# 数据挖掘——海藻数据分析

王恒怿　2120151038　2016.05

## 0.实验环境和使用软件

本次统计实验在 Windows 系统下，使用统计分析 R 软件进行。

## 1.读取实验数据 Analysis.txt

在 R 软件控制台编写如下指令

```
>algae<-read.table('L:/Analysis.txt',header=F,dec='.',col.names=c('season','size','speed','mxPH','mnO2','Cl','NO3','NH4','oPO4','PO4','Chla','a1','a2','a3','a4','a5','a6','a7'),na.string=c('XXXXXXX'))
```

## 2.数据摘要

数据读取完成后，在 R 软件控制台编写如下指令

```
>summary(algae)
```

得到如下的描述性统计，其中，NA's 表示缺失值的个数。

```
   season       size       speed        mxPH            mnO2             Cl              NO3              NH4               oPO4        
autumn:40   large :45   high  :84   Min.   :5.600   Min.   : 1.500   Min.   :  0.222   Min.   : 0.050   Min.   :    5.00   Min.   :  1.00  
spring:53   medium:84   low   :33   1st Qu.:7.700   1st Qu.: 7.725   1st Qu.: 10.981   1st Qu.: 1.296   1st Qu.:   38.33   1st Qu.: 15.70  
summer:45   small :71   medium:83   Median :8.060   Median : 9.800   Median : 32.730   Median : 2.675   Median :  103.17   Median : 40.15  
winter:62                           Mean   :8.012   Mean   : 9.118   Mean   : 43.636   Mean   : 3.282   Mean   :  501.30   Mean   : 73.59  
                                    3rd Qu.:8.400   3rd Qu.:10.800   3rd Qu.: 57.824   3rd Qu.: 4.446   3rd Qu.:  226.95   3rd Qu.: 99.33  
                                    Max.   :9.700   Max.   :13.400   Max.   :391.500   Max.   :45.650   Max.   :24064.00   Max.   :564.60  
                                    NA's   :1       NA's   :2        NA's   :10        NA's   :2        NA's   :2          NA's   :2       
      PO4              Chla               a1               a2               a3               a4               a5               a6               a7        
 Min.   :  1.00   Min.   :  0.200   Min.   : 0.00   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000  
 1st Qu.: 41.38   1st Qu.:  2.000   1st Qu.: 1.50   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000  
 Median :103.29   Median :  5.475   Median : 6.95   Median : 3.000   Median : 1.550   Median : 0.000   Median : 1.900   Median : 0.000   Median : 1.000  
 Mean   :137.88   Mean   : 13.971   Mean   :16.92   Mean   : 7.458   Mean   : 4.309   Mean   : 1.992   Mean   : 5.064   Mean   : 5.964   Mean   : 2.495  
 3rd Qu.:213.75   3rd Qu.: 18.308   3rd Qu.:24.80   3rd Qu.:11.375   3rd Qu.: 4.925   3rd Qu.: 2.400   3rd Qu.: 7.500   3rd Qu.: 6.925   3rd Qu.: 2.400  
 Max.   :771.60   Max.   :110.456   Max.   :89.80   Max.   :72.600   Max.   :42.800   Max.   :44.600   Max.   :44.400   Max.   :77.600   Max.   :31.600  
 NA's   :2        NA's   :12                                                                                                                        
```

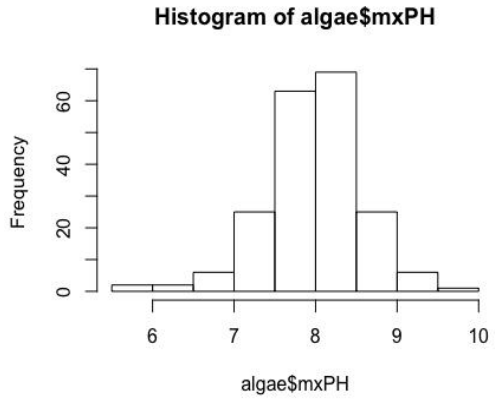
# 4.数据可视化

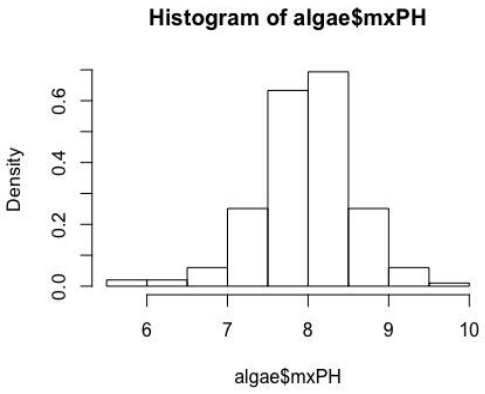
## 4.1 直方图

以 mxPH 为例，在 R 软件控制台编写如下指令

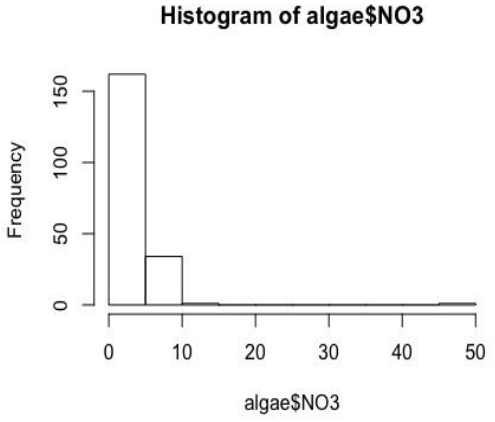
```
> hist(algae$mxPH,prob=T)
> hist(algae$mxPH)
```

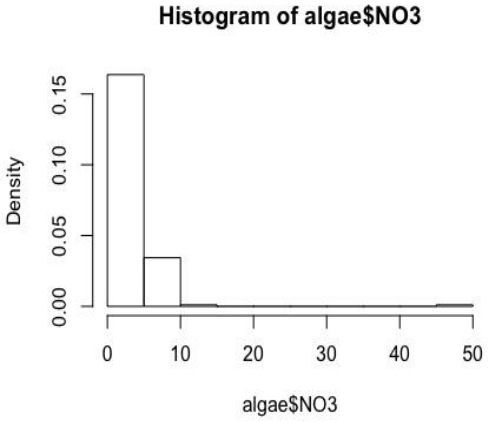
得到 mxPH 的直方图如下：




其他变量的直方图如下，因变量较多仅以举例的形式进行展示：变量 NO3



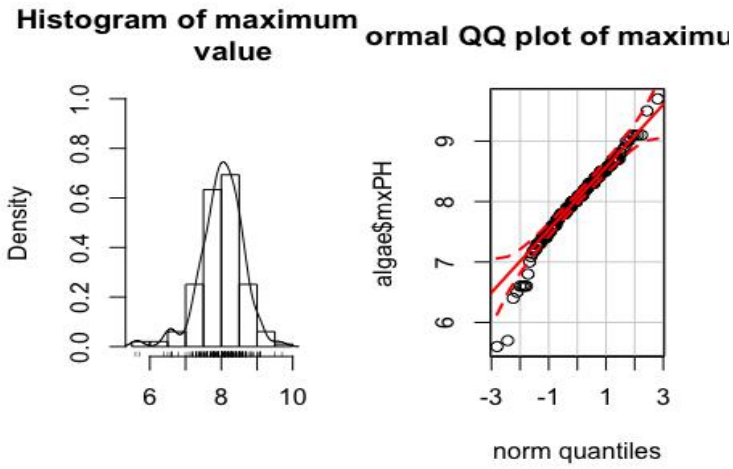




## 4.2 用 QQ 图检验正态分布

以 mxPH 为例，在 R 软件控制台编写如下指令

```
> install.package(pkgs='car')
> library(car)
> par(mfrow=c(1,2))
> hist(algae$mxPH,prob=T,xlab='',main='Histogram of maximum pH
   value',ylim=0:1)
> lines(density(algae$mxPH,na.rm=T))
> rug(jitter(algae$mxPH))
> qqPlot(algae$mxPH,main='Normal QQ plot of maximum pH')
> par(mfrow=c(1,1))
```

从而得到




从图中我们可以看出，变量 mxPH 基本符合正态分布

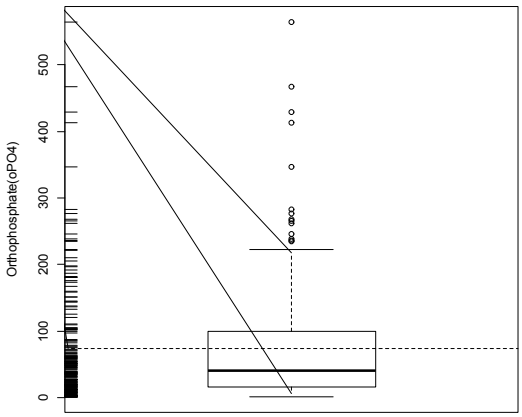
在其他变量中，我们进行同样的数据分析，mnO2 近似满足正态分布，其余的变量均不满足正态分布

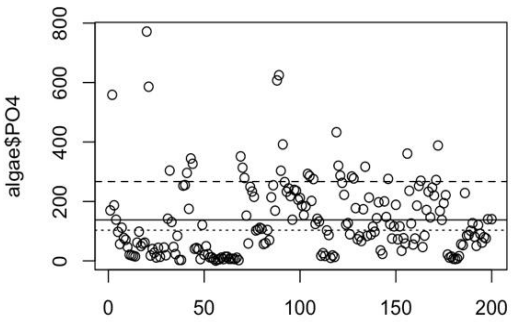
## 4.3 盒图与离群值识别

以 oPO4 为例，在 R 软件控制台编写如下指令

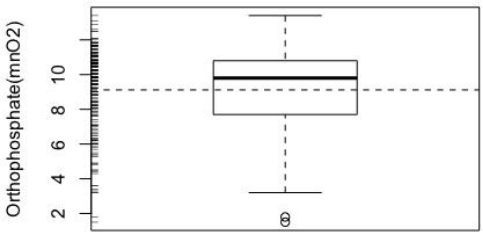
```
> boxplot(algae$oPO4,ylab="Orthophosphate(oPO4)")
> rug(jitter(algae$oPO4),side=2)
> abline(h=mean(algae$oPO4,na.rm=T),lty=2)
> plot(algae$oPO4,xlab="")
> abline(h=mean(algae$oPO4,na.rm=T),lty=1)
> abline(h=mean(algae$oPO4,na.rm=T)+sd(algae$oPO4,na.rm=T),lty=2)
> abline(h=median(algae$oPO4,na.rm=T),lty=3)
> identify(algae$oPO4)
> algae[algae$oPO4>19000,]
> algae[!is.na(algae$oPO4)&algae$oPO4>19000,]
```

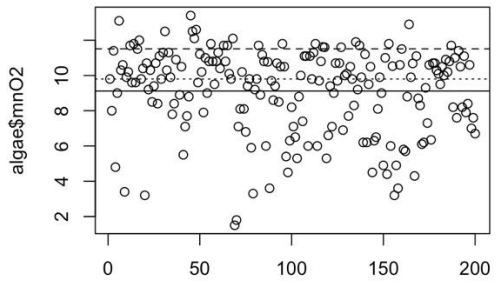
得到 oPO4 的盒图盒离群标示如下：




其他变量的盒图如下，因变量较多仅以举例的形式进行展示。变量 mnO2：



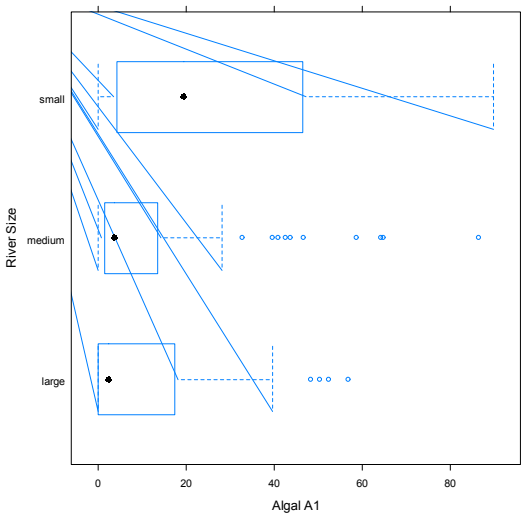




## 4.4 条件盒图

对于海藻 a1 和 River Size，在 R 软件控制台编写如下指令

```
> library(lattice)
> bwplot(size~a1,data=algae,ylab='River Size',xlab='Algal A1')
```

得到条件盒图如下：




# 5.数据缺失的处理

## 5.1 将缺失部分剔除

在 R 软件控制台编写如下指令

```
> install.package(pkgs='DMwR')
> library(DMwR)
> algae[!complete.cases(algae),]
> nrow(algae[!complete.cases(algae),])
> algae<-na.omit(algae)
> apply(algae,1,function(x) sum(is.na(x)))
> manyNAs(algae,0.2)
> algae<-algae[-manyNAs(algae),]
```

利用此方法，从原始数据中剔除了 16 条记录。

## 5.2 用最高频率值来填补缺失值

对于接近正态的分布来说，均值是最佳选择；对偏态分布或有离群值的分布而言，中位数通常是更好的代表数据中心趋势的指标；对于名义变量，通常采用众数。用以下函数完成填补所有缺失值：

在 R 软件控制台编写如下指令

```
> library(DMwR)
> data(algae)
> algae<-algae[-manyNAs(algae),]
> algae<-centralImputation(algae)
> summary(algae)
```

## 5.3 通过属性的相关关系来填补缺失值

通过计算相关性，NH4 和 NO3，PO4 和 oPO4 之间的相关性较大。在 R 软件控制台编写如下指令，主要针对 PO4 和 OPO4 进行数据处理

```
> library(DMwR)
> data(algae)
> synum(cor(algae[,4:18],use="complete.obs"))
> algae<-algae[-manyNAs(algae),]
> lm(formula=PO4~oPO4,data=algae)
> algae[28,"PO4"]<-42.897+1.293*algae[28,"oPO4"]
> algae[28,]
```

可以得到线性模型：PO4=42.897+1.293*oPO4，由于样本 62 和样本 199 含有过多的缺失数据已经剔除，所以仅样本 28 在 PO4 上有缺失值，可以用上面的线性关系来填补：

algae[28,"PO4"]<-42.897+1.293*algae[28,"oPO4"]

并用指令 algae[28,]：查看填补的记录。

## 5.4 通过数据对象之间的相似性来填补缺失值

采用欧式距离度量相似性。实验利用欧式距离来寻找与任何含有缺失值案例最相似的 10 个水样，并用它们填补缺失值。

在 R 软件控制台编写如下指令

```
> library(DMwR)
> data(algae)
> algae<-algae[-manyNAs(algae),]
> clean.algae<-knnImputation(algae,k=0,meth="median")
> summary(algae)
```

## 6.实验总结

本实验通过使用 R 软件对海藻数据进行了处理和分析。通过实验，主要了解学习了如下知识点：

1.熟悉了了实验环境，和 R 软件的基本操作
2.学习掌握了使用 R 软件读取、展示数据以及对数据的各种制图
3.学习掌握了利用软件进行基础的数据分析，如相关性计算等
4.学习掌握了利用软件对数据缺失项进行处理的四种方法
5.通过以上实验，了解了数据预处理和数据分析对数据挖掘的重要性