■■連載記事■

データをビジネスに活用する実践アナリティクス

<第 4 回>　データの構造を把握するクロス集計

梶山 昌之
株式会社ワイハット

1.　正しい分析を行うためには

前回の連載では，データを分析した結果は現場や管理者に役立つものでなくてはならず，そのためには，データ解析の手法に従って分析を行う必要があることを解説しました.

また，研究や評価の対象はほとんどの場合，多変量的な特性を持つため，最終的には多変量の解析を行う手順になります[1].

多変量解析で最も利用されている手法は重回帰分析であり, Excel や R 言語などを用いて簡単に分析できるようになりました.

しかしながら，複数の変数があれば，すぐにツールを使ってなんらかの結論を出してしまうというのは，大変に危険な行為です．何故なら，多変量解析では変数間に関係性があり，交絡や交互作用があるのが普通ですので，この用語の意味を理解し，適切な分析を行わなければ正しい分析を行ったとは言えないからです.

「交絡（Confunding)」,「交互作用（Interaction)」は聞きなれない用語だと思いますが，重回帰分析で登場する重要な用語です．これを反映した分析の具体的な方法については後日解説の予定ですが，ここでは，直観的にこの意味を理解していただくことにします.

2.　結果を混同させる交絡

最初に交絡について説明します．英語で confund は「混乱させる，混同する」という意味ですが，交絡とは 2 つ以上の原因が混じり合って分離できない状態を言います．例えば，新開発の稲の肥料の効果を知るために，隣接する 2 つの土地 A と B で，稲を生育する実験を行ったとします．実験の担当者は土地 A で従来の肥料 X を使用し，土地 B で新開発の肥料 Y を使用して，成長した稲の収穫量を比較しました．その結果，肥料 X より肥料 Y の方が，収穫量が多いことがわかりましたが，これで新開発の肥料の効果が証明されたと言っていいでしょうか.

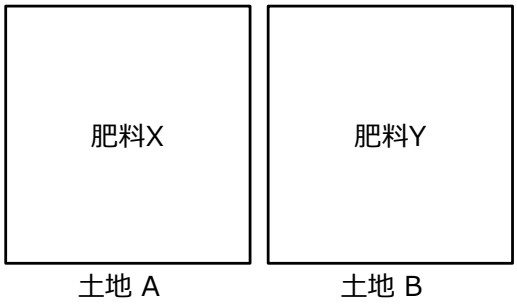

図 1　農場試験（交絡を無視）

2 つの土地は隣接していますが，土地 B は土地 A よりも日当たりがよく肥沃かもしれません．この場合は，肥料に差がない場合でも土地 B の稲はよく育つでしょう．これでは，何のために長い時間をかけて実験したのかわからなくなります.

実験の目的は肥料の効果を判断することですので，稲の生長に影響を与える他の要因の影響を取り除かなければ正しい判断はできません.

この場合は土地ですが，この様な要因は交絡要因または交絡因子と呼ばれます.

交絡因子の影響を取り除く方法の一つが乱塊法と呼ばれる方法です[2]．この場合は土地を小ブロックに分けて，図 2 に示すように肥料を配置したらどうでしょう.

土地 A と B で南北方向および東西方向に肥料 X と Y が同じ回数だけ配置されていますので土地による効果は相殺されて無視することができ，肥料のみの効果が観測されることになります.

以上は，分かり易い例として，農業試験の例で説明しましたが，医学では病気と原因の解明や薬の効果の検証などで交絡の影響を排除する工夫が

行われています.

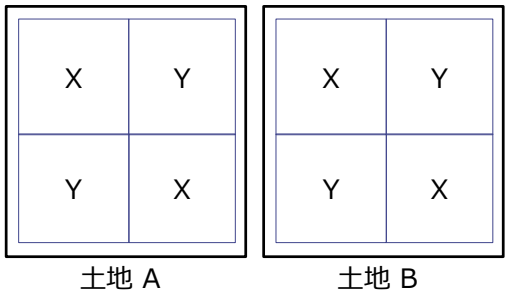

図 2　農場試験（交絡を考慮）

IT メトリクスの分析でも，生産性や品質に関係する要因としては，プラットフォーム，言語，業務や開発者のスキルなど様々な要因を考えることができ，これらの要因が相互に関係していますので，交絡の影響を反映した分析が必要になります.

例えば，Java は COBOL より生産性（FP 生産性）が高い言語ですが，プラットフォームの視点で見るとホスト系では COBOL が使用されることが多く，Web 系では Java が使用されることが多いという実態があり，その結果，Web 系の方がホスト系より生産性が高いという結論になります.

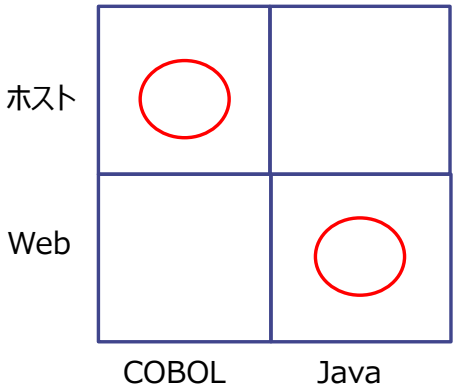

図 3　言語とプラットフォーム

すなわち，プラットフォームによる差を確認するために，プラットフォーム別の集計を行ったとしても，言語別の集計結果と概ね同じ結果になってしまいます．この場合は，言語とプラットフォームが交絡しており，2 つの要因の効果を区別して判定できないかもしれません.

3.　組み合わせの効果（交互作用）

次に交互作用について説明します．分かり易い例として，ダイエットにおける食事制限と運動の効果を考えます.

食事制限と運動はそれぞれがダイエットに効果があることは明らかです．ここで，食事制限によるダイエットの効果が 1kg/月であったとします.

この効果は，食事制限のみを行い運動量は変えていないときの効果です．同様に運動のみを行い，食事は制限していない場合の効果を 1kg/月とします.

それでは，食事制限と運動を同時に行ったときの効果はどうでしょうか．1+1=2 で 2kg/月かもしれません．しかし，要因の効果が相互に作用し，より強い効果が現れるかもしれません．その結果，例えば 3kg/月となることも考えられます.

すなわち，運動の効果が食事制限の有無によって異なるとき，運動と食事制限には交互作用があると言います.

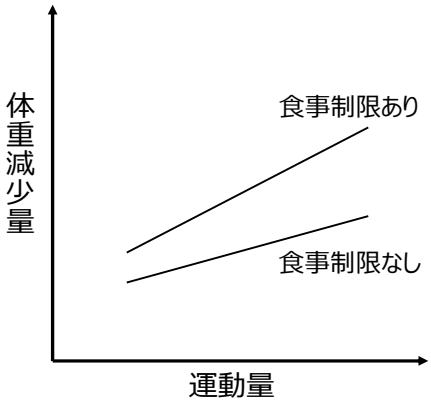

図 4　交互作用

4.　ドリンク飲料の売上データ分析

前節までに交絡と交互作用について説明したのは，1 つの要因と効果の関係を観測したとしても，それは他の要因の影響を受けた結果であるかもしれないことを理解するためです.

従って，このようなデータを対象として解析するためには，データの構造を明らかにする分析が必要です.

ここで，具体的な例としてドリンク飲料の売上データの分析を行ってみましょう.

あなたは，某飲料会社の企画室にいて今後の経営戦略を立案する立場であるとします．取扱う商品は，

ホットコーヒー（A），いちごミルク（B），
スポーツドリンク（C），ぶどうジュース（D），
その他の飲料（その他）

です.

分析のため昨年度の売上データを入手しました（表 1）．実務では数十万件以上のデータがあるかもしれませんが，ここでは 3834 件のデータを分析の対象としています.

また，実務では売上金額や購入者の年齢層などの項目もあると思いますが，説明を簡単にするため，売上個数といくつかの要因のみを取り上げていると考えてください.

この様なデータを前にして，経営戦略立案のヒントを得ることができるでしょうか．データは眺めているだけでは，いかなる関係性も掴むことはできません．すべてのデータに目を通したとしても，データに含まれる関係性を読み解くことは不可能でしょう.

表 1 ドリンク飲料の売上データ

No	販売日	販売月	商品	売上個数	地域
1	2015/1/1	1月	A	1	東京
2	2015/1/1	1月	A	1	名古屋
3	2015/1/1	1月	A	1	大阪
4	2015/1/1	1月	A	1	東京
5	2015/1/1	1月	C	1	大阪
6	2015/1/1	1月	C	1	名古屋
7	2015/1/1	1月	その他	1	大阪
8	2015/1/1	1月	その他	1	東京
9	2015/1/2	1月	A	1	大阪
10	2015/1/2	1月	A	3	東京
11	2015/1/2	1月	A	1	東京
12	2015/1/2	1月	A	1	東京
13	2015/1/2	1月	A	1	東京
14	2015/1/2	1月	A	1	大阪
15	2015/1/2	1月	A	1	東京
16	2015/1/2	1月	A	1	東京
17	2015/1/2	1月	A	3	東京
18	2015/1/2	1月	A	1	大阪
19	2015/1/2	1月	A	1	名古屋
20	2015/1/2	1月	C	1	東京

＊データの一部を示しています（全データ数: 3834)

このような場合に，まず行うべき作業がクロス集計です．分析対象となる要因として，売上月，商品，地域があり，各要因に着目して，月別売上，商品別売上，地域別売上を集計することも必要ですが，クロス集計ではこれらの要因間の関係も同時に示すことができます.

それでは Excel を使ってクロス集計を行う場合の具体的な操作を見ていきましょう．ここでは，Excel 2010 の場合で説明しますが，「挿入」タブ - 「ピボットテーブル」でクロス集計を実行することができます（図 5）.

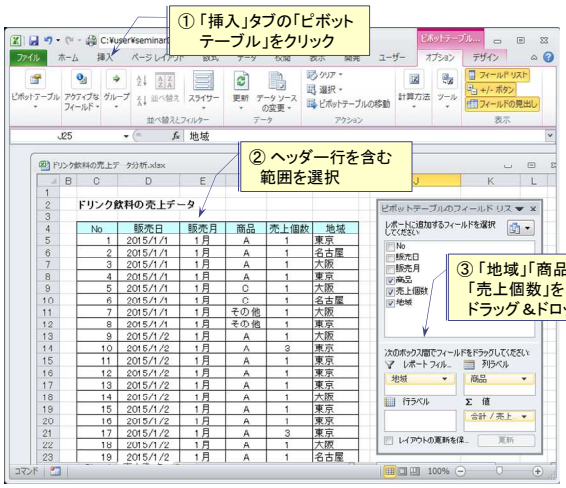

図 5　Excel のピボットテーブル

まず，「商品の売上に地域差があるか？」という課題について検討します．地域差があれば，地域差を反映した販売戦略が必要となるためです．データ項目に商品と地域がありますので，この二つの項目をレポートに追加します．また、集計の対象となる値としては売上個数を追加します.

その結果，地域別商品別売上個数（表 2）の集計結果が得られました.

表 2　地域別商品別売上個数

合計 / 売上個数	列ラベル					
行ラベル	A	B	C	D	その他	総計
東京	673	146	890	375	366	2450
大阪	343	161	435	207	210	1356
名古屋	250	269	361	162	159	1201
総計	**1266**	**576**	**1686**	**744**	**735**	**5007**

各地域の商品別の売上個数の集計結果から，どの地域でも売上個数が大きいのはスポーツドリンク（C）であることがわかります．しかし，地域ごとに売上個数が異なるので，地域毎の特性については読み解くことができません．また，数字のみでは直観的に理解することができません.

そこで，地域毎に商品別売上個数の構成比率を見てみることにします（表 3）．ピボットテーブルの計算結果は Excel のデータとして出力されていますので，さらにこれを加工して別の表を作ることができますので便利です.

表 3　地域別商品別売上個数比率

地域	A	B	C	D	その他	総計
東京	27.5%	6.0%	36.3%	15.3%	14.9%	100.0%
大阪	25.3%	11.9%	32.1%	15.3%	15.5%	100.0%
名古屋	20.8%	22.4%	30.1%	13.5%	13.2%	100.0%
総計	25.3%	11.5%	33.7%	14.9%	14.7%	100.0%

「複数の構成比を同時に比較する」グラフとしては帯グラフが最適です．その結果，図 6 が得られます．

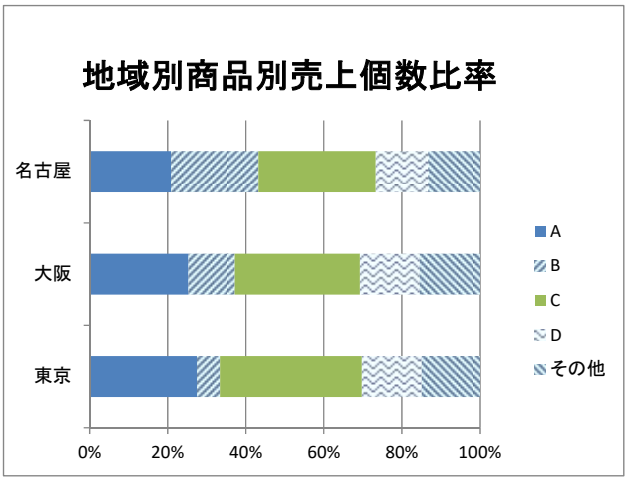

図 6　地域別商品別売上個数比率

図 6 を見ると，C，D，その他の構成比は大きな地域差がないが，ホットコーヒー（A）は東京では比率が高く，いちごミルク（B）は名古屋での比率が高いことがわかります．

この結果から，東京ではコーヒーの種類を増やして新たな市場を開拓するなどの戦略を考えることができます．また，帯グラフにより地域別の特性を直観的に把握することができました．

次に，「商品の売上に季節変動があるか？」という課題について検討します．販売月のデータがありますので，商品と販売月のクロス集計を行えば，簡単に結果を得ることができます（表 4）．

表 4　販売月別商品別売上個数

合計 / 売上個数	列ラベル					
行ラベル	A	B	C	D	その他	総計
1月	177	7	90	38	50	362
2月	164	7	83	40	61	355
3月	108	61	92	43	60	364
4月	88	165	83	33	57	426
5月	106	192	136	55	77	566
6月	105	79	175	43	80	482
7月	73	15	264	35	57	444
8月	85	13	272	37	55	462
9月	76	8	171	65	59	379
10月	86	11	115	122	53	387
11月	71	15	94	152	54	386
12月	127	3	111	81	72	394
総計	1266	576	1686	744	735	5007

「データの変化を見る」グラフとしては折れ線グラフが最適です．最初に総計の変化を見てみると，5 月にピークがありますが，季節による大きな変動はないように見えます（図 7）．

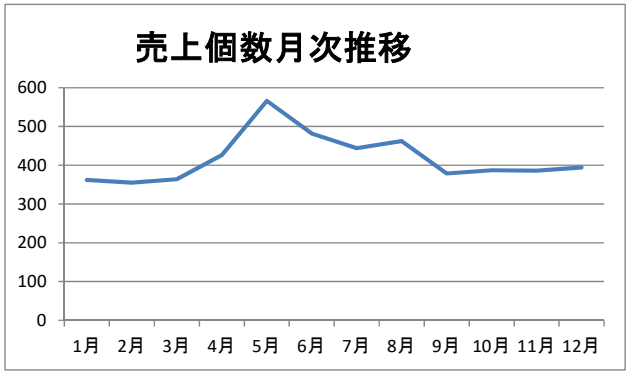

図 7　売上個数月次推移

ところが，同じデータを商品別にみると，商品により，変動のパターンが大きく異なることがわかります（図 8）．

ホットコーヒー（A）は冬場に需要が大きく，スポーツドリンク（C）は夏場がピークです．いちごの旬は 4 月～5 月ですが，いちごミルク（B）はこの時期に売れています．また，ぶどうの旬である 8 月～10 月には，ぶどうジュース（D）が売れています．季節を外れると売れなくなる商品についてはキャンペーンが必要かもしれません．

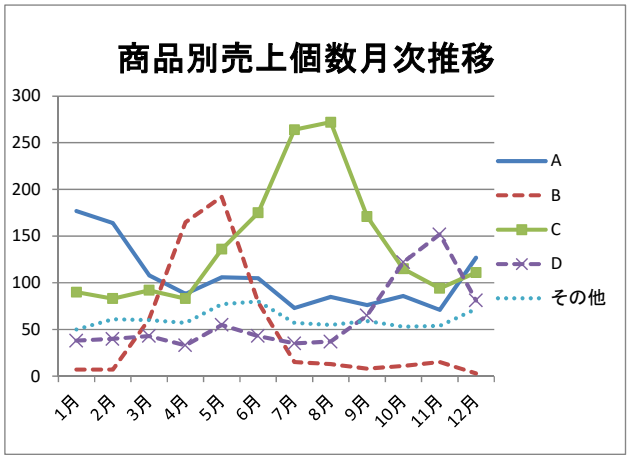

図 8　商品別売上個数月次推移

このようにデータ解析を行うことにより，数字の羅列に隠れて見えなかった実態が見えるようになりました．また，クロス集計はそのための強力な武器であることが理解できたと思います．

今回の売上データの分析では，ドリンクの売上には地域差があり，季節変動も影響していると結論しました．

しかし，ちょっと待ってください．この結果は，分析対象から収集した一部のデータの分析であり，すべてのデータを対象としたものではありません．

名古屋ではいちごミルクがよく売れると結論していますが，偶然の結果かもしれないのです．

「どの程度信頼のおける結果なのか」という事も，データ解析の重要なテーマです．次回はこの問いに応えるための考え方を解説の予定ですのでご期待ください．

参考文献

[1] 奥野忠一, 久米均, 芳賀敏郎, 吉沢正 著: 「多変量解析法（改訂版）」，日科技連, 1981.
[2] 田口玄一: 「第 3 版 実験計画法 上」，丸善株式会社, 1981.