医師のための統計入門 (草稿)

富山大学附属病院 初期臨床研修医 川口 真一

初版 2016.11.27
最終改訂 2017.02.05

目次

1 序

現代の臨床医療では、Evidence-Based Medicine (EBM) の名の下に、統計学的根拠に基づく診療を重視する風潮がある。この風潮に迎合するにせよ、あるいは批判して反発するにせよ、統計というものを知らなければ、どうにもならない。また、医学に限らず、実験科学の分野においては、実験結果を統計学的に解釈することが重要視されている。こうした事情から、臨床にせよ研究にせよ、医療や医学を担う医師にとっては、初等的な統計学は必須の教養であるといえる。

特に問題なのが、臨床診療ガイドラインの扱いである。多くのガイドラインは統計学的調査報告を根拠に作成されている。しかし統計学的調査というものは、実に主観的であり、恣意の入る余地が多い。従って、統計学をよく理解していなければ、ガイドラインを適切に解釈し運用することなど不可能である。

ところが現行の医学部医学科における教育では、統計学は極めて軽んじられている。理由は知らぬ。その結果として、若い医師の多くは統計について無知である。統計学的調査というものが、あたかも客観的で信頼できるものであるかのように誤解しているのである。その結果、不適切な診療行為が横行し、しかも、それを行っている医師には、その自覚がない。

遺憾ながら、医学科生や医師向けの優れた統計学の入門書は多くない。工学部などの人々が読むような教科書は、数学に疎い一般の医学科生や医師には難しすぎる。かといって「明日から役立つ」などと銘打ったアンチョコ本は、統計学の真髄から著しく乖離しており、統計をブラックボックス化して誤用させるだけの代物である。中庸を行く書物としては、新谷歩『今日から使える医療統計』(医学書院; 2015) がある。しかし、この書物は数式を廃して「わかりやすい」イメージを与えることに専念したために、厳格な議論からは外れてしまい、結局、統計学の本質には迫ることができていないように思われる。

私は、たまたま京都大学工学部で原子核物理学を学び、同大学大学院で原子炉物理学を修めた経験[*1]から、一般の医師に比べれば、統計学に長じている。むろん、理学部で統計を専門に研究している人々や、工学部で日々、統計を駆使している人々などに比べれば、私などは素人同然である。しかし、臨床医学・医療の現場と、本格的な統計学の片鱗とを共に修めた立場として、いわゆる学際的な観点から医学における統計を論じるには、私は適任であろう。

そこで、医学科生や若い医師のための統計学入門書を著そうというのが、本文書の趣旨である。統計学的な厳格さを保ちつつ、かつ、臨床医学で必要とされる範囲を大きく逸脱しないことを目標とする。たとえば「中心極限定理」は非常に重要であるが、その証明自体は、臨床医学に必要ではないから、割愛する。こういう態度は、理学部の諸君からすれば下劣であり、「だから医学部は程度が低い」などと言われ、侮蔑されるであろう。その通りである。我々は、崇高な学問を修める純粋科学者ではない。そのことに対する引け目と羞恥の心、そして基礎科学に生涯を捧げる科学者諸兄姉に対する敬意を、我々は忘れてはならぬ。

なお、本文書は、富山大学附属病院の同期研修医諸君と共に開催した統計学勉強会における議論の産物である。勉強会の参加者諸兄に、心より感謝申し上げる。

[*1] ただし博士課程は三年在籍した上で中途退学したので、博士の学位は有していない。肩書は「修士 (エネルギー科学)」である。弁明しておくが、私が中退したのは、学問上の諸問題を巡る衝突から、教授との関係を修復することが困難になったためである。私の学識が不足したために修了できなかったわけではない。

2　t 検定

2.1　t 検定の基本形

2.1.1　標準偏差

次のような状況について考えよう。

> 富山大学附属病院の K 研修医は、「杉谷 (富山大学附属病院の所在地) の猫の出生時体重は平均 110.0 g である。」と主張した。これに対し F 研修医は「そんなはずはない。杉谷の猫は、もっと小さいはずだ。」と考え、K 研修医の意見を否定する証拠を集めることにした。
>
> F 研修医は、病院周辺の山中で 101 匹の猫の出生に立ち会い、その出生時体重を調べた。すると、新生仔の平均体重は 100.0 g であった。「ほらみろ、君の意見は、間違いだ。」と言う F 研修医に対して、K 研修医は「110.0 g と 100.0 g の差などは、統計誤差の範疇だ。私の意見が間違っているということにはならない。」と反論した。F 研修医が「統計誤差とは、どういう意味だい。」と問えば、K 研修医は「つまり、偶然のばらつき、という意味だよ。」と答えた。
>
> もちろん F 研修医は、そのまま引き下がるようなことはしない。自分が集めた 101 匹の猫の出生時体重のばらつきを調べたところ、標準偏差は 30.0 g であった。

101 匹の猫の体重の標準偏差 σ というのは、次の式で計算される量である。

$$\sigma = \sqrt{\sum_{i=1}^{101} (w_i - \bar{w})^2} \tag{1}$$

ここで $\bar{w}$ というのは、F 研修医が調べた 101 匹の猫の、平均の出生時体重である。w_i は、i 番目の猫の出生時体重である。統計データのばらつきを示す指標としては、この標準偏差が用いられることが多い。式 (1) の意味を日本語で表現すれば、「それぞれの猫について、『調べた猫全体の平均出生時体重からのずれ』の 2 乗、の平均を標準偏差と呼ぶ」ということである。なぜ 2 乗するのか、という点には、深い意味はない。ただ、何もせずに単に足し算 ($\sum$ 記号) をすると常に $\sigma = 0$ になってしまうので、それを避けるため、ぐらいの意味である。だから、2 乗ではなく「平均からのずれの絶対値」などを用いても、悪くはない。あるいは、4 乗する、という発想も、あり得る。ただ、慣習的に 2 乗しているだけのことである。

ところで、本文書では「正規分布とは何か」というような話は、しない。それほど厳密な数学的議論をするわけではないから、漠然とグラフの形だけ想起していただければ充分である。なお、平均 m, 標準偏差 σ の正規分布のことを $N(m, \sigma^2)$ という記号で表すことにする。σ^2 というのは、分散と呼ばれる量であって、標準偏差の二乗として定義される。

2.1.2　中心極限定理

さて、杉谷の全ての猫について出生時体重の調査を行うことなど不可能なのだから、「真の平均出生時体重」というのは、我々には知り得ない、天の神様だけが知っている値である。F 研修医のデータでは、出生時体重は平均 100.0 g であったが、もちろん、この値と「真の平均体重」との間にはズレが存在する。その「ズレ」は、いかほどであろうか。この問題は非常に難しいのであるが、幸いなことに、過去の数学者が、とても難しい議論の末に、理論的な解答を導いてくれている。これが「中心極限定理」というものであり、次のような内容である。

N 個のサンプルを集めて平均を調べた場合、その「サンプルの平均」と「真の平均」とのズレは、次の式で表される正規分布に従って確率的に決まる。[*2]。

$$f(w) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{w^2}{2s^2}\right) \tag{2}$$

この正規分布は、平均が 0 であり、標準偏差 s は

$$s = \frac{\sigma}{\sqrt{N-1}} \tag{3}$$

と推定するのが妥当である[*3]。ここで σ はサンプルの標準偏差であり、式 (1) で計算した値である。

この定理を証明することは、多大な労力を要する割に臨床医学的にはあまり重要でないと思われるので、本文書では避ける。サンプル数が多ければ多いほど、サンプルの平均と真の平均は近い値をとりやすい、という直観的な事情だけ理解していれば充分であろう。なお、分母が N ではなく $N-1$ なのは、σ を計算する過程で w_i と $\bar{w}$ の差を使ったからである。すなわち、仮に $N=1$ であれば必ず $\sigma=0$ になるし、$N=2$ の場合は 2 つのデータの差だけが問題になるのであって、各々のデータの絶対値自体は σ に影響しない。このように、σ や s の計算に際しては、実際に有効なデータの数は N 個ではなく $(N-1)$ 個なのであって、それが式 (3) に反映されているのである。

さて、中心極限定理のいう「確率的に決まる」というのは、どういう意味か。これは、サンプルの平均 $\bar{w}$ が $w_1 < \bar{w} < w_2$ の範囲に収まる確率 p が、式 (2) の $f(w)$ を用いて

$$p = \int_{w_1}^{w_2} f(w)\mathrm{d}w \tag{4}$$

と表せる、という意味である。

なお、数学が苦手で、積分記号をみると嘔気を催す、というような人は、

$$p = f(\bar{w}) \tag{5}$$

であると認識してしまっても構わない。これは「確率」と「確率密度」を混同しているという点で数学的には誤りなのだが、臨床医療では、遺憾ながら、それほど厳密な数学的議論が要求されることは稀だからである。

2.1.3 95 % 信頼区間

さて、F 研修医が得た「平均 100.0 g」というデータは、どのくらい、信頼できるのだろうか。標準誤差 s を計算すると

$$s = \frac{\sigma}{\sqrt{N-1}} = \frac{30\mathrm{g}}{\sqrt{101-1}} = 3.0\mathrm{g} \tag{6}$$

となる。従って、標準誤差を用いて誤差範囲を示すことにすれば、杉谷の猫の平均出生時体重は「100.0 ± 3.0 g」ということになる。問題は、これを根拠に「杉谷の猫の平均出生時体重は 110.0 g だ」という K 研修医の主張を否定することはできるだろうか、という点である。換言すれば、標準誤差が 3.0 g である場合に、真の平均とサンプル平均が 10.0 g、つまり標準誤差の 3.3 倍もズレることは、あり得るだろうか、という問題である。

[*2] 言うまでもないことだが、この式は暗記するべきではない。必要な時に、教科書なり Wikipedia なりを調べれば済む話だからである。我々の頭脳は、こんなものの暗記より、もっと大事なことに使うべきである。

[*3] 統計学に長けている人々は、この説明に対して不満であろう。しかし不偏分散などを議論することは、一般的な医師の理解の範疇を超えると思われるので、敢えて割愛した。また、臨床医学的には、母集団は非常に大きいことが前提なので、簡略化のため、初めから近似しておいた。

ここで、このズレが正規分布に従う、という中心極限定理が活躍するのである。式 (2) と式 (4) に従って計算すると、95 % の確率で、ズレは標準誤差の 1.96 倍以内に収まることがわかる。数学が大好きではない読者は、「そんな計算、面倒くさくて、やってられないよ」と思うであろう。それでよろしい。昔の人も同じように思っていたらしく、「正規分布表」という便利なものが発明された。これは統計の教科書や理科年表に載っているし、あるいはインターネット上で検索すれば容易に手に入る。とにかく、この表をみれば、95 % に対応するのは 1.96 倍だ、と、すぐにわかるのである。これが t 検定である。

2.1.4 否定はできるが肯定はできない (t 検定における p 値)

さて、二人の研修医の対話は、どうなっただろうか。

> 以上のことから、F 研修医は K 研修医に対して「t 検定を行ったところ、95 % の信頼度で、君の考えは誤りであるといえる。」と主張した。すると K 研修医は「ウムム、わかった。私の考えが間違いであったことは認めよう。しかし、100.0 g というのは、やはり小さすぎるような気がする。本当の平均は 105.0 g ぐらいなのではないかね。」と抵抗を示した。

もし真の平均が 105.0 g であるならば、F 研修医のサンプル平均とのズレは 5.0 g であって、標準誤差の 1.67 倍である。正規分布表によれば、この範囲にズレが収まる確率は 90.5 % であるらしい。言い換えれば、ズレがこの範囲から外れる確率は 9.5 % である。この確率のことを「p 値」と呼ぶことにしよう[*4]。あまりキチンとした根拠はないのだが、伝統的に、こういう検定は 95 %、あるいは p 値でいえば 0.05 を閾値として用いることが多いから、$p = 0.095$ というのは、ちょっと大きい。

> F 研修医は「まぁ、信頼度 95 % で検定する限りは、その可能性は否定できないね。」と答えた。すると K 研修医は「つまり、100.0 g が正しい値だ、という証拠はないわけだね。フフフ。」と、いやらしい嗤いを浮かべた。

検定というのは、こういうものである。「110.0 g」という説を否定することはできるが、「100.0 g」という仮説を積極的に裏づけることは、できないのである。そのような観点から、検定においては「これから否定しようとしている仮説」のことを「帰無仮説」と呼ぶ。そして、相応の根拠に基づいて帰無仮説を否定することを「帰無仮説を棄却する」と表現する。

2.1.5 意味のない検定

鋭い人は気付いたであろうが、実は「杉谷の猫の平均出生体重は x g である」という仮説は、充分に多数のサンプルを集めれば、x の値によらず、統計的に否定することができる。というのも、N を大きくすれば、 s をいくらでも小さくできる一方、$\bar{w}$ が厳密に x と一致することは、現実には、まず起こり得ないからである。従って、「杉谷の猫の平均体重は x g である」という仮説が誤りであることは初めから明らかなのであって、わざわざ統計を調べて検定するまでもない。検定すること自体が無意味なのである。

これに対して、「杉谷の猫の平均出生体重は $x \pm d$ の範囲にある」という仮説は、検定する意義がある。もし「真の平均値」が $x \pm d$ の範囲に入っているならば、いくら N を大きくしても、この仮説を否定することはできないからである。

検定を行う際には、そもそも、それが意味のある検定なのかどうか、という点について注意しなければならない。

[*4] p 値を厳密に定義しようとすると、小難しい数学を駆使しなければならないので、ここでは曖昧に済ませる。

2.2 二群の平均値の比較

2.2.1 正規分布の再生性 (加算の場合)

ある日、富山大学附属病院の F 研修医は「どうも杉谷 (富山大学医学部の所在地) の猫の出生時体重は、五福 (富山大学工学部の所在地) の猫の出生時体重よりも、少しだけ重いような気がする。」と言い出した。すると K 研修医は「そんなことはないと思うが、もし本当にそう思うなら、統計をとってみてはどうかね。」と言った。

F 研修医が杉谷で多数の猫の出生時体重を測定したところ、平均 105.0 g, 標準誤差 3.0 g であった。また、五福でも同様に測定したところ、平均 102.0 g, 標準誤差 2.5 g であった。

まず我々が知りたいのは、杉谷の猫と五福の猫の出生時体重は等しいのかどうか、という問題である。そこで「杉谷と五福では、猫の出生時体重の『真の平均』は等しい」という帰無仮説を設定し、これを否定するための検定を行うのが適切であると思われる。言い換えれば、「真の平均」が等しい場合に、杉谷ではサンプル平均 105.0 g, 五福ではサンプル平均 102.0 g, というようなバラツキが生じることが、ありそうかどうかを調べれば良い。問題の状況からいって、なんとなく、t 検定が使えそうな感じがするであろう。しかし、今回は 2.1 節で考えた問題とは異なり、杉谷群にも五福群にも統計誤差があるのだから、そのままでは t 検定できない。

そこで「杉谷群と五福群の出生時体重の差」は何 g であるかを考えよう。当然、統計誤差つきで、平均と標準誤差を評価するのである。この問題は、一見、簡単そうにみえるかもしれないが、実はなかなかヤヤコシい。ただ幸いなことに、過去のエラい数学者が「正規分布同士を足し算した結果は、やはり正規分布である」という定理を証明してくれている。我々の例でいえば、中心極限定理より、杉谷群の真の出生時平均体重は 105.0 ± 3.0g, 五福群の真の出生時平均体重は 102.0 ± 2.5g という正規分布で与えられている。従って、両者の差も、何らかの正規分布で表現することができる、というのである。この正規分布の性質は、統計学の世界では「正規分布の再生性」と呼ばれている。

では、この「両群の真の平均の差」を表す正規分布の平均値は、いくらであろうか。直観的には「105.0 g と 102.0 g の差をとって 3.0 g ではないか」と思うであろう。実は、それで正しい。数学が大好きでキチンとした証明が気になる人は、ぜひ統計学の教科書を読んでいただきたい。一般の医師であれば、そこまで数学的にキチンとしている必要まではないと思うので、次の定理をメモ帳か何かに書いておいて、イザという時はチラリと読むことができるようにしておこう。すなわち、2 つの正規分布 $N(m_1, \sigma_1)$, $N(m_2, \sigma_2)$ で確率的に与えられる 2 つの変数を足し算した結果は正規分布 $N(m_s, \sigma_s)$ で確率的に与えられ

$$m_s = m_1 + m_2 \tag{7}$$

である。足し算と引き算は、数学的には符号が反転するだけで同じことである。

問題は、この新しい正規分布の標準偏差 σ_s である。誤差を持っているもの同士を足し算や引き算すれば、結果の誤差は、元の誤差よりも、いくぶん大きくなるであろうことは予想できる。だから、この新しい正規分布の標準偏差は、たぶん 3.0 g や 2.5 g よりも、少しだけ大きいはずである。これについても、数学的議論は医師の手には余るので、次の式をメモしておくことにしよう。

$$\sigma_s^2 = \sigma_1^2 + \sigma_2^2 \tag{8}$$

杉谷と五福の猫の例でいえば

$$\sigma_s = \sqrt{(3.0\text{g})^2 + (2.5\text{g})^2} = 3.9\text{g} \tag{9}$$

ということになる。

2.2.2 二群間の t 検定

さて、杉谷と五福の猫の件であるが、「出生時体重の真の平均の差」は、正規分布 $N(3.0\mathrm{g}, (3.9\mathrm{g})^2)$ で予想される。後は、2.1 節で議論した t 検定である。「真の平均の差は 0 である」という帰無仮説を検定すると、$p = 0.45$ となる。到底、帰無仮説を棄却することはできない。

> K 研修医は「君の予想は、間違っていたようだね」と言った。すると F 研修医は「そんなはずはない。これは、標本数が少なかったために統計誤差が大きくなってしまっただけであろう。」と言い、さらに多くの猫の出産現場を観察しに行った。
>
> 一ヶ月後、F 研修医は満面の笑みを浮かべて K 研修医に報告した。「さらに多くのデータを集めたところ、杉谷の猫は 106.0 ± 1.5g, 五福の猫は 101.5 ± 1.6g と出たよ。」

この新しいデータに基づいて両群の真の平均の差を考えると

$$m_s = 106.0\mathrm{g} - 101.5\mathrm{g} = 4.5\mathrm{g} \tag{10}$$

$$\sigma_s = \sqrt{(1.5\mathrm{g})^2 + (1.6\mathrm{g})^2} = 2.2\mathrm{g} \tag{11}$$

となる。これで t 検定を行えば、$p = 0.040$ となる。これならば、杉谷と五福では猫の出生時体重の平均が異なるといえそうである。

2.2.3 実用性の問題

統計学の議論だけでいうならば、上述のような方法で、二群の平均値に差があるかどうかを t 検定で比較することができる。しかし、よく考えると、この検定は臨床医学においては二つの理由で実用的ではない。

第一に、特に実際の患者について統計を取る場合には、二つの群が厳密に等しい平均値を持っていることなど、あり得ない。これは、比較する項目が死亡率であろうが、入院期間であろうが関係ない。患者を二つの群に分けた時に、本当に厳密に等しい平均値を持ち得るのは、ランダム割付し、かつ、全く同じ治療を実施した場合に限られる。多少なりとも違う薬や違う治療方法を用いたなら、少しは違いが出て当然なのである。そして少しでも違いが出るならば、標本数を充分に多くすることで、理論上、統計学的に検出可能である。従って、「両群に差があるかどうか」という問いに対しては、統計など調べなくても「差はある」と、自信を持って答えることができるのである。

もちろん、差があるという前提で「どちらの薬の方が優れているか」というようなことを調べたいのなら、統計を調べる意義はある。その場合、いずれ述べることになる「優越性試験」などを実施しなければならないのであって、t 検定では意味がない。

第二に、この検定は平均値だけを比較していることも、臨床医学では問題になる。先の猫の例でいえば、たとえば杉谷の猫の出生時体重は $N(105.0\mathrm{g}, 8.0\mathrm{g})$ の正規分布であり、五福の猫の出生時体重は $N(105.0\mathrm{g}, 1.0\mathrm{g})$ の正規分布であったとする。この場合、平均は両者とも等しいのだから、いくらたくさんの猫を集めて t 検定を行っても、「有意差なし」という結論になる。しかし、標準偏差は杉谷の方が圧倒的に大きく、つまり体重のバラツキが大きいのだから、出生時体重の分布は杉谷と五福で大きく異なる、とみるべきである。こうした相違が、t 検定では検出できないのである。

2.3 対応のある t 検定

3 独立性の検定

3.1 二項分布

3.2 二項分布に基づく t 検定

3.3 χ^2 検定

3.4 独立とは何か

以下、執筆中