岡本安晴　2001.10

カイ２乗検定：１次元度数分布

例えば、表１の形のデータに対して、目の数の頻度が出現比率がお互いに等しい母集団からのものと考えてよいかどうかというような検定について考える。

表１　サイコロを１００回振ったときの目の数の出現頻度

目の数	1	2	3	4	5	6
出現頻度	13	16	23	14	19	15

表１を一般的にあらわすと表２の形となる。

表 2　ｎ個の観察データをｋ個のカテゴリーに分類して数えたもの。f_iはカテゴリーｉに分類されたものの数。データがカテゴリーｉに分類されるものである確率をp_iとしたとき、観測度数の期待値は$e_i = np_i$となる。

	カテゴリー１	・　・　・	カテゴリーｋ
観察度数	f_1	・　・　・	f_k
確　　率	p_1	・　・　・	p_k

表２では、ｎ個のデータ（観測値）が、ｋ個のカテゴリーに分類されている。カテゴリーｉに分類されたものの観測度数が、f_iで表わされている。カテゴリーｉに分類される確率をp_iとおくと、観測度数の期待値e_iは次式で与えられる。

$$e_i = np_i$$

カイ２乗検定では、この期待値e_iを、データとして与えられた観測度数f_iと比べることによって、「確率がp_iで与えられる」という仮説の妥当性を調べる。

すなわち、次式

$$c^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \qquad (1)$$

で与えられる値c^2を、自由度ｋ－１のカイ２乗分布と比較することによって、帰無仮説「カテゴリーｉの確率はp_iである」の検定を行う。f_iとe_iの値の違いが大きいほど、式（1）で与えられるc^2の値も大きくなる。

有意水準aで帰無仮説「カテゴリーｉの確率はp_iである」を検定するときは、自由度 k-1 のカイ２乗分布に従う確率変数Ｘに対して、次式

$$P(X \geq c_a^2) = a$$

を満たす基準値c_a^2を求め、式（1）により算出されるc^2の値と比べる。式（1）により与えられるc^2の値がc_a^2の値より大きいとき、e_iとf_iの違いが大き過ぎるとして、有意水準aで帰無仮説「カテゴリーｉの確率はp_iである」は棄却される。

プログラム PChiSqr1S.dpr は、等確率

$$p_1 = p_2 = \cdots = p_k = \frac{1}{k}$$

を帰無仮説とするものである。このとき、期待値e_iは

$$e_i = np_i = \frac{n}{k}$$

となる。

プログラム PChiSqr1S.dpr を実行すると、図１のフォームが表示される。

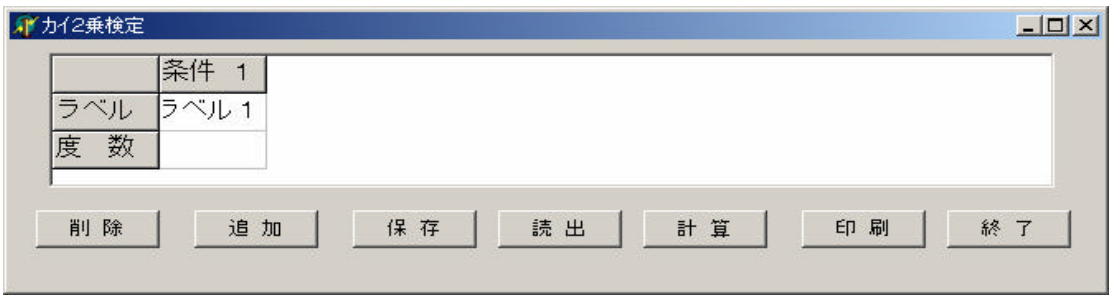

図１　起動時のフォーム

StringGrid コンポーネント内に設定するデータは、「追加」あるいは「削除」ボタンのクリックを適当に繰り返すことによりセルの列数をデータ数に合わせる。列を増やすときは、「追加」ボタンをクリックする。「追加」ボタンをクリックすると、アクティブなセルを含む行の下に空白行が挿入される。セルは、クリックされるとアクティブになる。

データの設定は StringGrid 内のすべてのセルに対して行わねばならない。データの設定されていない空白のセルがあってはならない。余分な列は、「削除」ボタンのクリックで削除することができる。削除は、アクティブなセルを含む列が除かれる。

図２はデータを設定した状態である。

図２　データを設定した状態

「保存」ボタンをクリックすると、設定したデータを保存することができる。「保存」ボタンをクリックすると、まず図３のダイアログボックスが表示される。

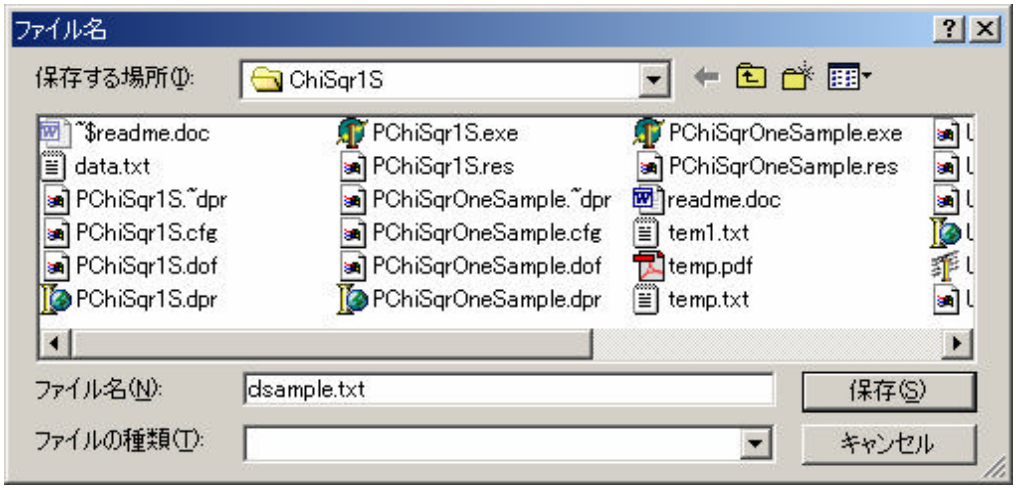

図３　データ保存用ファイル名の設定

ファイルの拡張子は何でもよいが、データはテキストファイルとして保存されるので、図３では拡張子を.txt としている。データ保存用のファイル名の設定後、図３の「保存」ボタンをクリックすると、設定した名前のファイルにデータが保存される。

保存したデータは、「読出」ボタンのクリックで読み込むことができる。「読出」ボタンをクリックすると、まず読み出すファイルの名前を設定するダイアログボックスが図４のように表示される。

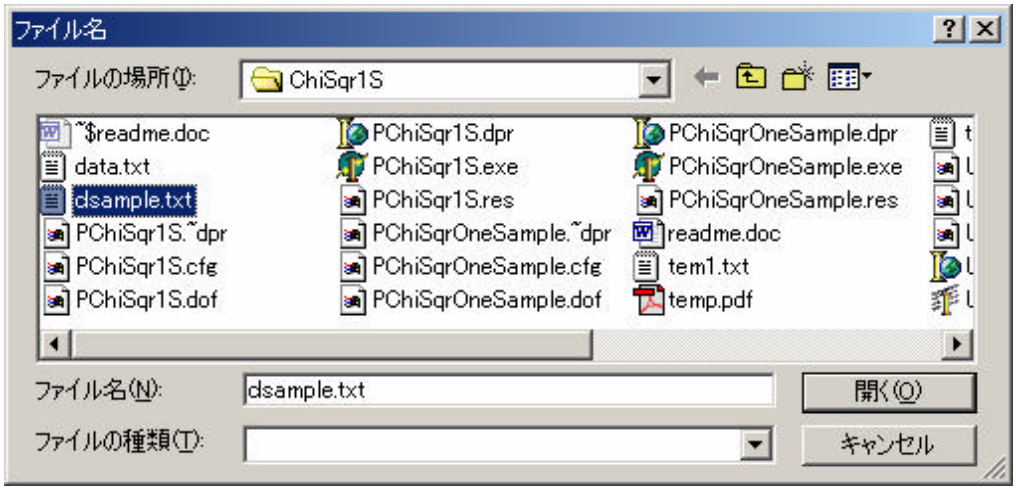

図４　データ読み込みファイル名の設定

ファイル名の設定後、図４の「開く」ボタンをクリックすると、設定したファイルからデータが読み込まれる。

設定されているデータは、「印刷」ボタンのクリックでプリンタに出力することができる。

図２のようにデータを設定した状態で「計算」ボタンをクリックすると、設定されたデ

ータに対するカイ２乗の値が算出される。「計算」ボタンのクリックで、まず図５のダイアログボックスが表示される。

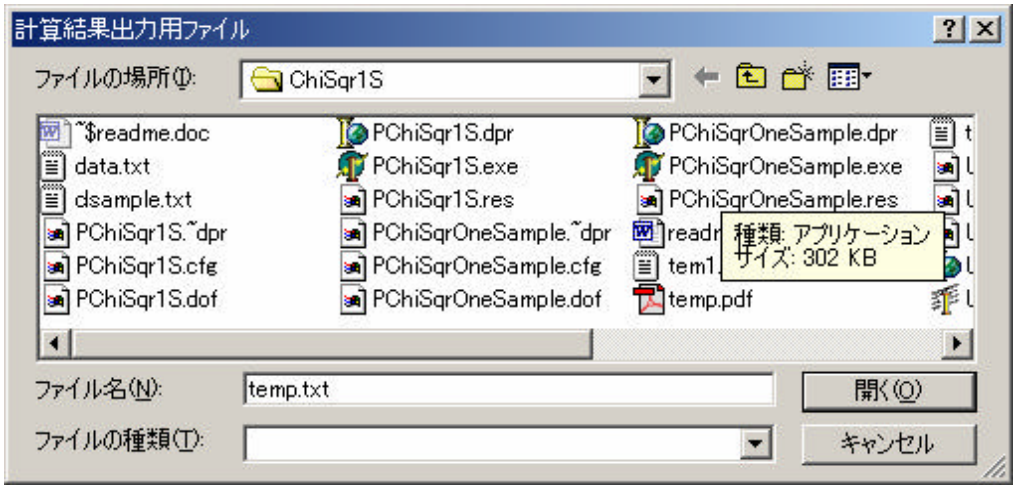

図５　計算結果出力用ファイル名の設定

図５で設定した名前のファイルに、計算結果がテキストファイルとして出力される。名前の設定後、「開く」ボタンをクリックすると計算が始り、計算が終了すると図６のフォームになる。

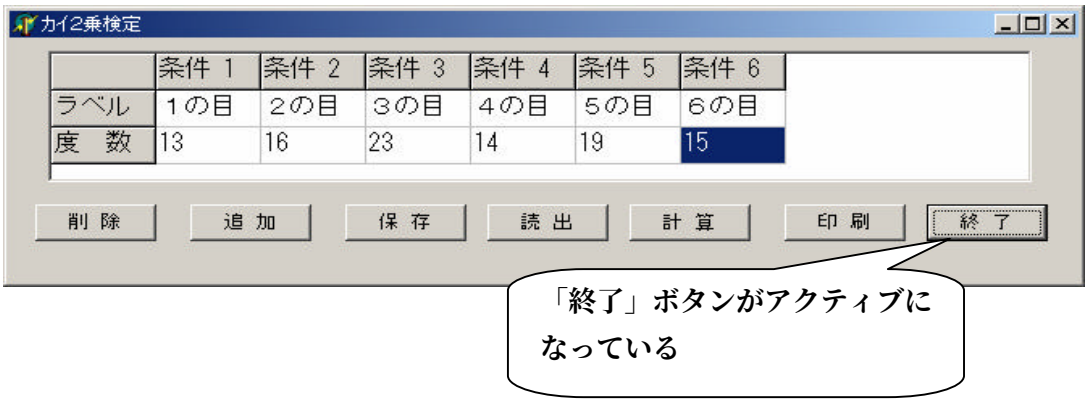

図６　計算終了時のフォーム

計算が終了すると「終了」ボタンがアクティブになる。「終了」ボタンのクリックでプログラムの実行が終了する。

プログラムの実行終了後、図５で設定した名前のファイルを開くとリスト１のようにな

っている。

リスト１　計算結果の出力例

```
データ
    13    16    23    14    19    15
総数 = 100

e =     16.7

Chi square =     4.16       df = 5
```

リスト１に示されているカイ２乗の値に対するｐ値は

$$p = P(c_5^2 > 4.16) \approx 0.527 > 0.05 = 5\%$$

となる。したがって、リスト１の結果の場合、有意水準５％で６つの目の出現確率には差が認められない。

表 2 の形のデータにカイ２乗検定を適用するとき、次の条件が満たされている必要があるとされている（Siegel & Castellan,Jr.,1988）。

（ａ）　自由度が１、すなわちカテゴリー数が２の場合は、どちらの期待値e_iも５以上でなければならない。

（ｂ）　自由度が２以上、すなわちカテゴリー数が３以上の場合は、期待値e_iの20%を越えるものが５より小さい、あるいはいずれか１つでも１より小さいものがあるということがあってはならない。

（ａ）あるいは（ｂ）の条件が満たされなければならないのは、データ数が十分に多いときに（1）式で計算される値の分布がカイ２乗分布で近似できるからである。（ａ）の条件が満たされていないときは、２項分布による検定などを用いる。（ｂ）の条件が満たされていないときは、いくつかのカテゴリーを１つにまとめて、e_iの値が十分に大きくなるよう

にする。カテゴリーの合併は、検定の目的を考えて行う。