データ分析による異常検知と発見的統制

Forensics事業部　西日本Forensicsグループ
公認情報システム監査人・システム監査技術者　西原則晶

▶ Noriaki Nishihara
ITベンダーで大規模ミッション・クリティカル・システムやSDKの研究開発に従事した後、2006年にEY入所。IT監査、CAAT、内部統制評価等に従事、14年よりForensics事業部にてデータ分析を活用した不正調査や会計監査などを支援。20年より西日本におけるForensicsサービスの提供を開始。情報システムコントロール協会（ISACA）大阪支部における理事、常務理事を歴任。

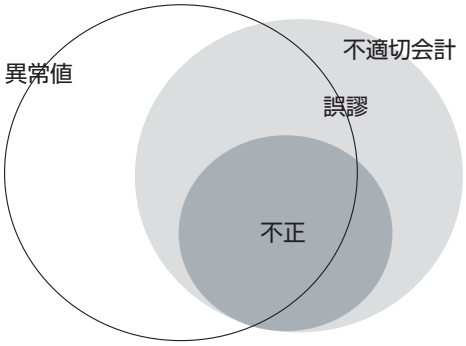
Ⅰ　はじめに

EY Forensicsはこれまで数多くの会計不正や品質偽装といった不正調査実務に携わっており、その際、データ分析を活用することで類似事案や件外事案を発見するということが多々あります。これは企業内に蓄積された多量のデータを分析し、通常とは異なる特徴を持つデータを見つける異常検知として行われています。＜図1＞に示されるように 、データ分析により検知される異常と実際の不正とは一致しておらず、現実には異常値が実際に不正であったかどうかについて、詳細な調査を行った上で判断することになります。ここでまず認識すべきことは、異常値の全てが不正であるというわけではないということです。

また、データ分析をモニタリング実務に取り入れる際には、まず発見的統制として構築していくことになります。異常を検知するためには、事後に他のデータと比較することで異常であるかどうかを判断することになるからです。

本シリーズでは、ファイナンス部門のDX／ファイナンス領域におけるデジタルの活用ポイントについて論述しています。本稿は、不正調査を主たる業務としているEY Forensicsによる前・後編の前編として、大量データ分析による異常検知のためにファイナンス部門のデータをどのように分析すべきか、そしてどのように発見的統制として活用していくのか、という点について解説します。

▶図1　異常値と不適切会計・不正の関係



出典：荒張 健 「データアナリティクスを活用した不正リスクモニタリング」企業会計 2021 Vol.73 No.10

Ⅱ　利用できるデータ

データ分析で利用するデータは電子データであることが前提となりますが、その電子データにもリレーショナル・データベース※に格納されているような構造化データもあれば、メール本文の日本語や電子帳票のような非構造化データもあります。一般的に非構造化データの活用は構造化データに比して難易度が高くなります。

企業内の構造化データを考えた際に、まず思い付くのは会計システムに含まれる会計データや業務システムのトランザクションデータではないでしょうか。モニタリング目的のデータ分析ではより上流である業務

※　関係データベースやRDBと表記されることもある。複数の項目を持つ多数のデータを列と行で構成された表形式で格納し、それぞれの表を項目で関連付けて管理するデータベースのこと。

システムデータの方が高度な分析ができそうな印象がありますが、実際には「データの質」により分析性能は変わってきます。例えば、経費の支払先をマスターと連携した支払先コードとして入力・保持されているデータと、摘要欄にフリーテキストで入力されているデータでは、支払先による分析の難易度は大きく異なります。また金額や単価においても、個数を「一式」として1、単価を支払総額、という取引実態を正しく反映していないデータでは、単価を使った分析はできないでしょう。

このような場合、データ分析に先駆けて業務システムや会計システムに関連する内部統制やIT統制、システム利用手順などの見直しを行うことは珍しいことではなく、データ分析の導入による重要な副次効果とも言えます。

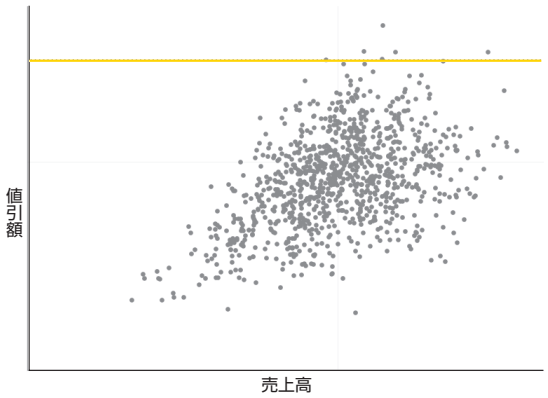
Ⅲ　何を「異常」とするのか

データ分析による異常検知で最も難しいのは、「何を異常とみなすのか」という一言に尽きます。特に不正発見を目的とした異常検知の場合、本当に見つけたい不正に起因する異常はレアケースであり「異常とは○○である」と言えるほどの事例を集めることができません。

そこで一般的には、「想定される通常と比較して異なるデータを異常とする（リスクシナリオ的）」「他の多くのデータと比較して異なるデータを異常とする（データ中心的）」という2つのアプローチにより異常を識別することになります。

前者のリスクシナリオ的アプローチの場合、業務内容や内部統制、ITシステムなどを理解した上で、正常な業務がどのようにデータに反映されるのかを検討します。例えば、「承認行為は申請者の上席者により行われる」という内部統制が有効に機能している場合、データ上の申請者IDと承認者IDは異なる値が入っているでしょう。また人事組織データと突き合わせれば、承認者IDは申請者IDと同じ組織かつ上位の職階である、ということまで確認できるかもしれません。このようにして正常の範囲を特定することで、そこから外れるデータを異常とみなしていきます。先ほどの例ですと「申請者IDと承認者IDは異なるが、承認者が申請者とは異なる部署の上席者である」というデータは正常ではない、つまり異常として検知することになります。リスクシナリオ的アプローチの場合、この閾値（しきい）を決定するためには現場における業務手順を細かく理解していく必要があります。

▶図2　プロットによる可視化の例



次に後者のデータ中心的アプローチでは、データの傾向を把握しその特徴からの乖離（かい）を見つけていきます。例えば、販売における値引きのようにある程度の裁量が許されているようなデータの場合、単純にルールに基づいた線引きは難しいかもしれません。このような例の場合、まずはデータをさまざまな観点で可視化していき、傾向や特徴の全体像を把握することがセオリーとなります。＜図2＞では、横軸を売上高、縦軸を値引額として1件の販売データを1つのデータ点としてその値引き状況を可視化した例です。傾向の見方はさまざまかと思いますが、次の特徴が見て取れるのではないでしょうか。

- ▶ データ点は各軸の値の範囲の中心に集中しており、極端に高額／低額の売上高／値引額は少ない
- ▶ 売上高が大きくなるほど、値引額が大きくなる
- ▶ 売上高が一定の値を超えると、値引額は売上高と連動して大きくならない

直感的にこのプロットから「他のデータと異なるデータ」を探すとなると、中央付近の密度の高いエリアからぽつんと離れているデータを拾い上げることができるかと思います。確かに直感的に異常値を抽出する手法もありますが、常に全てのデータを可視化して異常値を探していくというのも骨が折れますし、そもそも再現性が低く、なぜそのデータを異常値として抽出したのか客観的な説明が難しいという問題が起きます。

そこで、客観的な説明として数学的に説明できる値を用いることが考えられます。異常値を抽出する際に簡便的によく使われる統計値として、標準偏差があります。標準偏差はデータ集合のばらつきの程度をあらわす統計量です。標準偏差を使うことで平均値から大きく外れた値を異常として抽出する手順を考えます。＜図2＞の黄色の実線は値引額について標準偏差を2

倍し、平均値を加えた値を示しています。つまり、この黄色の実線よりも上にプロットされているデータについては、平均値からの平均的なばらつきの2倍以上のばらつきのあるデータであるということができます。

ここでは2倍という値を設定しましたが、それがどの程度稀（まれ）であるかはデータの分布にもよるため、実務では分析に用いるデータの質や分布状況、他の分析指標値との組み合わせなどから、現実的に検証可能なサンプル数まで絞り込めるように検討することが求められます。

Ⅳ　統計的な分析指標

異常検知に用いることができる統計的な分析指標には、標準偏差以外にも相関係数や回帰係数、変動係数、期待値残差などが考えられます（＜表1＞参照）。ここでは実際の不正調査でよく使われる変動係数について取り上げます。

会計不正調査においてデータ分析から異常検知をする際によく着目する会計数値として、単価があります。購買単価や販売単価など単価にも幾つか種類がありますが、共通する性質として、売買における単価には多少の値動きはあるもののあまり大幅に変化することはないという点が挙げられます。しかし実際の不正事例では、不正の実施者が商品の売買において単価を異常な範囲で調整することで不正のための資金を捻出することがあり、まさにこの行為の異常性が分析のポイントとなります。

ところが、単価の値動きを単純に比較しようとすると、通常の単価10,000円の商品が9,000円で取引される値動きと、100円の商品が90円で取引される値動きをどのように比較すればよいのかについて、すぐにはイメージできないのではないでしょうか。しかも取引件数は膨大でその件数にも違いがあるような状況では、比較可能な指標値を用意する必要があります。変動係数はこれを可能にする統計量で、金額の多寡や件数に関わらず、その数値の「ブレ具合」を比較可能な指標値として算出することが可能です。またこの変動係数を使った分析をする際、実務において検討すべき課題となるのは「どの単位に区切って計算するか」という点です。

例えば、ここに10年分の売買時の単価データがあるとします。これに対して品目ごとに変動係数を計算した場合、分かることは「この10年間でどの商品の単価がぶれ幅が大きいのか」という少し扱いにくい情報になります。異常検知の目的のためには、単に分析指標を計算するのではなく、目的に合致した結果を得られるように計算対象や範囲を工夫する必要があります。例えば、先ほどの計算では計算単位を品目ごととしていましたが、これを年度、部署、品目といった項目で細分化して計算します。こうすれば、同じ品目の変動係数を部署間で比較することで単価調整の部署による特徴が、また年度間で比較することでどの時期に単価調整が行われた可能性が高いのか、といったより目的に沿った分析が可能となります。

▶表1　分析に用いる一般的な統計量

変動係数	標準偏差を平均で標準化したもの。データ同士の散らばり方の比較に用いることができる。
相関係数	2変数の共分散を両方の標準偏差で割ったもの。2変数の関係性（どちらかが大きくなるともう一方も大きくなるなど）を-1から1の範囲で定量的にあらわすことができる。
回帰係数（1次）	2変数の一方の変数を用いて残りの変数を説明することを試み、2変数の関係を一次方程式であらわす直線関係で考えた際の係数。予測直線の傾きをあらわす。
期待値残差	観測値の実際の値と、回帰直線などから予測される値や確率変数の期待値との差。

出典：日本統計学会編　『統計学基礎』（東京図書）を基に筆者作成

Ⅴ　業務や内部統制に関する知見の活用

これまで述べてきたように、社内の業務データを使った分析を行う際にはデータの意味を理解するということが非常に重要になります。先ほどの例では変動係数を部署ごとに計算しましたが、担当者ごとに区切って計算、比較した方が見つけたい異常により詳細に迫ることができるかもしれません。この判断はデータと業務の関係をきちんと理解することで可能となります。

業務としてデータ分析をしていますと、PCのモニター画面にずっと向かって難しい計算や分析をしているのではないかとよく誤解されるのですが、データが生成された背景にある業務内容や内部統制の理解に時間を費やすことが多いのが実情です。分析対象となるデータの源泉である企業のITシステムはあくまでも業務ツールであり、そこには業務担当者の方々の考え、動き、日々の業務活動の過程がデータという形で写像として残っていきます。データ分析をするということは、その業務活動の写像であるデータを通じて、元の業務がどのように行われていたのかを推測する作業と言い換えることができるのかもしれません。

また、データ分析のためのITスキルや統計学の知識、内部統制や業務の知見などを一人の分析官だけで担う

ことは難しいのも実情です。より感度の高いデータ分析を行いたいと考えた場合、データ分析は複数の専門知識を持ったメンバーから構成されたチームで行うことが望ましく、例えば、データの処理に長けたIT専門家と社内の業務やルールに詳しい実務担当者の混成チームであれば、目的に合致した分析成果が期待できるのではないかと考えられます。

殊に現場担当者の知見はデータ分析に非常に大きな影響を与えることがあります。私が経験した事例ですが、販売データの分析で現場担当者の知見を得ずに進めていたため、分析作業が途中で頓挫したということがあります。分析対象としていた販売データは、販売担当者情報、承認者情報、入力担当者情報といった承認情報がとてもきれいに整備されたデータでした。業務システムのマニュアルや規程類、業務フロー文書などを確認すると、IDは個人ごとに付与されICカードによる認証システムも整備されるなど、非常に高度な職務分掌も存在していました。この情報を使ったデータ分析を行うことになったのですが、とある地方でのデータ入力現場を視察した際にこの分析は見直しを余儀なくされることとなりました。データ入力担当者が職員IDを兼ねたICカードを首に2枚かけていたのです。聞いてみると、そのうちの1枚は承認権限を持っている所長のカードで、毎朝出勤すると入力担当者がICカードを預かり、ノーチェックで承認行為を代行しているというのです。

こんなことは当然どのドキュメントを探しても載っていません。しかし現場をよく知る担当者に聞いてみると、「昔はよくあった」「地方だとまだやっているかもしれない」という情報をすぐに提供してくれました。現場を一番よく知っているのは現場の担当者です。空振りとなるデータ分析をしないためにも、業務や内部統制に詳しいメンバーの知見は非常に重要です。

Ⅵ　発見的統制としてのデータ分析

最後に、これまで論じてきたデータ分析をどのように発見的統制として利用するのか、ということについて整理します。

データ分析によるモニタリングは一度やれば終わりというものではなく、継続的な分析とフォローアップを行うために発見的統制として社内の内部統制システムに組み込むことが多くあります。併せて統制頻度も検討することになるのですが、データ分析を使った発見的統制の場合、異常値に対するフォローアップによってもいろいろな制約を受けることになります。例えば、発見された異常値について警告メールを送るような統制であればそれほど影響は受けませんが、異常値に対して非常に慎重な追加調査が求められるような場合には、抽出されるべき異常値の件数は限定される必要がありますし、統制頻度も頻繁に行うことも難しいでしょう。また、想定されるリスクインパクトが非常に大きく、早い発見と対応が必要という状況であれば、統制頻度は日次が望ましいということもあるでしょう。

別の視点からの考えもあります。モニタリングの自動化です。データ分析によるモニタリングはITシステム化して自動化しなければならないという誤解がまだ稀に聞こえてくることがあるのですが、どのような発見的統制として構築するのかという議論なしには検討できません。年次統制でよいということになれば、ITシステム構築による自動化は費用対効果との観点からはよい自動統制とは言えないでしょう。

また、内部統制全般に言えることですが、発見的統制として利用するデータ分析についても、定期的な見直し、再検討を行うことが重要です。これまで解説してきたように、データ分析は業務と非常に密接して構築されますので、業務の変化に合わせてデータ分析の変化も求められるからです。

Ⅶ　おわりに

データ分析による異常検知と発見的統制の実現のためにはIT、統計学、内部統制、会計、不正など多方面にわたる知見が必要であると言えます。そのためには、それぞれの専門領域を有するメンバーを集めたチームとして取り組むべきであり、またそうしなければ残念な結果になり得る、難易度の高いモニタリングであるとも言えます。

また今回は発見的統制としましたが、発見的統制を構築する際に得られる知見を蓄積していくことで、将来的には予防的統制としてのデータ分析を構築していくことも可能になると考えています。

お問い合わせ先
EY新日本有限責任監査法人
Forensics事業部　西日本Forensicsグループ
E-mail：noriaki.nishihara@jp.ey.com