データをビジネスに活用する実践アナリティクス

<第 9 回>　実務で役立つ分析の技術

梶山 昌之
株式会社ワイハット

前回の連載では，簡単な事例で 2 つの集団を比較する方法について学びました．また，対応のあるデータの場合には，グループ全体の平均値の変動よりは，個々の変動に着目したほうが，より変動を検出しやすくなることが理解できたと思います．

今回は，具体的で身近な例で，このような効果を確認する方法について解説したいと思います．

1.　悪玉コレステロール減少に効き目はあるか

健康に関するサプリメントや食品の広告を新聞やテレビでよく見かけますが，この中でも，効果を数字で示しているものもあります．

例えば，悪玉コレステロール（LDL-C: Low Density Lipoprotein Cholesterol）値を下げる効果がある食品を摂取した場合，摂取しない場合より低下していることを示す広告を見かけたことはないでしょうか．

あなたは，某食品会社の開発室にいて，開発中のサプリメントの効果を確認する立場であるとします．そこで，LDL-C 値を下げるサプリメントの効果を確認するための検証することになりました．

LDL-C は血管の内壁に付着しやすい性質を持ち，動脈硬化の原因になるため注意すべき指標です．政府統計によれば，男性の LDL-C 値（mg/dl）は平均 115.7, 標準偏差 30.1 です（図 1）．女性は 平均 120.2，標準偏差 32.4 です．

開発中のサプリメントはLDL-C値が高めの方を対象としています．そこで，LDL-C 値が平均より高い 30 名の方に被験者になっていただき，サプリメントの効果を確認することになりました．

検証の方法ですが，30 名の方全員に新サプリメントを 2 ヶ月摂取していただき，完了後に LDL-C 値の変化を測定する方法でよいでしょうか．

一見問題ないように見えますが，ここに落とし穴があります．重要なポイントは，LDL-C 値の低減は，必ずしも薬効による効果のみであるとは限らないということです．

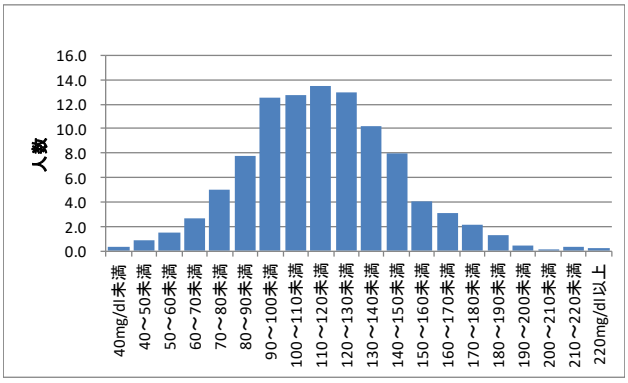

図 1　LDL-C 値の分布（男性）

※下記の資料によるデータを筆者がグラフ化したものです．
平成 22 年国民健康･栄養調査 血清 LDL－コレステロール値（(Friedewald の式）の分布（性･年齢階級別）［服薬者除外］

例えば，LDL-C 値が高めであることを知ってから食生活に気を配るようになり，その結果 LDL-C 値が下がるかもしれないのです．このように，LDL-C 値はサプリメントの効果以外にも，様々な原因で変動することが考えられますので，LDL-C 値が下がったとしても，これがサプリメントの効果であることを証明することができません．

そこで被験者を 2 つのグループ（A と B）に分け，グループ A には，新サプリメントと外観や味からは区別できないが薬効のない偽薬（プラセボ）を摂取いただき，グループ B には新サプリメントを摂取いただくことにします．

効果のない偽薬であっても，薬が効くと信じることによって，実際に効果が出るということがあります．これをプラセボ効果といいますが，仮にプラセボ効果があったとしても，この方式ではグループ A とグループ B の両方に同じ効果が働くこ

とになります．その状況で，グループ B にはより強い効果が観測できれば，グループ B の薬効を証明することができるという訳です．

以上のように結果を検証するために比較対象を設定した実験を対照実験といいますが，科学研究における実験の基礎となる考え方で，ビジネスにも広く応用可能です．

さて，実験開始前に全員の LDL-C 値を記録しました．これにより，個人別に効果を確認できるようになります．次に被験者をグループ A とグループ B に分けますが，この場合はランダムに分けることがポイントになります．ランダムに分けることにより特定の集団が持つ効果を排除します．実験が始まりましたら，毎日 3 回服用していただき，2 ヶ月後に，LDL-C 値を測定します．そのようにして得られた結果を表 1 に示します．

表 1　LDL-C 値対照実験

ID	グループ	実験前	実験後	低下量
1	B	173	162	11
2	A	124	128	−4
3	B	136	133	3
4	B	156	126	30
5	A	153	164	−11
6	B	149	158	−9
7	A	158	151	7
8	B	139	144	−5
9	A	165	152	13
10	B	144	127	17
11	A	177	167	10
12	A	191	177	14
13	B	141	119	22
14	A	134	130	4
15	A	149	140	9
16	B	169	137	32
17	B	191	173	18
18	A	124	125	−1
19	B	164	149	15
20	A	160	163	−3
21	A	155	152	3
22	A	188	209	−21
23	B	131	112	19
24	B	203	202	1
25	B	172	180	−8
26	B	196	189	7
27	A	143	122	21
28	A	171	183	−12
29	A	142	141	1
30	B	147	115	32

次に分析の方法ですが，今回はグループ A とグループ B それぞれで，LDL-C 値の変動を検証することにします．

実験前の計測値と実験後の計測値は対応のあるデータですので，各個人の LDL-C 値の変動に対して，意味のある変動があったかどうかの検証を行います．もし，新薬の効果がない場合，この値の平均は 0 に近い値になるはずです．この計算方法の詳細については，連載第 8 回の「対応のあるデータの検定」で解説しています．

表 2　LDL-C 値（t 検定のための計算）

統計量	低下量	
	グループA	グループB
サンプル数	15	15
平均	2.0	12.3
不偏分散	123.86	192.81
観測されたt値	0.696	3.440
p値（両側）	49.8%	0.4%

グループ A の被験者では，低下量の平均は 2.0 ですので，少し LDL-C 値が下がりました（表 2）．これは偶然ではなく意味のある変動なのかを検証する必要があります．

我々はグループ A のサプリメントが偽薬であることを知っていますので，LDL-C 値の低下が有意となった場合は，薬以外の別の効果が発生したと解釈することになります．

そこで，両側確率 p 値を計算すると，49.8%になりました．この意味は偽薬に効果がない場合でも，偶然の作用により 2.0 の効果が観測される確率が 49.8%あるということです．すなわち，変動は意味のある差であるとはいえないということになり，グループ A では LDL-C 値を低下させる効果は認められません．

一方，グループ B での低下量の平均は 12.3 であり，両側確率 p は 0.4%でした．この結果から，有意水準 1% でグループ B では LDL-C 値を低下させる効果があったといえることになります．

今回の場合は，グループ A では LDL-C 値が低下したとは認められなかったので，グループ B における新薬の効果を示すことができました．

しかし，グループ A でも LDL-C 値の低下が有意となった場合はどうなるでしょうか．グループ B でも新薬以外の効果が作用すると考えると，グループ B で得られた効果は，純粋に新薬の効果とはいえなくなってしまいます．

このような，問題に対応するためには，実験前の LDL-C 値と新薬の投与がどのように実験後の LDL-C 値に影響を与えるかをまとめて分析する手

法（回帰分析）を理解する必要があります．この手法は，実務では最も重要な手法ですので，正しく結果を解釈して使えるように，後日の連載でわかりやすく解説の予定です．

2. ドリンク飲料の嗜好性の分析

連載第 4 回（データの構造を把握するクロス集計）では，ドリンク飲料会社の例で，売上データから飲料の売上の予測や，地域差による嗜好の違いを確認する分析を行いました．しかし，この売上データの分析のみでは，どのような性別，年代の消費者がどのような商品を購入しているのかはわかりません．しかし，性別，年代別の商品戦略を立案するためには，このような情報が必須になります．

そこで，アンケートを実施して，このような情報を収集することになりました（図 2）．

アンケートを作成する際の留意点は，アンケートの目的を明確にし，どのように分析すればその目的を達成できるような結論を導くことができるかを十分に検討することです．今回のアンケートの目的は性別，年代別の商品戦略を立案することでした．

従って，性別，年代別にどのような商品が好まれているか，現在の商品の顧客に好まれている点，または改善点，要望などを知る必要があります．また，アンケートを作成する際には，どのようにデータを数値化するかということを想定しておくことも重要です．ここでデータの分析で必要になるデータの種類について学習しておきましょう．

ドリンク飲料に関する消費者アンケート

弊社の商品をお買い上げいただきありがとうございした。
今後の商品開発の参考にさせていただきますので、以下の、アンケートにお答えください。

Q1. あなたの性別をお知らせください。(いずれかを選択してください)
（男性）（女性）

Q2. あなたの年代をお知らせください。(いずれかを選択してください)
（10代未満）(20代）(30代）(40代）(50代）(60代以上）

Q3. 弊社の製品であなたが好きな商品をお知らせください。(複数選択可能です)

□ A コーヒー
□ B いちごミルク
□ C スポーツドリンク
□ D ぶどうジュース
□ X その他

Q4. 弊社の商品についての意見を自由にご記入ください。
[ ]

回答いただき、ありがとうございました。

図 2　ドリンク飲料に関する消費者アンケート例

データには量的変数と質的変数がありますが，Q1 と Q2 のデータはいずれも質的変数です．質的変数はカテゴリー変数とも呼ばれます．

Q1 の男性，女性はカテゴリー変数ですが，このカテゴリー変数の順序には意味がありません．このように順序に意味がないカテゴリー変数は「名義尺度」と呼ばれます．これに対して Q2 の変数は年代順に並べることができます．このように順序に意味があるカテゴリー変数は「順序尺度」と呼ばれます．

Q2 は選択肢にせずに年齢を直接入力させるようにすることもできますが，その場合は量的変数になります．Q2 をあえてカテゴリー変数にした理由は，年齢を数値入力にすると，今回のケースでは入力して貰えない可能性があるためです．

Q3 の回答は，複数の商品を選択できますので，多重回答と呼ばれます．商品が選択された場合は「好き」となり，選択されない場合は「好きではない」という意味になりますので名義尺度です．また，商品毎に一つの変数になります．

最後の Q4 は変数ではなく，テキスト（文字列）ですが，データマイニングの手法にはこのようなテキストを処理して，要望や苦情を把握する方法もあります．だだし，今回はテキストの処理方法については解説していません．

アンケートの方法にはアンケート用紙を使う方法，または Web でデータ取集する方法があります．

いずれの場合も，分析のためには収集したデータをコンピューターで処理しやすくするために表形式にまとめます．この作業をコード化といいます．今回の場合は，Excel を使い，表 3 に示す表にまとめました．

表 3　消費者アンケートの結果

No	Q1	Q2	Q3					Q4
	性別	年齢	商品					ご意見
			A	B	C	D	X	
1	男	60代	1	0	1	0	0	
2	男	20代	1	0	1	0	0	
3	男	40代	1	0	1	0	0	コーヒーが甘すぎる
4	男	40代	0	0	0	0	0	
5	女	30代	1	1	0	1	1	ぶどうは炭酸がいい
6	男	20代	0	1	0	0	0	
7	男	60代	0	0	0	1	0	
8	男	60代	0	1	0	1	1	やや値段が高い
9	女	40代	0	1	1	1	1	
10	女	40代	1	0	0	0	0	
11	男	10代	1	0	1	1	0	
12	男	20代	1	1	1	0	0	いちごミルクが好き
13	男	20代	0	0	0	0	0	
14	女	30代	1	0	1	1	1	
15	男	40代	0	0	0	0	0	
16	男	50代	1	1	1	0	0	
17	男	40代	0	1	0	1	1	
18	女	30代	0	1	0	0	0	自販機が少ない
19	女	30代	1	0	0	1	1	
20	男	30代	0	0	0	1	0	

※データの一部を表示しています（全データ数: 100）

ここで Q3 はカテゴリー変数ですが，選択（好き）を 1，非選択（好きではない）を 0 の数値に変換しています．この変数はダミー変数と呼ばれますが，この変換は多重回答の結果を処理するための重要なテクニックです．

アンケートの集計・分析でまず行うべきことは，

① 単純集計
② クロス集計

です．単純集計は変数毎に集計しそれを可視化する作業です．

例えば，Q1 で購入者の男女比を知ることができますが，約 6 割が男性です．この会社のドリンク飲料は男性の顧客が多く，女性が少ないことがわかります（図 3）．

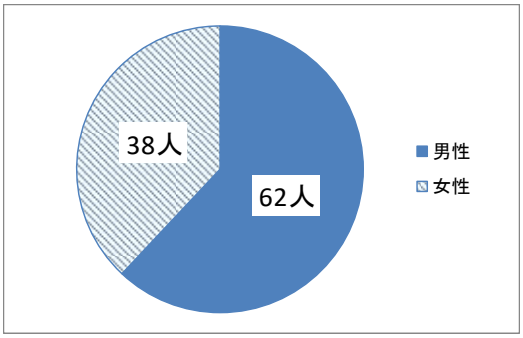

図 3　性別アンケート回答数

また，クロス集計を行えば，データの構造を掴むことができ，場合によっては，これだけの分析でも商品戦略につながる関係性を読み取ることができます．例えば，性別と年代でクロス集計を行い，性別年代別の分布を確認すれば，性別によって消費者の年代が異なることがわかります．特に女性の場合は，10 代から 20 代の若い世代の顧客が少ないようです．このことから，若い女性向けのキャンペーンを強化すれば売上の増加を期待できそうです（図 4）．

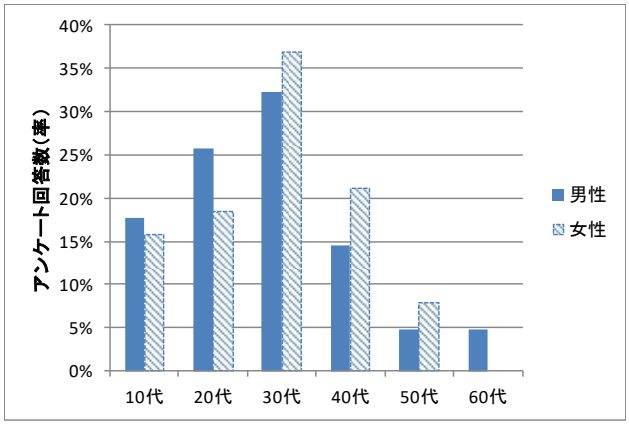

図 4　年代別アンケート回答数分布

Q3 については，まずは商品別に好きな飲料の分布を見てみましょう（図 5）．

最も人気がある商品は A（コーヒー），続いて，C（スポーツドリンク），D（ぶどうジュース）となりました．

概ね男女の分布パターンは類似しているので，商品別の男女差はないといってよさそうです．

それでは，年代の差についてはどうでしょうか．年代は時間的な変化ですので，ここでは折れ線グラフで可視化してみます（図 6）．

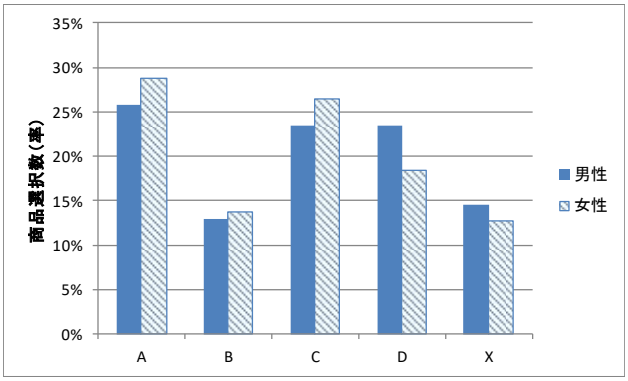

図 5　商品別好きな飲料の分布

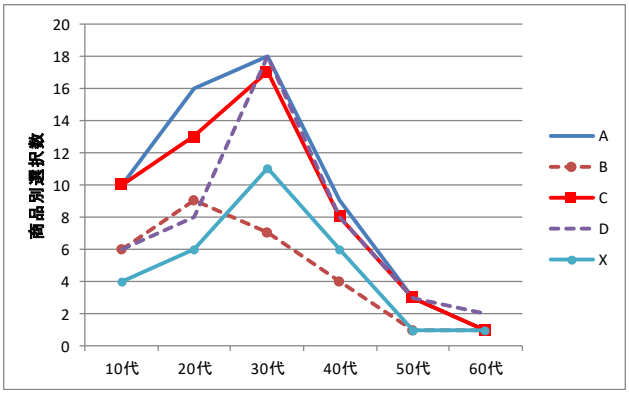

図 6　商品選択数年代分布

すると A と C は 30 代を中心として人気があり，年代分布のパターンが似ていることがわかります．B（いちごミルク）は 20 代で人気がありますが，他の世代ではあまり飲まれていないようです．

このグラフからわかることは，年代については，好まれる飲料に顕著な差があるということです．従って，キャンペーンも商品別に年代を反映したアプローチが必要になります．

以上は，性別，年代，商品を全体的に俯瞰した分析でしたが，次に個々の商品について分析することにしましょう．

例えば，A（コーヒー）に対する分析は，Q3 の A に対して Q1 と Q2 をクロス集計すれば得られます（図 7）．

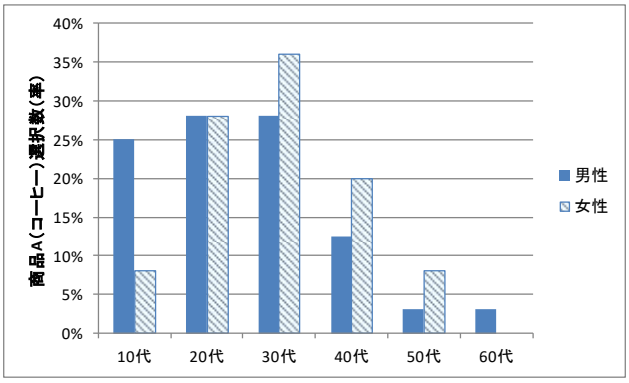

図 7　商品 A（コーヒー）選択数の年代別分布

この結果からわかることは，コーヒーは男女共に人気があるが，何故か 10 代の女性には好まれていないという点です．その原因を探るべく調査が必要です．商品パッケージのデザインが影響しているのかもしれません．

Q3 については，商品間の関係性を分析することができます．この方法の一つが相関分析です．相関分析は実務的にも非常に有効な手法であるため，その原理については後日の連載で解説の予定です．

今回は Excel を用いて相関分析を実行する方法を紹介します．

Q3 の A と B を例として説明すると，A と B の両方が選択されているケースが多い場合は，A を好む人は B も好む場合が多いということを意味します．反対に A が選択されている場合，B は選択されないケースが多い場合は，A を好む人は B を好まない傾向があると考えることができます．いずれの場合も，A の選択が B の選択と関係があるということであり，これを A と B には相関があるといいます．

変数 x と y の標本相関係数 r の定義式は以下の通りです．

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

すなわち，x と y の偏差積和を，x と y のそれぞれの偏差平方和の積の平方根で割ったものです．

相関係数は -1.0～1.0 の数値になり，通常はこの値が 0.4 以上の場合は正の相関があるといい，0.7 を超える場合は強い正の相関があるといいます．反対に，この値が -0.4 以下の場合は負の相関があるといいます．

相関係数は Excel の CORREL 関数で求めることができます．この関数では，分析の対象となる列を以下のように指定します．

CORREL（ 配列 1, 配列 2 ）

これを使って，ここでは商品 A~D の関係を調べることにします．商品 A～D のすべての組み合わせについて相関係数を求めると表 4 が得られます．

表 4　商品間の相関係数

商品	A	B	C	D
A	1.000	−0.043	0.661	0.095
B	−0.043	1.000	0.020	−0.027
C	0.661	0.020	1.000	−0.016
D	0.095	−0.027	−0.016	1.000

※対角行列ですので左下の数値は右上と同じ値になります．

この表から，任意の商品と同時に選択されることが多い商品を見つけることができます．例えば，A と C の相関係数は 0.661 ですので，正の相関があります．このような場合，店舗では A と C は近い場所に陳列する戦略が有効かもしれません．

一方，A と B の相関係数は -0.043 で負の相関となりますが，値は小さいので，両者は無相関と考えます．このことから，A が選ばれる理由と，B が選ばれる理由は異なるのではないかと推測できます．

簡単なアンケートですが，視点を変えて分析することにより，様々な関係性を把握できるようになることを実感いただけたでしょうか．

アンケートの分析でも，関係性を発見するためには，回帰分析や多変量解析など，変数間の関係を分析する手法が活用されます．

次回は，この分析の基礎となる回帰分析をわかりやすく解説したいと思いますので，ご期待ください．

最後に本稿の理解に役立つ文献を参考文献(今泉忠ら, 2012)として挙げます．

参考文献

今泉忠，田村義保，中西寛子，美添泰人 (2012). 日本統計学会公式認定 統計検定 2 級対応 統計学基礎. 東京図書.