Ｒを用いた生物統計解析入門

2006 年 5 月 26 日
日本計量生物学会
チュートリアルセミナー
（於・国立保健医療科学院）

群馬大学大学院医学系研究科
社会環境医療学講座　生態情報学
中澤　港
<nminato@med.gunma-u.ac.jp>

本日の目的

- 聴衆としては，SAS ユーザを想定
- 既に SAS で統計処理をする方法や，統計の理論はわかっている方が対象。したがって，統計の詳しい説明はしない。
- R の基本的な扱い方を説明する
- SAS でできることを R ではどのように実行するか，SAS と R の思想の違いを説明する

nminato@med.gunma-u.ac.jp

Rの起動と終了

- Windows では，アイコンをダブルクリックすると起動する。作業ディレクトリの .Rprofile が実行され，保存された作業環境 .RData が読まれる。
  - 作業ディレクトリの設定：起動アイコンのプロパティの「作業フォルダ（S）」に作業ディレクトリを指定する。環境変数 R_USER も同じ作業ディレクトリに指定するとよい（システムの環境変数または作業ディレクトリに .Renviron を置き，R_USER=”c:/work” などと書いておくと，それが優先される）
  - proxy 設定：起動アイコンのプロパティで，「起動コマンドのリンク先」末尾に --internet2 と付す
- 終了はプロンプトに対して q() と打つ

nminato@med.gunma-u.ac.jp

R のライブラリ

- 世界中に CRAN ミラー（日本では会津大と筑波大）。筑波大ミラーを規定値にするには，作業ディレクトリに .Rprofile テキストファイルを作り，options(repos="http://cran.md.tsukuba.ac.jp/")
- R バイナリに built-in なライブラリは，base, datasets, grDevices, graphics, grid, methods, splines, stats, stats4, tcltk, tools, utils（Windows 版は下記も）
- Recommended（将来全バイナリに built-in）が KernSmooth, MASS, boot, class, cluster, foreign, lattice, mgcv, nlme, nnet, rpart, spatial, survival
- search() でロード済み一覧，.packages(all.avail=T) でインストール済み一覧が表示される

nminato@med.gunma-u.ac.jp

Rはオブジェクト指向

- すべてがオブジェクト
- シンボルにオブジェクトを付値できる(シンボル，即ち変数名は，だいたい自由に付けられる。宣言，型定義も不要。規定の関数名さえオーバーライドできるものが多い。例外：NA への付値)
  - x <- 2; y <- function(a) { x <- x+a; x }
  - z <- function(a) { x <<- x+a }
  - y(5) と z(5) はどちらも 7 を返す。y(5) は x の値を変えないが，z(5) は x の値を変えてしまう
- 関数定義の中での変数はローカルなスコープをもつ(関数外には影響しない)

nminato@med.gunma-u.ac.jp

SAS のコードを変換するために

- 言語仕様上の違い
  - R では，大文字小文字が区別される（C や Lisp 似）
  - すべてが関数（Lisp と同じ）。SAS でいう DATA ステップと PROC ステップの区別はない
- データの読み込み
  - INFILE / INPUT に比べ，多様な読み込み関数
  - 高レベル：read.delim, read.csv, read.table など
  - 低レベル：scan
  - 特殊：foreign ライブラリの read.xport など
  - DBMS との連携：RODBC ライブラリが便利

nminato@med.gunma-u.ac.jp

RODBC ライブラリの利用例

- install.packages("RODBC")しておけば，DBMS から sqlQuery() で SQL 文でデータを読める
  - 例えば Excel のファイルをデータベースとして使う場合，作業ディレクトリに sample.xls があるとして，
    library(RODBC)
    conn <- odbcConnectExcel("./sample.xls")
    dat <- sqlFetch(conn,"Sheet1")
    odbcClose(conn)
    とすると，dat に sample.xls の Sheet1 の中身がデータフレームとして読み取れる（値のみ）。
  - Excel 以外には，odbcConnectAccess（MS Access），odbcConnectDbase（Dbase），odbcConnect（MySQL, PostgreSQL, Oracle）が接続可能。

nminato@med.gunma-u.ac.jp

SAS のデータを利用する方法

- SAS では，DATA ステップでデータを生成する際に，CARDS; で直接書いたり，INFILE でスペース区切りテキストファイルから読んだりした
- INFILE で読むデータの移行のためには
  - データを変換して read.delim() で読む
    - 元が表ならそこに戻ってタブ区切りテキスト化
    - スペース区切りテキストをエディタで開いて，欠損値である半角ピリオドを NA に置換し，スペースをタブコードに置換し，文字列として許されない文字は削除し，変数名を 1 行目に入れる
  - read.table() や scan() のオプションでそのまま読む

nminato@med.gunma-u.ac.jp

そのまま読み込む例(1)

```
DATA GIDRA;
INFILE 'ALLINFO.TXT' LRECL=300;
INPUT SAMPLE ID NAME $ SEX LP VIL
PLACE HEIGHT WEIGHT ARMC LC ASS BSS
LSS SBP DBP HR STIME HML PCV HB MCHC
GGTP ALP GLU BUN GOT GPT LDH TCPK
TG CHOL TPRO ALB FERRIT RW HDLCHOL
APOAI NA K SE CU FEICP FEBASO ZN AL CA
MG P SR S APF APV TF RETINOL ATOCOP
ACAROT BCAROT DUMMY;
IF TF=. THEN TIBC=.;
ELSE TIBC=1.4*(TF/76500)*55.8*10;
IF FEBASO=. THEN FEST=.;
ELSE FEST=FEBASO/100;
IF TIBC=. THEN TFSAT=.;
ELSE TFSAT=FEST/TIBC*100;
SELECT;
  WHEN (1<=VIL<=2) VGP=1;
  WHEN (3<=VIL<=8) VGP=2;
  WHEN (9<=VIL<=12) VGP=3;
  WHEN (VIL=13) VGP=4;
  OTHERWISE VGP=.;
END;
RUN;
```

```
gidra <-read.table("./allinfo.txt",header=F,
sep="¥x20",quote="",col.names=list(
"SAMPLE","ID","NAME","SEX","LP","VIL",
"PLACE","HEIGHT","WEIGHT","ARMC","LC",
"ASS","BSS","LSS","SBP","DBP","HR",
"STIME","HML","PCV","HB","MCHC","GGTP",
"ALP","GLU","BUN","GOT","GPT","LDH",
"TCPK","TG","CHOL","TPRO","ALB","FERRIT",
"RW","HDLCHOL","APOAI","NAT","K","SE",
"CU","FEICP","FEBASO","ZN","AL","CA","MG",
"P","SR","S","APF","APV","TF","RETINOL",
"ATOCOP","ACAROT","BCAROT","dummy"),
na.strings="¥.")

TIBC <- ifelse(is.na(gidra$TF), NA,
    1.4*(gidra$TF/76500)*55.8*10)
FEST <- ifelse(is.na(gidra$FEBASO), NA,
    gidra$FEBASO/100)
TFSAT <- ifelse(is.na(TIBC), NA, FEST/TIBC*100)
VGP <- cut(gidra$VIL,c(0,2,8,12,13)) # 因子型
gidra <-
    data.frame(gidra,TIBC=TIBC,FEST=FEST,TFSA
    T=TFSAT,VGP=VGP)
rm(TIBC); rm(FEST); rm(TFSAT); rm(VGP)
```


nminato@med.gunma-u.ac.jp

そのまま読み込む例(2)

```
DATA GIDRA;
INFILE 'ALLINFO.TXT' LRECL=300;
INPUT SAMPLE ID NAME $ SEX LP VIL
PLACE HEIGHT WEIGHT ARMC LC ASS BSS
LSS SBP DBP HR STIME HML PCV HB MCHC
GGTP ALP GLU BUN GOT GPT LDH TCPK
TG CHOL TPRO ALB FERRIT RW HDLCHOL
APOAI NA K SE CU FEICP FEBASO ZN AL CA
MG P SR S APF APV TF RETINOL ATOCOP
ACAROT BCAROT DUMMY;
IF TF=. THEN TIBC=.;
ELSE TIBC=1.4*(TF/76500)*55.8*10;
IF FEBASO=. THEN FEST=.;
ELSE FEST=FEBASO/100;
IF TIBC=. THEN TFSAT=.;
ELSE TFSAT=FEST/TIBC*100;
SELECT;
  WHEN (1<=VIL<=2) VGP=1;
  WHEN (3<=VIL<=8) VGP=2;
  WHEN (9<=VIL<=12) VGP=3;
  WHEN (VIL=13) VGP=4;
  OTHERWISE VGP=.;
END;
RUN;
```

```
gidra <- scan("./allinfo.txt",flush=T,what=list(
SAMPLE=0,ID=0,NAME="",SEX=0,LP=0,VIL=0,
PLACE=0,HEIGHT=0,WEIGHT=0,ARMC=0,LC=0,
ASS=0,BSS=0,LSS=0,SBP=0,DBP=0,HR=0,
STIME=0,HML=0,PCV=0,HB=0,MCHC=0,GGTP=0,
ALP=0,GLU=0,BUN=0,GOT=0,GPT=0,LDH=0,
TCPK=0,TG=0,CHOL=0,TPRO=0,ALB=0,FERRIT=0,
RW=0,HDLCHOL=0,APOAI=0,NAT=0,K=0,SE=0,
CU=0,FEICP=0,FEBASO=0,ZN=0,AL=0,CA=0,MG=0,
P=0,SR=0,S=0,APF=0,APV=0,TF=0,RETINOL=0,
ATOCOP=0,ACAROT=0,BCAROT=0,dummy=0),
sep="",na.strings=".",quote="")

TIBC <- ifelse(is.na(gidra$TF), NA,
    1.4*(gidra$TF/76500)*55.8*10)
FEST <- ifelse(is.na(gidra$FEBASO), NA,
    gidra$FEBASO/100)
TFSAT <- ifelse(is.na(TIBC), NA, FEST/TIBC*100)
VGP <- cut(gidra$VIL,c(0,2,8,12,13)) #  因子型
gidra <-
    data.frame(gidra,TIBC=TIBC,FEST=FEST,TFSA
    T=TFSAT,VGP=VGP)
rm(TIBC); rm(FEST); rm(TFSAT); rm(VGP)
```


nminato@med.gunma-u.ac.jp

グラフィックは簡単！

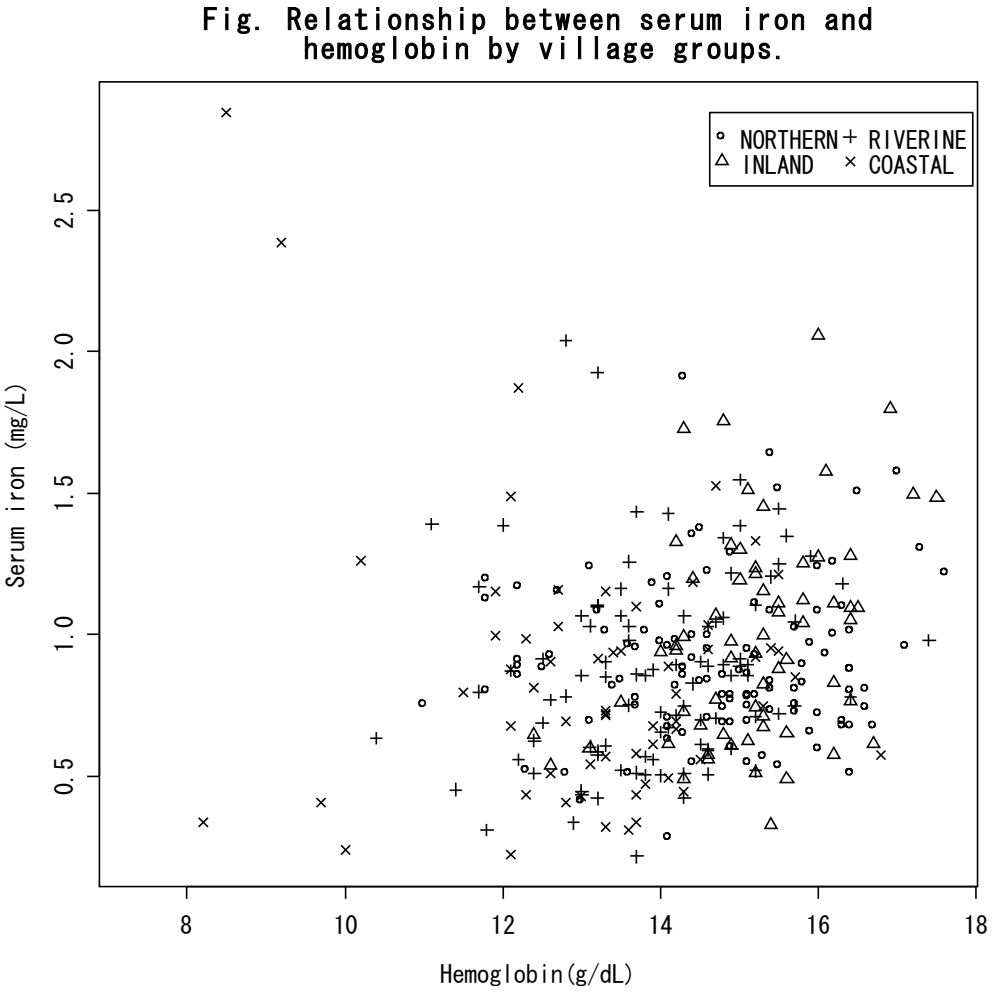
```
GOPTIONS DEVICE=LIPS3A4
GACCESS='SASGASTD>D:¥WORK¥LIPS3.BIN'
        HBY=0.8 CM CBY=BLACK FBY=SWISS;
AXIS1 LABEL=(F=SWISS A=90 R=0)
     LENGTH=14 CM OFFSET=(1 CM,1 CM)
      ORIGIN=(4 CM,4 CM) WIDTH=4;
AXIS2 LABEL=(F=SWISS A=0 R=0) LENGTH=14
     CM OFFSET=(1 CM,1 CM)
      ORIGIN=(4 CM,4 CM) WIDTH=4;
SYMBOL1 C=BLACK V=CIRCLE R=1;
SYMBOL2 C=BLACK V=DOT R=1;
SYMBOL3 C=BLACK V=HASH R=1;
SYMBOL4 C=BLACK V=STAR R=1;
SYMBOL5 C=BLACK V=SQUARE R=1;
PROC FORMAT;
VALUE VNM 1='NORTHN' 2='INLAND'
     3='RIVERN' 4='COASTL';
RUN;
PROC GPLOT;
 WHERE (1<=VGP<=4);
 FORMAT VGP VNM.;
 PLOT HB*FEST=VGP /HAXIS=AXIS2
     VAXIS=AXIS1;
RUN;
```

```
levels(VGP) <-
    c("NORTHERN","INLAND","RIVE
    RINE","COASTAL")
attach(gidra)
win.metafile("./hbfest.emf")
plot(HB,FEST,pch=as.integer(VGP),xla
    b="Hemoglobin(g/dL)",ylab="Seru
    m iron (mg/L)")
# legend の場所は対話的に
    legend(locator(1),...) の方がいい。
legend(mean(HB,na.rm=T),max(FEST,n
    a.rm=T),pch=1:4,legend=levels(VGP
    ),ncol=2)
title("Fig. Relationship between serum
    iron and¥n hemoglobin by village
    groups.")
dev.off()
detach(gidra)
```


nminato@med.gunma-u.ac.jp

グラフィック出力

- win.metafile("./hbfest.emf") でできた Windows メタファイルが右図。これは PowerPoint や OpenOffice.org の Draw や Impress で編集可能。
- 他に，postscript() や png() や pdf() が使える。



nminato@med.gunma-u.ac.jp

分布の正規性をチェックする

```
PROC UNIVARIATE
  NORMAL PLOT;
 VAR FEST TFSAT;
RUN;
```

```
library(Hmisc)
describe.data.frame(data.fram
  e(FEST=FEST,TFSAT=TFSAT))
detach(package:Hmisc)
summary(data.frame(FEST=FEST,
  TFSAT=TFSAT))
tapply(FEST,VGP,summary)
tapply(TFSAT,VGP,summary)
shapiro.test(FEST)
shapiro.test(TFSAT)
win.metafile("./normality.emf
  ")
layout(matrix(c(1,2),nrow=2))
qqnorm(FEST,main="Fig.
  Normality of serum iron
  (I/N/A/C/G method).")
qqnorm(TFSAT,main="Fig.
  Normality of transferrin
  saturation.")
dev.off()
```


nminato@med.gunma-u.ac.jp

UNIVARIATEに対応する出力

```
> describe.data.frame(data.frame(FEST=FEST,TFSAT=TFSAT))
data.frame(FEST = FEST, TFSAT = TFSAT)
 2  Variables      727  Observations
--------------------------------------------------------------------------
FEST
      n missing  unique    Mean     .05     .10     .25     .50     .75     .90
    336     391     289  0.8992  0.4295  0.5080  0.6710  0.8555  1.0860  1.3210
    .95
 1.5055
lowest : 0.214 0.221 0.236 0.280 0.307, highest: 1.926 2.039 2.059 2.387 2.849
--------------------------------------------------------------------------
TFSAT
      n missing  unique    Mean     .05     .10     .25     .50     .75     .90
    213     514     211   29.76   10.35   14.78   20.70   28.55   35.56   44.95
    .95
  55.67
lowest :   7.806   8.010   8.316   8.702   8.765
highest:  63.300  67.499  67.558  71.782 129.163
--------------------------------------------------------------------------
> summary(data.frame(FEST=FEST,TFSAT=TFSAT))
      FEST              TFSAT
 Min.   :  0.2140   Min.   :  7.806
 1st Qu.:  0.6710   1st Qu.: 20.703
 Median :  0.8555   Median : 28.551
 Mean   :  0.8992   Mean   : 29.755
 3rd Qu.:  1.0860   3rd Qu.: 35.559
 Max.   :  2.8490   Max.   :129.163
 NA's   :391.0000   NA's   :514.000
```


nminato@med.gunma-u.ac.jp

正規性の検定

```
> shapiro.test(FEST)
Shapiro-Wilk normality
  test

data:  FEST
W = 0.9375, p-value =
  1.103e-10

> shapiro.test(TFSAT)
Shapiro-Wilk normality
  test

data:  TFSAT
W = 0.8741, p-value =
  2.673e-12
```

Fig. Normality of serum iron (I/N/A/C/G method).

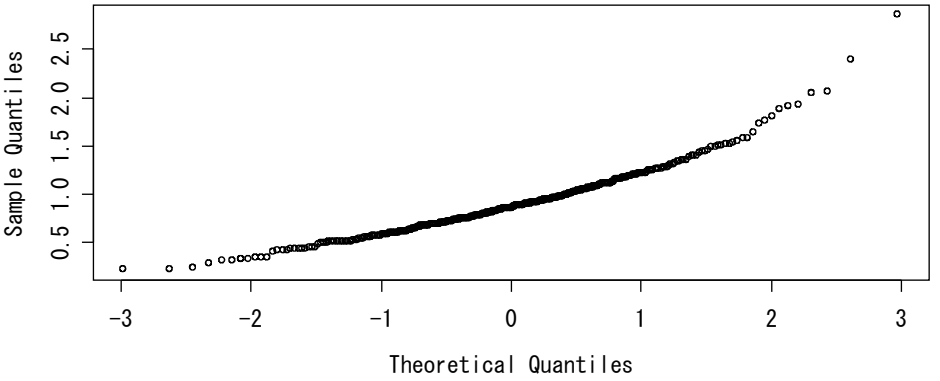
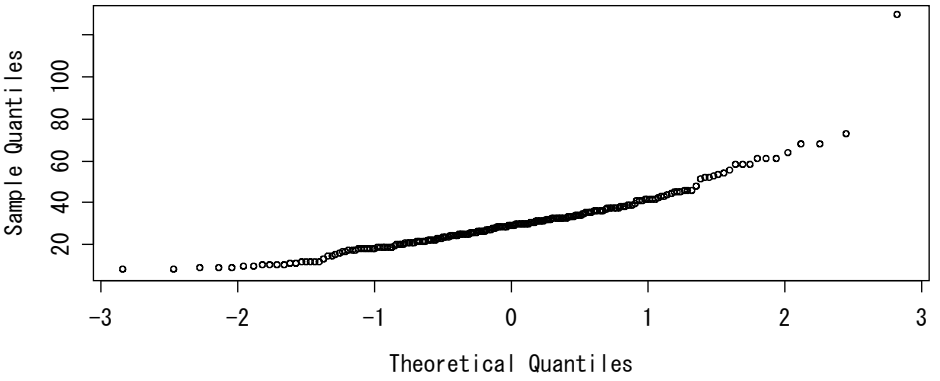


Fig. Normality of transferrin saturation.



nminato@med.gunma-u.ac.jp

分散分析

```
PROC GLM;
 WHERE HML IS NOT
  MISSING;
 CLASS HML;
 MODEL FEST=HML/SS2
  SOLUTION;
 MEANS HML/TUKEY LINES;
RUN;
```

```
gidrawhml <-
  subset(gidra,!is.na(HM
  L),drop=T)
attach(gidrawhml)
library(car)
stbyhml <- aov(FEST ~
  as.factor(HML))
summary(stbyhml)
TukeyHSD(stbyhml)
pairwise.t.test(FEST,as.
  factor(HML),method="ho
  lm")
Anova(lm(FEST ~
  as.factor(HML)))
detach(gidrawhml)
detach(package:car)
```


nminato@med.gunma-u.ac.jp

分散分析とTukeyの多重比較の結果

```
> stbyhml <- aov(FEST ~ as.factor(HML))
> summary(stbyhml)
                Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(HML)   3  6.214   2.071  19.238 1.589e-11 ***
Residuals      332 35.746   0.108
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
  1
> TukeyHSD(stbyhml)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = FEST ~ as.factor(HML))
$`as.factor(HML)`
          diff         lwr       upr     p adj
1-0 0.2361830  0.09550409 0.3768618 0.0001135
2-0 0.3377720  0.15362562 0.5219184 0.0000191
3-0 0.7206068  0.33815401 1.1030595 0.0000105
2-1 0.1015890 -0.11819042 0.3213685 0.6313713
3-1 0.4844238  0.08359563 0.8852520 0.0105365
3-2 0.3828348 -0.03523527 0.8009048 0.0860600
```


nminato@med.gunma-u.ac.jp

Holmの多重比較とType IIの平方和による分散分析の結果（carライブラリ）

```
> pairwise.t.test(FEST,as.factor(HML),method="holm")
Pairwise comparisons using t tests with pooled SD

data:  FEST and as.factor(HML)
  0       1       2
1 7.7e-05 -       -
2 1.6e-05 0.2335  -
3 1.1e-05 0.0059  0.0373
P value adjustment method: holm
> Anova(lm(FEST ~ as.factor(HML)))
Anova Table (Type II tests)

Response: FEST
                Sum Sq  Df F value     Pr(>F)
as.factor(HML)   6.214   3  19.238 1.589e-11 ***
Residuals       35.746 332
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
  1
```


nminato@med.gunma-u.ac.jp

クロス集計

```
DATA PLAC;
 INPUT MAL $ PLACE $ FR;
 CARDS;
  NEG INBEDNET 34
  POS INBEDNET 11
  NEG INROOM 59
  POS INROOM 28
  NEG INOPENKT 122
  POS INOPENKT 82
  NEG OUTBLDG 87
  POS OUTBLDG 41
  NEG BUSH 1
  POS BUSH 0
  NEG TERRACE 53
  POS TERRACE 30
  NEG BATH 22
  POS BATH 13
  NEG SEA 1
  POS SEA 0
  NEG NOTOBS 143
  POS NOTOBS 65
 ;
RUN;
 PROC FREQ DATA=PLAC;
  WEIGHT FR;
  TABLES PLACE*MAL / NOCOL NOROW
     NOPERCENT ALL;
 RUN;
```

```
options(error =
   quote({dump.frames(to.fi
   le=TRUE)}))
plac <-
   matrix(c(34,11,59,28,122
   ,82,87,41,1,0,53,30,22,1
   3,1,0,143,65),nr=2)
rownames(plac) <-
   c("NEG","POS")
colnames(plac) <-
   c("INBEDNET","INROOM","I
   NOPENKT","OUTBLDG","BUSH
   ","TERRACE","BATH","SEA"
   ,"NOTOBS")
print(plac)
fisher.test(plac,workspace
   =10000000)
```

大きなクロス集計のときは，作業領域を広げないと計算できない

nminato@med.gunma-u.ac.jp

クロス集計の出力

```
> print(plac)
    INBEDNET INROOM INOPENKT OUTBLDG BUSH TERRACE BATH SEA NOTOBS
NEG       34     59      122      87    1      53   22   1    143
POS       11     28       82      41    0      30   13   0     65
> fisher.test(plac,workspace=10000000)

    Fisher's Exact Test for Count Data

data:  plac
p-value = 0.4649
alternative hypothesis: two.sided
```


nminato@med.gunma-u.ac.jp

カプランマイヤ推定と生存関数

```
DATA SOL;
 INPUT GEN PERIOD CI;
 CARDS;
  1 101 1
  1 37 1
  1 22 1
  1 40 1
  1 15 1
  1 23 1
  1 24 1
  1 28 1
  2 17 1
  2 14 1
  2 22 1
  2 37 1
  2 12 1
  2 15 1
  2 19 1
  2 26 1
  2 29 0
  2 23 0
  2 20 0
  2 18 0
  2 9 0
  2 9 0
  2 3 0
  2 2 0
 ;
RUN;
PROC LIFETEST OUTSURV=SURV METHOD=KM
    PLOTS=(S,LLS);
 TIME PERIOD*CI(0);
 STRATA GEN; RUN;
PROC PRINT DATA=SURV; RUN;
```

```
sol <- data.frame(
GEN=c(rep(1,8),rep(2,16)),
PERIOD=c(101,37,22,40,15,23,24,28,1
   7,14,22,37,12,15,19,26,29,23,20,
   18,9,9,3,2),
CI=c(rep(1,16),rep(0,8)))
library(survival)
res <-
   survfit(Surv(PERIOD,CI)~as.facto
   r(GEN),data=sol)
summary(res)
print(res)
survdiff(Surv(PERIOD,CI)~as.factor(
   GEN),data=sol)
png("./LT.png",width=640,height=480
   )
par(family="sans",cex=1.2)
plot(res,lty=c(1,2),xlab="periods
   (months)",
ylab="Proportion never born 2nd
   baby",
main="Periods between 1st and 2nd
   birth for Solomon women.")
dev.off()
```


nminato@med.gunma-u.ac.jp

カプランマイヤ推定(1)

```
> summary(res)
Call: survfit(formula = Surv(PERIOD, CI) ~ as.factor(GEN), data =
   sol)
                as.factor(GEN)=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   15      8       1    0.875   0.117       0.6734        1.000
   22      7       1    0.750   0.153       0.5027        1.000
   23      6       1    0.625   0.171       0.3654        1.000
   24      5       1    0.500   0.177       0.2500        1.000
   28      4       1    0.375   0.171       0.1533        0.917
   37      3       1    0.250   0.153       0.0753        0.830
   40      2       1    0.125   0.117       0.0200        0.782
  101      1       1    0.000      NA           NA           NA
                as.factor(GEN)=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   12     12       1    0.917  0.0798        0.773        1.000
   14     11       1    0.833  0.1076        0.647        1.000
   15     10       1    0.750  0.1250        0.541        1.000
   17      9       1    0.667  0.1361        0.447        0.995
   19      7       1    0.571  0.1462        0.346        0.944
   22      5       1    0.457  0.1553        0.235        0.890
   26      3       1    0.305  0.1619        0.108        0.863
   37      1       1    0.000      NA           NA           NA
```

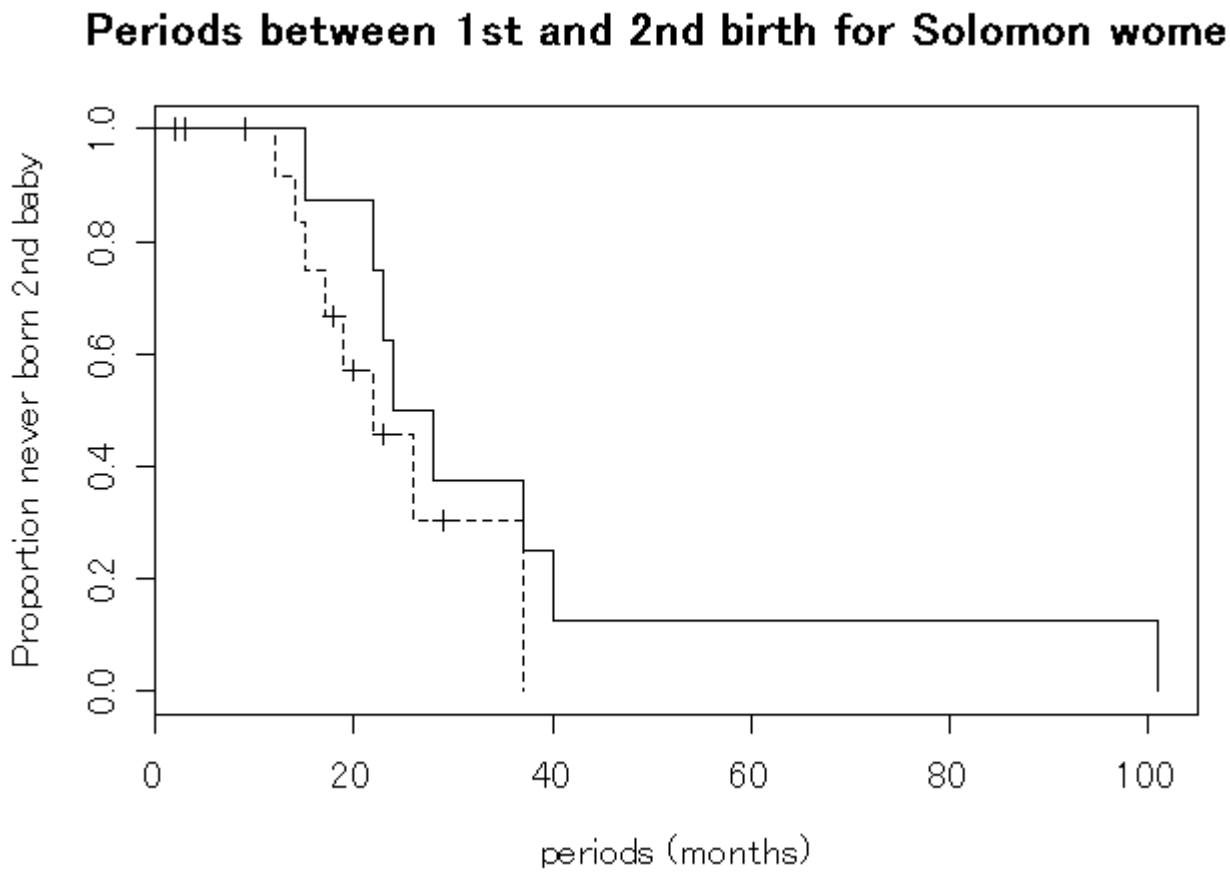

nminato@med.gunma-u.ac.jp

カプランマイヤ推定(2)

```
> print(res)
Call: survfit(formula = Surv(PERIOD, CI) ~ as.factor(GEN),
   data = sol)
                  n events median 0.95LCL 0.95UCL
as.factor(GEN)=1  8      8     26      23     Inf
as.factor(GEN)=2 16      8     22      17     Inf
> survdiff(Surv(PERIOD,CI)~as.factor(GEN),data=sol)
Call:
survdiff(formula = Surv(PERIOD, CI) ~ as.factor(GEN), data =
   sol)
                  N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(GEN)=1  8        8     9.78     0.323      1.03
as.factor(GEN)=2 16        8     6.22     0.508      1.03
 Chisq= 1  on 1 degrees of freedom, p= 0.311
```


nminato@med.gunma-u.ac.jp

生存関数のグラフ(png 出力の例)



nminato@med.gunma-u.ac.jp