入門 統計学（第２版）

― 検定から多変量解析・実験計画法・ベイズ統計学まで ―

章末問題解答

第1章

問1

a. 比率尺度，b. 名義尺度，c. 比率尺度，d. 順序尺度，e. 間隔尺度

問2

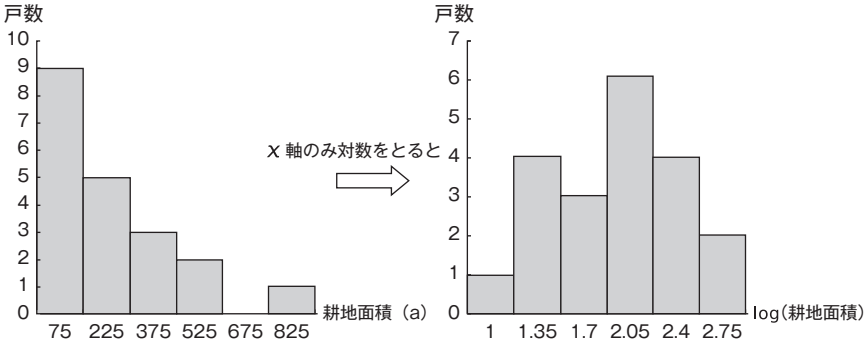
a.　まずは耕地面積の小さい順，または大きい順に並べ替えます。その後，階級の区間を決定します。耕地面積は最小値が15a，最大値が783a（7.83ha）なので，その差768aを5階級（データ数20の平方根に1を足して四捨五入）で割ると，1階級当たり153.6aとなるため，きりのよいところで階級区間を150aとします。そして，各階級に属する度数（農家戸数）をカウントして，相対度数と累積相対度数を計算します。

度数分布表

耕地面積 (a)	階級値 (a)	度数 (戸)	相対度数 (%)	累積相対度数 (%)
0 以上〜150 未満	75	9	45	45
150〜300	225	5	25	70
300〜450	375	3	15	85
450〜600	525	2	10	95
600〜750	675	0	0	95
750〜900	825	1	5	100
計		20	100	

その後，度数分布表から階級区間（耕地面積）と度数からヒストグラムを作成すると，次の図（左）のようになります。これを見ると左右非対称の分布になっており，耕地面積の小さな階級に農家が集中していることがわかります。もし，どうしても左右対称の山形に近づけたいならば，階級区間を等間隔にしないで，（耕地面積の）自然対数を取るという方法もあります（ExcelではLNという関数があります）。すると，大きい値は小さくなり，もともと小さい値はあまり変わらないため，図（右）のようにやや山形に近づきます。なお，柱と柱の間には隙間を作らないようにします。隙間がある図はヒストグラム（柱状図）ではなく**棒グラフ**と呼ばれ，質的データの度数分布を表すときに使います。

b. とc.　表計算ソフトで基本統計量を計算するには関数機能を使うと便利です。Excelでは，算術平均がAVERAGE，分散がVAR.P，標準偏差がSTDEV.P，相関係数がCORRELです。これらを使えば，表のような結果が得られます。



ヒストグラム

なお，変動係数のExcel関数はないので，自分で式（標準偏差÷平均）に従って計算してください。

	販売金額	耕地面積
平均	582.0	222.9
分散	566200.4	43608.6
標準偏差	752.5	208.8
変動係数	1.29	0.94
相関係数	0.84	

問３

幾何平均をExcelで計算するには，指数を示す^記号を使って，例えば＝(2 ＊ 4 ＊ 5 ＊ 7)^(1/4)と入力するか，GEOMEANという関数を利用します。

すると，算術平均はそれぞれ4.50と20.25，幾何平均は4.09と7.27になります。70という突出して大きな値を含むデータセットでも，幾何平均ならばそれほど影響を受けない（大きくならない）ことが確認できると思います。

第2章

問1

各値と平均との差（偏差）を標準偏差で割れば，次の表のようになります。平均はゼロ，標準偏差は（分散も）1にそろっていることが確認できます。

農家番号	農産物の販売金額（万円）	総経営耕地面積（a）
1	-0.24	-0.78
2	-0.75	-0.92
3	-0.14	0.68
4	0.55	-0.16
5	0.02	-0.42
6	-0.57	-1.00
7	-0.62	-0.89
8	-0.71	-0.59
9	-0.77	-0.25
10	-0.60	-0.73
11	3.21	2.68
12	-0.11	1.61
13	-0.51	-0.83
14	-0.70	-0.90
15	2.15	1.78
16	-0.77	-0.11
17	0.42	0.37
18	0.56	0.64
19	-0.18	-0.32
20	-0.24	0.13
平均	0.00	0.00
標準偏差	1.00	1.00

問2

Excelの関数を使って農産物の販売金額の歪度（= SKEW.P）を求めると2.03，尖度（= KURT）を求めると4.95，同様に総経営耕地面積の歪度を求めると1.24，尖度も1.24となります。どちらの変数（販売金額および耕地面積）の歪度，尖度ともゼロより大きいことから，正規分布に比べて右裾が長くなり，やや尖り気味であることがわかります。このように，標本サイズが小さい（データ数が少ない）場合には，正規分布の形状から離れやすくなります。なお，46ページの式で尖度を求めると販売金額が3.51，耕地面積が0.67となり，KURT関数の結果より少し小さくなります（歪度は同じ）。

問３

標準化した得点　：$(80-60)\div 10=2$

偏差値　　：$10\times(80-60)\div 10+50=70$

問４

標準正規分布表の$z=2.00$の確率を見てみると，0.0228となっています。この値は標準化した正規分布の上側の確率を示しているため，そのまま0.0228×100で，上位2.28％ぐらいの位置にいることがわかります。

問５

λ（試行回数n×生起確率p）から求めます。試行回数nは松戸市民の人口なので484600という大きな値になり，生起確率pは国民1人が1日あたりに食中毒になる確率なので，$20204\div(127370000\times 365)=4.346\times 10^{-7}$という大変小さな値になります。よって$\lambda=(484600)\times(4.346\times 10^{-7})=0.21061$となります。また，$x$は0回です。

これらをポアソン分布の関数式に代入します（ゼロの階乗は1です）。

$$e^{-\lambda}=2.718^{-0.21061}=0.810\cdots$$

したがって，松戸市において食中毒になる人が1人も発生しない日の起こる確率は約81％ということになります。意外と食中毒って起こらないものだと勘違いしないでください。これはあくまで厚生労働省にまで数値が上がってくるような食中毒（つまり飲食店などでの事件）であることをお忘れなく。

第3章

問1

母集団　　：新入生1万人の学力（試験結果）

標本　　：無作為抽出された新入生500人の学力（試験結果）

ユニバース　：1万人の新入生（人間そのものの集団）

標本数　　：1（標本サイズと間違わないように気をつけること）

問2

母分散 > 標本分散

【理由】56ページで示されている両方の計算式を見てみると，母平均μを使って計算する母分散よりも，標本平均$\overline{x}$を使って計算する標本分散の方が，式の分子は小さくなることがわかります（$\overline{x}$は標本自身から計算されているのに対して，μは母集団から計算されており，$\overline{x}$とはやや異なる値になるため）。

問3

個々のデータからなる集団のバラツキ具合を示す統計量が標準偏差なのに対して，標準誤差は標準偏差を標本サイズの平方根で割った値で，標本平均のバラツキ具合を示しています。よって，標準偏差は標本サイズと直接関係がないのに対して，標準誤差は標本サイズが大きくなればなるほど小さくなります（具体的には標本サイズが4倍になれば標準誤差は約半分になります）。そのため，標本から母集団の特性を推定したときの精度の低さ（誤差の大きさ）を示す指標として用いられるのです。

問4

相関係数を学んだ第1章では，まだ推測統計学の概念はありませんでした。しかし，相関係数においても，標本の相関係数から母集団における相関係数についていろいろと統計的に検討することができます（信頼区間の推定や無相関の検定など）。そのようなときに利用する統計量（具体的にはt値です）の**自由度は$n-2$となります**。平均を1つ使う度に自由度が1つ減るので，平均を2つ（$\overline{x}$と$\overline{y}$）使う相関係数の場合，自由度は2つ減るということですね。ただし，本書では，母相関係数の推定は扱っておりませんので（あまり推定する意味がないからです），あしからず……。

第4章

問1

この問題では，母分散は未知ですが標本サイズが大きい（目安として30以上）ため，標準正規分布（z分布）を使った区間推定を行っても差し支えないでしょう（t分布を使った場合と大差ありません）。さて，信頼係数95％の信頼限界の計算に使うz値は，本章の例題と同じように付録Ⅰに掲載した標準正規分布表から1.96であることが読み取れます。よって，$2.0 \pm 1.96 \times 1.0 \div \sqrt{100}$という計算によって，この島のスギの胸高周囲の母平均に対する信頼係数95％の信頼区間は（1.804m，2.196m）と推定できます。ちなみに，念のためt分布で区間推定したい場合には，t分布表には自由度が40までしかないので，Excel関数の=T.INV（0.025, 99）の絶対値を使って限界値を計算するか，オリジナルデータがあるならば分析ツールを使うことになります。その場合の信頼区間は（1.801m，2.199m）となり，ほんの少し広がります。

問2

この問題では，母分散は未知である上に標本サイズも20とそれほど大きくないため，t分布を使った区間推定を実施した方がよいでしょう。t分布を使う場合，信頼限界の計算に使用するt値は，同じ信頼係数でも自由度（$\nu = n - 1$）によって異なります。自由度は$20 - 1$で19となるので，信頼係数95％の信頼限界に使うt値は，付録Ⅱのt分布表の上側確率$p = 0.025$の列と，自由度$\nu = 19$の行がクロスするところの2.093となり，正規分布のときの1.96よりも大きくなるため，区間の幅が広くなることが予想されます。

さて，第1章の章末問題の農家データによると，20戸の農家の販売金額の標本平均は582万円，標本標準偏差は752.5万円でした。よって，$582 \pm 2.093 \times 752.5 \div \sqrt{19}$から，この地域の農家の農産物販売金額の母平均に対する信頼係数95％の信頼区間は（220.7万円，943.3万円）と推定されます。

問3

標本サイズが小さいので，AgrestiとCoullの式を使って母比率の信頼区間を求めてみましょう。なお，この場合の$\widehat{p}'$は，$(16 + 2) \div (20 + 4)$なので0.75となります。また，z値は信頼係数が95％なので1.96です。本章中の式に，これらの値を代入すると，

$$0.75 \pm 1.96\sqrt{\frac{0.75(1-0.75)}{20+4}}$$

で，0.75 ± 0.173の範囲となり，このペットショップの中で血液型がA型である猫の母比率に対する95％信頼係数の信頼区間は（57.7％，92.3％）であると推定されます。

実は，種によって若干の差はありますが，猫の血液型は，A型が80〜100％を占めていることがわかっています（次にB型が多く0〜20％，AB型は滅多にいません）。

問4

アンケート調査では母比率を推定することが多いため，許容できる推定の誤差（標本誤差）が1％となるような分析をするための標本サイズnは，次の母比率の推定式の誤差の部分から逆算して求めます（母比率pには誤差が最大となる0.5を入れておきます）。

$$1.96\sqrt{\frac{0.5 \times 0.5}{n}} = 0.01$$

すると，$n = 9\,604$という標本サイズが必要となることがわかります。さらに返信率が20％なのでその5倍に発送しなければなりません（大学で実施する市民向けの郵送調査では返信率は実際のところ20％程度です）。よって，この事例では最終的に約48 000件に発送する必要があることになります。

日本の代表的なテレビ視聴率調査会社の1つであるビデオリサーチ社の標本サイズが関東地方で900と案外小さいのも，このように精度を少し上げようとするだけでも必要となる標本サイズが急激に大きくなってしまい，コストの面で妥協せざるを得ないからでしょう。

第5章

問１

標準化変量z_iの２乗和（正規分布に従っているデータ全ての標準化変量を２乗して足し合わせたもの）です。

問２

母平均が既知の場合には標本サイズ（データの数）がそのままχ^2の自由度になりますが，未知の場合には他の統計量と同様に自由度は減少します。具体的には，標本サイズnから１を引いた値です（$\nu = n - 1$）。ただし，第11章で学ぶピアソンのχ^2検定で用いるχ^2は近似的なものなので，自由度の計算方法も異なります。

問３

農業所得とは農産物を売った後に各種の経費を差し引いた金額，つまり農家の手元に残る実際のお金のことです。よって農業所得率とは，その農業所得の農業粗収益（農業経営から得られた総収入額）に対する割合です。

この農業所得率は農家の経営分析指標としてよく用いられており，一般的に50％以上が望ましいとされていますが，近年では30％を下回る作目も多く見られます（ここでは比率そのものは問題としていません）。そして，こうした収益率のリスク指標として年ごとのバラツキの大きさが用いられるのです。つまり，特定の農家の農業所得率の年ごとのバラツキの幅を推定することは，その農家に対する経営診断の最初のステップの１つといえるでしょう。

さて，過去10年間（$n = 10$）の農業所得率の不偏標準偏差は５％（ということは不偏分散＝25）であったようなので，この農家の農業所得率に対する母分散σ^2の信頼区間を信頼係数90％で推測すると，該当する自由度９の２つのχ^2値（$p = 0.050$のχ^2値が16.919，$p = 0.950$のχ^2値が3.325）を使って，範囲を計算します。

$$\frac{(10-1)\times 25}{16.919} < \sigma^2 < \frac{(10-1)\times 25}{3.325}$$

よって，母分散σ^2の信頼区間は$(13.30, 67.67)$となるので，母標準偏差に対する信頼係数90％の信頼区間は，その平方根をとった$(3.65\,\%, 8.23\,\%)$と推定されます。

問4

t値の2乗と，分子の第1自由度が1のF値は等しい。

この性質があるために，2群の平均の差のt検定（第7章）と，2群に対して実施した分散分析（第8章）とが同じ結論になるのです。

第6章

問１

社会科学の分野では，標本が母集団から無作為に抽出されていること，つまり偏りを持って抽出されていないことが重視されます（サンプリングバイアスの問題）。というのも，調査対象が人間や企業などの場合が多いため，つい調査しやすさを優先してしまうからです。例えば，知り合いに頼んでしまったり，声をかけやすい人を中心に聞いてしまったりすることが考えられます。そのように標本が偏っていては，そこから導き出される答え（母集団に対する普遍的な結論）も偏ってしまいます。

そこで，この問題のように，母集団の分布と標本の分布が乖離しているのか（偏った標本なのか），それとも重なっているのか（偏っていない標本なのか）を統計的に検討することは大変重要な作業となります（第11章で学ぶ適合度の検定など，他にもいくつか方法はあります）。

この問題は全数調査によって母標準偏差（$\sigma = 11$）が与えられているので，z分布による母平均の検定が可能です。

手順①：仮説の設定

まず，「全国の農家の平均年齢（母平均$\mu_0 = 63$）とA君が調査した農家の平均年齢（標本平均$\overline{x} = 65$）の母平均μ）とでは差がない」という帰無仮説を立てます。この帰無仮説は，今回の調査で生じた2歳という差は，よく起こる（偶然のうち）という内容です。帰無仮説は一般的に，主張したい内容とは逆のつまらない内容となるのですが，棄却されない方がよいという珍しい事例です。一方，対立仮説は帰無仮説の逆の内容となるので，「全国の平均年齢と，調査した農家の平均年齢とでは差がある」，つまり標本が母集団を代表しているとはいえないことになります。A君が調査した農家の平均年齢（65歳）の真の値（母平均）をμ_0，全国の農家の平均年齢（63歳）をμとすると，仮説は次のように整理できます。

$$\begin{cases} \text{帰無仮説} H_0 : \mu(\overline{x} = 65\text{歳}) = \mu_0(63\text{歳}) \\ \text{対立仮説} H_1 : \mu(\overline{x} = 65\text{歳}) \neq \mu_0(63\text{歳}) \end{cases}$$

手順②：検定統計量の計算

標本平均$\overline{x}$の標準化変量zは，$z_{\overline{x}} = (\overline{x} - \mu_0)/(\sigma/\sqrt{n})$で求められます。

よって，（帰無仮説が正しい下での）検定統計量$z_{\bar{x}}$は，$(65-63)/(11/\sqrt{100}) = 1.82$となります。

手順③：確率の計算

帰無仮説が正しい下のz分布において，$z = 1.82$よりも極端な値が出る確率（p値）を計算します。t分布表では確率変数値と確率との対応が粗いため難しいですが，z分布表ならば細かい（z値の小数点が第2位までの確率が書かれています）ので可能です。そこで実際に，付録Ⅰの標準正規（z）分布表から，$z = 1.82$に対応する上（右）側確率を探してみると，0.0344（3.4％）であることがわかります。つまり，A君の調査した農家の平均年齢と日本の販売農家の平均年齢とが同じだという帰無仮説が正しい場合，2歳以上の差が出る確率は両側（正の差と負の差の両方を考える）で6.8％ということになります。

手順④：仮説の判定

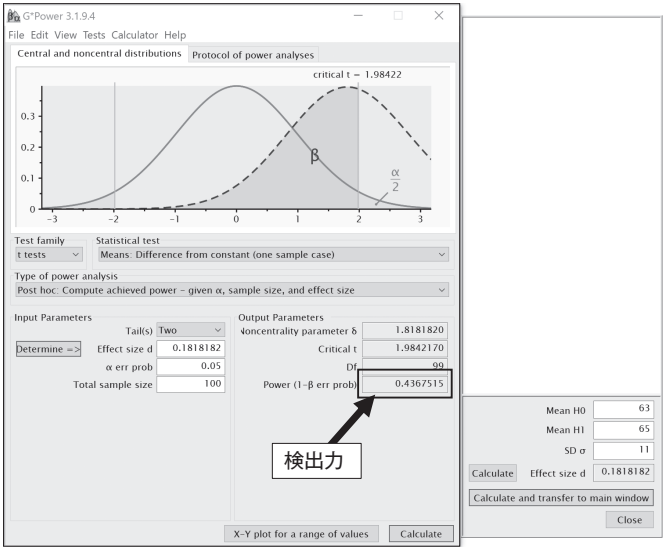
出る確率が6.8％の結果を，偶然と呼ぶべき（滅多に起きないことが起きた）か，当然と呼ぶべき（よく起きることが起きただけ）かを判断します。その判断基準（有意水準α）を，一般的には5％とします。

今回はp値（0.068）が有意水準α（0.05）よりも大きいため，よく起きることが起きただけと考え，帰無仮説を受容します。つまり，A君の調査した農家は母集団から無作為に抽出された標本と考えてもよい（「抽出による偏りは発生していなかった」）ことになります。ただし，帰無仮説が正しかったと断言はできません。あくまで今回のデータでは有意差が検出されなかっただけです。

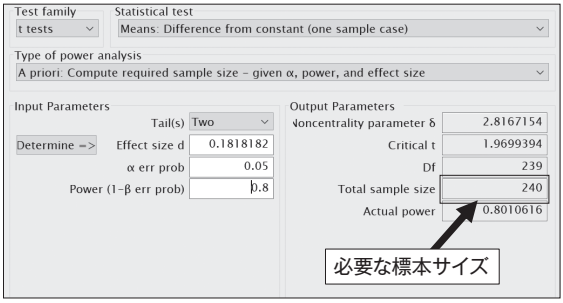
なお，本文で解説してきた伝統的な検定の判定方法，つまり限界値との比較もしてみましょう。z分布に自由度は関係ありませんから，有意水準5％（両側）の場合，限界値は常に± 1.96となります。よって，帰無仮説の分布において，検定統計量の1.82は，限界値の1.96よりも小さい（つまり母平均μ_0に近い内側）の「帰無仮説の受容域」にあるため，帰無仮説は棄却できないという判定になります。

問2

G*power（事後分析：Post hoc）を使って，問1の検出力（power）を推定すると，図のように0.44（44％）となります。つまり，第二種の過誤を犯す確率βが0.56と高い検定（あまり良くない検定）であったことがわかります。その理由は，効果量が0.18と小さかったからです。



次に，G*power（事前分析：A priori）で，この効果量のまま，有意水準５％の下で，検出力0.8を達成するための標本サイズを計算すると，240の標本を確保する必要があったことがわかります。ただし，この計算は，あくまで帰無仮説をちょうど良い標本サイズで棄却するためのものですから，本来，受容されることが期待される本事例では，こうした計算をすることはないでしょう（あくまで練習ということで出題しました）。



第7章

問1

このデータを整理すると，次のようになります。

あるガのマユの形成から羽化までの気温別日数

気　温	15℃	20℃
平均日数	34.0	25.0
不偏分散	4.67	4.67
標本サイズ	7	7

サナギでいる日数は，15℃で育てた場合と20℃で育てた場合で，平均で9日ほど後者の方が短くなっています。しかし，この9日という差は，本来は差がなかったにも関わらず，今回実験に使った7匹ずつの標本において偶然現れただけかもしれませんので，やはり統計的に検討した方が良いでしょう。

なお，母分散は未知で，ガは1度羽化したらサナギには戻らないので，検定は「“対応のない”2群の平均の差のt検定」（スチューデントのt検定）が良いでしょう。また，気温の高い方が必ず成長日数が短くなるという明確な根拠はないため，両側検定が望ましいでしょう。

手順①：仮説の設定

$$\begin{cases} \text{帰無仮説}H_0\text{：15℃と20℃では，サナギでいる日数の母平均に差はない} \\ \text{対立仮説}H_1\text{：15℃と20℃では，サナギでいる日数の母平均に差はある} \end{cases}$$

手順②：検定統計量の計算

スチューデントのt検定の統計量の式（バランスな場合）に，各値を代入すると7.79となります。

$$t_{\overline{x}_1-\overline{x}_2}=\frac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2+\hat{\sigma}_2^2}{n}}}=\frac{34-25}{\sqrt{\frac{4.67+4.67}{7}}}=7.79$$

（手順③の確率の計算はスキップして）手順④：仮説の判定

付録Ⅱのt分布表から限界値を読み取ります。自由度νは$2(n-1)$なので「12の行」と，両側5％の有意水準の半分の確率（p）となる「0.025の列」がクロス

する2.179が限界値となります。よって，検定統計量（7.79）は限界値（2.179）よりも大きいので，帰無仮説は棄却され，対立仮説が採択されます。つまり，15℃と20℃という気温の違いが，このガのサナギでいる日数に影響を与えたと，統計的に判断できることになります。

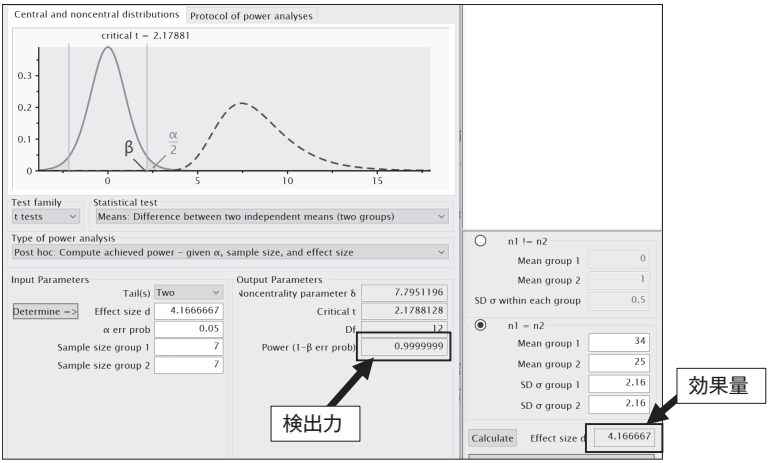
問2

1. 分析ツールの［t-検定: 等分散を仮定した2標本による検定］でp値（両側）を求めると，4.90447×10^{-6}と0に近い極めて小さい値であることがわかります。

	A	B	C
1	t-検定：等分散を仮定した２標本による検定		
2			
3		15℃	20℃
4	平均	34	25
5	分散	4.666666667	4.66666667
6	観測数	7	7
7	プールされた分散	4.666666667	
8	仮説平均との差異	0	
9	自由度	12	
10	t	7.794228634	
11	P(T<=t) 片側	2.45223E-06	
12	t 境界値 片側	1.782287556	
13	P(T<=t) 両側	4.90447E-06	
14	t 境界値 両側	2.17881283	

p値

2. G*powerによる事後分析の結果，効果量（Effect size d）が4.17と大きかったため，検出力（Power）も0.99と高かったことが確認できます。



3. 今回は，たまたま同じ不偏分散の大きさだったので，等分散の検定は不要ですが（検定統計量Fが1になることは明白です），念のため実施した結果を示しておきます。もちろん帰無仮説（等分散）は受容されますので，問1でスチューデントのt検定を適用したことに問題はありませんでした。

	A	B	C
1	F-検定: 2 標本を使った分散の検定		
2			
3		15℃	20℃
4	平均	34	25
5	分散	4.666667	4.666667
6	観測数	7	7
7	自由度	6	6
8	観測された分散比	1	
9	P(F<=f) 片側	0.5	
10	F 境界値 片側	0.233434	

第8章

問１

ひとくちに農家といっても，水田作経営，畑作経営，花き作経営，酪農経営など，様々です。この問題は，そうした営農類型の違いによって，農業所得も異なるのかどうかを検証するものです。問のデータから営農類型別に平均農業所得（1経営体あたり）を計算すると，水田作は60万円，果樹作は250万円，施設野菜作は500万円となっており，標本平均からは差はありそうですが，母平均においても差があるといってもよいでしょうか？

営農類型は，もっとも農産物の販売収入が高い部門で分類されますから，3群間の経営体に対応関係はありません。よって，「対応のない一元配置分散分析」を用いましょう。帰無仮説は「どの営農類型の農業所得も変わらない」となり，対立仮説は「営農類型によって農業所得は異なる」となります。

次は，Excel分析ツールによる分散分析表の出力結果です（ただし，この程度のデータならば手計算も可能ですね）。その結果，営農類型の違い（グループ間）から生じる分散は“97,400”，誤差（グループ内）から生じる分散は“9,400”となり，検定統計量Fはそれらを割り算した“10.36”となりました。5％のF分布表から限界値（第1自由度が2，第2自由度が3）を読み取ると“9.55”ですので，帰無仮説は棄却され，「営農類型によって農業所得は異なる」という対立仮説が採択されました。

分散分析：一元配置

変動要因	変動	自由度	分散	観測された分散比	P-値	F境界値
グループ間	194800	2	97400	10.36	0.045	9.55
グループ内	28200	3	9400			
	223000	5				

なお，このデータは著者が創作したものですが，各営農類型の平均は2018年の実際の値に近づけてあります。ちなみに，このデータ以外の営農類型も全て比較すると，一番農業所得が高いのは何だと思いますか？　実は，畜産が高くて，なかでも酪農とブロイラー養鶏経営がともに1,360万円で1位，続いて養豚経営が1,069万円となっています。ただし，畜産は専業農家であることが多いのに対して，水田作農家（平均56万円）は野菜なども作っていたり，本業は会社勤めだったりしますので，一概にどれが良いとはいえません。

問2

単に人種間の血圧の違いや塩負荷の影響を見るだけならば，それぞれの要因に対して「対応のない一元配置分散分析」を実施すれば良いでしょう。しかし，もしかしたら人種と塩負荷の間に交互作用があるかもしれませんので，できれば「(繰り返しのある）二元配置分散分析」を用いるべきでしょう。幸いなことに，水準の組合せごとに6つのデータが観測されていますので実施可能です。ただし，さすがにこのレベルのデータとなるとソフトウェアが必要でしょう。次はExcel分析ツールによる分散分析表の出力結果です。

分散分析：繰り返しのある二元配置

変動要因	変動	自由度	分散	観測された分散比	P-値	F境界値
標本	1067.11	1	1067.11	5.07	0.032	4.17
列	1350.39	2	675.19	3.21	0.055	3.32
交互作用	3399.39	2	1699.69	8.07	0.002	3.32
繰り返し誤差	6315.00	30	210.50			
合計	121.89	35				

これを見ると，5％有意水準では，塩負荷の違い（表では「標本」要因）による血圧の母平均に差はあるものの，人種（列要因）による差は（今回の標本では）あるとはいえないことがわかります。しかし，食塩の負荷の有無（表では「列」要因）と，人種×食塩負荷の交互作用については有意な差があることがわかりました。

つまり，塩分を多くとっている人たち，とくに黒人の人たちが高血圧になる傾向があることが，この実験・分析からはいえるのです。このような現象を，「黒人は血圧における食塩感受性が高い」といいます。アジア人や白人が太古の昔から食塩をたくさん摂取してきたのに対して，塩がなかなか手に入らないアフリカを起源とする黒人は，食塩の過剰摂取に対する耐性がまだ遺伝子には備わっていないのでしょう。また，アメリカの黒人にこの傾向が強いことから，塩分を体に貯めることができたおかげで灼熱の奴隷船で生き残れた者達の子孫であることが理由であるとの説もあります。実は，これと同様の実験を家森幸男先生（京都大学）らが実施しています。そのオリジナルデータが手に入らなかったので，著者が同じ分析結果になるようにデータを創作しました（ただし家森先生は塩分負荷（摂取）については，群間で「対応あり」のデータを収集しています)。

問３

- 相乗効果があるもの：なし
- 相殺効果があるもの：③と④
- 要因Ａの主効果のみあるもの：①

第9章

問1

この事例は，対応のない一元配置分散分析では，5％水準で有意（施肥効果あり）となりましたが，どの2群間（対）に差があるのかは不明でした。2種類の多重比較法で有意差のある対を見つけてみましょう。

まず，Bonferroni法は，有意水準（5％）を検定回数（3回）で割って厳しく調整します。t分布表では，切りの良い上側確率しか掲載されていないので，ソフトウェアを使いましょう。一般的には，Tukey法と同様，多重t検定を調整するのですが，Excel分析ツールでは実施できないので，ここでは普通のt検定を調整しましょう。分析ツールでは［t検定：等分散を仮定した2標本による検定］が普通のt検定（スチューデントのt検定）です。

有意水準のところ（分析ツールだとα(A)：の右側）に，0.05を3で割った0.0167を入力します。これを3対にそれぞれ実施します。すると以下のような結果になります（全群とも反復数は同じ2なので限界値も同じになります）。

- 対1-2：$|t$値$| = 3.54 <$ 限界値$= 7.64$ → 有意差なし
- 対1-3：$|t$値$| = 4.47 <$ 限界値$= 7.64$ → 有意差なし
- 対2-3：$|t$値$| = 2.24 <$ 限界値$= 7.64$ → 有意差なし

以上のように，普通のt検定をBonferroni法で調整した場合，いずれの対においても，今回のデータからは帰無仮説を棄却できませんでした（有意差のある対はありませんでした）。

次に，Tukey法を実施してみましょう（こちらはRコマンダーで可能です）。まずは3対の多重t検定の統計量を計算しましょう。対1-2の検定統計量tは，下記のように“2.5”となります。なお，$\widehat{\sigma}_e^2$は分散分析の検定統計量Fの分母（誤差分散）ですので，第8章で計算済みです（“4.0”になります）。

$$t_{\overline{x}_1-\overline{x}_2} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\widehat{\sigma}_e^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{2-7}{\sqrt{4\left(\frac{1}{2}+\frac{1}{2}\right)}} = -2.5$$

同様に計算すると，対1-3は“−5.0”，対2-3は“−2.5”となります。次に，この検定統計量の絶対値を，スチューデント化した範囲の分布（q分布）から取ってきた限界値$/\sqrt{2}$と比較します。上側確率5％のq分布表において，自由度 ν

はデータ総数（6）－群数（3）なので3の行と，群数$j=3$なので3の列がクロスした5.91を，$\sqrt{2}$で割った“4.18”が限界値となりますので，以下のような結果になります。

- 対1-2：$|t$ 値$|=2.5<$ 限界値$=4.18$ → 有意差なし
- 対1-3：$|t$ 値$|=5.0>$ 限界値$=4.18$ → 有意差あり
- 対2-3：$|t$ 値$|=2.5<$ 限界値$=4.18$ → 有意差なし

以上のように，Tukey法では，対1-3（肥料なしと肥料Bの群間）において有意差がでました。この結果からもBonferroni法より検出力が強いことがうかがえますが，多重t検定の統計量を使えば，（Bonferroni法でも）ぎりぎり5％水準で対1-3に有意差が出ます（p値$=0.046$）。

問2

「水準1の母平均」と「水準2と水準3の母平均」を比較するような“対比”はScheffe法でしか扱えません（事前に対比の内容と数を決めておけばBonferroni法でも可能ですが，あまり使われません）。

さて，この場合，対比係数は水準1が$c_1=1$，水準2が$c_2=-1/2$，水準3が$c_3=-1/2$となります。この対比係数を使って検定統計量Fを計算すると，次のように“9.37”となります。

$$F=\frac{(\sum c_j\overline{x}_j)^2/(j-1)}{\widehat{\sigma}_e^2\sum(c_j^2/n_j)}=\frac{\left(1\times 2-\frac{1}{2}\times 7-\frac{1}{2}\times 12\right)/(3-1)}{4\left(\frac{1^2}{2}+\frac{\left(-\frac{1}{2}\right)^2}{2}+\frac{\left(-\frac{1}{2}\right)^2}{2}\right)}=9.37$$

限界値は，有意水準5％のF分布表のなかの第1自由度$\nu_1=$ 群数$-1=3-1=2$の列と，第2自由度$\nu_2=$ データ総数$-$群数$=6-3=3$の行のクロスする“9.55”となります。

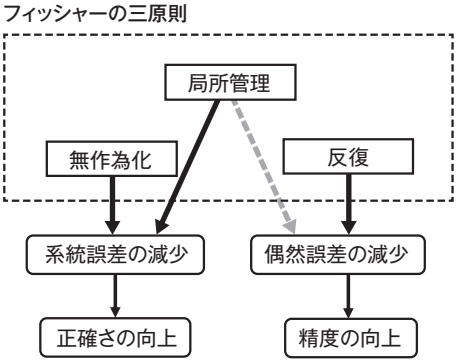
従いまして，限界値よりも検定統計量の方が（若干ではありますが）小さいので，有意水準5％では「水準1の母平均」と「水準2と水準3の母平均」に有意な差はあるとはいえないことになります。

なお，今回は1つの対比しか設定しませんでしたが，興味のある内容で無限に設定し（事後にデータを眺めながらでもOK），いくらでも検定を繰り返すことができるのがScheffe法の特徴です。

第10章

問1

反復の原則（③）は，大数の法則から偶然誤差を減らし，精度を向上させます。無作為化（②）は系統誤差を偶然誤差に転化させることで，正確さを向上させます。また，局所管理（①）も系統誤差を取り除いて正確さを向上させますが，実験環境がブロック内で均一化しますので，副次的な効果として偶然誤差を減らして精度向上にも役立ちます。



問2

a： 帰無仮説は主張したくない内容ですので「婦人は識別能力を持っていない」，対立仮説はその逆で「婦人は識別能力を持っている」となります。ですから，滅多に当てられないことを婦人が当てたら，帰無仮説（識別能力なし）を棄却して対立仮説（識別能力あり）を採択した方が合理的であると考えます。

b: フィッシャーの三原則のうちの「反復」にあたります。1杯ずつでは，識別能力を持っていなくても，50％の確率で正しく2組に分けることができてしまうため，もし正解しても，婦人が本当に識別できたのか，偶然当たっただけなのかを判定できません。しかし，4杯ずつならば，偶然当ててしまう確率は70分の1（1.43％）で極めて低い確率になるため，識別能力を持っていると考えた方が合理的です。

なお，異なる8杯のなかから4杯を（順序を考えずに）取り出す「組合せの数」は，${}_8C_4$ なので $8!\div 4!\,(8-4)!=(8\times7\times6\times5)\div(4\times3\times2\times1)=70$ となり，70通りとなります。

c: 処理別に4杯連続して供してしまうと，徐々に紅茶が冷めたり濃くなったりして，識別しやすく（あるいは逆に識別しにくく）なるのを防ぐためです。いいかえると，紅茶の温度や濃さという誤差が，処理と同じ（あるいは逆の）方向性を持って結果に交絡してしまわないようにするためです。

d: 反復と無作為化の原則に従っているので「完全無作為化法」になります。また，4日に分けて実験をする場合，日によって婦人の味覚が大きく変化したりするかもしれないので，実験日をブロック因子とした局所管理を導入し，「乱塊法」とすることなどが考えられます。

問3

例えば，要因aの主効果による群間変動を計算してみます。

群間変動は，反復数$\times\sum$(群平均 − 総平均)^{2}です。

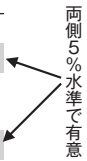
さて，要因aの水準1の群平均は6，水準2の群平均は2.75，総平均は4.375，反復数は4ですから，$4\times\{(6-4.375)^2+(2.75-4.375)^2\}$で，21.125となります。自由度は2 − 1で1ですから，そのまま要因分散となります。同様に誤差分散を計算すると，0.125となりますので，検定統計量Fはその割り算で169.0となります。このF値は，5％有意水準の限界値（$F_{(1,1)}=161.45$）よりも大きいので「要因aの主効果はある」と判定できます。

ほかの要因の変動を計算してみると，要因bは1.125，要因cは36.125，要因dは0.125，交互作用a × bは0.125，a × cは1.125となり，5％水準で有意となるのは要因cの主効果のみです（有意となる交互作用もありません）。

ちなみに，次がExcel分析ツールの［回帰分析］を使った計算結果です（Yにデータ，Xに誤差を除いた要因と交互作用を指定）。p値から，要因aとcのみが0.05よりも小さくなっている（有意である）ことが確認できます。

Excel分析ツールの回帰分析の結果

	係数	標準誤差	t	P-値
切片	-0.12	0.93	-0.13	0.915
a	-3.25	0.25	-13.00	0.049
b	0.75	0.25	3.00	0.205
a×b	0.25	0.25	1.00	0.500
c	4.25	0.25	17.00	0.037
a×c	0.75	0.25	3.00	0.205
d	0.25	0.25	1.00	0.500



第11章

問1

自分自身や研究室の先輩・仲間が行った調査ならばオリジナルの個票やデータが手に入るため，いろいろな統計的検定ができるでしょうが，場合によっては他の組織や人が行った調査結果を利用して分析しなければならないこともあります。このデータも著者が農林水産省のWebサイトから入手したものです（形式を一部修正しています）。表側はカテゴリカル（8水準），表頭は2値（2水準）で8×2の16セルのクロス集計で，度数の極端な偏りもないので，独立性の検定が適しています。

観測度数

	手入れあり	手入れなし	行和
北海道	19	17	36
東北	18	51	69
関東	89	336	425
北陸	9	31	40
東海	16	64	80
近畿	44	147	191
中国四国	20	88	108
九州沖縄	26	106	132
列和	241	840	1081

期待度数（列和×行和÷1081）

	手入れあり	手入れなし	行和
北海道	8.0	28.0	36
東北	15.4	53.6	69
関東	94.8	330.2	425
北陸	8.9	31.1	40
東海	17.8	62.2	80
近畿	42.6	148.4	191
中国四国	24.1	83.9	108
九州沖縄	29.4	102.6	132
列和	241	840	1081

検定統計量〔(観測度数-期待度数)^2/期待度数)〕

	手入れあり	手入れなし
北海道	15.005	4.305
東北	0.445	0.128
関東	0.349	0.100
北陸	0.001	0.000
東海	0.189	0.054
近畿	0.047	0.014
中国四国	0.691	0.198
九州沖縄	0.399	0.115

検定統計量 $x^2=$ 22.040

まず，観測度数表の周辺度数を使って，帰無仮説が正しい場合の期待度数の表を作成します。期待度数は，各セルにおいて行和×列和を総和で割って求めます。例えば，北海道の「手入れあり」の期待度数は $36 \times 241 \div 1081$ で8.0となります。

次に，各セルにおいて$(\text{観測度数} - \text{期待度数})^2/\text{期待度数}$で検定統計量の表

を作成します。最後に全てのセルの値を足し合わせた“22.04”が検定統計量であるピアソンのχ^2となります。

自由度は(8−1)×1で“7”ですので，有意水準5％の限界値をχ^2分布表から読み取ると“14.067”となります。よって，検定統計量（22.04）＞限界値（14.067）となり，表側と表頭は独立しているという帰無仮説は棄却され，関連しているという対立仮説が採択されます。これは，「地域によって森林の手入れには差がある」と，統計的にいえるということです。

Excel関数を使ってp値を算出してみましょう。Excelの適当なセルに，=CHISQ.TEST（実測値範囲，期待値範囲），としてEnterキーを押せば“0.0025”というp値が返ってきます（実測値範囲とは，観測度数の表のデータを範囲指定することです）。

また，効果量である連関係数（クラメールのV）を計算してみると，22.04 ÷ (1081 × 1)の平方根で“0.1428”となりますので，最大値の1からは程遠く，実質的な関連性はそれほど強いとはいえないことがわかります。

問2

この事例のように，「実際に観測された曜日別の売上の分布」と，「曜日は関係ないという理論の下で期待されると曜日別の売上の分布」のズレ具合を検証するのが，“適合度の検定”です。なお，このデータは，度数ではなく売上げ（金額）となっていますが，売上げも1円の度数と考えれば問題ないことがわかるでしょう。

さて，「売上げは曜日に関係ない」とい理論の下での期待度数はどのような分布になるでしょうか？　観測度数の週の合計は700万円なので，どの曜日もそれを7等分した100万円の売上げ（離散型の一様分布）となるはずです。

観測度数分布

日	月	火	水	木	金	土	合計
121	87	81	86	83	98	144	700

期待度数分布（700÷7）

日	月	火	水	木	金	土	合計
100	100	100	100	100	100	100	700

検定統計量〔(観測度数−期待度数)^2/期待度数)〕

日	月	火	水	木	金	土	合計(x^2)
4.41	1.69	3.61	1.96	2.89	0.04	19.36	33.96

ズレ具合

独立性の検定と同じように，曜日ごとに(観測度数 − 期待度数)^2 ÷ 期待度数を計算して，足し合わせたものが検定統計量(ピアソンのχ^2）となりますので，“33.96”となるはずです。自由度は7 − 1で“6”ですので，有意水準5％の限界値をχ^2分布表から読み取ると“12.592”となります。よって，検定統計量（33.96）＞限界値（12.592）となり，「観測度数と期待度数の分布は適合している」という帰無仮説は棄却されます。そして，「両分布は適合しない」→「この直売所の売上げは曜日によって差がある」という対立仮説が採択されます。

問3

アミノ酸の一種であるリジンの濃度が低い飼料で育てた豚の脂肪率の標本平均は，リジン濃度の高い飼料で育てた豚の1.3倍となっています。しかし，普通のt検定を実施しても，有意差は検出されません（$p = 0.82$）。それぞれ極端に大きな値が1つずつ含まれているために，tが小さく歪んでしまっているからだと思われます（実は，このデータは本文中の表11.1と同じです）。

観測値

0.6%	0.4%
5	10
3	8
4	9
2	6
1	7
93	98

並べ替え

0.6%	0.4%
1	6
2	7
3	8
4	9
5	10
93	98

順位を付ける

0.6%	0.4%
1	6
2	7
3	8
4	9
5	10
11	12

0.6%>0.4%の数

0.6%	0.4%
1	0
2	0
3	0
4	0
5	0
11	5

$U = 5$

そこで，外れ値があっても検出力が下がらないマン・ホイットニーのU検定を実施しましょう。まず，小さい順に並び替えて，順位を付けます。その後，0.6％の群よりも小さい順位のデータ数を数え，足し合わせれば検定統計量Uとなります（0.4％の群を基準としても良いですが，普通は小さい方を使います）。すると，Uは“5”となります。

標本サイズが小さいので，両側有意水準5％のU検定表から限界値を読み取りますと，5行目と5列目のクロスする“2”となります。

よって，検定統計量（5）＞限界値（2）となり，「2群の分布位置はズレていない→（2標本は）同じ母集団から抽出された」という帰無仮説は棄却され，「2群の分布はズレている→異なる母集団から抽出された」という対立仮説が採択されます。つまり，リジンの濃度の違いは，豚の筋肉中の脂肪含有率に影響を与えるといえます。

第12章

問1

9月の平均気温が21℃の地点は，観測データにはありません。そのため，回帰モデルを推定し，その式を使って紅葉最盛期を（内挿）予測します。

具体的には，気温が原因で紅葉が結果なので，説明変数xとして「9月の平均気温」，被説明変数yとして「紅葉の最盛期（10月1日からの日数）」の単回帰式を推定することになります。Excelの分析ツールで回帰分析を実施した結果を次に示します（できることならば正規方程式から計算してみましょう）。

概要

回帰統計	
重相関R	0.986
重決定R2	0.972
補正R2	0.968
標準誤差	2.730
観測数	10

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	2041.296	2041.296	273.982	0.000
残差	8	59.604	7.450		
合計	9	2100.900			

	係数	標準誤差	t	P値
切片	-48.376	6.012	-8.047	0.000
9月平均気温	4.602	0.278	16.552	0.000

まず，モデルの適合度である決定係数（重決定R2）は，0.97と大変高く，説明力が強いことがわかります（単回帰の場合は自由度修正済みの方でなくてOK）。また唯一の係数のt検定の結果においてもp値がほぼ0となり，5％有意水準で統計的に有意にゼロから離れた値となっていることから，良好な推定がなされたといえます。

よって，紅葉の最盛期の予測式は，次のようになります。yが10月1日からの日数，xが9月の平均気温です。

$$\hat{y} = -48.38 + 4.602x$$

さて，予測したいのは$x = 21$の地点の最盛期なので，$-48.38 + 4.602 \times 21$

となり，$y = 48.26$となります。つまり，「11月17日頃」に当該地点の最盛期がやってくると予測できるわけです。

実は，実際の関東地方の紅葉の予測式も，日本観光協会が保有している紅葉の見ごろに関する資料に基づき，同じように推定されています（ただし，定数項である切片は−47.69，回帰係数は4.62です）。

問2

2005年現在，農産物直売所は全国で約13538カ所あるといわれています。また，直売所に参加している農家数も約35万戸と推定されています。

この仮想データを重回帰分析にかけることによって，直売所の売上の要因を明らかにすることができます。また，推定された重回帰モデルを使えば特定の条件の直売所の売上高も簡単に予測できます。

早速，売上げを被説明変数y，その他の項目を説明変数x_1〜x_3として，重回帰分析を実施してみましょう。次の表は，分析ツールによる出力結果です。

概要

回帰統計	
重相関 R	0.998
重決定 R2	0.995
補正 R2	0.991
標準誤差	0.501
観測数	8

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	3	205.006	68.335	272.670	0.000
残差	4	1.002	0.251		
合計	7	206.009			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-0.879	0.435	-2.023	0.113	-2.087	0.328
店舗面積（㎡）	0.021	0.006	3.727	0.020	0.005	0.037
駐車可能台数	0.004	0.001	4.404	0.012	0.001	0.006
フリーマーケット（有=1）	5.490	0.658	8.341	0.001	3.663	7.318

これを見ると，自由度修正済み決定係数（補正R2）は1に近く，回帰係数も全て5％水準で統計的に有意と判定されていることから，良好なモデルが推定されたことがわかります。

推定結果の解釈をしてみましょう。店舗面積，フリーマーケット有，駐車可

能台数の3つの説明変数の偏回帰係数はどれもプラスとなっていることから，いずれも売上げを増加させる要因であることがわかります。フリーマーケットの（標準化していない）偏回帰係数は$\widehat{\beta}_2 = 5.490$ですから，店舗面積と駐車可能台数を一定とした場合，その存在は直売所の売上げを年に5490万円押し上げると推測できます。

どの説明変数がもっとも大きな売上要因なのかをみるため，標準偏回帰係数を計算してみましょう。変数ごとに標準化して再度回帰にかけても良いですが，偏回帰係数に(x_iの標準偏差 $\div$ yの標準偏差)を乗じてもOKです。後者の方法で計算してみると，店舗面積の標準偏回帰係数β_1^*は0.021×14.073で“0.295”，駐車可能台数β_2^*は0.004×69.298で“0.261”，フリーマーケットβ_3^*は5.490×0.095で“0.524”となり，売上げにもっとも強い影響を及ぼしている要因はフリーマーケットの有無であることがわかります。

農産物直売所の売上要因に関する既往研究を見てみますと，今回挙げた要因以外にも，農産物の品揃えやトイレの数，朝市の有無，レストランの有無，公園の有無などが売上げに影響することが指摘されています。

第13章

問1

自分で入力するか，オーム社Webページから「ロジスティック回帰（第13章末問題1）.RData」というファイルをダウンロードして，Rコマンダーの［一般化線形モデル］で推定すると，次のような出力が得られます。

```
Call:
glm(formula = 脳出血ダミー ~ 年齢 + 拡張期血圧, family = binomial(logit),
data = Dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
  -1.2655  -0.2665  -0.1780   0.1873   2.0344

Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -18.26671    8.75800  -2.086   0.0370 *
年齢             0.11150    0.05527   2.017   0.0436 *
拡張期血圧         0.14701    0.09422   1.560   0.1187
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.5256  on 19  degrees of freedom
Residual deviance:  8.5584  on 17  degrees of freedom
AIC: 14.558

Number of Fisher Scoring iterations: 6

> exp(coef(GLM.3))  # Exponentiated coefficients ('odds ratios')
     (Intercept)           年齢           拡張期血圧
0.00000001166456    1.11795781729661    1.15836279526126
```

出力結果から，推定されたロジスティック回帰モデルは次のようになります。

$$\widehat{p}_i = \frac{e^{-18.27+0.112\times 年齢 +0.147\times 拡張期血圧}}{1+e^{-18.27+0.112\times 年齢 +0.147\times 拡張期血圧}}$$

まず，回帰係数のz検定の結果（Pr(>|z|)）を見てみると，両変数（年齢と拡張血圧）の係数とも5％水準で有意に0から離れているといえます（＊が1つ付いています）。

回帰係数の符号を見てみると，年齢も拡張期血圧もプラスとなっていますので，加齢や高血圧が脳出血に罹る確率を上げる要因であることがわかります。

回帰係数を解釈するために，オッズ比$e^{\hat{\beta}}$を見てみましょう（Rコマンダーでは一番下のodds ratios)。すると，年齢のオッズは“$e^{0.112} = 1.11$”なので，拡張期血圧が一定だとすると，「年齢が1歳加わると，脳出血に罹らない確率に対する罹る確率（**リスク**）が1.118倍に増加する」ことがわかります。同様に，拡張期血圧のオッズ比は“$e^{0.147} = 1.16$”なので，年齢が一定だとすると，「拡張期血圧が1単位（mmHG)高くなると，脳出血のリスクが1.16倍に増加する」ことがわかります。

脳出血に罹る確率は推定されたロジスティック回帰モデルの年齢に60を，拡張期血圧に90を代入します。すると“0.839”となり，84％の確率で脳出血に罹ることが予測されます。罹るか罹らないかを判別すると，0.5が閾値ですので，脳出血に罹る方に分類されます（あくまで仮想データです)。

$$\widehat{p}_i = \frac{e^{-18.27+0.112\times\text{年齢}+0.147\times\text{拡張期血圧}}}{1+e^{-18.27+0.112\times\text{年齢}+0.147\times\text{拡張期血圧}}} = \frac{e^{-18.27+0.112\times60+0.147\times90}}{1+e^{-18.27+0.112\times60+0.147\times90}}$$
$$= \frac{5.229}{6.229} = 0.839$$

なお，この推定されたロジスティック回帰モデルの評価として，判別的中率（割合によるR^2）を計算してみましょう。20名全員の罹患確率を計算し，観測値（罹患した：1，罹患しなかった：0）と突き合わせてみると，5番目の人が罹患確率0.13なのにも関わらず脳出血に罹っていて，20番目の人が0.55なのに脳出血に罹っていないため，判別的中率は$18 \div 20$で90％となり，高い性能で判別できていると評価できます。

また，McFaddenの擬似決定係数は，$1 - (8.5584 \div 27.5256)$で“0.69”となり，本指標からもモデルの推定がそこそこ良好であったことがうかがえます。

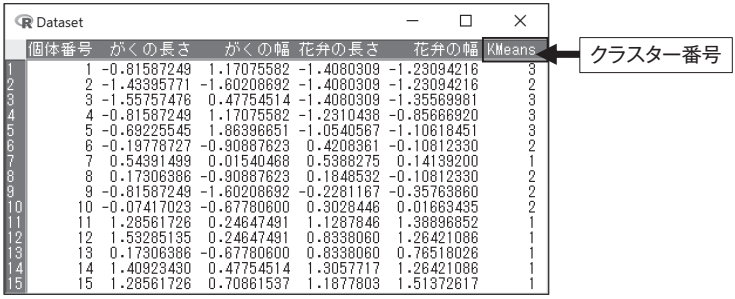
問2

「いずれがアヤメかカキツバタ」ということわざがあるぐらいなので，アヤメの種類というのは，素人には見分けがつかない代表格なのかもしれません。この問題で示したデータは，フィッシャーの1936年の論文に出てくる大変有名なものです（本来はもっと大きいデータなので，興味のある方はインターネット上でオリジナルデータを入手してみましょう。Rコマンダーの「datasets」というパッケージにも「iris」というデータセット名で入っています)。

さて，クラスター分析では，変数間は等分散でないと正しい分類ができなくなってしまう（大きい分散の変数の影響が強く出てしまう）ので，変数ごとに

標準化しておきましょう。

自分で入力して標準化するか，あるいはオーム社Webページから既に標準化してある「k-means法（第13章末問題2）.RData」というファイルをダウンロードして，Rコマンダーの［k-平均クラスタ分析］で推定（変数は4つ全てを選択します）すると，データセットの最右列（KMeans）に次のようなクラスター番号が出力されます（［データセットを表示］ボタンをクリックすると表示されます）。ただし，「オプション」タブで［作成するクラスタ数］を3に設定し，［データセットにクラスタを割り当てる］に☑を入れておいてください。



	個体番号	がくの長さ	がくの幅	花弁の長さ	花弁の幅	KMeans
1	1	-0.81587249	1.17075582	-1.4080309	-1.23094216	3
2	2	-1.43395771	-1.60208692	-1.4080309	-1.23094216	2
3	3	-1.55757476	0.47754514	-1.4080309	-1.35569981	3
4	4	-0.81587249	1.17075582	-1.2310438	-0.85666920	3
5	5	-0.69225545	1.86396651	-1.0540567	-1.10618451	3
6	6	-0.19778727	-0.90887623	0.4208361	-0.10812330	2
7	7	0.54391499	0.01540468	0.5388275	0.14139200	1
8	8	0.17306386	-0.90887623	0.1848532	-0.10812330	2
9	9	-0.81587249	-1.60208692	-0.2281167	-0.35763860	2
10	10	-0.07417023	-0.67780600	0.3028446	0.01663435	2
11	11	1.28561726	0.24647491	1.1287846	1.38896852	1
12	12	1.53285135	0.24647491	0.8338060	1.26421086	1
13	13	0.17306386	-0.67780600	0.8338060	0.76518026	1
14	14	1.40923430	0.47754514	1.3057717	1.26421086	1
15	15	1.28561726	0.70861537	1.1877803	1.51372617	1

これを見ると，7と11〜15の個体で第1クラスター，2，6，8〜10の個体で第2クラスター，1，3〜5の個体で第3クラスターが作られたことがわかります。

種明かしをしますと，実はフィッシャーの論文にはどの個体が何という種類かという外的基準が示されています（問題文では隠していました）。第1クラスターは「セトーサ」，第2クラスターは「ヴェルシコロール」，第3クラスターは「ヴィルジニカ」という種類だそうです。正解は，個体番号1〜5がセトーサ，6〜10がヴェルシコロール，11〜15がヴィルジニカでした。結果と正解を比較すると，そこそこ上手く分類できたことがわかります（個体番号2と7だけ間違って分類されました）。ただし，普通は，外的基準は存在しないので，こうした判別的中率は計算できません。

なお，メニュー［統計量］→［要約］→［相関行列］で変数間の相関係数を確認しますと，がくの長さ，花弁の長さ，花弁幅の間に0.9を超える強い相関があることがわかります。Rコマンダーのk-means法は，独立した変数に適したユークリッド距離しか使えないので，本来，クラスター分析の練習問題としてはあまり良くないデータだったといえます<(_ _)>。

第14章

問1

実は，Excelでも「ソルバー」という最適化問題を解くためのアドインを使えば，主成分分析や因子分析を実施できるのです（最尤法もできるのでロジスティック回帰も実施できます）が，かなり面倒ですのでRコマンダーを使ってみましょう。

データを自分で入力するか，オーム社Webページから「主成分分析（第14章末問題1）.RData」というファイルをダウンロードして，［データ］→［データセットのロード］で読み込みます。次に［統計量］→［次元解析］→［主成分分析］で起動したウィンドウの，［データ］タブで全ての変数を選択し，［オプション］タブで［相関行列の分析］と［スクリープロット］と［データセットに主成分得点を保存］の全てに☑を入れて［OK］ボタンを押します。最後に［保存する主成分数］を“2”に設定して［OK］ボタンを押すと，次のような出力が得られます。

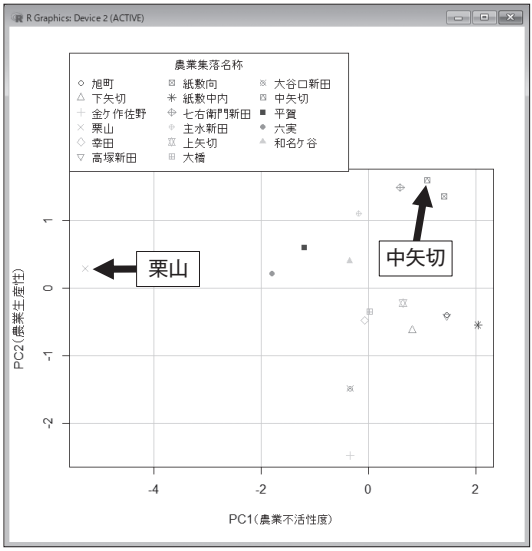
```
Component loadings（固有ベクトル）:
                    Comp.1    Comp.2   Comp.3   Comp.4   Comp.5
基幹的農業従事者率   0.045     0.898    0.426    0.011    0.088
経営耕地面積増加率  -0.422    -0.282    0.615    0.602    0.012
生産年齢人口率       0.466     0.142   -0.383    0.784    0.016
農家人口増加率      -0.552     0.151   -0.409    0.086    0.704
農家数増加率        -0.544     0.262   -0.354    0.117   -0.703

Component variances（固有値）:
    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
    2.696     1.073     0.693     0.470     0.066

Importance of components:
                                       Comp.1  Comp.2  Comp.3  Comp.4
Standard deviation（標準偏差）          1.642   1.036   0.832   0.685
Proportion of Variance（寄与率）        0.539   0.214   0.138   0.094
Cumulative Proportion（累積寄与率）     0.539   0.754   0.892   0.986
```

第2主成分までは固有値が1を超えており（スクリープロットでも確認しておきましょう），第2主成分までの累積寄与率が0.754ですので，代表性も高いといえます。

第1主成分の固有ベクトルを見てみると，耕地面積や農家の数，そして農家の人口の増加率が負になっていますので，得点の小さい集落ほど農業全般が元気であるといえそうです。よって，『農業不活性度』とネーミングできるで

しょう。第2主成分は，基幹的農業従事者（普段の主な状態が農業従事である者）率と生産年齢（15～64才）人口率の固有ベクトルが正で大きな値となっていることから，『農業生産性』とネーミングできるのではないでしょうか。

次に，データセットに保存された主成分得点を使って散布図（PC1を横軸，PC2を縦軸）を描くと，次のように15の集落が分類されます（作図の方法は本文参照）。これを見てみると，集落個別には，栗山（×）の農業が元気であることや，中矢切（△）の生産性が高いことがわかります。また，分類してみると，例えば右下の第4象限にある5つの集落は農業が全般的に元気がなく，生産性も低いことなどがわかります。



問2

因子分析もExcelで実行するのは大変ですので，Rコマンダーを使った結果を紹介します。なお，Rコマンダーでは“factanal関数”を使うようになっていますが，他にも“psy”や“psych”という因子分析用のパッケージがあります。

まず，データを自分で入力するか，オーム社Webページから「因子分析（第14章末問題2）.RData」というファイルをダウンロードして，［データ］→［データセットのロード］で読み込みます。次に本文と同様に，［統計量］→［次元解析］→［因子分析］を実行すると，次のような結果が出力されます（因子の回転は“プロマックス”，得点の推定方法は“回帰”で結構です）。

```
Uniquenesses:
  英語   国語   社会   数学   理科
0.586 0.492 0.005 0.612 0.005

Loadings:
     Factor1 Factor2
英語    0.637  -0.158
国語    0.705
社会    0.977   0.138
数学            0.625
理科    0.156   0.973

                Factor1 Factor2
SS loadings       1.886   1.385
Proportion Var    0.377   0.277
Cumulative Var    0.377   0.654

Factor Correlations:
        Factor1 Factor2
Factor1  1.0000 -0.0781
Factor2 -0.0781  1.0000

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 2.26 on 1 degree of freedom.
The p-value is 0.133
```

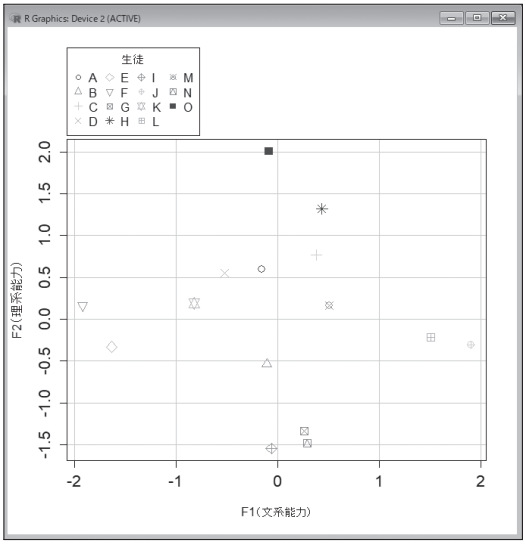
推定結果を見てみましょう。まず，英語と数学の独自性（Uniquenesses）の値が大きいことから，両科目については2つの共通因子では説明は十分ではないといえます。また，社会と理科については，factanal関数の下限値である0.005であるため，もしかしたら標本サイズが小さすぎて，不適解となって（共通性が1を超えて）いる可能性があります。

ただし，累積寄与率（Cumulative Var）は第2因子までで0.654ありますので，本モデルの説明力は悪くありません。また，最後の適合度検定もp値が0.133と大きいことから，推定モデルと観測データとが適合しているという帰無仮説は受容されています（共通因子は2つでOKということ）。

さて，因子負荷量（Loadings）を見てみると，両因子とも変数の負荷量にメリハリがあり，構造が単純で因子の解釈がしやすそうです。第1因子（Factor1）は英語と国語と社会の負荷量が正に大きく，第2因子（Factor2）は数学と理科の負荷量が正に大きくなっていることから，それぞれ『文系能力』と『理系能力』とネーミングできそうです。

因子得点を散布図にしてみると，次のようになります（描き方は本文の主成分分析で解説しています）。これを見てみると，文系，理系どちらの能力も高

い（あるいは低い）生徒は少なく，両能力は無関係である（直交している）ことがわかります。それは，先ほどの出力結果のFactor Correlations（因数間の相関係数）が0に近いことからも確認できます。つまり，理系能力が高いからといって文系科目が苦手というわけでも，理系能力が高いと文系科目も得意というわけでもないことがわかりました。



第15章

問1

不良品だったという結果をみて，それがC工場製であるという原因の確率を求めます。新型コロナウィルスの例題では，原因が感染・非感染の2種類の事象だけでしたが，本問題では，それが製造した3つの工場になる点がやや難しくなっております。一方，結果は，良品・不良品の2種類だけですので，新型コロナの陽性・陰性と同じです。

さて，不良品だったパネルがC工場製である事後確率P（C工場｜不良品）を求めるためのベイズの定理は，次のようになります。

$$P(\text{C工場}\,|\,\text{不良}) = \frac{P(\text{不良}\,|\,\text{C工場})P(\text{C工場})}{P(\text{不良}\,|\,\text{A工場})P(\text{A工場}) + P(\text{不良}\,|\,\text{B工場})P(\text{B工場}) + P(\text{不良}\,|\,\text{C工場})P(\text{C工場})}$$

それぞれ必要な確率を整理しましょう。まず，それぞれの工場で製造された確率は，全体に占める製造台数ですから簡単ですね。

$$P(\text{A工場}) = 0.5,\ P(\text{B工場}) = 0.3,\ P(\text{C工場}) = 0.2$$

次に，それぞれの工場で製造されたパネルが不良品の確率ですが，こちらも各工場の歩留りがわかっているので，その補数とすればよいのです。

$$P(\text{不良}\,|\,\text{A工場}) = 0.2,\ P(\text{不良}\,|\,\text{B工場}) = 0.3,\ P(\text{不良}\,|\,\text{C工場}) = 0.4$$

以上の確率値を，先ほどのベイズの定理に代入すると，つぎのように0.296となります。

$$P(\text{C工場}\,|\,\text{不良品}) = \frac{0.4 \times 0.2}{0.2 \times 0.5 + 0.3 \times 0.3 + 0.4 \times 0.2} = 0.296$$

よって，不良品だったパネルがC工場製である確率は29.6％となります。

問2

ロジスティック回帰モデルの式を，yがベルヌーイ分布（試行回数nが1の二項分布）に従うことを強調した形で書き換えると次のようになります（2値でない割合データの場合は$n = 1$ではありません）。なお，$\sim B$は二項分布に従うという意味です。

試行回数$n=1$　　p_i

$$y_i \sim B\left(1, \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)$$

よって，事前分布を設定すべきパラメータはαとβの2つのみとなり，正規分布に従う線形回帰より単純であることがわかります。