■■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 7 回>　改善効果を保証するには

梶山 昌之
株式会社ワイハット

1.　改善効果を保証するには

テレビやネットなどの広告で「2 週間で 6 キロ痩せた」などの事例を宣伝する広告をよく見かけますが，事例が事実であっても，それはダイエットの効果を保証するものではありません．そのため，やや小さな字で「痩身効果を保証するものではありません」という但し書きが表示されていると思います．

今回は，このような効果を検証する方法について，その基礎を学習します．実務では統計的に効果を把握する必要がある場面に多く遭遇しますので，この方法を学ぶことは重要です．

実際の計算は Excel や R（統計解析ツール）を使って行えばいいのですが，計算結果を得ても，その本質的な意味の理解には至りません．そこで，本連載では，その意味をできるだけ分かり易く理解し，ツールや理論を適切に使えるようになることを目標にしたいと思います．

2.　標準化

前回の連載では正規分布について解説しました．具体的な例として，成人女性の身長の分布は，概ね平均 160cm，標準偏差が 5cm であることを理解しました．正規分布に従う量の場合は，この 2 つの量で分布の特性を表すことができます．そこで，このような分布を代表する値を代表特性値と呼びます．

統計ではいろいろな単位を持つ量を扱います．例えば身長は cm，体重は kg ですが，統計の計算を行う際には，最初に「標準化」と呼ばれる次の変数変換を行います．

$$Z = (X - \mu) / \sigma$$

データを表す確率変数 X が平均μ，標準偏差σの正規分布に従う量であることがわかっているとき，上記の変数変換で得られる Z はどの様な分布に従うでしょうか．ここで注意すべき点は，平均μ（ミュー）と標準偏差 σ (シグマ）がギリシャ文字で書かれている点です．

例外はありますが，統計学では母数をギリシャ文字で表します．また，母数は確率変数ではなく定数です．例えば，日本の成人女性の身長の平均と標準偏差は，すべての成人女性の身長を測定して平均と標準偏差を計算すれば得られますが，これらの値はただ一つの値に決まります．この値が母数です．

今，　μ=160cm　σ=5cm　であったとすると，X がある値 x であるとき標準化した値は

$$z = (x-160)/5$$

になります．

正規分布に従う量に定数を加減しても，定数で乗除しても正規分布になる性質がありますので, Z は正規分布になります．また，その平均は 0，標準偏差は 1 になります．さらに, X は cm という単位を持っていたのですが, Z は単位のない値（無名数）になることにも注意してください．

このようにして標準化された Z の分布を標準正規分布と呼びます．ここで，Z という文字を使いましたが，統計学の慣例で Z は標準正規分布を意味します．

変数変換後の世界では，データの単位に依存しない統計的な性質のみを扱いますので，もとの変数のデータの単位に依存せず，データを統一的に取り扱うことができるようになります．このように変数変換し標準正規分布にして取り扱うことを標準化と呼びます．

ここで，再び身長の例で標準化を用いた計算を行ってみましょう．正規分布および身長の分布については前回の連載で解説していますので，そちらも参照してください．

「100 人の成人女性が集まりました．この中に身長 170cm 以上の女性は何人いると考えられるでしょうか.」

集まった方はスポーツ選手などの特別な集団ではなく，一般的な日本女性であり，平均 μ=160cm, 標準偏差 σ=5cm であることがわかっているものとします．x=170 ですので標準化すると，

z=(x-160)/5=(170-160)/5=2.0

となります．標準化された値がわかると，Z の分布値（パーセント点）に対応した確率を計算した数値表，または Excel で Z が 2.0 以上となる値を求めることができます．

標準化された値 z と確率の関係を図 1 に示します．ϕ(z)は標準正規分布の確率密度関数です．この図にいくつかの数値が示されていますが，これらの数値は覚えておくと便利です．

ここで，Pr{ } は{ }内の事象が発生する確率（probability）を意味します．確率は z の絶対値 |z| が任意の値を超える確率として示されていますが，このように両側の確率で示すのは，検定では基本的には両側確率を使用することに対応しています．

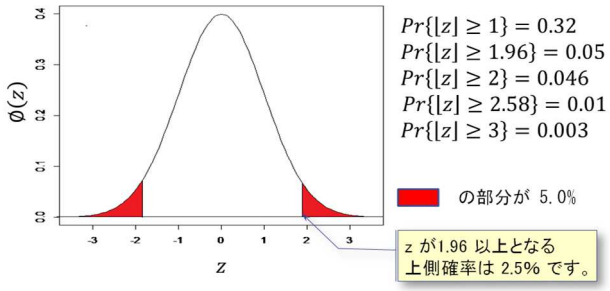

図 1　標準正規分布

解答には 2.0 以上となる確率が必要ですので，Pr{|z|≧2.0}=0.046 より 0.046/2=0.023 となります．

すなわち，100 名の場合は 2 名から 3 名が 170cm 以上となることがわかります．

3.　母集団とサンプリング

分析の対象となるすべての人や物の集まりが母集団です．実務で扱うデータは，分析対象となるすべてのメンバーの値を計測することは困難です．アンケートによるデータ，実験で得られたデータ，プロジェクトで計測されたデータ，いずれも，限られた時間と労力で収集されたデータですので，物事が持っている本質を知るためのすべてのデータを扱うわけではありません．このように母集団から抽出し，分析対象となった物を標本（サンプル）と呼びます（図 2）．

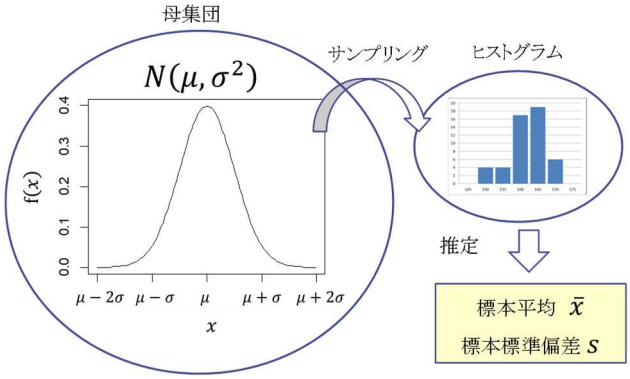

図 2　母集団とサンプリング

母集団の特性を表す値，例えば平均値の場合は母平均ですが，この値は神のみぞ知る値です．人間ができることは標本から得られたデータで母平均を推測することです．それ故，実務的には，母集団と標本の関係，母数と推定量の関係を理解することが，統計学の基本として重要です．

母数は神のみぞ知る値と言いましたが，母集団に含まれる全データを測定したわけではないが，十分に多くのデータを計測して得られた値は，実務的には，これを母数として扱う場合もあります．

ここで，次の問題を考えてみてください．

「ある集団で 5 名の身長のデータが得られました（表 1）．このデータは日本の成人女性の集団から得られたものと考えてよいでしょうか.」

表 1　ある集団の身長

No	身長
1	166.0
2	172.2
3	166.1
4	176.1
5	162.9
平均	168.7
標準偏差	5.4

平均身長は 168.7cm ですので，一般的な成人女性の身長にしては高すぎる様に思えます．しかし，たった 5 名のデータですので，たまたま，背が高

い人が抽出された結果なのかもしれません．この様な問題に答えるのが，母分散が既知の場合に使用できる z 検定です．

4. 平均の標準偏差

母集団からサンプリングされたデータ X_i は確率変数ですが，その平均も確率変数になります．例えば X_i (i=1～5) の平均（標本平均）

$$\bar{X}=(X_1+X_2+X_3+X_4+X_5)/5$$

は確率変数ですので，平均や標準偏差を求めることができます．さらに標本平均 の期待値は母平均に一致することを示すことができます．

$$\mathrm{E}[\bar{X}] = \mu$$

これは，例えば，5 つの標本を抽出しその平均を計算する作業を繰り返すと，その計算された平均の期待値（平均）は次第に母平均に近づいていくことをイメージすれば理解できると思います．

標本の大きさが 2 以上になりますと，$\bar{X}$ はより母平均に近い値になる場合が多くなり，平均のバラツキは小さくなるだろうということは想像できます．

例えば 100 名の平均は 160cm に非常に近い値になります．統計解析の理論によれば標本の大きさが n であるとき，分散はもとの値の 1/n，標準偏差は $1/\sqrt{n}$ になります．

$$\mathrm{V}[\bar{X}] = \frac{\sigma^2}{n}$$

$$\mathrm{D}[\bar{X}] = \frac{\sigma}{\sqrt{n}}$$

ここで，$\mathrm{V}[\bar{X}]$と$\mathrm{D}[\bar{X}]$はそれぞれ $\bar{X}$ の分散と標準偏差を示します．標本が 100 名の場合，$\bar{X}$ の標準偏差は σ =5 の 1/10，すなわち 0.5cm になります．

前節の例で，標本平均は 168.7cm でしたので，この値がどの程度母平均 160cm から乖離しているかを示すためには，この標本平均の標準偏差が $1/\sqrt{n}$ になっていることを反映する必要があります．5 名の場合の標準偏差は 2.236 (=5/$\sqrt{5}$) ですので，この標準偏差を用いて標準化すると

$$z = \frac{\bar{X}-\mu}{\sqrt{\sigma/n}} = \frac{168.7-160}{2.236} = 3.873$$

となります．図 1 を参照すれば，z が 3 の時，両側確率が 0.3% ですので，それより非常に小さい確率であり，成人女性の身長からのサンプリングと考えると，ほとんどありえない事態が発生したということになります．それ故，この場合は，5 名の身長のデータは，日本の成人女性の集団から得られたものとは言えないと言う結論になります．実は，この 5 名のデータは連載の第 6 回で登場した男性の身長のデータでした．

集団の性別を判別する情報として，身長のデータがあれば，非常に少ないサンプル数でも判別可能ということがわかります．

5. 母平均，母分散の推定

前節では，母分散（または母標準偏差）が既知の場合に使える z 検定について解説しましたが，実務では母分散が不明であり，得られたデータから推定しなければならないという場合の方が多いと思います．この場合も z 検定の場合に似ているのですが，いろいろと z 検定とは異なる状況が発生します．

母分散が不明という事は，得られたデータから母分散を推定しなければならないということです．

次の問題を考えてみましょう．

「技術研修でクラスの得点の平均が 50 点以上になることを目標として研修を続けているとします．テストを行い 8 名分の得点が得られました（表 2）．クラスの真の実力が平均 50 点以上になっていると言えるでしょうか．」

表 2　テストの得点

No	氏名	得点
1	A	48
2	B	54
3	C	58
4	D	47
5	E	58
6	F	56
7	G	57
8	H	65

この場合，母集団は 8 名のクラスの生徒が取る

得点の集合で，無限にある得点の仮想的な集合です．テストを無限に繰り返せば，真の実力が判明しますが，それは不可能です．真の実力が平均 40 点であったとしても，偶然の結果で，平均が 50 点以上になることもあります．そのため，統計的に実力を判定する必要があります．

この場合，標本平均 $\bar{X}$ は

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{8} X_i$$

ですので，母平均の推定値としては標本平均を使います．その理由は $\bar{X}$ の期待値は母平均に一致する性質を持つためです．

$$\mathrm{E}[\bar{X}] = \mu$$

このように期待値が母平均に一致する統計量を不偏推定量と呼びます．一方，$\bar{X}$ の分散 $\mathrm{V}[\bar{X}]$ は

$$\mathrm{V}[\bar{X}] = \frac{\sigma^2}{n}$$

でした．$\mathrm{X_i}$ の母分散が既知の場合は計算できますが，母分散が不明の場合，母平均を標本平均で置き換えて計算した分散を使用することが考えられます．これを標本分散と呼びますが，ここでは標本分散を Vs で表すことにします．

$$V_s = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

ここで Vs の期待値を計算すると，

$$\mathrm{E}[V_s] = \frac{n-1}{n}\sigma^2 = \left(1 - \frac{1}{n}\right)\sigma^2$$

となり，母分散より，すこし小さい値となってしまいます．そこで期待値が母分散に一致するように Vs を調整し，この値を s^2 とします．

$$s^2 = \frac{n}{n-1}V_s = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

この s^2 は

$$\mathrm{E}[s^2] = \sigma^2$$

となり，期待値が母数に一致しているので不偏推定量です．そこで s^2 を不偏分散と呼び標本分散と区別します．

不偏分散の平方根が標準偏差ですが，母標準偏差と明確に区別したい場合は，標本標準偏差と呼ぶ場合もあります．

何故 n-1 で割る計算になっているのでしょうか．バラツキを求めるために計算する偏差は，標本平均を使った場合，標本平均自身がバラツク量であるため，それを使って計算する分散のバラツキも大きくなると考えることができます．その分を調整したと考えるのが 1 つの考え方です．

多少理論的な説明としては，偏差は，

$X_1 - \bar{X},\ X_2 - \bar{X}, \ldots, X_n - \bar{X}$ であり，n 個のデータがあるのですが，$\bar{X}$ を一定とした場合，n 個の内 n-1 個の値が決まると，残りの 1 つは自由に決めることができず 1 つの値に決まってしまいます．つまり，自由に動ける数(自由度）は n-1 であるため平均は n-1 で割る計算になるとも言えます．

8 名のクラスの得点の場合で，母平均，母分散，母標準偏差を推定すると以下のようになります．

母平均の推定値 = 標本平均 = 55.4
母分散の推定値 = 不偏分散 = 33.70
母標準偏差の推定値 = 標本標準偏差 = 5.8

この節では，似たような用語がいろいろと登場するので，わかりにくかったと思います．しかし，次のステップに進むためには欠かせない重要な部分ですので，多少詳しく解説しました．

6. 母分散が不明の場合の標準化

第 2 節では母分散既知の場合の標準化について説明しました．これと同じ様に考えて，母分散が不明の場合は，母分散を不偏分散で置き換えて考えることができます．

$$\mathrm{t} = \frac{\bar{X} - \mu}{\sqrt{s^2/n}}$$

z は正規分布に従う量でしたので，t も正規分布に従う量と考えてよいのでしょうか．正規分布に従う量に対して定数を加減乗除した値は正規分布

になるのですが，不偏分散 s^2 は確率変数です．

そのため，t は正規分布にならず，ある特別な分布に従う量になります．この分布が t 分布です．t 分布を発見したのはギネスビール社に勤めていた統計学者のウィリアム・ゴセットですが，ある事情のため本名を隠して，スチューデントというペンネームでこの分布に関する論文発表したため，この分布はスチューデントの t 分布と呼ばれるようになりました．

実務では，ほとんどの場合，母分散は不明ですので，t 分布は頻繁に登場します．t 分布の形状は正規分布に似ていますが，正規分布より裾野がなだらかで，左右対称な分布になります（図 3）．また，t 分布は自由度のみで形状が決まる分布です．ここで，統計の計算で必要になる数値表を提示しておきます（表 3）．

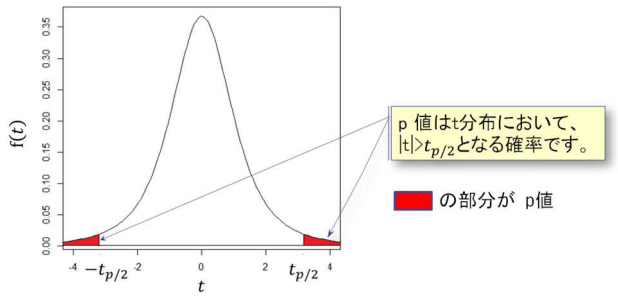

図 3　t 分布 (自由度 3)

表 3　t 分布表（パーセント点）

自由度	両側確率 P			
df	0.1	0.05	0.02	0.01
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947

以上で，前節で示した問題を解くための準備ができました．

クラスの得点の問題の場合，母集団の平均点が 50 点となっている場合を想定します．すなわちクラスの目標である 50 点を丁度達した値です．この目標値を μ_0 とします．すると，帰無仮説としては Ho: $\mu=\mu_0$ となります．これは母集団の平均は既に μ_0 に丁度達しているという意味になります．

収集するデータは 8 個のみですので，偶然の作用で標本平均は 50 点未満になることも 50 点以上になることもあります．

50 点から乖離する状況が発生し，それがめったに起きない確率であれば Ho を棄却することにします．

めったにない確率と言っても，具体的にその値を決めなければ判定できませんので，統計的慣例に習い，この確率を 5%とします．この確率は差に意味がある程度を表す確率ですので，有意確率とも言います．

観測されたデータにより計算された t 値（t_0）は

t_o=2.619

となります．数値表（表 3）から自由度が 7(=8-1)と両側 5%に対応する t 値を読み取ると， 2.365 になります．観測された t 値（t_0）はこの値より大きい値となりましたので，5%の有意水準で平均値は 50 点に等しいとは言えないという結論になります．

推定された平均値は 55.4 点であり，50 点より大きく，クラスの実力は既に目標に達していると判断できます．

ここで，クラスの成績は 50 点より大きいので，下限より低い場合の確率は考えなくてよいのではないかと考える方もおられるかと思いますが，帰無仮説として，Ho: $\mu=\mu_0$ としていますので，点数が低すぎる場合も仮説を棄却することになりますので，両側確率を用います．

成績が必ず向上する方向にしか変化しないと分かっている場合は，片側確率を使う方が判定には有利となりますが，その具体的な例については次回の連載にて解説の予定です．

本連載では，理論を理解するために，身近で分かり易い例を示していますが，t 検定は組織やプロジェクトにおける生産性や品質の指標の変動や，改善活動における効果の検証に使うことができますので，非常に有用な手法です．今回はその基礎を学びましたが，次回は両側検定と片側検定の使い分けや，より少ないデータまたは大きな差がな

い場合にも判定を可能にする方法について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]として挙げます.

参考文献

[1] 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.