# 研修資料 160323

福山平成大学　福井正康

## １．データ発生サブメニュー

メニュー「ツールーデータ生成」をクリックすると図１のような画面が表示される。

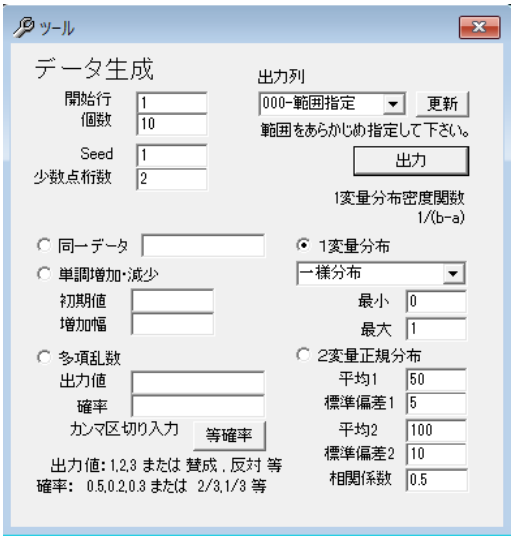



図１　データ発生画面

データ生成は、出力範囲または出力列を指定する。グリッド中の列を指定することもできるし、新たに列を追加して出力することもできる。列の指定はグリッド中を選択して「範囲指定」としてもよいし、列指定のコンボボックスで選択指定してもよい。出力データの個数はデフォルトで 10000 個であるが、行数が少ない場合は行数が限度である。乱数発生に再現性を与えるため、乱数の「Seed」を与えて発生させる。乱数の小数点以下桁数も指定する。

発生乱数の種類としては、図１にあるように、同一データ、単調増加・減少、多項分布、１変量分布、２変量正規分布があるが、それらのパラメータの値を指定して選択する。
１変量分布は分布の数が多いため、コンボボックスから分布の名前を選択して実行する。
２変量正規分布は、散布図を描くときに役に立つが、新規追加の場合は 2 列追加される。
講義資料などの作成には便利なメニューである。

[1] 四辻哲章, 計算機シミュレーションのための確率分布乱数生成法, プレアデス出版, 2010.

## ２．ＭＣＭＣ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理論について述べ、次にプログラムの利用法について説明する。

### 2.1　マルコフ連鎖モンテカルロ法による乱数発生

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。

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

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。

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

さて、ここで、推移核 $p(x|x')$ をこの提案分布確率密度 $q(x|x')$ と、ある確率 $\alpha(x|x')$ 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 $\alpha(x|x')$ 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。

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

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。

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

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

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

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

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

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

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

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

即ち、乱数を提案分布により発生させ、確率 $\alpha\left(x^{(t+1)}|x^{(t)}\right)$ 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 $\pi\left(x\right)$ 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。

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

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、 $q\left(x\,|\,x'\right)=q\left(x-x'\right)$ となり、条件付き確率は具体的に以下となる。

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

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

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

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

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、

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

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

となり、確率を決める式は以下となる

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

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。

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

$n$ 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 $\mu_i=0\ \left(i=1,\cdots,n\right)$ とする。

提案分布からの抽出確率は以下となる。

$$\alpha\left(x_i^{(t+1)}|x_1^{(t+1)},\cdots,x_{i-1}^{(t+1)},x_{i+1}^{(t)},\cdots,x_n^{(t)}\right)$$

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

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、

$$q\left(x_i^{(t+1)}|x_1^{(t+1)},\cdots,x_{i-1}^{(t+1)},x_i^{(t)},\cdots,x_n^{(t)}\right)$$

$$=\prod_{j=1}^{i-1}\frac{1}{\sqrt{2\pi}\sigma_j}e^{-\frac{x_j^{(t+1)\,2}}{2\sigma_j^2}}\times\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{\left(x_i^{(t+1)}-x_i^{(t+1)}\right)^2}{2\sigma_i^2}}\times\prod_{k=i+1}^{n}\frac{1}{\sqrt{2\pi}\sigma_k}e^{-\frac{x_k^{(t)\,2}}{2\sigma_k^2}}$$

$$=q\left(x_i^{(t)}|x_1^{(t+1)},\cdots,x_i^{(t+1)},x_{i+1}^{(t)},\cdots,x_n^{(t)}\right)$$

より、以下となる。

$$\alpha\left(x_i^{(t+1)}|x_1^{(t+1)},\cdots,x_{i-1}^{(t+1)},x_{i+1}^{(t)},\cdots,x_n^{(t)}\right)$$

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

独立連鎖の場合は同様であるので省略する。

## 2.2 プログラムの動作

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

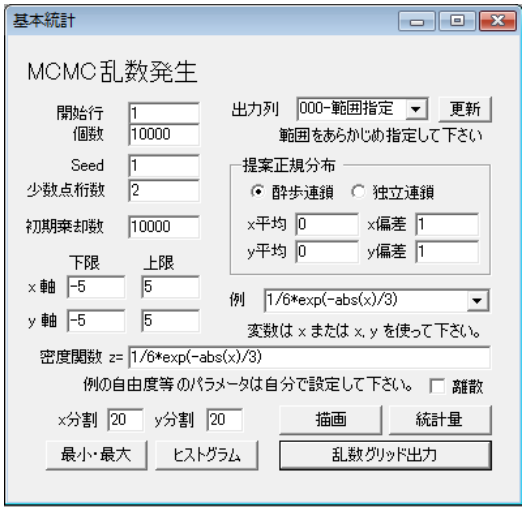



図 1　MCMC 乱数発生メニュー

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。

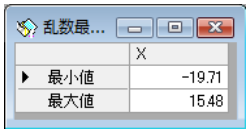



図 2　乱数発生の最小・最大

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。

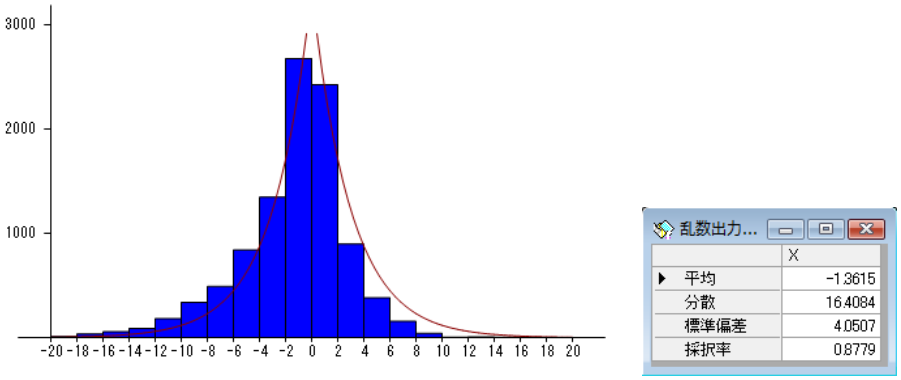



図 3a　乱数のヒストグラムと理論曲線（Seed=1）

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。

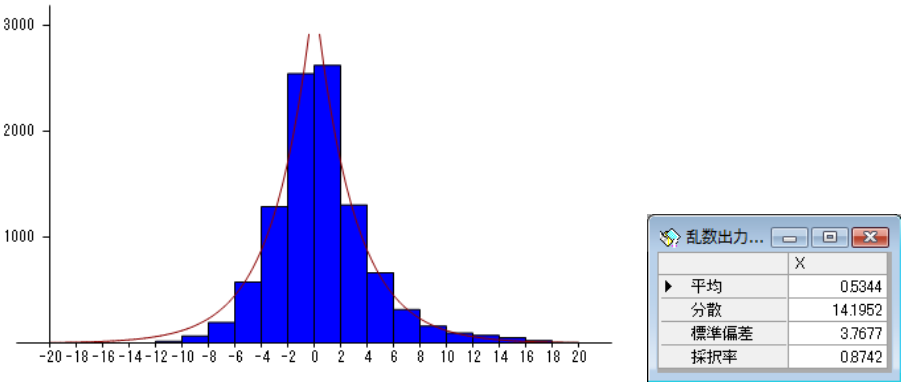



図 3b　乱数のヒストグラムと理論曲線（Seed=2）

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。

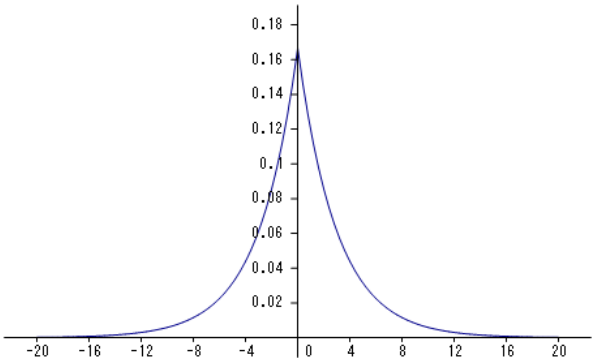



図 4　密度関数グラフ

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。

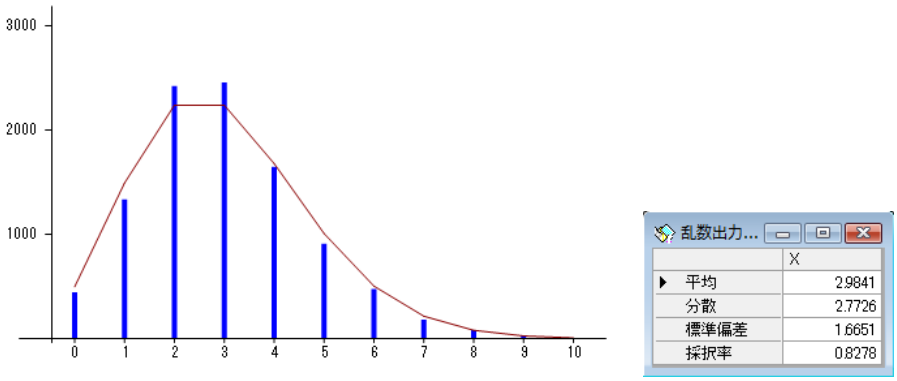



図 6　ポアソン分布

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。

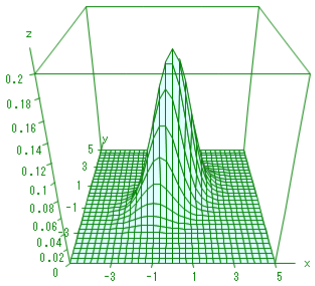


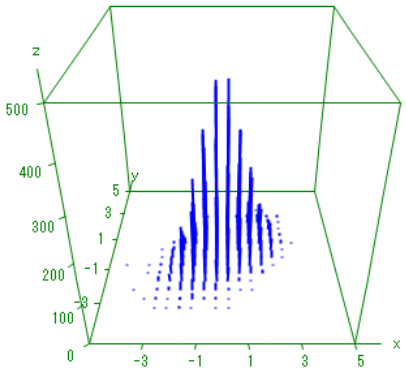

図 7　2 変量正規分布密度関数

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。


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。




乱数出力結果

| | 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。

[1] 豊田秀樹, マルコフ連鎖モンテカルロ法（統計ライブラリ）, 朝倉書店, 2008.

## ３．分布の検定

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－分布の検定］を選択すると図 1 のような分析メニューが表示される。

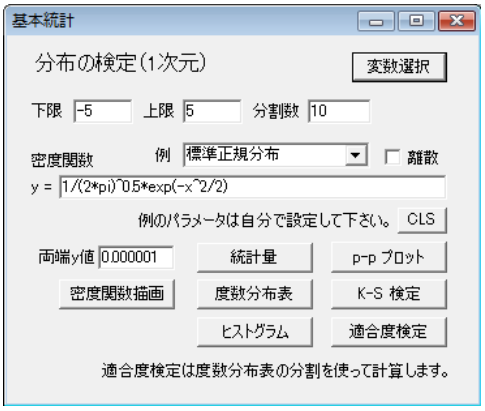



図 1　分析メニュー

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 のように選び、図 2b のようにパラメータと「下限」、「上限」を変更する。ここでは、自由度 3 の $\chi^2$ 分布を例にする。

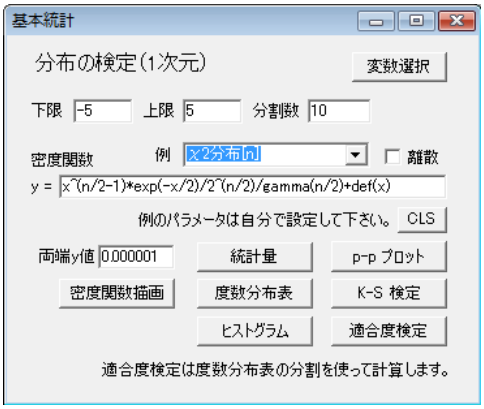



図 2a　密度関数の指定

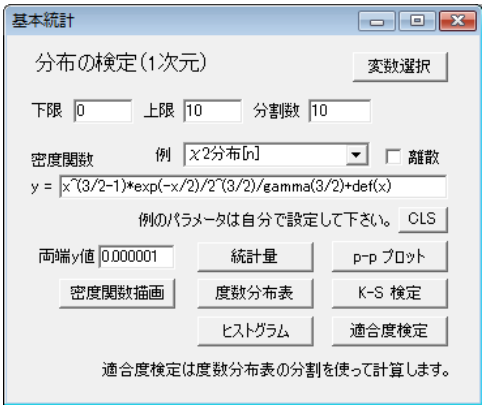



図 2b　パラメータと下限・上限の指定

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 | 0.0000 | 0.00 | 0.0000 |
| 0.0<=x<1.0 | 194 | 0.1940 | 198.72 | 0.1987 |
| 1.0<=x<2.0 | 216 | 0.2160 | 228.85 | 0.2288 |
| 2.0<=x<3.0 | 177 | 0.1770 | 180.78 | 0.1808 |
| 3.0<=x<4.0 | 134 | 0.1340 | 130.16 | 0.1302 |
| 4.0<=x<5.0 | 106 | 0.1060 | 89.67 | 0.0897 |
| 5.0<=x<6.0 | 63 | 0.0630 | 60.19 | 0.0602 |
| 6.0<=x<7.0 | 32 | 0.0320 | 39.71 | 0.0397 |
| 7.0<=x<8.0 | 28 | 0.0280 | 25.89 | 0.0259 |
| 8.0<=x<9.0 | 17 | 0.0170 | 16.72 | 0.0167 |
| 9.0<=x<10.0 | 11 | 0.0110 | 10.72 | 0.0107 |
| 10.0<=x<30.0 | 22 | 0.0220 | 18.56 | 0.0186 |
| 合計 | 1000 | 1.0000 | 999.97 | 1.0000 |

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 4.4 では「10.0<=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。

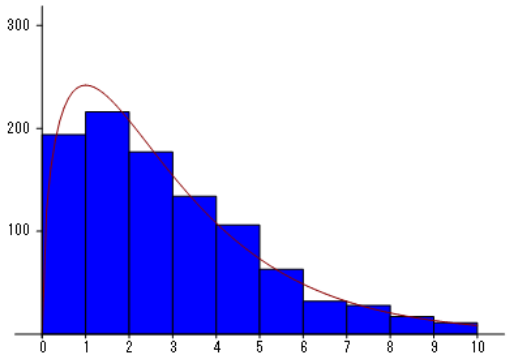



図 5　連続分布のヒストグラム

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。

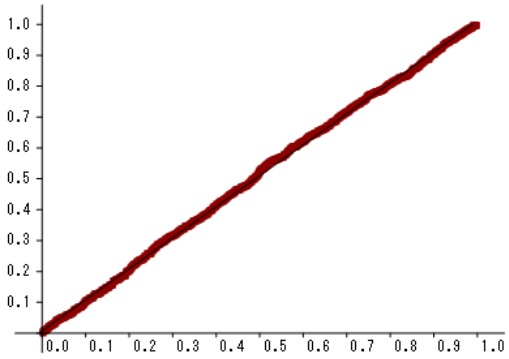



図 6　p-p プロット

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。

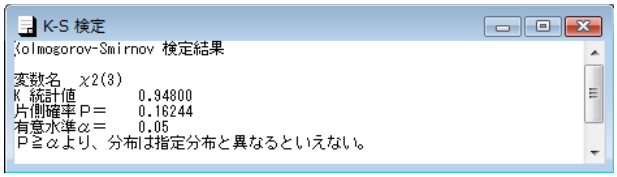



図 7　K-S 検定結果

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。

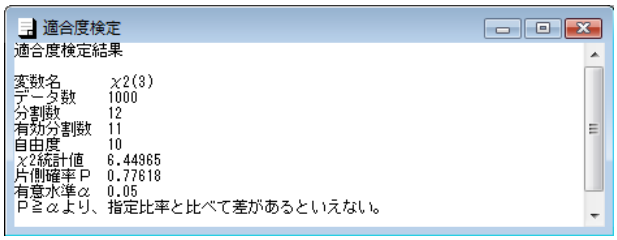



図 8　適合度検定結果

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。

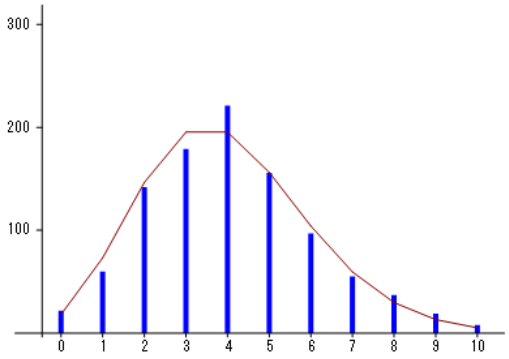
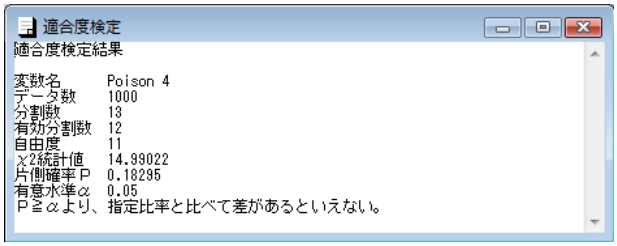



図 10　離散分布のヒストグラム

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。

## ４．異常検知について

製造現場における検査過程では多くのデータが測定されるが、正常なデータと異常なデータの迅速な選別は品質管理の上で非常に重要である。ここではその主要な理論について説明し、それを実践するプログラムを紹介する。理論はデータが多変量正規分布に従うと仮定される場合とそうでない場合を扱う。データが多変量正規分布に従う場合、マハラノビス距離の 2 乗を元にしたホテリングの t² 統計量に基づく判定法を用いる。また、多変量正規分布に従わない場合は、確率的な解釈も可能な混合正規分布モデルを仮定する方法を用いている。また、１次元データについては、ガンマ分布による異常検知の方法も加えている。

### 4.1　異常検知についてのプログラムの考え方

**多変量正規分布に基づく異常検知**

一般に $p$ 変数の多変量正規分布の密度関数は以下で与えられる。

$$f(\mathbf{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p\,|\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}{}^t(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

データ $D=\{\mathbf{x}_1,\cdots,\mathbf{x}_\lambda\}$ が与えられた場合の対数尤度関数は以下で与えられる。

$$L(\mu\ \Sigma\ D) = -\frac{p\ N}{2}\ \log(2\pi) - \frac{N}{2}\ \log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{\lambda=1}^{N}{}^t(\mathbf{x}_\lambda-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}\ (\mathbf{x}_\lambda-\boldsymbol{\mu}) \qquad (1)$$

我々は最尤法を用いて(2)式を最大化するが、その解は以下となる。

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{\lambda=1}^{N}\mathbf{x}_\lambda\ ,\quad \hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{\lambda=1}^{N}(\mathbf{x}_\lambda-\hat{\boldsymbol{\mu}})^t(\mathbf{x}_\lambda-\hat{\boldsymbol{\mu}})$$

ここで、同じ正規分布の確率変数 $\mathbf{x}'$ に対する異常度 $a(\mathbf{x}')$ を $-2\log f(\mathbf{x}'\,|\,\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$ を元に以下のように定義する。

$$a(\mathbf{x}') = {}^t(\mathbf{x}_\lambda-\hat{\boldsymbol{\mu}}\ \hat{\boldsymbol{\Sigma}}^{-1})\ \mathbf{x}'_\lambda-\hat{\boldsymbol{\mu}} \qquad (2)$$

ここで(2)式と $-2\log f(\mathbf{x}'\,|\,\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$ の差は定数であるので、評価関数として本質的な差はない。また(2)式は１次元変数の場合の変数の標準化の一般形である。

(2)式については、以下のように定数を掛けると、分布が自由度 $p, N-p$ の F 分布に従うことが知られている。

$$T^2 \equiv \frac{N-p}{(N+1)p}{}^t(\mathbf{x}'_\lambda-\hat{\boldsymbol{\mu}})\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}'_\lambda-\hat{\boldsymbol{\mu}}) \sim F_{p,N-p}$$

この $T^2$ をホテリング統計量という。

異常検知には、この統計量を使って確率の値を指定するか、直接 $T^2$ 値を指定して閾値とする。

**混合多変量正規分布に基づく異常検知**

$p$ 変数、$n$ 群混合多変量正規分布の密度関数は、群 $\alpha$ の確率密度関数

$$f_\alpha(\mathbf{x}\,|\,\boldsymbol{\mu}_\alpha,\boldsymbol{\Sigma}_\alpha)=\frac{1}{\sqrt{(2\pi)^p\,|\,\boldsymbol{\Sigma}_\alpha\,|}}\exp\left[-\frac{1}{2}{}^t(\mathbf{x}-\boldsymbol{\mu}_\alpha)\boldsymbol{\Sigma}_\alpha^{-1}(\mathbf{x}-\boldsymbol{\mu}_\alpha)\right]\quad(\alpha=1,\cdots,n)$$

を利用して以下で与えられる。

$$f(\mathbf{x}\,|\,\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_\alpha,\boldsymbol{\Sigma}_1,\cdots,\boldsymbol{\Sigma}_\alpha)=\sum_{\alpha=1}^{n}\pi_\alpha\,f_\alpha(\mathbf{x}\,|\,\boldsymbol{\mu}_\alpha,\boldsymbol{\Sigma}_\alpha)$$

$$=\sum_{\alpha=1}^{n}\frac{\pi_\alpha}{\sqrt{(2\pi)^p\,|\,\boldsymbol{\Sigma}_\alpha\,|}}\exp\left[-\frac{1}{2}{}^t(\mathbf{x}-\boldsymbol{\mu}_\alpha)\boldsymbol{\Sigma}_\alpha^{-1}(\mathbf{x}-\boldsymbol{\mu}_\alpha)\right]$$

ここに、$\pi_\alpha$ は群 $\alpha$ の生起確率である。

この密度関数に従うデータ $D=\{\mathbf{x}_1,\cdots,\mathbf{x}_\lambda\}$ による対数尤度は以下である。

$$L(\pi_1,\cdots,\pi_n,\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_1,\cdots,\boldsymbol{\Sigma}_n\,|\,D)=\sum_{\lambda=1}^{N}\log\left[\sum_{\alpha=1}^{n}\pi_\alpha\,f_\alpha(\boldsymbol{\mu}_\alpha,\boldsymbol{\Sigma}_\alpha\,|\,\mathbf{x}_\lambda)\right]$$

$$=\sum_{\lambda=}^{N}\log\left[\sum_{\alpha=1}^{n}\frac{\pi_\alpha}{\sqrt{(2\pi)^p\,|\,\boldsymbol{\Sigma}_\alpha\,|}}\exp\left\{-\frac{1}{2}{}^t(\mathbf{x}_\lambda-\boldsymbol{\mu}_\alpha)\boldsymbol{\Sigma}_\alpha^{-1}(\mathbf{x}_\lambda-\boldsymbol{\mu}_\alpha)\right\}\right]$$

最尤法を用いてこの対数尤度の最大値を求めるが、その際以下のアルゴリズムを利用する。

1）パラメータ $\pi_\alpha,\boldsymbol{\mu}_\alpha,\boldsymbol{\Sigma}_\alpha$ に初期値 $\hat{\pi}_\alpha,\hat{\boldsymbol{\mu}}_\alpha,\hat{\boldsymbol{\Sigma}}_\alpha$ を与える。

2）$\hat{\pi}_\alpha,\hat{\boldsymbol{\mu}}_\alpha,\hat{\boldsymbol{\Sigma}}_\alpha$ の値を用いて、各データの群 $\alpha$ への帰属度 $q_\alpha(\mathbf{x}_\lambda)$ を以下で求める。

$$q_\alpha(\mathbf{x}_\lambda)=\frac{\hat{\pi}_\alpha\,f_\alpha(\mathbf{x}_\lambda\,|\,\hat{\boldsymbol{\mu}}_\alpha,\hat{\boldsymbol{\Sigma}}_\alpha)}{\sum_{\beta=1}^{n}\hat{\pi}_\beta\,f_\beta(\mathbf{x}_\lambda\,|\,\hat{\boldsymbol{\mu}}_\beta,\hat{\boldsymbol{\Sigma}}_\beta)}$$

3）この帰属度を使い、新しいパラメータを以下のように決定する。

$$\hat{\pi}_\alpha=\frac{1}{N}\sum_{\lambda=1}^{N}q_\alpha(\mathbf{x}_\lambda),\qquad\hat{\boldsymbol{\mu}}_\alpha=\sum_{\lambda=1}^{N}q_\alpha(\mathbf{x}_\lambda)\mathbf{x}_\lambda\Big/\sum_{\lambda=1}^{N}q_\alpha(\mathbf{x}_\lambda),$$

$$\hat{\boldsymbol{\Sigma}}_\alpha=\sum_{\lambda=1}^{N}q_\alpha(\mathbf{x}_\lambda)(\mathbf{x}_\lambda-\hat{\boldsymbol{\mu}}_\alpha)^t(\mathbf{x}_\lambda-\hat{\boldsymbol{\mu}}_\alpha)\Big/\sum_{\lambda=1}^{N}q_\alpha(\mathbf{x}_\lambda)$$

4）新しいパラメータと元のパラメータを比較し、十分近ければ（プログラムではすべての成分が 0.001 未満）終了し、そうでなければ2）へ戻る。

この方法によって求めたパラメータを使って、異常判定には以下の指標を用いる。

$$a(\mathbf{x}')=-\ \mathrm{l}\sum_{\alpha=1}^{n}\hat{g}_\alpha\,f_\alpha\ \mathbf{x}'\ \hat{\boldsymbol{\mu}}_\alpha\ \hat{\boldsymbol{\Sigma}}_\alpha$$

判定基準はこの指標を小さい方から順番に並べ、分位点を閾値として決めるか、直接指標の閾値を指定する。

このモデルの適合度は赤池情報量基準 AIC、ベイズ情報量基準 BIC などを使って求める。今このモデルのパラメータ数を $M_n$ とすると、AIC と BIC はそれぞれ以下のように表現される。

$$AIC = -2\log(L(\Theta \mid D)) + 2M_n$$

$$BIC = -2\log(L(\Theta \mid D)) + M_n \log n, \qquad M_n = \frac{p}{2}(n+1)(n+2)$$

ここに、$L(\Theta \mid D)$ はパラメータを $\Theta$ で代表させて書いた対数尤度である。具体的には(8)式に求めたパラメータの値を代入したものである。適合度を求めるために、交差検証を使う方法も考えられるが、プログラムでは使用していない。

### 4.2　プログラムの利用法

メニュー［分析－ＯＲ－品質管理－異常検知］を選択すると、図 1 のようなメニューが表示される。

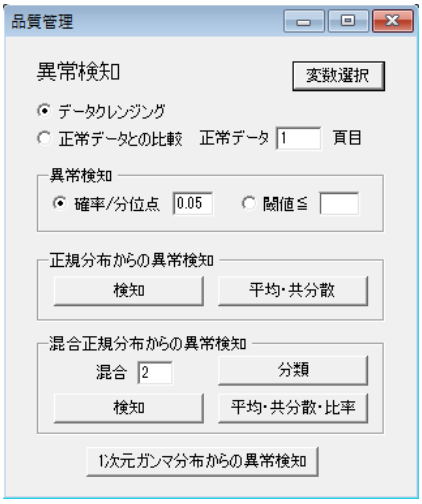



図 1　異常検知分析メニュー

このプログラムでは、グリッドエディタに表示されているデータから、異常データを選別する「データクレンジング」機能と、別頁にある正常データを用いて、現在表示されているデータの異常データを選別する「正常データとの比較」機能がある。後者の場合は、正常データがどの頁にあるかを指定しなければならない。

分析は実用的なレベルでは、(多変量)「正規分布からの異常検知」と（多変量)「混合正規分布からの異常検知」がある。前者は 1 つの正規分布からのデータのずれを検知するもので、マハラノビス距離を基にした手法である。これは分布が限定されているが、多くのデータでほぼ正規分布の仮定が成り立つと考えられるので、利用範囲は広い。後者は、正規分布が仮定できない場合で、しかもどのような分布になっているか予想が困難な場合に適用が可能である。これは、分布を複数の正規分布の重ね合わせとして考えるモデルで、いくつの正規分布の重ね合わせて考えると効果的かという判断まで可能である。

他に「１次元ガンマ分布からの異常検知」があるが、これは変数が 1 つの場合にしか適

用できないので、現実には利用しにくいかも知れない。以後、分かり易さと一般性を重視して２変数のデータを用いて、順を追ってこのプログラムについて説明をする。

図２に示すファイル「異常検知 1(正規分布).txt」の３頁目は、２変数で異常値を含んだデータである。

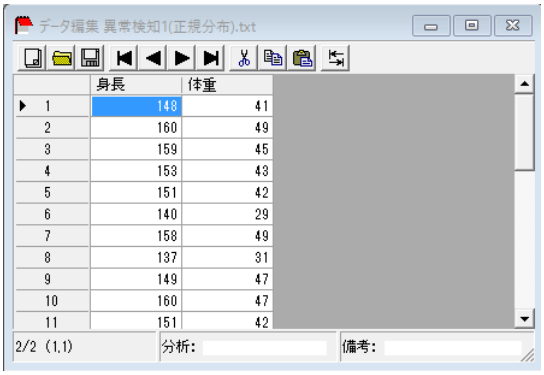



図２　異常値を含んだデータ

この中から異常値を検出するには、「データクレンジング」ラジオボタンを選び、「確率/分位点」ラジオボタンを選んで、異常値の確率値を指定する（この場合は確率値となる）。ここでは 5%に設定している。その後、正規分布からの異常検知の「検知」ボタンをクリックすると図３のような出力結果を得る。


図３　データクレンジング検知結果

出力は、このデータから求めた多変量正規分布の平均からのマハラノビス距離の 2 乗、それを元にしたホテリング統計量（F 分布の f 値）、その検定確率、異常かどうかの判定である。判定は正常と異常でそれぞれ 0 または 1 で出力される。

次に、「平均・共分散」ボタンをクリックすると、図４のように、平均や共分散等のパラメータ推定値等と共に、異常の判定に使われるホテリング統計量の閾値が出力される。利用者はこの値を参考にして、閾値としてホテリング統計量を用いてもよい。


図４　パラメータ推定値

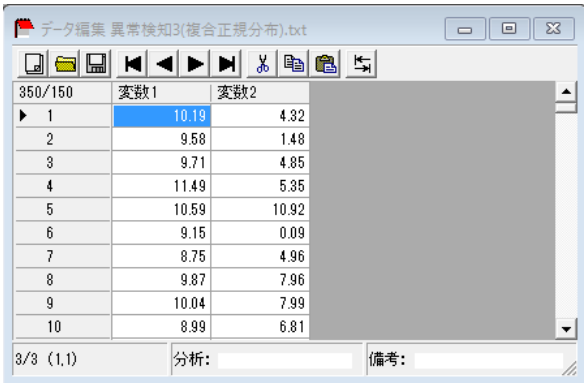
次に「正常データとの比較」ラジオボタンをクリックし、同じファイルの 1 頁目を開く、正常データは 2 頁目に入っているものとして、「正常データ」テキストボックスの中に 2 を入力する。「検知」ボタンをクリックすると、図 5 のように 2 頁目の正常データから求められるパラメータを元にした、異常検知結果が出力される。


| | Maha2乗 | 異常度(f値) | 確率 | 異常 |
|---|---|---|---|---|
| 1 | 2.295 | 1.036 | 0.368 | 0 |
| 2 | 0.484 | 0.219 | 0.805 | 0 |
| 3 | 10.957 | 4.948 | 0.014 | 1 |
| 4 | 0.368 | 0.166 | 0.848 | 0 |
| 5 | 2.384 | 1.077 | 0.354 | 0 |
| 6 | 0.645 | 0.292 | 0.749 | 0 |
| 7 | 2.763 | 1.248 | 0.303 | 0 |
| 8 | 2.048 | 0.925 | 0.408 | 0 |
| 9 | 2.412 | 1.089 | 0.350 | 0 |
| 10 | 6.417 | 2.898 | 0.072 | 0 |

図 5　正常データとの比較検知結果

出力項目については図 3 と同様である。「平均・共分散」ボタンをクリックした結果は、正常データを元にした結果であり、図 4 と同じ様式であるので省略する。

ファイル「異常検知 3(複合正規分布).txt」を図 6 のように読み込み、「データクレンジング」ラジオボタンを選択し、データ処理の結果を見てみよう。




| 350/150 | 変数1 | 変数2 |
|---|---|---|
| 1 | 10.19 | 4.32 |
| 2 | 9.58 | 1.48 |
| 3 | 9.71 | 4.85 |
| 4 | 11.49 | 5.35 |
| 5 | 10.59 | 10.92 |
| 6 | 9.15 | 0.09 |
| 7 | 8.75 | 4.96 |
| 8 | 9.87 | 7.96 |
| 9 | 10.04 | 7.99 |
| 10 | 8.99 | 6.81 |

3/3 (1,1)　分析:　備考:

図 6　非正規分布のデータ

このデータは、実際には 2 つの正規分布を合わせたものであるが、今の段階ではそれが分からないものとする。仮に「混合」テキストボックスを 2 とし、処理を進める。後にこの数字を変更して最も良いモデルを選択する。

「分類」ボタンをクリックすると、図 7 のように、2 つの群についてのデータの帰属度と分類結果が得られる。


| | 帰属度1 | 帰属度2 | 分類 |
|---|---|---|---|
| 343 | 0.999 | 0.001 | 1 |
| 344 | 1.000 | 0.000 | 1 |
| 345 | 0.999 | 0.001 | 1 |
| 346 | 0.984 | 0.016 | 1 |
| 347 | 0.999 | 0.001 | 1 |
| 348 | 0.999 | 0.001 | 1 |
| 349 | 0.999 | 0.001 | 1 |
| 350 | 1.000 | 0.000 | 1 |
| 351 | 0.000 | 1.000 | 2 |
| 352 | 0.000 | 1.000 | 2 |

図 7　レコード毎の帰属度と分類結果

この分類結果をコピーして、元のデータに図 8 のように貼り付け、

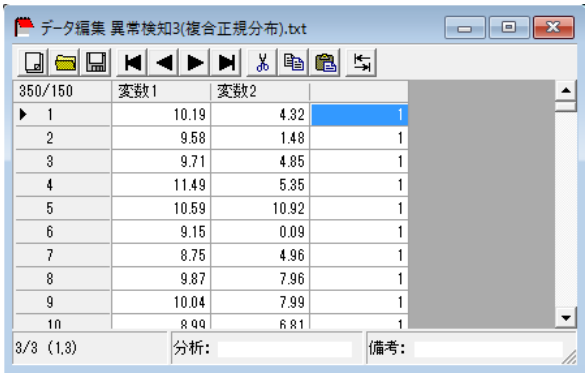



図 8　分類項目の貼り付け

分析「相関と回帰分析」の「先頭列で群分け」ラジオボタンを選択して、図 9 のような散布図を描くことも可能である。

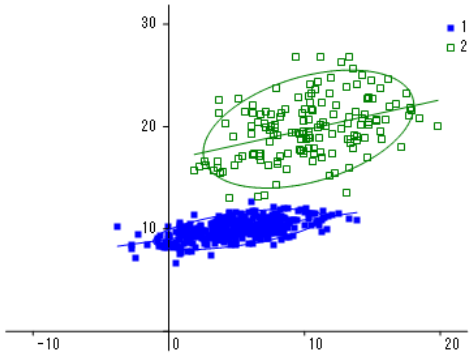



図 9　データの散布図

但し、このグラフには、2σ の確率楕円を加えてある。

次に「平均・共分散・比率」ボタンをクリックすると、図 10 のように 2 つの群の平均と共分散、群の生起確率の推測値などが表示される。この中で、一番下の BIC はモデル判定によく利用されるベイズ情報量基準と呼ばれるもので、この値が小さいほど良いモデルとして評価される。


図 10　2 群の場合のパラメータ推定値

現在の 2 群の場合は BIC=4860.110 であるが、3 群にすると 4888.488 となり、2 群の方が良いモデルであると判断される。これは、2 群を故意に作ったモデルであるので、当然の結果である。

最後に「検知」ボタンをクリックすると、図 11 のように異常度の値と判別結果が表示される。


| | 異常度(a値) | 異常 |
|---|---|---|
| 97 | 5.259 | 0 |
| 98 | 4.859 | 0 |
| 99 | 4.379 | 0 |
| 100 | 3.285 | 0 |
| 101 | 3.165 | 0 |
| 102 | 3.323 | 0 |
| 103 | 7.567 | 1 |
| 104 | 3.786 | 0 |
| 105 | 3.606 | 0 |
| 106 | 3.473 | 0 |

図 11　混合正規分布からの検知

判定には「確率/分位点」ラジオボックスの分位点を利用している。

1 次元ガンマ分布でも同様の結果の表示となるので、ここでは省略するが、ガンマ分布の場合は、2 つのパラメータが推測される。

**参考文献**

[1] 井出剛, 入門機械学習による異常検知, コロナ社, 2015.

## ５．生存時間分析について

生存時間分析は中途打ち切りを含むデータから死亡危険率や生存確率分布を予測する分析手法である。この分析は生物の生存時間だけでなく、機械の故障までの時間などにも利用できる。そのため、死亡という言葉は、あるイベントが発生するまでの時間とした方が的を得ているが、ここでは慣例的に使われてきた死亡や生存という言葉を使うことにする。

### 5.1　生存時間分析の基礎

時刻$t=0$に$l(0)$個の個体があり、死亡や観測打ち切りなどで、時刻$t$に個体数が$l(t)$個になっているものとする。時刻$t$からの単位時間の間に死亡する割合$p(t)=-\dfrac{dl(t)}{dt}$は、以下で与えられると仮定する。

$$-\frac{dl(t)}{dt}=\mu(t)l(t)$$

ここに$\mu(t)$は時刻$t$における死力という。

上式を時刻$t$と時刻$t+h$の間で定積分すると以下の関係を得る。

$$\log l(t+h)-\log l(t)=-\int_t^{t+h}\mu(\tau)\,d\tau=-\int_0^{h}\mu(t+\tau)\,d\tau$$

これより、

$$l(t+h)=l(t)\exp\left[-\int_0^{h}\mu(t+\tau)\,d\tau\right]$$

ここで、$p(h;t)=\exp\left[-\int_0^{h}\mu(t+\tau)d\tau\right]$とおくと、$p(h;t)$は時間$t\sim t+h$の間の期間生存率と呼ばれる。この期間生存率は、以下のようになる。

$$p(h;t)=\frac{l(t+h)}{l(t)}$$

同様にして、期間死亡率$q(h;t)$も以下のように与えられる。

$$q(h;t)=1-p(h;t)\equiv\frac{l(t)-l(t+h)}{l(t)}\equiv\frac{d(h;t)}{l(t)}$$

ここに$d(h;t)$は期間死亡数を表す。

特に、$h=1$とした区間生存率、区間死亡率を単に時刻$t$での生存率$p(t)$、死亡率$q(t)$という。

時刻$t$以降の生存時間の合計$T(t)$を個体の数で割った$e(t)$を平均余命という。

$$e(t)=\int_t^{\infty}d(\tau)d\tau\Big/l(t)=T(t)/l(t)$$

また、$t=0$での平均余命を平均寿命という。

死亡の発生までの時間を確率変数$T$とする確率分布を考え、その密度関数を$f(t)$、分布関数を$F(t)$とすると、これらには以下の関係がある。分布関数$F(t)$は累積死亡関数である。

$$F(t)=P(0\le T\le t)=\int_0^t f(\tau)\,d\tau$$

これに対して、時刻$t$まで生きる確率を表す関数を累積生存関数$S(t)$といい、以下で表す。

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^{\infty} f(\tau)d\tau$$

時刻$t$における死亡発生危険率をハザード関数（故障率関数）$\lambda(t)$といい、以下のように定義する。

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\log S(t)$$

死亡率$q(t)$は以下のように定義されるが、

$$q(t) = \int_t^{t+1} f(\tau)d\tau$$

時間の分割が小さい場合は、近似的にハザード関数の積分としても表される。

$$q(t) = \int_t^{t+1} \lambda(\tau)d\tau$$

このハザード関数を積分した累積ハザード関数$\Lambda(t)$は以下のように定義される。

$$\Lambda(t) = \int_0^t \lambda(\tau)d\tau = -\log S(t)$$

逆に累積生存関数は、以下のように表される。

$$S(t) = e^{-\Lambda(t)}$$

累積生存関数は$t \to \infty$で$S(t) \to 0$であるから、累積ハザード関数は$t \to \infty$で$\Lambda(t) \to \infty$でなければならない。

生存時間分布には、主に指数分布とワイブル分布が仮定される。

指数分布の確率密度関数は以下で与えられる。

$$f(t) = \lambda e^{-\lambda t} \quad (t \geq 0)$$

分布関数と累積生存関数はそれぞれ以下で与えられる。

$$F(t) = 1 - e^{-\lambda t}, \quad S(t) = e^{-\lambda t}, \quad (t \geq 0)$$

確率変数の平均、分散、標準偏差はそれぞれ以下で与えられる。

$$E[T] = \frac{1}{\lambda}$$

$$V[T] = \frac{1}{\lambda^2}$$

$$\sigma = \sqrt{V[T]} = \frac{1}{\lambda}$$

ハザード関数は定数で与えられる。

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

ワイブル分布の確率密度関数は以下で与えられる。

$$f(t) = a(b(t/b)^{a-1}\left[e^{-(t/b)^a}\right] \quad (t \geq 0)$$

分布関数と累積生存関数はそれぞれ以下で与えられる。

$$F(t) = 1 - \exp\left[-(t/b)^a\right], \qquad S(t) = \exp\left[-(t/b)^a\right] \qquad (t \geq 0)$$

確率変数の平均、分散、標準偏差はそれぞれ以下で与えられる。

$$E[T] = b\Gamma\left(1 + \frac{1}{a}\right)$$

$$V[T] = b^2\left[\Gamma\left(1 + \frac{2}{a}\right) - \Gamma^2\left(1 + \frac{1}{a}\right)\right]$$

$$\sigma = \sqrt{V[T]}$$

ハザード関数は以下で与えられる。

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{(a/b)(t/b)^{a-1}\exp\left[-(t/b)^a\right]}{\exp\left[-(t/b)^a\right]} = (a/b)(t/b)^{a-1} = at^{a-1}/b^a$$

実際のハザード関数は、初期段階で値が大きく、しばらく時間が経つと安定期に入り、最終的な段階でまた値が大きくなる。安定期では指数分布が使われ、初期段階ではワイブル分布がよく利用される。最終段階ではどちらの分布もあまり当てはまりが良くないと言われている。

### 5.2 実際のデータの取り扱い

観測対象 $\lambda = 1, \cdots, N$ に対して、生存時間を $t_\lambda = 0$ から $t_\lambda = T_\lambda$（打ち切りのないデータ）、$t_\lambda = 0$ から $t_\lambda = T_\lambda^+$（打ち切りのあるデータ、実際のデータでは 17+ 等と表記）とする。この終了時刻 $T_\lambda$ を 0 から順番に並べた時刻を $t_0 = 0, t_1, \cdots, t_m$（同一のものもある）とし、$t_m$ ですべて死亡および打ち切りが確認されたものとする。これに対して、一定の時間間隔で時刻を取る方法もある。各時点での生存数を $l_i$、$t_i < t \leq t_{i+1}$ の間に死亡した数を $d_i$、打ち切りになった数を $w_i$ とする。これらを使って、死亡のリスクにさらされた数を $r_i = l_i - w_i/2$ とする。

死亡の期間発生率 $q_i$ と期間生存率 $p_i$ は以下で与えられる。

$$q_i = d_i/l_i, \qquad p_i = 1 - q_i$$

累積生存関数 $S_i$、期間イベント発生確率 $f_i$、ハザード関数 $\lambda_i$ は以下のように計算される。

$$S_i = \prod_{k=0}^{i-1} p_k, \qquad f_i = q_i S_i, \qquad \lambda_i = f_i/S_i = q_i$$

このような累積生存関数の推定法を Kaplan-Meier の product-limit 推定法という。累積生存関数 $S_i$ のばらつきを表す標準誤差 $S.E.[S_i]$ は近似的に以下で与えられることが知られている。

$$S.E.[S_i] = S_{i-1}\sqrt{\sum_{k=0}^{i-1} \frac{d_k}{l_k(l_k - d_k)}}$$

平均生存時間 $\mu_i$ は以下で与えられる。

$$\mu_i = \sum_{k=0}^{i} S_k(t_{k+1} - t_k)$$

指数分布やワイブル分布の見極めは、累積ハザード関数に関する以下の関係を利用し、グラフが直線になるか否かで判断することができる。

指数分布　　$-\log S(t) = \lambda t$

ワイブルプンプ　　$\log(-\log S) = a\log(t/b) = a\log t - a\log b$

指数分布やワイブル分布のパラメータの推定は、以下の式によって与えられる。

指数分布　$S(t) = e^{-\lambda t}$

$$\lambda = -\sum_{i=0}^{m-1} t_i \log S_i \Big/ \sum_{i=0}^{m-1} t_i^2$$

ワイブル分布　$S(t) = \exp\left[-(t/b)^a\right]$

$t_i' = \log t_i$,　$S_i' = \log(-\log S_i)$　として、

$$a = \sum_{i=1}^{m-1}(t_i' - \bar{t}')(S_i' - \bar{S}') \Big/ \sum_{i=1}^{m-1}(t_i' - \bar{t}')^2 \quad , \quad b = \exp\left[-(\bar{S}' - a\bar{t}')/a\right]$$

分類数$G$の個体群について、生存時間データの差の検定を行うには以下の性質を用いる。第$r$分類群の$t_i$時点での期間死亡数を$d_i^r$、生存数を$l_i^r$として

$$O_r = \sum_{i=0}^{m-1} d_i^r \quad , \quad E_r = \sum_{i=0}^{m-1} l_i^r (d_i/l_i) \quad , \quad \text{ここに、} l_i = \sum_{r=1}^{G} l_i^r \quad , \quad d_i = \sum_{r=1}^{G} d_i^r$$

を計算し、以下の近似的な関係を用いて群間の差を検定する。

$$\chi^2 = \sum_{r=1}^{G} \frac{(O_r - E_r)^2}{E_r} \sim \chi_{G-1}^2$$

この検定を Peto & Peto の log-rank 検定という。

比例ハザードモデルはハザート関数に対して以下の仮定を行う。

$$\lambda(t \mid \mathbf{x}, \boldsymbol{\beta}) = \lambda_0(t) \exp(^t\mathbf{x}\boldsymbol{\beta}) \quad \text{ここに、} {}^t\mathbf{x}\boldsymbol{\beta} = \sum_{i=1}^{p} x_i\beta_i + \beta_0$$

Cox の比例ハザードモデルでは$\lambda_0(t)$と定数項$\beta_0$について議論しないが、ワイブル比例ハザードモデルでは

$$\lambda(t \mid \mathbf{x}, \boldsymbol{\beta}) = (a/b)\left(t/b\right)^{a-1} = at^{a-1}b^{-a} = at^{a-1}\exp(^t\mathbf{x}\boldsymbol{\beta})$$

として、時間に関してワイブル分布のハザード関数を仮定する。

Cox の比例ハザードモデルでは、尤度関数に対して近似的な部分尤度関数を考えて処理を行う。その対数尤度は以下で与えられる[3]。

$$\log L(\boldsymbol{\beta}) = \sum_{i=0}^{m-1}\left[\sum_{j\in D_i} {}^t\mathbf{x}_j\boldsymbol{\beta} - d_i \log\left(\sum_{j\in R_i} \exp(^t\mathbf{x}_j\boldsymbol{\beta})\right)\right]$$

ここに、$\boldsymbol{\beta}$ は定数項を除いた偏回帰係数ベクトル、$D_i$ は $t_i < t \le t_{i+1}$ で亡くなった個体の集合、$R_i$ は時刻 $t_i$ で生存が確認されている個体の集合である。これを最大化するようにニュートン・ラフソン法を使って $\boldsymbol{\beta}$ を求める。ここではそのための準備として以下の値を示しておく。

$$\mathbf{U} \equiv \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}) = \sum_{i=1}^{m-1} \left[ \sum_{i \in_i D} \mathbf{x}_j - d_i \sum_{\in_i R} w_j \mathbf{x}_j \Big/ \sum_{\in_i R} w_j \right]$$

$$\mathfrak{I} \equiv -\frac{\partial^2}{\partial \boldsymbol{\beta} \partial^t \boldsymbol{\beta}} \log L(\boldsymbol{\beta}) = \sum_{i=1}^{m-1} d_i \left[ \sum_{j \in R_i} w_j \mathbf{x}_j {}^t\mathbf{x}_j \Big/ \sum_{j \in R_i} w_j - \sum_{j \in R_i} w_j \mathbf{x}_j \sum_{j \in R_i} w_j {}^t\mathbf{x}_j \Big/ \left( \sum_{j \in R_i} w_j \right)^2 \right]$$

ここに $w_j = \exp({}^t\mathbf{x}_j \boldsymbol{\beta})$

この $\mathbf{U}$ をスコアベクトル、$\mathfrak{I}$ を情報行列という。$\boldsymbol{\beta}$ の推定値は以下の計算を繰り返して求める。

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

ここに右肩の添え字はニュートン・ラフソン法のループの段階を表している。

Weibull 比例ハザードモデルは、ハザード関数に対して以下の仮定を行う。

$$\lambda(t) = \frac{f(t)}{S(t)} = (a/b)(t/b)^{a-1} = a t^{a-1} / b^a = a t^{a-1} \exp({}^t\mathbf{x}\boldsymbol{\beta})$$

これより、$b = \exp(-{}^t\mathbf{x}\boldsymbol{\beta}/a)$ であるから、$\mu \equiv E[T] = b\,\Gamma(1+1/a)$ より、

$$\eta \equiv {}^t\mathbf{x}\boldsymbol{\beta} = -a \log b = -a \log(\mu / \Gamma(1+1/a))$$

となり、右辺が一般化線形モデルの連結関数となる。

この関係を用いて、累積生存関数と密度関数を求めると以下となる。

$$S(t) = \exp\left[ -(t/b)^a \right] = \exp\left[ -t^a \exp(-a\,{}^t\mathbf{x}\boldsymbol{\beta}) \right]$$

$$f(t) = a t^{a-1} \exp(-a\,{}^t\mathbf{x}\boldsymbol{\beta}) \exp\left[ -t^a \exp(-a\,{}^t\mathbf{x}\boldsymbol{\beta}) \right]$$

打ち切りデータと非打ち切りデータをそれぞれ $\delta_i = 0,1$ と区別し、尤度を求めると以下となる。添え字 $i$ について、ここでは個体の番号として使っている。

$$L(\alpha, \boldsymbol{\beta}) = \prod_{i=1}^{N} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

さらに、対数尤度は以下となる。

$$\log L(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^{N} \left[ \delta_i \log \lambda(t_i) + \log S(t_i) \right]$$

$$= \sum_{i=1}^{N} \left[ \delta_i \log\left( a t_i^{a-1} e^{{}^t\mathbf{x}_i\boldsymbol{\beta}} \right) - t_i^a e^{{}^t\mathbf{x}_i\boldsymbol{\beta}} \right]$$

$$= \sum_{i=1}^{N} \left[ \delta_i \left( \log a + (a-1)\log t_i + {}^t\mathbf{x}_i\boldsymbol{\beta} \right) - t_i^a e^{{}^t\mathbf{x}_i\boldsymbol{\beta}} \right]$$

これを微分して、スコアベクトル $\mathbf{U}$ と情報行列 $\mathfrak{I}$ をもとめると以下となる。

$$\boldsymbol{\beta}' = \begin{pmatrix} a \\ \boldsymbol{\beta} \end{pmatrix}, \qquad \mathbf{U} = \begin{pmatrix} \partial \log L/\partial a \\ \partial \log L/\partial \boldsymbol{\beta} \end{pmatrix}, \qquad \mathfrak{I} = -\begin{pmatrix} \partial^2 \log L/\partial a^2 & \partial^2 \log L/\partial a \partial^t \boldsymbol{\beta} \\ \partial^2 \log L/\partial a \partial \boldsymbol{\beta} & \partial^2 \log L/\partial \boldsymbol{\beta} \partial^t \boldsymbol{\beta} \end{pmatrix}$$

ここに

$$\frac{\partial}{\partial a} \log L = \sum_{i=1}^{N} \left[ \delta_i / a + (1 - \delta_i) \log t_i e^{a \log t_i + {}^t \boldsymbol{\beta} \mathbf{x}_i} \right]$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log L = \sum_{i=1}^{N} \left[ \delta_i \mathbf{x}_i - \mathbf{x}_i e^{a \log t_i + {}^t \boldsymbol{\beta} \mathbf{x}_i} \right]$$

$$\frac{\partial^2}{\partial a^2} \log L = \sum_{i=1}^{N} \left[ -\delta_i / a^2 - (\log t_i)^2 e^{a \log t_i + {}^t \boldsymbol{\beta} \mathbf{x}_i} \right]$$

$$\frac{\partial^2}{\partial a \partial \boldsymbol{\beta}} \log L = -\sum_{i=1}^{N} (\log t_i \mathbf{x}_i e^{a \log t_i + {}^t \boldsymbol{\beta} \mathbf{x}_i})$$

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial^t \boldsymbol{\beta}} \log L = -\sum_{i=1}^{N} \mathbf{x}_i {}^t \mathbf{x}_i e^{a \log t_i + {}^t \boldsymbol{\beta} \mathbf{x}_i}$$

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。

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

ここに右肩の添え字はニュートン・ラフソン法のループの段階を表している。

求められたパラメータを使って、個人の予想寿命を以下のように求めることができる。

$$\mu \equiv E[T] = b\Gamma(1 + 1/a), \quad b = \exp(\mathbf{x}\boldsymbol{\beta}/a)$$

この値を実際の寿命と比較することで相関係数等を求めることもできる。

### 5.3 プログラムの利用法

メニュー［分析－多変量解析他－生存時間分析］を選択すると、図 1 のような分析実行メニューが表示される。

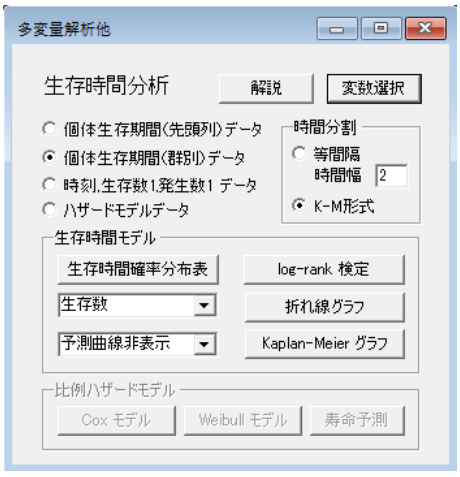



図 1　生存時間分析実行メニュー

この分析のデータ形式は大きく分けて 3 種類ある。１つは個体の生存時間を元にしたデータで、先頭列で分類される形式とすでに群別に並べられている形式に分けられる。これらの形式は基本統計のデータ形式に類似している。次に、すでに生命表に近い形式になっ

ているデータである。これは、観測時刻、その時点での生存個体数、その時点より後で次の時点までに死亡する期間発生数が、すでに表の形式になっているデータである。生存個体数と期間発生数は複数組入力が可能である。詳しくはサンプルを見てもらいたい。最後は、ハザードモデルデータで、重回帰分析などと同様の形式である。最初と最後の形式で、通常のデータと異なる部分は、観測の打ち切りデータが含まれる点である。打ち切りデータは、観測を打ち切られた時点の数値の後ろに+記号を付けて表す。観測が打ち切られた際の扱いは、生存数から打ち切られたデータ数の半分を引いて、死亡リスクに晒されたデータ数として処理している[1]。

最初に図 2 の単独データを元に説明をする。

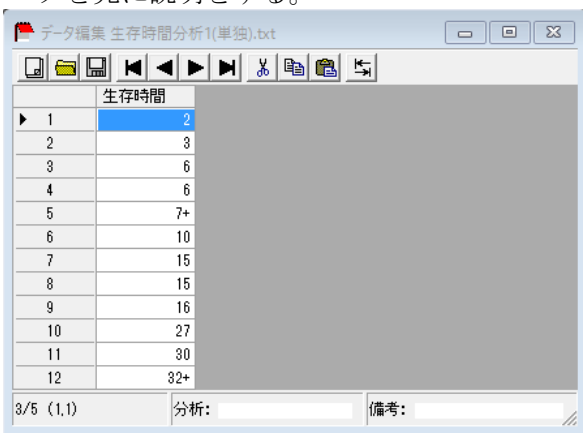
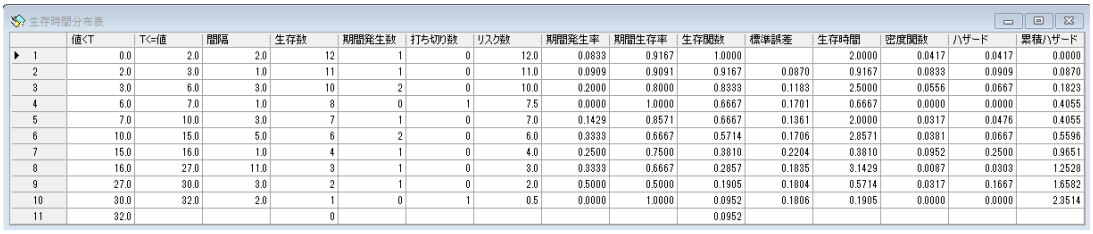



図 2　単独データ（生存時間分析 1(単独).txt 3 頁目）

このデータでは、2 個体が観測を打ち切られている。

「個体生存時間(群別)データ」ラジオボタンを選択し、変数選択を実行して、「生存時間分布表」ボタンをクリックすると図 3 のような結果が表示される。




| | 値<T | T<=値 | 間隔 | 生存数 | 期間発生数 | 打ち切り数 | リスク数 | 期間発生率 | 期間生存率 | 生存関数 | 標準誤差 | 生存時間 | 密度関数 | ハザード | 累積ハザード |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.0 | 2.0 | 2.0 | 12 | 1 | 0 | 12.0 | 0.0833 | 0.9167 | 1.0000 | | 2.0000 | 0.0417 | 0.0417 | 0.0000 |
| 2 | 2.0 | 3.0 | 1.0 | 11 | 1 | 0 | 11.0 | 0.0909 | 0.9091 | 0.9167 | 0.0870 | 0.9167 | 0.0833 | 0.0909 | 0.0870 |
| 3 | 3.0 | 6.0 | 3.0 | 10 | 2 | 0 | 10.0 | 0.2000 | 0.8000 | 0.8333 | 0.1183 | 2.5000 | 0.0556 | 0.0667 | 0.1823 |
| 4 | 6.0 | 7.0 | 1.0 | 8 | 0 | 1 | 7.5 | 0.0000 | 1.0000 | 0.6667 | 0.1701 | 0.6667 | 0.0000 | 0.0000 | 0.4055 |
| 5 | 7.0 | 10.0 | 3.0 | 7 | 1 | 0 | 7.0 | 0.1429 | 0.8571 | 0.6667 | 0.1361 | 2.0000 | 0.0317 | 0.0476 | 0.4055 |
| 6 | 10.0 | 15.0 | 5.0 | 6 | 2 | 0 | 6.0 | 0.3333 | 0.6667 | 0.5714 | 0.1706 | 2.8571 | 0.0381 | 0.0667 | 0.5596 |
| 7 | 15.0 | 16.0 | 1.0 | 4 | 1 | 0 | 4.0 | 0.2500 | 0.7500 | 0.3810 | 0.2204 | 0.3810 | 0.0952 | 0.2500 | 0.9651 |
| 8 | 16.0 | 27.0 | 11.0 | 3 | 1 | 0 | 3.0 | 0.3333 | 0.6667 | 0.2857 | 0.1835 | 3.1429 | 0.0087 | 0.0303 | 1.2528 |
| 9 | 27.0 | 30.0 | 3.0 | 2 | 1 | 0 | 2.0 | 0.5000 | 0.5000 | 0.1905 | 0.1804 | 0.5714 | 0.0317 | 0.1667 | 1.6582 |
| 10 | 30.0 | 32.0 | 2.0 | 1 | 0 | 1 | 0.5 | 0.0000 | 1.0000 | 0.0952 | 0.1806 | 0.1905 | 0.0000 | 0.0000 | 2.3514 |
| 11 | 32.0 | | | 0 | | | | | | 0.0952 | | | | | |

図 3a　生存時間分布表結果

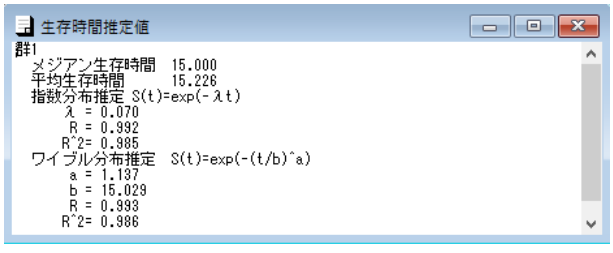



図 3b　生存時間推定値

図 3a では、様々な指標が区切られた時点毎に表示されている。ここで特に大切な指標は、「生存関数」と「ハザード」である。これらはそれぞれ、その時点まで生存している確率とその時点での死亡の危険率の意味を持つ。図 3b ではこのデータを元にした、メジアン生

存時間、平均生存時間、指数分布を仮定したパラメータ推定値、ワイブル分布を仮定したパラメータ推定値が表示される。図 3b の「S(t)= 」の式は生存関数の数式を表している。当てはめの良さは、指数分布では累積ハザード関数 $-\log S$ の実測値（表から求めた値）と $t$ の値の相関係数（後者で $\lambda t$ の値を用いても同じ）、ワイブル分布では累積ハザード関数の対数 $\log(-\log S)$ の実測値（表から求めた値）と $\log t$ の値の相関係数（後者で $a\log(t/b)$ の値を用いても同じ）を求めて表示している。

図 3a の生存時間分布表の中で、生存数、累積生存関数、ハザード関数、累積ハザード関数については、コンボボックスで設定して、「折れ線グラフ」ボタンをクリックすると表示される。ここでは累積生存関数とハザード関数についてのグラフを図 4a と図 4b に示す。

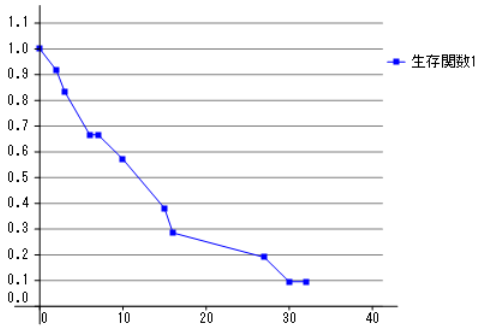



図 4a　生存関数

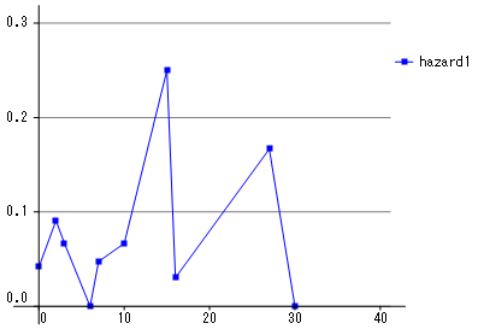



図 4b　ハザード関数

また、同じコンボボックスで「指数分布確認」または「ワイブル分布確認」を選択すると、図 5a と図 5b のような図が表示される。

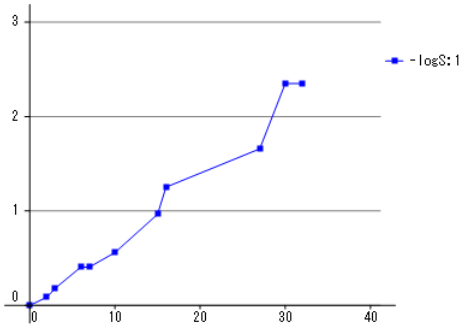



図 5a　累積生存関数

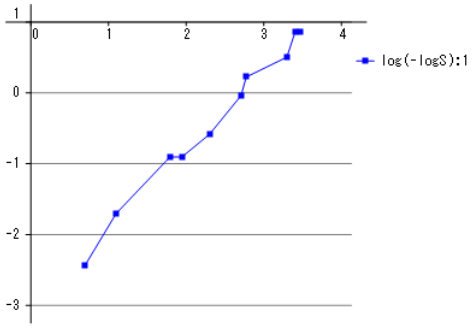
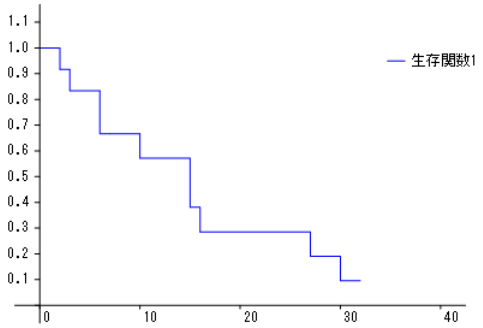



図 5b　ハザード関数

生存時間が指数分布またはワイブル分布に従うならば、それぞれの累積生存関数の時間依存性からこの点列は直線状に並ぶ。指数分布はワイブル分布の特殊な場合であるので、指数分布が成り立つ場合はワイブル分布も成り立つ。

生存時間関数の Kaplan-Meier 推定のグラフは、「Kaplan-Meier グラフ」ボタンをクリックして表示される。その際、左のコンボボックスで指定して、指数分布またはワイブル分布の予想曲線を描くこともできる。予想曲線のないグラフと、ワイブル分布の予想曲線を付けて描いたグラフを図 6a と図 6b に示す。



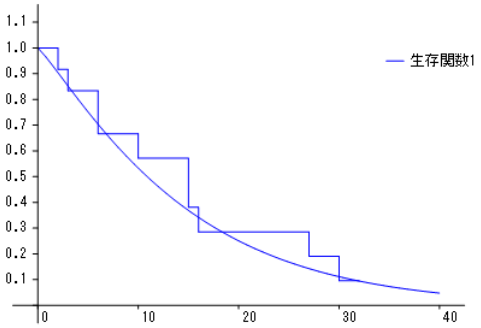



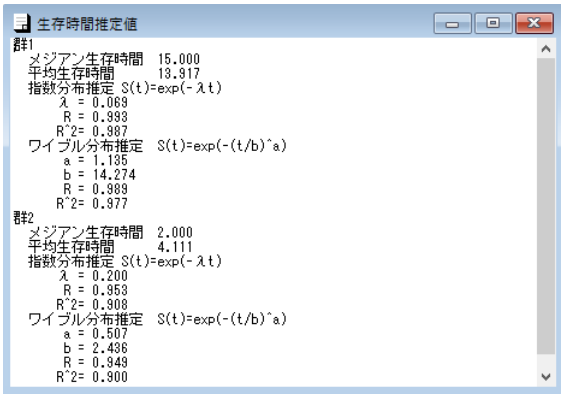

図 6a　Kaplan-Meier 生存関数グラフ　図 6b　予想曲線付き Kaplan-Meier グラフ

複数群の生存時間分布表は、先頭列で群分けデータ（生存時間分析 2(2 群比較).txt）または群別データを元に図 7a と図 7b のように縦に並べて表示される。

生存時間分布表

| | 値<T | T<=値 | 間隔 | 生存数 | 期間発生数 | 打ち切り数 | リスク数 | 期間発生率 | 期間生存率 | 生存関数 | 標準誤差 | 生存時間 | 密度関数 | ハザード | 累積ハザード |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.0 | 1.0 | 1.0 | 12 | 0 | 0 | 12.0 | 0.0000 | 1.0000 | 1.0000 | | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 2 | 1.0 | 2.0 | 1.0 | 12 | 1 | 0 | 12.0 | 0.0833 | 0.9167 | 1.0000 | 0.0000 | 1.0000 | 0.0833 | 0.0833 | 0.0000 |
| 3 | 2.0 | 3.0 | 1.0 | 11 | 1 | 0 | 11.0 | 0.0909 | 0.9091 | 0.9167 | 0.0870 | 0.9167 | 0.0833 | 0.0909 | 0.0870 |
| 4 | 3.0 | 6.0 | 3.0 | 10 | 2 | 0 | 10.0 | 0.2000 | 0.8000 | 0.8333 | 0.1183 | 2.5000 | 0.0556 | 0.0667 | 0.1823 |
| 5 | 6.0 | 7.0 | 1.0 | 8 | 1 | 0 | 8.0 | 0.1250 | 0.8750 | 0.6667 | 0.1701 | 0.6667 | 0.0833 | 0.1250 | 0.4055 |
| 6 | 7.0 | 9.0 | 2.0 | 7 | 0 | 0 | 7.0 | 0.0000 | 1.0000 | 0.5833 | 0.1627 | 1.1667 | 0.0000 | 0.0000 | 0.5390 |
| 7 | 9.0 | 10.0 | 1.0 | 7 | 1 | 0 | 7.0 | 0.1429 | 0.8571 | 0.5833 | 0.1423 | 0.5833 | 0.0833 | 0.1429 | 0.5390 |
| 8 | 10.0 | 15.0 | 5.0 | 6 | 2 | 0 | 6.0 | 0.3333 | 0.6667 | 0.5000 | 0.1684 | 2.5000 | 0.0333 | 0.0667 | 0.6931 |
| 9 | 15.0 | 16.0 | 1.0 | 4 | 1 | 0 | 4.0 | 0.2500 | 0.7500 | 0.3333 | 0.2041 | 0.3333 | 0.0833 | 0.2500 | 1.0986 |
| 10 | 16.0 | 22.0 | 6.0 | 3 | 0 | 0 | 3.0 | 0.0000 | 1.0000 | 0.2500 | 0.1667 | 1.5000 | 0.0000 | 0.0000 | 1.3863 |
| 11 | 22.0 | 27.0 | 5.0 | 3 | 1 | 0 | 3.0 | 0.3333 | 0.6667 | 0.2500 | 0.1250 | 1.2500 | 0.0167 | 0.0667 | 1.3863 |
| 12 | 27.0 | 30.0 | 3.0 | 2 | 1 | 0 | 2.0 | 0.5000 | 0.5000 | 0.1667 | 0.1614 | 0.5000 | 0.0278 | 0.1667 | 1.7918 |
| 13 | 30.0 | 32.0 | 2.0 | 1 | 1 | 0 | 1.0 | 1.0000 | 0.0000 | 0.0000 | 0.1596 | 0.0000 | 0.0000 | 0.5000 | 0.0000 |
| 14 | 32.0 | | | 0 | | | | | | 0.0000 | | | | | |
| 1 | 0.0 | 1.0 | 1.0 | 9 | 4 | 0 | 9.0 | 0.4444 | 0.5556 | 1.0000 | | 1.0000 | 0.4444 | 0.4444 | 0.0000 |
| 2 | 1.0 | 2.0 | 1.0 | 5 | 1 | 0 | 5.0 | 0.2000 | 0.8000 | 0.5556 | 0.2981 | 0.5556 | 0.1111 | 0.2000 | 0.5878 |
| 3 | 2.0 | 3.0 | 1.0 | 4 | 2 | 0 | 4.0 | 0.5000 | 0.5000 | 0.4444 | 0.2070 | 0.4444 | 0.2222 | 0.5000 | 0.8109 |
| 4 | 3.0 | 6.0 | 3.0 | 2 | 0 | 0 | 2.0 | 0.0000 | 1.0000 | 0.2222 | 0.2772 | 0.6667 | 0.0000 | 0.0000 | 1.5041 |
| 5 | 6.0 | 7.0 | 1.0 | 2 | 0 | 0 | 2.0 | 0.0000 | 1.0000 | 0.2222 | 0.1386 | 0.2222 | 0.0000 | 0.0000 | 1.5041 |
| 6 | 7.0 | 9.0 | 2.0 | 2 | 1 | 0 | 2.0 | 0.5000 | 0.5000 | 0.2222 | 0.1386 | 0.4444 | 0.0556 | 0.2500 | 1.5041 |
| 7 | 9.0 | 10.0 | 1.0 | 1 | 0 | 0 | 1.0 | 0.0000 | 1.0000 | 0.1111 | 0.2095 | 0.1111 | 0.0000 | 0.0000 | 2.1972 |
| 8 | 10.0 | 15.0 | 5.0 | 1 | 0 | 0 | 1.0 | 0.0000 | 1.0000 | 0.1111 | 0.1048 | 0.5556 | 0.0000 | 0.0000 | 2.1972 |
| 9 | 15.0 | 16.0 | 1.0 | 1 | 0 | 0 | 1.0 | 0.0000 | 1.0000 | 0.1111 | 0.1048 | 0.1111 | 0.0000 | 0.0000 | 2.1972 |
| 10 | 16.0 | 22.0 | 6.0 | 1 | 1 | 0 | 1.0 | 1.0000 | 0.0000 | 0.0000 | 0.1048 | 0.0000 | 0.0000 | 0.1667 | 0.0000 |
| 11 | 22.0 | 27.0 | 5.0 | 0 | | | | | | 0.0000 | | | | | |
| 12 | 27.0 | 30.0 | 3.0 | | | | | | | | | | | | |
| 13 | 30.0 | 32.0 | 2.0 | | | | | | | | | | | | |
| 14 | 32.0 | | | | | | | | | | | | | | |

図 7a　2 群の生存時間分布表



生存時間推定値

```
群1
  メジアン生存時間  15.000
  平均生存時間      13.917
  指数分布推定 S(t)=exp(-λt)
       λ = 0.069
       R = 0.993
    R^2= 0.987
  ワイブル分布推定  S(t)=exp(-(t/b)^a)
       a = 1.135
       b = 14.274
       R = 0.989
    R^2= 0.977
群2
  メジアン生存時間  2.000
  平均生存時間      4.111
  指数分布推定 S(t)=exp(-λt)
       λ = 0.200
       R = 0.953
    R^2= 0.908
  ワイブル分布推定  S(t)=exp(-(t/b)^a)
       a = 0.507
       b = 2.436
       R = 0.949
    R^2= 0.900
```

図 7b　2 群の生存時間推定値

これ以外に、もっと群の違いを比較できる方法を考えて行きたい。

複数群の累積生存関数と Kaplan-Meire 累積生存関数グラフを図 8 と図 9 に示す。

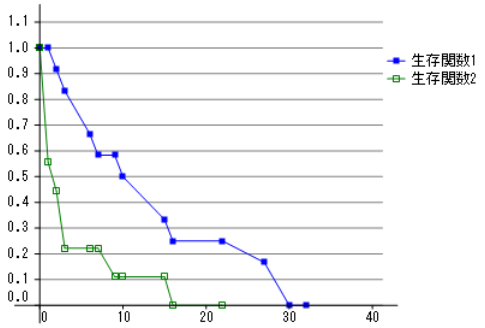



図 8　2 種類の累積生存関数グラフ

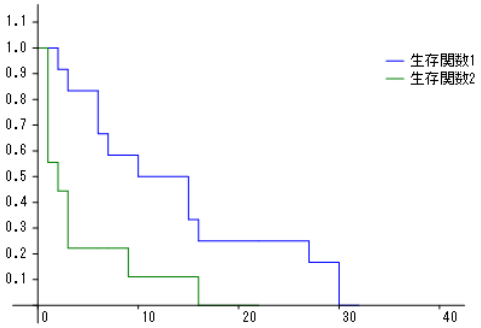



図 9　2 種類の Kaplan-Meier グラフ

複数群の累積生存関数間の差の log-rank 検定結果は、「log-rank 検定」ボタンをクリックすると図 10 のように表示される。

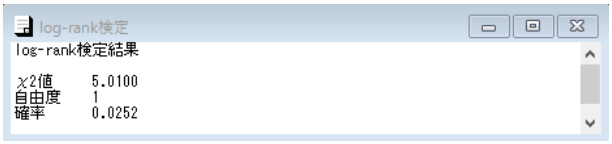



図 10　log-rank 検定結果

最後に、比例ハザードモデルの分析結果について示しておく。データは図 11 のような重回帰分析などと同じデータ形式である。

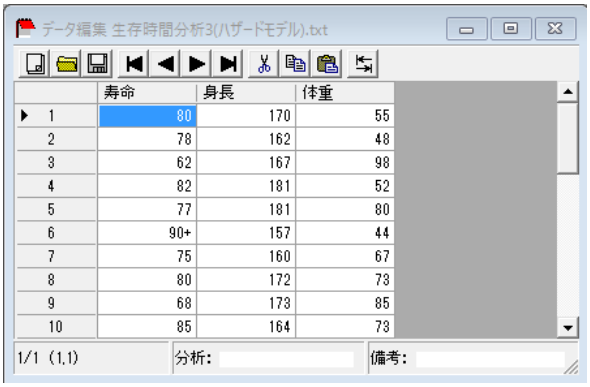



図 11　比例ハザードモデルデータ（生存時間分析 3(ハザードモデル).txt）

ハザードモデルでは Cox 比例ハザードモデルと Weibull 比例ハザードモデルを組み込んでいる。ハザード関数について、2 つのモデルとも以下の形を仮定する。

$$\lambda(t \mid \mathbf{x}, \boldsymbol{\beta}) = \lambda_0(t) \exp({}^t\mathbf{x}\boldsymbol{\beta}) \quad \text{ここに、} {}^t\mathbf{x}\boldsymbol{\beta} = \sum_{i=1}^{p} x_i \beta_i + \beta_0$$

Cox 比例ハザードモデルは $\lambda_0(t)$ や $\beta_0$ の推定は行わないが、分布の形に依存しない利点がある。Weibull ハザードモデルでは、時間部分にワイブル分布を仮定し、その 1 つのパラメータを説明変数で推定するという一般化線形モデルの形式を採用している。

$$\lambda(t \mid \mathbf{x}, \boldsymbol{\beta}) = (a/b)\left(t/b\right)^{a-1} = at^{a-1}b^{-a} = at^{a-1}\exp({}^t\mathbf{x}\boldsymbol{\beta})$$

「Cox モデル」ボタンをクリックした結果を図 12 に、「Weibull モデル」ボタンをクリックした結果を図 13 に示す。


Cox比例ハザードモデル結果

| | 偏回帰係数 | 標準偏差 | 両側確率 | 2.5%下限 | 2.5%上限 | EXP(b) |
|---|---|---|---|---|---|---|
| ▶ 身長 | -0.0247 | 0.0246 | 0.3165 | -0.0729 | 0.0236 | 9.756E-01 |
| 体重 | 0.0461 | 0.0154 | 0.0027 | 0.0159 | 0.0763 | 1.047E00 |

図 12　Cox 比例ハザードモデル結果


Weibull比例ハザードモデル結果

| | 偏回帰係数 | 標準偏差 | 両側確率 | 2.5%下限 | 2.5%上限 | EXP(b) |
|---|---|---|---|---|---|---|
| ▶ a | 11.7941 | 1.6723 | 0.0000 | 8.5163 | 15.0718 | |
| 身長 | -0.0274 | 0.0239 | 0.2512 | -0.0741 | 0.0194 | 9.730E-01 |
| 体重 | 0.0546 | 0.0157 | 0.0005 | 0.0237 | 0.0854 | 1.056E00 |
| 切片 | -50.7044 | 8.4355 | 0.0000 | -67.2380 | -34.1709 | 9.536E-23 |

図 13　Weibull 比例ハザードモデル

最後に Weibull 比例ハザードモデルが予想する生存時間の平均値と実際の観測値との比較を行ってみる。「寿命予測」ボタンをクリックすると図 14a と図 14b の結果が示される。


Weibull寿命予測

| | 寿命 | 寿命予測 | 残差 | b推定値 |
|---|---|---|---|---|
| 21 | 90 | 80.748 | 9.252 | 84.318 |
| 22 | 62 | 68.517 | -6.517 | 71.546 |
| 23 | 81 | 76.920 | 4.080 | 80.320 |
| 24 | 74 | 80.942 | -6.942 | 84.520 |
| 25 | 71 | 76.931 | -5.931 | 80.332 |
| 26 | 78 | 79.262 | -1.262 | 82.766 |
| 27 | 61 | 67.425 | -6.425 | 70.406 |
| 28 | 82 | 80.186 | 1.814 | 83.731 |
| 29 | 81 | 81.506 | -0.506 | 85.109 |
| 30 | 83 | 79.089 | 3.911 | 82.585 |
| R | 0.718 | R^2 | 0.516 | |

図 14a　寿命予測

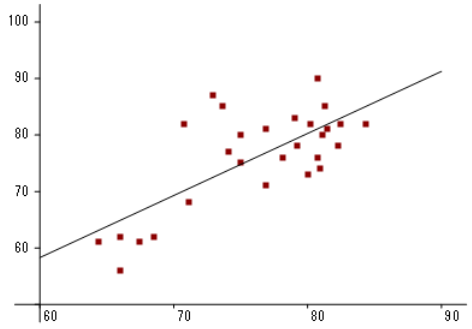



図 14b　実測/予測散布図

これには非打ち切りデータのみが用いられている。また、寿命予測の結果の最後に、予測値と実測値の相関係数の値とその 2 乗の値を表示している。

**参考文献**

[1] 打波守, Excel で学ぶ生存時間解析, オーム社, 2005.

[2] 柳井晴夫, 高木廣文編著, 多変量解析ハンドブック, 現代数学社, 1986.

[3] Annete J. Dobson, 田中豊他訳, 一般化線形モデル入門 原著第 2 版, 共立出版, 2008.