

身近で活用できるデータサイエンスとは

京都府中小企業特別技術指導員／(有)坂井経営技術研究所 代表取締役 坂井 公一

最近、データサイエンスという言葉をよく見聞しますが、ICT分野の先進的な企業が扱っている印象があり、中小企業さんには縁遠い感じがするのではないのでしょうか。データサイエンスとは以下の技術を総称します。

- 数理統計学を基礎とする高度なデータ分析技術
- 経営やマーケティングの知見と合わせてビジネスに生かすデータ処理技術

具体的には、Excelにより、重回帰分析、クラスター分析、ロジスティック回帰分析、自己相関関数分析、 χ^2 検定、マハラノビス距離分析、最適化分析などが実行可能です（一部はアドインソフトを必要とします）。

1. 自己相関関数を活用した機械部品の需要予測

以下に自己相関関数がある機械部品の需要予測に活用した事例を示します。営業が注文を取ってくる高額の商品と異なり、商社、代理店、ホームセンターなどで販売される汎用機械部品の販売については、発注数の情報を1月以上前から得ることは難しく、しかも月ごとで2倍以上の差がある場合、何らかの周期性を発見し、あらかじめ精度の高い予測ができれば人員計画や素材発注を効率的にできます。

図1は、ある企業さんの2008年6月から2013年6月までの5年間の出荷個数をグラフ化したものです。（2008年から2009年はリーマンショックの影響で市況が全般に悪化している）

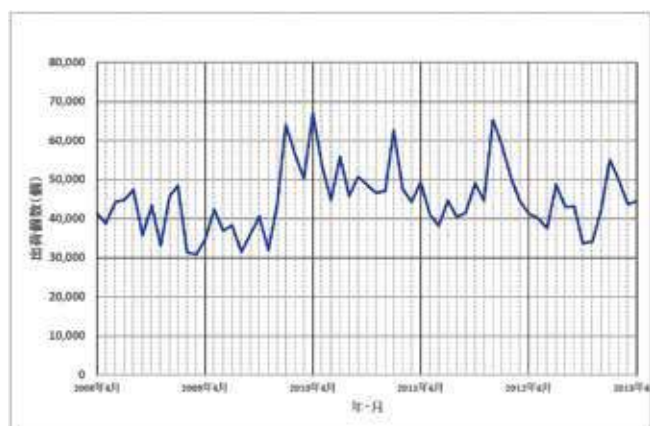


図1. 2008年6月から2013年6月までの出荷実績

月ごとで見ると、3万個/月から7万個弱/月まで、大きく変動していることがわかります。厳しいコスト要求と納期遵守の中で、工数を1分単位で管理し、製品や原材料の在庫を極小に抑える経営努力をしている企業さんにはこの月ごとの大きな受注変動は経営の大きなリスクになっています。

自己相関関数はある時系列データとそのデータ自身にラグタイム（遅れ時間）を設けたデータとの相関を分析するもので、データに含まれる周期性を検出するのに有効な手法です。

Excelで演算する場合はデータをずらしながら、元データとの相関係数を演算し、横軸をラグタイム（遅れ時間）、縦軸を相関係数としてグラフ化します。図2に自己相関係数の演算結果を示します。

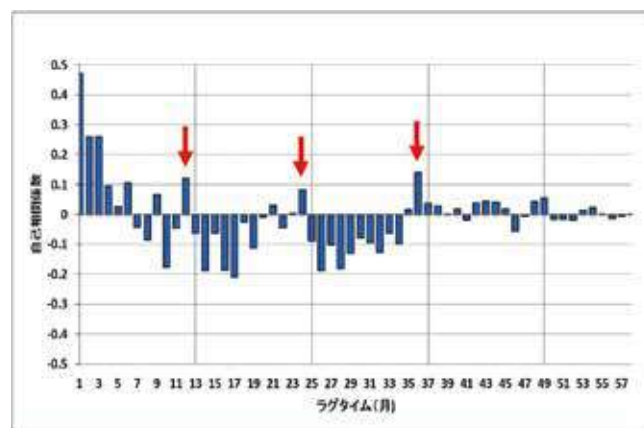


図2. 2008年6月から5年間の自己相関係数

赤い矢印で示した12月、24月、36月の箇所に正の相関が顕著に出ており、結局12か月の基本周期であることがわかります。

基本周期が判明すると、それに基づき以後の予測数を計算することができます。本例では2008年6月から2013年6月までのデータに対して移動平均法を用いて算出した2014年の予測値と2014年の実績データを比較し、その結果を図3に示します。

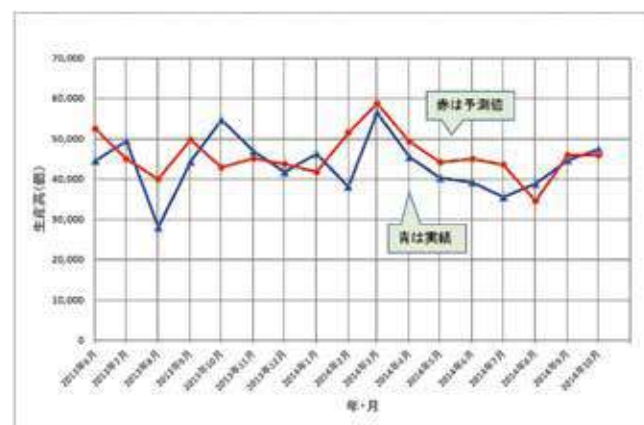


図3. 過去データから求めた予測値と実績値

このようにして、集積している過去データから今後の生産計画を立てるのに有効な予測データを求めることができます。

データサイエンスは開発から販売まで、業種や専門を問わず、あらゆる場面で有効に使えます。

2. ベイズ統計による予測分析

最近の医療技術の進歩により、以前は初期では発見が難しかった病気も新たな試験法で発見できるようになり早期治療が可能になってきました。最新の血液検査法で、ある特定の病気にかかっている人は95%の確率で陽性と正しく診断されるが、治験データではその病気にかかっていない人も5%の率で陽性と誤診断される

リスクがあると説明を受けたとします。また同年代の人がその病気にかかる比率は10,000人当たり100人のデータがあるとします。

あなたが2週間後に受け取った血液検査の結果は、その病気について「陽性、要精密検査」と書かれていました。もしこのような結果を受け取れば、あなたは大変落ち込むと思いますが、実際に精密検査で病気と診断される確率はベイズ理論で以下のような数式で求めることができます。

まずベイズ理論では、結果(D)とその仮定(H)について以下の式が基本となっています。

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

上式でP(H|D)は、データDが得られた時の仮定(原因)がHで有ることを示し、P(H|D)はデータDの原因の確率と考えられます。

データDが二つの仮説H₁とH₂から生まれる場合には、ベイズの定理は次式に展開できます。

$$P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D|H_1)P(H_1)+P(D|H_2)P(H_2)}$$

ここで、D、H₁、H₂は以下を意味します。

D:血液検査でこの病気にかかっている(陽性)と判定される

H₁:この病気にかかっている

H₂:この病気にはかかっていない

この定義からP(D|H₁)とP(D|H₂)は以下の数字になります。

- P(D|H₁)は陽性と判断され、実際に病気にかかっている人で、P(D|H₁)=0.95
- P(D|H₂)は陽性と判断されたが、実際には病気にかかっていない人で、P(D|H₂)=0.05

求めたいのは、陽性と判断され、実際に病気にかかっている人の率P(H₁|D)であり、上式にそれぞれの数字を入れて計算すると以下のように計算できます。

$$P(H_1|D) = \frac{0.01 \times 0.95}{0.95 \times 0.01 + 0.05 \times 0.99} = 0.161$$

これより実際に病気にかかっている可能性は約16%と求まります。

最初に説明を受けた「病気にかかっている人は95%の確率で陽性と診断される」という数字と16%の数字には大きな差があります。もちろん精密検査を受けないといけないのですが、95%と16%の差は気分的に大きな違いがあります。

この理論を提唱したのはイギリスの牧師であり、数学者であったThomas Bayes(1710-1761)で、ベイズ統計学の始祖です。品質管理に広く用いられる「有意水準5%」などの厳密な統計学とは異なり、柔軟な適用性を持っていることが特徴です。インターネットの世界では電子メールで特定の単語を抽出して迷惑メールをはじき出すスパムフィルターや、現在の検索サイトの高いヒット率を支える基本理論に使われています。またネット通販でおなじみの、購入履歴や閲覧履歴からのレコメンデーション機能として応用されています。

3. マハラノビス距離による判別分析法

新たに取得したサンプルについて、すでにグルーピングされた集団のどれに最も近いかを判別する手法にマハラノビス距離法があります。これは単位系の異なる多次元のパラメータであっても、ばらつきを含んで既知のグループとの距離を比較し、どのグルー

プに属するか決める手法です。分散を加味するので正規化ユークリッド距離とも呼ばれます。

簡単な例で説明すると、あるスポーツ用品の発売前に対価を払って男女合わせて21名のモニターを募り、デザイン、機能、希望販売価格など、多くの情報を収集したとします。匿名であっても個人情報保護を明記の上で、アンケートの最初に身長、体重、性別や年代の記載を求めたとして、集計すると1名だけ性別が未記載のアンケートが見つかりました。21名の身長と体重のデータを散布図に記載すると図4が得られました。

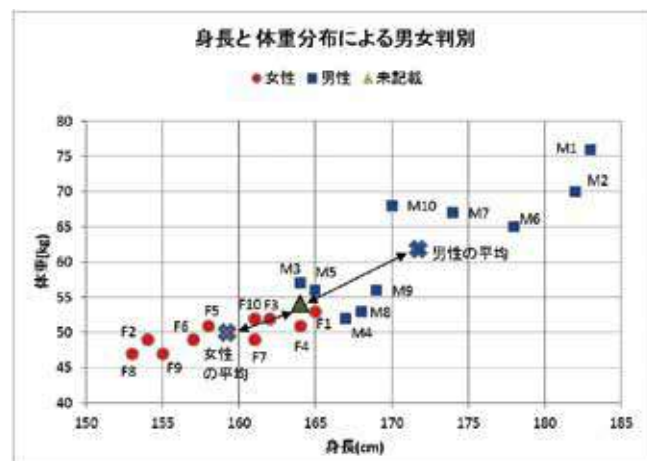


図4. 男女21名の身長と体重の分布図

図4では女性10名は●F、男性10名は■M、未記載の1名は▲で表します。図中の★マークは男性の、★マークは女性の身長、体重の平均値を示します。それぞれの平均値と性別未記入の▲との距離を両矢印↔で示します。

男性の身長、体重の平均値は172cm、62kg、女性の身長、体重の平均値は159cm、50kgですが、それらの分散を含んでマハラノビス距離を計算すると、男性の平均値との距離は1.83、女性の平均値との距離は2.36と求まり、男性のグループとの距離が近いと判断されます。図4の見かけの距離は女性の平均値との距離が近く見えますが、分散を含んで計算すると距離が遠く見える男性グループの方に分類される結果となりました。

上記は2次元の簡単な例を示しましたが、この手法は新規に取得されたデータが、多次元で表現される既存の多数のグループのどれに最も近いかを統計的に求める、判別分析の有効な手法の一つです。

終わりに

企業の開発、生産、販売現場には多くのデータが蓄積しているはずですが、日常業務で収集したデータには大きな付加価値があり、その多くは汎用の表計算ソフトExcelで高度な分析ができますので、是非データの有効な活用を期待します。



坂井 公一 氏

昭和49年同志社大学大学院 工学研究科電気工学専攻修士課程終了。同年松下電器産業株式会社(現パナソニック株式会社)入社。平成17年上記を退職、翌年(有)坂井経営技術研究所設立。平成19年同志社大学理工学部生命医科学部嘱託講師。平成31年上記を退職。