分子生物第３講座/構造生物学研究センター　　2009 年 2 年生コンピュータ実習
文責　成田哲博

第六回テキスト補遺
統計学とエラー超入門

どちらが速い？

A 君はある蛋白質 B の ATP 分解速度を調べた。25℃と 30℃でそれぞれ一回ずつ測定し、25℃では、5.4 分子/sec、30℃では 5.0 分子/sec であった(図 1A)。蛋白質 B の ATP 分解速度は、25℃と 30℃のどちらが速いと言えるだろうか？

このとき、正しい自然科学者や技術者は”わからない”と答えなければならない。なぜなら、その測定にどの程度の誤差があるのか、一回の測定ではわからないからである。A 君はさらに同じ測定を 4 回(図 1B)、9 回(図 1C)、16 回(図 1D)繰り返した。各回のデータは下表に示す。繰り返し実験を行うことで、測定値のばらつきの大きさ、測定値の分布の中心が正確にわかってくる。測定値のばらつきの大きさは一回の測定の信頼性、すなわち誤差の大きさを表しており、標準偏差で評価される。分布の中心は平均値で評価できる。二つの測定条件における測定値が違うか違わないかは、標準偏差、平均、測定数の三つから、判定できる。この場合は、A 君は 16 回の測定を通じて、ようやく ATP 分解速度は 30℃のほうが速いと結論できた。本テキストでは、その理論的背景と実際の手法を概説する。

	25℃(分子/sec)	30℃(分子/sec)
1 回目	5.4	5.0
2 回目	3.8	6.4
3 回目	5.0	7.2
4 回目	5.6	5.8
5 回目	5.8	5.6
6 回目	6.7	4.5
7 回目	5.9	5.8
8 回目	5.7	5.6
9 回目	3.5	5.6
10 回目	5.9	7.4
11 回目	6.4	6.8
12 回目	5.8	6.1
13 回目	3.9	5.2

14 回目	6.4	5.4
15 回目	4.2	4.2
16 回目	5.2	7.0

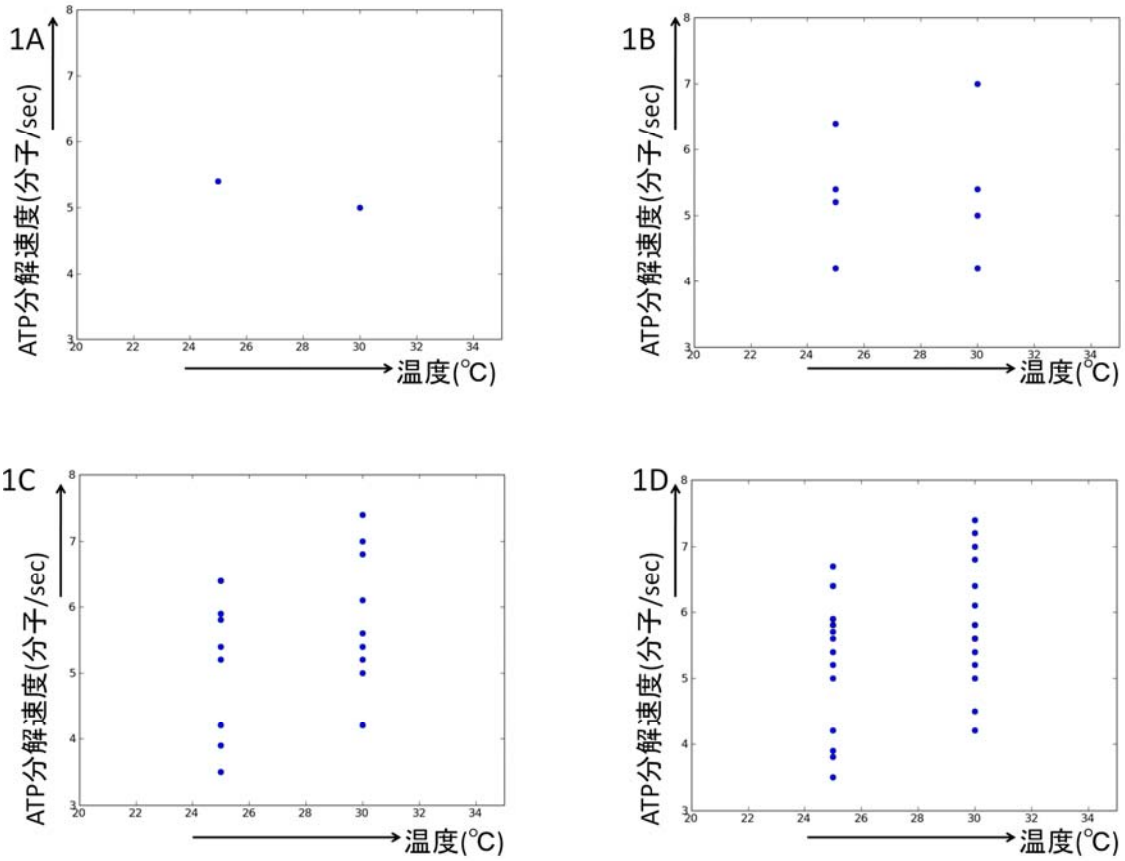


図 1: A 君の実験結果。A: 一回目まで。B:四回目まで。C:九回目まで。D:十六回目まで。

確率分布関数

あるパラメータで表せる事象が起こる確率を、そのパラメータの関数として表現したのが確率分布関数である。また、そのパラメータを確率変数と呼ぶ。たとえば、さいころを振ったときに 1 が出る確率は 1/6 であるから、確率分布関数を f、確率変数(さいころの目)を X とすると、

$P(X=1) \equiv f(1)=1/6$

である。P(評価式)は、評価式が真になる確率を表し、この場合、確率分布関数 f(n)は、P(X=n)と定義される。

$P(X=n) \equiv f(n) = 1/6,\ n=1,2,3,4,5,6$

となる。一般には、X が加算集合{$x_1,x_2,\cdots,$}の中の値をとる場合、確率の定義から

$f(x_k) \geqq 0$　　式 1

$\sum_k f(x_k) = 1$　　式 2

が成立する。この場合、f(x)は離散型の確率分布関数と呼ばれる。X がこのような離散値ではなく、連続値をとる場合にも適用できる。確率変数 X のとる値が、

$P(a < X < b) = \int_a^b f(x)dx$　　式 3

のように表せる場合、f(x)は連続型の確率分布関数と呼ばれる。a と b が極めて近い場合は、

$P(x < X < x + dx) = f(x)dx$　式 4

のようになる。つまり、f(x)は、x と x+dx の間に X が入る確率が f(x)dx であることを示し、f(x)が一種の密度を表しているのがわかる。そのため、連続型の確率分布関数は、確率密度関数と呼ばれる。離散型における式 1 、 2 と同様に、確率の定義から、

$f(x) \geqq 0$　式 5

$\int_{-\infty}^{\infty} f(x)dx = 1$ 式 6

が成立する。

期待値

確率と確率変数を掛けた総和を取ったものを期待値と呼ぶ。確率変数 X の期待値を E(X)と表記することにする。さいころの例では、

$$E(X) = f(1) \times 1 + f(2) \times 2 + f(3) \times 3 + f(4) \times 4 + f(5) \times 5 + f(6) \times 6$$

$$= \frac{1}{6} * (1 + 2 + 3 + 4 + 5 + 6)$$

$$= 3.5$$

となる。離散的確率分布では、

$E(X) = \sum_k x_k f(x_k)$　　式 7

となる。確率密度関数では、これを積分に直せばよい。

$E(X) = \int_{-\infty}^{\infty} xf(x)dx$　　式 8

期待値には、その定義式から、以下のような性質がある。

$E(c) = c$　　式 9

$E(X+c) = E(X) + c$　　式 10

$E(cX) = cE(X)$　　式 11

$E(X+Y) = E(X) + E(Y)$　　式 12

式 9-11 は自明であろう。式 12 について、離散的な場合に証明してみる。

$$E(X+Y) = \sum_i \sum_j (x_i + y_j) P(X = x_i, Y = y_j)$$

$$= \sum_i \sum_j x_i P(X = x_i, Y = y_j) + \sum_i \sum_j y_j P(X = x_i, Y = y_j)$$

$$= \sum_i x_i \sum_j P(X = x_i, Y = y_j) + \sum_j y_j \sum_i P(X = x_i, Y = y_j)$$

$$= \sum_i x_i P(X = x_i) + \sum_i y_j P(Y = y_j)$$

$$= E(X) + E(Y)$$

ここで、$\sum_j P(X = x_i, Y = y_j)$は、X=$x_i$ かつ Y=y_j の確率をすべての可能な j に対して和をとったものである。したがって、X=x_i であれば、Y はとりうるどの値でも良いという事象が起こる確率であり、P(X=x_i)と等しくなる。$\sum_i P(X = x_i, Y = y_j)$についても同様に、P(Y=$y_j$)と等しい。

また、X と Y に相関が無い場合には、

$E(XY) = E(X)\,E(Y)$　式 13

も成立する。

分散と標準偏差

確率分布の中心は期待値で求められるが、分布の広がりは期待値だけではわからない。確率分布がどの程度広がっているかを表すパラメータが分散と標準偏差である。確率変数 X, 分布の期待値 E(X)が μ であるとする。分散 V(X)の定義は以下のようになる。

$V(X) = E((X-\mu)^2)$　式 14

つまり、期待値からのずれの二乗の期待値である。分布の幅が広いほど、この値は大きくなる。分散の平方根をとったものが、標準偏差である。この分布の標準偏差をσ(X)とすると、

$\sigma(X) = \sqrt{V(X)}$ 式 15

である。例として、さいころの目の分散と標準偏差を計算してみよう。式 14 のまま計算するのは煩雑である。式 14 を式 9-12 を用いて変形すると、

$$V(X) = E((X-\mu)^2)$$
$$= E(X^2 - 2\mu X + \mu^2)$$
$$= E(X^2) - 2\mu E(X) + E(\mu^2)$$
$$= E(X^2) - 2\mu^2 + \mu^2$$
$$= E(X^2) - \mu^2$$

となる。実際の計算では、この変形の結果である、

$V(X) = E(X^2) - \mu^2$ 式 16

を用いて計算することが多い。さいころの目の場合、分散は、式 16 から、

$$V(X) = E(X^2) - \mu^2 = \sum_{i=1}^{6} f(i)i^2 - 3.5^2 = \sum_{i=1}^{6} \frac{1}{6} i^2 - 3.5^2 = \frac{1}{6}(1+4+9+16+25+36) - 3.5^2$$
$$= 2.917$$

標準偏差は、式 15 から、

$$\sigma(X) = \sqrt{V(X)} = 1.708$$

となる。

二つの確率変数 X と Y の間に全く相関がない場合には、分散の加法定理

$V(X+Y)=V(X)+V(Y)$　式 17

が成立する。また、式 14 から、c が定数の場合

$V(cX) = c^2V(X)$　式 18

である。

誤差と正規分布

実験におけるあらゆる測定値は誤差を含む。実験者のオペレーションは毎回完全に同じには決してならないし、測定機そのものの誤差、温度、湿度、振動など様々な要因が絡む。測定値は、

測定値 = 真の値 + 誤差

で表され、誤差はある確率分布をもった確率変数として扱うことができる。真の値が μ であるばあい、測定値の確率分布は、ほとんどの場合、平均値 μ、標準偏差 σ の正規分布

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$　式 19

で良く近似できることが知られている(図 2)。測定値の誤差の大きさは、この確率分布の標準偏差 σ で表される。正規分布は以下の性質がある。

$P(\mu - \sigma < X < \mu + \sigma) = 0.6827$

$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$

$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$

つまり、誤差の大きさが σ であるばあい、測定値は 95.45 %の確率で真の値 $\pm 2\sigma$ の範囲内に存在することになる。

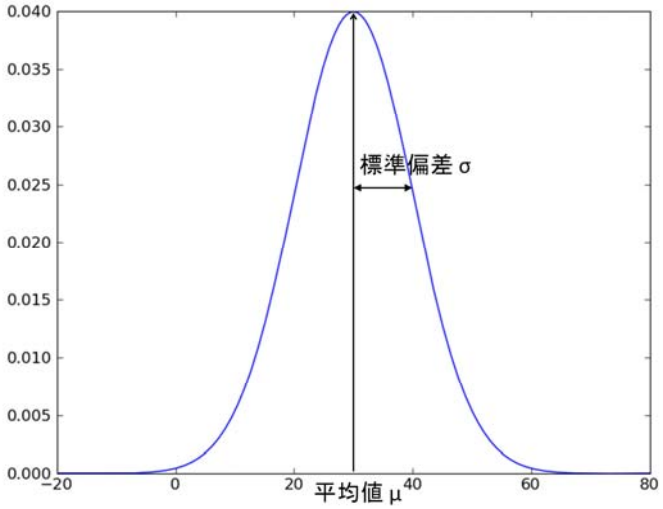


図 2:平均値 30, 標準偏差 10 の正規分布の例

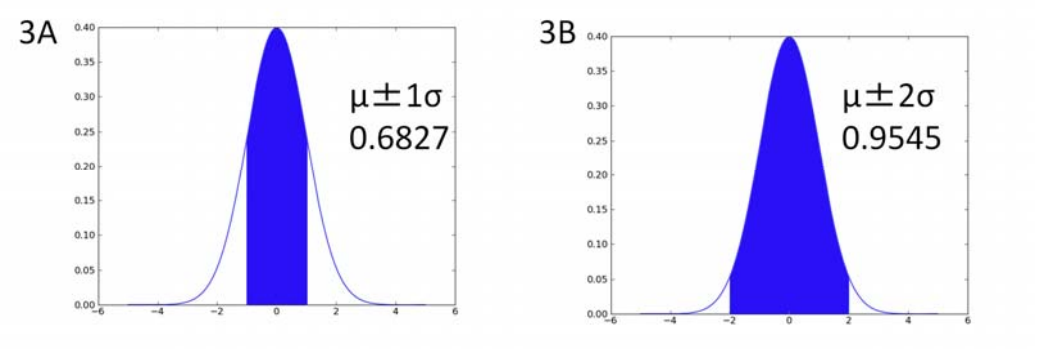


図 3: A: 正規分布(μ=0, σ=1)において、$\mu \pm \sigma$ の範囲が全体に対して占める領域を青で、全体に対する比率(=試行によって確率変数がその範囲内に入る確率)を右上の数値で表した。同様に、B: $\mu \pm 2\sigma$、C: $\mu \pm 3\sigma$

平均

実験において知りたいのは、真の値 μ であるが、実験から推定した値がどの程度信頼できるかを知るには、誤差の大きさ σ がわからなければならない。この μ と σ を推定するには、何回か同じ実験を行い、その測定値の平均とばらつきから推定値を計算する。まず、μ の推定値として用いられるのが平均値である。期待値 μ、標準偏差 σ の確率分布を持つ事象の試行を n 回繰り返した場合の平均値$\overline{X}$、試行による測定値を$\{X_1, X_2, X_3, X_4 \cdots\}$とする。測定値 X_k はそれぞれ、期待値 μ、標準偏差 σ の分布を持つ確率変数である。それぞれの試行は独立であるとする。

$$\overline{X} = \frac{\sum_k^n X_k}{n}$$

である。この平均の期待値$E(\overline{X})$は、式 9,12 を用いて、

$$E(\overline{X}) = E\left(\frac{\sum_k^n X_k}{n}\right) = \frac{\sum_k^n E(X_k)}{n} = \frac{\sum_k^n \mu}{n} = \mu$$

となり、分布の期待値の推定値として使えることがわかる。また、その分散は、式 17,18,15 から、

$$V(\overline{X}) = V\left(\frac{\sum_k^n X_k}{n}\right) = \frac{\sum_k^n V(X_k)}{n^2} = \frac{\sum_k^n \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

したがって、標準偏差は

$\sigma(\overline{X}) = \sqrt{V(\overline{X})} = \frac{\sigma}{\sqrt{n}}$ 式 20

この式 20 から、データをたくさんとればとるほど、平均値の標準偏差＝平均値の誤差は小さくなっていくことがわかる。

標本偏差

前項で、試行数 n の場合、平均値の標準偏差が$1/\sqrt{n}$に比例して小さくなっていくことを示した。σ がわかれば、平均値がどの程度のばらつき＝誤差を持っているのかがわかることになる。期待値 μ、標準偏差 σ の確率分布を持つ事象の試行を n 回繰り返した場合の平均値$\overline{X}$、試行による測定値を$\{X_1, X_2, X_3, X_4\cdots\}$とする。測定値 X_k はそれぞれ、期待値 μ、標準偏差 σ の分布を持つ確率変数である。σ を計算するにあたって、真の期待値 μ を用いることはできず、代わりに μ の推定値$\overline{X}$を用いる必要がある。有限回数の測定値から分散を推定するには、以下の式で定義される不偏分散 s^2 を用いる。s を標本偏差と呼ぶ。

$s^2 = \frac{\sum_k^n (X_k - \overline{X})^2}{n-1}$ 式 21

s^2 は、その期待値が、元の確率分布の分散 σ^2 と一致する。分母が n では無く、n-1 であることに注意。n-1 でないと、元の確率分布の分散の推定値にならない。n が小さい場合、n と n-1 で数十%もの差が出るので、n-1 で割ることは重要である。式 21 の期待値が、元の確率分布の分散 σ^2 と一致することを
証明してみよう。

$$
\begin{aligned}
&(n-1)E(s^2) = E(\sum_k^n (X_k - \overline{X})^2) \\
&= E(\sum_k^n (X_k - \frac{1}{n}\sum_j^n X_j)^2) \\
&= E\left(\sum_k^n X_k^2 - \frac{2}{n}\sum_k^n\sum_j^n X_k X_j + \frac{1}{n^2}\sum_k^n\sum_j^n\sum_i^n X_j X_i\right) \\
&= \sum_k^n E(X_k^2) - \frac{2}{n}\sum_k^n\sum_j^n E(X_k X_j) + \frac{1}{n}\sum_j^n\sum_i^n E(X_j X_i) \\
&= \sum_k^n E(X_k^2) - \frac{1}{n}\sum_k^n\sum_j^n E(X_k X_j)
\end{aligned}
$$

式　22

ここで、式 13 から、

$k \neq j$ のとき、$E(X_k X_j) = E(X_k)E(X_j) = \mu^2$ 式 23

また、i=j のときは、X_k, X_j は独立でないため、式 13 は成立せず、

$k = j$ のとき、$E(X_k X_j) = E\left(X_k{}^2\right)$ 式 24

また、X_k はすべて同じ確率分布を持つので、その確率分布をもつ確率変数を X と定義すると、

$E\left(X_k{}^2\right) = E(X^2)$ 式 25

式 23,24,25 から、

$$\sum_k^n \sum_j^n E(X_k X_j) = \sum_k^n E\left(X_k^2\right) + \sum_{k \neq j} E\left(X_k X_j\right) = nE(X^2) + n(n-1)\mu^2$$

式 26

式 22 に式 26,25 を代入

$$(n-1)E(s^2) = \sum_k^n E\left(X_k^2\right) - \frac{1}{n}\sum_k^n \sum_j^n E(X_k X_j)$$
$$= nE(X^2) - E(X^2) - (n-1)\mu^2$$
$$= (n-1)(E(X^2) - \mu^2)$$
$$= (n-1)V(X)$$

式 27

最後の行では、式 16 を用いた。V(X)は確率分布の分散である。したがって、s^2 の期待値は、もとの確率分布の分散の推定値として使えることがわかった。また、式 21 を変形すると、

$$s^2 = \frac{1}{n-1}\left(\sum x_i{}^2 - \frac{1}{n}\left(\sum x_i\right)^2\right)$$ 式 28

実際に標本偏差を計算する場合はこの式 28 を用いることが多い。

統計的有意な差

ここまでで、数学的な準備は整った。平均値の推定標準偏差は、試行数 n の場合、式 20 の σ のかわりに標本偏差 s を用いて $s/\sqrt{n}$と表すことができ、$s/\sqrt{n}$は、その測定条件のエラー(誤差の大きさ)と呼ぶ。

さて、ここで平均値の誤差は、エラーを標準偏差とする正規分布に従うと考えて良い。したがって、測定条件 1 における平均値が a_1、エラーが b_1 の場合、測定条件 1 の真の値は、a_1-$2b_1$ と a_1+$2b_1$ の間に、95.45 %の確率で存在すると考えて良い。

測定条件 1 と、平均値が a_2、エラーが b_2 の測定条件 2 の値が統計的に異なるかどうかは、t 検定などのさまざま検定方法による。しかし、簡単には、それぞれの測定条件で三回以上の計測を行っていて、a_1 と a_2 の差が、b_1+b_2 よりも大きければ、二つの測定条件の値は統計的に有意と見なしても良い。たとえば、それぞれの測定条件で三回ずつ測定した結果、b_1=b_2 で、a_1-a_2=2 b_1 (a_1 と a_2 の差がちょうど b_1+b_2)の場合、測定条件 1 の真の値が測定条件 2 の真の値よりも大きい確率は、t 検定によれば 96.5 %程度になる。同じエラー、平均値

でも、測定回数を多くとった結果であれば、標本偏差の信頼性が上がっていくので、この確率は大きくなる。四回ずつなら約 98.5 %である。一方、二回測定の場合、この確率は 90.8 %程度になり、著しく低い。このため、統計的に有意な差かどうかを調べるためには、最低一条件 3 回以上の測定が一般に必要である。ただし、ここで注意してほしいのは、四回の測定でエラーが b という場合、標本偏差は 2b である。また、二回の測定でエラーが b と言う場合、標本偏差は$\sqrt{2}$b である。同じエラーでも測定回数が何回の結果かによって、もとのデータの標本偏差は異なる。

エラーバーとグラフ

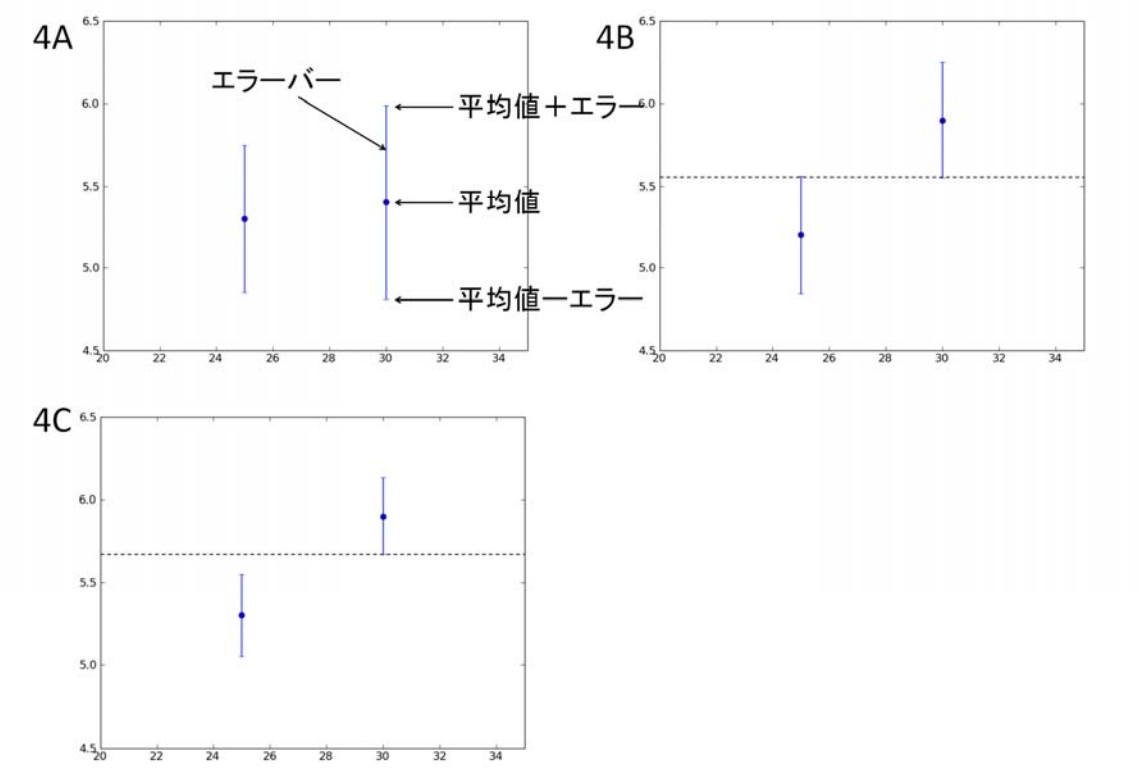
前項で述べた条件（a_1 と a_2 の差が、b_1+b_2 よりも大きければ、二つの測定条件の値は統計的に有意と見なしても良い）を一目で判断するために、グラフにエラーバーをつけて表示するということが良く行われる(図 4A)。平均値 a、エラーb の場合、a の上下に長さ b のバーをつける。このエラーバーが重ならなければ、二つの測定条件の値は有意な差があるということになる。次の項で実際に見てみよう。

最初の例

では、最初の例に戻ろう。A 君の実験で、4 回、9 回、16 回までの測定値を用いた平均値、標本偏差、エラー(平均値の推定標準偏差, 標本偏差/$\sqrt{n}$) を計算すると以下のようになる。

		4回目まで			9回目まで			16回目まで	
	平均	標本偏差	エラー	平均	標本偏差	エラー	平均	標本偏差	エラー
25℃	5.3	0.9	0.45	5.2	1.08	0.36	5.3	0.99	0.25
30℃	5.4	1.18	0.59	5.9	1.05	0.35	5.9	0.93	0.23

これをグラフにすると図 4 のようになる。4A は 4 回目まで、4B は 9 回目まで、4C は 16 回目までのデータを示した。9 回目まではエラーバーがぎりぎり重なる領域がある。16 回目になると重なる領域が無くなり、二つの測定条件は明かに有意な差があると言える。なお、わかりやすいように、図 4B, 4C においては 30℃の下側エラーバーに合わせて点線を引いた。



参考文献

統計学入門　東京大学教養学部統計学教室編　東京大学出版会