College Analysis レファレンスマニュアル

－ 基本統計 －

目次

１．概要

統計処理ソフトウェアは、様々な機関で、人力と時間をかけて、数え切れないほど多く作成されており、個人が作るものにはおのずと限界がある。しかし、統合的な教育プログラムを作るという立場からは避けて通れない道であり、その際ある種の独自性を打ち出す必要もある。

統計処理プログラムは一般に個々の分析プログラムの集合体となっており、ユーザーは必要に応じてそれらを選択して使い分ける。しかし、統計に不慣れな初心者にとってはどの分析をどのように利用するか、その判断こそが最も難しい。しかし、自分が行おうとする分析の位置付けが明確に示され、その指針がプログラム中にあれば、判断の手助けとなり、安心感を持って分析が実行できるに違いない。特に統計学の講義を受講している学生にとっては、このガイドラインが必要であろう。

分析の位置付けを明らかにするという考え方は主に検定手続きの中で実現されている。検定の体系（異論のある方もおられるかも知れないが）を図式化したメニューをダイアログボックスとして示し、その中から自分の利用する分析手法を選択する。この考え方は特に目新しいものではないが、必ずや学習の手助けになるものと信じる。

このシステム中で利用できる統計処理手法は、「２次元グラフ」、「３次元グラフ」、「分布と確率」、「密度関数グラフ」、「量から質変換」、「データ標準化」、の統計処理に関するユーティリティと、「質的データの集計」、「量的データの集計」、「質的データの検定」、「量的データの検定」、「相関係数と回帰分析」、「トレンドの検定」、「標本数の決定」、「区間推定」、という集計と検定、に分けられている。また、「質的データの検定」と「量的データの検定」は、さらに細かい具体的な分析手法に分かれている。

欠損値データの処理方法、有意水準の指定と片側・両側検定の区別、エディタからの変数の選択については、共通の設定項目としてコマンドボタンにより各分析から簡単に設定できるようになっている。これらには適当なデフォルト値が与えられ、初心者でも分析に不都合が生じないようになっている。ここではまず、集計と検定から話を始め、次に統計処理に関するユーティリティに進んで行く。

具体的な統計分析について説明を始める前に、欠損値の処理、有意水準の設定、エディタ上の変数の選択方法に関する設定事項について述べる。実行画面は図１で与えられるが、このメニューは各分析から共通に呼び出され、この中で指定された設定はプログラムの実行中値が保持される。

図 2.1.1 初期設定画面

欠損値の除去方法は、選択された変数についてのレコード単位の除去、データ毎の個別の除去、統計手法に応じた自動選択がある。有意水準の設定については、片側検定、両側検定、検定手法に応じて標準的なものを選択する自動選択がある。例えば、χ^2 検定とＦ検定は片側検定であり、ｔ検定その他については両側検定である。その数値は、パーセント表示で入力するが、デフォルトは 5%になっている。もちろん集計等のように有意水準に無関係なものについて、この値は無視される。

変数選択によって、エディタ上のデータから利用される変数が選ばれるが、左上のコンボボックスで変数名を選択することによって、それが左下のリストボックスに現れる。変数の選択順は分析によって意味を持つので（例えば順回帰分析で、最初の変数は目的変数等）、選択した変数の順番を入れ替えるためのボタンが用意されている。このメニューは単に変数だけ選択する分析では、左半分だけ表示されるようになっており、すべての分析で汎用的に利用される。

２．質的データの集計

分類データを対象とする質的データの集計画面は、メニュー［分析－基本統計－質的データの集計］を選択することによって図１のように表示される。

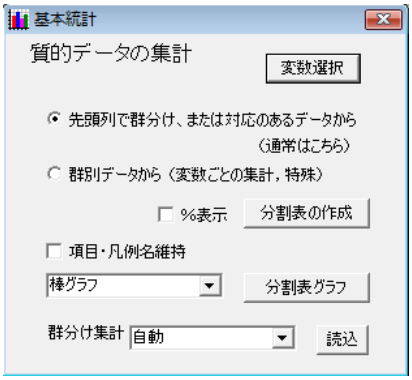
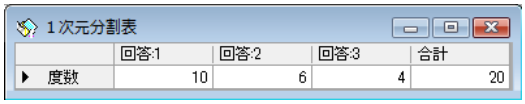


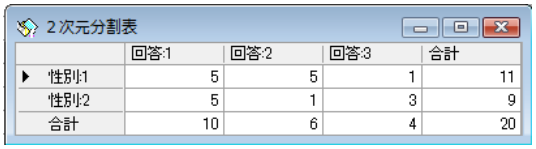
図１ 質的データの集計画面

分析画面で「分割表の作成」ボタンをクリックすることにより、項目ごとにデータ数を集計され、分割表が作られる。１つの変数を選んだ場合の１次元分割表と２つの変数を選んだ場合の２次元分割表の例を図２と図３に示す。分割表の表示の際、「%表示」チェックボックスにチェックを入れると、横方向の割合を%で表示する。



	回答:1	回答:2	回答:3	合計
▶ 度数	10	6	4	20

図２ １次元分割表



	回答:1	回答:2	回答:3	合計
▶ 性別:1	5	5	1	11
性別:2	5	1	3	9
合計	10	6	4	20

図３ ２次元分割表

「賛成」、「反対」など、データが文字列で表わされている場合でも集計が可能である。行と列の関係は設定の変数選択の順番で決まる。現在、分割は２次元分割表までである。これらの分割表は、質的データの検定のところでも作成することができる。これらの表示はグリッド表示の機能によって、簡単に行と列を入れ替えることもできる。

分割表は、コンボボックスからグラフの種類を選択し、「分割表グラフ」ボタンをクリックすると、

グラフとして表示することができる。グラフの種類には、棒グラフ、積み重ね棒グラフ、横棒グラフ、積み重ね横棒グラフ、横帯グラフ、0/1 回答横棒グラフ、円グラフ、がある。図 4 に棒グラフと円グラフ、図 5 に 2 つの変数の選択順を変えた積み重ね棒グラフを示す。

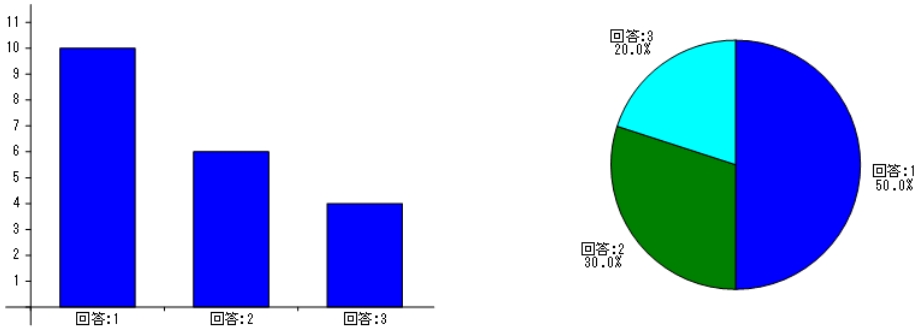


図 4 棒グラフと円グラフ

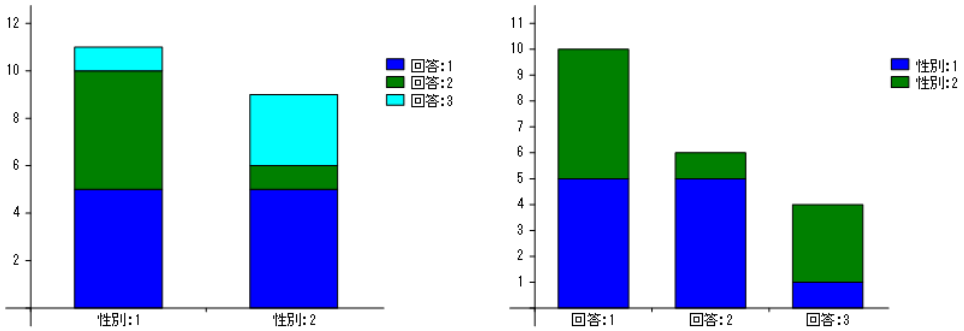


図 5 積み重ね棒グラフ

変数名はデフォルトのままであるが、グラフのメニュー（「項目名変更」、「データ・凡例名変更」）によって変数名や凡例名を付け替えることもできる。

0/1 回答横棒グラフは、複数の変数が 0/1 で回答されている複数回答などの場合に、それぞれの変数の 1 を選択した人の割合を横棒で表わすグラフである。必要な変数をすべて選択し、「群別データから」ラジオボタンを選択して実行すると結果の表示は例えば図 6 のようになる。ここではグラフメニュー（「%表示[ON/OFF]」によって横軸を%表示にしている。

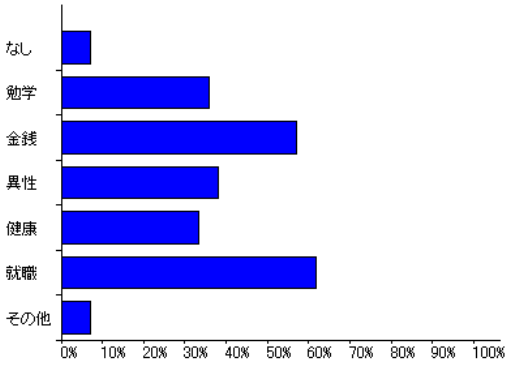


図６0/1 回答横棒グラフ

分析メニューの「群分け集計」は２つの群で円グラフなどを分けて表示する場合に利用される。

３．量的データの集計

量的データの集計の分析画面は、メニュー［分析－基本統計－量的データの集計］を選択すると図１のように示される。

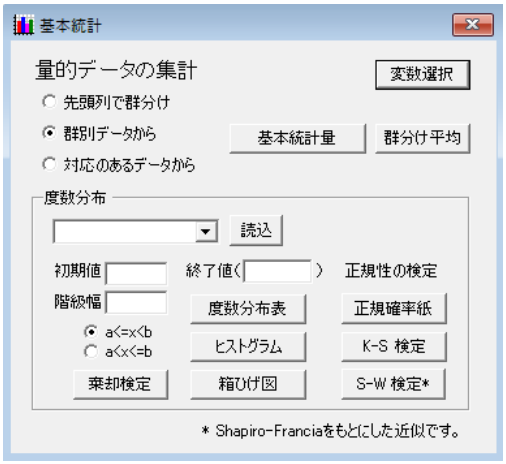


図１ 量的データ集計画面

変数選択で必要な変数を選択して「基本統計量」ボタンをクリックすると、図２のような結果が表示される。ここでは、１つの変数だけ選択したが、複数選択したり、「先頭列で群分け」ラジオボタンを選んで、ある変数で分けて表示することもできる。

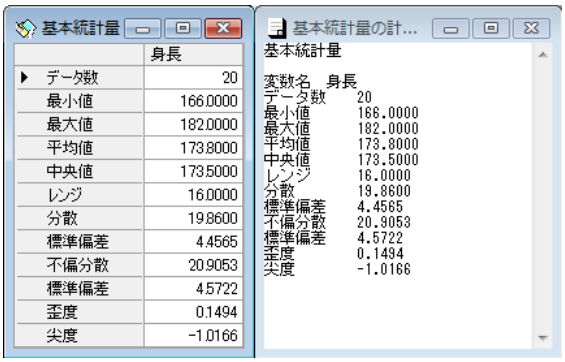


図２ 基本統計量

ここで、基本統計量という言葉は、分布の中心を表す指標に用いられることが多いので、本来は要約統計量とした方が良いのかも知れない。「群分け平均」ボタンは、「先頭列で群分け」ラジオボタンが選択されている場合、群ごとの平均値を見易く並べたものである。

基本統計量の定義は以下の通りである。

データ数　n

平均値　$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

中間値　Me

最大値　$\max\{x_i\}$

最小値　$\min\{x_i\}$

範囲　$\max\{x_i\} - \min\{x_i\}$

分散　$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

不偏分散　$u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

標準偏差　s　または　u

歪度　$a_3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$

尖度　$a_4 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4$

量的データの分布型を見るために、度数分布グループボックス内で、「読込」ボタンで表示用の変数を設定し、度数分布表とヒストグラムを図 3 と図 4 のように表示させることができる。

度数分布表　支出

	度数	相対度数(%)	累積度数	累積相対度数(%)
10<=x<20	8	4	8	4
20<=x<30	14	7	22	11
30<=x<40	36	18	58	29
40<=x<50	58	29	116	58
50<=x<60	48	24	164	82
60<=x<70	25	12.5	189	94.5
70<=x<80	8	4	197	98.5
80<=x<90	3	1.5	200	100

図 3 度数分布表

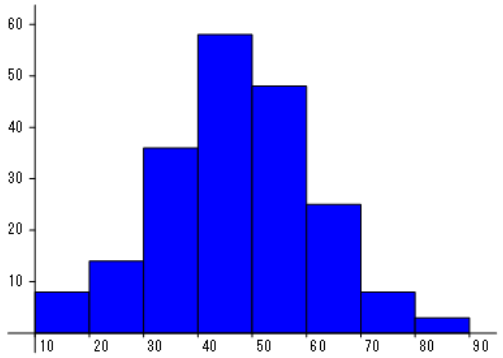


図 4 ヒストグラム

度数分布表には、度数・相対度数・累積度数・累積相対度数が含まれる。設定は自動になっているが、初期値、分割幅、終了値を指定してもよい。

箱ひげ図は、分布の比較を行う場合などに利用する簡易的な分布の表示法である。図 5 と図 6 に先頭列で群分けして比較した 2 つのデータについてのヒストグラムと箱ひげ図をそれぞれ示す。

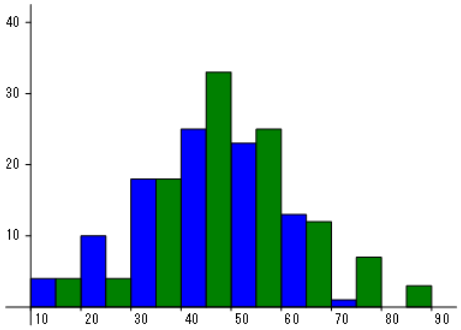


図 5 比較のためのヒストグラム

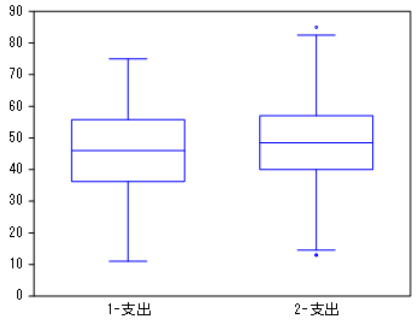


図 6 箱ひげ図

ヒストグラムは、度数分布グループボックス内で「読込」を行い、コンボボックスで「すべて」を選択する。箱ひげ図の箱の中央は平均値、箱の下と上は 25%、75% 分位点、ひげの最小は、データの最小値または-3σ 値の大きい方、最大は、データの最大値または 3σ 値の小さい方で、はみ出したデータは丸印で表わす。

データの正規性を見るために、「正規確率紙」による正規性の確認の方法（Q-Q プロットとも呼ぶ）も用意されている。これは特にデータ数が少なく、ヒストグラムが使えないような場合に有効である。図 7 に実行画面を示す。

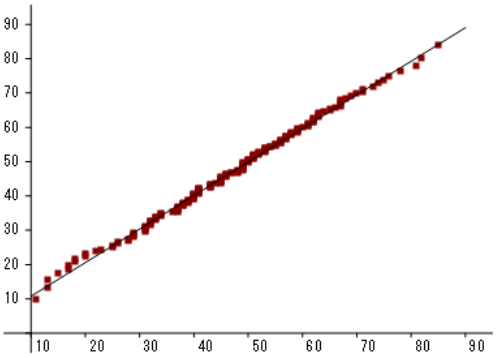


図 7 正規確率紙の方法

また、正規性の確認については、コルもゴロフ・スミルノフの検定（K-S 検定）やシャピロ・ウィルクの検定（K-S 検定）の近似の方法（作者の勉強不足で申し訳ありません）が含まれている。特に後者は、データ数があまり多くない場合に有効である。図 8 と図 9 に実行画面を示す。

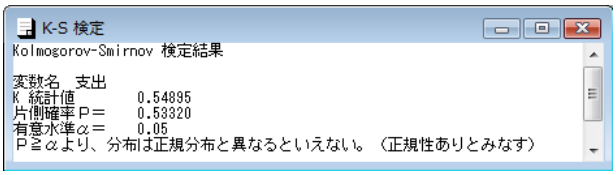


図 8 K-S 検定

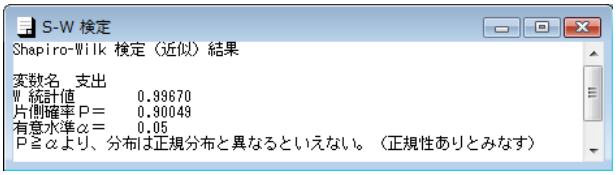


図 9　（近似）S-W 検定

データの中に飛び離れた値があり、これを分析から除くべきかどうか調べる必要がある場合、ここでは Grubbs-Smirnov 棄却検定が利用できる。飛び離れたデータが最大値 $x_{\max}$ である場合、それを除いたデータが正規分布かどうかまず確認する。正規分布の場合、以下の統計量 $T_{\max}$ を求め、

$$T_{\max} = \frac{x_{\max} - \bar{x}}{u}$$

それと全データ数を用いて数表から検定確率を調べる [1]。ここに、$\bar{x}$ と u はそれぞれ全データを用いた平均値と、不偏分散からの標準偏差である。

データが正規分布でない場合、対数正規分布も確認する。対数正規分布の場合は、データに対数変換を行って上と同様の検定を行う。正規分布でも対数正規分布でもない場合は、一応元データを用いて検定を行ってはいるが、信頼性はない。

飛び離れたデータが最小値 $x_{\min}$ である場合も全く同様に、以下の統計量 $T_{\min}$ を利用する。

$$T_{\min} = \frac{\bar{x} - x_{\min}}{u}$$

４．質的データの検定

質的指標の検定手順については、図１の分類を用いた。データ数の少ない場合など、この考え方が利用できないこともあるが、その対応は今後の課題とする。

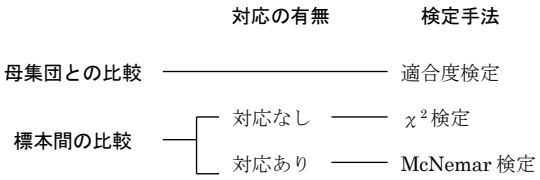


図１ 質的指標に関する検定手法の分類

利用者に検定手法の位置付けを明確に認識させるために、分析を選択するメニューを一般的な統計ソフトで見られる羅列的なものとせず、図１の形式をそのままメニュー化した。

具体的な実行画面を図２に示す。

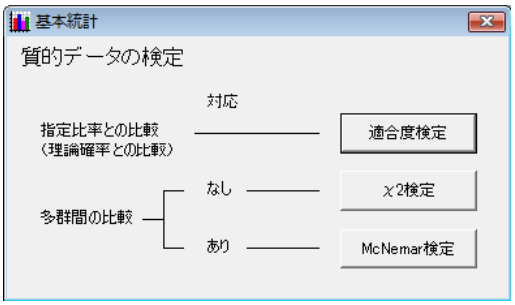


図２ 質的指標の検定画面

図２の検定のコマンドボタンから具体的な分析メニューが呼び出される。利用する分布公式については、図１の検定手法に応じて以下のようにまとめられる。

適合度検定

標本数 n，事象 i の出現回数 n_i，事象 i の母比率 p_i

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2_{k-1} \text{ 分布}$$

χ^2 検定

標本数 n，要因 i 事象 j の出現回数 n_{ij}，$n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$，$n_{i\cdot} = \sum_{j=1}^{r} n_{ij}$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(\left|n_{ij} - n_{i\cdot}n_{\cdot j}/n\right| - 1/2\right)^2}{n} \sim \chi^2_{(r-1)(s-1)} \text{ 分布}$$

特に、$r = k = 2$ のとき、$\chi^2 = \dfrac{n\left(\left|n_{11}n_{22} - n_{12}n_{21}\right| - n/2\right)^2}{n_{\cdot 1}n_{\cdot 2}n_{1\cdot}n_{2\cdot}} \sim \chi^2_1$ 分布

McNemar 検定

群・対照群の要因の有無別数（有有 a，有無 b，無有 c，無無 d）

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \sim \chi^2_1 \text{ 分布}$$

適合度検定について、図 3 に実行画面を示す。

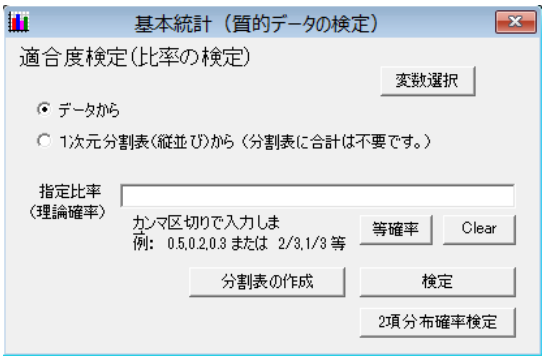


図 3 適合度検定画面

一般に、質的指標の検定には 2 種類の検定用データが考えられる。1 つは調査票等から直接入力されたデータで、それを元に分割表の作成や検定が行われる。また既に分割表を作成している場合には、その分割表を利用して検定を実施することも考えられる。実際の調査等では前者の形式が多くなるであろうが、講義用としては後者の場合も必要である。それゆえ、このプログラムでは質的指標の検定の際、どちらかのデータ形式を選択するようになっている。前者のデータの場合、分割表だけを作る場合もあると考えられるので、これらの検定メニューからも分割表が作れるようになっている。

実測値と比較する理論確率については、カンマ区切りで入力する。例えば、0.5, 0.3, 0.2 のような小数表示と 1/3, 1/3, 1/3 のような分数表示が可能である。メニューには注意書きを多く加え、分かり易さを高めている。等確率の場合、「等確率」ボタンをクリックすると、簡単に設定できる。

適合度検定の分析結果の例を図 4 に示す。

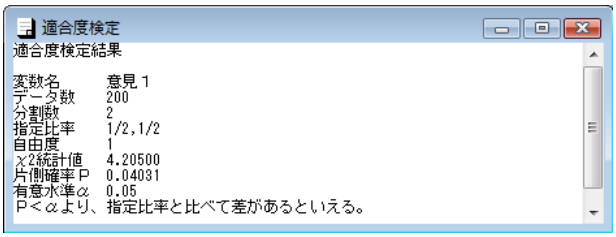


図 4 適合度検定結果

2 次元分割表の比率の検定を行うχ²検定の実行画面を図 5 に示す。

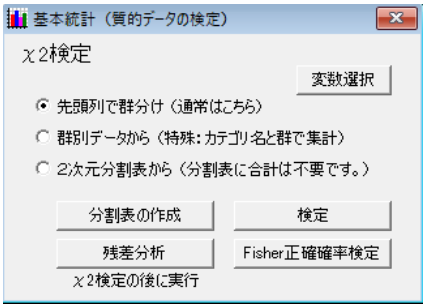


図 5 χ²検定画面

通常のデータの場合は「先頭列で群分け」を使い、分割表から求める場合は「２次元分割表から」を用いる。通常はこの２つで「検定」ボタンをクリックすれば事足りる。「群別データから」は、変数間のデータの比率の比較に用いる。変数１と変数２で、1 と 2 のデータがある場合、「先頭列で群分け」の集計結果は表 1 のようになり、「群別データから」の集計結果は表 2 のようになる。通常は表 1 のような集計をする。

表 1 「先頭列で群分け」の集計

	変数２が１	変数２が２
変数１が１	*a*	*B*
変数１が２	*c*	*D*

表 2 「群別データから」の集計

	1	2
変数１	*a+b*	*c+d*
変数２	*a+c*	*b+d*

χ²検定結果の画面を図 6 に示す。

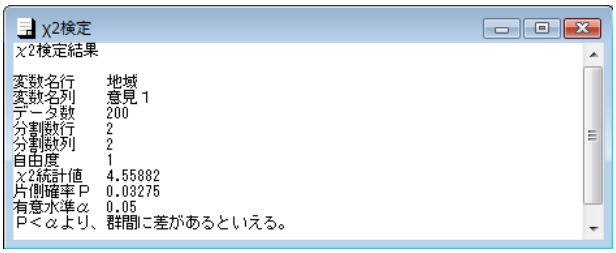


図 6 χ^2 検定結果

χ^2 検定は基本的に分割表の 1 つのマスが 10 以上の時に利用するのが望ましい。しかし、データ数が少ない場合で、２×２分割表の場合に限り、「Fisher 正確確率検定」が利用できる。その分析結果を図 7 に示す（データは上のものと異なる）。

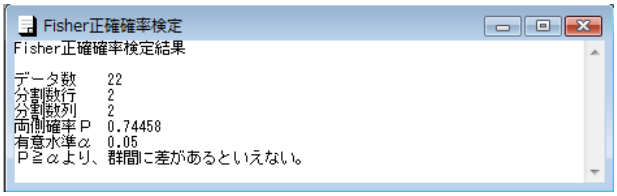


図 7 Fisher 正確確率検定結果

結果表示には検定結果の数値表示の他に、初心者の学習用に、例えば「標本値と理論値とを比べて差があるといえない。」のような検定結果を言葉にした表現や、標本数に関する利用上の注意等を加えている。

残差分析は χ^2 検定後に行う多重比較の一種である。ここでは、標準的な Haberman の残差分析を用いている。これはセル i, j に対して以下の基準化残差 e_{ij} の以下の性質を利用している。

$$e_{ij} = \frac{n_{ij} - n_{i\cdot}n_{\cdot j}/n}{\sqrt{(n_{i\cdot}n_{\cdot j}/n)(1-n_{i\cdot}/n)(1-n_{\cdot j}/n)}} \sim N(0,1)$$

2 項分布確率とフィッシャーの正確確率検定について

適合度検定は多項分布の近似を使った理論であるが、2 項分布に関しては正確な確率を求められるようにしておくことは意味がある。例えば、納品された商品の故障については、故障率が小さい場合は、たくさん発生することはない。これに対して適合度検定は、ある程度の（少なくとも 10 以上）故障例を必要とし、それ以下だと確率値に誤差が生じる。そのため、2 項分布による正確な確率値の計算は、品質管理などにおいて有効である。また、2×2 分割表における Fisher の正確確率検定も、少数の例数を扱う場合に重要である。

我々は、これらの確率計算を見直し、適合度検定と χ^2 検定のプログラムの中に組み込んだ。その際、これらの中に含まれる階乗の計算をスターリングの公式を用いて対数で実行し、大きな例数にも対応できるようにした。これによって、正確な確率と近似である χ^2 検定確率との比較もできるようになった。

ここではまず、2 項分布を用いて、適合度検定と同じ確率を計算することを考える。理論確率を p 、データ数を n 、事象の出現数を x とするとき、2 項分布では事象の出現確率は以下で与えられる。

$$P(x) = {}_nC_x p^x (1-p)^{n-x}$$

今事象の出現数が $\hat{x}$ であった場合、適合度検定に相当する確率 $Q(\hat{x})$ は以下のように求められる。

$$Q(\hat{x}) = \sum_{P(x)\le P(\hat{x})} {}_nC_x p^x (1-p)^{n-x} \quad \text{ここに、} P(\hat{x}) = {}_nC_{\hat{x}} p^{\hat{x}} (1-p)^{n-\hat{x}}$$

この領域は x が少ない場合と多い場合に分かれ、適合度検定に相当する検定確率は両側の確率を足したものになる。傾向がはっきりしている場合はどちらか一方になり、より偏りが大きい側の片側検定となる。

フィッシャーの正確確率検定は表 1 の分割表を基にする。

表 1　2×2 分割表

	列群 1	列群 2	合計
行群 1	x	$r_1 - x$	r_1
行群 2	$c_1 - x$	$x - r_1 + c_2$ $(= x - c_1 + r_2)$	r_2
合計	c_1	c_2	n

合計を固定して考えると、その度数の自由度は 1 になる。その 1 つの度数を x とすると、 x は以下の範囲で与えられる。

$$a \le x \le b, \qquad a = \max\{0, r_1 - c_2, c_1 - r_2\}, \quad b = \min\{r_1, c_1\}$$

この分割表を用いると、実現確率 $P(x)$ は超幾何分布の確率として以下のように与えられる。

$$P(x) = \frac{x!(r_1 - x)!(c_1 - x)!(x - r_1 + c_2)!}{n! r_1! r_2! c_1! c_2!}$$

観測された度数を $\hat{x}$ 、その場合の実現確率を $P(\hat{x})$ として、 χ^2 検定で与えられる検定確率 $Q(\hat{x})$ は上で定義した a, b を用いて以下のようになる。

$$Q(\hat{x}) = \sum_{P(x)\le P(\hat{x})} \frac{x!(r_1-x)!(c_1-x)!(x-r_1+c_2)!}{n!r_1!r_2!c_1!c_2!}$$

この領域も適合度検定のときと同様に、x が少ない場合と多い場合に分かれ、χ^2 検定に相当する検定確率は両側の確率を足したものになる。傾向がはっきりしている場合はどちらか一方になり、偏りの大きい側の片側検定となる。

確率の計算には、階乗が多く含まれているため、度数が大きくなると非常に大きな数の計算になり、場合によっては計算機の演算範囲を超えることもある。そのため、確率計算は一度対数を取って行い、計算結果である確率を再度元に戻す。

超幾何分布の式では、まず以下を計算する。

$$\begin{aligned}\log P(x) = {} & \log x! + \log(r_1-x)! + \log(c_1-x)! + \log(x-r_1+c_2)! \\ & - \log n! - \log r_1! - \log r_2! - \log c_1! - \log c_2!\end{aligned}$$

各項の対数内の数値が大きい場合、計算には以下の Starling の公式を用いる。

$$\log n! \cong n\log n - n + \frac{1}{2}\log(2\pi n)$$

計算した後、

$$P(x) = \exp(\log P(x))$$

で元に戻しておく。

５．量的データの検定

5.1 概要

量的指標の場合には図 1.1 の分類法と検定手法を用いる。特に、ノンパラメトリック検定についての他の分析手法や、適用限界についてのさらに細かい分類は今後の課題とする。

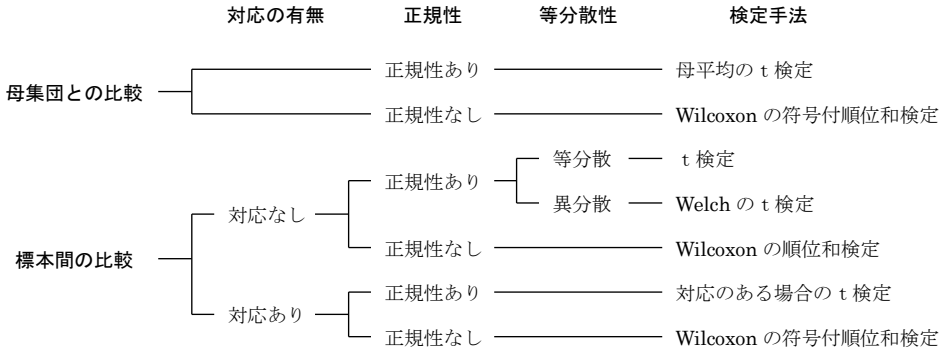


図 1.1 量的指標に関する検定の分類

質的指標と同様に、量的指標に関しても検定の位置付けを明確にするために、図 7 の様式を持った検定メニューが用意されている。その実行画面は図 1.2 で与えられる。

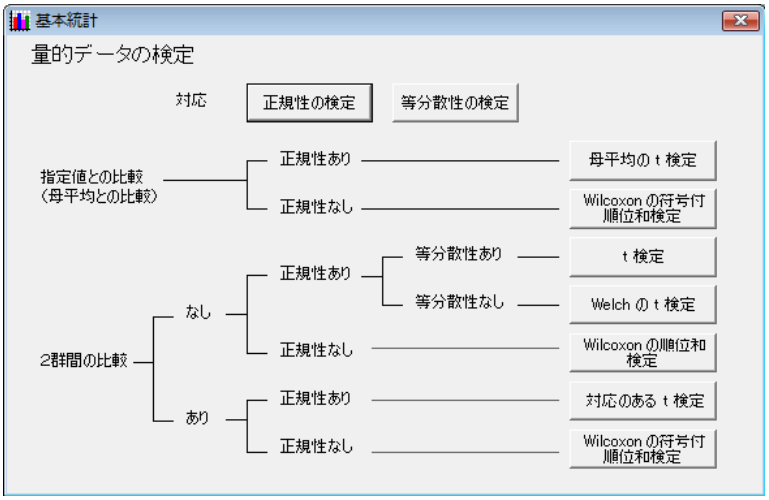


図 1.2 量的指標の検定画面

このメニューでは、右端の検定手法だけでなく、分類項目である正規性の検定や等分散性の検定も選択できるようになっている。

ここでは検定手法を母集団との比較と標本間の比較とに分け、標本間の比較については、それらの間の対応の有無によってさらに分類する。

量的指標の検定の基本性質は、パラメトリック検定とノンパラメトリック検定を分ける分布の正規性であるが、これらの見極めのために正規性の検定が必要である。そのために、ここでは目視的方法と数値的方法の 2 通りを用意する。

目視的方法としては、データ数が多い場合に使われる、度数分布表やヒストグラムから正規性を見る方法、またデータ数が少ない場合に利用される、正規確率紙による方法が用意されている。グラフは正規確率紙へのプロットに準じて、データの個数を n、あるデータの順位を i としてその累積確率を $i/(n+1)$で与え、データの数値と、この累積確率から得られる標準正規分布の検定値とで分布図を描く。これに回帰直線を加え、直線状への並びを見易くする。

正規性の数値的な検定方法としては Kolmogorov-Smirnov 検定と Shapiro-Wirk 検定に近い近似的検定法があるが、後者を使うことが多い（量的データ集計の部分を参照）。

5.2 指定値との比較

指定値との比較に関して、その手法を以下にまとめる。

母平均の t 検定

標本数 n，標本平均 $\bar{x}$，不偏分散 u^2，母平均 μ

$$t = \frac{\sqrt{n}|\bar{x} - \mu|}{u} \sim t_{n-1} \text{分布}$$

Wilcoxon の符号付順位和検定

データ x_i，中間値 μ，$z_i = x_i - \mu$

$|z_i|$ の昇順に 0 を除いて順位 r_i を付け、z_i の正負で 2 群に分類

各群の順位和 R_r，R_s の中で小さい方を選択 $R = \min(R_r, R_s)$

標本数が少ないとき（$z_i \neq 0$ の例数 < 10）

数表の利用

標本数が多いとき（$z_i \neq 0$ の例数 ≥ 10）

$$z = \frac{R - n(n+1)/4}{n(n+1)(2n+1)/24} \sim N(0,1) \text{分布}$$

非正規性の場合の検定は分布の対称性を仮定して、Wilcoxon の符号付順位和検定を採用した。またこの検定において、同順位の場合は順位平均を用いるが、同順位が多く含まれる場合の補正は今後の課題とする。

データに正規性があり、指定値と比較する場合の検定手法、母平均の t 検定について、その分析画面を図 2.1 に示す。

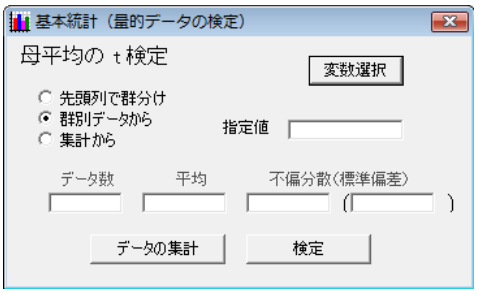


図 2.1 母平均の t 検定画面

指定値のところに比較する値を入れて、「検定」ボタンをクリックする。「集計から」のときは、データ数や平均、不偏分散（または標準偏差どちらか）に値を入力しておく。図 2.2 に母平均の t 検定の検定結果画面の例を表示する。

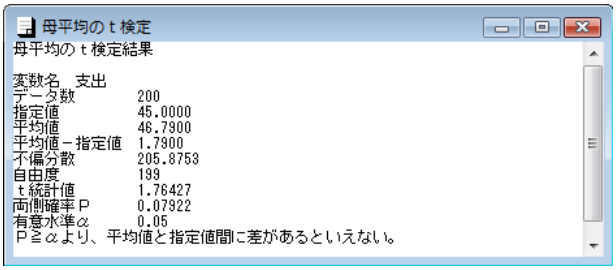


図 2.2 母平均の t 検定の検定結果

データに正規性がない場合は、Wilcoxon の符号付き順位和検定となる。同じ名前の分析が、対応のあるデータの場合にもあるので、間違わないように注意する必要がある。その分析画面を図 2.3 に示す。

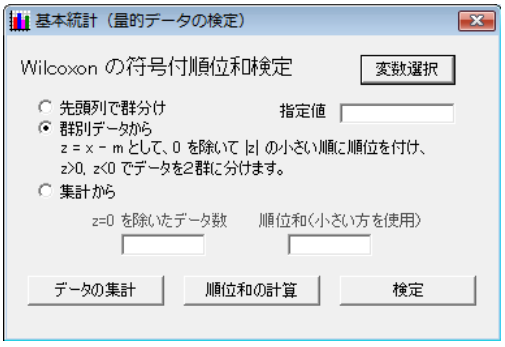


図 2.3 Wilcoxon の符号付き順位和検定画面

ここでも比較する値を「指定値」に入れて「検定」ボタンをクリックする。出力結果を図 2.4 に示す。

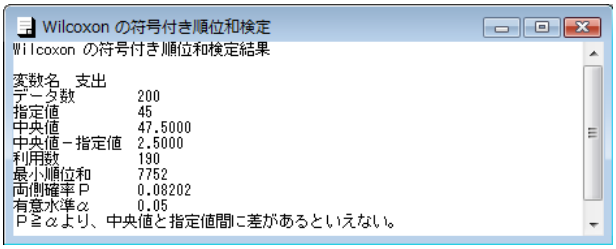


図 2.4 Wilcoxon の符号付き順位和検定結果

5.3 ２群間の比較（対応のない場合）

２群間の比較の場合は、対応のある場合とない場合とに分類する。対応とは、２つの群に同じ対象（同じように設定された対象の場合もある）がいるかどうかで判断する。例えば、入試で国語と英語を比較する場合、同じ人が両方受験しているので、対応があるとする。また、男女別に比較する場合は、同じ人が両方の群にはいないので、対応はないとする。

対応がない場合、正規性の検定を行い、正規分布ならさらに等分散性を検定する必要がある。これらの分類による具体的な検定手法は以下にまとめる。正規性の認められない場合は Wilcoxon の順位和検定を用いる。

F 検定（等分散性の検定）

標本数 n_1, n_2，不偏分散 u_1^2, u_2^2 （$u_1^2 > u_2^2$）

$$F = \frac{u_1^2}{u_2^2} \sim F_{n_1-1,\, n_2-1} \text{ 分布}$$

（student の）t 検定

標本数 n_1, n_2，標本平均 $\bar{x}_1, \bar{x}_2$，不偏分散 u_1^2, u_2^2

$$t = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\dfrac{(n_1-1)u_1^2 + (n_2-1)u_2^2}{n_1 + n_2 - 2}}} \sim t_{n_1+n_2-2} \text{ 分布}$$

Welch の t 検定

標本数 n_1, n_2，標本平均 $\bar{x}_1, \bar{x}_2$，不偏分散 u_1^2, u_2^2

自由度 $d = \dfrac{1}{\dfrac{c^2}{n_1 - 1} + \dfrac{(1-c)^2}{n_2 - 1}}$，$c = \dfrac{u_1^2/n_1}{u_1^2/n_1 + u_2^2/n_2}$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{u_1^2/n_1 + u_2^2/n_2}} \sim t_d \text{ 分布}$$

Wilcoxon の順位和検定

標本数 n_1, n_2 （$n_1 \le n_2$），標本 x_i^1, x_j^2

標本の昇順に順位 r_i を付け、標本数の少ない群の順位和を求める。

$$W = \sum_{i=1}^{n_1} r_i$$

標本数が少ない場合（$n_2 \le 20$）

文献 5), 6) 等の数表を利用

標本数が多い場合（$n_2 > 20$）

$$Z = \frac{W - \frac{1}{2} n_1 (n_1 + n_2 + 1)}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \sim N(0,1) \text{ 分布}$$

対応のない２標本の比較の場合、データの読み込み方法は、先頭列で群分け、群別データから、集計からの３種類用意する。正規性が認められた場合の等分散性の検定画面を図 3.1 に示す。

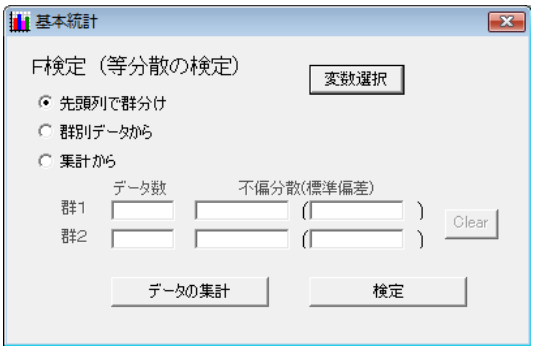


図 3.1 等分散性の検定

図 3.2 に等分散性の検定結果の例を示す。

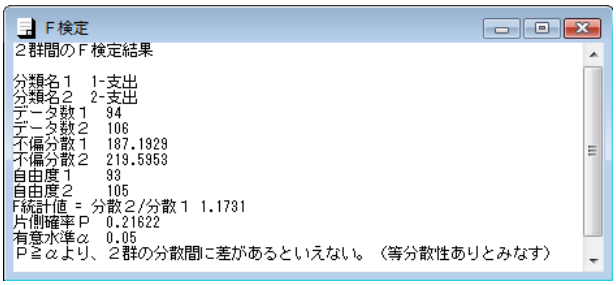


図 3.2 等分散性の検定結果

正規性と等分散性が認められた場合の t 検定の検定画面を図 3.3 に示す。

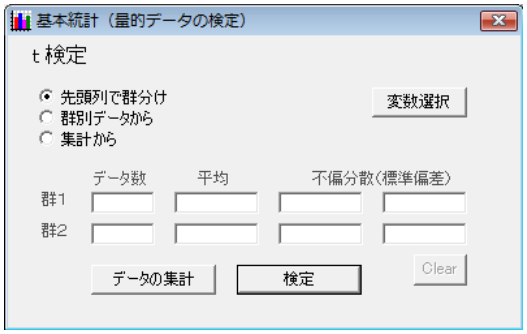


図 3.3　t 検定画面

t 検定の出力結果を図 3.4 に示す。

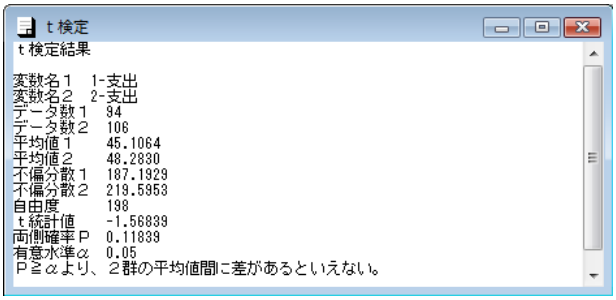


図 3.4　t 検定結果

データに正規性があり、等分散性がない場合の Welch の t 検定の画面を図 3.5 に示す。

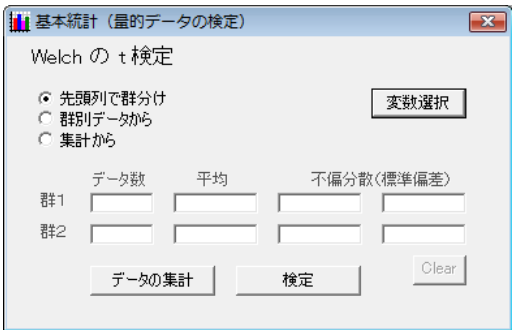


図 3.5 Welch の ｔ検定結果

Welch の ｔ検定の出力結果を図 3.6 に示す。

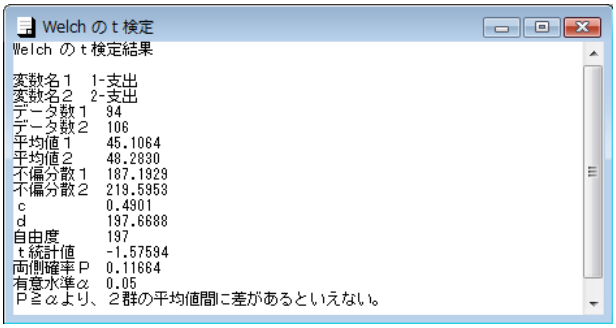


図 3.6 Welch の ｔ検定結果

データに正規性がない場合、Wilcoxon の順位和検定を利用するが、その画面を図 3.7 に示す。

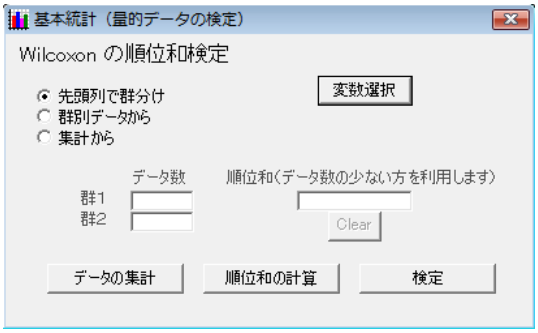


図 3.7 Wilcoxon 順位和検定画面

Wilcoxon の順位和検定の実行結果を図 3.8 に示す。

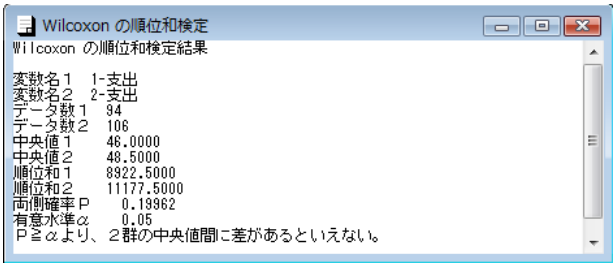


図 3.8 Wilcoxon 順位和検定結果

5.4 ２群間の比較（対応がある場合）

対応のある場合の検定手法を以下にまとめる。

対応がある場合の t 検定

例数 n，標本差 z_i，平均 $\bar{z}$，不偏分散 u_z^2

$$t = \frac{\sqrt{n}\,|\,\bar{z}\,|}{u_z} \sim t_{n-1} \text{ 分布}$$

Wilcoxon の符号付き順位和検定

標本差 z_i をもとにする。

データ x_i, y_i，中間値 $z_i = x_i - y_i$

$|\,z_i\,|$ の昇順に 0 を除いて順位 r_i を付け、z_i の正負で 2 群に分類

各群の順位和 R_r，R_s の中で小さい方を選択 $R = \min(R_r, R_s)$

標本数が少ないとき（$z_i \neq 0$ の例数 < 10）

　　数表の利用

標本数が多いとき（$z_i \neq 0$ の例数 ≥ 10）

$$z = \frac{R - n(n+1)/4}{n(n+1)(2n+1)/24} \sim N(0,1) \text{ 分布}$$

対応のあるデータの正規性は、対応する 2 つのデータの差を取ったものを使って判定する。そのため、図 4.1 の正規性の検定画面で、「対応のあるデータから」ラジオボタンを選択する。

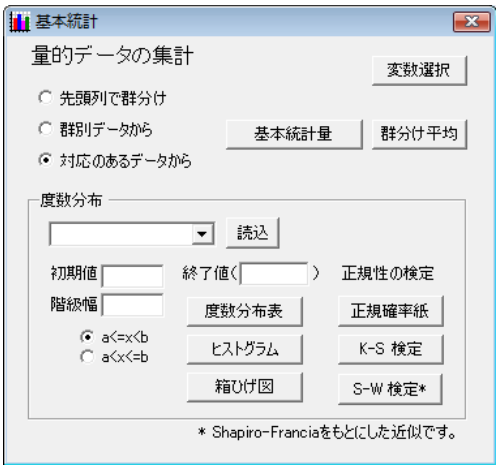


図 4.1 正規性の検定

対応のある場合の正規性の検定結果は図 4.2 のように示される。

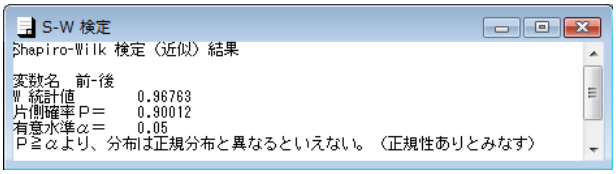


図 4.2 対応のある場合の正規性の検定結果

正規性の検定で正規性が認められた場合の、対応のある t 検定の検定画面を図 4.3 に示す。

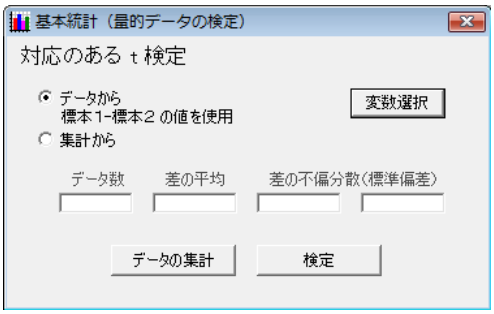


図 4.3 対応のある t 検定画面

対応のある t 検定の検定結果を図 4.4 に示す。

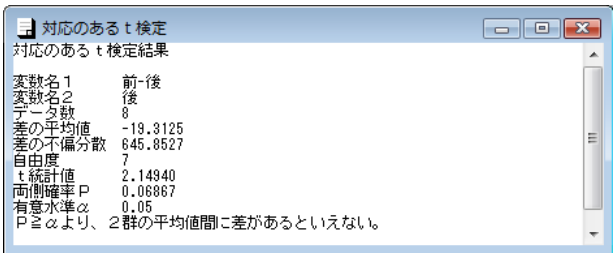


図 4.4 対応のあるｔ検定結果

正規性が認められなかった場合の、Wilcoxon 符号付き順位和検定の検定画面を図 4.5 に示す。

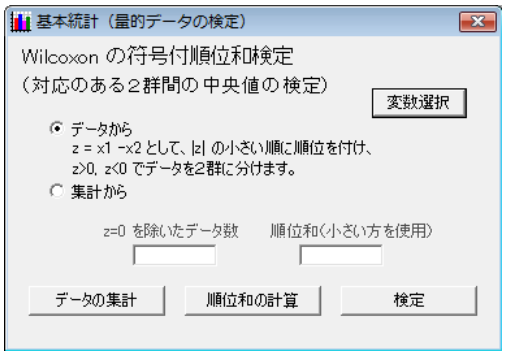


図 4.5 Wilcoxon 符号付き順位和検定画面

分析実行画面を図 4.6 に示す。

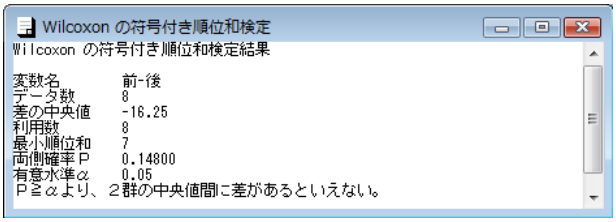


図 4.6 Wilcoxon 符号付き順位和検定結果

６．相関係数と回帰分析

6.1　相関係数と回帰係数の検定

相関係数については、正規性が認められる場合の Pearson の相関係数及び、正規性が認められない場合の Spearman の順位相関係数について求めており、無相関か否かの検定を行っている。また、回帰分析については、回帰式と重相関係数、及び寄与率について求め、回帰係数の有効性について、残差の正規性を仮定して検定を行っている。また、結果表示には回帰直線も含めた分布図も利用する。具体的な公式については以下にまとめる。

Pearson の相関係数

標本数 n，相関係数 r

$$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2} \text{ 分布}$$

Spearman の相関係数の検定

標本数 n，群ごとの順位による順位相関係数 r_s

$$t = \frac{|r_s|\sqrt{n-2}}{\sqrt{1-r_s^2}} \sim t_{n-2} \text{ 分布}$$

回帰分析

標本平均 $\bar{x}, \bar{y}$，不偏分散 u_x^2, u_y^2，相関係数 r

$$y = ax + b, \quad a = r\frac{u_y}{u_x}, \quad b = \bar{y} - r\frac{u_y}{u_x}\bar{x}$$

重相関係数 R　　実測値 y_i と予測値の相関係数

寄与率　　R^2

説明変数は１つだけに限り、複数の場合は重回帰分析として多変量解析に含まれている。

回帰分析の検定については、表中では表しにくいので、ここで簡単にふれておく。目的変数を y、説明変数を x とし、これらの間に、関係式 $y = ax + b + \varepsilon$ があると仮定する。ここに予測式は $Y = ax + b$ であり、残差は $\varepsilon \sim N(0, \sigma^2)$ 分布とする。

回帰係数の有効性の検定は、データ数 n，残差変動 $EV = \sum_{i=1}^{n}(y_i - Y_i)^2$，説明変数の不偏分散 u_x^2 として、以下の関係を用いる。

$$t_a = \frac{a}{\sqrt{\frac{EV}{n-2}\Big/(n-1)u_x^2}} \sim t_{n-2}\text{分布} \qquad t_b = \frac{b}{\sqrt{\frac{EV}{n-2}\left(\frac{1}{n}+\frac{\bar{x}^2}{(n-1)u_x^2}\right)}} \sim t_{n-2}\text{分布}$$

単回帰分析の場合に前者の検定は、残差変動に対する回帰変動の有効性を検定する、回帰式の有効性の検定と一致する。

メニュー［分析－基本統計－相関と回帰分析］を選択すると、図 1 の分析画面が表示される。

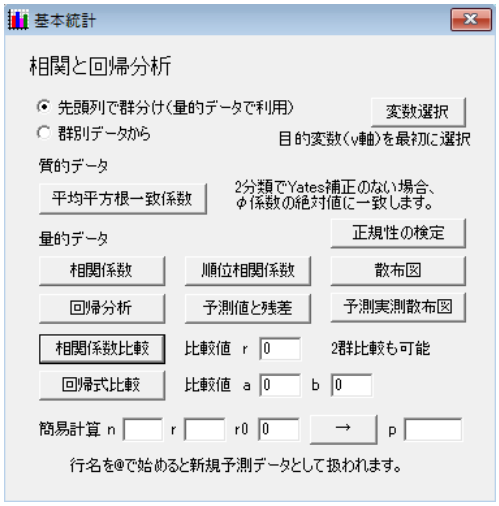


図 1 相関と回帰分析画面

2 つの変数を選択して、「相関係数」ボタンをクリックすると、図 2 のような、相関係数とその検定結果（相関 0 と比較）が表示される。相関係数は、2 変数が多変量正規分布する場合に用いられる。

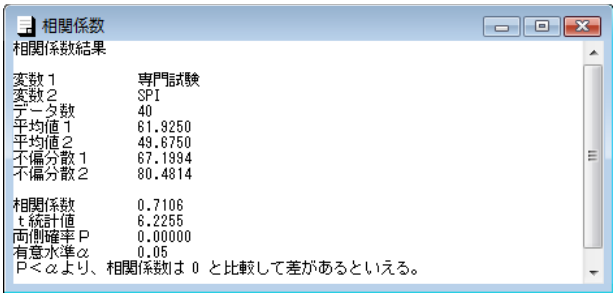


図 2 相関係数結果

2 変数のトレンドの相関を見る場合は、Wilcoxon の順位相関係数を利用する。「順位相関係数」ボ

タンをクリックした場合の結果を、図 3 に示す。

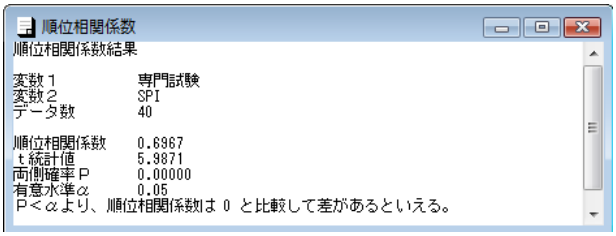


図 3 Wilcoxon の順位相関係数結果

3 つ以上の変数を選択して、「相関係数」ボタンをクリックすると、図 4 のように、表形式で相関係数とその検定値が表示される。「順位相関係数」でも同様である。

	判定	専門試験	SPI
▶ 判定	1.0000	-0.4356	-0.7979
専門試験	-0.4356	1.0000	0.7106
SPI	-0.7979	0.7106	1.0000
検定確率			
判定		0.0050	0.0000
専門試験	0.0050		0.0000
SPI	0.0000	0.0000	

図 4　3 変数以上の相関係数表示画面

図 1 のメニューで「散布図」ボタンをクリックすると、図 5 のような散布図が表示される。

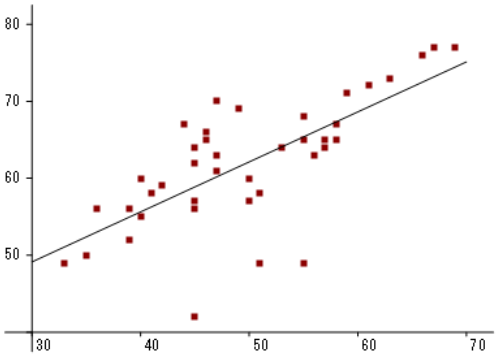


図 5 散布図

グラフの「設定」メニューで、データラベルを付けたり、回帰直線を消したりすることができる。「先頭列で群分け」ラジオボタンを選び、最初に群分け変数を選んで、散布図を描くと図 6 のような多重散布図となる。

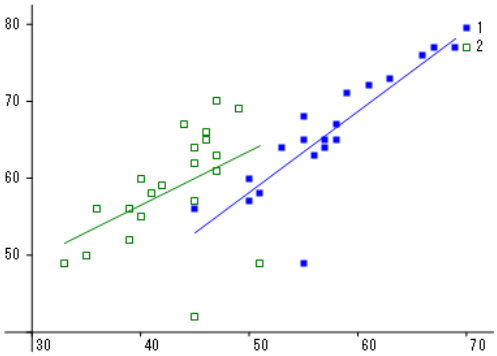


図 6 多重散布図

この群分け機能は相関係数や次に述べる回帰分析でも有効である。

回帰分析の計算結果と回帰係数の検定結果は、「回帰分析」ボタンをクリックすると図 7 のように表示される。

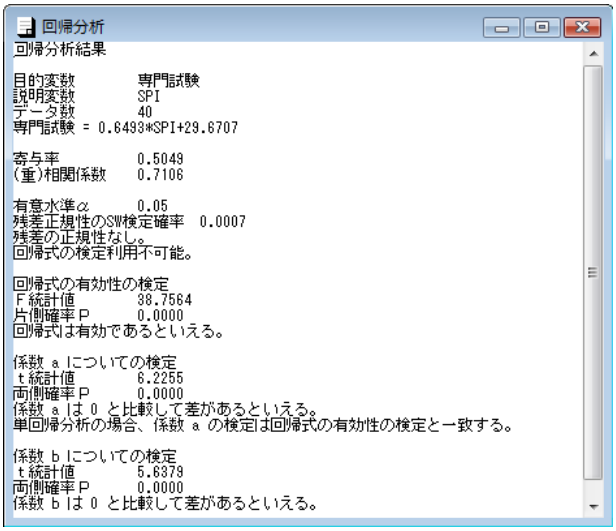


図 7 回帰分析結果

回帰分析による予測値は「予測値と残差」ボタンをクリックすると図 8 のように表示される。

予測値と残差

	実測値	予測値	残差
1	55	55.643	-0.643
2	57	58.889	-1.889
3	63	60.188	2.812
4	71	67.980	3.020
5	72	69.278	2.722
6	64	64.084	-0.084
7	59	56.942	2.058
8	42	58.889	-16.889
9	76	72.525	3.475
10	60	62.136	-2.136

図 8 予測値と残差

予測値と実測値でグラフを描くと図 9 のようになる。実測値が縦軸、予測値が横軸である。

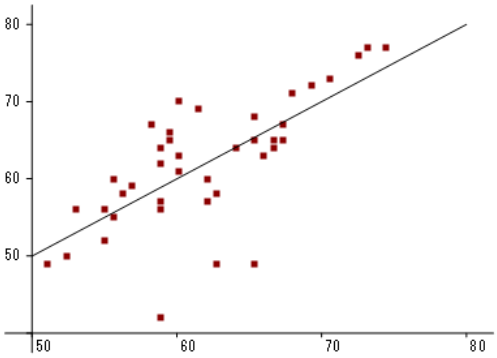


図 9 予測値と実測値の散布図

6.2　2 群間の相関係数と回帰係数の比較

これまで College Analysis の相関と回帰分析では、相関係数と回帰係数は 0 との比較の場合だけを考えてきた。しかし、相関係数や回帰式が同じかどうかを調べることも多くなると考え、検定を加えることにした。

相関係数と母相関係数の比較では、データ数を n 、標本相関係数を r 、母相関係数を ρ として、以下の関係を利用する。

$$T = \frac{\frac{1}{2}\log\frac{1+r}{1-r} - \frac{1}{2}\log\frac{1+\rho}{1-\rho}}{1/\sqrt{n-3}} \sim N(0,1)$$

２群の相関係数の比較では、データ数を n_1, n_2 、標本相関係数を r_1, r_2 として、以下の関係を利用する。

$$T=\frac{\frac{1}{2}\log\frac{1+r_1}{1-r_1}-\frac{1}{2}\log\frac{1+r_2}{1-r_2}}{\sqrt{1/(n_1-3)+1/(n_2-3)}}\sim N(0,1)$$

回帰係数と母回帰係数の比較では、データ数を n 、標本回帰式を $y=ax+b$、母回帰式を $y=\alpha x+\beta$ として、以下の関係を利用する。

勾配係数の比較　$T_a=(a-\alpha)\sqrt{SS_x/V_E}\sim t_{n-2}$

定数係数の比較　$T_b=\dfrac{b-\beta}{\sqrt{V_E\left(1/n+\bar{x}^2/SS_X\right)}}\sim t_{n-2}$

ここに

$$\bar{x}=\frac{1}{n}\sum_{\lambda=1}^{n}x_\lambda,\quad \bar{y}=\frac{1}{n}\sum_{\lambda=1}^{n}y_\lambda$$

$$SS_x=\sum_{\lambda=1}^{n}x_\lambda^2-n\bar{x}^2,\quad SS_y=\sum_{\lambda=1}^{n}y_\lambda^2-n\bar{y}^2,\quad SS_{xy}=\sum_{\lambda=1}^{n}x_\lambda y_\lambda-n\overline{xy}$$

$$V_E=\frac{1}{n-2}\left[SS_y-(SS_{xy})^2/SS_x\right]$$

2 群の回帰係数の比較では、データ数を n_1, n_2 、標本回帰式を $y=a_1x+b_1$，$y=a_2x+b_2$ として、まず、以下の関係を利用して勾配係数の比較を行う。

$$F_a=\left[(\Delta_2/\Delta_1)-1\right](n_1+n_2-4)\sim F_{1,\,n_1+n_2-4}$$

勾配係数が異なるとすると、回帰式はそのまま使われ、勾配係数が等しいとすると、以下の関係を利用して定数係数の比較を行う。

$$F_b=\left[(\Delta_3/\Delta_2)-1\right](n_1+n_2-3)\sim F_{1,\,n_1+n_2-3}$$

ここで、定数係数が異なるとすると $a=(SS_{xy1}+SS_{xy2})/(SS_{x1}+SS_{x2})$，$b_i=\bar{y}_i-a\bar{x}_i$ として、回帰式は以下を与える。

$$y=ax+b_1,\quad y=ax+b_2$$

定数係数が同じとすると $a=SS_{xy}/SS_x$，$b=\bar{y}-a\bar{x}$ として、回帰式は同一に以下で与える。

$$y=ax+b$$

ここに、$i=1,2$ として以下の関係を用いた。

$$\bar{x}_i=\frac{1}{n_i}\sum_{\lambda=1}^{n}x_{i\lambda},\quad \bar{y}_i=\frac{1}{n_i}\sum_{\lambda=1}^{n}y_{i\lambda}$$

$$SS_{xi}=\sum_{\lambda=1}^{n}x_{i\lambda}^2-n_i\bar{x}_i^2,\quad SS_{yi}=\sum_{\lambda=1}^{n}y_{i\lambda}^2-n_i\bar{y}_i^2,\quad SS_{xyi}=\sum_{\lambda=1}^{n}x_{i\lambda}y_{i\lambda}-n_i\bar{x}_i\bar{y}_i$$

$$\Delta_1=\left[SS_{y1}-(SS_{xy1})^2/SS_{x1}\right]+\left[SS_{y2}-(SS_{xy2})^2/SS_{x2}\right]$$

$$\Delta_2=SS_{y1}+SS_{y2}-\frac{(SS_{xy1}+SS_{xy2})^2}{SS_{x1}+SS_{x2}}$$

$$\Delta_3=SS_y-(SS_{xy})^2/SS_x$$

以下の図 10 のようなデータを用いて、図 1 の分析メニューで先頭列で群分けとして、「相関係数比較」ボタンをクリックすると図 11 のような分析結果が表示される。

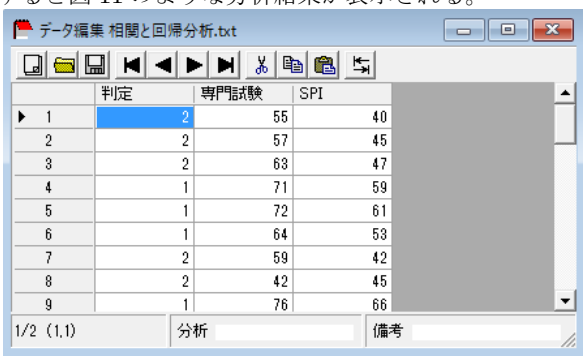


図 10　相関係数と回帰係数の比較データ

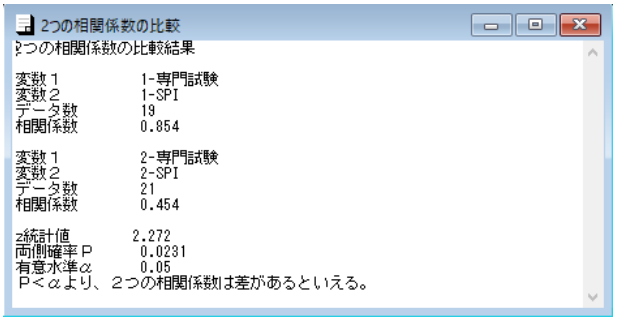


図 11　相関係数の比較分析結果

また、「回帰式比較」のボタンをクリックすると、図 12 のような結果が表示される。

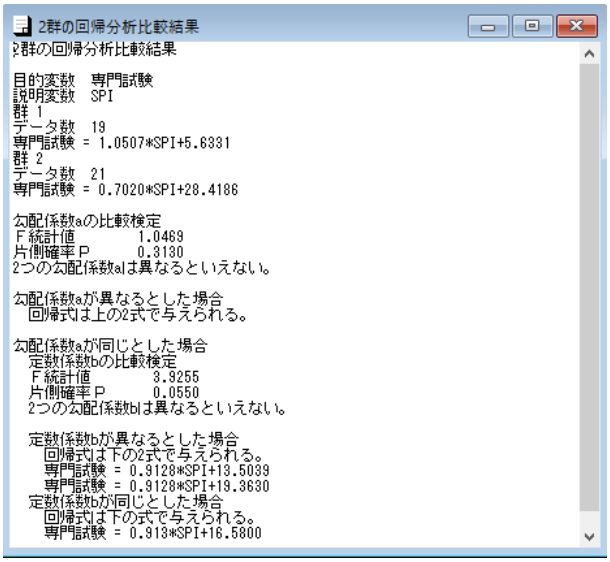


図 12　回帰式の比較結果

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.

7．トレンドの検定

7.1　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。まず、質的なデータに対する比率のトレンドの検定について説明する [2]。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。

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

しかし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' を用いる場合が多い。

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

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。

7.2 プログラムの利用法

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分析実行画面が表示される。

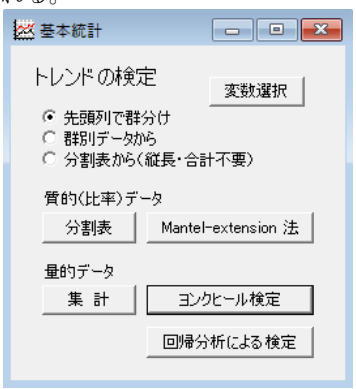


図 1　トレンドの検定分析実行画面

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。

図 2a のような分割表画面の質的データに対して、データ形式を「分割表から」として「Mantel-extension 法」ボタンをクリックすると図 2b のような結果表示画面が示される。

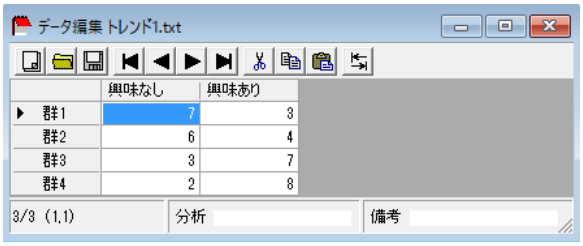


図 2a　分割表データ例

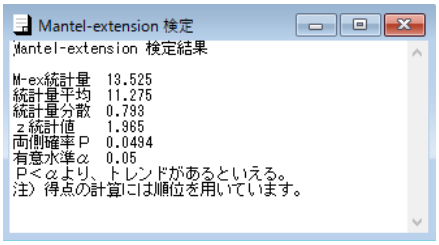


図 2b　Mantel-extension 検定結果

量的データについては、図 3a のようなデータに対して、データ形式を「先頭列で群分け」として「ヨンクヒール検定」ボタンをクリックすると、図 3b のような結果表示画面が得られる。また同じデータに対して、「回帰分析による検定」ボタンを押すと図 3c のような画面が示される。

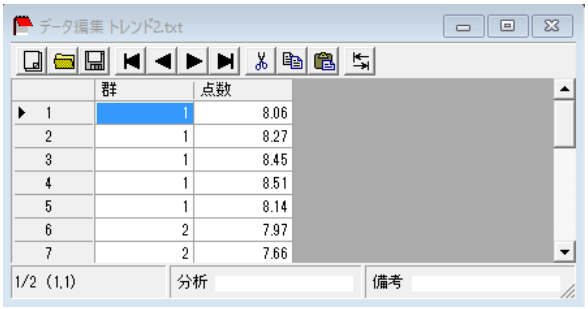


図 3a　トレンドの検定量的データ例

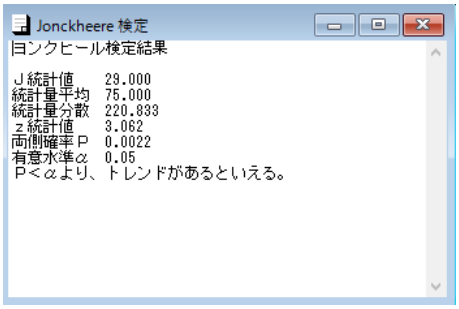


図 3b　ヨンクヒール検定結果

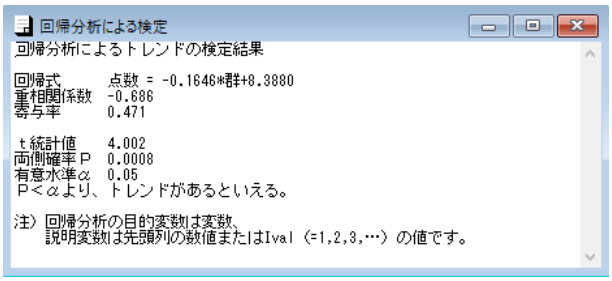


図 3c　回帰分析による検定結果

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.

８．標本数の決定

標本数の決定については、正規性が認められる場合に限定し、母比率の検定と母平均の検定のために必要なデータ数を求める。具体的な公式は以下にまとめる。

母比率の検定用

母比率 p，標本比率 $\hat{p}$

$$n = \frac{\chi_1^2(\alpha)\, p(1-p)}{(\hat{p}-p)^2}$$

母平均の検定用（両側）

母平均 μ，母分散 σ^2，標本平均 $\bar{x}$

$$n = \frac{Z(\alpha/2)^2 \sigma^2}{|\bar{x}-\mu|^2}$$

但し、母平均を求める検定に必要な標本数は、数が多いものとして近似的に標準正規分布の検定統計値を利用している。ここに、$\chi_1^2(\alpha)$ は自由度 1 の χ^2 分布の上側確率 α の検定統計値であり、$Z(\alpha/2)$ は標準正規分布の上側確率 $\alpha/2$ の検定統計値である。

質的指標で分割数が３以上の場合や２群間の差の検定及び、正規性を持たない場合等の標本数の決定については今後の課題とする。図 1 に標本数の決定の画面を示すが、入力には母集団の統計量と、データを収集した場合の予想値とを用いる。標本数の決定に関しては、予想値によるところが大きいので、多くの検定手法への対応は特に重要であるとは考えない。

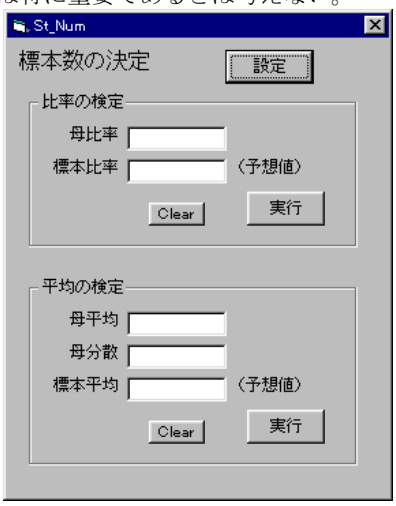


図 1 標本数の決定

９．区間推定

区間推定についても正規性が認められる場合に限定する。求める推定値は、母比率、母平均、母分散とした。具体的な手法については、以下にまとめる。

母比率の推定

標本数 n，標本比率 $\hat{p}$

$$\hat{p} \pm Z(\alpha/2)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

母平均の推定

標本数 n，標本平均 $\bar{x}$，不偏分散 u^2

$$\bar{x} \pm \frac{u}{\sqrt{n}} t_{n-1}(\alpha/2)$$

母分散の推定

標本数 n，不偏分散 u^2，母平均 σ^2

$$\frac{(n-1)u^2}{\chi^2_{n-1}(\alpha/2)} \le \sigma^2 \le \frac{(n-1)u^2}{\chi^2_{n-1}(1-\alpha/2)}$$

ここに、前節で説明した表式を除いて、$t_{n-1}(\alpha/2)$ は自由度 $n-1$ の t 分布の上側確率 $\alpha/2$ の検定統計値である。表式の簡単化のために、母比率と母平均については上限と下限を示すこととする。

入力は調査データからの入力と統計量からの入力と２種類持っておけばよい。

メニュー［分析－基本統計－区間推定－比率の推定］を選択すると、図１のような母比率の推定のための分析画面が表示される。

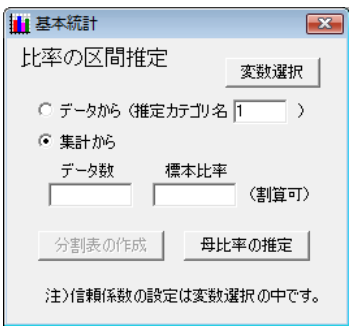


図１ 比率の推定画面

「集計から」の場合はデータ数と比率を入力して「母比率の推定」ボタンをクリックする。「データ

から」の場合は、変数を選択し、比率を推定するカテゴリの名前をテキストボックスに記入しておく。結果は図 2 のようになる。

図 2 母比率の推定結果

メニュー［分析－基本統計－区間推定－平均と分散の推定］を選択すると、図 3 のような平均と分散の推定のための分析画面が表示される。

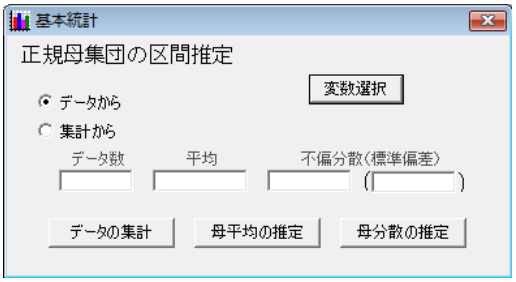


図 3 平均と分散の推定

「母平均の推定」ボタンをクリックした場合の結果を図 4 に示す。

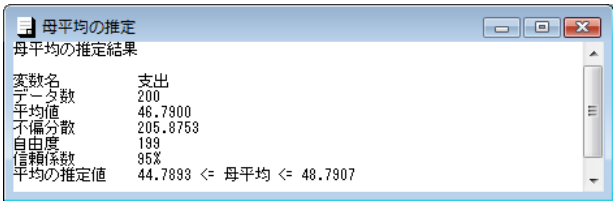


図 4 母平均の推定結果

「母分散の推定」ボタンをクリックした場合の結果を図 5 に示す。


母分散の推定
母分散の推定結果
変数名 支出
データ数 200
平均値 46.7900
不偏分散 205.8753
自由度 199
信頼係数 95%
分散の推定値 170.7319 <= 母分散 <= 253.1738
標準偏差の推定値 13.0664 <= 標準偏差 <= 15.9114

図 5 母分散の推定結果

１０．２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2 次元グラフ]を選択すると、図 1 のような分析画面が表示される。

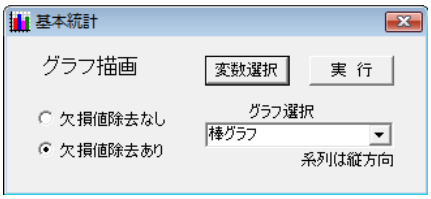


図 1 2 次元グラフ描画画面

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。

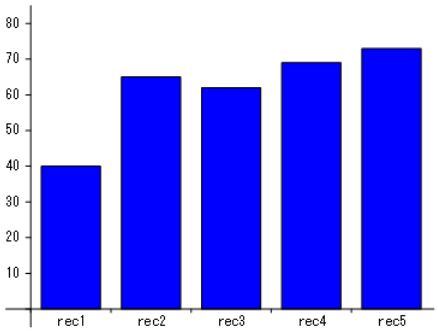


図 2a 棒グラフ（1 変数）

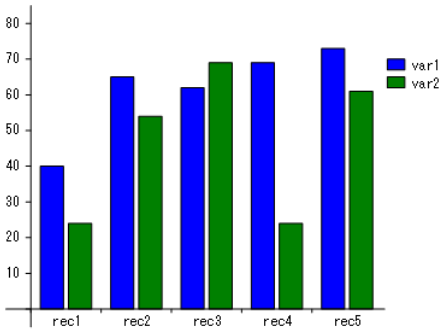


図 2b 棒グラフ（2 変数）

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。

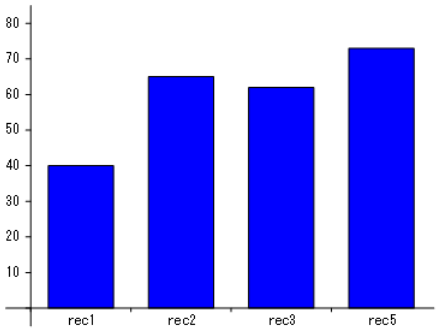


図 3a 棒グラフ（欠損値除去あり）

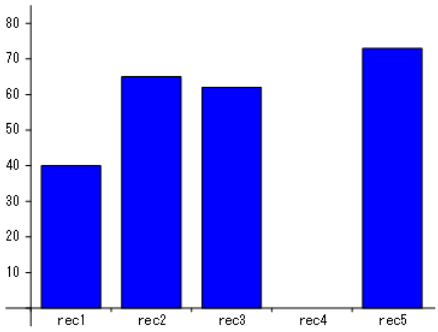


図 3b 棒グラフ（欠損値除去なし）

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。

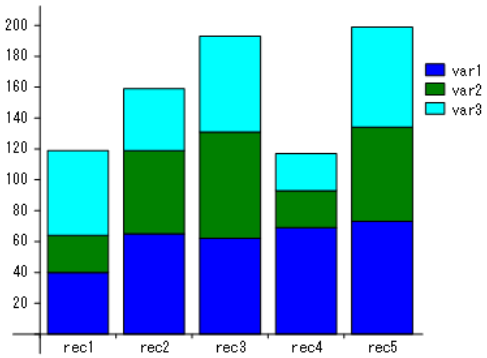


図 4 積重ね棒グラフ

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。

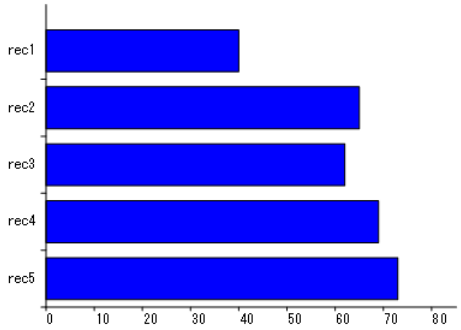


図 5a 横棒グラフ（1 変数）

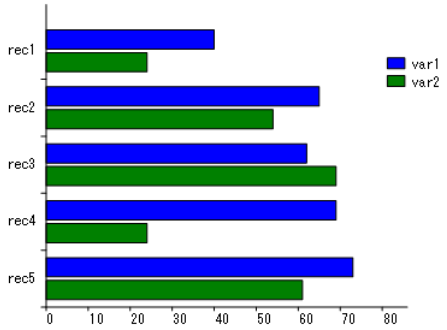


図 5b 横棒グラフ（2 変数）

変数を 3 つ選んだ積重ね横棒グラフの例を図 6 に描く。

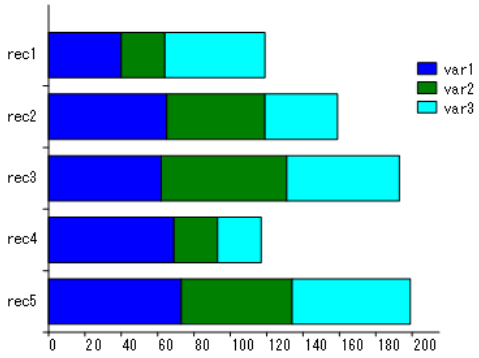


図 6 積重ね横棒グラフ

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。

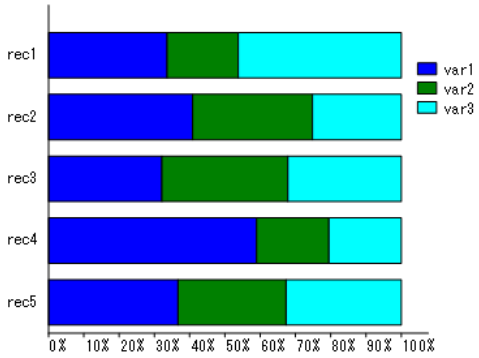


図 7 帯グラフ

立体棒グラフの例を図 8 に示す。

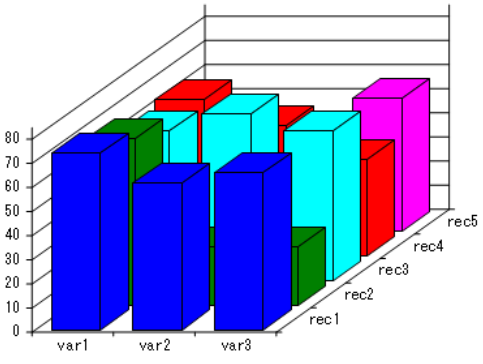


図 8 立体棒グラフ

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。

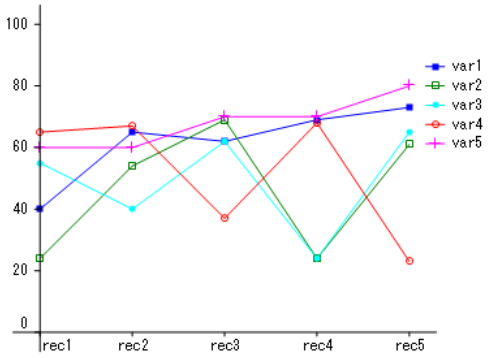


図 9 折れ線グラフ

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。

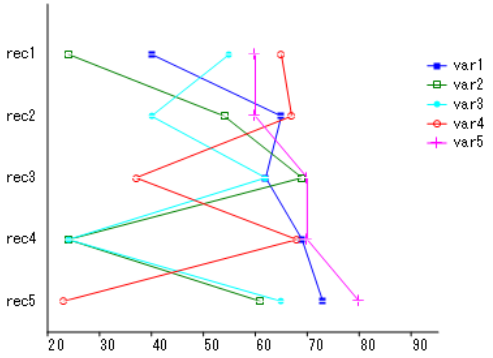


図 10 横折れ線グラフ

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。

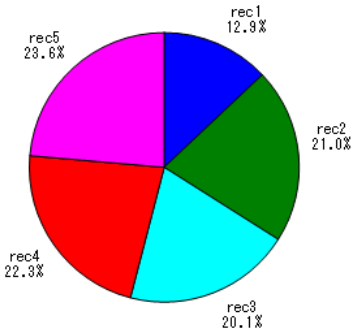


図 11 円グラフ

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。

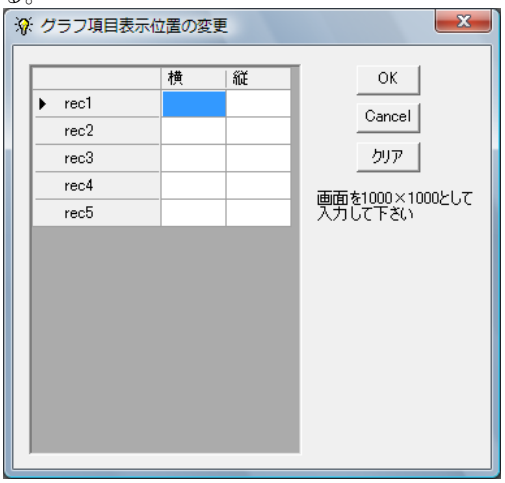


図 12 項目名位置変更

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。

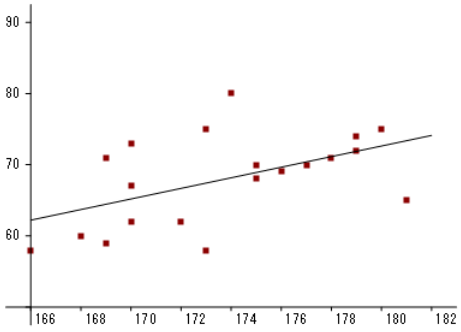


図 13a 散布図（回帰直線）

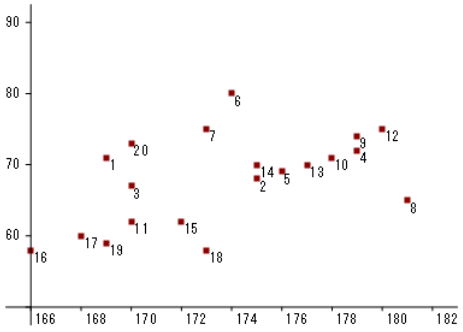


図 13b 散布図（データラベル）

変数を３つ選んだレーダーチャートの例を図 14 に示す。

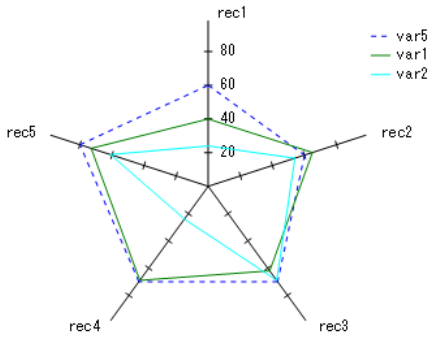


図 14 レーダーチャート

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。

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。

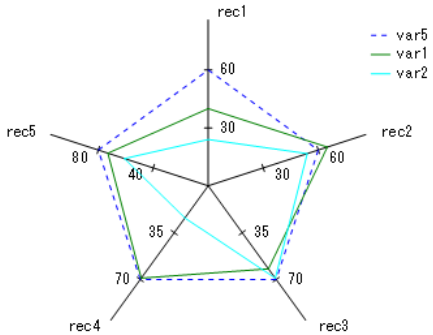


図 15 比較レーダーチャート

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。

１１．３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。

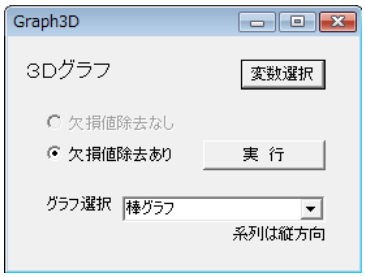


図１3D グラフ描画画面

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。

棒グラフの例を図２に示す。

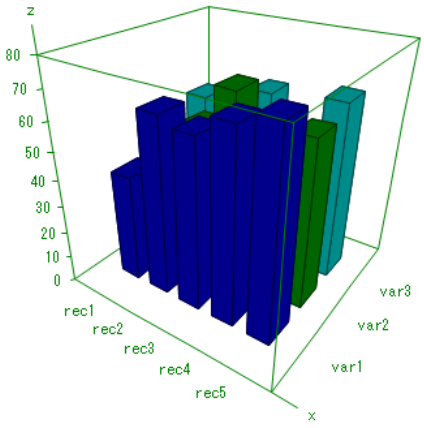


図２3D 棒グラフ

散布図の例を図３に示す。

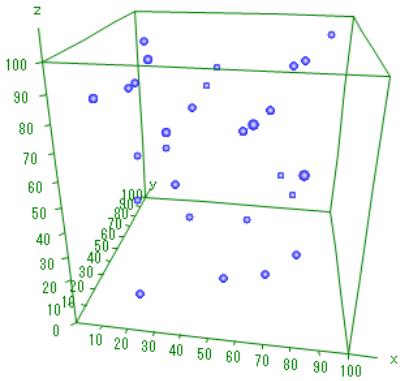


図 3 3D 散布図

１２．統計ユーティリティ

12.1 分布と確率

基本的な分布関数について、教育用に検定値から上側確率及び、上側確率から検定値を求める必要があり、簡単な計算メニューを加える。具体的な実行画面は、図 1.1 で与えられる。

図 1.1 検定値と確率画面

ここではパラメトリックな検定に利用される、標準正規分布、χ^2分布、F 分布、t 分布について、結果が求められる。値か確率かに数値を入力し、「→」か「←」ボタンをクリックして他方を求める。

12.2 密度関数グラフ

メニュー［基本統計－密度関数グラフ］を選択すると、標準正規分布、χ^2 分布、F 分布、 t 分布について、密度関数のグラフを描くことができる。図 2.1 にその描画画面を示す。

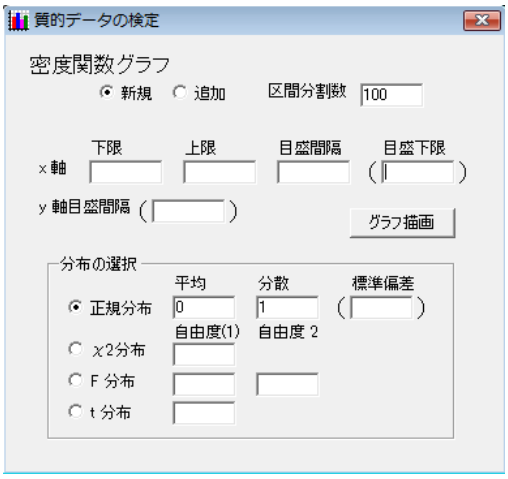


図 2.1 密度関数グラフ

x 軸の下限と上限、目盛間隔を入力し、分布を選択して、必要な場合は自由度を入力して、「グラフ描画」ボタンをクリックする。標準正規分布の出力画面を図 2.2 に示す。

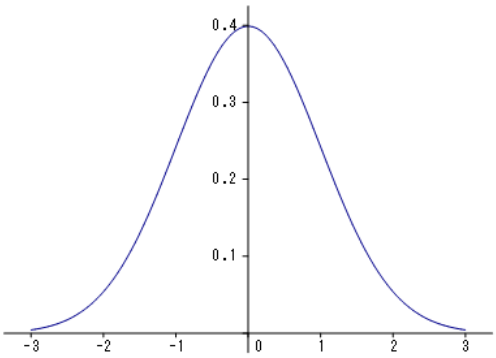


図 2.2 標準正規分布密度関数

χ^2 分布等では、自由度を変えていくつもグラフを表示したい場合がある。そのときは、始めに「新規」ラジオボタンでグラフを表示した後、「追加」ボタンで自由度を変えて描画して行く。図 2.3 に自由度を 1, 2, 3, 4 とした場合の χ^2 分布の密度関数を示す。

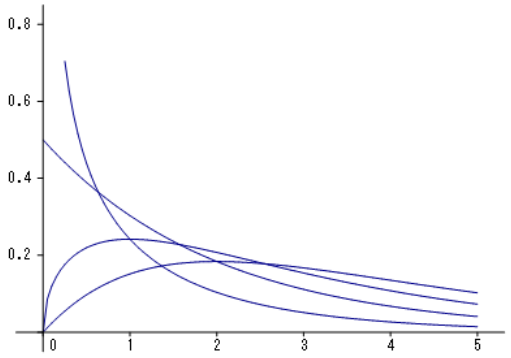


図 2.3 χ^2 分布密度関数（自由度 1, 2, 3, 4 ）

12.3 量から質変換

データ処理では量的データを区間を区切って、分類データのように使うことがある。例えば身長 170cm 未満と以上に分ける等がその例である。メニュー［基本統計ー量から質変換］を選ぶと、図 3.1 のような量から質変換ツールが表示される。

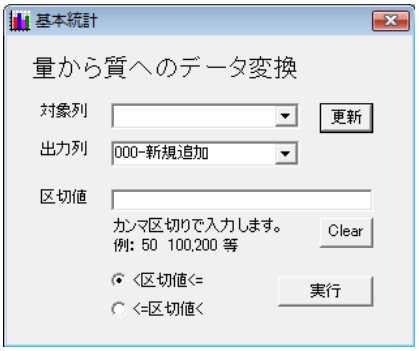


図 3.1 量から質変換ツール

変換したい変数を「対象列」コンボボックスで選択し、出力列を設定して、「区切値」を指定する。例えば上の 170cm の例だと、”170” と入力する。上の設定では新しい列を追加してそこに 170 未満は 1、170 以上は 2 と出力される。未満を以下と変えることもできる。また、160 と 170 で区切って 3 つに分類する場合、”160,170” とカンマ区切りで入力する。結果は、1, 2, 3 の 3 区分となる。新しく作ったこのデータを元に差の検定を行ってもよい。

12.4 データの標準化

多変量解析ではデータを平均 0、（不偏）分散 1 に標準化して分析を実行することが多い。例えば

主成分分析や正準相関分析の相関行列モデルなどがその例である。当初我々はこの標準化の機能を各分析に持たせようと考えたが、今後も多くの分析で利用されることが考えられるので、別個に独立させることにした。図 4.1 にその実行画面を示す。

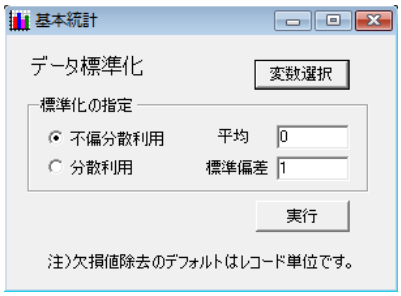


図 4.1 データ標準化実行画面

標準化では分散を固定する場合と不偏分散を固定する場合が考えられるのでメニューにその選択肢を設けている。また、例えば偏差値のように平均と標準偏差の値を 0 と 1 以外に指定する場合もあるので、これらは利用者が設定できるようにした。結果は選択された変数のみを対象として実行する。出力例を図 4.2 に示す。

	体重_std	身長_std
1	0.4782	-1.0677
2	0.0078	0.2953
3	-0.1489	-0.8405
4	0.635	1.204
5	0.1646	0.5225
6	1.8892	0.0681
7	1.1053	-0.159
8	-0.4625	1.6583
9	0.9485	1.204
10	0.4782	0.9768

図 4.2 データの標準化結果

この結果をエディタに貼り付けることにより、そのまま標準化されたデータとして利用することができる。

１３．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。

13.1　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。

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

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、

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

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

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

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

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

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。

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

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

ここで、 $\mu = 0$ の場合は $q(x^{(t)} \mid x^{(t+1)}) = q(x^{(t+1)} \mid x^{(t)})$ 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、

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

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

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。

13.2　ハミルトニアン・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。

13.3　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 のようなメニューが表示される。

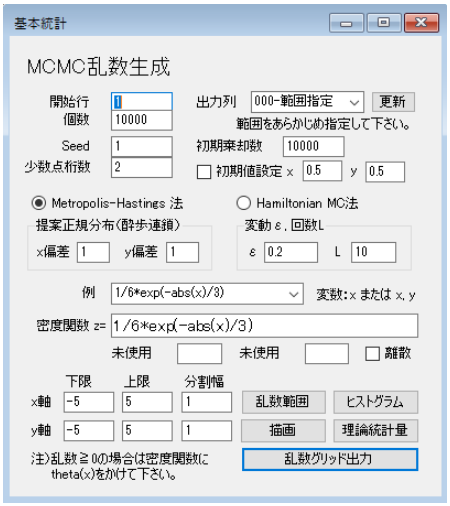


図 1 MCMC 乱数発生メニュー

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。

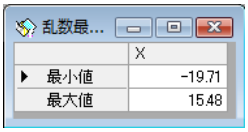


図 2 乱数発生の最小・最大

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。

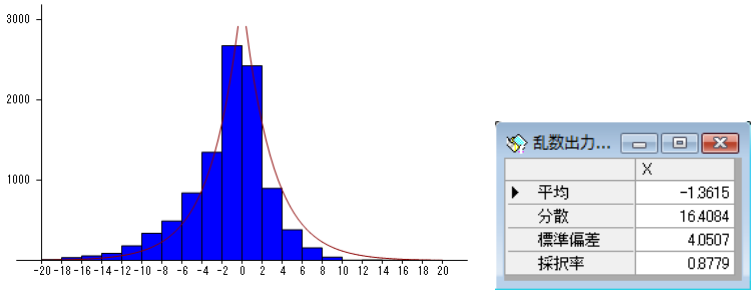


図 3a 乱数のヒストグラムと理論曲線（Seed=1）

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。

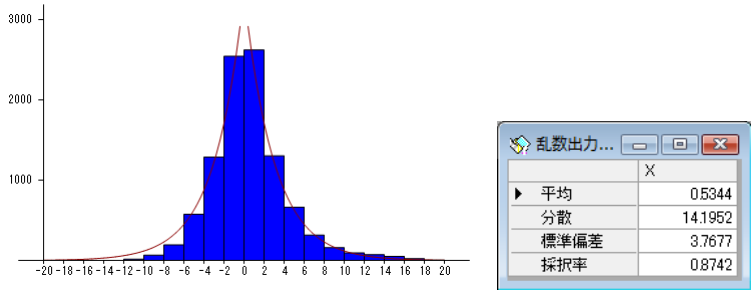


図 3b 乱数のヒストグラムと理論曲線（Seed=2）

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。

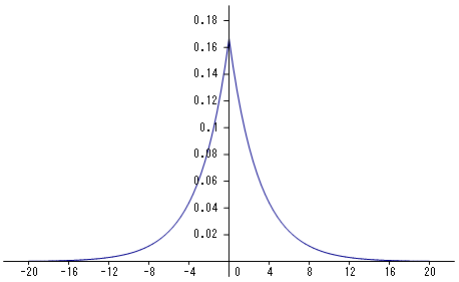


図 4 密度関数グラフ

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。

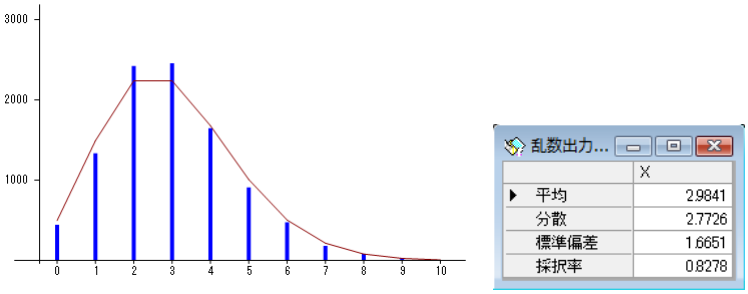


図 6 ポアソン分布

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。

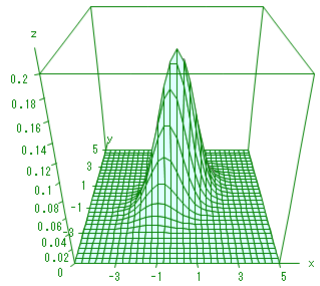


図 7 2 変量正規分布密度関数

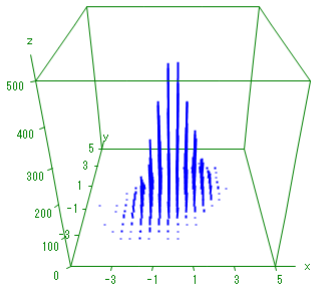
次に、「統計量」ボタンをクリックすると、図 8 に示されるような結果が表示される。

統計量

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。



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９２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．分布の検定

14.1　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。

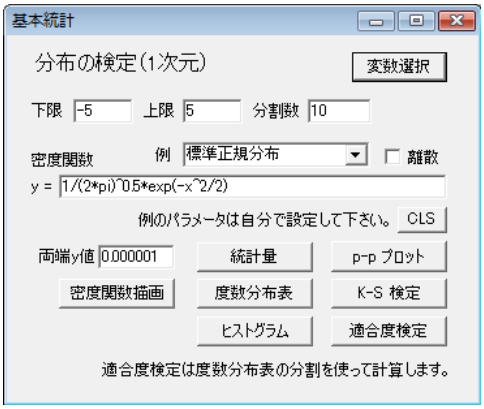


図 1 分析メニュー

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 χ^2 分布を例にする。

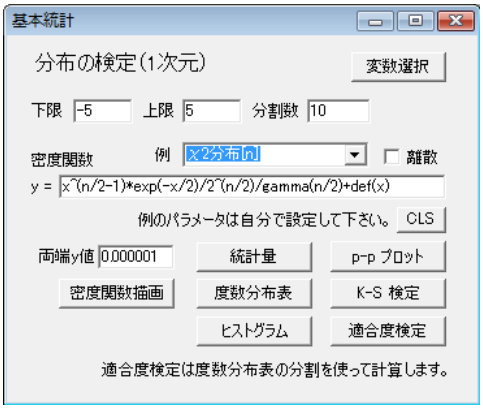


図 2a 密度関数の指定

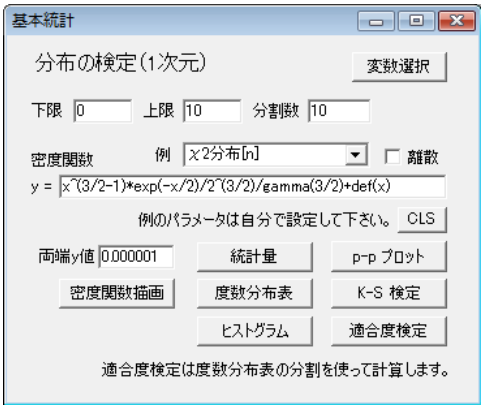


図 2b パラメータと下限・上限の指定

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。

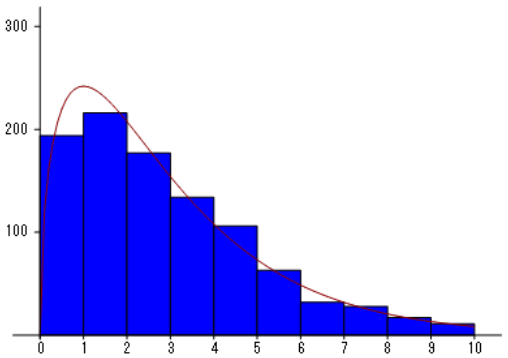


図 5 連続分布のヒストグラム

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。

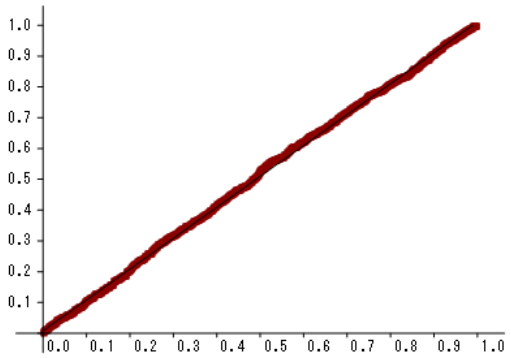


図 6 p-p プロット

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。

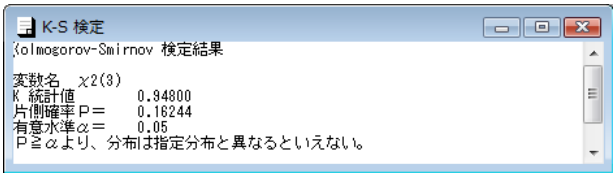


図 7 K-S 検定結果

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。

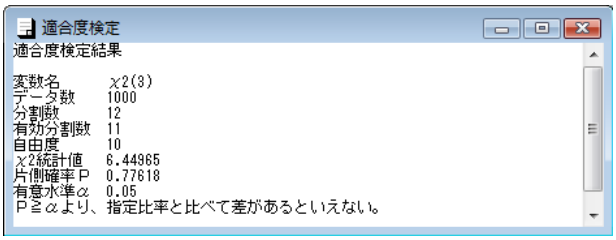


図 8 適合度検定結果

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	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。

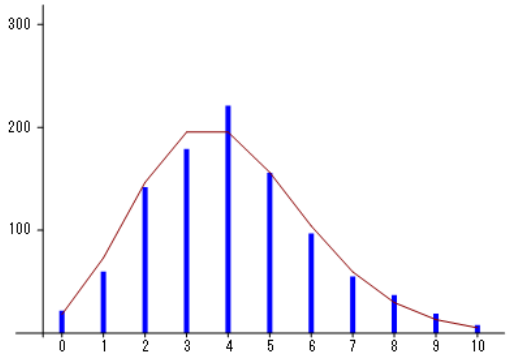


図 10 離散分布のヒストグラム

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。

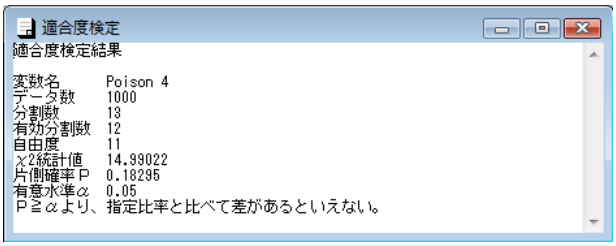


図 11 適合度検定結果

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。

14.2. パラメータの最尤推定

得られたデータ x_λ $(\lambda = 1, \cdots, N)$ 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、メニュー［分析－基本統計－分布と検定］で表示される図 1 の分析実行メニューに、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。

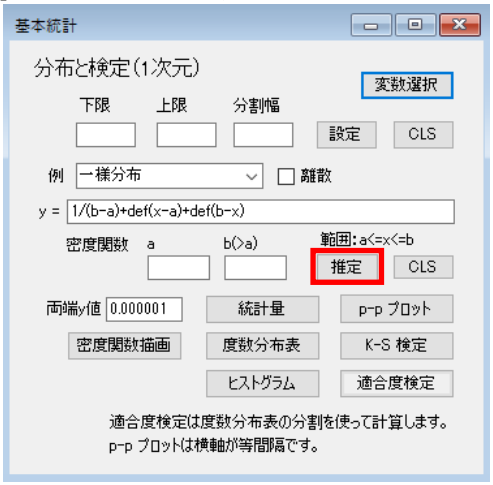


図 1　分布と検定実行メニュー

ここでは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。

正規分布 $(-\infty < x < \infty)$

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

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 $\mathfrak{I}$

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

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

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

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。

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

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

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

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

注） $\partial^2 \log L / \partial \sigma^2 = -\dfrac{2N}{\sigma^2}$ $\partial^2 \log L / \partial \sigma^2 = 4\sigma^2 \partial^2 \log L / \partial (\sigma^2)^2$

パラメータの推定にはどちらの値を用いるべきだろうか。

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

$\chi^2 \sim \chi_n^2$ 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 $(1\leq)\, n-5 \leq n_{\max} \leq n+5$ 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： $f(x) = \frac{1}{B(n_1/2, n_2/2)}\left(\frac{n_1}{n_2}\right)^{n_1/2}\frac{x^{n_1/2-1}}{(1+xn_1/n_2)^{(n_1+n_2)/2}}$

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

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

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

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

これを元に、ぶれが大きいので、$(1\leq) n_i - 20 \leq n_{i\max} \leq n_i + 20$ 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：$V[X]=\dfrac{n}{n-2}$ を利用して、 $n=\dfrac{2V[X]}{V[X]-1}$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

尤度関数：

$$= a^N e^{N\beta} \prod_{\lambda=1}^{N} x_\lambda^{a-1} \exp[-x_\lambda^a e^\beta]$$

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

$$= (a-1) \sum_{\lambda=1}^{N} \log x_\lambda - e^\beta \sum_{\lambda=1}^{N} x_\lambda^a + N \log a + N\beta$$

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

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

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

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

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

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： $P(x) = e^{-a}\, a^{x} \big/ x!$

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

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。

１５．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 の基本統計に加えることにする。

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 に示す分析メニューが表示される。

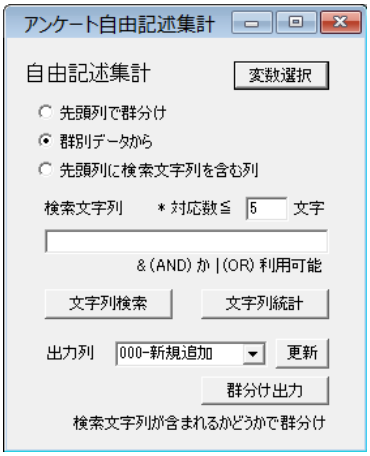


図 1 自由記述集計分析メニュー

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。

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。これらについては順に説明して行く。

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。

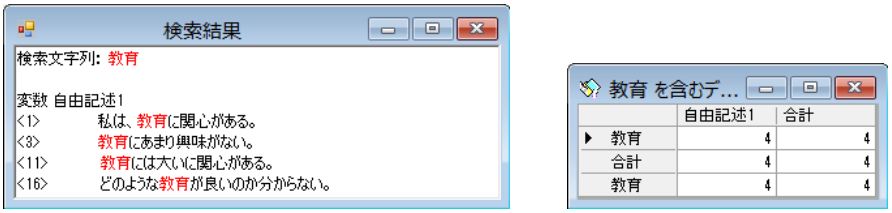


図 3 「群別データから」での検索結果１

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。

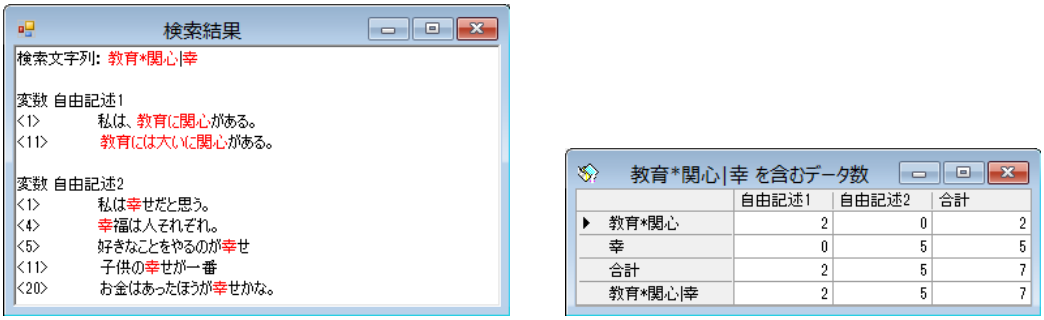


図 4 「群別データから」での検索結果 2

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。

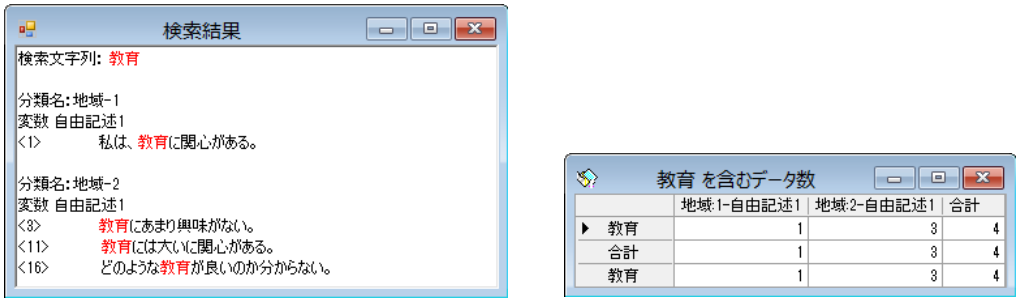


図 5 「先頭列で群分け」での検索結果

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。

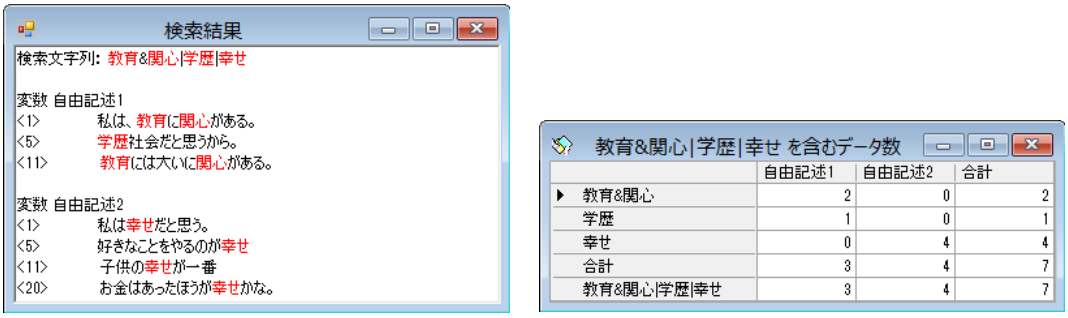


図 6「先頭列に検索文字列を含む列」での検索結果

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、出力列を「新規追加」のまま「群分け出力」ボタンをクリックすると、図 7 のように、新たな列が追加され、選択文字列を含むレコードに 1、含まないレコードに 0 が出力される。

図 7 群分け出力結果

これを用いると、２つの検索文字列の相関などを求めることが可能となる。And で検索した結果を見るより、関係が分かり易くなると思われる。

我々はアンケートにある自由記述欄をある程度数値的に検討できるようになるプログラムを College Analysis に追加した。本格的なテキストマイニングの機能について、現在は考えていないが、今後必要になる可能性もある。これは、現在棚上げ状態にある質的研究のためのツールと連動して考えて行く必要があるだろう。

１６．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のようなメニューが表示されるが、これは元のメニューと変わらない。

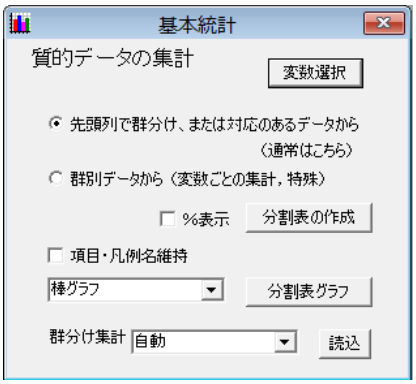


図１ 質的データの集計メニュー

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。

	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。

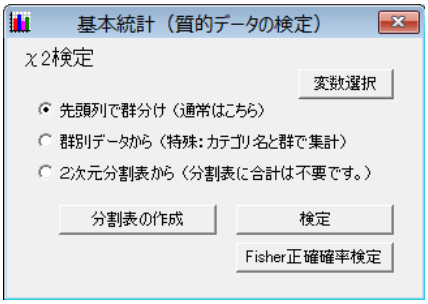


図 3　χ^2検定メニュー

しかし、まとめて変数を選んだ場合は、テキスト表示と違い、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

ここで、集計結果では 0 を入れていた部分は、検定では省略され、2 行 2 列の分割表として処理されていることが、社会活動の自由度(行数-1)×(列数-1)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おこなった。例えば、t 検定のメニューは、図 5 のように与えられ、変更はないが、

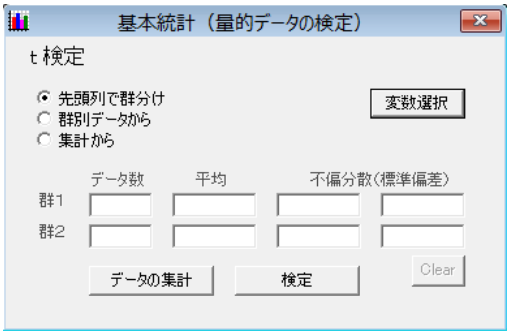


図 5　t 検定メニュー

「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、量的データ検定メニューに検定を自動選択するボタンを加えた。

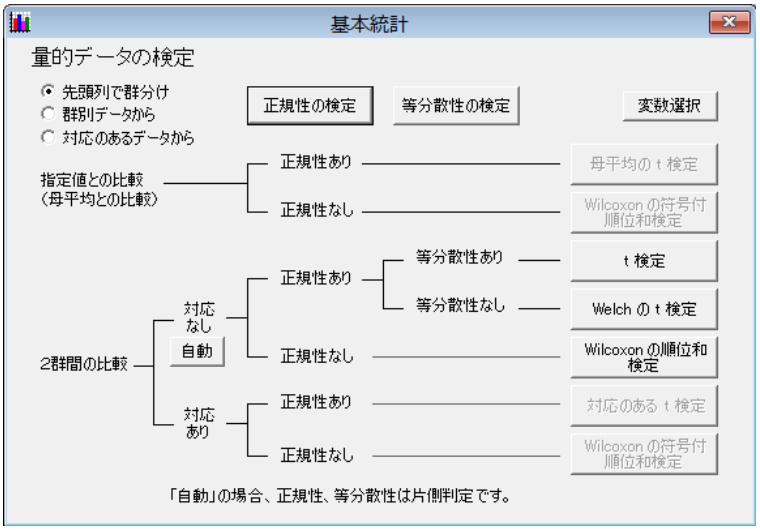
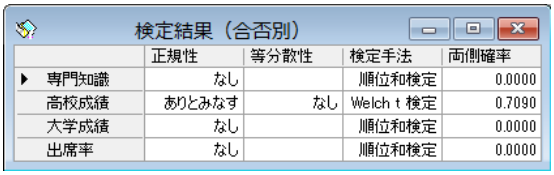


図 7 量的データ検定メニュー

変数を上の 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。

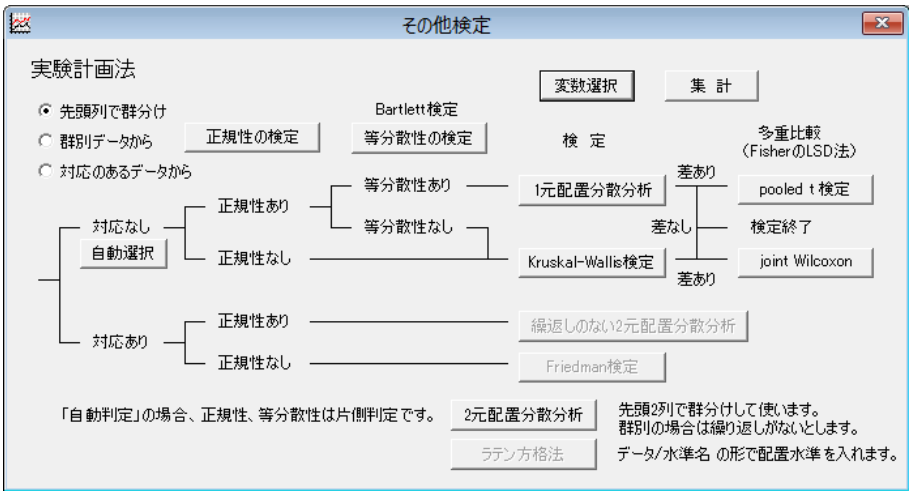
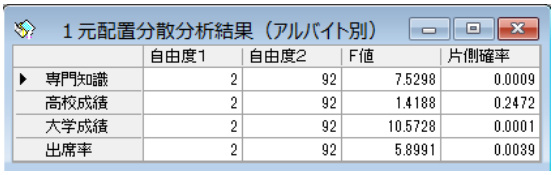


図 9 実験計画法メニュー

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 10 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 10 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。

１７．層別分割表の検定

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)]。

17.1　計算方法

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。

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。

$$D \equiv \frac{\sqrt{N-1}\left(x_{11}x_{22} - x_{12}x_{21}\right)}{\sqrt{m_1 m_2 n_1 n_2}} \sim N\left(0,1\right)$$

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

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

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。

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

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

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

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。

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

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。

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

17.2 プログラムの利用法

これらの検定について、我々の作成したソフトの利用法について説明する。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 の実行メニューが表示される。

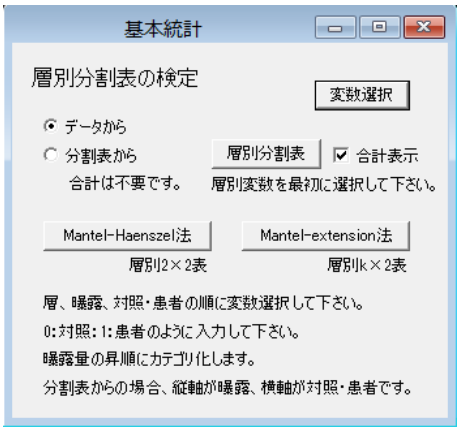


図 1 実行メニュー

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。

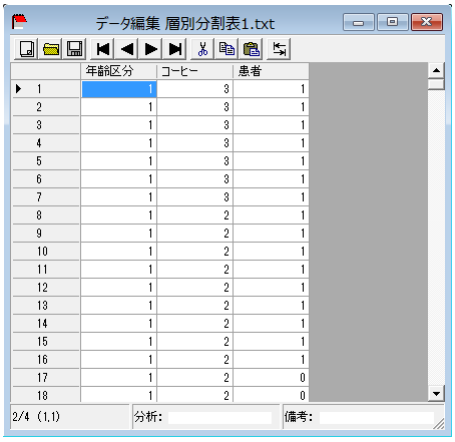


図 2 「データから」を選択した場合のデータ形式

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。このデータは曝露変数について 0 ～ 3 の 4 区分に分類されており、Mantel-extension 法に用いるデータである。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。

「合計表示」のチェックボックスにチェックを入れて、「層別分割表」ボタンをクリックすると、図 3 のような合計を含む分割表が得られる。

	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 3 合計を含む層別分割表

「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4のような合計を含まない分割表が得られる。

	1:患者-0	1:患者-1	2:患者-0	2:患者-1	3:患者-0	3:患者-1
▶ コーヒー-0	16	2	19	5	21	4
コーヒー-1	47	9	50	21	55	22
コーヒー-2	24	9	25	22	31	22
コーヒー-3	19	7	14	11	15	10

図 4 合計を含まない層別分割表

この分割表の形式は、検定を行う際の、ラジオボタン「分割表から」を選択した場合のデータ形式でもある。分割表から検定を行う際に合計は不要である。

曝露変数が 2 分類の場合、交絡因子を調整したオッズ比の検定である Mantel-Haenszel 法が利用可能である。同名のボタンをクリックすると計算結果が図 5 のように表示される。

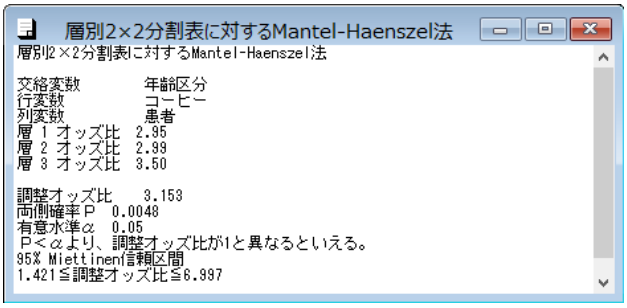


図 5 層別 Mantel-Haenszel 法計算結果

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 Mantel-extension 法が利用可能である。同名のボタンをクリックすると計算結果が図 6 のように表示される。

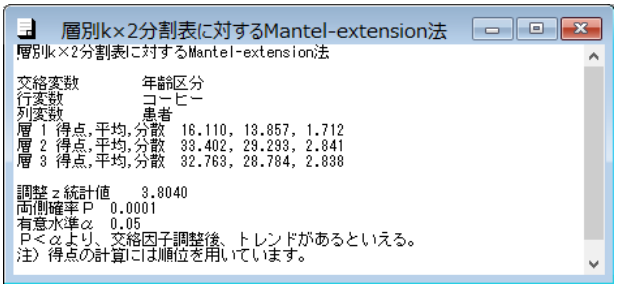


図 6 層別 Mantel-extension 法計算結果

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。

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.