Homework Assignment 1

TJ Sipin and Preeti Kulkarni

January 23, 2022

Background

High concentrations of certain harmful algae in rivers constitute a serious ecological problem with a strong impact not only on river lifeforms, but also on water quality. Being able to monitor and perform an early forecast of algae blooms is essential to improving the quality of rivers.

With the goal of addressing this prediction problem, several water samples were collected in different times duringa period of approximately 1 year. For each water sample, different chemical properties were measured as well as the frequency of occurence of seven harmful agle. Some other characteristics of the water collection process were also stored, such as the season of the year, the river size, and the river speed.

Goal

We want to understand how these frequencies are related to certain chemical attributes of water samples as well as other characteristics of the samples (like season of the year, type of river, etc.)

Data Description

The data set consists of data for 200 water samples and each observation in the available datasets is in effect an aggregation of several water samples collected from the same river over a period of 3 months, during the same season of the year. Each observation contains information on 11 variables. Three of these variables are nominal and describe the season of the year when the water samples to be aggregated were collected, as well as the size and speed of the river in question. The eight remaining variables are values of different chemical parameteres measured in the water samples forming the aggregation, namely: Maximum pH value, Minimum value of O_2 (oxygen), Mean value of Cl (chloride), Mean value of NO_3^- (nitrates), Mean value of NH_4^+ (ammonium), Mean of PO_4^3 (orthophosphate), Mean of total PO_4 (phosphate) and Mean of chlorophyll.

Associated with each of these parameters are seven frequency numbers of different harmful algae found in the respective water samples. No information is given regarding the names of the algae that were identified.

1. ***Descriptive summary statistics***

a. We count the number of observations in each size.

size	**count**
<chr>	<int>

large	45
medium	84
small	71

3 rows

b. Check to see if there are missing values. Then calculate the mean and variance of each chemical. What can we notice about the magnitude of the two quantities for different chemicals?

```
##     season              size              speed               mxPH
##  Length:200         Length:200         Length:200         Min.   :5.60
##  Class :character   Class :character   Class :character   1st Qu.:7.70
##  Mode  :character   Mode  :character   Mode  :character   Median :8.06
##                                                           Mean   :8.01
##                                                           3rd Qu.:8.40
##                                                           Max.   :9.70
##                                                           NA's   :1
##       mnO2              Cl              NO3              NH4
##  Min.   : 1.50   Min.   :  0.2   Min.   : 0.05   Min.   :    5
##  1st Qu.: 7.72   1st Qu.: 11.0   1st Qu.: 1.30   1st Qu.:   38
##  Median : 9.80   Median : 32.7   Median : 2.67   Median :  103
##  Mean   : 9.12   Mean   : 43.6   Mean   : 3.28   Mean   :  501
##  3rd Qu.:10.80   3rd Qu.: 57.8   3rd Qu.: 4.45   3rd Qu.:  227
##  Max.   :13.40   Max.   :391.5   Max.   :45.65   Max.   :24064
##  NA's   :2       NA's   :10      NA's   :2       NA's   :2
##       oPO4             PO4             Chla              a1
##  Min.   :  1.0   Min.   :  1.0   Min.   :  0.20   Min.   : 0.00
##  1st Qu.: 15.7   1st Qu.: 41.4   1st Qu.:  2.00   1st Qu.: 1.50
##  Median : 40.1   Median :103.3   Median :  5.47   Median : 6.95
##  Mean   : 73.6   Mean   :137.9   Mean   : 13.97   Mean   :16.92
##  3rd Qu.: 99.3   3rd Qu.:213.8   3rd Qu.: 18.31   3rd Qu.:24.80
##  Max.   :564.6   Max.   :771.6   Max.   :110.46   Max.   :89.80
##  NA's   :2       NA's   :2       NA's   :12
##        a2              a3              a4              a5
##  Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.00
##  Median : 3.00   Median : 1.55   Median : 0.00   Median : 1.90
##  Mean   : 7.46   Mean   : 4.31   Mean   : 1.99   Mean   : 5.06
##  3rd Qu.:11.38   3rd Qu.: 4.92   3rd Qu.: 2.40   3rd Qu.: 7.50
##  Max.   :72.60   Max.   :42.80   Max.   :44.60   Max.   :44.40
##
##        a6              a7
##  Min.   : 0.00   Min.   : 0.0
##  1st Qu.: 0.00   1st Qu.: 0.0
##  Median : 0.00   Median : 1.0
##  Mean   : 5.96   Mean   : 2.5
```

```
##  3rd Qu.: 6.92   3rd Qu.: 2.4
##  Max.   :77.60   Max.   :31.6
##
## [1] "Mean:"
##    mnO2      Cl     NO3     NH4    oPO4     PO4    Chla
##   9.118  43.636   3.282 501.296  73.591 137.882  13.971
## [1] "Variance:"
##        mnO2          Cl          NO3          NH4         oPO4          PO4
##       5.832    2210.222      14.986 4106410.930    8561.134   16685.144
##        Chla
##     409.307
```

There are missing values. The variances of the min value of O_2 and the mean of value of NO_3^- are all quite low, with the remaining chemicals having incredibly high variances, specifically NH_4 and PO_4. On the other hand, the means of the above chemicals are also low, with the addition of the mean of chlorophyll.

I'm interested in the relationship between the mean and variance of each, since there seems to be a correlation between the magnitudes of the mean and variance of each chemical.

```
## [1] "Variance / Mean:"
##      mnO2        Cl       NO3       NH4      oPO4       PO4      Chla
##    0.6397   50.6510    4.5654 8191.5921  116.3346  121.0102   29.2965
```

c. Mean and Variance is one measure of central tendency and spread of data. Median and Median Absolute Deviation are alternative measures of central tendency and spread.

For a univariate data set $X_1, X_2, \ldots, X_n$, MAD is defined as the median of the absolute deviations from the data's median:

$$\text{MAD} = \text{median}(|X_i - \text{median}(X)|)$$

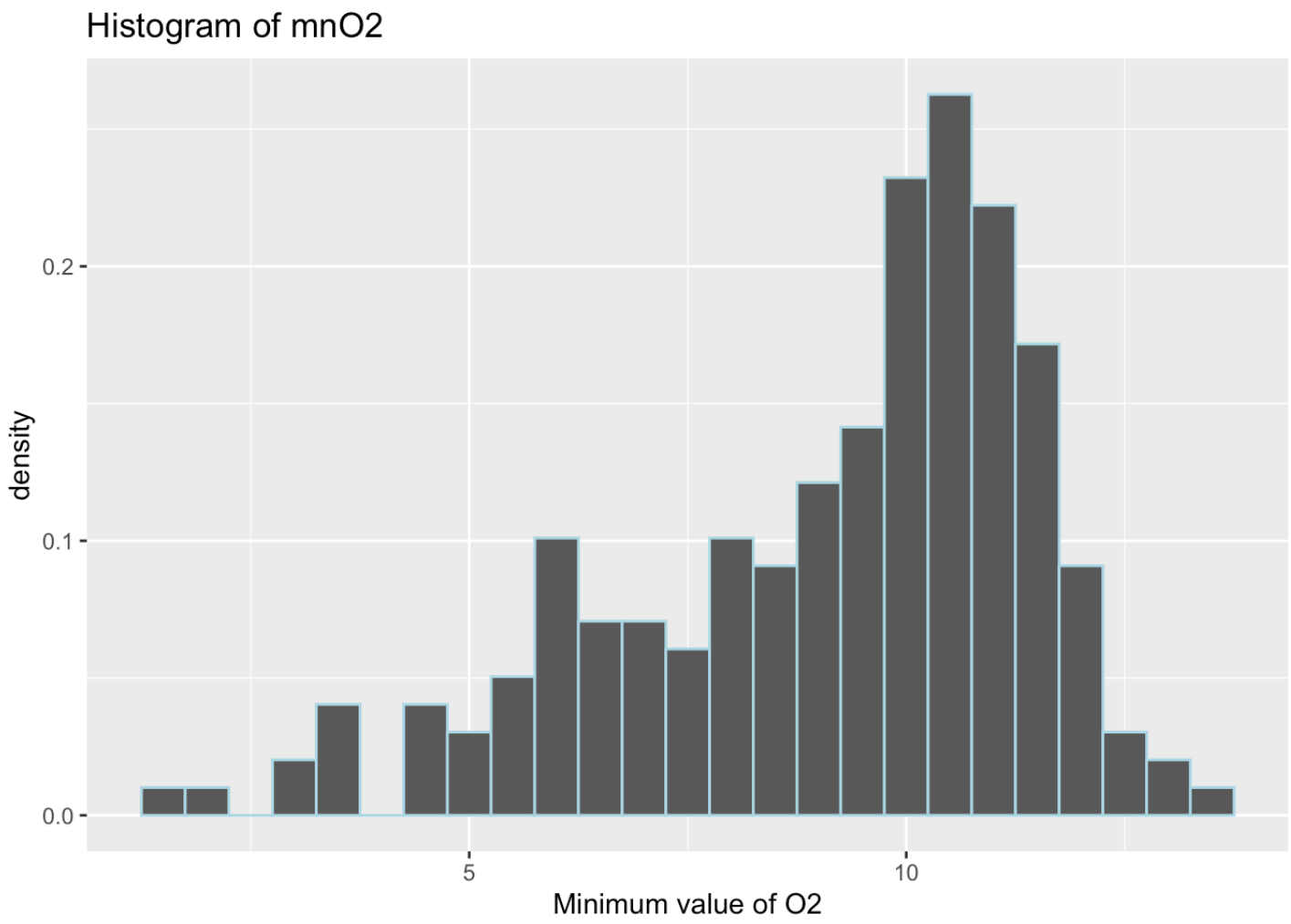
We want to compute the median and MAD of each chemical and compare the two sets of quantities.

```
## [1] "Median:"
##    mnO2      Cl     NO3     NH4    oPO4     PO4    Chla
##   9.800  32.730   2.675 103.166  40.150 103.285   5.475
## [1] "MAD:"
##    mnO2      Cl     NO3     NH4    oPO4     PO4    Chla
##   2.053  33.250   2.172 111.618  44.046 122.321   6.672
```

For each chemical, the mean and median are similar only when the variance is relatively low. For example, the mean and median of the mininum value of O_2 is similar with a variance of 5.832. On the other hand, the mean of NH_4^+ is 501.296, the median is 103.166 (a difference of about 400), with a variance of $4,106,410.930$. The variance and MAD differ similarly to the mean and median in regards to the fact that the MAD is lower than the variance. In fact, the MAD of each chemical are much lower than the variance. Of course, they all scale similarly, as in those with a small variance also has a small MAD.

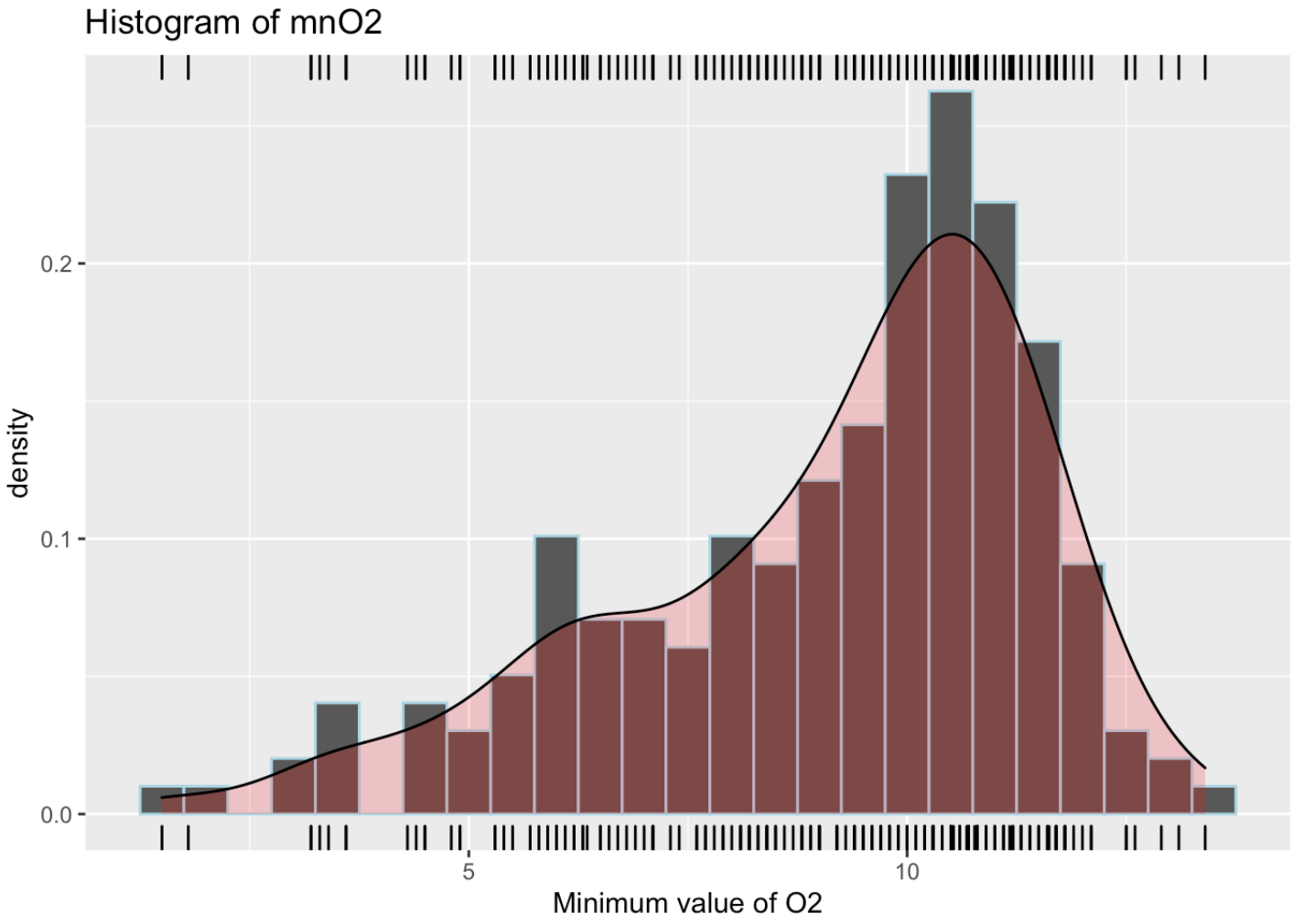
2. ***Data visualization*** Ploting a histogram, scatter plot, boxplot, Q-Q plot

a. We produce a histogram of mnO_2. Is the distribution skewed?

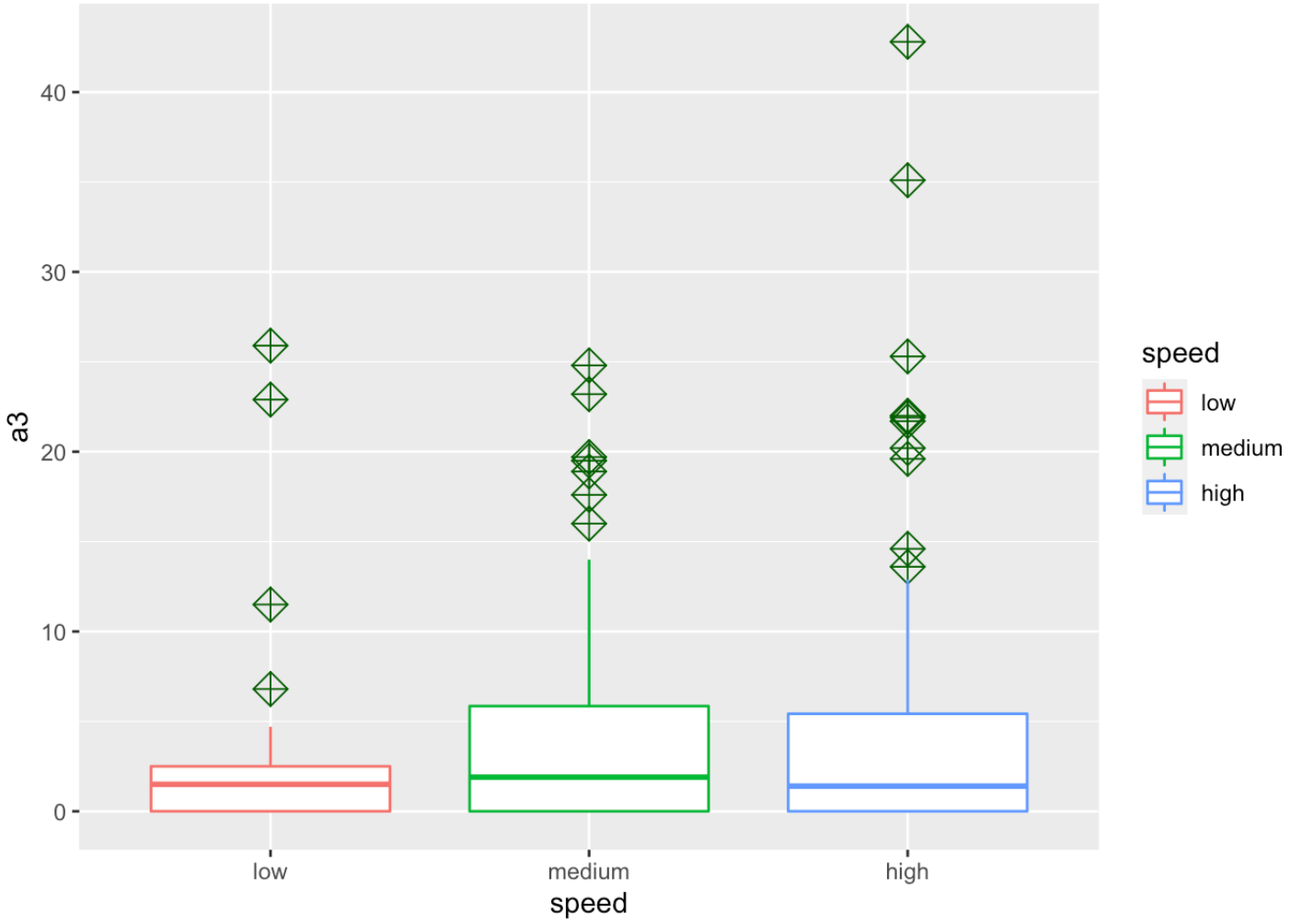


We see that the distribution has a negative skew since it has a long tail in the negative direction, and is indicative of a median larger than the mean.

b. Adding a density curve and rug plots to the histogram



c. Creating a boxplot for a_3 grouped by $speed$.



It seems that the higher the speed of the river, the higher the algal frequencies for a_3 , though it seems like the range of the medium speed is higher than the range of the high speed. However, we see that the outliers of the high speed are in higher frequencies of a_3 than those of the medium speed. Additionally, the distribution of low has a negative skew, while the distributions of medium and high have positive skews.

3. ***Dealing with missing values***

a. Count how many observations contain missing values and how many missing values in each variable.

```
nrow(na.omit(algae)) # There are 200 - 184 = 16 observations with missing values
sapply(algae, function(x) sum(is.na(x))) # Shows the amount of NA values for each var
iable
```

```
## [1] 184
## season    size   speed    mxPH    mnO2      Cl     NO3     NH4    oPO4     PO4    Chla
##      0       0       0       1       2      10       2       2       2       2      12
##     a1      a2      a3      a4      a5      a6      a7
##      0       0       0       0       0       0       0
```

There are 184 rows without NA values so that means there are 16 observations with missing values. Above shows the amount of missing values for each variable.

b. **Removing observations with missing values**

```
algae.del = algae %>%
  filter(across(everything(), ~!is.na(.x)))
length(complete.cases(algae.del))
```

```
## [1] 184
```

As expected, there are 184 observations in algae.del.

4. The bias-variance tradeoff:

$$\mathbb{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \text{var}(\hat{f}(\mathbf{x}_0)) + \left[\text{Bias}(\hat{f}(\mathbf{x}_0))\right]^2 + \text{var}(\epsilon)$$

a. The terms $\text{var}(\hat{f}(\mathbf{x}_0))$ and $\left[\text{Bias}(\hat{f}(\mathbf{x}_0))\right]^2$ represent the reducible error since $\hat{f}(\cdot)$ is obtained on a training set. The term $\text{var}(\epsilon)$ is irreducible since it is simply random noise.

b. Since $\text{var}(\hat{f}(\mathbf{x}_0))$ and $\left[\text{Bias}(\hat{f}(\mathbf{x}_0))\right]^2$ are reducible, then in theory they can be reduced to 0. Thus, the lower bound of $\mathbb{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right]$ is $0 + 0 + \text{var}(\epsilon) = \text{var}(\epsilon)$.