



算数ができる統計データ分析入門

講演：中央大学名誉教授
統計科学研究所所長 杉山 高一

気温変化などの時系列データ、成分比較のグラフ、アンケート結果の比較グラフ、今年から高校で必修になった「箱ひげ図」など統計にも様々な表現があります。

二つの変量間の関連性について、考えてみましょう。例えば喫煙と肺癌の関連性は有名です。大学入学時の成績と卒業時の成績というのはどうでしょう。こうした関連性は「散布図」に表すことができます。ただ、図で表して関連性があるように見えても、必ずしも関連があるとは限りません。そういう例は世の中にはたくさんあります。因果関係があるかどうかは、統計学の範疇ではないのです。



このような 2 つの指標間の関連性を、統計では「相関係数」で表します。ペアになったデータの関連性の強さを表す方法です。一方のデータが増えると、もう片方も増えるのが正の相関、一方が増える時に、もう片方が減る場合は負の相関があると言います。しかしこの方法で表せるのは直線的な関連だけで、例えば半円のような曲線的な関連は把握できません。

相関係数はゼロと 1、(またはゼロと -1) の間の値をとり、最も強い関連では、散布図の点が一直線上に並びます。相関の強さが「中間」というのはどのあたりでしょうか。(正の相関の場合) 0 と 1 の間の 0.5 ではなくて、2 乗して 0.5 になる値、0.71 辺りが中間であると専門家は説明しています。

では例えば相関係数が 0.7 だ、ということの意味は何かと言えば、それはなかなか難しい問題です。



長期投資仲間通信「インベストライフ」

散布図を眺めて、関連性があると感じられるのが 0.7 辺りから上でしょうか。ただ、この「中間」というのも、0.9 だと主張する人もいて、単純ではありません。

株価もそうですが、時系列のデータから、先を予測することはできるのでしょうか。データの動きの規則性を見つけて把握しようというわけです。ではどうすれば、動きを捉えられるのでしょうか。

そこで、時刻「 t 」での値と $(t-1)$ での値の関係を見ます。時系列で、1 つ手前の値とのペアを作ります。そしてその相関係数を見ます。これが「1次の自己相関係数」です。もし関連性が高ければ、ひとつ前のデータは情報を持っていることになります。次に2 つ手前 $(t-2)$ とペアを作って、同様に相関係数を(2次の自己相関係数)、更に同様に、3 つ手前 $(t-3)$ との関係も見ます(3次の自己相関係数)。このように t と $(t-1)$ 、 $(t-2)$ 、 $(t-3)$ …で予測するとどうなるか、それぞれの相関係数を用いて計算する方法が、「自己回帰」です。相関を上手く使って先を予測する方法です。

試してみると、予測値が実際のデータを、結構きちんと追いかけているように見えます。自己回帰の計算は、最小二乗法、又はユールウォーカー法という呼ばれる方法を用いて計算します。「R」という有名なフリー・ソフトウェアが、世界的に使われています。

講演では、多くの統計の例を挙げて、グラフを見ながら解説していただきました。堅苦しそうなテーマでしたが、ユーモアたっぷりのお話で、楽しく学ぶことができました。

生活に密着した統計学

対談：杉山 高一、馬淵 治好、岡本 和久

岡本| 杉山先生とは、私が 1960 年代の中ごろ、ニューヨークのコロンビア大学というところに留学していたとき、キャンパスで知り合って以来の仲です。当時は日本食を食べたくても、高いし、だいたい日本レストランもそんなになかった。そこで電気炊飯器を持ってきていた杉山先生の寮の部屋に押しかけ、ごはんを炊き、それにお醤油とバターを乗せて食べたりしたものでした。その節はお世話になりました(笑)。

杉山| なつかしいですね。

岡本| 統計というと選挙のときの「当確」の出し方がいろいろな議論を呼びます。あの統計処理はどうのようにされているのですか？



長期投資仲間通信「インベストライフ」

杉山| 統計データ分析の精度は、160万でも1000万でも1億でも変わらないです。ほぼ母集団は無限と考えますからね。僕の所で博士号を取った者が、毎日新聞の調査部で選挙が来る度に猛烈に働いています。杉並区は50万の有権者がいますが、新聞社の取るデータは300です。でも、実際には300だと厳しいですね。ちょっと伯仲している候補者のいる所は、少し無理して500とります。それでもなかなか当たらないものです。新聞社の予想をみると、20とか30とかざれています。一方、出口調査では4000とか5000とか、凄い数をとります。だから誤差が1%くらいなのです。ランダムかどうかは少し不安があるけれど、すごく正確な数字が出ますね。東京都の都知事選が何年かに一回ありますよね。あの時は800取るとかなり正確なところがでます。テレビの視聴率は600ですね。1千万人位の世帯が見ているわけですよね。NHKの大河ドラマが視聴率何%になったと言いますが、誤差がかなりあるなと見ていています。前のデータがあるのでそれを見て補正しながらやっているのでしょうか。残念ながらやり方がインターネットを見ても書いてありません。製薬会社が開発した薬が効くのかどうかというのを、厚生労働省が実際に患者に投与して、そのデータで『確かに効く』という報告書が出れば認可しますが、その時のデータは、1億3000万人の日本人に対して100人です。100人でいいのかという問題はありますが、大体100人ですね。データの数を決めるのは難しいですね。そして、系列もどこまで過去に遡って取ればいいのか難しいです。ところで株価ってよくわからないんですけど、株って値が変わりますよね。

岡本| はい、それは変わりすぎます(笑)。ただ株式の場合には、基本的に企業の持っている本源的な価値というものがあります。1社だけだと潰れてしまうという会社もあるでしょうが、全ての会社を併せてみて見ると、それなりに価値は徐々に増加していきます。価値の増加を反映して、短期で乱高下する株式市場も、長い目で見れば右肩上がりになっています。ここに分散投資と長期投資という二つの大切なルールが成り立つ理由があります。でも、短期株価をみると、思惑、需給関係、心理面によって大きくぶれてしましますよね。

杉山| 平均値は安定しているから、平均値で考えなさいというのは、正しいと思います。

岡本| ただ、時間的に言うと、平均の所に留まっているというのはなかなかなくて、ぶれている時間の方が長いですよね。

杉山| 岡本さんは、株式を平均値の動きでずっと見ていますよね。平均的な所では一つの系列であり、それはそれで面白いかなと思います。



長期投資仲間通信「インベストライフ」

岡本| さっき医学の話がちょっとありました。株式市場だとバー・ローゼンバーグという人は、もともと医学分野の出身で、一つの薬を与えた時の患者毎の薬効がどういう風に違うのかという研究を応用して、ポートフォリオの分析ツールを作ったりしています。株価がどういう要因にどういう反応度を示しているのか、それによってリスクを推計していくという、モデルを作って大成功した人がいます。医学と株式市場はアプローチ的に類似しているところがあるのでしょうね。因果関係の話で言うと、サザエさんの視聴率と株式市場の関係を研究した方もいます。日曜日夜放映のサザエさんの視聴率が高い時は、みんな自宅で食事をしているから景気が悪いのです。景気が良くなると、日曜日は外食するから視聴率が下がるそうです。ですから、サザエさんの視聴率が下がると株価が上がるというわけです。また、馬渓さんの『父ちゃんの立場指数』という有名な指標があります(笑)。デパートの衣料品の売上で、紳士物の売上の伸びと、女物の売上の伸びの違いですね。



馬渓| 紳士物の売り上げは不況の時に、大きく落ちますが、景气回復期の時は紳士物が大きく戻ります。奥さんは、自分の服は不況でも好況でもあまり単価を変えないので。その点、お父さんは、不況の時は圧迫されてお母さんから『あまり服は買わないでね』と言われて減ってしまいます。

岡本| お父さんの立場指数、今の状況はどうなのですか？

馬渓| 今はプラスです。ひと頃よりは、お父さんの立場は回復しています。しかし、紳士物は伸びていますが、伸びが伸び悩んでいます。実はこの指標は、景気の循環を見ようとして作ったものなのです。実際、2002年の始め、ITバブルが崩壊して一番悪かった時にボトムがあったので合致しています。その後2回ボトムがあり、2011年の秋と2012年の夏がボトムでした。2011年秋は、イタリアの財政懸念で株価が悪かったのと、昨年の中頃はスペイン問題で株式市場が低迷した。株価の底と大体一致しているのでそういう意味では面白いかなと思います。ただ最近は株価の勢いが強すぎて父ちゃんの立場が株価についていけないですね。やはりまだ、実際の雇用環境はそんなに改善していない。失業率は低下していますが、所定外労働時間がマイナスです。普通、残業時間は先行指標なのです。まず企業は



長期投資仲間通信「インベストライフ」

従業員をすぐ増やすよりは、今いる従業員に働いてもらい、残業を増加した後から雇用が改善します。今、残業時間がマイナスなのが怖いなというのと、雇用の伸びを見ると、正社員は全部マイナスなのです。パートは増えていますけどね。やはり、まだトータルの労働コストを抑制したくて、雇用は増やすけどパート中心という感じが非常に強く出ています。2002年の景気の回復の時は、パートが先行していたので、景気が回復していると言われながら回復感がなかったですね。今は所得の方の改善をやや欠いたままで、少し気分として浮かび上がっているだけなので、株価の上昇は実体経済と比べると早すぎるなという気がしますね。気分が明るくなって、その結果、更に景気が明るくなることはありうることですが、株価はちょっと行きすぎだと感じます。でもそれは、水準が高すぎるというよりは、スピードとして上がりすぎですね。行き過ぎが更に行き過ぎるということがマーケットではよくあるので、そういう点でこの勢いがまだ続くのかもしれません。しかし、実際との乖離は出ています。

岡本| 先生は、海外の大学でずっと教えられてきましたけど、外国の学生さんは、日本の学生さんと比較すると、早い時期に自分のやりたい分野に対してめどがたっている気がしますがその辺はどうですか？

杉山| アメリカの学生さんは、借金して大学に行きますよね。そういう点の意識は高いでしょうね。自分の将来のことを考えて、自分の借りたお金をどこにどういう風に投資して自分を高めていくのか。日本の場合は、大体親が全部学費を出しますよね。日本も奨学金を利用する学生さんもたくさんいます。今、奨学金は月10万円、有利子で簡単に借りられますよね。4年間で480万円。理系だと大学院に行くのでプラス200万円。ただ大学を卒業する時に700万円位の借金は、ちょっと気の毒ですよね。

岡本| 今、アメリカでは大きな借金をしてようやくMBAをとった学生さんの就職口がなく、返済ができなくて困っています。ちょうどブームがあったのでMBAを持っている人が山ほどいるわけですよ。企業もそうそう雇用を増やす環境ではないから、大学を出ているのに就職口がない。でもローンだけはローンと背中に背負っている状態ですよね。ある意味、高学位の人に対する需給関係が崩れてきてしまっている。例えば中国やインドで大量の高学位の人達が大量生産されてきているわけですよね。学位だけでは就職口がなくなっています。昭和初期の「大学を出たけれど」という時代と全く同じ状態ですよね。国民が少し豊かになるとみんな大学を出るけど、実態経済の方はそんな膨らんでないから就職口がなく余ってしまう。これと同じ現象がグローバルな規模で起きている気がしますね。

杉山| インドというのは、統計学の最先端の国ですね。私が学んでいた1960年代にはどこの大学に行ってもインド人の統計の教授が何人かいいました。統計学はイギリスとインドとアメリカが



長期投資仲間通信「インベストライフ」

盛んですね。日本は、20年前の日本統計学会60周年の時に、様々な企業のいろいろな分野で活躍してもらう為に、4年間学べる統計学科を持つべきだという議論があがりました。しかし実際は、その議論を聞いた韓国の教授が帰国して、今、韓国には統計学科ができます。私は、4~5年前に高麗大学に行って大学院の生徒向けに講演をしてきましたが、半分位は女性の参加者でしたね。統計学は女性に人気があるようです。しかし残念なことに、日本には統計学科がどこの大学にもありません。向こうでは4年間、統計のアンケート調査の作り方等きっちり勉強した後、更にもう2年間勉強して、いろいろな企業で統計学を駆使しているのですけれどね。ただ日本人って賢いから、調査しなくても大体調べてきますけどね。僕たちの頃は韓国と日本だと、比較にならないほどアジアでは日本がトップをいっていましたが、そういう後継者が出てこないというのは厳しいです。レベル的には、日本はもはやトップではないですね。2年前に、マレーシアのクアラルンプールの大学に行きましたが、驚くべきことにそこに統計学科がありました。学生さんは100人位の定員があり、先生方は40人位いて、大学院教育までしっかりやっています。日本にはないのに、マレーシアにこんな立派な統計学科があるのかと印象的でした。アジアの他の国もそうでしょうね。インドは昔から統計学科の最先端の国なので、インド人の統計学者はすごく多いです。ただ最近のインドは情報系の人が多いですよね。今は統計から情報系にシフトしていくのかもしれませんね。

岡本| 最近は統計学の本が良く売れたりしているようですが、このようなことをきっかけに日本でも統計の大切さがもっと認知され、活用されるようになることが必要ですね。益々のご活躍を祈っています。本日は大変多岐に渡るお話をありがとうございました。