

睡眠覚醒リズム解析ソフト

～ カイ二乗分布グラフ ～

補足資料

目次

□ はじめに	2
□ 概略	3
□ 統計学の基礎知識	4
1、自由度について	4
2、標本数と自由度の関係	5
3、有意確率	6
□ 本プログラムについて	8
1、考えている標本	8
2、考えている平均	9
3、青線と黒線の関係	10
4、カイ二乗値を考えることで、周期性が導出できるわけ	11
□ 棄却域の境界値(パーセント点)	11
□ 参考文献	13

- 睡眠覚醒リズム解析ソフト ～～ カイ二乗分布グラフ
本ソフトでカイ二乗検定に関する問合せが多い為、説明を補足します。

□ はじめに

書籍等でカイ二乗値を学ぶと、次の二つの計算式を見つけ、混乱を誘う場合が多いようですが、先ず、本ソフトにおけるカイ二乗の対応について、説明します。

- (1) **カイ二乗 = $\Sigma \{ (\text{観測値}) - (\text{期待値}) / \text{期待値} \}$**
(2) **カイ二乗値 = (標本サイズ × 標本分散) / 母分散**

式(1)は「 **適合度検定などに用いられる時のカイ二乗値** 」です。

カイ二乗検定としてもっとも良く利用されるもので、
この式を用いた検定にはピアソンの適合度検定という名前がついています。
インターネット等で「カイ二乗検定」をキーワードに検索をすると、
たいていの場合にはこの式を用いた計算に出会うようです。

式(2)は「 **標準正規分布に従う母集団の母分散の検定** 」に使用されます。

**本ソフトで用いているカイ二乗値は、
こちらのカイ二乗値に対応しています。**

この点に注意してください。

※1 どちらも、検定で用いる指標がカイ 2 乗分布となることから、カイ 2 乗値という呼び方をしています。共通することとしては、どちらもデータのばらつきに関する検定として使われていることがあげられます。

※2 (1)の式において、 $(X - \bar{X}) / \sigma$ で変数の標準化を行い、二乗和をとると

$$\Sigma((X - \bar{X}) / \sigma)^2$$

$\bar{X}$: (標本) 平均, σ : 母分散

となり、(2)の式に帰着します。

□ 概略

□ 黒線

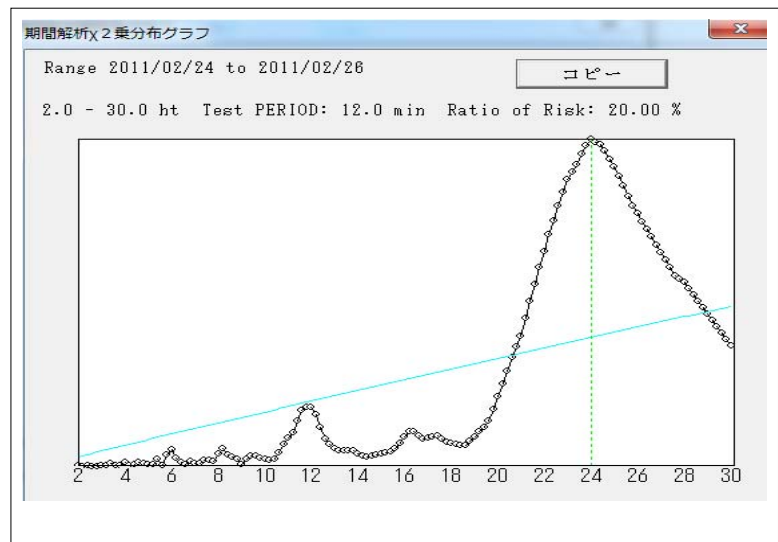
横軸・・・時刻

縦軸・・・カイ二乗値

□ 青線

横軸・・・自由度

縦線・・・カイ二乗値



□1 上記のように、二つの線の横軸は異なっていることに注意が必要です。

青線は自由度のみにしか依らない値であるので、黒線の値には依存していません。

□2 二つのグラフは、計算過程で自由度がそろうように計算しています(後述)。

□3 今回計算している「カイ二乗値」とは、

$$\chi^2 = (n - 1)s^2 / \sigma^2$$

のことです。ここで、

$$X_1, X_2, X_3, \dots, X_n$$

は標本のことであり、

$$\bar{X} = (X_1 + X_2 + \dots + X_n) / n$$

は標本平均、

$$s^2 = \frac{1}{n - 1} \{ (X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \}$$

は標本分散(不偏分散とも呼ばれています)、 σ^2 は母分散をあらわします。

□ 統計学の基礎知識

1、 自由度について

独立した確率変数の個数に対応。

これだけではわかりにくいので、以下に具体例を交えつつ説明します。

例1

「1」、「2」とかかれた二枚のカードがあるとします。

A、Bという二人が一枚ずつカードを取るとしましょう。(この段階では、A、B以外の人からは、どちらが何番のカードを持っているのかわかりません。)

ここで、

「 Aが「1」のカードを持っている 」

という情報が入ると、自動的に

「 Bが「2」のカードを持っている 」

ことが決定されます。この場合の自由度は「1」であるといえます。

□1 今回のケースでは、確率変数は

- ・ 「1」と書かれたカード
- ・ 「2」とかかれたカード

の二つですが、

- ・ カードが「1」か「2」しかない

という条件を考慮すると、独立した確率変数は1つしかないことになります。

□2 もし、カードにかかれた数字がまったく未知である場合(「1」か「2」の二枚しかないという仮定をはずすと、一枚のカードが決まっても、もう一枚のカードが何であるかわからないので、この場合は独立した確率変数が2つあることになります(自由度2))。

例2

「1」、「2」、「3」とかかれた三枚のカードがあるとします。

A、B、Cという三人が一枚ずつカードを取るとしましょう。(この段階では、A、B、C以外の人からは、どちらが何番のカードを持っているのかわかりません。)

ここで、

「 Aが「3」のカードを持っている 」

という情報が入っても、まだB、Cのもつカードが何であるかは決まりません。

この情報に加えて、

「 Cが「2」のカードを持っている 」

という二つ目の情報が与えられると、

「 Bが「1」のカードを持っている 」

ということが決定されます。この場合の自由度は「2」であるといえます。

2、 標本数と自由度の関係

・標本数・・・n

・自由度・・・f

とすると、

$$f = n - 1 \quad (1)$$

という関係が成立します。詳しい説明は[1]や[2]を参考にすると良いでしょう。
ここでは概略のみ述べます。

先述したように、自由度は「独立した確率変数の個数」に対応します。

今回の場合には、標本数が確率変数に対応します。

また、抽出した標本を、

$$X_1, X_2, X_3, \dots, X_n$$

とし、標本平均を

$$\bar{X} = (X_1 + X_2 + \dots + X_n) / n$$

とすると、

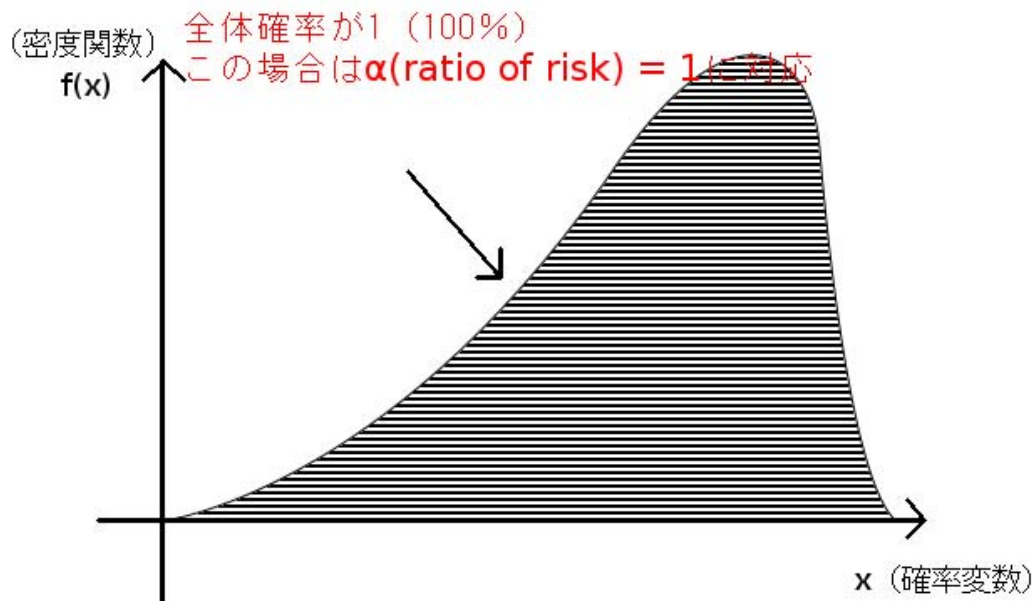
$$(X_1 - \bar{X}) + (X_2 - \bar{X}) + \dots + (X_n - \bar{X}) = 0$$

という拘束条件が一つあるため、独立した確率変数は n-1 個です。

つまり、上記の式(1)が成立します。

3、 有意確率

ソフト内では「 α : ratio of risk 」というのがこの値に対応します。
この α は確率をあらわす。どのような確率のことを言っているのかということを、
以下で説明します。



上記のような密度関数を対象に話を進めます。

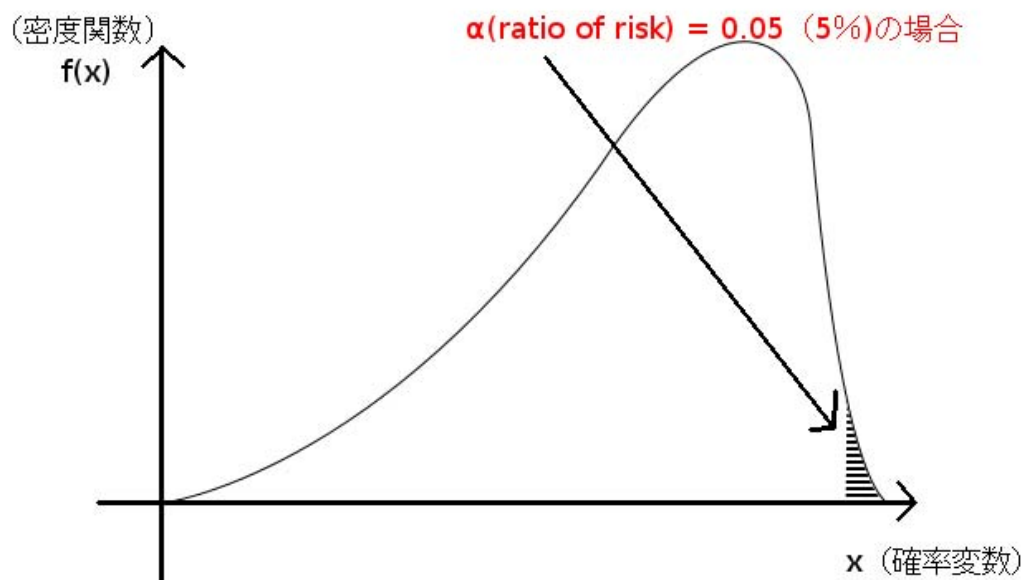
この場合のグラフの下にある面積が、全体確率をあらわします。
確率は $0 \sim 1$ (0%~100%) までの値をとり、この場合は全体をとっているので
確率 1 をあらわします。
ちなみに、この場合は α (: ratio of risk) が 1 である場合に対応しています。

□(補足) 確率密度関数を積分すると、積分区間内における確率となります。
そのため、図を描いたときに、

「 面積 = 確率をあらわしている 」

という点を理解していると、図からの情報を読み取る助けとなるでしょう。(補足終了)

次に α (: ratio of risk) が 0.05 (5%) である場合を見てみましょう。

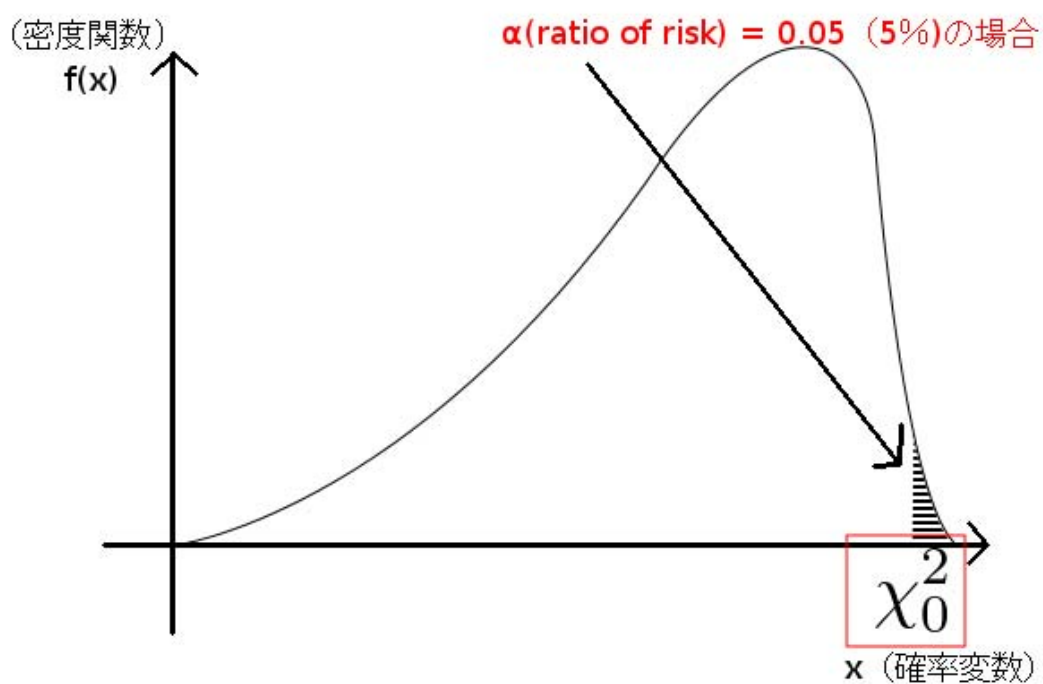


α (ratio of risk) が 0.05 (5%) というのは、要するに全体(の面積)を 1 (100%) としたときに、右から 0.05 (5%) 分の面積をあらわします。

ここで、ちょうど 0.05 (5%) となるときの横軸の値こそ、

「棄却域の境界値 (100α パーセント点と呼ばれます) 」

をあらわします。(この値のことをしばしば「 χ_0^2 」と書きます。) つまり、



ということになります。

□ 本プログラムについて

要点としては、以下の4点があげられます

- ・ 何を標本として用いているのか
- ・ 標本平均について
- ・ 青線と黒線の関係
- ・ なぜカイ二乗値を用いると周期性が導出できるのか

以下、これらを順に説明することといたします。

1、 考えている標本

一口に標本といっても、何をさしているのかあいまいです。

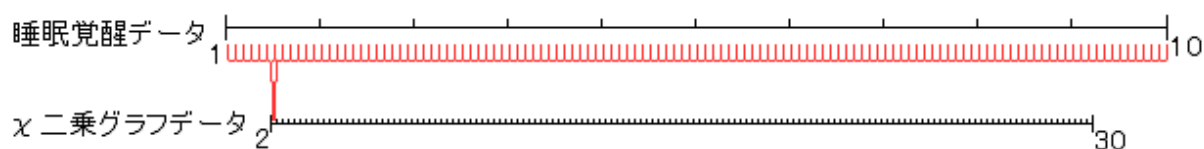
今回考えている標本について、少々詳しく説明したいと思います。

今回は 10 日分のデータを考えることにします。

- ・ データ数 480 個 (30 分 1 データ)
- ・ 刻み幅 (test period) : 0.2 (12 分ごと)
- ・ 0 : おきている、1 : 寝ている

$x_1, x_2, \dots, x_{480}$ で、480 個のデータをあらわすことにします。

例 1



時刻 2 における実測値 :

10 日分の睡眠覚醒データ (上の図) に注目しましょう。

一番最初のデータから、2 ずつずらしながら最後のデータまで足し合わせます。

30 分に 1 データあるので、2 時間では 4 つのデータ飛ばしごとに足していくことになります。これを取れるところまでとって、平均をとると、

$$(x_1 + x_5 + x_9 + \dots + x_{473} + x_{477}) / 120$$

これを一つの**実測値 X_1** として使うこととなります。

同様に、刻み幅 (test period) だけずらした位置から、同じように 2 時間分ずつずらしながら計算を行い、平均を取ったものを次の実測値 X_2 としてもちいます。

ここでどれだけ実測値を考えるかが、自由度を決めていることと同値であります。

注意すべき点としては、30 分に 1 データしかないので、

12 分刻みでとった場合に、データが存在しないケースがあることです。

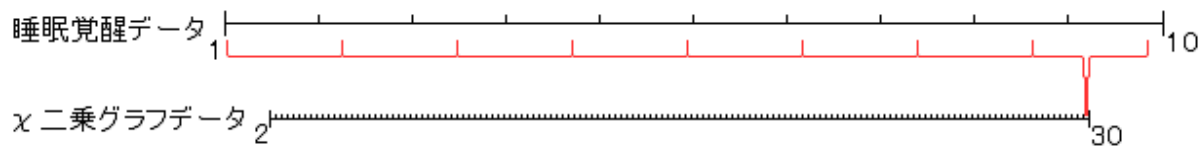
(たとえば、0, 30 60. 90. ...というデータに対して、12, 24 などはデータを持ちません)。

こういった場合には、15 分を境に四捨五入して考えています。

つまり、**12 に関しては 0 のデータ**を使い、**24 に関しては 30 のデータ**を使います。

例2

時刻 30 における実測値：



同様に 10 日分の睡眠覚醒データ(上の図)に注目しましょう。

一番最初のデータから、2 ずつずらしながら最後のデータまで足し合わせます。

30 分に 1 データあるので、30 時間では 60 個のデータ飛ばしごとに足していくことになります。

これを取れるところまでとって、平均をとると、

$$(x_{_1} + x_{_61} + x_{_121} + \dots + x_{_361} + x_{_421}) / 8$$

これを一つの実測値 X_1 として使うこととなります。

同様に、刻み幅(test period)だけずらした位置から、同じように 30 時間分ずつ、ずらしながら計算を行い、平均を取ったものを次の実測値 X_2 としてもちい、繰り返します。

上の式を見てわかるように、**時刻が大きくなるにつれ、一つ一つの標本の精度が下がっていることがわかります(平均を取る際の分母が小さくなっている)。**

そのため、時刻を大きくするに従い、標本数を多くとることで、時刻ごとの精度差を小さくしようとしているのです。(自由度を大きくとっていることにも対応します)。

2、 考えている平均

これは単純に、得られたすべてのデータの平均値のことです。

例として、

- ・ 0：おきている状態
- ・ 1：寝ている状態

という情報が、30 分に 1 データあるとします。

これを 5 日分とると、すべてのデータ数は 240 個となります(一日 48 個のデータ)。

人の睡眠時間が 1 日平均 8 時間程度であると仮定するなら、

一日 24 時間であることを考えると、480 個のデータの平均は、約 1/3 程度であると予測されます(データにより当然変動する)。

注意点としては、**上記1の標本の平均を取っているわけではない点。**

これは何を意味するかというと、

$$\text{「 すべてのデータの平均値 } = \text{ 今回考えている標本平均 } \text{」}$$

という暗黙の仮定していることに対応します。

3、 青線と黒線の関係

青線で考えている自由度は、時刻 1 において自由度 0 (ここまでは意味のない値)。刻み幅ごとに 1 ずつ増えていくようにしている。

(たとえば、刻み幅 0.2 (test period : 0.2) である場合には、時刻 1.2 で自由度 1、時刻 1.4 で自由度 2、時刻 1.6 で自由度 3、・・・となります)

それぞれの自由度において、設定された有意確率 (ソフト内の ratio of risk に対応) を用いて、棄却域の境界値 (パーセント点) パーセント点を導出してプロットしています。

黒線については、それぞれの時刻において、

$$\chi^2 = (n - 1)s^2 / \sigma^2$$

を使ってカイ二乗値を計算してプロットしています。

その際、重要な点は、

「青線の自由度と同じ自由度

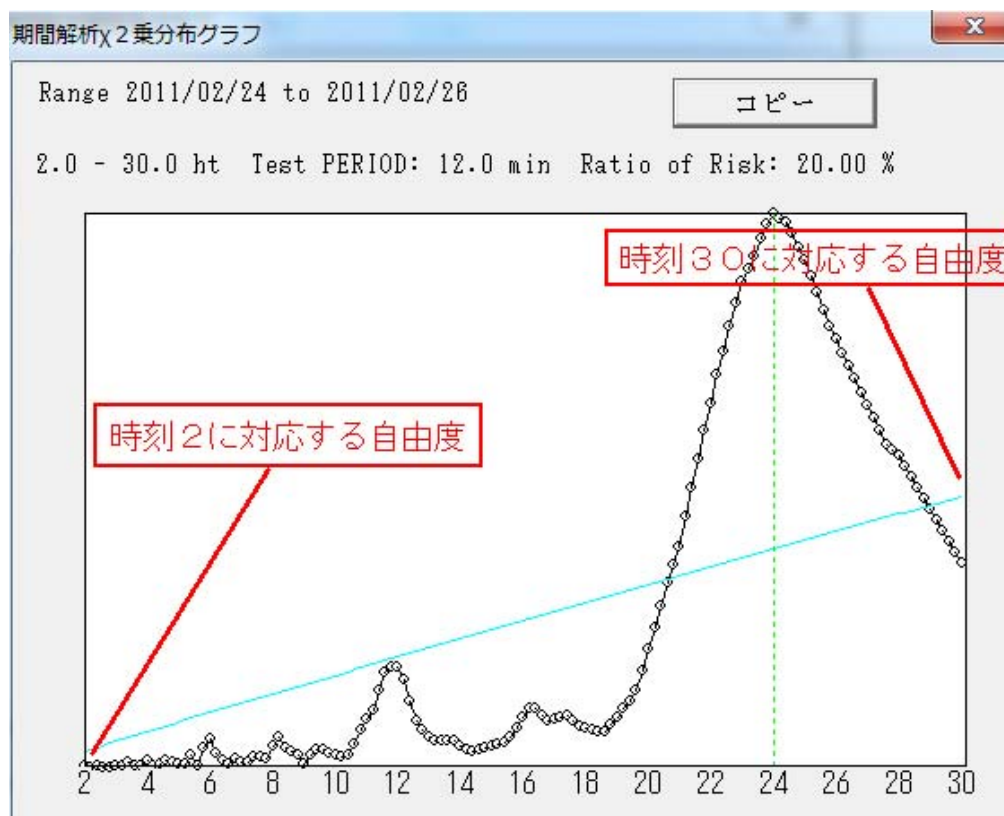
となるように、とってくる標本の個数を調節しながらカイ二乗値を計算している。」

ということが言えます。つまり、時刻 1 においては、もってくる標本の個数は 2 つ。

刻み幅が 0.2 (12 分) である場合には、時刻 1.2 において標本の個数は 3 つ。

以下、刻み幅ずつ考える標本の個数を 1 ずつ増やしていくことで、

青線と黒線の自由度をそろえてプロットしています。



4、 カイ二乗値を考えることで、周期性が導出できるわけ

簡単に説明すると、これは、

「 周期性がある = 周期時間後に、同じ状態にある可能性が高い 」

という性質を用いています。

- ・ 0：おきている状態
- ・ 1：寝ている状態

という情報が、30 分に 1 データあるとします。

これを 10 日分とすると、すべてのデータ数は 480 個になります(一日 48 個のデータ)。

人の睡眠時間が 1 日平均 8 時間程度であると仮定するなら、

一日 24 時間であることを考えると、480 個のデータの平均は、約 1/3 程度であると予測されます。

周期性を持つデータの場合には、

実測値としては 0、もしくは 1 の値である可能性が高く、

これと期待値(上記した 1/3 に近い値)との差をとり、

2 乗和をとった際に、もっとも大きな値になりやすいことがわかります。

□ 棄却域の境界値（パーセント点）

[1]付録の表Ⅲの近似を求める方法はいくつか考案されています。

今回プログラム内部では次の方法で計算を行っています。

```
static double Xalpha( double alpha )
{
    static double c0 = 2.515517, c1 = 0.802853, c2 = 0.010328;
    static double d1 = 1.432788, d2 = 0.189269, d3 = 0.001308;
    double z, bu, bd;

    z = pow( -2.0 * log( alpha ), 0.5 );
    bu = c0 + c1 * z + c2 * pow( z, 2.0 );
    bd = 1.0 + d1 * z + d2 * pow( z, 2.0 ) + d3 * pow( z, 3.0 );
    return( z - bu / bd );
}

static double ComputeCHI( double alpha, int freedom )
{
    double value1, value2, value3;
    value1 = 2.0 / (9.0 * (double)freedom);
    value2 = pow(value1, 0.5);
    value3 = Xalpha(alpha);
    return((double)freedom * pow((1.0 - value1 + value2 * value3), 3.0));
}
```

以下はプログラムの一部抜粋です。

上記プログラムで行っていることを、数式を使い書き直してみましょう。

$$Z_0 = z - \frac{c_0 + c_1 z + c_2 z^2}{1 + d_1 z + d_2 z^2 + d_3 z^3}$$

まずは上部の計算部で、

ただし、

$$z = \sqrt{-2 \ln(\alpha)}$$

$$c_0 = 2.515517, c_1 = 0.802853, c_2 = 0.010328$$

$$d_1 = 1.432788, d_2 = 0.189269, d_3 = 0.001308$$

これはヘスティング(Hessting)の最良近似式と呼ばれている計算式です。

明記されている書籍を見つけることができなかったため、参考になりそうなウェブサイトを[3],[4]として載せておきました。

- ・ 自由度 2 の場合には、ここで得た値を初期値としてニュートン法を使い、パーセント点を導出します。
少し補足しておくと、ヘスティング(Hessting)の最良近似式は標準正規分布 $N(0, 1)$ に従う場合のパーセント点導出に使われるものです。
ではなぜカイ二乗分布を考えているのにこの近似式を使うかというと、これは以下のカイ二乗値の定義自体から来ています。

definition1.

$$Z_1, Z_2, \dots, Z_k$$

を、標準正規分布 $N(0, 1)$ に従う確率変数であるとする、

$$\chi^2 = Z_1^2 + Z_2^2 + \dots + Z_k^2$$

確率変数 χ^2 が従う確率分布を自由度 k の χ^2 分布という。

この定義より、標準正規分布 $N(0, 1)$ の変数近似が必要である理由がわかります。

- ・ 自由度3以上の場合は、「Wilson-Hilferty」の近似式より導出します。

$$f \times \left(1.0 - \frac{2}{9f} + z_0 \sqrt{\frac{2}{9f}} \right)^3$$

※ ここでfは自由度をあらわします。

いずれの場合においても重要な点は、

「 自由度と、有意確率(今プログラムにおいては、ratio of risk。つまり危険率のこと)のみにしか依存していない 」

ということです。

黒線の値を元に、なんらかの計算を行った結果ではないという点に注意が必要でしょう。

以上の計算から、[1]の表Ⅲ (p.386)の近似が得られます。

□ 参考文献

- [1] PG.ホエール著, 浅井 晃, 村上 正康訳(1978)「入門数理統計学」, 培風館
- [2] 東京大学教養学部統計学教室編(2011(第 31 刷))「統計学入門」, 東京大学出版会
- [3] <http://www.snap-tck.com/room04/c01/stat/stat99/stat9901.html>
- [4] http://www.geocities.jp/ikuro_kotaro/koramu/tyokika.htm