◆ 記述統計の復習：R を用いた計算技法

R = A Statistical Computing Environment

1) **C 言語よりも手軽.**
   なぜなら，対話型の言語である.
   R では，S 言語を使用する.
   S 言語は，C 言語と同じ所（ベル研）で生まれる（C 言語と類似）.

2) **スプレッドシート・アプリケーション（代表：Excel ）より柔軟.**
   なぜなら，コマンドによる記述である.
   したがって，実行の手順などを理解しやすい（第３者に説明しやすい）.
   つまり，複雑な手順を記述しやすい（GUI アプリケーションは，
   この点で，柔軟さに欠ける）

3) 入手先：　`http://www.r-project.org/`
   `http://cran.md.tsukuba.ac.jp/`

4) 起動して，データファイルのアクセスを確保（MacOSX の場合）：

```
R : Copyright 2003, The R Foundation for Statistical Computing
Version 1.8.1  (2003-11-21), ISBN 3-900051-00-3

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.

> getwd()
[1] "/Users/tk"
> setwd("/Volumes/working/studies2004/lect3")
> list.files()
[1] "conbini.txt" "math.txt"    "metal.txt"   "p57.txt"
```

5) 起動して，データファイルのアクセスを確保（Windows の場合）：

```
R : Copyright 2004, The R Foundation for Statistical Computing
Version 2.0.0  (2004-10-04), ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.

> getwd()
[1] "C:/Program Files/R/rw2000"
> setwd("D:/works04/da")
> list.files()
[1] "conbini.txt" "math.txt"    "metal.txt"   "p57.txt"
>
```

演習室では，
以下のように指定

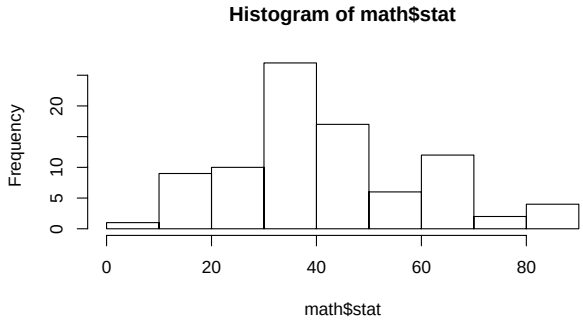
```
> setwd("Z:/")
```

6）データの読み込み：

```
> math <- read.table("math.txt", header=TRUE)
> dim(math)
[1] 88  5
> names(math)
[1] "dyna" "vect" "alge" "anal" "stat"
```

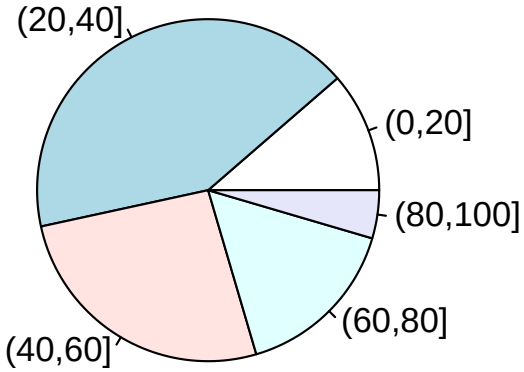
7）データの図示【hist, stem】：

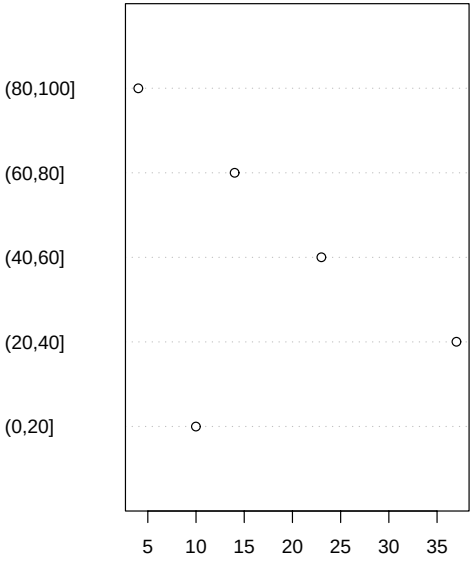
```
> hist(math$stat)
> stem(math$stat)

  The decimal point is 1 digit(s) to the right of the |

  0 | 9
  1 | 45778889
  2 | 012455699
  3 | 00111223333445566777 99
  4 | 000000011234445555 6679
  5 | 0113346
  6 | 11233447888
  7 | 033
  8 | 1111
```



8）データの図示【pie, dotchart】：

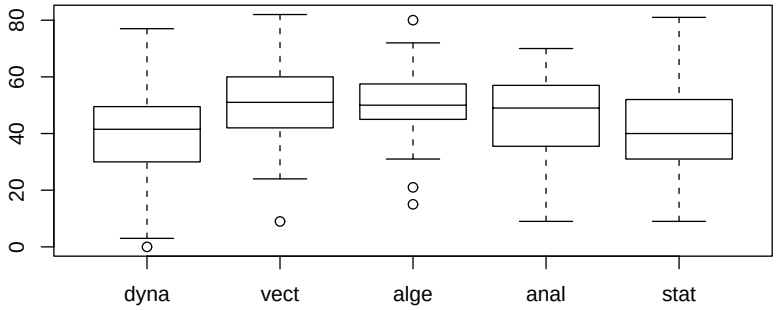
```
> (math.stat.tab <- table(cut(math$stat, breaks=seq(0,100, by=20))))

  (0,20]  (20,40]  (40,60]  (60,80] (80,100]
      10       37       23       14        4
> pie(math.stat.tab)
> dotchart(math.stat.tab)
```

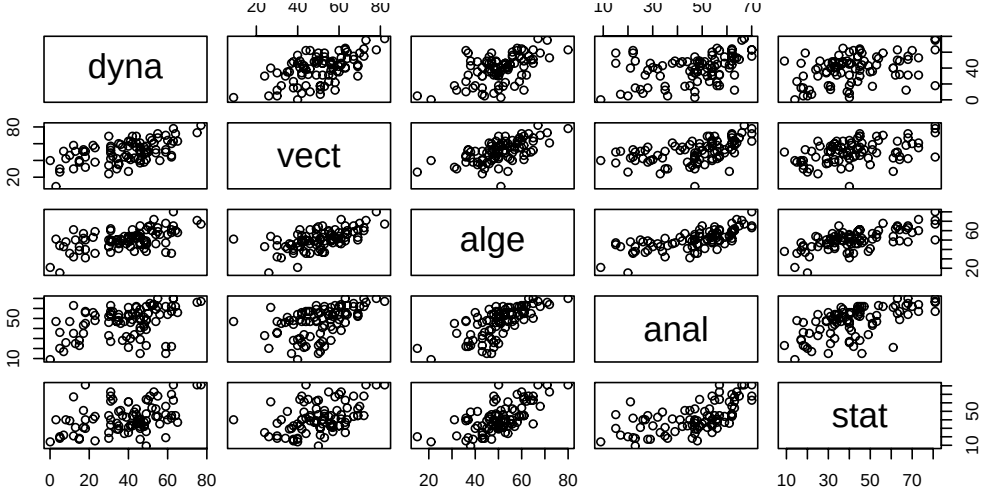




9）データの図示【boxplot, pairs】：

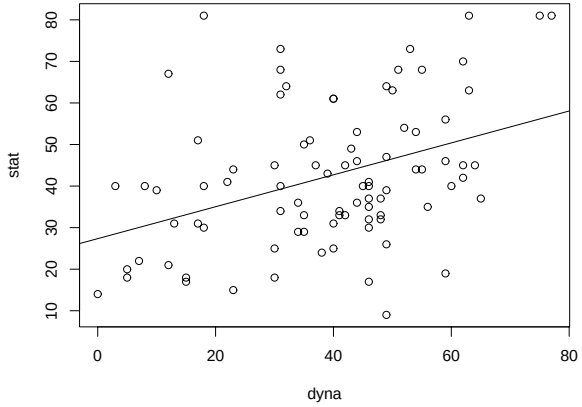
```
> boxplot(math)
> boxplot(math,
+ horizontal=TRUE)
> pairs(math)
```





10）データの図示【plot, abline】：

```
> cov(math)
         dyna      vect      alge      anal      stat
dyna 305.7680 127.22257 101.57941 106.27273 117.40491
vect 127.2226 172.84222  85.15726  94.67294  99.01202
alge 101.5794  85.15726 112.88597 112.11338 121.87056
anal 106.2727  94.67294 112.11338 220.38036 155.53553
stat 117.4049  99.01202 121.87056 155.53553 297.75536
> cor(math)
> # cf. t(scale(math)) %*% scale(math)/(88 - 1)
> # cf. crossprod(scale(math))/(88 - 1)
>
> plot(math[,c("dyna","stat")])
> cor(math[,c("dyna","stat")])
          dyna      stat
dyna 1.0000000 0.3890993
stat 0.3890993 1.0000000
> abline(lm(stat ~ dyna, data=math))
```



11）データの要約値：

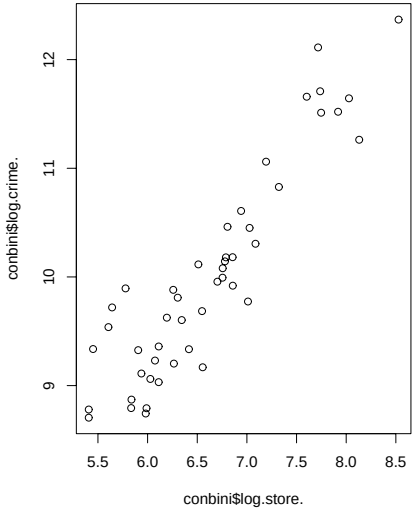
```
> summary(math)
      dyna            vect            alge            anal            stat
 Min.   : 0.00   Min.   : 9.00   Min.   :15.00   Min.   : 9.00   Min.   : 9.00
 1st Qu.:30.00   1st Qu.:42.00   1st Qu.:45.00   1st Qu.:35.75   1st Qu.:31.00
 Median :41.50   Median :51.00   Median :50.00   Median :49.00   Median :40.00
 Mean   :38.95   Mean   :50.59   Mean   :50.60   Mean   :46.68   Mean   :42.31
 3rd Qu.:49.25   3rd Qu.:60.00   3rd Qu.:57.25   3rd Qu.:57.00   3rd Qu.:51.50
 Max.   :77.00   Max.   :82.00   Max.   :80.00   Max.   :70.00   Max.   :81.00
> summary(math$stat)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   31.00   40.00   42.31   51.50   81.00
> quantile(math$stat)
  0%  25%  50%  75% 100%
 9.0 31.0 40.0 51.5 81.0
> apply(math, 2, mean)
    dyna     vect     alge     anal     stat
38.95455 50.59091 50.60227 46.68182 42.30682
> apply(math, 2, sd)
    dyna     vect     alge     anal     stat
17.48622 13.14695 10.62478 14.84521 17.25559
```

注意：

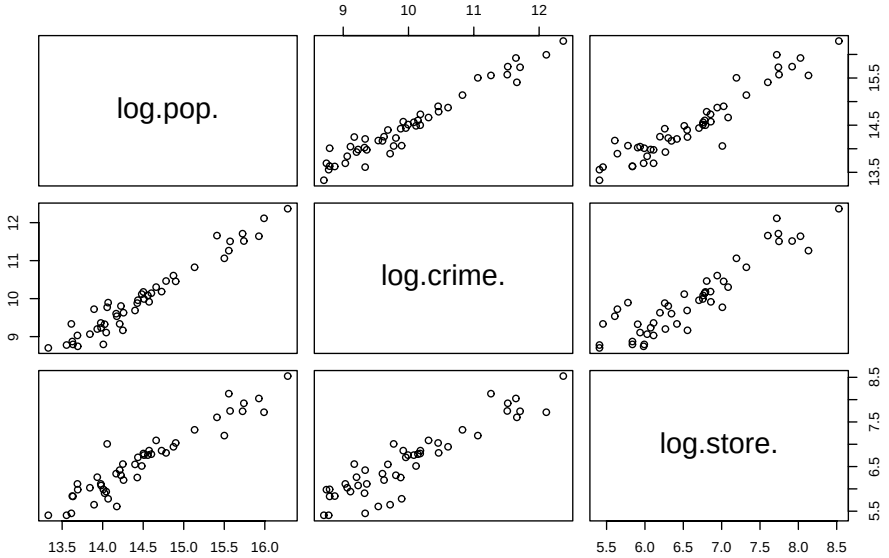
```
> boxplot(math$stat)$stats
     [,1]
[1,]    9
[2,]   31
[3,]   40
[4,]   52
[5,]   81
> fivenum(math$stat)
[1]  9 31 40 52 81
```

12）コンビニ・データ：

```
> conb <- read.table("conbini.txt", header=TRUE)
> attach(conb)
> log(pop)
 [1] 15.55463 14.20868 14.16582 ...
> conbini <- data.frame(log(pop), log(crime), log(store))
> names(conbini)
[1] "log.pop."   "log.crime." "log.store."
> cor(conbini)
            log.pop. log.crime. log.store.
log.pop.   1.0000000  0.9729749  0.9463297
log.crime. 0.9729749  1.0000000  0.9161085
log.store. 0.9463297  0.9161085  1.0000000
>
> plot(conbini$log.store., conbini$log.crime.)
```



13）コンビニ・データ（続き）：

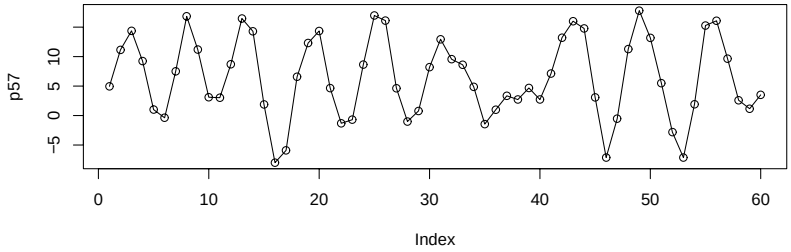



```
> pairs(conbini)
> res1 <- lm(log.crime. ~ log.pop., data=conbini)
> res2 <- lm(log.store. ~ log.pop., data=conbini)
> cor(res1$residuals, res2$residuals)
[1] -0.06225971
```

偏相関係数 … 犯罪数とコンビニ店舗数はあまり関係がない（いずれも，人口と密接に関係するため）

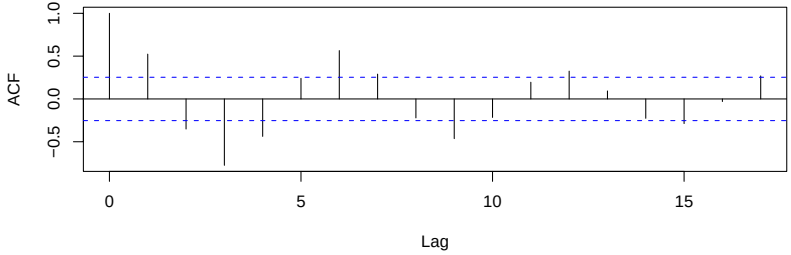
14）時系列データ：

```
> p57 <- scan("p57.txt")
Read 60 items
> op <- par(mfrow=c(2,1))
> plot(p57, type="o") # type="l", type="b", etc
> acf(p57)
> par(op)
```



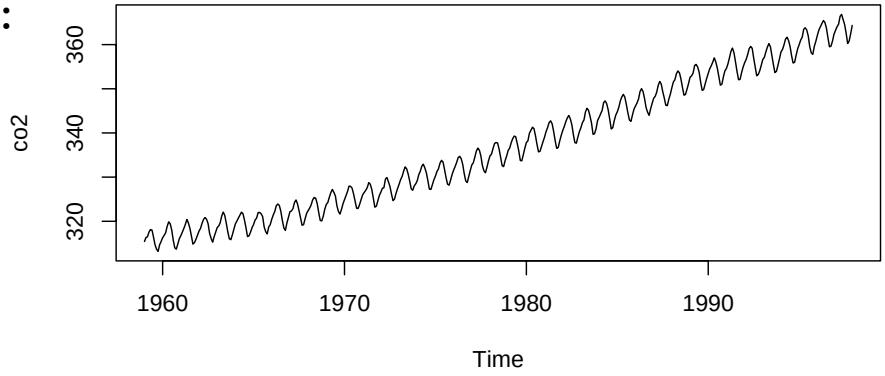
15）短いデータの読み込み：

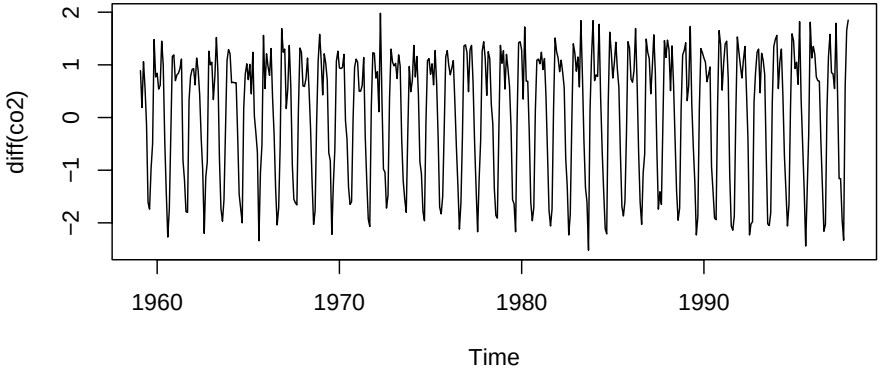
```
> # text p.6
> galton <- scan()
1: 15.3
2: 16.0
3: 15.6
4: 16.3
5: 16.0
6: 17.3
7: 17.5
8:
Read 7 items
>
```



16）内蔵データの読み込み：

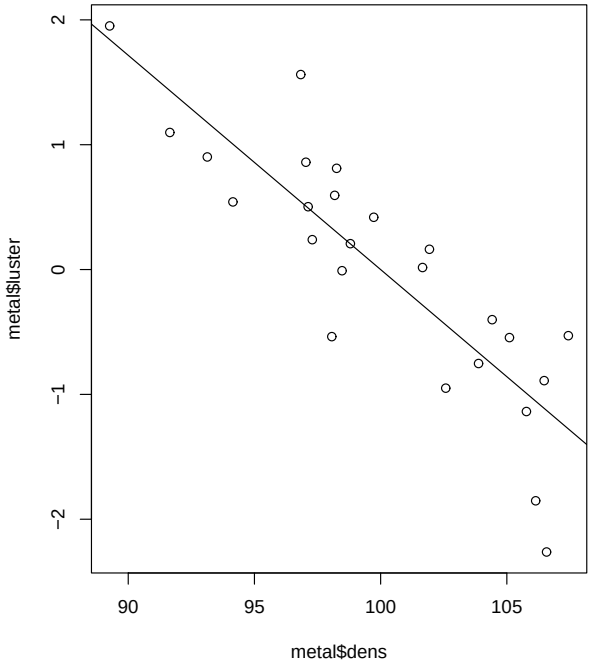
```
> ?co2
> data(co2)
> par(mfrow=c(2,1))
> plot(co2)
> plot(diff(co2))
> data()
```





17）メタル・データ：

```
> metal <- read.table("metal.txt", header=TRUE)
> metal[1:5,]
    dens temp2 luster
1 106.57 57.81 -2.262
2  98.07 47.63 -0.538
3  97.04 45.80  0.860
4  98.18 47.03  0.594
5 101.66 50.74  0.017
> plot(metal$dens, metal$luster)
> res <- lm(luster ~ dens, data=metal)
> res

Call:
lm(formula = luster ~ dens, data = metal)

Coefficients:
(Intercept)         dens
    17.1623      -0.1716

> abline(res)
```



18）応用：メタル・データ（分散解析）：

```
> (ssr <- sum((res$fitted.values - mean(res$fitted.values))^2))
[1] 17.66932
> (sse <- sum(res$residual^2))
[1] 6.330017
> sse + ssr
[1] 23.99934
> sum((metal$luster - mean(metal$luster))^2)
[1] 23.99934
>
> anova(res)
Analysis of Variance Table

Response: luster
          Df  Sum Sq Mean Sq F value    Pr(>F)
dens       1 17.6693 17.6693  64.201 4.175e-08 ***
Residuals 23  6.3300  0.2752
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

19）応用：メタル・データ（線形モデル）：

```
> summary(res)

Call:
lm(formula = luster ~ dens, data = metal)

Residuals:
     Min       1Q   Median       3Q      Max
-1.13432 -0.27695  0.01055  0.35210  1.01978

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.16233    2.14450   8.003 4.26e-08 ***
dens        -0.17162    0.02142  -8.013 4.18e-08 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 0.5246 on 23 degrees of freedom
Multiple R-Squared: 0.7362,  Adjusted R-squared: 0.7248
F-statistic:  64.2 on 1 and 23 DF,  p-value: 4.175e-08
```

20）応用：メタル・データ（線形モデル，続き）

```
> par(mfrow=c(2,2))
> plot(res)
```