1 章練習問題解答例

【練習問題 1】

偏差平方和 $SS = \sum_{i=1}^{n}(x_i - \bar{x})^2$は、$\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2$のように簡便な式で算出できるが、

簡便式を導く過程を自ら導きなさい。

【解答例】

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \quad \text{なぜなら、} 2\bar{x}\sum_{i=1}^{n} x_i = 2\bar{x}n\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) \\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 \quad \text{なぜなら、} n\bar{x}^2 = n\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2
\end{aligned}
$$

【練習問題 2】

あるマウス系統の離乳時体重 (g) を以下に示した (データは省略)。これらのデータをもとにして、平均値、偏差平方和、分散、標準偏差、変動係数および標準誤差を求めなさい。

【解答例】

観測値数　：　$n = 10$

平均値　：　$\bar{x} = (9.4 + 9.2 + \cdots + 9.2)/10 = 9.30$

偏差平方和　：　$SS_x = 9.4^2 + 9.2^2 + \cdots + 9.2^2 - (93.0)^2/10 = 865.26 - 864.90 = 0.36$

不偏分散　：　$v_x = 0.36/(10-1) = 0.04$

標準偏差　：　$SD = \sqrt{v_x} = \sqrt{0.04} = 0.20$

変動係数　：　$CV = 0.2/9.3 = 0.022$

標準誤差　：　$SE = 0.2/\sqrt{10} = 0.063$

【R のコード・出力例】

```
> x <- c(9.4,9.2,9.0,9.0,9.5,9.4,9.6,9.3,9.4,9.2)
> cat("データ数=",length(x),"\n",
+     "平均値=",mean(x),"\n",
+     "偏差平方和=",sum((x-mean(x))^2),"\n",
+     "不偏分散=",var(x),"\n",
+     "標準偏差=",sd(x),"\n",
+     "変動係数=",sd(x)/mean(x),"\n",
+     "標準誤差=",sd(x)/sqrt(length(x)),"\n")
データ数= 10
 平均値= 9.3
 偏差平方和= 0.36
 不偏分散= 0.04
 標準偏差= 0.2
 変動係数= 0.02150538
 標準誤差= 0.06324555
>
```

【補足事項】

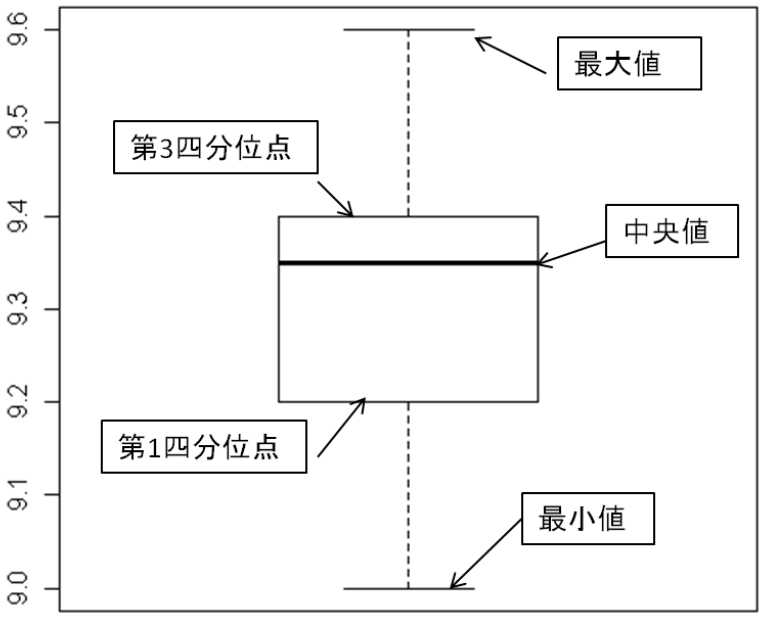
R の関数 summary( ) を使えば

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00    9.20    9.35    9.30    9.40    9.60
>
```

のように、平均値 (Mean)、中央値 (Median)、最小値 (Min.)、最大値 (Max.)、第 1(1st Qu.) および第 3 四分位数 (3rd Qu.) が求まる。また、それらの統計量の関係は

```
> boxplot(x)
```

と入力することにより下記のような箱ひげ図 (ボックスプロット) により簡単に表示できる。



中央値が平均値より上側 (第 3 四分位数寄り) にあり、分布が左側 (小さい方) に裾を引いた分布をしていることが分かる。

なお、箱の高さは範囲 (第 3 四分位数 − 第 1 四分位数) を示している。箱から上下に描かれた点線の先端の横棒は範囲 × 1.5 倍を超えない観察値中の最大値および最小値 (図中の最大値および最小値が対応している) の観察値を示しており、これを超える観察値ははずれ値としてその数値がプロットされる。

【練習問題 3】

血糖値を測定するための検量線を作成する目的で、2 名の測定者により比色計 (OD=540) を用いて、グルコースを定量含む標準試薬の吸光度を測定し、この 2 名の測定者による観測値を表に示した (データは省略)。補正値 (実測値からブランクを引いた値) を用いて、各測定者による検量線を描き、測定者によりどのような違いがあるのか、また、検量線を実用に供するようにするにはどのようにすればよいかを述べなさい。

【解答例】

R を用いて測定者 1(m1) と測定者 2(m2) による吸光度の実測値をブランクにより補正した補正値により検量線を描き、両者を比較する。

【R のコード・出力例】

```
> st <- c(0,50,100,150,200,300)
> st
[1]   0  50 100 150 200 300
> (m1 <- c(0,0.062,0.120,0.146,0.180,0.280))
[1] 0.000 0.062 0.120 0.146 0.180 0.280
> (m2 <- c(0,0.056,0.082,0.156,0.223,0.290))
[1] 0.000 0.056 0.082 0.156 0.223 0.290
>
```

※上記のように式全体を括弧で囲むと左辺の変数の値が自動的に出力される。

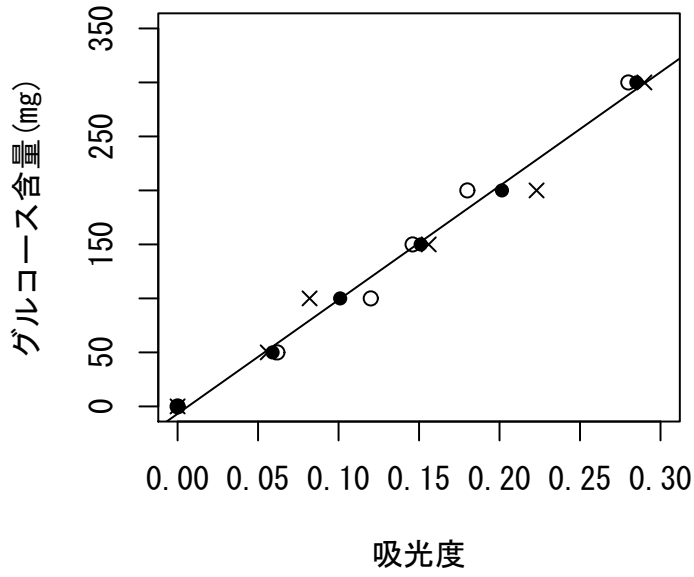
```
> (m1 <- c(0,0.062,0.120,0.146,0.180,0.280))  と
> m1 <- c(0,0.062,0.120,0.146,0.180,0.280)
> m1
```

は同じ結果となる。

グルコース含量と両者の吸光度をプロットすると図のようになり、測定者 1(○) は測定者 2(×) より低濃度で吸光度を高め，高濃度で低めに測定する傾向にあることがわかる。これだけのデータではどちらがより真の値を代表しているかは判断できないので、両者の平均をとり、プロット (●) し、1 次直線を当てはめ実線で表示した。

この処理を施すための R のコードを以下に示した。

```
> plot(m1,st,xlim=c(0,0.3),ylim=c(0,350),type="n",xlab="absorbance",
+      ylab="glucose(mg)")
> points(m1,st,pch=1,cex=1)
> points(m2,st,pch=4,cex=1)
>
> m<-(m1+m2)/2
> points(m,st,pch=16,cex=1)
> abline(lm(st~m))
```



濃度が未知の検体については、吸光度を測定し、このグラフからグルコース含量を読み取り、推定すればよいが、読み取り時の誤差が生じる危険性もある。

実際には、吸光度と標準試薬含量との関係を最もよく説明できる 1 次回帰式として検量線を導出して、吸光度に対応するグルコース含量を推定することが一般である。詳細は 8 章の相関と回帰の章を学んで自ら回帰式を導いてもらいたいが、R の関数 lm( ) を用いれば切片 (intercept) と回帰係数が求まる。すなわち、図中の●をつなぐ直線は、

$$\text{グルコース含量の推定値}\ (mg) = 1054.626 \times \text{吸光度} - 6.844$$

の 1 次回帰式を描いたものである。この処理のための R のコードを以下に示す。

```
> lm(st~m)

Call:
lm(formula = st ~ m)

Coefficients:
(Intercept)            m
     -6.844     1054.626

>
```

また、得られた回帰式の検量線としての精度 (実測値と推定値の相関 (重相関係数、決定係数あるいは寄与率と呼ばれる)：Multiple R-squared=99.74%) は以下のように評価できる。

```
> summary(lm(st ~ m))

Call:
lm(formula = st ~ m)
```

```
Residuals:
      1       2       3       4       5       6
 6.8441 -5.3789  0.3268 -2.4045 -5.6631  6.2756

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.844      4.394  -1.558    0.194
m           1054.626     27.051  38.986 2.59e-06 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 6.187 on 4 degrees of freedom
Multiple R-squared: 0.9974,     Adjusted R-squared: 0.9967
F-statistic:  1520 on 1 and 4 DF,  p-value: 2.586e-06

>
```

以上、興味のある方は試してほしい。

【練習問題 4】

3 品種の小麦を 1 区画 10a の圃場で栽培し、区画当たりの収穫量 (kg/10a) を調べた。区画および品種ごとの収穫量に関する代表値ならびに標準偏差を求め、それぞれの要因ごとにどのような差異があるか検討しなさい。ただし、1 区画の圃場内での品種はランダムに配置されているとする。

【解答例】

R を用いて、まず、品種ごとのデータをベクトル (変数) に格納し、要約統計量、箱ヒゲ図、散布図などを描き、色々な角度からデータの特長を検討する。

【R のコード・出力例】

まず、品種ごとのデータをベクトル (変数) に格納する。

```
A<-c(8,14,12,8,16,11)
B<-c(9,11,10,7,11,9)
C<-c(16,17,14,12,18,13)
```

個々の変数を用いて、平均値と標準偏差はそれぞれ、mean( )、sd( ) で計算できる。たとえば品種 A についての平均値と標準偏差は以下のように計算する。

```
> mean(A)
[1] 11.5
> sd(A)
[1] 3.209361
```

R には非常に便利な関数が多数揃っている。このような複数の項目からなるデータをひとまとめにして扱うにはデータフレームが便利である。そこで、ベクトル A、B、C をまとめてひとつのデータフレーム (breed) に格納する。

```
> breed <- data.frame(A,B,C)
> breed
   A  B  C
1  8  9 16
2 14 11 17
3 12 10 14
4  8  7 12
5 16 11 18
6 11  9 13
>
```

作成したデータフレームを用いて要約統計量を求めてみる。summary(データフレーム名) とすると、3 品種の要約統計量が一度に求まる。

```
> summary(breed)
       A               B               C
 Min.   : 8.00   Min.   : 7.00   Min.   :12.00
 1st Qu.: 8.75   1st Qu.: 9.00   1st Qu.:13.25
 Median :11.50   Median : 9.50   Median :15.00
 Mean   :11.50   Mean   : 9.50   Mean   :15.00
 3rd Qu.:13.50   3rd Qu.:10.75   3rd Qu.:16.75
 Max.   :16.00   Max.   :11.00   Max.   :18.00
>
```

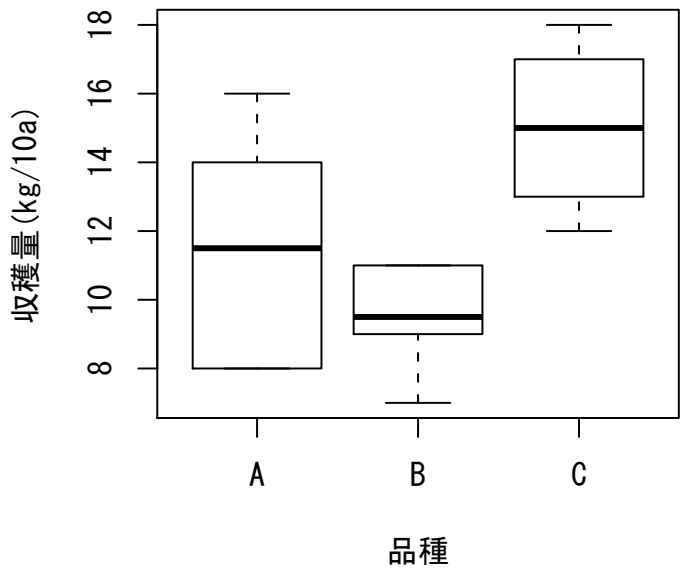
要約統計量では、値範囲 (最小値・最大値)、四分位数、メディアン、平均値が一度に表示される。標準偏差については要約統計量の出力に含まれていないので、sd(データフレーム名) として求める。

```
> sd(breed)
       A        B        C
3.209361 1.516575 2.366432
>
```

平均値を比較すれば、品種 C の収穫量が最も優れ，品種 B が劣っていることが分かる。変動係数については、sd( ) と mean( ) を組み合わせて、以下のように計算する。

```
> sd(breed)/mean(breed)*100
       A        B        C
27.90749 15.96395 15.77621
>
```

変動係数の結果から、品種 A の変異が大きいことがわかる。また、このような関係は以下のようにグラフ化 (箱ヒゲ図：boxplot を利用) すれば容易に把握できる。




```
> boxplot(breed,xlab="品種",ylab="収穫量 (kg/10a)")
```

同様に、区画ごとの平均値や標準偏差も mean( )、sd( ) を用いて以下のように求めることができる。

```
> b1<-c(8,9,16)
> mean(b1)
```

```
[1] 11
> sd(b1)
[1] 4.358899
>
```

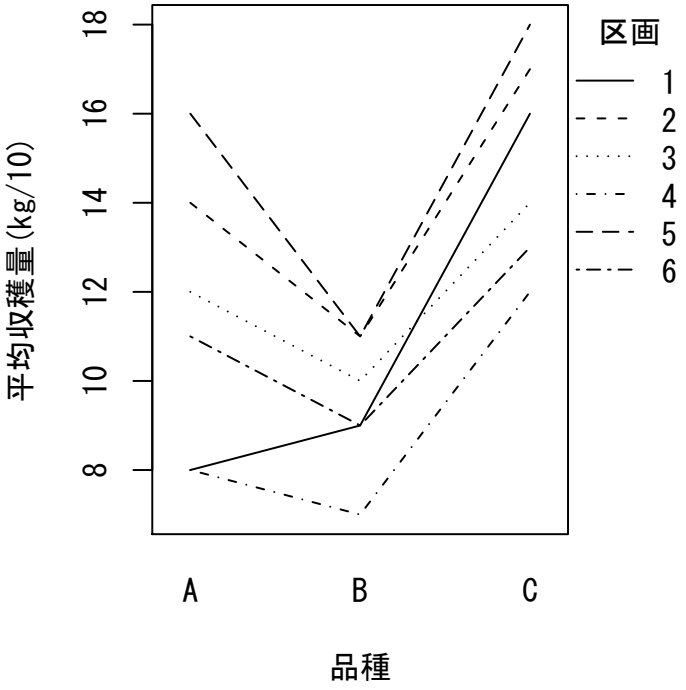
なお、区画ごとのデータベクトルを別途入力しなくても、データフレームを行列に変換して、for ループを用いると、区画ごとの平均値、標準偏差は以下のようにして、一度に求めることができる。

```
> lot <- as.matrix(breed)
> for(i in 1:6){
+   cat("区画",i,"\t 平均値=",mean(lot[i,]),"\t 標準偏差=",sd(lot[i,]),"\n")
+ }
区画 1  平均値= 11        標準偏差= 4.358899
区画 2  平均値= 14        標準偏差= 3
区画 3  平均値= 12        標準偏差= 2
区画 4  平均値= 9         標準偏差= 2.645751
区画 5  平均値= 15        標準偏差= 3.605551
区画 6  平均値= 11        標準偏差= 2
>
```

次に、品種と区画の収穫量への影響を把握するために、R を使って別のアプローチを試してみよう。まず、ベクトル A、B、C から収穫量のベクトル gain を作成し、gain の各要素に対応した品種のラベル (nbreed) と区画のラベル (block) のベクトルをそれぞれ作成する。

```
> gain <- c(A,B,C)
> gain
 [1]  8 14 12  8 16 11  9 11 10  7 11  9 16 17 14 12 18 13
> nbreed <- c(rep("A",6),rep("B",6),rep("C",6))
> nbreed
 [1] "A" "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "B" "C" "C" "C" "C" "C" "C"
> block <- rep(seq(1:6),3)
> block
 [1] 1 2 3 4 5 6 1 2 3 4 5 6 1 2 3 4 5 6
>
```

品種と区画ごとの収穫量の関係を下記のように interaction.plot( ) を用いて図示すると、品種と区画の収穫量への影響が容易に把握できる。




```
> interaction.plot(nbreed,block,gain)
```

この図から、品種 C は区画 1、3、6 において他の品種の収穫量と区画ごとの順位が入れ代わっていることが一目瞭然である。品種と区画の組み合わせによって収穫量が異なる影響がうかがわれるが、品種の特性を客観的に評価するためには、7 章で学ぶ分散分析法を適用することが必要である。

2章練習問題解答例

【練習問題 1】

ある樹木の葉のふちには棘があり、1 枚の葉あたりの棘数には 10 本から 16 本の変異がある。各棘数の葉が占める割合について、次の表のような数値が分かっている (データは省略)。葉あたりの棘数の平均値と分散を求めなさい。分散の計算方法は、本章のコラムを参照すること。

【解答例】

平均値：

平均値は、棘数の期待値から

$$\mu = 10 \times 0.09 + 11 \times 0.10 + \cdots + 16 \times 0.07 = 12.94$$

として得られる。

分散：

分散は、平均値からの偏差の 2 乗の期待値に等しい。

したがって

$$\begin{aligned}\sigma^2 &= (10 - 12.94)^2 \times 0.09 + (11 - 12.94)^2 \times 0.10 + \cdots + (16 - 12.94)^2 \times 0.07 \\ &= \quad 2.77\end{aligned}$$

となる。

【R のコード・出力例】

```
> togesu <- c(10, 11, 12, 13, 14, 15, 16)
> prop <- c(0.09, 0.10, 0.21, 0.23, 0.19, 0.11, 0.07)
> toge.mean <- sum(togesu*prop)
> toge.mean
[1] 12.94
> toge.var <- sum(((togesu-toge.mean)^2)*prop)
> toge.var
[1] 2.6764
>
```

【練習問題 2】

次のデータ (データは省略) は、ある里山で採集したオオクワガタの雄 20 個体の体長である。この里山に生息するオオクワガタの雄 (母集団) の体長に関する平均値と分散の不偏推定量を求め、結果をアプリケーションソフトウェア R で確認しなさい。なお、採集は体長について無作為に行ったものとする。

【解答例】

標本平均値 $\bar{x}$ は母平均値 μ の不偏推定量なので、20 個のデータの平均値を求めればよい。すなわち

$$\begin{aligned} \text{平均値}:\bar{x} &= \frac{x_1 + x_2 + \cdots + x_{20}}{20} \\ &= \frac{51.2 + 63.5 + \cdots + 42.9}{20} \\ &= 50.645 \end{aligned}$$

である。

一方、分散の不偏推定量は偏差平方和を $n-1$(データ数 -1) で割った値となる。すなわち

$$\begin{aligned} \text{分散の不偏推定量}:\hat{\sigma}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\} \\ &= \frac{1}{19}\left\{51.2^2 + 63.5^2 + \cdots + 42.9^2 - \frac{(51.2 + 63.5 + \cdots + 42.9)^2}{20}\right\} \\ &= 39.24366 \end{aligned}$$

【R のコード・出力例】

上記の式に従って R で計算を行う。

```
> rawData <- c(51.2,63.5,48.9,58.1,55.4,41.7,40.8,53.1,60.5,51.5,
+              46.9,47.7,51.4,52.9,49.9,48.2,41.1,56.3,50.9,42.9)
> n.data <- length(rawData)
> x.mean <- sum(rawData)/n.data
> x.var <- sum((rawData - x.mean)^2)/(n.data-1)
> cat("平均値=",x.mean,"\t 分散=",x.var,"\n")
平均値= 50.645     分散= 39.24366
>
```

【補足事項】

分散の計算は

```
> x.var <- (sum(rawData^2)-sum(rawData)^2/n.data)/(n.data-1)
> x.var
[1] 39.24366
```

としても、同じ結果が得られる。また、R の関数 mean( )、var( ) を用いると平均値、不偏分散が直接求まる。

```
> mean(rawData)
[1] 50.645
> var(rawData)
[1] 39.24366
```

【練習問題 3】

上のオオクワガタのデータを cm 単位で表したら平均値と分散はどのように変化するかを考えなさい。また実際に平均値と分散を計算して、その考えが正しいことを確認しなさい。

【解答例】

cm 単位で表わしたデータを $x_{(cm)}$ とし、平均値と分散をそれぞれ $\bar{x}_{(cm)}$ と $\sigma^2_{(cm)}$ とする。また、mm 単位で表わしたデータを $x_{(mm)}$ とし、平均値と分散をそれぞれ $\bar{x}_{(mm)}$ と $\sigma^2_{(mm)}$ とする。

一般に a を定数、x を変数とすると、ax の期待値は $E[ax] = aE[x]$ である。したがって

$$\bar{x}_{(cm)} = E[x_{cm}] = E[x_{(mm)}/10] = E[x_{(mm)}]/10 = \bar{x}_{(mm)}/10$$

である。すなわち、cm 単位で表わしたときの平均値は mm 単位で表わしたときの平均値の 1/10 になる。

同様に

$$\begin{aligned}\sigma^2_{(cm)} &= E\left[\left(x_{(cm)} - \bar{x}_{(cm)}\right)^2\right] \\ &= E\left[\left(x_{(mm)}/10 - \bar{x}_{(mm)}/10\right)^2\right] \\ &= E\left[\left(x_{(mm)} - \bar{x}_{(mm)}\right)^2\right]/100 = \sigma^2_{(mm)}/100\end{aligned}$$

である。すなわち、cm 単位で表わしたときの分散は mm 単位で表わしたときの分散の 1/100 になる。

実際にオオクワガタのデータでも、【練習問題 2】と同様に計算すると上の結果が確認できる。

【R のコード・出力例】

```
> rawData <- c(5.12, 6.35, 4.89, 5.81, 5.54, 4.17, 4.08, 5.31, 6.05, 5.15,
+              4.69, 4.77, 5.14, 5.29, 4.99, 4.82, 4.11, 5.63, 5.09, 4.29)
> n.Data <- length(rawData)
> rawData
 [1] 5.12 6.35 4.89 5.81 5.54 4.17 4.08 5.31 6.05 5.15 4.69 4.77 5.14 5.29
[15] 4.99 4.82 4.11 5.63 5.09 4.29
> rawData.mm <- rawData*10
> rawData.mm
 [1] 51.2 63.5 48.9 58.1 55.4 41.7 40.8 53.1 60.5 51.5 46.9 47.7 51.4 52.9
[15] 49.9 48.2 41.1 56.3 50.9 42.9
> x.mean <- sum(rawData)/n.Data
> x.var <- sum((rawData-x.mean)^2)/(n.Data-1)
> xmm.mean <-sum(rawData.mm)/n.Data
> xmm.var <- sum((rawData.mm-xmm.mean)^2)/(n.Data-1)
> cat("平均値 (cm)=",x.mean,"\t 分散 (cm)=",x.var,"\n")
平均値 (cm)= 5.0645         分散 (cm)= 0.3924366
> cat("平均値 (mm)=",xmm.mean,"\t 分散 (mm)=",xmm.var)
平均値 (mm)= 50.645         分散 (mm)= 39.24366
>
```

この結果から、cm 単位の平均値は mm 単位の平均値の 1/10、cm 単位の分散は mm 単位の分散の 1/100 となっていることがわかる。

3 章練習問題解答例

【練習問題 1】

前翅の長さが平均 27.0mm、分散 1.21 の正規分布に従うモンシロチョウの集団において、前翅が 25.5mm 以下の個体が占める割合を求めなさい。また、この集団で前翅の長さについて上位 20%の個体群は、何 mm 以上の前翅の長さを示すか。まず、正規分布表を使って答えを求め、R を利用して答えが正しいことを確認しなさい。

【解答例】

変数 X の分布が、平均 μ、分散 σ^2 の一般の正規分布 $N(\mu, \sigma^2)$ に従う場合、以下に示した Z 変換を施すと、得られた変数 Z は標準正規分布 $N(0, 1)$ に従う。

$$Z \text{ 変換} : Z = \frac{X - \mu}{\sigma}$$

そこで、Z 変換を用いて、前翅の長さ 25.5mm を標準正規分布の値に変換する。

$$z_0 = \frac{25.5 - 27.0}{\sqrt{1.21}} = \frac{-1.5}{1.1} = -1.364$$

標準正規分布は左右対称の分布なので、標準正規分布表から 1.364 の値に対応した確率を求めると 0.0863 となる。すなわち、前翅が 25.5mm 以下の個体が占める割合は 8.63%である。

累積確率が 0.8 となる標準正規分布の切断点は 0.8416 である。Z 変換の式を X について解いた以下の式を用いて

$$\begin{aligned} X &= \mu + z\sigma \\ x &= 27.0 + 0.8416 \times \sqrt{1.21} = 27.0 + 0.8416 \times 1.1 = 27.93 \end{aligned}$$

すなわち、前翅の長さについて上位 20%の個体群は、27.93mm 以上の前翅の長さを示す。

【R のコード・出力例】

```
> z <- (25.5-27.0)/sqrt(1.21)
> pnorm(z)
[1] 0.08634102
> x <- 27.0+qnorm(0.8)*sqrt(1.21)
> x
[1] 27.92578
>
```

【練習問題 2】

上のモンシロチョウの集団を母集団として、20 個体を無作為抽出したときの前翅の長さの標本平均は、どのような分布に従うか。

【解答例】

平均 μ、分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従う母集団から無作為に抽出した標本の大きさ n の標本平均値 $\bar{x}$ は、平均 μ、分散 σ^2/n の正規分布 $N(\mu, \sigma^2/n)$ に従う。

すなわち、モンシロチョウの集団を母集団として、20 個体を無作為抽出したときの前翅の長さの標本平均値は平均 27.0、分散 1.21/20=0.0605 の正規分布 $N(27.0, 0.0605)$ に従う。

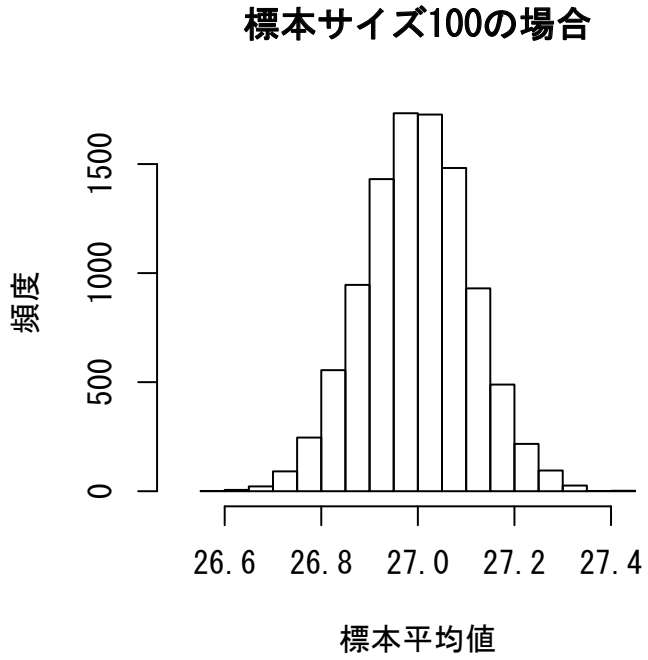
【練習問題 3】

以下に示した R のプログラム (R のコード・出力例を参照) は、練習問題 1 のモンシロチョウの母集団から 100 個体を無作為抽出して標本平均を求める試行を 10,000 回繰り返して行い、10,000 個の標本平均の平均値と分散を求めるとともに、10,000 個の標本平均値のヒストグラムを描くためのものである。このプログラムを実行して、得られた標本平均の平均値と分散を 3.4.1 で示した理論から期待される値と比較しなさい。また、同様の比較を、標本数を $n = 1,000$ とした場合についても行いなさい。

【解答例】

このプログラムを実行して得られた平均値は 26.99831、分散は 0.01217254 であった (100 個体を抽出した場合)。ただし、このプログラムでは正規乱数を用いているので、プログラムを実行するたびに値は異なったものとなる。

ヒストグラムを以下に示した。



標本サイズ (無作為抽出を行う個体数) を $n = 1,000$ とした場合の、平均値は 26.99944、標本平均の分散は 0.001202334 であった。ヒストグラムを次ページに示した。

3.4.1 で示した理論から期待される値は標本のサイズにかかわらず、標本平均値の期待値 (平均) は母平均 μ、分散は σ^2/n である。上記の結果では、標本サイズが 100 の場合 ($E[\bar{x}] = 26.99831$)、1,000 の場合 ($E[\bar{x}] = 26.99944$) とも、母平均 27.0 に非常に近い値となっている。

一方、分散の値は標本サイズが 100 の場合は 0.01217254、1,000 の場合は 0.001202334 であり、それぞれ 1.21/100、1.21/1000 に非常に近い値となっている。

【R のコード・出力例】

```
> 標本平均 <- numeric(length=10000)       # 10,000 個の標本平均を格納する場所を確保
> for(i in 1:10000){                      #{ }に囲まれた処理を 10,000 回繰り返す
+ 標本 <- rnorm(n=100,mean=27.0,sd=1.1)   # 100 個体の標本抽出
+ 標本平均[i] <- mean(標本)               # 標本平均を計算して格納
```

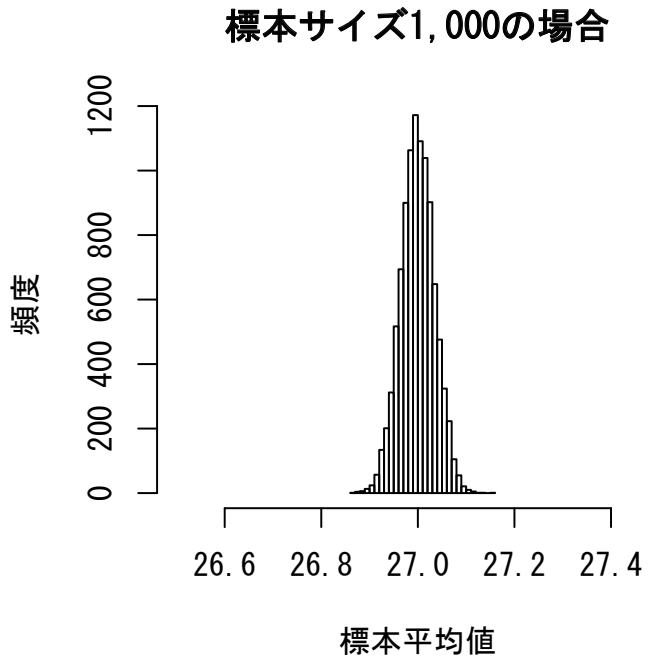
```
+ }
> mean(標本平均)                              #10,000 個の標本平均の平均値の計算
[1] 26.99831
> var(標本平均)                               #10,000 個の標本平均の分散の計算
[1] 0.01217254
> hist(標本平均,26,28,xlab="Sample Means",
+      ylab="Frequency",main=NULL)          #10,000 個の標本平均のヒストグラムを作成
>
```

【補足事項】

標本サイズを 1,000 に変えて実行する場合は、「標本 <- rnorm(n=100,mean=27.0,sd=1.1)」の部分を、「標本 <- rnorm(n=1000,mean=27.0,sd=1.1)」と変更する。

なお、このプログラムのように、変数名に日本語を用いてもよい。

標本サイズ 1,000 の場合のヒストグラムを以下に示す。標本サイズ 100 の場合に比べ分布が 27.0 の周りに集中していることが分かる。



【練習問題 4】

中国が原産の梅山豚 (メイシャントン) は、多産系のブタ品種として知られる。この品種の 1 頭の雌が 1 回の分娩で 13 頭の子ブタを産んだ。13 頭の子ブタのうち、5 頭が雄である確率を求めなさい。なお、性比は 1：1 とする。

【解答例】

二項分布を用いて、以下のように計算する。

$$\begin{aligned} p_5 = P(x=5) &= {}_{13}C_5(\frac{1}{2})^5(1-\frac{1}{2})^{(13-5)} \\ &= \frac{13!}{5!(13-5)!}(\frac{1}{2})^{13} \\ &= \frac{1287}{8192} = 0.1571 \end{aligned}$$

13 頭の子ブタのうち、5 頭が雄である確率は 0.1571 である。

【R のコード・出力例】

R では、${}_nC_k$ は choose(n,k) で求まる。

```
> choose(13,5)*(0.5)^5*(1-0.5)^(13-5)
[1] 0.1571045
>
```

【補足事項】

R には二項分布の確率関数 pbinom( ) が用意されている。

$$P(X \leq x) = pbinom(x, n, prob) = \sum_{k=1}^{x} {}_nC_k(prob)^k(1-prob)^{n-k}$$

すなわち、pbinom(x,n,prob) を用いると $X \leq x$ までの累積確率が求まる。
よって、13 頭の子ブタのうち、5 頭が雄である確率を求めるのであれば、

```
> pbinom(5,13,0.5)-pbinom(4,13,0.5)
[1] 0.1571045
>
```

とする。

【練習問題 5】

貯蔵豆類の害虫であるヨツモンマメゾウムシとブラジルマメゾウムシの 2 種について、成虫の雌をそれぞれの種ごとに、アズキ 50 粒を入れたシャーレのなかで 12 時間産卵させた。その後、アズキ 1 粒ごとに産卵された卵を数えて、下の表のような結果を得た (データは省略)。それぞれの種について、図 3.5 のような図を描き、ポアソン分布から期待される頻度と比較して、2 種の産卵習性について考察しなさい。

【解答例】

ヨツモンマメゾウムシ、ブラジルマメゾウムシの平均値、および産卵数の頻度は以下のようになる。

実測値から求めた頻度

		産卵数 (頻度)					
種　名	平均値	0	1	2	3	4	5
ヨツモンマメゾウムシ	1.00	0.12	0.76	0.12	0.00	0.00	0.00
ブラジルマメゾウムシ	0.66	0.52	0.34	0.10	0.04	0.00	0.00

一方、ポアソン分布の確率関数はヨツモンマメゾウムシ、ブラジルマメゾウムシの期待頻度を計算すると以下のようになる。なお、λ の値には、それぞれの平均値を用いる。

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda} \quad (k = 0, 1, \cdots, 5)$$

で与えられるので、

ポアソン分布から求めた期待頻度

		産卵数 (頻度)					
種　名	平均値	0	1	2	3	4	5
ヨツモンマメゾウムシ	1.00	0.3679	0.3679	0.1839	0.0613	0.0153	0.0031
ブラジルマメゾウムシ	0.66	0.5169	0.3411	0.1126	0.0248	0.0041	0.0005

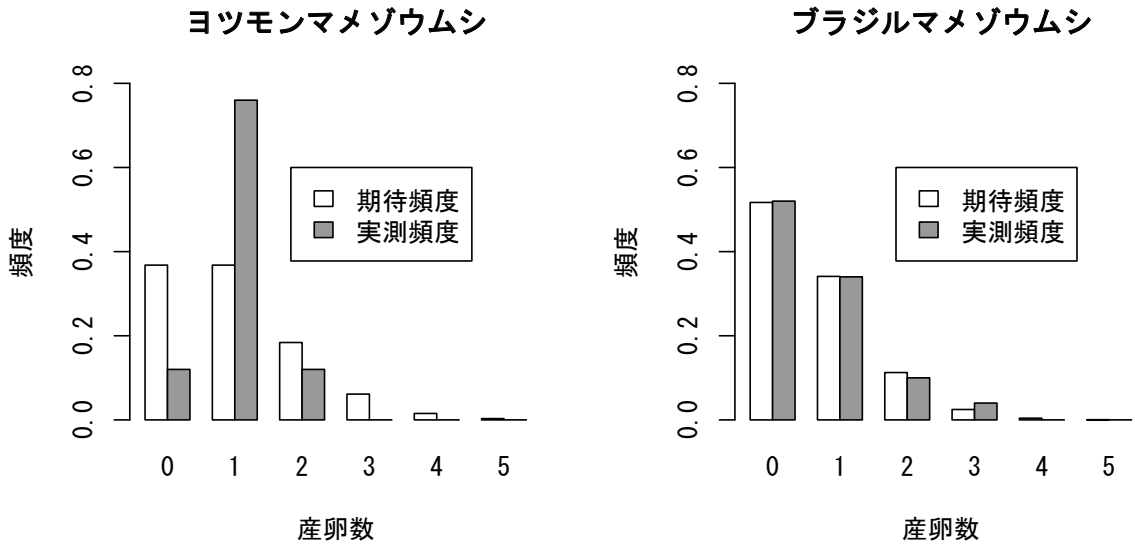
ヨツモンマメゾウムシとブラジルマメゾウムシの頻度をグラフに表すと次ページのようになる。

考察

まず、ヨツモンマメゾウムシのグラフから、本種はすべての豆に均等に産卵する傾向があることが分かる。すなわち、一度、卵を産んだ豆は避けて産卵する習性がある。このような習性は、成虫が産卵を規制する物質を豆に付着させることによるものと考えられている。一方、ブラジルマメゾウムシのグラフはポアソン分布に近く、本種がランダムに豆に産卵していることを示している。

【R のコード・出力例】

```
> x <- 0:5
> y.y <- c(6,38,6,0,0,0)
> y.b <- c(26,17,5,2,0,0)
> mean.y <- sum(x*y.y)/sum(y.y)
> mean.b <- sum(x*y.b)/sum(y.b)
> #平均値の出力
> mean.y
[1] 1
> mean.b
```




```
[1] 0.66
> #ポアソン分布を用いた頻度の計算
> p.y <- (mean.y)^x*exp(-mean.y)/factorial(x)
> p.b <- (mean.b)^x*exp(-mean.b)/factorial(x)
> #期待頻度の出力
> p.y
[1] 0.367879441 0.367879441 0.183939721 0.061313240 0.015328310 0.003065662
> p.b
[1] 0.5168513345 0.3411218808 0.1125702207 0.0247654485 0.0040862990
[6] 0.0005393915
> #実測頻度の計算
> f.y <- y.y/sum(y.y)
> f.b <- y.b/sum(y.b)
> #実測頻度の出力
> f.y
[1] 0.12 0.76 0.12 0.00 0.00 0.00
> f.b
[1] 0.52 0.34 0.10 0.04 0.00 0.00
> #グラフ出力
> layout(matrix(c(1,2),ncol=2,byrow=T))
> yFreq <- matrix(c(p.y,f.y),ncol=6,byrow=T)
> bFreq <- matrix(c(p.b,f.b),ncol=6,byrow=T)
> xname <- factor(0:5)
> col1 <- "white"
> col2 <- "gray60"
> barplot(yFreq,beside=T,names.arg=xname,
+   col=c(col1,col2),xlab="産卵数",ylab="頻度",
+   main="ヨツモンマメゾウムシ",ylim=c(0,0.8))
> legend(3*4,0.6,legen=c("期待頻度","実測頻度"),
```

```
+        fill=c(col1,col2),col=c("black","black"))
> barplot(bFreq,beside=T,names.arg=xname,
+   col=c(col1,col2),xlab="産卵数",ylab="頻度",
+   main="ブラジルマメゾウムシ",ylim=c(0,0.8))
> legend(3*4,0.6,legen=c("期待頻度","実測頻度"),
+        fill=c(col1,col2),col=c("black","black"))
>
>
```

4章練習問題解答例

【練習問題 1】

自由度 17 の t 分布で $t_0 \geq 2.110$ となるような確率はいくらか。

【解答例】

t 分布表の自由度 17 の行の値を見ると、右側確率 $p = 0.025$ の値が相当する。よって、$t_0 \geq 2.110$ となる確率は 0.025 である。

【R のコード・出力例】

```
> 1-pt(2.110,17)
[1] 0.02499106
>
```

【練習問題 2】

自由度 9 の t 検定において，両側検定を仮定し有意水準に 5%を設定した場合の棄却限界値の絶対値はいくらか。

【解答例】

t 分布表の、自由度 9 の行の値を見る。右側確率 $p = 0.025$ の値が 2.262 となっているので、棄却限界値の絶対値は 2.262 である。

【R のコード・出力例】

```
> qt(0.025,9)
[1] -2.262157
> qt(0.975,9)
[1] 2.262157
>
```

【練習問題 3】

自由度 5 の t 検定において，片側検定を仮定し有意水準に 1%を設定した場合の棄却限界値の絶対値はいくらか。

【解答例】

有意水準 1%の片側検定なので、自由度 5、右側確率 $p = 0.01$ の交わるところが棄却限界値の絶対値となる。すなわち、t(5; 0.01)=3.365 である。

【R のコード・出力例】

```
> qt(0.99,5)
[1] 3.36493
>
```

【練習問題 4】

トウモロコシの草丈 (cm) に関して二つの処理区より下記のデータを得た (データは省略)。A 区と B 区の草丈に差があるといえるか。5%水準で検定しなさい。

【解答例】

帰無仮説 H_0、対立仮説 H_1 を以下のように設定する。

帰無仮説 H_0：A 区と B 区の草丈に差はない。

対立仮説 H_1：A 区と B 区の草丈に差がある。

通常の t 検定

$$t_0 = \frac{307 - 252}{\sqrt{\frac{336+948}{7+6-2}(\frac{1}{7} + \frac{1}{6})}} = 9.150 > t(11; 0.05/2) = 2.201$$

※帰無仮説が破棄されるので、A 区と B 区の草丈に有意な差があるといえる。

ウェルチの t 検定

$$t_0 = \frac{307 - 252}{\sqrt{\frac{56}{7} + \frac{189.6}{6}}} = 8.740$$

$$\text{自由度}：\varphi = \frac{(\frac{56}{7} + \frac{189.6}{6})^2}{\frac{56^2}{7^2 \times (7-1)} + \frac{189.6^2}{6^2 \times (6-1)}} = 7.454$$

$t(7; 0.05/2) = 2.365$、$t(8; 0.05/2) = 2.306$、$q = \frac{\frac{1}{7.454} - \frac{1}{8}}{\frac{1}{7} - \frac{1}{8}} = 0.513$ より

$$t_0 = 8.740 > t(7.454; 0.05/2) = 0.513 \times 2.365 + (1 - 0.513) \times 2.036 = 2.336$$

※帰無仮説が破棄されるので、A 区と B 区の草丈に有意な差があるといえる。

【R のコード・出力例】

通常の t 検定

```
> a <- c(308,319,313,304,296,302,307)
> b <- c(236,260,245,275,251,245)
> a.mean <- mean(a)
> b.mean <- mean(b)
> a.var <- var(a)
> b.var <- var(b)
> a.n <- length(a)
> b.n <- length(b)
> a.ss <- sum((a-a.mean)^2)
> b.ss <- sum((b-b.mean)^2)
> v <- (a.ss+b.ss)/((a.n-1)+(b.n-1))
> t0 <- abs(a.mean-b.mean)/sqrt(v*(1/a.n+1/b.n))
> cat("A 区のデータ数、平均値、分散＝",a.n,a.mean,a.var,"\n")
A 区のデータ数、平均値、分散＝ 7 307 56
> cat("B 区のデータ数、平均値、分散＝",b.n,b.mean,b.var,"\n")
B 区のデータ数、平均値、分散＝ 6 252 189.6
```

```
> cat("A 区の偏差平方和=",a.ss,",B 区の偏差平方和=",b.ss,"\n")
A 区の偏差平方和= 336 ,B 区の偏差平方和= 948
> t0
[1] 9.150177
> pt(t0,a.n+b.n-2)
[1] 0.9999991
>
```

ウェルチの t 検定

```
> t0 <- abs(a.mean-b.mean)/sqrt(a.var/a.n+b.var/b.n)
> df <- (a.var/a.n+b.var/b.n)^2/(a.var^2/(a.n^2*(a.n-1))+(b.var^2/(b.n^2*(b.n-1))))
> q <- (1/df-1/(a.n+1))/(1/a.n-1/(a.n+1))
> t.0.975 <- q*qt(0.975,7)+(1-q)*qt(0.975,a.n+1)
> cat("t 値=",t0,", 自由度=",df,",t(7.454;0.05/2)=",t.0.975,"\n")
t 値= 8.740074 , 自由度= 7.453988 ,t(7.454;0.05/2)= 2.336062
>
```

【補足事項】

R の t.test( ) を用いてウェルチの検定を行うと以下のようになる。

```
> t.test(a,b)

        Welch Two Sample t-test

data:  a and b
t = 8.7401, df = 7.454, p-value = 3.546e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 40.30153 69.69847
sample estimates:
mean of x mean of y
      307       252
```

【練習問題 5】

ウシの呼吸数 (回/分) について 6 頭から下記のデータを得た (データは省略)。暑熱時の呼吸数は寒冷時より有意に多いといえるか。5%水準で検定しなさい。

【解答例】

帰無仮説 H_0、対立仮説 H_1 を以下のように設定する。

帰無仮説 H_0：暑熱時と寒冷時の呼吸数に差はない。

対立仮説 H_1：暑熱時の呼吸数は寒冷時より多い。

暑熱時と寒冷時の差は以下のようになる。

	個体 A	個体 B	個体 C	個体 D	個体 E	個体 F
暑熱時	37	35	32	38	34	36
寒冷時	20	25	19	23	24	21
差	17	10	13	15	10	15

$$t_0 = \frac{13.333}{\sqrt{\frac{8.267}{6}}} = 11.359 > t(5; 0.05) = 2.015$$

※暑熱時の呼吸数は寒冷時より有意に多いといえる。

【R のコード・出力例】

```
> hot <- c(37,35,32,38,34,36)
> cold <- c(20,25,19,23,24,21)
> diff <- hot-cold
> t0 <- mean(diff)/sqrt(var(diff)/length(hot))
> cat("t0=",t0,",t(5;0.05)=",qt(0.95,length(hot)-1),"\n")
t0= 11.35924 ,t(5;0.05)= 2.015048
>
```

【補足事項】

R の t.test( ) を使う場合は、上記の diff を引数に与える。

```
> t.test(diff)

        One Sample t-test

data:  diff
t = 11.3592, df = 5, p-value = 9.25e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 10.31602 16.35065
sample estimates:
mean of x
 13.33333
```

5 章練習問題解答例

【練習問題 1】

以下のデータ (データは省略) は、ソバのある品種の 20 個体について、1,000 粒の実の重さ (単位：g) を測った結果である。標本平均と不偏分散を求めなさい。さらに、母平均と母分散の信頼度 90%および 95%の信頼区間を求めなさい。

【解答例】

標本平均値は 13.56、不偏分散は 2.63 である。

母分散が未知として t 分布を用いて母平均の信頼区間を求めると、母平均の 95%信頼区間は [12.80086,14.31814]、90%信頼区間は [12.93276,14.18624] となる。

母分散の信頼区間は χ^2 分布を用いて、95%信頼区間は [1.519646,5.605322]、90%信頼区間は=[1.656206,4.934648] となる。

【R のコード・出力例】

```
> rawData <- c(12.61,14.51,14.37,13.10,12.14,13.70,11.70,14.31,15.39,13.56,
+              10.53,14.40,12.29,15.22,15.52,15.49,10.72,11.79,14.17,15.67)
> cat("標本平均値=",rawData.mean,"\t 不偏分散=",rawData.var,"\n")
標本平均値= 13.5595        不偏分散= 2.627573
> cnt <- length(rawData)
> df <- cnt-1
> #母平均に関する信頼区間の推定
> alpha5 <- 0.05
> alpha10 <- 0.10
> (t95 <- qt(1-alpha5/2,df))
[1] 2.093024
> (t90 <- qt(1-alpha10/2,df))
[1] 1.729133
> ci.95 <- t95*sqrt(rawData.var/cnt)
> ci.90 <- t90*sqrt(rawData.var/cnt)
> cat("母平均：95%信頼区間=[",rawData.mean-ci.95,"\t,",rawData.mean+ci.95,"]\n")
母平均：95%信頼区間=[ 12.80086  , 14.31814 ]
> cat("母平均：90%信頼区間=[",rawData.mean-ci.90,"\t,",rawData.mean+ci.90,"]\n")
母平均：90%信頼区間=[ 12.93276  , 14.18624 ]
>
> #母分散に関する信頼区間の推定
> chi5.l <- qchisq(alpha5/2,df)
> chi5.u <- qchisq(1-alpha5/2,df)
> var5.low <- df*rawData.var/chi5.u
> var5.upper <- df*rawData.var/chi5.l
> cat("母分散：95%信頼区間=[",var5.low,"\t,",var5.upper,"]\n")
母分散：95%信頼区間=[ 1.519646  , 5.605322 ]
>
> chi10.l <- qchisq(alpha10/2,df)
> chi10.u <- qchisq(1-alpha10/2,df)
> var10.low <- df*rawData.var/chi10.u
```

```
> var10.upper <- df*rawData.var/chi10.l
> cat("母分散：90%信頼区間=[",var10.low,"\t,",var10.upper,"]\n")
母分散：90%信頼区間=[ 1.656206  , 4.934648 ]
```

【補足事項】

R では t 検定のための関数 t.test を用いて、信頼区間を求めることもできる。本来 t.test は単一のデータベクトルに対しては帰無仮説：母平均=0 の検定のために用いられるが、出力の中に、信頼区間も表示される (特に引数で指定しなければ、両側検定、母分散未知の条件下での 95%信頼区間推定となる。90%信頼区間が必要な場合は、引数 conf.level=0.90 を指定する)。

$$t.test(rawData, conf.level = 0.95)$$

のように、引数としてデータベクトルと、信頼区間を与えると、以下に示すように 95 percent confidence interval の箇所に信頼区間が表示される。
例

```
> t.test(rawData,conf.level=0.95)

        One Sample t-test

data:  rawData
t = 37.4094, df = 19, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 12.80086 14.31814
sample estimates:
mean of x
  13.5595

> t.test(rawData,conf.level=0.90)

        One Sample t-test

data:  rawData
t = 37.4094, df = 19, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 12.93276 14.18624
sample estimates:
mean of x
  13.5595

>
```

【練習問題 2】

母分散 σ^2 が 1.0 である正規分布に従う母集団から標本抽出を行い、母平均の 95%信頼区間を推定する際、推定した区間の幅を 1.0 以下にするために必要な標本の大きさ n の最小値を求めなさい。

【解答例】

母分散が既知の場合、母平均 μ の信頼度 $100(1-\alpha)\%$の信頼区間は

$$\left[\bar{x} - z(\alpha/2)\frac{\sigma}{\sqrt{n}}, \bar{x} + z(\alpha/2)\frac{\sigma}{\sqrt{n}}\right]$$

として求めることができる。推定した区間の幅を w 以下にするためには、信頼区間の幅が対称であることから

$$2 \times z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq w$$

を満たす最小の n を求めればよい。

この式を変形すると

$$\begin{aligned} 2 \times z(\alpha/2)\frac{\sigma}{w} &\leq \sqrt{n} \\ \left(2 \times z(\alpha/2)\frac{\sigma}{w}\right)^2 &\leq n \\ 4 \times \{z(\alpha/2)\}^2\frac{\sigma^2}{w^2} &\leq n \end{aligned}$$

ここで、σ_2=1.0、w=1.0、$z(\alpha/2) = z(0.025) = 1.960$ より

$$\begin{aligned} n &\geq 4 \times 1.960^2\frac{1.0}{1.0^2} \\ &\geq 4 \times 3.8416 = 15.3664 \end{aligned}$$

すなわち、推定した区間の幅を 1.0 以下にするために必要な標本の大きさ n の最小値は 16 である。

【R のコード・出力例】

```
> sigma <- 1.0
> z <- qnorm(0.975)
> w <- 1.0
> n <- ceiling(4*z^2*sigma/w)
> n
[1] 16
>
```

【補足事項】

母分散 σ^2 が 1.0 である正規分布に従う母集団から標本の大きさ 16 の標本抽出を行い、母平均の 95%信頼区間を推定すると以下のようになる。

```
> sigma <- 1.0
> z <- qnorm(0.975)
> n <- 16
> w <- z*sqrt(sigma/n)
> cat("95%信頼区間=±",w,"\t 幅=",2*w,"\n")
```

```
95%信頼区間=± 0.489991          幅= 0.979982
> n <- 15
> w <- z*sqrt(sigma/n)
> cat("95%信頼区間=±",w,"\t 幅=",2*w,"\n")
95%信頼区間=± 0.5060605         幅= 1.012121
>
```

すなわち、区間の幅を 1.0 以下にするには標本の大きさは 16 以上でなければならない。

【練習問題 3】

母平均 10、母分散 4 の正規分布に従う母集団から大きさ 25 の標本を抽出したところ、標本平均は 10.20、不偏分散は 4.20 であった。母分散 4 が既知として母平均 μ の 95%信頼区間を推定しなさい。

【解答例】

練習問題 2 で示したように、母分散が既知の場合は標準正規分布が利用でき、95%信頼区間は、$z(0.025) = 1.960$ なので

$$\left[\bar{x} - 1.960 \times \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.960 \times \frac{\sigma}{\sqrt{n}}\right]$$

で求まる。

標本平均 $\bar{x} = 10.20$、母分散 $\sigma^2 = 4$、標本の大きさ 25 を代入して計算すると

$$\begin{aligned} 1.960 \times \frac{\sigma}{\sqrt{n}} &= 1.960 \times \frac{\sqrt{4}}{\sqrt{25}} \\ &= 1.960 \times \frac{2}{5} \\ &= 0.784 \end{aligned}$$

よって、95%信頼区間は $[10.20 - 0.784, 10.20 + 0.784]$、すなわち、$[9.416, 10.984]$ である。

【R のコード・出力例】

```
> sigma <- 4.0
> z <- qnorm(0.975)
> n <- 25
> w <- z*sqrt(sigma/n)
> ave <- 10.20
> cat("95%信頼区間=[",ave-w,"\t,",ave+w,"]\n")
95%信頼区間=[ 9.416014  , 10.98399 ]
>
```

【練習問題 4】

練習問題 3 について、母分散が未知として母平均および母分散の 95%信頼区間を推定しなさい。

【解答例】

母分散が未知なので、$z(\alpha/2)$ の代わりに $t(n-1;1-\alpha/2)$、母分散 σ^2 の代わりに不偏分散 $\hat{\sigma}^2$ を用いた、以下の式で信頼度 $100(1-\alpha)\%$の信頼区間を求める。

$$\left[\bar{x} - t(n-1;1-\alpha/2) \times \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + t(n-1;1-\alpha/2) \times \frac{\hat{\sigma}}{\sqrt{n}}\right]$$

標本平均 $\bar{x} = 10.20$、不偏分散 $\hat{\sigma^2} = 4.2$、標本の大きさ $n = 25$、$t(n-1;1-\alpha/2) = t(24;0.975) = 2.064$ を用いて計算すると

$$\left[10.20 - 2.064 \times \frac{\sqrt{4.2}}{\sqrt{25}}, 10.20 + 2.064 \times \frac{\sqrt{4.2}}{\sqrt{25}}\right] = [9.354, 11.046]$$

母分散既知の場合の 95%信頼区間 $[9.416, 10.984]$ に比べ範囲が広くなっているが、これは母分散の代わりに不偏分散 (これは、標本から推定している) を用いたことにより情報の精度が低下したためである。

【R のコード・出力例】

```
> sigma <- 4.2
> ave <- 10.20
> n <- 25
> t <- qt(0.975,n-1)
> w <- t*sqrt(sigma/n)
> cat("95%信頼区間=[",ave-w,"\t,",ave+w,"]\n")
95%信頼区間=[ 9.354053  , 11.04595 ]
>
```

【練習問題 5】

練習問題 3 について、標本の大きさが 25 ではなく 100 として抽出を行った結果が、標本平均 10.20、不偏分散 4.20 であったとして、練習問題 4 と同様に、母分散が未知として母平均および母分散の 95%信頼区間を推定しなさい。

【解答例】

標本の大きさ n を 25 から 100 に、また $t(n-1;1-\alpha/2)$ の値を $t(n-1;1-\alpha/2) = t(99;0.975) = 1.984$ 代えるだけで、練習問題 4 で用いた計算式、計算過程がそのまま利用できる。

すなわち

$$\left[10.20 - 1.984 \times \frac{\sqrt{4.2}}{\sqrt{100}}, 10.20 + 1.984 \times \frac{\sqrt{4.2}}{\sqrt{100}}\right] = [9.793, 10.607]$$

標本の大きさ $n = 25$ の母分散未知の場合の 95%信頼区間 $[9.354, 11.046]$ に比べ信頼区間の幅が狭くなっているが、これは、標本の大きさが大きくなったことにより母分散の推定精度が向上したことによる。また、標本の大きさ $n = 25$ の母分散既知の場合の 95%信頼区間 $[9.416, 10.984]$ に比べても範囲が狭くなっているが、標本平均は正規分布 $N(\mu, \sigma^2/n)$ に従うことから、標本の大きさが大きくなったことにより、標本平均の分布の分散が小さくなったことによる。

【R のコード・出力例】

```
> sigma <- 4.2
> ave <- 10.20
> n <- 100
> t <- qt(0.975,n-1)
> w <- t*sqrt(sigma/n)
> cat("95%信頼区間=[",ave-w,"\t,",ave+w,"]\n")
95%信頼区間=[ 9.793357  , 10.60664 ]
>
```

【練習問題 6】

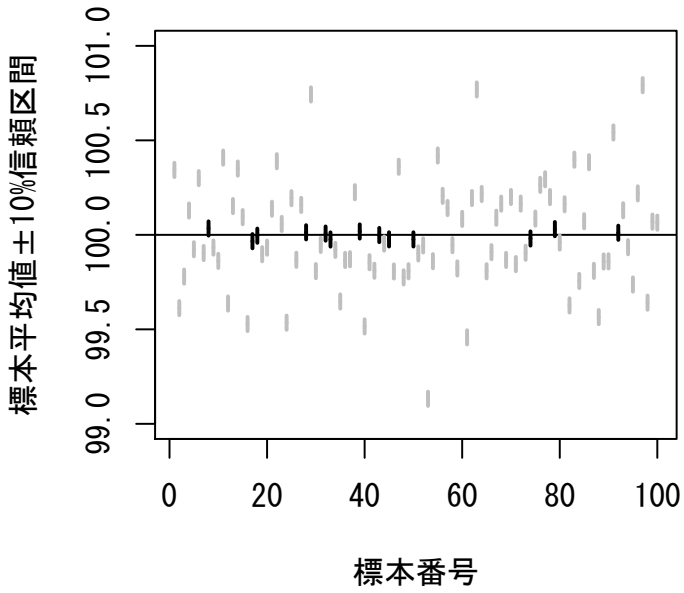
信頼区間の推定に用いる信頼水準としては、95%あるいは 99%が一般的である。信頼水準が小さいほど信頼区間の幅は小さくなるので、たとえば信頼水準として 10%を用いるほうがより狭い範囲で母平均の区間推定ができる。しかしながら、信頼水準 10%の信頼区間の推定は実用的ではない。その理由を考えなさい。

【解答例】

信頼度 $100(1-\alpha)\%$の信頼区間とは、標本の大きさを固定し、標本抽出を k 回繰り返して標本平均を k 個得たとき、これらの標本平均を使って求められた k 個の信頼区間のうち、母平均 μ を区間内に含むものが $k \times (1-\alpha)$ 個であることが期待されることを意味している。

たとえば、信頼度 10%の信頼区間を 100 個求めた場合、それらの信頼区間の内、母平均 μ を含むものは 10 個程度 (全体の 1 割程度) しかないことを意味しているので、いくら得られた信頼区間の範囲が狭いとしても、その区間内に母平均 μ が含まれている可能性は低いのでほとんど役に立たない。

母平均 100、母分散 9 の正規分布に従う母集団から標本の大きさ 100 の標本を 100 個求め、それぞれの標本平均と 10%信頼区間を求めた結果を以下に示す。



灰色のバーは、信頼区間に母平均 μ を含まないもの。
黒色のバーは、信頼区間に母平均 μ を含まれているもの

このグラフでも明らかなように大半の信頼区間には母平均 μ が含まれていない。たとえば、89 番目の標本から推定した 10%信頼区間は $[99.8214, 99.8968]$ であり、非常に狭い範囲であるが、母平均 $\mu = 100$ を含んでいない。

【R のコード・出力例】

```
> mu <- 100                       #母平均
> sigma2 <- 9                     #母分散
> sigma <- sqrt(sigma2)
> n <- 100                         #標本サイズ
> alpha=0.55
> z.value <- qnorm(alpha)         #切断点の値
> z.range <- z.value*sigma/sqrt(n) #
>
> n.rep <- 100                    #標本抽出の反復回数
> ave <- numeric(n.rep)           #標本平均のベクトル
> lower <- numeric(n.rep)         #下限値のベクトル
> upper <- numeric(n.rep)         #上限値のベクトル
> flag <- numeric(n.rep)          #範囲外のフラグ
> color <- c("black","red")
>
> ########  標本抽出の実行 #####
> for(i in 1:n.rep){
+   ave[i] <- mean(rnorm(n,mean=mu,sd=sigma))  #乱数を利用した標本平均の計算
+   lower[i] <- ave[i]-z.range                 #下限値の計算
+   upper[i] <- ave[i]+z.range                 #上限値の計算
+   if((lower[i]>mu)||(upper[i]<mu)){          #母平均が信頼区間に含まれるか判定
+     flag[i]=1
+   }
+ }
> cat("母平均を含まない件数=",sum(flag),"\n")
母平均を含まない件数= 90
>
> ########  結果のグラフ表示 ####
> x.axis <- 1:n.rep
> plot(x.axis,ave,ylim=c(99,101),xlab="標本番号",ylab="標本平均± 10%信頼区間",
+      pch="",col=color[flag+1])
> abline(100,0)
> for(i in 1:n.rep){
+  segments(i,ave[i],i,upper[i],col=color[flag[i]+1]) #信頼区間の上限値の表示
+  segments(i,ave[i],i,lower[i],col=color[flag[i]+1]) #信頼区間の下限値の表示
+ }
>
> upper[89]
[1] 99.8968
> lower[89]
[1] 99.8214
>
```

6 章練習問題解答例

【練習問題 1】

実験計画法の基礎となっているフィッシャーの 3 原則について説明しなさい。

【解答例】

実験誤差の評価と制御のための提唱された原則であり、1) 反復、2) 無作為化、3) 局所管理のことである。それぞれ、誤差分散の評価、定誤差の克服、定誤差の除去を目的としている。

【練習問題 2】

コムギ 5 品種について、さび病抵抗性を比較するための実験を行うものとする。すべての実験は同一の温室内で同一サイズの鉢 25 個を用いて同時に実施する。同じ温室内の鉢は比較的均一な条件で管理できるため、管理法の違いや鉢の配置などによる定誤差は生じないものと考えられる。完全無作為化法で 25 個の鉢に五つの品種を 5 回割り付けなさい。$*1$ を利用すること ($*1$ は省略)。

【解答例】

実験順序を無作為に決める方法

各処理がひとつひとつの実験単位に割当てられる確率が等しい条件のもとで実験の順序をきめることを実験順序の無作為化という。そのために一般に乱数 (random numbers) を利用することが多い。乱数を利用して鉢に 5 品種を配置する。

手順 1：25 個の鉢の位置を決め、一連番号をつけておく。

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

手順 2：25 個の乱数を発生させ、最初の 5 個を品種 1 に、次の 5 個を品種にと割り付ける。

R のコード・出力例に示した実行結果を整理すると、品種ごとの鉢の位置の対応は以下のようになる。

品種 A ： $5, 1, 15, 6, 3$

品種 B ： $23, 20, 16, 24, 13$

品種 C ： $4, 7, 8, 25, 12$

品種 D ： $2, 19, 21, 17, 14$

品種 E ： $11, 9, 10, 18, 22$

すなわち、

1 A	2 D	3 A	4 C	5 A
6 A	7 C	8 C	9 E	10 E
11 E	12 C	13 B	14 D	15 A
16 B	17 D	18 E	19 D	20 B
21 D	22 E	23 B	24 B	25 C

乱序数の作成に乱数を用いているので、常にこの結果が得られるわけではない。

【R のコード・出力例】

```
> n<-25
> data<-1:n
```

```
> x<-runif(n)
> y<-cbind(data,x[1:n])
> y<-y[order(y[,2]),]
> data<-y[,1]
> cat("乱序数:",data,"\n")
乱序数: 5 1 15 6 3 23 20 16 24 13 4 7 8 25 12 2 19 21 17 14 11 9 10 18 22
```

【練習問題 3】

オオムギ 5 品種の生産力検定試験を行うものとする。畑は南北 20m × 東西 50m であり、東から西に向けて水位が高くなっていることがわかっている。1 品種あたりの生産力を測定するためには 40m^2(4m × 10m) 程度の面積が必要であるとする。乱塊法による試験設計を行いなさい。ブロック内の品種の割付には $*1$ を利用すること ($*1$ は省略)。畑の周縁効果はないものとする。

【解答例】

手順１：

実験区画 (20m × 50m) を r 個の同一面積のブロックに分ける。ブロックは局所管理のためであり、ブロック内をできるだけ均一に管理し、定誤差の大部分をブロック間差として取り除くためである。この場合、東西の水位勾配に対して垂直に 5 ブロックに分けている。東西に対してブロックの長さができるだけ短いほうよいのであるが作業効率も考慮してブロック長は 10m とした。

手順 2：

各ブロック内を処理数に応じて区分する。ここでは品種数 5 に対応して 5 等分する。

I	II	III	IV	V
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5

この図は上側を北とした場合の各ブロック (I〜V) の配置、並びにブロック内の区画番号 (1〜5) を表している。

手順 3：

各ブロック内で品種 A,B,C,D,E をランダムに貼り付ける。「R のコード・出力例」に示すような方法で、ブロックごとに 1〜5 までの乱数を発生させ、順に A〜E の品種を割り当てる。「R のコード・出力例」に基づいた割り振りは以下のようになる。

I	II	III	IV	V
B	D	A	E	D
E	B	E	A	E
C	E	B	C	B
A	C	C	B	C
D	A	D	D	A

ここで練習問題 2 の完全無作為化との違いを見てみる。練習問題 2 の鉢の位置 1 から 5 を本練習問題における 1 ブロック目、以下順に 5 ずつ 1 つのブロックとして、21 から 25 を 5 ブロック目とする。練習問題 2 の配置では、1、2 ブロックにが 4 回、4、5 ブロック目に B が 4 回表れており、水位勾配の影響 (定誤差) が分別できないことになる。一方、本練習問題の場合、ブロック内には重複なく A〜E の 5 つの品種が無作為に割り当てられている。本練習問題で用いたブロック化の考え方は、グロースキャビネット等において多段を用いる場合など温度などの庫内環境の均一性が確保されないときもにも有効である。

【R のコード・出力例】

```
> block <- matrix(rep(0,25),ncol=5)
> n <- 5
> data <- 1:5
> breed <- c("A","B","C","D","E")
> for(i in 1:5){
+  x <- runif(n)
+  y <- cbind(data,x[1:n])
+  y <- y[order(y[,2]),]
+  block[,i] <- breed[y[,1]]
+ }
> block
     [,1] [,2] [,3] [,4] [,5]
[1,] "B"  "D"  "A"  "E"  "D"
[2,] "E"  "B"  "E"  "A"  "E"
[3,] "C"  "E"  "B"  "C"  "B"
[4,] "A"  "C"  "C"  "B"  "C"
[5,] "D"  "A"  "D"  "D"  "A"
```

【練習問題 4】

緬羊における飼料の嗜好性の評価方法の一つに採食量がある。緬羊を用いてオーチャードグラス 1 番乾草 (以下 A)、同 2 番乾草 (B)、アルファルファ1 番乾草 (C) および同 2 番乾草 (D) について自由採食量を測定することにした。採食量は 4 歳齢去勢緬羊 4 頭に各飼料を 1 期 7 日間, 朝夕 30 分間ずつ採食させ 1 日の平均採食量を算出した。4×4 ラテン方格法にしたがって試験設計を行いなさい。

【解答例】

4^2 型ラテン方格では以下の 4 種類の標準方格がある。標準型の方格を入れ替えると 4^2 型では 576 通りの方格を作ることができる。

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

A	B	C	D
B	A	D	C
C	D	B	A
D	C	A	B

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

A	B	C	D
B	D	A	C
C	A	D	B
D	C	B	A

4 種類の飼料に 4 頭を割り付ける。1 期 7 日間を 4 期繰り返す。576 種類から無作為に選んでも良いが、基準型を左端のものとし、行と列を無作為に入れ替えてもよい。

手順 1：行の入れ替え。乱数を用いて行を無作為に入れ替える。
手順 2：列の入れ替え。乱数を用いて列を無作為に入れ替える。

「R のコード・出力例」に示した方法で、行、列を入れ替えた結果を以下に示す。

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

C	D	A	B
B	C	D	A
D	A	B	C
A	B	C	D

A	B	C	D
D	A	B	C
B	C	D	A
C	D	A	B

【R のコード・出力例】

```
> #基本形のブロックの作成
> block <- matrix(c("A","B","C","D",
+                   "B","C","D","A",
+                   "C","D","A","B",
+                   "D","A","B","C"),ncol=4,byrow=T)
> block
     [,1] [,2] [,3] [,4]
[1,] "A"  "B"  "C"  "D"
[2,] "B"  "C"  "D"  "A"
[3,] "C"  "D"  "A"  "B"
[4,] "D"  "A"  "B"  "C"
> n <- 4
```

```
> nseq <- 1:4
> #乱数発生の関数
> randRC <- function(k,data)
+ {
+  x <- runif(k)
+  y <- cbind(data,x[1:k])
+  y <- y[order(y[,2]),]
+  return(y)
+ }
> yy <- randRC(n,nseq)
> yy[,1]
[1] 3 2 4 1
> #行の入れ替え
> block.row <- rbind(block[yy[1,1],],block[yy[2,1],],
+                    block[yy[3,1],],block[yy[4,1],])
> block.row
     [,1] [,2] [,3] [,4]
[1,] "C"  "D"  "A"  "B"
[2,] "B"  "C"  "D"  "A"
[3,] "D"  "A"  "B"  "C"
[4,] "A"  "B"  "C"  "D"
> yy <- randRC(n,nseq)
> yy[,1]
[1] 3 4 1 2
> #列の入れ替え
> block <- cbind(block.row[,yy[1,1]],block.row[,yy[2,1]],
+                block.row[,yy[3,1]],block.row[,yy[4,1]])
> #最終的な結果
> block
     [,1] [,2] [,3] [,4]
[1,] "A"  "B"  "C"  "D"
[2,] "D"  "A"  "B"  "C"
[3,] "B"  "C"  "D"  "A"
[4,] "C"  "D"  "A"  "B"
>
```

7 章練習問題解答例

【練習問題 1】

分散分析では、データの構造についていくつかの前提条件が仮定されている。それらをあげ、簡潔に説明しなさい。

【解答例】

分散分析における仮定は、以下のようなものである。

1) **線形性**：観測値は、全平均、因子の水準効果、誤差など、その構成要素の加算的な和として表される。すなわち、観測値は構成要素の線形式で表される。例えば、一元配置の分散分析では、観測値 y_{ij} の構成は、全平均を μ、因子の i 番目の水準の効果を g_i、i 番目の水準内の j 番目の誤差を e_{ij} で示せば、$y_{ij} = \mu + g_i + e_{ij}$ と仮定される。

2) **誤差の正規性**：誤差は、無限の大きさの正規母集団からの標本である。

3) **誤差の不偏性**：誤差の期待値は 0 である。

4) **誤差の等分散性**：誤差の母分散の大きさは、各水準について等しい。

5) **誤差の独立性**：因子の水準効果と水準内の誤差とは互いに独立であり、水準効果と誤差との間に関連性はない。

これらのうち、線形性と誤差の独立性は、分散分析を行ううえでとくに重要な前提条件である。

【練習問題 2】

因子 A の水準数が a で、各水準における実験の繰返し数が n の一元配置法による観測値を y_{ij} とする。このとき、総変動 $[SS_{\mathrm{T}} = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2]$ は、因子 A による変動 $[SS_{\mathrm{A}} = \sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})^2]$ と誤差変動 $[SS_{\mathrm{E}} = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2]$ とに分解され、$SS_{\mathrm{T}} = SS_{\mathrm{A}} + SS_{\mathrm{E}}$ が成り立つことを証明しなさい。
ヒント：$(y_{ij}-\bar{y}_{..}) = (\bar{y}_{i.}-\bar{y}_{..}) + (y_{ij}-\bar{y}_{i.})$ の両辺を平方し、総和 $(\sum_{i=1}^{a}\sum_{j=1}^{n})$ をとりなさい。また、$\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.}) = (y_{i1}-\bar{y}_{i.}) + (y_{i2}-\bar{y}_{i.}) + \cdots + (y_{in}-\bar{y}_{i.}) = (y_{i1}+y_{i2}+\cdots+y_{in}) - n\bar{y}_{i.}$ を利用しなさい。

【解答例】

$(y_{ij}-\bar{y}_{..}) = (\bar{y}_{i.}-\bar{y}_{..}) + (y_{ij}-\bar{y}_{i.})$ の両辺を平方すれば

$$(y_{ij}-\bar{y}_{..})^2 = (\bar{y}_{i.}-\bar{y}_{..})^2 + 2(\bar{y}_{i.}-\bar{y}_{..})(y_{ij}-\bar{y}_{i.}) + (y_{ij}-\bar{y}_{i.})^2$$

を得る。そこで、両辺についてすべての観測値の総和をとると

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})^2 + 2\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})(y_{ij}-\bar{y}_{i.}) + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$$

が成り立つ。ここで

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})(y_{ij}-\bar{y}_{i.}) = \sum_{i=1}^{a}\left\{(\bar{y}_{i.}-\bar{y}_{..})\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})\right\}$$

と表されるが

$$\begin{aligned}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.}) &= (y_{i1}-\bar{y}_{i.}) + (y_{i2}-\bar{y}_{i.}) + \cdots + (y_{in}-\bar{y}_{i.}) \\ &= (y_{i1}+y_{i2}+\cdots+y_{in}) - n\bar{y}_{i.} \\ &= (y_{i1}+y_{i2}+\cdots+y_{in}) - n\left(\frac{y_{i1}+y_{i2}+\cdots+y_{in}}{n}\right) \\ &= 0\end{aligned}$$

であるので、結果的に

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$$

と表され、$SS_{\mathrm{T}} = SS_{\mathrm{A}} + SS_{\mathrm{E}}$ が成り立つ。

【練習問題 3】

水稲の籾重に対する 3 種の肥料 (A_1、A_2 および A_3) の影響を調べるために、コシヒカリを用いて一元配置法によるポット試験を行い、個体 (株) あたりの籾重を計測して次のようなデータを得た (データは省略)。試験は、1 ポットあたり 1 株として温室内で行い、ポットの配置は完全無作為化法に従った。

(1) この場合の分散分析における帰無仮説と対立仮説を述べなさい。

(2) 肥料による変動の有意性を調べなさい。また、有意である場合には、テューキーの HSD 法による多重比較検定を行いなさい。

【解答例】

(1) 分散分析 (一元配置) における帰無仮説 H_0 と対立仮説 H_1 は、

帰無仮説 H_0:$\mu_{A_1} = \mu_{A_2} = \mu_{A_3}$(3 種の肥料群のいずれの平均のあいだにも差はない)
対立仮説 H_1:3 種の肥料群の平均のいずれかのあいだに差がある

である。

(2) 肥料の水準数、実験の繰返し数および観測値の総数を、それぞれ a、n および N とすると、$a = 3$、$n = 5$ および $N = a \times n = 15$ であり

$$\sum_{i=1}^{3}\sum_{j=1}^{5} y_{ij} = \sum_{i=1}^{3} y_{i\cdot} = y_{\cdot\cdot} = 12 + 12 + 10 + \cdots + 8 = 52 + 40 + 47 = 139$$

$$\sum_{i=1}^{3}\sum_{j=1}^{5} y_{ij}^2 = 12^2 + 12^2 + 10^2 + \cdots + 8^2 = 1319$$

である。よって、修正項は、$CT = y_{\cdot\cdot}^2/N = 139^2/15 = 1288.0666$ となるので

$$\text{総変動：} SS_{\mathrm{T}} = \sum_{i=1}^{3}\sum_{j=1}^{5} y_{ij}^2 - CT = 1319 - 1288.0666 = 30.9333$$

$$\text{肥料間変動：} SS_{\mathrm{A}} = \sum_{i=1}^{3} \frac{y_{i\cdot}^2}{5} - CT = \frac{1}{5}\left(52^2 + 40^2 + 47^2\right) - 1288.0666 = 14.5333$$

$$\text{誤差変動：} SS_{\mathrm{E}} = SS_{\mathrm{T}} - SS_{\mathrm{A}} = 30.9333 - 14.5333 = 16.4$$

であり、自由度は

$$SS_{\mathrm{T}}\text{の自由度} \ : \ \varphi_{\mathrm{T}} = N - 1 = 15 - 1 = 14$$
$$SS_{\mathrm{A}}\text{の自由度} \ : \ \varphi_{\mathrm{A}} = a - 1 = 3 - 1 = 2$$
$$SS_{\mathrm{E}}\text{の自由度} \ : \ \varphi_{\mathrm{E}} = \varphi_{\mathrm{T}} - \varphi_{\mathrm{A}} = 14 - 2 = 12$$

である。

肥料および誤差の平均平方は、それぞれ

$$MS_{\mathrm{A}} = SS_{\mathrm{A}}/\varphi_{\mathrm{A}} = 14.5333/2 = 7.2666$$
$$MS_{\mathrm{E}} = SS_{\mathrm{E}}/\varphi_{\mathrm{E}} = 16.4/12 = 1.3666$$

であるので、分散比は

$$F_0 = MS_{\mathrm{A}}/MS_{\mathrm{E}} = 7.2666/1.3666 = 5.3170$$

として求められる。

よって、$F(2,12;0.05)=3.89<F_0=5.3170<F(2,12;0.01)=6.93$ より、帰無仮説は棄却される。

以上の結果、分散分析表は次のように表示される。

分散分析表

変動因	自由度	偏差平方和	平均平方	分散比
肥　料	2	14.5333	7.2666	5.3170*
誤　差	12	16.4	1.3666	

* :p<0.05

ここでは、肥料による変動に有意性が認められたので、テューキーの HSD 法による多重比較検定を行う。

$$\sqrt{\frac{MS_E}{n}}=\sqrt{\frac{1.3666}{5}}=0.5228$$

であり、有意水準を 5%($\alpha=0.05$) とすれば、$a=3$、$a(n-1)=3(5-1)=12$ より、Q 表から $Q(3,12;0.05)=3.7729$ であるので、有意となる最小の差は

$$D(\alpha)=D(0.05)=0.5228\times 3.7729=1.9724$$

である。よって

$$\begin{aligned} D(0.05) &= 1.9724<|\bar{y}_{1.}-\bar{y}_{2.}|=|10.4-8|=2.4 \\ |\bar{y}_{1.}-\bar{y}_{3.}| &= |10.4-9.4|=1<1.9724 \\ |\bar{y}_{2.}-\bar{y}_{3.}| &= |8-9.4|=1.4<1.9724 \end{aligned}$$

であるので、A_1 群の平均値と A_2 群の平均値のあいだに有意差あり、A_1 群の平均値と A_3 群の平均値のあいだに有意差なし、A_2 群の平均値と A_3 群の平均値のあいだに有意差なし、と判定され、検定結果は次表のようにまとめられる。

肥料についての多重比較検定の結果

肥　料	A_1	A_2	A_3
平均値 (g)	10.4^a	8^b	9.4^{ab}

^{a,b} 同じ肩文字をもたない平均値のあいだに有意差あり (p<0.05)

【R のコード・出力例】

```
> fertilizer <-factor(c(rep("A1",5),rep("A2",5),rep("A3",5)))   #肥料の水準を指定
> record <- c(12,12,10,9,9,9,8,7,8,8,11,10,9,9,8)     #観測値を入力
> anova.result <- aov(record ~ fertilizer)           #分散分析を実行
> summary(anova.result)                             #分散分析の結果を表示
            Df  Sum Sq Mean Sq F value  Pr(>F)
fertilizer   2 14.5333  7.2667  5.3171 0.02221 *
Residuals   12 16.4000  1.3667
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1
> TukeyHSD(anova.result)                            #TukeyHSD を実行
  Tukey multiple comparisons of means
```

```
    95% family-wise confidence level

Fit: aov(formula = record ~ fertilizer)

$fertilizer
      diff       lwr        upr     p adj
A2-A1 -2.4 -4.372536 -0.4274641 0.0178296
A3-A1 -1.0 -2.972536  0.9725359 0.3949642
A3-A2  1.4 -0.572536  3.3725359 0.1828408

>
```

【補足事項】

R の関数 aov( ) を用いず、解答例で示した計算手順を R で実行した例を以下に示す。

```
> fertilizer <-factor(c(rep("A1",5),rep("A2",5),rep("A3",5)))
> record <- c(12,12,10,9,9,9,8,7,8,8,11,10,9,9,8)
> data.mat <- matrix(record,ncol=3,byrow=F)
> n.total <- length(record)
> CT <- sum(data.mat)^2/n.total
> SST <- sum(data.mat^2)-CT
> n.row <- dim(data.mat)[1]
> n.col <- dim(data.mat)[2]
> sum.col <- colSums(data.mat)
> SSA <- sum(sum.col^2)/n.row-CT
> SSE <- SST-SSA
> cat("修正項 (CT)=",CT,"\n","総変動 (SST)=",SST,"\n",
+     "肥料間変動 (SSA)=",SSA,"\n","誤差変動 (SSE)=",SSE,"\n")
修正項 (CT)= 1288.067
 総変動 (SST)= 30.93333
 肥料間変動 (SSA)= 14.53333
 誤差変動 (SSE)= 16.4
> df.T <- n.total-1
> df.A <- n.col-1
> df.E <- df.T-df.A
> MS.A <- SSA/df.A
> MS.E <- SSE/df.E
> F.value <- MS.A/MS.E
> F.Prob <- pf(F.value,df.A,df.E,lower.tail=F)
> cat("総変動の自由度=",df.T,"\t 肥料間変動の自由度=",df.A,
+     "\t 誤差変動の自由度=",df.E,"\n")
総変動の自由度= 14       肥料間変動の自由度= 2    誤差変動の自由度= 12
> cat("肥料間の平均平方=",MS.A,"\t 誤差の平均平方=",MS.E,"\n")
肥料間の平均平方= 7.266667       誤差の平均平方= 1.366667
> cat("F 値=",F.value,"\t 有意確率=",F.Prob,"\n")
F 値= 5.317073    有意確率= 0.02220766
> cat("分散分析表\n",
```

```
+     "-----------------------------------------------------------------\n",
+     "変動因\t 自由度\t 偏差平方和\t 平均平方\tF 値\t 有意確率\n",
+     "-----------------------------------------------------------------\n",
+     "肥　料\t",df.A,"\t",round(SSA,4),"\t",round(MS.A,4),"\t",
+      round(F.value,4),"\t",round(F.Prob,4),"\n",
+     "誤　差\t",df.E,"\t",round(SSE,4),"\t","\t",round(MS.E,4),"\n",
+     "-----------------------------------------------------------------\n")
分散分析表
 ------------------------------------------------------------------
 変動因 自由度   偏差平方和       平均平方        F 値      有意確率
 ------------------------------------------------------------------
 肥　料   2      14.5333         7.2667     5.3171     0.0222
 誤　差  12      16.4            1.3667
 ------------------------------------------------------------------
> SEA <- sqrt(MS.E/n.row)
> A.mean <- colMeans(data.mat)
>
> for(i in 1:2){
+   for(j in (i+1):3){
+     diff.abs <- abs(A.mean[i]-A.mean[j])
+     HSD <- ptukey(diff.abs/SEA,n.col,df.E,lower.tail=F)
+     cat(i," vs ",j,"\t 平均値の差の絶対値=",diff.abs,"\t 有意確率=",HSD,"\n")
+   }
+ }
1  vs  2        平均値の差の絶対値= 2.4 有意確率= 0.01782962
1  vs  3        平均値の差の絶対値= 1   有意確率= 0.3949642
2  vs  3        平均値の差の絶対値= 1.4 有意確率= 0.1828408
> cat("平均値=",A.mean,"\n")
平均値= 10.4 8 9.4
>
```

【練習問題 4】

6 品種のオオムギの収量 (kg/a) を比較した乱塊法による圃場試験のデータ (6 章の表 6.3) について、分散分析を実施しなさい (データは省略)。

【解答例】

このようなデータの分散分析は、繰り返しのない二元配置の分散分析法によって行う。

ブロックを因子 A、品種を因子 B とすると、ブロックの水準数は 3、品種の水準数は 6 であるので、$a = 3$、$b = 6$、$N = 18$ であり、$CT = y_{\cdot\cdot}^2/N = 609^2/18 = 20604.5$ より

$$
\begin{aligned}
SS_{\mathrm{T}} &= \sum_{i=1}^{3}\sum_{j=1}^{6} y_{\cdot\cdot}^2 - CT = (40^2 + 22^2 + 50^2 + \cdots + 49^2) - 20604.5 = 23961 - 20604.5 = 3356.5 \\
SS_{\mathrm{A}} &= \sum_{i=1}^{3} \frac{y_{i\cdot}^2}{6} - CT = \frac{1}{6}(215^2 + 205^2 + 189^2) - 20604.5 = 20661.8333 - 20604.5 = 57.3333 \\
SS_{\mathrm{B}} &= \sum_{j=1}^{6} \frac{y_{\cdot j}^2}{3} - CT = \frac{1}{3}(102^2 + 60^2 + \cdots + 156^2) - 20604.5 = 23823 - 20604.5 = 3218.5 \\
SS_{\mathrm{E}} &= SS_{\mathrm{T}} - SS_{\mathrm{A}} - SS_{\mathrm{B}} = 3356.5 - 57.3333 - 3218.5 = 80.6666
\end{aligned}
$$

と求められる。各変動の自由度は

$$
\begin{aligned}
\varphi_{\mathrm{T}} &= N - 1 = 18 - 1 = 17 \\
\varphi_{\mathrm{A}} &= a - 1 = 3 - 1 = 2 \\
\varphi_{\mathrm{B}} &= b - 1 = 6 - 1 = 5 \\
\varphi_{\mathrm{E}} &= \varphi_{\mathrm{T}} - \varphi_{\mathrm{A}} - \varphi_{\mathrm{B}} = 17 - 2 - 5 = 10
\end{aligned}
$$

であるので、平均平方は

$$
\begin{aligned}
MS_{\mathrm{A}} = SS_{\mathrm{A}}/\varphi_{\mathrm{A}} &= 57.3333/2 = 28.6666 \\
MS_{\mathrm{B}} = SS_{\mathrm{B}}/\varphi_{\mathrm{B}} &= 3218.5/5 = 643.7 \\
MS_{\mathrm{E}} = SS_{\mathrm{E}}/\varphi_{\mathrm{E}} &= 80.6666/10 = 8.0666
\end{aligned}
$$

であり、分散比は

$$
\begin{aligned}
F_{0(\mathrm{A})} &= MS_{\mathrm{A}}/MS_{\mathrm{E}} = 28.6666/8.0666 = 3.5537 \\
F_{0(\mathrm{B})} &= MS_{\mathrm{B}}/MS_{\mathrm{E}} = 643.7/8.0666 = 79.7975
\end{aligned}
$$

となる。そこで

$$
\begin{aligned}
F_{0(\mathrm{A})} &< F(2, 10; 0.05) = 4.10 \\
F(5, 10; 0.01) &< F_{0(\mathrm{B})} = 79.7975
\end{aligned}
$$

より、分散分析の結果は次のようにまとめられる。

分散分析表

変動因	自由度	偏差平方和	平均平方	分散比	
ブロック	2	57.3333	28.6666	3.5537	
品　　種	5	3218.5000	643.7000	79.7975	**
誤　　差	10	80.6666	8.0666		

** : $p < 0.01$

よって、ブロック間の変動には有意性は認められず、品種間の変動にのみ、1%水準で有意性が認められる。

なお、6 品種について、テューキーの HSD 法による多重比較検定 (有意水準：5%) を行った結果は、次のとおりである。

品種についての多重比較検定の結果

品　種 *	B_1	B_2	B_3	B_4	B_5	B_6
平均値	34^b	20^c	46^a	37^b	14^c	52^a

^{a,b,c} 同じ肩文字をもたない平均値間に有意差あり ($p<0.05$)

*:B_1：ミノリムギ、B_2：シュンライ、B_3：ミユキオオムギ、
B_4：東山皮 86 号、B_5：東山皮 93 号、B_6：新系 93 である。

【R のコード・出力例】

```
> block <- factor(c(rep(c("I","II","III"),6)))
> breed <- factor(c(rep("Mino",3),rep("Syun",3),rep("Miyu",3),
+                   rep("Hi86",3),rep("Hi93",3),rep("Sh93",3)))
> record <- c(40,32,30,22,22,16,50,46,42,35,38,38,13,15,14,55,52,49)
> Barley.data <- data.frame(Block=block,Breed=breed,Record=record)
> anova.result <- aov(Record ~ .,data=Barley.data)
> summary(anova.result)
            Df Sum Sq Mean Sq F value   Pr(>F)
Block        2   57.3    28.7  3.5537  0.06825 .
Breed        5 3218.5   643.7 79.7975 9.94e-08 ***
Residuals   10   80.7     8.1
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1
> breed.THSD <- TukeyHSD(anova.result,"Breed")
> breed.THSD
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Record ~ ., data = Barley.data)

$Breed
          diff         lwr        upr     p adj
Hi93-Hi86  -23 -31.0546408 -14.945359 0.0000185
Mino-Hi86   -3 -11.0546408   5.054641 0.7823975
Miyu-Hi86    9   0.9453592  17.054641 0.0269149
Sh93-Hi86   15   6.9453592  23.054641 0.0007364
Syun-Hi86  -17 -25.0546408  -8.945359 0.0002627
Mino-Hi93   20  11.9453592  28.054641 0.0000647
Miyu-Hi93   32  23.9453592  40.054641 0.0000009
Sh93-Hi93   38  29.9453592  46.054641 0.0000002
Syun-Hi93    6  -2.0546408  14.054641 0.1867750
Miyu-Mino   12   3.9453592  20.054641 0.0040618
Sh93-Mino   18   9.9453592  26.054641 0.0001617
```

```
Syun-Mino  -14 -22.0546408  -5.945359 0.0012728
Sh93-Miyu    6  -2.0546408  14.054641 0.1867750
Syun-Miyu  -26 -34.0546408 -17.945359 0.0000060
Syun-Sh93  -32 -40.0546408 -23.945359 0.0000009

>
```

【練習問題 5】

次のデータ (データは省略) は、アフリカのある種のハエを効率よく捕捉するための野外設置型トラップについて、デザイン (4 種：A, B, C, D) を処理因子とし、採集地および採集日を二つのブロック因子とするラテン方格法による調査記録である。ここでの観測値の値は、採集個体数を c として、$c+1$ を常用対数変換した値として与えられている。このデータについて分散分析を行い、トラップ・デザインの効果の有意性を検定しなさい。

【解答例】

二つのブロック因子を採集地 (行) と採集日 (列)(それぞれ因子 A および B) とすると、$n = 4$、$N = n \times n = 4 \times 4 = 16$ であり、修正項は、$CT = y_{\cdot\cdot}^2/N = 34^2/16 = 72.25$ である。

そこで、各変動は

$$
\begin{aligned}
SS_{\mathrm{T}} &= \sum_{i=1}^{4}\sum_{j=1}^{4} y_{\cdot\cdot}^2 - CT = (2.8^2 + 1.6^2 + 2.2^2 + \cdots + 2.3^2) - 72.25 = 73.84 - 72.25 = 1.59 \\
SS_{\mathrm{A}} &= \sum_{i=1}^{4} \frac{y_{i\cdot}^2}{4} - CT = \frac{1}{4}(8.5^2 + 7.8^2 + 8.9^2 + 8.8^2) - 72.25 = 72.435 - 72.25 = 0.185 \\
SS_{\mathrm{B}} &= \sum_{j=1}^{4} \frac{y_{\cdot j}^2}{4} - CT = \frac{1}{4}(9.1^2 + 7.8^2 + 8.4^2 + 8.7^2) - 72.25 = 72.475 - 72.25 = 0.225 \\
SS_{\mathrm{C}} &= \sum_{k=1}^{4} \frac{y_{\cdot\cdot(k)}^2}{4} - CT = \frac{1}{4}(9.2^2 + 7.6^2 + 7.5^2 + 9.7^2) - 72.25 = 73.185 - 72.25 = 0.935 \\
SS_{\mathrm{E}} &= SS_{\mathrm{T}} - SS_{\mathrm{A}} - SS_{\mathrm{B}} - SS_{\mathrm{C}} = 1.59 - 0.185 - 0.225 - 0.935 = 0.245
\end{aligned}
$$

であり、自由度は

$$
\varphi_{\mathrm{A}} = \varphi_{\mathrm{B}} = \varphi_{\mathrm{C}} = n - 1 = 4 - 1 = 3、\ \varphi_{\mathrm{E}} = (n-1)(n-2) = (4-1)(4-2) = 6
$$

であるので、各平均平方は

$$
\begin{aligned}
MS_{\mathrm{A}} &= SS_A/\varphi_{\mathrm{A}} = 0.185/3 = 0.0616 \\
MS_{\mathrm{B}} &= SS_{\mathrm{B}}/\varphi_{\mathrm{B}} = 0.225/3 = 0.075 \\
MS_{\mathrm{C}} &= SS_{\mathrm{C}}/\varphi_{\mathrm{C}} = 0.935/3 = 0.3116 \\
MS_{\mathrm{E}} &= SS_{\mathrm{E}}/\varphi_{\mathrm{E}} = 0.245/6 = 0.0408
\end{aligned}
$$

として算出される。分散比は

$$
\begin{aligned}
F_{0(\mathrm{A})} &= MS_{\mathrm{A}}/MS_{\mathrm{E}} = 0.0616/0.0408 = 1.5102 \\
F_{0(\mathrm{B})} &= MS_{\mathrm{B}}/MS_{\mathrm{E}} = 0.075/0.0408 = 1.8367 \\
F_{0(\mathrm{C})} &= MS_{\mathrm{C}}/MS_{\mathrm{E}} = 0.3116/0.0408 = 7.6326
\end{aligned}
$$

と求めらるので

$$
\begin{aligned}
F_{0(\mathrm{A})} &= 1.5102 < F(3,6;0.05) = 4.76 \\
F_{0(\mathrm{B})} &= 1.8367 < F(3,6;0.05) = 4.76 \\
F(3,6;0.05) &= 4.76 < F_{0(\mathrm{C})} = 7.6326 < F(3,6;0.01) = 9.78
\end{aligned}
$$

より、分散分析の結果は次のようにまとめられる。

分散分析表

変動因	自由度	偏差平方和	平均平方	分散比
採集地 A	3	0.185	0.0616	1.5102
採集日 B	3	0.225	0.0750	1.8367
トラップ C	3	0.935	0.3116	7.6326*
誤　　差	6	0.245	0.0408	

*:p<0.05

よって、二つのブロック因子 (採集地および採集日) のそれぞれによる変動に有意性は認められないが、トラップ・デザインの効果には 5%水準で有意差が認められる。

そこで、トラップ・デザインの平均値間の差の検定をテューキーの HSD 法によって行えば、以下のような結果を得る。

トラップ・デザインについての多重比較検定の結果

トラップ・デザイン	A	B	C	D
平均値 *	2.300^{ab}	1.900^{b}	1.875^{b}	2.425^{a}

^{a,b} 同じ肩文字をもたない平均値間に有意差あり (p<0.05)

*:常用対数変換値

【R のコード・出力例】

```
> daycode <- factor(c(rep("Day1",4),rep("Day2",4),rep("Day3",4),rep("Day4",4)))
> site <- factor(c(rep(c("Site1","Site2","Site3","Site4"),4)))
> designcode <- factor(c("D","B","A","C",
+                        "C","D","B","A",
+                        "A","C","D","B",
+                        "B","A","C","D"))
> record <-c(2.8,1.8,2.4,2.1,1.6,2.1,1.8,2.3,2.2,1.6,2.5,2.1,1.9,2.3,2.2,2.3)
> Fly.data <- data.frame(Site=site,Day=daycode,Design=designcode,Record=record)
> anova.result <- aov(Record ~.,data=Fly.data)
> summary(anova.result)
            Df  Sum Sq Mean Sq F value  Pr(>F)
Site         3 0.18500 0.06167  1.5102 0.30485
Day          3 0.22500 0.07500  1.8367 0.24102
Design       3 0.93500 0.31167  7.6327 0.01799 *
Residuals    6 0.24500 0.04083
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1
> Design.HSD <- TukeyHSD(anova.result,"Design")
> Design.HSD
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Record ~ ., data = Fly.data)

$Design
      diff         lwr        upr     p adj
B-A -0.400 -0.89463321 0.09463321 0.1094514
```

```
C-A -0.425 -0.91963321 0.06963321 0.0887120
D-A  0.125 -0.36963321 0.61963321 0.8179008
C-B -0.025 -0.51963321 0.46963321 0.9978834
D-B  0.525  0.03036679 1.01963321 0.0392225
D-C  0.550  0.05536679 1.04463321 0.0322404

>
```

8 章練習問題解答例

【練習問題 1】

1 章の練習問題 3 のデータにおいて，2 人の測定者による実測値間の相関係数を求めなさい。

【解答例】

小数点以下を 3 桁までにして求めると

$$r = \frac{0.301 - \frac{1.226 \times 1.221}{6}}{\sqrt{(0.297 - \frac{1.226^2}{6})(0.308 - \frac{1.221^2}{6})}} = 0.979$$

【R のコード・出力例】

```
> x <- c(0.073,0.135,0.193,0.219,0.253,0.353)
> y <- c(0.069,0.125,0.151,0.225,0.292,0.359)
> xy.cov <- sum(x*y)-sum(x)*sum(y)/length(x)
> x.var <- sum(x*x)-sum(x)^2/length(x)
> y.var <- sum(y*y)-sum(y)^2/length(y)
> r.xy <- xy.cov/sqrt(x.var*y.var)
> cat("相関係数=",r.xy,"\n")
相関係数= 0.9742607
```

R の相関係数を求める関数 cor( ) を用いると

```
> cor(x,y)
[1] 0.9742607
```

【練習問題 2】

下表 (表は省略) は平成 7 年から 10 年間の東北地方と九州地方の水稲の作況指数を示したものである。これら 2 地域の作況指数を使い相関係数を求めなさい。

【解答例】

$$r = \frac{97566 - \frac{987 \times 988}{10}}{\sqrt{(97877 - \frac{987^2}{10})(98156 - \frac{988^2}{10})}} = 0.101$$

【R のコード・出力例】

ここでは、共分散を求める R の関数 cov( ) と分散を求める R の関数 var( ) を用いて

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

の計算式から相関係数を求めてみる。

```
> x <- c(96,103,103,97,103,104,102,101,80,98)
> y <- c(106,104,100,103,85,103,104,102,96,85)
> r <- cov(x,y)/sqrt(var(x)*var(y))
> cat("相関係数=",r,"\n")
相関係数= 0.1009637
```

R の相関係数を求める関数 cor( ) を用いると

```
> cor(x,y)
[1] 0.1009637
```

【練習問題 3】

練習問題 2 の作況指数が有意な相関関係にあるのかどうか検定しなさい。

【解答例】

帰無仮説 H_0、対立仮説 H_1 を以下のように設定する。

帰無仮説 H_0：東北と九州の作況指数は相関関係にない。

対立仮説 H_1：東北と九州の作況指数は相関関係にある。

検定統計量 t_0 は、相関係数を r、データ数 (ペアの数) を n とすると以下の式で与えられる。

$$t_0 = r\sqrt{\frac{n-2}{1-r^2}}$$

すなわち

$$t_0 = 0.101 \times \sqrt{\frac{10-2}{1-0.101^2}} = 0.287 < t(8; 0.05/2) = 2.306$$

※東北と九州の作況指数は相関関係にない。

【R のコード・出力例】

以下の計算に用いたベクトル x とベクトル y は問題 2 と同じものである。

```
> r <- cor(x,y)
> t0 <- r*sqrt((length(x)-2)/(1-r^2))
> cat("検定統計量 t0=",t0,",t(8;0.05/2)=",qt(0.975,length(x)-2),"\n")
検定統計量 t0= 0.2870351 ,t(8;0.05/2)= 2.306004
>
```

【補足事項】

R の関数 cor.test( ) を用いると、相関係数の推定と有意性の検定が行える。

計算例

```
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = 0.287, df = 8, p-value = 0.7814
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5645508  0.6869227
sample estimates:
      cor
0.1009637
```

【練習問題 4】

1%水準で有意な相関関係にあるというためには、15 組のデータより得た相関係数の場合、絶対値はいくら以上必要か求めなさい。

【解答例】

自由度 $\varphi = 13$ より、$t(13; 0.01/2)$=3.012 である。したがって、棄却限界値に達する相関係数は

$$r\sqrt{(\frac{15-2}{1-r^2})} = 3.012$$

を r について解けばよい。すなわち、r の絶対値としては 0.641 以上必要となる。

$$\begin{aligned} r\sqrt{\frac{\varphi}{1-r^2}} &= t_0 \text{とおくと} \\ r^2(\frac{\varphi}{1-r^2}) &= t_0^2 \\ r^2\varphi &= t_0^2 - r^2 t_0^2 \\ (\varphi + t_0^2)r^2 &= t_0^2 \\ r &= t_0\sqrt{\frac{1}{\varphi + t_0^2}} \end{aligned}$$

【R のコード・出力例】

上記の式変形を用いて R で計算すると以下のようになる。

```
> df <- 15-2
> t0 <- qt(0.995,df)
> r <- t0/sqrt(t0^2+df)
> r
[1] 0.6411448
```

【練習問題 5】

肉用牛における分娩後 30 日までの母ウシの平均乳量 (kg/日) とその子ウシの体重の増加量 (kg/日) について 6 組の親子から下記のデータを得た (データは省略)。母ウシの乳量を独立変数、子ウシの体重の増加量を従属変数として回帰式を求めなさい。

【解答例】

回帰係数 (b)、y 切片 (a) は以下の式で与えられる。

$$
\begin{aligned}
b &= \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \\
a &= \hat{y} - b\hat{x}
\end{aligned}
$$

よって、表の数値を用いて以下のように回帰式が求まる。

$$
\begin{aligned}
b &= \frac{13.6 - \frac{24.5 \times 3.28}{6}}{102.07 - \frac{24.5^2}{6}} = 0.102 \\
a &= 0.547 - 0.102 \times 4.083 = 0.131 \\
\hat{y} &= 0.131 + 0.102x
\end{aligned}
$$

【R のコード・出力例】

```
> yield <- c(3.2,4.1,4.5,3.6,4.1,5.0)
> gain <- c(0.46,0.58,0.61,0.49,0.51,0.63)
> b <- cov(yield,gain)/var(yield)
> a <- mean(gain)-b*mean(yield)
> cat("y=",a,"+",b,"x\n")
y= 0.1306163 + 0.1018899 x
>
```

【練習問題 6】

練習問題 5 の回帰式の寄与率を求め、乳量と増体量が有意な回帰関係にあるか検定しなさい。さらに乳量が 4.0kg のときの子ウシの増体量の推定値とその 95%信頼区間を求めなさい。

【解答例】

全体 (SS_y)、回帰 (SS_{reg}) および誤差 (SS_e) の偏差平方和は以下のようになる。

$$\begin{aligned} SS_y &= \sum y_i^2 - \frac{(\sum y_i)^2}{n} \\ SS_{reg} &= b^2(\sum x_i^2 - \frac{(\sum x_i)^2}{n}) \\ SS_e &= SS_y - SS_{reg} \end{aligned}$$

また、寄与率 (R^2) は

$$R^2 = \frac{SS_{reg}}{SS_y}$$

表の数値を用いて、それぞれの偏差平方和を求めると以下のようになる。

$$\begin{aligned} SS_y &= 1.8172 - \frac{3.28^2}{6} = 0.024133 \\ SS_{reg} &= 0.102^2 \times (102.07 - \frac{24.5^2}{6}) = 0.021103 \\ SS_e &= 0.024133 - 0.021103 = 0.00303 \\ R^2 &= \frac{0.021103}{0.024133} = 0.874 \end{aligned}$$

帰無仮説 H_0、対立仮説 H_1 を以下のように設定する。

帰無仮説 H_0：乳量と増体量は有意な回帰関係にない。

対立仮説 H_1：乳量と増体量は有意な回帰関係にある。

上記で求めた偏差平方和から分散分析表を作成すると以下のようになる。

要因	自由度	偏差平方和	平均平方	分散比
全体	5	0.024133	0.0048266	
回帰	1	0.021103	0.021103	27.859
誤差	4	0.00303	0.0007575	

$F(1,4;0.05) = 7.709$ であるので、乳量と増体量は有意な回帰関係にあるといえる。
乳量が x_i のときの増体量の推定値 $\hat{y_i}$ とその 95%信頼区間は、以下の式で与えられる。

$$\begin{aligned} \hat{y_i} &= \hat{a} + \hat{b}x_i \\ v_x &= \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{6}}{n-1} \\ SE &= \sqrt{\{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{v_x(n-1)}\} \times MS_e} \end{aligned}$$

この式に数値を当てはめると

$$\begin{aligned} \hat{y_i} &= 0.131 + 0.102 \times 4.0 = 0.539 \\ v_x &= \frac{102.07 - \frac{24.5^2}{6}}{5} = 0.4057 \\ SE &= \sqrt{\{1 + \frac{1}{6} + \frac{(4.0 - 4.083)^2}{0.4057 \times 5}\} \times 0.0007575} = 0.0298 \end{aligned}$$

$t(4; 0.05/2) = 2.776$ から
$0.539 \pm 2.776 \times 0.0298 = 0.539 \pm 0.083$kg/日となる。

【R のコード・出力例】

```
> ss.y <- sum((gain-mean(gain))^2)
> ss.reg <- b^2*sum((yield-mean(yield))^2)
> ss.e <- ss.y-ss.reg
> r2 <- ss.reg/ss.y
> df.t <- length(yield)-1
> df.e <- df.t -1
> ms.y <- ss.y/df.t
> ms.reg <- ss.reg
> ms.e <- ss.e/df.e
> fvalue <- ms.reg/ms.e
> cat("分散分析表\n",
+ "要因 \t 自由度 \t 偏差平方和 \t 平均平方 \t 分散比\n",
+ "全体\t",df.t,"\t",ss.y,"\t",ms.y,"\n",
+ "回帰\t 1 \t",ss.reg,"\t",ms.reg,"\t",fvalue,"\n",
+ "誤差\t",df.e,"\t",ss.e,"\t",ms.e,"\n")
分散分析表
 要因     自由度          偏差平方和       平均平方          分散比
 全体     5       0.02413333      0.004826667
 回帰     1       0.02105724      0.02105724      27.38184
 誤差     4       0.003076089     0.0007690222
> cat("F 値：F(1,4;0.05)=",qf(0.95,1,4),"\n")
F 値：F(1,4;0.05)= 7.708647
> y.hat <- a+b*4.0
> v.x <- var(yield)
> se <- sqrt((1+1/length(yield)+(4.0-mean(yield))^2/(v.x*df.t))*ms.e)
> t0 <- qt(0.975,df.e)
> cat("95%信頼区=",y.hat,"±",t0*se,"\n")
95%信頼区= 0.5381758 ± 0.08328528
> cat("下限値=",y.hat-t0*se,"\n")
下限値= 0.4548906
> cat("上限値=",y.hat+t0*se,"\n")
上限値= 0.6214611
>
```

【補足事項】

回帰式の推定および分散分析は、R の lm( ) および aov( ) を用いて以下のように計算できる。

```
> reg <- lm(gain ~ yield)
> summary(reg)

Call:
lm(formula = gain ~ yield)
```

```
Residuals:
        1         2         3         4         5         6
 0.003336  0.031635  0.020879 -0.007420 -0.038365 -0.010066

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.13062    0.08031   1.626  0.17919
yield        0.10189    0.01947   5.233  0.00637 **
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 0.02773 on 4 degrees of freedom
Multiple R-squared: 0.8725,     Adjusted R-squared: 0.8407
F-statistic: 27.38 on 1 and 4 DF,  p-value: 0.006372

> summary(aov(gain ~ yield))
            Df    Sum Sq   Mean Sq F value   Pr(>F)
yield        1 0.0210572 0.0210572  27.382 0.006372 **
Residuals    4 0.0030761 0.0007690
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1
>
```

9 章練習問題解答例

【練習問題 1】

ミツバチの訪花性を決める物質を調べる目的で、45 頭のミツバチを供試して 4 方向から異なる花香成分 (A,B,C,D) が流れる装置を使い、選択性を調べた。各花香成分に誘引された個体数 (観測度数) は以下のとおりであった (データは省略)。四つの花香成分のあいだで選択性に差があるかどうかを有意水準 5%として検定しなさい。

【解答例】

帰無仮説 H_0：花香成分の誘引効果は同じである。理論度数：A = B = C = D

対立仮説 H_1：成分によって誘引効果が異なる。

有意水準：5%両側検定

各花香成分の観測度数、期待度数は以下のようになる。

花香成分	A	B	C	D	合計
観測度数	10	20	10	5	45
期待度数	11.25	11.25	11.25	11.25	45

$np_i = E_i$ として，観察数を (O_i) として

$$\chi_O^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

を求める。

$$\begin{aligned} \chi_O^2 &= \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(10-11.25)^2}{11.25} + \cdots + \frac{(5-11.25)^2}{11.25} \\ &= 10.56 \quad 、 \quad df = k - 1 = 4 - 1 = 3 \end{aligned}$$

$10.56 > \chi^2(3; 0.05) = 7.815$、従って帰無仮説のもとでで得られたような数値が生じる確率はきわめて小さい。帰無仮説は棄却され、ミツバチは特定の花香成分を好むと結論できる。

【R のコード・出力例】

(1) **上記の計算手順をそのまま実行する場合**

```
> Expect <- 45/4   #期待値の計算
> Expect
[1] 11.25
> 観測度数 <- c(10,20,10,5)   #観測度数を行列にする
> 観測度数
[1] 10 20 10  5
> χ二乗要素 <- (観測度数-11.25)^2/11.25   #観測度数と期待値との差を計算する。
> χ二乗要素
[1] 0.1388889 6.8055556 0.1388889 3.4722222
> χ二乗 <- sum(χ二乗要素)   #χ二乗値を計算する。
> χ二乗
```

```
[1] 10.55556
> qchisq(0.95,3)   #上側確率 95 %の χ 二乗値
[1] 7.814728
>
```

(2)χ^2 検定用の関数 chisq.test( ) を用いる場合

```
> data1 <- c(10,20,10,5)
> chisq.test(data1)

        Chi-squared test for given probabilities

data:  data1
X-squared = 10.5556, df = 3, p-value = 0.01439

>
```

【補足事項】

同じ期待度数が想定される場合 (A:B:C:D=1:1:1:1) は chisq.test の引数には、上記のように度数のみを与えればよい。

期待度数が異なる場合は、以下に示したように、それぞれの期待度数の確率を引数 p で指定する。

例）

```
> o <- c(151422, 149233, 152455, 147356) # 観測度数を o とする。
> prob <- c(1,1,1,1)/4 # 理論確率を prob とする。
> chisq.test(o,p=prob) # o と prob を用いたカイ二乗検定。

        Chi-squared test for given probabilities

data:  o
X-squared = 103.7451, df = 3, p-value < 2.2e-16

>
```

【練習問題 2】

表 9.4 で示した 2×2 分割表における χ_0^2 値が

$$\chi_0^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

として計算できることを示しなさい。

【解答例】

	B_1	B_2	計
A_1	a	b	$a+b=R_1$
A_2	c	d	$c+d=R_2$
計	$a+c=C_1$	$b+d=C_2$	$n=a+b+c+d$

a の期待度数 E_a は $E_a = \frac{C_1R_1}{n}$、b の期待度数 E_b は $E_b = \frac{C_2R_1}{n}$、c の期待度数 E_c は $E_c = \frac{C_1R_2}{n}$、d の期待度数 E_d は $E_d = \frac{C_2R_2}{n}$ となる。

期待度数と観察度数の偏りを

$$\chi_0^2 = \frac{(a-E_a)^2}{E_a} + \frac{(b-E_b)^2}{E_b} + \frac{(c-E_c)^2}{E_c} + \frac{(d-E_d)^2}{E_d}$$

として求める。

この式の各期待値 E を a, b, c, d, n で置き換えると

$$\begin{aligned}
\chi_0^2 &= \frac{(a-\frac{C_1R_1}{n})^2}{\frac{C_1R_1}{n}} + \frac{(b-\frac{C_2R_1}{n})^2}{\frac{C_2R_1}{n}} + \frac{(v-\frac{C_1R_2}{n})^2}{\frac{C_1R_2}{n}} + \frac{(d-\frac{C_2R_2}{n})^2}{\frac{C_2R_2}{n}} \\
&= \frac{(na-C_1R_1)^2}{nC_1R_1} + \frac{(nb-C_2R_1)^2}{nC_2R_1} + \frac{(nc-C_1R_2)^2}{nC_1R_2} + \frac{(nd-C_2R_2)^2}{nC_2R_2}
\end{aligned}$$

ここで

$$\begin{aligned}
na - C_1R_1 &= (a+b+c+d)a - (a+c)(a+b) \\
&= (a+c) + (b+d)a - (a+c)(a+b) \\
&= (a+c)(a-a-b) + (b+d)a \\
&= -ab - bc + ab + ad \\
&= ad - bc
\end{aligned}$$

以下、同様にして $nb - C_2R_1 = -(ad-bc)$、$nc - C_1R_2 = -(ad-bc)$、$nd - C_2R_2 = (ad-bc)$

よって

$$\begin{aligned}
\chi_0^2 &= (ad-bc)^2 \left\{ \frac{1}{nC_1R_1} + \frac{1}{nC_2R_1} + \frac{1}{nC_1R_2} + \frac{1}{nC_2R_2} \right\} \\
&= (ad-bc)^2 \left\{ \frac{C_2R_2 + C_1R_2 + C_2R_1 + C_1R_1}{nC_1C_2R_1R_2} \right\}
\end{aligned}$$

ここで

$$\begin{aligned}
C_2R_2 + C_1R_2 + C_2R_1 + C_1R_1 &= R_2(C_2+C_1) + R_1(C_2+C_1) \\
&= (C_1+C_2)(R_1+R_2) \\
&= n^2 \quad (\text{なぜなら、} C_1+C_2=n\text{、} R_1+R_2=n)
\end{aligned}$$

よって

$$\begin{aligned}\chi_0^2 &= \frac{n(ad-bc)^2}{C_1C_2R_1R_2} \\ &= \frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}\end{aligned}$$

【練習問題 3】

ソバの 2 品種に冠水処理を行ったところ、幼根が褐変したのち、幼根基部から新たに不定根が発生した種子と発生しない種子が以下の数で認められた (データは省略)。品種間で不定根発生程度に違いがあるといってよいかを 5%水準で検定しなさい。なお、不定根とは主根や側根以外の器官から形成される根のことである。

【解答例】

帰無仮説 H_0：品種間で不定根発生程度に差がない。

対立仮説 H_1：品種間で不定根発生程度に差がある。

χ_0^2 統計量を計算すると

$$\chi_0^2 = \frac{95 \times (22 \times 33 - 26 \times 14)^2}{48 \times 47 \times 36 \times 59} = 2.598$$

$2.598 < \chi^2(1; 0.05) = 3.841$ より、帰無仮説が採択できる。すなわち品種間には不定根発生程度に差がないといえる。

【R のコード・出力例】

(1) 計算手順をそのまま実行する場合

```
> (観測度数 <-c(22,26,14,33))
[1] 22 26 14 33
> (度数行列 <- matrix(観測度数,ncol=2,byrow=T))
     [,1] [,2]
[1,]   22   26
[2,]   14   33
> (行和  <- matrix(rowSums(度数行列),ncol=1))
     [,1]
[1,]   48
[2,]   47
> (列和 <- matrix(colSums(度数行列),ncol=2))
     [,1] [,2]
[1,]   36   59
> (総和 <- sum(観測度数))
[1] 95
> (期待度数 <- 行和 %*% 列和 / 総和)
         [,1]     [,2]
[1,] 18.18947 29.81053
[2,] 17.81053 29.18947
> カイ二乗 <- sum((度数行列 - 期待度数)^2/期待度数)
> カイ二乗
[1] 2.598048
> #自由度 1 のカイ二乗分布で下側確率 0.95 となるカイ二乗値を求める
> qchisq(0.95,1)
[1] 3.841459
```

(2)χ^2 検定用の関数 chisq.test( ) を用いる場合

```
> (観測度数 <- matrix(c(22,26,14,33),ncol=2,byrow=T))
     [,1] [,2]
[1,]   22   26
[2,]   14   33
> chisq.test(観測度数)

        Pearson's Chi-squared test with Yates' continuity correction

data:  観測度数
X-squared = 1.961, df = 1, p-value = 0.1614

>
```

【練習問題 4】

表 9.5 のデータを用いて χ^2 検定を実施しなさい。

【解答例】

帰無仮説 H_0：栽培体系と雑草種の出現数は独立である。

対立仮説 H_1：栽培体系と雑草種の出現数には関連がある。

i 行 j 列の期待値 E_{ij} は

$$E_{ij} = \frac{n_i \times T_j}{n}$$

である。$k \times m$ 個ある E_{ij} のうち $E_{ij} < 5$ のものが $0.2 \times k \times m$ 個以上あるとき (この例では $0.2 \times 4 \times 5 = 0.2 \times 20 = 4$ 個) には他の関係ある群と一緒にしてしまう。それ以下であればそのままとして χ^2 を計算する。ただし $E_{ij} < 1$ はひとつでもあってはいけない。χ^2 統計量は以下のように計算する。

$$\chi^2_O = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}、\text{自由度は } df = (k-1)(m-1)$$

表 5.9 の例では、$E_{ij} < 5$ は 1 区だけなのでこのまま計算する。

$$\chi^2_O = \frac{(6-7.4)^2}{7.4} + \frac{(15-22.8)^2}{22.8} + \cdots + \frac{(20-38)^2}{38} + \frac{(199-222.5)^2}{222.5} = 169.8$$

$\chi^2((4-1)(5-1); 0.05) = \chi^2(12; 0.05) = 21.026 < 169.8 = \chi^2_O$

従って帰無仮説は棄却され、栽培体系によって雑草種の出現分布は異なると結論できる。

【R のコード・出力例】

```
> data <- matrix(c(6,15,1,7,72,
+                  32,67,3,82,203,
+                  32,49,21,10,223,
+                  18,142,7,20,199), ncol=5, byrow=T)
> data
     [,1] [,2] [,3] [,4] [,5]
[1,]    6   15    1    7   72
[2,]   32   67    3   82  203
[3,]   32   49   21   10  223
[4,]   18  142    7   20  199
>
> chisq.test(data)

        Pearson's Chi-squared test

data:  data
X-squared = 170.0662, df = 12, p-value < 2.2e-16

Warning message:
In chisq.test(data) :  カイ自乗近似は不正確かもしれません
>
```

このような Warning message はセルの度数が 5 以下のものを含む場合に出力される。

【補足事項】

あらかじめ、以下のような csv ファイルを Excel などで作成しておくと、read.csv( ) と chisq.test( ) を用いて実行できる。

例：csv ファイルが c:/R.csv である場合

```
>data <- read.csv("c:/R.csv")
>chisq.test(data)
```

イネ科, シロザ, タデ類, ナギナタコウジュ, ハコベ
6,15,1,7,72
32,67,3,82,203
32,49,21,10,223
18,142,7,20,199

C:/R.csv の内容

10 章練習問題解答例

【練習問題 1】

表 10.1 に示した京都市内のナミテントウの色紋型別の個体数の記録 (データは省略) のうち、2008 年のデータについて、四つの色紋型それぞれの 95%信頼区間を求めなさい。なお、計算は 10.2.2 項で示した簡便な方法で行いなさい。

【解答例】

簡便法による $100(1-\alpha)\%$ 信頼区間推定は、当該色紋型の頻度を $\hat{p}$、総個体数を n とした時、標準正規分布の切断点 $z(\alpha/2)$ を用いて

$$[\hat{p} - \frac{z(\alpha/2)}{\sqrt{n}}\sqrt{\hat{p}(1-\hat{p})}, \hat{p} + \frac{z(\alpha/2)}{\sqrt{n}}\sqrt{\hat{p}(1-\hat{p})}]$$

で求まる。たとえば、二紋型についての 95%信頼区間は、総個体数が 2,006、二紋型の個体数が 1,409 なので、$\hat{p} = 1409/2006 = 0.702$、$z(\alpha/2) = z(0.975) = 1.960$、$n = 2006$ より

$$\begin{aligned} \frac{z(\alpha/2)}{\sqrt{n}}\sqrt{\hat{p}(1-\hat{p})} &= \frac{1.960}{\sqrt{2006}} \times \sqrt{0.702(1-0.702)} \\ &= 0.0438 \times \sqrt{0.702 \times 0.298} \\ &= 0.0438 \times 0.457 = 0.0200 \end{aligned}$$

よって、95%信頼区間は

$$[0.702 - 0.020, 0.702 + 0.020] = [0.682, 0.722]$$

となる。他の色紋型についても、同様の計算を行えばよい。
以下の R のコード・出力例では、4 種類の色紋型について、簡便法および厳密な方法による 95%信頼区間を計算した。

【R のコード・出力例】

```
> rawData <- c(1409,255,80,262)
> type.name <- c("二紋型","四紋型","斑 型","紅 型")
> total.n <- sum(rawData)
> Ratio <- rawData/total.n
> Ratio
[1] 0.70239282 0.12711864 0.03988036 0.13060818
> z95 <- qnorm(0.975)
> type.no <- length(rawData)
> z95.rootN <- z95/sqrt(total.n)
> p.limit <- numeric(type.no)
> c.limit <- numeric(type.no)
> for(i in 1:type.no){
+   p.limit[i] <- z95.rootN*sqrt(Ratio[i]*(1-Ratio[i]))
+   c.limit[i] <- z95.rootN*sqrt(Ratio[i]*(1-Ratio[i])+z95.rootN^2/4)
+ }
> #簡便法
> for(i in 1:type.no){
+   cat(type.name[i],":[",Ratio[i]-p.limit[i],"\t",Ratio[i]+p.limit[i],"]\n")
+ }
```

```
二紋型 :[ 0.6823853      0.7224004 ]
四紋型 :[ 0.1125418      0.1416955 ]
斑 型 :[ 0.03131738      0.04844334 ]
紅 型 :[ 0.1158621       0.1453542 ]
>
> #厳密な推定法
> p.div <- 1+z95.rootN^2
> for(i in 1:type.no){
+   cat(type.name[i],":[",((Ratio[i]-z95.rootN^2/2)-c.limit[i])/p.div,
+       "\t",((Ratio[i]-z95.rootN^2/2)+c.limit[i])/p.div,"]\n")
+ }
二紋型 :[ 0.6801025      0.7200868 ]
四紋型 :[ 0.1113396      0.1405004 ]
斑 型 :[ 0.0302486       0.04744835 ]
紅 型 :[ 0.1146540       0.1441517 ]
>
```

【補足事項】

R には母比率の検定を行うための関数 binom.test( ) が用意されている。この関数を用いると検定の他、95%信頼区間の推定値も得られる。一般的な書式は binom.test(当該項目の度数, 総数) である。たとえば、上記の二紋型の 95%信頼区間を求めるなら

$$binom.test(1409, 2006)$$

とする。95%以外の信頼限界を設定する場合は引数 conf.level にその値を指定する。たとえば、90%信頼区間を推定するのであれば

$$binom.test(1409, 2006, conf.level = 0.90)$$

とする。

bino.test を用いた色紋型ごとの 95%信頼区間推定の例を以下に示す。なお、type.no、、type.name、rawData、total.n は「R のコード・出力例」で示したものを用いている。

```
> for(i in 1:type.no){
+   cat("\n 色紋型：",type.name[i],"\n")
+   print(binom.test(rawData[i],total.n))
+
+ }

色紋型： 二紋型

        Exact binomial test

data:  rawData[i] and total.n
number of successes = 1409, number of trials = 2006, p-value <
2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.6818506 0.7223419
```

```
sample estimates:
probability of success
             0.7023928
```

色紋型： 四紋型

```
        Exact binomial test

data:  rawData[i] and total.n
number of successes = 255, number of trials = 2006, p-value <
2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1128454 0.1424912
sample estimates:
probability of success
             0.1271186
```

色紋型： 斑 型

```
        Exact binomial test

data:  rawData[i] and total.n
number of successes = 80, number of trials = 2006, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.03174694 0.04939091
sample estimates:
probability of success
            0.03988036
```

色紋型： 紅 型

```
        Exact binomial test

data:  rawData[i] and total.n
number of successes = 262, number of trials = 2006, p-value <
2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1161608 0.1461444
sample estimates:
probability of success
             0.1306082
```

【練習問題 2】

表 10.1 に示した京都市内のナミテントウの色紋型別の個体数の記録 (データは省略) について、紅型以外の色紋型に年代変化があったかどうかを有意水準 5%として検定しなさい。

【解答例】

1953 年の標本集団の当該色紋型の個体数とその比率を n_1, p_1、2008 年の標本集団の当該色紋型の個体数とその比率を n_2, p_2 とする。帰無仮説 H_0 および対立仮説 H_1 は以下のようになる。

帰無仮説 H_0: $p_1 = p_2$

対立仮説 H_1: $p_1 \neq p_2$

対立仮説が $H_1 : p_1 \neq p_2$ なので、両側検定を行う。有意水準 5%の棄却限界値は $z(0.05/2) = 1.960$ である。

検定統計量 z_0 は以下の式より計算する。

$$z_0 = \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

二紋型について検定統計量 z_0 を計算すると、$n_1 = 2494, p_1 = 0.637$、$n_2 = 2006, p_2 = 0.702$ なので

$$z_0 = \frac{|0.637 - 0.702|}{\sqrt{\frac{0.637(1-0.637)}{1590} + \frac{0.702(1-0.702)}{2006}}} = 4.113$$

$z_0 = 4.113 > z(0.025) = 1.960$ なので、帰無仮説は棄却され、二紋型の比率には年代変化が認められる。

【R のコード・出力例】

```
> data1950 <- c(1590,394,128,382)
> data2008 <- c(1409,255,80,262)
> type.name <- c("二紋型","四紋型","斑 型","紅 型")
> total.1950n <- sum(data1950)
> total.2008n <- sum(data2008)
> y1950.p <- data1950/total.1950n
> y2008.p <- data2008/total.2008n
> z95 <- qnorm(0.975)
> z0 <- numeric(length(type.name))
> test.flag <- rep(T,4)
> for(i in 1:length(type.name)){
+  z0[i] <- abs(y1950.p[i]-y2008.p[i])/
+               sqrt(y1950.p[i]*(1-y1950.p[i])/total.1950n+
+                    y2008.p[i]*(1-y2008.p[i])/total.2008n)
+  if(z0[i]>z95) test.flag[i]=F
+ }
> for(i in 1:length(type.name)){
+  if(test.flag[i]){
+    cat(type.name[i],":帰無仮説を採択  z0=",z0[i],"<=z(0.05/2)=",z95,"\n")
+  }
+  if(!test.flag[i]){
+    cat(type.name[i],":帰無仮説を棄却  z0=",z0[i],">z(0.05/2)=",z95,"\n")
+  }
```

```
+ }
二紋型 :帰無仮説を棄却   z0= 4.622893 >z(0.05/2)= 1.959964
四紋型 :帰無仮説を棄却   z0= 2.960651 >z(0.05/2)= 1.959964
斑 型 :帰無仮説を採択   z0= 1.841543 <=z(0.05/2)= 1.959964
紅 型 :帰無仮説を棄却   z0= 2.164650 >z(0.05/2)= 1.959964
>

>
```

この結果から、斑型を除く 3 つの色紋型で比率には年代変化が認められる。

【補足事項】

R には母比率間の差の検定として、prob.test( ) が用意されている。この関数を用いた検定結果を以下に示す。

```
> data2008 <- c(1409,255,80,262)
> data.mat <- cbind(data1950,data2008)
> data.n <-c(2494,2006)
> type.name <- c("二紋型","四紋型","斑 型","紅 型")
> for(i in 1:4){
+   cat("\n",type.name[i],"\n")
+   print(prop.test(data.mat[i,],data.n,correct=F))
+ }
```

二紋型

```
        2-sample test for equality of proportions without continuity
        correction

data:  data.mat[i, ] out of data.n
X-squared = 21.0413, df = 1, p-value = 4.495e-06
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.09236255 -0.03736294
sample estimates:
   prop 1    prop 2
0.6375301 0.7023928
```

四紋型

```
        2-sample test for equality of proportions without continuity
        correction

data:  data.mat[i, ] out of data.n
X-squared = 8.5788, df = 1, p-value = 0.003401
alternative hypothesis: two.sided
95 percent confidence interval:
```

```
 0.01043071 0.05129030
sample estimates:
   prop 1    prop 2
0.1579791 0.1271186
```

斑 型

```
        2-sample test for equality of proportions without continuity
        correction

data:  data.mat[i, ] out of data.n
X-squared = 3.302, df = 1, p-value = 0.0692
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.0007358311  0.0236214645
sample estimates:
    prop 1     prop 2
0.05132318 0.03988036
```

紅 型

```
        2-sample test for equality of proportions without continuity
        correction

data:  data.mat[i, ] out of data.n
X-squared = 4.614, df = 1, p-value = 0.03171
alternative hypothesis: two.sided
95 percent confidence interval:
 0.002133183 0.042985671
sample estimates:
   prop 1    prop 2
0.1531676 0.1306082

>
```

【練習問題 3】

ある種苗会社が出荷するダイコンの種子には、従来 5%の割合で発芽しない不良な種子が含まれていた。ある年に、この種苗会社からダイコンの種子を購入して 400 粒の種子を播いたところ、30 粒が発芽しなかった。種子の品質が変化したと判断してよいだろうか。有意水準を 5%として検定しなさい。

【解答例】

母比率の検定を行う。不良な種子が含まれている比率を p とすると、検定の帰無仮説 H_0 と対立仮説 H_1 は

帰無仮説 H_0:$p = 0.05$

対立仮説 H_1:$p \neq 0.05$

である。$n = 400$ 粒の種子のうち、30 粒が発芽しなかったので、不良な種子の比率 $\hat{p} = 30/400 = 0.075$ である。

帰無仮説のもとで、$\hat{p}$ が近似的に平均 $\mu = p$、分散 $\sigma^2 = p(1-p)/n$ の正規分布 $N(p, p(1-p)/n)$ に従うとすると、$\hat{p}$ の Z 変換は

$$z_0 = \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{p(1-p)}} = \frac{\sqrt{400}(0.075 - 0.5)}{\sqrt{0.05(1-0.05)}} = 2.294$$

有意水準を 5%とした、両側検定を行う場合、$z(0.05/2) = 1.960$ が棄却限界値となる。従って、$z(0.05/2) = 1.960 < z_0 = 2.294$ なので、帰無仮説は棄却される。すなわち、種子の品質が変化したと判断できる。

【R のコード・出力例】

```
> num <- 400
> p.ratio <- 0.075
> p.value <-0.05
> z0 <- sqrt(num)*(p.ratio-p.value)/sqrt(p.value*(1-p.value))
> z5 <- qnorm(1-p.value/2)
> if(z0<z5){
+   cat("帰無仮説を採択：z=",z0,"<=z(0.05/2)=",z5,"\n")
+ }else{
+   cat("帰無仮説を棄却：z=",z0,">z(0.05/2)=",z5,"\n")
+ }
帰無仮説を棄却：z= 2.294157 >z(0.05/2)= 1.959964
>
>
```

【補足事項】

binom.test を用いて、直接検定することができる。この場合の書式は binom.test(当該項目の度数, 総数,p=帰無仮説に用いる比率) とする。

binom.test( ) を用いた実行結果を以下に示す。p-value = 0.02843 なので、5%水準で帰無仮説が棄却される。

```
> num <- 400
> obs.num <- 30
> p.value <- 0.05
```

```
> binom.test(obs.num,num,p=p.value)

        Exact binomial test

data:  obs.num and num
number of successes = 30, number of trials = 400, p-value = 0.02843
alternative hypothesis: true probability of success is not equal to 0.05
95 percent confidence interval:
 0.05117168 0.10533764
sample estimates:
probability of success
                 0.075

>
```

【練習問題 4】

上の問題の種苗会社に、いくつかの農家から「今年は、ダイコンの発芽率が例年よりも悪い」と苦情がでている。このことを、上の結果から調べるにはどのような検定を行えばよいか。実際に有意水準5%で検定を行いなさい。

【解答例】

練習問題 3 では、対立仮説として $H_1 : p \neq 0.05$ を用意したが、この問題では「例年より発芽率が悪い」ということを検定するため、対立仮説として $H_1 : p > 0.05$ を採用する。すなわち、帰無仮説、対立仮説を示せば以下のようになる。

帰無仮説 H_0: $p = 0.05$

対立仮説 H_1: $p > 0.05$

対立仮説をこのように設定した場合は、検定は片側検定となる。そのため、棄却限界値として $z(0.05) = 1.645$ を用いる。

問題 2 で求めた検定統計量 z_0 は 2.294 なので、$z(0.05) = 1.645 < z_0 = 2.294$ であり、帰無仮説は棄却される。すなわち、有意確率 5%で「今年は、ダイコンの発芽率が例年よりも悪い」といえる。

【R のコード・出力例】

binom.test を用いた結果を示す。片側検定を行うので、引数として alternative = "greater"を与える。

```
> num <- 400
> obs.num <- 30
> p.value <- 0.05
> binom.test(obs.num,num,p=p.value,alternative = "greater")

        Exact binomial test

data:  obs.num and num
number of successes = 30, number of trials = 400, p-value = 0.01903
alternative hypothesis: true probability of success is greater than 0.05
95 percent confidence interval:
 0.0544982 1.0000000
sample estimates:
probability of success
                 0.075

>
```

11 章練習問題解答例

【練習問題 1】

ブタの 1 産あたり分娩頭数 (一腹産子数) に関する下記のデータ (データは省略) の中央値、第 1 四分位数および第 3 四分位数を求め、正規分布と見なしてよいか検定しなさい。

【解答例】

データを小さい順に並びかえると以下のようになる。

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
分娩頭数 (頭)	8	8	9	9	9	10	10	10	15

中央値：$i = (1 - 0.5 + 0.5 \times 9) = 5$ から $x_5 = 9$

第 1 四分位数：$i = (1 - 0.25 + 0.25 \times 9) = 3$ から $x_3 = 9$

第 3 四分位数：$i = (1 - 0.75 + 0.75 \times 9) = 7$ から $x_7 = 10$

帰無仮説 H_0、対立仮説 H_1 を以下のように設定する。

帰無仮説 H_0：データは正規分布する

対立仮説 H_1：データは正規分布しない

平均からの偏差の 2 乗和：$d^2 = 35.556$
平均からの偏差の 3 乗和：$d^3 = 129.8025$
平均からの偏差の 4 乗和：$d^4 = 764.823$

歪度 $a_3 = \frac{\frac{129.8025}{9}}{(\frac{35.556}{9})^{\frac{3}{2}}} = 1.837$、尖度 $a_4 = \frac{\frac{764.823}{9}}{(\frac{35.556}{9})^2} - 3 = 2.445$

$$JB = 9(\frac{1.837^2}{6} + \frac{2.445^2}{24}) = 7.304 > \chi^2(2; 0.05) = 5.9915$$

※ データは正規分布ではない

【R のコード・出力例】

```
> pig <- c(8,9,10,8,10,15,9,10,9)
> pig.sort <- sort(pig)
> pig.n <- length(pig.sort)
> pig.med <- (1-0.5+0.5*pig.n)
> pig.1qt <- (1-0.25+0.25*pig.n)
> pig.3qt <- (1-0.75+0.75*pig.n)
> cat("中央値\t 第 1 四分位数\t 第 3 四分位数\n",
+     pig.sort[pig.med],"\t",pig.sort[pig.1qt],"\t",pig.sort[pig.3qt],"\n")
中央値　第 1 四分位数　　第 3 四分位数
 9       9       10
>
> d2 <- sum((pig-mean(pig))^2)
> d3 <- sum((pig-mean(pig))^3)
> d4 <- sum((pig-mean(pig))^4)
> a3 <- (d3/pig.n)/(d2/pig.n)^(3/2)
```

```
> a4 <- (d4/pig.n)/(d2/pig.n)^2
> cat("歪度=",a3,"\t 尖度=",a4,"\n")
歪度= 1.836720  尖度= 5.444883
> jb <- pig.n*(a3^2/6+(a4-3)^2/24)
> cat("JB=",jb,"\tchi-square(2;0.05)=",qchisq(0.950,2),"\n")
JB= 7.301856    chi-square(2;0.05)= 5.991465
```

【補足事項】

中央値、第 1 四分位数、第 3 四分位数は R の summary( ) で求まる。

```
> pig <- c(8,9,10,8,10,15,9,10,9)
> summary(pig)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.000   9.000   9.000   9.778  10.000  15.000
>
```

また、歪度、尖度は R の package e1071(これは、http://cran.r-project.org/から入手しておく必要がある) の skewness( )：歪度、kurtosis( )：尖度で求めることができる。

```
> library(e1071)
> skewness(pig,type=1)
[1] 1.836720
>
```

ジャック・ベラ検定 (Jarque-Bera test) は package tseries の中で、jarque.bera.test( ) として用意されている。package tseries は package quadprog、package zoo と依存関係にあるので、jarque.bera.test( ) を用いる場合はあらかじめこれら三つの package をインストールしておかなければならない。

```
> library(tseries)
> jarque.bera.test(pig)

        Jarque Bera Test

data:  pig
X-squared = 7.3019, df = 2, p-value = 0.02597
```

【練習問題 2】

モルモットに 2 種類の飼料 (飼料 A と B) を与え、嗜好性に対する試験を行ったところ、15 頭のうち 11 頭が飼料 A を好み、残りは飼料 B を好んだ。モルモットの嗜好性に差があるといえるか。

【解答例】

帰無仮説 H_0、対立仮説 H_1 を以下のように設定する。

帰無仮説 H_0：飼料の嗜好性に差はない

対立仮説 H_1：飼料の嗜好性に差はある

飼料 B を好んだモルモットの数を $n_B = 4$ とすると

$$P_0 = 2 \times \frac{\sum_{i=0}^{4} {}_{15}C_i}{2^{15}} = 2 \times \frac{1 + 15 + 105 + 455 + 1365}{32768} = 0.118$$

※ 飼料の嗜好性に差があるとはいえない

【R のコード・出力例】

$$\sum_{i=0}^{r} {}_{n}C_i$$

を計算すための自作関数 cumsum(n,r) をまず定義する。

```
cumsum <- function(n,r){
 sum <- 0
 for(i in 0:r){
  sum <- sum+choose(n,i)
 }
 return(sum)
}
```

この関数 cumsum(15,4) で $\sum_{i=0}^{4} {}_{15}C_i$ が求まるので、P_0 は以下のように計算できる。

```
> cumsum <- function(n,r){
+  sum <- 0
+  for(i in 0:r){
+   sum <- sum+choose(n,i)
+  }
+  return(sum)
+ }
> p0 <- 2*cumsum(15,4)/2^(15)
> p0
[1] 0.1184692
>
```

【補足事項】

R の binom.test( ) を用いて検定することができる。この場合、引数は成功した試行数と試行総数である。ここでは、使用 B を好んだものを成功した試行数とみなすと以下のようになる。

```
> binom.test(4,15)

        Exact binomial test

data:  4 and 15
number of successes = 4, number of trials = 15, p-value = 0.1185
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.07787155 0.55100324
sample estimates:
probability of success
             0.2666667
```

【練習問題 3】

練習問題 2 と同じ試験を 50 頭のモルモットに対して行った場合、5%水準で飼料 A の嗜好性が高いというためには、少なくとも何頭のモルモットが飼料 A に対して嗜好性を示す必要があるか求めなさい。

【解答例】

片側検定を考えていることから、正規分布における棄却限界値は $z = 1.645$ であり

$$z_0 = \frac{|n_b - \frac{50}{2}| - 0.5}{\sqrt{\frac{50}{4}}}$$

より、$n_B = 19$ のとき、$z_0 = 1.556$、$n_B = 18$ のとき $z_0 = 1.838$ が得られる。したがって、$n_b \leq 18$ のとき、有意な差があると結論できるため、少なくとも 32 頭が A に対して嗜好性を示す必要がある。

【R のコード・出力例】

$$\begin{aligned}
\frac{|n_b - \frac{n}{2}| - |\frac{1}{2}|}{\sqrt{\frac{n}{4}}} &> z \\
|n_b - \frac{n}{2}| &> z\sqrt{\frac{n}{4}} + \frac{1}{2} \\
n_b - \frac{n}{2} &< -z\sqrt{\frac{n}{4}} - \frac{1}{2}
\end{aligned}$$

または

$$\begin{aligned}
n_b - \frac{n}{2} &> z\sqrt{\frac{n}{4}} + \frac{1}{2} \\
\text{よって } n_b &< -z\sqrt{\frac{n}{4}} + \frac{n-1}{2}
\end{aligned}$$

または

$$n_b > z\sqrt{\frac{n}{4}} + \frac{n+1}{2}$$

```
> n <- 50
> z <- qnorm(0.95)
> nb <- -z*sqrt(n/4)+(n-1)/2
> nb2 <- z*sqrt(n/4)+(n+1)/2
> cat(nb,"\t",nb2,"\n")
18.68456        31.31544
>
```

【練習問題 4】

脳の機能を活性化させると宣伝されている物質がある。近交系マウス 20 頭を無作為に 2 群に分け，その物質を含んだ飼料と含まない飼料を一定期間摂取させた。その後、すべてのマウスを同時に迷路に入れ、活性化物質を摂取したマウスの脱出順位を以下のように得た。この物質に効果があると判定できるか検定しなさい。
脱出順位：1 位，2 位，4 位，6 位，8 位，10 位，12 位，13 位，14 位，17 位

【解答例】

帰無仮説 H_0、対立仮説 H_1 を以下のように設定する。

帰無仮説 H_0：この活性化物質に効果はない

対立仮説 H_1：この活性化物質に効果はある

活性化物質を給与されたグループの順位和は $RS_1 = 1+2+4+6+8+10+12+13+14+17 = 87$ である。さらに

$$
\begin{aligned}
E(RS_1) &= \frac{1}{2} \times 10 \times (10+10+1) = 105 \\
V(RS_1) &= \frac{1}{12} \times 10 \times 10 \times (10+10+1) = 175 \\
z_0 &= \frac{|87-105|-0.5}{\sqrt{175}} = 1.323
\end{aligned}
$$

となるが、正規分布における片側 5%の棄却限界値は 1.645 であるので、この活性化物質に効果があるとはいえない。

【R のコード・出力例】

```
> escRank <-c(1,2,4,6,8,10,12,13,14,17)
> nt <-20
> na <- length(escRank)
> rs1 <- sum(escRank)
> rs1.mean <- 1/2*na*(nt+1)
> rs1.var <- 1/12*na*(nt-na)*(nt+1)
> cat("E(RS1)=",rs1.mean,"\tV(RS1)=",rs1.var,"\n")
E(RS1)= 105     V(RS1)= 175
> z0 <- (abs(rs1-rs1.mean)-0.5)/sqrt(rs1.var)
> cat("z0=",z0,"\tz(0.95)=",qnorm(0.95),"\n")
z0= 1.322876    z(0.95)= 1.644854
>
```

【練習問題 5】

7 組のウシの親子について、子ウシの生時の体重が小さい順に番号を振り、さらにその母ウシの出産時日齢の若いものから順に番号を振った以下のデータを得た (データ省略)。これら 2 変数の相関を求め、出産時日齢と生時体重が有意な関係にあるか検定しなさい。

【解答例】

順位相関 r_s は以下の式で与えられる。

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum (R_{1i} - R_{2i})^2$$

よって、このデータの順位相関は

$$r_s = 1 - \frac{6}{7 \times (7^2-1)} \times (1+1+0+1+4+4+1) = 0.786$$

となる。帰無仮説 H_0、対立仮説 H_1 を以下のように設定する。

帰無仮説 H_0：出産時日齢と生時体重は有意な相関関係にない

対立仮説 H_1：出産時日齢と生時体重は有意な相関関係にある

$$t_0 = 0.786 \times \sqrt{\frac{7-2}{1-0.786^2}} = 2.843 > t(5; 0.05/2) = 2.571$$

より、出産時日齢と生時体重は有意な相関関係にあるといえる。

【R のコード・出力例】

```
> bw <- c(1,2,3,4,5,6,7)
> bd <- c(2,1,3,5,7,4,6)
> n <- length(bw)
> rs <- 1-6/(n*(n^2-1))*sum((bw-bd)^2)
> cat("順位相関係数 rs=",rs,"\n")
順位相関係数 rs= 0.7857143
> t0 <- rs*sqrt((n-2)/(1-rs^2))
> cat("t0=",t0,"\t t(5;0.05/2)=",qt(0.975,n-2),"\n")
t0= 2.840188     t(5;0.05/2)= 2.570582
>
```

【補足事項】

R でスピアマンの順位相関係数 (Spearman's rank correlation coefficient) を求める場合は相関係数を求める関数 cor( ) の引数に method="spearman"を指定する。このとき、bw、bd は順位データではなく測定された値そのものでもよい。

相関係数の検定 cor.test( ) についても同様に、引数に method="spearman"を指定する。実行結果を以下に示す。

```
> cor(bw,bd,method="spearman")
[1] 0.7857143
> cor.test(bw,bd,method="spearman")

        Spearman's rank correlation rho
```

```
data:  bw and bd
S = 12, p-value = 0.04802
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7857143
```

12 章練習問題解答例

【練習問題 1】

ベクトルの要素の和を求める関数として sum( ) がある。また、ベクトルの要素数を求める関数として length( ) がある。x <- c(1,2,3,4,5,6,7,8,9,10) であるとき、sum( ) と length( ) を用いて x の平均を求めなさい。また、mean(x) を用いて、求めた平均が正しいことを確認しなさい。

【解答例】

平均値は

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

で求まるので、R では sum(x)/length(x) とする。

【R のコード・出力例】

```
> x <- c(1,2,3,4,5,6,7,8,9,10)
> x.ave <- sum(x)/length(x)
> x.ave
[1] 5.5
> mean(x)
[1] 5.5
>
```

【練習問題 2】

不偏分散の定義式に従って、上記のベクトル x の分散を求めなさい。また、var(x) を用いて、求めた分散が正しいことを確認しなさい。sum(x*x) とすると、x の要素の平方和が求まる。

【解答例】

不偏分散の定義式は

$$\hat{\sigma}^2 = \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n})$$

で求まるので、R では (sum(x*x)-(sum(x)^2/length(x)))/(length(x)-1) とする。

【R のコード・出力例】

```
> x.var <-(sum(x*x)-(sum(x)^2/length(x)))/(length(x)-1)
> x.var
[1] 9.166667
> var(x)
[1] 9.166667
>
```

【練習問題 3】

対応のない標本の平均値の比較 (4.4 節) のサラブレット種とアラブ種の例題データ (データは省略) について差の検定を行いなさい。

【解答例】

等分散を仮定した場合の検定は R では t.test(ベクトル 1, ベクトル 2,var.equal=T) とする。var.equal=T を省略したり、t.test(ベクトル 1, ベクトル 2,var.equal=F) とすると、等分散を仮定しないウェルチの t 検定が実行される。

このデータについて、等分散を仮定した場合の検定統計量 t は 4.3692、有意確率は 0.003277 であり、両集団間の平均値間の差は 1%水準で有意である。

【R のコード・出力例】

```
> x <- c(160,168,158,165,161)
> y <- c(153,155,149,152)
> t.test(x,y,var.equal=T)

        Two Sample t-test

data:  x and y
t = 4.3692, df = 7, p-value = 0.003277
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.656745 15.643255
sample estimates:
mean of x mean of y
   162.40    152.25
```

【補足事項】

等分散を仮定しない場合 (ウェルチの t 検定) の R の実行結果は以下のようになる。

```
> x <- c(160,168,158,165,161)
> y <- c(153,155,149,152)
> t.test(x,y,var.equal=F)

        Welch Two Sample t-test

data:  x and y
t = 4.622, df = 6.701, p-value = 0.002712
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.909866 15.390134
sample estimates:
mean of x mean of y
   162.40    152.25

>
```

【練習問題 4】

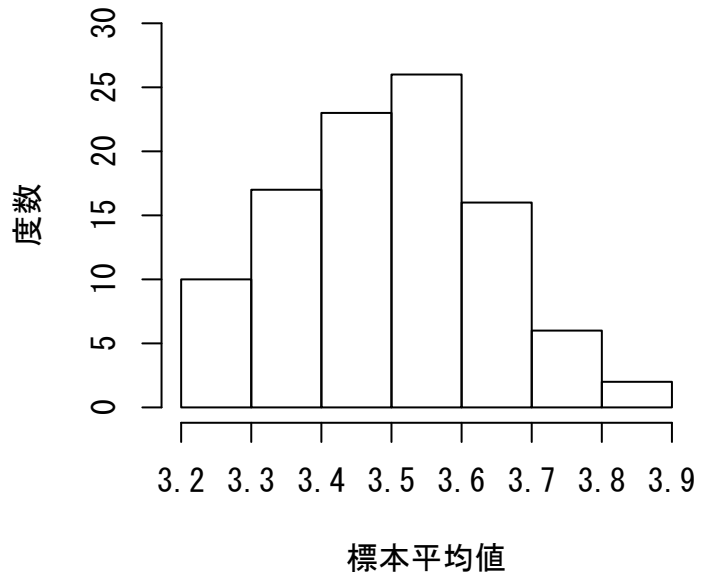
標本の大きさ 100 の一様乱数 (範囲 [1,6]) に従う標本を 100 個作成し、それぞれの標本平均を計算しそのヒストグラムを作成しなさい。なお、R での繰返し処理は以下のように for 文を用いて表すことができる。
for (変数 in 1:終了値){
処理
}
たとえば、処理を 100 回繰り返すのであれば
for (k in 1:100){
処理
}
とする。

【解答例】

標本の大きさ 100 の一様乱数 (範囲 [1,6]) に従う標本は runif(100,1,6) で生成できる。
100 個の標本平均値を格納するベクトルは x.ave<-numeric(100) と宣言してあらかじめ用意しておく。
ヒストグラムは hist(x.ave) で描ける。

【R のコード・出力例】

```
> x.ave <- numeric(100)
> for(i in 1:100){
+   x.ave[i] <- mean(runif(100,1,6))
+ }
> hist(x.ave,ylim=c(0,30),main=NULL,
+   xlab="標本平均値",ylab="度数")
>
```



【練習問題 5】

大きさ 100 の一様乱数 (範囲 [0,10]) のベクトル x を作成し、次に x の個々の要素から $y = x + e$ となるベクトル y を作成しなさい。このとき e は平均=0、標準偏差=1 に従う正規乱数とすること。得られた (x, y) の散布図を描き、x と y の相関係数を求めなさい。

【解答例】

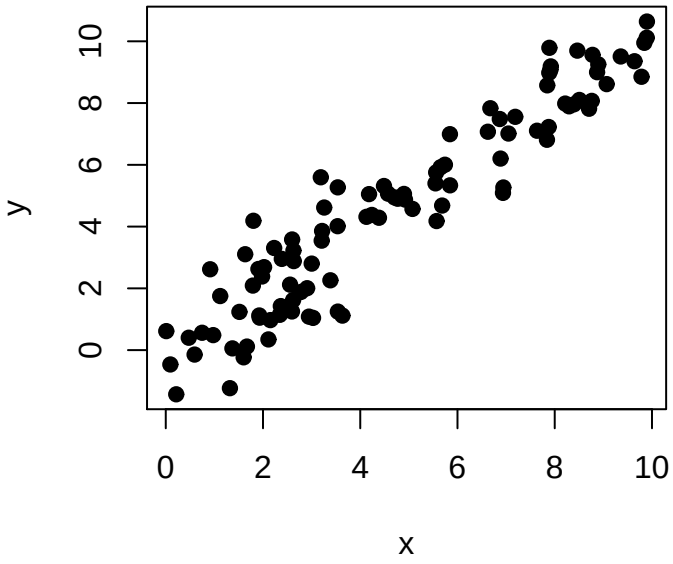
標本の大きさ 100 の正規乱数 (平均=0, 標準偏差=1) に従う標本は rnorm(100) で生成できる。

rnorm( ) の一般的な形式は rnorm(個数,mean=平均,sd=標準偏差) であるが、mean、sd を省略すると、それぞれ mean=0、sd=1 となる。

散布図は plot(ベクトル 1, ベクトル 2) で描ける。相関係数は cor(ベクトル 1, ベクトル 2) で求まる。cor(行列) とすると、総当たりの相関行列が得られる。

【R のコード・出力例】

```
> x <- runif(100,0,10)
> e <- rnorm(100)
> y <- x+e
> plot(x,y,pch=19)
> cor(x,y)
[1] 0.94489
>
```



【練習問題 6】

練習問題 5 で作成した x,y を用いて、定義式に従って y を従属変数、x を独立変数とした直線回帰式を求めなさい。

【解答例】

回帰係数 b はσ_{xy}/σ_x^2 で得られる。また、切片 a は$\overline{y} - \hat{b}\overline{x}$ で得られる。

これらの式を R で表すと、b < -cov(x,y)/var(x)、a <- mean(y)-b*mean(x) である。なお、lm( ) を用いて、回帰式を求めることもできる。

得られた回帰式は $y = -0.3212 + 1.0457x$ である。

【R のコード・出力例】

```
> b <- cov(x,y)/var(x)
> a <- mean(y)-b*mean(x)
> a
[1] -0.3212298
> b
[1] 1.045657
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    -0.3212       1.0457

>
```

13 章練習問題解答例

【練習問題 1】

8 章の例題データ [ブタの 30kg 到達日齢 (x) と 90kg 到達日齢 (y) のデータ] について、正規方程式 $\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{y}$ を作成し、この正規方程式の解 $\hat{\mathbf{b}}$ を求めなさい。

【解答例】

行列表記で $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ を表すと

$$\begin{bmatrix} 158 \\ 149 \\ 143 \\ 153 \\ 147 \end{bmatrix} = \begin{bmatrix} 1 & 75 \\ 1 & 67 \\ 1 & 65 \\ 1 & 71 \\ 1 & 72 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 5 & 350 \\ 350 & 24564 \end{bmatrix}、\mathbf{X}'\mathbf{y} = \begin{bmatrix} 750 \\ 52575 \end{bmatrix}$$

よって、正規方程式 $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$ は

$$\begin{bmatrix} 5 & 350 \\ 350 & 24564 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 750 \\ 52575 \end{bmatrix} \text{となる。}$$

$$(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} = \begin{bmatrix} 76.763 & -1.094 \\ -1.094 & 0.016 \end{bmatrix} \text{より}$$

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 76.763 & -1.094 \\ -1.094 & 0.016 \end{bmatrix} \begin{bmatrix} 750 \\ 52575 \end{bmatrix} = \begin{bmatrix} 67.969 \\ 1.172 \end{bmatrix}$$

【R のコード・出力例】

```
> X<- matrix(c(1,1,1,1,1,75,67,65,71,72),nrow=5,ncol=2)
> y <- c(158,149,143,153,147)
> XTX <- t(X) %*% X
> XTy <- t(X) %*% y
> b <- solve(XTX) %*% XTy
> b
          [,1]
[1,] 67.968750
[2,]  1.171875
```

【練習問題 2】

重回帰分析の例題 (表 13.1 のデータ) について、回帰式に独立変数を二つに制限した場合、体高、体長、胸囲のうちどの二つを取り上げるのがよいか検討しなさい。

【解答例】

ここでは、R を用いた分析方法を示す。データは以下に示すカンマ区切りのテキストファイル (mulreg.csv) から読み込むことにする。

```
Height,Length,Chest,Weight
1.27,1.45,1.80,475
1.28,1.50,1.86,490
1.31,1.50,1.98,580
1.28,1.52,1.96,557
1.32,1.50,1.87,540
1.27,1.46,1.86,440
1.28,1.53,1.84,459
1.27,1.50,1.82,470
```

mulreg.csv の内容

三つの独立変数から二つの独立変数を選ぶ選び方は全部で ${}_3C_2 = 3$ 通り。すなわち (Height,Length)、(Height,Chest)、(Length,Chest) のそれぞれの組み合わせについて、重回帰分析を行い AIC を比較する。AIC のもっとも小さな組み合わせを持つ回帰式を最良の回帰式とする。計算には R の lm( ) と AIC( ) を用いる (【R のコード・出力例】参照)。

三つの独立変数を全て考慮したモデルと、上記の三つのモデルの寄与率、AIC をまとめると以下のようになる。

モデル	寄与率	AIC
体高＋体長＋胸囲	0.817	88.652
体高＋体長	0.564	94.535
体高＋胸囲	0.817	86.737
体長＋胸囲	0.669	92.079

この結果から、独立変数として体高と胸囲を考慮したモデルが最もあてはまりが良いことがわかる。

【R のコード・出力例】

```
> mulreg.data <- read.csv("mulreg.csv")
> lm.full <- lm(Weight~.,data=mulreg.data)
> AIC(lm.full)
[1] 88.65167
> lm.HL <- lm(Weight~Height+Length,data=mulreg.data)
> AIC(lm.HL)
[1] 94.53532
> lm.HC <- lm(Weight~Height+Chest,data=mulreg.data)
> AIC(lm.HC)
[1] 86.73658
> lm.LC <- lm(Weight~Length+Chest,data=mulreg.data)
> AIC(lm.LC)
```

```
[1] 92.07922
> summary(lm.HL)

Call:
lm(formula = Weight ~ Height + Length, data = mulreg.data)

Residuals:
    Min      1Q  Median      3Q     Max
-43.461 -26.782  -5.016  12.429  57.021

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2220.6     1016.4  -2.185   0.0716 .
Height        1830.7      745.7   2.455   0.0494 *
Length         248.2      503.3   0.493   0.6395
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 36.28 on 6 degrees of freedom
Multiple R-squared: 0.5646,     Adjusted R-squared: 0.4195
F-statistic: 3.891 on 2 and 6 DF,  p-value: 0.08252

> summary(lm.HC)

Call:
lm(formula = Weight ~ Height + Chest, data = mulreg.data)

Residuals:
    Min      1Q  Median      3Q     Max
-35.168 -14.573   1.857  13.047  27.154

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1906.3      590.9  -3.226   0.0180 *
Height        1208.2      522.0   2.315   0.0599 .
Chest          455.4      153.1   2.975   0.0248 *
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 23.52 on 6 degrees of freedom
Multiple R-squared: 0.817,      Adjusted R-squared: 0.756
F-statistic: 13.39 on 2 and 6 DF,  p-value: 0.006132

> summary(lm.LC)

Call:
lm(formula = Weight ~ Length + Chest, data = mulreg.data)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-42.637 -20.573   1.236  16.922  42.282

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -956.1      643.8  -1.485   0.1881
Length         228.5      436.8   0.523   0.6197
Chest          594.2      189.8   3.131   0.0203 *
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 31.65 on 6 degrees of freedom
Multiple R-squared: 0.6686,     Adjusted R-squared: 0.5582
F-statistic: 6.053 on 2 and 6 DF,  p-value: 0.03639

>
```

【補足事項】

変数選択法による重回帰分析のために R には step( ) が用意されている。

以下に step( ) を用いた解答例を示す。

```
> step(lm.full)
Start:  AIC=61.11
Weight ~ Height + Length + Chest

         Df Sum of Sq    RSS    AIC
- Length  1      31.2 3320.5   59.2
<none>                3289.3   61.1
- Height  1    2722.5 6011.8   64.5
- Chest   1    4608.9 7898.1   67.0

Step:  AIC=59.2
Weight ~ Height + Chest

         Df Sum of Sq    RSS    AIC
<none>                3320.5   59.2
- Height  1    2965.5 6285.9   62.9
- Chest   1    4897.7 8218.2   65.4

Call:
lm(formula = Weight ~ Height + Chest, data = mulreg.data)

Coefficients:
(Intercept)       Height        Chest
    -1906.3       1208.2        455.4
```

【練習問題 3】

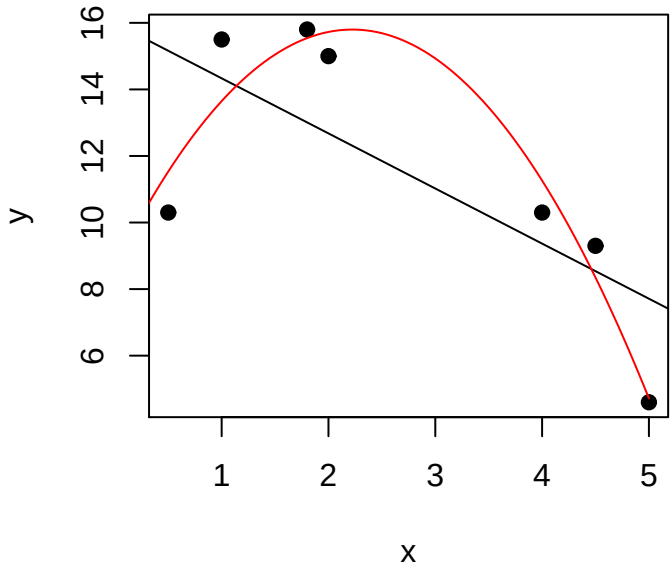
以下のデータ (データは省略) について直線回帰式と二次回帰式のいずれがよく当てはまるかを検討し、その回帰式を求めなさい。

【解答例】

直線回帰式：$y = 15.9890 - 1.6555x$、寄与率 $= 0.5169$

2 次回帰式：$y = 8.6969 + 6.3983x - 1.4367x^2$、寄与率 $= 0.9277$

このデータに対しては明らかに 2 次回帰式の方があてはまりがよい。
データの分布とそれぞれの回帰式をグラフに示すと以下のようになる。



【R のコード・出力例】

2 次回帰式の場合、x^2 の項が必要となるが、lm( ) の中のモデルの記述で、lm(y ~ x+I(x*x)) とすると、明示的に x^2 の項を用意しておかなくてもよい。なお、lm の中での式の記述で単に x*z とするとこれは変数 x と変数 z の交互作用を表す。それぞれの要素の積を記述する場合は関数 I( ) を用いて I(x*z) とする。

```
> x <- c(0.5,1.0,1.8,2.0,4.0,4.5,5.0)
> y <- c(10.3,15.5,15.8,15.0,10.3,9.3,4.6)
> lm.single <- lm(y~x)
> lm.quad <- lm(y~x+I(x*x))
> summary(lm.single)

Call:
lm(formula = y ~ x)

Residuals:
```

```
      1       2       3       4       5       6       7
-4.8613  1.1665  2.7909  2.3220  0.9329  0.7607 -3.1116

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.9890     2.2592   7.077 0.000871 ***
x            -1.6555     0.7157  -2.313 0.068653 .
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 3.14 on 5 degrees of freedom
Multiple R-squared: 0.5169,     Adjusted R-squared: 0.4203
F-statistic:  5.35 on 1 and 5 DF,  p-value: 0.06865

> summary(lm.quad)

Call:
lm(formula = y ~ x + I(x * x))

Residuals:
      1       2       3       4       5       6       7
-1.2324  1.8505  0.2574 -0.7286 -0.9665  0.9449 -0.1253

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.6969     1.8156   4.790  0.00871 **
x             6.3893     1.7162   3.723  0.02042 *
I(x * x)     -1.4367     0.3015  -4.766  0.00887 **
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 1.359 on 4 degrees of freedom
Multiple R-squared: 0.9277,     Adjusted R-squared: 0.8915
F-statistic: 25.65 on 2 and 4 DF,  p-value: 0.005233

>
> plot(x,y,pch=19)
> abline(lm.single)
> curve(8.697+6.389*x-1.437*x*x,0,5,add=T)
>
```

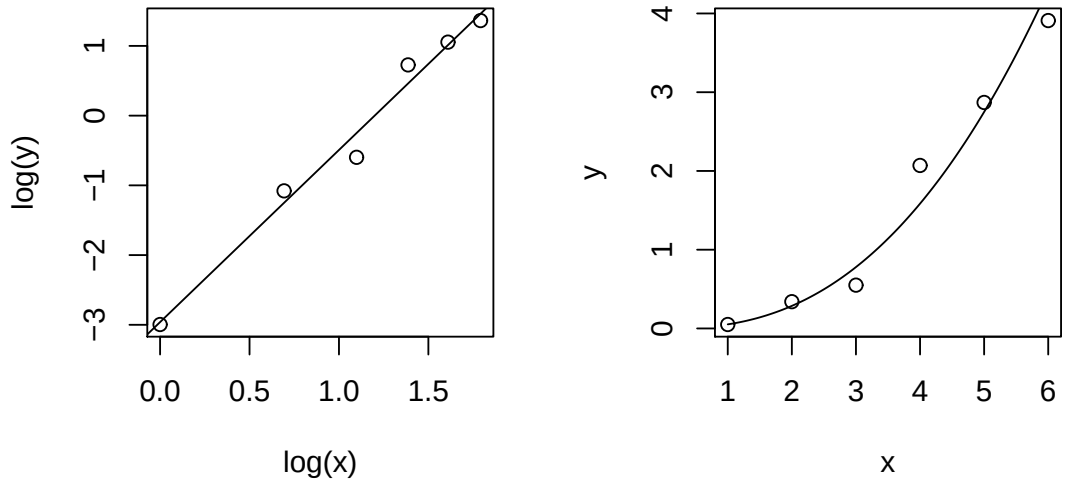
【練習問題 4】

以下のデータ (データは省略) に対して、アロメトリー (相対成長) 式 ($y = ax^b$) を当てはなさい。
ヒント：アロメトリー式の両辺対数をとると線形式となる

【解答例】

アロメトリー式 $y = ax^b$ の推定に当たってはまず、両辺対数をとって $\log y = \log a + b \log x$ と式変形する ($\log$ は自然対数)。 $\log y = Y, \log a = A, \log x = X$ で表すと、この式は $y = A + bX$ となり、単回帰式として処理が行える。

この結果、$Y = -2.9607 + 2.4860X$ と推定されるので (寄与率=0.9828)、元のアロメトリー式に戻すと $y = 0.05178x^{2.4680}$ となる ($e^{-2.9607} = 0.05178$ である)。このデータの散布図と推定曲線をグラフに示すと以下のようになる (グラフの左側は x 軸、y 軸とも対数をとったもの、右側は変換を行っていな元のデータの散布図である)。



【R のコード・出力例】

```
> x <- 1:6
> y <- c(0.05,0.34,0.55,2.07,2.87,3.91)
> arometory <- lm(log(y) ~log(x))
> summary(arometory)

Call:
lm(formula = log(y) ~ log(x))

Residuals:
       1        2        3        4        5        6
-0.03499  0.17126 -0.34845  0.26694  0.04299 -0.09775

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.9607     0.2047  -14.46 0.000133 ***
log(x)        2.4680     0.1635   15.10 0.000112 ***
```

```
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 0.2422 on 4 degrees of freedom
Multiple R-squared: 0.9828,     Adjusted R-squared: 0.9784
F-statistic:   228 on 1 and 4 DF,  p-value: 0.0001121
>
> par(mfrow=c(1,2))
> plot(log(x),log(y))
> abline(arometory)
> plot(x,y)
> curve(exp(-2.9607)*x^(2.4680),1,6,add=T)
```

【練習問題 5】

分散分析の例題 (表 13.2 のデータ) について、品種と産次の交互作用を考慮した分散分析を行いなさい。

【解答例】

計算には R の aov( ) を用いる。基本的には、図 13.3 に示したコードと同じであるが avo の関数内の記述を anova.result <- aov(rhs ~ breed*parity,data=Yield.data) と変更する。rhs ~ breed*parity は rhs ~ breed+parity+breed*parity と同じ内容となる。すなわち、チルダ (~) の右側に、交互作用のみを記述すると、交互作用として考慮する主効果が自動的にモデルに含まれる。

分散分析の結果を以下に示す。

交互作用を考慮した分散分析結果 (平方和：Type I)

変動因	自由度	平方和	平均平方	F 値	有意確率
品種	1	0.9375	0.9375	6.7568	0.06009
産次	2	4.6961	2.3480	16.9228	0.01117
品種 × 産次	2	0.3764	0.1882	1.3565	0.35505
誤差	4	0.5550	0.1387		

【R のコード・出力例】

```
> Breed <- factor(c(1,2,2,1,2,2,1,1,2,2))
> Parity <- factor(c(1,1,1,3,3,3,5,5,5,5))
> Yield <- c(5.0,6.2,6.4,6.6,6.4,7.3,6.7,7.2,7.8,7.9)
> Yield.data <- data.frame(breed=Breed,parity=Parity,rhs=Yield)
> anova.result <- aov(rhs ~ breed*parity,data=Yield.data)
> summary(anova.result)
             Df Sum Sq Mean Sq F value  Pr(>F)
breed         1 0.9375  0.9375  6.7568 0.06009 .
parity        2 4.6961  2.3480 16.9228 0.01117 *
breed:parity  2 0.3764  0.1882  1.3565 0.35505
Residuals     4 0.5550  0.1387
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1
>
```

【補足事項】

同じモデルに対して平方和の計算を Type III で行った場合の R のコード・出力例を以下に示す。この場合、パッケージ car がインストールされていなければならない。

```
> library(car)
> Breed <- factor(c(1,2,2,1,2,2,1,1,2,2))
> Parity <- factor(c(1,1,1,3,3,3,5,5,5,5))
> Yield <- c(5.0,6.2,6.4,6.6,6.4,7.3,6.7,7.2,7.8,7.9)
> Yield.data <- data.frame(breed=Breed,parity=Parity,rhs=Yield)
> Anova.result <- Anova(lm(rhs ~ breed*parity,data=Yield.data),Type="III")
> Anova.result
Anova Table (Type II tests)
```

```
Response: rhs
             Sum Sq Df F value  Pr(>F)
breed        1.6019  1 11.5453 0.02733 *
parity       4.6961  2 16.9228 0.01117 *
breed:parity 0.3764  2  1.3565 0.35505
Residuals    0.5550  4
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1
>
```

14章1節練習問題解答例

【練習問題 1】

次の式は、生物個体の成長に関する一般モデルの一つである [West ら,*Nature,* **413** ,628,(2001)]。このモデルを表 14.1 のデータ (データ省略) に当てはめ、パラメータ m_0、M および a を推定しなさい。

$$\left(\frac{y_t}{M}\right)^{1/4} = 1 - \left\{1 - \left(\frac{m_0}{M}\right)^{1/4}\right\} e^{-at/(4M^{1/4})}$$

ただし、m_0、M および a は、それぞれ出生時体重、成熟体重および係数である。なお、R で計算する場合には、この式をあらかじめ $y_t = \cdots$ の形に変形して記述する必要がある。

【解答例】

West の非線形成長曲線は

$$y_t = M \times \left[1 - \left\{1 - \left(\frac{m_0}{M}\right)^{1/4}\right\} \times e^{-at/(4M^{1/4})}\right]^4$$

y_t ： t 月齢での体重測定値

m_o ： 出生時体重

M ： 成熟体重

a ： 係数

と表される。

パラメータ m_0、M および a の推定には R の関数 nls( ) を用いる。nls( ) は非線形モデル分析のための関数であり、推定すべきパラメータ m_o、M および a の初期値を start 引数で与えなければならない。ここでは、初期値として、$M = 400$、$m_o = 20$、および $a = 1$ を用いる。

上記の非線形成長曲線の式の記号を以下に示す R のコードではそのまま用いた (ただし、体重測定値 y、月齢 t、生時体重 mo、成熟体重 M、係数 a とした)。

【R のコード・出力例】

```
> y <- c(27.7, 33.8, 46.5, 64.7, 87.0, 112.3, 254.4, 369.7, 428.0,
+        436.1, 420.8, 412.1, 426.0, 447.8, 415.2, 416.8) #体重の入力
> t <- c(0, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55) #月齢の入力
> #West model (west et al., 2001)
> West.result <- nls(y ~ M*(1-(1-(mo/M)^(1/4))*exp(-a/(4*M^(1/4))*t))^4,
+                start=list(M=400, mo=20, a=1))
> summary(West.result)

Formula: y ~ M * (1 - (1 - (mo/M)^(1/4)) * exp(-a/(4 * M^(1/4)) * t))^4

Parameters:
   Estimate Std. Error t value Pr(>|t|)
M  431.3852     6.5566   65.79  < 2e-16 ***
mo   6.7134     3.8809    1.73    0.107
a    3.2259     0.2592   12.45 1.35e-08 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1
```

```
Residual standard error: 16.41 on 13 degrees of freedom

Number of iterations to convergence: 8
Achieved convergence tolerance: 4.047e-06

> West.predict <- predict(West.result)
> summary(West.result)

Formula: y ~ M * (1 - (1 - (mo/M)^(1/4)) * exp(-a/(4 * M^(1/4)) * t))^4

Parameters:
   Estimate Std. Error t value Pr(>|t|)
M  431.3852     6.5566   65.79  < 2e-16 ***
mo   6.7134     3.8809    1.73    0.107
a    3.2259     0.2592   12.45 1.35e-08 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 16.41 on 13 degrees of freedom

Number of iterations to convergence: 8
Achieved convergence tolerance: 4.047e-06

> AIC(West.result)
[1] 139.6177
> cor(West.predict,y)^2
[1] 0.9929034
> plot(t,y,main="Cattle weight",xlab="Age(m)",ylab="weight")
> points(West.predict~t,typ="l")
>
```

計算結果をまとめると以下のようになる。ここで、パラメータ M、m_0 および a の値は上記の summary(West.result) の出力から得られる。また、決定係数は cor(West.predict,y)^2、AIC は AIC(West.result) の出力値である。

計算結果

モデル	決定係数	AIC	$\hat{M}$	$\hat{m_0}$	$\hat{a}$
West	0.993	139.62	431.39	6.71	3.23

データの分布と推定した曲線を次ページの図 1 に示した。この処理は上記の

```
> plot(t,y,main="Cattle weight",xlab="Age(m)",ylab="weight")
> points(West.predict~t,typ="l")
```

の出力である。

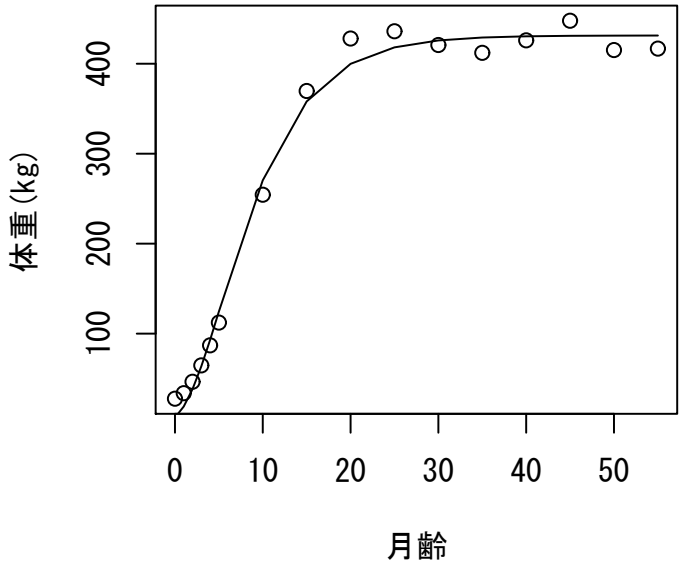


図 1: West の非線形成長曲線のプロット図

【練習問題 2】

1 章の表 1.4 に示した日本人女性の身長と体重 (データは省略) に、表 14.2 の非線形成長モデルを当てはめ、発育様相を図示しなさい。なお、22〜24 か月齢は 22 か月齢として計算すること。

【解答例】

R での計算例、計算結果の概要ならびに測定値と当てはめたモデルのプロット図 2 は、「R のコード・出力例」の項に示した。なお、ここでは、モデルの当てはめに R の nls 関数を利用し、デフォルトとして用意されているガウス・ニュートンアルゴリズムを用いたが、このアルゴリズムでは設定した初期値の値によって収束しないことがあるので、留意を要する。また、この練習問題における体重のデータの場合には、Richards モデルを収束させるためには、［A=50, B=-3×10^5, k=1, M=-0.1］のような極端な初期値を設定する必要がある。

身長および体重の分析結果は次のとおりである。

計算結果 (身長)

モデル	決定係数	AIC	$\hat{A}$	$\hat{B}$	$\hat{k}$	$\hat{M}$
Brody	0.978	73.54	162.61	1.15	0.22	
Logistic	0.983	68.62	161.32	2.33	0.28	
Gompertz	0.981	71.13	161.90	1.63	0.25	
Bertalanffy	0.980	71.94	162.12	0.48	0.24	
Richards	0.999	19.23	158.2	-4.239×10^5	1.05	-0.05

計算結果 (体重)

モデル	決定係数	AIC	$\hat{A}$	$\hat{B}$	$\hat{k}$	$\hat{M}$
Brody	0.956	76.22	57.83	1.84	0.16	
Logistic	0.980	64.44	54.09	14.57	0.34	
Gompertz	0.970	70.51	55.35	4.97	0.25	
Bertalanffy	0.966	72.47	55.98	1.18	0.22	
Richards	0.998	33.93	51.96	-3.600×10^6	1.14	-0.11

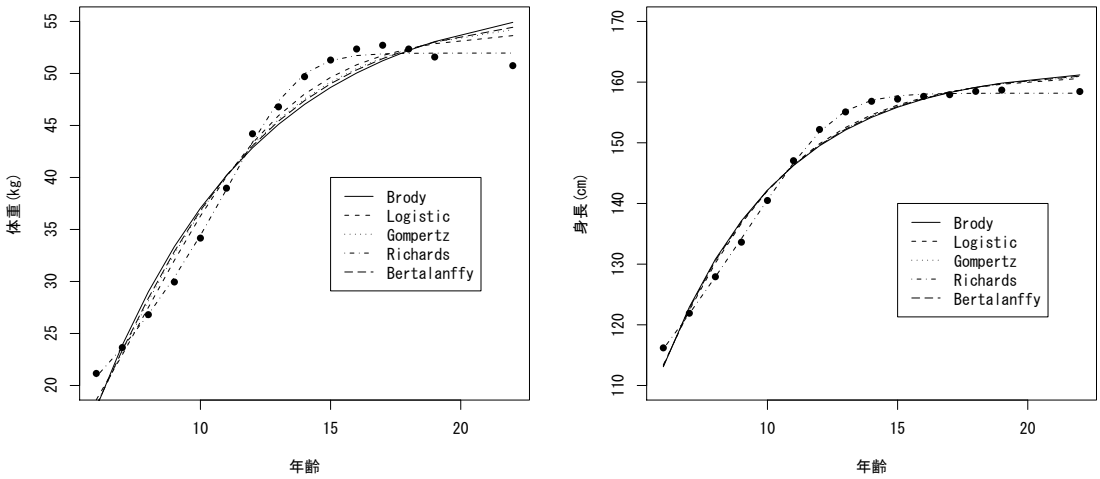


図 2: 測定値と当てはめた非線形成長モデルのプロット図

【R のコード・出力例】

```
>#データフレームの用意
> Height <- c(116.20,121.91,127.92,133.62,140.49,147.05,152.19,155.09,
+             156.84,157.20,157.65,157.93,158.48,158.69,158.45)
> Weight <- c(21.16,23.65,26.81,29.95,34.17,38.97,44.20,46.79,49.69,51.29,
+             52.35,52.71,52.35,51.58,50.75)
> age    <- c(6,7,8,9,10,11,12,13,14,15,16,17,18,19,22)
> Woman.data <- data.frame(age=age,height=Height,weight=Weight)
>#身長についての 5 つのモデルの当てはめ
> #Brody model (Brody, 1945)
> Brody.height.result <- nls(height ~ A*(1-B*exp(-K*age)),
+ start=list(A=160, B=1, K=0.1),data=Woman.data)
> #Logistic model (Robertson, 1923)
> Logistic.height.result <- nls(height ~ A/(1+B*exp(-k*age)),
+ start=list(A=160, B=10, k=0.1),data=Woman.data)
> #Gompertz model(Rogers et al., 1987)
> Gompertz.height.result <- nls(height ~ A*exp(-B*exp(-k*age)),
+ start=list(A=160, B=1, k=0.1),data=Woman.data)
> #Richards model (Nahashon, 2006)
> Richards.height.result <- nls(height ~ A*(1-B*exp(-k*age))^M,
+ start=list(A=155, B=-8, k=0.2, M=-0.2),data=Woman.data)
> #von Bertalanffy model (von Bertalanffy, 1957)
> Bertalanffy.height.result <- nls(height ~ A*(1-B*exp(-k*age))^3,
+ start=list(A=160, B=0.1, k=0.1),data=Woman.data)
> #グラフの作成
> Brody.height.predict <- predict(Brody.height.result)
> Logistic.height.predict <- predict(Logistic.height.result)
> Gompertz.height.predict <- predict(Gompertz.height.result)
> Richards.height.predict <- predict(Richards.height.result)
> Bertalanffy.height.predict <- predict(Bertalanffy.height.result)
```

```
> plot(Woman.data$age,Woman.data$height,main="Height",
+ xlab="Age(y)",ylab="height")
> matplot(c(6,22),c(110,170),main="Growth curve",
+         xlab="Age(y)",ylab="height",type="n")
> points(Brody.height.predict~age,type="l",col="red",lwd=2)
> points(Logistic.height.predict~age,type="l",col="blue",lwd=2)
> points(Gompertz.height.predict~age,type="l",col="green",lwd=2)
> points(Richards.height.predict~age,type="l",col="black",lwd=2)
> points(Bertalanffy.height.predict~age,type="l",col="pink",lwd=2)
> points(height~age,pch="*",data=Woman.data)
> legend(15,140,legend=c("Brody","Logistic","Gompertz","Richards","Bertalanffy"),
+        col=c("red","blue","green","black","pink"),
+        lty=c(1))
>#それぞれのモデルの推定結果の表示 (summary を用いる)
> summary(Brody.height.result)

Formula: height ~ A * (1 - B * exp(-K * age))

Parameters:
   Estimate Std. Error t value Pr(>|t|)
A 162.61445    1.92438  84.502  < 2e-16 ***
B   1.15016    0.19688   5.842 7.94e-05 ***
K   0.22147    0.02768   8.002 3.75e-06 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 2.405 on 12 degrees of freedom

Number of iterations to convergence: 9
Achieved convergence tolerance: 6.646e-06

> summary(Logistic.height.result)

Formula: height ~ A/(1 + B * exp(-k * age))

Parameters:
   Estimate Std. Error t value Pr(>|t|)
A 161.31762    1.34710 119.752  < 2e-16 ***
B   2.33174    0.40140   5.809 8.36e-05 ***
k   0.28467    0.02619  10.869 1.44e-07 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 2.041 on 12 degrees of freedom

Number of iterations to convergence: 10
Achieved convergence tolerance: 4.478e-06
```

```
> summary(Gompertz.height.result)

Formula: height ~ A * exp(-B * exp(-k * age))

Parameters:
   Estimate Std. Error t value Pr(>|t|)
A 161.89622    1.59888 101.256  < 2e-16 ***
B   1.62844    0.27963   5.824 8.17e-05 ***
k   0.25270    0.02693   9.385 7.08e-07 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 2.219 on 12 degrees of freedom

Number of iterations to convergence: 10
Achieved convergence tolerance: 4.559e-06

> summary(Richards.height.result)

Formula: height ~ A * (1 - B * exp(-k * age))^M

Parameters:
    Estimate Std. Error t value Pr(>|t|)
A  1.582e+02  1.714e-01 922.580  < 2e-16 ***
B -4.239e+05  5.566e+05  -0.762    0.462
k  1.051e+00  9.949e-02  10.563 4.26e-07 ***
M -4.658e-02  5.273e-03  -8.834 2.51e-06 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 0.3844 on 11 degrees of freedom

Number of iterations to convergence: 28
Achieved convergence tolerance: 6.645e-06

> summary(Bertalanffy.height.result)

Formula: height ~ A * (1 - B * exp(-k * age))^3

Parameters:
   Estimate Std. Error t value Pr(>|t|)
A 162.11808    1.69783  95.486  < 2e-16 ***
B   0.48279    0.08281   5.830 8.09e-05 ***
k   0.24221    0.02717   8.913 1.22e-06 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1
```

```
Residual standard error: 2.28 on 12 degrees of freedom

Number of iterations to convergence: 10
Achieved convergence tolerance: 4.31e-06

>#それぞれのモデルについての AIC の算出
> AIC(Brody.height.result)
[1] 73.54243
> AIC(Logistic.height.result)
[1] 68.62319
> AIC(Gompertz.height.result)
[1] 71.12748
> AIC(Richards.height.result)
[1] 19.23418
> AIC(Bertalanffy.height.result)
[1] 71.94274
>
>#それぞれのモデルについての寄与率の算出
> cor(Brody.height.predict,Woman.data$height)^2
[1] 0.9775266
> cor(Logistic.height.predict,Woman.data$height)^2
[1] 0.983821
> cor(Gompertz.height.predict,Woman.data$height)^2
[1] 0.980872
> cor(Richards.height.predict,Woman.data$height)^2
[1] 0.9994735
> cor(Bertalanffy.height.predict,Woman.data$height)^2
[1] 0.9798016
>
>#体重についての 5 つのモデルの当てはめ
> #Brody model (Brody, 1945)
> Brody.weight.result <- nls(weight ~ A*(1-B*exp(-K*age)),
+ start=list(A=50, B=1, K=0.1),data=Woman.data)
> #Logistic model (Robertson, 1923)
> Logistic.weight.result <- nls(weight ~ A/(1+B*exp(-k*age)),
+ start=list(A=50, B=10, k=0.1),data=Woman.data)
> #Gompertz model(Rogers et al., 1987)
> Gompertz.weight.result <- nls(weight ~ A*exp(-B*exp(-k*age)),
+ start=list(A=50, B=1, k=0.1),data=Woman.data)
> #Richards model (Nahashon, 2006)
> Richards.weight.result <- nls(weight ~ A*(1-B*exp(-k*age))^M,
+ start=list(A=50, B=-3e+05, k=1., M=-0.1),data=Woman.data)
> #von Bertalanffy model (von Bertalanffy, 1957)
> Bertalanffy.weight.result <- nls(weight ~ A*(1-B*exp(-k*age))^3,
+ start=list(A=50, B=1, k=0.1),data=Woman.data)
> #グラフの作成
```

```
> Brody.weight.predict <- predict(Brody.weight.result)
> Logistic.weight.predict <- predict(Logistic.weight.result)
> Gompertz.weight.predict <- predict(Gompertz.weight.result)
> Richards.weight.predict <- predict(Richards.weight.result)
> Bertalanffy.weight.predict <- predict(Bertalanffy.weight.result)
>
> plot(Woman.data$age,Woman.data$weight,main="Weight",
+ xlab="Age(y)",ylab="Weight")
>
> matplot(c(6,22),c(20,55),main="Growth curve",xlab="Age(y)",ylab="weight",type="n")
> points(Brody.weight.predict~age,type="l",col="red")
> points(Logistic.weight.predict~age,type="l",col="blue")
> points(Gompertz.weight.predict~age,type="l",col="green")
> points(Richards.weight.predict~age,type="l",col="black")
> points(Bertalanffy.weight.predict~age,type="l",col="pink")
>
> points(weight~age,pch="*",data=Woman.data)
> legend(15,40,legend=c("Brody","Logistic","Gompertz","Richards","Bertalanffy"),
+        col=c("red","blue","green","black","pink"),
+        lty=c(1))
>#それぞれのモデルの推定結果の表示 (summary を用いる)
> summary(Brody.weight.result)

Formula: weight ~ A * (1 - B * exp(-K * age))

Parameters:
  Estimate Std. Error t value Pr(>|t|)
A 57.83472    3.27303  17.670 5.89e-10 ***
B  1.84334    0.40846   4.513  0.00071 ***
K  0.16348    0.03467   4.716  0.00050 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 2.629 on 12 degrees of freedom

Number of iterations to convergence: 12
Achieved convergence tolerance: 3.947e-06

> summary(Logistic.weight.result)

Formula: weight ~ A/(1 + B * exp(-k * age))

Parameters:
  Estimate Std. Error t value Pr(>|t|)
A 54.08659    1.18193  45.761 7.74e-15 ***
B 14.56719    3.73867   3.896  0.00212 **
k  0.33914    0.03294  10.295 2.61e-07 ***
```

```
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 1.775 on 12 degrees of freedom

Number of iterations to convergence: 14
Achieved convergence tolerance: 3.314e-06

> summary(Gompertz.weight.result)

Formula: weight ~ A * exp(-B * exp(-k * age))

Parameters:
  Estimate Std. Error t value Pr(>|t|)
A 55.34713    1.82821  30.274 1.06e-12 ***
B  4.97202    1.19593   4.157  0.00133 **
k  0.24929    0.03379   7.378 8.53e-06 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 2.173 on 12 degrees of freedom

Number of iterations to convergence: 12
Achieved convergence tolerance: 5.352e-06

> summary(Richards.weight.result)

Formula: weight ~ A * (1 - B * exp(-k * age))^M

Parameters:
    Estimate Std. Error t value Pr(>|t|)
A  5.196e+01  3.006e-01 172.861  < 2e-16 ***
B -3.596e+06  1.200e+07  -0.300  0.77006
k  1.140e+00  2.385e-01   4.781  0.00057 ***
M -1.107e-01  2.711e-02  -4.083  0.00181 **
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 0.6275 on 11 degrees of freedom

Number of iterations to convergence: 15
Achieved convergence tolerance: 3.576e-06

> summary(Bertalanffy.weight.result)

Formula: weight ~ A * (1 - B * exp(-k * age))^3
```

```
Parameters:
  Estimate Std. Error t value Pr(>|t|)
A 55.98201    2.17341  25.758 7.15e-12 ***
B  1.17827    0.27585   4.271  0.00109 **
k  0.22015    0.03405   6.466 3.09e-05 ***
---
Signif. codes:  0 ‘ *** ’ 0.001 ‘ ** ’ 0.01 ‘ * ’ 0.05 ‘ . ’ 0.1 ‘   ’ 1

Residual standard error: 2.32 on 12 degrees of freedom

Number of iterations to convergence: 14
Achieved convergence tolerance: 6.365e-06

>#それぞれのモデルについての AIC の算出
> AIC(Brody.weight.result)
[1] 76.2209
> AIC(Logistic.weight.result)
[1] 64.44308
> AIC(Gompertz.weight.result)
[1] 70.51081
> AIC(Richards.weight.result)
[1] 33.93666
> AIC(Bertalanffy.weight.result)
[1] 72.46707
>
>#それぞれのモデルについての寄与率の算出
> cor(Brody.weight.predict,Woman.data$weight)^2
[1] 0.9560442
> cor(Logistic.weight.predict,Woman.data$weight)^2
[1] 0.980139
> cor(Gompertz.weight.predict,Woman.data$weight)^2
[1] 0.970057
> cor(Richards.weight.predict,Woman.data$weight)^2
[1] 0.997705
> cor(Bertalanffy.weight.predict,Woman.data$weight)^2
[1] 0.9658294
```

【練習問題 3】

市販の牛肉 417 サンプルから脂肪を抽出し、いくつかの脂肪酸の組成をガスクロマトグラフにより測定した結果、下記の相関係数を得た (相関係数の表は省略)。これら五つの脂肪酸の相関係数に対して主成分分析を適用し、各主成分の特徴と形質の分類を検討しなさい。

【解答例】

固有ベクトルは

$$\begin{pmatrix} -0.4182751 & -0.51690879 & 0.1884612 & 0.7088396 & -0.1409969 \\ -0.5200952 & -0.20876935 & 0.3755776 & -0.6341430 & -0.3777835 \\ 0.1937091 & -0.68586230 & -0.6466746 & -0.2394871 & -0.1285593 \\ -0.4231845 & 0.46766217 & -0.5782785 & 0.1463334 & -0.4963743 \\ 0.5813248 & 0.01027860 & 0.2661387 & 0.1290019 & -0.7579476 \end{pmatrix}$$

固有値は、

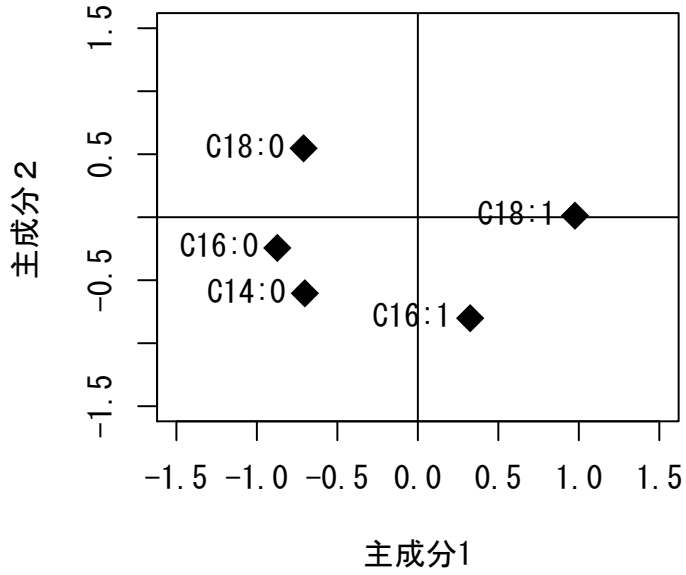
$$\lambda = \begin{pmatrix} 2.817258596 \\ 1.366527842 \\ 0.567855648 \\ 0.242189117 \\ 0.006168797 \end{pmatrix}$$

であるので、1 より大きい第 2 主成分までを取り上げることにする。第 2 主成分までの累積寄与率は、0.8367573 であり、5 形質がもつバラツキの 84%を考慮していることになる。第 1 主成分と C14:0 の因子負荷量は、−0.7020618 であり、第 2 主成分と C14:0 の因子負荷量は −0.604259 となる。同様に残りの脂肪酸も計算し、因子負荷量を表と図で表せば次のようになる。

主成分	C14:0	C16:0	C16:1	C18:0	C18:1
第 1 主成分	-0.702	-0.873	0.325	-0.710	0.976
第 2 主成分	-0.604	-0.244	-0.802	0.547	0.012

これより，横軸の第 1 主成分は C16:0 や C18:0 のような飽和脂肪酸と C16:1 や C18:1 のような不飽和脂肪酸とを区別している成分であり，縦軸の第 2 主成分は炭素数の違いを表している主成分であるといえる。さらに，C14:0 と C16:0 は近くに位置し，他の脂肪酸と比較的似通った相関関係をもつ形質であることも明らかとなった。

因子負荷量の図および計算に用いた R のコードとその実行例は次ページに示した。



【R のコード・出力例】

```
> #相関行列を作成
> cmat <- matrix(c(1,0.692,0.146,0.132,-0.641,
+          0.692,1,-0.189,0.342,-0.816,
+          0.146,-0.189,1,-0.465,0.203,
+          0.132,0.342,-0.465,1,-0.767,
+          -0.641,-0.816,0.203,-0.767,1),ncol=5,byrow=T)
> evv <- eigen(cmat)  #固有値の計算
> evv    #計算結果の表示
$values
[1] 2.817258596 1.366527842 0.567855648 0.242189117 0.006168797

$vectors
           [,1]        [,2]       [,3]       [,4]       [,5]
[1,] -0.4182751 -0.51690879  0.1884612  0.7088396 -0.1409969
[2,] -0.5200952 -0.20876935  0.3755776 -0.6341430 -0.3777835
[3,]  0.1937091 -0.68586230 -0.6466746 -0.2394871 -0.1285593
[4,] -0.4231845  0.46766217 -0.5782785  0.1463334 -0.4963743
[5,]  0.5813248  0.01027860  0.2661387  0.1290019 -0.7579476

>
```

evv$values が固有値、evv$vectors が固有ベクトルの出力結果である。

【補足事項】

R には主成分分析のための関数 princomp() と prcomp() が標準で用意されている。データから直接主成分分析を行う場合、これらの関数が利用できる (本問題のように相関行列のみが与えられている場合はこれらの関数は利用できない)。

表 14.5 の横綱の身長と体重のデータについて princomp() を用いた例を示す。

【R のコード・出力例】

```
> height <-c(192,184,192,180,185,203,189,189,181,200)
> weight <-c(153,147,237,134,154,233,143,211,151,150)
> data <- matrix(c(height,weight),ncol=2)
> #主成分分析の実行
> data.pca <- princomp(data,cor=T,scores=T)
> #主成分分析の結果表示
> summary(data.pca)
Importance of components:
                          Comp.1    Comp.2
Standard deviation     1.2419447 0.6764417
Proportion of Variance 0.7712133 0.2287867
Cumulative Proportion  0.7712133 1.0000000
> #固有値の計算
> latent.v <- (data.pca$sdev)^2
> latent.v
   Comp.1    Comp.2
1.5424266 0.4575734
> #因子負荷量の計算
> pc1.factor <- sqrt(latent.v[1])*data.pca$load     ings[,1]
> pc1.factor
[1] 0.8781875 0.8781875
> pc2.factor <- sqrt(latent.v[2])*data.pca$loadings[,2]
> pc2.factor
[1] -0.4783165  0.4783165
> #固有ベクトルの表示
> data.pca$loadings

Loadings:
     Comp.1 Comp.2
[1,]  0.707 -0.707
[2,]  0.707  0.707

               Comp.1 Comp.2
SS loadings       1.0    1.0
Proportion Var    0.5    0.5
Cumulative Var    0.5    1.0
> #主成分得点の出力
```

```
> data.pca$scores
          Comp.1      Comp.2
 [1,] -0.1005593 -0.59155830
 [2,] -0.9996196  0.08057823
 [3,]  1.4879073  0.99690829
 [4,] -1.6382529  0.22754333
 [5,] -0.7690476  0.11475064
 [6,]  2.4924639 -0.15893082
 [7,] -0.5842619 -0.48606206
 [8,]  0.7016397  0.79983947
 [9,] -1.2185778  0.45081891
[10,]  0.6283082 -1.43388769
>
```

princomp() では、固有値は出力されない。固有値は summary() の出力欄の Standard deviation を 2 乗したものとして得られる。上記の例では固有値は

```
latent.v <- (data.pca$sdev)^2
```

として求めている。また、因子負荷量を出力されていないので、固有値と固有ベクトルを用いて計算しなければならない。上記の例では

```
> #因子負荷量の計算
> pc1.factor <- sqrt(latent.v[1])*data.pca$loadings[,1]
> pc1.factor
[1] 0.8781875 0.8781875
> pc2.factor <- sqrt(latent.v[2])*data.pca$loadings[,2]
> pc2.factor
```

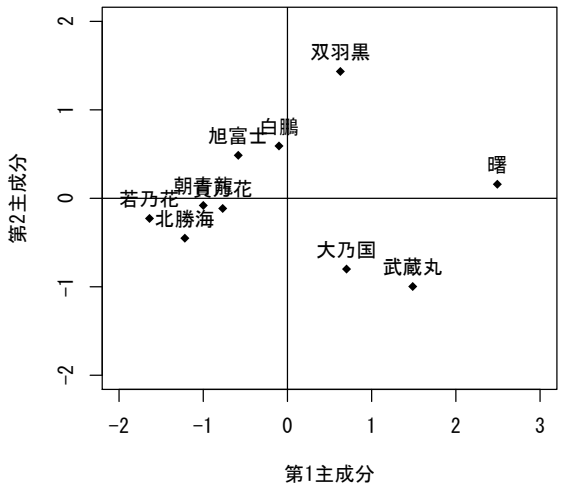
として、第 1 主成分および第 2 主成分の因子負荷量を計算している。なお、princomp の結果オブジェクト名 (上記の例では data.pca) の後に$sdev、$loadings を付けると、それぞれ、標準偏差、固有ベクトルが参照できる。

第 14 章 2.1 節の計算結果と R の計算結果では、第 2 主成分の固有ベクトルの符号が逆になっているが、これは用いたアルゴリズムの相違などで生じることがあるが、結果の解釈には影響を与えない。

主成分得点の分布を次ページに図で示した (第 14 章 2.1 節の計算結果と合わすために、第 2 主成分の得点の符号を反転している)。

作図のための R のコードを以下に示す。

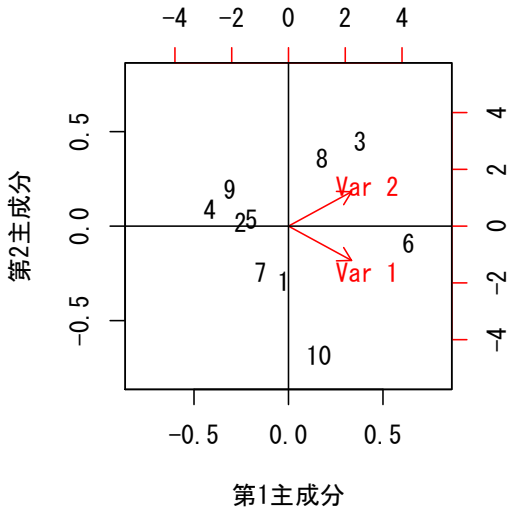
```
> name <-c("白鵬","朝青龍","武蔵丸","若乃花","貴乃花",
+          "曙","旭富士","大乃国","北勝海","双羽黒")
>
> plot(data.pca$scores[,1],data.pca$scores[,2]*-1,
+      xlim=c(-2,3),ylim=c(-2,2),pch=18,
+      xlab="第 1 主成分",ylab="第 2 主成分")
> text(data.pca$scores[,1],data.pca$scores[,2]*-1,name,pos=3)
> abline(h=0)
> abline(v=0)
```



R では、主成分分析の結果の視覚的な表示のために、biplot() と screeplot() が用意されている。いずれも引数には princomp の結果オブジェクトを与える。

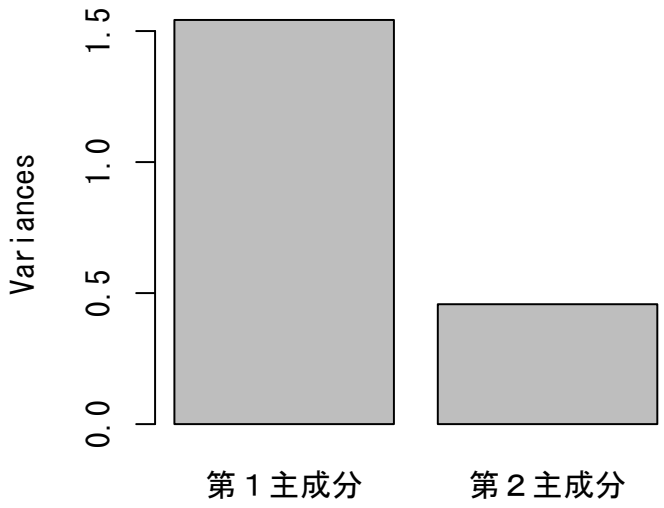
biplot() による作図のための R コードを以下に示した。biplot() はデータの主成分得点の分布と、変数の固有ベクトルを同時に表示する。

```
> biplot(data.pca,xlim=c(-0.8,0.8),ylim=c(-0.8,0.8))
> abline(h=0)
> abline(v=0)
```



一方、screeplot() は、各主成分軸の分散の大きさを棒グラフで表示する。

```
>screeplot(data.pca)
```



14章2節練習問題解答例

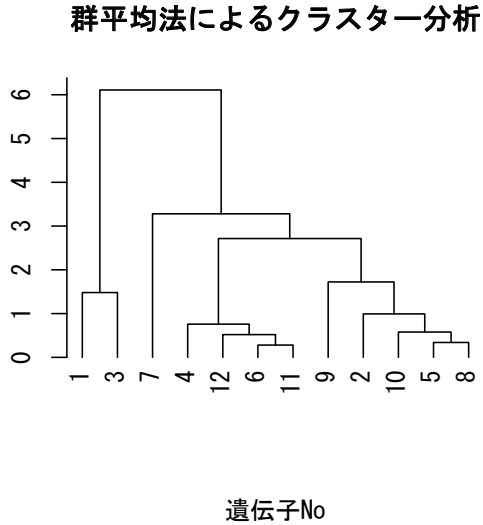
【練習問題4】

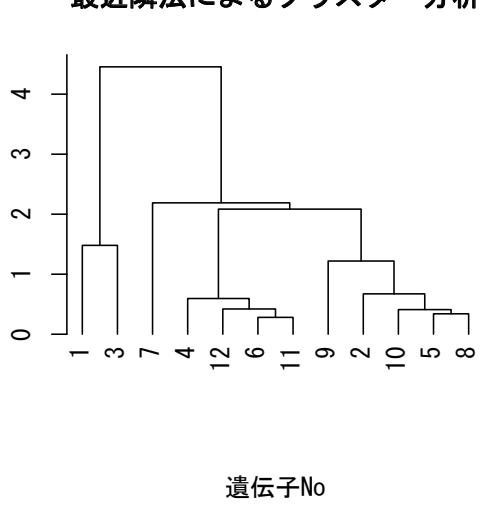
マイクロアレイ法による実験から得られた遺伝子の発現量に関する以下のデータ (データは省略) に対して、R を用いて群平均法、最近隣法および最遠隣法によるクラスター分析を行い、三つの方法間で結果を比較しなさい。

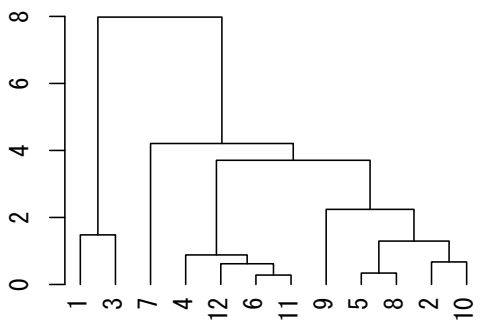
【解答例】

R を用いて群平均法、最近隣法と最遠隣法で分析すると、以下のようなデンドログラムが得られる。

この問題では、最遠隣法で作成されるデンドログラムと、群平均法および最近隣法で作成されるデンドログラムの間で違いが認められ、異なる過程でクラスターが形成されていることが分かる。一方、群平均法と最近隣法では結果に大きな違いは認められないが、データの構成によっては、これら二つの方法の間でも結果に大きな違いを生じることがある。







【R のコード・出力例】

```
> rawData <- matrix(c(1,5.31,6.21,
+ 2,3.12,2.33,
+ 3,3.98,6.86,
+ 4,1.01,0.09,
+ 5,2.14,2.08,
+ 6,0.61,0.53,
+ 7,4.29,0.48,
+ 8,1.83,2.22,
+ 9,1.03,3.14,
+ 10,2.54,1.99,
+ 11,0.33,0.51,
+ 12,0.13,0.14),ncol=3,byrow=T)
> rawData
      [,1] [,2] [,3]
 [1,]    1 5.31 6.21
 [2,]    2 3.12 2.33
 [3,]    3 3.98 6.86
 [4,]    4 1.01 0.09
 [5,]    5 2.14 2.08
 [6,]    6 0.61 0.53
 [7,]    7 4.29 0.48
 [8,]    8 1.83 2.22
 [9,]    9 1.03 3.14
[10,]   10 2.54 1.99
[11,]   11 0.33 0.51
[12,]   12 0.13 0.14
> rawData[,1] <- factor(rawData[,1])    #遺伝子番号を factor に変換
```

```
> rawData.dist <- dist(rawData[,2:3])          #ユークリッド距離で距離行列を作成
> round(rawData.dist,3)
       1     2     3     4     5     6     7     8     9    10    11
2  4.455
3  1.480 4.611
4  7.480 3.077 7.393
5  5.206 1.011 5.122 2.288
6  7.372 3.089 7.171 0.595 2.178
7  5.820 2.189 6.388 3.303 2.680 3.680
8  5.294 1.295 5.114 2.282 0.340 2.084 3.013
9  5.267 2.241 4.748 3.050 1.535 2.644 4.208 1.219
10 5.048 0.672 5.078 2.439 0.410 2.420 2.311 0.746 1.898
11 7.569 3.331 7.324 0.799 2.396 0.281 3.960 2.275 2.722 2.660
12 7.980 3.706 7.745 0.881 2.794 0.618 4.174 2.686 3.132 3.038 0.421
> rawData.cluster <- hclust(rawData.dist,method="average")  #群平均法でクラスター分析
> plot(rawData.cluster,hang=-1,
+       main="群平均法によるクラスター分析",
+       xlab="遺伝子 No",
+       sub="")                              #デンドログラム (樹形図) の出力
>
>
> rawData.cluster <- hclust(rawData.dist,method="single")  #最近隣法でクラスター分析
> plot(rawData.cluster,hang=-1,
+        main="最近隣法によるクラスター分析",
+        xlab="遺伝子 No",
+        sub="")                               #デンドログラム (樹形図) の出力
>
> rawData.cluster <- hclust(rawData.dist,method="complete")  #最遠隣法でクラスター分析
> plot(rawData.cluster,hang=-1,
+       main="最遠隣法によるクラスター分析",
+       xlab="遺伝子 No",
+       sub="")                              #デンドログラム (樹形図) の出力
>
>
```

14章3節練習問題解答例

【練習問題5】

表現型値を選抜基準とした選抜において、選抜の正確度、すなわち表現型値と相加的遺伝子型値の相関が遺伝率の平方根になることを示しなさい。

【解答例】

表現型値を P、相加的遺伝子型値を A、非相加的遺伝子型値を含む環境偏差を e とすれば、

$$P = A + e$$

である。選抜の正確度 (r_{PA}) は

$$r_{PA} = \frac{COV(P, A)}{\sqrt{V(P) \cdot V(A)}} = \frac{COV(A + e, A)}{\sqrt{V(P) \cdot V(A)}} = \frac{V(A)}{\sqrt{V(P) \cdot V(A)}} = \sqrt{\frac{V(A)}{V(P)}}$$

と書けるので、遺伝率 $h^2 = V(A)/V(P)$ の平方根になる。

【練習問題 6】

ガラパゴス諸島に生息する鳥のダーウィン・フィンチは、1977 年に起こった干ばつで多くの個体が死亡した。この鳥は種子を食料としているが、干ばつのときに食料として利用できた種子は固く大きいものに限られた。このため干ばつ後に生き残った個体は、干ばつ前の個体に比べてくちばしが大きい傾向があった。干ばつ前の集団のくちばしの幅の平均値を 9.40mm、干ばつ後の集団のくちばしの幅の平均値を 9.96mm としたとき、どのような進化的応答が期待できるか。なお、くちばしの幅の遺伝率は 0.4 とする。

【解答例】

選抜差 ΔP は、干ばつ後の集団のくちばしの幅の平均値と干ばつ前の集団のくちばしの幅の平均値の差、すなわち

$$\Delta P = 9.96 - 9.40 = 0.56$$

として得られる。進化的応答 ΔG は、遺伝的改良量と同じようにして選抜差と遺伝率 h^2 の積、すなわち

$$\Delta G = h^2 \Delta P = 0.4 \times 0.56 = 0.224(mm)$$

として得られる。

【練習問題 7】

ブロイラーのある系統で、5 から 9 週齢のあいだの体重の増加量 (増体量：グラム単位) を選抜によって改良するものとしよう。この系統における 5 から 9 週齢の増体量の遺伝率は $h^2 = 0.4$、表現型分散は $V(P) = 12,100$ であるとする。親世代で雌雄それぞれ 100 羽について増体量を測定し、雄では上位 10 羽、雌では上位 40 羽を選抜するものとしよう。したがって、雄の選抜率は $p_{\mathrm{m}} = 10/100 = 0.1$、雌の選抜率は $p_{\mathrm{f}} = 40/100 = 0.4$ である。この選抜によって子世代に期待される遺伝的改良量を求めなさい。

【解答例】

相加的遺伝分散 $V(A)$ は

$$V(A) = V(P)h^2 = 121000 \times 0.4 = 4840$$

である。また、雄と雌の選抜強度をそれぞれ i_{m} と i_{f} とすれば表 14.9 より、$i_{\mathrm{m}} = 1.755$ および $i_{\mathrm{f}} = 0.966$ が得られる。したがって、集団全体での選抜強度 (i) は

$$i = \frac{i_{\mathrm{m}} + i_{\mathrm{f}}}{2} = 1.361$$

となる。これらより、遺伝的改良量 (ΔG) が

$$\Delta G = ih\sqrt{V(A)} = 1.361 \times \sqrt{0.4} \times \sqrt{4840} = 59.88$$

として得られる。

【R のコード・出力例】

```
> parent.num <- 100 #親世代の頭数
> select.num <- c(10,40) #親の選抜頭数
> select.ratio <- select.num/parent.num #選抜率
> i.m <- dnorm(qnorm(select.ratio[1],lower.tail=F))/select.ratio[1]
> i.f <- dnorm(qnorm(select.ratio[2],lower.tail=F))/select.ratio[2]
> cat("雄の選抜強度=",i.m,"\t 雌の選抜強度=",i.f,"\n")
雄の選抜強度= 1.754983  雌の選抜強度= 0.9658563
>
>
> var.p <- 12100 #表現型分散
> h2 <- 0.4  #遺伝率
> i.parent <- (i.m+i.f)/2 #込にした親の選抜強度
> var.A <- h2*var.p
> delta.G <- i.parent*sqrt(h2*var.A)
>
>
> cat("遺伝的改良量=",delta.G,"\n")
遺伝的改良量= 59.85847
```

【練習問題 8】

子どもの表現型値の中間親への回帰係数が、遺伝率の推定値を与えることを示しなさい。

【解答例】

父親を添字 S、母親を添字 D で表わし、表現型値 P を相加的遺伝子型値 A と非相加的遺伝子型値を含む環境偏差 e に分解すれば、

$$\begin{aligned} P_{\mathrm{S}} &= A_{\mathrm{S}} + e_{\mathrm{S}} \\ P_{\mathrm{D}} &= A_{\mathrm{D}} + e_{\mathrm{D}} \end{aligned}$$

であるから、中間親 $\bar{P}$ は

$$\bar{P} = \frac{P_{\mathrm{S}} + P_{\mathrm{D}}}{2} = \frac{A_{\mathrm{S}} + A_{\mathrm{D}}}{2} + \frac{e_{\mathrm{S}} + e_{\mathrm{D}}}{2}$$

である。親の交配はランダムに行われたとすると、$COV(P_{\mathrm{S}}, P_{\mathrm{D}}) = Cov(A_{\mathrm{S}}, A_{\mathrm{D}}) = 0$ である。したがって、

$$V(\bar{P}) = \frac{1}{4}V(P_{\mathrm{S}} + P_{\mathrm{D}}) = \frac{1}{2}V(P)$$

である。また、子どもを添字 O で表わせば、子どもの表現型値は

$$P_{\mathrm{O}} = A_{\mathrm{O}} + e_{\mathrm{O}} = \frac{A_{\mathrm{S}} + A_{\mathrm{D}}}{2} + e_{\mathrm{O}}$$

と書ける。親の相加的遺伝子型値と子どもが受けた環境効果の間、および親が受けた環境効果と子どもが受けた環境効果の間には相関はないものとすれば、

$$COV(\bar{P}, P_{\mathrm{O}}) = \frac{1}{4}V(A_{\mathrm{S}} + A_{\mathrm{D}}) = \frac{1}{2}V(A)$$

である。

これらより、子どもの表現型値の中間親への回帰係数 b は、

$$b = \frac{COV(\bar{P}, P_{\mathrm{O}})}{V(\bar{P})} = \frac{V(A)}{V(P)} = h^2$$

となる。

【練習問題 9】

あるブタの品種の体長について遺伝率を推定するために、30 頭の雄のそれぞれに 5 頭の雌を交配し、各雌から生まれた 3 頭の子ブタについて測定した。得られたデータに分散分析を施して、以下のような結果を得た (データは省略)。このデータから遺伝率を推定しなさい。

【解答例】

父親、母親および残差の分散成分を、それぞれ σ_{S}^2、σ_{D}^2 および σ_{w}^2 とすると平均平方の構成は、

$$\begin{aligned} 15\sigma_{\mathrm{S}}^2 + 3\sigma_{\mathrm{D}}^2 + \sigma_{\mathrm{w}}^2 &= 7.055 \\ 3\sigma_{\mathrm{D}}^2 + \sigma_{\mathrm{w}}^2 &= 4.280 \\ \sigma_{\mathrm{w}}^2 &= 2.870 \end{aligned}$$

である。これらより、

$$\begin{aligned} \sigma_{\mathrm{S}}^2 &= \frac{7.055 - 4.280}{15} = 0.185 \\ \sigma_{\mathrm{D}}^2 &= \frac{4.280 - 2.870}{3} = 0.470 \\ \sigma_{\mathrm{w}}^2 &= 2.870 \end{aligned}$$

を得る。σ_{D}^2 のほうが σ_{S}^2 よりも大きいので、優性分散あるいは共通環境分散は無視できないものと考えられる。したがって、遺伝率は半きょうだい間の共分散を用いて

$$h^2 = \frac{4\sigma_{\mathrm{S}}^2}{\sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{D}}^2 + \sigma_{\mathrm{w}}^2} = \frac{4 \times 0.185}{0.185 + 0.470 + 2.870} = 0.210$$

として推定できる。

【R のコード・出力例】

```
> varmat <- matrix(c(15,3,1,0,3,1,0,0,1),ncol=3,byrow=T)
> varvec <- matrix(c(7.055,4.280,2.870),ncol=1,byrow=T)
> varsol <- solve(varmat) %*% varvec
> varsol #父親、母親、残差の分散成分
      [,1]
[1,] 0.185
[2,] 0.470
[3,] 2.870
> h2 <- 4*varsol[1]/sum(varsol)
> h2
[1] 0.2099291
>
```

【補足事項】

通常、ブタでは産まれた子どもは母親に哺育させる。このために、母親間の哺育能力の違いが共通環境効果を生じさせる原因となる。この例では、σ_{D}^2 のほうが σ_{S}^2 よりも大きい原因としては、共通環境分散が最も重要であると考えられる。

【練習問題 10】

ある自殖性作物の二つの品種間の交雑によって得た F_1 世代の草丈の分散は 5.0 であった。また、F_1 世代から自殖で得た F_2 および F_3 世代の分散は、それぞれ 30.0 および 35.0 であった。エピスタシス分散を無視して、広義および狭義の遺伝率を求めなさい。

【解答例】

F_1、F_2 および F_3 世代の分散は、相加的遺伝分散 ($V(\mathrm{A})$)、優性分散 ($V(\mathrm{D})$) および環境分散 ($V(\mathrm{E})$) を用いて次のように表わされる。

$$
\begin{aligned}
V(\mathrm{F}_1) &= V(\mathrm{E}) = 5.0 \\
V(\mathrm{F}_2) &= V(\mathrm{A}) + V(\mathrm{D}) + V(\mathrm{E}) = 30.0 \\
V(\mathrm{F}_3) &= \frac{3}{2}V(\mathrm{A}) + \frac{3}{4}V(\mathrm{D}) + V(\mathrm{E}) = 35.0
\end{aligned}
$$

これらより、$V(\mathrm{A}) = 15.0$ および $V(\mathrm{D}) = 10.0$ を得る。したがって、広義の遺伝率 (h^2_{B}) は

$$
h^2_{\mathrm{B}} = \frac{V(\mathrm{A}) + V(\mathrm{D})}{V(\mathrm{A}) + V(\mathrm{D}) + V(\mathrm{E})} = \frac{25.0}{30.0} = 0.83
$$

となる。また、狭義の遺伝率 (h^2) は

$$
h^2 = \frac{V(\mathrm{A})}{V(\mathrm{A}) + V(\mathrm{D}) + V(\mathrm{E})} = \frac{15.0}{30.0} = 0.50
$$

として得られる。

【R のコード・出力例】

```
> varmat <- matrix(c(0,0,1,1,1,1,3/2,3/4,1),ncol=3,byrow=T)
> varvec <- c(5,30,35)
> varsol <- solve(varmat) %*% varvec
> varsol #相加的遺伝分散、優性分散、環境分散
     [,1]
[1,]   15
[2,]   10
[3,]    5
> (h2b <- (varsol[1]+varsol[2])/sum(varsol)) #広義の遺伝率
[1] 0.8333333
> (h2n <- varsol[1]/sum(varsol)) #狭義の遺伝率
[1] 0.5
>
```

【練習問題 11】

イネ科の一年生植物ソルガム (モロコシ) の 9 品種について、品種内で 2 個体の収量を調べて品種比較試験を行い、次のような分散分析の結果を得た (データは省略)。この結果から、広義の遺伝率を推定しなさい。

【解答例】

品種および誤差の分散成分を、それぞれ σ_{g}^2 および σ_{e}^2 とすると、平均平方の構成は

$$\begin{aligned} 9\sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{e}}^2 &= 104.4 \\ \sigma_{\mathrm{e}}^2 &= 24.2 \end{aligned}$$

となる。これより $\sigma_{\mathrm{g}}^2 = 8.91$ を得る。遺伝分散を $V(\mathrm{G})$、環境分散を $V(\mathrm{E})$ とすると

$$\begin{aligned} V(\mathrm{G}) &= \sigma_{\mathrm{g}}^2 \\ V(\mathrm{E}) &= \sigma_{\mathrm{e}}^2 \end{aligned}$$

であるから、広義の遺伝率 (h_{B}^2) は

$$h_{\mathrm{B}}^2 = \frac{V(\mathrm{G})}{V(\mathrm{G}) + V(\mathrm{E})} = \frac{\sigma_{\mathrm{g}}^2}{\sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{e}}^2} = \frac{8.91}{24.2 + 8.91} = 0.27$$

として推定できる。

【練習問題 12】

卵用鶏の系統において、産卵率、飼料要求率、卵重について以下のデータが得られている (データは省略)。

(1) このデータから、表現型分散共分散行列 $\mathbf{P}$、遺伝分散共分散行列 $\mathbf{G}$ を求めなさい。

(2) 三つの形質を相対的経済価値に従って改良するための選抜指数における重み付け値を求めなさい。

【解答例】

産卵率、飼料要求率および卵重を、それぞれ添字 X、Y、Z で表す。

(1) 表現型標準偏差から三つの形質の表現型分散が

$$\begin{aligned} V(\mathrm{P_X}) &= 10^2 = 100.0 \\ V(\mathrm{P_Y}) &= 3^2 = 9.0 \\ V(\mathrm{P_Z}) &= 4^2 = 16.0 \end{aligned}$$

として得られる。これらと各形質の遺伝率から相加的遺伝分散が

$$\begin{aligned} V(\mathrm{A_X}) &= V(\mathrm{P_X})h^2_{\mathrm{X}} = 30.0 \\ V(\mathrm{A_Y}) &= V(\mathrm{P_Y})h^2_{\mathrm{Y}} = 1.8 \\ V(\mathrm{A_Z}) &= V(\mathrm{P_Z})h^2_{\mathrm{Z}} = 8.0 \end{aligned}$$

として得られる。

たとえば、産卵率と飼料要求率の間の表現型共分散は、二つの形質の表現型分散と表現型相関 ($r_{\mathrm{P}} = -0.7$) から

$$COV(\mathrm{P_X}, \mathrm{P_Y}) = r_{\mathrm{P}}\sqrt{V(\mathrm{P_X})V(\mathrm{P_Y})} = -0.7 \times \sqrt{100 \times 9} = -21.0$$

として求められる。他の形質の組み合わせについても同様の計算を行うと、表現型分散共分散行列 $\mathbf{P}$ が

$$\mathbf{P} = \begin{bmatrix} 100.0 & -21.0 & -4.0 \\ -21.0 & 9.0 & -3.6 \\ -4.0 & -3.6 & 16.0 \end{bmatrix}$$

として得られる。

また、産卵率と飼料要求率の間の相加的遺伝共分散は、二つの形質の相加的遺伝分散と遺伝相関 ($r_{\mathrm{A}} = -0.6$) から

$$COV(\mathrm{A_X}, \mathrm{A_Y}) = r_{\mathrm{A}}\sqrt{V(\mathrm{A_X})V(\mathrm{A_Y})} = -0.6 \times \sqrt{30.0 \times 1.8} = -4.409$$

として得られる。他の形質の組み合わせについても同様の計算を行うと、遺伝分散共分散行列 $\mathbf{G}$ が

$$\mathbf{G} = \begin{bmatrix} 30.0 & -4.409 & -3.098 \\ -4.409 & 1.8 & -1.518 \\ -3.098 & -1.518 & 8.0 \end{bmatrix}$$

として求められる。

(2) 相対的経済価値のベクトル $\mathbf{a}$ は

$$\mathbf{a} = \begin{bmatrix} 3.0 \\ -7.0 \\ 5.0 \end{bmatrix}$$

である。したがって、選抜指数中の重み付け値のベクトルが

$$\begin{aligned} \mathbf{b} &= \mathbf{P^{-1}Ga} \\ &= \begin{bmatrix} 100.0 & -21.0 & -4.0 \\ -21.0 & 9.0 & -3.6 \\ -4.0 & -3.6 & 16.0 \end{bmatrix}^{-1} \begin{bmatrix} 30.0 & -4.409 & -3.098 \\ -4.409 & 1.8 & -1.518 \\ -3.098 & -1.518 & 8.0 \end{bmatrix} \begin{bmatrix} 3.0 \\ -7.0 \\ 5.0 \end{bmatrix} \\ &= \begin{bmatrix} 1.27 \\ 0.44 \\ 3.00 \end{bmatrix} \end{aligned}$$

として得られる。

【R のコード・出力例】

```
> (P <- matrix(c(100,-21,-4,
+                 -21,9,-3.6,
+                 -4,-3.6,16), nrow=3))  #行列 P へ値を代入
     [,1]  [,2] [,3]
[1,]  100 -21.0 -4.0
[2,]  -21   9.0 -3.6
[3,]   -4  -3.6 16.0
> (G <- matrix(c(30,-4.409,-3.098,
+                -4.409,1.8,-1.518,
+                -3.098,-1.518,8), nrow=3)) #行列 G へ値を代入
       [,1]   [,2]   [,3]
[1,] 30.000 -4.409 -3.098
[2,] -4.409  1.800 -1.518
[3,] -3.098 -1.518  8.000
> a <- c(3,-7,5)                # ベクトル a へ値を代入
> b <- solve(P) %*% G %*% a  # 重み付け値の計算
> b
          [,1]
[1,] 1.2664954
[2,] 0.4418741
[3,] 2.9992955
>
>
```