鈴木義一郎著「J 言語による統計分析」ノート (J 6 対応版)

2006 年 9 月 22 日

目次

はじめに

鈴木義一郎著「J 言語による統計解析」1966 森北出版は、添付されていた DISK に膨大な統計プログラムが入っていた。

AIC を中心に据えた画期的な著作であるが、残念ながら絶版となり新たに入手できなくなった。

J 言語は当時のバージョンから最新の J601 に進化しているが、関数面では大幅な改定ではなく、少しずつ改良が加えられている。新たに加わった関数を用いなくても十分実用になるので、J6 で動くように最小限の改良を加え、グラフィックスを付けて「基本統計パッケージ」として提供するものである。

改良とグラフィックスの追加、レファレンスとしてのとりまとめは日本 APL 協会会員の志村の責任で行った。

豊富な関数の使い方と解を得る方法に絞ったので、原著の理論的記述は著者の他の著書で補ってほしい。

stadt_J

file	stat_j_suzuki.ijs stat_j_data.ijs stadt_j_foreign.ijs	
J の入手とインストール	http://www.jsoftware.com からダウンロードする インストーラに従って「Y」と打つ。	WIN/2000 以降 MAC/OSX LINUX(BSD OK) POCKET PC
load	RUN → FILE → stadt_j.ijs	サイレント ロード 他のファイルは stat_j_suzuki.ijs の先頭の行に書いておくと良い。 load require の差は CTRL+W で load は指定ファイルの load を繰り返す。 本格的には PROJECT を組む

SCRIPT の編集	FILE → OPEN で編集画面に load して CTRL (+)W で active になる。 変更も CTRL+W で取り込む F12 や ALT+TAB で IJX と IJS が変わる。	

J6 の関数の変更 (対応しないと動かないもの

	J5 まで	J6
引数 動詞の引用	`x. y. m. n. u. v.`	`x y m n u v` y=.y. などが支障になる。
文字化	`7.5 ":`	`7j5 ":`

1　統計データの縮約

1.1　統計データの性格と縮約値

1.2　最大値と最小値

区分	Script	単項 Monad	両項 Dyad
最大値	max=:3 : 0 >./y. : x.>.y.)	max 6 4 8 8	6 max 4 6
最小値	min=:3 : 0 <./y. : x.<.y.)	min 6 4 8 4	6 min 4 4

この関数は複素数は扱えない。

```
  1j2 max 2j1
|domain error: max
|   x.    >.y.
```

最大値、最小値を求め昇順に並べる。

```
order=:/:~
```

[*1]

*DS*0　NB. 1991 年の東京の春分の日を挟む 25 日間の最低気温のデータ

[*1] order=:sort=:/:~とすることが多い

```
  # DS0    NB.  # is take n
25

  5 5 $ DS0
3.4  5.5  7.7 10.2  8.5
7.6 10.1 11.5  6.9    2
2.3  5.9  7.6  1.8  3.3
1.7  2.6  6.6 10.3  9.4
7.2  8.8 10.2 10.7 10.8
```

```
  max DS0       NB. maximum
11.5

   min DS0      NB. minimum
1.7

  5 5  $ order DS0
NB. find order /up sort
 1.7  1.8    2  2.3  2.6
 3.3  3.4  5.5  5.9  6.6
 6.9  7.2  7.6  7.6  7.7
 8.5  8.8  9.4 10.1 10.2
10.2 10.3 10.7 10.8 11.5
```

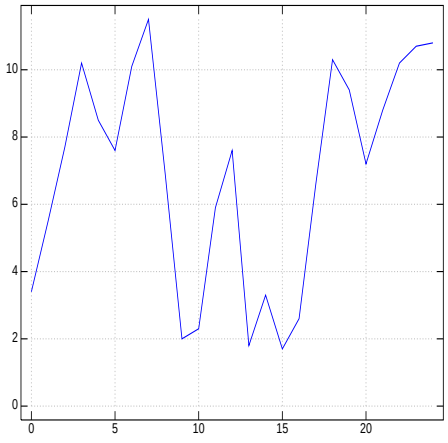

図 1　東京の 1991 年 3 月の気温

1.3 範囲と四分位偏差

四分位偏差	rngq=:3 : 0 r=:a,b,b+a=:<.-:b=:<.0.5+-:#y. ({:-{.)-:+/"1(r{y),.(<:r){y=:/:˜y.)
範囲	rng=:>./ - <./

四分位偏差	データを昇順に並べて $(b+a), (b+a+1)$ 番目の要素の平均から $(a, a+1)$ 番目の要素の平均を引いた値	rngq 2 4 6 8 10 12 14 16 8 NB. **四分位偏差** rng DS0 9.8 NB. **範囲・レンジ** rngq DS0 6.8 NB. **四分位偏差**	

1.4 平均値

算術平均	$\frac{1}{n}\sum_{i=1}^{x} x_i$	mean=:+/ % #	mean >:i.5 3
幾何平均	$\sqrt[n]{\Pi_{i=1}^{n} X_i}$	meang=:# %: */	meang >:i.5 2.60517

調和平均	$\frac{n}{\sum_{i=1}^{n}\frac{1}{X_i}}$	meanh=:mean&.(%"_)	meanh >:i.5 2.18978

1.5 標準偏差

標準偏差	$V(x)=\frac{1}{n}\sum_{i=1}^{n}(x_i-M(x))^2$	sd >:i.5 1.41421 NB. **標準偏差** mesd >:i.5 3 1.41421 NB. **算術平均と標準偏差**

```
ss=:[: +/ *:&(- mean)  NB. 偏差平方和

sd=:%:&(ss % #)        NB. 標準偏差  standard deviation

mesd=:mean , sd   NB. mean and sd
```

1.6 相関比

m 個のデータ $(x_1, x_2 \cdots x_m)$ の平均を $M(x)$, 分散を $V(x)$,
n 個のデータ $(y_1, y_2 \cdots y_n)$ の平均を $M(y)$, 分散を $V(y)$,
2 組のデータを合わせた複合データ z の平均 $M(z)$ と分散 $V(z)$ は次のようになる

$$M(z)=\alpha M(x)+(1-\alpha)M(y)$$
$$V(z)=\alpha V(x)+(1-\alpha)V(y))+(\alpha(M(x)-M(z))^2+(1-\alpha)(M(y)-M(z))^2$$

全分散 = 級内分散 + 級間分散

$$\text{相関比 } h=\frac{\text{級間分散}}{\text{全分散}}=\frac{\text{級間分散}}{\text{級内分散}+\text{級間分散}}$$

相関比	('KT';'KZ';'ST';'SZ') ,.{KT,KZ,ST,:SZ +--+---------------+ |KT|80 70 60 90 100| +--+---------------+ |KZ|50 60 70 80 90 | +--+---------------+ |ST|78 79 82 81 80 | +--+---------------+ |SZ|70 69 71 68 72 | +--+---------------+ KT **太郎の** 5 **回の国語のテスト** KZ **次郎の** 5 **回の国語のテスト** ST **太郎 数学** SZ **次郎 数学**	corro KT;KZ within variance : 200.00 between variance : 25.00 total variance : 225.00 correration ratio : 0.11 corro ST;SZ within variance : 2.00 between variance : 25.00 total variance : 27.00 correration ratio : 0.93 国語の相関比 0.11 に比べて算数の相関比 0.93 は大きく数学の成績に関しては両者に顕著な差がある。

```
  corro=:3 : 0
   m=:+/(k=:(+/%#)&>y.)*w=:n%+/n=:#&>y.
   v=:(((+/"1*:(>y.)-k)%n),:*:k-m)+/ . * w
   w=:'within variance   :',8j2":{.v
   b=:'between variance  :',8j2":bv=:}.v
   t=:'total variance    :',8j2":v=:+/v
   w,b,t,:'correration ratio :',8j2":bv%v
)
```

1.7 移動平均と指数平滑化

移動平均		3 7 $ 5 mav DS0 7.06 7.9 8.82 9.58 8.92 7.62 6.56 5.72 4.94 3.92 4.18 4.06 3.4 3.2 4.9 6.12 7.22 8.46 9.18 9.26 9.54 # 5 mav DS0 21 5 plot_mav DS0 pd 'eps temp\mav.eps'

指数平滑化	$\bar{x_1} = x_1$ $\bar{x_2} = w\bar{x_1} + (1-w)x_2$ $\bar{x_3} = w\bar{x_2} + (1-w)x_3 = w^2x_1 + w(1-w)x_2 + (1-w)x_3$	ウエイトの値は 0.6 とする `5j2 ": 5 5 $  0.6 smexp DS0` `3.40 4.24 5.62 7.45 7.87` `7.76 8.70 9.82 8.65 5.99` `4.51 5.07 6.08 4.37 3.94` `3.04 2.87 4.36 6.74 7.80` `7.56 8.06 8.91 9.6310.10` `5j2 ":`は文字化と桁数指定

移動平均は次数分-1 個のデータが落ちる。plot では工夫が必要。偶数次数を指定したときは中心化を行うと比較したりグラフを合わせることが出来る。

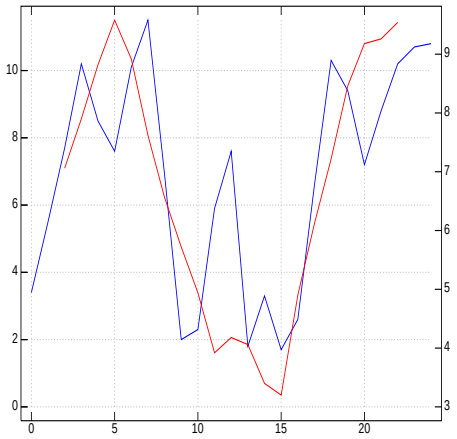

図 2　東京の 1991 年 3 月の気温と移動平均

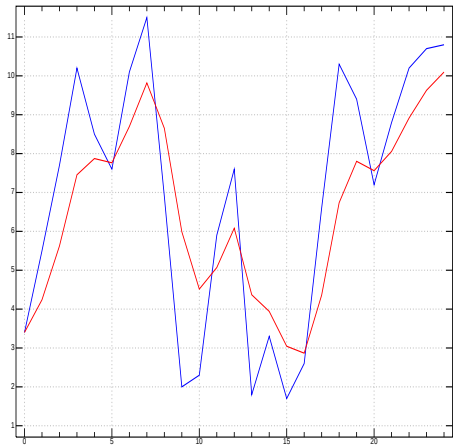

図 3　東京の 1991 年 3 月の気温と指数平滑化

移動平均	mav=:+/\ % [
指数平滑化	smexp=:4 : 0 s=:{.Y=:y. while. 1<#Y do. s=:s,(x.*{:s)+(1-x.)*{.Y=:}.Y end.)

```
plot  DS0,: 0.6 smexp DS0
   pd 'eps temp/smexp.eps'
```

1.8　相関係数

$M(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$ NB. mean

$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - M(x))^2$ NB. valiance

$x_{01} = \dfrac{(x_i - M(x))}{\sqrt{V(x)}}$ NB. 標準化

$R(x,y) = \dfrac{1}{n}\sum_{i=1}^{n} x_{01} \cdot y_{01}$ NB. 相関係数

```
   DS1     NB. 前期の成績
8 8 6 6 6 4 4 3 3 2
   DS2     NB. 後期の成績
9 5 9 7 5 7 4 7 4 3


   stand DS1    NB. 標準化
1.5 1.5 0.5 0.5 0.5 _0.5 _0.5 _1 _1 _1.5
   stand DS2
1.5 _0.5 1.5 0.5 _0.5 0.5 _1 0.5 _1 _1.5
   DS1 cor DS2     NB. 相関係数
0.525

   DS3  NB. 札幌 山形 東京 大阪 鹿児島の緯度
43.05 38.25 35.68 34.68 31.57
   DS4  NB. 1 月の各都市の最低気温
_8.9 _4.2 0.5 2.2 2.4


   DS3 cor DS4 NB. マイナスの相関
_0.962308
```

相関係数	`cor=:+/@([ *&stand ]) % #@]`

1.9 回帰直線

$Y = (y_1, y_2, \cdots y_n)$ n 次のベクトル

$B = \dfrac{XY'}{XX'}$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \\ 1 & x_n \end{bmatrix}$$

```
   DS4 reg DS3
38.2955 _1.08867
```

$f = 38.2955 - 1.08867x$

緯度 (DS3) が一度高くなると、最低気温の平年値 (DS4) が一度下がる。

回帰 OLS	`reg=;[ %. 1&(,"0)@]`	same as `reg=: [ %. 1&,.@]` `reg=: 4 : 'x %. 1 ,. y'`

1.10 分類と集計

2 cly DS0 NB. 2 刻みで分類する

order NB. sort

order 2 cly DS0

2 3 4 6 5 4 6 6 4 1 2 3 4 1 2 1 2 4 6 5 4 5 6 6 6

acum 分類

```
acum order 2 cly DS0  NB. 集計表
1 3
2 4
3 2
4 6
5 3
6 7
```

	2 table DS0 1 3 2 4 3 2 4 6 5 3 6 7	grf order 2 cly DS0 +-+-------+ |1|*** | |2|**** | |3|** | |4|****** | |5|*** | |6|*******| +-+-------+

2　確率と確率分布

2.1　不確実な現象と統計的現象

2.2　標本空間と確率空間

標本空間 統計的現象の記号化された結果の集合
確率空間 標本空間の確率まで対応している結果の集合

2.3　事象の確率

全事象	U	起こりうる全ての結果の集合
空事象	ϕ	決して起こらない事象
余事象	$\bar{E},$	事象 E に対して、E が起こらないという事象、E の補集合
和事象	$E \cup F$	事象 E, F の少なくとも一方が起こる事象
積事象	$E \cap F$	E, F がともに起こる事象
排反事象	$E \cup F = \phi$	ある試行によって起こる事象で、片方が起これば、他方は決して起こらない事象

2.4　相互排反性と加法定理

Working Example

加法法則	$p(E \cup F) = p(E) + p(F) - P(E \cap F)$	
加法定理	$p(E \cup F) = p(E) + p(F)$	排反事象の場合 和事象の確率は各々の事象の確率の和で与えられる

全部で 8 頭出走する競馬レース	贔屓の馬の勝つ (*success* S) 確率	$p(s) = \frac{1}{8}$
ドサ競馬なので何回も走る	2 回のレースで一回勝つ確率 $p(S_1 \cup s_2) = p(S_1) + p(S_2) - P(S_1 \cap S_2)$	$\frac{1}{8} + \frac{1}{8} = \frac{16}{64}$ ではなく (相互排反的でなく同じ馬が勝つことはある) $\frac{1}{8} + \frac{1}{8} - \frac{1}{64} = \frac{15}{64}$

2.5 事象の独立性と乗法定理

乗法法則	$P(E \cap F) = P(E) \cdot P(E \mid F)$	
条件付き確率	$P(E \mid F) = \frac{P(E \cap F)}{P(E)}$	
独立	$P(F) = P(F \mid E)$	
乗法定理	$P(E \cap F) = P(E) \cdot P(F)$	E *cdot* F が独立なとき

乗法定理 積事象の確率は各の事象の積で与えられる。

2.6 整数乱数

```
   rnd 10  NB. 100 までの整数乱数を 10 個生成
46 55 79 52 54 39 60 57 60 94

   rnd 3 10    NB. 3 行 10 列の整数乱数
46 78 13 18 51 92 78 60 90 62
31 16 60 64 64 71 13  3 76 26
25 77 68 48 42 91 99 97 99  9

   10 rnd 10   NB. 0 から 9 までの整数乱数を 10 個生成
1 9 7 8 2 4 6 5 0 3
```

2.7 サイコロ投げの実験

```
   dice 10  NB. サイコロを$n=10$回振る
3 6 6 5 6 4 4 3 1 6

]  D=. dice 5 10  NB. 50 回

1 1 5 3 3 2 1 6 1 6
4 6 1 3 2 5 2 6 5 4
5 1 6 3 4 2 5 4 4 4
2 1 4 6 6 1 6 2 4 4
3 3 3 2 4 2 6 5 6 3

   +/ D     NB. 縦の合計
15 12 19 17 19 12 20 23 20 21

   +/ "1 D  NB. 横の合計
29 38 38 36 37

   acum ,D  NB. 目の分類
1  8
5  6
3  8
2  8
6 10
4 10
```

2.8 カード抽出の実験

```
card 5   NB. 5 枚
C4
```

CT
D6
S4
S6

cardb 13 NB. 13 枚を分類

```
+--+--+--+--+--+
|C7|C8|  |  |  |
+--+--+--+--+--+
|D2|D3|D5|DJ|DA|
+--+--+--+--+--+
|H4|H5|  |  |  |
+--+--+--+--+--+
|S6|S9|SJ|SA|  |
+--+--+--+--+--+
```

2.9　2 項係数とパスカルの三角形

組み合わせの数	2!5 10	5 個から 2 個選び出す組み合わせの数
2 項係数	bic 5 1 5 10 10 5 1	
Pascal triangle パスカルの三角形	bic "0 i.5 1 0 0 0 0 1 1 0 0 0 1 2 1 0 0 1 3 3 1 0 1 4 6 4 1	
同じ	pascal 5 1 1 1 1 2 1 1 3 3 1 1 4 6 4 1	

2.10　確率変数と確率分布

確率変数		
期待値	$E(X) = \mu = \sum_{i=1}^{n} x_i \times p_i$	
分散	$V(X) = \sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 \times p_i$	

標準偏差	$\sigma = \sqrt{\sigma^2}$	
確率関数	$p(x_i) = p_i \ \ i = 1, 2, 3 \cdots, k,$	

2.11 確率変数の独立性

X と Y が独立な場合は

$V\{X + Y\} = X\{X\} + X\{Y\}$

2.12 2 項分布

*2

ベルヌイ試行	$p(s) = P$ $P(F) = 1 - p = q$	一回の試行で表か裏かどちらか (出る確率は問わない)
2 項分布	$p(X = k) =_n C_k p^k q^{n-k}$	
X の期待値	$E(X) = np$	
X の分散	$V(X) = npq$	

4 binom 0.5 NB. n=4,p=0.5 の 2 項分布の確率関数

```
0.0625 0.25 0.375 0.25 0.0625
```

10 bden 0.1 NB. n が大きく p が小さい場合の0.001以下をくくる関数

```
0.348678 0.38742  0.19371 0.0573956 0.0111603 0.00148803 0.000146903
```

4 bgf 0.5 NB. グラフ表示

```
+-+------------------------------+
|0|*****                         |
|1|********************          |
|2|******************************|
|3|********************          |
|4|*****                         |
+-+------------------------------+
```

*2 囲碁の世界では近頃コミが 5.5 目から 6.5 目に変更され、大コミ時代を迎え、黒が先行逃げ切りから戦闘的に変わった。白黒のごく僅かな勝率差が調整された。確率では僅差からポアソンのような希に起こる確率まで取り扱われる。

```
   4 bmev 0.5        NB. 2 項分布の平均と分散$(n=4,p=0.5)$
2 1
```

2.13 ポアソン分布

n(実験回数) が大きく,p(成功の確率) が小さい場合の 2 項分布は $m = np$(2 項分布の平均) を平均とするポアソン分布で近似できる。

2 項分布 $B(n, p)$ の確率	$p(x) =_n C_k p^k q^{n-k}$ $p(x+1) = p(x) \times \frac{p}{1-p} \cdot \frac{n-x}{x+1}$	
$x = 0$ の場合	$p(0) = (1-p)^n$	
$m = np$ と置くと $\left[1 - \frac{m}{n}\right]^n$	$p(0) = e^{-m}$	m に比べて n の値が大きいときは e^{-m} で近似できる
n の値が大きいとき	$p(x+1) = p(x) \times \frac{m}{x+1}$ と近似できる	
poison distribution	$p(0) = e^{-m}$ $p(1) = \frac{m}{!} e^{-m}$ $p(2) = \frac{m^2}{2!} e^{-m}$ $p(3) = \frac{m^3}{3!} e^{-m}$	

ポアソン分布の期待値と分散	$E(x) = V(x) = m$	

pden 1 NB. **平均が** 1 **のポアソン分布の** 0 **から** 6 **までの値**

```
0.367879 0.367879 0.18394 0.0613132 0.0153283 0.00306566 0.000594185
```

100 bden 0.01 NB. n=100 p=0.01 **の** 2 **項分布**

```
0.366032 0.36973 0.184865 0.0609992 0.0149417 0.00289779 0.000534534
```

2.14 正規分布

試行回数が非常に大きい場合、2 項分布 $B(n, p)$ の確率

$P(X = np + d) = P(Y = p + \frac{d}{n})$

と言う値が

$\frac{1}{\sqrt{2\pi npq}} \times exp(-\frac{d^2}{2npq})$

という値で近似できる (中心極限定理)

標準正規分布 $\psi_0(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

Z=標準正規分布

$X = \mu + \sigma Z$ と置くと

$E(X) = \mu,\ V(X) = \sigma^2$

変数 X を 平均 μ, 分散 σ の正規分布は $X \sim N(\mu,\ \sigma^2)$ と表す

標準正規分布の確率密度関数を求める関数

ndens 0 0.5 1 2 NB. [0,0.5 1,2] **での標準正規分布の確率密度関数の値**

```
0.398942 0.352065 0.241971 0.053991
```

0 1 nden 0 0.5 1 2 NB. **左引数で与えた平均と分散の値を持つ正規分布の確率密度**

```
0.398942 0.352065 0.241971 0.053991
```

```
   1 4 nden 1 2 3 4   NB. 同上
0.199471 0.176033 0.120985 0.0647588

   ndfs 1   NB. 標準正規の確率関数の (0,1) の範囲の積分
0.341344

   ndf1 1.96 NB. 同じく y(右引数) までの積分
0.975002

   1 ndf2 2  NB. 標準正規の確率密度の (1,2) の範囲の積分
0.135905
```

2.15 (整数型) 正規乱数

```
   nrnd 10
43 31 30 28 28 28 36 35 30 35
   nrne 20
5 5 5 5 7 4 5 8 5 7 6 7 8 6 8 4 7 6 5 8
   acum order nrne 100
4 13
5 26
6 31
7 20
8  9
9  1
```

2.16　正規分布に関連した分布

カイ二乗分布	Z は標準正規分布 $N(0,1)$ に従う確率変数とする。 $C_k^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2$ $E(Z_k^2) = 1,\ V(Z_k^2) = 2,\ i = 1, 2, \cdots, k$ $E(C_k^2) = k,\ V(C_k^2) = 2k$	
t 分布	$t = \dfrac{1}{\sqrt{\dfrac{c^2}{k}}}$ $E(t) = 0,\ V(t) = \dfrac{k}{k-2},\ (k > 2)$	
F 分布	$F = \dfrac{\dfrac{C_1^2}{k_1}}{\dfrac{C_2^2}{k_2}}$ $E(F) = \dfrac{k_2}{k_2 - 2},\ k_2 > 2$ $V(F) = \dfrac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}$	

3　推定・検定の考え方

3.1

硬貨を 10 回投げて 3 回表が出た。表の出る確率を p とすると、
表の出る回数 X は 2 項分布 $B(10, p)$ に従う。

```
 (i.6),. 6{. |: ;("1) ,. 10 bden L:0 {@>0.2 0.25 0.3 0.35  0.4
```

```
表   p=0.2       0.25       0.3      0.35       0.4
0  0.107374 0.0563135 0.0282475 0.0134627 0.00604662
1  0.268435  0.187712  0.121061 0.0724917  0.0403108
2   0.30199  0.281568  0.233474  0.175653   0.120932
3  0.201327  0.250282  0.266828   0.25222   0.214991
4 0.0880804  0.145998  0.200121  0.237668   0.250823
5 0.0264241 0.0583992  0.102919   0.15357   0.200658
```

3 回表の出る場合 p が 0.3 の時が最大=最尤推定
対数に取れば最大対数尤度

3.2　硬貨投げ実験の出現比率

ベルヌイ試行 ある試行の結果が成功 (表 S) か失敗 (裏 F) の 2 通りである試行。
$S = p, F = q = 1 - p$ とする。
S の起こる範囲はほとんど次の範囲で、回数を増やすと p の値に近ずく。

FS の出現範囲	$p \pm 3 \times \sqrt{\frac{pq}{n}}$	pqn=: 4 : 0 NB. x is n (100) NB. yb is p q (0.5 0.5) TMP=.3 *%: ({.y)*({: y)% x (y ,0)+1 _1 1* TMP)

		100 pqn 0.5 0.5 NB. 100 times S=0.5 F=0.5 0.65 0.35 0.15 NB. 65<->35 range +-15 1000 pqn 0.5 0.5 NB. 1000times 0.547434 0.452566 0.0474342 NB. 55<->45 +-5 10000 pqn 0.5 0.5 0.515 0.485 0.015

3.3 比率に対する区間推定

p 目的関数 或る性質や属性を持つ集団の理論比率

n 標本

X 標本の中で当該属性を持つ者の個数

$P = \frac{X}{n}$ 観測比率

X は 2 項分布 $B(n, p)$ に従い、近似的に正規分布 $N(np, npq)$ に従う。

$P(-2 \leq \frac{X - np}{\sqrt{npq}} \leq 2) = 0.95$

$P(np - 2\sqrt{npq} \leq X \leq np + 2\sqrt{npq}) = 0.95$

未知の p, q を $P, Q(= 1 - p)$ で置き換える

$P - 2\sqrt{\frac{PQ}{n}} \leq p \leq P + 2\sqrt{\frac{PQ}{n}}$

不等式を p について解くと

$\frac{X+2}{n+4} - 2D \leq p \leq \frac{X+2}{n+4} + 2D$

$D = \sqrt{((n - X)(X + 1) + X)/n/(n + 4)}$

95% 信頼区間	$\frac{X+2}{n+4} \pm 2D$	civl=:4 : '(x.+2+(-,+)+:%:(x.+(y.-x.)*x.+1)%y.)%y.+4'

x サンプル数 y 母集団 95% 信頼区間

275 civl 400
0.639485 0.731802

3.4 比率の差の区間推定

p_1, p_2	理論比率を持つ 2 つの母集団
m, n	標本数
x, y	その標本の当該属性を持つ個数
$P = \frac{x}{m}, Q = \frac{y}{n}$	観測比率
$d = p_1 - p_2$	2 組の理論比率の差

b 2 組の理論比率の差 $d = p_1 - p_2$ に対する信頼区間を求める。

x, y は 2 項分布に従い、近似的に正規分布に従う。

$P - Q$ は平均が $d = p_1 - p_2$, 分散が $p_1\frac{(1-p_1)}{m} + p_2\frac{(1-p_2)}{n}$ の正規分布で近似できる。

$P(d - 2wb \leq P - Q \leq d + 2w) = 0.95$

$W = \sqrt{p_1(1-p_{1)}/m + p_2(1-p_{2)}/n}$

$P(P - Q - 2W \leq d \leq P - Q + 2W) = 0.95$

$w = \sqrt{P(1-P)/m + Q(1-Q)/n}$

d に対する 95% 信頼区間	$P - Q \pm 2W$	civld=:4 : 0 w=:x*(m-X=:{.x.)%(m=:{:x.)^3 w=:+:%:w+Y*(n-Y=:{.y.)%(n=:{:y.)^3 1 3 { |.((X%m)-Y%n)+w,-w) NB.slightly modified

x left x, m

y right y, n

```
   100 400 civld 90 400
```

```
_0.035156 0.085156
```

NB. 信頼区間は 0 を含んでいるので $d = 0$ の可能性があり、2 組の比率に差がない。

```
  100 400 civld 70 400
0.0173914 0.132609
```

NB. $p_1 > p_2$ 両者の比率に差がある。

3.5 平均の推定

正規分布の場合 95% の的中率で次の不等式が成り立つ。

$$-2 < \frac{\sqrt{n}}{\sigma}(M(x) - \mu) + 2\frac{\sigma}{\sqrt{n}}$$

$$M(x) - 2\frac{\sigma}{\sqrt{n}} < \mu < M(x) + 2\frac{\sigma}{\sqrt{n}}$$

母平均 σ^2 に対する 95% 信頼区間	$M(x) \pm 2\frac{\sigma}{\sqrt{n}}$	estim=:3 : 0 m=:(+/d=:(p=:?10)+nrne n)%n=:y. w=:+:(-s),s=:%:(+/*:d-m)%n e=:'point estimation :',6j2":m c=:'confidence interval :',6j2":m+w e,c,:'population mean :',3":p+6)

乱数を用いているので、実験の都度答えは変化する。

```
   estim 50
point estimation    : 11.96
confidence interval :  9.73 14.19
population mean     : 12
   estim 100
point estimation    : 14.18
confidence interval : 11.91 16.45
population mean     : 14
```

母集団の分散 σ^2 が分からない場合 (こちらの方が普通) は標本分散 s^2 で代用する。必ずしも正規分布でない標本でも、利用できる。

3.6 平均の差に対する区間推定

3.7 理論と実際

$$\chi^2 = \frac{(x-np)^2}{np} + \frac{(n-x-n(1-p)^2)}{n(1-p)} = \left(\frac{x-np}{\sqrt{np(1-p)}} \right)^2$$

$Z = (\frac{x-np}{\sqrt{np(1-p)}})^2$ は平均 0, 標準偏差 1 の標準正規分布に従う。

$P(\chi^2 \geq 4) = P(|Z| \geq 2) = 0.05$

χ^2 は理論と実際のずれの値で 4 を超えるようなら実際の値と理論値がずれている。

3.8 一様性の検定

離散形の一様分布と χ^2 の値

サイコロを 60 回投げた時の出目。このサイコロは正しいか。

```
   (>:i.6),. 12 8 10 13 8 9
```

```
1 12
2  8
3 10
4 13
5  8
6  9

 (10- 12 8 10 13 8 9)        NB. 10-x
_2 2 0 _3 2 1

 (10- 12 8 10 13 8 9)^2
4 4 0 9 4 1                  NB. (10-x)^2

 +/ 10%~(10- 12 8 10 13 8 9)^2
2.2                          NB. +/ ((10-x)^2)%10

   chigf
3 : '(+/*:y.-e)%e=:(+/y.)%#y.'

   chigf 12 8 10 13 8 9
2.2
```

サイコロ投げの自由度は 5、χ^2_5 は 11 なので 11 を超えるようならサイに問題がある。

3.9 傾向性の検定

傾向がないとする仮説　$H_0 : P_{01} = p_{02} = \cdots = p_{0k}$

傾向があるとする仮説　$H_1 : p_{01} \leq p_{02} \leq \cdots \leq p_{0k}$

観測された比率　$p_i = \dfrac{x_i}{n_i}\ ,\ (i = 1, 2, \cdots, k)$

χ^2値

$$\chi^2 = \sum n_i \frac{(p_1 - p_{0i})^2}{p_{0i}(1 - p_{0i})} = \sum \left(\frac{x_i - n_i p_{0i}}{\sqrt{n_i p_{0i}(1 - p_{0i})}} \right)^2$$

理論比率 $P01$ が分からないとき。$\bar{P}$ で代用する。

$$\chi^2 = \sum n_i \frac{(P_i - \bar{P})^2}{\bar{P}(1-\bar{P})}$$

$$\bar{P} = \frac{P_1 + p_2 + \cdots + p_K}{k}$$

Example

18 才から 24 才の調査人数と不眠症の数は 534 人と 150 人。加齢に従い不眠症は増加するか。

```
(18,25,35,45,55,65,75),.DS5,.DS6,.(DS6%DS5)
18 534 150 0.280899
25 746 250 0.335121
35 784 264 0.336735
45 705 302 0.428369
55 443 238 0.537246
65 299 176 0.588629
75  70  36 0.514286
```

χ^2 の値は大きく 傾向が無いとする仮説 H_0 は棄却され、年齢と関係があると判断される。

```
   DS6 chitr DS5
158.702
```

```
   chitr=:4 : ' (+/y.**:p-q)%q*1-q=:(+/%#)p=:x.%y.'
```

3.10 適合度検定

$$\chi^2 = \sum_{i=1}^{k} n_i \frac{(x_i - np_i)^2}{np_i}$$

textbfExample 男女出生比率 1.06 : 1 で観測結果 男 480 人、女 420 人出生

```
((1.06%2.06), 1%2.06) testgf 480 420
1.26943
```

4 を超えないので差異は認められない。

```
testgf=:4 : ' +/(*:y.-t)%t=:x.*+/y.'
```

3.11　分割表の独立性の検定

新薬の薬候比較 (2 × 2 分割表)
適合度検定 95% の値は約 4

```
 ('';'NEW';'OLD'),. ('CARE';'NOT'), {@> DS7
+---+----+---+
|   |CARE|NOT|
+---+----+---+
|NEW|353 |166|
+---+----+---+
|OLD|304 |104|
+---+----+---+


     2 2 $ 'e';'f';'g';'h'
+-+-+
|e|f|
+-+-+
|g|h|
+-+-+
```

$$\chi^2 = \frac{(e-e_0)^2}{e_0} + \frac{(f-f_0)^2}{f_0} + \frac{(g-g_0)^2}{g_0} + \frac{(h-h_0)^2}{h_0} + = \frac{(e+f+g+h)(eh-fg)^2}{(e+f)(e+g)(f+h)(g+h)}$$

2×2 分割表	chic=:3 : 0 (+/,y.)*(*:-/ . *y.)%*/(+/"1 y.),+/y.)	chic DS7 4.66718
一般の分割表	testc=:3 : 0 p=:,(+/"1 y.)*/+/y. +/(*:p-(*+/),y.)%p*+/,y.)	testc DS7 4.66718

2×3 分割表

```
   testc 2 3 $ 45 23 20 12 20 30
17.3285
```

4　情報量規準とモデル選択

4.1　統計的方法とモデル

4.2　理論モデルと経験モデル

4.3　対数尤度と最大対数尤度

4.4　カルバック・ライブナーの情報量

4.5　正規分布に対する KL 情報量

4.6　真のモデルからの近さ

4.7　情報量規準 AIC

最大対数尤度	$MLL = xlogp + (n-x)log(1-p)$
情報量規準	$AIC(M) = -2 \times MLL(M) + 2 \times k$

4.8　出生性比のモデル選択

3.7 の男女出生比率の例

経験モデル	$MLL = xlog(x/n) + (n-x)log(1-x/n) = xlogx + (n-x)log(n-x) - nlogn$ $AIC(P) = -2 \times MLL + 2 \times 1$
理論モデル	$MLL = xlog(p) + (n-x)log(1-p)$ $AIC(Q) = -2 \times MLL$

```
   ratio=:4 : 0
   A0=:-+:+/((k=:{.y.),--/y.)*^.(x.),1-x.
   A1=:2-+:(+/(*^.)k,n-k)-(*^.)n=:}.y.
   A0=:'theoretical model :',8j2":A0
   A0,:'empirical model   :',8j2":A1
)
```

9 出生数 00 人中男が 480 人出生

```
 0.5  ratio 480 900
theoretical model : 1247.66
empirical model   : 1245.66
```

理論モデル (男女の出生率は同じ) は取ることが出来ない

```
 (1.06%2.06)  ratio 480 900
theoretical model : 1244.93
empirical model   : 1245.66
```

今度は理論モデル (男 1.06) の方が優れている

4.9 薬の治癒効果

	独立モデル			従属モデル		
-	B	B'	Total	B	B'	TOTAL
A	pq	P(1-q)	p	p_{11}	p_{21}	p
A'	(1-p)q	(1-p)(1-q)	1-p	P_{21}	p_{22}	1-p
Total	q	1-q	1	q	1-q	1

	分類基準	(I)	周辺度数
分類基準	a	b	a+b
(II)	c	d	c+d
周辺度数	a+c	b+d	a+b+c+d

独立モデル	p, q に対する最尤推定 P, Q は $P = \frac{a+b}{n},\ Q = \frac{a+c}{n}$	$a + b + c + d = n$

従属モデル	$p_{11}, p_{12}, p_{21}, p_{22}$ に対する最尤推定 $p_{11} = \frac{a}{n}, p_{12} = \frac{b}{n}, p_{21} = \frac{c}{n}, p_{22} = \frac{d}{n}$	

独立モデル	$MLL0 = alogPQ+blogP(1-Q)+clog(1-P)Q+dlog(1-P)(1-Q)$ $AIC(IM) = -2 \times MLL0 + 2 \times 2$
従属モデル	$MLL1 = alog(a/n) + blog(b/n) + clog(c/n) + dlog(d/n)$ $AIC(DM) = -2 \times MLL1 + 2 \times 3$

```
   cont=:3 : 0
   C=:(n=:+/,y.)*(*:-/ .*y.)%*/d=:(+/"1 y.),+/y.
   A2=:4-+:(+/(*^.)d)-+:(*^.)n
   A3=:6-+:(+/(*^.),y.)-(*^.)n
   A2=:'AIC(indep) :',8j2":A2
   A3=:'AIC(dep)   :',8j2":A3
   ('chi-square :',8j2":C),A2,:A3
)

   cont DS7
chi-square :    4.67
AIC(indep) : 2394.25
AIC(dep)   : 2391.55
```

従属モデルの方が AIC の値が小さい。新薬の方が治癒効果が大きい。

4.10 正規分布のモデル選択

(M1)	$\mu_{01} = \mu_{02}$ $\sigma_{01}^2 = \sigma_{02}^2$	$MLL1 = LL(\mu, \sigma) = -\frac{m+n}{2}(1 + log2\pi\sigma^2)$ $AIC(M1) = -2 \times MLL1 + 2 \times 2$
(M2)	$\mu_{01} \neq \mu_{02}$ $\sigma_{01}^2 = sigma_{01}^2$	$MLL2 = LL(\mu_1, mu_2, \sigma) = -\frac{m+n}{2}(1 + log2\pi\sigma^2)$ $AIC(M2) = -2 \times MLL2 + 2 \times 3$

AIC の大小を比較して、小さい値を取るモデルを選択する。

4.11 肥料の効果の比較

4.11.1 例題

3 種の肥料 ABC を 4 つの地域１，２，３，４に与えて収穫量を測定した。肥料や地域差で, 収穫量に差があるか。

	肥料	地域			
DAT	区分	1	2	3	4
DS8	A	25	18	21	24
DS9	B	17	13	16	14
DS10	C	24	20	26	22

```
   DS8 aic1 DS9
49.4855 18.5 17.25   NB. AIC  Mean-left  Mean-right Cov
   DS8 aic2 DS9
41.5785 22 15 5
```

A B はモデル 2 でのサンプルデータとみなせる。

```
   DS8 aic1 DS10
41.6774 22.5 6.5
   DS8 aic2 DS10
43.3637 22 23 6.25
```

MODEL1 の方が AIC が小さいので、平均は違っているとはいえない。

```
   aic1=:4 : 0      NB. Model1
   m=:(+/z)%n=:$z=:(x.),y.
   v=:((+/*:z)%n)-*:m
```

```
   (4+n*1+^.(o.2)*v),m,v
)

aic2=:4 : 0          NB. Model2
   m=:((+/x.),+/y.)%n=:($x.),$y.
   v=:((+/*:(x.),y.)-+/n**:m)%+/n
   (6+(+/n)*1+^.(o.2)*v),m,v
)
```

4.12 分散の値も異なるかも知れない場合

(M3)	$\mu_{01} = \mu_{02}$ $\sigma_{01}^2 \neq \sigma_{02}^2$	$MLL3 = -\frac{m+n}{2}(log2\pi) - \frac{m}{2}log\sigma_1^2 - \frac{n}{2}log2\sigma 2^2 - \frac{m+n}{2}$ $AIC(M3) = -2 \times MLL3 + 2 \times 3 = d(1 + log2\|pi) + mlog\sigma_1^2 + nlog\sigma_2^2 + 2 \times 3$ (d=m+n) 総サンプル数
(M4)	$\mu_{01} \neq \mu_{02}$ $\sigma_{01}^2 \neq \sigma_{02}^2$	$MLL4 = -\frac{m+n}{2}(log2\pi) - \frac{m}{2}log\sigma_1^2 - \frac{n}{2}log2\sigma 2^2 - \frac{m+n}{2}$

```
   aic3=:4 : 0
   m1=.{.m=.x.(,.&((+/%#)@(],.*:@])))y.
   A=.+/(1+|.a=.(]%+/)n=.x.,&$y.)*m1
   B=.(+:*/m1)++/(|.a)*m2=.{:m
   (C=.+/a*m1*|.m2);v=.m2-*:m1
   m=.((-C),B,(-A),1)pnm(+/m1*|.v)%+/v
   (6++/n*1+^.v*o.2),m,v=.v+*:m1-m
)

   aic4=:4 : 0
   m=.((+/x.),+/y.) % n=.($x.),$y.
   v=.(((+/*:x.),+/*:y.)%n)-*:m
```

```
   (8+(+/n)*1+^.(o.2)**/v),m,v
)

   DS8 aic3 DS9
48.3428 15.3354 51.9174 2.61247

   DS8 aic4 DS9
54.1526 22 15 7.5 2.5
```

5　分散分析

5.1　測定値ばらつきの分解

メーカー A,B,C が各 1000 個入りの箱 4 箱抽出し不良品を数えたら次のとおりであった。

会社	不良	品	の	個数
A	12	8	10	6
B	14	10	6	6
C	4	6	4	10

5.1.1　J

```
    decomp1 datav
+-----------+-+-----------+-+-----------+
|12  8 10  6|=|8 8 8 8    |+| 4  0  2 _2|
|14 10  6  6| |8 8 8 8    | | 6  2 _2 _2|
| 4  6  4 10| |8 8 8 8    | |_4 _2 _4  2|
+-----------+-+-----------+-+-----------+
```

（原データ）（ばらつかない部分）（ばらつく部分）

```
+-----------+-+-----------+-+-----------+
| 4  0  2 _2|=| 1  1  1  1|+| 3 _1  1 _3|
| 6  2 _2 _2| | 1  1  1  1| | 5  1 _3 _3|
|_4 _2 _4  2| |_2 _2 _2 _2| |_2  0 _2  4|
+-----------+-+-----------+-+-----------+
```

（ばらつく部分）（メーカーのばらつき）（変動誤差）

```
   av1 datav
24 2 12
88 9 9.77778
 F =1.23
```

要因	偏差平方和	自由度	分散	分散比
級間変動	24	2	12	F=1.23
級内 (誤差) 変動	88	9	9.77778	

F 値は分布表の $F_{(}0.05) = 4.26$ (自由度　2　9）より小さいので メーカー二より不良品率が異なるとはいえない。

5.2　分散比と分散分析

5.3　2 要因の分散分析

5.3.1　例題

3 種の肥料 ABC を 4 つの地域１，２，３，４に与えて収穫量を測定した。肥料や地域差で, 収穫量に差があるか。

肥料　地域

	1	2	3	4
A	25	18	21	24
B	17	13	16	14
C	24	20	26	22

5.3.2　J

```
   decomp2 datav2
+-----------+-+-----------+-+-----------+
|25 18 21 24|=|20 20 20 20|+| 5 _2  1  4|
|17 13 16 14| |20 20 20 20| |_3 _7 _4 _6|
|24 20 26 22| |20 20 20 20| | 4  0  6  2|
+-----------+-+-----------+-+-----------+


+-----------+-+-----------+-+-----------+
| 2  2  2  2|+|2 _3 1 0   |+| 1 _1 _2  2|
|_5 _5 _5 _5| |2 _3 1 0   | | 0  1  0 _1|
| 3  3  3  3| |2 _3 1 0   | |_1  0  2 _1|
+-----------+-+-----------+-+-----------+


   av2 datav2
```

5.3.3 結果と解説

２要因分散分析表

	偏差２乗和	自由度	分散	分散比
肥料間	152	2	76	F=25.3333
地域間	42	3	14	F=4.66667
残差	18	6	3	

自由度（２，６）の $F_{0.005} = 14.55$, 自由度（３，６）の $F_{0.05} = 4.76$

肥料間の有意差は大きいが地域間の差は認められない。

5.4 １要因モデルの AIC 分散分析

AIC A Information Criterion, Akaike’s Information Criterion

AIC の値は小さいほうを採用する。

5.5 AIC 一要因分析

モデル１　要因の違いで観測地に差はない。

モデル２　要因の差が認められる。

5.5.1 肥料に関する要因分析

```
    ava0 datav2
72.5147 17.6667 20

   ava1 datav2
61.3678 5 20 2 _5 3
```

Model	AIC	σ^2	μ	a_i
Model 1	72.5147	17.6667	20	
Model 2	61.3678	5	20	2 _5 3

Model 2 の AIC が小さく, 要因の差が認められる。

5.5.2 地域に関する要因分析

データを転置して地域データとする。

```
   ava1 |: datav2
72.5147 17.6667 20


   ava1 |: datav2
75.8652 14.1667 20 2 _3 1 0
```

Model	AIC	σ^2	μ	a_i
Model 1	72.5147	17.6667	20	
Model 2	75.8652	14.1667	20	2 _3 1 0

Model2 の AIC の値のほうが大きく Model1 を採用すべきで地域差は認められない。

5.6 AIC 2 要因分析

```
    ava datav2
MODEL M1 : 72.51 17.67 20.00
MODEL M2 : 61.37 5.00 20.00 2.00 _5.00 3.00
MODEL M3 : 75.87 14.17 20.00 2.00 _3.00 1.00 0.00
MODEL M4 : 52.92 1.50 20.00 2.00 _5.00 3.00 2.00 _3.00 1.00 0.00
```

Model	AIC	σ^2	μ	$\alpha_1(\beta_i)$
MODEL M1	72.51	17.67	20.00	
MODEL M2	61.37	5.00	20.00	2.00 _5.00 3.00
MODEL M3	75.87	14.17	20.00	2.00 _3.00 1.00 0.00
MODEL M4	52.92	1.50	20.00	2.00 _5.00 3.00 2.00 _3.00 1.00 0.00

model4 で AIC が最小になっている。一要因分析と異なり肥料に地域要因を加味したモデルが採用される。

6　回帰分析

6.1　回帰分析とは

線形関数	$y = a_0 + b_0 x + \epsilon$	
誤差項	ϵ	線形関係では説明しきれない諸々の変動部分
回帰係数	$b = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum y_i^2)}$ $a = \frac{1}{n}((\sum y_i) - b(\sum x_i))$	
推定したモデル	$y = a + bx$	
推計値	$z_i = a + bx_i$	
残差	$e_i = y_i - z_i$	
残差平方和	$Q = \sum e_i^2 = \sum (y_i - z_i)^2$	
残差分散	$V(e) = \frac{Q}{n} = \frac{1}{n} \sum e_i^2$	
相関係数	$R = \frac{\sum (x_i - M(x))(y_i - M(y))}{\sqrt{\sum (x_i - M(x))^2) \sum (y_i - M(y))^2}}$	
決定係数	$R^2 = \frac{V(z)}{V(y)} = 1 - \frac{V(e)}{V(y)}$ $R = \frac{(\sum (x_i - M(x))(y_i - M(y)))^2}{\sum (x_i - M(x))^2) \sum (y_i - M(y))^2}$	
重相関係数	$\sqrt{R^2}$	決定係数の正の平方根

6.2 重回帰モデル

k 個の説明変数	$(x_1, x_2, \cdots, x_k)$	
回帰モデル	$y = b_{00} + b_{01}x_1 + b_{02}x_2 + \cdots + b_{0k}x_k + \epsilon$	
DATA	$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_k \end{bmatrix}$	
	$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{24} & \cdots & x_{k2} \\ 1 & x_{13} & x_{23} & \cdots & x_{k3} \\ \cdots & \cdots & \cdots & \cdots\cdots & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$	
推計モデル	$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$	
正規方程式	$b' = (y'X)(X'X)^{-1}$	
残差平方和	$Q = (y - Xb)'(y - Xb) = yy' - (y'X)b$	
最大対数尤度	$MLL = -\frac{n}{2}log\frac{Q}{n}$	
情報量規準	$AIC = nlog\frac{Q}{n} + 2 \times (k + 1)$	

6.3 重回帰モデル

6.4 職員数と収入保険料

```
   add i.5
1 0
1 1
1 2
1 3
1 4
```

 (i.15),. HOKEN,.GAIMU,.NAIMU

A (保険会社) B 保険料収入 (百億円) C(外務員数) D(内務員数) 各百人

A	B	C	D
0	332	759	137
1	209	568	113
2	208	612	106
3	128	385	81
4	106	335	81
5	85	257	54
6	79	160	67
7	64	133	34
8	52	77	38
9	42	134	24
10	40	100	29
11	40	111	30
12	35	47	22
13	27	93	28
14	15	34	20

```
   HOKEN regb GAIMU
1.77662 0.377228

   HOKEN ssr GAIMU
4960.61

   HOKEN rega GAIMU
91.0185

   var GAIMU
0.154088 1.35859e_6
```

```
   HOKEN regt GAIMU
0.231693 16.5677

   HOKEN rega NAIMU
99.1186

   HOKEN rega GAIMU+NAIMU
90.1853

   HOKEN rega GAIMU,:NAIMU
92.0601

   HOKEN reg_all GAIMU+NAIMU
res. variance : 312.839
co. of det.(%):  95.722
value of AIC  :  90.185
reg-coeff. and t-values:
    _3.797     0.325
    _0.493    17.056


   HOKEN reg GAIMU+NAIMU
_3.79653 0.325326

   HOKEN reg0 GAIMU+NAIMU
318.687 88.4631 0.318078 26.9186


 plot {|: 1 0 {"1 a
  pd 'eps temp\hoken_1.eps'
 plot {|: 2 0 {"1 a
  pd 'eps temp\hoken_2.eps'
```

6.5　多項式回帰モデル

多項式	$y = c_{00} + c_{01}t + \cdots + c_{0k}t^k + \epsilon$ $\begin{bmatrix} S_0 & S_1 & \cdots & S_k \\ S_1 & S_2 & \cdots & S_{k+1} \\ \cdots & \cdots & \cdots & \cdots \\ S_k & S_{k+1} & \cdots & S_{2k} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \cdots \\ c_k \end{bmatrix} = \begin{bmatrix} T_0 \\ T_1 \\ \cdots \\ T_k \end{bmatrix}$ $S_i = \sum_{t=1}^{n} t_t^i$ $T_i = \sum_{t=1}^{n} t_t^i \cdot y_i$	
残差平方和	$Q_k = \sum y^2 - (T_0 c_0 + c_1 T_1 + \cdots + c_k T_k)$	
対数最大尤度	$MLL = -\frac{n}{2} log \frac{Q_k}{n}$	
情報量規準	$AIC = 2 \times MLL + 2 \times (k+1)$	

```
   ]Y1 =. 1 dep T
172.6 2360.9
   ]Y2=. 2 dep T
172.6 2360.9 42788.1
   ] X1=. 1 indep T
 25  325
325 5525
   ]X2=. 2 indep T
  25    325      5525
 325   5525    105625
5525 105625 2.15365e6
   Y1 %. X1
5.733 0.0900769
   1 pregb T
5.733 0.0900769
   2 pregb T
9.20887 _0.682339 0.0297083
   T reg_all Z1
res. variance :   9.718
```

```
co. of det.(%):   4.161
value of AIC  :  60.849
reg-coeff. and t-values:
     5.733     0.090
     4.278     0.999


   2 preg T
co. of det.(%):  22.90
value of AIC  :  57.41
value of MLE  :   9.21  _0.68   0.03
```

6.6 自己回帰モデル

DATA 自己回帰式	$(x_1, x_2, \cdots, +x_n)$ $t_i = b_{k1}t_{i-1} + b_{k2}t_{i-2} + \cdots + b_{kk}t_{i-k} + \epsilon$ $i = k+1, k+2, \cdots, n$	
平均を求め X_i を変換する	$M)x) = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$ $t_i = x_i - M(x),\ \ i = 1, 2, \cdots, n$ $\begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1k} \\ S_{12} & S_{22} & \cdots & S_{2k+1} \\ \cdots & \cdots & \cdots & \cdots \\ S_{1k} & S_{1k} & \cdots & S_{kk} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_k \end{bmatrix} = \begin{bmatrix} T_1 \\ T_2 \\ \cdots \\ T_k \end{bmatrix}$ $\mathrm{S}_{rs} = \sum_{i=k+1}^{n} = t_{i-r} \times t_{i-s}$ $\mathrm{T}_k = \sum_{i=t+1}^{n} = t_i \cdot t_{i-r}$	

残差平方和	$Q_k = T_0 - (b_1T_1 + b_2T_2 + \cdots + b_kT_k)$	
対数最大尤度	$MLL = -\frac{n-k}{n} log \frac{Q_k}{n-k}$	
情報量規準	$AIC = 2 \times MLL + 2 \times k)$	
	$= (n-k) \times log \frac{Q_k}{n-k} + 2 \times k$	
予測値	$t_{n+1} = b_1t_n + b_2t_{n-1} + \cdots + b_[kt_{n-k+1}$	
	$t_{n+2} = b_1t_{n+1} + b_2t_n + \cdots + b_[kt_{n-k+2}$	

```
   2 adep i.5
0 1 2
   2 aindep i.5
_1 _2
 0 _1
 1  0
   2 aregb T
0.91021 _0.388917
   2 areg T
6.904 0.504805 41.6 0.91021 _0.388917
```

7　主成分分析

7.1　大きいことはイイことだ

条件 $x^2 + y^2 = 1$ のもとで $Q(x, y) = ax^2 + 2bxy + cy^2$ を最大 (または最小) にする。

$Q = [x\ y] = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$ を

$Q = [u\ v] = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$ に変換する

7.2　2 変数の場合の最大・最小問題

7.3　ラグランジュの未定係数法

7.4　2 次形式

7.5　固有値と固有ベクトル

7.6　直交行列

7.7　最大 · 最小化と固有値問題

7.8　主成分分析とは

7.9　美女のプロポーション

多変量解析重相関については, 回帰の項参照

多次元、多変量のデータを縮約し、整理すると新しい視点で物事が見られる。

数値データの整理では、重相関分析、判別分析、主成分分析などがある。カテゴリーデータ（yes,maybe,no) も規準化すれば数値データと同様に取り扱え、数量化理論が適用できる。

数量化理論には、I II III IV 類などがある。

今回は、主成分分析を実行してみよう。

ミスユニバース日本代表の５サイズ（NO2-NO25 回）

```
   pp
165 53 86 56 92
160 47 84 52 92
166 55 86 64 89
164 56 90 60 95
168 55 87 56 87
164 54 87 57 92
168 54 94 58 97
169 55 88 57 92
169 53 86 58 93
166 56 84 57 90
165 53 85 55 90
163 49 84 59 90
164 52 87 58 90
167 53 86 59 88
169 58 89 60 90
169 51 84 60 90
166 50 86 59 87
168 53 88 60 88
165 54 88 62 90
167 50 88 58 89
170 55 88 60 90
168 57 84 62 92
168 56 85 62 94
163 52 83 60 88
```

単位が異なるので、相関行列を求める。
分散共分散行列から求める方法もある。

```
   corv pp
        1 0.542494  0.326565   0.372425  0.0115011
 0.542494         1    0.3131   0.449056   0.241926
```

```
 0.326565   0.3131         1  0.0605705   0.424285
 0.372425 0.449056 0.0605705          1 _0.0649716
0.0115011 0.241926  0.424285 _0.0649716          1
```

５サイズの相関行列　相関係数はあまり高くない。
相関行列の固有値と固有ベクトルを求める。
上段が固有値。固有値とは、マトリクスの比例定数である。
下段は、縦に、それぞれの固有値に即応する固有ベクトル。

```
   > dgeev_jlapack_ corm pp

2.14898 1.30735 0.675427 0.509665 0.358586

0.521033 _0.236935 _0.534739   0.27597  0.557029
0.570928 _0.104455  0.198549  0.481168 _0.626247
0.419977  0.483214 _0.401878 _0.591852 _0.279874
0.399578 _0.488062  0.513792 _0.551293  0.185005
0.25792  0.679155  0.499144  0.195203   0.43009
```

もとの５サイズデータを基準化する。規準化とは、平均０分散１のデータに変換することである。

```
   stand pp
 _0.53298 _0.145407  _0.22758  _1.06586  0.566029
 _2.59613  _2.47193  _1.06788  _2.64005  0.566029
 _0.12035  0.630099  _0.22758   2.08253 _0.668943
_0.945609   1.01785   1.45301  0.508333     1.801
 0.704909  0.630099  0.192568  _1.06586  _1.49226
_0.945609  0.242346  0.192568 _0.672312  0.566029
 0.704909  0.242346   3.13361 _0.278763   2.62431
  1.11754  0.630099  0.612716 _0.672312  0.566029
  1.11754 _0.145407  _0.22758 _0.278763  0.977686
 _0.12035   1.01785  _1.06788 _0.672312 _0.257286
```

```
 _0.53298 _0.145407 _0.647729  _1.45941 _0.257286
 _1.35824  _1.69642  _1.06788  0.114785 _0.257286
_0.945609  _0.53316  0.192568 _0.278763 _0.257286
 0.292279 _0.145407  _0.22758  0.114785   _1.0806
  1.11754   1.79336   1.03286  0.508333 _0.257286
  1.11754 _0.920913  _1.06788  0.508333 _0.257286
 _0.12035  _1.30867  _0.22758  0.114785  _1.49226
 0.704909 _0.145407  0.612716  0.508333   _1.0806
 _0.53298  0.242346  0.612716   1.29543 _0.257286
 0.292279  _1.30867  0.612716 _0.278763 _0.668943
  1.53017  0.630099  0.612716  0.508333 _0.257286
 0.704909    1.4056  _1.06788   1.29543  0.566029
 0.704909   1.01785 _0.647729   1.29543   1.38934
 _1.35824  _0.53316  _1.48803  0.508333   _1.0806
```

基準化データに固有ベクトルを内積演算すると、各年次代表の左から、縦に第一主成分得点、第２主成分得点、、が得られる。

```
      (stand pp)  +/ . * v
    _0.7362  0.936127 0.0824939  0.615734 _0.0958763
   _4.12136   2.03024  0.252701  0.292094   0.155811
   0.861054  _1.61799   1.01701 _0.873998  _0.300371
    1.36629    1.7949   1.28395 _0.559852  _0.702183
_0.00288101  _0.63305  _1.62171  0.680054  _0.894831
  _0.396108   1.00434  0.413484  0.222809  _0.613332
    2.38716   3.24024 _0.421466 _0.877535    0.44099
    1.07669  0.678023 _0.781621  0.730085   0.175482
   0.544456  0.440489 _0.190222  0.717664    1.14618
   _0.26507 _0.440424  0.221754   1.40899  _0.640632
   _1.28225  0.366024 _0.361812  0.920647  _0.405195
    _2.1452 _0.247758   0.74919 _0.672574   0.515251
  _0.893963  0.334108 0.0507581 _0.528014  _0.408964
  _0.259151 _0.953949 _0.574104 _0.128828  _0.125957
```

```
  2.17669  _0.37585 _0.523853  0.229547  _0.806268
_0.255225  _1.10744 _0.218527  0.166855    1.48148
 _1.24446  _1.01425 _0.789895 _0.882781   0.195636
 0.466001 _0.837749 _0.930248 _0.729246 _0.0584787
 0.569255 _0.409947  0.624044   _1.1575  _0.491132
_0.621462 0.0452568  _1.13949 _0.888565   0.471596
   1.5511 _0.555129 _0.806617 0.0323629   0.269655
  1.38491  _1.07768   1.27941   0.89922   0.294372
  1.55233    _0.275   1.44452  0.624694   0.773712
 _1.71261  _1.32352  0.940251 _0.241859  _0.376939
```

第一主成分は、（縦に見る）総じてプラスであるが、身長、体重のウエイトが大きいので、”Total　Volume” と解釈できる。

第二種成分は、Ｂ　Ｈにウエイトがあるので、「　　」と解釈できる。

第一主成分の固有値と、第二主成分の固有値で、3.45/5.0 で全体の７０％は説明できる。

```
2.14898 1.30735 0.675427 0.509665 0.358586
```

第一主成分　第二主成分

```
0.521033 _0.236935 _0.534739   0.27597  0.557029
0.570928 _0.104455  0.198549  0.481168 _0.626247
0.419977  0.483214 _0.401878 _0.591852 _0.279874
0.399578 _0.488062  0.513792 _0.551293  0.185005
0.25792  0.679155  0.499144  0.195203   0.43009
```

0mm