Contents

学生成绩分析实例

读入学生成绩

```
#setwd("E:\\DataAnalysis")
scores <- read.table("scores.txt", header=TRUE, row.names="num")
head(scores)
```

```
##     class chn math   eng  phy  chem politics  bio history  geo pe
## 158     3  99  120 114.0 70.0 49.50     50.0 49.0    48.5 49.5 60
## 442     7 107  120 118.5 68.6 43.00     49.0 48.5    48.5 49.0 56
## 249     4  98  120 116.0 70.0 47.50     47.0 49.0    47.5 49.0 60
## 573     9 102  113 111.5 70.0 47.00     49.0 49.0    49.0 49.5 60
## 310     5 103  120 111.5 70.0 44.75     46.5 48.0    48.0 48.0 60
## 613    10  98  120 113.0 70.0 46.75     47.5 47.5    47.0 48.5 60
```

```
str(scores)
```

```
## 'data.frame':    599 obs. of  11 variables:
##  $ class   : int  3 7 4 9 5 10 8 2 5 9 ...
##  $ chn     : num  99 107 98 102 103 ...
##  $ math    : int  120 120 120 113 120 120 120 117 120 118 ...
##  $ eng     : num  114 118 116 112 112 ...
##  $ phy     : num  70 68.6 70 70 70 70 68.6 70 64.4 66.5 ...
##  $ chem    : num  49.5 43 47.5 47 44.8 ...
##  $ politics: num  50 49 47 49 46.5 47.5 46.5 50 49 47.5 ...
##  $ bio     : num  49 48.5 49 49 48 47.5 47.5 48 48.5 48.5 ...
##  $ history : num  48.5 48.5 47.5 49 48 47 47.5 48 47.5 46.5 ...
##  $ geo     : num  49.5 49 49 49.5 48 48.5 48 48.5 49 49.5 ...
##  $ pe      : int  60 56 60 60 60 60 60 56 56 52 ...
```

```
names(scores)
```

```
##  [1] "class"    "chn"      "math"     "eng"      "phy"      "chem"
##  [7] "politics" "bio"      "history"  "geo"      "pe"
```

```
attach(scores)
```

给出数据的概略信息

```
summary(scores)
```

```
##      class             chn              math             eng
##  Min.   : 1.000   Min.   : 26.50   Min.   :  3.00   Min.   : 15.00
##  1st Qu.: 3.000   1st Qu.: 78.25   1st Qu.: 84.00   1st Qu.: 74.00
##  Median : 6.000   Median : 84.00   Median :100.00   Median : 93.00
##  Mean   : 5.519   Mean   : 83.24   Mean   : 93.98   Mean   : 85.55
##  3rd Qu.: 8.000   3rd Qu.: 89.00   3rd Qu.:111.00   3rd Qu.:103.50
##  Max.   :10.000   Max.   :107.00   Max.   :120.00   Max.   :118.50
##       phy             chem          politics          bio
##  Min.   : 7.00   Min.   : 8.00   Min.   :15.0   Min.   :14.00
##  1st Qu.:49.00   1st Qu.:27.75   1st Qu.:39.5   1st Qu.:40.00
##  Median :58.80   Median :37.00   Median :43.5   Median :44.00
##  Mean   :54.15   Mean   :34.63   Mean   :42.0   Mean   :42.17
##  3rd Qu.:64.40   3rd Qu.:42.50   3rd Qu.:45.5   3rd Qu.:46.00
##  Max.   :70.00   Max.   :49.75   Max.   :50.0   Max.   :50.00
##     history           geo              pe
##  Min.   : 8.00   Min.   :10.50   Min.   :48.00
##  1st Qu.:32.50   1st Qu.:43.00   1st Qu.:52.00
##  Median :39.00   Median :45.50   Median :56.00
##  Mean   :36.82   Mean   :43.92   Mean   :53.86
##  3rd Qu.:43.00   3rd Qu.:47.00   3rd Qu.:56.00
##  Max.   :49.00   Max.   :50.00   Max.   :60.00
```

```
summary(scores$math)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00   84.00  100.00   93.98  111.00  120.00
```

选择某行，求一个学生总分

```
child <- scores['239',]
sum(child)
```

```
## [1] 647.45
```

```
scores.class4 <- scores[class==4,]     # 挑出 4 班的
```

求每个班的平均数学成绩

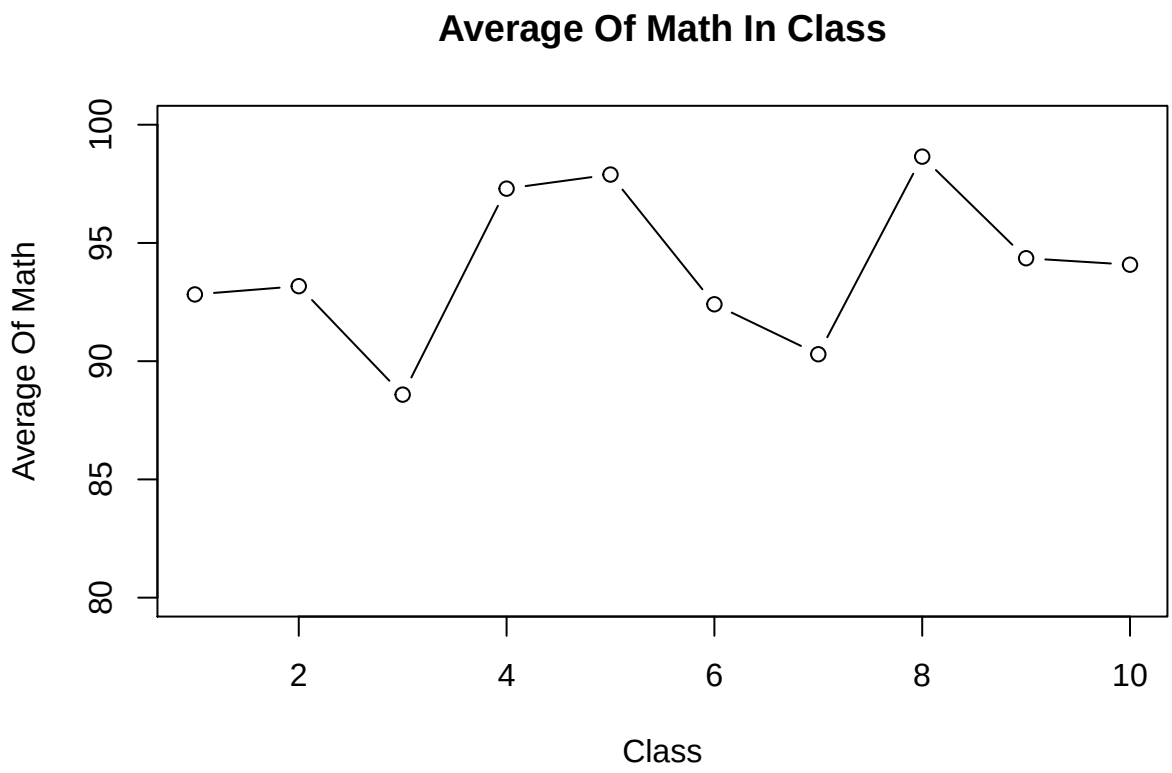
```
aver <- tapply(math, class, mean)
aver
```

```
##        1        2        3        4        5        6        7        8
## 92.82258 93.17241 88.58621 97.29688 97.89286 92.40678 90.29310 98.65000
```

```
##         9        10
## 94.35484 94.08065
```

画条曲线看看每个班的数学平均成绩

```
plot(aver, type='b', ylim=c(80,100), main="Average Of Math In Class", xlab="Class", ylab=
```



生成数据的二维列联表

```
table(math,class)
```

```
##       class
## math   1 2 3 4 5 6 7 8 9 10
##    3   0 0 0 0 0 0 1 0 0  0
##    9   1 0 0 0 0 0 0 0 0  0
##    10  1 0 1 0 0 0 0 0 0  0
##    18  0 0 0 1 0 1 0 0 1  0
##    19  0 0 0 0 0 0 1 0 0  0
##    20  0 0 1 0 0 0 0 1 0  0
##    22  0 0 1 0 0 0 0 0 0  0
##    24  0 0 0 0 0 0 0 0 1  0
##    26  0 0 0 1 0 0 0 0 0  0
##    29  0 0 0 0 0 0 0 1 0  0
##    32  0 0 1 0 0 0 0 0 0  0
```

```
##    34  0 0 0 0 0 1 0 0 0  0
##    35  0 0 0 1 1 0 0 0 0  1
##    36  0 0 1 0 0 1 1 0 0  0
##    38  0 0 0 0 0 0 1 0 0  0
##    40  1 0 0 0 0 0 0 0 1  0
##    41  0 0 0 1 0 0 0 0 0  1
##    42  0 0 0 0 1 0 0 0 0  1
##    43  0 0 1 0 0 0 0 0 0  0
##    44  0 0 0 0 0 0 0 0 0  1
##    45  0 0 0 0 0 0 1 0 0  1
##    46  0 1 0 0 0 0 0 0 0  0
##    47  0 1 1 0 1 1 2 0 0  0
##    49  0 0 0 0 0 0 0 0 0  1
##    51  0 1 0 0 0 0 0 0 0  0
##    52  0 0 0 1 0 0 0 0 0  0
##    53  0 0 1 0 0 0 0 0 0  0
##    56  0 1 0 0 0 0 0 0 0  1
##    57  0 1 1 0 0 0 0 0 1  0
##    58  2 1 0 0 0 0 0 0 0  0
##    59  1 0 0 1 1 2 0 0 0  1
##    60  0 1 2 0 0 0 1 0 0  0
##    61  0 1 0 0 0 0 0 0 0  0
##    62  1 0 0 0 0 0 0 0 0  0
##    63  0 1 1 0 0 0 0 0 0  1
##    64  0 0 0 0 0 0 0 0 1  0
##    65  1 0 0 0 0 0 0 0 1  1
##    66  0 0 0 0 0 1 0 2 0  0
##    67  0 0 0 0 0 0 2 0 2  0
##    68  0 0 0 0 0 1 0 0 1  0
##    69  2 0 0 0 0 1 2 0 1  0
##    70  0 0 0 0 2 0 1 0 1  1
##    71  0 0 1 0 0 0 0 1 0  0
##    72  0 1 0 0 0 0 1 2 1  1
##    73  2 1 0 0 0 1 0 0 0  0
##    74  0 0 0 1 1 0 1 0 0  1
##    75  0 1 1 0 0 2 0 0 1  1
##    76  0 0 1 0 0 1 0 0 1  0
##    77  0 0 1 0 0 0 1 0 0  0
##    78  0 0 0 0 2 1 1 1 0  0
##    79  1 0 0 0 0 1 1 0 1  0
##    80  0 1 1 2 0 0 0 0 0  1
##    81  0 0 0 0 0 1 1 0 1  0
##    82  0 2 0 1 0 1 0 0 1  1
##    83  2 0 0 1 0 1 0 0 1  1
##    84  0 1 1 0 0 1 0 1 1  0
##    85  0 1 1 0 0 0 0 1 0  2
##    86  0 0 1 1 0 1 0 0 1  0
##    87  0 0 1 1 1 2 0 1 0  0
##    88  0 1 0 0 0 0 1 0 1  3
##    89  2 2 0 2 0 0 0 3 1  0
##    90  3 0 1 1 1 0 3 1 1  1
##    91  1 0 0 0 0 3 0 1 1  0
##    92  2 0 0 1 2 0 0 0 1  0
##    93  0 3 1 0 3 1 1 2 0  0
```

```
##    94   1 0 0 1 3 1 1 0 2   0
##    95   3 0 1 3 0 2 1 3 1   2
##    96   0 1 2 3 0 0 2 3 1   1
##    97   2 2 2 0 2 1 2 1 0   2
##    98   3 2 2 1 1 3 1 2 0   0
##    99   2 2 1 1 1 0 0 2 0   0
##    100 1 4 1 1 2 2 3 1 2   0
##    101 2 1 1 0 1 1 0 2 1   1
##    102 0 1 4 3 0 0 0 1 2   2
##    103 0 0 0 3 0 2 1 1 0   0
##    104 2 3 0 2 1 0 0 0 0   3
##    105 3 1 2 1 2 0 0 2 0   2
##    106 1 4 0 1 3 1 2 1 1   2
##    107 3 0 2 1 0 2 0 1 0   1
##    108 0 0 3 1 5 0 0 0 3   1
##    109 1 1 3 1 5 1 2 1 1   0
##    110 1 1 0 2 1 0 2 1 0   1
##    111 3 2 0 5 2 2 3 0 4   2
##    112 1 1 4 3 5 0 3 0 2   1
##    113 2 2 0 2 0 1 0 1 3   3
##    114 1 0 0 1 1 2 2 2 2   2
##    115 2 2 1 4 0 3 0 2 1   2
##    116 0 0 1 0 1 1 2 3 3   2
##    117 1 3 2 1 0 1 1 3 0   4
##    118 3 0 1 2 2 4 2 3 3   0
##    119 0 1 0 2 0 0 0 2 2   1
##    120 1 1 2 2 2 3 4 4 3   4
```

求 4 班每一科的平均成绩

```
subjects <- c('chn','math','eng','phy','chem','politics','bio','history','geo','pe')
sapply(scores[class==4, subjects], mean)
```

```
##       chn      math       eng       phy      chem politics       bio   history
## 83.10938 97.29688 85.60156 54.30469 34.67969 42.41406 41.79688 36.77344
##       geo        pe
## 44.24219 54.31250
```

求各班各科的平均成绩

```
aggregate(scores[subjects], by=list(class), mean)
```

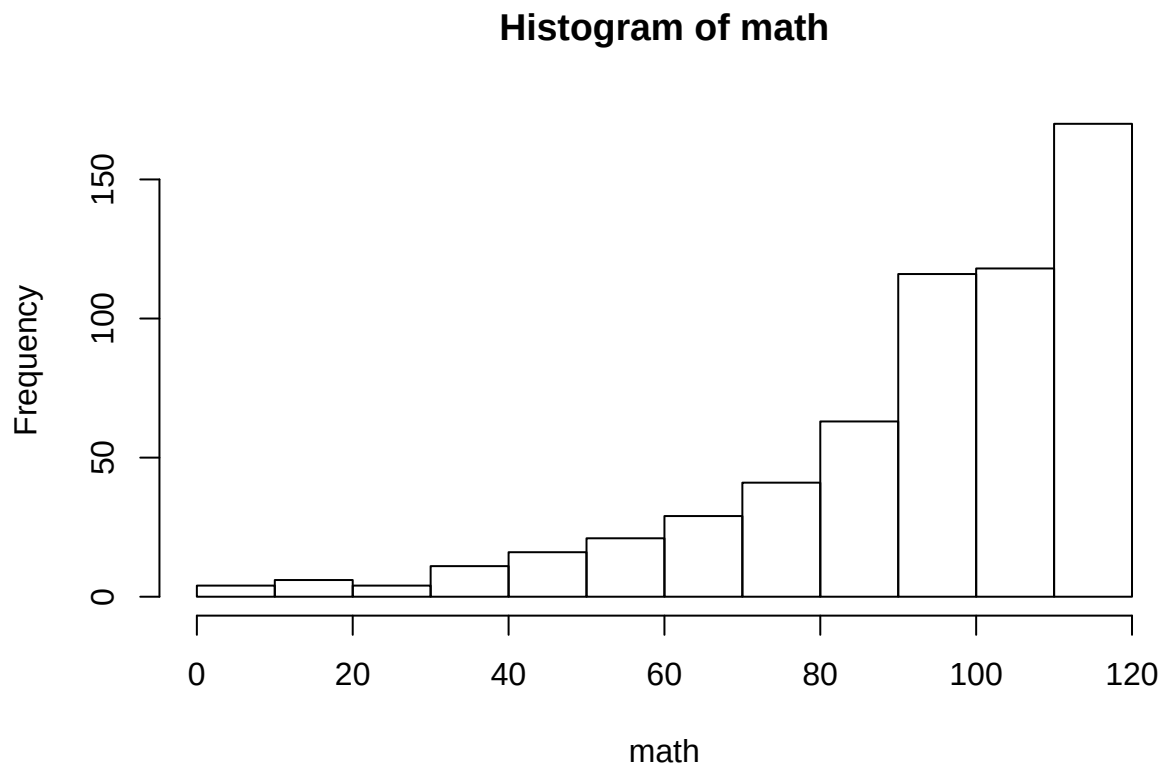
```
##    Group.1      chn     math      eng      phy     chem politics      bio
## 1        1 82.98387 92.82258 92.45161 56.04516 34.95161 42.57258 42.29839
## 2        2 81.57759 93.17241 85.01724 54.39483 34.60776 43.13793 42.05172
## 3        3 82.62069 88.58621 82.46552 51.59483 32.33190 41.99138 41.59483
## 4        4 83.10938 97.29688 85.60156 54.30469 34.67969 42.41406 41.79688
## 5        5 84.74107 97.89286 83.66964 56.10000 33.91518 42.05357 42.57143
## 6        6 83.14407 92.40678 78.57627 51.74068 33.36864 40.64407 41.55932
## 7        7 83.01724 90.29310 87.00862 51.75172 33.98276 41.63793 42.51724
## 8        8 83.65833 98.65000 86.91667 56.02333 36.07917 41.70000 42.40833
## 9        9 83.20968 94.35484 86.48387 54.29516 36.11694 41.94355 42.72581
```

```
## 10      10 84.33871 94.08065 86.66774 55.08548 36.01210 41.86290 42.22581
##     history      geo       pe
## 1  37.03226 43.44355 54.12903
## 2  38.59483 43.60345 54.68966
## 3  35.49138 42.97414 54.55172
## 4  36.77344 44.24219 54.31250
## 5  37.77679 43.96429 54.00000
## 6  34.46610 43.37288 53.22034
## 7  37.46552 44.22414 53.72414
## 8  37.84167 44.81667 52.93333
## 9  36.07258 44.30645 53.48387
## 10 36.78226 44.14516 53.61290
```

```
## aggregate
```

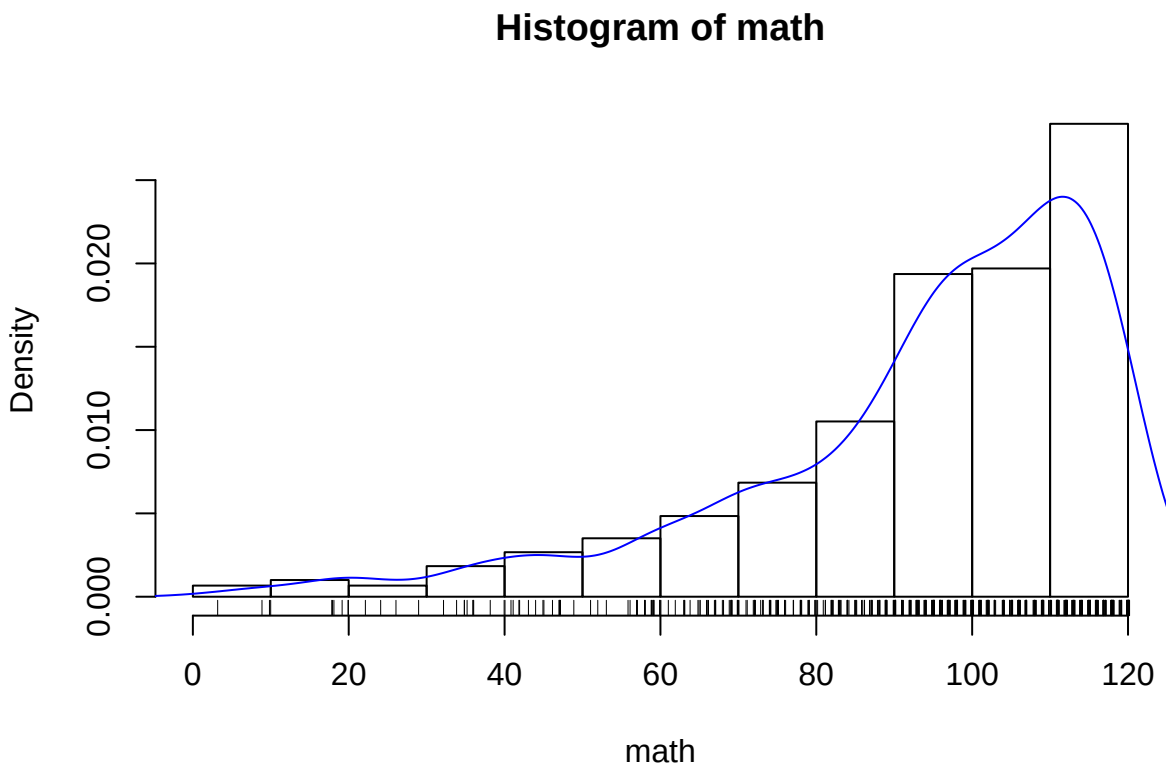
看看数学成绩的分布图

```
hist(math)
```



默认是按频数形成的直方图，设置 freq 参数可以画密度分布图。

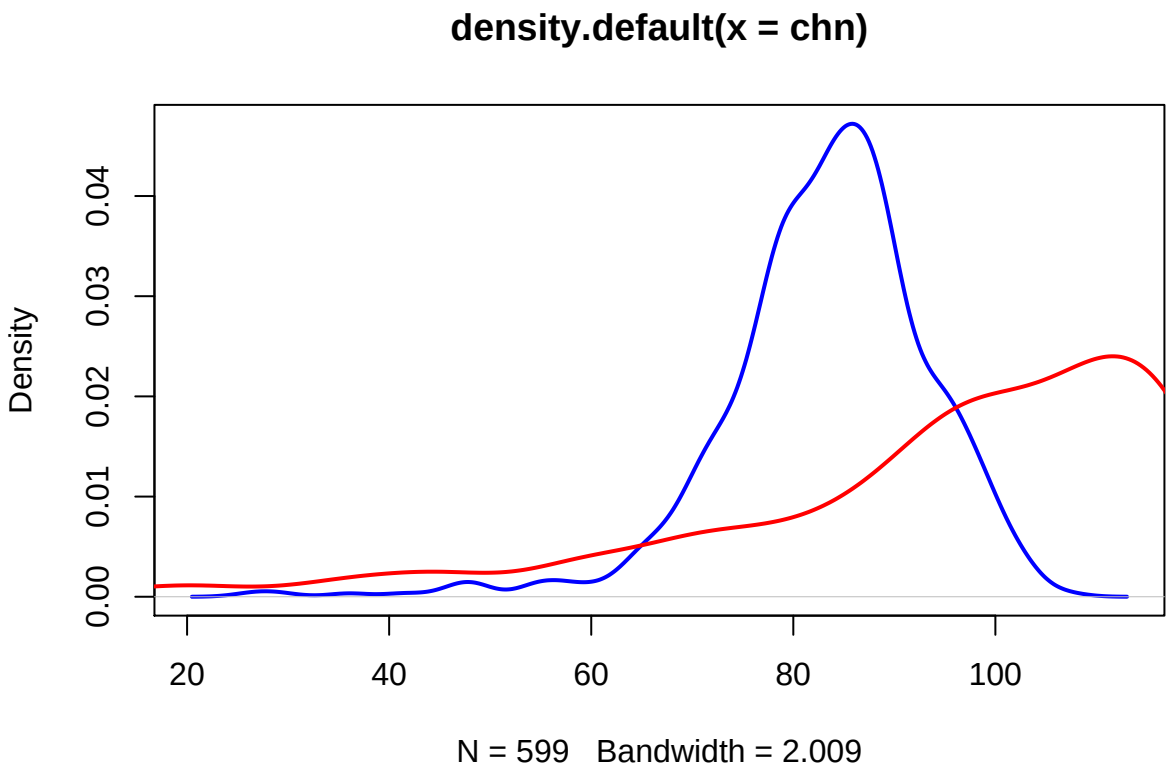
```
hist(math, freq=FALSE)
lines(density(math), col='blue')
rug(jitter(math))
```




```
# 轴须图，在轴旁边出现一些小线段，jitter 是加噪函数
```

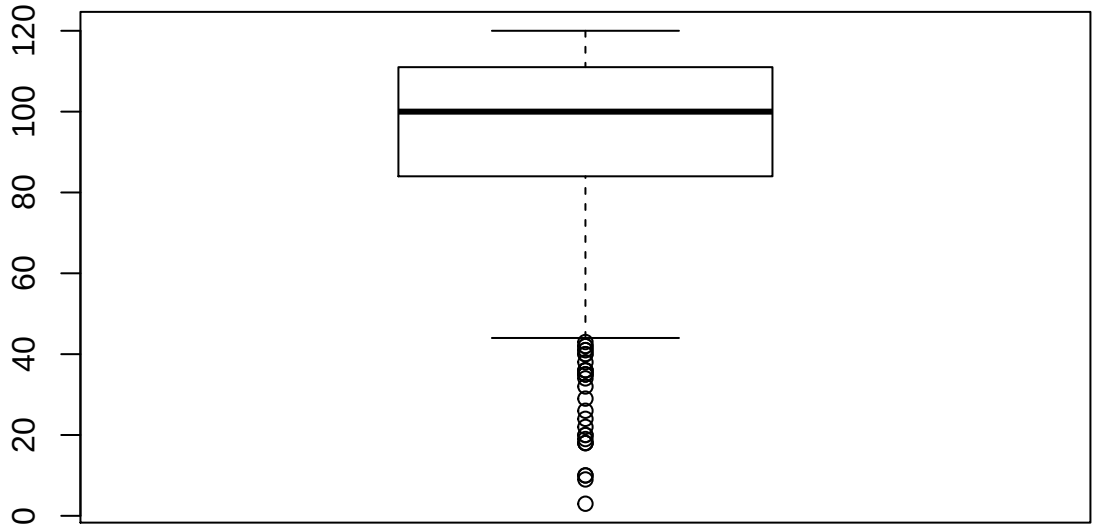
核密度图

```
plot(density(chn), col='blue', lwd=2)
lines(density(math), col='red', lwd=2)
```



箱线图

```
boxplot(math)
```



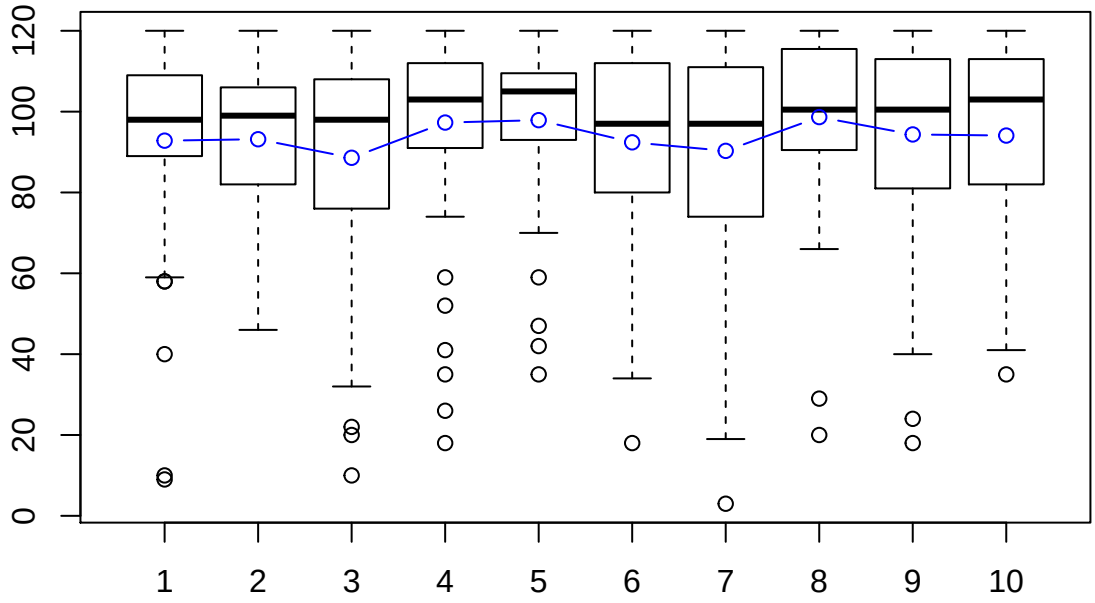

```
boxplot.stats(math)
```

```
## $stats
## [1]  44  84 100 111 120
##
## $n
## [1] 599
##
## $conf
## [1]  98.25696 101.74304
##
## $out
##  [1] 38 42 35 40 43 36 41 40 36 18 26 36 42 32 41 29 18 24 10 20 34 19 10
## [24]  3 35 20 35 18 22  9
```

```
# 这个函数可以看到画出箱线图的具体的数据值
```

并列箱线图，看各班的数据分布情况

```
boxplot(math ~ class, data=scores)
#Add Average
lines(tapply(math,class,mean), col='blue', type='b')
```



可以看出 2 班没有拖后腿的，4 班有 6 个拖后腿的

看看各科成绩的相关性

```
cor(scores[,subjects])
```

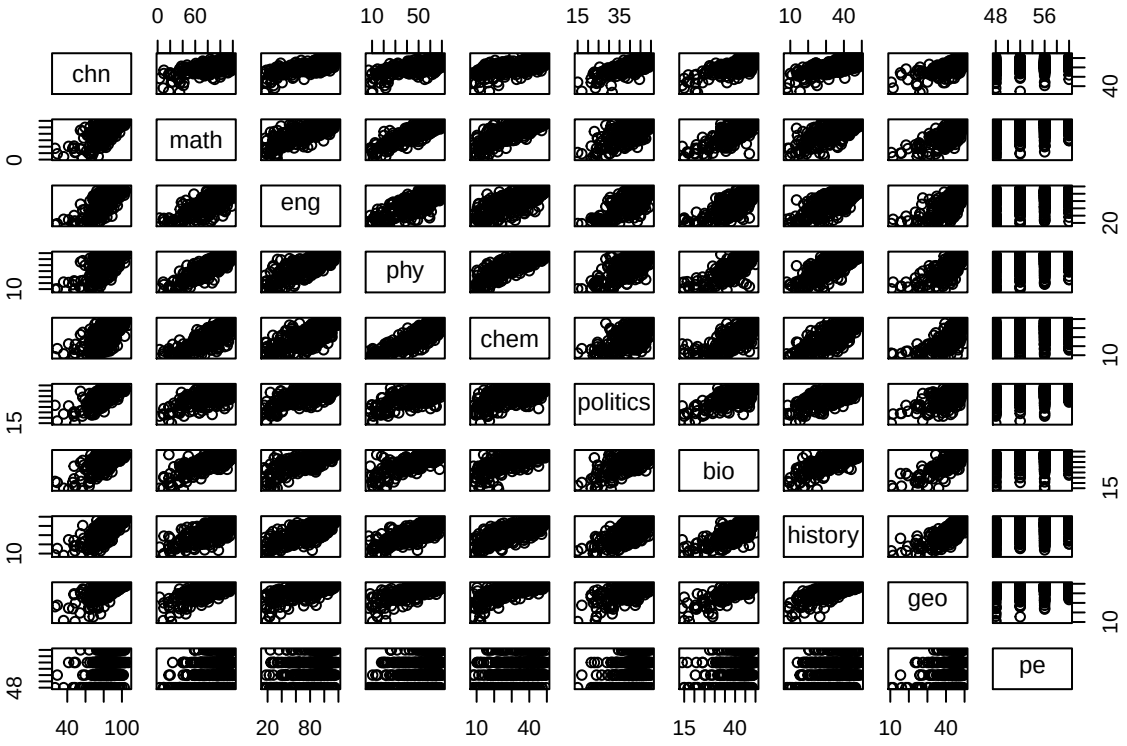
```
##                chn      math       eng       phy      chem  politics
## chn      1.0000000 0.6588126 0.7326778 0.6578172 0.6271155 0.7257003
## math     0.6588126 1.0000000 0.8079255 0.8860467 0.8304643 0.7090681
## eng      0.7326778 0.8079255 1.0000000 0.8170998 0.7868710 0.7498946
## phy      0.6578172 0.8860467 0.8170998 1.0000000 0.8615512 0.7081717
## chem     0.6271155 0.8304643 0.7868710 0.8615512 1.0000000 0.6441334
## politics 0.7257003 0.7090681 0.7498946 0.7081717 0.6441334 1.0000000
## bio      0.6902282 0.7951987 0.7731044 0.8077105 0.7578770 0.7071181
## history  0.6971145 0.7732791 0.7948219 0.8100599 0.7993298 0.7192860
## geo      0.6438662 0.7723853 0.7265406 0.7814152 0.7264814 0.6906930
## pe       0.2712453 0.3300249 0.3159347 0.3251233 0.2769066 0.3033607
##                bio   history       geo        pe
## chn      0.6902282 0.6971145 0.6438662 0.2712453
## math     0.7951987 0.7732791 0.7723853 0.3300249
## eng      0.7731044 0.7948219 0.7265406 0.3159347
## phy      0.8077105 0.8100599 0.7814152 0.3251233
## chem     0.7578770 0.7993298 0.7264814 0.2769066
## politics 0.7071181 0.7192860 0.6906930 0.3033607
## bio      1.0000000 0.7771735 0.8382525 0.2428081
## history  0.7771735 1.0000000 0.7731044 0.2708434
```

```
## geo       0.8382525 0.7731044 1.0000000 0.2605251
## pe        0.2428081 0.2708434 0.2605251 1.0000000
```

可以看出：数学和物理的相关性达 88%，物理和化学成绩的相关性达 86%。

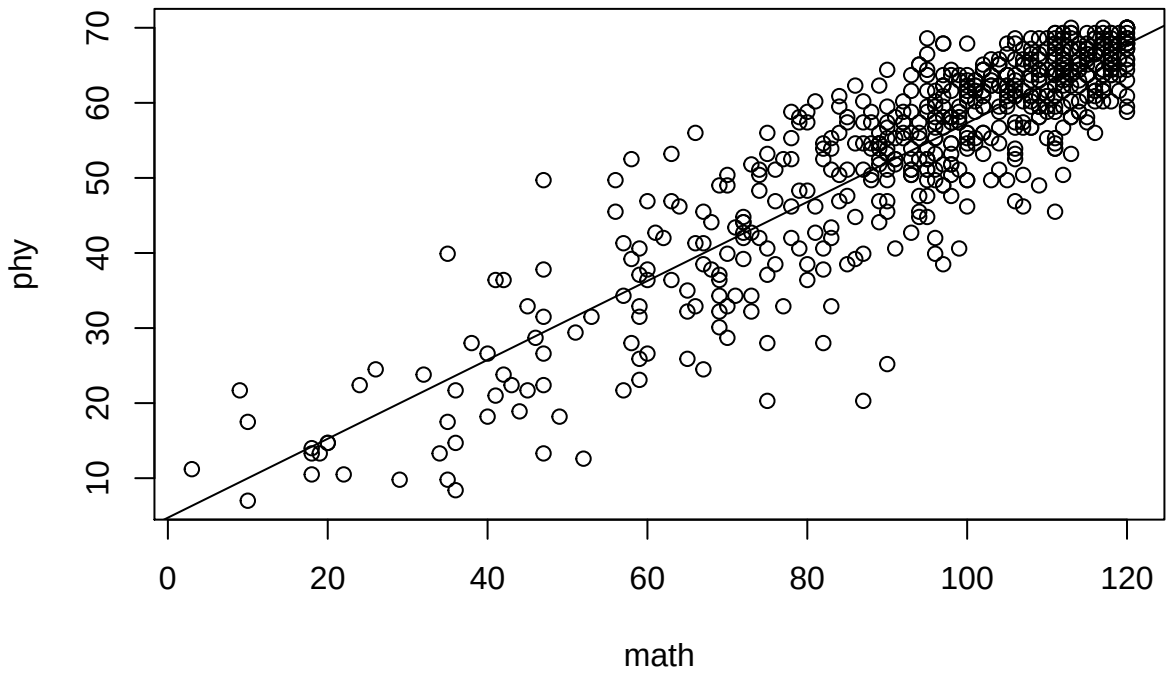
画个图出来看看

```
pairs(scores[,subjects])
```



详细看看数学和物理的线性相关性

```
cor_phy_math <- lm(phy ~ math, scores)
plot(math, phy)
abline(cor_phy_math)
```




```
cor_phy_math
```

```
##
## Call:
## lm(formula = phy ~ math, data = scores)
##
## Coefficients:
## (Intercept)         math
##      4.7374       0.5258
```

也就是说拟合公式为：phy = 0.5258 * math + 4.7374, 为什么是 0.52？因为数学最高分为 120，物理最高分为 70