College Analysis 総合マニュアル

－ 基本統計２ －

目次

１１．２次元グラフ

11.1　２次元グラフ

これは主に統計で利用するグラフを集めたもので、グラフ表示の際に集計は行わない。メニュー［ファイルー基本統計ー集計グラフー2 次元グラフ］を選択すると、図 1 のような分析実行画面が表示される。

図 1　2 次元グラフ実行画面

グラフの種類は、棒グラフ、積重ね棒グラフ、横棒グラフ、積重ね横棒グラフ、帯グラフ、立体棒グラフ（2D）、折れ線グラフ、横折れ線グラフ、円グラフ、散布図、レーダーチャート、比較レーダーチャート、である。

グラフ選択で「棒グラフ」を選択し、変数を 1 種類選んで、「実行」ボタンをクリックすると、図 2a のようなグラフが表示される。また、変数を 2 種類選ぶと図 2b のようなグラフになる。

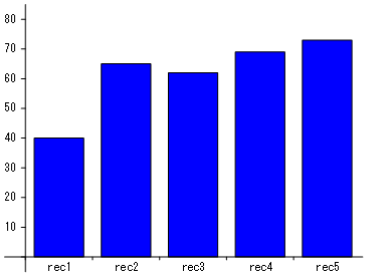


図 2a　棒グラフ（1 変数）

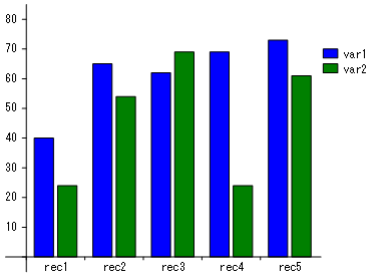


図 2b　棒グラフ（2 変数）

図 2b はグラフの「設定」メニューで、凡例を追加している。また、グラフの横軸の項目名や凡例名は、グラフの「編集」メニューで、「項目名変更」や「データ・凡例名変更」によって変更することができる。また、「画面コピー」でグラフをクリップボードに保存でき、ワープロ等に貼り付けて利用できる。

欠損値除去のラジオボタンで、「欠損値除去あり」を選択した場合のグラフを図 3a に、「欠損値除去なし」を選択した場合のグラフを図 3b に示す。

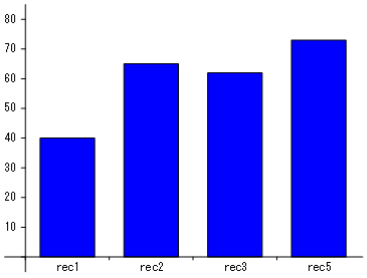


図 3a　棒グラフ（欠損値除去あり）

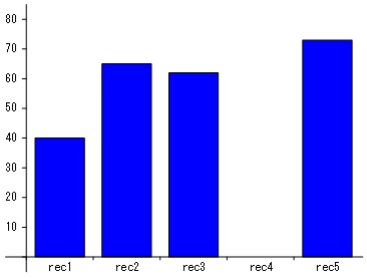


図 3b　棒グラフ（欠損値除去なし）

以後それぞれのグラフで、欠損値の除去の有無による違いがあるので、実際に操作してみて欲しい。

変数を３つ選んだ場合の「積重ね棒グラフ」の例を図 4 に示す。

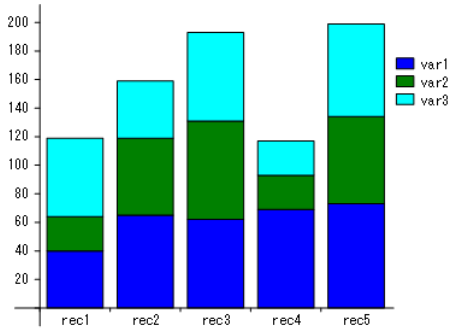


図 4　積重ね棒グラフ

変数を１つ選んだ横棒グラフを図 5a に、２つ選んだ横棒グラフを図 5b に示す。

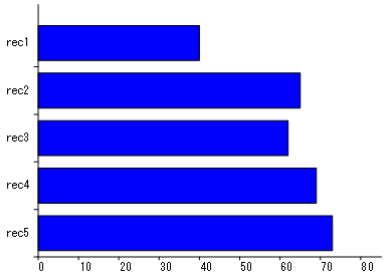


図 5a　横棒グラフ（1 変数）

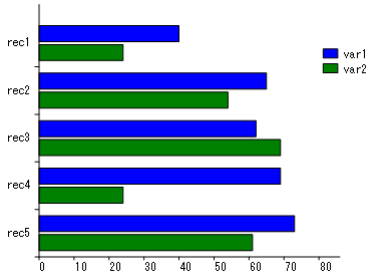


図 5b　横棒グラフ（2 変数）

変数を 3 つ選んだ積重ね横棒グラフの例を図 6 に描く。

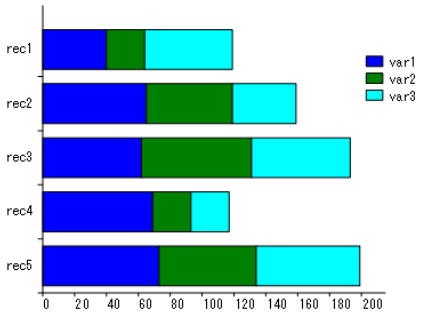


図 6　積重ね横棒グラフ

積重ね横棒グラフの右端に揃えたものが帯グラフである。帯グラフの例を図 7 に示す。

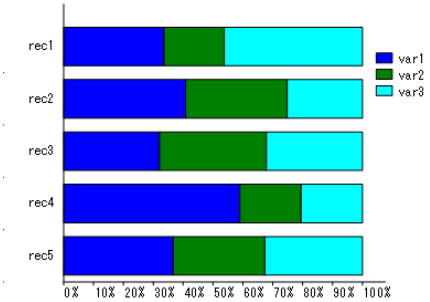


図 7　帯グラフ

立体棒グラフの例を図 8 に示す。

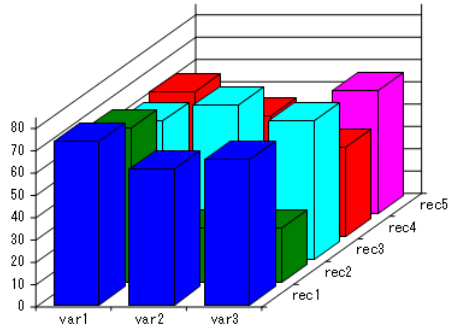


図 8　立体棒グラフ

3 次元グラフに含まれる 3D 棒グラフとは異なり、これには遠近感を付けていない。そのため、意外に棒の高さが比較し易いように思われる。

折れ線グラフの例を図 9 に示す。

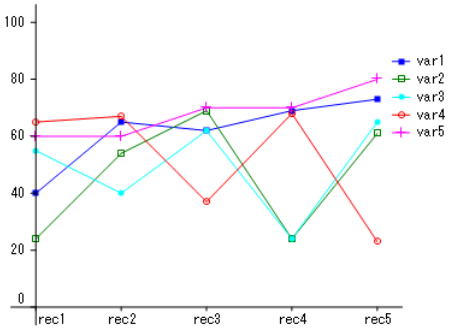


図 9　折れ線グラフ

ここで、縦軸はグラフのメニュー［設定－軸設定］によって、最小値 0、最大値 100、目盛間隔 20 に設定した。

折れ線グラフの縦横を変えたものが、横折れ線グラフで、例を図 10 に示す。

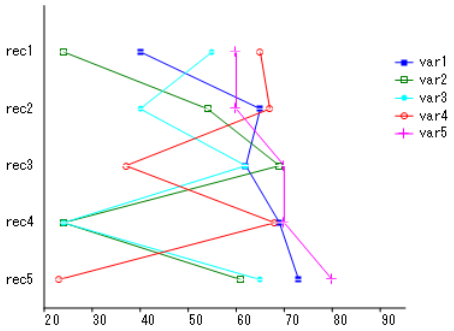


図 10　横折れ線グラフ

これは、ユーザーのリクエストにより、特殊な用途向けに作ったグラフである。

円グラフの例を図 11 に示す。

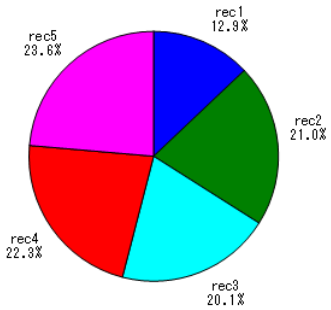


図 11　円グラフ

円グラフの文字位置は、メニュー［編集－項目名位置変更］で表示される図 12 のメニューで、標準位置からずらすことができる。

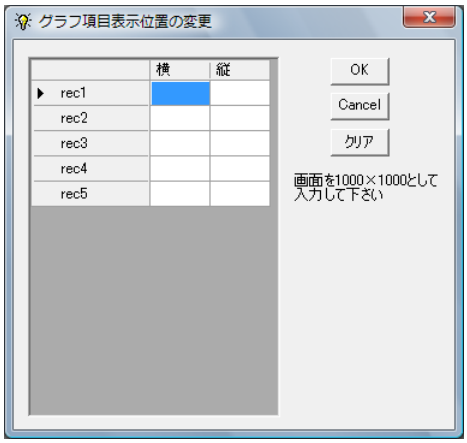


図 12　項目名位置変更

回帰直線の付いた散布図の例を図 13a に、メニュー「設定」の「回帰直線[ON/OFF]」で回帰直線を取って、「データラベル[ON/OFF]」でラベルを付けた例を図 13b に示す。

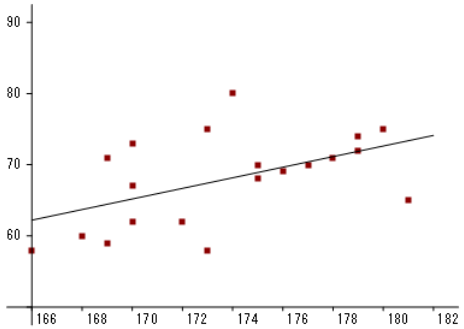


図 13a　散布図（回帰直線）

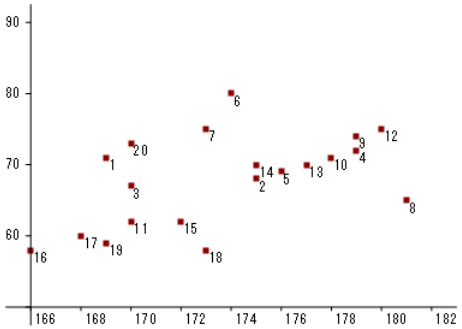


図 13b　散布図（データラベル）

この他にも「設定」メニューで「正規楕円半径」で半径を選択すると、指定された確率楕円が表示される。

変数を３つ選んだレーダーチャートの例を図 14 に示す。レーダーチャートはすべての軸目盛が揃った図である。レーダーチャートには目標値と個々のデータが含まれるが、鎖線で描かれたものが目標値である。

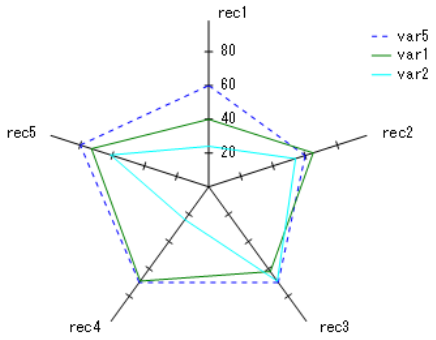


図 14　レーダーチャート

次に、変数を３つ選んだ比較レーダーチャートの例を図 15 に示す。比較レーダーチャートは目標値に対する達成率を表す図で、目標値が同じ半径で描かれている。

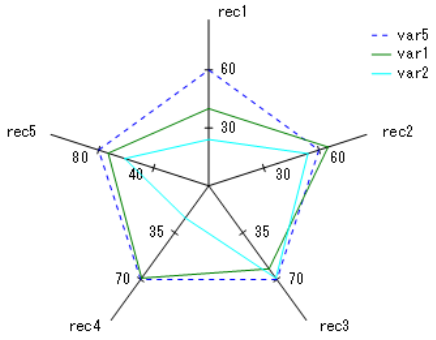


図 15　比較レーダーチャート

誤差付き折れ線グラフは、変数として、平均値１、誤差 1（標準偏差または標準誤差)、平均値 2、誤差 2、・・・の形式で入力する。レコードは例えば年度や実験番号などの比較する時系列的な分類値で、図 16 のように表示される。

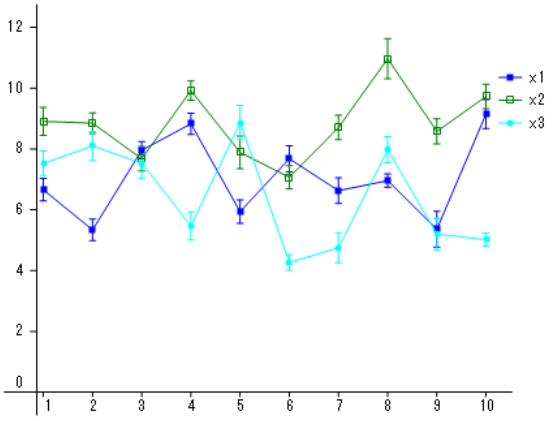


図 16　誤差付き折れ線グラフ

同じく、時々見られる誤差付き棒グラフは、変数として平均値と誤差（標準偏差または標準誤差）が取られるが、レコードは比較する分類である。結果は図 17 のように表示される。

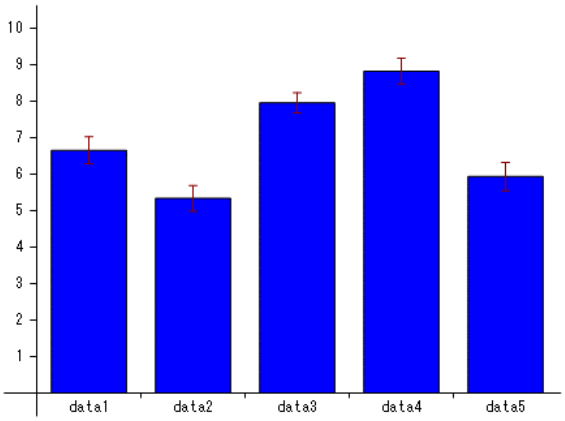


図 17　誤差付き棒グラフ

2 値折れ線グラフは異なった値を左右両側の軸で表示して、2 つの変数の変化を比較するようなグラフである。結果は図 18 のように表示される。

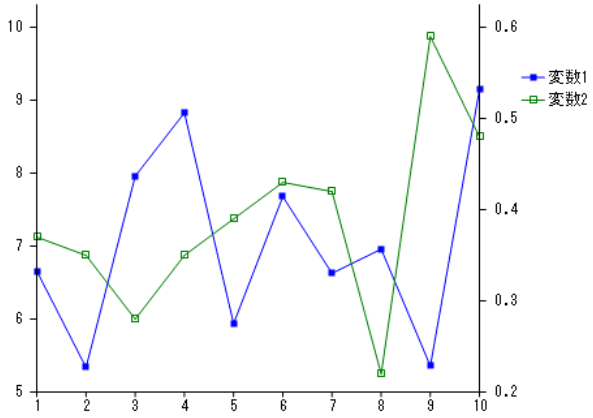


図 18　2 値折れ線グラフ

グラフの機能については、総合マニュアル（ツール）の中にも色の設定や文字の追加などの説明があるので参照して欲しい。

11.2　特殊グラフ

1)　散布図・関数グラフ

同時に表示された複数の散布図の中に、予測曲線を散布図の数だけ表示したい場合がある。１つだけなら、C.Analysis の予測曲線を計算するプログラムに散布図と予測曲線を同時に表示する機能が付いているが、複数の散布図と予測曲線を同時に表示できるものはない。しかし、個々の分析プログラムにそのような複雑な機能を付けることは困難であるし、無駄でもある。そのため、我々は散布図とグラフを別にし、単純に複数の散布図と複数の関数を独立に表示するプログラムを開発した。関数部分は自由に表示できるため、個々の分析プログラムの結果を混ぜて使用することもできる。

メニュー［分析－基本統計－集計グラフ－散布図・関数グラフ］を選択すると図 1 のような実行画面が表示される。

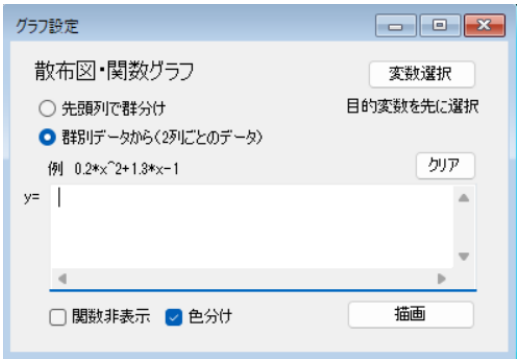


図 1　分析実行画面

これは単純な複数の散布図とグラフを表示するだけのプログラムであるが、うまく活用すると効果的である。以後利用法を述べる。

分析に使うデータを図 2 に示す。

図 2　散布図・関数グラフ.txt

この中の年収と支出を地域で分類し、支出を年収の対数関数グラフで表すことにする。もちろん予測する関数形は異なってもよいがここでは同一とする。また、線形回帰のプログラムから、予め変数を変換して実行することも考えられるが、ここでは使わない。

まず、データを地域 1 と地域 2 に分離する。メニュー［ツール－データ形式変換］を選択すると、図 3 のような実行画面が表示される。

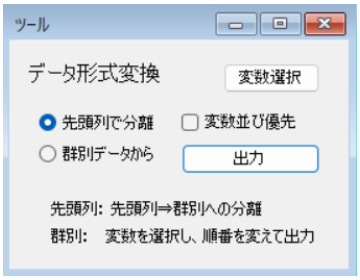
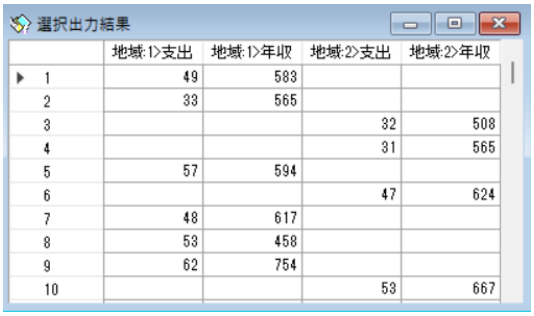


図 3　データ形式変換実行画面

群を分けたいときは、「先頭列で群分け」を選択して、それを含めて分けて出力したい変数を選択する。ここでは、地域、支出、年収の順に選択する。「出力」ボタンをクリックすると図 4 のような結果が表示される。



	地域:1>支出	地域:1>年収	地域:2>支出	地域:2>年収
1	49	583		
2	33	565		
3			32	508
4			31	565
5	57	594		
6			47	624
7	48	617		
8	53	458		
9	62	754		
10			53	667

図４　群分けされたデータ

これは一見、データをずらしただけのように見えるが、ソフトはデフォルトで欠損値を削除してくれるので、分けた場合と同等である。これをエディタの前のデータの後ろに付けて分析を実行することもできるし、別のページに貼り付けて利用することもできる。ここでは後者を選ぶことにする。

ここで予測式を得るために、メニュー［分析－基本統計－非線形回帰分析］を選択し、図５の分析実行画面を表示させる。

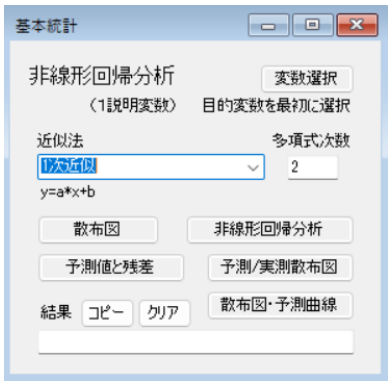


図５　非線形回帰分析実行画面

変数選択として図４の最初の２つの変数（地域 1）を選択し、「近似法」として「対数近似」を選択する。その後「非線形回帰分析」ボタンをクリックすると図６のような結果を得る。

y=a*log(x)+b	推定値	標準誤差	z統計量	確率値	95%下限	95%上限
a	54.3621	5.9781	9.0936	0.0000	42.6453	66.0789
b	-302.7156	38.3937	-7.8845	0.0000	-377.9658	-227.4653
実測・予測 R	0.622	R^2	0.387			

図６　非線形回帰分析結果出力

この結果は「結果」テキストボックスに反映されている。これをコピーして、図１の「散布図・関数グラフ」の関数入力用のテキストボックスに貼り付ける。これを地域２についてもう一度繰り返し、図７の状態にする。

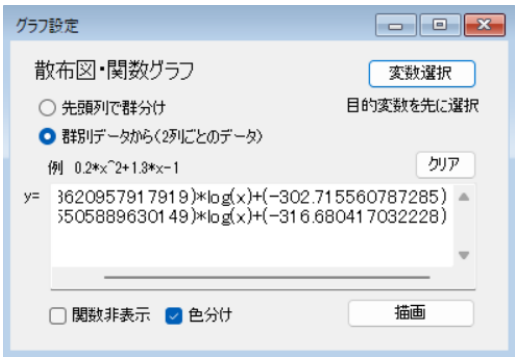


図７　関数の入力された画面

最後に「変数選択」ですべてを選び、「描画」ボタンをクリックすると図８のようなっ結果が得られる。

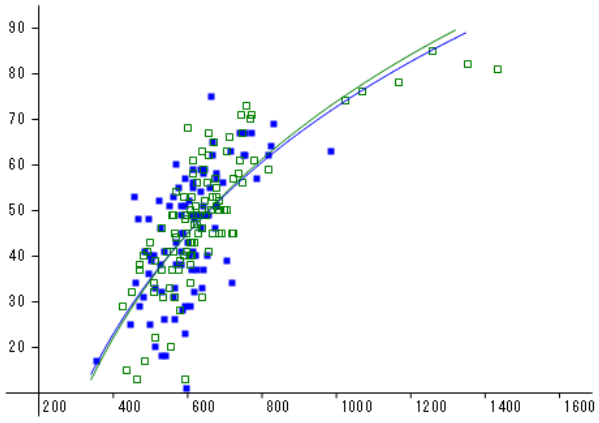


図８　描画結果

２つの散布図とその予測曲線が同時に描かれている。

動画

１２．３次元グラフ

３次元グラフは、3 次元空間上に表示されるグラフで、３Ｄビューアによって表示されるため、自由に回転させたり、近づけたりすることができる。3 次元グラフの描画画面を図１に示す。

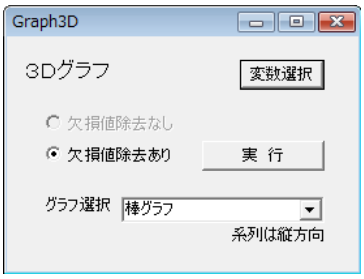
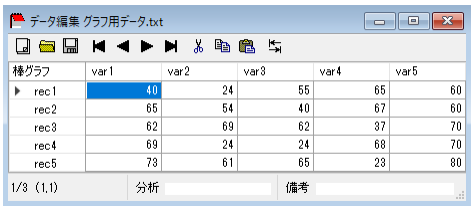


図１　3D グラフ描画画面

このメニューは、まだ開発中のもので、分析は、棒グラフと散布図しかない。

3D 棒グラフのデータと例を図 2 に示す。



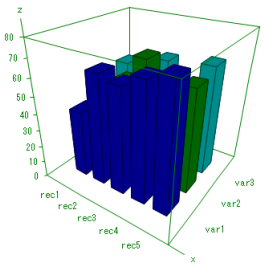
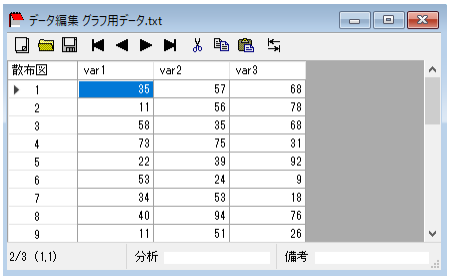


図２　3D 棒グラフ

3D 散布図のデータと例を図 3 に示す。



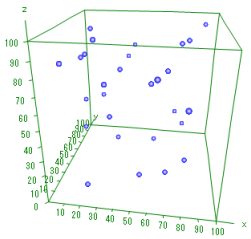


図３　3D 散布図

これらのグラフについて標準の 3D 描画以外の機能は全くないので今後作っていかなければならない。

１３．トレンドの検定

13.1　トレンドの検定とは

トレンドの検定とはある順番に群を並べた場合に、その群のデータについての比率や平均値などの統計量が次第に大きくまたは小さくなってゆく傾向の有無を調べることである。質的なデータに対する比率のトレンドの検定では Mantel-extension 法が利用される。量的データに関する Jonckheere の順位和検定は分布によらない検定である。

メニュー［分析－基本統計－その他の検定－トレンドの検定］を選択すると図 1 のようなにトレンドの検定の実行画面が表示される。

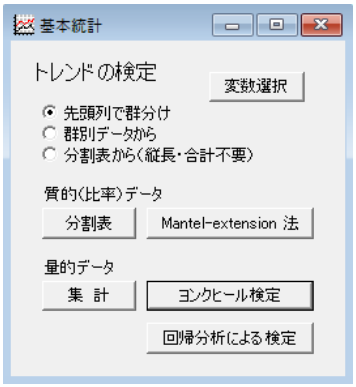


図 1　トレンドの検定実行画面

このメニューにはデータ形式の選択ボタンと「変数選択」ボタンがあるが、これらの使い方はこれまでの統計分析のものと同じである。

今、図 2 のようなデータがあり、それを「先頭列で群分け」として分割表にすると図 3 のようになったとする。

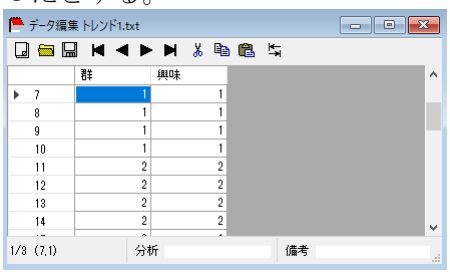


図 2　質的検定データ

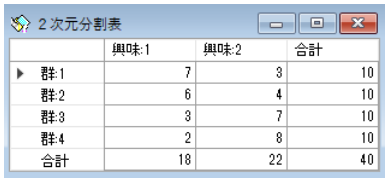


図 3　分割表

この分割表を見ると群:1 から群:4 まで、興味ありの比率が上がっているように見えるが、これは偶然か否かを検定する。このデータに対して、「Mantel- extension 法」ボタンをクリックすると図 4 のような結果画面が示される。

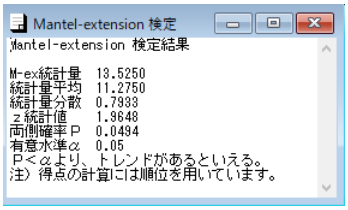


図 4　Mantel-extension 検定結果

このデータにはトレンドがあると判定された。この検定については、図 5 や図 6 のデータ

形式も利用できる。それぞれ、群別データと分割表データである。

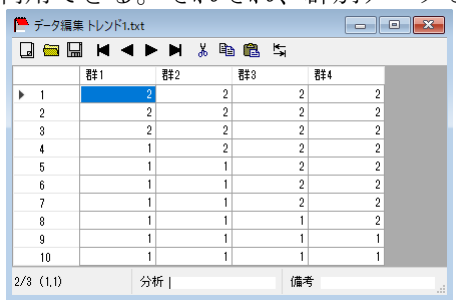


図 5　群別データ

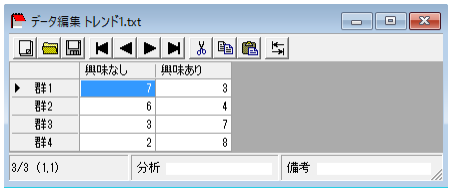


図 6　分割表データ

分割表データには合計は不要である。

量的データについては、図 7 のようなデータを考える。

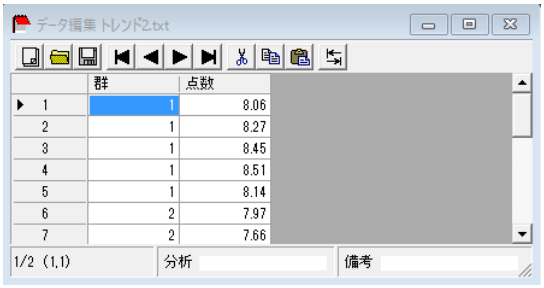


図 7　量的検定データ

「先頭列で群分け」ラジオボタンを選択し、「ヨンクヒール検定」ボタンをクリックすると、図 8 のような結果が得られる。

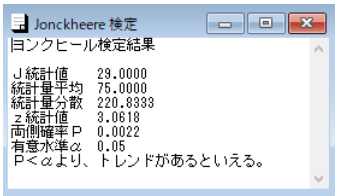


図 8　ヨンクヒール検定結果

量的データのトレンドの検定には、他に回帰分析による検定もある。これは群を強制的に 1, 2, 3, …とし、その数字を使ってデータを予測する回帰式を作る方法である。この場合は傾きのパラメータが 0 かどうかを検定することになる。「回帰分析による検定」ボタンをクリックした場合の結果を図 9 に示す。

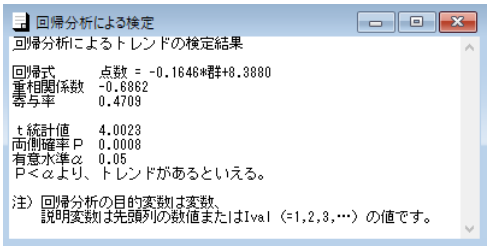


図 9　回帰分析による検定結果

13.2 トレンドの検定の理論

比率のトレンドの検定では Mantel-extension 法が利用されるが、これには以下のように表される統計量 Z または Z' が用いられる。

群 i （$i=1,2,3,\cdots m$）の個体数を n_i，反応した個体数を r_i として以下の量を考える。

$$O=\sum_{i=1}^{m} r_i X_i\ ,\quad E=\left(r\sum_{i=1}^{m} n_i X_i\right)\Big/N\ ,\quad V=\frac{r(N-r)}{N^2(N-1)}\left[N\left(\sum_{i=1}^{m} n_i X_i^{\,2}\right)-\left(\sum_{i=1}^{m} n_i X_i\right)^2\right]$$

ここに、$r=\sum_{i=1}^{m} r_i$，$N=\sum_{i=1}^{m} n_i$ である。また X_i については、最も簡単に $X_i=i$ とした。

これらを用いて漸近的に標準正規分布に従う統計量 Z を計算する。

$$Z=\frac{O-E}{\sqrt{V}}\xrightarrow[n_i\to\infty]{} N(0,1)$$

実用上は以下のような Yates の連続補正項を加えた統計量 Z' を用いる場合も多いが、

$$Z'=\frac{|O-E|-1/2}{\sqrt{V}}\xrightarrow[n_i\to\infty]{} N(0,1)\text{ の正の部分}$$

層別 Mantel-extension 法との関連でここでは採用していない。

量的データに関する Jonckheere の順位和検定は分布によらない検定で、以下のように計算される統計量 Z または Z' を用いる。但し n_i と N についてはこれまでの定義と同じである。

i 群のデータ $x_{i\lambda}$ と j 群（$i<j$）のデータ $x_{j\mu}$ について、$x_{i\lambda}<x_{j\mu}$ なら w_{ij} を 1 増やし、$x_{i\lambda}=x_{j\mu}$ なら w_{ij} を $1/2$ 増やすという処理を群 i と群 j に含まれるすべてのデータについて行う。これは近似的な同順位の処理を行った Wilcoxon の順位和を計算することに等しい。この w_{ij} をすべての i,j（$i<j$）について合計し、以下の量を求める。

$$J=\sum_{i<j} w_{ij}\ ,\quad E=\left(N^2-\sum_{i=1}^{m} n_i^2\right)\Big/4\ ,\quad V=\left[N^2(2N+3)-\sum_{i=1}^{m} n_i^2(2n_i+3)\right]\Big/72$$

これらを用いて漸近的に標準正規分布する以下の統計量 Z を計算する。

$$Z=\frac{J-E}{\sqrt{V}}\xrightarrow[n_i\to\infty]{} N(0,1)$$

しかし実用上は上と同様に Yates の連続補正を加えた統計量 Z' を用いる場合が多い。

$$Z'=\frac{|J-E|-1/2}{\sqrt{V}}\xrightarrow[n_i\to\infty]{} N(0,1)\text{ の正の部分}$$

群 i（$i=1,2,\cdots,m$）の数値 i を説明変数にして、データ $x_{i\lambda}$ を目的変数にする回帰分析もトレンドの検定として考えることができる。即ち、以下のような回帰モデルを考える。

$$x_{i\lambda}=a\cdot i+b+u_\lambda\ ,\quad u_\lambda\ \sim\ N(0,\sigma^2)\ ,$$

これを用いて $a\neq 0$ の検定を行い、群の並びでデータの値に傾向性が見られるか調べる。この回帰式の検定については参考文献 6) に詳しいのでここでは省略する。

参考文献

[1] 新版 医学への統計学，古川俊之，丹後俊郎，朝倉書店，1993.

１４．マルコス連鎖モンテカルロ法による乱数発生

共分散構造分析やベイズ統計などで有力な手法として利用されるマルコフ連鎖モンテカルロ法（MCMC）について、その性質を調べるために乱数発生のプログラムを作成した。発生した乱数はヒストグラムで表示され、理論分布と比較することができ、そのままデータとしてグリッドに出力することもできる。最初にマルコフ連鎖モンテカルロ法の Metropolis-Hastings 法について述べ、次にプログラムの利用法について説明する。

14.1　マルコフ連鎖モンテカルロ法とは

過去のデータから順次確率的に決まって行く時系列のデータを $x^{(t)}$ とし、このデータの従う分布（の密度関数）を $f^{(t)}(x)$ とする。このデータの決定過程を確率過程というが、マルコフ連鎖とは、1 期先のデータ $x^{(t+1)}$ が、それまでのデータの履歴によらず、$x^{(t)}$ の値だけから推移確率（推移核） $p\left(x^{(t+1)}|x^{(t)}\right)$ によって、決まるような確率過程をいう。ある条件のマルコフ連鎖に従うデータは、時間と共に、一定の分布 $f(x)$ に推移することが知られている。この性質を利用して乱数を発生させる方法がマルコフ連鎖モンテカルロ法である。

ここでは密度関数 $y=f(x)$ の乱数を発生させる問題を考える。下図のように、まずある点 x_1 を起点として平均 0 の正規分布（特にこれに限らないが非対称の分布の場合は下の式が異なる）の乱数を 1 つ発生し、その点を x_2 とする。その確率を p とする。

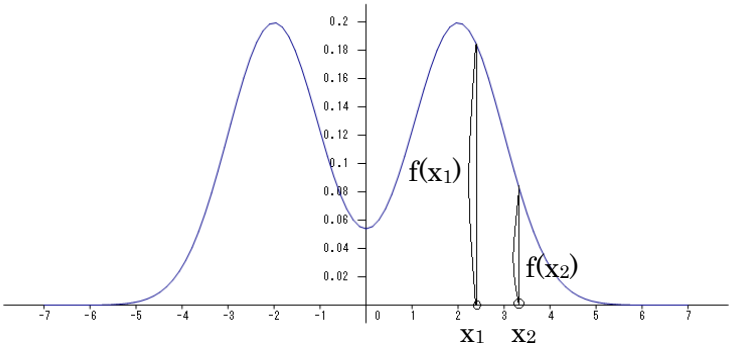


図 1　乱数発生

次にこの値と 2 点での関数の高さから、以下のような量を計算する。

$$\alpha=\begin{cases}\min\left[f(x_2)/f(x_1),1\right] & \text{分母}>0\\ 1 & \text{分母}=0\end{cases}$$

コンピュータで、[0, 1]区間の一様乱数を発生させ、その値がこの α より小さいなら x_2 を採択し、大きいなら、改めて x_1 より再度やり直す。（これは確率 α で x_2 を採択すると言ってもよい。）

これを続けて行くと、最終的に密度関数が $y=f(x)$ で与えられる乱数に収束することになる。なぜなら x_1 の点から $x_2\pm d/2$ の点に遷移する確率は $f(x_2)/f(x_1)d$ 、 x_2 の点から $x_1\pm d/2$ の点に遷移する確率は d になり、両者の比 $f(x_2)/f(x_1):1=f(x_2):f(x_1)$ は、推移する先の点の密度関数の高さの比になる。これがすべての 2 点で成り立っているため、

各点の出現比率は密度関数の高さに比例する。即ち、これは密度関数で与えられる分布の乱数を発生したことになる。正しい乱数にするために、最初のいくつかの点（関数にもよるが数千点以上）は捨てることが望ましい。

ここでは、Metropolis-Hastings 法について述べたが、Hamiltonian マルコフ連鎖モンテカルロ法というのもある。詳細は最後に節に書いておく。

14.2　プログラムの利用法

メニュー［分析－基本統計－MCMC 乱数発生］を選択すると、図 1 のような実行画面が表示される。

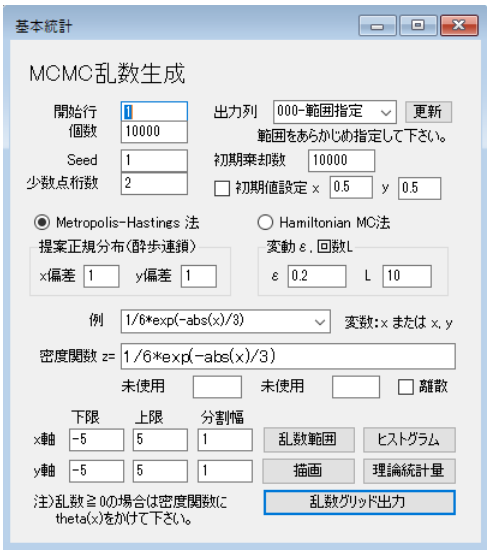


図 1　MCMC 乱数発生実行画面

乱数発生には基本的にメトロポリス・ヘイスティングス法とハミルトニアン・モンテカルロ法があり、メニューで変更できる。

プログラムを利用する際、まず「密度関数」テキストボックスに、出力させる目的分布の乱数の密度関数を入力する。「例」のコンボボックスにサンプルが入っているので、それを参考にしてもらいたい。ここではまず、密度関数 = 1/6*exp(-abs(x)/3) の 1 次元の例を用いて説明を行う。

目的分布の密度関数を入力したら、描画範囲の x 軸の上限と下限を入力する。この範囲はあくまで描画する際の表示範囲で、乱数発生はこれにとらわれない。乱数の発生範囲は、「最大・最小」ボタンで、図 2 のように表示される。

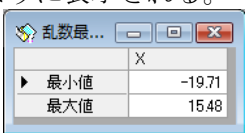


図 2　乱数発生の最小・最大

描画範囲が不明の場合はこの結果を参考にしてもよい。

描画範囲として下限-20 と上限 20 を入力したら、まず、「ヒストグラム」ボタンで図 3a

のようなヒストグラムを描いてみる。

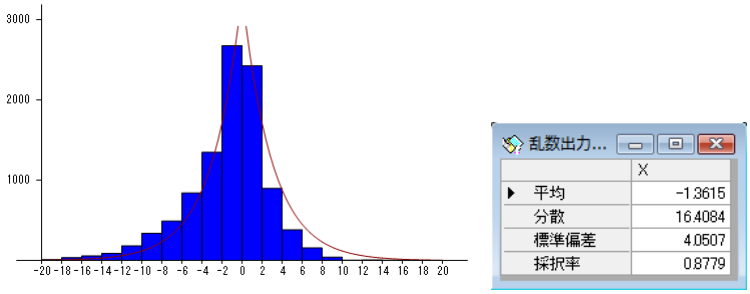


図 3a　乱数のヒストグラムと理論曲線（Seed=1）

ヒストグラムと同時に出力した乱数の統計量も表示される。採択率は、Metropolis-Hastings アルゴリズムの抽出率をいう。

図 3a の中の曲線は目的分布の密度関数を利用した理論値である。この場合少しずれているが、乱数の「Seed」を変えることによって分布が異なってくる。例として、図 3b に Seed = 2 の場合を示す。

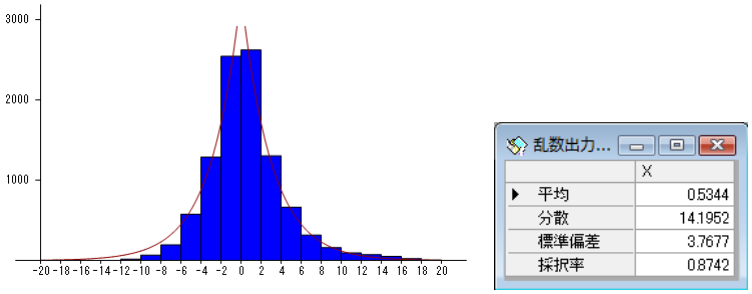


図 3b　乱数のヒストグラムと理論曲線（Seed=2）

ヒストグラムの階級幅は「x 分割」の数によって決まる。この場合、範囲が 40 で x 分割数が 20 であるので階級幅は 2 になっている。

密度関数の形は、「描画」ボタンで見ることができる。但し、１変量関数グラフのプログラムを利用するので、そのメニューが表示されるが、その中の「グラフ描画」ボタンをクリックすると図 4 のようなグラフが表示される。

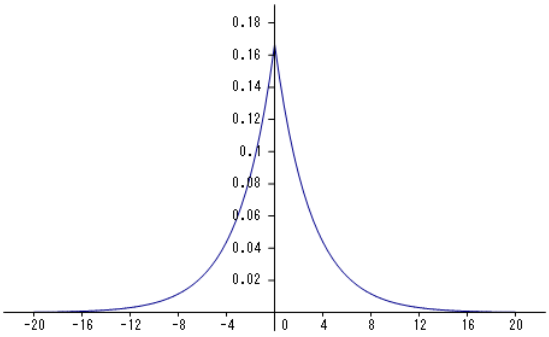


図 4　密度関数グラフ

密度関数から求められる、平均、分散、標準偏差は、「統計量」ボタンで図 5 のように表示される。

図 5　統計量結果

目的分布の関数形のみ分かって、スケールが不明の場合は、定数の部分に表示された値（1／面積）を掛けておけばよい。乱数発生はスケールにはよらないので、特に掛けておく必要もない。

提案分布については、酔歩乱数の場合、平均は 0 とし、標準偏差は目的分布のものより小さくしておくと無難である。提案分布の標準偏差を大きくして行くと乱数の尖度が小さくなる傾向があるので、適当な標準偏差を選ぶことは重要である。また独立連鎖の場合、提案分布の平均と標準偏差を目的分布に合わせておくと無難である。

以上のようにして求めた乱数は、データとしてグリッドに出力できる。予め複数行のグリッドを用意しておき、「出力列」コンボボックスで「範囲指定」を選び、列を選択して、「乱数グリッド出力」ボタンをクリックする。また、「出力列」で「新規追加」を選択すると、新しい列を追加して乱数を出力する。これは、メニュー「ツールーデータ発生」の乱数発生と同じである。

次に離散的な乱数発生について説明する。例えば「例」で、ポアソン分布を選択すると、「密度関数」テキストボックスには、密度関数＝ exp(-λ)*λ^x/fact(x) が表示され、右下の「離散」チェックボックスにチェックが入る。離散分布の場合は、この「離散」チェックボックスのチェックが重要である。密度関数にはパラメータλが含まれているが、利用者はこれを書き換えて適当な値にする。例えば、λを 3 とすると、exp(-3)*3^x/fact(x) となる。発生された最小値と最大値は「最小・最大」ボタンをクリックすることにより、0 と 9 であるから、「下限」を 0、「上限」を 10 にして、「ヒストグラム」ボタンをクリックすると図 6 のようになる。

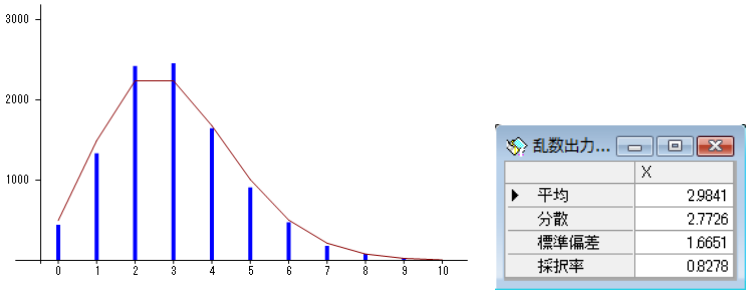


図 6　ポアソン分布

現在のバージョンでは、離散分布は 1 次元の場合だけに対応している。また、「描画」ボタンは離散分布に対応していない。

次に 2 次元の分布について見る。変数は x と y で与える。例として、密度関数のコンボボックスで 2 変量正規分布を選ぶと、以下のような 2 変量正規分布の密度関数の式が表示

される。

密度関数＝ 1/(2*pi*(1-r^2)^0.5)*exp(-(x^2-2*r*x*y+y^2)/2/(1-r^2))

ここで、r は相関係数を表す。例えば r を 0.5 と書き換えて、「描画」ボタンをクリックし、表示された 2 変量関数グラフのメニューで、そのまま「グラフ描画」ボタンをクリックすると、図 7 のような密度関数グラフが表示される。

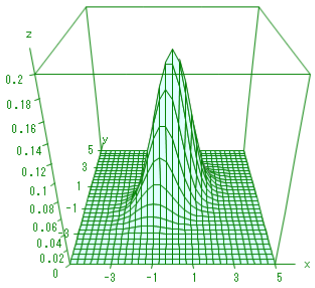


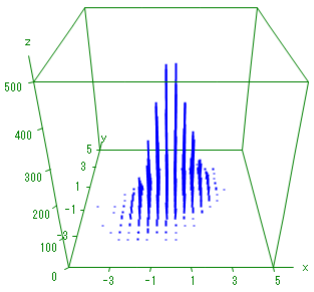
図 7　2 変量正規分布密度関数

次に、「統計量」ボタンをクリックすると、図 8 に示されるような結果が表示される。

面積(s)	1.0000
定数(1/s)	1.0000
X平均	0.0000
X分散	1.0000
X標準偏差	1.0000
Y平均	0.0000
Y分散	1.0000
Y標準偏差	1.0000
相関係数	0.5000

図 8　統計量結果

出力される乱数の分布を見るために「ヒストグラム」ボタンをクリックすると図 9 のような 2 変量ヒストグラムが表示される。



	X	Y
平均	0.0137	0.0115
分散	1.0023	0.9579
標準偏差	1.0012	0.9787
採択率	0.6672	0.6620

図 9　2 変量ヒストグラム

2 変量の場合のグリッドへの乱数出力は、2 列同時に出力されるので注意を要する。

14.3 マルコフ連鎖モンテカルロ法の理論

1) Metropolis-Hastings 法

時刻 t に値 x が確率 $\pi^{(t)}(x)$ で生じる、ある確率変数 X について、この値が、時刻 t と共に変化して行く過程 $x^{(1)},x^{(2)},\cdots,x^{(t)},\cdots$ を確率過程という。マルコフ連鎖は、この確率過程が時刻 t まで実現した後に、時刻 $t+1$ での値 $x^{(t+1)}$ の発生確率 $P(X=x^{(t+1)}\,|\,x^{(1)},\cdots,x^{(t)})$ が時刻 t の値 $x^{(t)}$ だけによって決まるものをいう。すなわち、

$$P(X=x^{(t+1)}\,|\,x^{(1)},\cdots,x^{(t)})=P(X=x^{(t+1)}\,|x^{(t)})$$

である。

$$p(x^{(t+1)}\,|x^{(t)})\equiv P(X=x^{(t+1)}\,|x^{(t)})$$

とすると、この $p(x^{(t+1)}\,|x^{(t)})$ は推移核と呼ばれる。値が離散的で有限個の場合、推移核はある有限な定数行列（推移行列）となる。マルコフ連鎖が既約的、正回帰的、かつ非周期的であるとき、エルゴード的であると言われ、以下の性質を満たすことが知られている。

$$\lim_{t\to\infty}\pi^{(t)}(x)=\pi(x)$$

ここに $\pi(x)$ はある不変分布である。即ち、どの状態から出発しても、$t\to\infty$ ではある状態 $\pi(x)$ に収束する。この状態を利用すると、以下の関係が成り立つことが分かる。

$$\pi(x^{(t+1)})=\int\pi(x^{(t)})\,p(x^{(t+1)}\,|\,x^{(t)})dx^{(t)}$$

マルコフ連鎖が不変分布になっているための十分条件は隣接する 2 つの時刻 $t,t+1$ に対して以下の詳細つり合い条件が成り立つことである。

$$\pi\left(x^{(t)}\right)p\left(x^{(t+1)}|x^{(t)}\right)=\pi\left(x^{(t+1)}\right)p\left(x^{(t)}|x^{(t+1)}\right)$$

我々はある提案分布により乱数を発生させ､ある条件に従ってこの詳細つり合い条件を満たすようにデータをサンプリングする。我々の提案分布の密度関数を $q(x_1\,|\,x_2)$ とすると、通常この分布は詳細つり合い条件を満たさない。

$$\pi(x^{(t)})q(x^{(t+1)}|x^{(t)})\neq\pi(x^{(t+1)})q(x^{(t)}|x^{(t+1)})$$

さて､ここで､推移核 $p\left(x|x'\right)$ をこの提案分布確率密度 $q\left(x|x'\right)$ と､ある確率 $\alpha\left(x|x'\right)$ を用いて以下のように表す。

$$p(x|x')=cq(x|x')\alpha(x|x')$$

ここに c は定数である。これは提案分布によって発生させた乱数を確率$\alpha\left(x|x'\right)$で選別して推移核の定数倍に一致させようとするものである。

この関係を詳細つり合い条件に代入すると定数 c の自由度を残して以下となる。

$$\pi(x^{(t)})q(x^{(t+1)}|x^{(t)})\alpha(x^{(t+1)}|x^{(t)})=\pi(x^{(t+1)})q(x^{(t)}|x^{(t+1)})\alpha(x^{(t)}|x^{(t+1)})$$

確率の$\alpha\left(x|x'\right)$値は 0 から 1 の範囲で、以下のように決めれば良いことが分かる。

$\pi(x^{(t)})q(x^{(t+1)}|x^{(t)})=\pi(x^{(t+1)})q(x^{(t)}|x^{(t+1)})$ のとき、

$$\alpha(x^{(t+1)}|x^{(t)})=1,\ \ \alpha(x^{(t)}|x^{(t+1)})=1$$

$0\le\pi(x^{(t)})q(x^{(t+1)}|x^{(t)})<\pi(x^{(t+1)})q(x^{(t)}|x^{(t+1)})$ のとき、

$$\alpha(x^{(t+1)}|x^{(t)})=1,\ \ \alpha(x^{(t)}|x^{(t+1)})=\frac{\pi(x^{(t)})q(x^{(t+1)}|x^{(t)})}{\pi(x^{(t+1)})q(x^{(t)}|x^{(t+1)})}<1$$

$\pi(x^{(t)})q(x^{(t+1)}|x^{(t)})>\pi(x^{(t+1)})q(x^{(t)}|x^{(t+1)})\geq 0$ のとき、

$$\alpha(x^{(t+1)}|x^{(t)})=\frac{\pi(x^{(t+1)})q(x^{(t)}|x^{(t+1)})}{\pi(x^{(t)})q(x^{(t+1)}|x^{(t)})}<1,\ \ \alpha(x^{(t)}|x^{(t+1)})=1$$

これを $\alpha(x^{(t+1)}|x^{(t)})$ についてまとめると以下となる。

$$\alpha(x^{(t+1)}|x^{(t)})=\begin{cases}\min\left[\dfrac{\pi(x^{(t+1)})q(x^{(t)}|x^{(t+1)})}{\pi(x^{(t)})q(x^{(t+1)}|x^{(t)})},1\right] & \text{分母}>0\\ 1 & \text{分母}=0\end{cases}$$

即ち、乱数を提案分布により発生させ、確率 $\alpha(x^{(t+1)}|x^{(t)})$ によって抽出すれば、目的の分布に従う乱数を得ることができる。この方法を Metropolis - Hastings アルゴリズムという。

さて、任意の密度関数 $\pi(x)$ からの乱数を得るために、提案分布として我々のプログラムでは正規分布を考える。その確率密度関数は以下である。

$$q(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

この乱数の発生法について､酔歩的に前時刻の位置を中心として発生させる場合と前回とは全く独立に発生させる場合を考える。前者を酔歩連鎖、後者を独立連鎖と呼ぶ。

酔歩連鎖では、状態 x' から状態 x への推移は、x' を中心として上の正規分布を発生させるので、 $q(x\,|\,x')=q(x-x')$ となり、条件付き確率は具体的に以下となる。

$$q(x^{(t)}\,|\,x^{(t+1)})=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x^{(t)}-x^{(t+1)}-\mu)^2}{2\sigma^2}}$$

$$q(x^{(t+1)}\,|\,x^{(t)})=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x^{(t+1)}-x^{(t)}-\mu)^2}{2\sigma^2}}$$

ここで、 $\mu=0$ の場合は $q(x^{(t)}\,|\,x^{(t+1)})=q(x^{(t+1)}\,|\,x^{(t)})$ となることから、確率を決める式は以下となる。

$$\frac{\pi(x^{(t+1)})q(x^{(t)}|x^{(t+1)})}{\pi(x^{(t)})q(x^{(t+1)}|x^{(t)})}=\frac{\pi(x^{(t+1)})}{\pi(x^{(t)})}$$

次に独立連鎖の場合は、これまでの位置に関係なく、上の乱数を発生させるので、

$$q(x^{(t)}\,|\,x^{(t+1)})=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x^{(t)}-\mu)^2}{2\sigma^2}}$$

$$q(x^{(t+1)}\,|\,x^{(t)})=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x^{(t+1)}-\mu)^2}{2\sigma^2}}$$

となり、確率を決める式は以下となる

$$\frac{\pi(x^{(t+1)})q(x^{(t)}|x^{(t+1)})}{\pi(x^{(t)})q(x^{(t+1)}|x^{(t)})}=\frac{\pi(x^{(t+1)})e^{-\frac{(x^{(t)}-\mu)^2}{2\sigma^2}}}{\pi(x^{(t)})e^{-\frac{(x^{(t+1)}-\mu)^2}{2\sigma^2}}}$$

この関係は、離散分布の場合にも適用され、我々は正規分布から得られた値を、小数点以下１桁目の四捨五入により整数化して、提案分布として利用している。

次にこれを変数が複数ある場合に拡張する。時系列データを $x_i^{(t)}$ とし、提案分布として我々は独立な正規分布を考える。

$$q(x_1,\cdots,x_n)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}}$$

n 変数の場合も、1 変数の場合と同様に、酔歩連鎖と独立連鎖を考える。特に酔歩連鎖では $\mu_i=0\ \ (i=1,\cdots,n)$ とする。

提案分布からの抽出確率は以下となる。

$$\alpha=\begin{cases}min\left[\dfrac{\pi(\cdots,x_i^{(t+1)},x_{i+1}^{(t)},\cdots)q(x_i^{(t)}|\cdots,x_i^{(t+1)},x_{i+1}^{(t)},\cdots)}{\pi(\cdots,x_{i-1}^{(t+1)},x_i^{(t)},\cdots)q(x_i^{(t+1)}|\cdots,x_{i-1}^{(t+1)},x_i^{(t)},\cdots)},1\right] & \text{分母}>0\\ 1 & \text{分母}=0\end{cases}.$$

ここで、変数の順番を変えて次の時点の乱数を求めたとしても、抽出された乱数の分布には影響がないことが知られている。

具体的に提案分布として上の独立な正規分布を考えると、酔歩連鎖の場合、

$$\begin{aligned}&q(x_i^{(t+1)}|x_1^{(t+1)},\cdots,x_{i-1}^{(t+1)},x_i^{(t)},\cdots,x_n^{(t)})\\&=\prod_{j=1}^{i-1}\frac{1}{\sqrt{2\pi}\sigma_j}e^{-\frac{x_j^{(t+1)2}}{2\sigma_j^2}}\times\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(x_i^{(t+1)}-x_i^{(t+1)})^2}{2\sigma_i^2}}\times\prod_{k=i+1}^{n}\frac{1}{\sqrt{2\pi}\sigma_k}e^{-\frac{x_k^{(t)2}}{2\sigma_k^2}}\\&=q(x_i^{(t)}|x_1^{(t+1)},\cdots,x_i^{(t+1)},x_{i+1}^{(t)},\cdots,x_n^{(t)})\end{aligned}$$

より、以下となる。

$$\begin{aligned}&\alpha(x_i^{(t+1)}|x_1^{(t+1)},\cdots,x_{i-1}^{(t+1)},x_{i+1}^{(t)},\cdots,x_n^{(t)})\\&=\begin{cases}min\left[\dfrac{\pi(x_1^{(t+1)},\cdots,x_i^{(t+1)},x_{i+1}^{(t)},\cdots,x_n^{(t)})}{\pi(x_1^{(t+1)},\cdots,x_{i-1}^{(t+1)},x_i^{(t)},\cdots,x_n^{(t)})},1\right] & \text{分母}>0\\ 1 & \text{分母}=0\end{cases}\end{aligned}$$

独立連鎖の場合は同様であるので省略する。

2) Hamiltonian モンテカルロ法による乱数発生

ここでは新しく導入した Hamiltonian （ハミルトニアン）モンテカルロ法について説明する。MCMC による乱数発生では、初期値の設定は重要である。Metropolis-Hastings（MH）法の酔歩乱数では、正規分布を使って最尤値に 1 歩ずつ近づけていくために、初期値が最

尤値から離れた位置だと大きな標準偏差が必要である。しかし、最尤値に近いところで良い精度を出そうとすると適当な大きさの標準偏差が必要となる。これらの相反する条件を解決する手法として期待されるのが Hamiltonian Monte Carlo（HMC）法である。

HMC 法は、変数を $q_\alpha\ (\alpha=1,2,\cdots,n)$ とした目的の分布と、変数を $p_\alpha\ (\alpha=1,2,\cdots,n)$ とした独立な標準正規分布を合成した分布の密度関数を、力学のハミルトニアン H の関数式 e^{-H} とみなし、ハミルトニアン（エネルギー）の保存則を利用して変数 q_α を決めて行く方法である。

今、発生させたい乱数の密度関数を $f(\mathbf{q})$ とし、それに独立な標準正規分布の密度関数を $g(\mathbf{p})=1/(2\pi)^{n/2}\exp\left(-\sum p_\alpha^2/2\right)$ とすると、合成関数 $f(\mathbf{q})g(\mathbf{p})$ は以下のようになる。

$$f(\mathbf{q})g(\mathbf{p})\sim\exp\left[-h(\mathbf{q})-\sum p_\alpha^2/2\right]\equiv\exp[-H(\mathbf{q},\mathbf{p})]$$

ここに、$h(\mathbf{q})=-\log f(\mathbf{p})$ はポテンシャルエネルギー、$\sum p_\alpha^2/2$ は質点の運動エネルギーに相当する。但し、質点の質量はすべて 1 としている。このハミルトニアンの元で、運動方程式は以下となる。

$$\frac{dp_\alpha}{dt}=-\frac{\partial H}{\partial q_\alpha},\qquad \frac{dq_\alpha}{dt}=\frac{\partial H}{\partial p_\alpha}=p_\alpha$$

この運動に際して、ハミルトニアンは以下のように不変である。

$$\frac{dH}{dt}=\sum_{\alpha=1}^{n}\left[\frac{dq_\alpha}{dt}\frac{\partial H}{\partial q_\alpha}+\frac{dp_\alpha}{dt}\frac{\partial H}{\partial p_\alpha}\right]=\sum_{\alpha=1}^{n}\left[-\frac{dq_\alpha}{dt}\frac{dp_\alpha}{dt}+\frac{dp_\alpha}{dt}\frac{dq_\alpha}{dt}\right]=0$$

ハミルトニアンの不変性から、２つの時点 $t,t'\ (t<t')$ で関数間の関係は以下となる。

$$f(\mathbf{q}')g(\mathbf{p}')=f(\mathbf{q})g(\mathbf{p})$$

ここに、上式では以下のように時間 t',t が略されている。

$$\mathbf{q}'=\mathbf{q}(t'),\mathbf{p}'=\mathbf{p}(t'),\quad \mathbf{q}=\mathbf{q}(t),\mathbf{p}=\mathbf{p}(t)$$

我々は変数 $\mathbf{q}$ を初期値として与え、独立な n 個の正規乱数を発生させ、それを変数 $\mathbf{p}$ とする。これらを使ってハミルトンの運動方程式を解き、新しい変数 $\mathbf{q}',\mathbf{p}'$ を求める。その際、位置 $\mathbf{q}$ で $\mathbf{p}\pm\mathbf{d}/2$ の乱数を発生させる確率は $g(\mathbf{p})\mathrm{d}^n$ であるため、位置 $\mathbf{q}'$ の近傍に到達する確率も $g(\mathbf{p})\mathrm{d}^n$ である。またこの過程を逆にたどることを考えると、位置 $\mathbf{q}'$ で $\mathbf{p}'\pm\mathbf{d}/2$ の乱数を発生させ、位置 $\mathbf{q}$ の近傍に到達する確率は $g(\mathbf{p}')\mathrm{d}^n$ であるため、位置 $\mathbf{q}$ から位置 $\mathbf{q}'$ に到達する確率とその逆の確率の比は $g(\mathbf{p}):\mathrm{g}(\mathbf{p}')$ となる。ここで、上に述べた関係 $f(\mathbf{q}')g(\mathbf{p}')=f(\mathbf{q})g(\mathbf{p})$ を使うと、この比は $f(\mathbf{q}'):f(\mathbf{q})$ となり、到達する位置の発生させたい密度関数の大きさに比例することになる。これがすべての２つの位置の間で成り立っていることから、$\mathbf{q}$ の値が得られる確率は $f(\mathbf{q})$ に比例する。これは密度関数 $f(\mathbf{q})$ で乱数が発生したことになる。

この手法はマルコフ連鎖を意識して利用しているわけではないが、関連を考えてみよう。マルコフ連鎖では 1 つの状態 $(\mathbf{p},\mathbf{q})$ から他の状態 $(\mathbf{p}',\mathbf{q}')$ に推移する場合、推移は推移核 $S(\mathbf{q}',\mathbf{p}'|\mathbf{q},\mathbf{p})$ を用いて以下の形で表される。

$$f(\mathbf{q}')g(\mathbf{p}') = S(\mathbf{q}',\mathbf{p}'|\mathbf{q},\mathbf{p})f(\mathbf{q})g(\mathbf{p})$$

運動が可逆過程であることから、推移も可逆的となり、

$$f(\mathbf{q})g(\mathbf{p}) = S(\mathbf{q},\mathbf{p}|\mathbf{q}',\mathbf{p}')f(\mathbf{q}')g(\mathbf{p}')$$

これらの関係より、

$$S(\mathbf{q}',\mathbf{p}'|\mathbf{q},\mathbf{p}) = S(\mathbf{q},\mathbf{p}|\mathbf{q}',\mathbf{p}') = 1 \quad \text{（確率 1 でこの推移が起こる）}$$

$$S(\mathbf{q}',\mathbf{p}'|\mathbf{q},\mathbf{p})f(\mathbf{q})g(\mathbf{p}) = S(\mathbf{q},\mathbf{p}|\mathbf{q}',\mathbf{p}')f(\mathbf{q}')g(\mathbf{p}')$$

となり、詳細つり合い条件は自動的に満たされる。

この推移を実際に計算するには、オイラー法を拡張したリープ・フロッグ法を用いる。その際、微分を差分で置き換えるため誤差が生じ、以下のような関係になるとする。

$$f(\mathbf{q}')g(\mathbf{p}') = r\,f(\mathbf{q})g(\mathbf{p})$$

これを補正するために MH 法の考え方を利用する。

$$\alpha(\mathbf{q}',\mathbf{p}'|\mathbf{q},\mathbf{p}) = \begin{cases} \min\left[\dfrac{f(\mathbf{q}')g(\mathbf{p}')}{f(\mathbf{q})g(\mathbf{p})} = r\ ,1\right] & \text{分母} > 0 \\ 1 & \text{分母} = 0 \end{cases}$$

即ち、リープ・フロッグ法で新しく得られた変数 $\mathbf{q}'$ については、上に与えた確率 $\alpha(\mathbf{q}',\mathbf{p}'|\mathbf{q},\mathbf{p})$ で採択の可否を決める。これは 1 に近い値のため、採択率はかなり高くなる。

最後に、オイラー法とリープ・フロッグ法の計算法を与えておく。

n 次元オイラー法

$$p_\alpha(t+1) = p_\alpha(t) - \varepsilon\,\partial h/\partial q_\alpha$$

$$q_\alpha(t+1) = q_\alpha(t) + \varepsilon\,p_\alpha(t)$$

n 次元リープ・フロッグ法

$$p_\alpha(2t+1) = p_\alpha(2t) - (\varepsilon/2)\,dh/dq_\alpha\big|_{q(2t)}$$

$$q_\alpha(2t+2) = p_\alpha(2t) + \varepsilon\,p_\alpha(2t+1)$$

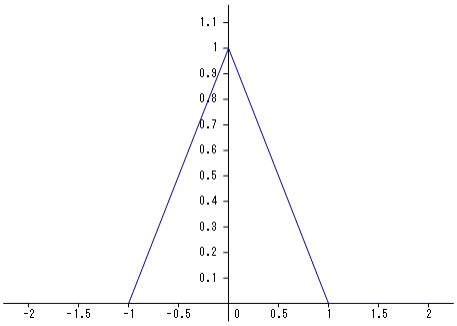
$$p_\alpha(2t+2) = p_\alpha(2t+1) - (\varepsilon/2)\,dh/dq_\alpha\big|_{q(2t+2)}$$

実際の計算での HMC 法の使い勝手はどうであろうか。標準正規乱数を発生させるごとにリープ・フロッグ法を用いるため、計算量（特に微分の部分）がかなり多くなる。採択率は上がるが、その分計算量が増えるため、計算時間は MH 法に比べて長くなっている。しかし、乱数の精度から見ると改良されているのではないかと思われる。

次に初期値が最尤値から離れている場合の収束性について、これまでの計算では、遠くの最尤値まで速く収束するようには感じられない。むしろ初期値に対して最尤値が離れている場合は、密度関数が計算誤差で 0 となってしまうところが問題のように思われる。これについては MH 法も HMC 法もあまり変わらないように思う。

問題１

以下のような密度関数を持つ乱数を MH 法と Hamiltonian 法で 10,000 個の乱数を発生させ比較せよ。但し、乱数の Seed は 1、ヒストグラムの下限:-1、上限:1、分割幅:0.2 として、後はデフォルトのままとする。



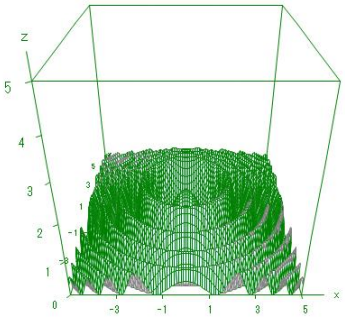
問題２

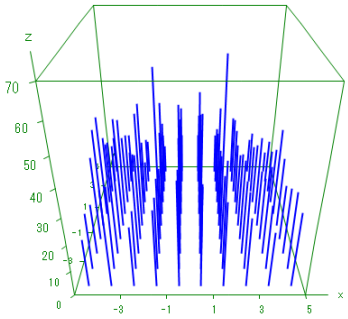
MCMC によって以下の 2 次元の密度関数に従う乱数発生を試みる。乱数の Seed は 1 とすること。

$$f(x) \sim (1-\sin(x^2+y^2))\exp\left(-\sqrt{x^2+y^2}/5\right)$$

１）密度関数を描画せよ（左図)。但し、規格化は不要である。
範囲は x:[-5,5]，y[-5,5]，z:[0,5] とし、区間分割数は「100」に変えよ。

２）MH 乱数のヒストグラムを描画せよ（右図)。範囲は x:[-5,5]，y[-5,5]。これでは、はっきりした結果は見えて来ない。





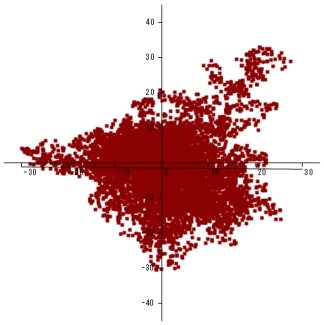
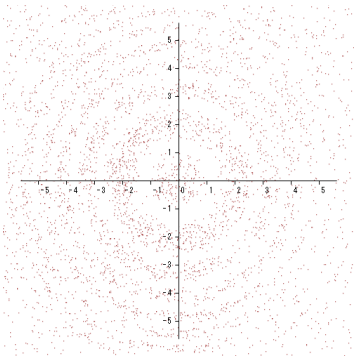
３）以下の理論値を求めよ。

x 平均　［　　］　y 平均　［　　］

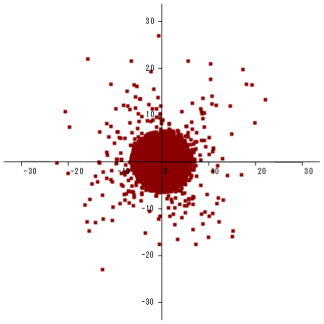
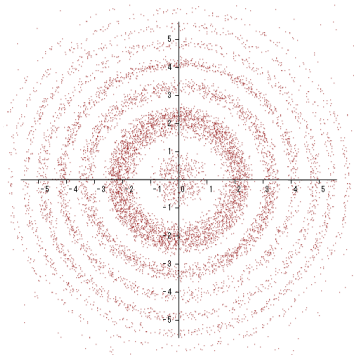
x 標準偏差［　　］　y 標準偏差［　　］

４）乱数をグリッドに 10,000 個発生し（棄却数 1,000)、メニュー［分析－基本統計－集計グラフ－2 次元グラフ］の中にある「散布図」で描画せよ（左図)。

５）散布図のメニュー［設定ーポイントサイズ］を 0.2 にし、範囲を x:[-5,5], y[-5,5]とする（右図）。うっすらと縞模様が出ている。

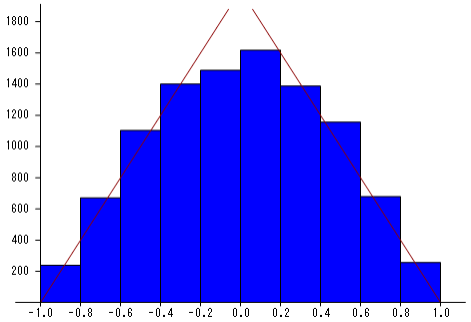
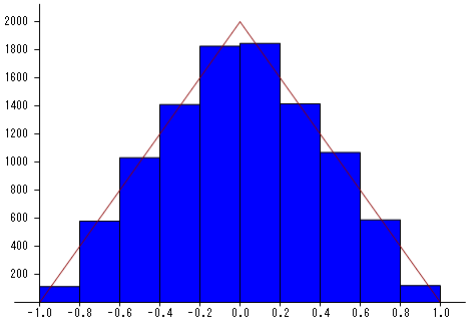
６）問題４）、５）の処理を時間がかかるが、HMC で行うとどうなるか。

これを見ると、HMC は MH に比べて中央に集中している。時間はかかるが、良い精度を持っているのが分かる。

問題１解答

密度関数を　(1-abs(x))*theta(1-abs(x))　として、MH 法と Hamiltonian 法を試す。

MH 法　　Hamiltonian 法

これを見ると、Hamiltonian 法の方が精度よく生成されていることが分かる。

【動画 g171119_1.mp4】

問題２解答

密度関数を　(1-sin(x^2+y^2))*exp(-(x^2+y^2)^0.5/5)　とする。

３）以下の理論値を求めよ。

x 平均　　［-1.449］　　y 平均　　［ 0.623 ］

x 標準偏差［ 8.620］　　y 標準偏差［ 8.229 ］

【動画 g171119_2.mp4】

１５．分布の検定

15.1　分布の検定とは

乱数データが与えられている場合、それが本当に自分が求める分布に従っているかどうか調べることは重要である。ここではこの分布の検定法について説明する。College Analysis でメニュー［分析－基本統計－ユーティリティー分布の検定］を選択すると図 1 のような実行画面が表示される。

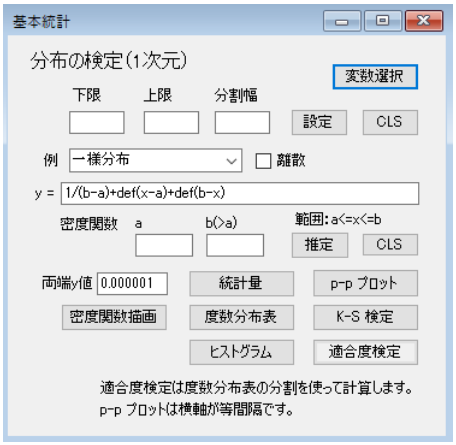


図 1　実行画面

データは縦 1 列でグリッドエディタに入力されたものを使う。「変数選択」で、検定するデータの変数を 1 つ選択し、メニューの「y =」テキストボックスに密度関数の形を数式で入力する。よく知られた分布の場合は、上の「例」コンボボックスから図 2a のように選ぶ。ここでは χ^2 分布を選択している。さらに、「設定」ボタンで「下限」、「上限」を変更し、「推定」ボタンでパラメータを推定する。ここでは図 2b のように、χ^2 分布として自由度 3 と推定している。

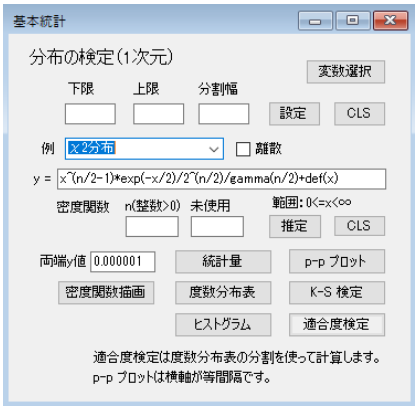


図 2a　密度関数の指定

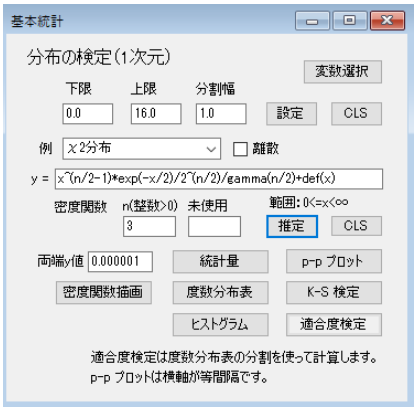


図 2b　パラメータと下限・上限の指定

連続パラメータの場合は、推定値がパラメータの欄に設定され、推定値とその標準偏差などが別途グリッド出力される。

密度関数の性質を見るために、「統計量」ボタンをクリックすると図 3 の結果を得る。

	データ	理論値
最小(全確率)	0.0300	1.0000
最大(1/全確率)	15.0900	1.0000
平均	3.0830	3.0000
分散	6.2460	6.0000
標準偏差	2.4992	2.4495

図 3　統計量

これはデータを用いた統計量と統計量の理論値との比較である。但し、最小（全確率）と最大（1/全確率）は、データでは最小と最大、理論値では全確率と 1/全確率を表す。

次に「度数分布表」ボタンをクリックするとデータと理論値の度数分布の比較が、図 4 のように表示される。

	度数	比率	理論度数	理論比率
領域なし	0	0.000	0.00	0.000
0.0<=x<1.0	194	0.194	198.72	0.199
1.0<=x<2.0	216	0.216	228.85	0.229
2.0<=x<3.0	177	0.177	180.78	0.181
3.0<=x<4.0	134	0.134	130.16	0.130
4.0<=x<5.0	106	0.106	89.67	0.090
5.0<=x<6.0	63	0.063	60.19	0.060
6.0<=x<7.0	32	0.032	39.71	0.040
7.0<=x<8.0	28	0.028	25.89	0.026
8.0<=x<9.0	17	0.017	16.72	0.017
9.0<=x<10.0	11	0.011	10.72	0.011
10.0<=x<11.0	9	0.009	6.84	0.007
11.0<=x<12.0	3	0.003	4.34	0.004
12.0<=x<13.0	2	0.002	2.75	0.003
13.0<=x<14.0	3	0.003	1.73	0.002
14.0<=x<15.0	4	0.004	1.09	0.001
15.0<=x<16.0	1	0.001	0.68	0.001
16.0<=x<30.0	0	0.000	1.13	0.001
合計	1000	1.000	999.97	1.000

図 4　連続分布の度数分布表

合計を除く一番上と一番下は、「下限」と「上限」に指定された領域以外についての度数と比率の和である。ここで領域外の範囲は、密度関数の高さが分析メニューの「両端 y 値」で指定された値より小さくなった点までを計算する。図 4 では「16.0<=x<30」の 30 がその点である。

次に、分析メニューで「ヒストグラム」をクリックすると、上の度数分布表の「下限」と「上限」の範囲内のデータと理論的な密度曲線が図 5 のように表示される。

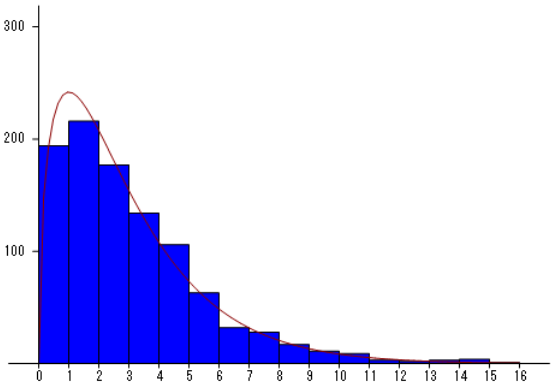


図 5　連続分布のヒストグラム

度数分布表やヒストグラムにより、定性的な分布の検討ができる。

次にもう少し、分布との一致を見易くするために、分析メニューの「p-p プロット」をク

リックする。結果は図 6 のようになる。

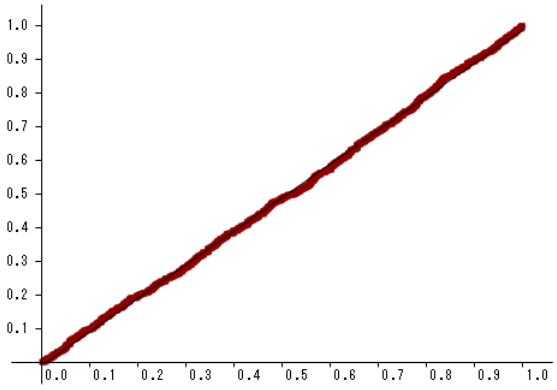


図 6　p-p プロット

これは、データと理論値の適合性を見るための直線で、適合が良ければプロットはこの図のように直線状に並ぶ。これは正規性の検定の「正規確率紙」の方法（一般に q-q プロットと呼ぶ）に類似するもので、縦軸が累積確率、横軸が理論的な確率である。（現在のバージョンでは、縦軸と横軸の役割が逆になっている。）

p-p プロットを数値的に検定する方法がコルモゴロフースミルノフ（Kolmogorov-Smirnov）検定である。これは略して、K-S 検定と呼ばれる。この検定はプロットがこの直線から最大どれ位離れているかで適合の検定確率を求める。分析メニューで「K-S 検定」ボタンをクリックすると図 7 のような結果が得られる。

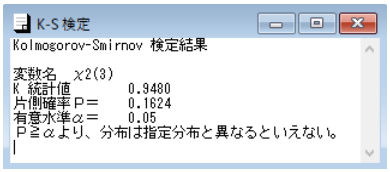


図 7　K-S 検定結果

また分布の検定には、図 4 の度数分布表をもとに、度数分布が理論比率に合っているかどうかを調べる適合度検定がある。これは分析メニューの「適合度検定」ボタンをクリックして得られる。分割は、度数分布表で与えられる分割を利用する。但し、理論比率が 0 の部分は分析から除外する。結果を図 8 に示す。

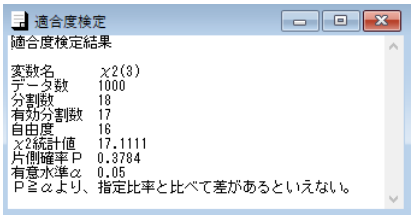


図 8　適合度検定結果

この適合度検定は離散的な分布に対しても適用できる。分析メニューの離散チェックボックスにチェックを入れた後に「度数分布表」ボタンをクリックして表示される、λ=4 のポアソン分布に対する度数分布表を図 9 に示す。

度数分布表 Poison 4

	度数	比率	理論度数	理論比率
-1<=x<=-1	0	0.0000	0.00	0.0000
x=0	22	0.0220	18.32	0.0183
x=1	60	0.0600	73.26	0.0733
x=2	142	0.1420	146.53	0.1465
x=3	179	0.1790	195.37	0.1954
x=4	221	0.2210	195.37	0.1954
x=5	156	0.1560	156.29	0.1563
x=6	97	0.0970	104.20	0.1042
x=7	55	0.0550	59.54	0.0595
x=8	37	0.0370	29.77	0.0298
x=9	19	0.0190	13.23	0.0132
x=10	8	0.0080	5.29	0.0053
11<=x<=17	4	0.0040	2.84	0.0028
合計	1000	1.0000	1000.00	1.0000

図 9 離散分布の度数分布表

これを「ヒストグラム」で表わすと図 10 のようになる。

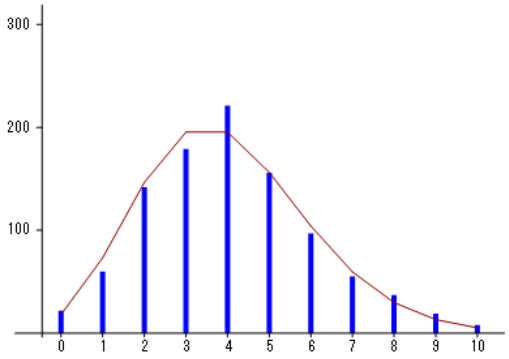


図 10 離散分布のヒストグラム

この乱数について「適合度検定」を実行すると図 11 のような結果となる。

適合度検定

適合度検定結果

変数名	Poison 4
データ数	1000
分割数	13
有効分割数	12
自由度	11
χ2統計値	14.99022
片側確率 P	0.18295
有意水準α	0.05

P≧αより、指定比率と比べて差があるといえない。

図 11 適合度検定結果

最後に、連続分布の場合は、「密度関数描画」ボタンで、関数描画用のメニューが表示され、関数グラフを描くことができる。

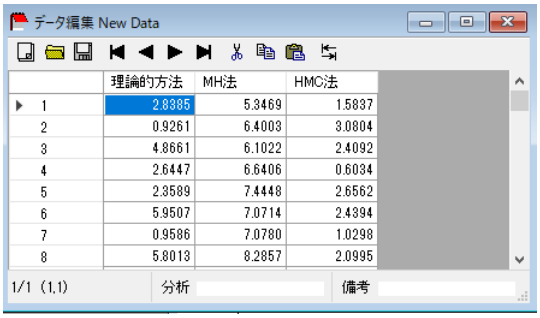
仮説検定を利用する場合、検定結果から、分布と異なることは示されるが、指定された分布になるという保証はない。特に、データ数が少ない場合には、有意差を見出すことが困難なため、注意を要する。また、連続分布の場合、分割数をいくつにするのか、どこに分割の境界を持ってくるのかで、検定結果が変わる場合もある。いろいろな場合で試して、総合的に確信を得る以外に方法はないのではなかろうか。

【動画 g171119_3.mp4】

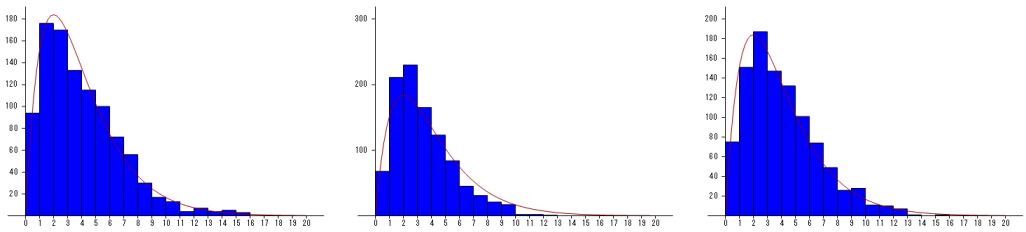
問題

1）自由度 4 の$\chi 2$ 分布のデータを［分析－基本統計－ユーティリティーデータ生成］で 1000 個発生させよ。これは理論的な方法である。

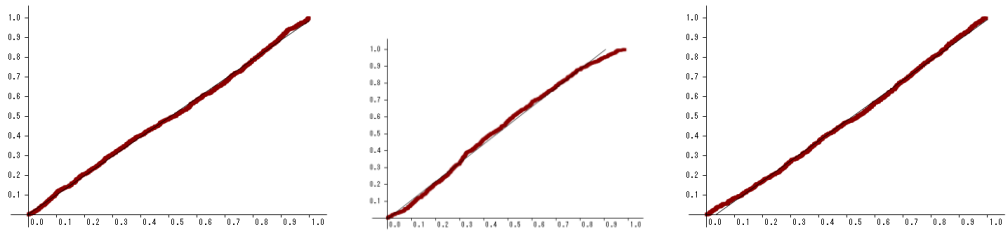
2）自由度 4 の$\chi 2$ 分布のデータを［分析－基本統計－ユーティリティーMCMC 乱数発生］の MH 法と HMC 法で 1000 個発生させよ。



3）データの上限を 20 として 3 つの方法のヒストグラムを描け。



4）3 つの方法の p-p プロットを描け。



5）K-S 検定で調べた検定確率を書け。

理論的方法［　　］　MH 法［　　］　HMC 法［　　］

6）適合度検定で調べた検定確率を書け。

理論的方法［　　］　MH 法［　　］　HMC 法［　　］

問題解答（乱数発生法の精度がよく分かる）

5）K-S 検定で調べた検定確率を書け。

理論的方法［0.1805］　MH 法［0.0000］　HMC 法［0.0231］

6）適合度検定で調べた検定確率を書け。

理論的方法［0.5241］　MH 法［0.0000］　HMC 法［0.2388］

15.2　パラメータの最尤推定法の理論

得られたデータ $x_\lambda\ (\lambda=1,\cdots,N)$ が、特定の分布に従うかどうかを調べる際、分布のパラメータが既知であるとは限らない。そのため、多くの場合、与えられたデータを用いて各種分布のパラメータを推定し、その下で検定の問題を考えることになると思われる。そこで、前節図 1 の分析実行画面に、パラメータを自動的に推定する機能を加えた。分布を選んで「推定」ボタンをクリックすると左のテキストボックスに推定値が表示される。

ここではまず簡単に最尤法の推定について解説し､分布毎にパラメータを推定するための方法を具体的に与えておく。

簡単な最尤法の考え方

準備

$$0=\frac{\partial^2}{\partial\beta^2}\prod_{i=1}^{N}\int f(x_i\mid\beta)\mathbf{dx}=\frac{\partial^2}{\partial\beta^2}\int\prod_{i=1}^{N}f(x_i\mid\beta)\mathbf{dx}=\frac{\partial^2}{\partial\beta^2}\int\exp\left[\sum_{i=1}^{N}\log f(x_i\mid\beta)\right]\mathbf{dx}$$

$$=\frac{\partial}{\partial\beta}\int\exp\left[\sum_{j=1}^{N}\log f(x_j\mid\beta)\right]\frac{\partial}{\partial\beta}\sum_{i=1}^{N}\log f(x_i\mid\beta)\mathbf{dx}$$

$$=\frac{\partial}{\partial\beta}\int\prod_{j=1}^{N}f(x_j\mid\beta)\frac{\partial}{\partial\beta}\sum_{i=1}^{N}\log f(x_i\mid\beta)\mathbf{dx}$$

$$=\int\prod_{j=1}^{N}f(x_j\mid\beta)\left[\frac{\partial}{\partial\beta}\sum_{i=1}^{N}\log f(x_i\mid\beta)\right]^2\mathbf{dx}+\int\prod_{j=1}^{N}f(x_j\mid\beta)\frac{\partial^2}{\partial\beta^2}\sum_{i=1}^{N}\log f(x_i\mid\beta)\mathbf{dx}$$

これを書き換えて、

$$E\left[\left\{\frac{\partial}{\partial\beta}\sum_{i=1}^{N}\log f(x_i\mid\beta)\right\}^2\right]=-E\left[\frac{\partial^2}{\partial\beta^2}\sum_{i=1}^{N}\log f(x_i\mid\beta)\right]$$

最尤法

尤度関数：　$L=\prod_{i=1}^{N}f(x_i\mid\beta)$

対数尤度関数：$\log L=\sum_{i=1}^{N}\log f(x_i\mid\beta)$

$\frac{\partial}{\partial\beta}\log L=\sum_{i=1}^{N}\frac{\partial}{\partial\beta}\log f(x_i\mid\beta)=0$　より、パラメータの推定を行う。

一方

$$\frac{\partial}{\partial\beta}\log L\simeq\frac{\partial}{\partial\beta}\log L_{\beta_0}+(\beta-\beta_0)\frac{\partial^2}{\partial\beta^2}\log L_{\beta_0}=0$$

$$\beta-\beta_0\simeq-\left(\frac{\partial^2}{\partial\beta^2}\log L_{\beta_0}\right)^{-1}\frac{\partial}{\partial\beta}\log L_{\beta_0}$$

ここで、

$$U_{\beta_0} \equiv \frac{\partial}{\partial \beta} \log L_{\beta_0}, \quad \mathfrak{I}_{\beta_0} \equiv -\frac{\partial^2}{\partial \beta^2} \log L_{\beta_0} \quad \text{（情報行列）}$$

とすると求める値は、以下の関係を繰り返し使って、求めることになる。

$$\beta_1 = \beta_0 + \mathfrak{I}_{\beta_0}^{-1} U_{\beta_0} \to \hat{\beta}$$

準備で求めた式より、$V[U] = E[U^2] \simeq \mathfrak{I}_{\hat{\beta}}$ であるから、上の式を元に、

$$V[\beta] \simeq \mathfrak{I}_{\hat{\beta}}^{-1} V[U] \mathfrak{I}_{\hat{\beta}}^{-1} \simeq \mathfrak{I}_{\hat{\beta}}^{-1} \mathfrak{I}_{\hat{\beta}} \mathfrak{I}_{\hat{\beta}}^{-1} = \mathfrak{I}_{\hat{\beta}}^{-1}$$

となり、推定値の分散が求められる。

正規分布 $(-\infty < x < \infty)$

密度関数： $f(x \mid \mu, \sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp[-(x-\mu)^2 / 2\sigma^2]$

尤度関数： $L = \dfrac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\dfrac{1}{2\sigma^2} \sum_{\lambda=1}^{N} (x_\lambda - \mu)^2 \right]$

対数尤度： $\log L = -\dfrac{1}{2\sigma^2} \sum_{\lambda=1}^{N} (x_\lambda - \mu)^2 - \dfrac{N}{2} \log(2\pi\sigma^2)$

$$\partial \log L / \partial \mu = \frac{1}{\sigma^2} \sum_{\lambda=1}^{N} (x_\lambda - \mu) = 0 \qquad \mu = \frac{1}{N} \sum_{\lambda=1}^{N} x_\lambda$$

$$\partial \log L / \partial \sigma^2 = \frac{1}{2\sigma^4} \sum_{\lambda=1}^{N} (x_\lambda - \mu)^2 - \frac{N}{2\sigma^2} = 0 \qquad \sigma^2 = \frac{1}{N} \sum_{\lambda=1}^{N} (x_\lambda - \mu)^2$$

以上で解析的に求めることが可能であるが、プログラムでは練習問題としてニュートン・ラフソン法を用いて計算を試している。ニュートン・ラフソン法の一般的な方法は以下である。

スコアベクトル $\mathbf{U}$ と情報行列 $\mathfrak{I}$

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \qquad \mathbf{U} = \begin{pmatrix} \partial \log L / \partial \mu \\ \partial \log L / \partial \sigma^2 \end{pmatrix}, \qquad \mathfrak{I} = -\begin{pmatrix} \partial^2 \log L / \partial \mu^2 & \partial^2 \log L / \partial \mu \partial \sigma^2 \\ \partial^2 \log L / \partial \mu \partial \sigma^2 & \partial^2 \log L / \partial (\sigma^2)^2 \end{pmatrix}$$

最尤推定の計算法（ニュートン・ラフソン法）

$$\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} + (\mathfrak{I}^{(m-1)})^{-1} \mathbf{U}^{(m-1)}$$

最尤推定法の結果

$$\mathbf{b} = \boldsymbol{\beta} + \mathfrak{I}^{-1} \mathbf{U} \sim N(\boldsymbol{\beta}, \mathfrak{I}^{-1})$$

情報行列を計算するために対数尤度の 2 階微分を与えておく。

$$\partial^2 \log L / \partial \mu^2 = -\frac{N}{\sigma^2}$$

$$\partial^2 \log L / \partial \mu \partial \sigma^2 = -\frac{1}{\sigma^4} \sum_{\lambda=1}^{N} (x_\lambda - \mu) \to 0$$

$$\partial^2 \log L / \partial (\sigma^2)^2 = -\frac{1}{\sigma^6} \sum_{\lambda=1}^{N} (x_\lambda - \mu)^2 + \frac{N}{2\sigma^4} \to -\frac{N}{2\sigma^4}$$

初期値は $\mu_0 = 0,\ \sigma_0^2 = 1$ を用いている。

ここで情報行列の逆行列$\mathfrak{I}^{-1}$の対角成分は、各パラメータ推定値の分散の値を示している。具体的に与えると以下となる。

$$\mathfrak{I}^{-1} \to \begin{pmatrix} \sigma^2/N & 0 \\ 0 & 2\sigma^4/N \end{pmatrix}$$

これまでは分散σ^2の推定値を求めたが、標準偏差σの推定値としてはどのように変わるのだろうか。

$$\partial \log L/\partial\mu = \frac{1}{\sigma^2}\sum_{\lambda=1}^{N}(x_\lambda - \mu) = 0 \qquad \mu = \frac{1}{N}\sum_{\lambda=1}^{N} x_\lambda$$

$$\partial \log L/\partial\sigma = 2\sigma\,\partial \log L/\partial\sigma^2 = 2\sigma\left[\frac{1}{2\sigma^4}\sum_{\lambda=1}^{N}(x_\lambda-\mu)^2 - \frac{N}{2\sigma^2}\right]$$

$$= \frac{1}{\sigma^3}\sum_{\lambda=1}^{N}(x_\lambda-\mu)^2 - \frac{N}{\sigma} = 0 \qquad \sigma^2 = \frac{1}{N}\sum_{\lambda=1}^{N}(x_\lambda-\mu)^2$$

$$\partial^2 \log L/\partial\mu^2 = -\frac{N}{\sigma^2}$$

$$\partial^2 \log L/\partial\mu\partial\sigma = 2\sigma\left[\partial^2 \log L/\partial\mu\partial\sigma\right] = -\frac{2}{\sigma^3}\sum_{\lambda=1}^{N}(x_\lambda-\mu) \to 0$$

$$\partial^2 \log L/\partial\sigma^2 = \frac{\partial}{\partial\sigma}\left[\frac{1}{\sigma^3}\sum_{\lambda=1}^{N}(x_\lambda-\mu)^2 - \frac{N}{\sigma}\right] = -\frac{3}{\sigma^4}\sum_{\lambda=1}^{N}(x_\lambda-\mu)^2 + \frac{N}{\sigma^2} \to -\frac{2N}{\sigma^2}$$

以上より、

$$\mathfrak{I}^{-1} \to \begin{pmatrix} \sigma^2/N & 0 \\ 0 & \sigma^2/2N \end{pmatrix}$$

結果には、標準偏差の分散としてこちらを使うことにする。

対数正規分布 $(0 < x < \infty)$

密度関数： $f(x\,|\,\mu,\sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}\,x}\exp[-(\log x-\mu)^2/2\sigma^2]$

平均、分散との関係

$E[X] = e^{\mu+\sigma^2/2}$， $V[X] = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$

尤度関数： $L = \dfrac{1}{(2\pi\sigma^2)^{N/2}}\prod_{\lambda=1}^{N}\dfrac{1}{x_\lambda}\exp\left[-\dfrac{1}{2\sigma^2}\sum_{\lambda=1}^{N}(\log x_\lambda-\mu)^2\right]$

対数尤度： $\log L = -\dfrac{1}{2\sigma^2}\sum_{\lambda=1}^{N}(\log x_\lambda-\mu)^2 - \sum_{\lambda=1}^{N}\log x_\lambda - \dfrac{N}{2}\log(2\pi\sigma^2)$

ここではσについて推定値を求める。

$$\partial \log L/\partial\mu = \frac{1}{\sigma^2}\sum_{\lambda=1}^{N}(\log x_\lambda-\mu) = 0 \qquad \mu = \frac{1}{N}\sum_{\lambda=1}^{N}\log x_\lambda$$

$$\partial \log L/\partial\sigma = \frac{1}{\sigma^3}\sum_{\lambda=1}^{N}(\log x_\lambda-\mu)^2 - \frac{N}{\sigma} = 0 \qquad \sigma^2 = \frac{1}{N}\sum_{\lambda=1}^{N}(\log x_\lambda-\mu)^2$$

以上で解析的に求めることが可能である。対数正規分布についてはこれを使う。この推定値は正規分布に対して $x_\lambda \to \log x_\lambda$ としたものである。推定値はこれを使ったものと考える。

$$\partial^2 \log L/\partial\mu^2 = -\frac{N}{\sigma^2}$$

$$\partial^2 \log L/\partial\mu\partial\sigma = -\frac{2}{\sigma^3}\sum_{\lambda=1}^{N}(\log x_\lambda - \mu) \to 0$$

$$\partial^2 \log L/\partial\sigma^2 = -\frac{3}{\sigma^4}\sum_{\lambda=1}^{N}(\log x_\lambda - \mu)^2 + \frac{N}{\sigma^2} \to -\frac{2N}{\sigma^2}$$

以上より、

$$\Im^{-1} \to \begin{pmatrix} \sigma^2/N & 0 \\ 0 & \sigma^2/2N \end{pmatrix} \qquad \text{但し、} \ \sigma^2 = \frac{1}{N}\sum_{\lambda=1}^{N}(\log x_\lambda - \mu)^2$$

χ^2 分布 $(0 < x < \infty)$ パラメータが離散的

密度関数： $f(x) = \dfrac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1}\exp(-x/2)$

尤度関数： $L = \dfrac{1}{2^{Nn/2}\Gamma(n/2)^N}\prod_{\lambda=1}^{N} x_\lambda^{n/2-1}\exp(-x_\lambda/2)$

対数尤度： $\log L = \dfrac{n-2}{2}\sum_{\lambda=1}^{N}\log(x_\lambda) - \dfrac{1}{2}\sum_{\lambda=1}^{N}x_\lambda - \dfrac{Nn}{2}\log 2 - N\log\Gamma(n/2)$

$\chi^2 \sim \chi_n^2$ のとき、$E(\chi^2) = n$ の性質を用いて、

$n = \lfloor x + 0.5 \rfloor$　　注）$\lfloor x \rfloor$ は x を越えない最大の整数

これを元に $(1 \le) \ n-5 \le n_{\max} \le n+5$ の範囲で最大の対数尤度を与える自由度 $n_{\max}$ を求めている。

F 分布 $(0 < x < \infty)$

密度関数： $f(x) = \dfrac{1}{B(n_1/2, n_2/2)}\left(\dfrac{n_1}{n_2}\right)^{n_1/2}\dfrac{x^{n_1/2-1}}{(1+xn_1/n_2)^{(n_1+n_2)/2}}$

尤度関数： $L = \dfrac{1}{B(n_1/2, n_2/2)^N}\left(\dfrac{n_1}{n_2}\right)^{Nn_1/2}\prod_{\lambda=1}^{N}\dfrac{x_\lambda^{n_1/2-1}}{(1+x_\lambda n_1/n_2)^{(n_1+n_2)/2}}$

対数尤度：
$$\log L = \left(\frac{n_1}{2}-1\right)\sum_{\lambda=1}^{N}\log(x_\lambda) - \frac{n_1+n_2}{2}\sum_{\lambda=1}^{N}\log(1+x_\lambda n_1/n_2) + \frac{Nn_1}{2}\log(n_1/n_2) - N\log B(n_1/2, n_2/2)$$

$E[X] = \dfrac{n_2}{n_2-2} \ \ (n_2 > 2)$, $V[X] = \dfrac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)} \ \ (n_2 > 4)$　　を利用して、

$$n_2 = \frac{2E[X]}{E[\mathrm{X}]-1} \quad , \quad n_1 = \frac{2n_2^2(n_2-2)}{(n_2-2)^2(n_2-4)V[X]-2n_2^2}$$

これを元に、ぶれが大きいので、$(1\le)n_i - 20 \le n_{i\max} \le n_i + 20$ の範囲で対数尤度を最大化する $n_{i\max}$ を求めている。

t 分布 $(-\infty < x < \infty)$

密度関数： $f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}$

尤度関数： $L = \left(\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\right)^N \prod_{\lambda=1}^{N}\left(1+\frac{x_\lambda^2}{n}\right)^{-(n+1)/2}$

対数尤度： $\log L = -\frac{n+1}{2}\sum_{\lambda=1}^{N}\log\left(1+\frac{x_\lambda^2}{n}\right) + N\log\left(\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\right)$

平均： $E[X]=0$

分散： $V[X] = \frac{n}{n-2}$ を利用して、 $n = \frac{2V[X]}{V[X]-1}$

これを元に $(1\le)\, n-5 \le n_{\max} \le n+5$ の範囲で最大の対数尤度を与える自由度 $n_{\max}$ を求めている。

ガンマ分布 $(0 < x < \infty)$

密度関数： $f(x) = \frac{1}{b^a\Gamma(a)}x^{a-1}\exp(-x/b)$

尤度関数： $L = \frac{1}{[b^a\Gamma(a)]^N}\prod_{\lambda=1}^{N}x_\lambda^{a-1}\exp(-x_\lambda/b)$

対数尤度： $\log L = (a-1)\sum_{\lambda=1}^{N}\log x_\lambda - \frac{1}{b}\sum_{\lambda=1}^{N}x_\lambda - Na\log b - N\log\Gamma(a)$

$\partial\log L/\partial a = \sum_{\lambda=1}^{N}\log x_\lambda - N\log b - N\,\Gamma'(a)/\Gamma(a)$

$\partial\log L/\partial b = 1/b^2\sum_{\lambda=1}^{N}x_\lambda - N\,a/b$

$\partial^2\log L/\partial a^2 = -N\left[\Gamma''(a)/\Gamma(a) - \Gamma'(a)^2/\Gamma(a)^2\right]$

$\partial^2\log L/\partial a\partial b = -N/b$

$\partial^2\log L/\partial b^2 = -2/b^3\sum_{\lambda=1}^{N}x_\lambda + N\,a/b^2$

初期値は $a_0 = 0.5,\ b_0 = 0.5$ を用いている。

逆ガンマ分布 $(0 \le x \le 1)$

密度関数： $f(x) = \dfrac{b^a}{\Gamma(a)} x^{-a-1} \exp(-b/x)$

尤度関数： $L = \left[b^a / \Gamma(a) \right]^N \prod_{\lambda=1}^{N} x_\lambda^{-a-1} \exp(-b/x_\lambda)$

対数尤度： $\log L = -(a+1) \sum_{\lambda=1}^{N} \log x_\lambda - b \sum_{\lambda=1}^{N} (1/x_\lambda) + Na \log b - N \log \Gamma(a)$

$\partial \log L / \partial a = -\sum_{\lambda=1}^{N} \log x_\lambda + N \log b - N \Gamma'(a)/\Gamma(a)$

$\partial \log L / \partial b = -\sum_{\lambda=1}^{N} (1/x_\lambda) + N\, a/b$

$\partial^2 \log L / \partial a^2 = -N \left[\Gamma''(a)/\Gamma(a) - \Gamma'(a)^2/\Gamma(a)^2 \right]$

$\partial^2 \log L / \partial a \partial b = N/b$

$\partial^2 \log L / \partial b^2 = -N\, a/b^2$

初期値は $a_0 = 0.5,\ b_0 = 0.5$ を用いている。

ベータ分布 $(0 \le x \le 1)$

密度関数： $f(x) = \dfrac{x^{a-1}(1-x)^{b-1}}{B(a,b)} = \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$

尤度関数： $L = \dfrac{1}{B(a,b)} \prod_{\lambda=1}^{N} x_\lambda^{a-1}(1-x_\lambda)^{b-1}$

対数尤度： $\log L = (a-1) \sum_{\lambda=1}^{N} \log x_\lambda + (b-1) \sum_{\lambda=1}^{N} \log(1-x_\lambda) - N \log B(a,b)$

$\partial \log L / \partial a = \sum_{\lambda=1}^{N} \log x_\lambda - N \dfrac{B(a,b)_a}{B(a,b)}$

$\partial \log L / \partial b = \sum_{\lambda=1}^{N} \log(1-x_\lambda) - N \dfrac{B(a,b)_b}{B(a,b)}$

$\partial^2 \log L / \partial a^2 = -N \left(\dfrac{B(a,b)_{aa}}{B(a,b)} - \dfrac{B(a,b)_a^2}{B} \right)$

$\partial^2 \log L / \partial a \partial b = -N \left(\dfrac{B(a,b)_{ab}}{B(a,b)} - \dfrac{B(a,b)_a B(a,b)_b}{B(a,b)^2} \right)$

$\partial^2 \log L / \partial b^2 = -N \left(\dfrac{B(a,b)_{bb}}{B(a,b)} - \dfrac{B(a,b)_b^2}{B} \right)$

初期値の設定で、平均値が 0 に近い場合は 1,5、1 に近い場合は 5,1、0.5 に近い場合は 0.5, 0.5 などを使う。小さい方から大きい方へ近づけて行くことは問題ないが、大きい方から小さい方へ近づけて行く際にはエラーが出る。

ワイブル分布 $(0<x<\infty)$（失敗例）

通常のa,bを使って最尤法を試みた。

密度関数： $f(x)=(a/b)\left(x/b\right)^{a-1}\exp\left[-\left(x/b\right)^{a}\right]$

尤度関数： $L=(a/b)^{N}\prod_{\lambda=1}^{N}\left(x_\lambda/b\right)^{a-1}\exp\left[-\left(x_\lambda/b\right)^{a}\right]=a^{N}b^{-Na}\prod_{\lambda=1}^{N}x_\lambda^{a-1}\exp[-x_\lambda^{a}b^{-a}]$

対数尤度： $\log L=(a-1)\sum_{\lambda=1}^{N}\log x_\lambda-b^{-a}\sum_{\lambda=1}^{N}x_\lambda^{a}+N\log a-Na\log b$

$$\partial\log L/\partial a=\sum_{\lambda=1}^{N}\log x_\lambda+\log b\cdot b^{-a}\sum_{\lambda=1}^{N}x_\lambda^{a}-b^{-a}\sum_{\lambda=1}^{N}\log x_\lambda\cdot x_\lambda^{a}+N/a-N\log b$$

$$\partial\log L/\partial b=ab^{-a-1}\sum_{\lambda=1}^{N}x_\lambda^{a}-Na/b$$

$$\partial^2\log L/\partial a^2=-(\log b)^2b^{-a}\sum_{\lambda=1}^{N}x_\lambda^{a}+2\log b\cdot b^{-a}\sum_{\lambda=1}^{N}\log x_\lambda\cdot x_\lambda^{a}-b^{-a}\sum_{\lambda=1}^{N}(\log x_\lambda)^2\cdot x_\lambda^{a}-N/a^2$$

$$\partial^2\log L/\partial a\partial b=(1-a\log b)b^{-a-1}\sum_{\lambda=1}^{N}x_\lambda^{a}+ab^{-a-1}\sum_{\lambda=1}^{N}\log x_\lambda\cdot x_\lambda^{a}-N/b$$

$$\partial^2\log L/\partial b^2=-a(a+1)b^{-a-2}\sum_{\lambda=1}^{N}x_\lambda^{a}+Na/b^2$$

この方法は、収束が思うように行かず、エラーとなった。

ワイブル分布 $(0<x<\infty)$　再度

上記の失敗を踏まえ、生存時間分析で用いたパラメータの推定法を利用する。

密度関数： $f(x)=(a/b)\left(x/b\right)^{a-1}\exp\left[-\left(x/b\right)^{a}\right]$

尤度関数：
$$L=(a/b)^{N}\prod_{\lambda=1}^{N}\left(x_\lambda/b\right)^{a-1}\exp\left[-\left(x_\lambda/b\right)^{a}\right]=a^{N}b^{-Na}\prod_{\lambda=1}^{N}x_\lambda^{a-1}\exp[-x_\lambda^{a}b^{-a}]$$
$$=a^{N}e^{N\beta}\prod_{\lambda=1}^{N}x_\lambda^{a-1}\exp[-x_\lambda^{a}e^{\beta}]$$

$$\log L(a,b)=\sum_{\lambda=1}^{N}\left[\left(\log a+(a-1)\log x_\lambda+\beta\right)-x_\lambda^{a}e^{\beta}\right]$$
$$=(a-1)\sum_{\lambda=1}^{N}\log x_\lambda-e^{\beta}\sum_{\lambda=1}^{N}x_\lambda^{a}+N\log a+N\beta$$

$$\frac{\partial}{\partial a}\log L=\sum_{\lambda=1}^{N}\log x_\lambda-e^{\beta}\sum_{\lambda=1}^{N}\log x_\lambda\cdot x_\lambda^{a}+N/a$$

$$\frac{\partial}{\partial\beta}\log L=-e^{\beta}\sum_{\lambda=1}^{N}x_\lambda^{a}+N$$

$$\frac{\partial^2}{\partial a^2}\log L = -e^{\beta}\sum_{\lambda=1}^{N}(\log x_\lambda)^2 x_\lambda^a - N/a^2$$

$$\frac{\partial}{\partial a\partial\beta}\log L = -e^{\beta}\sum_{\lambda=1}^{N}\log x_\lambda\, x_\lambda^a$$

$$\frac{\partial^2}{\partial \beta^2}\log L = -e^{\beta}\sum_{\lambda=1}^{N} x_\lambda^a$$

初期値は$a_0 = 2,\ \beta = 2$を用いている。

指数分布 $(0 < x < \infty)$

密度関数： $f(t) = a\exp(-ax)$　　（$x \geq 0$）

尤度関数： $L = a^N\prod_{i=1}^{N}\exp(-ax_\lambda) = a^N\exp\left(-a\sum_{\lambda=1}^{N}x_\lambda\right)$

対数尤度： $\log L = N\log a - a\sum_{i=1}^{N}x_\lambda$

$$\frac{\partial}{\partial a}\log L = \frac{N}{a} - \sum_{\lambda=1}^{N}x_\lambda = 0 \qquad a = N\Big/\sum_{\lambda=1}^{N}x_\lambda$$

$$\frac{\partial^2}{\partial a^2}\log L = -\frac{N}{a^2}$$

この逆数は、推定値の分散を与える。（今回は使わない）

推定値は解析的に求まるが、練習問題として最尤法を用いてみる。

初期値は$a = 0.1$を用いている。

ポアソン分布 $(0 < x < \infty)$，整数

確率関数： $P(x) = e^{-a}\,a^x/x!$

尤度関数： $L = \prod_{\lambda=1}^{N}e^{-a}\,a^{x_\lambda}/x_\lambda! = e^{-Na}\prod_{\lambda=1}^{N}a^{x_\lambda}/x_\lambda!$

対数尤度： $\log L = -Na + \log a\sum_{\lambda=1}^{N}x_\lambda - \sum_{\lambda=1}^{N}\log x_\lambda!$

$$\partial\log L/\partial a = -N + \frac{1}{a}\sum_{\lambda=1}^{N}x_\lambda$$

$$\partial^2\log L/\partial a^2 = -\frac{1}{a^2}\sum_{\lambda=1}^{N}x_\lambda$$

初期値は$a = 0.1$を用いている。

2 項分布 $(0 < x < \infty)$，整数

まず以下の関係を使って、度数nを求める。

$E[X] = np$，　$V[X] = npq$

$$n = \frac{E[X]^2}{E[X]-V[X]}$$

次に最尤法を使って、確率 p を求める。

確率関数： $P(x) = {}_nC_x p^x (1-p)^{n-x}$

尤度関数： $L = \prod_{\lambda=1}^{N} {}_nC_{x_\lambda} p^{x_\lambda} (1-p)^{n-x_\lambda}$

対数尤度： $\log L = \sum_{\lambda=1}^{N} \log {}_nC_{x_\lambda} + \log p \sum_{\lambda=1}^{N} x_\lambda + \log(1-p) \sum_{\lambda=1}^{N} (n-x_\lambda)$

$$\partial \log L / \partial p = \frac{1}{p} \sum_{\lambda=1}^{N} x_\lambda - \frac{1}{1-p} \sum_{\lambda=1}^{N} (n-x_\lambda)$$

$$\partial^2 \log L / \partial p^2 = -\frac{1}{p^2} \sum_{\lambda=1}^{N} x_\lambda - \frac{1}{(1-p)^2} \sum_{\lambda=1}^{N} (n-x_\lambda)$$

これも解析的に解を求めることができるが、最尤法の演習とする。

初期値は $p = 0.5$ を用いている。

１６．自由記述集計

アンケートなどで自由記述欄を設けた際、その文章を検索してキーワードを見出し、その出現頻度を求め、文中でのキーワード同士の連携関係を求めることはテキストマイニングの初歩として重要である。我々はデータの特殊な集計法として、この自由記述文の検索と集計プログラムを College Analysis の基本統計に加えている。

本格的な大量データのテキストマイニングには自動的な形態素解析が必須であり、現在の我々のシステムでは不可能である。しかし、規模の小さな自由記述データでは分析者の判断によるキーワード抽出が可能であり、これを利用したデータ処理はある程度可能である。このプログラムはこれらの分析を行うためのツールである。

メニュー［分析－基本統計－自由記述集計］を選択すると、図１に示す実行画面が表示される。

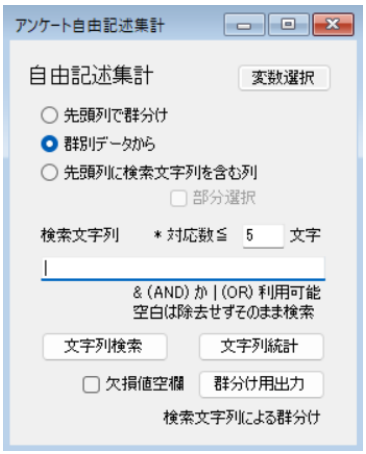


図１　自由記述集計実行画面

データは図 2 のように、自由記述データと数値や記号を混在させてもよい。右端の「検索 1」の列は、元々のデータではなく、検索のために追加した列で、利用法は後で説明する。

	地域	年収	支出	意見1	意見2	自由記述1	自由記述2	検索1
1	1	583	49	2	3	私は、教育に関心がある。	私は幸せだと思う。	教育&関心
2	1	565	33	2	3			学歴
3	2	508	32	1	3	教育にあまり興味がない。		幸せ
4	2	565	31	2	1		幸福は人それぞれ。	
5	1	594	57	2	3	学歴社会だと思うから。	好きなことをやるのが幸せ	
6	2	624	47	1	1		人それぞれ	
7	1	617	48	2	1	重要なのは本人のやる気だ。		
8	1	458	53	2	3			
9	1	754	62	2	1			
10	2	667	53	2	1			
11	2	470	37	1	1	教育には大いに関心がある。	子供の幸せが一番	
12	2	578	28	2	3			
13	2	592	13	2	3			
14	2	723	45	2	2			
15	1	674	46	2	3			
16	2	676	51	1	3	どのような教育が良いのか分からない。	警察に捕まりさえしなければ。	

図２　自由記述集計のデータ形式

通常の基本統計の集計と同じように、集計の形式を「群別データから」、「先頭列で群分け」、「先頭列に文字列を含む列」の中から選択することができるが、最初の２つは基本的に通常の基本統計の集計の場合と同じである。

最初に「群別データから」の場合について、検索文字列で "教育" と指定し（両側の""は入力しない）、変数選択で自由記述１だけを選択して、「文字列検索」ボタンをクリックすると図３左、「文字列統計」ボタンをクリックすると図３右のような検索結果を得る。

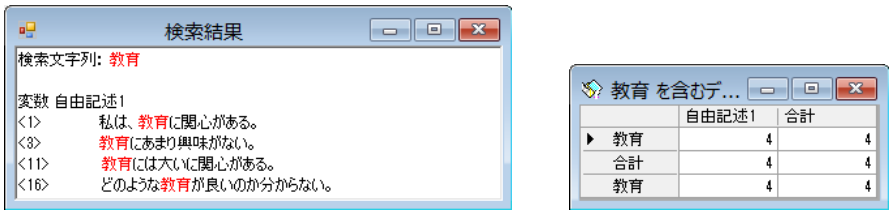


図３　「群別データから」での検索結果１

検索文字列には、& (and)、| (or) やワイルドカード「*」が利用できる。ワイルドカードは * が何文字までに対応するかを「* 対応数≦」として指定することができる。例えば、検索文字列に、"教育*関心|幸" と指定し、変数選択で自由記述１と自由記述２を選択して、「文字列検索」ボタンをクリックすると、図４左、「文字列統計」ボタンをクリックすると図４右のような検索結果を得る。

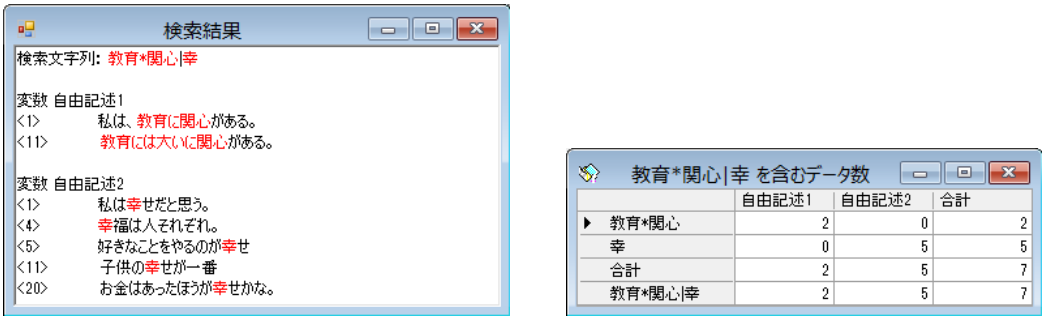


図４　「群別データから」での検索結果２

集計の形式が「群別データから」であるため、２つの変数は独立に検索対象になっている。また、図３及び図４の右側の表で、合計の下に検索文字列が表示されているが、これは、合計までが or で分けて検索した結果、その下が検索文字列でそのまま検索した結果を表している。一般に上の合計と下の結果は異なるが、今の場合は同じ数になっている。図３のように and や or を使わない場合は全く同じものが表示されている。

次に集計方法として「先頭列で群分け」を選択し、検索文字列で "教育" と指定し、変数選択で地域と自由記述１を選択して、「文字列検索」ボタンをクリックすると、図５左、「文字列統計」ボタンをクリックすると図５右のような検索結果を得る。

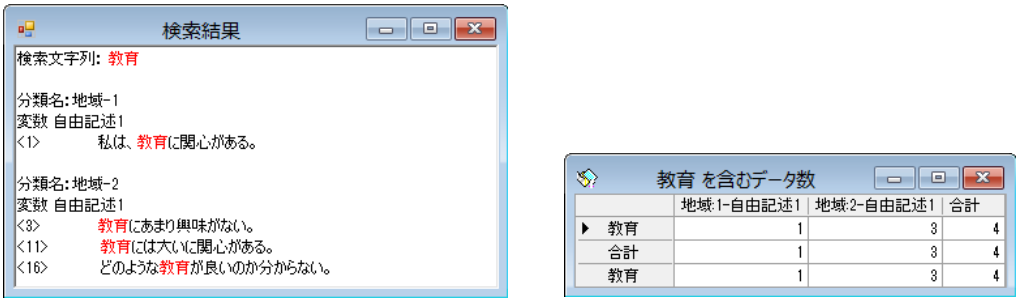


図 5　「先頭列で群分け」での検索結果

これは地域による群分けを実行した後で検索を実行した結果である。

最後に「先頭列に検索文字列を含む列」では、変数選択で、例えば図 2 の検索 1 の列を最初に選択し、検索対象とする列を次に選択する。この例では、検索文字列で "教育&関心|学歴|幸せ" と指定し、その後の変数を群別データからで選択したものとすることと同等である。例えば変数選択で、検索 1、自由記述 1、自由記述 2 を選択して、「文字列検索」ボタンをクリックすると、図 6 左、「文字列統計」ボタンをクリックすると図 6 右のような検索結果を得る。

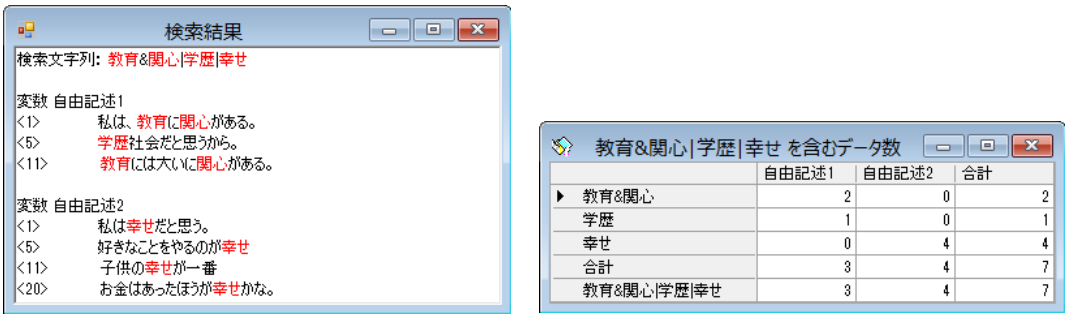


図 6　「先頭列に検索文字列を含む列」での検索結果

次に、集計の形式を「群別データから」として、変数選択で自由記述 1 を選び、検索文字列を "教育" として、「群分け用出力」ボタンをクリックすると、グリッド出力に選択文字列を含むレコードに 1、含まないレコードに 0（欠損値は 0 か空欄が選択可能）が出力される。

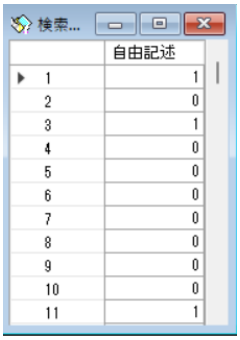


図 7　群分け用出力

これはグリッド出力メニュー［エディタ全列追加］により、図 8 のようにグリッド編集画面の最後に容易に貼り付けることができる。

	地域	年収	支出	意見1	意見2	自由記述1	自由記述2	検索1	自由記述1:...
1	1	583	49	2	3	私は、教育に...	私は幸せだと...	教育&関心	1
2	1	565	33	2	3			学歴	0
3	2	508	32	1	3	教育にあまり...		幸せ	1
4	2	565	31	2	1		幸福は人そ...		0
5	1	594	57	2	3	学歴社会だ...	好きなことを...		0
6	2	624	47	1	1		人それぞれ		0
7	1	617	48	2	1	重要なのは...			0
8	1	458	53	2	3				0
9	1	754	62	2	1				0
10	2	667	53	2	1				0
11	2	470	37	1	1	教育には大...	子供の幸せ...		1

図 8　群分け出力結果

これを用いると、自由記述欄にある検索文字列を使っているかどうかで分けて、他のデータの集計や分けたデータ間で検定をすることが可能になる。

　最後に英文などで利用し易くなるように、空白はそのまま残して検索することにしました。注意して下さい。

１７．検定の効率化

統計の処理や検定では、１つ１つの項目の性質を見極め、十分検討しながら処理を行うことが重要であるが、質問項目の多いアンケート調査などでは、最初にある程度の結果を出し、有意差の出そうなものを見つけて、後で詳しく調べたいと考えることがある。今回この方法を実現するために、χ^2検定、2 群間の量的データの検定、実験計画法の中に、複数の処理を一度に行う機能を追加した。ここでは、簡単な以下の例を元にこれらの機能を紹介する（検定の効率化.txt）。

１）合否（１：合格，２：不合格・質）
２）クラブ活動（３段階・質）
３）アルバイト（３段階・質）
４）社会活動（２段階・質）
５）専門知識（点数・量）
６）高校成績（点数・量）
７）大学成績（点数・量）
８）出席率（%表示・量）

メニュー［基本統計ー質的データの集計］を選択すると、図 1 のような実行画面が表示される。これは元の画面と同じである。

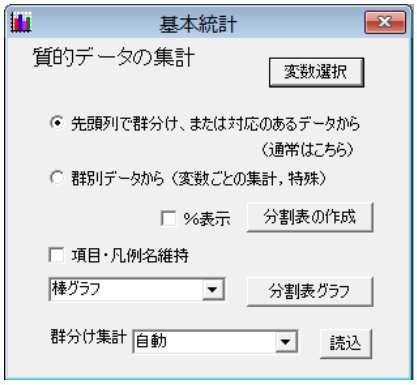


図１　質的データの集計実行画面

2 次元分割表を描くには通常２つの質的データを選択するが、処理を一度に行う場合は、例えば、１）合否～４）社会活動までまとめて選択する。その後「分割表の作成」ボタンをクリックすると、以下のように、先頭列（最初に選んだ変数）を元に１つの分割表が横１行にまとまって表示される。


複数変数２次元分割表

	クラブ活動-1	クラブ活動-2	クラブ活動-3	合計	アルバイト-1	アルバイト-2	アルバイト-3	合計	社会活動-1	社会活動-2	合計
合否-1	20	15	11	46	8	27	11	46	22	24	46
合否-2	7	26	16	49	19	27	3	49	27	22	49
合計	27	41	27	95	27	54	14	95	49	46	95

図２　まとめて表示された 2 次元分割表

χ^2検定についても、図 3 のようにメニューの上では変更がない。

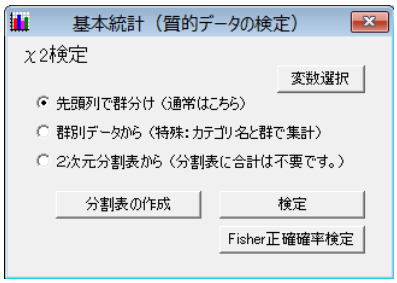


図３　χ^2検定実行画面

しかし、まとめて変数を選んだ場合は、テキスト表示と異なり、図 4 のようなグリッド表示となる。

	自由度	χ2値	片側確率
クラブ活動	2	8.34169	0.01544
アルバイト	2	7.07138	0.02914
社会活動	1	0.25379	0.61442

図４　まとめて表示されたχ^2検定結果

ここにクラブ活動とアルバイトは 2×3 分割表、社会活動は 2×2 分割表として処理されていることが自由度から分かる。その他の質的なデータの集計や検定については、データの形式からまとめて処理することがないと思われるので、変更を加えていない。

量的なデータについては、対応のない 2 群間の比較と 1 元配置実験計画法の問題に機能追加を行った。例えば、ｔ検定の実行画面は、図 5 のように与えられ、変更はないが、

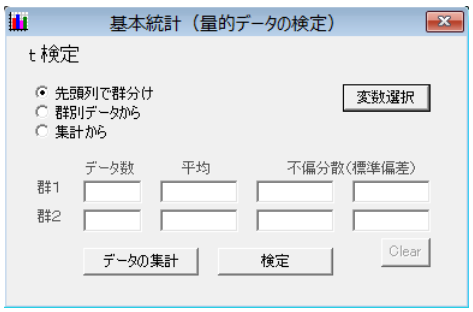


図５　ｔ検定実行画面

「先頭列で群分け」で、通常 2 つの変数を選ぶところを、群分けする変数１）合否に続いて５）専門知識～８）出席率のように複数の変数を選んで、「検定」ボタンをクリックすると、図 6 に示されるように一括で処理される。

	自由度	t値	両側確率
専門知識	93	11.7621	0.0000
高校成績	93	0.3775	0.7067
大学成績	93	5.7350	0.0000
出席率	93	4.4273	0.0000

図６　まとめて表示されたｔ検定結果

Welch のｔ検定や Wilcoxon の順位和検定でも同様の機能追加がなされている。

さて、量的データの検定では、データの分布によって検定方法を変えるのが一般的である

ので、このようにすべてｔ検定で行うのは好ましくない。そこで我々は、図 7 のように、量的データの検定実行画面に検定を自動選択するボタンを加えた。

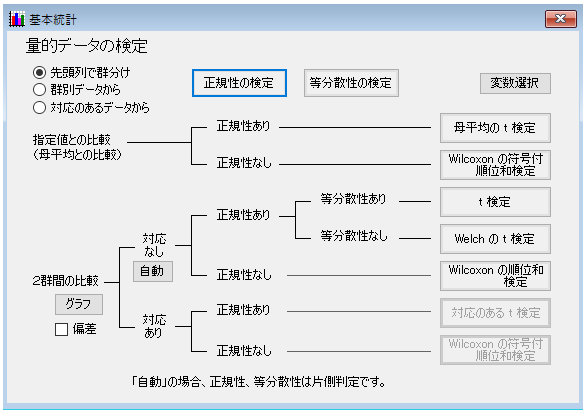


図 7　量的データ検定実行画面

変数を上で述べたｔ検定の場合と同じように選び、対応なしの下の「自動」ボタンをクリックすると、図 8 のように、検定が自動検索される様子が示され、結果が表示される。

	正規性	等分散性	検定手法	両側確率
▶ 専門知識	なし		順位和検定	0.0000
高校成績	ありとみなす	なし	Welch ｔ 検定	0.7090
大学成績	なし		順位和検定	0.0000
出席率	なし		順位和検定	0.0000

図 8　2 群間の比較検定自動検索結果

ここで、正規性の検定には S-W 検定（このプログラムの場合は近似）、等分散性の検定には F 検定が片側確率で利用されている。群別データの場合は、選択した複数の変数を、条件を変えた 1 つの変数として考えるので、結果は 1 行で表示される。他の検定については、データの形式から、一括で処理することがないのでこれまで通り 1 種類ずつ処理する。

ここで 2 群間の比較を考えたので、3 群以上の比較についても同様の機能拡張を行う必要がある。これは 1 元配置の実験計画法の問題である。メニュー［多変量解析－実験計画法］を選択して表示される実行画面を図 9 に示す。

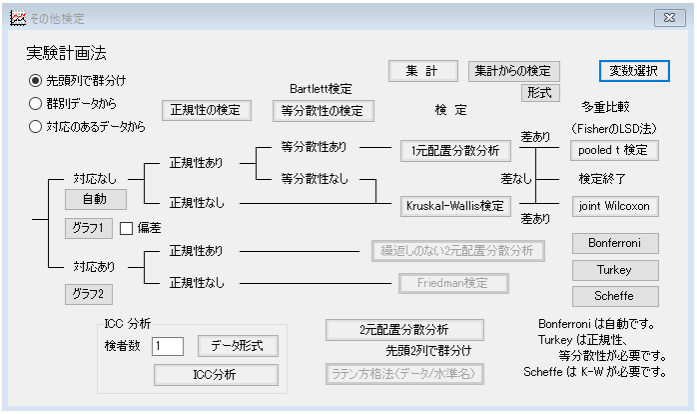
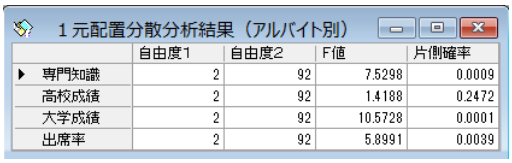


図 9　実験計画法実行画面

この中で、先頭列で群分けの場合、1 元配置分散分析と Kruskal-Wallis 検定では一括処理が可能である。例えば、群分けする変数３）アルバイトに続いて５）専門知識～８）出席率と複数の変数を選んで、「1 元配置分散分析」ボタンをクリックすると、図 10 に示されるように一括で処理した結果が表示される。



	自由度1	自由度2	F値	片側確率
▶ 専門知識	2	92	7.5298	0.0009
高校成績	2	92	1.4188	0.2472
大学成績	2	92	10.5728	0.0001
出席率	2	92	5.8991	0.0039

図 10　まとめて表示された 1 元配置分散分析結果

実験計画法でもデータの分布によって検定方法を変えるので、図 9 のメニューでも検定を自動選択するボタンを加えてある。1 元配置分散分析と同じ変数を選択し、図 9 の「自動選択」ボタンをクリックすると図 11 の結果が表示される。

	正規性	等分散性	検定手法	片側確率
▶ 専門知識	ありとみなす	ありとみなす	1元配置分...	0.0009
高校成績	ありとみなす	なし	K-W 検定	0.4862
大学成績	なし		K-W 検定	0.0000
出席率	ありとみなす	ありとみなす	1元配置分...	0.0039

図 11　1 元配置検定自動検索結果

他の検定については、データの形式から、一括で処理することがないのでこれまで通り 1 種類ずつ処理する。

１８．層別分割表のオッズ比検定

18.1　層別分割表のオッズ比検定とは

質的データ同士の関係を調べるための基本的な統計手法は 2 次元分割表に基づく検定である。例えばたばこ摂取の度合いにより、ある疾病の罹患状況に差があるかどうか調べるといった場合、たばこ摂取の有無による差を見る場合はオッズ比の検定（ほぼ通常の χ^2 検定と同様）を行い、たばこの用量－反応関係を調べる場合は Mantel-extension 法などのトレンドの検定手法を利用する。しかし、これは本当に正しいのであろうか。疾病の原因は、たばこだけとは限らないし、日頃の生活管理にも影響される。例えば、喫煙しない人が、健康のために毎日の適度な運動習慣を持っているということはないであろうか。この例のように２次元分割表における見かけの差の背後に結果に影響を及ぼす交絡因子（背景因子）が存在することがある。この交絡因子の影響を調整して分割表の有意差を検定する手法が層別分割表の検定である [1]。

ここで取り扱う検定手法は、層別 2×2 分割表に対する Mantel-Haenszel 法と層別 Mantel-extension 法である。前者は交絡因子を調整したオッズ比（相対危険度）の違い、後者は交絡因子を調整した用量－反応関係を検定する方法である。ここでは、Mantel-Haenszel 法の利用の前段階として、単独のオッズ比検定も分析に含めている。

18.2　プログラムの利用法

層別分割表の検定について、プログラムの利用法を説明する。メニュー［分析－基本統計－層別分割表の検定］をクリックすると、図 1 の実行画面が表示される。ラジオボタン「データから」を選択すると、図 2 のようなデータからの読み込みになる。

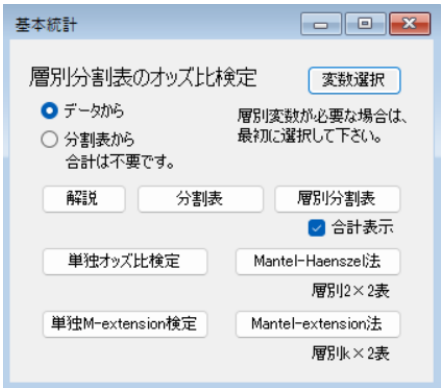


図 1　分析実行画面

最初に Mantel-Haenszel 法について図 2 のデータを元に利用法を説明する。

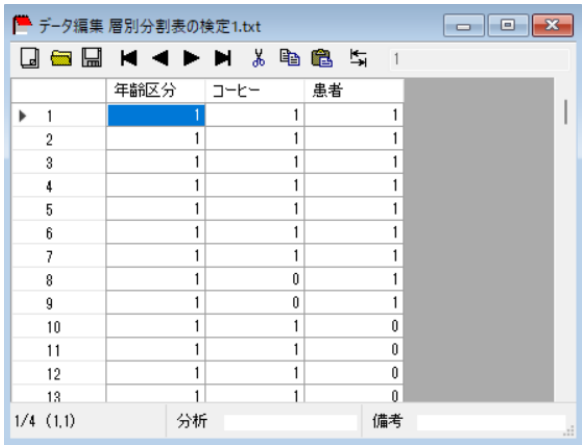


図 2　層別分割表の検定 1.txt（p1）

単独のオッズ比検定については、変数をコーヒー、患者の順に 2 つ選んで「単独オッズ比検定」のボタンをクリックすると、図 3 のような結果を得る。

図 3　単独オッズ比検定結果

この結果はイェーツ補正を含まない χ^2 検定の結果 p=0.0062 の結果とほぼ一致する。これは同一のことを 2 つの方法で分析しているからである。χ^2 検定では同時に 3 つ以上の変数を選んで複数の検定を行うことができるが、単独オッズ比検定でも同様である。

Mantel-Haenszel 法は、層別に単独のオッズ比検定を実行し、それを平滑する方法を取る。変数選択は、交絡因子（年齢区分）、曝露変数（コーヒー）、患者・対照変数（患者）の順に選ぶ。「合計表示」のチェックボックスをチェックし、「層別分割表」ボタンをクリックすると図 4 のように層（年齢区分）で分けた分割表が表示される。

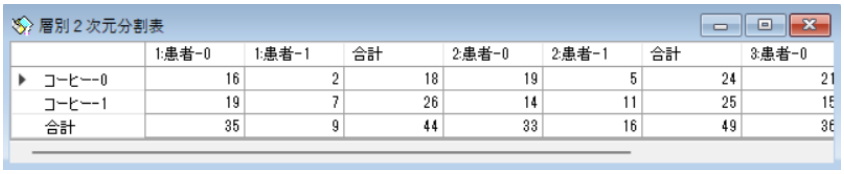


図 4　層別分割表結果（合計付）

「合計表示」のチェックボックスのチェックを外し、「層別分割表」ボタンをクリックすると、図 5 のような合計を含まない分割表が得られる。この分割表は分析実行画面で「分割表から」ラジオボタンを選択することにより、データとして使用できる。

図 5　層別分割表結果（合計なし）

「Mantel-Haenszel 法」ボタンをクリックすると分析結果が図 6 のように示される。

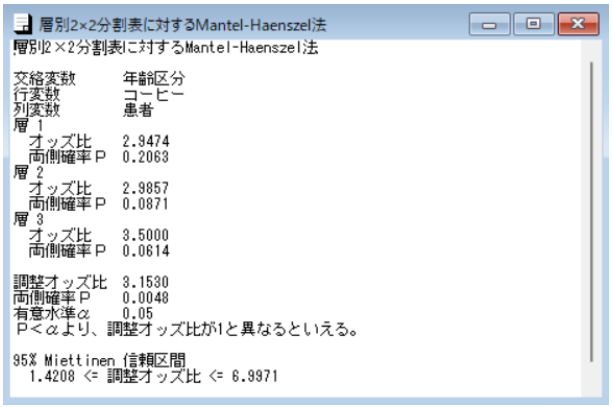


図 6　Mantel-Haenszel 法

これは層別の結果を 1 つに調整する分析である。分析事項画面の「単独 M-extension 検定」

次に単独の Mantel-extension 法について説明する。データは図 7 を用いる。

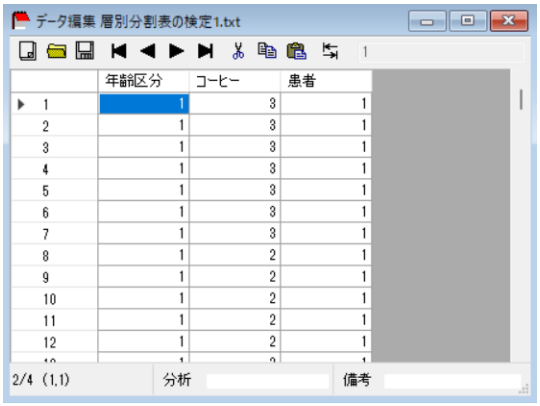


図 7　層別分割表の検定 1.txt（p2）

変数選択は曝露変数（コーヒー）、患者・対照変数（患者）の順に選ぶ。このデータは曝露変数について 0 ～ 3 の 4 区分に分類されている。Mantel-Haenszel 法では曝露変数について 2 区分のデータが用いられるが、データの形式や選択法は同じである。

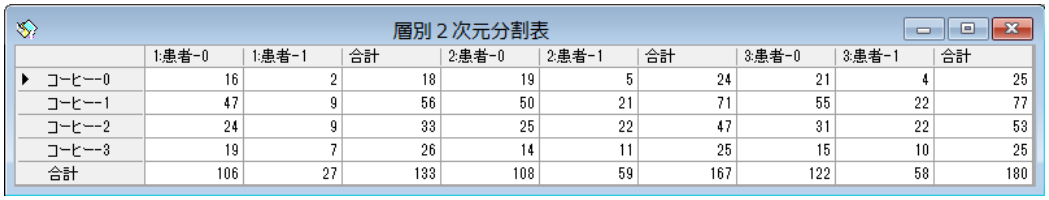
「単独 M-extension 検定」ボタンをクリックすると図 8 に示す分析結果が表示される。

図 8　単独 Mantel-extension 検定結果

対照・患者データを複数選ぶと行が追加される。

次に層別 Mantel-extension 法について説明する。変数選択は、交絡因子（年齢区分）、曝露変数（コーヒー）、患者・対照変数（患者）の順に選ぶ。

「層別分割表」ボタンをクリックすると、図 9 のような合計を含む分割表が得られる。合計を含まない分割表は、「分割表から」ラジオボタンを選択することにより、データとして使用できる。

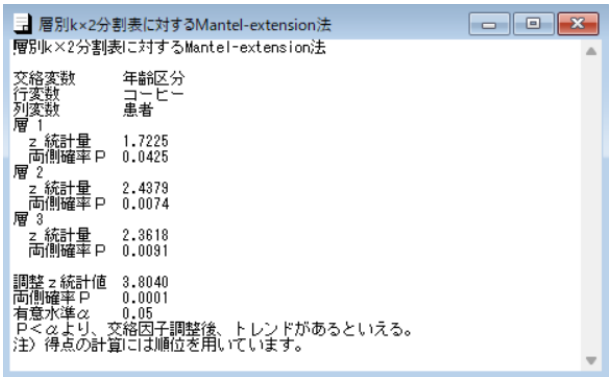


層別２次元分割表

	1:患者-0	1:患者-1	合計	2:患者-0	2:患者-1	合計	3:患者-0	3:患者-1	合計
▶ コーヒー-0	16	2	18	19	5	24	21	4	25
コーヒー-1	47	9	56	50	21	71	55	22	77
コーヒー-2	24	9	33	25	22	47	31	22	53
コーヒー-3	19	7	26	14	11	25	15	10	25
合計	106	27	133	108	59	167	122	58	180

図 9　合計を含む層別分割表

ここのように曝露変数が 2 分類以上の場合、交絡因子を調整したトレンドの検定である層別 Mantel-extension 法が利用可能である。「Mantel-extension 法」のボタンをクリックすると計算結果が図 10 のように表示される。



層別k×2分割表に対するMantel-extension法

```
層別k×2分割表に対するMantel-extension法

交絡変数        年齢区分
行変数          コーヒー
列変数          患者
層 1
  z 統計量      1.7225
  両側確率P     0.0425
層 2
  z 統計量      2.4379
  両側確率P     0.0074
層 3
  z 統計量      2.3618
  両側確率P     0.0091

調整ｚ統計値    3.8040
両側確率P       0.0001
有意水準α       0.05
P＜αより、交絡因子調整後、トレンドがあるといえる。
注）得点の計算には順位を用いています。
```

図 10　層別 Mantel-extension 法計算結果

ここで用いた例や計算結果は、参考文献 1）の中に与えられたものである。

18.3　層別分割表の検定の理論

層別 2×2 分割表に対する Mantel-Haenszel 法と層別 Mantel-extension 法について説明する。前者は交絡因子を調整したオッズ比（相対危険度）の違い、後者は交絡因子を調整した用量－反応関係を検定する方法である。

オッズ比の検定

患者－対照調査で、要因の有無により、表 1 のような分割表が得られたとする。

表 1 オッズ比検定のための 2×2 分割表

	対照	患者	合計
要因無	x_{11}	x_{12}	m_1
要因有	x_{21}	x_{22}	m_2
合計	n_1	n_2	N

このデータに対して患者群と対照群のオッズ比の観測値 RR は以下で与えられる。

$$RR \equiv \frac{x_{22} / x_{12}}{x_{21} / x_{11}} = \frac{x_{11} x_{22}}{x_{12} x_{21}}$$

オッズ比の検定について、帰無仮説 H_0 と対立仮説 H_1 は以下で与えられる。

H_0： $RR=1$

H_1： $RR\neq 1$

この検定には以下の関係を利用する。

$$D\equiv\frac{\sqrt{N-1}\left(x_{11}x_{22}-x_{12}x_{21}\right)}{\sqrt{m_1m_2n_1n_2}}\sim N\left(0,1\right)$$

オッズ比 RR の $\left(1-\alpha\right)\times 100\%$ 信頼区間は以下で与えられる。

$$RR^{1-Z(\alpha/2)/|D|}\leq RR\leq RR^{1+Z(\alpha/2)/|D|}$$

これを Miettinen の検定に基づく信頼区間という。

次はこの検定から交絡因子の影響を取り除く方法を述べる。交絡因子がある場合、集計には表 2 の層別 2×2 分割表を用いる。

表 2 交絡因子を調整したオッズ比検定のための層別 2×2 分割表

	第 1 階層			…	第 K 階層		
	対照	患者	合計	…	対照	患者	合計
要因無	x_{111}	x_{112}	m_{11}	…	x_{K11}	x_{K12}	m_{K1}
要因有	x_{121}	x_{122}	m_{12}	…	x_{K21}	x_{K22}	m_{K2}
合計	n_{11}	n_{12}	N_1	…	n_{K1}	n_{K2}	N_K

我々は交絡因子の階層数を K とし、各階層に対して表 1 の 2×2 分割表を考える。その際 Mantel-Haenszel による調整されたオッズ比は以下で与えられる。

$$RR_{MH}\equiv\frac{\sum_{k=1}^{K}x_{k11}x_{k22}/N_k}{\sum_{k=1}^{K}x_{k12}x_{k21}/N_k}$$

調整されたオッズ比について $RR_{MH}=1$ の検定は以下の性質を利用する。

$$D\equiv\frac{\sum_{k=1}^{K}x_{k22}-\sum_{k=1}^{K}\left(m_{k2}n_{k2}/N_k\right)}{\sqrt{\sum_{k=1}^{K}\frac{m_{k1}m_{k2}n_{k1}n_{k2}}{N_k^2\left(N_k-1\right)}}}\sim N\left(0,1\right)$$

近似的にこの検定はイェーツ補正をしない χ^2 検定に等しい。

オッズ比 RR_{MH} の Miettinen の検定に基づく $\left(1-\alpha\right)\times 100\%$ 信頼区間は以下で与えられる。

$$RR_{MH}^{1-Z(\alpha/2)/|D|}\leq RR_{MH}\leq RR_{MH}^{1+Z(\alpha/2)/|D|}$$

用量反応関係の検定

続いて、表 3 で与えられる用量－反応関係検定のための r×2 分割表について述べる。

表 3 用量－反応関係検定のための r×2 分割表

	対照	患者	合計
用量 1	x_{11}	x_{12}	m_1
用量 2	x_{21}	x_{22}	m_2
⋮	⋮	⋮	⋮
用量 r	x_{r1}	x_{r2}	m_r
合計	n_1	n_2	N

これはトレンドの検定としてすでに取り上げてある問題であるが、交絡因子調整の前段階として再度公式を与えておく。帰無仮説 H_0 と対立仮説 H_1 は以下で与えられる。

H_0：$OR_1 = 1 = OR_2 = \cdots = OR_r$（トレンドなし）

H_1：$OR_1 = 1 \le OR_2 \le \cdots \le OR_r$ または $OR_1 = 1 \ge OR_2 \ge \cdots \ge OR_r$（トレンドあり）

この検定のためにはまず、合計得点 O 、合計得点の平均 E 、合計得点の分散 V を計算する。

$$O \equiv \sum_{j=1}^{r} x_{j2} X_j$$

$$E \equiv \left(n_2 \sum_{j=1}^{r} m_j X_j \right) / N$$

$$V \equiv \frac{n_2 (N - n_2)}{N^2 (N-1)} \left\{ N \left(\sum_{j=1}^{r} m_j X_j^2 \right) - \left(\sum_{j=1}^{r} m_j X_j \right)^2 \right\}$$

ここで X_j は用量 j 群への得点を表す。これには $1 \sim r$ の値を与えるなど、何種類かの与え方があるが、我々は以下のような j 群の順位 R_j を用いている。

$$X_j \equiv R_j / N = \left(\sum_{i=1}^{j-1} m_i + \frac{n_j + 1}{2} \right) / N$$

これらの量を用いて以下の性質を利用する。

$$Z = \frac{O - E}{\sqrt{V}} \sim N(0,1)$$

単独で検定を行う場合は、イェーツ補正を加えることも多いが、ここでは省略している。

次に交絡因子がある場合の分割表を表 4 に示す。

表 4 交絡因子を調整した用量－反応関係検定のための r×2 分割表

	第 1 階層			…	第 K 階層		
	対照	患者	合計	…	対照	患者	合計
用量 1	x_{111}	x_{112}	m_{11}	…	x_{K11}	x_{K12}	m_{K1}
用量 2	x_{121}	x_{122}	m_{12}	…	x_{K21}	x_{K22}	m_{K2}
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮
用量 r	x_{1r1}	x_{1r2}	m_{1r}	…	x_{Kr1}	x_{Kr2}	m_{Kr}
合計	n_{11}	n_{12}	N_1	…	n_{K1}	n_{K2}	N_K

この検定のためにはまず層別の合計得点 O_k 、合計得点の平均 E_k 、合計得点の分散 V_k を計算する。

$$O_k \equiv \sum_{j=1}^{r} x_{kj2} X_j$$

$$E_k \equiv \left(n_{k2} \sum_{j=1}^{r} m_{kj} X_j \right) / N_k$$

$$V_k \equiv \frac{n_{k2}\left(N_k - n_{k2}\right)}{N_k^2\left(N_k - 1\right)} \left\{ N_k \left(\sum_{j=1}^{r} m_{kj} X_j^2 \right) - \left(\sum_{j=1}^{r} m_{kj} X_j \right)^2 \right\}$$

ここで X_j は j 群への得点を表す。得点の与え方にはいくつかの方法があるが、我々は以下のような j 群の順位 R_j を用いた方法を取っている。

$$X_j \equiv R_j / \sum_{k=1}^{K} N_k = \left\{ \sum_{i=1}^{j-1} \sum_{k=1}^{K} m_{ki} + \frac{1}{2} \left(\sum_{k=1}^{K} n_{kj} + 1 \right) \right\} / \sum_{k=1}^{K} N_k$$

トレンドの検定にはこれらの値を用いた以下の性質を利用する。

$$Z = \left\{ \sum_{k=1}^{K} \left(O_k - E_k \right) \right\} / \sqrt{\sum_{k=1}^{K} V_k} \sim N\left(0,1\right)$$

ここで述べた層別分割表の検定にはイェーツ補正は加えられていない。χ^2 検定や単独の Mantel-extension 法との整合性については今後考察する必要がある。

参考文献

1) 新版医学への統計学, 古川俊之監修, 丹後俊郎著, 朝倉書店, 1993.

１９．非線形回帰分析

College Analysis には非線形最小 2 乗法のプログラムが含まれているが、この分析は機能が多い分複雑で、初心者が見た場合、分かりにくく感じるのではないかと思われる。我々はこの欠点をカバーするために、最も簡単な 1 変数についての非線形回帰分析を別に作り直した。特徴としては後に述べるように、候補の非線形関数が選べることである。ここではこの分析について少し紹介しておく。

メニュー［分析－基本統計－非線形回帰分析］を選択すると図 1 のような分析実行画面が表示される。

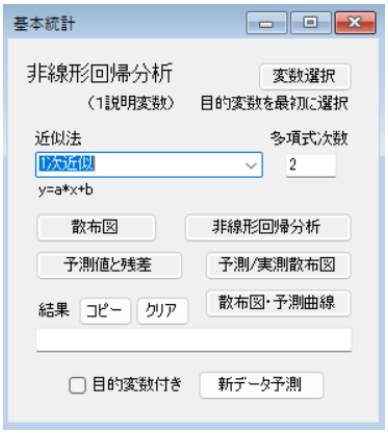


図 1　非線形回帰分析実行画面

使い方は非線形最小 2 乗法と同様であるが、この場合は、近似する関数として予めよく知られた関数が用意されている。例えば、非線形回帰分析 1.txt のデータを用いて、「対数近似」を行った例を図 2 に示す。

y=a*log(x)+b	推定値	標準誤差	z統計量	確率値	95%下限	95%上限
a	55.8599	3.2452	17.2129	0.0000	49.4993	62.2204
b	-312.2585	20.9772	-14.8856	0.0000	-353.3731	-271.1439
実測・予測 R	0.734	R^2	0.539			

図 2　非線形回帰分析実行結果（非線形回帰分析 1.txt）

この結果から、「散布図・予測曲線」を求めた結果を図 3 に示す。

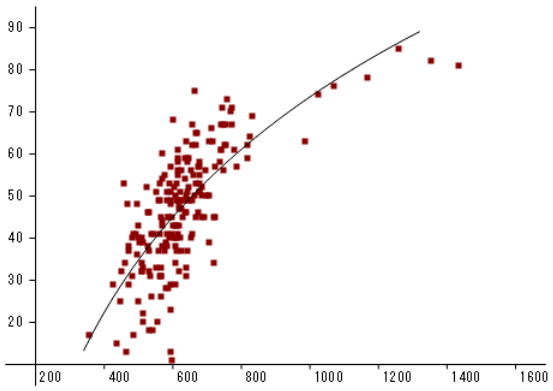


図 3　散布図・予測曲線実行結果

分析実行画面の一番下にある「結果」テキストボックスは、分析のパラメータ推定が終わると図 4 のような状態で、パラメータが代入された式が示されている。

図 4　結果テキストボックス

ここでは小数点以下何桁かは決めず、計算精度そのままで表示している。これはこの式をコピーして他のプログラムで計算した場合、誤差をできるだけ少なく結果を表示するためである。この結果の有効な使い方は 11.2 節 特殊グラフのところで説明する。

予測について

機械学習などでは、訓練データに対して独立なテストデータを用いて目的変数の予測を行う。図 1 の分析実行画面の一番下に「新データ予測」ボタンがあるが、これは一度非線形回帰分析を実行した後、新たなデータを選択して予測を行うためのボタンである。テストデータに目的変数を含む場合は「目的変数付き」チェックボックスにチェックを入れて、「新データ予測」ボタンをクリックする。テストデータで予測値を計算し、適合性を R^2 の値で示してくれる。

２０．対数尤度関数の視覚化

いくつかの変数によるモデルを用いたパラメータの推定の問題では、1990 年ごろまでは、最小 2 乗法が用いられることが多かったが、それ以降は最尤法が多く用いられるようになった。これにはコンピュータの発達が大きく寄与し、最近ではパソコンを使って簡単に最尤法の計算ができるようになった。特に共分散構造分析では、複雑な最尤法のパラメータの推定値も可能になった。しかし、最尤法は初心者にとって直感的に分かり易い手法ではない。そしてその分かりにくさの原因の一つとして考えられるのが、尤度関数のイメージを掴みにくいということである。尤度関数は多くの実測値とパラメータから作られる関数であるが、利用者にはそれが、実測値が与えられた下でのパラメータの関数、と感じられているであろうか。実際最尤法を使っている人でも、尤度関数をパラメータの関数として表示してみた人は少ないだろう。

そこで我々は少しでもイメージをつかめるように、尤度関数をできるだけ簡単に視覚化するツールを作ってみようと考えた。もちろんソフトを使って分析用のパラメータを推定するだけなら、このようなことを考える必要はない。このプログラムの面白さは、「尤度関数はきれいな山のような形をしているのだろうか」、「データ数が多くなるとパラメータの推定の幅は縮まるけれど、尤度関数の形や幅はどうなるんだろう」などと余計なことを考える利用者にこそ伝わると思う。

対数尤度関数を表示する際の困難は、対数尤度関数が多くの座標値で与えられた対数密度関数の合計である点である。例えば、１つの関数の形が $\log f(x;a)$ で与えられる場合を考える。これをパラメータ a の関数と考えると、計算には、まず数式の中に x の値を（文字列として）代入し、その後 a の値を少しずつ変化させて代入してゆくと、その関数形は描画点の個数だけの計算によって容易に描かれる。しかし、尤度関数では以下のように関数形が $\log f(x_\lambda;a)$ の λ についての和になっている。

$$\log L=\sum_{\lambda=1}^{N}\log f(x_\lambda;a)=\log f(x_1;a)+\log f(x_2;a)+\cdots+\log f(x_N;a)$$

このため、最初に x_λ の値を（文字列として）入れると、関数式が長くなり過ぎて計算が困難になる。これを解決するためには、まず a の値を決めて、関数形 $\log f(x_\lambda;a)$ の中に（文字列として）代入し、関数形を定める。その後 1 つずつ x_λ の値を代入して計算、それを足すことで１つの a に対する対数尤度関数の値を定める。これは各 a に対する関数値を求めるために通常の関数計算を N 回行うことになり、データ数が多い場合には計算時間がかかる。

次に、対数尤度関数は対数をとった密度関数を N 回足したものであるので、データ数 N に依存する。そのため、我々は対数尤度関数をそのまま表示する場合と、データ数で対数尤度関数を割った場合を考えてみた。元々我々は対数尤度関数の形を議論する場合、データ 1 つ当たりの形が有効であると考えていた。データの数によって対数尤度関数の大きさが大きく左右されるのは問題があると理由もなしに考えたからである。しかしこれは大きな誤りで

あった。データ数の増加は対数尤度関数の高さを高くする。これが対数尤度関数の尖り方に影響を与え、パラメータ推定の標準誤差に、$1/\sqrt{N}$ で影響を与える。このため対数尤度関数の幅を議論するときには、平均化されたものでなく、本来の対数尤度関数を利用すべきであった。

最後に、グラフを描く際に問題になるのが、いかに簡単にグラフ化するかである。このグラフ化の目的は初学者の最尤法の理解のためのものであるから、表示の手順が困難では意味がない。そのため、分析をいくつかに絞り、利用するデータ数を簡単に選べて、グラフの表示範囲もできれば自動化したい。グラフの推定点（極大点）はデータ数により変化する。そのため、推定点を画面上で固定し、描画範囲を調整することによってグラフの形状を細かく見ることができるようにした。

ここではまず初心者のために、最小 2 乗法と最尤法の違いについて回帰分析を例に説明する。その後、このプログラムの中で使う、正規分布のパラメータの推定とロジスティック回帰分析の最尤法による解法について説明する。次に最尤法の数値計算法について簡単に説明し、計算の過程で求められる推定パラメータの標準誤差について言及する。

また、プログラムの利用法の節では、プログラムを使って最初に述べた尤度関数についての疑問に答えて行く。また最後の節で実際に計算を進めるための計算式をまとめておく。

20.1　回帰分析における最小 2 乗法と最尤法

回帰分析を例に最小 2 乗法と最尤法を考えてみよう。回帰分析は目的変数 y_λ を説明変数 x_λ で以下のように予測する分析である。

$$y_\lambda = ax_\lambda + b + u_\lambda \qquad (\lambda = 1, 2, \cdots, N) \tag{1}$$

ここに u_λ は予測の誤差である。 a, b はパラメータであるが、どのようにして求められるのだろうか。このパラメータの値を求めるには通常最小２乗法という方法が使われる。それにはまず以下のような量 D を考える。

$$D = \sum_{\lambda=1}^{N} (y_\lambda - ax_\lambda - b)^2 = \sum_{\lambda=1}^{N} u_\lambda^2$$

この D は、目的変数値と説明変数値との差の２乗の合計で、言い換えれば予測誤差の２乗の合計である。尤（もっと）もらしいパラメータの値を求めるには、この D が最小（この場合極小）になるようにパラメータの値を決める必要がある。すなわち、この D をパラメータ a と b で微分し、以下のようにそれぞれを 0 にして求める。

$$\frac{\partial D}{\partial a} = -2\sum_{\lambda=1}^{N} x_\lambda (y_\lambda - ax_\lambda - b) = 0 \tag{2}$$

$$\frac{\partial D}{\partial b} = -2\sum_{\lambda=1}^{N} (y_\lambda - ax_\lambda - b) = 0 \tag{3}$$

(3)式から、

$$\sum_{\lambda=1}^{N} y_\lambda - a\sum_{\lambda=1}^{N} x_\lambda - Nb = 0$$

すなわち、

$$b = \bar{y} - a\bar{x}, \quad \text{ここに、} \ \bar{y} = \frac{1}{N}\sum_{\lambda=1}^{N} y_\lambda, \quad \bar{x} = \frac{1}{N}\sum_{\lambda=1}^{N} x_\lambda$$

これを(2)式に代入して、

$$\sum_{\lambda=1}^{N} x_\lambda[(y_\lambda - \bar{y}) - a(x_\lambda - \bar{x})] = 0$$

さらに

$$\sum_{\lambda=1}^{N} \bar{x}[(y_\lambda - \bar{y}) - a(x_\lambda - \bar{x})] = 0$$

という関係を利用すると、以下が得られる。

$$\sum_{\lambda=1}^{N} (x_\lambda - \bar{x})[(y_\lambda - \bar{y}) - a(x_\lambda - \bar{x})] = 0$$

これを整理すると、

$$s_{xy} - as_x^2 = 0, \quad \text{ここに、} \ s_{xy} = \frac{1}{N}\sum_{\lambda=1}^{N} (x_\lambda - \bar{x})(y_\lambda - \bar{y}), \ s_x^2 = \frac{1}{N}\sum_{\lambda=1}^{N} (x_\lambda - \bar{x})^2$$

以上より、

$$\hat{a} = s_{xy} / s_x^2, \quad \hat{b} = \bar{y} - \hat{a}\bar{x} \qquad (4)$$

という答えが得られる。この方法が最小２乗法である。実際の分析ではさらに誤差が平均０、分散σ^2の正規分布に従う（$u_\lambda \sim N(0, \sigma^2)$）という仮定を加えて、推定値$\hat{a}, \hat{b}$の分散や分布について考える。

パラメータ$\hat{a}$については、

$$\hat{a} = s_{xy} / s_x^2 = a + \frac{1}{s_x^2 N}\sum_{\lambda=1}^{N} (x_\lambda - \bar{x})\varepsilon_\lambda$$ より、正規分布となるがその平均と分散は以下で与えられる。

$$E[\hat{a}] = a + \frac{1}{s_x^2 N}\sum_{\lambda=1}^{N} (x_\lambda - \bar{x})E[\varepsilon_\lambda] = a$$

$$V[\hat{a}] = \frac{1}{s_x^4 N^2}\sum_{\lambda=1}^{N}\sum_{\lambda'=1}^{N} (x_\lambda - \bar{x})(x_{\lambda'} - \bar{x})E[\varepsilon_\lambda \varepsilon_{\lambda'}] = \frac{1}{s_x^4 N^2}\sum_{\lambda=1}^{N} (x_\lambda - \bar{x})^2 \sigma^2 = \frac{\sigma^2}{s_x^2 N}$$

ここに、σ^2については以下で与えられる$\hat{\sigma}^2$で推測する。

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} \hat{u}_\lambda^2 = \frac{1}{N}\sum_{i=1}^{N} (y_\lambda - \hat{a}x_\lambda - \hat{b})^2$$

パラメータ$\hat{b}$についても、

$$\hat{b} = \bar{y} - \hat{a}\bar{x} = a\bar{x} + b + \bar{\varepsilon} - \hat{a}\bar{x} = b - (\hat{a} - a)\bar{x} + \bar{\varepsilon}$$
$$= b - \frac{\bar{x}}{s_x^2 N}\sum_{\lambda=1}^{N}(x_\lambda - \bar{x})\varepsilon_\lambda + \bar{\varepsilon}$$

より、正規分布となるがその平均と分散は以下で与えられる。

$$E[\hat{b}] = b$$
$$V[\hat{b}] = V\left[\frac{\bar{x}}{s_x^2 N}\sum_{\lambda=1}^{N}(x_\lambda - \bar{x})\varepsilon_\lambda - \bar{\varepsilon}\right]$$
$$= \bar{x}^2 V[\hat{a}] + V[\bar{\varepsilon}] - \frac{2\bar{x}}{s_x^2 N}\sum_{\lambda=1}^{N}(x_\lambda - \bar{x})E[\varepsilon_\lambda \bar{\varepsilon}]$$
$$= \frac{\bar{x}^2\sigma^2}{s_x^2 N} + \frac{\sigma^2}{N} - \frac{2\bar{x}}{s_x^2 N}\sum_{\lambda=1}^{N}(x_\lambda - \bar{x})\frac{\sigma^2}{N} = \frac{\bar{x}^2\sigma^2}{s_x^2 N} + \frac{\sigma^2}{N} = \frac{(\bar{x}^2 + s_x^2)\sigma^2}{s_x^2 N}$$

次に、最尤法を考えてみよう。今度は最初に誤差u_λの分布を考える。ここでは一般的に確率変数u_λの密度関数を$f(u_\lambda, c)$とする。ここにcはその確率変数の分布に特有のパラメータである。ここでは１つだけ書いたが、もちろん複数のパラメータを持つ場合もある。密度関数は確率的に最も起こり易いところで最大の値を取る。

この密度関数を使って、以下の尤度(likelyhood)関数と呼ばれるものを作る。

$$L = \prod_{\lambda=1}^{N} f(u_\lambda; c) = \prod_{\lambda=1}^{N} f(y_\lambda - a x_\lambda - b, c) \qquad (5)$$

これには実測データy_λ, x_λが含まれているが、尤度関数は、実測データが最も起こり易いパラメータの値を取ったときに最大（極大）の値となる。我々はこの性質を使ってパラメータを推測することができる。そのために以下の微分を考え、これをパラメータの連立方程式とみなして最小２乗法のときのように解を求める。

$$\frac{\partial L}{\partial a} = 0,\ \frac{\partial L}{\partial b} = 0,\ \frac{\partial L}{\partial c} = 0$$

分布の種類にもよるが、この計算は一般に数式を用いては難しいので、コンピュータを利用することになる。これが最尤法と呼ばれる方法である。最尤法の利点はコンピュータを使うため、分布の種類によらず計算が可能となるところである。また、密度関数を用いるため、確率的な解釈がし易く、最も確率が大きくなるパラメータの値ということで、納得できる結果が得られる点である。

原理的には尤度関数Lをそのまま用いて計算は可能であるが、一般に、尤度関数は多くの 1 未満の数を掛けることになるので、値が非常に小さくなる可能性がある。数値計算では値が極端に小さくなると、計算誤差が問題になり正確な数値が求められなくなることがある。そこで尤度関数Lをそのまま使うのではなく、極大が同じパラメータの値になる対数尤度関数$\log L$がよく利用される。即ち、以下となる。

$$\log L = \log \prod_{\lambda=1}^{N} f(y_\lambda - ax_\lambda - b, c) = \sum_{i=1}^{N} \log f(y_\lambda - ax_\lambda - b, c) \tag{6}$$

$$\frac{\partial \log L}{\partial a} = 0,\ \frac{\partial \log L}{\partial b} = 0,\ \frac{\partial \log L}{\partial c} = 0$$

さて、一般に数式を用いては難しいと言ったが、計算できる特別な場合を示しておく。それは上でも述べた誤差が正規分布 $u_\lambda \sim N(0, \sigma^2)$ の場合である。

平均 0 で、分散 σ^2 の正規分布の密度関数は、c を σ と考えて以下のように表される。

$$f(u_\lambda; \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-u_\lambda^2 / 2\sigma^2\right] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-(y_\lambda - ax_\lambda - b)^2 / 2\sigma^2\right]$$

ここで、$\exp(x)$ は、ネーピアの数 e (=1.71828...)を用いた指数関数 e^x のことである。これを使って尤度関数を作ってみよう。

$$L = \prod_{\lambda=1}^{N} f(u_\lambda; \sigma) = \frac{1}{(2\pi)^{N/2}\sigma^N} \prod_{\lambda=1}^{N} \exp\left[-(y_\lambda - ax_\lambda - b)^2 / 2\sigma^2\right]$$

この尤度関数を使うと対数尤度関数は以下のようになる。

$$\log L = \sum_{\lambda=1}^{N} \log f(u_\lambda; \sigma) = -\frac{1}{2\sigma^2} \sum_{\lambda=1}^{N} (y_\lambda - ax_\lambda - b)^2 - N \log \sigma - \frac{N}{2} \log(2\pi)$$

この対数尤度関数をパラメータで微分する。

$$\frac{\partial}{\partial a} \log L = \frac{1}{\sigma^2} \sum_{\lambda=1}^{N} x_\lambda (y_\lambda - ax_\lambda - b) = 0$$

$$\frac{\partial}{\partial b} \log L = \frac{1}{\sigma^2} \sum_{\lambda=1}^{N} (y_\lambda - ax_\lambda - b) = 0$$

$$\frac{\partial}{\partial \sigma} \log L = \frac{1}{\sigma^3} \sum_{\lambda=1}^{N} (y_\lambda - ax_\lambda - b)^2 - \frac{N}{\sigma} = 0$$

下の２つの式は最小 2 乗法の制約式と同じで同じ結果を与える。また、最後の式から σ の推定値が求まる。

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{\lambda=1}^{N} (y_\lambda - \hat{a}x_\lambda - \hat{b})^2 = \frac{1}{N} \sum_{\lambda=1}^{N} \hat{u}_\lambda^2 \tag{13}$$

これは推定された誤差の分散の値になる。

最尤法ではパラメータの分布は多変量正規分布となるが、その分散は、以下で与えられる情報行列 $\mathcal{J}$ の逆行列の対角成分で与えられることが知られている。

$$\mathcal{J} = -\begin{pmatrix} \dfrac{\partial^2}{\partial a^2} \log L & \dfrac{\partial^2}{\partial a \partial b} \log L \\ \dfrac{\partial^2}{\partial a \partial b} \log L & \dfrac{\partial^2}{\partial b^2} \log L \end{pmatrix}$$

これを計算してみよう。

$$\frac{\partial^2}{\partial a^2} \log L = \frac{1}{\sigma^2} \frac{\partial}{\partial a} \sum_{\lambda=1}^{N} x_\lambda (y_\lambda - ax_\lambda - b) = -\frac{1}{\sigma^2} \sum_{\lambda=1}^{N} x_\lambda^2 \equiv -\frac{N(s_x^2 + \bar{x}^2)}{\sigma^2}$$

$$\frac{\partial^2}{\partial b^2}\log L=\frac{1}{\sigma^2}\frac{\partial}{\partial b}\sum_{\lambda=1}^{N}(y_\lambda-ax_\lambda-b)=\frac{-N}{\sigma^2}$$

$$\frac{\partial^2}{\partial a\partial b}\log L=\frac{1}{\sigma^2}\frac{\partial}{\partial a}\sum_{\lambda=1}^{N}(y_\lambda-ax_\lambda-b)=\frac{-1}{\sigma^2}\sum_{\lambda=1}^{N}x_\lambda=-\frac{N\overline{x}}{\sigma^2}$$

これより、

$$\mathcal{I}=\frac{N}{\sigma^2}\begin{pmatrix}s_x^2+\overline{x}^2 & \overline{x}\\ \overline{x} & 1\end{pmatrix}$$

$$\mathcal{I}^{-1}=\frac{\sigma^2}{N((s_x^2+\overline{x}^2)-\overline{x}^2)}\begin{pmatrix}1 & -\overline{x}\\ -\overline{x} & s_x^2+\overline{x}^2\end{pmatrix}=\frac{\sigma^2}{Ns_x^2}\begin{pmatrix}1 & -\overline{x}\\ -\overline{x} & s_x^2+\overline{x}^2\end{pmatrix}$$

これより以下を得る。

$$V[\hat{a}]=\frac{\sigma^2}{Ns_x^2},\qquad V[\hat{b}]=\frac{\sigma^2(\overline{x}^2+s_x^2)}{Ns_x^2}$$

これは最小 2 乗法で得た結果と一致する。

20.2　正規分布のパラメータ推定

もう一つ最尤推定法の簡単な例を考えてみる。我々は平均 μ と分散 σ^2 を推測する場合に以下のような推定量を用いる。

$$\overline{x}=\frac{1}{N}\sum_{\lambda=1}^{N}x_\lambda\to\mu,\qquad s^2=\frac{1}{N}\sum_{\lambda=1}^{N}(x_\lambda-\overline{x})^2\to\sigma^2 \tag{1}$$

これは、確率密度関数が大きな値をとる領域からはデータが多く集まることを考えると、補遺 1 の(A1.1)や(A1.2)の定義と矛盾しないが、この妥当性を、正規分布に基づく最尤法を使って考えてみる。まず、変数 x_λ が正規分布 $N(\mu,\sigma^2)$ に従うとするとその確率密度関数は以下となる。

$$f(x_\lambda;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-(x_\lambda-\mu)^2/2\sigma^2\right]$$

これを使って尤度関数 L と対数尤度関数 $\log L$ を作ると以下となる。

$$L=\prod_{\lambda=1}^{N}f(x_\lambda;\mu,\sigma)=\frac{1}{(2\pi)^{N/2}\sigma^N}\prod_{\lambda=1}^{N}\exp\left[-(x_\lambda-\mu)^2/2\sigma^2\right] \tag{2}$$

$$\log L=\sum_{\lambda=1}^{N}\log f(x_\lambda;\mu,\sigma)=-\frac{1}{2\sigma^2}\sum_{\lambda=1}^{N}(x_\lambda-\mu)^2-N\log\sigma-\frac{N}{2}\log(2\pi) \tag{3}$$

これから、パラメータ μ,σ を推定してみよう。

$$\partial\log L/\partial\mu=\frac{1}{\sigma^2}\sum_{\lambda=1}^{N}(x_\lambda-\mu)=0\qquad\qquad \mu=\frac{1}{N}\sum_{\lambda=1}^{N}x_\lambda$$

$$\partial \log L/\partial \sigma = \frac{1}{\sigma^3}\sum_{\lambda=1}^{N}(x_\lambda-\mu)^2 - \frac{N}{\sigma} = 0 \qquad \sigma^2 = \frac{1}{N}\sum_{\lambda=1}^{N}(x_\lambda-\mu)^2$$

ここでは(1)で与えた推定量の通りの結果が得られている。

20.3　ロジスティック回帰分析のパラメータ推定

よく使われる 2 値ロジスティック回帰分析とは、事象の出現確率を説明変数の 1 次式のロジスティック関数で推測するモデルである。説明変数が 1 つの 2 値ロジスティック回帰分析の事象の出現確率 p_λ は以下で与えられる。

$$p_\lambda = \frac{e^{z_\lambda}}{1+e^{z_\lambda}} \quad , \quad z_\lambda = ax_\lambda + b \qquad (1)$$

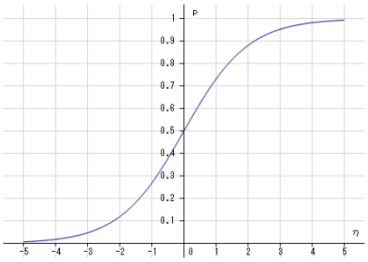


図 1 ロジスティック関数

また、事象の出現と非出現を $y_\lambda = \{1, 0\}$ で表すとその尤度関数は(1)式の p_λ を用いて、以下で与えられる。これはベルヌーイ分布の確率関数を掛けたものである。

$$L = \prod_{\lambda=1}^{N} p_\lambda^{y_\lambda}(1-p_\lambda)^{1-y_\lambda}$$

ここにロジスティック関数の形状は図 1 で与えられ、変域が $(-\infty,\infty)$ で、値域が $(0,1)$ であり、確率の値域と一致している。また対数尤度関数は以下で与えられる。

$$\log L = \sum_{\lambda=1}^{N} y_\lambda \log p_\lambda + \sum_{\lambda=1}^{N}(1-y_\lambda)\log(1-p_\lambda)$$

この式に(1)式を代入した式は補遺に譲るが、これを微分してパラメータ a,b を求める計算は式の上では不可能で、次節で述べる数値計算を必要とする。

20.4　最尤法の数値計算

対数尤度関数作成後、ニュートン・ラフソン法を用いてパラメータの推定を行うが、その計算を簡単に示しておく。但しパラメータは２つと仮定し、以下のような行列表現を用いる。

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

対数尤度をパラメータで微分してスコアベクトル $\mathbf{U}$ と情報行列 $\mathbf{J}$ を求めると以下となる。

$$\mathbf{U} = \begin{pmatrix} \partial \log L/\partial b_1 \\ \partial \log L/\partial b_2 \end{pmatrix}, \quad \mathbf{J} = -\begin{pmatrix} \partial^2 \log L/\partial b_1^2 & \partial^2 \log L/\partial b_1 \partial b_2 \\ \partial^2 \log L/\partial b_2 \partial b_1 & \partial^2 \log L/\partial b_2^2 \end{pmatrix}$$

これらを用いてニュートン・ラフソン法でパラメータの推定を行う。ニュートン・ラフソン法は以下のような計算ステップを繰り返し、パラメータの推定値を求める。

$$\mathbf{b}^{(m+1)} = \mathbf{b}^{(m)} + (\mathbf{J}^{(m)})^{-1}\mathbf{U}^{(m)}$$

ここで右肩の (m) や $(m+1)$ は、ニュートン・ラフソン法の計算のステップを表す。ニュートン・ラフソン法では、最初にパラメータの推定値に近い値を何らかの方法で与えて $\mathbf{b}^{(0)}$ と

し、それを用いて $\mathbf{U}^{(0)}, \mathbf{J}^{(0)}$ を計算し、上の繰り返し計算を実行する。繰り返し計算は、$\mathbf{b}^{(m)}$ の値が落ち着いたら終了する。また、この情報行列 $\mathbf{J}$ の逆行列 $\mathbf{J}^{-1}$ の対角成分はパラメータの推定値の分散を与えることが知られている。

20.5 標準誤差の幾何的意味

ここでは対数尤度関数の形状の特徴を一般的に見てみよう。2 変数関数の形状は推定値（極値点）$(\hat{a}, \hat{b})$ の近傍で以下のように与えられる。

$$G = \log L \simeq \frac{1}{2}(a-\hat{a})^2 G_{aa} + (a-\hat{a})(b-\hat{b})G_{ab} + \frac{1}{2}(b-\hat{b})^2 G_{bb} + c_0$$

ここに、c_0 は対数尤度関数の頂点の高さである。また、

$$G_{aa} = \partial^2 G/\partial a^2\Big|_{a=\hat{a},b=\hat{b}},\quad G_{bb} = \partial^2 G/\partial b^2\Big|_{a=\hat{a},b=\hat{b}},\quad G_{ab} = \partial^2 G/\partial a\partial b\Big|_{a=\hat{a},b=\hat{b}}$$

これを利用すると、情報行列 $\mathbf{J}$ とその逆行列 $\mathbf{J}^{-1}$ は以下のように求められる。

$$\mathbf{J} = -\begin{pmatrix} G_{aa} & G_{ab} \\ G_{ab} & G_{bb} \end{pmatrix},\qquad \mathbf{J}^{-1} = \frac{1}{|\mathbf{J}|}\begin{pmatrix} -G_{bb} & G_{ab} \\ G_{ab} & -G_{aa} \end{pmatrix}$$

情報行列の行列式 $|\mathbf{J}|$ は以下となる。

$G_{ab} = 0$ のとき、$|\mathbf{J}| = G_{aa}G_{bb}$

$G_{ab} \neq 0$ のとき、$|\mathbf{J}| = G_{aa}G_{bb} - G_{ab}^2 = (1-\rho^2)G_{aa}G_{bb}$

ここに、$\rho = G_{ab}/\sqrt{G_{aa}G_{bb}}$

情報行列の逆行列はパラメータの分散、共分散行列になるから、ρ はパラメータの相関係数である。

ここで、パラメータの標準誤差を $\hat{\sigma}_a, \hat{\sigma}_b$ とすると、

$$\hat{\sigma}_a^2 = -G_{bb}/|\mathbf{J}| = -1/[(1-\rho^2)G_{aa}]$$

$$\hat{\sigma}_b^2 = -G_{aa}/|\mathbf{J}| = -1/[(1-\rho^2)G_{bb}]$$

これらより、

$$G_{aa} = -1/(1-\rho^2)\hat{\sigma}_a^2$$

$$G_{bb} = -1/(1-\rho^2)\hat{\sigma}_b^2$$

$$G_{ab} = \rho/(1-\rho^2)\hat{\sigma}_a\hat{\sigma}_b$$

以上のことから、対数尤度関数 $\log L$ は以下となる。

$$\log L = -\frac{(a-\hat{a})^2}{2(1-\rho^2)\hat{\sigma}_a^2} + \frac{\rho(a-\hat{a})(b-\hat{b})}{(1-\rho^2)\hat{\sigma}_a\hat{\sigma}_b} - \frac{(b-\hat{b})^2}{2(1-\rho^2)\hat{\sigma}_b^2} + c_0$$

次に、特別な点での対数尤度関数の値について見てみる。上の結果から、

$a = \hat{a} \pm 1/\sqrt{-G_{aa}} = \hat{a} \pm \sqrt{1-\rho^2}\hat{\sigma}_a,\ b = \hat{b}$ のとき、　$\log L \simeq -1/2 + c_0$

$a = \hat{a},\ b = \hat{b} \pm 1/\sqrt{-G_{bb}} = \hat{b} \pm \sqrt{1-\rho^2}\hat{\sigma}_b$ のとき、　$\log L \simeq -1/2 + c_0$

$a = \hat{a} + \sqrt{1-\rho^2}\hat{\sigma}_a t,\ b = \hat{b} \pm \sqrt{1-\rho^2}\hat{\sigma}_b t$ のとき、　$\log L \simeq -(1 \mp \rho)t^2 + c_0$

上の範囲は広い可能性があるので、正確には範囲を $\alpha\ (0 < \alpha \ll 1)$ 倍にした場合、対数尤度関数は（上の値$\times\alpha^2 + c_0$）になるということであるが、以後、標準誤差は十分小さいとして上の式をそのまま使うことにする。

次に、対数尤度関数の頂点での軸に沿った曲率を考えてみよう。今、中心が $(\hat{a},\hat{b})$ でパラメータ a の軸に沿った半径 R_a の半円形の関数を考える。

$$f(a)=\sqrt{R_a^2-(a-\hat{a})^2}$$

この２回微分は、

$$\frac{d^2}{da^2}f(a)=-\frac{1}{\sqrt{R_a^2-(a-\hat{a})^2}}-\frac{(a-\hat{a})^2}{[R_a^2-(a-\hat{a})^2]^{3/2}}$$

これより $a=\hat{a}$ のところでは、

$$\left.\frac{d^2}{da^2}f(a)\right|_{a=\hat{a}}=-\frac{1}{R_a}$$

これと $G_{aa}=-1/(1-\rho^2)\hat{\sigma}_a^2$ より、頂点での曲率を合わせると、

$$R_a=-1/G_{aa}=(1-\rho^2)\hat{\sigma}_a^2$$

同様にして、

$$R_b=-1/G_{bb}=(1-\rho^2)\hat{\sigma}_b^2$$

これより、パラメータ間の相関がない場合、対数尤度関数の頂点での軸に沿った曲率半径は標準誤差の 2 乗になる。ここで少し例を考えてみよう。

正規分布のパラメータ推定

正規分布のパラメータ推定では、2 章で述べたように、平均と分散の推定値は以下のように与えられる。

$$\hat{\mu}=\frac{1}{N}\sum_{\lambda=1}^{N}x_\lambda=\bar{x},\qquad \hat{\sigma}^2=\frac{1}{N}\sum_{\lambda=1}^{N}(x_\lambda-\mu)^2=s^2$$

また、情報行列 $\mathbf{J}$ とその逆行列 $\mathbf{J}^{-1}$ は以下で与えられる。

$$\mathbf{J}=\begin{pmatrix}N/s^2 & 0\\ 0 & 2N/s^2\end{pmatrix},\qquad \mathbf{J}^{-1}=\begin{pmatrix}s^2/N & 0\\ 0 & s^2/2N\end{pmatrix}$$

対数尤度関数は、特別の点で以下の値をとる。

$\mu=\bar{x}\pm s/\sqrt{N}$, $\sigma=s$ のとき、　　$\log L\simeq -1/2+c_0$

$\mu=\bar{x}$, $\sigma=s\pm s/\sqrt{2N}$ のとき、　　$\log L\simeq -1/2+c_0$

$\mu=\bar{x}+ts/\sqrt{N}$, $\sigma=s\pm ts/\sqrt{2N}$ のとき、　　$\log L\simeq -t^2+c_0$

回帰分析におけるパラメータ推定

回帰分析のパラメータ推定では、3 章で述べたように、パラメータの推定値は以下のように与えられる。

$$\hat{a}=\sum_{\lambda=1}^{N}(y_\lambda-\bar{y})(x_\lambda-\bar{x})\Big/\sum_{\lambda=1}^{N}(x_\lambda-\bar{x})^2=s_{xy}/s_x^2$$

$$\hat{b}=\bar{y}-\hat{a}\bar{x}$$

$$\hat{\sigma}^2=\frac{1}{N}\sum_{\lambda=1}^{N}(y_\lambda-\hat{a}x_\lambda-\hat{b})^2=\frac{1}{N}\sum_{\lambda=1}^{N}\hat{u}_\lambda^2$$

情報行列 $\mathbf{J}$ とその逆行列 $\mathbf{J}^{-1}$ は以下で与えられる。

$$\mathbf{J}=\frac{N}{\hat{\sigma}^2}\begin{pmatrix} s_x^2+\bar{x}^2 & \bar{x} & 0 \\ \bar{x} & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \qquad \mathbf{J}^{-1}=\frac{\sigma^2}{Ns_x^2}\begin{pmatrix} 1 & -\bar{x} & 0 \\ -\bar{x} & s_x^2+\bar{x}^2 & 0 \\ 0 & 0 & 1/2 \end{pmatrix}$$

ここでパラメータ a, b, σ の標準誤差と a, b の相関係数は以下となるが、a, b と σ との相関係数は 0 である。

$$\hat{\sigma}_a=\frac{\hat{\sigma}}{\sqrt{N}s_x}, \quad \hat{\sigma}_b=\frac{\hat{\sigma}\sqrt{s_x^2+\bar{x}^2}}{\sqrt{N}s_x},$$

$$\hat{\sigma}=\sqrt{\frac{1}{N}\sum_{\lambda=1}^{N}(y_\lambda-\hat{a}x_\lambda-\hat{b})^2}=\sqrt{\frac{1}{N}\sum_{\lambda=1}^{N}\hat{u}_\lambda^2}$$

$$\rho=-\bar{x}\Big/\sqrt{s_x^2+\bar{x}^2}$$

対数尤度関数は、特別の点で以下の値をとる。

$a=\hat{a}\pm\sqrt{1-\rho^2}\hat{\sigma}_a,\ b=\hat{b},\ \sigma=\hat{\sigma}$ のとき、 $\log L\simeq -1/2+c_0$

$a=\hat{a},\ b=\hat{b}\pm\sqrt{1-\rho^2}\hat{\sigma}_b,\ \sigma=\hat{\sigma}$ のとき、 $\log L\simeq -1/2+c_0$

$a=\hat{a}+\sqrt{1-\rho^2}\hat{\sigma}_a t,\ b=\hat{b}\pm\sqrt{1-\rho^2}\hat{\sigma}_b t,\ \sigma=\hat{\sigma}$ のとき、 $\log L\simeq -(1\mp\rho)t^2+c_0$

20.6　プログラムの利用法

メニュー［分析－基本統計－ユーティリティー尤度関数グラフ］を選択すると図 1 の分析実行画面が表示される。

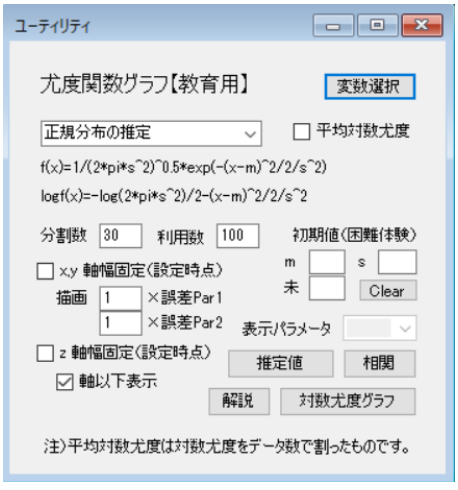


図 1　分析実行画面

このプログラムは実用の計算を行うものではなく、最尤法の尤度関数について視覚的に体験するものである。そのため、利用できる分析は、「正規分布の推定」、「回帰分析」及び、「0/1 ロジスティック回帰分析」の 3 つの推定問題である。さらに、回帰分析は説明変数が

１つ、0/1 ロジスティック回帰分析は目的変数が 0/1 のデータで、説明変数が１つという制約が付く。実用的な問題は、多変量解析の中の「重回帰分析」や「ロジスティック回帰分析」で解決できる。これは真に教育的なプログラムである。

まず始めに正規分布のパラメータ推定の問題を考える。図 2 に示すデータ（尤度関数グラフ 1.txt）の変数「y」を利用し、最尤法を用いて平均値と標準偏差を推定する問題を考える。

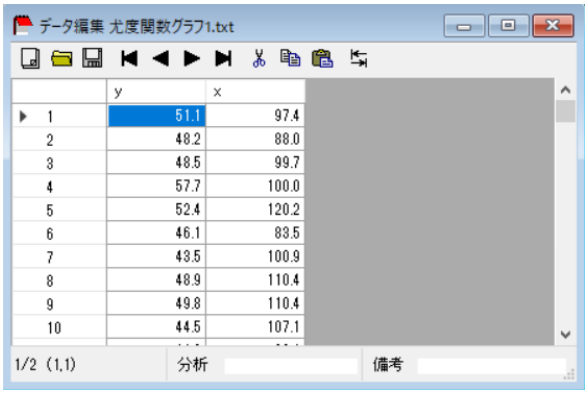


図 2　正規分布のパラメータの推定データ

データの数は 500 あるが、まず先頭から 100 個使って最尤法の結果を求める。分析実行メニューの「変数選択」で「y」を選択し、「利用数」をデフォルトの 100 に設定する。「推定値」ボタンをクリックすると、図 3 に示す推定結果を得る。

図 3　平均と標準偏差の推定結果

ここでは、結果が分かっているので、その結果の近くをニュートン・ラフソン法の初期値にしている。真にニュートン・ラフソン法を体験する希望があれば、「初期値（困難体験）」の部分に初期値を与えて実行してみることをお勧めする。正しい結果を得るのにかなり苦労すると思う。特に最尤法による回帰分析のパラメータ推定などは至難である。我々は同じ結果が出るものは最も結果が出しやすい手法を使うことを基本にし、最尤法でどうしても難しい場合は、MCMC 乱数発生などで、尤度関数が最も大きくなる点の近傍を求めて初期値にしている。これは紙上の計算通りにはいかない数値計算の難しいところである。

データを 100 個使った、この問題の対数尤度関数を描いてみよう。x 軸に平均値のパラメータ、y 軸に標準偏差のパラメータを取り、対数尤度関数の表示領域として図 3 で与えた結果を使い、「推定値±1×標準誤差」の領域をとる。幅が標準誤差の何倍かは、分析実行メニューの「領域」のところで設定できる。ここではデフォルトの 1 としている。図 4 に「対数尤度グラフ」ボタンをクリックした結果を示す。

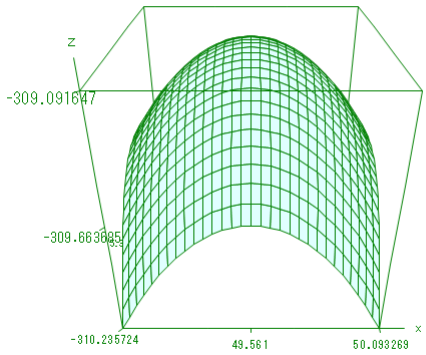


図 4　データ数 100 個の対数尤度関数グラフ

これは推定値を中心とした素直な山型をしている。

次に、データ数を減らした場合や増やした場合、この形がどのように変わるか見てみよう。データ数 50 個、200 個の対数尤度関数を描いてみる。z 軸の幅を不変にして、結果を図 5 と図 6 に示す。

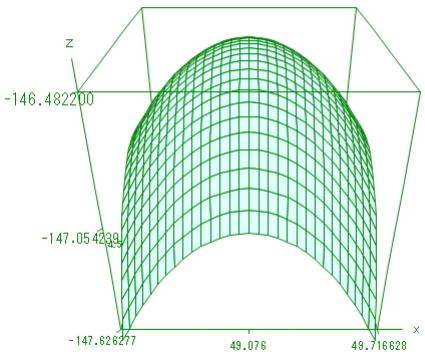


図 5　対数尤度関数（50 個）

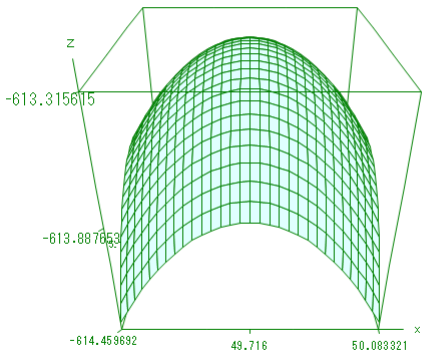


図 6　対数尤度関数（200 個）

これを見ると、z 軸の幅は変わっていないが、横軸の幅は、推定値の標準誤差の値に設定してあるので、標準誤差の値に応じて変わっている。

ここで、実際の標準誤差の値を求めておこう。結果を表 1 に示す。

表 1　データによる標準誤差の変化

	標準誤差（50 個）	標準誤差（100 個）	標準誤差（200 個）
平均(m)	0.6406	0.5323	0.3673
標準偏差(s)	0.4530	0.3764	0.2597

これを見ると、パラメータ推定の標準誤差はデータが増えるに従って表 1 のように小さくなって行く。そのため、対数尤度関数のグラフは、データ数が増えるに従って細くなってい

くことが分かる。

これをはっきりと見るため、図 4 で与えた z 軸と横軸の幅を固定してグラフを描いてみる。結果を図 7 と図 8 に示す。

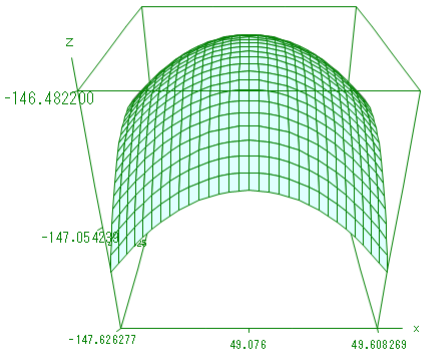


図 7　対数尤度関数（50 個）

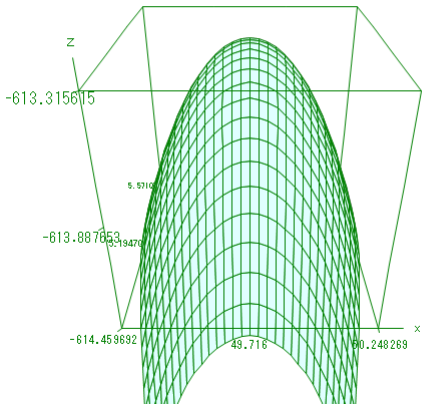


図 8　対数尤度関数（200 個）

以上述べたように、標準誤差の値は対数尤度関数の曲がり方（尖り方）で与えられることが分かる。

次に描画の範囲を拡げて、推定値±5×標準誤差の領域をとる。データ数 100 で描かれる対数尤度関数は図 9 の通りである。

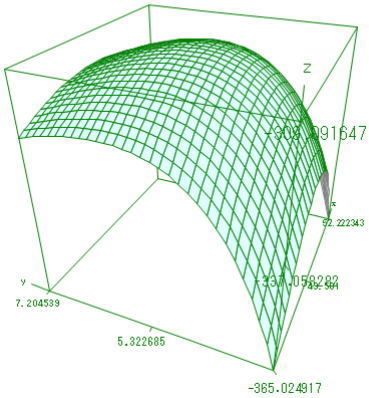


図 9　領域を拡げた対数尤度関数（100 個）

推定値±1×標準誤差の範囲では平均と分散が同じような形状を取っていたが、範囲を拡げると違いがはっきり見えてくる。標準偏差は推定値より大きな領域と小さな領域で非対称である。

次に、回帰分析のパラメータ推定について調べてみる。まず分析の種類を「回帰分析」に設定する。図 2 のデータをそのまま用いて、目的変数を「y」、説明変数を「x」として最尤

法で推定する。パラメータは回帰係数 2 つと誤差の標準偏差１つで３つである。パラメータの推定値は数式から計算できるので、簡単のため、その結果を初期値としている。初期値を「初期値（困難体験）」で別に指定する場合は、なかなかよい結果が得られないことを覚悟する必要がある。推定結果を図 10 に示す。

パラメータの最尤推定

	推定値	標準誤差	2.5%下限	2.5%上限
a	0.2364	0.0191	0.1990	0.2738
b	26.1178	1.9195	22.3555	29.8801
s	4.3861	0.1387	4.1142	4.6579

図 10　回帰のパラメータと誤差標準偏差の推定結果

対数尤度関数はパラメータの中から 2 組を選んで表示する。まず、分析実行メニューの「表示パラメータ」で回帰係数の２つ「a,b」を選択する。もう１つのパラメータは推定値をとるものとして計算する。表示範囲を推定値±1×標準誤差として、データの「利用数」を 100 個で対数尤度関数を表示すると図 11 のようになる。

我々は対数尤度関数が山のような形になっているものと考えていたが、これは壁のような形状である。この壁はこのサイズではどこが推定点か見分けがつかない。しかし、最尤法で答えが出ている以上この図の中央が推定点である。そのため、z 軸幅を固定して、x,y 軸の幅を推定値±5×標準誤差のように拡げてみる。見やすくするため「分割数」を 50 に増やし、「軸以下表示」のチェックを外した結果を図 12 に示す。

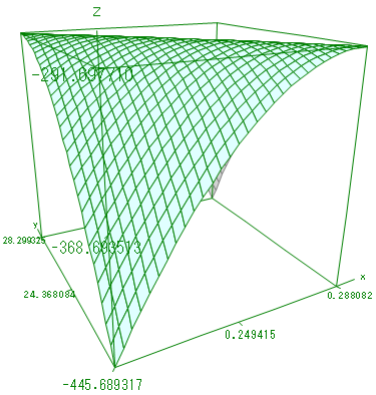


図 11　回帰係数の対数尤度関数

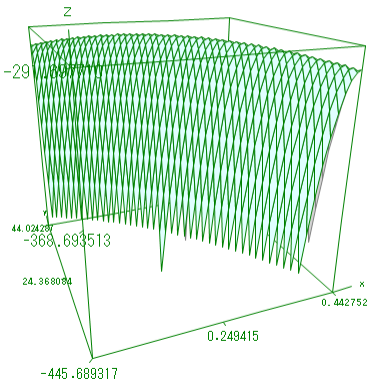


図 12　回帰係数の推定点

これをみると稜線がわずかにカーブしており、中央が最も高いことが分かる。回帰係数の対数尤度関数は壁のような形状であるが、やはり高い部分が存在することが分かった。この壁は描画領域に関わらず存在する。

この結果は、5 節からパラメータの相関係数によっていると考えられるので、この相関係数 ρ と、相関がある場合の描画範囲に掛ける係数 $\sqrt{1-\rho^2}$ を求めてみる。実行画面の「相関」ボタンをクリックすると、図 13 のような結果が表示される。

	a	b	s
a	1.0000	−0.9935	0.0000
b	−0.9935	1.0000	0.0000
s	0.0000	0.0000	1.0000
描画範囲	aとbに掛ける	0.1138	

図 13　相関係数と描画範囲に掛ける係数

確かに相関係数は -1 に近い値となっている。そこで、「範囲」を±0.1×0.1138×標準誤差、「z 軸幅」を 0.1^2 として、データ数 100 の対数尤度関数を描いてみる。結果を図 14 に示す。

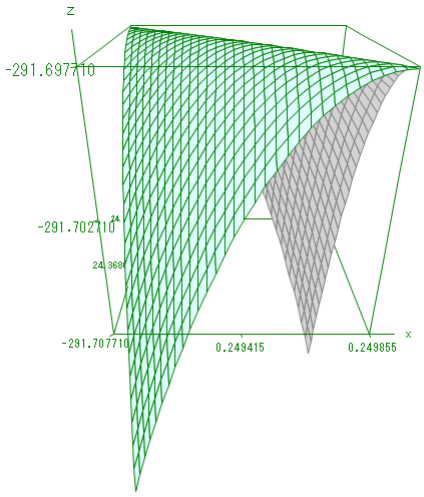


図 14　相関のあるパラメータ間の対数尤度関数

これは 6 節で述べた結果とよく一致している。

最後に表示パラメータを「a,s」に設定し、「範囲」を推定値±1×標準誤差として対数尤度関数を表示すると図 15 のようになる。

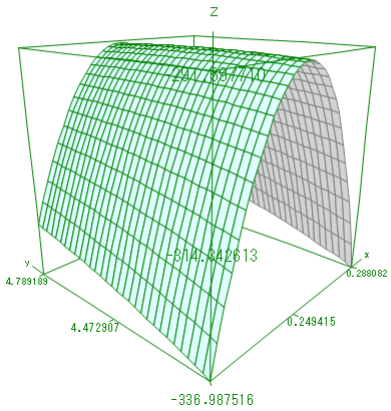


図 15　回帰係数と誤差標準偏差の対数尤度関数

これは回帰係数の対数尤度関数と同じような壁の構造であるが、方向が誤差標準偏差の方向に連なる壁である。この形状は表示範囲と関係があると思われるので、「範囲」をパラメータ a は±0.1×0.1138×標準誤差、パラメータ s は±0.1×標準誤差、z 軸幅を 0.1^2 に固定

して対数尤度関数を表示する。結果を図 16 に示す。

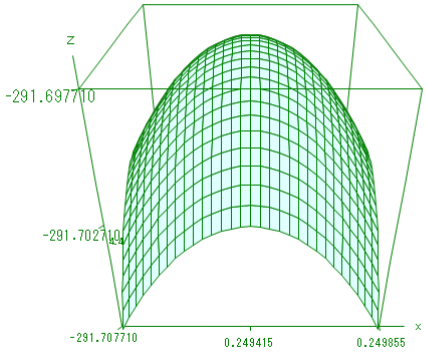


図 16　前節のように設定した回帰係数と誤差標準偏差の対数尤度関数

これを見ると、パラメータ間の相関が 0 の形状で、6 節で述べた結果と一致している。ロジスティック回帰分析の対数尤度関数については、回帰分析と同様であるので省略する。

20.7　最尤法の理論のまとめ

正規分布の推定

ここでは正規分布のパラメータを最尤法で推定するための方法をまとめておく。

密度関数：$f(x\,|\,\mu,\sigma^2)=\dfrac{1}{\sqrt{2\pi\sigma^2}}\exp[-(x-\mu)^2/2\sigma^2]$

尤度関数：$L=\dfrac{1}{(2\pi\sigma^2)^{N/2}}\exp\left[-\dfrac{1}{2\sigma^2}\sum_{\lambda=1}^{N}(x_\lambda-\mu)^2\right]$

対数尤度：$\log L=-\dfrac{1}{2\sigma^2}\sum_{\lambda=1}^{N}(x_\lambda-\mu)^2-\dfrac{N}{2}\log(2\pi\sigma^2)$

スコアベクトル$\mathbf{U}$と情報行列$\mathbf{J}$　（$\mathbf{b}=\boldsymbol{\beta}+\mathbf{J}^{-1}\mathbf{U}\sim N(\boldsymbol{\beta},\mathbf{J}^{-1})$）

$$\boldsymbol{\beta}=\begin{pmatrix}\mu\\ \sigma\end{pmatrix},\qquad \mathbf{U}=\begin{pmatrix}\partial\log L/\partial\mu\\ \partial\log L/\partial\sigma\end{pmatrix},\qquad \mathbf{J}=-\begin{pmatrix}\partial^2\log L/\partial\mu^2 & \partial^2\log L/\partial\mu\partial\sigma\\ \partial^2\log L/\partial\mu\partial\sigma & \partial^2\log L/\partial\sigma^2\end{pmatrix}$$

$$\partial\log L/\partial\mu=\frac{1}{\sigma^2}\sum_{\lambda=1}^{N}(x_\lambda-\mu)=0 \qquad \mu=\frac{1}{N}\sum_{\lambda=1}^{N}x_\lambda$$

$$\partial\log L/\partial\sigma=\frac{1}{\sigma^3}\sum_{\lambda=1}^{N}(x_\lambda-\mu)^2-\frac{N}{\sigma}=0 \qquad \sigma^2=\frac{1}{N}\sum_{\lambda=1}^{N}(x_\lambda-\mu)^2$$

以上で解析的に求めることが可能であるが、プログラムではニュートン・ラフソン法を用いて計算を試している。

$$\partial^2\log L/\partial\mu^2=-\frac{N}{\sigma^2}$$

$$\partial^2 \log L/\partial\mu\partial\sigma = -\frac{2}{\sigma^3}\sum_{\lambda=1}^{N}(x_\lambda - \mu)$$

$$\partial^2 \log L/\partial\sigma^2 = -\frac{3}{\sigma^4}\sum_{\lambda=1}^{N}(x_\lambda - \mu)^2 + \frac{N}{\sigma^2}$$

初期値は収束を早めるために $\mu^{(0)} = 0.9\bar{x},\ \sigma^{(0)} = 0.9s$ としている。

回帰分析の推定

ここでは回帰分析のパラメータを最尤法で推定するための方法を具体的に与えておく。しかし、回帰分析は通常、最小 2 乗法を用いて計算を行う。

密度関数： $f(y,x;a,b,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-(y-ax-b)^2/2\sigma^2\right]$

尤度関数： $L = \frac{1}{(2\pi)^{N/2}\sigma^N}\prod_{\lambda=1}^{N}\exp\left[-(y_\lambda - ax_\lambda - b)^2/2\sigma^2\right]$

対数尤度： $\log L = -\frac{1}{2\sigma^2}\sum_{\lambda=1}^{N}(y_\lambda - ax_\lambda - b)^2 - N\log\sigma - \frac{N}{2}\log(2\pi)$

スコアベクトル $\mathbf{U}$ と情報行列 $\mathbf{J}$ （$\mathbf{b} = \boldsymbol{\beta} + \mathbf{J}^{-1}\mathbf{U} \sim N(\boldsymbol{\beta}, \mathbf{J}^{-1})$）

$$\boldsymbol{\beta} = \begin{pmatrix} a \\ b \\ \sigma \end{pmatrix}, \qquad \mathbf{U} = \begin{pmatrix} \partial \log L/\partial a \\ \partial \log L/\partial b \\ \partial \log L/\partial \sigma \end{pmatrix},$$

$$\mathbf{J} = -\begin{pmatrix} \partial^2 \log L/\partial a^2 & \partial^2 \log L/\partial a\partial b & \partial^2 \log L/\partial a\partial \sigma \\ \partial^2 \log L/\partial a\partial b & \partial^2 \log L/\partial b^2 & \partial^2 \log L/\partial b\partial \sigma \\ \partial^2 \log L/\partial a\partial \sigma & \partial^2 \log L/\partial b\partial \sigma & \partial^2 \log L/\partial \sigma^2 \end{pmatrix}$$

この対数尤度関数をパラメータで微分する。

$$\partial \log L/\partial a = \frac{1}{\sigma^2}\sum_{\lambda=1}^{N} x_\lambda(y_\lambda - ax_\lambda - b)$$

$$\partial \log L/\partial b = \frac{1}{\sigma^2}\sum_{\lambda=1}^{N}(y_\lambda - ax_\lambda - b)$$

$$\partial \log L/\partial \sigma = \frac{1}{\sigma^3}\sum_{\lambda=1}^{N}(y_\lambda - ax_\lambda - b)^2 - \frac{N}{\sigma}$$

本文で述べたように、最適解は解析的に求めることが可能であるが、プログラムではニュートン・ラフソン法を用いて計算を試している。

$$\partial^2 \log L/\partial a^2 = -\frac{1}{\sigma^2}\sum_{\lambda=1}^{N} x_\lambda^2$$

$$\partial^2 \log L/\partial a\partial b = -\frac{1}{\sigma^2}\sum_{\lambda=1}^{N} x_\lambda$$

$$\partial^2 \log L/\partial a\partial\sigma = -\frac{2}{\sigma^3}\sum_{\lambda=1}^{N} x_\lambda (y_\lambda - ax_\lambda - b)$$

$$\partial^2 \log L/\partial b^2 = -\frac{N}{\sigma^2}$$

$$\partial^2 \log L/\partial b\partial\sigma = -\frac{2}{\sigma^3}\sum_{\lambda=1}^{N} (y_\lambda - ax_\lambda - b)$$

$$\partial^2 \log L/\partial\sigma^2 = -\frac{3}{\sigma^4}\sum_{\lambda=1}^{N} (y_\lambda - ax_\lambda - b)^2 + \frac{N}{\sigma^2}$$

0/1 ロジスティック回帰分析の推定

ここでは目的変数が 0/1 となるロジスティック回帰分析のパラメータを最尤法で推定するための方法をまとめておく。

確率関数： $p^y(1-p)^{1-y} \quad y=\{0,1\}$ ベルヌーイ分布

尤度関数： $L = \prod_{\lambda=1}^{N} p_\lambda^{y_\lambda}(1-p_\lambda)^{1-y_\lambda}$

対数尤度： $\log L = \sum_{\lambda=1}^{N} y_\lambda \log p_\lambda + \sum_{\lambda=1}^{N}(1-y_\lambda)\log(1-p_\lambda)$

ここで、

$$p_\lambda = \frac{e^{z_\lambda}}{1+e^{z_\lambda}}, \quad z_\lambda = ax_\lambda + b, \quad \text{さらに、} z_\lambda = \log\frac{p_\lambda}{1-p_\lambda}, \quad e^{z_\lambda} = \frac{p_\lambda}{1-p_\lambda}$$

スコアベクトル $\mathbf{U}$ と情報行列 $\mathbf{J}$ （$\mathbf{b} = \boldsymbol{\beta} + \mathbf{J}^{-1}\mathbf{U} \sim N(\boldsymbol{\beta}, \mathbf{J}^{-1})$）

$$\boldsymbol{\beta} = \begin{pmatrix} a \\ b \end{pmatrix}, \qquad \mathbf{U} = \begin{pmatrix} \partial \log L/\partial a \\ \partial \log L/\partial b \end{pmatrix}, \qquad \mathbf{J} = -\begin{pmatrix} \partial^2 \log L/\partial a^2 & \partial^2 \log L/\partial a\partial b \\ \partial^2 \log L/\partial a\partial b & \partial^2 \log L/\partial b^2 \end{pmatrix}$$

最初に以下を求め、

$$\frac{\partial p_\lambda}{\partial a} = \frac{\partial z_\lambda}{\partial a}\frac{\partial p_\lambda}{\partial z_\lambda} = x_\lambda \frac{e^{z_\lambda}(1+e^{z_\lambda}) - e^{z_\lambda}e^{z_\lambda}}{(1+e^{z_\lambda})^2} = \frac{x_\lambda e^{z_\lambda}}{(1+e^{z_\lambda})^2} = x_\lambda p_\lambda (1-p_\lambda)$$

$$\frac{\partial p_\lambda}{\partial b} = \frac{\partial z_\lambda}{\partial b}\frac{\partial p_\lambda}{\partial z_\lambda} = \frac{e^{z_\lambda}(1+e^{z_\lambda}) - e^{z_\lambda}e^{z_\lambda}}{(1+e^{z_\lambda})^2} = \frac{e^{z_\lambda}}{(1+e^{z_\lambda})^2} = p_\lambda (1-p_\lambda)$$

この関係を用いて、対数尤度関数の微分を得る。

$$\frac{\partial \log L}{\partial a} = \sum_{\lambda=1}^{N}\frac{y_\lambda}{p_\lambda}\frac{\partial p_\lambda}{\partial a} - \sum_{\lambda=1}^{N}\frac{1-y_\lambda}{1-p_\lambda}\frac{\partial p_\lambda}{\partial a} = \sum_{\lambda=1}^{N}\frac{y_\lambda - p_\lambda}{p_\lambda(1-p_\lambda)}\frac{\partial p_\lambda}{\partial a} = \sum_{\lambda=1}^{N}(y_\lambda - p_\lambda)x_\lambda,$$

$$\frac{\partial \log L}{\partial b} = \sum_{\lambda=1}^{N}\frac{y_\lambda - p_\lambda}{p_\lambda(1-p_\lambda)}\frac{\partial p_\lambda}{\partial b} = \sum_{\lambda=1}^{N}(y_\lambda - p_\lambda),$$

$$\frac{\partial^2 \log L}{\partial a^2} = -\sum_{\lambda=1}^{N}\frac{\partial p_\lambda}{\partial a}x_\lambda = -\sum_{\lambda=1}^{N}x_\lambda^2 p_\lambda(1-p_\lambda)$$

$$\frac{\partial^2 \log L}{\partial a \partial b} = -\sum_{\lambda=1}^{N} \frac{\partial p_\lambda}{\partial b} x_\lambda = -\sum_{\lambda=1}^{N} x_\lambda p_\lambda (1-p_\lambda)$$
$$\frac{\partial^2 \log L}{\partial b^2} = -\sum_{\lambda=1}^{N} \frac{\partial p_\lambda}{\partial b} = -\sum_{\lambda=1}^{N} p_\lambda (1-p_\lambda)$$

複数の説明変数を持つ一般的なモデルについては、総合マニュアル・多変量解析２の２値ロジスティック回帰分析の章に詳しい。

謝辞

このプログラムは淵上由衣花氏の卒業論文用に作成したものである。通常ではあまり考えないようなものであるが、議論を進めることで最尤法の理解、特にデータ数と対数尤度関数の形状との関係が明白になった。心より感謝します。

参考文献

[1] 福井正康，淵上由衣花，尤度関数の視覚化，日本教育情報学会第 36 回年会論文集，(2020) 332-333，（札幌学院大学，2020/8/22-23）

２１．罹患率の推測

著者が長い間利用させていただいている「医学への統計学 第３版」[1]で、頻度に関する推測の章が設けられ、検定や区間推定が詳しく論じられている。この部分はC.Analysisで不足している分野であるので、この中から特に罹患率に関する分野を抜き出し、保健医療系の学生のために補強することにした。死亡率についても同様であるので、ここでは罹患率を代表として使うことにする。ここでの理論は参考文献[1]をまとめている。

21.1　母罹患率に関する推測

罹患率を I 、ある集団を一定期間追跡して、対象の個体ごとの観測期間の和（人年など）を T 、一定観測期間の中で罹患した人数を r とする。罹患率があまり高くないとすると、罹患人数の分布は以下のポアソン分布に従う。

$$\Pr(r \mid \lambda) = \frac{\lambda^r}{r!} e^{-\lambda} \qquad \text{ここに、} \lambda = IT$$

ポアソン分布の平均と分散は以下で与えられる。

$$E[r] = \lambda = IT\ ,\quad Var[r] = \lambda = IT$$

これより罹患率の点推定値を次のように定義すると、

$$\hat{I} = r/T$$

標準誤差は以下となる。

$$Var[\hat{I}] = Var[r]/T^2 = r/T^2$$

母罹患率の検定

$H_0 : I = I_0,\ H_1 : I \neq I_0$　として、検定の両側確率 p は以下のように与えられる。

$$p = 2\left[1 - \frac{1}{2}\Pr(r \mid I_0T) - \sum_{j=0}^{r-1} \frac{(I_0T)^j}{j!} e^{-I_0T}\right] = 2\left[1 - \frac{(I_0T)^r}{2r!} e^{-I_0T} - \sum_{j=0}^{r-1} \frac{(I_0T)^j}{j!} e^{-I_0T}\right]$$

母罹患率の区間推定

母罹患率の $100(1-\alpha)\%$ 信頼区間は以下のように与えられる。

$$\frac{\chi^2_{2r}(1-\alpha/2)}{2T} \le I \le \frac{\chi^2_{2r+2}(\alpha/2)}{2T}$$

これらの他に正規分布に基づく近似的な方法もあるが、プログラムでは上の方法を用いる。

21.2　罹患率の比較に関する推測

２つの群の罹患率をそれぞれ I_1, I_2 、２つ群を一定期間追跡して、対象の個体ごとの観測期間の和（人年など）を T_1, T_2 、一定観測期間の中で罹患した人数を r_1, r_2 とする。２つの群の罹患率の推定値は以下で与えられる。

$$\hat{I}_1 = r_1/T_1\ ,\quad \hat{I}_2 = r_2/T_2$$

罹患率の比較の検定

$H_0 : I_1 = I_2 = I\ ,\quad H_1 : I_1 \neq I_2$ として、検定には帰無仮説を利用した以下の関係が使われる。

$$Z = \frac{\hat{I}_1 - \hat{I}_2}{\sqrt{(r_1 + r_2)/(T_1 T_2)}} \sim N(0,1)$$

罹患率の差の区間推定

罹患率の差の$100(1-\alpha)\%$信頼区間は、

$$SE \equiv SE(I_1 - I_2) = \sqrt{\hat{I}_1/T_1 + \hat{I}_2/T_2}$$

の関係を用いて、

$$\hat{I}_1 - \hat{I}_2 - Z(\alpha/2)SE \le I_1 - I_2 \le \hat{I}_1 - \hat{I}_2 + Z(\alpha/2)SE$$

罹患率の比の区間推定

罹患率の比の$100(1-\alpha)\%$信頼区間は、

$$SE \equiv SE\left[\log\left(\hat{I}_1/\hat{I}_2\right)\right] = \sqrt{1/r_1 + 1/r_2}$$

の関係を用いて、

$$\exp\left[\log\left(\hat{I}_1/\hat{I}_2\right) - Z(\alpha/2)SE\right] \le I_2/I_1 \le \exp\left[\log\left(\hat{I}_1/\hat{I}_2\right) + Z(\alpha/2)SE\right]$$

21.3　罹患割合に関する推測

2 つの群の罹患割合をそれぞれp_1, p_2、２つ群の観測対象者数をn_1, n_2、一定観測期間の中で罹患した人数をr_1, r_2とする。2 つの群の罹患割合の推定値は以下で与えられる。

$$\hat{p}_1 = r_1/n_1,\ \hat{p}_2 = r_2/n_2$$

罹患割合の検定

罹患割合の違いの検定には以下の関係を利用する。

$$Z = \frac{|\hat{p}_1 - \hat{p}_2| - (1/n_1 + 1/n_2)/2}{\sqrt{\bar{p}(1-\bar{p})(1/n_1 + 1/n_2)}} \sim N(0,1)$$

ここに、$\bar{p} = (r_1 + r_2)/(n_1 + n_2)$である。

罹患割合の差の信頼区間

罹患割合の差の$100(1-\alpha)\%$信頼区間は、

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

の関係を用いて以下となる。

$$\hat{p}_1 - \hat{p}_2 - Z(\alpha/2)SE \le p_1 - p_2 \le \hat{p}_1 - \hat{p}_2 + Z(\alpha/2)SE$$

罹患割合の比の信頼区間

罹患割合の比の$100(1-\alpha)\%$信頼区間は、

$$SE\left[\log(\hat{p}_1/\hat{p}_2)\right] = \sqrt{\frac{1-\hat{p}_1}{n_1\hat{p}_1} + \frac{1-\hat{p}_2}{n_2\hat{p}_2}} = SE$$

の関係を用いて以下となる。

$$\exp\left[\log\left(\hat{p}_1/\hat{p}_2\right) - Z(\alpha/2)SE\right] \le p_2/p_1 \le \exp\left[\log\left(\hat{p}_1/\hat{p}_2\right) + Z(\alpha/2)SE\right]$$

21.4　プログラムの利用法

メニュー［基本統計－医療関連手法－罹患率の推測］を選択すると図 1 のような実行画面が表示される。

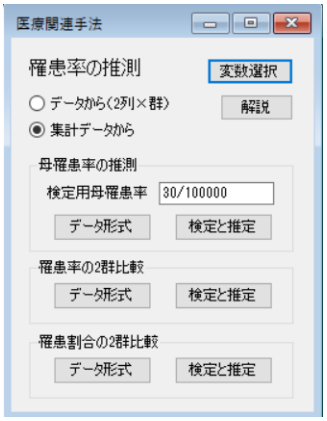


図 1　罹患率の推測実行画面

このプログラムは、罹患率と罹患割合について推測を行うものである。罹患率は発生数を対象者の観測時間合計（人年など）で割ったものであり、罹患割合は発生数を対象者数で割ったものである。ここでは母罹患率に関する検定と推定、罹患率の 2 群比較に関する検定と差と比の推定、罹患率ではないが、罹患割合の 2 群比較に関する検定と差と比の推定が求められる。特に罹患割合の比較は、χ^2 検定としてすでにプログラムに組み込まれているが、ここではそこで求めていない差や比の信頼区間も表示される。

母罹患率の推測のデータには図 2 のような 2 つの形式がある。

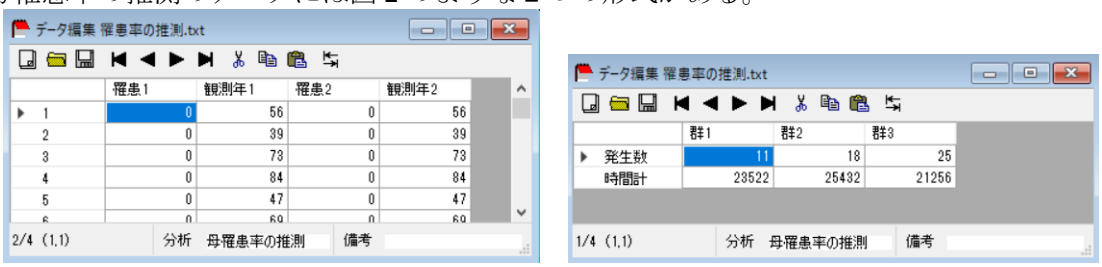


図 2　母罹患率の推測のデータ形式（罹患率の推測.txt）

左は「データから」の形式で、変数選択では 2 列ごとに読み込むため、全部選択すると 2 種類の罹患率の推測を行うことになる。右は「集計データから」の形式で、1 列ごとに読み込むため、全部選択すると 3 種類の推測を行うことになる。母罹患率を指定して行う比較検定では、「検定用母罹患率」テキストボックスの中に比較する母罹患率を入力しておく。入力の形式は式で入力することもできる。何も入力されていない場合は、0 と解釈される。

図 2 の右側のデータを用いて「検定と推定」ボタンをクリックした結果を図 3 に示す。

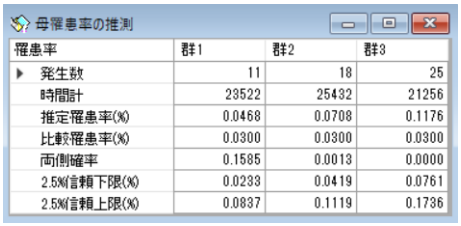


図 3　母罹患率の推測の実行結果

ここで、罹患率は小さな値になることがあるので%表示にしている。

次に罹患率の比較について説明する。図 4 にデータ形式を示す。但し、「データから」の形式は図 2 と同じである。

図 4　罹患率の比較のデータ形式

右のデータを元に「検定と推定」を実行した結果を図 5 に示す。

図 5　罹患率の比較の実行結果

最後に罹患割合について説明する。図 6 にデータ形式を示すが、左の図の 2 つの群で一般にデータ数は異なる。

図 6　罹患割合の比較のデータ形式

右のデータを元に「検定と推定」を実行した結果を図 7 に示す。

図 7　罹患割合の比較の実行結果

参考文献

[1] 丹後俊郎，古川俊之監修，医学への統計学【第 3 版】，朝倉書店, 2013.

２２．ROC 曲線

ROC（Receiver Operatorating Characteristic）曲線は 2 値の結果（例えば陽性と陰性）に対するある説明変数による診断の有効性を検討する手段である。説明変数にある閾値（カットオフポイント）を考え、その値より上が陽性、下が陰性と診断する（逆も考えられる）。しかしこの診断には間違いが生じる。例えば、陽性と診断して実は陰性、陰性と診断して実は陽性となることである。特に陰性に対して陽性と診断する割合を擬陽性率と呼ぶ。また、陽性に対して正しく陽性と診断する割合を（真）陽性率と呼ぶ。カットオフポイントとして説明変数のデータ値をとり、その値を変えることで、この陽性率と擬陽性率の変化を表示するグラフを ROC 曲線という。例えばあるカットオフポイントで表 1 の分割表の結果を得たとする。

表 1　診断結果と実測結果

	実測陽性	実測陰性	合計
予測陽性	a	b	a+b
予測陰性	c	d	c+d
合計	a+c	b+d	N (=a+b+c+d)

ここで、陽性率は a/(a+c)、擬陽性率は b/(b+d) である。陽性率は 1 に近い方が良く、擬陽性率は 0 に近い方が良い。

22.1　プログラムの利用法

ここでは実際の画面を見ながら ROC 曲線について説明する。メニュー［分析－基本統計－医療関連手法－ROC 曲線］を選択すると以下の分析実行画面が表示される。

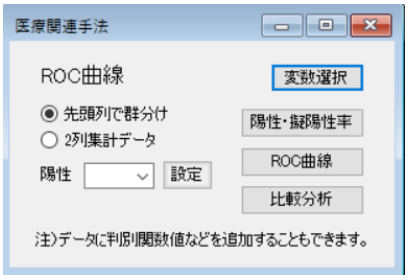


図 1　ROC 曲線実行画面

分析用のデータは、図 2a か図 2b のような形式である。

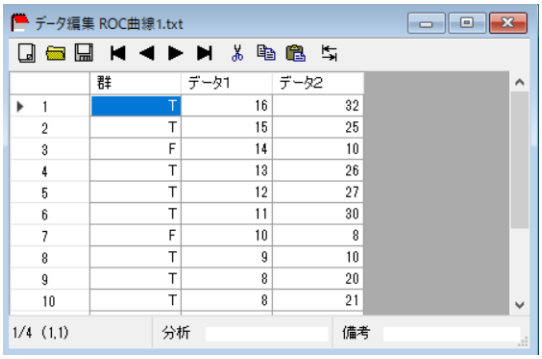


図 2a　先頭列で群分け

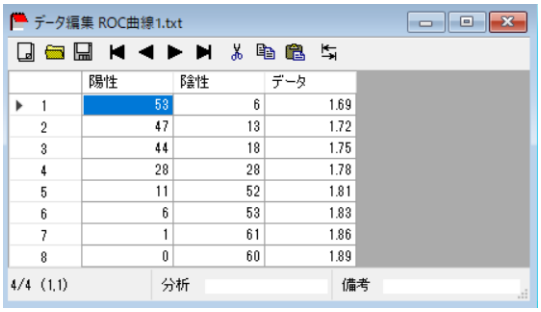


図 2b　2 列集計形式

図 2a のデータは、2 群間の差の検定、判別分析、数量化Ⅱ類、2 値ロジスティック回帰分析などで利用される基本的なデータ形式である。図 2b は 2 値ロジスティック回帰分析にも使われている形式で、最初の 2 列に 2 つの群のデータ数が与えられた集計データである。この形式の集計データは他の分析にも組み込んで行く必要がある。

「先頭列で群分け」のデータの変数をすべて選択して、「設定」ボタンで陽性の設定を「T」とし、「陽性・擬陽性率」ボタンをクリックすると図 3 に示す結果が表示される。

陽性率と擬陽性率

	データ1	陽性率(+)	擬陽性率(+)	データ2	陽性率(+)	擬陽性率(+)
1	16.000	0.100	0.000	32.000	0.100	0.000
2	15.000	0.200	0.000	25.000	0.600	0.000
3	14.000	0.200	0.200	10.000	1.000	0.600
4	13.000	0.300	0.200	26.000	0.500	0.000
5	12.000	0.400	0.200	27.000	0.300	0.000
6	11.000	0.500	0.200	30.000	0.200	0.000
7	10.000	0.500	0.400	8.000	1.000	1.000
8	9.000	0.600	0.400	10.000	1.000	0.600
9	8.000	0.900	0.600	20.000	0.900	0.200
10	8.000	0.900	0.600	21.000	0.800	0.000

図 3　陽性・擬陽性率出力結果

これは入力データ順に、そのデータの値をカットオフ値にした場合の陽性率と擬陽性率を与えたものである。ここで変数名「陽性率(+)」と「擬陽性率(+)」の後ろに「(+)」が付いているが、これはデータの大きい方を陽性にしていることを示している。データの小さい方を陽性にする場合は後ろに「(-)」が付く。陽性率も擬陽性率も陽性の予測は指定データ以上（小さい方は以下）として計算している。

ROC 曲線はこのデータを変数ごとに並び替えて求める。そのままの設定で「ROC 曲線」ボタンをクリックすると図 4 に示すグラフが表示される。横軸が擬陽性率、縦軸が陽性率である。

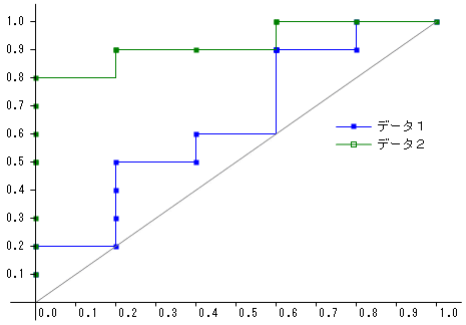


図 4　ROC 曲線描画結果

変数として 2 つを選択しているので、2 本の ROC 曲線が表示されている。ROC 曲線は理想的な判別の点 (0,1) にどれだけ近いか、また (0,0) と (1,1) を結ぶ判別不能の線からどれだけ離れているかなどによって評価される。

この ROC 曲線に対する評価は、「比較分析」ボタンをクリックすることで、図 5 のように示される。


比較分析

	AUC	最小距離	カットオフ値	Youden	カットオフ値
▶ データ1	0.640	0.539	11	0.300	8
データ2	0.920	0.200	21	0.800	21
AUC比較検定	z値	1.511	両側確率	0.131	

図 5　比較分析結果

ここでは「AUC」（Area Under the Curve）、(0,1)からの「最小距離」、「Youden Index」などが示されている。AUC は折れ線グラフと x 軸の間の面積を与えており、0.5≦AUC≦1.0 の範囲をとる。AUC は大きな値ほど有効性が高いと評価される。(0,1)からの最小距離は理想的な判別の点からの距離で、小さな値ほど有効性が高い。また、Youden Index は図に示した斜めの線からの y 軸方向の距離を表す。これは大きな値ほど有効性が高い。最小距離と Youden Index にはそれぞれデータを分ける値であるカットオフ値が付いている。

2 つのデータを並べて比較するとき、同じ対象である場合は AUC についての比較検定が可能である。3 行目にその結果が与えられている。この理論は 2 節で説明されているが、同じ対象（同じ数で同じ陽性率）以外ではこれは表示されない。

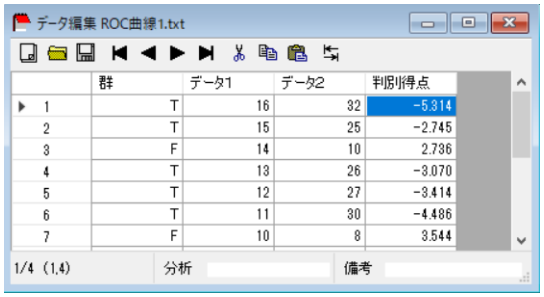
ROC 曲線を用いた分析は 1 つの変数が対象であるが、判別分析による判別関数やロジスティック回帰分析の確率など､スコアを求める分析のデータを用いればそれらの比較が可能である。ただ、これらは元々多変量による直接分類が目的で作られた手法であるので、ROC 曲線を用いる必要があるのかは疑問であるが、視覚的に比較するのは興味深い。例として判別分析で求めた判別関数値を加えて分析を行ってみる。

図 2a のデータによる判別関数値は、メニュー［分析－多変量解析他－判別手法－判別分析］による 2 群の「判別得点」で簡単に求められる。それを図 6a に示す。このデータを例えば［編集－エディタ指定列追加］などでグリッドエディタに追加すると図 6b のようなデータとなる。


判別得点

	所属群	判別得点	判別群
▶ 1	T	-5.314	T
2	T	-2.745	T
3	F	2.736	F
4	T	-3.070	T
5	T	-3.414	T
6	T	-4.486	T
7	F	3.544	F
8	T	2.836	F
9	T	-0.785	T
10	T	-1.149	T

図 6a　判別得点結果


データ編集 ROC曲線1.txt

	群	データ1	データ2	判別得点
▶ 1	T	16	32	-5.314
2	T	15	25	-2.745
3	F	14	10	2.736
4	T	13	26	-3.070
5	T	12	27	-3.414
6	T	11	30	-4.486
7	F	10	8	3.544

1/4 (1,4)　分析　備考

図 6b　追加されたデータ

これを用いて ROC 曲線を描いたグラフが図 7 で、比較分析を行った結果が図 8 である。少しであるが、結果の向上が視覚的によく分かる。

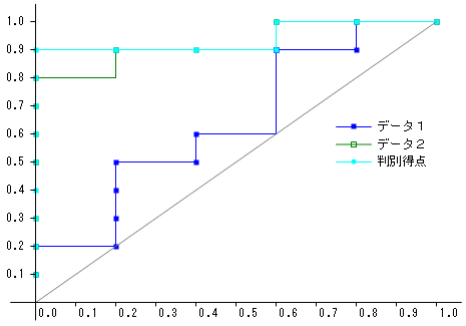


図 7　ROC 曲線結果

比較分析

	AUC	最小距離	カットオフ値	Youden	カットオフ値
データ1	0.640	0.539	11	0.300	8
データ2	0.920	0.200	21	0.800	21
判別得点	0.940	0.100	−0.785	0.900	−0.785

図 8　比較分析結果

22.2　ＲＯＣ曲線の検定

2 つの指標 A, B を使った ROC 曲線で、指標の有効性の基準の 1 つである AUC について、プログラムで使った比較検定の理論を与えておく[2]。但し、ここでは陽性群と陰性群の数は指標 A、指標 B 共にそれぞれ、m, n とする。

指標 A のデータのうち、陽性群のデータを x_i^a （$i=1,\cdots,m$）、陰性群のデータを y_j^a（$j=1,\cdots,n$）、指標 B のデータのうち、陽性群のデータを x_i^b （$i=1,\cdots,m$）、陰性群のデータを y_j^b （$j=1,\cdots,n$）とする。ここでは指標 A, B ともに、陽性群は陰性群より高い値をとるものと仮定している。プログラムで指標 A についての AUC は ROC 曲線の下側の面積を直接求めているが、以下のような方法でも求めることができる。

$$AUC(A)=\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(x_i^a,y_j^a) \qquad \text{ここに、} \psi(x,y)=\begin{cases}1 & x>y\\ 0 & x\le y\end{cases}$$

この ψ 関数の定義は参考文献[2]とは少し異なることを注意しておく（検討の余地あり）。

AUC の分散 V_A は以下のように与えられる。

$$V_A=\frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}\psi(x_i^a,y_j^a)-AIC(A)\right)^2$$
$$+\frac{1}{n(n-1)}\sum_{j=1}^{n}\left(\frac{1}{m}\sum_{i=1}^{m}\psi(x_i^a,y_j^a)-AIC(A)\right)^2$$

また、指標 A と指標 B の AUC の共分散は以下の通りである。

$$S_{AB}=\frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}\psi(x_i^a,y_j^a)-AIC(A)\right)\left(\frac{1}{n}\sum_{j=1}^{n}\psi(x_i^b,y_j^b)-AIC(B)\right)$$
$$+\frac{1}{n(n-1)}\sum_{j=1}^{n}\left(\frac{1}{m}\sum_{i=1}^{m}\psi(x_i^a,y_j^a)-AIC(A)\right)\left(\frac{1}{m}\sum_{i=1}^{m}\psi(x_i^b,y_j^b)-AIC(B)\right)$$

これらの関係を利用すると、統計量 z は標準正規分布に従う。

$$z=\frac{AUC(A)-AUC(B)}{\sqrt{V_A+V_B-2S_{AB}}}\sim N(0,1)$$

この関係を用いて、2 つの指標 A, B について、AUC の比較検定が可能となる。

参考文献

[1] 山田 実、浅井 剛、土井剛彦、「メディカルスタッフのためのひと目で選ぶ統計手法」、羊土社（2018）

[2] 2 つの予測モデルどっちが良いの?（AUC の差の検定）, https://qiita.com/sz_dr/items/96e9306979cb1832d120 （2021/2/28 取得）

２３．傾向スコアマッチング

ある対象データについて、2 群間に差があるかどうかの判定は、2 群間の差の検定を用いて調べることができるが、その分類と対象データの両方に影響を与える交絡因子（バイアス要因）がある場合、検定結果が真に 2 群の分類による差かどうか明らかではない。その場合には、交絡因子の影響を除去するために、2 群の間で交絡因子の影響をそろえる操作が行われる。その 1 つの方法が傾向スコアによるデータのマッチングである。

まず 2 群の分類を、交絡因子を使って説明する分析を考える。これには判別分析やロジスティック回帰分析などが考えられる。そのときの判別得点や判別確率を傾向スコアと呼ぶ。２つの群のデータで、この傾向スコアが近いものは同じような交絡因子を持つと考え、2 つの群から最も近いものを 1 つずつ取り出してペアを作り、データから取り除く。残りのデータを使ってこの操作を繰り返し、すべての傾向スコアの差が指定された値より大きくなったら取り出しを終わる。この操作で取り出された 2 組のデータは交絡因子が平均的に等しいと考えられるので、このデータだけを使って差の検定を実施する。これにより交絡因子の影響は抑えられるものと考える。以後実際のプログラムを使って検討してみよう。

今回作成したプログラムは、傾向スコアを利用してデータのマッチングを行うプログラムである。傾向スコアを作成するのは判別分析や 2 値ロジスティック回帰分析など、他の分析に譲る。

現在のある動作の可否で、1 年後にあるデータの得点（検査の値など）に差が出るかどうか検定する問題を考える。ただこれらに影響を与える交絡因子があるとして、性別、年齢、ある要因の有無を考えておく。これらのデータを図 1 に与える。

	動作	得点	性別	年齢	要因
1	1	2.4	1	71	0
2	0	2.4	1	71	0
3	1	1.8	0	66	0
4	1	4.5	0	75	1
5	1	2.3	1	67	0
6	1	4.8	0	64	1
7	1	5.4	1	70	1
8	1	2.1	1	79	0
9	1	5.4	1	70	1
10	0	4.5	0	74	1
11	0	4.8	0	67	1

図 1　元データ

図 1 の動作の可否と得点を用いて 2 群の検定を実行すると、図 2 のような結果を得る。

動作	正規性	等分散性	検定手法	両側確率
得点	なし		順位和検定	0.0117

図 2　2 群の差の検定結果

この結果は、動作の可否でみた真の得点の差とはいえない。そこには交絡因子の違いが含まれている。動作による真の差を見るためにはこの違いを平滑化しなければならない。

そこで、動作の可否を性別、年齢、要因で説明するロジスティック回帰分析を実行し、結果の予測確率を図 1 のデータに加えてみる。この予測確率が傾向スコアである。結果を図 3 に示す。

	動作	得点	性別	年齢	要因	予測確率
1	1	2.4	1	71	0	0.452
2	0	2.4	1	71	0	0.452
3	1	1.8	0	66	0	0.681
4	1	4.5	0	75	1	0.369
5	1	2.3	1	67	0	0.660
6	1	4.8	0	64	1	0.860
7	1	5.4	1	70	1	0.658
8	1	2.1	1	79	0	0.130
9	1	5.4	1	70	1	0.658
10	0	4.5	0	74	1	0.420
11	0	4.8	0	67	1	0.764

図 3　傾向スコアを加えたデータ

この傾向スコアが、３つの交絡因子を代表する変数である。

メニュー［分析－基本統計－ユーティリティー傾向スコアマッチング］を選択すると、分析実行画面が図 4 のように表示される。

図 4　傾向スコアマッチング実行画面

変数選択として、動作（分類）、予測確率（傾向スコア）、得点（比較変数）の順に選択し（得点の部分は選択しなくてもよいし、多く選択してもよい）、「マッチング」ボタンをクリックすると図 5 のような結果が得られる。

	レコード0	レコード1	スコア0	スコア1	スコア差	0>得点	1>得点
9	43	13	0.117	0.117	0.000	1.6	1.6
10	45	68	0.874	0.874	0.000	5.3	5.3
11	53	31	0.787	0.787	0.000	2.3	2.3
12	54	50	0.529	0.529	0.000	1.9	1.9
13	58	33	0.397	0.397	0.000	5.0	5.0
14	64	37	0.237	0.237	0.000	1.5	1.5
15	66	4	0.369	0.369	0.000	4.5	4.5
16	73	51	0.169	0.169	0.000	1.5	1.5
17	75	41	0.258	0.258	0.000	5.1	5.1
18	70	69	0.473	0.475	-0.002	4.5	1.9
19	20	93	0.582	0.579	0.003	1.9	4.9
20	87	26	0.503	0.506	-0.003	5.0	2.4
21	88	8	0.141	0.130	0.011	1.6	2.1

図 5　傾向スコアマッチング結果

これは、動作の可否の 2 つの群で傾向スコアの類似したデータのマッチング結果である。マッチングは、2 群の傾向スコアを合わせた標準偏差の「0.2」倍までを取っている。これはメニューの中で指定する。

この結果を確認したのち、「利用データ」ボタンをクリックすると、図 6 のような結果が表示される。

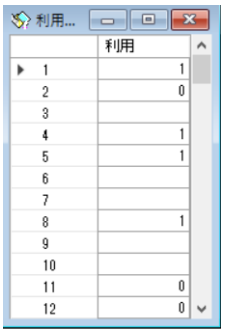


図 6　利用データ

このデータを元のグリッドエディタに貼り付けた結果が図 7 である。

図 7　利用データ結果を貼り付けたデータ

「利用」の分類で得点を分けて 2 群の差の検定を行った結果が図 8 である。

図 8　傾向スコアでマッチングした 2 群の差の検定結果

「利用」データを用いて交絡因子の平均を比較してみる（量的データの集計の簡易統計量）。図 9 に結果を示す。

図 9　傾向スコアでマッチングした交絡因子の基本統計量

これを見るとマッチングにより、うまく平滑化が行われている様子が分かる。

２４．中心極限定理

24.1　実験・中心極限定理

メニュー［分析－基本統計－ユーティリティー－データ生成］または、グリッドエディタのメニュー［ツール－データ生成］を選択すると図 0 のような分析実行画面が表示される。

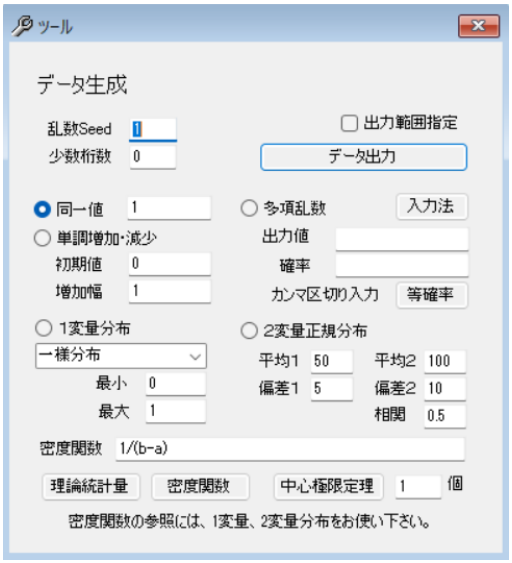


図 0　分析実行画面

この中で、「1 変数分布」の中から分布を選択し、平均をとる個数を入力して「中心極限定理」ボタンをクリックすると、中心極限定理確認用のデータがグリッドエディタに出力される。ここで、「出力列」等は通常のデータ生成に準じる（ツールの巻参照）。

中心極限定理は、以下の形で与えられる。

独立な確率変数、$X_i\ (i=1,2,\cdots,n)$ が、平均 μ_i、分散 σ_i^2 のある条件に従う一般的な確率分布に従うとき、以下となる。

$$\lim_{n\to\infty}\sum_{i=1}^{n}(X_i-\mu_i)\Big/\sqrt{\sum_{i=1}^{n}\sigma_i^2}\sim N(0,1)$$

最も有名な形式は、各確率変数の平均と分散が等しく値を持つときに、n 個の標本データの平均を求める場合である。

独立な確率変数 $X_1, X_2, \cdots, X_n$ が、平均 μ、分散 σ^2 の同一の確率分布に従うとき、

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i=\frac{1}{n}(X_1+X_2+\cdots+X_n)\underset{n\to\infty}{\sim}N(\mu,\sigma^2/n)$$

これは一口に言えば､平均と分散が値を持つどんな分布でもデータを多く集めて平均を取れば､それが性質のよく知られた正規分布になり､しかも分散も小さくできるということである。分散が小さくなるということは、平均を推定するのに精度が上がるということになる。

この章では後者の形の中心極限定理の成立を何種類かの分布で確かめてみることにする。使う分析は［分析－基本統計－ユーティリティ］の「データ生成」と「MCMC 乱数生成」で与える。

まず「データ生成」で、標準正規分布の乱数を 1000 個発生させ、それを図 1 のようにヒストグラムに描いてみよう。この標準正規分布の乱数はどのようにして作られているのだろうか。コンピュータには正規乱数を発生させる機能はない。基本的に 0 から 1 の一様分布が出力される。この一様分布は平均が 1/2、分散が 1/12 の図 2 のような分布である。それを連続して 12 個集めて 6 を引いたのが図 1 である。

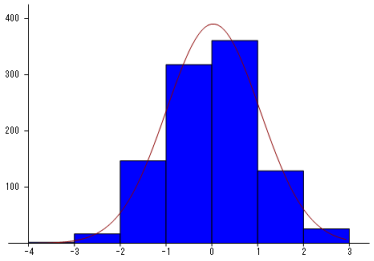


図 1　標準正規乱数

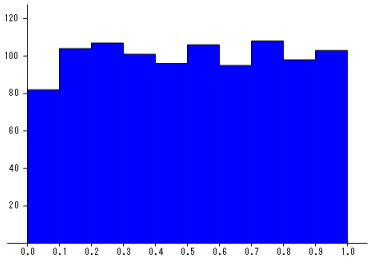


図 2　一様分布

ここではこの一様分布を用いてデータ数 5 個と 10 個で中心極限定理を試してみよう。

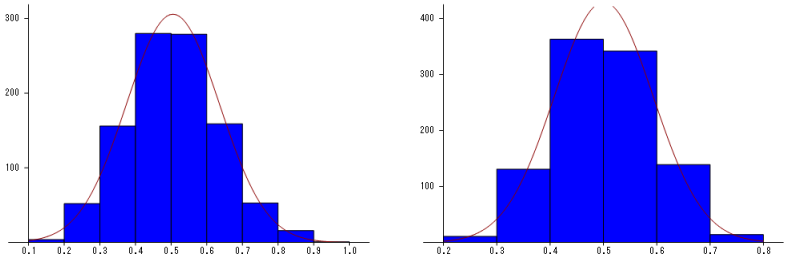
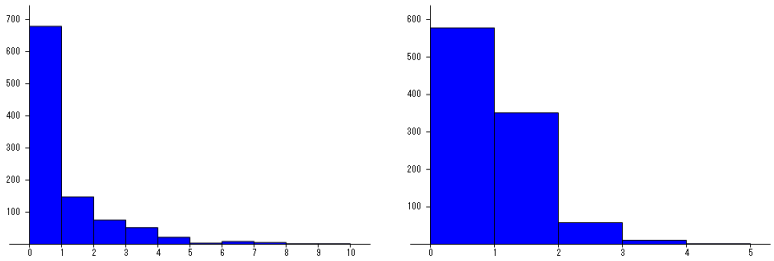


図 2　5 個と 10 個のデータの平均

5 個からすでに正規分布に近づき始めている。実際 12 個の平均を取ると図 1 の形になることが示せる。ここでは正規性の確認のために、正規分布の密度関数を重ねているが、数値による確認も可能である。試してみてもらいたい。

次に少し極端な分布の中心極限定理を、元の分布、5 個、10 個、100 個の平均で比較してみる。

自由度 1 の χ^2 分布（平均 1, 分散 2）



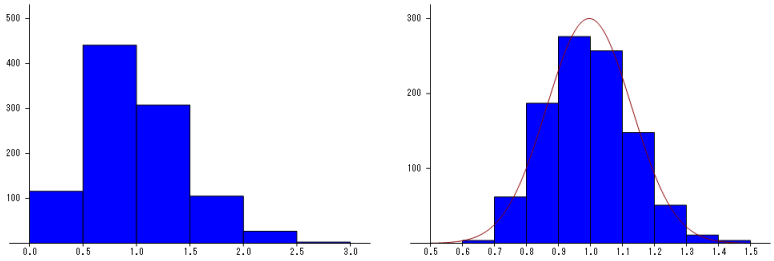


図 3　元の分布、5 個、10 個、100 個の平均

n=1, p=0.7 の 2 項分布（平均 0.7, 分散 0.21）

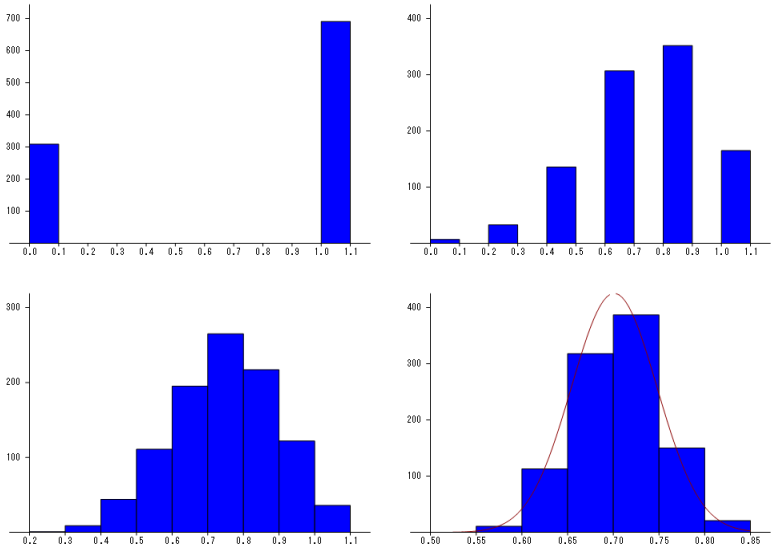


図 4　元の分布、5 個、10 個、100 個の平均

うまく正規分布に近づいていることが分かる。

さて、中心極限定理は条件にあるように、平均と分散が与えられている分布が対象である。しかし、分布の中には裾野が拡がって分散が無限大になるものもある。そのときどんな結果になるのだろうか。例として自由度 1 の t 分布を用いてみる。 t 分布は自由度 3 以上のときに分散が確定するような分布であるが、図 5 に見るように、自由度 1 の場合も見た目は正規分布に似てる（但し、0 から離れたところはカットしてある）。

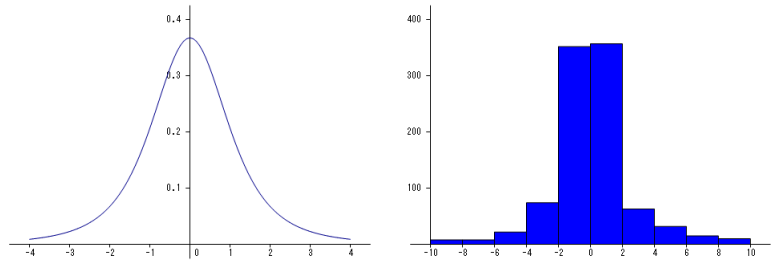


図 5　自由度 1 の t 分布

自由度 1 の t 分布について、5 個、10 個、100 個の平均を比較した結果を図 6 に示す。

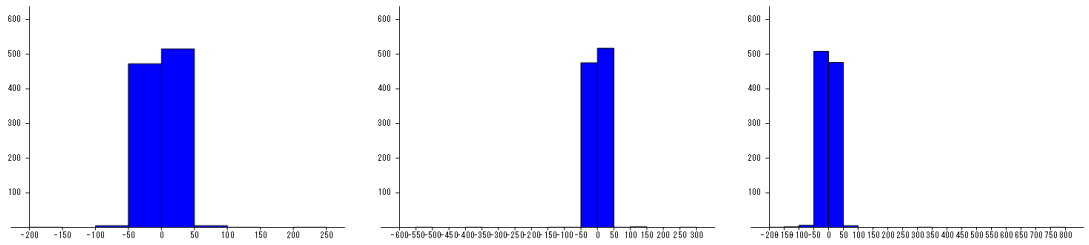


図 6　5 個、10 個、100 個の平均

これを見ると大きな値や小さな値が見られ、正規分布とは言えない。このデータに正規確率紙の方法（q-q プロット）を使った結果が図 7 である。

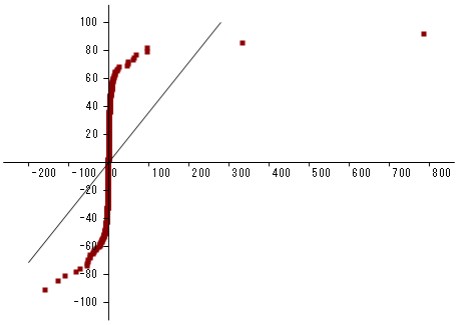


図 7　正規確率紙の方法（q-q プロット）

直線からかなりずれていることが分かる。

これまでの結果をみると、分散が求められない例外的な場合を除いて、中心極限定理が成り立っていることが分かる。ここまで、乱数発生法は数式を使った昔ながらの擬似乱数発生法を使ってきたが、ここでは、乱数発生のもう一つの方法である MCMC を使ってみよう。これは密度関数が任意の形をしたものに対応させる乱数発生法である。我々は正規分布をずらして 2 つ重ねた二峰分布について調べてみる。

二峰分布

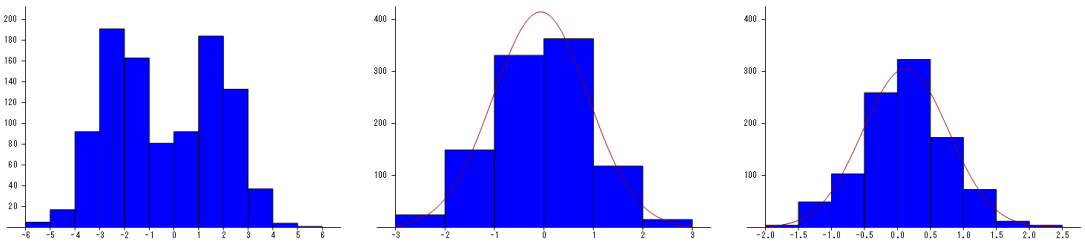


図 8　元の分布、5 個、10 個の平均

中心極限定理はうまく成り立っています。ところが乱数の平均を取る順番を入れ替えると図 9 のように収束が非常に悪くなります。

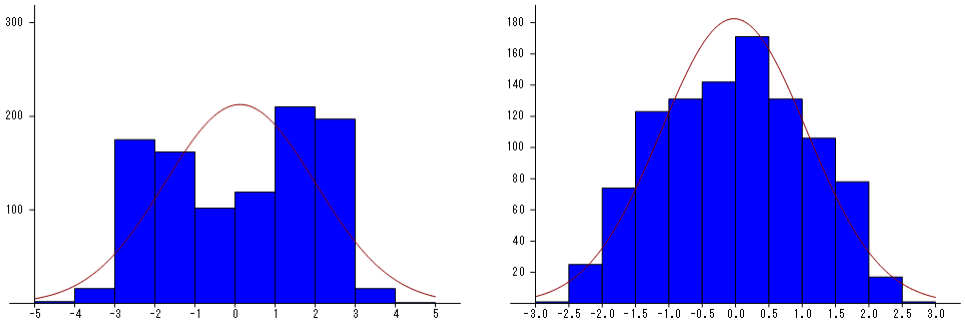


図 9　連続したデータの 10 個、100 個の平均

これはどういうことかというと、MCMC で作った乱数は、全体としては正しい乱数だが、連続する近いデータの間には相関がある。この結果はその影響が示された例である。利用に際して十分注意が必要であろう。

24.2　中心極限定理の理論

ここでは参考文献［1］に従って、中心極限定理の理論について見てみよう。

平均 μ、分散 σ^2 の任意の母集団から、大きさ n の標本を任意抽出した。その標本を $x_i(i=1,\cdots,n)$、n 個の標本平均を $\overline{x}_n=(x_1+x_2+\cdots+x_n)/n$ としたとき、以下を証明する。

$$y_n=(\overline{x}_n-\mu)\Big/\left(\sigma/\sqrt{n}\right)=\sum_{i=1}^{n}(x_i-\mu)\Big/\left(\sqrt{n}\sigma\right)\underset{n\to\infty}{\sim}N(0,1)$$

今、$u_i=(x_i-\mu)\Big/\left(\sqrt{n}\sigma\right)$ として、$E[(x_i-\mu)^n/\sigma^n]$ が有限と仮定すると、u_i の積率母関数 $\varphi_i(t)$ は、$E[u_i]=0$、$E[u_i^2]=1/n$、$E[u_i^3]\sim O(n^{-3/2})$、…、より、

$$\begin{aligned}\varphi_i(t)&\equiv E[e^{tu_i}]=1+E[u_i]t+E[u_i^2]t^2/2!+E[u_i^3]t^3/3!+\cdots\\&=1+t^2/(2n)+O(n^{-3/2})\\&=1+[t^2+\varepsilon(n)]/(2n)\qquad\lim_{n\to\infty}\varepsilon(n)=0\end{aligned}$$

$u_i(i=1,\cdots n)$ はそれぞれ独立であることから、y_n の積率母関数 $\varphi_n(t)$ は、

$$\varphi_n(t)=E\left[\exp\left(t\sum_{i=1}^{n}u_i\right)\right]=\prod_{i=1}^{n}E\left[\exp\left(tu_i\right)\right]=\prod_{i=1}^{n}\varphi_i(t)$$

$$=\prod_{i=1}^{n}\left[1+\frac{(t^2+\varepsilon(n))/2}{n}\right]=\left[1+\frac{(t^2+\varepsilon(n))/2}{n}\right]^n$$

ある N 以上では、$-\Delta\le\varepsilon(N)\le\Delta$ であることから、$n>N$ に対して、

$$\left[1+\frac{(t^2-\Delta)/2}{n}\right]^n<\varphi_n(t)<\left[1+\frac{(t^2+\Delta)/2}{n}\right]^n$$

よって、

$$e^{(t^2-\Delta)/2}<\lim_{n\to\infty}\varphi_n(t)<e^{(t^2+\Delta)/2}$$

より、

$$\lim_{n\to\infty}\varphi_n(t)=e^{t^2/2}$$

ここで、$e^{t^2/2}$ が標準正規分布の積率母関数であることから、

$$y_n \underset{n\to\infty}{\sim} N(0,1)$$

注）標準正規分布の積率母関数

$$\varphi(t)=E[e^{tx}]=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{tx}e^{-x^2/2}dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-(x-t)^2/2}e^{t^2/2}dx$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-(x-t)^2/2}dx\times e^{t^2/2}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-y^2/2}dy\times e^{t^2/2}=e^{t^2/2}$$

参考文献

[1] 安倍 齊, 応用数理統計学入門, 培風館, 1985