

標準偏差(SD : standard deviation)と標準誤差(SE : standard error of mean)と信頼区間(CI: Confidence interval) (070705)

SDとSEの違いを知る

多くの医師がその違いを理解していないらしい。あまり馴染みもないし、どこでどう使われているのやらと言った感じかもしれない。意味など知らなくても、診療は出来る。そう思うかもしれないが、MR の示すパンフレットにはだまされ続けるかもしれない…。知っておいて損はない。しかし、正直にいうと、このあたりは苦手分野なので、以下の説明は全く見当違いである可能性もある。適宜修正が必要。

標準偏差(SD)とは

標準偏差はデータのバラつきを示す一つの指標である。数字が大きいほどバラつきが大きい。例えば、ある都市 C の医師の身長を考えてみる。A 群 150、160、170、180、190cm の 5 人よりも、B 群 160、165、170、175、180cm の 5 人のほうがバラつきが少ないことは直感的に分かると思う。それを数字で表すわけである。どうやって数字で表そうか？昔の人も色々悩んだと思うが、それぞれの数字から平均値を引いた値の 2 乗を人数で割ってその平方根を取る方法が採用されたわけだ(その理由については後述)。どちらの群も平均値が 170 なのでそれぞれ以下のように計算される。(ルートの記号は最後まで伸びていると考えてください)

標本の標準偏差を求める式は $\sqrt{(\text{個々のデータ}-\text{平均})^2}$ を足したもの/(データの個数-1)なので、以下のように計算できる。

A 群

$$\begin{aligned} & \sqrt{[(150-170)^2 + (160-170)^2 + (170-170)^2 + (180-170)^2 + (190-170)^2] / 4} \\ &= \sqrt{(400 + 100 + 0 + 100 + 400) / 4} \\ &= \sqrt{250} \\ &\approx 15.8 \end{aligned}$$

B 群

$$\begin{aligned} & \sqrt{[(160-170)^2 + (165-170)^2 + (170-170)^2 + (175-170)^2 + (180-170)^2] / 4} \\ &= \sqrt{(100 + 25 + 0 + 25 + 100) / 4} \\ &= \sqrt{50} \end{aligned}$$

≒7.9

標準偏差はA群で大きく、ばらつきが大きいことが数字で表されたわけだ。なんとも面倒なことをしていると思うかもしれない。

実はこの標準偏差はただのばらつきを示す指標ではない。母集団が正規分布に従うとき、全体の約 68%が平均±標準偏差の間にあり、約 87%が平均±1.5 標準偏差の間にあり、約 95%が平均±2 標準偏差の間にあることが分かっている。この仕組みについて私は十分には理解していないが、とにかくそういうことが数学的に証明されているのである。成長曲線なんかで+2SD とか-2SD とかの線が引かれているが、それはこのことを意味している。(この例のように標本が 5 例だと正規分布しないのでそうはいかないのだが…。)

標準誤差(SE)とは

標準誤差とは母集団のデータの平均値を推定する幅を示すものである。SE= SD/√nと示される。ここで、前述のある都市Cの医師の身長平均はどれくらいなのか？そんなことを考えてみたい。都市 C の医師全員を調べれば文句はないが、それが可能だと今回は話が進まないで、例によって標本調査をしようとするわけである。仮に、5 人の医師を無作為に調査したところ、A 群のような結果になったとする。

標準誤差は標準偏差/√標本数と表されるので、以下のように計算できる。

$$\begin{aligned} & \sqrt{250}/\sqrt{5} \\ & =\sqrt{50} \\ & \approx 7.07 \end{aligned}$$

さて、何をやっているのかだんだん意味が分からなくなってきたが、大事なのはここからである。A 群という一つの標本から母集団の平均値を推定する場合に、幅を持たせて推定しようとするのが区間推定の考え方である。この区間のことを信頼区間という。よく、95%信頼区間という言葉が聞かれますが、これは母集団の真の値(この場合は平均値)がこの区間に含まれる確率が 95%というわけである。

95%信頼区間は一般に $x \pm 1.96 \times SE$ として示される。1.96 がどこから来たのか考えてみると、正規分布している場合、-1.96SD から+1.96SD に全体の 95%が入っているということが分かる。信頼区間は標本の数値で無く、平均値のある区間を示しているので、SD が SE に変わっただけである (SE= SD/√nなので、nが増えるほど生データよりは幅の狭い、先のとがった正規分布になるは

ずである)。面倒くさいので $\pm 2 \times SE$ としてもいいと思うが、変人扱いされるだけと思うので1.96としておいたほうが良い。ちなみに99%信頼区間であれば ± 2.58 とすればよい。

ここで注意が必要なのだが、実は n が少ない場合には1.96という数字は使えない。正規分布をする場合には1.96を使うわけであるが、標本の数 n が120を超えるような場合にだけしか使えない。 n が少ないと正規分布するという仮定が成り立たないわけである。この場合は t 分布表を用いなければならない。ネットで t 分布表で検索すると多くの表が検索できる。自由度4、有意水準0.05の場所をみると、2.13とある。この数字を使うわけである。そうすると、95%信頼区間は以下のように計算できる。

$$\begin{aligned} & 170 \pm 2.13 \times \sqrt{50} \\ & = 170 \pm 15.1 \end{aligned}$$

つまり、都市Cの医師の身長は平均154.9cmから185.1cmの間にある可能性が高いと考えられるわけだ。 n が少ないからすごくアバウトになっていることが分かる。 n が多くなると、正規分布の幅はどんどん狭くなる(ここでもう一度 $SE = SD / \sqrt{n}$ を思い出そう)。よって信頼区間も狭くなる。

自由度とは $n-1$ のことである。昔、ルービックキューブという玩具が流行したが、6面のうち、5面だけ色を揃えることはできない。考えれば当たり前だが、5面揃えると、6面目も自動的にそろってしまう。最後の数字は結局自動的に決まってしまう性格なのである。自由な数字は $n-1$ まで。自由度とはそんなものである(と勝手に理解している)。

参考文献

1. [高橋 信.マンガでわかる統計学.東京,オーム社,2004](#)
2. [中村 好一.基礎から学ぶ楽しい疫学.東京,医学書院,2002](#)