Homework #1

Madhav Viswesvaran

January 26, 2025

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
##
## -- Column specification --------------------------------------------------------
## cols(
##   season = col_character(),
##   size = col_character(),
##   speed = col_character(),
##   mxPH = col_double(),
##   mnO2 = col_double(),
##   Cl = col_double(),
##   NO3 = col_double(),
##   NH4 = col_double(),
##   oPO4 = col_double(),
##   PO4 = col_double(),
##   Chla = col_double(),
##   a1 = col_double(),
##   a2 = col_double(),
##   a3 = col_double(),
##   a4 = col_double(),
##   a5 = col_double(),
##   a6 = col_double(),
##   a7 = col_double()
## )
```

1. Descriptive Summary Statistics

a)

```r
#number of observations in each season
library(dplyr)
algae %>%
  summarize(.by = season,n = n())
```

```
## # A tibble: 4 x 2
##   season     n
##   <chr>  <int>
```

```
## 1 winter    62
## 2 spring    53
## 3 autumn    40
## 4 summer    45
```

b) There are missing values for some of the chemicals.

```r
#Calculating mean and variance for each chemical

algae %>%
  summarize(across(c(4:11), list(mean = ~ mean(.x, na.rm = TRUE),
                                 var = ~ var(.x, na.rm = TRUE)))) %>%
  pivot_longer(cols = everything(),
               names_to = c("variable", "statistic"),
               names_sep = "_",
               values_to = "value") %>%
  mutate(value = format(value, scientific = FALSE, digits = 6))
```

```
## # A tibble: 16 x 3
##    variable statistic value
##    <chr>    <chr>     <chr>
##  1 mxPH     mean      "      8.011734"
##  2 mxPH     var       "      0.357969"
##  3 mnO2     mean      "      9.117778"
##  4 mnO2     var       "      5.718089"
##  5 Cl       mean      "     43.636279"
##  6 Cl       var       "   2193.171725"
##  7 NO3      mean      "      3.282389"
##  8 NO3      var       "     14.261756"
##  9 NH4      mean      "    501.295828"
## 10 NH4      var       "3851584.684865"
## 11 oPO4     mean      "     73.590596"
## 12 oPO4     var       "   8305.849930"
## 13 PO4      mean      "    137.882101"
## 14 PO4      var       "  16639.384545"
## 15 Chla     mean      "     13.971197"
## 16 Chla     var       "    420.082735"
```

```r
#
```

It looks like NH4 has extremely large mean and variance relative to the other chemicals.

```r
#Calculating median and mad for each chemical
algae %>%
  summarize(across(c(4:11), list(median = ~ median(.x, na.rm = TRUE),
                                 mad = ~ mad(.x, na.rm = TRUE)))) %>%
  pivot_longer(cols = everything(),
               names_to = c("variable", "statistic"),
               names_sep = "_",
               values_to = "value") %>%
  mutate(value = format(value, scientific = FALSE, digits = 6))
```

```
## # A tibble: 16 x 3
##    variable statistic value
##    <chr>    <chr>     <chr>
##  1 mxPH     median    "   8.060000"
##  2 mxPH     mad       "   0.504084"
```

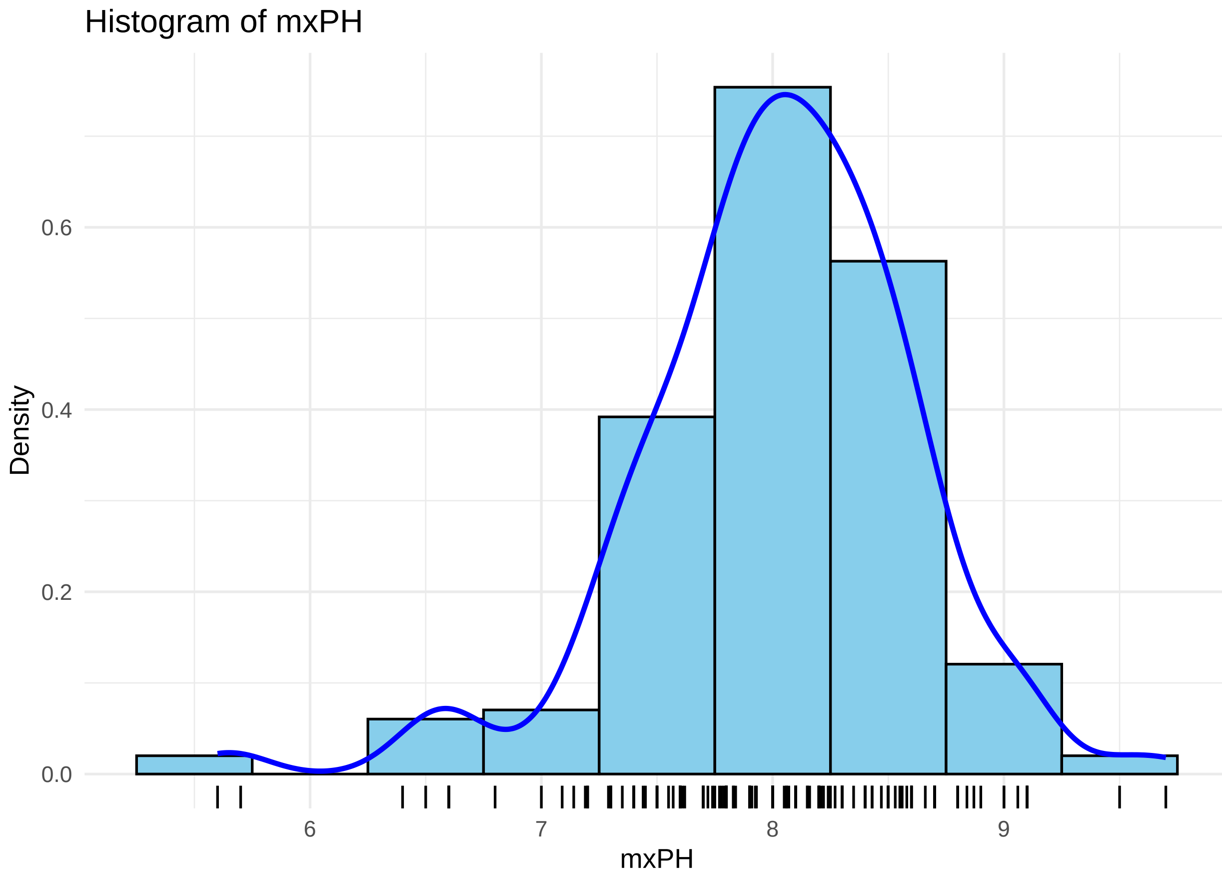
```
##   3 mnO2     median    "  9.800000"
##   4 mnO2     mad       "  2.053401"
##   5 Cl       median    " 32.730000"
##   6 Cl       mad       " 33.249529"
##   7 NO3      median    "  2.675000"
##   8 NO3      mad       "  2.172009"
##   9 NH4      median    "103.166500"
##  10 NH4      mad       "111.617548"
##  11 oPO4     median    " 40.150000"
##  12 oPO4     mad       " 44.045822"
##  13 PO4      median    "103.285500"
##  14 PO4      mad       "122.321172"
##  15 Chla     median    "  5.475000"
##  16 Chla     mad       "  6.671700"
```

The median and MAD tend to be pretty similar for the chemicals expect for mxPH and mnO2, with NH4 still having the biggest MAD and median

2 - Data visualization

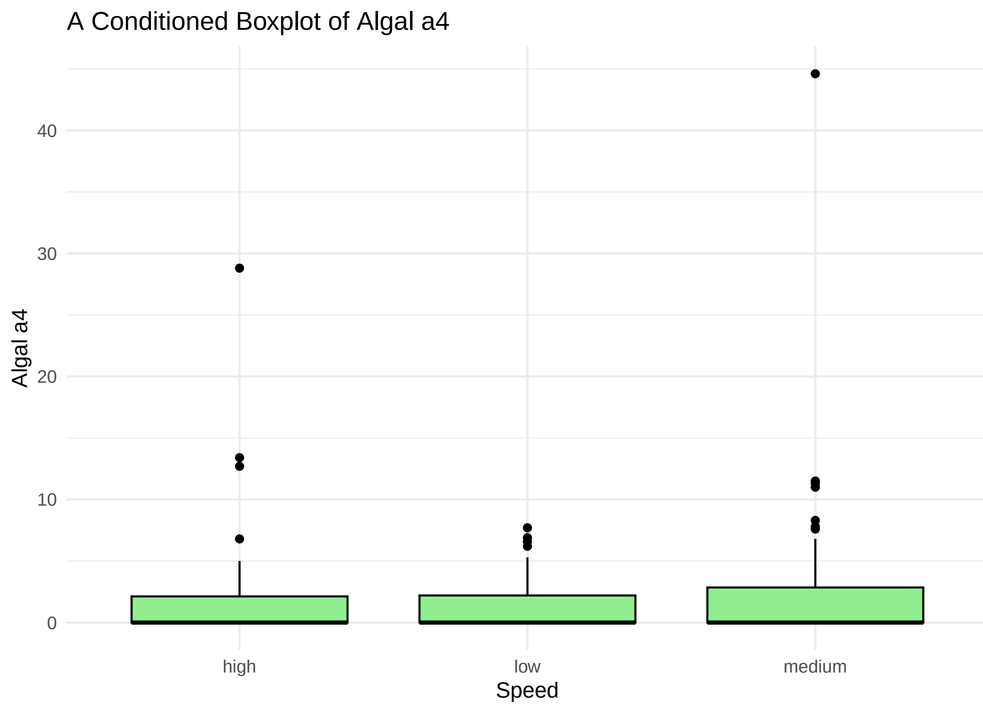
a)

```r
#creating a histogram
ggplot(algae, aes(x = mxPH)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.5, fill = "skyblue", color = "black") +
  geom_density(color = "blue", linewidth = 1) +
  geom_rug() +
  ggtitle("Histogram of mxPH") +
  xlab("mxPH") +
  ylab("Density") +
  theme_minimal()
```



The distribution slightly skews left.

```
#boxplot
ggplot(algae, aes(x = speed, y = a4)) +
  geom_boxplot(fill = "lightgreen", color = "black") +
  ggtitle("A Conditioned Boxplot of Algal a4") +
  xlab("Speed") +
  ylab("Algal a4") +
  theme_minimal()
```



It appears that the high and medium speeds have some outlier for algal a4. ## 3 - Missing Values a)

```r
#table with na values for each column
num_rows_with_na <- sum(apply(is.na(algae), 1, any))
num_rows_with_na
```

```
## [1] 16
```

```r
algae %>%
  summarize(across(everything(), ~ sum(is.na(.)), .names = "count_{col}")) %>%
  pivot_longer(cols = everything(), names_to = "column", values_to = "na_count")
```

```
## # A tibble: 18 x 2
##    column       na_count
##    <chr>           <int>
##  1 count_season        0
##  2 count_size          0
##  3 count_speed         0
##  4 count_mxPH          1
##  5 count_mnO2          2
##  6 count_Cl           10
##  7 count_NO3           2
##  8 count_NH4           2
##  9 count_oPO4          2
## 10 count_PO4           2
## 11 count_Chla         12
## 12 count_a1            0
## 13 count_a2            0
## 14 count_a3            0
## 15 count_a4            0
## 16 count_a5            0
## 17 count_a6            0
## 18 count_a7            0
```

b) 16 observations contain missing values, and the table shows the number of missing values by variable.

```r
algae.del <- algae[complete.cases(algae), ]
#View(algae.del)
```

algae.del has 184 observations.

4 - Bias Variance Tradeoff

a) The terms that represent reducible error are $\mathrm{Var}(\hat{f}(x_0))$ and $[\mathrm{Bias}(\hat{f}(x_0))]^2$ The term that represents irreducible error is $\mathrm{Var}(e)$
b) In the bias-variance tradeoff we know that the variance and bias are non-negative terms because they are squared, therefore even if the bias and variance are 0, the expected test error is still at least equal to the irreducible error, but in most cases it will be equal to the irreducible error plus some bias and variance since they are nonnegative.