■■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 5 回>　比率に関する統計的判断と分割表の利用

梶山 昌之
株式会社ワイハット

1.　信頼性のある報告とは

前回の連載では，2 つの集団の差を発見する方法について解説しましたが，差があったとしても，偶然の結果かもしれず，本当に意味のある差なのかどうかを検証することも重要であることを説明しました．

どのように検証したらよいのか．今回はその問いに答えるための考え方について解説します．

日本に住む外国人に「あなたは日本が好きですか」と聞いたところ，67%の方が「好き」と答えました．皆さんは，このアンケートの結果をどのように思いますか．日本が好きな外国人が多いと聞いていたが，はやりそうだったのかと感じたでしょうか．

しかし，このアンケートの報告は少し変です．何人の外国人に聞いたのかが報告されていません．

そこで，アンケートの人数を確認すると，なんと 3 人の回答を集計しただけのものでした．

1 番目：好き，2 番目：嫌い，3 番目：好き

この 3 人の結果から 67%（=2/3）というわけです．たった 3 人の結果では，信頼できないことは直観的にわかりますが，それでは，何人に聞けば信頼できる結果と言えるのでしょうか．この問いに答えるのが，検定と推定の考え方です．

日本に住む外国人は，在留外国人統計によれば 2015 年の時点で約 220 万人ですが，このうち半数以上が日本を好きと答えれば，「日本が好きな外国人の方が多い」と判断することにしましょう．

今，分析の対象にしている集団は日本に住むすべての外国人ですので，このような集団を母集団と呼びます．

もし，すべての在留外国人に答えていただければ，好きと答えた方の数の全体に対する割合 p が得られます．p はただ 1 つに決まる数値であり，このような数値を母数と言います．

p を求めるためには，すべての在留外国人に質問する必要がありますが，現実的にはそれを実施することは困難です．そこで外国人をランダムに何人か選んで質問することになります．

以上の状況は沢山の赤玉と白玉がガラポン抽選機に入っていて，ガラポンで出てきた玉の色を確認することと同じです（図 1）．

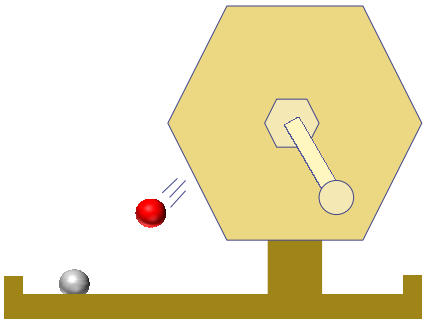

図 1　ガラポン抽選機

白が日本を好き，赤が日本を嫌いに対応するものとします．抽選機を 3 回廻した結果は，

[白赤白]

となりました．

ここで，3 つの玉を取り出したとき，1 つが赤となる確率はいくつかを考えてみましょう．実際のところ，赤と白の比率はわからないので，最初は，それぞれ同じ数だけ抽選機に入っているものとします．すなわち p=0.5 と考えます．このような仮定を置くと現象の出現確率を計算できるようになるというのが重要なポイントです．

その仮定のもとに，1 回目が白である確率は 0.5,

2 回目が赤である確率も，ほぼ 0.5 になります.（最初に白が出たので，2 回目の時は 1 個だけ赤の玉が多いと考えなければならないのですが，沢山の玉が入っていれば，ほとんど 0.5 なので，ここでは 0.5 とみなして計算します.）

3 回目が白である確率は 0.5 ですので，[白赤白]となる確率は 0.125（= 0.5×0.5×0.5）です.

3 回のうち 1 回が赤となる場合は

[赤白白]，[白赤白]，[白白赤]

の 3 通りの組み合わせが考えられますが，それぞれの組合せは同じ確率で起きますので，3 回のうち 1 回が赤となる確率は 0.375（= 3×0.125）になります.

すなわち，赤と白の数が等しいとしても，3 回のうち 1 回が赤となる現象は約 40%発生します. 従って，珍しいことが発生したわけではなく，たまたま，白が出た回数が多かっただけと考えることもできます.

そこで，さらに抽選機を回すと，

[白赤白白白白赤白白白]

となりました．すなわち，10 回のうち 2 回で赤が出ました．この例では，2 回目と 7 回目が赤となりましたが，このパターンの事象が発生する確率は約 0.00098（=0.5^10）です.

また 10 回のうち 2 回が赤となる場合の数は，1 回目と 2 回目が赤，1 回目と 3 回目が赤 ･･･ という様にいろいろな場合を数え挙げたものですので，n 個のうち k 個を選ぶ組合せの数を求める公式

$$ {}_nC_k = \frac{n!}{k!\,(n-k)!} $$

より求めます.

その結果，45 通り（=10×9/2）になります．これを使って，10 回のうち 2 回が赤となる確率を算すると 4.4%（=0.00098×45）になります.

すなわち，100 回のうち 4 回ほどしか起きない確率であり，そのようになったのは最初に設定した赤と白の玉の数が等しいという仮定が妥当ではないのかもしれません.

ところで，どの程度ならば，あまり起きない確率と考えるかですが，これは理論的に出てくる数字ではなく，最初に決めておくべき数字です.

統計学の慣例では，この値（有意水準）は 5%または 1%にします．通常は 5%にしますが，差がないのに差があると間違った判断をする危険を少なくするために 1%にする場合もあります.

以上の様に，得られたデータが仮説として設定した関係を覆すほど意味のある差（有意差）があるかどうかを判断することを検定と言います．仮説とは母集団に関する宣言と理解してください.

この場合，最終的には「白が多い」または「日本を好きな人が多い」ということを示したいのですが，「白と赤の数は等しい」または「日本を好きな人と嫌いな人は等しい」とういう仮説を設定しました．これを否定することで，両者に差があることを示そうとしたわけです.

この様に，否定する（無に帰する）ことを目的として設定される仮説を帰無仮説と呼びます.

その反対は「白と赤の数は等しくない」ですが，これを対立仮説と言います.

ここでは，5%の水準を用いることにしましょう. このケースでは 4.4%しか発生しない現象が発生しており，p=0.5 という帰無仮説は捨てられる（棄却される）ことになります．その結果，白（日本が好き）となる確率 p の推定値は，得られたデータから 0.8（=8/10）となります．すなわち，白は 80%含まれていると考えられることを統計的に示すことができました.

以上の内容で重要なポイントは，赤と白の割合が等しいという仮説に対する検定を行い，意味のある差が認められる場合に，白の割合の推定を行っていることです.

最初のアンケートの事例では，たった 3 人の回答から，67%と推定していますが，検定の結果，意味のある差として認められない場合は，推定しても，その値は信頼できないものになります.

2. 分割表による独立性の検定

前回の連載では「商品の売上は地域差があるか？」という課題を取扱いました．最初に，同じ課題を単純化して考えることにします.

例えば，ある企業は地域 A1，A2 で商品 B1，B2 を販売しているものとします．地域 A1，A2 で商品の種別による売れ行きに違いがあるのでしょうか.

販売データは地域と商品の属性を持ちます．1 週間分の売上結果集計した結果は以下の通りです.

観測度数

属性	B1	B2	行計
A1	14	6	20
A2	5	10	15
列計	19	16	35

この様に，2 つの属性に関係あるかどうかを調べるために作成した表は分割表（またはクロス集計表）と呼ばれます．この場合のデータは，2 行 2 列の表ですので 2×2 分割表です．表のデータをよく観察すると地域 A1 では商品 B1 がよく売れ，地域　A2 では商品 B2 がよく売れているように見えますが，偶然の結果かもしれません．

ここで，商品の売上は地域には影響されないと考えてみましょう．この場合，商品と地域は「独立」であると言います．また，複数の要因間に関係があるか否かを調べるための検定を独立性の検定と言います．

両者は独立であるという仮説を置けば，地域別売上数と商品別売上数の比率から，地域別商品別の売上数を推定できます．

例えば, 地域 A1 での売上数の合計は 20 個です．商品 B1 の売上は全体の 19/35 ですので，地域 A1 での商品 B1 の売上は 10.86（＝20×19/35）となります．この様にして期待度数の表が得られます．

期待度数

属性	B1	B2	行計
A1	10.86	9.14	20
A2	8.14	6.86	15
列計	19	16	35

次に観測度数と期待度数の差を求めます．商品の売上に地域差がある場合，この値は正または負の値となります．

観測度数−期待度数

属性	B1	B2	行計
A1	3.14	−3.14	0
A2	−3.14	3.14	0
列計	0	0	0

ただし，商品と地域が無関係の場合でも，偶然に観測度数と期待度数に差が出る場合もありますので，この差が十分に期待度数から離れていることを統計的に示す必要があります．そのための判断に用いられる値が以下に示すカイ 2 乗値です．

$$\chi_o{}^2 = \sum_{行}\sum_{列} \frac{(観測度数 - 期待度数)^2}{期待度数}$$

(観測度数−期待度数)^2/期待度数

属性	B1	B2	行計
A1	0.91	1.08	1.99
A1	1.21	1.44	2.65
列計	2.12	2.52	4.64

カイ 2 乗値は観測値と理論値の乖離を測る統計量で，カイ 2 乗分布と呼ばれる確率分布に従うことが知られています．

カイ 2 乗分布は左右非対称の分布で自由度によって形状が異なります．自由度というのは独立に選べる変数の数ですが，ここでは，分布のパラメータとして自由度という数値があるとだけ理解いただければよいと思います．

また，カイ 2 乗分布は自由度のみで決まる分布ですので，自由度がわかれば，5%または 1%の有意水準に対応するカイ 2 乗値を求めるための表が利用できます（表 1）．

表 1　有意水準とカイ 2 乗値

自由度	上側確率 p	
	5%	1%
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.72
12	21.03	26.22
13	22.36	27.69
14	23.68	29.14
15	25.00	30.58
16	26.30	32.00
17	27.59	33.41
18	28.87	34.81
19	30.14	36.19
20	31.41	37.57

図 2 は自由度 8 のカイ 2 乗分布ですが，右側の

グレーの部分（上側確率）が 5%になるようなカイ 2 乗値は表 1 から 15.51 と読み取れます. すなわち, 実績データから計算されたカイ 2 乗値がこの値以上ならば, 偶然ではない差があると判定します.

ここでは, 期待度数に対して観測度数に大きな差があるかどうかに着目するので, 片側確率の表を掲載しました. 分割表の場合は, 行数 r と列数 c で決まる値 f=(r-1)(c-1) が自由度になります. 2×2 分割表の場合, 自由度 f=(2-1)×(2-1)=1 ですので, 自由度 1 の 5%の値を読みとり, カイ 2 乗値は 3.84 になります.

一方, 観測値から計算されたカイ 2 乗値は 4.64 であり, 3.84 を超えました. これは, 商品の売上は地域差があることを意味しています.

どの程度の地域差なのかは,「観測度数-期待度数」の表を見れば明らかです. 地域 A1 では商品 B1 の売れ行きがよく, 地域 A2 では商品 B2 の売れ行きがよいことがわかりました.

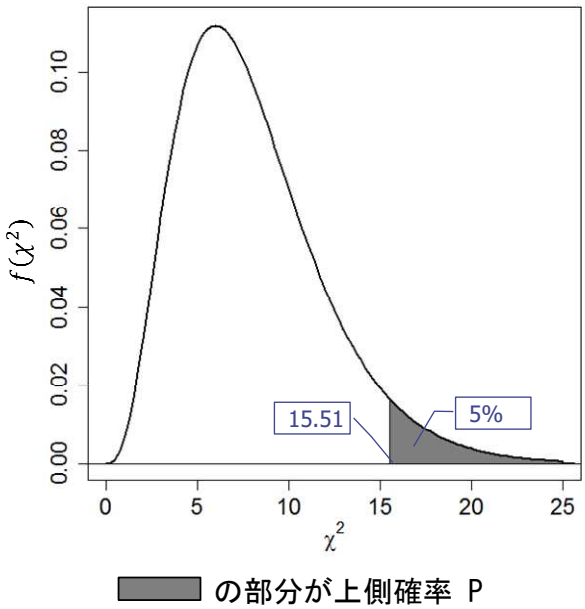

図 2　自由度 8 のカイ 2 乗分布

3.　商品の売上には地域差があるか

2 章では 2 つの地域および 2 つの商品で両者の関係があるかを検討しました. 3 つ以上の地域および商品の場合はどうなるでしょうか.

前回の連載では, 次の地域別商品別売上数のデータを対象として分析しました. これを使って説明しましょう.

観測度数

属性	A	B	C	D	その他	行計
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
列計	1266	576	1686	744	735	5007

取り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.

前回の分析では,「複数の構成比を同時に比較する」帯グラフで商品の売上には地域差があると判断しました. グラフによる判断は分かり易いのですが, 偶然の結果, 差があるように見えただけかもしれないという疑問が残ります. そこでカイ 2 乗検定の登場です.

カイ 2 乗検定の手順に従って作成した, [観測度数-期待度数」と[観測されたカイ 2 乗値]の計算を示します.

観測度数−期待度数

属性	A	B	C	D	その他	行計
東京	53.53	−135.85	65.01	10.95	6.35	0
大阪	0.14	5.01	−21.60	5.51	10.95	0
名古屋	−53.67	130.84	−43.41	−16.46	−17.30	0
列計	0	0	0	0	0	0

(観測度数−期待度数)^2/期待度数

属性	A	B	C	D	その他	行計
東京	4.63	65.48	5.12	0.33	0.11	75.67
大阪	0.00	0.16	1.02	0.15	0.60	1.94
名古屋	9.48	123.90	4.66	1.52	1.70	141.26
列計	14.11	189.54	10.81	2.00	2.41	218.86

3×5 分割表ですので, 自由度 f=(3-1)(5-1)=8 となりますので, カイ 2 乗表から 5%の値を読み取ると 15.51 になります.

一方, 観測されたカイ 2 乗値は 218.88 ですので, 商品の売上には地域差があるという結論になります.

[観測度数-期待度数]の表で具体的にその内容を見てみますと, 商品 B（いちごミルク）は名古屋での売れ行きがよく東京では少ない. 商品 A（コーヒー）と商品 C（スポーツドリンク）については東京での売れ行きがよいことがわかります.

4.　データ数の影響

前節の分析は総売上数が 5007 個のデータを対象としたものでした. ここでデータ数が少ない場合はどのような結果になるか検討してみましょう.

データ数が少なくても商品の売上げには地域差があるという結論になるでしょうか.

そこで，地域別商品別の比率が同じで売上数を1/20 として 251 個のデータを分析してみます.

観測度数

属性	A	B	C	D	その他	行計
東京	34	7	45	19	18	123
大阪	17	8	22	10	11	68
名古屋	13	13	18	8	8	60
列計	64	28	85	37	37	251

(観測度数-期待度数)^2/期待度数

属性	A	B	C	D	その他	行計
東京	0.22	3.29	0.27	0.04	0.00	3.83
大阪	0.01	0.02	0.05	0.00	0.10	0.17
名古屋	0.35	5.94	0.26	0.08	0.08	6.71
列計	0.57	9.26	0.58	0.12	0.18	10.71

分析の結果，観測されたカイ 2 乗値は 10.71 となり，有意水準 5%のカイ 2 乗値 15.51 より小さくなりました．この場合は，「商品の売上に地域差があるとは言えない」という結論になってしまいます.

このように分析対象となるデータ数が少ない場合は，差があっても，その差を検出することができません．反対に，データ数が多い場合は，少しの差であってもそれを意味のある差として検出します.

ビッグデータの分析では，大量のデータを対象にすることが多くなりますが，差を検出できたとしても，その差を反映した対策を行うことが実務的にはどの程度の効果があるかを確認することも重要です.

「2 週間で 6kg 痩せた」，「悪玉コレステロール値が下がった」など，世の中には商品の効用を謳う広告があふれていますが，その真偽は判断できるでしょうか．次回は，これらの本質を理解するための，基礎的な統計の知識を解説の予定で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[1]～[2] として挙げます.

参考文献

[1] 奥野忠一，久米均，芳賀敏郎，吉沢正:「多変量解析法（改訂版）」，日科技連, 1981.

[2] 今泉忠，田村義保，中西寛子，美添泰人:「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, 2012.