社会基盤基礎演習（２年）：要点ノート

教科書：統計学入門，東京大学教養学部統計学教室 編，東京大学出版会.
担当：北野 利一（２４号館３階３１９号室；kitano@nitech.ac.jp）

【１章】 統計学の考え方

(p.2)　統計学とは何か？　～　現象の法則性を考察するための道具.

(p.7)　近代統計学理論のコンセプト
「部分」が**正しく選ばれ**ていれば，
それをもとに，「全体」を知ることが論理的に可能.

「部分」～ 標本；「全体」～ 母集団 （cf.「全数調査」と「標本調査」の違い）

「部分」と「全体」を結びつけるものが「モデル」である.
「部分」が**正しく選ばれている**か，否かは，
母集団を記述する「モデル」から抽出された「部分」であるか，否かという問題である.
また，その**論理**の考察には，以下のような概念が必要である.
「統計量」,「統計的推測」,「推定」,「誤差」,「確率」など

（例１：テキスト）標本相関係数と，相関係数の分布
（ちなみに，式 (1.2) は，このテキストの中で最も難解な式である．p.209 では，近似的に扱う）

$$p(r)=\frac{\left(1-\rho^2\right)^{\frac{n-1}{2}}\left(1-r^2\right)^{\frac{n-4}{2}}}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)}\sum_{k=0}^{\infty}\frac{(2\rho r)^k}{k!}\left\{\Gamma\left(\frac{n-1+k}{2}\right)\right\}^2$$

（例２）標本平均（標本例＝{68, 70, 57, 90, 68, 32, 13}）
・６個までのデータの平均と，７個めを含めた平均の違い.
・まだ見ぬ８個めのデータを含めた平均は？？？（この考え方が，「推測統計」である）

(p.8)　データ（資料，標本）～ 量的／質的，次元，時系列，クロスセクション
なお，測定（観測）の際には，尺度を用いる．尺度の分類（名義／順序／間隔／比） (p.27)

(p.13)　それぞれの統計手法には，それらが用いられるべき**理由**がある.
統計計算パッケージを盲目的に利用するのは，重大な誤り！
コンピュータは，「計算」してくれるけれども，「分析（結果の解釈）」はしてくれません.

【２章】 １次元データ

(p.17)　記述統計　～ データを整理・要約.（cf. 統計的推測，p.176）

(p.19)　ヒストグラム（度数分布表の図示化）（スタージェスの公式）
累積度数グラフ（ヒストグラムとの関係），累積度数相対グラフ（縦軸が区間 [0,1] となる）
（２つの累積度数相対グラフの比較 ＝ ローレンツ曲線およびジニ係数,
ただし，これらは２次元データへの適用）

(p.28)　代表値（分布の位置に関する要約）
・平均（算術平均／幾何平均／調和平均）　cf. 相加相乗の不等式（相加平均 ≧ 相乗平均）
・メディアン（中央値＝Ｑ２）　・モード（最頻値）　　・外れ値（！これも，代表値？）

(p.35)　散らばりの尺度
・レンジ（＝最大値 － 最小値）　　・４分位偏差（＝（Ｑ３－Ｑ１）／２）
・標準偏差（＝√分散；式 (2.10 & 11), p.37）　　・変動係数（変異係数）CV（＝標準偏差／平均）

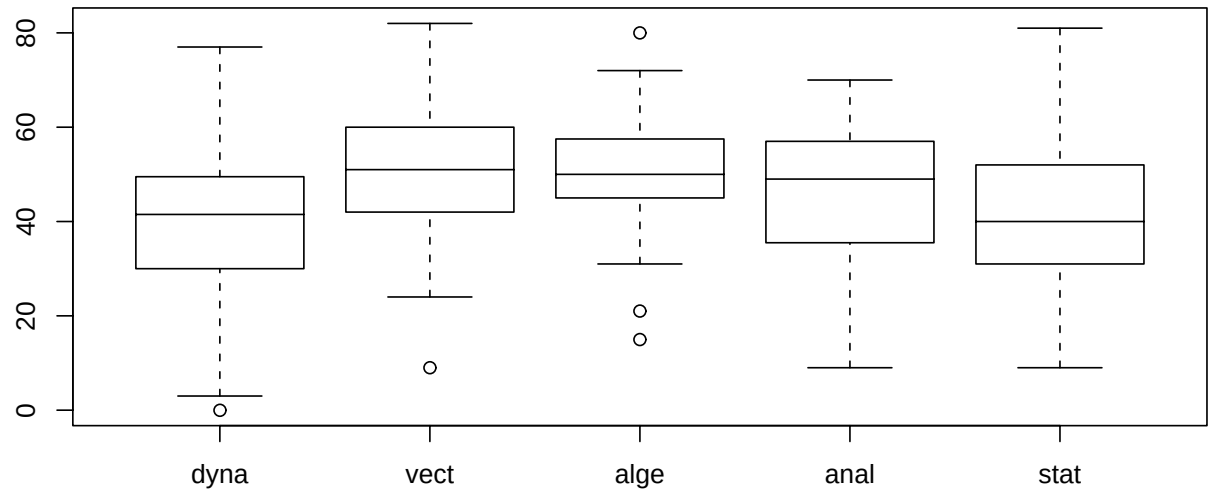
(演習)

computing software: R [available at http://www.r-project.org/]

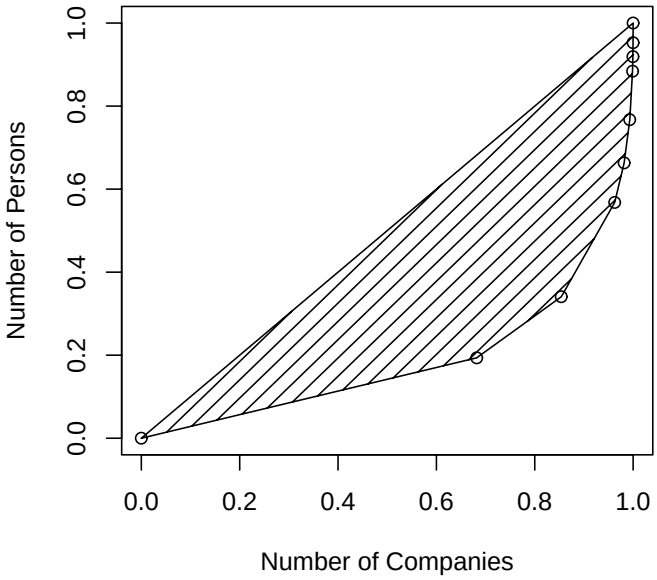
```
R : Copyright 2003, The R Development Core Team
Version 1.7.1  (2003-06-16)

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.

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.
```

```
> # 1 #
> a <- c(68, 70, 57, 90, 68, 32, 13)
> mean(a)
[1] 56.85714
> sum(a)
[1] 398
> sum(a)/7
[1] 56.85714
> a
[1] 68 70 57 90 68 32 13
> a[-7]
[1] 68 70 57 90 68 32
> sum(a[-7])/6
[1] 64.16667
> q()
Save workspace image? [y/n/c]:

> # 2 #
> getwd()                          # w_orking d_irectory
[1] "MacOSX3:Users:tk:rm171:"
> setwd("MacOSX3:Users:tk:exercises")
> precip <- scan("precip.txt")   # example of time series data
Read 29 items
> precip
 [1] 43.40 53.02 63.52 45.93 48.26 50.51 49.57 43.93 46.77 59.12 54.49 47.38
[13] 40.78 45.05 50.37 54.91 51.28 39.91 53.29 67.59 58.71 42.96 55.77 41.31
[25] 58.83 48.21 44.67 67.72 43.11
> #
> # exercise 5.4 in Probability Concepts in Engineering Planning and Design
> # by Ang. A. H-S. and W. H. Tang
> # annual precipitation (unit: inch) at Esopus River (NY)
> #
```

```
> length(precip)
[1] 29
> year <- 1918:1946
> plot(year, precip, type="b")
> abline(h=mean(precip), lty=2)
> hist(precip)
> hist(precip, breaks=seq(38, 68, by=2))
> hist(precip, breaks=seq(38, 68, len=7))
> summary(precip)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  39.91   44.67   49.57   50.70   54.91   67.72
> range(precip)
[1] 39.91 67.72
> IQR(precip)
[1] 10.24
> 44.67 + 10.24
[1] 54.91
> sd(precip) # s_tandard d_eviation
[1] 7.702964
> var(precip)
[1] 59.33566
> .Last.value
[1] 59.33566
> sqrt(.Last.value)
[1] 7.702964
> (d2 <- (precip - m.precip)^2)
 [1]  53.32524721   5.37120583 164.29051617  22.77593341   5.96538514
 [6]   0.03702307   1.28236100  45.86558859  15.46387824  70.85575755
[11]  14.34580927  11.03843341  98.45429548  31.94978169   0.11049893
[16]  17.70378169   0.33360583 116.47619548   6.69560238 285.19056790
[21]  64.12143686  59.94497134  25.68042996  88.21743686  66.05765755
[26]   6.21212652  36.39001617 289.59824031  57.64474721
> sum(d2)/length(precip) # by definition text p.37
[1] 57.2896
> sum(d2)/(length(precip) - 1) # another important def., discussed later
[1] 59.33566

> # 3 #
> math <- read.table("math.txt", header=TRUE) # example of cross sectional data
> 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
> apply(math, 2, mean)
    dyna     vect     alge     anal     stat
38.95455 50.59091 50.60227 46.68182 42.30682
```



```
> boxplot(math)
> boxplot(math, horizontal=T)
```




```
> # advanced
> hist(math$alge)
> ?hist
> ?nclass.Sturges
> nclass.Sturges(math$alge)
[1] 8
> nclass.FD(math$alge)
[1] 12
> hist(math$alge, breaks="FD")
> length(math$alge)
[1] 88
> log(88)/log(2) +1  # text p.22
[1] 7.459432
> table(cut(math$alge, breaks=seq(0,100, by=10)))

 (0,10]  (10,20]  (20,30]  (30,40]  (40,50]  (50,60]
      0        1        1       11       33       27
(60,70]  (70,80]  (80,90] (90,100]
     12        3        0        0
> table(cut(math$alge, breaks=seq(0,100, by=10), right=F))

 [0,10)  [10,20)  [20,30)  [30,40)  [40,50)  [50,60)
      0        1        1       10       28       30
[60,70)  [70,80)  [80,90) [90,100)
     15        2        1        0
>
> stem(math$alge)

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

  1 | 5
  2 | 1
  3 | 1266677889
  4 | 011333334556666777788899999
  5 | 000000111223333344455666778899
  6 | 000111123455578
  7 | 12
  8 | 0

>

> # 4 # Lorenz curve in text p.26-27 (Table 2.3)
> n.companies <- c(4428, 1118, 705, 124, 75, 36, 4.5, 2.4, 1.1)
> n.persons <- c(9486, 7214, 11134, 4648, 5103, 5734, 1706, 1651, 2320)
> (sum(n.companies) -> total.comp)
[1] 6494
> (sum(n.persons) -> total.pers)
[1] 48996
> (cumsum(n.companies)/total.comp -> crf.comp)
[1] 0.6818602 0.8540191 0.9625808 0.9816754 0.9932245
[6] 0.9987681 0.9994610 0.9998306 1.0000000
> (cumsum(n.persons)/total.pers -> crf.pers)
[1] 0.1936076 0.3408442 0.5680872 0.6629521 0.7671034
[6] 0.8841334 0.9189526 0.9526492 1.0000000
> plot(c(0, crf.comp), c(0, crf.pers),
+      xlab="Number of Companies",
+      ylab="Number of Persons")
> polygon(c(0, crf.comp), c(0, crf.pers), dens=10)
```

【３章】　２次元データ

１）散布図（２元とも量的データの場合）

・相関関係（直線関係＝比例関係を**前提**）

・表現（強い／弱い相関，正／負の相関，無相関／完全相関）

２）分割表（クロス表）（質的データが含まれる場合）

（表側と表頭の同時度数分布表）

（量的データの標本数の多い場合には，散布図より分割表にする方が有効）

３）相関係数（ピアソンの積率相関係数）

散布図（量的データ）における要約値の１つ

◆ 定義式＝式 (3.1) p.49 あるいは式 (1.1) p.6 ：

$$r_{xy} = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2}\sqrt{\sum\left(y_j - \bar{y}\right)^2}}$$

（**注意点**：この係数を用いて，相関があるか，ないかについて議論するには，
変量が正規分布に従っていることが前提となる）

４）因果関係（原因と結果，相関関係とは異なることに注意）

５）みかけ上の相関

例１： 飲食店が多いと，金融機関の店舗が多い？　例２： コンビニが多いと，犯罪が多い？

６）偏相関係数（なぜ，このような式になるのかの説明は，９）回帰を用いて示す）

$$r_{xy,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1-r_{xz}^2}\sqrt{1-r_{yz}^2}}$$

７）順位相関係数（順位のついた**質的データ**に対して）

・スピアマンの順位相関係数（順位の積率相関係数）

・ケンドールの順位相関係数

ただし，タイ（同順位）のあるデータには，注意！（タイが少なければ，順位の平均を使用）

８）自己相関関数　　（時系列データの周期特性を考察）

理解へのヒント： $\frac{1}{T_0}\int_0^{T_0}\cos(\omega_0 t)\cos(\omega_0 t+\varphi)dt = \frac{1}{2}\cos(\varphi)$

９）回帰（単回帰　cf. 重回帰（複数の説明変数）；　なお，検定を含めた詳しい議論は１３章で行う）

・説明変数と被説明変数の線形関係を導く（切片と勾配を算出）

・誤差の平方和（$\varepsilon^2 = \sum_i \left\{y_i - \left(a + b\,x_i\right)\right\}^2$）の最小化（**最小自乗法**）

＝　**正規方程式**（式 (3.9) p.59）

$$\left.\begin{aligned} \frac{\partial}{\partial a}\varepsilon^2 &= -\sum_i \left\{y_i - \left(a + b\,x_i\right)\right\} = 0 \\ \frac{\partial}{\partial b}\varepsilon^2 &= -\sum_i x_i\left\{y_i - \left(a + b\,x_i\right)\right\} = 0 \end{aligned}\right\} \Leftrightarrow \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}$$

・勾配係数と相関係数との関係（式 (3.15) p.61）

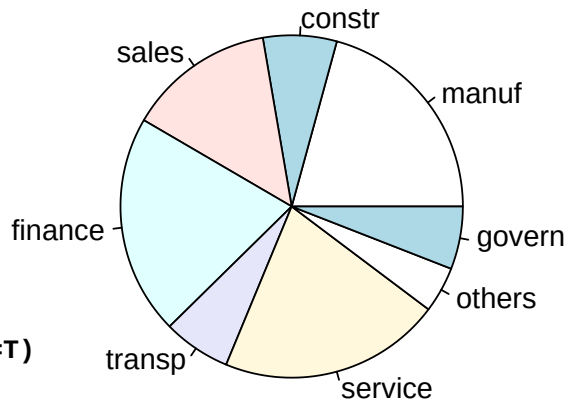
・決定係数（寄与率： 回帰の効果）

全体の変動　＝　回帰により説明できる変動　＋　誤差変動　（図 3.25 参照）

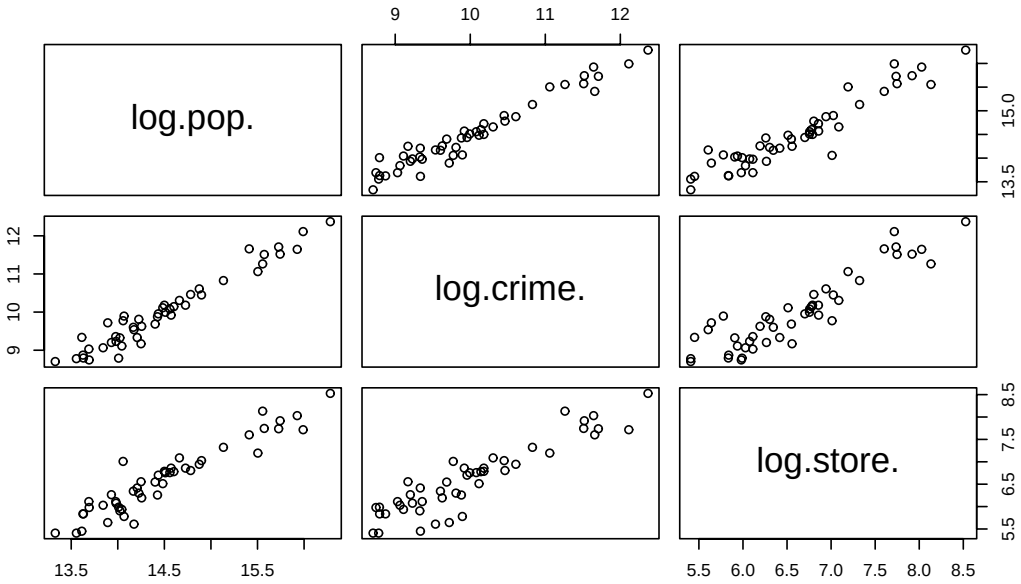
（言い換えれば，　...　　誤差と予測値は，直交する）

(演習)

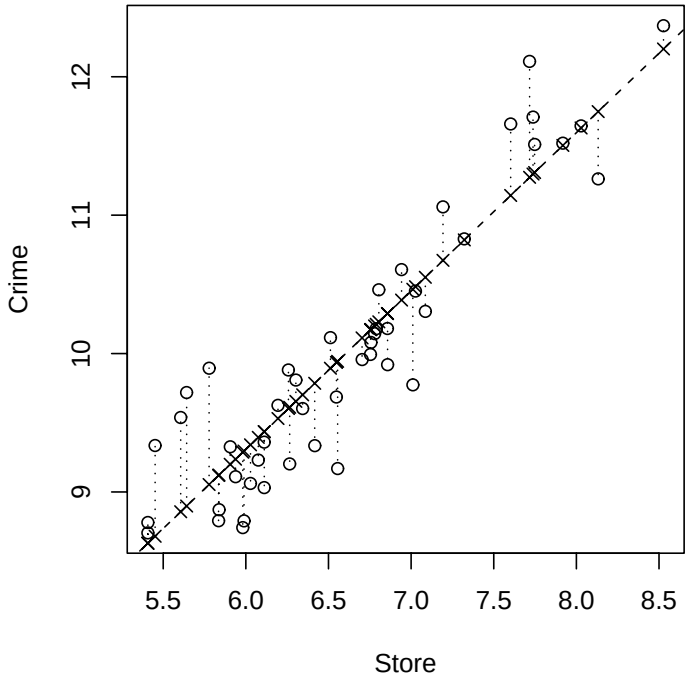
```
1:   > # 1 # visit http://www.nikkenren.com/handbook/index3.html (ver. 2004)
2:   > seisan <- c(102.3, 34.3, 68.5, 102.3, 31.5, 103.5, 21.4, 29.2)
3:   > syugyou <- c(1178, 604, 1133, 232, 496, 2055, 391, 227)
4:   > pie(seisan, labels=gyokai <- c("manuf", "constr", "sales", "finance",
     + "transp", "service", "others", "govern"))
5:   > plot(seisan, syugyou)
6:   > cor( seisan, syugyou)
     [1] 0.597209
7:   > d.seisan  <- seisan  - mean(seisan)
8:   > d.syugyou <- syugyou - mean(syugyou)
9:   > plot(d.seisan,         d.syugyou)
10:  > sum( d.seisan    *     d.syugyou )/sqrt(
     + sum( d.seisan^2) * sum(d.syugyou^2) )
     [1] 0.597209
11:  > # this makes little sense, ...
12:  > # see lines 32-42 in the next page
```

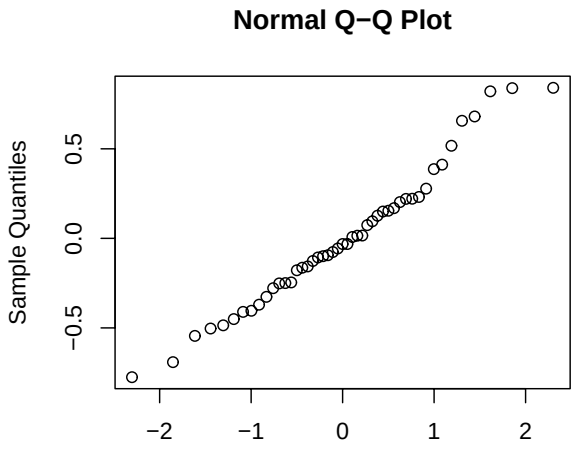


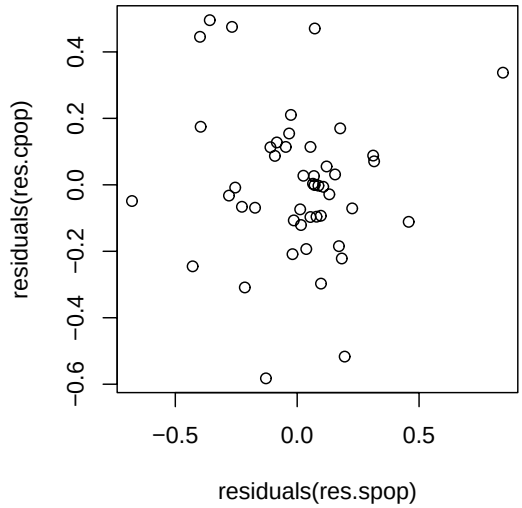

```
13:  > # 2 #
14:  > conb <- read.table("conbini.txt", header=T)
15:  > names(conb)
     [1] "pop"   "crime" "store"
16:  > pop
     Error: Object "pop" not found
17:  > attach(conb)
18:  > pop
      [1]  5692321  1481663  1419505  2328739  1213667  1256958  2133592  ...
19:  > par(mfrow=c(1,2))
20:  > hist(pop); hist(log(pop))

21:  > conbini <- data.frame(log(pop), log(crime), log(store))
22:  > par(mfrow=c(1,1))
23:  > plot(conbini[[1]], conbini[[2]])
24:  > pairs(conbini)
25:  > 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
26:  > (cor(log(crime), log(store)) -> cor.cs)
     [1] 0.9161085
27:  > cor(crime, store) # similar value but not better ...
     [1] 0.9148999
28:  > cor(rank(crime), rank(store))
     [1] 0.8812374
29:  > cor(rank(log(crime)), rank(log(store))) # obviously, ...
30:  > cor.test(crime, store)
31:  > cor.test(crime, store, method="spearman")
```

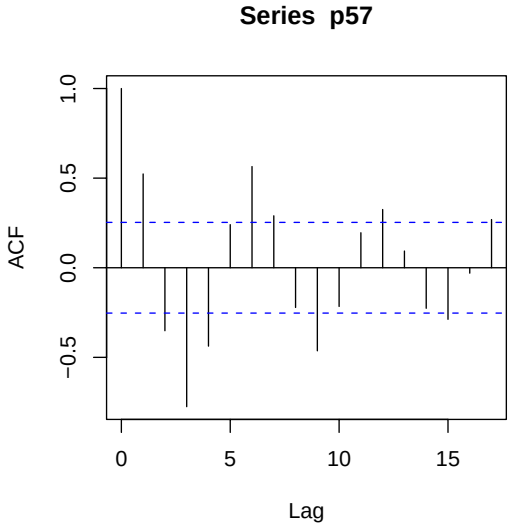


```
> # 3 #
> cor(rank(seisan), rank(syugyou)) # spearman
> rank(seisan)
 [1] 6.5 4.0 5.0 6.5 3.0 8.0 1.0 2.0
> cor.test(seisan, syugyou, method="spearman")

> # 4 #  avoid the tie for the explanation ...
> sum(seisan)
 [1] 493
> seisan[4] <- 102.4 # ~ 493.1 * 0.208 > 102.3 ~ 493.1 * 0.207
> seisan
 [1] 102.3  34.3  68.5 102.4  31.5 103.5  21.4  29.2
> cor.test(seisan, syugyou, method="spearman")
> cor.test(seisan, syugyou, method="kendall")
> matrix(c(rank(seisan), rank(syugyou)), byrow=T, nr=2,
+ dimnames=list(c("seisan","syugyou"), gyokai)
+ )[,order(seisan)]
        others govern transp constr sales manuf finance service
seisan       1      2      3      4     5     6       7       8
syugyou      3      1      4      5     6     7       2       8

> # 5 #
> plot(conbini$log.store, conbini$log.crime, xlab="Store", ylab="Crime")
> lm(log.crime. ~ log.store., data=conbini) -> res.cs
> res.cs

 Call:
 lm(formula = log.crime. ~ log.store., data = conbini)

 Coefficients:
 (Intercept)   log.store.
       2.441        1.145

> abline(res.cs, lty=2)
> nrow(conbini)
 [1] 47
> s.x  <- sum(conbini$log.store)
> s.y  <- sum(conbini$log.crime)
> s.xy <- sum(conbini$log.crime *
              conbini$log.store)
> s.xx <- sum(conbini$log.store^2)
> mat <- matrix(
+  c(47, s.x, s.x, s.xx), nc=2)
          [,1]      [,2]
 [1,]  47.0000  310.7895
 [2,] 310.7895 2082.9446
> solve(mat) %*% c(s.y, s.xy)
          [,1]
 [1,] 2.440790
 [2,] 1.144578
> (crime.f <- predict(res.cs))
        1         2         3
11.748296  9.785239  9.701854 ...

> points(conbini$log.s, crime.f, pch=4)
> segments(conbini$log.s, crime.f,
+          conbini$log.s,
+             conbini$log.crime, lty=3)
> qqnorm(residuals(res.cs))
> 
> # advanced
> crossprod(matrix(c(rep(1, 47),
+        conbini$log.store), nc=2))
          [,1]      [,2]
 [1,]  47.0000  310.7895
 [2,] 310.7895 2082.9446
> lsfit(log(crime), log(store))
```

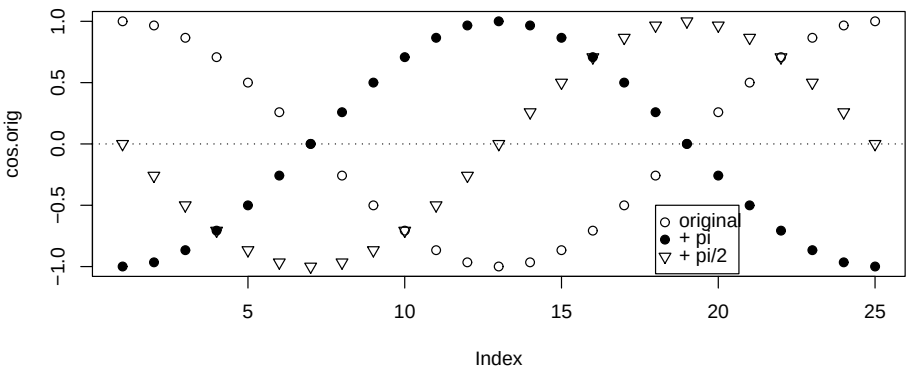









```
63:  > # 6 #
64:  > lm(log.crime. ~ log.pop., data=conbini) -> res.cpop
65:  > lm(log.store. ~ log.pop., data=conbini) -> res.spop
66:  > plot(residuals(res.spop), residuals(res.cpop))
67:  > cor( residuals(res.spop), residuals(res.cpop))
     [1] -0.06225971
68:  > # cf.
69:  > cor.cp <- cor(conbini$log.c, conbini$log.p)
70:  > cor.sp <- cor(conbini$log.s, conbini$log.p)
71:  > (cor.cs - cor.cp * cor.sp)/sqrt((1 - cor.cp^2) * (1 - cor.sp^2))
     [1] -0.06225971
72:  >
73:  > # advanced
74:  > par(mfrow(2,1)) -> op
75:  > qqnorm(res.cpop)
76:  > qqnorm(res.spop)
77:  > par(op)
78:  > cor.test(residuals(res.spop), residuals(res.cpop))

79:  > # 7 #
80:  > acf(p57) -> res
81:  > as.vector(res$acf)                    # another definition is used in R.
      [1]  1.00000000  0.52322256 -0.35142142 -0.77527647 -0.43742543  0.24042305
      [7]  0.56461222  0.28974677 -0.22164656 -0.46356294 -0.21596706  0.19524342
     [13]  0.32493090  0.09323200 -0.22630062 -0.28774449 -0.03028531  0.26890559
82:  > as.vector(res$acf * 60/(60 - 0:17))    # these are the same as those in text.
      [1]  1.00000000  0.53209074 -0.36353940 -0.81608049 -0.46867010  0.26227969
      [7]  0.62734691  0.32801521 -0.25574603 -0.54536816 -0.25916047  0.23907358
     [13]  0.40616363  0.11901957 -0.29517472 -0.38365932 -0.04129815  0.37521710
82:  >
83:  > # cf.
84:  > cos.orig <- cos(2*pi*0:24/24)
85:  > plot(cos.orig)
86:  > points(cos.b <- cos(2*pi*(0:24/24) + pi), pch=19)
87:  > points(cos.c <- cos(2*pi*(0:24/24) + pi/2), pch=25)
88:  > abline(h=0, lty=3)
89:  > legend(18,-.5, c("original", "+ pi", "+ pi/2"), pch=c(1,19,25))
90:  > cor(cos.orig, cos.orig)
     [1] 1
91:  > cor(cos.orig, cos.b)
     [1] -1
92:  > cor(cos.orig, cos.c)
     [1] 1.423511e-17
93:  > #, which stands for
94:  > # zero.
```



【４章】 確率

１）ランダムネスの法則

・コインを投げて次に表が出るか裏が出るか，　...　いいあてることはできない．
・１０回中，表が２回，裏が８回であるコインは，公平か否か？　...　という判断はできる．

◆ 偶然の法則（**誕生日問題** ＝ 練習問題 4.4 p.85）

同じ誕生日である人が含まれる確率が Q であるための集団のサイズ N は，
近似的に，次式で得られる．したがって，約 23 人の集団では，五分五分である．

$$N \cong \sqrt{-2c\log(1-Q)} \quad (c = 365)$$

２）順列と組み合わせ

順列：　$_nP_r = \dfrac{n!}{(n-r)!}$ ；　組み合わせ：　$_nC_r = \dfrac{n!}{r!\,(n-r)!}$

＊階乗（0! は 1 と定義）とガンマ関数（p.125 も参照）

＊計算のコツ：スターリングの公式（大きな値の階乗）

$$n! \cong \sqrt{2n\pi}\, n^n e^{-n}$$

３）条件付き確率

同時確率　＝　周辺確率　＊　条件付確率　　（乗法定理　式 (4.13) ）

cf. 加法定理（式 (4.9) p.80）

４）独立性（無相関（ p.138 ）とは異なる！）（p.143 も参照）

周辺確率　＝　条件付確率　（条件に影響されない）　　となるので，
同時確率　＝　周辺確率　＊　周辺確率　　が成立する．

（同時確率，周辺確率などの概念は，p.134, 135 も参照）

５）ベイズの定理

原因に対する結果の確率 Ｐ（Ａ｜Ｈ）をもとに
結果に対する原因の確率 Ｐ（Ｈ｜Ａ）を知る．

Ｐ（Ｈ｜Ａ）＝ Ｐ（Ｈ）＊Ｐ（Ａ｜Ｈ）／Ｐ（Ａ）

（乗法定理を変型したもの，つまり，式(4.13) の右辺と式 (4.14) の右辺が等しいことから導かれる）

◆ ガン診断：練習問題 4.7

ガンである人が，ガン診断の検査の結果，陽性となる確率： $P(A|C) = 0.95$,
ガンでない人が，ガン診断の検査の結果，陰性となる確率： $P(\overline{A}|\overline{C}) = 0.95$,
ガンに罹る確率： $P(C) = 0.005$ とすれば，この検査法により陽性となった時，ガンである確率は？
（予想以上に，よくない検査法であることがわかる．なぜか？改善のためには，どうすればよいか？）

【５章】 確率変数

・乱数（疑似乱数，一様乱数，指数乱数，正規乱数，...）
・期待値とその性質（式 (5.25) p.96）　　・分散とその性質（式 (5.29) p.98）

特に，　$E(cX) = cE(X)$;　$V(cX) = c^2 V(X)$　であることに注意．

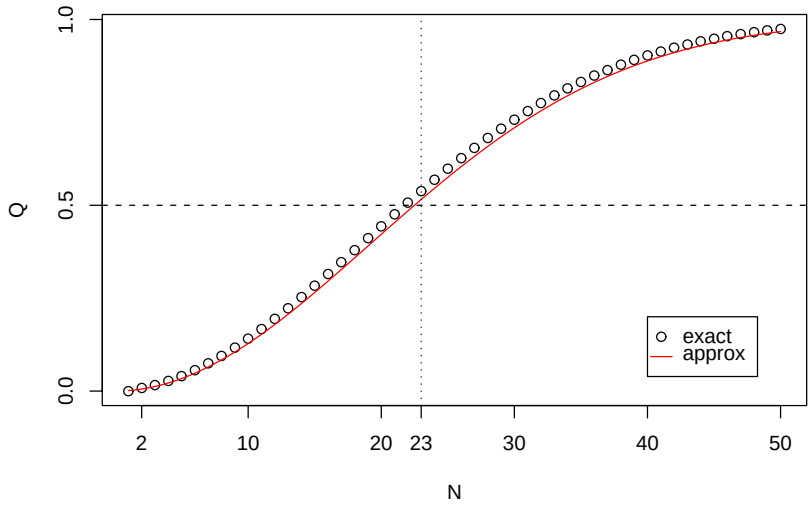
（また，期待値の加法性は，常に成立する．分散の加法性は，常に成立しない．
独立（無相関で十分）であれば，分散の加法性は成立する．p.148）

・モーメント（積率），モーメント母関数（cf. 特性関数 ... ただし，このテキストでは扱っていない）
・チェビシェフの不等式（例えば，大数の法則を示す時に用いる p.160 ）

（この章の内容は，非常に重要であるが，やや概念的である．６章以降にて，関連する具体的な内容とともに説明する）

(演習)

```
> # 1 #
> q <- numeric(50)
> for (j in 2:50) q[j] <- 1 - prod(1 - 1:j/365)
> q
 [1] 0.000000000 0.008204166 0.016355912 0.027135574 0.040462484 0.056235703
 [7] 0.074335292 0.094623834 0.116948178 0.141141378 0.167024789 0.194410275
     ...
> approx <- function(n) 1 - exp(-n^2/2/365)
> plot(q, xlab="N", ylab="Q", axes=F)
> axis(1, c(2,10,20,23,30,40,50))
> axis(2, c(0, .5, 1))
> box()
> curve(approx, 1,50, col="red", add=T)
> abline(h=.5, lty=2)
> abline(v=23, lty=3)
> legend(40, .2, c("exact", "approx"),
+        pch=c(1,NA), lty=c(NA,1), col=c("black","red"))
> q[23]
[1] 0.5383443
> approx(23)
[1] 0.5155095
>
> # advanced
> ?pbirthday
> pbirthday(23)
[1] 0.530137
> pbirthday(88, coin=3)
[1] 0.5699803
> qbirthday(.5, coin=9)
[1] 754
```




```
> # 2 #
> gamma(11)
[1] 3628800
> prod(1:10)
[1] 3628800
> sqrt(2*10*pi)*(10/exp(1))^10
[1] 3598696

> lg <- function(n) (n + 1/2) * log(n) - n + log(2*pi)/2 # Stirling's formula
> curve(lgamma, .01, 50)
> points(log(gamma(1:50)))
> curve(lg, 10, 50, lty=3, add=T)

> # 3 #
> pr.A.C <- 0.95; pr.notA.notC <- 0.95
> pr.C <- 0.005
> pr.C * pr.A.C/(pr.C * pr.A.C + (1 - pr.C) * (1 - pr.notA.notC))
[1] 0.08715596
> pr.A.C <- 0.9995; pr.notA.notC <- 0.9995 # pr.C = 0.005
> pr.C * pr.A.C/(pr.C * pr.A.C + (1 - pr.C) * (1 - pr.notA.notC))
[1] 0.9094631
> pr.A.C <- 0.95  # pr.notA.notC  = 0.9995,  pr.C = 0.005
> pr.C * pr.A.C/(pr.C * pr.A.C + (1 - pr.C) * (1 - pr.notA.notC))
[1] 0.905193
> pr.C <- 0.5; pr.notA.notC <- 0.95 # pr.A.C   = 0.95
> pr.C * pr.A.C/(pr.C * pr.A.C + (1 - pr.C) * (1 - pr.notA.notC))
[1] 0.95

> # 4 #
> qqnorm(rnorm(100))
> qqnorm(rnorm(10000))   #and qqnorm(rnorm(10))
> qqnorm(rexp(100))      #and qqnorm(runif(100))
>
> sample(1:6, 10, rep=T) # dice simulator
 [1] 5 3 3 3 2 6 2 5 3 1
```

【６章】（具体的現象を記述する）確率分布

《おもな離散型の確率分布》

１）超幾何分布　～ 捕獲再捕獲法（全体数の推測）　◆ 密度関数（式 (6.1) p.109）

◎ ２つの抽出法（復元抽出と非復元抽出）

２）２項分布

（非復元抽出としての超幾何分布を，復元抽出として扱えば，２項分布となる．
または，総数が無限大であれば，非復元抽出は復元抽出に近似できる．式 (6.5) にて極限）

例：日最高気温がある値を超える年間発生数，コイン投げによる表の発生数など

・ベルヌーイ試行
・多項分布
◆ 密度関数（式 (6.6) p.111），期待値と分散（式 (6.8) p.112）
■ ポアソン分布への近似，正規分布への近似（図 6.2 参照）

３）ポアソン分布

（２項分布の近似：ポアソンの小数の法則，
観測数＝大量，確率＝希少として近似する（ただし，発生率＝観測数＊確率＝一定値））
（２項分布に従う現象には，時間の分割（ブロック）があるのに対し，
ポアソン分布に従う現象には，ブロックはない．）

例：交通事故件数，高波の年間発生数など（リスクや安全性に関連する現象への適用範囲は多い）

◆ 密度関数（式 (6.9) p.114），　『 期待値 ＝ 分散 』

４）幾何分布（待ち時間分布）

＝ 初めての成功までの失敗の回数（失敗が連続する時間）の分布

（cf. 所与の時間内の成功数の分布 ＝ ２項分布）

◆ 期待値と分散（式 (6.13) p.117），『 期待値 ＝ 成功率の逆数 』

例： 再現期間 ＝ 災害（例えば，あるレベル以上の高波）が発生する平均周期．

幾何分布の期待値が成功率の逆数となることは，容易に（キチンとした計算をしなくても）
納得できる（が，計算でも確認すること）．

５）負の２項分布（＝ k 回めの成功までの失敗の回数の分布）

（または，混合ポアソン過程の一例：発生率が変動する場合）

負の２項分布の期待値および分散（式 (6.16) p.119）は，幾何分布の期待値および分散の k 倍．

６）（離散）一様分布（例：公平なサイコロ）

◆ 期待値と分散（式 (6.18) p.119）

《おもな連続型の確率分布》

1）正規分布（ガウスの誤差関数）　　◆ 密度関数 （式 (6.19a) p.120）
（確率変数の和や平均の分布に適用できる ＝ 中心極限定理.
この定理は，大数の法則も含め，後に詳しく検討　p.157, p.162）

2）指数分布（待ち時間分布）◆ 密度関数と分布関数 （式 (6.25) と式 (6.26) p.123）
＝ 時間単位は連続的である（時間の計量にブロックがない）.
cf. 相当する離散型の分布 ＝ 幾何分布
cf. 所与の時間内の成功数の分布 ＝ ポアソン分布
例：耐用年数，寿命，災害との遭遇までの年数など適用できる現象は多い

3）ガンマ分布（指数分布の一般化）　　◆ 密度関数 （式 (6.31) p.125）
cf. 相当する離散型の分布 ＝ 負の２項分布
例：システムダウンまでの待ち時間など （ガンマ関数という少し高級な数学を用いるものの，
計算が簡単であるので，待ち時間という現象に限らず，いろんな現象に適用することが多い.）

・再生性
確率変数の和の分布が，もとの分布と同一の分布族であること （p.149-151）
二項分布，ポアソン分布，正規分布，ガンマ分布，負の二項分布について，再生性が成立.

・モーメント母関数
定理証明用の道具である．例えば，再生性の証明 (p.150)，中心極限定理の証明 (p.165）で用いる．
なお，重要な性質として，独立な確率変数の和に対し，式 (7.28) p.145 が成立.

他に，経験的な観点から，現象を説明するのに用いられるものとして，以下のものが代表的である.

4）ベータ分布　　◆ 密度関数 （式 (6.34) p.126）
（確率的主観の分布：ベイズの事前確率 p.84；図 6.13 参照）

5）コーシー分布　　◆ 密度関数 （式 (6.38) p.128）
（正規分布の形状に類似するが，広がりが大きく，分散がない.
期待値もない！！ 線対称なのに期待値がないとは，いかに？！）

6）対数正規分布　　◆ 密度関数 （式 (6.39) p.128）
log X が正規分布に従う場合の *X* の分布

7）ワイブル分布　　◆ 密度関数 （式 (6.43) p.129）

以下は，現象に関連づける理論があるが，その理論がやや高度であるもの.

8）パレート分布　　◆ 密度関数 （式 (6.41) p.129）
（あるレベルを超える高波の波高分布など）

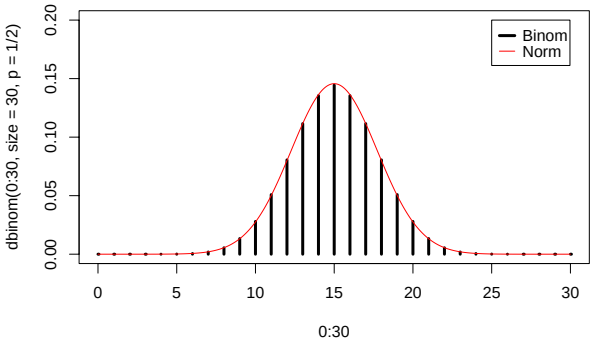
9）極値分布（年最大波高の分布）
極値 I 型分布は，ガンベル分布あるいは二重指数分布としても知られる.

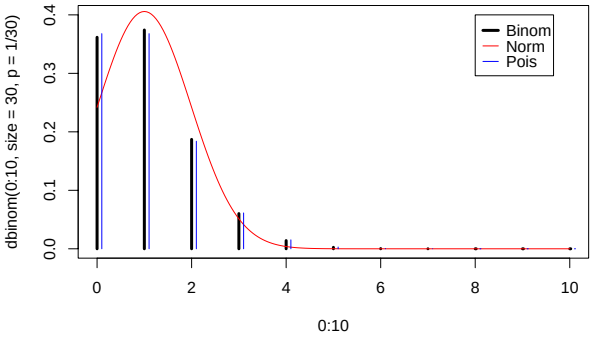
◆ 密度関数：　$f(x)=\frac{1}{\sigma}\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)-\exp\left(\frac{x-\mu}{\sigma}\right)\right\}$

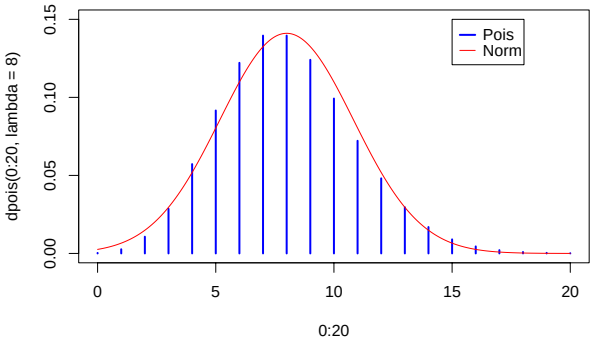
(演習)

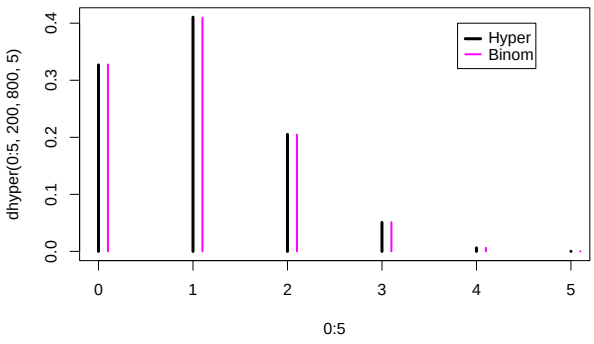
```
    > # 1 #
1:  > plot(0:30, dbinom(0:30, size=30, p=1/2), type="h", ylim=c(0, .2), lwd=3)
2:  > x <- seq(0,30, by=.05)
3:  > lines(x, dnorm(x, mean=15, sd=sqrt(30/4)), col="red")
4:  > legend(25,.2, c("Binom", "Norm"), lwd=c(3,1), col=c("black","red"))

5:  > # 2 #
6:  > plot(0:10, dbinom(0:10, size=30, p=1/30), type="h", ylim=c(0, .4), lwd=3)
7:  > x <- seq(0,10, by=.05)
8:  > lines(x, dnorm(x, mean=1, sd=sqrt(30*29)/30), col="red")
9:  > lines(0:10 +.1, dpois(0:10, lambda=1), col="blue", type="h")
10: > legend(8,.4, c("Binom", "Norm", "Pois"), lwd=c(3,1,1),
    +          col=c("black", "red",  "blue"))

11: > # 3 #
12: > plot(0:20, dpois(0:20, lambda=8), col="blue", type="h", ylim=c(0, .15), lwd=2)
13: > x <- seq(0,20, by=.05)
14: > lines(x, dnorm(x, mean=8, sd=sqrt(8)), col="red")
15: > legend(15,.15, c("Pois", "Norm"), lwd=c(2,1), col=c("blue","red"))

16: > # 4 #
17: > plot(0:5, dhyper(0:5, 200, 800, 5), type="h", lwd=3) # text p.110
18: > lines(0:5 +.1, dbinom(0:5, size=5, p=200/1000), type="h", col="magenta", lwd=2)
19: > legend(3.8,.4, c("Hyper", "Binom"), lwd=c(3,2), col=c("black","magenta"))

20: > # 5 #
21: > rbinom(10, size=1, p=.5)      # Bernoulli trials
     [1] 0 1 0 0 1 0 0 1 0 1
22: > sum(.Last.value)
    [1] 4
23: > sum(rbinom(10, size=1, p=.5))
    [1] 8
24: > rbinom(1, size=10, p=.5)      # just the concept of “reproductive”
    [1] 4
25: > (rbinom(100, size=10, p=.5) -> dat.1)
      [1] 7 7 1 2 7 6 5 3 6 4 4 6 4 6 3 4 2 7 3 5 6 4 3 6 4 4 6 5 8 9 6 6 6 4
     [35] 5 4 8 3 1 7 4 5 3 6 6 4 7 7 6 6 8 2 4 4 4 4 2 7 5 6 6 5 5 3 6 2 5 7
     [69] 3 3 5 5 6 5 0 5 4 8 5 4 9 5 6 3 4 7 6 3 7 5 5 4 4 5 5 7 4 1 4 5
26: > hist(dat.1)
27: > rpois(10, lambda=2)
     [1] 0 2 1 1 5 3 5 1 0 1
28: > sum(.Last.value)
    [1] 19
29: > rpois(1, lambda=20)
    [1] 14

30: > # 6 #
31: > # data obtained at the site:
32: > # http://www.data.kishou.go.jp/yohou/typhoon/statistics/index.html
33: >
34: > typh <- read.table("typhoonL.txt", header=T)
35: > names(typh)
     [1] "Jan"   "Feb"   "Mar"   "Apr"   "May"   "Jun"   "Jul"   "Aug"
     [9] "Sep"   "Oct"   "Nov"   "Dec"   "total"
36: > row.names(typh)
     [1] "1951" "1952" "1953" "1954" "1955" "1956" "1957" "1958"  ...
    [51] "2001" "2002" "2003"
37: > factor(typh$total) -> n.typh
38: > summary(n.typh)
     0  1  2  3  4  5  6
     3  6 13 15  9  5  2
39: > plot(n.typh)
40: > hist(typh$total, br=0:8-.5, axes=F, prob=T); axis(1, 0:7); axis(2)
41: > mean(typh$total)
    [1] 2.830189
42: > lines(0:7, dpois(0:7, lambda=2.83), type="b")
43: > legend(6,.3, "Fit", lty=1, pch="o")
```

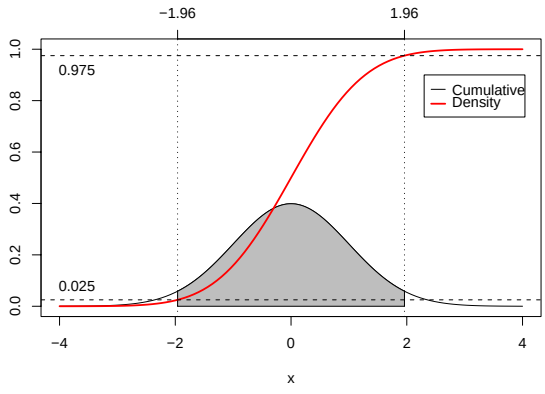
```
> # 7 #
> ?phyper
> ?pbinom
> ?ppois
> ?pgeom
> ?pnbinom

> # 8 #
> ?pnorm
> ?pexp
> ?pgamma
> ?pbeta
> ?punif
> ?pcauchy
> ?plnorm
> ?pweibull

> # advanced
> # package: evd & mvtnorm
> # download them from:
> # http://www.r-project.org/
> #
> library(evd)
> ?pgpd
> ?pgev
> ?pgumbel
> library(mvtnorm)
> ?pmvnorm
>
> # distributions in Chapter 10
> ?pchisq
> ?pt
> ?pf

> # more advanced
> ?plogis
> ?pwilcox

> # 9 #
> curve(dnorm, -4,4,
+ ylab="", ylim=c(0,1))
> x <- seq(-1.96,1.96, by=.02)
> polygon(c(-1.96,x,1.96),
+         c(0,dnorm(x),0),
+ col="gray")
> curve(pnorm, -4,4,
+ col="red", lwd=2, add=T)
> abline(h=c(.025, .975), lty=2)
> abline(v=1.96*c(-1,1),  lty=3)
> text(rep(-3.7, 2), c(.08,.92),
+ c("0.025","0.975"))
> axis(3, 1.96*c(-1,1))
> legend(2.3,.9,
+ c("Cumulative","Density"),
+ lwd=1:2, col=c("black","red"))
```



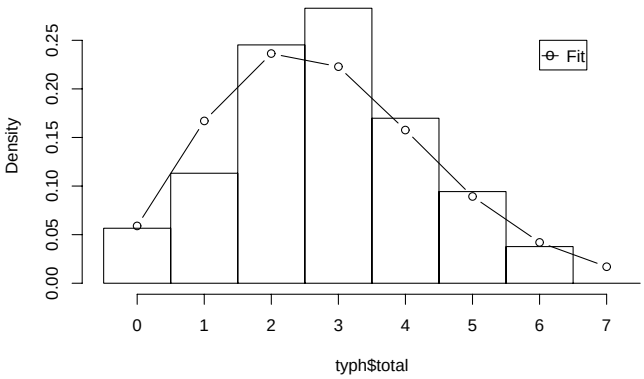




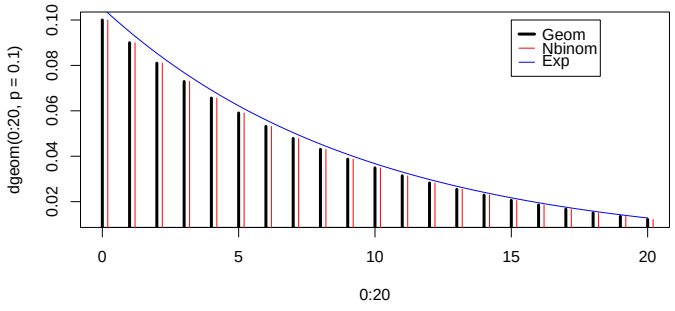


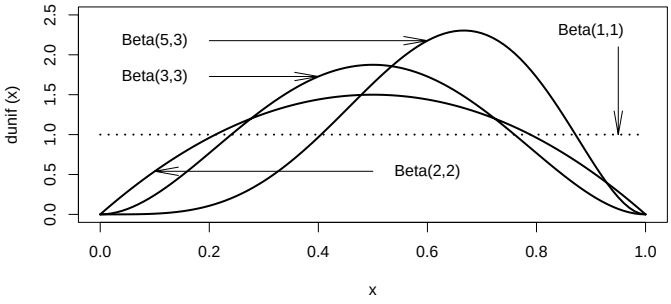


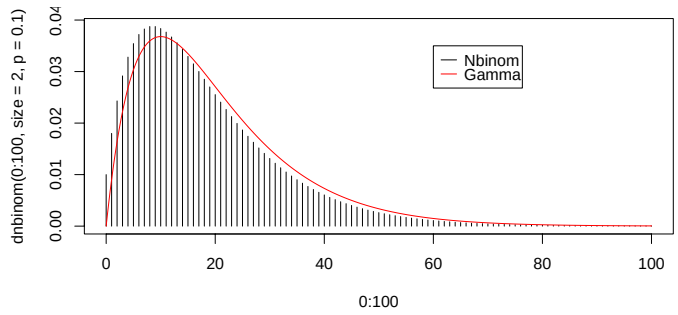
Histogram of typh$total

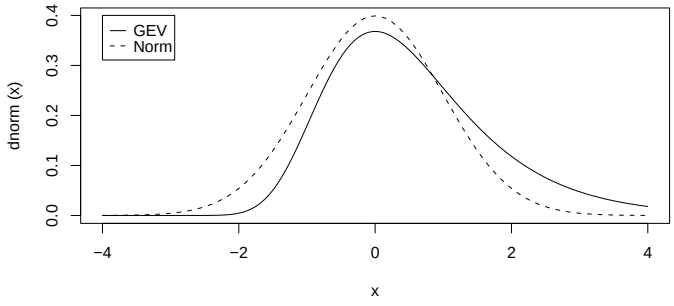


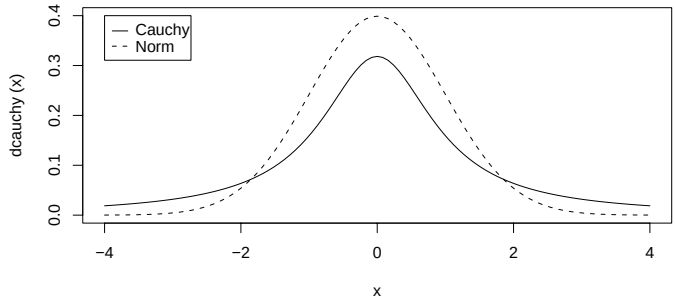
```
> # 10 #
> x <- seq(0,20, by=.01)
> plot(0:20, dgeom(0:20, p=.1), type="h", lwd=3)
> lines(x, dexp(x, rate=-log(1-.1)), col="blue")
> lines(0:20 +.2, dnbinom(0:20, size=1, p=.1), type="h", col="red")
> legend(15, .1, c("Geom", "Nbinom", "Exp"),
+ lwd=c(3,1,1), col=c("black", "red", "blue"))

> # 11 #
> x <- seq(0,100, by=.1)
> plot(0:100, dnbinom(0:100, size=2, p=.1), type="h")
> lines(x, dgamma(x, shape=2, scale=1/.1), col="red")
> legend(60, .035, c("Nbinom", "Gamma"), lty=rep(1,2), col=c("black", "red"))

> # 12 # Fig. 6.14 in text p.127
> curve(dcauchy, -4,4, ylim=c(0,.4))
> curve(dnorm,   -4,4, lty=2, add=T)
> legend(-4,.4, c("Cauchy", "Norm"), lty=1:2)

> # 13 # Fig. 6.13 in text p.127
> x <- 0:100/100
> curve(dunif, 0,1, lty=3, lwd=2, ylim=c(0,2.5))
> lines(x, dbeta(x, 2,2), lwd=2); text(.6, dbeta(.1, 2,2), "Beta(2,2)")
> lines(x, dbeta(x, 3,3), lwd=2); text(.1, dbeta(.4, 3,3), "Beta(3,3)")
> lines(x, dbeta(x, 5,3), lwd=2); text(.1, dbeta(.6, 5,3), "Beta(5,3)")
> arrows(.5, dbeta(.1, 2,2), .1, dbeta(.1, 2,2), angle=10)
> arrows(.2, dbeta(.4, 3,3), .4, dbeta(.4, 3,3), angle=10)
> arrows(.2, dbeta(.6, 5,3), .6, dbeta(.6, 5,3), angle=10)
> text(.9, 2.3, "Beta(1,1)"); arrows(.95, 2.1, .95, 1.0, angle=10)

> # 14 #
> curve(dnorm, -4,4, lty=2)
> lines(x, dgev(x))
> legend(-4,.4, c("GEV", "Norm"), lty=1:2)

> # 15 #
> curve(dgpd, 0.001,5, xlim=c(-1,5))
> curve(dgev,    -1,5, lty=3, add=T)
> legend(3,1, c("GPD", "GEV"), lty=c(1:3))
```

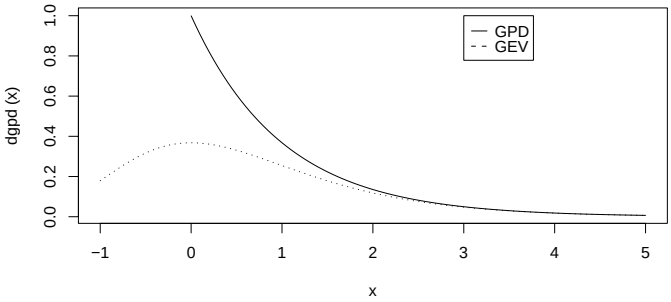












【７・８章】 多次元の確率分布，大数の法則と中心極限定理

《おもな多次元確率分布》

＊ ２次元正規分布　　◆ 密度関数 （式 (7.34) p.147）

・同時確率分布，同時密度関数，周辺分布，条件付密度関数
・共分散（共分散をそれぞれの標準偏差で除すことにより規格化したものが，相関係数）
・相関係数（正規分布であれば，相関係数＝ゼロ：無相関は，独立を意味する）
独立であれば無相関であるけれども，逆のこと（＝無相関であれば独立）は，一般的に成立しない.

＊一般の多次元分布 について，ここでは，その確率変数の和についての性質を検討する.

◎ 同一の確率分布に従う独立な確率変数の和 S_n の期待値および分散（式 (7.40)）

（コラム p.149 も参照）

A) $$E(S_n) = n\mu, \quad V(S_n) = n\sigma^2$$

◎ 同一の確率分布に従う独立な確率変数の相加平均 $\overline{X}$ の期待値および分散（式 (7.41)）

B) $$E(\overline{X}) = \mu, \quad V(\overline{X}) = \frac{\sigma^2}{n}$$

・相加平均（＝ 標本平均）： $\overline{X}_n = S_n / n$ ； $S_n = \sum X_i = X_1 + X_2 + \cdots + X_n$

○ チェビシェフの不等式（p.105）

C) $$\text{Prob.}\left(|X - \mu| < k\sigma \right) \geq 1 - \frac{1}{k^2}$$

☆ 大数の法則（p.157, 160）

標本平均は，標本サイズ n を十分に大きくとれば，
母平均 μ にいくらでも近い値をとることが確率的に高いことを示す.

D) $$\text{Prob.}\left(|\overline{X}_n - \mu| < \varepsilon \right) \to 1 \quad (n \to \infty)$$

★ 中心極限定理（大数の法則の精密化，p.162）

E, F) $$S_n \sim N\left(n\mu, n\sigma^2 \right) \text{ or } \quad \overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n} \right) \text{ or}$$

G) $$\text{Prob.}\left(a < |\overline{X}_n - \mu| < b \right) \to \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx \quad (n \to \infty)$$

【９章】 標本分布

・記述統計（p.17）と推測統計（p.176） ・母集団と標本（p.176）
・パラメトリックとノンパラメトリック（p.179, 180）
・母数と統計量（p.182） 注：統計量は標本の要約値であり，未知パラメータを含まない
・標本平均（式 (9.3) p.183，式 (7.41) p.149）
・標本分散（記述統計の分散（式 (2.10) p.37）とは異なる．不偏分散（式 (9.6) p.184）を用いる）

【１０章】 正規分布から抽出された標本（の標本分布とその性質）

６章では，具体的な現象（モデル化したい現象）を記述する分布を扱ったのに対し，ここでは，モデル化により生じる推測誤差（＝統計的変動）の分布について扱う．ここでは，最も基本となる推測誤差が正規分布に従う場合をとりあげ，その標本分布の性質がテーマである．

なお，具体的現象を記述する分布の場合には，分布（あるいは，密度関数）の関数表現を明示しているけれど，本章で扱う標本分布については，１）その表現が複雑であること，２）データ解析で実際的に必要となるのは，分布関数の表現式ではなく，分布の性質（特に，パーセント点など）であることから，ここでは，その関数表現をあえて示していない．

１）標本分散の標本分布：

標本分散 s^2 について，母分散 σ^2 と標本サイズ n を用いて表される量

$$\chi^2 = (n-1)s^2/\sigma^2$$

は，自由度 $n-1$ のカイ自乗分布に従う．

a) カイ自乗分布の期待値は，自由度に等しい．

b) 大数の法則から，どういうことがいえるか？考えよ．

２）標本平均の標本分布：

２－１）分散が既知の場合

正規分布の再生性より，標本平均 $\overline{X}$ は正規分布に従う．

２－２）分散が未知の場合

母平均 μ に対し，標本平均 $\overline{X}$ を標本分散 s^2 により規格化して表される量

$$t = \frac{\overline{X} - \mu}{\sqrt{s^2/n}}$$

は，自由度 $n-1$ の ティ分布（スチューデント分布ともいう）に従う．

a) 標本サイズの増大とともに，ティ分布は，正規分布に近付く．

b) ２標本問題（２つの標本の母平均が一致するか，否か？という問題）に応用．

３）標本分散の比の標本分布：

母分散が等しいことを前提にすれば，

２つの標本分散 s_1^2 および s_2^2 （標本サイズ： m および n ）の比

$$F = \frac{s_2^2}{s_1^2}$$

は，自由度 $(m-1, n-1)$ のエフ分布（フィッシャー分布ともいう）に従う．

a) 単刀直入に，母分散が一致するか，否か？という問題に応用．

b) さらに，回帰が有効か，否か？（回帰により説明できる変動の分散が，誤差分散に等しい場合は，その回帰は有効でないはず）

４）標本相関係数の標本分布：

フィッシャーの z 変換により，変換した量が正規分布に近似的に従う．

a) 無相関の検定に応用．

b) 厳密には，式 (1.2) p.7 に従う．