Narwhal: Development of de novo Reference intervals from blood sampling results

Sandie Black

April 5, 2017

Introduction: Narwhals (Monodon monoceros) are medium sized toothed whales, mesopredators of key importance to the Canadian Arctic marine ecosystem. They are also of great sociocultural importance to Canada's Inuit. During the summers of 2004, 2005 and 2009-2012 narwhal were captured in two locations on northwest Baffin Island (Nunavut, Canada) using previously described methods (Dietz et al, 2001), and handled for the attachment of satellite transmitters. During the handling period, blood samples were collected from each whale from the ventral fluke vascular bundles. Blood smears were made for indirect white blood cell counts, and samples were centrifuged to separate cellular components from serum. Separated serum was frozen at -80C (in a liquid nitrogen dry shipper, Wharton Inc) for storage and transport. All serum samples and prepared slides were submitted to Idexx laboratories(Canada Corp., Markham, ON) for analysis.

There is very little information in the literature regarding reference intervals (RI) for narwhal. One paper representing hemogram information from three animals (MacNeill, 1975), and a second with values from a single animal(DeMonte and Pilleri, 1983) represent all the previous published values found through a comprehansive literature search. RI form an integral part of baseline health information, and are an important component of both clinical diagnostic testing and decisions. Reference intervals are developed from values collected from comparable healthy individuals within a population, and care must be taken to both carefully identify and select individuals for inclusion in RI, and to detect and remove outliers as outlined in Friedrichs (et al, 2012). The following project uses data gathered during six Arctic field seasons (Raw.data.narwhal.csv) which consists of the results from 44 individual narwhal for 22 different measures of blood cell counts, electrolytes, protein and enzymes. The consensus guidelines for development of de novo RI in veterinary species (ASVCP, 2011) will be followed. This resource contains robust methodology for the calculation of RI when sample sizes are small.

In our study, the selection of animals from which reference values were collected was random, represented by subadult and adult narwhal which were entrapped in a net. All animals were assessed as healthy based on physical exam, behaviour during handling, and on release. Definition of the reference population with inclusion and exclusion criteria is of key importance in the development of RI. Analytical errors were minimized through protocols for collection and storage of samples, and the use of a consistent, well recognized diagnostic laboratory for all analyses (Friedrichs et al, 2012)

To begin a statistical analysis of reference values collected from a population of interest, the first step suggested is to prepare and examine histograms which illustrate the data distribution and the presence of potential outlier data points (Friedrichs et al, 2012). As an example we will examine one column of data composed of measures of phosphorus and conduct summary statistics and graphs to examine the data for normal (Gaussian) distribution.

#Install packages we will need for graphing and manipulating data:  
install.packages("dplyr",repos = "http://cran.us.r-project.org")

##   
## The downloaded binary packages are in  
## /var/folders/\_w/7cb4\_27n4r10cgyy4grkgxph0000gn/T//Rtmpz6wfFo/downloaded\_packages

install.packages("ggplot2",repos = "http://cran.us.r-project.org")

##   
## The downloaded binary packages are in  
## /var/folders/\_w/7cb4\_27n4r10cgyy4grkgxph0000gn/T//Rtmpz6wfFo/downloaded\_packages

library(dplyr)

##   
## Attaching package: 'dplyr'

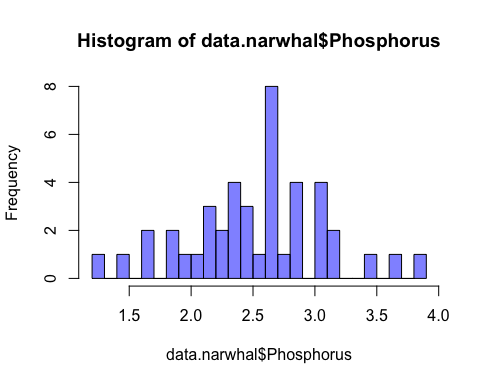
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

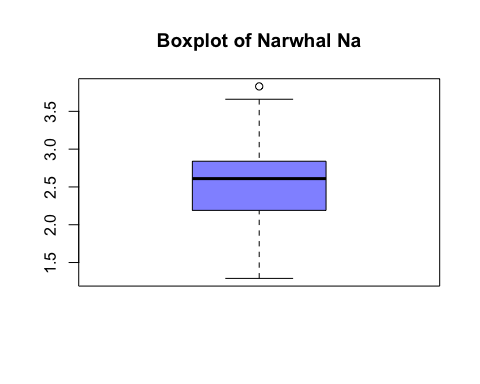
library(ggplot2)  
  
#Examine the data in more detail for phosphorus:  
summary(data.narwhal$Phosphorus)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.290 2.190 2.610 2.526 2.840 3.830 1

hist(data.narwhal$Phosphorus, breaks = 20, col = rgb(0,0,1,0.5))



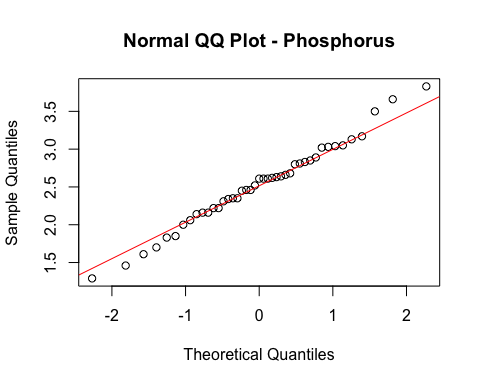
boxplot(data.narwhal$Phosphorus, col = rgb(0,0,1,0.5), main = "Boxplot of Narwhal Na")



shapiro.test(data.narwhal$Phosphorus)

##   
## Shapiro-Wilk normality test  
##   
## data: data.narwhal$Phosphorus  
## W = 0.98866, p-value = 0.9427

qqnorm(data.narwhal$Phosphorus, main = "Normal QQ Plot - Phosphorus")  
qqline(data.narwhal$Phosphorus, col = "red")

 Looking at the boxplot, there may be one outlier value on the upper end, quartiles look even; the Shapiro-Wilk test with p= 0.94 (> .05) retains Null hypothesis of normally distrubuted data. The QQ plot shows relatively good alignment with the theoretical quantile line in the middle ranges of the data, with several potential outliers at the minimum and maximum value ends.

The next step is to examine the Phosphorus data points for the presence of outliers on each end of the data. ASVCP suggest using Dixon's outlier range statistic. This will require the downloading of the outliers package in R. Both the Dixon test and Grubb's test in R will only calculate for samples between 3 and 30 however, so we will use the "chi-squared test for outliers" function which is less sensitive but will run for sample sets greater than 30 observations. The NA's are removed before the chi-squared test is run, and the test is run for outliers on both tails of the data.

install.packages("outliers",repos = "http://cran.us.r-project.org")

##   
## The downloaded binary packages are in  
## /var/folders/\_w/7cb4\_27n4r10cgyy4grkgxph0000gn/T//Rtmpz6wfFo/downloaded\_packages

library(outliers)  
  
Phosphorus <- data.narwhal$Phosphorus[!is.na(data.narwhal$Phosphorus)]  
  
#Chi-Sq Tests for Outliers for phosphorus  
chisq.out.test(Phosphorus,variance = var(Phosphorus),opposite = TRUE)

##   
## chi-squared test for outlier  
##   
## data: Phosphorus  
## X-squared = 5.0447, p-value = 0.0247  
## alternative hypothesis: lowest value 1.29 is an outlier

chisq.out.test(Phosphorus,variance = var(Phosphorus),opposite = FALSE)

##   
## chi-squared test for outlier  
##   
## data: Phosphorus  
## X-squared = 5.6184, p-value = 0.01777  
## alternative hypothesis: highest value 3.83 is an outlier

Using a level of significance equal to 0.05, we see that the p-value at the lower end of the data is less at 0.025 and thus accept the minimum value in the data set as an outlier. Similarly at the upper end, the p-value is 0.018, and thus reject the upper value as an outlier. We will order the data and drop the lowest and highest values, and run the Chi-square test again

Phosphorus<-sort(Phosphorus, decreasing = FALSE)  
Phosphorus1 <- c(Phosphorus[2:42])  
  
chisq.out.test(Phosphorus1,variance = var(Phosphorus1),opposite = TRUE)

##   
## chi-squared test for outlier  
##   
## data: Phosphorus1  
## X-squared = 4.7747, p-value = 0.02888  
## alternative hypothesis: lowest value 1.46 is an outlier

chisq.out.test(Phosphorus1,variance = var(Phosphorus1),opposite = FALSE)

##   
## chi-squared test for outlier  
##   
## data: Phosphorus1  
## X-squared = 5.4398, p-value = 0.01968  
## alternative hypothesis: highest value 3.66 is an outlier

In both cases now (upper and lower limits of the data), the p-value for the chi-square outlier test is remains below 0.05, and the alternative hypothesis will again be accepted.

Phosphorus2 <- c(Phosphorus1[2:40])  
  
chisq.out.test(Phosphorus2,variance = var(Phosphorus2),opposite = TRUE)

##   
## chi-squared test for outlier  
##   
## data: Phosphorus2  
## X-squared = 4.4772, p-value = 0.03435  
## alternative hypothesis: lowest value 1.61 is an outlier

chisq.out.test(Phosphorus2,variance = var(Phosphorus2),opposite = FALSE)

##   
## chi-squared test for outlier  
##   
## data: Phosphorus2  
## X-squared = 5.142, p-value = 0.02335  
## alternative hypothesis: highest value 3.5 is an outlier

Once more our p-values from the chi-square outlier test remain below 0.05, and suggest the lowest and highest values are outliers, so the test will be repeated a thrid time with those values removed

Phosphorus3 <- c(Phosphorus2[2:38])  
  
chisq.out.test(Phosphorus3,variance = var(Phosphorus3),opposite = TRUE)

##   
## chi-squared test for outlier  
##   
## data: Phosphorus3  
## X-squared = 2.8782, p-value = 0.08979  
## alternative hypothesis: highest value 3.17 is an outlier

chisq.out.test(Phosphorus3,variance = var(Phosphorus3),opposite = FALSE)

##   
## chi-squared test for outlier  
##   
## data: Phosphorus3  
## X-squared = 4.5943, p-value = 0.03208  
## alternative hypothesis: lowest value 1.7 is an outlier

The upper limit p-value is now considerably above 0.05, so the alternative hypothesis will be rejected. The lower limit p-value however still remains below 0.05, and the outlier will be removed.

Phosphorus4 <- c(Phosphorus3[2:37])  
  
chisq.out.test(Phosphorus4,variance = var(Phosphorus4),opposite = TRUE)

##   
## chi-squared test for outlier  
##   
## data: Phosphorus4  
## X-squared = 2.9986, p-value = 0.08334  
## alternative hypothesis: highest value 3.17 is an outlier

The vector Phosphorus4 now contains the normally distributed reference values with outliers removed which will be selected to produce a reference interval for serum phosphorus in adult narwhals. The standard format for RI is as follows: n(number of reference values in the set), mean +/- standard deviation, median, range. We can write code to create a vector of these values, and a second vector to represent the column names.

RI\_Phosphorus <- c("Phosphorus", "mmol/L", mean(Phosphorus4),sd(Phosphorus4), median(Phosphorus4),range(Phosphorus4), length(Phosphorus4))  
  
RI\_Phosphorus

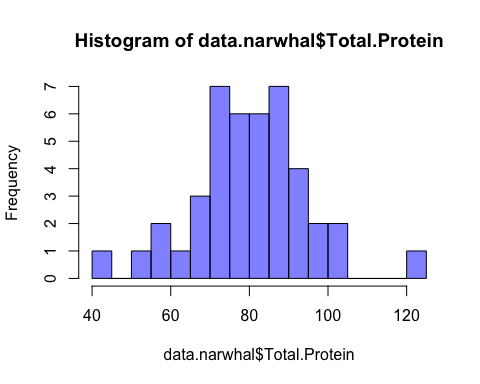
## [1] "Phosphorus" "mmol/L" "2.54333333333333"   
## [4] "0.361891854406416" "2.61" "1.83"   
## [7] "3.17" "36"

We will construct RI for a second measure, and begin to construct a table of RI for Narwhal

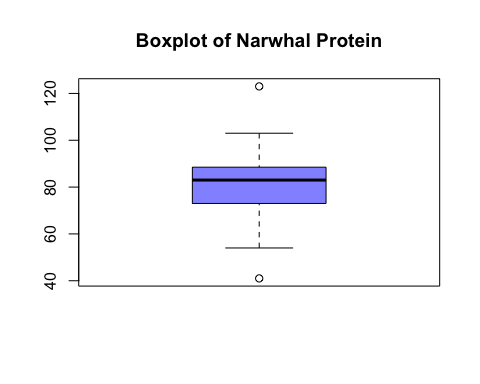
summary(data.narwhal$Total.Protein)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 41.00 73.00 83.00 80.88 88.50 123.00 1

hist(data.narwhal$Total.Protein, breaks = 20, col = rgb(0,0,1,0.5))



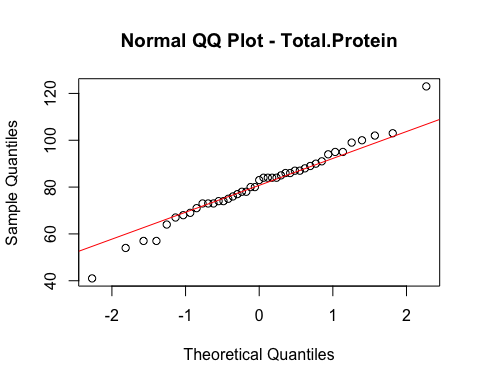
boxplot(data.narwhal$Total.Protein, col = rgb(0,0,1,0.5), main = "Boxplot of Narwhal Protein")



shapiro.test(data.narwhal$Total.Protein)

##   
## Shapiro-Wilk normality test  
##   
## data: data.narwhal$Total.Protein  
## W = 0.97952, p-value = 0.6285

qqnorm(data.narwhal$Total.Protein, main = "Normal QQ Plot - Total.Protein")  
qqline(data.narwhal$Total.Protein, col = "red")

 Similar to the results for Phosphorus, in the boxplot, there appear to be outliers on both ends, quartiles look even; the Shapiro-Wilk test with p= 0.62 (> .05) retains Null hypothesis of normally distrubuted data. The QQ plot shows relatively good alignment with the theoretical quantile line in the middle ranges of the data, with several potential outliers at the minimum and maximum value ends.

Protein <- data.narwhal$Total.Protein[!is.na(data.narwhal$Total.Protein)]  
  
#Chi-Sq Tests for Outliers for protein  
chisq.out.test(Protein,variance = var(Protein),opposite = TRUE)

##   
## chi-squared test for outlier  
##   
## data: Protein  
## X-squared = 7.2838, p-value = 0.006958  
## alternative hypothesis: lowest value 41 is an outlier

chisq.out.test(Protein,variance = var(Protein),opposite = FALSE)

##   
## chi-squared test for outlier  
##   
## data: Protein  
## X-squared = 8.122, p-value = 0.004373  
## alternative hypothesis: highest value 123 is an outlier

Outliers identified at both tails of the data

Protein<-sort(Protein, decreasing = FALSE)  
Protein1 <- c(Protein[2:42])  
  
chisq.out.test(Protein1,variance = var(Protein1),opposite = TRUE)

##   
## chi-squared test for outlier  
##   
## data: Protein1  
## X-squared = 3.3854, p-value = 0.06578  
## alternative hypothesis: highest value 103 is an outlier

chisq.out.test(Protein1,variance = var(Protein1),opposite = FALSE)

##   
## chi-squared test for outlier  
##   
## data: Protein1  
## X-squared = 4.9575, p-value = 0.02598  
## alternative hypothesis: lowest value 54 is an outlier

Outlier identified only on the low end of the data

Protein2 <- c(Protein1[2:41])  
  
chisq.out.test(Protein2,variance = var(Protein2),opposite = FALSE)

##   
## chi-squared test for outlier  
##   
## data: Protein2  
## X-squared = 4.6173, p-value = 0.03165  
## alternative hypothesis: lowest value 57 is an outlier

Outlier identified again on the low end of the data, two next values are the same, so both removed

Protein3 <- c(Protein2[3:40])  
  
chisq.out.test(Protein3,variance = var(Protein3),opposite = TRUE)

##   
## chi-squared test for outlier  
##   
## data: Protein3  
## X-squared = 3.4318, p-value = 0.06395  
## alternative hypothesis: lowest value 64 is an outlier

The vector Protein3 now contains the normally distributed reference values with outliers removed which will be selected to produce a reference interval for serum phosphorus in adult narwhals.

RI\_Protein <- c("Protein", "g/L", mean(Protein3),sd(Protein3), median(Protein3),range(Protein3), length(Protein3))  
  
RI\_Protein

## [1] "Protein" "g/L" "82.7894736842105"  
## [4] "10.1426524990017" "84" "64"   
## [7] "103" "38"

Now we can bind these two vectors into a data frame to begin to construct the table of RI for Narwhal

RI\_Table<-rbind(RI\_Phosphorus,RI\_Protein)  
  
colnames(RI\_Table)<- c("Analyte", "Units", "Mean","S.D.", "Median", "lower\_limit", "upper\_limit", "n")  
RI\_Table

## Analyte Units Mean S.D.   
## RI\_Phosphorus "Phosphorus" "mmol/L" "2.54333333333333" "0.361891854406416"  
## RI\_Protein "Protein" "g/L" "82.7894736842105" "10.1426524990017"   
## Median lower\_limit upper\_limit n   
## RI\_Phosphorus "2.61" "1.83" "3.17" "36"  
## RI\_Protein "84" "64" "103" "38"

knitr::kable(RI\_Table, digits = 2, caption = "Reference Intervals for Narwhal")

Reference Intervals for Narwhal

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Analyte | Units | Mean | S.D. | Median | lower\_limit | upper\_limit | n |
| RI\_Phosphorus | Phosphorus | mmol/L | 2.54333333333333 | 0.361891854406416 | 2.61 | 1.83 | 3.17 | 36 |
| RI\_Protein | Protein | g/L | 82.7894736842105 | 10.1426524990017 | 84 | 64 | 103 | 38 |

sessionInfo()

## R version 3.3.2 (2016-10-31)  
## Platform: x86\_64-apple-darwin13.4.0 (64-bit)  
## Running under: OS X El Capitan 10.11.6  
##   
## locale:  
## [1] en\_CA.UTF-8/en\_CA.UTF-8/en\_CA.UTF-8/C/en\_CA.UTF-8/en\_CA.UTF-8  
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] outliers\_0.14 ggplot2\_2.2.1 dplyr\_0.5.0   
##   
## loaded via a namespace (and not attached):  
## [1] Rcpp\_0.12.8 digest\_0.6.11 rprojroot\_1.1 assertthat\_0.1   
## [5] plyr\_1.8.4 grid\_3.3.2 R6\_2.2.0 gtable\_0.2.0   
## [9] DBI\_0.5-1 backports\_1.0.4 magrittr\_1.5 scales\_0.4.1   
## [13] evaluate\_0.10 highr\_0.6 stringi\_1.1.2 lazyeval\_0.2.0   
## [17] rmarkdown\_1.3 tools\_3.3.2 stringr\_1.1.0 munsell\_0.4.3   
## [21] yaml\_2.1.14 colorspace\_1.3-2 htmltools\_0.3.5 knitr\_1.15.1   
## [25] tibble\_1.2

The reference values for the remaining 20 analytes can be examined for normal distribution in the same manner. It is likely that some may need to be transformed when distribution is not normal before reference intervals can be built. Most publicaitons report RI in tabular format, the statistical information on selection of reference values from the sample, and the elimination of outliers is described in the methods. Even with sample sets of less than 50 individuals, fairly robust reference intervals can be created. This is of value particularly in wildlife work where numbers of individuals sampled may be hampered by access, remoteness, weather and other challenges. RI are important the evaluation of physiologic and pathologic conditions which affect animal health (Norman et al, 2012) and are part of a health baseline needed to allow monitoring of wildlife populations.

Literature cited:

De Monte T, Pilleri G. 1983. Hematological parameters in the Naruhal Monodon monoceros. Investigations on Cetacea 15: 143

Dietz R, Heide-Jorgensen MP, Richard PR, Acquarone M. 2001. Summer and fall movements of narwhals (Monodon monoceros) from northeastern Baffin Island towards northern Davis Strait. Arctic 54(3): 244-261.

Friedrichs KR, Harr KE, Freeman KP, Szladovits B, Walton RM, Barnhart KF, Blanco-Chavez J. 2012. ASVCP reference interval guidelines: determination of de novo reference intervals in veterinary species and other related topics. Veterinary Clin. Path. 41(4):441-453.

MacNeill AC. 1975. Blood values for some captive cetaceans.Canadian Veterinary Journal 16(7):187-193.

Norman SA, Goertz CEC, Burek KA, Quakenbush LT, Cornick LA, Romano TA, Spoon T, Miller W, Beckett LA, Hobbs RC. 2012. Seasonal Hematology and serum chemistry of wild beluga whales (Delphinapterus leucas) in Bristol Bay, Alaska, USA. Journal of Wildlife Disease 48(1): 21-32.

```