統計分析のすすめ
～とりあえず試してみましょう～

埼玉県総務部統計課　平井　毅

1　はじめに

現状を客観的に認識し、将来に向けた政策を検討するためには、統計データの分析は欠かせません。

しかしながら、特に事務職に多い「文系」出身者には、数式を伴う統計分析を「基本から体系的に」習得するのはやや敷居が高いかもしれません。

（筆者も文系です。数学はかなり苦手です。）

本稿では、「とりあえず」を合言葉に、統計分析に役立ちそうな考え方や方法を、ごく簡単に（なるべく数式を使わず）ご紹介したいと思います。

＜本稿の構成＞
1 はじめに
2 とりあえずやってみましょう
　－「散布図」から「主成分分析」まで －
　(1) グラフにしてみる
　(2) 散布図
　(3) 回帰分析
　(4) 主成分分析（多変量解析の一例）
3 様々な分析など（各論）
　(1) 季節調整 － 時系列分析に役立ちます －
　(2) 人口 － 将来人口の予測 －
　(3) 地図にしてみましょう
4 おわりに

2　とりあえずやってみましょう
－「散布図」から「主成分分析」まで －

（1）グラフにしてみる

統計データは、普通、表として提供されています。

そのままでも色々なことがわかりますが、「とりあえず」グラフにして眺めてみましょう。

グラフには様々な種類がありますが、「とりあえず」の段階では深く考える必要はありません。

エクセルのグラフ機能では、グラフの種類や仕様を簡単に切り替えることができます。

いろいろ試してみることで、分析の手がかりが得られるかもしれません。

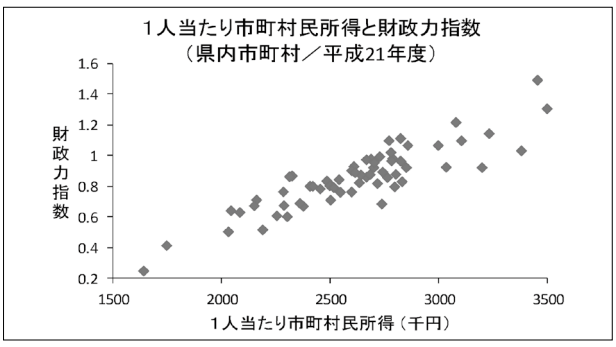
（2）散布図

統計分析では、「複数の項目の関係」をみることがよくあります。

例えば、「「通勤時間」と「女性の就業率」との関係」、「「１人当たり市町村民所得」と「財政力指数」」というような分析です。

こうした「２種類の項目の関係（傾向）」を直感的に見るためには、散布図が便利です。

下はその例です。



（3）回帰分析

①　単回帰分析

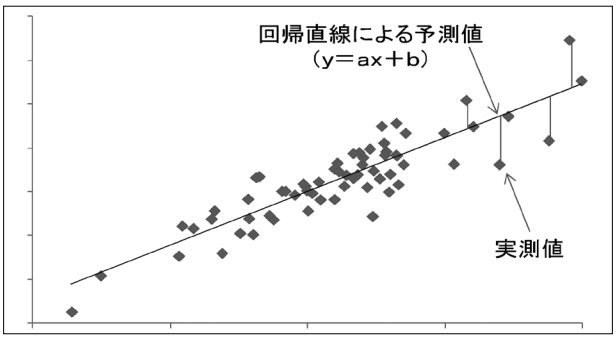
散布図をみて、全体の傾向に最もよくあてはまる直線（回帰直線）を引くことを考えます。

（曲線の方がよくあてはまる場合もありますが、なるべく簡単な例で考えてみます。）

正確さにこだわらなければ、目分量で線を引く、という方法もありますが、ここでは、より客観的な

方法として、「最小２乗法」をみてみましょう。

最小２乗法とは、回帰直線の方程式ｙ＝ａｘ＋ｂについて、方程式から求められるｙ（予測値）と、散布図上の実際の値（実測値）との「誤差」の総和を最小にする、という考え方に基づいています。

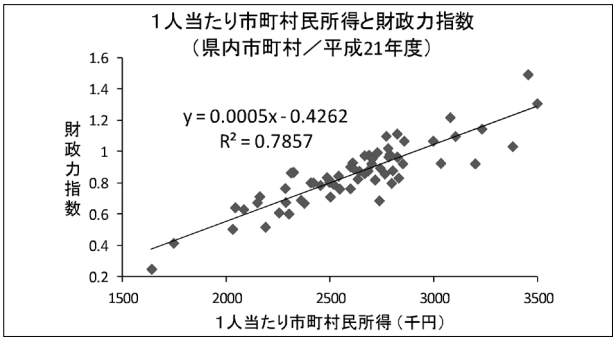


具体的には、実測値と予測値の差の２乗の合計が最小になるようなａ、ｂを求めることになります。

実測値と予測値の差を２乗にする理由
単純に「実測値と予測値の差」とすると、＋の場合もあればーの場合もあり、＋とーが打ち消し合うため、「誤差の総和を最小とする」ための計算としては使えません。
なお、実測値と予測値の差の「絶対値」をとり、その総和を最小にするという方法も考えられますが、計算が複雑になるため、２乗の総和で考える方が簡単です。

これは、「最小値」を与える条件を求める問題なので、手計算では「微分」の知識が必要ですが、実務上は、エクセルの関数、分析ツール、ソルバー（後述）等で求めることができます。

また、エクセルの散布図にも、回帰直線を自動で引く機能があります。（※ 曲線も可能です。）



なお、回帰分析は、２つの項目の相関関係をみるものであり、必ずしも因果関係を示しているものではないことには注意が必要です。

②　重回帰分析

単回帰分析は１つの項目（説明変数）で他の１つの項目を説明するものですが、説明変数が複数ある場合を重回帰分析と呼びます。

単回帰方程式は２次元平面上の「線」として表すことができるのに対し、説明変数が２つの場合の回帰方程式は３次元空間内の「面」としてイメージすることができます。（なお、説明変数が３つ以上の場合、図形としてはイメージ困難ですが、数式としては同様に考えることができます。）

実務的には、エクセルの分析ツールが便利です。

（４）主成分分析（多変量解析の一例）

分析対象の項目（変数）が複数ある場合の統計分析を総じて多変量解析といいます。

多変量解析には様々な種類がありますが、何らかの数学的な方法により「複数の項目（変数）の間に潜む関係」を抽出するという共通点があります。

（(3)②の重回帰分析も多変量解析の一種です。）

ここでは、多変量解析の例として、「主成分分析」について簡単にみてみましょう。

①　主成分分析とは（目的）

例えば、県内の市町村の特徴を比較し、なるべく簡潔に説明することを考えてみましょう。

比較の元になるデータとして、人口増加率、一人当たり所得、生産年齢人口比率、男女別就業率など複数の指標が得られているとします。

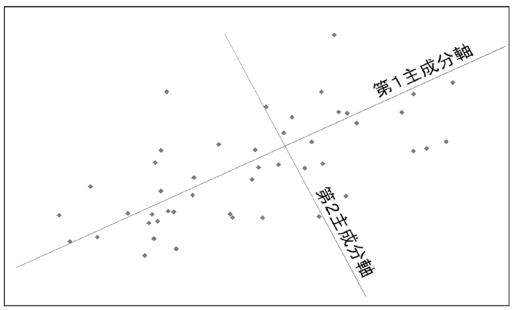
この場合、各指標をそれぞれ比較するのも一つの方法ですが、項目数が多くなるほど、全体的な特徴を把握して説明するのが難しくなります。

主成分分析は、これらの指標を合成した新たな評価軸を作成し、なるべく少ない評価軸でデータを総合的に説明することを目指すものです。

②　直感的な説明（図形的に）

直感的には、以下のように、「データの分散が最も大きくなるように軸を合成（変換）し、新たな評価軸を設定する手法」と理解することができます。

1) まず、ｎ種類の分析項目をもつデータの集まりについて、ｎ次元の空間に広がった「散布図」を考えます。
（４次元以上はイメージが難しいので、とりあえずは３次元で考えてもかまいません。）
その点の広がりを見て、最も分散している方向に軸を通します。この軸を「第１主成分軸」とします。
2) 次に、この第１主成分軸に直交する平面を考えます。
この平面の真上から見ると、第１主成分軸は１個の点に見えますので、この平面に投影されている各点は、第１主成分軸で説明できる要素を除いたものとなります。
この平面の真上から見て点が最も分散している方向に２本目の軸を通します。この軸を「第２主成分軸」とします。



3）さらに、この２本の軸に「直交する」平面（３次元ではイメージ困難ですが、「数学的に」考えられるものとします。）を考えます。2）と同様に、項目の数に応じて軸を順次設定していくことができます。

これらの軸は、性質上、第１主成分軸が最も各データの特徴（違い）の説明力が高く、第２以降の説明力はそれより低くなります。

各主成分を見て、例えば「地域の活性化度」など、わかりやすい名称をつけると説明しやすくなります。
（ただし、解釈できない場合もあります。）

③　数式の説明（ごく簡単に）

ｘ、ｙ、ｚの３種類の変量（項目）がある一連の統計データの例で考えてみます。（何種類でも考え方は同じですが、簡単な例で見てみます。）

以下のような合成変量ｐを考えます。

$p = ax + by + cz$

（ただし、$a^2 + b^2 + c^2 = 1$）※

※ $a^2 + b^2 + c^2 = 1$として、ｐの大きさを制限します。
((a，b，c）をベクトルとすると、長さは１になります。)

ｐは、ｘ、ｙ、ｚにそれぞれａ、ｂ、ｃという重みづけをして変換したものということができます。

このｐを「主成分」と呼び、その「分散」が最大になるようなａ、ｂ、ｃを「主成分負荷量」と呼びます。これらを求めるのが主成分分析です。

＜分散（不偏分散）＞
各標本の値と標本平均との差（偏差）をそれぞれ２乗し、その合計を「標本数－1」（自由度）で割ったもの。（※）
※）なお、標本＝母集団の場合、「標本数」で割ります。
※）分散の平方根は「標準偏差」です。

④　「ソルバー」を利用した主成分分析

主成分を数学的に正攻法で求めるのは、文系にはかなり荷が重い作業ですが、実務上は、エクセルの「ソルバー」により求めることができます。

「ソルバー」とは、「指定された範囲で最適な解（最大値､最小値など）を求める」ツールで、エクセルに標準添付されています。（※ 初期状態ではオフ）

この「ソルバー」を使って、第１主成分を求めてみたのが以下の例です。

県内市町村の比較（※表の一部）

主成分分析 （第1主成分） ※ 各個別データの値は標準化済み

因子	a	b	c	d	e	平方和
主成分負荷量	0.499	0.543	0.531	−0.394	−0.139	1.000

	人口増加率（H17～22） x_1	1人あたり市町村民所得 x_2	財政力指数 x_3	人口当たり保育所定員数 x_4	人口当たり製造品出荷額等 x_5	主成分 p
分散	1	1	1	1	1	2.857
さいたま市	0.75	2.11	0.87	−0.77	−0.58	2.365
川越市	0.48	0.66	1.05	−0.88	−0.05	1.511
熊谷市	−0.23	0.19	0.60	0.68	0.32	−0.002
川口市						1.753
行田						0.55
秩父						0.60
所沢						0.94
飯能						0.17
加須						0.08
本庄						0.20
東松						0.13
春日						0.25
狭山						0.52
羽生						0.603
鴻巣市	−0.07	0.04	−0.17	−0.28	−0.20	0.040
深谷市	−0.36	−0.37	−0.11	1.72	0.80	−1.225
上尾市	0.28	0.30	0.68	−0.57	−0.33	0.935

【ソルバーの設定】
主成分の分散を「目的セル」とし、この値が最大値をとるように「主成分負荷量」の値を変化させる。
制約条件として、平方和＝1 とする。

※ 各個別データは標準化済みです。

※　各項目（変量）の単位やスケールがそろっていないような場合には、あらかじめ標準化した方が良いようです。
標準化後の値＝(各データの値－平均)／標準偏差
→標準化後の平均は０､標準偏差は１になります。
（※ この値を10 倍して50 を足すと「偏差値」）

また、第２主成分以下についても、ほぼ同様に、順次ソルバーで求めることができます。

なお、第２主成分は、第１主成分が取りこぼした情報を対象とするため、各項目（ｘ、ｙ、……）を以下のように変換してから求めます。（第３主成分以下も同様）

ｘ'＝ｘ－ａｐ、ｙ'＝ｙ－ｂｐ、……

（ａ、ｂ……は主成分負荷量、ｐは主成分）

第２主成分（※表の一部）

第２主成分（※表の一部）

因子	a	b	c	d	e	平方和
主成分負荷量	0.075	0.035	0.263	0.163	0.947	1.000

	人口増加率（H17～22）	1人あたり市町村民所得	財政力指数	人口当たり保育所定員数	人口当たり製造品出荷額等	主成分
	x_1	x_2	x_3	x_4	x_5	p
分散	0.28827	0.15873	0.19591	0.55564	0.94435	1.005
さいたま市	−0.43	0.83	−0.38	0.16	−0.25	−0.312
川越市	−0.27	−0.16	0.25	−0.29	0.16	0.142
熊谷市	−0.22	0.19	0.61	0.68	0.32	0.565
川口市	−0.06	0.10	0.12	0.31	−0.25	−0.154
行田市	−0.31	0.06	0.14	−0.11	−0.03	−0.030
秩父市	−0.11	−0.36	−0.04	−0.51	−0.54	−0.624
所沢市	−0.51	0.48	0.35	0.64	−0.40	−0.204
飯能市	−0.29	0.25	0.13	0.22	−0.19	−0.125
加須市	0.24	0.24	−0.05	0.51	0.17	0.254
本庄市	0.67	0.24	0.53	1.86	0.07	0.563
東松山市	−0.45	−0.14	0.25	−0.36	−0.20	−0.224
春日部市	−0.18	−0.34	−0.15	−0.69	−0.58	−0.726
狭山市	−0.76	0.16	0.88	−0.08	1.47	1.558

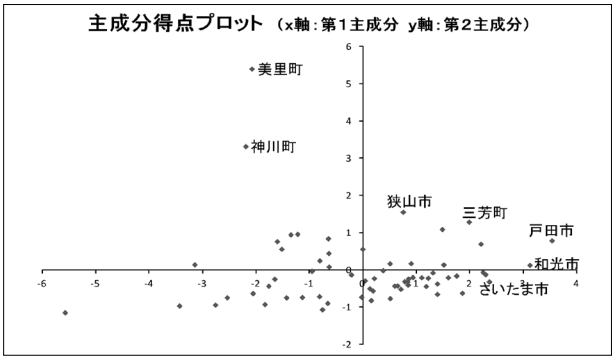
⑤　主成分分析結果の「解釈」

－「主成分負荷量」に着目－

各主成分表の「主成分負荷量」の「絶対値」の大きさにより、その主成分の特徴がわかります。

今回の例では、第１主成分は、「人口増加率」「１人あたり市町村民所得」「財政力指数」の主成分負荷量の絶対値が大きいことから、例えば「まちの発展度」、第２主成分は、「人口当たり製造品出荷額等」の主成分負荷量の絶対値が大きいことから、「工業特化度」などと表現できるかもしれません。

こうした特徴は、グラフからも確認することができます。第１、第２の各主成分（主成分得点）をプロットしたのが下のグラフです。



⑥　主成分分析の「説明力」

－寄与率・累積寄与率－

今回の例では、第１主成分の分散は「２．８５７」です。各項目の分散を合計すると「５」ですから、その57％を占めるということになります。

各項目の「分散」は、その項目の「特徴」とみなすことができますので、第１主成分で、全体の特徴の57％を説明できるということになります。

同様に、第２主成分の分散は「１．００５」ですから、「１．００５／５」で、全体の分散の20％ ということになり、全体の特徴の20％を説明できるということになります。

こうした「主成分の分散」の「各項目の分散の合計」に対する比率を「寄与率」といいます。

また、第ｋ主成分までの寄与率の合計を「累積寄与率」といい、通常、70～80％を超えれば十分とされています。今回の例では、第２主成分までの累積寄与率は57％＋20％＝77％であり、分析としてはまずまずの結果ということができます。

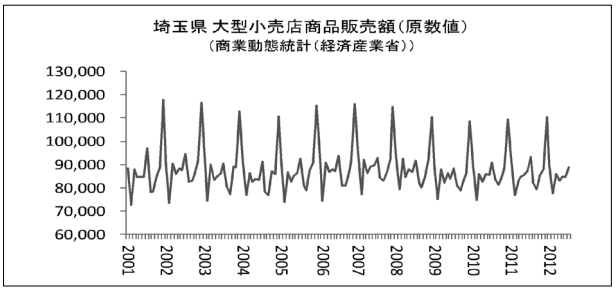
３　様々な分析など（各論）

（１）季節調整　－時系列分析に役立ちます－

①　季節変動

月次の統計データでは、季節的な変動が大きく、前月との比較が難しいようなことがよくあります。

以下はその例です。



②　前年同月比（簡易な季節調整法）

こうした場合、原数値を「前年同月比」に置き換えて比較するのが簡単なので、よく行われています。

ただし、「前年同月比」には短所もあります。

ｉ）短所１：前年の不規則変動の影響

前年同月に何らかの不規則変動があると、前年同

月比の値もその影響を受けてしまいます。

ii）短所２：トレンドの変化に遅れやすい

過去１年の間にトレンドの方向が変化しているような場合、前年同月比にその変化が反映されるのは遅れることになります。

③　より精度の高い季節調整法

より精度の高い季節調整法として、何らかの統計的処理により季節要素を除去することを目的としたツールがいろいろ公開されています。

中でも、アメリカの商務省センサス局が開発し、無償でインターネットに公開している「X－１２－ＡＲＩＭＡ」は広く使われており、埼玉県でも景気動向指数などの作成の際に利用しています。

④　Ｘ－１２－ＡＲＩＭＡの基本的な仕組み（１）

－12 か月周期の変動の除去（移動平均法）－

Ｘ－１２－ＡＲＩＭＡは、その前身の「Ｘ１１」を発展させたもので、Ｘ１１はおおむね次のような「移動平均法」の考え方を基本としています。

1）原数値の「中心化12か月移動平均」
　各月の「６か月前から５か月先までの平均」と「５か月前から６か月先までの平均」の平均を求める。
　→「季節要素」と「不規則要素」が除去され、「トレンド」に相当する数値が得られる。
2) 原数値を上記「トレンド」で割る。
　→「季節要素×不規則要素」が得られる。
3) 上記2）に対し、各月の縦の移動平均（各月について年を串刺しにした平均）を求める。
　→「不規則要素」が除去され、「季節要素」が得られる。
4) 原数値をiiiの「季節要素」で割る。
　→「季節調整値」が得られる。
　（＝「トレンド」×「不規則要素」）

実際には、上記の一連の計算は自動的に行われるため、途中の計算を意識することはありません。

また、繰り返し計算、異常値補正、将来値予測等により、さらに精緻化されています。

⑤　Ｘ－１２－ＡＲＩＭＡの基本的な仕組み（２）

－ 稼働日要因の補正（ＲｅｇＡＲＩＭＡ）－

移動平均法により、12か月周期の季節変動は相当な精度で補正できますが、実務上はもう一つ補正したい要素として「稼働日要因」が残っています。

稼働日要因とは、各月の曜日別の日数、祝日数、うるう年かどうか等で、これらの状況は毎年異なります。稼働日要因は狭義の季節要素ではありませんが、これらの影響を補正することにより、データの傾向がより理解しやすくなります。

Ｘ－１２－ＡＲＩＭＡには、「ＲｅｇＡＲＩＭＡ」という手法で稼働日要因を補正する機能があります。

RegARIMA：Regression（回帰式）＋ ARIMA
ARIMA：Auto Regression Integrated Moving Average
　詳しい説明は省きますが、RegARIMA は、回帰分析と移動平均法の組み合わせを基本とした調整手法です。

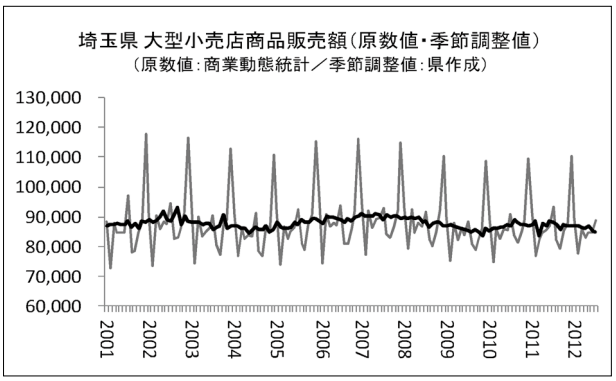
稼働日要因の影響は、分析対象となる統計データにより様々です。Ｘ－１２－ＡＲＩＭＡの利用に当たっては、どの要素を補正の対象とするか、あらかじめ検証し、適切に設定しておく必要があります。

検証の手順をここで詳しく説明する余裕はありませんが、基本的には、考えうる全ての組み合わせについて、統計的なあてはまりの良さを検証し、最も良い組み合わせを採用することとなります。

（実務上は、この検証作業が最も面倒で、時間もかかります。）

⑥　季節調整結果（例）

Ｘ－１２－ＡＲＩＭＡにより、以下のように季節調整値を得ることができます。



⑦　Ｘ－１２－ＡＲＩＭＡの参考書など

Ｘ－１２－ＡＲＩＭＡは便利なツールですが、日

本語のわかりやすい参考書に乏しいのが現状です。

とりあえず、参考になりそうな資料をいくつかご紹介します。必要な方は探してみてください。

・「職業安定業務統計季節調整値の改善について」（労働政策研究・研修機構／２００６／ネット）
　第1章「季節調整の現状」は必読です。

・「季節調整法センサス局法Ｘ−１２−ＡＲＩＭＡの適用における日本型曜日調整の有効性」（奥本佳伸／２００１／ネット）
　若干古い部分もありますが、稼働日要因の検証作業の具体例として参考になります。

・「入門季節調整」（有田帝馬／東洋経済新報社）
　最近出た本です。（２０１２年）

（2）人口− 将来人口の予測 −

①　人口分析の重要性

統計分析では、「人口当たり」や「世帯当たり」の数量を求めるような分析はよく行われます。

また、逆に、１人当たりの平均値に人口を乗じて全体を推計するようなこともよくあります。

人口分析はそれ自体重要ですが、他の分析の基礎となるという意味でも重要ということができます。

②　人口予測の考え方（コーホート法）

人口に関する各種のパラメータ（年齢・性別階層（コーホート）ごとの出生（出産）率、死亡率、人口移動率など）は、通常、急には変化しません。

そのため、各コーホートについてこれらの現在値や将来値を設定すると、将来人口が予測できます。

（各パラメータは、現在値の固定、回帰分析による推測等により、適宜設定します。）

こうした考えに基づく人口予測法を「コーホート法」（コーホート要因法）と呼びます。

コーホート法による人口予測は、人口ピラミッドの各階層が微妙に変化しつつ上にずれていく過程をイメージすると、考え方が理解できます。

③　人口予測の手順

概ね以下のような手順により、エクセル等で比較的簡単に行うことができます。

1) 国勢調査等のデータから、基準となる年の年齢階層別・男女別人口の表を作成します。

2) 市町村別（都道府県別）の生命表から、年齢階層別・男女別の生残率の時系列表（過去〜将来）を作成します。

生命表が存在しない将来の生残率は、何らかの方法で延長推計します。

3) 人口動態統計から、母親の年齢階層別の男女別出生率の時系列表を作成します。

将来の出生率は、何らかの方法で延長推計します。

4) 生残率表と出生率表から、年齢階層別・男女別の封鎖人口（地域間の移動がない場合の人口の理論値）の時系列表（過去〜現在）を作成します。

5) 国勢調査による実際の人口の時系列表と上記の封鎖人口表を比較し、年齢階層別・男女別の人口移動率の時系列表を作成します。

将来の人口移動率は、何らかの方法で延長推計します。

6) 各表を利用して、将来推計人口を順次計算します。

説明では面倒な感じもしますが、各パラメータの予測（エクセルのｔｒｅｎｄ関数等で求められます。）を除き、それぞれの計算は単なる四則計算です。

ぜひ挑戦してみてはいかがでしょうか。

④　コーホート法による人口予測の活用

将来人口予測に関して公的にオーソライズされたデータとしては、国立社会保障・人口問題研究所（社人研）によるものが有名です。

県統計課においても、外部に公表するような重要な分析の際にはこちらを参照していますが、社人研の人口予測は更新まで時間がかかるため、内部的な軽易な分析等に関しては、課内で人口予測ツールを試作して対応している場合もあります。

なお、平成25年３月に、県内市町村用の簡易将

来人口予想ツールを試作し、各市町村（統計担当課）に配布します。可能であれば、ぜひこちらも活用してください。

（3）地図にしてみましょう

①　統計データの地図化

統計データを地図化すると、地域ごとの特性が一目でわかるため、自治体における政策の検討・検証においては大きな効果を発揮します。

ここでは、行政や教育の（一部の）現場で広く利用されている無料ＧＩＳソフト「ＭＡＮＤＡＲＡ」について、簡単にご紹介します。

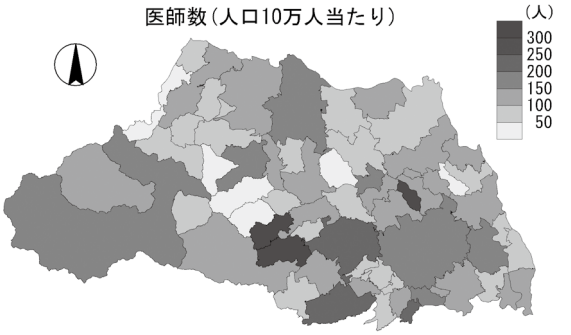
②　無料ＧＩＳソフト「ＭＡＮＤＡＲＡ」

埼玉大学教育学部の 谷 謙二 准教授が作成、公開されている無料のＧＩＳソフトです。

http://ktgis.net/mandara/

＜主な機能＞

・　エクセル等の地域統計データを読み込み、地図化することができます。



・　標準装備の地図のほか、ユーザー自ら地図を作成・編集したり、外部の地図データを取り込んだりして利用することができます。

・　複数の地図を重ねあわせることができます。

・　時間概念を付与することができるため、時系列での切り替え表示もできます。

・　地図の出力形式として一般的な「シェープファイル方式」のほか、ＨＴＭＬ形式（ウェブに対応した形式）や「グーグルアース」の形式に出力することもできます。

③　「ＭＡＮＤＡＲＡ」の活用

地域に関する統計を利用した資料を作成する際には、地図で表示するとわかりやすくなることが多いため、県統計課では、この「ＭＡＮＤＡＲＡ」を各種の資料作成によく使用しています。

なお、平成25年３月に、県内市町村の「小地域（町丁字）区分地図」のＭＡＮＤＡＲＡファイルを試作し、各市町村（統計担当課）に配布します。可能であれば、ぜひこちらも活用してください。

４　おわりに

本稿でご紹介した手法やツールは、多種多様な統計分析手法のごく一部でしかありません。統計分析を本格的に行うためには、体系的に学ぶ方が良いのはもちろんです。

しかしながら、そのことが多くの事務系職員を「統計分析」から遠ざけている面も否めません。

本稿では、主にこうした事務系の職員に向けて、やや（かなり）「怖いもの知らず」の紹介や説明を試みてみました。

厳密さという点ではいろいろ難もありますが、多くは統計課での実例等を元にしており、その意味でそれなりに実績のある内容になっています。

本稿を通じて、少しでも統計分析に興味をもっていただき、政策に活かしていただければ幸いです。