データをビジネスに活用する実践アナリティクス

<第 6 回>　正規分布は統計解析の基本

梶山 昌之
株式会社ワイハット

1.　重要な正規分布

前回の連載では，統計的に信頼性のある報告とはどのようなものかについて解説しました．

「東京ではコーヒーが良く売れる」と言っても，データ数が少ない場合は偶然の結果かもしれません．

地域により商品の売上に差があるかという課題ではクロス集計とカイ 2 乗検定を用いた方法を学習しました．

ビジネスの分析では，各種の効果を比較検討する場面に遭遇します．その最も基本的な分析は，データが正規分布に従っている場合の比較です．そこで，今回は正規分布に対する理解を深めることにしましょう．

正規分布は左右対称のベル型の分布ですが，身近な例としては身長の分布を考えると分かり易いと思います．

統計理論の多くは正規分布を基礎として発展していますので，その基本的概念の習得は各種の統計理論や手法を理解するために必要になります．また，様々な自然現象に登場する分布ですので，最も興味深い分布です．

2.　パチンコ分布

ここで，図 1 のパチンコで遊ぶことを想像してみてください．突然パチンコが登場するのを不思議に思うかもしれませんが，このパチンコで正規分布を直観的に理解することができます．

パチンコ台は上から下まで 8 段の釘があります．釘に当たると左右等確率ではじかれ，次の段の釘に当たるようになっています．

一番下には玉を受けるスロットがあり，スロットには 0 から 8 までの番号が書かれています．

スロットに溜まる玉はどのような形状になるでしょうか．

どのスロットも同じ程度に玉が入ると思いますか．それとも，真ん中のスロットに一番多く玉が入ると考えますか．

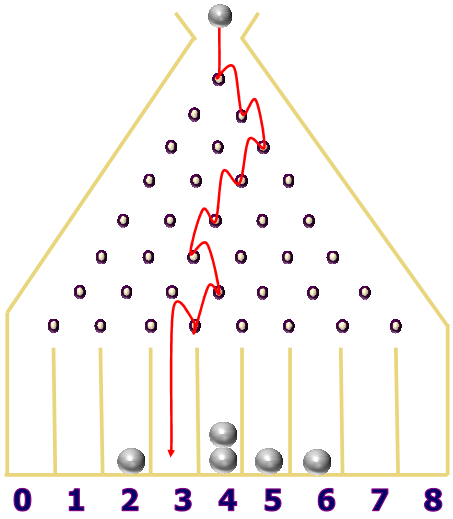

図 1　パチンコ（8 段）

そこで，玉が各スロットに入る確率を計算してみましょう．スロット 0 に玉が入るのは，すべての釘で玉が左側にはじかれた場合ですので，そのようなことが起きる確率は $(1/2)^8=1/256$ になります．

スロット 1 に玉が入るのは，8 つの釘のうち 1 つで右側にはじかれ，その他の釘ではすべて左にはじかれた場合です．そのようなことが起きるのは 8 通りですので，スロット 1 に玉が入る確率は 8/256 です．

スロット 2 に玉が入る確率は，8 つのうち 2 つの釘で右にはじかれ，その他の釘では左にはじかれる場合ですので，これは 8 個のうち 2 個を選ぶ場合の数を求める問題，すなわち組合せの問題になります（前回の連載を参照してください）．組合せの数を求める公式を使って計算した結果は表 1 の通りです．

これを図に表すとベル型の分布になります（図

2). 玉はランダムに左右にはじかれているのですが，不思議なことにスロットに溜まる玉はあるベル型の形状に近づいていきます！

表 1　パチンコ分布（確率の計算）

スロット	組合せの数	確率
0	1	0.004
1	8	0.031
2	28	0.109
3	56	0.219
4	70	0.273
5	56	0.219
6	28	0.109
7	8	0.031
8	1	0.004

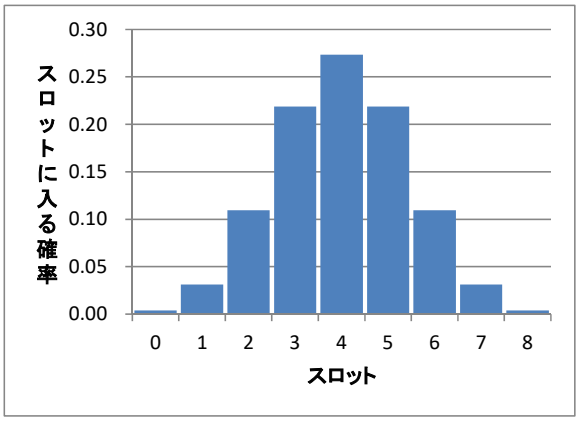

図 2　パチンコ分布

1 段目にある釘で左にはじかれた場合は 0，右にはじかれた場合は 1 となる変数を X_1 とします．同様に i段目にある釘に対応する変数を Xi とすると，玉が入ったスロット Y は

$$Y=X_1+X_2+X_3+X_4+X_5+X_6+X_7+X_8$$

となります. ここで, 任意の段における変数 Xi はランダムに 0 または 1 に変化し，その前段または後段の影響は受けない点に注意してください．

このような場合，Y の分布（またはその平均の分布）は，加算される値の数 n（この場合は 8) が大きくなるほど，正規分布に近づいていくことが知られています．ここで重要なことは，Xi は互いに独立でなければいけませんが，どのような分布に従う量でもかまわないということです．

n が大きくなると，その平均の分布が正規分布に近づきます．そのことを理論的に示した定理が中心極限定理（CLT: Central Limit Theorem）と呼ばれているものです．ここでは，正規分布に近づく性質が数学的に証明されていることを理解いただければよいと思います．

個々の値の和またはその平均として考えられる量は日常生活でも沢山見つけることができます．それらの量が従う分布は正規分布になると想定できることになります．

3.　身長の分布で正規分布を理解しよう

正規分布に従う量を取り扱う感覚を養うため，身長の分布で考えることにします．身長は非常に身近な量であるため，正規分布を実感で把握できます．

身長は男女で異なりますが，ここでは，成人女性の身長の分布が平均 160cm，標準偏差 5cm の正規分布に従うものとして考えることにします（図 3）．偏差とはそれぞれのデータの平均からの隔たりのことですが，標準偏差とは平均的な隔たりを表す量と考えてください．

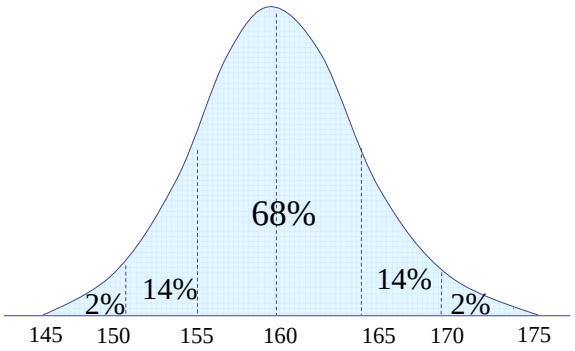

図 3　身長の分布

正規分布はベル型の分布ですので，凹の部分から凸の部分に替わる点，または凸の部分から凹の部分に変わる点（変曲点）があります．

正規分布では平均から標準偏差の大きさだけ離れた位置が変曲点になります．

成人女性の身長の分布の場合は, 155cm と 165cm が変曲点です．

凸の部分には約 7 割(68%) の人が入りますので，この範囲の人は平均的と感じることでしょう．

成人女性の身長の場合は 155cm～165cm になります．これを超える高さの人を背が高い人と呼んでも不自然ではありません．

また，平均から標準偏差の 2 倍離れた範囲は 150cm～170cm ですが，この範囲には 95.4%の人が

入ります．170cm を超える人は 100 人のうち 2 人くらいしかいませんので，とても背が高いと表現することでしょう．このように，私たちが集団を区分し，程度を表すときに使う「やや，とても，非常に」という表現は，その発生する確率と対応しています．

「センミツ」という言葉がありますが，製造業では 1000 個の 3 個の不良品が出るという意味で使われています．

この数字は「めったにない」こと，または「異常な」という表現に対応しており，正規分布では平均から標準偏差の 3 倍以上離れた場合に相当します．

標準偏差の 3 倍の範囲は 145cm～175cm ですが，この範囲には 99.7%が入ります．この範囲で 175cm を超える人は 1000 人のうち 1～2 人しかいませんので，非常に背が高い人ということになります．ちなみに，現時点の日本女子バレーの選手の平均身長は 175cm ですので，半数は非常に背が高い人が選手になっていると言えます．

事象の発生頻度を判断の基準にする考え方は，製造業における装置の異常の検出，学校では成績の評価，プロジェクト管理では組織の評価基準の設定，医学では病気の判定基準として使われています．

4. 正規分布に近似する 2 項分布

ここで，再びパチンコの例で考えることにします．今度は，玉は等確率で左右にはじかれるのではなく，ある確率 p で右にはじかれるように作られているものとしましょう．そして，玉が左にはじかれた場合は失敗，右にはじかれた場合は成功と考えます．すると，玉が入ったスロットは 8 回の試行で成功した回数ということになります．このように結果が成功か失敗（2 項）のいずれかである独立な n 回の試行で成功する数 X が従う分布の確率関数は，

$$f(x) = B(n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

ここで，

$$\binom{n}{x} = {}_nC_x = \frac{n!}{x!\,(n-x)!}$$

となりますが，これが 2 項分布です．この式で B は 2 項分布（Binomial distribution）を表しています．

最初に登場したパチンコ分布は p=0.5 の 2 項分布でした．2 項分布の平均 μ（ミュー）と標準偏差 σ（シグマ）は

$$\mu = np$$
$$\sigma = \sqrt{np(1-p)}$$

ですが，この公式を使って，p=0.5 のパチンコ分布の平均と標準偏差を求めると，

$$\mu = 8 \cdot 0.5 = 4$$
$$\sigma = \sqrt{8 \cdot 0.5 \cdot (1-0.5)} = \sqrt{2} = 1.41$$

となります．ここで，図 2 とこの数値を見比べてみてください．X=4 に分布のピークがあり，X=5.41 (=4+1.41) で分布の凸が凹に変わっていることが確認できると思います．

この例のように，2 項分布はある条件を満たせば正規分布に近似します．

しかし，どのような場合でも正規分布とみなせるわけではありません．例えば p=0.2 の場合は中心が左に移動し左右非対称の右裾が長い分布になります（図 4）．

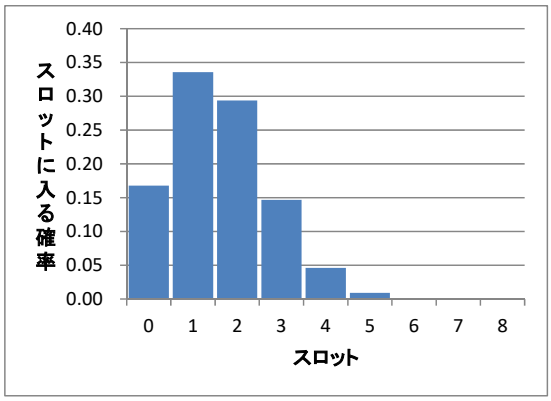

図 4　パチンコ分布（p=0.2）

それでは，どんな場合に正規分布とみなせるかと言うと，p が 0.5 以下であれば，

「np が 3 を超えると 2 項分布は正規分布とみなしてよい」

というのが，統計学で慣例として使われている規則です．3 という値は理論的に導かれたものではないのですが，2 項分布と正規分布の値を比較し

た結果から得られた統計的知見です.

パチンコ（8 段）の例で，np は 4 ですので，正規分布とみなしてよいという結論になります.

何故，このような知識が重要になるかと言いますと，製造業では検査個数に対する不良個数，IT 分野ではテストケース実施数に対する検出欠陥数など, いずれも離散量を扱うことも多いのですが, それを正規分布として扱うことにより，分析や管理が簡単になるためです.

ここで,「コインを 100 回投げたときに，40 回以下が表になる確率はいくつでしょう.」というクイズに答えてみてください.

筆者がこのクイズを多くの方に訊ねて直観的に答えてもらった経験では，40%位という回答が多いようでした.

「10 回投げたときに 4 回以下が表になる確率」は 38%ですので，40%位という直観は正しいのかもしれません.

しかし，100 回になりますと，これを 2 項分布で計算するのは大変であり，暗算でできるようなものではありません.

ところが，正規分布の知識と 2 項分布の正規近似の規則を利用すれば，簡単に答えを得ることができます.

この場合，np=100・0.5=50 ですので，表が出る回数は正規分布に従うとみなせます. np(1-p)=100・0.5・0.5=25 ですので，標準偏差は 5 です. 従って, 標準偏差の 2 倍の範囲を考えると, 表が出る回数は 95.4%が 40～60 の間に入ることになります. 従って 40 回以下が表になる確率は約 2.3% (=(1-0.954)/2) となります. 直観とはかなり違った値になりました.

5. 正規性の判定

前節までの解説で正規分布とはどのような分布なのかを, 直観的に学ぶことができたと思います.

ここで注意すべき点があります. 図 2 では非常に綺麗に正規分布に適合している図が描かれていますが，これは実際にスロットに溜まった玉の分布を表しているわけではなく，理論値としてこのような分布になることを表しています. 試行回数が多ければ，この分布型に近い結果が得られますが，例えば，10 回程度の試行では左右対称になるとは限らず，正規分布とはかけ離れた形状が現れることもあります.

実務では分析の対象になる量が従う分布は不明の場合が殆どですので，収集したデータから，その分析対象の量が従っている分布を推定することが，分析の第一歩になります.

例えば，成人の男女の身長のデータを各々10 名ずつ収集し，基本的な統計量を計算した結果が得られたとします（表 2）.

この結果から身長が正規分布に従っていることを示すことができるでしょうか. または，正規分布ではない可能性もありますので，そのことを示すにはどのようにしたらよいでしょうか.

正規分布の特徴は左右対称であり，ある一定の尖りぐあいの山を持つベル型の形状になることです.

この左右対称性と山の尖りぐあいを表す統計量が「ひずみ」と「とがり」です. 分布が正規分布に従っている場合，これらの値はいずれも 0 になります. ただし,「とがり」については正規分布であれば 3 とする定義もありますが，0 になるように定義するのが一般的であると考えてください. 分布の形状が左右対称ではなく，山が左に寄り右裾が長い形状のとき,「ひずみ」は正の値になり, 山が右に寄った場合は負の値になります.「とがり」はこの値が大きいほど分布が尖っていることを表しています. 統計的慣例では,「ひずみ」または「とがり」の値が -1.5～1.5 の範囲を超えると「正規分布に従っているとは言えない」と判断します.

表 2　身長のデータ

No	男性	女性
1	166.0	159.1
2	172.2	158.3
3	166.1	169.8
4	176.1	163.5
5	162.9	153.0
6	173.8	162.5
7	177.6	161.7
8	171.2	159.5
9	176.8	153.2
10	168.1	151.8
平均	171.1	159.2
標準偏差	5.1	5.6
ひずみ	−0.2	0.3
とがり	−1.3	0.0

さて，収集した 10 名の身長の例では,「ひずみ」「とがり」のいずれも-1.5~1.5 の範囲になりましたので「正規分布とみなしてよい」と判定されました.

尚，「ひずみ」「とがり」は，歪度（Skewness），尖度（Kurtosis）とも言います．計算は Excel の関数に SKEW，KURT がありますので，これを利用するのが便利です．

ここで，平均 μ，標準偏差 σ の正規分布の確率密度関数の式を示しておきます．

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

この式には，π，e などの数学定数が登場しますので，「統計は難しそう」と感じた方もおられるかと思います．ただし，実務的にはこの式の理論を理解する必要はありませんのでご安心ください．

尚，X が平均 μ，分散 σ^2 の正規分布に従うことを表すのに

$$X \sim N(\mu, \sigma^2)$$

と表現します．標準偏差を 2 乗したものを分散といいます．N は正規分布（Normal distribution）を表しており，～（波ダッシュ）は「に従う」とう意味で使われています．

正規分布への適合性を判定する方法には，もっと高度な数学を使う方法もありますが，実務的には，この「ひずみ」と「とがり」による判定で十分です．何故なら，実務で扱うデータは完全な正規分布に従うものでもなく，分布型が不明の場合は多少正規分布との乖離があったとしても，正規分布に従うものとして扱う方が実務的であるためです．

男女 10 件の身長の例では「正規分布とみなしてよい」という結果になりましたが，さらにデータを収集した結果，「正規分布に従っているとは言えない」と判断が変わる場合もあります．その場合は，どのような分布に従っているかを探索する分析に入ればよいのです．

今回は正規分布を中心とした解説を行いましたが，実務では対数正規分布などの非対称の分布も登場します．これらのデータを対象とした分析を行う場合でも，正規分布で得た知識が基本となります．

次回は，正規分布を基本として，改善の効果の判定を具体的にどのように行うのかを解説する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.