each sex separately, these data do not support a relationship across the board. It may be possible to separate out infants in A1 and A2 from adults, but this will be not be error free. We will explore this possibility in the second data analysis.

6) Not having information on abalone locations, sampling depths, nutrition, predation, prior harvesting or disease further complicates the picture. In the final analysis the investigators ended up with variation they could not explain, and were unsuccessful in determining how to predict age with precision based on physical dimensions alone.

Another study might have different results.