

1. はじめに

統計分析とは何をするのでしょうか？ 何の役にたつのでしょうか？

ざっくりと言うと 統計分析は 違いを明らかにする道具 です。

例えば学校である学科の試験をし、AクラスとBクラスに違いがあるか判断する場合を考えます。

統計分析では両クラスの統計量(平均値、分散値)を計算し議論します。

統計量という基準に基づくことで、違いの有無を 明確にします。

この明確な基準がないと、

- ・AクラスとBクラスで明らかに差があるのに、担任がごねてその差を認めない。
指導法の改善が必要なのに、改善を図ろうとしない。
- ・AクラスがわずかにBクラスを上回ったが、実際は大差ないにもかかわらず、
Aクラスの担任が殊更に優位を主張しすぎる。

というようなことが起きます。

統計分析が発展した背景には、農業試験場で肥料の効果や、栽培法の効果を検討していた研究者が、それぞれの違いを少ない実験コストで明らかにするニーズがあったといわれています。

現在では実験が伴うあらゆる分野(自然科学分野、社会科学分野)で、統計分析の利用は必須です。

学校教育の現場でも、これを効果的に用いることで、授業効果の判断などに利用できます。

統計分析には、以下のメニューがあります。

- ・平均値、分散値の区間推定
- ・平均値、分散値の仮説検定
- ・分散分析
- ・相関分析
- ・適合度の検定

一方 多変量解析は生じている現象の相関関係や因果関係を数量化して、現象の要因分析や予測の目的に利用されます。以下のようなメニューがあります。

- ・回帰分析
- ・主成分分析
- ・判別分析
- ・因子分析
- ・正準相関分析
- ・数量化手法(I類～IV類)

などです。

2. 統計分析の考え方（その1）

まず、統計分析から説明します。

統計分析は、データから統計量(平均値、分散値)を計算し、前提としている確率分布を評価します。

まず 身近な確率現象について説明します。例えば

- ・サイコロを投げて、偶数の目が出る数を数える

を考えます。

サイコロを振って偶数の目が出る確率は0.5です。（奇数の目が出る確率も0.5です）

実際には、各面のデザインの微妙な違いが、出現確率に相違をもたらすようですが、

ここでは すべての面の出現確率は同じと思われるサイコロで実験することになります。

さて サイコロを10回投げて、偶数の目が出る回数を数えたとします。

正しいサイコロですから、その回数は5が一番多いだろうと予想されます。

一般に生起確率 p (0.0～1.0)の現象を n 回(独立に)試行して、その現象が k 回発生する確率 Pr は

$$Pr[X = k] = \binom{n}{k} p^k (1-p)^{n-k}, \quad \binom{n}{k} = \frac{n!}{(n-k)!k!}$$

と計算されます。この確率分布は 2 項分布と呼ばれます。

上の例、つまり10回試行して、5回発生する確率は $p=0.5$ 、 $n=10$ 、 $k=5$ として、

$$Pr[X = 5] = \binom{10}{5} 0.5^5 (1-0.5)^{10-5}, \quad \binom{10}{5} = \frac{10!}{(10-5)!5!}$$

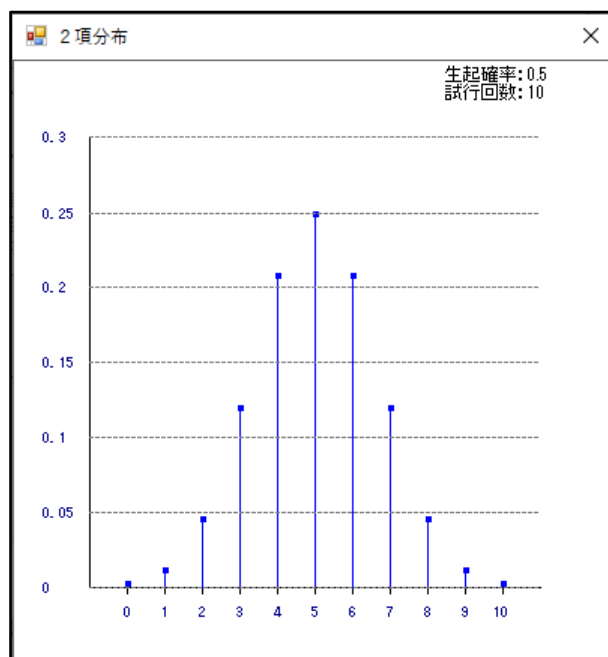
となります。

上式を用いて $X=0, 1, 2, \dots, 10$ の確率 Pr を計算すると下のようになります。

X	0	1	2	3	4	5	6	7	8	9	10	合計
Pr	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001	1.000

表 1. 2 項分布の計算例 （ 生起確率 $p = 0.5$ 、試行回数 $n = 10$ ）

これをグラフに描画すると以下のようです。



この表とグラフを見て

- $X=10$ ($X=0$ と同じ) の時の確率は 0.001 と 非常に小さい

- $X=9$ ($X=1$ と同じ) の時の確率も 0.01 と これも小さい

とわかります。

ですので、 $X=9$ または 10 ($X=0$ または 1 と同じ)が発生したら、非常にまれなことが生じたと考えて間違いありません。

もしこの現象が頻繁に起きるなら、このサイコロは正しくなくて、妙な仕掛けがされている？

と疑うでしょう。偶数の目が出る生起確率 p が0.5であるという前提条件を疑うのです。

では、 $X=5$ ではどうでしょう？ 高い確率で出現するかなと思ったりしますが、表によれば

- $X=5$ の時の確率は 0.246

意外と“小さい”感じがするでしょうか？

ですが、 $X=5$ のまわりの $X=3, 4, 6, 7$ を見てみなすと、

- $X=3, 7$ の時の確率は 0.117

- $X=4, 6$ の時の確率は 0.205

$X=5$ の周辺を含め $X=3\sim 7$ のいずれかが生じる確率を合計すると 0.89 となります。

つまり、 $X=5$ のみのピンポイントの確率を論ずるのでなく、周辺の分布も考慮すればそれなりに高い確率になるんだと 納得していただくことになります。

ここで唐突ですが、有意水準という言葉を紹介します。前提となる確率分布の正否を判断する基準として利用されるもので、通常 0.05 または 0.01 が利用されます。

実際に生じた現象に係る確率を計算し、

- 有意水準より小さければ、前提となる確率分布は正しくない。

- 有意水準以上 ならば 、前提となる確率分布は正しくない とは いえない。

と判断します。ここでは有意水準を0.05とします。

このような基準を設けないと、

- わずか 0.001 の確率でも 0 じゃないんだから、その希少な事象が生じて、

前提条件(生起確率 p が0.5とする)を否定してはいけない。

とか、おかしい議論に傾きます。

これが 声の大きい人の発言だったりすると、ややこしいです。

また、有意水準以下の現象が生じるならば前提条件を疑うわけですが、確実に怪しいというわけでもありません。実際には、前提条件は正しいのに、有意水準以下の確率現象がたまたま生じることも ないわけではない ので、この場合 残念ながら 前提条件は却下されてしまいます。判断ミスと言わざるを得ないのですが、でもその判断ミスの確率は、有意水準以下であることを理解してください。

さて 上記の例で どうも偶数の目が出る確率が大きそうだ、つまり生起確率 $p > 0.5$ ではないか？ という状況の下で 考えることにします。

$X=10$ の確率は0.001で、有意水準0.05よりかなり小さいので、この現象が生じたら、前提条件はおかしい(生起確率 $p=0.5$ はおかしい)と判断します。

$X=10$ の現象が生じた事実を疑うのではなく、生起確率 $p=0.5$ とする前提を疑うわけです。

では $X=9$ の確率は0.01で、有意水準0.05より小さいから、この場合も、同様に確率分布を疑うことになりそうですが、一つ 考慮すべきことがあります。

前提条件が正しいと主張する立場から こんな異議が出ます。

*$X=9$ の確率が0.01で小さいのは仕方ないが、 $X=9$ のみの確率で論じるのは不公平だ。だって $X=5$ の時は周辺の確率を含めていてずるくないか？
それなら $X=9$ の確率に、せめて $X=10$ の確率も 含めて計算してくれ。
それでも 有意水準0.05 より小さいなら 納得するけど。*

という異議です。

この $X=9$ の確率だけでなく、 $X=10$ の確率を含めたものを 有意確率(p 値)といいます。

もう少しフォーマルに言うと、「データから計算された統計量を含め より極端な統計量が観測される確率を合計したもの」が 有意確率(p 値) です。

そこで 前提が正しいと譲らない 立場の方に あえて譲歩し、この提案を受け入れることにします。

$X=9$ の確率は0.01、 $X=10$ の確率は0.001で、合計して0.011。つまり

有意確率は 0.011で、有意水準0.05より小さいですから、生起確率 p が0.5という

前提(仮説)はおかしいことになります。これを 前提の仮説を棄却する と言います。

この場合は、異議を唱えた側の言い分は却下されました。

では $X=8$ が発生した場合はどうでしょう？ $X=8$ の確率は0.044で単独では有意水準

0.05より小さいですが、 $X=8, 9, 10$ の確率を合計すると 0.055 となります。

つまり有意確率=0.055は、有意水準を上回ることになります。

ですので、この場合は仮説を棄却できない、前提条件を否定できません。

異議を唱えた側の言い分がこの場合は認めてもらえることになります。

10回程度の少ない実験なんで $X=8$ が生じても、まーそれほどおかしくないかな . . . と なんとなく思ったりもします。

ここまでは サイコロを10回転がす場合を考えました。 復習しますとこうなります。

・ $X=9, 10$ ($X=0, 1$ も同様)が出現した場合、その 有意確率 は 有意水準 を下回るので、前提の仮説(生起確率 $p=0.5$)を 棄却する。

・ $X=8$ ($X=2$ も同様)が出現した場合、その 有意確率 は 有意水準 を上回るので、前提の仮説(生起確率 $p=0.5$)は 棄却できない。

統計分析はこんな感じで実際に生じた現象から統計量を求め、それにかかわる確率(有意確率)を計算して判断します。

さてもう少しこの話におつきあいください。今度はサイコロを100回転がす場合を考えてみます。

そして80回 偶数だった、つまり $X=80$ が出現したら、どうでしょうか？

先ほどの $n=10, X=8$ の現象を認めたのだから、今回の $n=100, X=80$ も同じ現象だから、前提条件を認めるべきだ。

と考えるでしょうか？ それとも、

$n=10, X=8$ の現象を認めたのは n が小さいかったからで、 n が 100ともなれば様子が違うから、あり得ない話。つまり前提条件は棄却すべきだ。

と考えるでしょうか？ おそらく後者に同意されるでしょう。

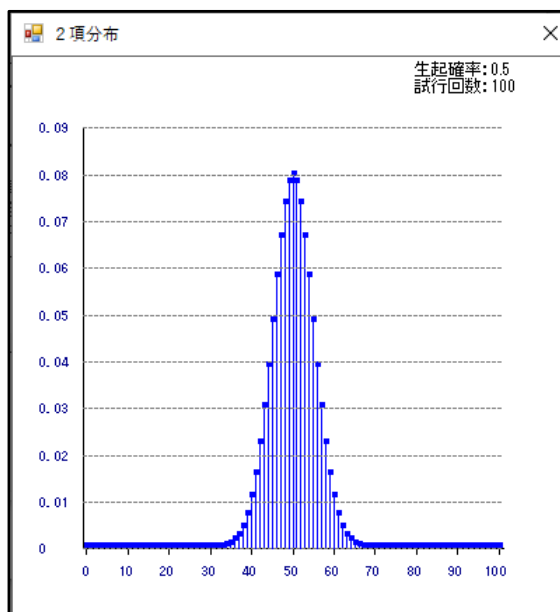
でも、ちゃんと理由づけをしないとイケません。 やはり ここでも確率を計算する必要があります。

試行回数 n が100の場合の、確率を計算すると、以下のようになります。

$n=100$ ですと数が多いので、グループ分けして確率を表示します。

X	0 ~ 9	10 ~ 19	20 ~ 29	30 ~ 39	40 ~ 49	50	51 ~ 60	61 ~ 70	71 ~ 80	81 ~ 90	91 ~ 100	合計
Pr	0.0	0.0	0.0	0.017	0.443	0.080	0.443	0.017	0.0	0.0	0.0	1.000

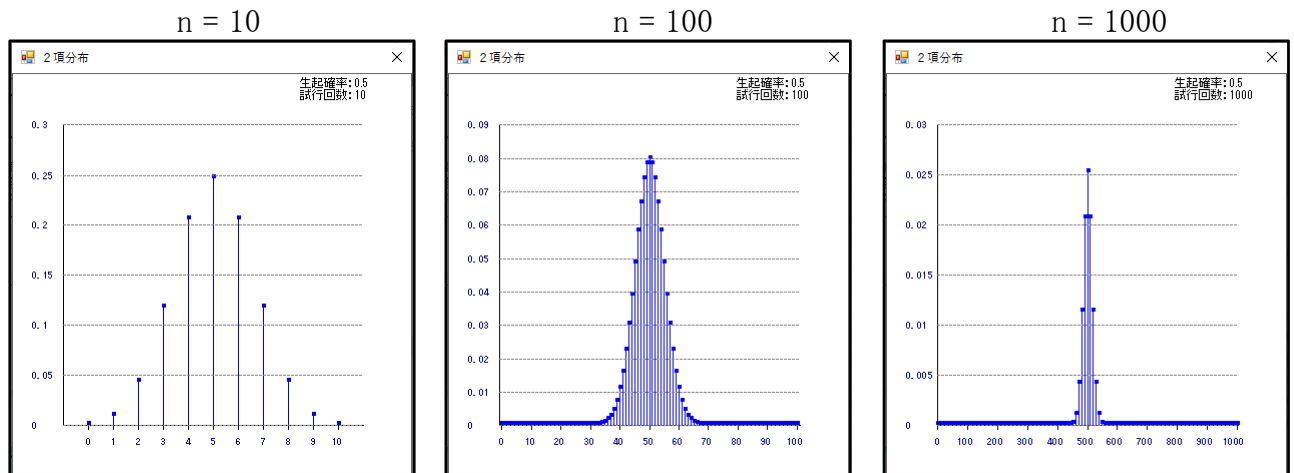
グラフに描画したのが右図です。



これで、 $X=80$ の場合の有意確率($X=80\sim 100$ の確率の合計)を計算すると、ほとんど 0.0 です。
したがって、 $n=100$ 、 $X=80$ が発生した場合、前提の仮説(生起確率 $p=0.5$)は棄却されること
になります。

また、 $X=30\sim 70$ の間の確率を計算すると ほとんど 1.0 です。

グラフを見ると n が大きくなるにつれ、真ん中に集中して尖った形状になるのが判ります。



$n=10$ 、 $X=8$ の場合 と $n=100$ 、 $X=80$ の場合 は、“同じ割合だから似たような状況”と考えて
しまいがちですが、グラフを見てわかるように分布の様子が違うので、仮説の判定は異なってきます。

統計分析では、検討している対象の母集団平均や分散がいくらか、あるいは信頼できる範囲はどうか、
また 複数のグループ間で 統計量が同じなのか異なるのか といった分析を行うのに利用されます。

ここまでのサイコロを転がす例はそれほど日常的でないので、やや関心が薄かったかもしれません。
ですので、もう少し関りがありそうな分野に話題を移します。

ある国家試験では 55 問が出題され、33 問以上正解すれば合格です。
各問は選択形式の 4 択です。つまり 4 つ問題文が提示され、正解は必ず 1 つという決まり
です。

ここで、受験生の正答率がどの程度なら合格? の目安を考えてみます。

まず全く勉強しない学生が、当てずっぽで、例えば四角い鉛筆を転がすとか、4つのくじをランダムに引くとかして回答したとします。

もちろん、難しいと思いますが、合格率はどのくらいでしょう？

この場合 各問についての正答率は 0.25 です。

この学生が 55 問中 33 問を正解する確率は、

$$\Pr[X = 33] = \binom{55}{33} 0.25^{33} (1 - 0.25)^{55-33}, \quad \binom{55}{33} = \frac{55!}{(55-33)!33!}$$

で、これは 先に説明した 2項分布 です。

そして、この時の確率 $\Pr[X = 33]$ はほぼ 0 (正確には $3.14e-08$) です。

しかし、33 問以上正解 が 合格なので、

ならば 34 問、35 問 ~ 55 問 正解 のいずれかに引っかかるのでは

という きわめて薄い期待 を抱いたとして、それぞれの確率 $\Pr[X = 34] \sim \Pr[X = 55]$

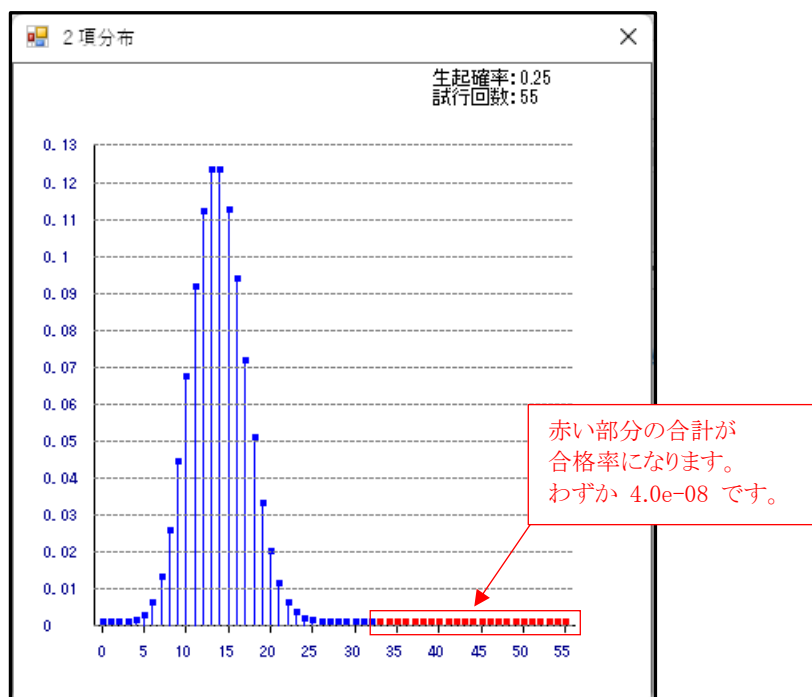
を計算すると、予想通りほとんど0です。(先の $3.14e-08$ より、さらにさらに小さい)

いくら それらを足し合わせたところで 彼の合格率は ほぼ 0 です。

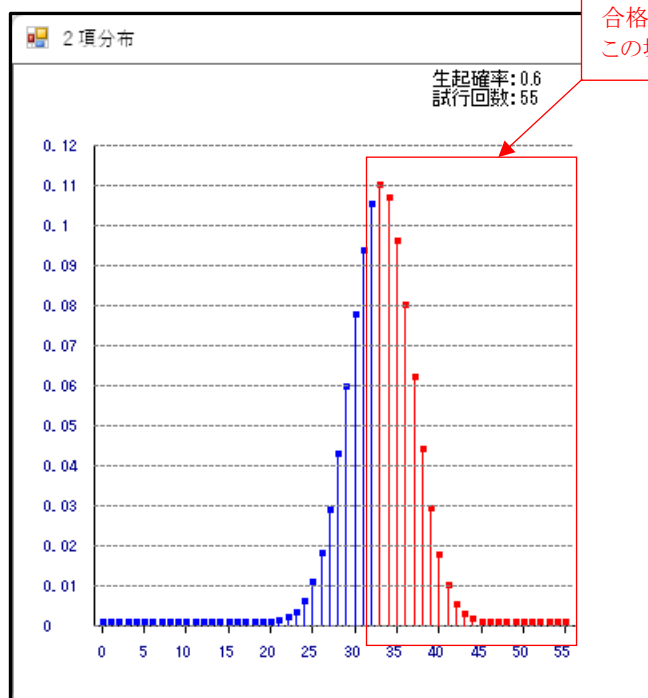
正答率が 0.25 なので、正解数は 13、14 あたりの確率が 一番高いわけですが、

これではどうあがいても、合格に遠いことが明らかなです。 甘くないです。

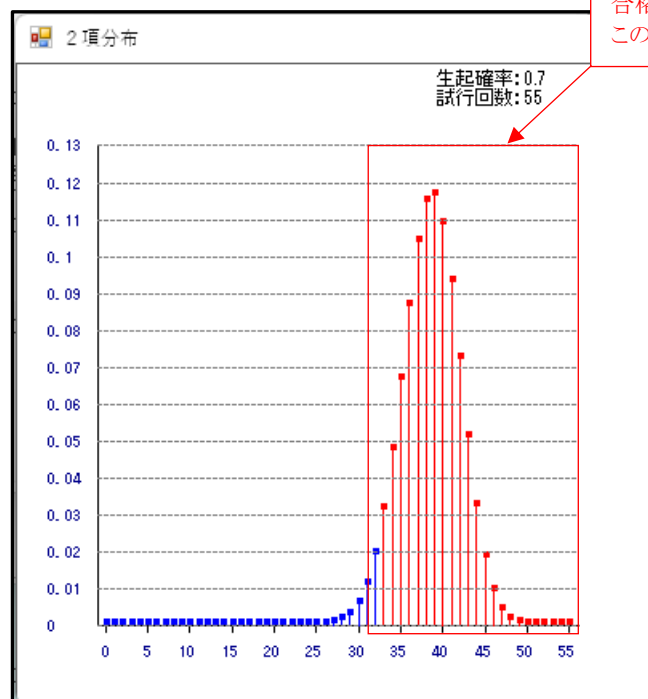
その様子は 以下のグラフで読み取れます。



それでは、(55 問中 33 問を正解が合格なので) 正答率が 0.6 ならば どうでしょう。
その場合の合格率は 0.55823 です。合格？ 不合格？ 微妙です。
勉強の結果、正答率が 0.6 に達したからといっても、合格圏内だと安心することはできません。
その場合の様子は、以下で読み取れます。



では、正答率が 0.7 であればどうでしょう。合格率はアップし、0.95842 と かなり確実になります。
その場合の様子は、以下で読み取れます。



こんなふうに、私たちにとって身近な話題が 2 項分布 で説明出来るのです。

＊) この計算は、メニュー「統計計算→簡易確率計算→2 項分布の確率計算」を選択すると、確認できます。

3. 統計分析の考え方（その2）

ここでは、統計分析手法の一つ、2つの群のデータの平均値の検定について説明します。

最初の説明でも紹介しましたが、2つのクラスの試験結果を例に説明します。

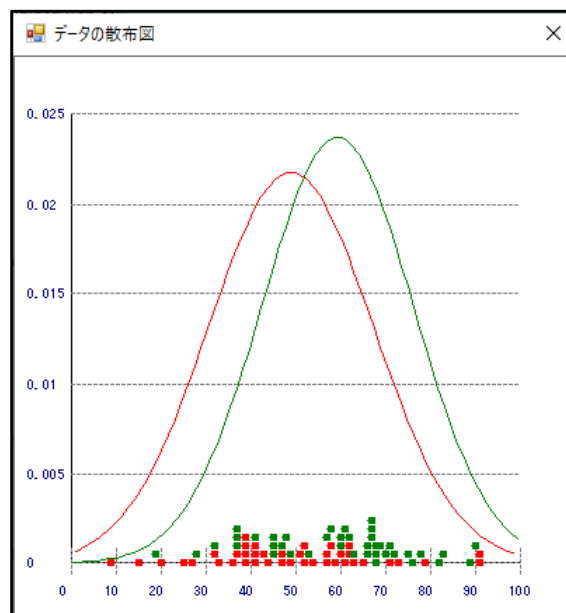
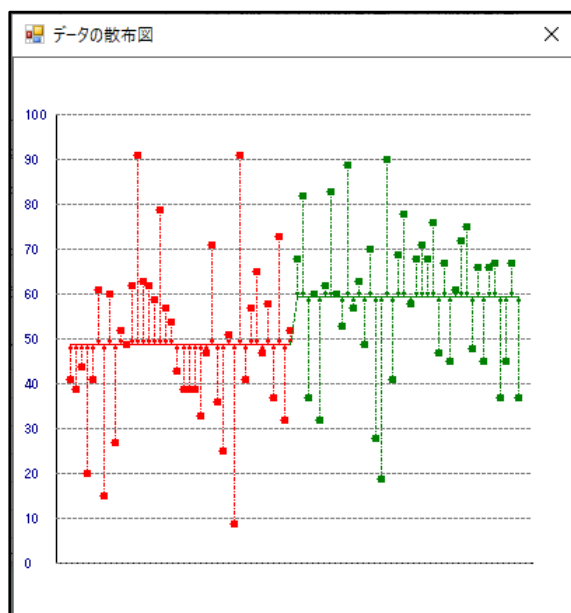
以下の表にそれぞれの点数が表示されています。第1群がAクラス、第2群がBクラスです。

（なお、ここでは両クラス同じデータ数ですが、必ずしも同数である必要はありません。）

第1群		
No	ID	値
1	01-01	41
2	01-02	39
3	01-03	44
...
38	01-38	73
39	01-39	32
40	01-40	52

第2群		
No	ID	値
1	02-01	68
2	02-02	82
3	02-03	37
...
38	02-38	45
39	02-39	67
40	02-40	37

数字だけではわかりづらいので、グラフで評価します。



左のグラフの説明をします。生徒各人の点数が縦軸で、以下のようです。

- ・ 赤は第1群Aクラスの各自の点数を順に表示、赤の横線は平均値を表示しています。
各点から平均値に垂線が引かれています。平均点からどれだけ乖離しているか示すものです。
- ・ 緑は第2群Bクラスの各自の点数を順に表示、緑の横線は平均値を表示しています。
各点から平均値に垂線が引かれています。平均点からどれだけ乖離しているか示すものです。

右のグラフの説明をします。 生徒各人の点数が横軸で、以下のようです。

- ・ 赤は第 1 群Aクラスの各自の点数を表示しています。
- ・ 緑は第 2 群Bクラスの各自の点数を表示しています。
- ・ 同じ数値の点の場合、横軸上で位置が重なり見えづらくなるので、少しずらして表示しています。
- ・ グラフの赤い曲線と緑の曲線は、第 1 群、第 2 群のデータがそれぞれ正規分布からの標本であると想定した時の確率分布を表示したもので、縦軸は確率値です。曲線の頂上の位置は平均値を、山のなだらかさで分散値を表現しています。

グラフから、第 1 群Aクラスの平均値より、第2群 B クラスの平均値が大きいです。

一方、山のなだらかさは、第 1 群と第2群とで大きな違いはありません。（第1群はややなだらかなだらかな山は分散が大きく、急峻な山は分散が小さいです。

もう少し詳しく見てみますと、第 1 群Aクラスには高成績の生徒 (90 点より大) がいる反面、低成績の生徒 (10 点未満) もいます。 一方第2群Bクラスには高成績も低成績もいませんが、平均点は上です。両クラスの生徒の生来のポテンシャルは同じとして、これから推察されるのは、

- ・ 第1群Aクラスの先生の教え方は、高スキルの生徒に力点が置かれている。
できる子は伸びるが、そうでない生徒に対する指導が十分ではない。
- ・ 第2群 B クラスの先生の教え方は、中スキルの生徒に力点が置かれている。
全体的なレベルが維持できているが、すごく優秀な子は第1群Aクラスに比べて少ない。

といったようなことかと思われます。

どちらが いい 悪い の判断は 置いといて、特別な計算をしなくても、データをグラフに表示することで、ある程度 状況を識別することができます。

次に、この 2 群の平均値の検定です。 2 群の平均値は等しいかどうかの判断をします。

第 1 群Aクラス の 平均値は 約 49 で、分散値は 約 335 です。

第 2 群Bクラス の 平均値は 約 59 で、分散値は 約 283 です。

だから 二つの平均値は明らかに違う と 直ちに 判断しないでください。

そもそも データの平均値が ぴったり 一致することなど あり得ません。

この試験データは、生徒の能力値をある時点で測定したものです。

本当の能力値は、試験を繰り返し初めて知ることができるものと考えれば、このデータは、母集団の中の 1 つの標本と考えることができます。

母集団平均値は未知ですが、それぞれは標本平均値とそう変わらず、標本平均値を中心に幅を持って存在するだろうと予想できます。

標本平均値がよっぽど離れていれば別ですが、両者の幅を持った部分に重なりがあれば、標本平均値のピンポイントの値のみで 大小の比較、相違の有無を 単純に 判断するのは早計です。

ここで、

- ・第1群Aクラスのデータの個数を n_1 、第2群 B クラスのデータの個数を n_2
- ・第1群Aクラスの標本の平均値を X_1 、第2群 B クラスの標本の平均値を X_2
- ・第1群Aクラスの標本の不偏分散を U_1^2 、第2群 B クラスの標本の不偏分散を U_2^2
- ・第1群Aクラスの母集団の平均値を μ_1 、第2群 B クラスの母集団の平均値を μ_2

(不偏分散については、「平均、分散」の項をご覧ください。)

としますと、 μ_1 と μ_2 とは未知ですが、 μ_1 は X_1 の近傍、 μ_2 は X_2 の近傍にあると考えられます。

そして、

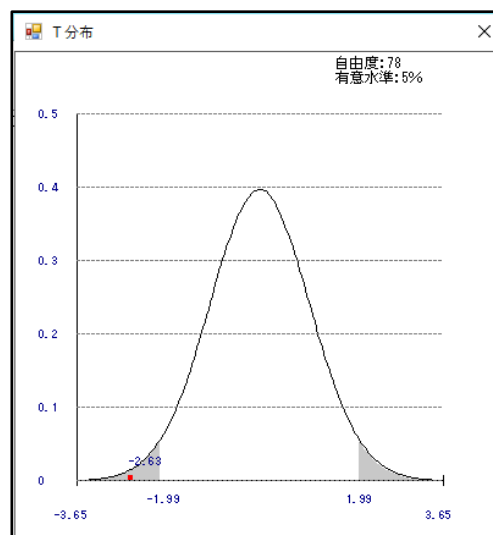
$$\text{統計量 } T = \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) U_{12}}}, \quad U_{12}^2 = \frac{(n_1 - 1)U_1^2 + (n_2 - 1)U_2^2}{n_1 + n_2 - 2}$$

は 自由度 $n_1 + n_2 - 2$ の t 分布に従うことが判っています。これを 利用します。

ここで t 分布についての説明は省略しますが、上記の統計量 T は $\mu_1 = \mu_2$ とすると計算できます。

例題のデータで計算すると $T = -2.6$ となります。また 自由度は 78 となります。

統計量 T は、自由度78の T 分布に従うので、下図の確率分布に従うことになります。



この曲線は確率密度を表現したもので(確率密度関数といいます)、この曲線を $(-\infty, +\infty)$ にわたって積分した値(面積)は 1.0 です。

曲線の左右の領域はグレーで表現されていますが、グレー部分の面積の合計は有意水準の値(通常5%とか1%を指定することが多い) になります。

$\mu_1 = \mu_2$ という前提が正しければ、計算された T は 上図のグレーでない部分 $(-1.99, 1.99)$ に入るはずですが。

一方 計算された T が上図のグレー部分に入るなら、確率的に起こりにくい状況が発生したので、 $\mu_1 = \mu_2$ という前提が怪しいと考えます。T がグレー部分に入った事実を疑うのではありません。

上の計算例では $T = -2.6$ で、グレー部分に入ったので、 $\mu_1 = \mu_2$ とは言えないと結論します。これを 帰無仮説($\mu_1 = \mu_2$ を前提とする)を 棄却する、別の言い方では“有意である”と言います。この言い方は、帰無仮説が棄却されるのは嬉しくて、棄却されないが残念 という雰囲気を伝えるのですが、それは どうでもいいことです。

ここまでの一連の手続きは、メニューの

「統計解析→平均値の検定→2群のデータの平均値の検定(等分散を仮定)」を
を選択することで、実行できます。

平均値の検定

2群のデータの平均値の検定(等分散を仮定) (StudentのT検定)

有意水準 α (%) : 5 ☒ 両側 ☐ 片側検定(左) ☐ 片側検定(右)

計算結果

自由度 : 78

平均値 1 : 49.025 (不偏)分散値 1 : 335.2558

平均値 2 : 59.4 (不偏)分散値 2 : 282.6564

T 値 : -2.639701 帰無仮説の採択域 : (-1.991351 , 1.991351)

P値 (%) : 1.001729

結 果 : 有意 : 帰無仮説(2群の平均値は等しい)を棄却する

☒ 先頭行をラベルとして使用 直接入力可能

第 1 群のデータ 第 2 群のデータ

	NO	ID	値
	36	01-36	58
	37	01-37	37
	38	01-38	73
	39	01-39	32
	40	01-40	52
*			

	NO	ID	値
	1	02-01	68
	2	02-02	82
	3	02-03	37
	4	02-04	60
	5	02-05	32
	6	02-06	62
	7	02-07	82

データを元に 統計的に正しい判断をするには、先に説明したような、
「統計量 T」を計算して「.....」となるわけですが、そのような計算理論を
理解せずとも、とりあえず結果だけを知りたい時に利用できるのが StatToolsCS
システムです。

「計算理論を理解せずとも」と書きましたが、「T値」の意味、「帰無仮説を棄却する」の
意味などを理解していないと、計算結果の評価に苦慮します。

「使用法」ボタンを押すと、チュートリアルが表示されますので、それを参照しながら
結果を評価できるように設計されています。

4. 多変量解析の紹介

StatToolsCS の多変量解析には、

- ・ 回帰分析
- ・ 主成分分析
- ・ 因子分析
- ・ 判別分析
- ・ 正準相関分析
- ・ 数量化 1 類～4 類

などがありますが、ここでは その中の一つ 判別分析を紹介します。

以下を例に説明します。

ある病院の 血液検査の結果です。

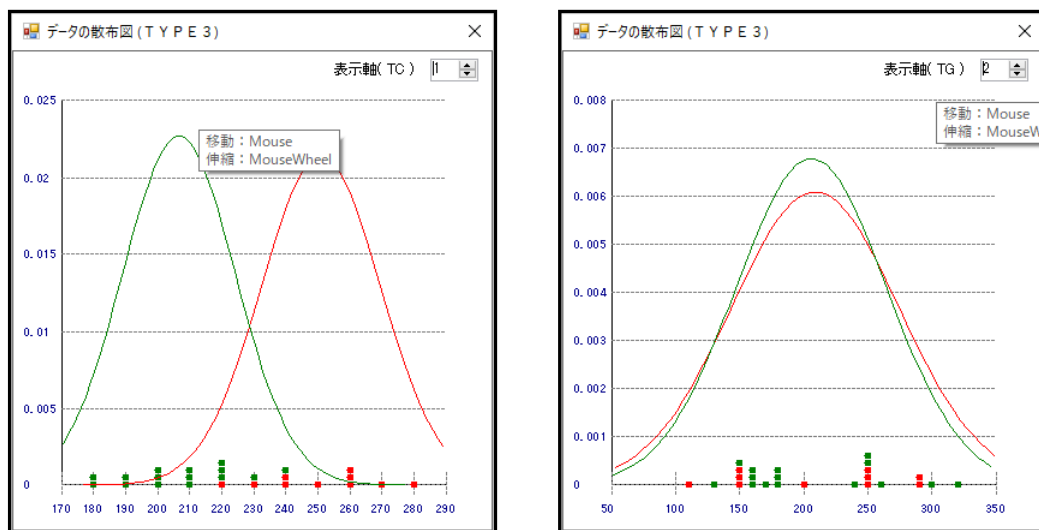
T C（総コレステロール）、T G（中性脂肪）が記述されています。

2 グループに分けていて、左は 動脈硬化症患者 で 右は 正常者です。

動脈硬化症			正常		
ID.	TC	TG	ID.	TC	TG
動脈硬化-1	220	110	正常-1	180	130
動脈硬化-2	230	150	正常-2	180	150
動脈硬化-3	240	150	正常-3	190	160
動脈硬化-4	240	250	正常-4	190	180
動脈硬化-5	250	200	正常-5	200	160
動脈硬化-6	260	150	正常-6	200	170
動脈硬化-7	260	250	正常-7	200	240
動脈硬化-8	260	290	正常-8	210	160
動脈硬化-9	270	250	正常-9	210	180
動脈硬化-10	280	290	正常-10	210	250
			正常-11	220	180
			正常-12	220	260
			正常-13	220	300
			正常-14	230	250
			正常-15	240	320

新たな検査データが与えられたとして、動脈硬化症患者、正常者のどちらに属するのかを判断する という状況を考えます。

T C（総コレステロール）、T G（中性脂肪）のデータを個々に調べると、こんな状況です。



左の図は T C の分布を示しています。

赤は 動脈硬化症患者のデータ、 緑は 正常者 のデータです。

これを見ると、動脈硬化症患者の方が 数値が高いことが分かります。

しかし、真ん中あたりでは、正常者のデータ と かぶっている部分があり、

いくつ以上なら 動脈硬化症患者、 いくつ以下なら 正常者 という 閾値を

明確に 決めかねます。

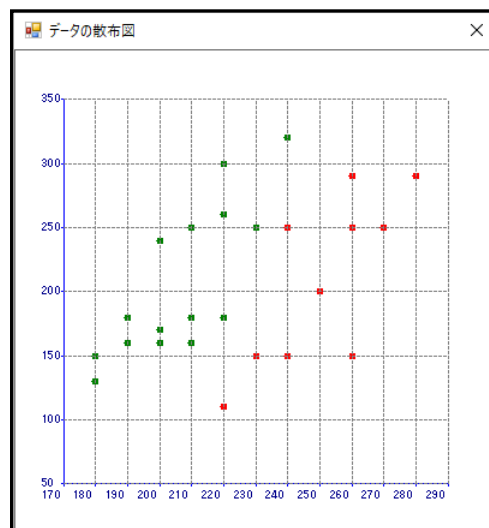
右の図は T G の分布を示しています。

先と同様 赤は 動脈硬化症患者のデータ、 緑は 正常者 のデータです。

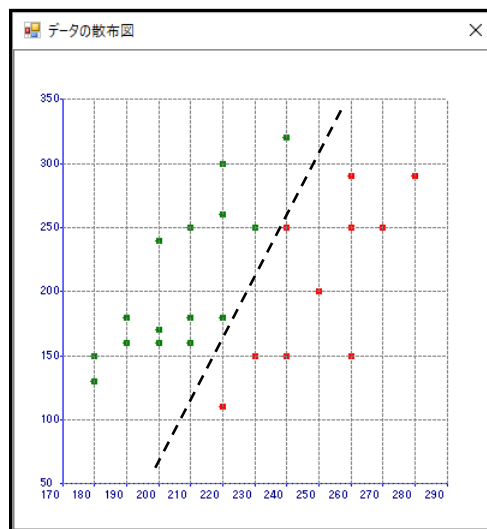
これを見ると、動脈硬化症患者のデータ と 正常者のデータは ほとんど かぶっており、

これでは T Gから識別のための情報は ほとんど得られない と思えてしまいます。

次に、2次元でデータをプロットしてみます。 横軸はT Cの値、縦軸はT Gの値です。

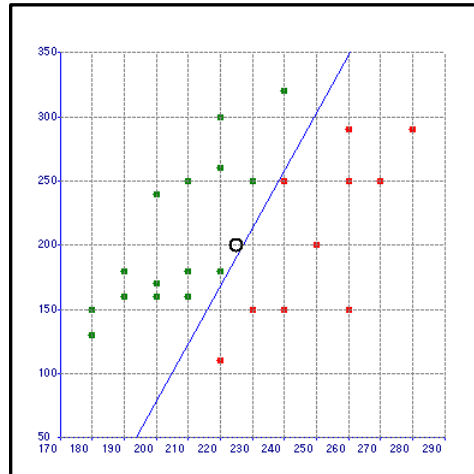


1つの変量で見たときには分からなかったのですが、2次元でデータを見ると、
様子が違って見えます。 例えば、



のように 斜めに線を引くと、 赤の点、緑の点が 分離できそうです。
2変量の 判別分析は このようなアイデアで 分離を図るものです。
データの分布から、最も分離の程度が良い 分離直線（線形判別式と言ってます）
を計算するのが、判別分析です。

新たな 検査データが与えられたとき、それを 上のグラフにプロットして、
分離線 の どちらに 位置するかがわかれば、識別可能です。



変量が3以上になると、ますます 判別は難しくなりますが、判別分析を用いることで、直感的に 判別しやすくなります。

例えば、13人の患者の検診データがあります。(名前等は仮)

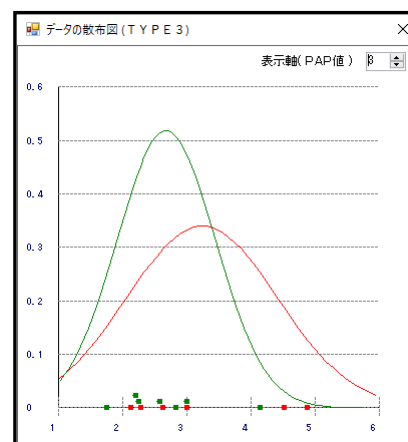
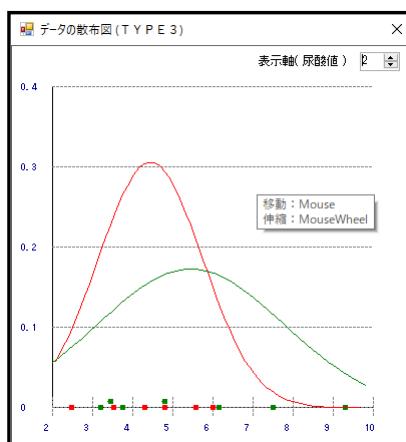
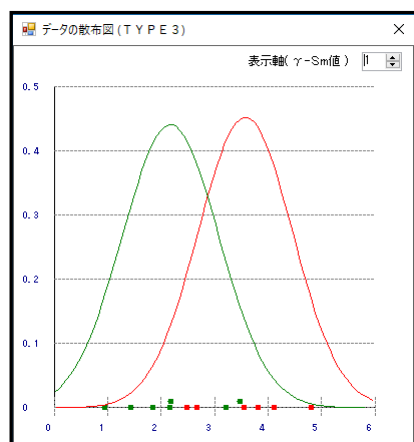
各自の 検診データは γ -Sm 値、尿酸値、PAP 値 の3種類で、

IDの佐藤～鬼頭の6人は前立腺癌を発症しており、土居～瀬尾の7人は前立腺肥大を発症しています。

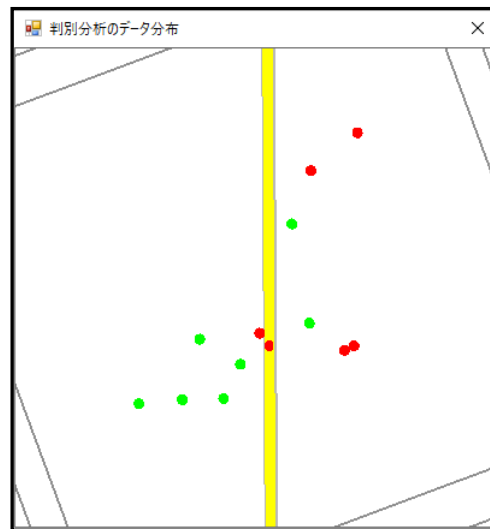
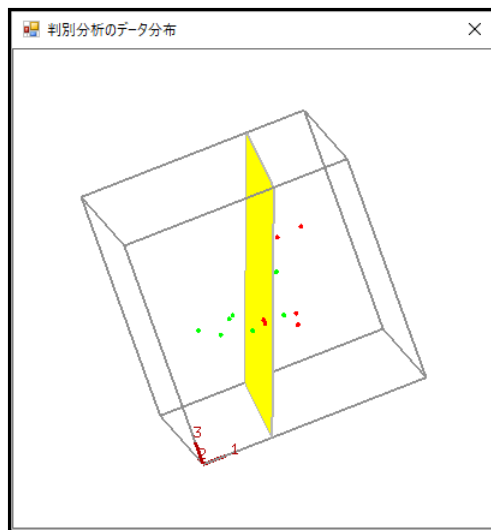
前立腺癌			
ID	γ -Sm 値	尿酸値	PAP 値
佐藤	4.12	6.00	4.52
石井	3.82	5.58	2.13
深井	2.67	4.30	2.64
佐山	3.55	3.55	2.29
尾上	2.49	2.49	3.00
鬼頭	4.81	4.81	4.88

前立腺肥大			
ID	γ -Sm 値	尿酸値	PAP 値
土居	3.21	3.21	2.83
新井	0.95	7.5	2.25
新川	3.47	3.47	4.15
小室	2.16	9.3	1.76
杉山	2.18	3.75	2.59
田中	1.43	6.15	2.21
瀬尾	1.85	4.8	3.01

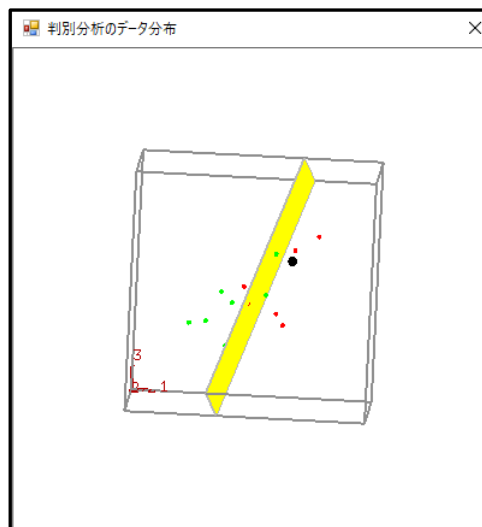
2変量の場合と同様に、1つの変量だけに注目すると、2つを分離するのは難しそうです。



判別計算を行い、3次元空間で判別の為の平面を表示すると、分離の様子を直感的に認識できるようになります。



新たな検査データが与えられたとき、それを上の3Dグラフに表示して、分離平面のどちらに位置するかがわかります。




5. システムのインストールと起動方法

圧縮ファイル DistStatToolsCS.zip を 適当なフォルダにコピーしてください。










名前	更新日時	種類	サイズ
 DistStatToolsCS.zip	2022/09/29 12:47	圧縮 (zip 形式) フォ...	17,146 KB

ファイル DistStatToolsCS.zip をダブルクリックすると、DistStatToolsCS というフォルダが表示されます。

名前	種類
 DistStatToolsCS	ファイル フォルダー

そのフォルダをクリックしてコピーし、しかるべきフォルダで貼り付け（解凍）をすると、“DistStatToolsCS” というフォルダができます。

フォルダ “DistStatToolsCS” をクリックすると、

名前	更新日時	種類
 Documents	2022/09/26 17:29	ファイル フォルダー
 LOG	2022/09/29 10:56	ファイル フォルダー
 sample_data	2022/09/29 11:23	ファイル フォルダー
 dsdIntrMedLibraryRelease.dll	2022/09/29 9:44	アプリケーション拡張
 dsdLibraryRelease.dll	2022/09/29 9:42	アプリケーション拡張
 glut32.dll	2019/11/30 16:13	アプリケーション拡張
 msvcp140.dll	2018/10/01 11:14	アプリケーション拡張
 msvcr120_clr0400.dll	2018/10/01 11:00	アプリケーション拡張
 StatToolsCS.exe	2022/09/29 9:45	アプリケーション

StatToolsCS.exe という1つの実行形式ファイル、5つのdllファイル、Documents と sample_data という2つのフォルダができます。

StatToolsCS.exe をクリック（又はダブルクリック）することで、解析ツールが起動されます。

実行を開始すると、LOG フォルダに計算時のログファイルが書き出されます。

万が一計算時にエラーが発生した場合のログが書かれますが、問題なく作業が終了したならば、中にあるログファイルは削除しても問題ありません。

また sample_data というフォルダには、入力データのサンプルとなる Excel ファイルが格納されています。チュートリアルで説明に利用しているデータが記述されています。