連載講座

栄養学研究者のための統計的データ解析入門

——第2回 推定と検定の枠組み——

中 里 溥 志^{*1}, 青 江 誠 一 郎²

Introduction to Statistical Data Analysis for Nutrition Research

——Part 2 Statistical Inference——

Hiroshi Nakazato^{*1} and Seiichiro Aoe²

2 推定と検定の枠組み

2.1 確率分布

(1) 確率変数と確率密度関数

[例1] のデータについて作成した図1.3の幹葉表示から、表2.1のような度数表を作成する。表2.1の相対度数について、ヒストグラムを作成すると、図2.1のようになる。

[例1] のデータの大きさは90であるが、もし同じ条件にある非常に多数の成人男子について、HDL-コレステロール値を測定したデータがあったとしたら、区間の幅を小さくしてヒストグラムを作成すると、図2.2のような曲線に近づくものと仮定する。

図2.2のグラフでは、曲線と横軸に囲まれた部分の面積が1になるように、縦軸のスケールを決めている。このとき、HDL-コレステロール値を x とし、曲線を表す関数を $f(x)$ と記して、“確率密度関数”という。

$$f(x) \geq 0 \quad (2.1)$$

$$\int_{-\infty}^{\infty} f(x) dx = 1 \quad (2.2)$$

である。

ここで、[例1] と同じ条件にある非常に多数の成人男子から、ランダムに一人を抽出したとして、その人のHDL-コレステロール値を X とする（まだ測定はしていない）。 X は、次のような性質をもった変数であると考える。

- 1) X の取りうる値の範囲は、実数である。

- 2) 実際にどんな値になるかは、測定が終わるまでは不確定であるが、測定が終われば確定する。

- 3) X の実現値（得られる測定値）が x 以下である確率 $P\{X \leq x\} = F(x)$ （“下側確率”）が、次のように定められている。

$$P\{X \leq x\} = F(x) = \int_{-\infty}^x f(x) dx \quad (2.3)$$

以上のような変数 X を“確率変数”， $F(x)$ を“分布関数”という。 X の実現値が x より大きい確率 $P\{x < X\}$ （“上側確率”）は、 $1 - F(x)$ になる。また、確率変数 X の実現値が a より大きく、かつ b 以下である確率

$$P\{a < X \leq b\} = F(b) - F(a) = \int_a^b f(x) dx \quad (2.4)$$

は、図2.3のグラフで斜線を引いた部分の面積になる。

確率変数 X と、確率密度関数 $f(x)$ または分布関数 $F(x)$ を対応させたものを、“確率分布”という。

(2) 確率分布の平均値と分散

データの分布の中心の位置を表す指標に平均値、ばらつきを表す指標に分散を用いたように、確率分布についても平均値 $E(X)$ と分散 $V(X)$ を次のように定義する。

[例1] のデータが個々の測定値ではなく、表2.1のような度数表で与えられていたら、平均値は区間の中心値と相対度数の積和

$$22.0 \times 0.0111 + 27.0 \times 0.0333 + \dots + 112.0 \times 0.0111 = 50.9$$

で計算することができる。個々の測定値から計算した平

* 連絡・別刷請求先 (Corresponding author)

¹ (171-0041 東京都豊島区千川 2-29-3)

2-29-3 Senkawa, Toshima-ku, Tokyo 171-0041, Japan

² 雪印乳業株式会社開発企画室 (350-1165 埼玉県川越市南台 1-1-2)

Products Planning Department, Snow Brand Milk Products Co., Ltd., 1-1-2 Minamidai, Kawagoe, Saitama 350-1165, Japan

表 2.1 HDL-コレステロール値の度数表

No.	区間 (mg/dL)	中心値 (mg/dL)	度数	相対度数
1	19.5- 24.5	22.0	1	0.0111
2	24.5- 29.5	27.0	3	0.0333
3	29.5- 34.5	32.0	2	0.0222
4	34.5- 39.5	37.0	10	0.1111
5	39.5- 44.5	42.0	13	0.1444
6	44.5- 49.5	47.0	17	0.1889
7	49.5- 54.5	52.0	16	0.1778
8	54.5- 59.5	57.0	10	0.1111
9	59.5- 64.5	62.0	6	0.0667
10	64.5- 69.5	67.0	4	0.0444
11	69.5- 74.5	72.0	5	0.0556
12	74.5- 79.5	77.0	0	0.0000
13	79.5- 84.5	82.0	1	0.0111
14	84.5- 89.5	87.0	0	0.0000
15	89.5- 94.5	92.0	0	0.0000
16	94.5- 99.5	97.0	0	0.0000
17	99.5-104.5	102.0	0	0.0000
18	104.5-109.5	107.0	1	0.0111
19	109.5-114.5	112.0	1	0.0111
計			90	1.0000

図 2.1 HDL-コレステロール値のヒストグラム

均値と少し異なるのは、測定値の代わりに区間の中心値を使っているためである。

度数表からデータの平均値 $\bar{x}$ は、区間 i の中心値を x_i 、相対度数を f_i とすると

$$\bar{x} = \sum_i x_i f_i \quad (2.5)$$

求められるので、この和を積分に置き換えれば

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \quad (2.6)$$

で、確率分布の平均値 $E(X)$ が定義できる。同様にして、分散 $V(X)$ は

$$V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx \quad (2.7)$$

になる。

平均値 $E(X)$ の $E(\)$ という記号は一般に、確率変数 X の関数 $g(X)$ に、その“期待値”

図 2.2 HDL-コレステロール値の確率密度関数のグラフ

図 2.3 確率密度関数のグラフ

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx \quad (2.8)$$

を対応させる働きを表す記号（“演算子”）である。この記号を使うと、分散を

$$V(X) = E((X - E(X))^2) \quad (2.9)$$

と書くことができる。平均値 $E(X)$ は、正確には確率変数 X の期待値であるが、平均値と同じ意味で期待値と呼ぶことがある。

2.2 正規分布

(1) 正規分布と標準正規分布

これまで確率密度関数を $f(x)$ と表しただけで、具体的にどんな関数を想定しているのかには触れずにきた。しかし、推定や検定という統計的な手法を適用するためには、データに分布を想定すること、つまり $f(x)$ を具体的に定めることが必要になる。

“正規分布”とは、確率密度関数が

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad (2.10)$$

である確率分布のことである。この式に含まれる二つの“母数”（“パラメータ”ともいう） μ と σ^2 の値を定めると正規分布が決まるので、これを $N(\mu, \sigma^2)$ と表す。 μ を“母平均”、 σ^2 を“母分散”と呼んでいる。式 (2.10) を式 (2.6) と式 (2.7) に代入して積分の計算を行うと、 $E(X) = \mu$ 、 $V(X) = \sigma^2$ であることが確かめられ

図 2.4 $N(\mu, \sigma^2)$ の確率密度関数のグラフ

る。 $f(x)$ のグラフを、図 2.4 に示す。 $f(x)$ は、 $x=\mu$ を軸にして左右対称、 $x=\mu$ で最大値、 x が μ から左右に離れるに従って減少し、 $x=\mu\pm 3\sigma$ で極めて小さな値をとる、ベル型のグラフである。

$N(\mu, \sigma^2)$ に従う確率変数 X に

$$U = (x - \mu) / \sigma \quad (2.11)$$

という変換を行うと、 U は $N(0, 1)$ に従う。この確率分布を“標準正規分布”と呼んで、その確率密度関数を $\phi(u)$ 、分布関数を $\Phi(u)$ で表す。

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\} \quad (2.12)$$

で、グラフは図 2.5 のようになる。標準正規分布の上側確率 $1 - \Phi(u)$ を、 $Q(u)$ で表す。

u の値を与えて下側確率 $\Phi(u)$ または上側確率 $Q(u)$ を求める、あるいは逆に $\Phi(u)$ または $Q(u)$ を与えて u の値を求めるための、統計数値表やコンピュータ・ソフトの関数が用意されている。これらは下側か上側のどちらか一方なので、使用にあたっては注意が必要である。標準正規分布は x 軸について左右対称だから、 $\phi(-u) = \phi(u)$ 、 $Q(-u) = \Phi(u)$ 、 $\Phi(-u) = Q(u)$ なので、統計数値表には $u \geq 0$ の値しかないが、実用上は支障がない。

(2) 正規 Q-Q プロット

この講座で説明する推定と検定の手法は、すべて正規分布を想定して導出されている。ところが、栄養学の研究でとられている計量値のデータには、常に正規分布が想定できるわけではない。むしろ、[例 1] の HDL-コレステロール値のように、そうでない場合の方が多いので、データに正規分布を想定してよいかを検証しておく必要がある。歪度と尖度を計算して両者が 0 に近いことを確かめる、幹葉表示を作成して正規分布の確率密度関数のグラフと比較するという方法がある。しかし、データの大きさが 50 程度以上ないと、歪度と尖度の値や幹葉表示の形が安定しない。そこでデータの大きさが 20 程度以上あれば、適用可能な方法を紹介する。大きさが 10 程度のデータについて、正規分布を想定してよいかを検証する有効な方法はないといってよい。

“正規 Q-Q プロット（正規確率プロット）”は、大き

図 2.5 $N(0, 1)$ の確率密度関数のグラフ

さが 20 程度以上の計量値データ

$x_1, x_2, \dots, x_n$; 大きさは n

についてその分布をグラフ表示し、正規分布を想定してよいか、データのなかに外れ値がないかを吟味するための手法である。

【基本手順】

手順 1 データを昇順に並べ換える。

$x_{(1)}, x_{(2)}, \dots, x_{(n)}$; 大きさは n

手順 2 すべての測定値について、測定値 $x_{(i)}$ の“累積比率” p_i を、次の式で計算する。 i は昇順に並べ換えた後の順位で、データに同じ値の測定値が含まれている場合も、機械的に自然数を付与する（同順位や平均順位にはしない）。

$$p_i = i / (n + 1) \quad (2.13)$$

手順 3 すべての測定値について、累積比率 p_i に対応する“正規スコア” u_i ($\Phi(u) = p_i$ となる u の値) を、統計数値表かコンピュータ・ソフトの関数を使って求める。

手順 4 x を横軸、 u を縦軸にとって、 $x_{(i)}$ と u_i を対にしたプロットを作成する。すべての点がほぼ直線に並んでいれば、データに正規分布を想定することができる。

【数値例】

[例 1] の HDL-コレステロール値について、正規 Q-Q プロットを描いてみる。

手順 1 データを昇順に並べ換える。

20	28	28	29	32	34	35	36	36	37
37	37	38	38	39	39	40	40	40	41
42	42	43	43	43	43	44	44	44	45
45	45	46	46	47	47	47	47	48	48
48	48	48	49	49	49	50	50	50	51
52	52	52	52	52	52	52	53	53	53
54	54	55	55	56	57	57	57	57	57
58	59	60	60	62	63	64	64	65	65
67	67	70	70	71	74	74	82	109	110

手順 2 測定値の累積比率を $i / (90 + 1)$ で計算する。

0.011	0.022	0.033	0.044	0.055
0.066	0.077	0.088	0.099	0.110

0.121	0.132	0.143	0.154	0.165
0.176	0.187	0.198	0.209	0.220
0.231	0.242	0.253	0.264	0.275
0.286	0.297	0.308	0.319	0.330
0.341	0.352	0.363	0.374	0.385
0.396	0.407	0.418	0.429	0.440
0.451	0.462	0.473	0.484	0.495
0.505	0.516	0.527	0.538	0.549
0.560	0.571	0.582	0.593	0.604
0.615	0.626	0.637	0.648	0.659
0.670	0.681	0.692	0.703	0.714
0.725	0.736	0.747	0.758	0.769
0.780	0.791	0.802	0.813	0.824
0.835	0.846	0.857	0.868	0.879
0.890	0.901	0.912	0.923	0.934
0.945	0.956	0.967	0.978	0.989

手順3 累積比率の正規スコアを求める。日本規格協会の「簡約統計数値表（すべての表が上側確率で作られている）」を使うなら、 $Q=1-p_i$ とおいた $u(Q)$ の値として、“正規分布のパーセント点”の表から求める。表は $0 \leq Q < 0.5$ の範囲しかないので、 $p_i < 0.5$ の場合は $-u(p_i)$ の値にする。 $p_i = 0.5$ では、 $u_i = 0$ である。コンピュータ・ソフトの関数では、 p_i から直ちに正規スコアを出力するものが多い。

-2.29	-2.01	-1.84	-1.71	-1.60
-1.51	-1.43	-1.35	-1.29	-1.23
-1.17	-1.12	-1.07	-1.02	-0.97
-0.93	-0.89	-0.85	-0.81	-0.77
-0.74	-0.70	-0.66	-0.63	-0.60
-0.56	-0.53	-0.50	-0.47	-0.44
-0.41	-0.38	-0.35	-0.32	-0.29
-0.26	-0.23	-0.21	-0.18	-0.15
-0.12	-0.10	-0.07	-0.04	-0.01
0.01	0.04	0.07	0.10	0.12
0.15	0.18	0.21	0.23	0.26
0.29	0.32	0.35	0.38	0.41
0.44	0.47	0.50	0.53	0.56
0.60	0.63	0.66	0.70	0.74
0.77	0.81	0.85	0.89	0.93
0.97	1.02	1.07	1.12	1.17
1.23	1.29	1.35	1.43	1.51
1.60	1.71	1.84	2.01	2.29

手順4 HDL-コレステロール値を横軸，正規スコアを縦軸にとって，図2.6を作成する。

[解説と補足]

i) 累積比率を i/n ではなく， $i/(n+1)$ で求めるのは，中央値と50パーセント点が一致する，最大値の累積比率が1になって正規スコアが $+\infty$ になるのを回避できるなど，取扱いの上で都合のよい性質を持たせるためである。

図2.6 HDL-コレステロール値の正規Q-Qプロット

ii) コンピュータ・ソフトを使えば，データを昇順に並べ換えずに，データを昇順に並べ換えたときの順位を直接，関数で求めることができる。しかし，データに同じ値の測定値が含まれていた場合の取り扱いが，ソフトによって異なるので，注意が必要である。

[結果の解釈]

図2.6から，[例1]のデータについて，次のことがいえる。

- i) プロットした点は直線上に並んでいないから，データに正規分布を想定することは無理である。
- ii) 上側に長い裾を引いていて，二つの飛び離れた測定値がある。

(3) 対数変換

この講座で説明する手法は，すべて正規分布が想定できることを前提にしているのに，栄養学の研究で処理しなければならないデータは，必ずしもこの前提を満足していない。この現実はどう対処したらよいかには，次の三つの対策がある。

- 1) 前提を満足していないデータに適用したとき，解析結果にどのような影響が現れるかが検討されていて，結論の妥当性が保たれる，つまり“頑健性（ロバストネス）”を持っていることがわかっている手法を選択する。
- 2) データに“変数変換”を行って，変換値に正規分布が想定できるようにする。
- 3) データに“順位変換”を行って，“ノンパラメトリックな手法”を適用する（この講座では紙数の関係でその説明は割愛する）。

変数変換のなかで一番よく使われる“対数変換”を，[例1]のデータに適用してみる。HDL-コレステロール値の，常用対数をとる。

1.78	1.66	1.81	1.83	1.74
1.70	1.79	1.68	1.81	1.65

図 2.7 HDL-コレステロール値の対数変換値の正規 Q-Q プロット

1.57	1.85	1.63	1.59	1.64
1.68	1.85	1.45	1.71	1.76
1.61	1.62	1.60	1.73	1.59
1.81	1.67	1.72	1.76	1.57
1.68	1.56	1.91	1.67	1.70
1.76	1.63	1.30	1.60	1.72
1.57	1.56	1.45	1.72	1.72
1.83	1.67	1.77	1.58	1.80
1.69	1.65	1.87	1.69	2.04
1.51	1.68	1.46	1.72	2.04
1.85	1.72	1.81	1.53	1.72
1.76	1.74	1.67	1.62	1.66
1.71	1.76	1.54	1.73	1.63
1.60	1.65	1.64	1.75	1.72
1.72	1.72	1.87	1.58	1.63
1.64	1.68	1.78	1.76	1.69

この変換したデータについて、正規 Q-Q プロットを作成すると、図 2.7 が得られる。図 2.6 と比較すれば、点はかなりよく直線の上に並んでいる。歪度は -0.023 、尖度は 1.481 なので、分布は左右対称であるが両側の裾は少し重い。しかし、この程度であれば変換したデータに、正規分布を想定して差し支えないだろう。

変換したデータの平均値は 1.692 なので、変換をもどすと 49.2 になる。これは変換前のデータの平均値 51.0 より、中央値 49.0 に近い。変換したデータの標準偏差は 0.117 で、変動係数は 0.069 である。変換前のデータの標準偏差は 14.4 、変動係数は 0.282 だから、対数変換によって変動係数は $1/4$ になった。

現在、栄養学の研究でとられているデータの多くは、化学物質の含量や酵素の活性などの生化学的な測定値である。これらの測定値は負の値をとらないし、データの分布は上側の裾が重いのが普通で、ときに非常に大きな外れ値が現れる。このようなデータに対数変換を行う

と、変換したデータには正規分布を想定できるようになる場合が多い。このときもとのデータに想定している確率分布を、“対数正規分布”と呼んでいる。

上側の裾が重い分布のデータを対数変換をすれば、常に正規分布が想定できるわけではない。対数変換を適用すると、変換後のデータの分布の下側の裾が重くなってしまう場合は、もっとおだやかな方法として、測定値の平方根をとる“平方根変換”がある。また、対数変換を行ってもまだ上側の裾が重い場合、より強力な方法として、測定値の逆数をとる“逆数変換”をほどこすことができる。逆数をとると大小が逆転するから、グラフ標示や結果の解釈にあたっては注意が必要である。

データに含まれる測定値に 0 がある場合、そのままでは対数変換や逆数変換を適用することができない。そこで対数変換では、測定値全体に一定の数（たとえば 1 ）を加えてから変換を行うか、 0 の測定値を他の測定値より小さな値（たとえば測定値の検出限界）に置き換えてから変換を適用することが試みられている。逆数変換でも同じことはできるが、加える数や置き換える数の選び方による影響が大きいため注意が必要である。いずれにせよ、 0 の測定値が何個も含まれているデータを、変数変換によって正規分布が想定できるようにすることには無理がある。

栄養学の研究でとられるデータにも、下側の裾が重い分布をするものもある。ヒトの出生時の体重の分布は、その典型である。ヒトや実験動物に、生化学的な測定値を低下させる効果のある薬物を投与した場合、薬物に対する反応の個体差から、その測定項目のデータは下側の裾が重い分布をすることがある。

2.3 母平均の区間推定

【例 1】のデータとほぼ同じ条件にある非常に多数の成人男子について、HDL-コレステロール値を測定してその平均値を求めたらどんな値になるか、つまり母平均（データに想定した確率分布の平均値）を、データから推定する問題を考える。

データの平均値を母平均の推定値にするのが普通であるが、それがどの程度の信頼性を持つのか明らかにしておくことが望ましい。そのために、母平均の“信頼区間”を求める“区間推定”が行われる。

(1) 母平均の信頼区間を求める手順

計量値のデータ

$x_1, x_2, \dots, x_n$; 大きさは n

が得られたとき、このデータに正規分布 $N(\mu, \sigma^2)$ が想定できるとする。

【基本手順】

手順 1 平均値 $\bar{x}$ 、平方和 S とその自由度 ν 、分散 V 、標準偏差 SD 、標準誤差 SE を求める。 SE は、次の式で計算する。

$$SE = \sqrt{V/n} = SD/\sqrt{n}$$

図 2.8 データの分布と平均値の分布

図 2.9 $N(0,1)$ と t 分布

手順2 自由度 ν の t 分布の上側 2.5% 点の値 $t(\nu, 0.025)$ を、統計数値表かコンピュータ・ソフトの関数を使って求める。表にない自由度に対しては、 ν についての線形補間を次の式で行う。 ν_L は表で ν より小さくて ν に最も近い自由度、 ν_U は表で ν より大きくて ν に最も近い自由度である。

$$\begin{aligned} t(\nu, 0.025) &= t(\nu_L, 0.025)(\nu - \nu_L) / (\nu_U - \nu_L) \\ &\quad + t(\nu_U, 0.025)(\nu_U - \nu) / (\nu_U - \nu_L) \end{aligned} \quad (2.14)$$

手順3 母平均の推定値 $\hat{\mu}$ と、その“信頼係数”95%の“信頼区間”を、次の式で計算する。

$$\begin{aligned} \hat{\mu} &= \bar{x} \\ (\hat{\mu} - t(\nu, 0.025)SE, \hat{\mu} + t(\nu, 0.025)SE) \end{aligned} \quad (2.15)$$

[数値例]

[例1] の HDL-コレステロール値を対数変換した変換値のデータに、正規分布 $N(\mu, \sigma^2)$ を想定して、母平均の 95% 信頼区間を求める。

手順1 $\bar{x} = 1.692$,

$$S = 1.22030, \nu = 89, V = 0.01371$$

$$SD = 0.117, SE = 0.0124$$

手順2 「簡約統計数値表」の“ t 分布のパーセント点”から t 分布の上側 2.5% 点の値 $t(80, 0.025) = 1.990$ と $t(120, 0.025) = 1.980$ を求めて、 $t(89, 0.025)$ を線形補間する。

$$\begin{aligned} t(89, 0.025) &= 1.990 \times (89 - 80) / (120 - 80) \\ &\quad + 1.980 \times (120 - 89) / (120 - 80) \\ &= 1.982 \end{aligned}$$

手順3 $\hat{\mu} = \bar{x} = 1.692$

$$t(\nu, 0.025)SE = 1.982 \times 0.0124 = 0.025$$

$$(1.692 - 0.025, 1.692 + 0.025)$$

$$= (1.667, 1.717)$$

[解説と補足]

i) 信頼区間を CI と略記することがある。信頼区間の両端を“信頼限界”という。

[結果の解釈]

i) 変換値のデータに想定した正規分布の母平均の推定

値は 1.692 で、母平均は 1.668 と 1.716 の間にあると考える。

ii) 推定値の変換をもどすと 49.2 で、信頼区間の変換をもどすと (46.5, 52.1) になる。これは母幾何平均の推定値とその 95% 信頼区間である。

(2) 区間推定の仕組み

データに正規分布 $N(\mu, \sigma^2)$ を想定するということは、各測定値は互いに独立に $N(\mu, \sigma^2)$ に従う確率変数 $X_1, X_2, \dots, X_n$ の実現値であると考えることである。データの平均値 $\bar{x}$ に対応する確率変数の関数

$$\bar{X} = (X_1 + X_2 + \dots + X_n) / n \quad (2.16)$$

も正規分布をして、その平均値と分散は

$$E(\bar{X}) = \mu \quad (2.17)$$

$$V(\bar{X}) = \sigma^2 / n \quad (2.18)$$

になる。つまり、 $\bar{x}$ は $N(\mu, \sigma^2/n)$ に従う。

データの分布と平均値の分布を、 $n=4$ の場合についてグラフに示すと、図 2.8 のようになる。

$\bar{X}$ に

$$U = (\bar{X} - \mu) / (\sigma / \sqrt{n}) \quad (2.19)$$

という変換を行うと、その分布は標準正規分布 $N(0,1)$ になる。式 (2.19) の $\sigma / \sqrt{n}$ を、データの関数である SE に置き換えた

$$t = (\bar{X} - \mu) / SE \quad (2.20)$$

の分布は、自由度 ν の“ t 分布”をする。

$N(0,1)$ と $\nu=3$ の t 分布のグラフを、図 2.9 に示す。 t 分布は x 軸について左右対称で、中心部分は $N(0,1)$ よりやや小さく、両側の裾はやや重い。 ν が大きくなるにつれて $N(0,1)$ に近づき、 $\nu > 25$ でほぼ一致する。

式 (2.20) で表される確率変数 t の絶対値が、 t 分布の上側 2.5% 点の値 $t(\nu, 0.025)$ 以上になる確率は

$$P\{|t| \geq t(\nu, 0.025)\} = 0.05$$

で、書き改めると

$$P\{|\bar{X} - \mu| \geq t(\nu, 0.025)SE\} = 0.05$$

になる。これから、95% 信頼区間が得られる。

$$\begin{aligned} P\{\bar{X} - t(\nu, 0.025)SE < \mu \\ < \bar{X} + t(\nu, 0.025)SE\} = 0.95 \end{aligned} \quad (2.21)$$

この式は、同一条件でデータをとる実験を繰り返して、(1) で説明した手順で母平均の信頼区間を求める試

行を続けたら、およそ95%の信頼区間が母平均を含み、およそ5%は母平均を含まないことを意味している。現実には、統計学の理論が保証しているわけではないが、1回の実験で得られたデータから母平均の信頼区間を求めて、母平均はほぼ確実にそこに含まれると考えている。

区間推定で信頼係数を95%にとるのは、慣行としてのルールである。信頼係数が高いことは望ましいが、そうすると信頼区間の幅が広がって実用上の意味を失うから、両者のかねあいで95%が採用されている。

(3) 母平均の信頼区間の頑健性

母平均の信頼区間を求める(1)の手順は、“データの正規性”(データに正規分布を想定できること)を前提にしている。もし、この前提を満足していないデータに適用したら、どのような結果になるだろうか。

正規分布を前提にしなくても、 $\bar{X}$ について式(2.17)と式(2.18)が成立する。 $\bar{X}$ の分布は $N(\mu, \sigma^2/n)$ にはならないが、 n が大きくなると $N(\mu, \sigma^2/n)$ に近づいてゆく。 n がどの程度大きくなれば $N(\mu, \sigma^2/n)$ とみなしてよいかは、データの分布がどれだけ正規分布から外れているかによって決まる。ほとんどのデータは、 n が10以上であれば、 $\bar{X}$ の分布を $N(\mu, \sigma^2/n)$ とみなすことができる。そのとき、式(2.19)の分布は $N(0, 1)$ とみなして差し支えないから、式(2.20)の分布は t 分布に近いと考えられる。

以上のことから、母平均の信頼区間を求める(1)の手順は、正規性の前提に対して頑健である。[例1]のHDL-コレステロール値のデータを、対数変換せずに95%信頼区間を求めると

$$\hat{\mu} = \bar{x} = 51.0$$

$$t(\nu, 0.025)SE = 1.982 \times 1.53 = 3.0$$

$$(51.0 - 3.0, 51.0 + 3.0) = (48.0, 54.0)$$

になる。データに正規分布が想定できなくても、 n が十分に大きいから、この信頼区間は妥当である。

2.4 母平均の検定

日本人の健康な成人男子について、非常に多くの人のHDL-コレステロール値を測定して得られた平均値が、47.5 mg/dL (架空の数値)であったとする。このとき、[例1]の事業所に勤務している健康な成人男子の平均値と、日本人の健康な成人男子との平均値の差は、統計学的に意味のあるものなのか、それとも偶然によるものなのかを、データから確かめる問題を考える。

この問題は、「[例1]のデータに想定した確率分布の平均値は、日本人の健康な成人男子との平均値に等しい」という仮説(これを“帰無仮説”と呼んで、 H_0 で表す)を、「[例1]のデータに想定した確率分布の平均値は、日本人の健康な成人男子との平均値と異なっている」という仮説(これを“対立仮説”と呼んで、 H_1 で表す)に対して“検定”する(データから“検定統計

量”の値を求めて、“棄却限界値”と比較する)という手続きで処理される。

(1) 母平均の検定を行う手順

計量値のデータ

$$x_1, x_2, \dots, x_n; \text{大きさは } n$$

が得られたとき、このデータに正規分布 $N(\mu, \sigma^2)$ が想定できるとする。このとき

$$\text{帰無仮説 } H_0: \mu = \mu_0 \quad (\mu_0 \text{ は既知の常数}) \text{ を,}$$

$$\text{対立仮説 } H_1: \mu \neq \mu_0$$

に対して検定する。

[基本手順]

手順1 平均値 $\bar{x}$ 、平方和 S とその自由度 ν 、分散 V 、標準偏差 SD 、標準誤差 SE を求める。

手順2 自由度 ν の t 分布の上側2.5%点の値 $t(\nu, 0.025)$ を、統計数値表かコンピュータ・ソフトの関数を使って求める。表にない自由度に対しては、 ν についての線形補間を行う。

手順3 検定統計量 t_0 を、次の式で計算する。

$$t_0 = |\bar{x} - \mu_0| / SE \quad (2.22)$$

手順4 $t(\nu, 0.025)$ の値を“有意水準”5%の棄却限界値として、 $t_0 \geq t(\nu, 0.025)$ であれば、帰無仮説 H_0 を棄却して、母平均 μ は μ_0 とは異なっていると判断する。 $t_0 < t(\nu, 0.025)$ であれば、帰無仮説 H_0 を採択する。

[数値例]

[例1]のHDL-コレステロール値のデータに、正規分布 $N(\mu, \sigma^2)$ を想定して

$$\text{帰無仮説 } H_0: \mu = 47.5 \text{ を,}$$

$$\text{対立仮説 } H_1: \mu \neq 47.5$$

に対して検定する。

手順1 $\bar{x} = 51.0$,

$$S = 1.22030, \nu = 89, V = 207.0$$

$$SD = 14.4, SE = 1.53$$

手順2 $t(\nu, 0.025) = 1.982$

手順3 $t_0 = |51.0 - 47.5| / 1.53 = 2.288$

手順4 $2.288 \geq 1.982$ であるから、帰無仮説 H_0 を有意水準5%で棄却する。

[解説と補足]

i) 対立仮説 $H_1: \mu \neq \mu_0$ は、「 $\mu > \mu_0$ または $\mu < \mu_0$ 」と同等なので“両側対立仮説”と呼び、 H_1 に対してを H_0 を検定することを“両側検定”という。区間推定と両側検定の手順を比較すれば明らかなように、 μ_0 が95%信頼区間に含まれないということと、 $H_0: \mu = \mu_0$ が両側検定で有意水準5%で棄却されることは同等である。

ii) 対立仮説 $H_2: \mu > \mu_0$ に対して H_0 を検定するには、検定統計量を $t_0 = (\bar{x} - \mu_0) / SE$ にして、 $t_0 \geq t(\nu, 0.05)$ であれば、 H_0 を棄却する。

対立仮説 $H_3: \mu < \mu_0$ に対して H_0 を検定するには、検定統計量を $t_0 = (\mu_0 - \bar{x}) / SE$ にして、 $t_0 \geq t(\nu, 0.05)$

であれば、 H_0 を棄却する。

H_2 や H_3 を“片側対立仮説”と呼び、 H_2 や H_3 に対して H_0 を検定することを“片側検定”という。

- iii) 帰無仮説を採択するということは、 H_0 が正しいと主張するのではなく、データは H_0 が誤りであるとする証拠にはならないという意味である。
- iv) 検定のことを“有意差検定”と呼んで、帰無仮説が棄却された場合は“有意差あり”，採択された場合は“有意差なし”ということがある。

[結果の解釈]

- i) [例1]のデータに想定した確率分布の平均値は、日本人の健常な成人男子との平均値47.5 mg/dLとは異なっている。したがって、HDL-コレステロール値について、この事業所の健常な成人男子は、日本人の健常な成人男子全体からランダムに抽出された集団とみなすことはできない。
- ii) 正規分布が想定できる、対数変換した変換値のデータに、母平均の検定を適用した方が望ましい。そのためには、日本人の健常な成人男子との平均値も、対数変換した変換値のデータについて求められていなければならない。未変換のデータから求めた平均値47.5の対数1.677を μ_0 として、[例1]の対数変換した変換値のデータを検定するのは誤りである。

(2) 有意性, 信頼区間, p 値

検定という方法は、帰無仮説が棄却されるか採択されるかという二者択一の結果を導くので、適用にあたっては注意が必要である。

対立仮説が真であるのに、帰無仮説が採択される理由としては、次のことが考えられる。

- 1) 検出しようとする差に対して、データの大きさが小さすぎる (例数が不足している)。
- 2) 検出しようとする差に対して、データのばらつきが大きすぎる。
 - ①データに想定すべき分布が対数正規分布のように片側の裾が重いのに、適切な変数変換をしていないので、データのなかに極端な外れ値が含まれている。
 - ②データのなかに、測定値の記入間違いや入力誤りのような、“異常値”が含まれている。

③実験そのものに、条件の管理が不十分である、測定法が不安定であるなどの問題がある。

こうした面からの検討をせずに、有意差がなかったというだけで対立仮説の検証をあきらめてしまうと、価値の高い研究成果を逃してしまう危険がある。反対に帰無仮説が棄却されても、検出している差は実質的に意味のない僅少な差である可能性もある。

これに対して区間推定という方法は、検定より多くの情報を提供してくれる。[例1]のHDL-コレステロール値のデータから、検定で $H_0: \mu=47.5$ が有意水準5%で棄却されるということと、区間推定で μ の95%信頼区間が(48.0, 54.0)であるということを比較してみる。47.5は区間(48.0, 54.0)に含まれないから、検定の手順を経なくても $H_0: \mu=47.5$ が棄却されることがわかる。そればかりでなく、[例1]のデータに想定した確率分布の母平均の推定値51.0と47.5の差3.5の95%信頼区間は(0.5, 6.5)であることがわかる。この差とその信頼区間が、実質的にどのような意味を持つかを検討することは、栄養学研究の課題になるだろう。

以上のような事情で、バイオサイエンスの研究で統計的データ解析を適用するなら、検定よりも区間推定を行うことが望ましいというのが、最近の“biostatistics”の動向である。

しかし、正規性の検定、分布の適合度の検定、分割表の独立性の検定など (この講座では割愛している) のように、未知母数で表現できない問題もある。このような場合には区間推定は適用できない。そこで、有意差があったかなかったか (*が付いたか付かなかったか) という二者択一の結果ではなく、“ p 値”を求めることが薦められている。 p 値とは、帰無仮説のもとで検定統計量がデータから求めた検定統計量の実現値を越える確率、つまり本当は差がないのに観測された差が偶然で生ずる確率のことで、 p 値が0.05以下であることと、帰無仮説が有意水準5%で棄却されることは同等である。

[例1]のデータで $H_0: \mu=47.5$ のもとで t_0 が2.288を超える確率は、 t 分布の2.288に対する上側確率と-2.288に対する下側確率の和になるから、 p 値は0.0246である。

(つづく)