Homework Assignment 1 PSTAT 131

Quinlan Wilson and Jack Guo (both 131)

October 22, 2025

```
## Warning: package 'knitr' was built under R version 4.4.3

## Warning: package 'dplyr' was built under R version 4.4.3

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

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

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

## Warning: package 'ggplot2' was built under R version 4.4.3
```

```r
algae <- read.table("algaeBloom.txt", col.names=
                      c('season','size','speed','mxPH','mnO2','Cl','NO3','NH4',
                        'oPO4','PO4','Chla','a1','a2','a3','a4','a5','a6','a7'),
                    na = "XXXXXXX")
glimpse(algae)
```

```
## Rows: 200
## Columns: 18
## $ season <chr> "winter", "spring", "autumn", "spring", "autumn", "winter", "su~
## $ size   <chr> "small", "small", "small", "small", "small", "small", "small", ~
## $ speed  <chr> "medium", "medium", "medium", "medium", "medium", "high", "high~
## $ mxPH   <dbl> 8.00, 8.35, 8.10, 8.07, 8.06, 8.25, 8.15, 8.05, 8.70, 7.93, 7.7~
## $ mnO2   <dbl> 9.8, 8.0, 11.4, 4.8, 9.0, 13.1, 10.3, 10.6, 3.4, 9.9, 10.2, 11.~
## $ Cl     <dbl> 60.80, 57.75, 40.02, 77.36, 55.35, 65.75, 73.25, 59.07, 21.95, ~
## $ NO3    <dbl> 6.238, 1.288, 5.330, 2.302, 10.416, 9.248, 1.535, 4.990, 0.886,~
## $ NH4    <dbl> 578.00, 370.00, 346.67, 98.18, 233.70, 430.00, 110.00, 205.67, ~
## $ oPO4   <dbl> 105.00, 428.75, 125.67, 61.18, 58.22, 18.25, 61.25, 44.67, 36.3~
## $ PO4    <dbl> 170.00, 558.75, 187.06, 138.70, 97.58, 56.67, 111.75, 77.43, 71~
## $ Chla   <dbl> 50.000, 1.300, 15.600, 1.400, 10.500, 28.400, 3.200, 6.900, 5.5~
## $ a1     <dbl> 0.0, 1.4, 3.3, 3.1, 9.2, 15.1, 2.4, 18.2, 25.4, 17.0, 16.6, 32.~
## $ a2     <dbl> 0.0, 7.6, 53.6, 41.0, 2.9, 14.6, 1.2, 1.6, 5.4, 0.0, 0.0, 0.0, ~
## $ a3     <dbl> 0.0, 4.8, 1.9, 18.9, 7.5, 1.4, 3.2, 0.0, 2.5, 0.0, 0.0, 0.0, 2.~
## $ a4     <dbl> 0.0, 1.9, 0.0, 0.0, 0.0, 0.0, 3.9, 0.0, 0.0, 2.9, 0.0, 0.0, 0.0~
## $ a5     <dbl> 34.2, 6.7, 0.0, 1.4, 7.5, 22.5, 5.8, 5.5, 0.0, 0.0, 1.2, 0.0, 1~
## $ a6     <dbl> 8.3, 0.0, 0.0, 0.0, 4.1, 12.6, 6.8, 8.7, 0.0, 0.0, 0.0, 0.0, 0.~
## $ a7     <dbl> 0.0, 2.1, 9.7, 1.4, 1.0, 2.9, 0.0, 0.0, 0.0, 1.7, 6.0, 1.5, 2.1~
```

1. Descriptive summary statistics

(a)

```
algae %>%
  group_by(season) %>%
  summarize(n = n())
```

```
## # A tibble: 4 x 2
##   season     n
##   <chr>  <int>
## 1 autumn    40
## 2 spring    53
## 3 summer    45
## 4 winter    62
```

(b)

```
c(Missing_Vals = sum(is.na(algae)))
```

```
## Missing_Vals
##           33
```

```
chemicals <- c("mxPH", "mnO2", "Cl", "NO3", "NH4", "oPO4", "PO4", "Chla")

mean_and_var <- function(x) {
  mean_x <- mean(x, na.rm = T)
  var_x <- var(x, na.rm = T)

  return(c(Mean = mean_x, Variance = var_x))
}

sapply(algae[, chemicals], mean_and_var)
```

```
##           mxPH  mnO2      Cl    NO3       NH4    oPO4     PO4   Chla
## Mean     8.012 9.118   43.64  3.282     501.3   73.59   137.9  13.97
## Variance 0.358 5.718 2193.17 14.262 3851584.7 8305.85 16639.4 420.08
```

The means and the variances differ significantly between chemicals. Where mxPH and mnO2 have small variances and NH4, PO4, and oPO4 have massive ones.

(c)

```
median_and_MAD <- function(x) {
  median_x <- median(x, na.rm = T)
  MAD_x <- median(abs(x - median_x), na.rm = T)

  return(c(median = median_x, MAD = MAD_x))
}

sapply(algae[, chemicals], median_and_MAD)
```

```
##        mxPH  mnO2    Cl   NO3    NH4  oPO4   PO4  Chla
## median 8.06 9.800 32.73 2.675 103.17 40.15 103.3 5.475
## MAD    0.34 1.385 22.43 1.465  75.29 29.71  82.5 4.500
```

Comparing the mean and the median we can see that the means are typically higher than the medians. Additionally they medians and MAD's for the chemicals appear to be more calm suggesting that there are outliers in the observations. The only chemicals where this isn't the case are from mxPH and mnO2 where their means and medians are close along with not having extreme variances.

2. Data visualization

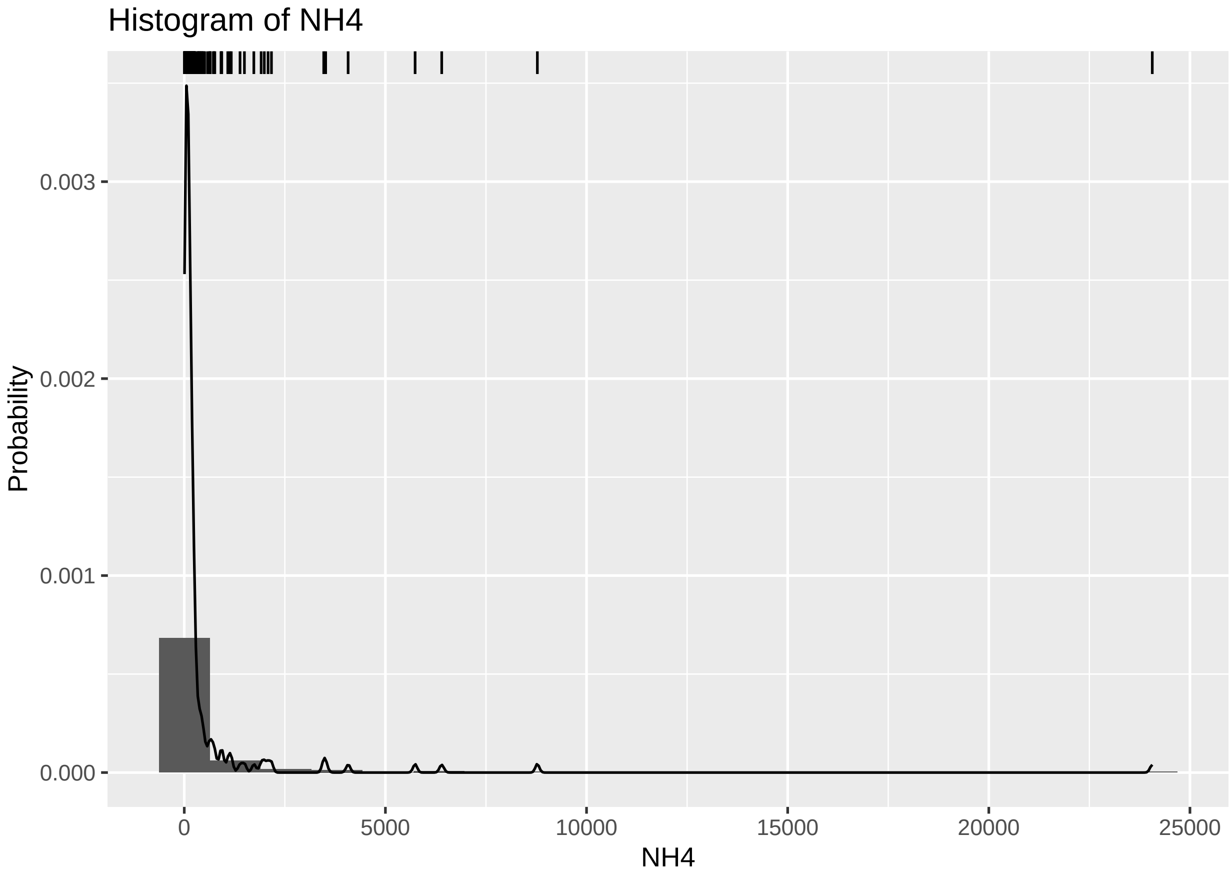
(a)

The distribution is heavily skewed right for NH4

(b)

```r
algaeBloom<-algae
algaeBloom <- na.omit(algaeBloom)
algaeBloom$NH4 <- as.numeric(algaeBloom$NH4)

ggplot(algaeBloom, aes(x = NH4)) +
  geom_histogram(aes(y = ..density..), bins = 20) +
  labs(
    x="NH4",
    y="Probability",
    title="Histogram of NH4"
  ) +
  geom_density() +
  geom_rug(sides='t')
```

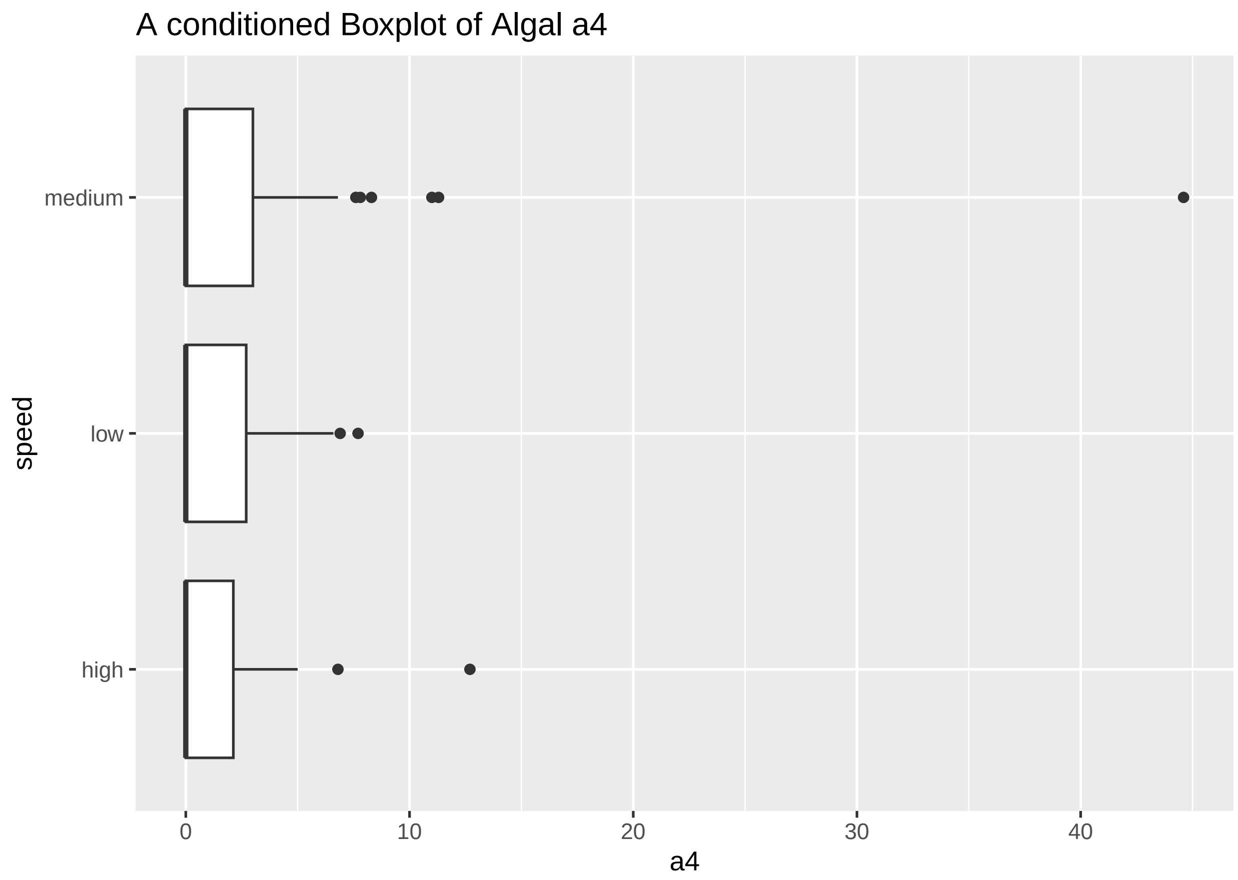


(c)

The box plots for all three speeds are skewed right.

```
a4_speed <- algaeBloom %>%
  select(a4, speed) %>%
  group_by(speed)

ggplot(a4_speed, aes(x=a4, y=speed)) +
  geom_boxplot() +
  labs(
    title="A conditioned Boxplot of Algal a4"
  )
```



3. Dealing with missing values

(a)

```
sum(!complete.cases(algae))
```

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

```
sapply(algae, function(x) sum(length(which(is.na(x)))))
```

```
## 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
```

(b)

```
algae.del <- algae %>%
  filter(complete.cases(.))

nrow(algae.del)
```

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

4.

(a)

The reducible error terms are

$$Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2$$

and the irreducible error term is

$$Var(\epsilon)$$

(b)

$$\begin{aligned}
&= E[(f(x_0) + \epsilon - \hat{f}(x_0))^2] \\
&= E[(f(x_0) - \hat{f}(x_0))^2] + 2E[\epsilon(f(x_0) - \hat{f}(x_0))] + E[\epsilon^2] \\
&= E[(f(x_0) - \hat{f}(x_0))^2] + E[\epsilon^2] \\
&= Var(\hat{f}(x_0)) + [Bias(\hat{(}f)(x_0))]^2 + Var(\epsilon)
\end{aligned}$$

Since $Var(\hat{f}(x_0)) + [Bias(\hat{(}f)(x_0))]^2 \geq 0$ that means $E[(f(x_0) - \hat{f}(x_0))^2] \geq Var(\epsilon)$.