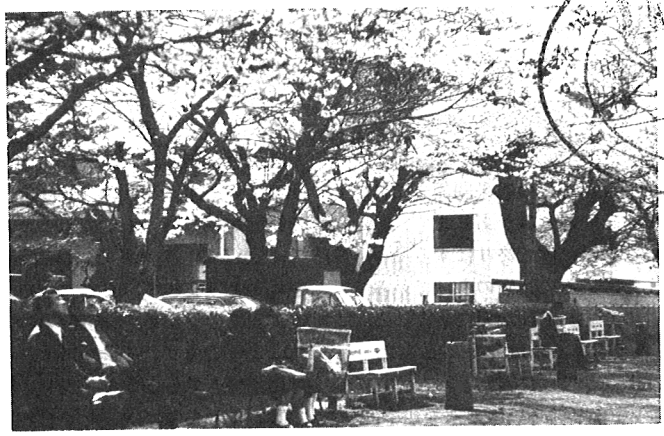


花　見

春らんまん，おだやかな陽気にさそわれて，花見にでかける。さわやかな微風が頬をつたわり，あたり一面から新生のいぶきが感じられ，私たちの気持ちもはずんでくる。

花見と言えば，桜である。しかし，桜をめでることが花見のすべてではない。「酒なくて，何のおのれが桜かな」にもあるように，桜にかこつけて遊び興じることも又，花見にちがいない。そこに桜があるだけで，私たちはなごんだ一時を過ごすことができる。

桜は，その華やかさのなかにも，移ろいやすさを合せもつ。満天をおおっていた花びらが，散り時になると一様に散りはじめ，二・三日で散りつくしてしまう。その散りぎわのあざやかさに驚かされるが，同時に人の世の移ろいやすさを想いおこさせる。そうした移ろいやすいものを惜しむ気持ちが，私たちを花見にかりたてることも事実だろう。

４月のおもな行事

○１日	昭和55年国勢調査茨城県実施本部設置
	全国統計大会茨城分局設置
○１～４日	学校基本調査市町村説明会（水戸市・神栖町・阿見町・下妻市）
○７日	茨城県常住人口調査（３月１日現在）公表予定
	全国統計大会班長会議
○８～11日	学校基本調査高校・特殊学校説明会（水戸市・神栖町・阿見町・下妻市）
○10日	物価指数（水戸市３月）速報公表予定
○10～11日	都道府県統計主管庶務主任者会議（行政管理庁）
○14～18日	学校保健統計調査小・中・高・幼説明会（水戸市・日立市・神栖町・阿見町・下妻市）
○14～15日	関東ブロック庶務主任者会議（神奈川県）
○15日	物価指数（茨城県２月）公表予定
○16日	全国統計大会参事会（関東ブロック統計主管課長，東京都）
○17日	昭和55年度通商産業省全国統計主管課長会議（通産省）
○17～19日	第１回産業連関表研究会（予定）
○18日	国勢調査全国統計主管課長会議（総理府統計局）
○21日	学校基本調査専修学校説明会（水戸市）
○24～25日	茨城県統計調査員研修会（山梨県）
○28～29日	関東ブロック統計主管課長会議（群馬県）

●シリーズ「統計研修から」

統計データの見方・表わし方 (6)

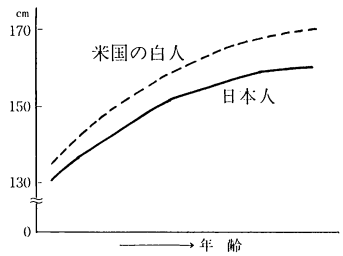
――比率の解釈のための手法――

1．比率の解釈 ――混同要因をめぐって――

統計データを見ていくとき，なんらかの因果関係を頭において，それを説明しようという問題意識があります。数字が何パーセントあったという事実だけで終りにする場合もありますが，そこにとどまらず，その数字のもっている意味の解釈まですすめたいものです。それが面白いところであり，同時に難かしいところでもあるわけです。そこで，今回は，あることを結論づけようと思ったら，どんなデータをもってこなければいけないのか，あるいはどんな比率，あるいはもっと一般的な指標を使わなければいけないのか，これらのことを中心に考えていきます。まず，〔例―1〕をみてください。

〔例―1〕 日本人および米国の白人の学童の年齢別平均身長を国連の統計書から拾ったところ，次の図のようになっていました。

これによって「学童の身長の伸び方は人種によって違う」と結論できますか。



文中の「米国の白人」「国連の統計書」「人種」に気をつけて，例題を考えてください。グラフを見ると，年齢が増えれば身長が伸びていく様子が描いてあります。(実際にはこれほど違いませんが，事例ですから誇張して書いてあります。)グラフを見れば，日本人と米国の白人の身長の伸び方が違うのは明らかです。この場合，データは国連の統計書から取ったとあります。国連では，自ら調査するのでなく，各国の政府にデータを出してもらうわけです。各国とも子供の身長の伸びなどは学校で必ず調査していますから，ここではサンプルが多いとか少ないとかは気にしなくてもよいと思います。

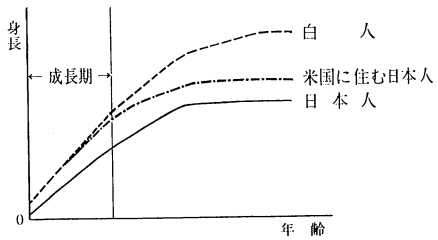
単にグラフを見て，身長の伸び方が違うということで終りにしたのでは，データの読み方としては余りにも単純です。もっと関心をもつ人は，こんなに違うのはなぜなのか議論したくなるはずです。この場合のデータは，一方は「日本人」，他方は「米国に住む白人」です。日本には日本人以外はあまりいませんから，日本の統計は大体「日本人」ということになります。しかし，アメリカのように人種がたくさんいる国では，統計は白人とそれ以外の人種に分けて出しています。そういう意味で，国連の統計書では「白人」の統計が載っているのでしょう。そこで，このデータを見て，『なる程，アメリカには白人以外にもいるが，一応白人を代表とみなして，日本人と白人は違うんだ，つまり人種によって違うんだ』と結論づけたのが，この事例です。『人種によって違う』と言うには，もっと多くの人種について対比すべきでしょうが，説明の本質をつかむために，ここでは2つの人種に限ったのだと了解してください。それにしても，『人種によって違う』という言葉をもってくることには疑問があります。その答はわかっていても，なぜだということを几帳面に説明できなければいけません。図の2つの線をみて，この差を人種による違いと言い切っていいでしょうか。日本人と米国の白人のうち，後者はわざわざ「米国の」(＝米国に住む)と書いてあります。前者も当然，正確には「日本に住む日本人」となります。そうすると，2つの線の違いは，「米国に住む」ことと「日本に住む」ことの違いなのか，それとも「白人」と「日本人」の違いなのか不明です。言い換えれば，地域環境(生活環境)の違いか，人種の違いかわからないということです。どちらかわからないのに，一方だけを言ってはいけません。このデータからは，『人種によって違う』は言い過ぎなのです。統計の専門家として，データから言えることは何かということを判断しなければなりません。

今のデータは，日本に住む日本人のデータと米国に住む白人のデータです。この2つのデータを対比するのですから，ここから考えられることは，日本と米国という地域差(あるいは慣習差)が身長に差をもたらすか，また，日本人と白人の人種差が身長に差をもたらすか，ということです。そうだとすれば，『□によって違う』の□に入れるのは，これだけのデータでは地域差・人種差のどちらとも言えません。これで，この例題の答は一応終りにしてもよいのですが，さらに，どちらによる差かわかるようにするにはどうすればよいかまで話を進めましょう。

そのためにはどうすればよいのでしょうか。この場合，答は簡単だと思います。それは，米国なみの生活習慣をもっている日本人の情報をもってくればよいわけです。生活環境が同じですから，人種による違いかどうかわかります。例えば，米国へ行って暮らしている二世の人がいますから，そういう人の情報をもってくるのです。米国に住む日本人を含めた図―1のようなデータがあれば，答が出てくるのです。つまり，身長が伸びざかりの頃は地域環境(生活環境)の影響が大きく，大人になると遺伝の方が大きな要因として表われるのです。

総理府統計局参事官　上田尚一

図―1　白人と日本人の身長



事例のように，環境の要因と人種の要因という２つの要因が重なっているときは，正しい結論を出せないのです。人種の差を議論する観点にたつと，日本に住んでいる，米国に住んでいるという環境の違いが邪魔になります。逆に，地域環境の影響を議論する観点にたつと，日本人，白人という人種の違いが邪魔になります。一方を議論しようとすると，他方が邪魔になるわけです。このように，議論しようとする目的に対して，邪魔になっている要因がよくあります。それを混同要因と言います。対比しようとする要因に重なっている別の要因のことです。この例題で強調しておきたいのは，統計データをみるとき，混同要因が混じっているかどうかに気をつけないと正しい解釈ができないということです。以下，いくつか例題がありますが，いずれも混同要因に対してどのような対処をするかという事例です。

２．集団の細分とクロス集計

〔例―２〕をみてください。

〔例―２〕「男性ドライバーと女性ドライバーをくらべると，どちらが交通事故を起こしやすいか」という問に答えるため，運転免許者台帳によって選んだサンプル250人について，過去１年に事故を起こしたことのあるものの数を調べたところ，男では200人中50人，女では50人中10人でした。

この結果によって，「男性の方が女性より事故を起こしやすい」と結論してよいでしょうか。

この例題で与えられているデータは，表―1のとおりです。

表―１　男女別の交通事故率

	計	事故あり	事故なし	事故率
男	200 人	50 人	150 人	25 %
女	50	10	40	20

男と女とでは，どちらが事故が多いか調べてみようと誰かが考えたと思ってください。(例題のデータはモデル化してあり，実際のデータはこんな数字ではありません。) この場合，運転の上手・下手を議論するのですから，運転に伴う事故だと考えてください。そうすると，事故率は男25%，女20%ですから，男性ドライバーは女性ドライバーより危険だということになります。しかし，先程の混同要因を考えて解釈すれば，これは正しくありません。例題をよく読むと，ヒントは与えられます。サンプルは「運転免許者台帳から選んだ」と書いてあります。ということは，運転免許者台帳に載っていても，自分では車を持っていない人，持っていてもめったに運転しない人が含まれているわけです。交通事故につながるのは，運転するからであり，運転しなければ事故は起こらないのです。男200人，女50人のデータには，男・女という要因の他に，ペーパードライバーか常に運転している人か，更に，運転している人でも走行距離の長い人か短い距離しか運転しない人か，という要因が混同している可能性があります。そう考えれば，このデータが男女の比較になっているのかどうか少しあやしくなります。男も女も車の運転距離が同等ならよいのですが，そうではありません。可能性としては，男の方が走行距離が長く，女の方は走行距離が短い人が多いと考えられます。これは男と女の比較ではなく，走行距離の長い人と短い人の比較になっているのかも知れません。

この事故率の解釈は２つあります。１つは，男は女より事故率が高いという解釈です。もう１つは，走行距離の長い人が短い人よりも事故率が高いという解釈です。例題のデータからは，このどちらであるかわかりません。まさに２つの要因が混同している可能性があるわけです。一般に，事故を起こす危険率は，危険にさらされている時間または走行距離に比例します。腕前のよい人でも長い時間(距離)走っていれば事故を起こす可能性はあるし，下手な人でも余り走らなければ事故を起こさないはずです。そういう意味では，危険率を議論するとき，人間の属性区分によって答を出そうとしても，話は必ず危険にさらされている時間の長短というものが混同していると考えなければなりません。

事例の答はわかったと思いますが，混同要因があるとすれば，その混同要因も組み合わせた集計をしなければならないということです。統計の集計を行うとき，よくクロス集計というのを行います。そのクロス集計が必要な理由の１つがここにあります。〔例―２〕は，これだけのデータからはどちらとも答は出せませんから，このデータを更に細分化して集計してみるわけです。表―２のような集計が必要になるわけです。

表—2　男女別・走行距離別の交通事故率

	距離	計	事故あり	事故なし	事故率
		人	人	人	%
男	長い				
	短い				
女	長い				
	短い				

統計調査の結果を出すのに，なぜクロス集計が必要なのかは，1つは要因が混同していると解釈がしにくいからです。解釈をするためには，どうしてもクロス集計をしなければならないわけです。言い換えれば，混同要因が予想されるときは必ずクロス集計してみることです。そういう意味で，クロス集計は分析手法とみるべきものです。クロス集計の役割には，もう1つあります。それについては〔例—3〕をみてください。

〔例—3〕　次の表で洋服の買い方区分をクロス集計した効果如何。
〈注〉　収入不詳のサンプルが多いことに注意。

洋服の買い方　　百分比

	収入階級				
	1,000ドル以下	1,000～2,000ドル	2,000～3,000ドル	3,000ドル以上	不詳
オーダーメイド	32	43	49	56	55
レディメイド	49	34	25	15	16
両方	19	23	26	29	29
計	100	100	100	100	100
（実数）	(300)	(800)	(400)	(200)	(300)

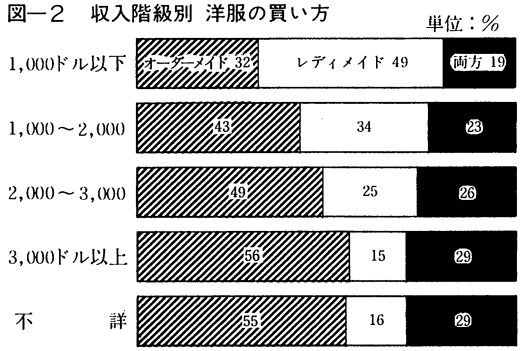
出典：ザイテル『数に語らせる』

よく統計調査でも「不詳」が問題になります。不詳がたくさんあると，データは扱いにくいわけです。不詳をなくすことが望ましいわけですが，調査の種類（世論調査，意識調査，市場調査など）によっては，ある程度の不詳はやむを得ないわけです。そういう調査では，不詳のデータをどう扱うかが大変重要な問題です。不詳を無視して解釈をくだすことはできません。

この場合，（　　）書きにあるように，収入階級が不詳の人が300人，全体の15％もいます。そこで，この不詳という区分に属する人が，収入の高い人か低い人かをわからないと困るわけです。

それを判断するために，たとえば収入階級の高低と関係ありそうな洋服の買い方という情報を組み合わせ集計してみると，不詳の性格がわかります。これをグラフに表わせば，図—2のようになります。

図—2　収入階級別 洋服の買い方



このグラフから，不詳のグループが，3,000ドル以上のグループと大変よく似ているのがわかるわけです。そういう意味で，不詳を判断するテクニックとしても，クロス集計が役に立つということです。なお，この例は，ザイテル著・安田三郎訳『数に語らせる』から引用したものです。この本は，ここで説明している「データの見方」について解説した最初の書物で，発売当時は多くの人に読まれたものです。

3．比率の標準化 ——総合指標——

最後に，〔例—4〕をみてください。

〔例—4〕　次の表にもとづいて，都市と農村の死亡率を比較せよ。

都市と農村の死亡率比較

年齢階級	全国		A市		B村	
	N	D	N	D	N	D
	千人	人	千人	人	千人	人
合計	2,000	76,940	50	1,741	50	2,587
0歳	40	1,200	1	30	1	35
1～4歳	120	240	3	6	3	12
5～19歳	600	3,000	18	75	15	60
20～39歳	750	10,500	20	250	15	180
40～59歳	340	17,000	6	440	10	500
60～79歳	150	45,000	2	940	6	1,800

〈注〉　N＝人口数，D＝年間死亡者数

前例でみたように，混同要因が存在する場合，クロス集計すること，言い換えれば集団を細分してみることが必要となります。このように，集団を細分化することが，論理上必要です。細分すると数字が細かくなってきますが，正しい解釈をだすためには必要なことなのです。〔例－4〕は死亡率の問題です。A市は都市でB村は農村と考えれば，都市と農村の死亡率を比べるとどちらが高いか，という問題意識にたちます。まず，年齢を考えずにトータルで計算すると，

粗死亡率

A 市　$1741^{人} / 50^{千人} = 34.8‰$

B 村　$2587^{人} / 50^{千人} = 51.7‰$

になります。(年齢別の違いを無視したあらっぽい死亡率を粗死亡率と言います。)農村は都市と比べて老人が多いわけですから，農村の粗死亡率が高いのは当たり前です。死亡率というものを考えるとき，年齢は大きな混同要因です。そこで，年齢別に分けて計算しなければならないということになります。年齢別に分けてはじめて，A市とB村の比較ができるわけです。(死亡率を比べるということは，ある意味で地域衛生・環境衛生の良し悪しのモノサシになります。)混同要因を除去するためには，クロス集計して年齢別の比率を出さなければなりません。

ここまではすでに述べたとおりですが，もう1つの考え方として，年齢差が多いために死亡率が高いという影響を除去するため，粗死亡率に修正系数を掛けようという考え方があります。死亡率を計算し直すという考え方から，それを訂正死亡率(あるいは標準化死亡率)と言います。簡単な修正の仕方を1つだけ説明しておきます。まず，年齢別の死亡率を出して，その平均をつくります。ただし，A市とB村の粗死亡率が違う理由は人口の年齢構成が違うために起こった問題ですから，A市・B村の年齢構成ではなく，例えば全国の人口の年齢構成をもってくるのです。現実にそこに住んでいる人の年齢構成ではないものをもってくるという意味では，架空になります。しかし，対比しようとする意図に立っていえば合理的なことです。全国における年齢別人口の割合をもってきて，こういう割合でA市・B村に人が住んでいると仮定すれば，どれだけ死ぬかが計算できるわけです。数字のうえでは年齢別死亡率を出して，それに人口のウェイトを掛けて平均する。ただし，そのときの人口ウェイトとしてある標準の人口(例えば全国の人口)をもってくるということです。そうすれば，A市54.4‰，B村37.6‰という数字が得られます。この違いは，人口の年齢構成は同じですから，死亡率の違いと言えるわけです。

死亡率を計算するには，この標準化という概念が一般化しています。年齢構成が違うと議論しにくいから，それを標準化する(もち論，先程のクロス集計もその基本ですが，別な考え方として標準化する)というテクニックも出てくるわけで，どんな分野でも適用できる基本的な考え方だと言えます。(ここでは触れませんが，物価指数の場合でも同じことをやっています。その場合は，年齢構成でなく購入品目の数量別割合を標準化するわけです。)

表－3　年齢階級別死亡率

年齢階級	全国の人口構成 (1)	A市 D/N (2)	A市 標準化死亡者 (1)×(2)	B村 D/N (3)	B村 標準化死亡者 (1)×(3)
	千人	‰	人	‰	人
0　歳	40	30.0	1,200	35.0	1,400
1～4歳	120	2.0	240	4.0	480
5～19歳	600	4.2	2,520	4.0	2,400
20～39歳	750	12.5	9,375	12.0	9,000
40～59歳	340	73.3	24,922	50.0	17,000
60～79歳	150	470.0	70,500	300.0	45,000
合　計	2,000	—	108,757	—	75,280
平　均	—	54.4 $\left(\frac{108,757^{人}}{2,000^{千人}}\right)$		37.6 $\left(\frac{75,280^{人}}{2,000^{千人}}\right)$	

編集子より；このシリーズは，上田先生が昭和54年3月に総理府統計研修所で講義されたものを収録・編集したものです。上田先生の全講義を紹介することはできませんでしたが，今回でひとまず終わります。御多忙中にもかかわらず御校閲いただいた上田先生には，この紙面を借りて御礼申し上げます。収録者は高野，編集者は斉藤でした。

なお，「統計データの見方・表わし方」の姉妹版として，上田尚一編著『統計グラフの見方使い方』(東洋経済新報社)をおすすめします。

簡単な傾向線計算法……

——バートレット法——

(1) 競馬や競輪(本県にもありますネ)の1着，2着を当てるために，人間はあらゆるデータ……過去の戦績，血統，その日の気象条件に至るまで……を集め，分析し，判断し，かつ祈りにも似た気持ちをこめてお金を賭ける訳ですが，それが必ず当たるという保証は何もないのです。

「予測」というのは，このように無視できない程の大きさの誤差や偏りを持っているものなのです。それでも良いから未来について知りたいという時に，予測の手法が必要となってきます。

予測の手法で最も簡単なのは，**目分量**で直線や曲線の傾向線をあてはめる手法でしょう。しかし，あまり使われていないのは，作成者の経験や力量によって精度に差がでてしまうことと，多少なりとも主観的な要素が含まれてしまうことによるようです。

割にポピュラーな手法なのは，**最小二乗法**を使って直線あるいは曲線の傾向線を数学的に算出するものでしょう。これは，

$$y = a + bt \quad (\text{直線})$$
$$y = a + bt + ct^2 \quad (\text{二次曲線})$$
$$\vdots$$

という関係式を想定し，この関係式の値と各データの値との差を二乗して全期にわたって足し合わせたものが最小になるような係数 a，b，c,……を求める，という手法です。

（最小二乗法については，本誌1977年6月号参照。この手法の具体的な例については，本誌1978年4月号の「時系列の分析(下)」をみてください。）

最小二乗法を使った手法の場合，短期間の傾向をみるには直線の傾向線を算出すればこと足りる場合が多いようです。この場合の計算式は次のとおりです。

$$\begin{cases} \Sigma Y = na + b\Sigma t \\ \Sigma tY = a\Sigma t + b\Sigma t^2 \end{cases} \quad \cdots\cdots ①$$

これでは面倒だという人のために簡便法があります。その方法にはいくつかありますが，ここでは**バートレット法**を紹介しましょう。直線の傾向線の場合有効な方法です。

まずデータの全期間を概略3等分します。その際，必ず両端のグループが奇数のデータ数を含むようにします。両端のグループの平均値(y)の差を両端のグループの中心の年号の差で割り，勾配 b を求めます。a は y の全体の平均から，b とデータの全期間の中心を乗じたものを差し引いて求めます。n はデータ数です。これを式で表わせば次のようになります。

$$\begin{cases} \bar{y} = \sum_{i=1}^{n} \dfrac{y_i}{n} \\ \bar{t} = \sum_{i=1}^{n} \dfrac{t_i}{n} \\ b = \dfrac{y_3 - y_1}{t_3 - t_1} \\ a = \bar{y} - b \cdot \bar{t} \end{cases} \quad \cdots\cdots ②$$

一見すると，こちらの方がむずかしそうですが，実際には計算の手間がずいぶん助かるのです。

(2) 人口推計の場合を例にとって，実際にデータを計算してみましょう。表—1が①の式による場合，表—2が②の式による場合です。どちらが計算しやすいかは一目瞭然でしょう。

表—1　最小二乗法の場合

年次(t)	人口(Ya)	t^2	$t \times Ya$
昭和45年	2,144千人	2,025	96,480
46	2,181	2,116	100,326
47	2,211	2,209	103,917
48	2,250	2,304	108,000
49	2,294	2,401	112,406
50	2,342	2,500	117,100
51	2,378	2,601	121,278
52	2,416	2,704	125,632
53	2,462	2,809	130,486
54	2,508	2,916	135,432
495	23,186	24,585	1,151,057

注) 人口(茨城県)は各年10月1日現在。

①の式に代入して

$$23,186 = 10a + 495b \quad \cdots\cdots ⓐ$$
$$1,151,057 = 495a + 24,585b \quad \cdots\cdots ⓑ$$

ⓑ÷495 − ⓐ÷10

$$6.77 = 0.17b$$
$$\therefore b = 39.82$$

ⓐにこの値を代入して

$$23,186 = 10a + 19,710.9$$
$$\therefore a = 347.51$$

自治研修所　伊　藤　　宰

従ってこの場合の傾向線(直線)は,

$Yb = 347.51 + 39.82\,t$

となる。tを順次代入していけば推計値Ybが得られる。

例えば, 昭和55年(10月1日)の人口推計値は,

$Yb = 347.51 + 39.82 \times 55$

$= 2,537.61$

となる。

表—2　バートレット法の場合

年次 (t)	人口 (Ya) (千人)	平均 (y)
昭和45年	2,144	
46	2,181	$y_1 = 2,178.7$
47	2,211	
48	2,250	
49	2,294	
50	2,342	
51	2,378	
52	2,416	
53	2,462	$y_3 = 2,462.0$
54	2,508	
495	23,186	

注) 人口(茨城県)は各年10月1日現在。

②の式に代入して,

$$\bar{y} = \frac{23,186}{10} = 2,318.60$$

$$\bar{t} = \frac{495}{10} = 49.50$$

$$b = \frac{2,462.00 - 2,178.70}{53 - 46}$$

$$= \frac{283.30}{7}$$

$$= 40.47$$

これらの値を$a = \bar{y} - b \cdot \bar{t}$に代入すると,

$a = 2,318.60 - 49.50 \times 40.47$

$= 2,318.60 - 2,003.27$

$= 315.33$

従ってこの場合は,

$Yc = 315.33 + 40.47t$

となる。

例えば, 昭和55年(10月1日)の人口推計値は,

$Yc = 315.33 + 40.47 \times 55$

$= 2,541.18$

となる。

(3) 最小二乗法, バートレット法いずれの場合でも, 各々の式にtの値を代入していけば, それぞれの推計値Yb, Ycが得られます。これを図示したのが図—1です。もとのデータが直線的なので, 傾向線とほとんど重なってしまいました。

いずれにせよ, 傾向線を出したいが計算が面倒だなと思った時には, このバートレット法を利用するとよいでしょう。

図—1　最小二乗法, バートレット法による人口推計