現場の濃度分析技術者のための
データ解析の基礎知識

臨床検査（生化学，ELISA），治験検査（薬物濃度測定），環境検査（濃度測定）
食品検査（理化学，残留農薬，合成抗菌剤）などの濃度測定における
平均値の計算から検量線や方法間比較までの本質を理解する

秋山　功

2009年11月 E1.0.3
改訂　2018年　3月 E2.0.2β

はじめに

濃度分析技術者を対象として，市販の統計学の多くの本からは見え難いデータを扱う上での基本的で大切なことで，実際に測定の現場で問題になったデータ解析の方法などについて，最も単純な数値例や図を用いて解説します。

「はじめての統計学」とか「誰でもできる統計処理」といった内容ではありません。統計処理の入門やハンドブック的な内容でもなく，逆に数理統計学的な内容でもありません。

平均値の計算方法，検査のスケジューリング方法，有意差検定でデータを幾つ集めれば良いのか，2 方法の比較方法，測定精度の計算，検量線の最適な回帰式と重み付けの選択はどのようにするのかなど，現場に即した「実践的」で「本質的」な内容です。本文は一見偏りがあるように見えますが，内容は広い範囲に及んでいます。入門的な本では記載されていない内容も含まれます。関連することで楽しく興味を引くと思われる，遊びのような内容も取り入れました。

また，よくある統計的過誤を犯さないない為の内容ともなっています。

最尤法やモーメント法や線形モデルなどは解り難いと思いますが，データ解析の基本です。簡単に努力しなで理解できる統計学やデータ解析の本はありません。実際に現場の広範囲な問題に対応しようとすれば，多くの本を読み多少の努力はしかたがありません。

良くある「誰でもできる統計処理」のような本で述べているような，定型的なデータでないと計算できない，実践で応用できない役立たないデータ解析では意味がありません。

統計学を専門とはしていない分析技術者が作成したので，統計学の専門家が問題としない内容でも，現場で必要と感じ，知ってもらいたい事柄を記載しました。

分析は専門でも，統計学は専門家ではないので，本文中の間違い，問題点の指摘，意見，感想などを頂ければ幸いです。

秋山　功

E-mail: isao_akiyama@kpd.biglobe.ne.jp

本文について

なるべくデータ解析の基本的なことから記載しました。第 I 部の第 9 章までが特に基本的な内容ですが，どの章からでも読めるように記載しました。

式には番号を付けていません。説明で必要な場合は同じ式を何回も記載しています。

必要な場合は，同様に同じような説明を繰り返しています。

式が多く難しいように感じますが，式の部分は読み飛ばしても概要は理解できるように説明したつもりです。

さらに，難しいと思う章を読み飛ばしても問題ありません。

現場で必要と思う部分だけを読んでもなるべく理解できるように，各章は独立して読むことができようにしたつもりです。

最尤法やモーメント母関数は，聞きなれない場合は特に解り難いかも知れませんが，基本をなすものとして説明しました。

本文中の注）や参考の部分で，理解し難いものは読み飛ばして下さい。

正確に理解出来なくても，一度全ての文章に目を通してみて下さい。一度目を通してから，現場で必要になった時に再度読み返して下さい。

本文では従来の市販の本などでは無かった，下記の試みを行いました。

1）同じ職場の同僚に，1 対 1 で説明する時のような記述を試みました。

2）多くの入門的な統計学の本では記載されていない，実際の分析の現場でよく問題になること，基本的で重要なこと，勘違いし易いこと，質問を受け共に考えたことについて述べました。

3）文章は少なくし，手順書ではない，数式での導出や証明でもない，なるべく簡単な架空の数値例，シミュレーション，図，実際のデータなどから納得できるような説明を試みました。

4）多少わき道でも，楽しいく興味を引くと思われる内容も記載しました。

主な内容

(a) 基本的な計算（平均値，標準偏差，相関係数など）
(b) 間違え易い計算（シンプソンのパラドクス，ベイズの定理など）
(c) 統計量の質問（不偏標準偏差の計算方法，相関係数の 2 乗とは何かなど）
(d) 検査工程の最適化（検査のスケジューリングについて）
(e) 計算方法の説明（ISO17025 の不確かさの伝播則の計算方法など）
(f) 統計手法（有意差検定，最小 2 乗法，最尤法，モーメント法など）
(g) 統計モデルの選択（AIC，検量線の回帰式の選択，重み付けの選択など）

○本文中「エクセル」は，表計算ソフト Microsoft Excel です。

目次

第Ⅰ部　基本的な事柄

分析現場で日常的に使用する，平均，標準偏差，変動係数，相関係数，回帰式，正規分布や統計用語などの基本的なことと使用上の注意点を述べます。

第１章　基本的な計算と用語，記号の確認

1.1　基本的な計算

平均，分散，標準偏差，相関係数などの基本的な計算を簡単な数値例で確認しておきす。

Σ とΠによる略記

Σは総和の略記号で，Πは総積の略記号として使用します。

例えば，データ２，３，４が得られたとします。

データ	2	3	4

添え字を付け $x_1=2, x_2=3, x_3=4$ ならば，

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 2+3+4 = 9$$

$$\prod_{i=1}^{3} x_i = x_1 \times x_2 \times x_3 = 2\times 3\times 4 = 24$$

と計算します。

平均値

平均値 $\bar{x}$ の計算は，先ほどの総和の略記号Σを使用して，$x_1=2, x_2=3, x_3=4$ では

データ数ｎ＝3 で

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \frac{2+3+4}{3} = 3$$

となります。

平均値は重要で基本ですので,「第２章　平均値の計算方法を考える」で再度考えてみます。

偏差平方和（平方和）

標本のバラツキを表す偏差平方和Ｓは，平均値と各データの差の平方の和で，

$x_1=2, x_2=3, x_3=4$ では

$$S = \sum_{i=1}^{n} (x-\bar{x})^2$$

$$= (2-3)^2 + (3-3)^2 + (4-3)^2$$

$$= 2$$

となります。本文では大文字のＳを使用します。

分散

標本の不偏分散Ｖ（または s²）は，平方和を自由度で割ったもので，$x_1=2, x_2=3, x_3=4$ では

$$V=\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$$

$$=\frac{(2-3)^2+(3-3)^2+(4-3)^2}{3-1}=1$$

となります。ｎではなくｎ－1 で割る意味については「第５章　標準偏差はｎ－1 で割っても不偏推定量にならない」で説明します。

標準偏差

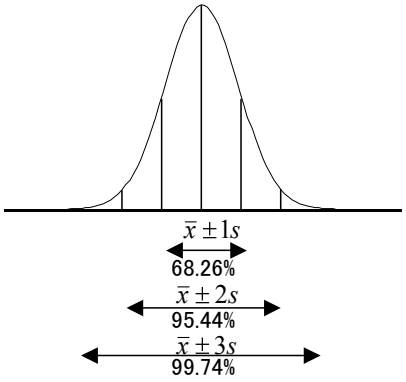
バラツキの目安として日常的に使用する標本標準偏差 S.D.（s.d.または s）は，分散の平方根で $x_1=2, x_2=3, x_3=4$ では

$$s=\sqrt{V}$$

$$=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$=\sqrt{\frac{(2-3)^2+(3-3)^2+(4-3)^2}{3-1}}=1$$

と計算します。



正規分布に従う場合は，右の図に示す関係があります。正規分布は釣鐘状の分布で，「第８章　正規分布について」で詳細に述べます。

良く質問されるエクセルでの標本標準偏差 s.d.の関数は　**=STDEV(数値 1:数値 2)**です。

変動係数（相対標準偏差）

変動係数 C.V.（c.v.）または相対標準偏差 RSD はバラツキを相対的に表すもので，標準偏差を平均値で割ったものです。通常は百分率で表し，精密度の目安として日常的に使用します。$x_1=2, x_2=3, x_3=4$ では，標準偏差が１，平均値が 3 ですから

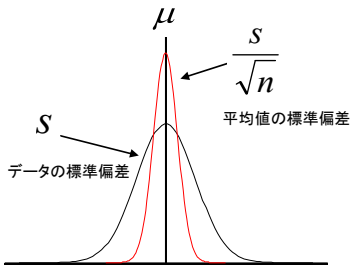
$$c.v.\%=\frac{s}{\bar{x}}\times 100$$

$$=\frac{1}{3}\times 100=33(\%)$$

となります。変動係数を理解するために「第８章　正規分布について」の参考３に，さらに説明を加えてあります。

標準誤差

標本から計算した平均値もバラツキがあります。平均値の標準偏差である標準誤差 S.E.(s.e.)は，



$$S.E.=\frac{s}{\sqrt{n}}$$

で計算します。

$x_1=2, x_2=3, x_3=4$ の標準誤差は，標準偏差を $\sqrt{n}$ で割ったものですから，

$$S.E.=\frac{s}{\sqrt{n}}=\frac{\sqrt{\frac{(2-3)^2+(3-3)^2+(4-3)^2}{3-1}}}{\sqrt{3}}=0.58$$

となります。

２重測定ではｎ＝2 より，１回の測定値よりも $\frac{s}{\sqrt{2}}=\frac{s}{1.41}$ だけ精密度が良くなります。

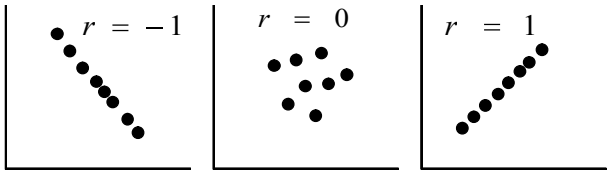
「第１０章　誤差伝播の法則」の参考２に，$\sqrt{n}$ で割る理由を説明してあります。

相関係数

２つの標本の相関係数ｒは

$$r\equiv\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$$

で計算します。



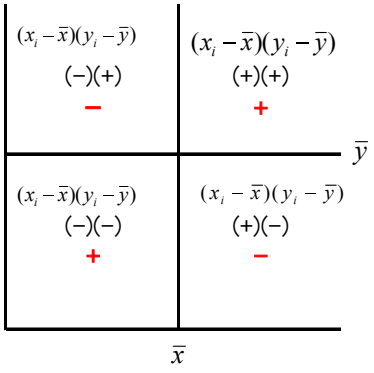
相関係数は２標本の相関の程度として－１≦ｒ≦１の値を取り，ｒを 2 乗した R^2 は検量線の適合の程度などの指標として使用します。

ｒの計算式の分母は，範囲を±１に規格するためのものです。

上記式の分子の部分に注目すると，相関係数の±は右下の図のようになります。

下記のデータでの相関係数の計算を示します。

	X	Y
1	2	2
2	3	3
3	4	10
平均値	3	5



$$r=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$$

$$=\frac{(2-3)(2-5)+(3-3)(3-5)+(4-3)(10-5)}{\sqrt{\{(2-3)^2+(3-3)^2+(4-3)^2\}\times\{(2-5)^2+(3-5)^2+(10-5)^2\}}}=\frac{3+0+5}{\sqrt{38}}=0.918$$

尚，原点回帰や曲線回帰での相関係数などは考えません。[注1)]

「第６章　相関係数ｒとＲの２乗とは何か」で，相関係数と決定係数の詳しい説明をします。

回帰式

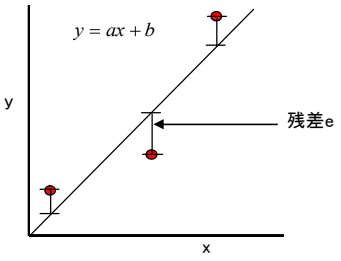
相関図や検量線の回帰式は，一般に最小２乗法で求めます。

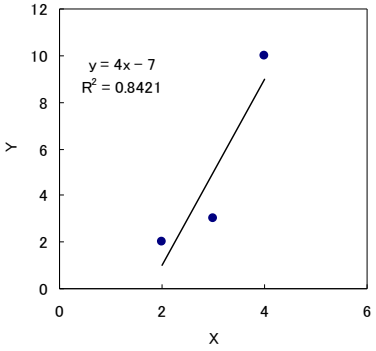
最小２乗法は下記の図の残差ｅの２乗の和，残差平方和Seを最小にするものです。

$$Se=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}\{y_i-(ax_i+b)\}^2$$

下記のデータをエクセルの「分析ツール」で１次回帰すると，次のような結果が得られます。[注2)]

	X	Y
1	2	2
2	3	3
3	4	10
平均値	3	5





概要

回帰統計	
重相関 R	0.917663
重決定 R2	0.842105
補正 R2	0.684211
標準誤差	2.44949
観測数	3

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	32	32	5.333333333	0.260147
残差	1	6	6		
合計	2	38			

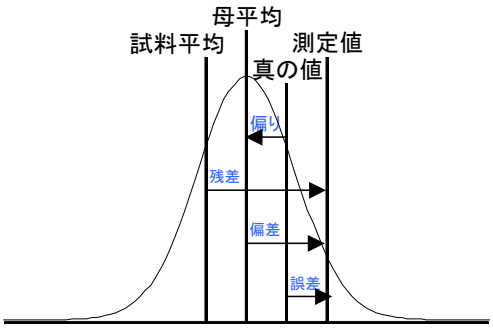
検量線については「第１６章　検量線の重み付きと回帰式の選択方法」で，詳細に記載しました。

注 1）中村 永友　土屋 高宏：「焦点をもつ回帰直線群の推定とその周辺」応用統計学会 Vol.36,No.1(2007),31-50

注 2）Microsoft Excel「エクセル」の「分析ツール」の使用方法は，「エクセル」の「Microsoft Excel のヘルプ」を見て下さい。

1.2　用語

　ばらつき，系統誤差，偶然誤差，不確かさ，誤差，偏り，精度，正確さ，精密度，精度などさらに多くの用語があり使用されます。JIS　Z8103 の付図も示しました。



JIS　Z8103 の基本的な表現も書き写しておきます。

ばらつき	測定値の大きさがそろっていないこと。また，ふぞろいの程度。
系統誤差	測定結果にかたよりを与える原因によって生じる誤差。
偶然誤差	突き止められない原因によって起こり，測定値のばらつきとなって現れる誤差。
不確かさ	合理的に測定量に結びつけられ得る値のばらつきを特徴づけるパラメータ。これは測定結果に付記される。
正確さ	かたよりの小さい程度。
精密さ，精密度	ばらつきの小さい程度
精度	測定結果の正確さと精密さを含めた，測定量の真の値との一致の度合い。 JIS　Z8101 では精確さ，総合精度のこと

1.3　なるべく修正項 CT は使用しない

古い統計学の本だけでなく最新の本や通知法などでも，計算が簡単であるとして修正項 CT を使用しているものがありますが，精度が落ちるのでなるべく定義に従って計算すべきです。[注 1)]

標準偏差を例に示します。

$$s = \sqrt{V} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x\right)^2}{n}\right)}$$

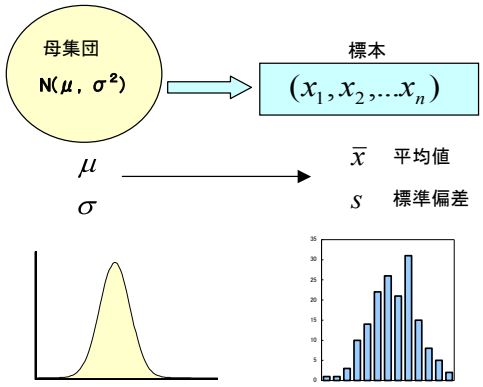
CT=
（全データの合計）^{2}
全データ数n

よく分散分析で計算が簡単であるとして，今でも修正項 CT を使用している本，文献がありますが，パソコンが普及しているので修正項 CT を使用する必要はなくなっています。数値処理上，大きい数から大きい数を引き小さい数を求めることは，桁落ちの可能性があり，あまり好ましくはありません。

注 1）奥村 晴彦：「パソコンによるデータ解析入門」技術評論社（1986 年）

1.4　記号と関数

母集団とは，ある集合の全体で，その一部を標本として取り出していると考えます。母集団と標本の関係は下記の図のようになっています。



統計学では約束として，なるべく母集団に関することはギリシャ文字α，βなどを使用し，標本からの統計量はラテン文字 a，b などを使用します。これは，違いを明確にするためです。例えば，母標準偏差はσで標本標準偏差は s を使用します。相関係数は一般に，母相関係数はρで，標本相関係数は r を使用します。

母集団のパラメータθ（平均μ，分散σ^2など）の推定値は ^ を使う慣わしです。$\hat{\theta}$なら，母集団のパラメータの推定値を意味します

sup は上限（最大のもの），inf は下限（最小のもの）で，i.i.d.は独立に同一分布に従うことを表し，～は「…に従う」と言う意味で通常は使用します。

平均μ，分散σ^2の正規分布は $N(\mu, \sigma^2)$ と表しますが，例えば

$$x_i \sim \text{i.i.d. } N(\mu, \sigma^2)$$

ならば，変数x_iは独立に平均μ分散σ^2の正規分布に従うことを示しています。

さらに，近似は≒を使用する場合もありますが，一般的には近似は≈や≅を使用します。∝は比例するの意味で使用します。また，定義として明確にしたい時は：＝を使用する場合もありますが，本文では≡を使用します。

ガンマ関数とベータ関数[注1)]

統計学ではガンマ関数Γ（x），やベータ関数Ｂ（α,β）などがよく出てきます。見慣れないと難しく感じますが,これらの関数の数値はエクセルや数学ソフトでも計算できます。

簡単に説明しておきます。

ｎ！はｎの階乗で,

$n!=n\times(n-1)\times(n-2)\times...\times2\times1$

と計算します。ｎが０の場合は０！＝1と約束します。[注2)]

異なるｎ個のものの並べ方（順列）はｎ！です。

例えばABCDの4個の並べ方は４！で，$4!=4\times3\times2\times1=24$となり24通りあります。

この階乗を実数，複素数に拡張したものがガンマ関数であると見ることができます。

ガンマ関数は

$$\Gamma(x)\equiv\int_0^\infty u^{x-1}e^{-u}du \qquad x>0$$

ですが,

$\Gamma(n)=(n-1)!$

$\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$

の関係があります。

例えば

$\Gamma(5)=(5-1)!=4\times3\times2\times1=24$

です。

ベータ関数は

$$B(\alpha,\beta)\equiv\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \qquad \alpha>0,\beta>0$$

で定義されます。

注１）E.アルティン（上野 健爾　訳，解説）:「ガンマ関数入門」日本評論社（2002年）

注2）１は順列，組合わせ，変換，などを変えない，つまり，１は変化しないという意味を持ちます。

ギリシャ文字

数式を正しく読むのは意外と難しいものです。ギリシャ文字の読み方を示します。

A	α	ｱﾙﾌｧ	N	ν	ﾆｭｰ
B	β	ﾍﾞｰﾀ	Ξ	ξ	ｸﾞｻﾞｲ
Γ	γ	ｶﾞﾝﾏ	O	o	ｵﾐｸﾛﾝ
Δ	δ	ﾃﾞﾙﾀ	Π	π	ﾊﾟｲ
E	ε	ｲﾌﾟｼﾛﾝ	P	ρ	ﾛｰ
Z	ζ	ｾﾞｰﾀ	Σ	σ	ｼｸﾞﾏ
H	η	ｲｰﾀ	T	τ	ﾀｳ
Θ	θ	ｼｰﾀ	Y	υ	ｳﾌﾟｼﾛﾝ
I	ι	ｲｵﾀ	Φ	ϕ	ﾌｧｲ
K	κ	ｶｯﾊﾟ	X	χ	ｶｲ
Λ	λ	ﾗﾑﾀﾞ	Ψ	ψ	ﾌﾟｻｲ
M	μ	ﾐｭｰ	Ω	ω	ｵﾒｶﾞ

参考　本文の記号について

①本文中で紛らわしいかも知れませんが，同じ意味で使用したものを上げておきます。不偏分散は V または s^2 を使用しました。その平方根である標本標準偏差は，S.D.(s.d.)または s を使用しました。

変動係数 C.V.(c.v.)と相対標準偏差 RSD は同じです。

②ネイピア数 e の指数部分が複雑な場合は exp を使用する場合があります。

e^x と exp(x)は同じです。使用する数学記号のイメージ，雰囲気などから，本文ではなるべく e を使用しました。しかし，本文の「第 8 章　正規分布について」の参考で，2 変量正規分布の式は見難いので exp を使用しました。

③)関数 $f(x)$ の $x=a$ での微分は $f'(a)$　の他に $\frac{d}{dx}f(a)$, $\frac{df(a)}{dx}$　や

$\left.\frac{df(x)}{dx}\right|_{x=a}$　などの記号も使用します。

また，変数が x, y である場合に，y を定数として扱い，変数 x で微分する場合は偏微分と言いますが，∂ を使用して

$\frac{\partial f(x,y)}{\partial x}$　などを使用します。

④組み合わせで n 個から m 個選ぶ場合，一般的（国際的）に $\binom{n}{m}$ を使用しますが，解り易さから高校で習う ${}_nC_m$ を使用しました。

1.5　計算不能について

精密度の指標として，普段よく c.v.（変動係数）または RSD（相対標準偏差）とも呼ばれる統計量を計算します。

c.v.%は

$$c.v.\% = \frac{s.d.}{\bar{X}} \times 100$$

と計算します。s.d.は標本標準偏差で，$\bar{X}$は標本平均です。

コントロール検体の３濃度 10，5，0（Blank）を各５回測定して，下記の表を得たとします。

n＝5

平均濃度$\bar{X}$	s.d.	c.v.%
10	1.8	18
5	1	20
0	0.5	0

この表には誤りがあります。平均濃度 0 のときの c.v.%は

$$c.v.\% = \frac{s.d.}{\bar{X}} \times 100 = \frac{0.5}{0} \times 100$$

で，0 で割ることになります。

平均濃度 0 のときの c.v.%は計算「不能」になります。計算できないので「0」ではなく，

平均濃度$\bar{X}$	s.d.	c.v.%
10	1.8	18
5	1	20
0	0.5	―

「－」とすべきです。[注1)] 0 や負の数を含むデータがある場合の c.v.%は考えません。

注１）簡単に上の例で「不能」であること示します。

$$\frac{0.5}{0} \times 100 = b$$

b が何であるかを考えます。割り算を掛け算に直すと

$b \times 0 = 0.5 \times 100$

となります。b がどのような数であっても 0 を掛けると 0 になり，0.5×100 にできません。

このため，b がどのような数でも式が成り立たないので，b は「不能」です。

また，形式的ですが，$\frac{0}{0} = b$ ならば，$0 = b \times 0$ で，b がどんな数でも式は成り立つので，「不定」です。

平均濃度$\bar{X}$	S.D.	C.V.%
10	1.8	18
5	1	20
0	0	―

また，表で s.d.が 0 の場合も c.v.％＝0 ではなく，

$c.v.\% = \frac{s.d.}{\bar{x}} \times 100 = \frac{0}{0} \times 100 =$ 不定 です。

0 の取り扱いについて幾つかを示すと下の表のようになります。[注１）]

$\frac{0}{a} = 0$

$\frac{a}{0} =$ 不能

$\frac{0}{0} =$ 不定

$a^0 = 1$

$0^0 =$ 不定

0！＝1！＝1

log 1= 0

log 0=不能　　（a≠0）

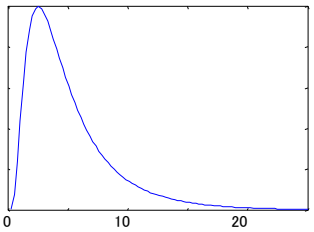
$\frac{0}{0} =$ 不定 ですが，同様に $\frac{\infty}{\infty}$，$\infty - \infty$，$\infty \times 0$，∞^0 なども不定形です。

また，臨床検査の基準値（正常値）は対数正規分布することが多く，対数変換して正規分布にすることがありますが，データに 0 がある場合は対数変換ができません。

y ＝log 0

このためデータに 1 を加え変換し，後で引いたりします。

y ＝log (x+1)



このように 0 を避ける必要があり，0 により一貫した計算ができない例はたくさんあります。

注１）土基 善文：「x の x 乗のはなし」日本評論社（2002 年）

参考　2 の逆数は 1／2 で，3 の逆数は 1／3 です。1 の性質は 1 の逆数は 1 自身ですので実数での単数です。1×1＝1 となるからで，1 は幾ら掛けても 1 です。0 は何を掛けても 1 に出来ないので，通常は 0 の逆数は存在しません。このことが 0 と 1 の特徴ともなっています。

数学で a÷0＝∞を導入して処理する場合や，無限遠点∞を考え 0 で割ることを回避するとか，他にも∞と理解することがありますが，数値を示すことができないことに変わりはありません。∞は数値ではありません。

0^0 は「不定」ですが，$0^0 = 1$ と約束する場合もあります。また，数学ソフトよって，ここで述べた不定，不能，などが 1 や 0 で返される場合があり，注意が必要です。大切なのは，実際のデータ解析で不都合なことが起きていないかを調べることです。

ところで，0 は偶数ですが，自然数に 0 を入れるか，0 を平方数とするかなどは流儀の問題のようです。

第２章　平均値の計算方法を考える

2.1　重み付き平均

数回測定して，その平均値を求めることはよく行います。平均値の計算について改めて考えてみます。

通常使用する平均は**算術平均**で，

$$\bar{x} = \frac{x_1 + x_2 + x_3 \ldots x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

と計算します。Σは和の略記号です。

ある検体のある物質を，A の分析機器で５回測定し，確認のために B の機器で３回測定し，C の機器で１回測定して，下記の表を得たとします。

機器	測定値					平均値
A	5	4	5	3	5	4.4
B	4	3	5			4
C	5					5

この検体の平均値はどのように計算すべきか考えてみます。

各 A，B，C の平均値の平均値は

$$\frac{4.4 + 4 + 5}{3} = 4.5$$

です。このように，平均値の平均値を使用することはよくあるのではないでしょか。

しかし，A は５回の測定で，C は１回の測定です。これを同等に扱うのは少し疑問を感じます。そこで，機器 A，B，C に明確な差が見られないのならば，何らかの「重み」（または「荷重」）を付けるのが良いのではないかと考えられます。

データ数を重み w_i として使用してみます。各機器の平均値を x_1，x_2，x_3 とすると

$$\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{w_1 + w_2 + \ldots + w_n}$$

$$= \frac{w_1}{w_1 + w_2 + \ldots + w_n} x_1 + \frac{w_2}{w_1 + w_2 + \ldots + w_n} x_2 + \ldots + \frac{w_n}{w_1 + w_2 + \ldots + w_n} x_n$$

で計算するのが，「**重み付き平均値**」です。

先ほどの例で，データ数での重み付き平均を計算すると，全データ数は９で，各データ数は５，３，１なので，データ数で重みを付けて

$$\frac{5}{9}\times 4.4+\frac{3}{9}\times 4+\frac{1}{9}\times 5=4.3$$

となります。

機器 A，B，C に明確な号機間差が見られないのであれば，データ数による「重み付き平均値」の 4.3 を採用すべきであると考えられます。

この重み付き平均は，全データの総平均と同じです。

$$\frac{5+4+5+3+5+4+3+5+5}{9}=4.3$$

A，B，C の精度（精密度）が異なるときに，その分散を重みにして，重み付き平均を計算することもできます。この場合は，重みとしては標準偏差ではなく，一般的に分散の逆数を使用します。[注 1)]

尚，各機器によるデータ数が同じならば，重み付き平均と通常の平均は同じなので，重み付き平均は通常の平均を拡張したものとなっています。

「重み」は，検量線の重み付き回帰式の「重み」と同じです。重み付き回帰式については，「第１６章　検量線の重み付きと回帰式の選択方法」で述べます。

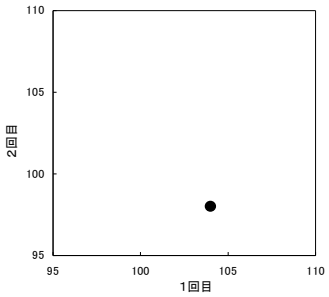
注１）最尤法から正規分布に従うと仮定すると，分散の逆数 $1/\sigma^2$ を重みとすると最尤推定量となります。このため，分散の逆数 $1/\sigma^2$ を掛けて平均値を計算します。データ数で重み付けることの妥当性も最尤法から導けます。最尤法については本文「第１３章　最尤法について」で述べます。

2.2　平均値と最小２乗法

データの**期待値**として，次のような例を考えてみます。[注 1)]

ある検査で異常値が出て 104 であったとします。再検の必要性があるとして検査したら，2 回目は 98 であったとします。104 と 98 で真の値はどこにあるのかを考えます。

104 と 98 のデータの情報を失うことなく図にするために，1 回目を x 軸（横軸），2 回目を y 軸（縦軸）にし，プロットしてみます。



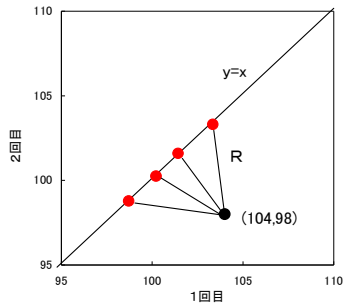
この図から，２回のデータの真の値を推定することを考えてみます。

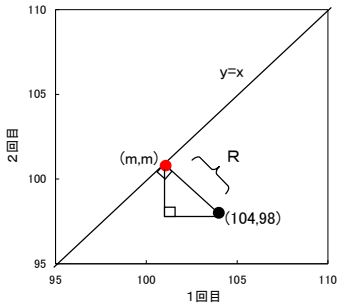
下の図を見て下さい。真の値は 1 回目と 2 回目は同じ検体なので，同じ値になる直線 y＝x 上にあるはずです。点（104,98）と y＝x と交わる線分 R で最短となる値が，真の値の推定値になると考えられます。

R の距離は三平方の定理（ピタゴラスの定理）から

$$R = \sqrt{(x-m)^2 + (y-m)^2}$$

で計算できます。





注 1）奥村 晴彦：「パソコンによるデータ解析入門」技術評論社（1986 年）

Ｒが最短となる，つまり，Ｒが最小となるｍを求めるにはどのようにしたらよいのでしょうか。

式にデータを入れると

$$R^2 = (104 - m)^2 + (98 - m)^2$$

となり，

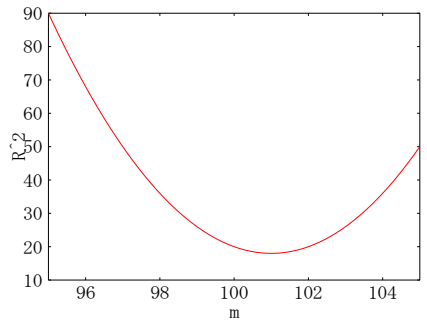
（104 - m）^{2} と（98 - m）^{2} の和は，残差平方和です。

展開すると

$$R^2 = 2m^2 - 404m + 2042$$

になります。

これは 2 次曲線で，R^2 を縦軸，ｍを横軸にすると下の図が得られます。



このＲ^{2}が最小となるのは凹の極値を求めることです。つまり，微分して 0 となる値を求めてみます。[注１）]

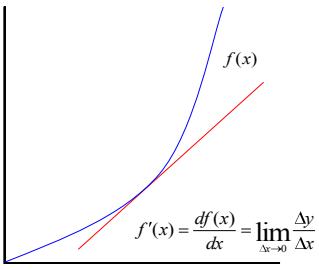
$$\frac{d}{dm}(2m^2 - 404m + 2042) = 0$$

$$4m - 404 = 0$$

$$m = 101$$

注１）参考　**微分公式**

関数 $f(x)$	導関数 $f'(x)$
x^n	nx^{n-1}
$\sin x$	$\cos x$
$\cos x$	$-\sin x$
e^x	e^x
$\log x$	$\frac{1}{x}$



微分

よって，R^2が最小となるmは 101 になります。先ほどの式に代入してみます。
Rは

$$R=\sqrt{(x-m)^2+(y-m)^2}=\sqrt{(98-101)^2+(104-101)^2}$$
$$=3\sqrt{2}$$

となります。

1回目 104 と2回目 98 で真の値に近いと考えられる値，m＝101 が得られました。つまり，真の値は 101 と推定できます。

ところで，面倒な微分をしなくても，この 101 は平均値に他なりません。

$$m=\frac{98+104}{2}$$
$$=\frac{1}{2}\times 98+\frac{1}{2}\times 104$$
$$=101$$

ここで示したのは，最小2乗法の原理そのものです。[注1)]

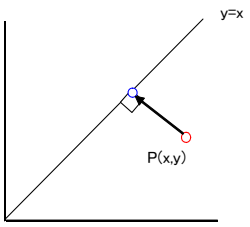
つまり，残差の2乗が最小になる関数を求めるのが最小2乗法です。

2つのデータの平均値は2次平面のデータを1次元に落としたと見ることができ，平均値は情報の損失を最小に抑えて次元を下げたものと言えます。

注1）P（x，y）からy＝xへの垂線の足，つまり，影を落とした点である正射影（m，m）が平均値です。

ここて示した回帰法は，回帰式との垂線を最小にするので，直交回帰法（線形関係式）と呼ばれるものです。

線形関係式については「第１７章　測定値の比較検討方法（Deming 法)」で説明します。



2.3　平均を積分で表す

平均値（ｘの期待値 E[ｘ]）は

$$E[x] \equiv \sum_{i=1}^{n} P_i x_i \qquad \sum_{i=1}^{n} P_i = 1$$

と書けます。Pi は確率です。

例えばサイコロでは各目の出る確率は１／６と考えられます。サイコロの目は１から６が出るので，平均値は各確率に出る数を掛けると

$$\sum_{i=1}^{n} P_i x_i = \frac{1}{6}\times 1 + \frac{1}{6}\times 2 + \frac{1}{6}\times 3 + \frac{1}{6}\times 4 + \frac{1}{6}\times 5 + \frac{1}{6}\times 6 = 3.5$$

と，計算できます。１／６は確率であり，１／６を重みと考えることもできます。

$$\sum_{i=1}^{n} P_i = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$$

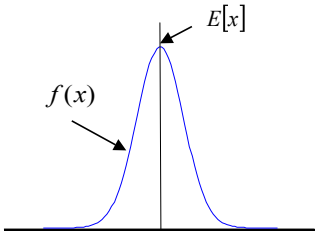
全部の確率を加えると１になります。このことは確率として成り立つ必要条件です。

連続分布での平均値は，和を積分に変えることにより，

$$E[x] \equiv \int_{-\infty}^{\infty} x f(x) dx \qquad \int_{-\infty}^{\infty} f(x) dx = 1$$

となります。平均値は積分により，圧縮，要約した指標ともいえます。

正規分布に従うならば右の図のようになります。



算術平均は

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

でした。

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}x_1 + \frac{1}{n}x_2 + \ldots + \frac{1}{n}x_n = \sum_{i=1}^{n}\frac{1}{n}x_i = \sum_{i=1}^{n} p_i x_i$$

ともなっています。

例えば「第１章　基本的な計算と用語，記号の確認」で述べたように，

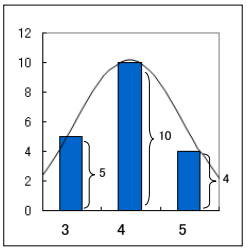
$x_1 = 2, x_2 = 3, x_3 = 4$ の平均は

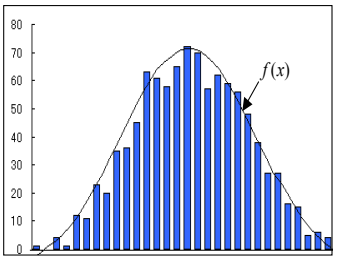
$$\bar{x} = \sum_{i=1}^{n}\frac{1}{n}x_i = \frac{1}{3}\times 2 + \frac{1}{3}\times 3 + \frac{1}{3}\times 4 = 3$$

です。

多くの測定データの誤差を調べると一般的には正規分布になります。
同じことをまた述べますが，下記のようなデータが得られたとします。
データ数を増やせば，$f(x)$ に近似すると考えられます。

データ	頻度	Pi
3	5	5/19
4	10	10/19
5	4	4/19





このデータの平均値は，頻度を確率，または重みと考えることができ，データ数は

$n = 5 + 10 + 4 = 19$

ですので，平均値は

$$\sum_{i=1}^{n} P_i x_i = \frac{5}{19} \times 3 + \frac{10}{19} \times 4 + \frac{4}{19} \times 5 = 3.9$$

です。連続分布を考えると

$$E[x] \equiv \int_{-\infty}^{\infty} x f(x) dx$$

となります。さらに

$$\sum_{i=1}^{n} P_i = \frac{5}{19} + \frac{10}{19} + \frac{4}{19} = 1$$

で，合計するれば確率なので 1 になりますが，連続分布の積分でも

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

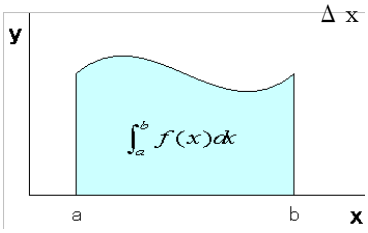
となることが理解できます。

参考 **定積分**

区間 a≦x≦b の面積 S

$$S = \lim_{\Delta x \to 0} \sum_{i=0}^{n} f(x_i) \Delta x = \int_a^b f(x) dx$$

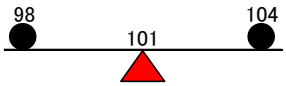
Δ x ： x の微細な幅



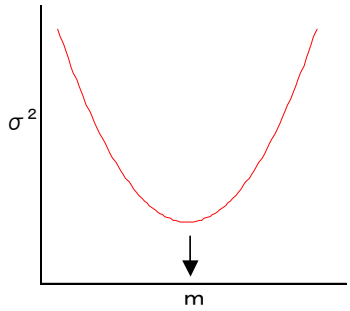
積分のイメージ

関数 $f(x)$	$\int f(x) dx$
x^n （n は -1以外の実数）	$\frac{1}{n+1} x^{n+1}$
$\frac{1}{x}$	$\log\|x\|$
$\sin x$	$-\cos x$
$\cos x$	$\sin x$
e^x	e^x

また，平均値は重心としてとらえることもできます。



ところで，バラツキである分布の幅の大きさを表すのは分散です。
分散から見ると，先ほど残差の 2 乗を最小にすることで平均値を求めましたが，分散 σ^2 を最小にするものが平均値mになります。
つまり，平均値を分布の代表とする理由は，その周りの分散を最小にするためです。[注１）]



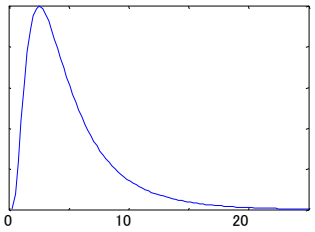
ここで示した下記の式は重要なので，再度記載しておきます。

$$E[x] \equiv \int_{-\infty}^{\infty} xf(x)dx \qquad \int_{-\infty}^{\infty} f(x)dx = 1$$

注１）小針 晛宏：「確率・統計入門」岩波書店（1973 年）

2.4　幾何平均

データが右の図のように対数分布する場合は，算術平均が重心とはなっていません。



もし，対数分布であるならば，**幾何平均**（相乗平均）を使用します。
幾何平均は

$$\bar{x} = \sqrt[n]{x_1 x_2 \ldots x_n}$$

と計算します。

$x_1, x_2 \ldots x_n$ で 1 つでも 0 があると 0 になってしまいますので，0 を避けなくてはなりません。
対数を取ると下記のようになります。積が和になり，通常の平均の計算方法と同じになります。対数変換した分布は正規分布になります。計算した後で対数を戻します。

$$\log \bar{x} = \log \sqrt[n]{x_1 x_2 \ldots x_n} = \frac{\log x_1 + \log x_2 + \ldots + \log x_n}{n}$$

算術平均（相加平均）と幾何平均（相乗平均）は

$$\frac{x_1 + x_2 + \ldots + x_n}{n} \geq \sqrt[n]{x_1 x_2 \ldots x_n}$$

x は負でない実数とする

の不等式が成り立ちます。このことは対数正規分布の上の図から，左に傾いていることからも想像できます。$x_1, x_2 \ldots x_n$ で 1 つでも 0 があると 0 になり，「自明な不等式」になります。

平均値は普段使用しますが，その重み付けなど，意外と計算方法は悩むものです。
平均値を計算するときに，データ数による重みや，分散による重みが妥当な場合があります。

平均値の歴史は古く，いつ誰が発見しかは不明のようですが，この統計量は強力で感覚的にも説得力があり，誰でもが普段使用している統計量です。

簡単に思える平均値について，無駄な難しいことを述べたように感じるかもしれませんが，ここで述べた，重みの考え方，微分して極値を求めること，Σを∫に変えたこと，分布を考えたことは，これから述べることを理解する上で大切です。

参考１　**期待値Ｅ**（Expectation）は平均値で近似できます。

E[ ]，V[ ]は一種の演算子として扱うことが可能です。V[ ]は分散です。

下記の式でＣは定数で，$\mathrm{cov}(X,Y)$は共分散です。共分散については本文中の「第６章　相関係数の２乗とは何か」で述べます。

$$E[X \pm Y] = E[X] \pm E[Y]$$
$$E[CX] = CE[X]$$
$$E[X \times Y] = E[X] \times E[Y]$$
$$V[CX] = C^2 V[X]$$
$$V[X \pm C] = V[X]$$
$$V[X \pm Y] = V[X] + V[Y] \pm 2\,\mathrm{cov}(X,Y)$$

この計算の基本則に従い，平均と分散の変形が簡単に行えます。

例えば，$E[X \pm Y] = E[X] \pm E[Y]$　ですから，X＋Y の平均は，X の平均と Y の平均を後で加えても良いことになり，

$$\frac{1}{n}\sum(x_i \pm y_i) = \frac{1}{n}\sum x_i \pm \frac{1}{n}\sum y_i$$

となります。

参考２　下記の不等式

$$\frac{x_1 + x_2 + \ldots + x_n}{n} \geq \sqrt[n]{x_1 x_2 \ldots x_n}$$

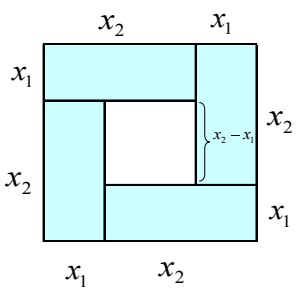
は高校でも習いますが，証明は多くの方法が知られているようです。ｎ＝2 についての証明の一つを紹介しておきます。

これは数学セミナー10　2008（日本評論社）「もっと問題を解いてみよう　相加・相乗平均」中島 匠に記載されていたものです。

$$\frac{x_1 + x_2}{2} \geq \sqrt{x_1 x_2}$$

の証明は下記のようになります。

x_1とx_2の不等式は入れ換えても変わらないから$x_1 < x_2$として，右の図が得られます。



右の図の面積から

$(x_1 + x_2)^2 \geq 4x_1 x_2$

が読み取れ，両辺を 4 で割り，平方根をとると

$$\frac{x_1 + x_2}{2} \geq \sqrt{x_1 x_2}$$　　となります。

$(x_1 + x_2)^2 = 4x_1 x_2$になるのは，$x_1 = x_2$の場合であることも解ります。

第３章　シンプソンのパラドックス

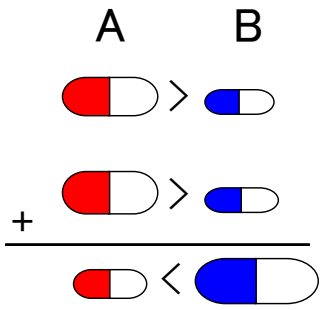
これから述べることはデータを扱う上での大切な事柄を含んでいます。また，少し不思議な現象です。この現象は多くの人を悩ませ，今も問題となっていて，完全に解決されてはいないようです。

ここで示す例は「aha! Gotcha 2」Martin Gardner （竹内 郁雄　訳）日経サイエンス社　の「１１　ミス・ロンリーハート」を元にしています。

元々はシンプソンのパラドクスとして知られているものです。Sympson, E, H, Journal of the Royal Statistical Society , B,vol14 p238-241,1951

薬剤 A と B があり，ある疾患に投与して治った比率を男女について調べたとします。

	薬剤 A	薬剤 B	効果
男性	$\frac{5}{11}$	$\frac{3}{7}$	A>B
女性	$\frac{6}{9}$	$\frac{9}{14}$	A>B
合計	$\frac{11}{20}$	$\frac{12}{21}$	A<B



男性では薬剤 A で 11 人中 5 人，薬剤 B で 7 人中 3 人が治っています。

A＝5／11＝0.45

B＝3／7＝0.43

で，A の方が治る確率が高いことになります。

女性でも薬剤 A で 9 人中 6 人，薬剤 B で 14 人中 9 人が治っていて，

A＝6／9＝0.67

B＝9／14＝0.64

なので，A の方が治る確率が高いと言えます。

男女ともに薬剤 A の方が治る確率が高いので，治療効果は A＞B だと考えられます。

しかし，実際に A と B の薬剤で治った男女の人数を合わせてみると，

A（5＋6）／（11＋9）＝11／20＝0.55

B（3＋9）／（7＋14）＝12／21＝0.57

で A＜B と逆転し，B の方が治る確率が高くなってしまします。

性別を無視すると，B の薬剤の方が治療効果は高いと言えます。

男性 A＞B と女性 A＞B を加えて，合計では逆転し A＜B になる「パラドクス」が起きます。このよう現象はよく起き，治験の臨床試験でも現れることがあります。性別を考慮す

ると A＞B であるのに，全体では A＜B で効果があることに，計算上，統計上の間違いはありません。

このパラドクスは実際にトランプでも確認できます。つまり，トランプの賭けで，この比率で２人で勝負すると A＞B になり，次にトランプを合わせると，逆転して A＜B となり B が勝つ，不思議な現象が起きます。

原因は，各データ数による「重み」にあります。下記の表で「合計」の比率の A＝0.55　B＝0.57　は，「第２章　平均値の計算方法を考える」で示した「重み付き平均」です。

データ数は 10 倍しても同じです。

	薬剤 A　比率	薬剤 B　比率	
男性	$\frac{5}{11}$ =0.45	$\frac{3}{7}$ =0.43	A>B
女性	$\frac{6}{9}$ =0.67	$\frac{9}{14}$ =0.64	A>B
合計	$\frac{11}{20}$ =0.55	$\frac{12}{21}$ =0.57	A<B

重み付き平均は

$$\frac{w_1}{w_1+w_2}x_1+\frac{w_2}{w_1+w_2}x_2$$

です。人数を重み w_i として薬剤 A の平均は，

$$\frac{11}{20}\times 0.45+\frac{9}{20}\times 0.67=0.55$$

薬剤 B の平均は

$$\frac{7}{21}\times 0.43+\frac{14}{21}\times 0.64=0.57$$

です。

つまり，人数による重みがかかり，全体では逆の結果となっています。逆転した結果について，結論，判断を下すのは難しくなります。

シンプソンのパラドクスは実験をする場合の考慮すべき問題点を示しています。[注１）]

シンプソンのパラドクスを回避する基本的な方法は，偏りのないデータを集めることです。重みが同じになるように試験計画を立てばよいですが，シンプソンのパラドクスが生じたときの解釈はそのケースにより異なり困難です。

この現象があることを知らないと，誤った結論を述べる可能性もあります。

注１）ランダム化については「第７章　小標本のランダム化の注意点」を参考にして下さい。

参考 1　**分数**の計算について

シンプソンのパラドクスは分数の次のような問題を含んでいます。分数は小学校で習いますが難解です。

① 調査で 50／100（人）と 1／2（人）は同じ 0.5 でも，100 人を調べた方が信頼できます。つまりデータ解析では，安易な約分や小数にしない方が誤りを犯す危険性は少ないといえます。データ数に意味があります。

② 1／2 ＋ 1／3 ならば，通分して 1／2 ＋ 1／3＝3／6 ＋ 2／6＝5／6 と計算します。
しかし，今回の計算は，1／2（○●） ＋1／3（○●●）＝ 2／5（○○●●●）と計算しているように見え，子供に説明すると混乱を招きます。

遠山啓が推奨した分数での「量から数」への考えや，森毅が「数の現象学」朝日選書（1989 年）で述べているように，分数には多くの混乱を招く概念があります。

「分割分数」と「量分数」の違いを明確にしないと混乱の元となるとよく言われますが，分数は理解し難いものです。「分割分数」「量分数」「仮分数」「真分数」「帯分数」などの「表現」は混乱の元です。

参考 2　**三すくみ**について

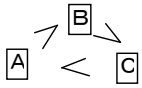
「三すくみ」とは強さが　なめくじ＞蛇＞蛙＞なめくじ　のような関係にあるものです。シンプソンのパラドクスは，三すくみのような非推移的な関係にあります。

ブラッドリー・テリーのモデル（Bradley-Terry's model）

プロ野球，相撲などの勝敗で相性（苦手な相手など）があると「三すくみ」の関係になります。
下の表では勝数からＡが強く，Ｃが一番弱いことになりますが，A>B>C>A の関係があります。
Ａは一番弱いＣに負けています。

	A	B	C	勝数
A	－	8	3	11
B	2	－	8	10
C	7	2	－	9
負数	9	10	11	30

関係：



（トーナメント方式ならば，一番弱いＣが優勝する可能性があります。）
このような表の場合に「三すくみ」の関係になっているのかをブラッドリー・テリーのモデルで調べることができます。計算方法は下記の本などに記載されていますが，適合度検定などでは「三すくみ」の関係があるかを注意する必要があります。

参考：東京大学教養学部統計学教室編：「基礎統計学 III　自然科学の統計学」東京大学出版会（1992 年）

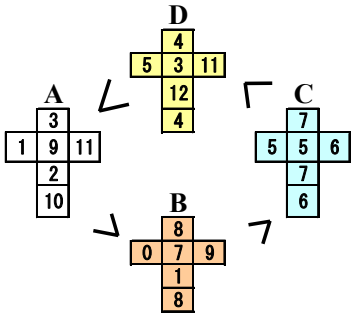
参考 3　**エフロンのダイスのパラドックス**について

「三すくみ」の関係にある楽しい不思議なサイコロ（ダイス）を紹介します。(読み飛ばしてもかまいません。)

これはエフロン(Efron)が考案したもので，エフロンは統計学者で Bootstrap 法(1979 年)の考案者として有名です。

下記のサイコロＡからＤを作製してみてください。このサイコロで賭けをすれば勝てます。

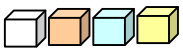
Ａ＞Ｂ＞Ｃ＞Ｄ＞Ａ…で強くなっています。サイコロの展開図と実際に作製したサイコロを下記に示します。



ルール：サイコロを振って大きい数が出た方が勝ちとします。

手順

１）相手に好きな，一番強いと思う A から D のサイコロを１つ選んでもらう。



Ａ　Ｂ　Ｃ　Ｄ

２）もし相手がＤなら，Ｃを選ぶ。同様に，相手がＣならＢを選び，相手がＢならＡ選び，相手がＡならＤを選びます。

３）サイコロを振って大きい数が出た方が勝ちとします。

最初に選ぶのは相手であるため公平な賭けと思うが，相手は負けてしまう。

確率とルール

サイコロの各数の組み合わせを計算すると確かに 11/17＝0.647 で勝てます。
しかし，p=0.647 の勝ちでは，必ず勝つわけではありません。
そこで確率を計算してみました。
３回先に勝った方が勝ちにするルールだと，76%で勝てます。

第４章　ベイズの定理 - 間違え易い確率の計算 -

少し間違え易い確率の計算を示します。ある人がある感染症の検査を受け（＋）になった時に，どの程度感染の可能性があるのか知りたい場合があります。国内でのその感染者の割合は 0.１％程度であるとします。

ある感染症で A の検査をした時に陽性（＋）率を調べて下記の表を得たとします。

	感染者	健常人	合計
A 検査＋	90	5	100
A 検査－	10	95	100
合計	100	100	200

このときに，表の数値から感染者で陽性（＋）になる確率は

$$\frac{90}{100}\times 100 = 90(\%)$$

で，90%正確に陽性あると判断できます。健常人ならば

$$\frac{5}{100}\times 100 = 5(\%)$$

で 5%の確率で擬陽性になります。このためほぼ確実な検査と言えます。
さらに，検定をしてみると p＜0.01 となり，当然有意差が認められ，この検査は有効であると判断されます。

ある人が病院に行き，この検査で陽性（＋）になった時に，上記の結果から間違える確率は 5%（擬陽性）で，感染していれば 90%陽性になるので，「感染している確率は 9 割」です。このため，自分は陽性（+）であると判断したとします。

この表は感染者と判っていている人を調べて 90%が（＋）で，健常者を調べて 95%が（－）で正しいことを示しています。

一見正しいように思えますが，この判断は大きな間違いを犯しています。この誤りを説明します。

感染者の割合が 0.1%程度である情報を見落としています。99.9%の人は感染者ではではありません。

陽性（＋）であった時の事後確率を計算する必要があります。

表のデータと感染者の割合が 0.1 %程度あることを考慮すると，後で説明しますがベイズの定理から下記のように計算できます。

$$P(\theta_i|x)=\frac{p(\theta_i)p(x|\theta_i)}{\sum p(\theta_j)p(x|\theta_j)}=\frac{p(\theta_1)p(x|\theta_1)}{p(\theta_1)p(x|\theta_1)+p(\theta_2)p(x|\theta_2)}$$

$$=\frac{\frac{90}{100}\times 0.001}{\frac{90}{100}\times 0.001+\frac{5}{100}\times 0.999}=0.018$$

検査で陽性（＋）でも感染者である確率は 1.8%でしかないことになります。
つまり，検査が陽性（＋）でも，この検査のみでは感染している確率は非常に低いことになります。

ベイズの定理

上記のような計算が必要な場面はよくあります。ベイズの定理からベイズ統計へとつながりますが，基本となるベイズの定理を簡単に説明します。[注 1)]
ベイズの定理は「条件付き確率」の定理で，事後確率を計算しています。式は下記のようになります。

$$P(\theta_i|x)=\frac{p(\theta_i)p(x|\theta_i)}{\sum p(\theta_j)p(x|\theta_j)}$$

$p(\theta_i)$ は θｉ が起きる確率で， $p(x|\theta_i)$ は事象 θｉ が起こった条件下での，ｘの起きる条件付き確率を表しています。

先ほどの例も簡単ですが，理解するために，さらに簡単な例を示します。

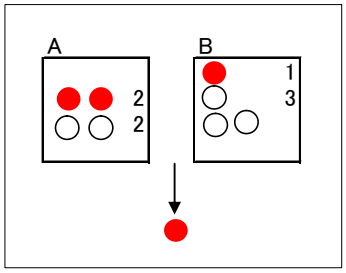
注１）ベイズの定理を基本としてベイズ統計と呼ばれる統計手法（推定，検定など）があります。ベイズ統計の信仰者はベイジアンと呼ばれます。ベイズ統計は，本文の「第１３章　最尤法について」で述べる尤度原理の理解が必要です。ベイズ統計に関する本を紹介しておきます。

渡辺 洋：「ベイズ統計学入門」福村出版（1999 年）

中妻 照雄：「入門ベイズ統計学」朝倉書店（2007 年）

例

A の箱に赤球が２個と白球が２個あり，B の箱に赤球が１個と白球が３個あるとします。A と B の箱は外見が同じで区別がつかないとします。球を１つ取り出してみたら，赤球であったとします。この赤球は A の箱からか B の箱からか判らないとして，A の箱から取り出したとする確率を計算してみます。

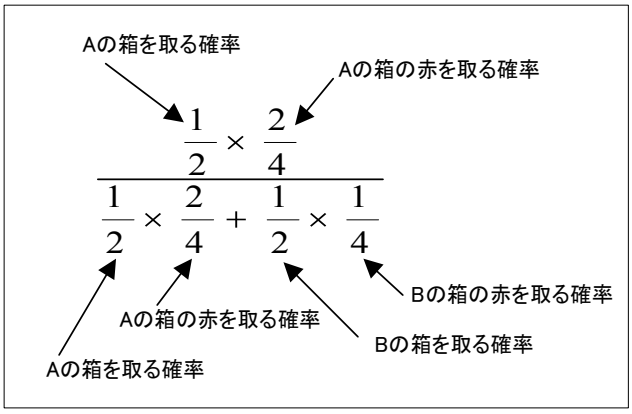


$$P(\theta_i|x)=\frac{p(\theta_i)p(x|\theta_i)}{\sum p(\theta_j)p(x|\theta_j)}$$

$$=\frac{p(\theta_1)p(x|\theta_1)}{p(\theta_1)p(x|\theta_1)+p(\theta_2)p(x|\theta_2)}=\frac{\frac{1}{2}\times\frac{2}{4}}{\frac{1}{2}\times\frac{2}{4}+\frac{1}{2}\times\frac{1}{4}}=\frac{2}{3}=0.67$$

より，A の箱から取り出した確率は 67%です。A の箱から取り出した可能性が高いことになります。

少し解り難いので式に説明を加えておきます。下の数値の意味をよく見ると理解が容易で納得できると思います。



例で示した赤球を得た後で，A の箱であるか B の箱であるか，A の母集団か B の母集団かの，尤（もっとも）らしさを計算したことになります。

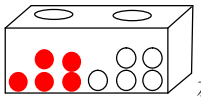
ここで述べたベイズの定理は，「第１３章　最尤法について」の尤度関数の話につながります。[注1）]

注１）ベイズ統計は「主観確率」を取り込むことが可能です。つまり，経験や感も取り込めます。
頻度論ではこのよな扱いはしません。濃度測定でいつも使用している統計処理のほとんどが，頻度論的な扱いをしています。
本文「第１３章　最尤法について」で述べる最尤法は，ここで示したベイズの定理から導けます。

参考　データ解析で確率を計算する必要がある場合がありますので, 確率計算の基本的なものを説明します。

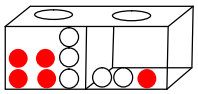
確率の計算をすると，人の感覚と少し異なる場合があります。簡単な問題を２つ考えてみます。

問１　入り口が２つある箱があり，箱の中には合計で赤球５個，白球５個が入って場合は，赤球を取る確率はいつでも 1／2 のように感じます。



左の箱の中から，赤球を取る確率は $\frac{5}{10}=0.5$　です。（確率）

もしも，箱の中を２つに区切り，赤球と白球のバランスを下記のように崩すと，箱の中の合計は赤球５個，白球５個でも，赤球を取り出す確率は変わります。



赤球を取る確率は $\frac{1}{2}\times\frac{4}{7}+\frac{1}{2}\times\frac{1}{3}=0.45$　（乗法定理，加法定理）

で１／2 の確率ではなくなります。入り口は２つでも，中にあるのは半々なのだから，１／2 であると思う予測がハズレます。感覚的には少し不思議な感じがします。

問２　サイコロで１の目の出方は 1／６の確率であると予想されます。

６回投げて，１の目が一度は出るのだろうか。一度でも出たら終了します。賭けるなら，6 回投げて１の目が出るのか，出ないのか，どちらに賭けますか。

この問いに，一回投げたときは 1／６の確率だから一度は出るとする意見と，６回投げるのだから賭けは 1／２になる，または，出ない確率の方が高いなど，意見は分かれます。

１回投げて１でない確率は 5／6 なので，６回投げて一度も出ない確率は積になり，

$\left(\frac{5}{6}\right)^6=0.33$

となります。このことから，

１－0.33＝0.67　（余事象）

67%で１の目が一度は出ることになります。

（1 回目に 1 が出る確率，2 回目にはじめて 1 が出る確率と，加えていっても計算結果は同じです。）

一度は１の目が出る方に賭けた方が有利です。

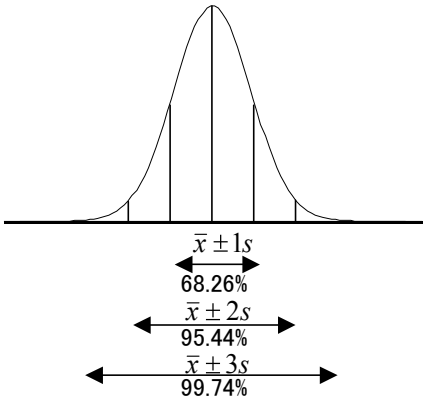
（注）６回投げて，１の目が１回だけ出る確率ではありません。

１の目が１回だけ出る確率は，$\frac{1}{6}\times 6\times\left(\frac{5}{6}\right)^5=0.40$　になります。

第５章　標準偏差はｎ－1 で割っても不偏推定量にならない

標準偏差ｓ（または s.d.）について考えてみます。

普段使用するｎ－1 で割る標準偏差が，偏りを補正した標準偏差であると思っている人がいますが誤りです。



関数電卓やエクセルなどの表計算ソフトで簡単に標準偏差は計算できます。ｎ で割るのか，ｎ－1 で割るか質問を受けることがありますが，ｎでもｎ－1 でもない，「不偏」標準偏差を計算することが必要な場合があります。

濃度測定の場合，ほとんどが母集団の一部の標本を測定していると考えられ，不偏分散の平方根の意味で，標本標準偏差のｎ－1 で計算した方が良いと考えられます。

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

この方法をｎ－１法とします。

ただし，標準偏差はｎ=20 以下ならｎ－1 で割ると言う人がいますが，これは誤りです。ｎ＞20 でもｎ－1 で計算します。

よく質問されるエクセルでの計算は，下記の関数を使用します。

=STDEV(数値 1:数値 2)

ｎ－1 で「分散」を計算した場合は偏りがない不偏推定量ですが，その平方根である「標準偏差」は不偏推定量ではありません。

つまり，普段よく使用するｎ－1 法の標準偏差には偏りがあります。このことはあまり知られていないようです。

5.1　不偏推定量

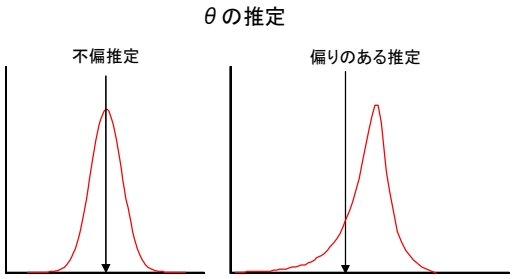
まず，不偏推定量とは何かについて説明します。**不偏推定量**とは，母集団の推定値として偏りがないことです。

つまり，母集団の母数（パラメータ） θ である母平均 μ や母分散 σ^2 の推定値として，標本平均 $\bar{x}$ や標本分散 V を計算します。
これらに偏りがない場合は，不偏推定量と呼びます。

母数 θ を推定する統計量 T の期待値を E[T]とすると，

$$E[T]-\theta=\hat{\theta}-\theta=0$$

となるものが不偏推定値です。
良い推定値として右の図からも，当然偏りのない不偏推定量を使用すべきです。



当然，正規分布の標本平均 $\bar{x}$ は，母平均 μ の推定値として偏りがない不偏推定量です。

5.2　不偏分散

標準偏差の前に分散から述べます。**不偏分散**が n－1 で割ることにより得られることを簡単な例で示します。
正規分布は $N(\mu,\ \sigma^2)$で示すように，母平均 μ と母分散 σ^2 で決まる分布です。分散 σ^2 の平方根が標準偏差 σ です。

分散は

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

で計算します。
母平均 μ は一般に不明ですから，標本データからの推定値の標本平均 $\bar{x}$ を使用すると，

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

となります。
計算に母平均 μ の代わりに標本平均 $\bar{x}$ を使用しても，問題はないように思えます。
しかし，上記の式の s^2 では母分散 σ^2 を少なめに推定することになります。

例えば，母平均μ＝100 であることがあらかじめ解っているとして，下記のデータが得られたとします。

No.	データ
1	105
2	100
3	110
4	99
5	100
標本平均	102.8

母平均μ＝100 を使用して分散を実際に計算すると，

$$s_{\mu}{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-100)^2 = 25.2$$

となります。

しかし，標本平均＝102.8 で分散を計算すると

$$s_{\bar{x}}{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-102.8)^2 = 17.4$$

で，少なめに計算されます。

この例からも得られたデータの平均を使用すると，必ず分散が小さく計算されます。

母平均μ＝100 は通常は不明なので，標本平均から分散を推定する場合は補正するためにｎではなく，ｎ－１で分散を計算します。

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

これを**不偏分散**と呼びます。

少し正確にｎ－1 であることを示します。

$s_{\mu}{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$ と $s_{\bar{x}}{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ は一致しないことを示しましたが，

データがＮ（μ，σ^2）に従う場合，得られたデータの平均値 $\bar{x}$ の，平均値と分散は

Ｎ（μ，σ^2／ｎ）に従います。[注1)]

注１）標準誤差として「第１章　基本的な計算と用語，記号の確認」で示しました。

このことが成り立つことは，「第１０章　誤差伝播の法則－不確かさの計算－」の最後の参考で，誤差伝播の法則から導く方法を述べます。

$$E\left[s_{\bar{x}}{}^2\right]=E\left[\frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^2\right]=E\left[\frac{1}{n}\sum\{(x-\mu)-(\bar{x}-\mu)\}^2\right]$$
$$=E\left[(x-\mu)^2\right]-E\left[(\bar{x}-\mu)^2\right]$$
$$=\sigma^2-\frac{\sigma^2}{n}$$
$$=\frac{n-1}{n}\sigma^2$$

で,

$$E\left[s_{\bar{x}}{}^2\right]\neq\sigma^2$$

となり,

$$E\left[\frac{1}{n-1}\sum(x-\bar{x})^2\right]=\sigma^2$$

であることが示せます。

以上より，ｎではなくｎ－1 を使用します。分散の計算はｎ－１を使用すると「不偏推定量」になります。

標本平均で母平均を推定したことにより，自由度が減り，ｎ－１で分散を計算したと考えることもできます。

もしも，母平均μが解っていれば母平均μを使用し，ｎ－１を使用する必要はありません。また，全てのデータで計算した場合はｎで割ります。通常は母集団の１部の標本から母集団を推定するので，ｎ－１で計算します。

標準偏差は分散の平方根なので,

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

です。これを，**不偏分散平方根（標本標準偏差）**と言います。

何の問題もないように思えますが，このｎ－1 法の標本標準偏差は不偏推定量にはなりません。

つまり，分散はｎ法から，ｎ―1 法で補正できましたが，標準偏差はｎ―1 法でも補正できません。このことは後のシミュレーションなどでも示します。

5.3　不偏標準偏差を計算する

標準偏差の偏りを補正した**不偏標準偏差**は，下記の式で計算する必要があります。[注1）]

$$\sigma=\frac{\sqrt{n-1}}{\sqrt{2}}\cdot\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Γ（）：ガンマ関数

下記の式で計算される標本標準偏差を標準偏差として通常使用していますが，不偏推定量ではなく，偏りのある標準偏差を使用していることになります。

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

標準偏差の計算はｎで割るよりｎ－1 で割る方が良いが，ｎ－1 でも偏りのない好ましい推定値とはなっていません。

注１）母標準偏差の推定式は

$$\sigma=\frac{\sqrt{n-1}}{\sqrt{2}}\cdot\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

です。

この式はχ²分布に従うことを利用して導くことができす。つまり，

$$\chi^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

はχ²分布に従いますが，χ²分布は下記の式です。

$$\begin{cases} f_v(\chi^2)=\dfrac{1}{2^{v/2}\Gamma\left(\frac{v}{2}\right)}\left(\chi^2\right)^{v/2-1}e^{-\chi^2/2} & \chi^2>0 \\ f_v(\chi^2)=0 & \chi^2\le 0 \end{cases}$$

このことから母標準偏差の上記の推定式が導けます。

参考 **χ²分布**は下記のように定義します。

『独立に標準正規分布Ｎ（０，１²）に従う変数を u_i としたとき

$u_i \sim$ i.i.d. N(0,1²)

その２乗の和

$$\chi^2=u_1^2+u_2^2+\ldots+u_n^2=\sum_{i=1}^{n}u_i^2$$

の従う分布を，自由度ｎのχ²分布といいます。』

確率密度関数は先ほどの関数です。

$$\begin{cases} f_v(\chi^2)=\dfrac{1}{2^{v/2}\Gamma\left(\frac{v}{2}\right)}\left(\chi^2\right)^{v/2-1}e^{-\chi^2/2} & \chi^2>0 \\ f_v(\chi^2)=0 & \chi^2\le 0 \end{cases}$$

通常計算する標準偏差に下記式の係数

$$C=\frac{\sqrt{n-1}}{\sqrt{2}}\cdot\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$$

を掛ければ，不偏分散と同様に母標準偏差の
不偏推定値が得られます。補正係数Ｃのｎ＝10 以下の値
を右の表に示しました。

n（データ数）	補正係数 C
2	1.2533
3	1.1284
4	1.0854
5	1.0638
6	1.0509
7	1.0424
8	1.0362
9	1.0317
10	1.0281

不偏標準偏差は右の表の係数を使用すれば，下記の式で
計算できます。

$$s_c = C\cdot\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

標準偏差の計算はｎで割るよりはｎ－1 で割る方が良いが，それでも少なめの標準偏差を計算していますので，ここに示す不偏標準偏差ならば，少なめに見積もることはありません。

例として，下記のデータで標準偏差を，ｎ法，ｎ－１法，Ｃでの補正
（ｎ＝5 からＣ＝1.0638 で補正）で計算した値を示します。

	データ
1	5.22
2	4.81
3	5.10
4	5.21
5	4.91

平均値	5.05

標準偏差の計算値の比較

標準偏差を n法で計算	0.16
標準偏差を n-1 法で計算	0.18
標準偏差の不偏推定値 （Cで補正）	0.19

5.4　シミュレーションによる標準偏差の計算の比較

シミュレーションでよく使用するｎ－１法よりも，不偏標準偏差が優れていることを確認してみます。

正規乱数で平均μ＝100，標準偏差σ＝2.0 のデータを 10 個，５回発生させて各標準偏差を算出し，その平均を計算してみました。[注 1)]

X=100　　σ=2.0

	1回目	2回目	3回目	4回目	5回目	
1	100.6	100.5	98.5	98.3	100.2	
2	101.1	99.3	98.9	97	101.5	
3	101.7	102.7	100	102.3	97.4	
4	101.2	100.5	98	97.2	101.3	
5	97.8	98.5	99.9	101.5	98.8	
6	99.4	101.1	97.1	98.5	99.7	
7	96.7	104.9	102.7	99.3	100.4	
8	98.5	102.4	96.2	97.2	99.1	
9	99	102.1	98.6	98	99.7	
10	99.2	99.2	102.4	103.3	97.3	
N=10						S.D.の平均
S.D.(n法)	1.53	1.87	1.99	2.17	1.37	1.79
S.D.(n−1法)	1.61	1.97	2.1	2.29	1.45	1.88
不偏S.D.	1.66	2.03	2.16	2.36	1.49	1.94

母標準偏差の推定はｎ法→ｎ―1 法→補正係数による不偏標準偏差の順でσ＝2.0 に近く，良い推定値が得られることが解ります。

実際のデータ解析では，10 個程度のデータでの標準偏差の計算ではバラツキがあり，どの方法でも大差ないとも言えます。このため，どの方法で計算するかは趣向の問題のようにも感じます。

しかし，注意すべきことは，上の表からも解るように，よく精度を調べるのに 10 個程度のデータを使用することがありますが，多少小さく見積もっている可能性があります。例えば，残留農薬検査の一斉分析で，一度に 200 農薬をｎ＝5 で測定して，各農薬の標準偏差をｎ―1 法で計算し，相対標準偏差 RSD の平均を計算したとすると，この例と同様に明らかに少なく見積もってしまいます。

注 1）シミュレーションのデータ数が少ないと思われますが，可能ならばさらにデータを増やして確認してみて下さい。

第 5 章　標準偏差はｎ－1 で割っても不偏推定量にならない

通常は，不偏分散の平方根の意味で，標本標準偏差のｎ－1 法で計算します。

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

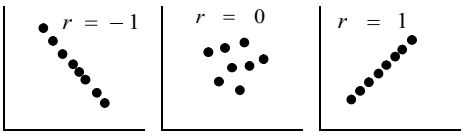
しかし，この標本標準偏差は不偏推定量ではなく，小さく見積もっていますので，下記の不偏標準偏差を計算する場合もあります。[注1）]

$$s_C = \frac{\sqrt{n-1}}{\sqrt{2}} \cdot \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

注１）石居 進：「生物統計学」培風館（1975 年）では，ｎ≦10 では不偏標準偏差を使用するのが普通であると述べています。

第６章　相関係数ｒとＲの２乗とは何か

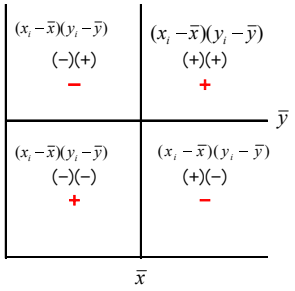
6.1　相関係数の２乗とは



相関係数ｒは－1≦ｒ≦1 の範囲で，0 が無相関で，1 なら完全に直線上に並ぶことを示していることは，よく知られています。しかし，エクセルなどで R^2 が計算されたり，文献の回帰分析で R^2 が記載されていたりします。この「相関係数の２乗」とは何でしょうか。

相関係数は**共分散** Cov(x,y)と言うｘとｙの関係を示す統計量を，－1≦ｒ≦１になるように規格化したものです。

共分散は $\sum(x_i-\bar{x})(y_i-\bar{y})/n$ で計算します。



右の図より，共分散の式で＋及び－を考えると，相関係数の－1≦ｒ≦１の意味が理解できます。
相関係数の計算は電卓やエクセルなどで行えますが，相関係数は

$$r \equiv \frac{\text{共分散}}{x\text{の標準偏差}\times y\text{の標準偏差}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

と定義されます。分母により規格化されています。[注1)]

この相関係数の２乗の R^2 は，「**寄与率**」または「**決定係数**」と呼ばれます。
標本データを y_i として回帰式からの推定値を Y_i とすると

$$R^2 = \frac{\sum(Y_i-\bar{Y})^2}{\sum(y_i-\bar{y})^2} = \frac{S_{YY}}{S_{yy}} = \frac{SR}{ST} = 1-\frac{Se}{ST}$$

S_{YY}：回帰式（直線回帰式，高次回帰式，重回帰式等）の推定値Ｙの偏差平方和　SR：回帰平方和

S_{yy}：従属変数ｙの偏差平方和　Se:残差平方和　ST：全平方和

となります。

注１）共分散の期待値は

$$\mathrm{cov}(x,y) = \sigma_{xy} = E[[x-E(x)][y-E(y)]]$$　です。

面倒な平方和の計算をしなくても，エクセルの「分析ツール」で相関係数は計算できます。次の例で説明します。

	x	y
1	1	2
2	2	6
3	3	7
4	4	10
5	5	14

エクセルの「分析ツール」で回帰分析を行うと，下記のような，相関係数，決定係数，分散分析表が出力されます。

回帰統計	
重相関 R	0.9850366
重決定 R2	0.970297
補正 R2	0.960396
標準誤差	0.8944272
観測数	5

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	78.4	78.4	98	0.0021923
残差	3	2.4	0.8		
合計	4	80.8			

変動：平方和

r ＝0.985 が得られたとすると，R^2は相関係数を 2 乗したもので 0.985×0.985＝0.97 となります。

この R^2＝0.97 に 100 を掛けると%になり，寄与率は 97%となります。寄与率は，実験計画の分散分析の寄与率と同じです。全体のデータのバラツキで回帰式により説明できる割合，つまり寄与する割合を示しています。

また，上の分散分析表の「回帰」と「合計」の値から

$$R^2 = \frac{SR}{ST} = \frac{78.4}{80.8} = 0.97$$

と計算することもできます。

分散分析表に寄与率%（決定係数%）を書き込むと，

分散分析表

	自由度	変動	分散	観測された分散比	有意 F	寄与率(%)
回帰	1	78.4	78.4	98	0.0021923	97
残差	3	2.4	0.8			3
合計	4	80.8				100

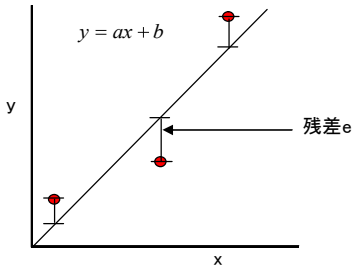
変動：平方和

になります。

ＳＴ（総平方和）＝ＳＲ（回帰平方和）＋Ｓｅ（残差平方和）

80.8（総平方和）＝78.4（回帰平方和）＋2.4（残差平方和）

100%＝97%＋3%



R^2は，独立変数ｘで従属変数ｙとしたときに，ｘの値でｙの値が決定できる割合を示す係数と言えます。

相関関係を問題にする場合はｒを使用しますが，検量線などの回帰式などで，回帰式の適合の良し悪し，説明できる割合，寄与する割合などを示す場合は，R^2を使用します。

注 1）原点回帰 $y = ax$ は注意が必要です。

注意すべきことは，定義の違いで，原点回帰では相関係数が負になることがあります。

相関分析と回帰分析は異なり，通常は原点回帰の相関係数は考えません。

原点回帰の場合はデータの重心が $(\bar{x}, \bar{y})$ を通らないためで，下記の式を使用する場合があります。

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - Y_i)^2}{\sum_{i=1}^{n} y_i^2}$$

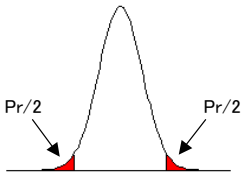
参考：中村 永友　土屋 高宏：「焦点をもつ回帰直線群の推定とその周辺」応用統計学会 Vol.36,No.1(2007),31-50

6.2　相関係数の検定

相関係数から，相関が有ると言えるのかどうかを知りたい場合があります。
相関係数の検定のための表を下記に示します。
自由度はφ＝n－2 で，両側確率の値が表になっています。
例えば，20 個のデータでは，自由度φ＝18 の所が両側検定の時の値です。[注１）]
φ＝18 で相関係数が r ＝0.3783 よりも強ければ，相関があると言えます。

相関係数の検定

φ \ Pr	0.1	0.05	0.02	0.01
10	0.4973	0.5760	0.6581	0.7079
11	0.4762	0.5529	0.6339	0.6835
12	0.4575	0.5324	0.6120	0.6614
13	0.4409	0.5139	0.5923	0.6411
14	0.4259	0.4973	0.5742	0.6226
15	0.4124	0.4821	0.5577	0.6055
16	0.4000	0.4683	0.5424	0.5897
17	0.3887	0.4555	0.5285	0.5751
18	0.3783	0.4438	0.5155	0.5614
19	0.3687	0.4329	0.5034	0.5487
20	0.3598	0.4227	0.4921	0.5368
25	0.3233	0.3809	0.4451	0.4869
30	0.2960	0.3494	0.4093	0.4487
35	0.2746	0.3246	0.3810	0.4182
40	0.2573	0.3044	0.3578	0.3932
50	0.2306	0.2732	0.3218	0.3541
60	0.2108	0.2500	0.2948	0.3248
70	0.1954	0.2319	0.2737	0.3017
80	0.1829	0.2172	0.2565	0.2830
90	0.1726	0.2050	0.2422	0.2673
100	0.1638	0.1946	0.2301	0.2540
近似式	$\frac{1.645}{\sqrt{\phi+1}}$	$\frac{1.960}{\sqrt{\phi+1}}$	$\frac{2.326}{\sqrt{\phi+2}}$	$\frac{2.576}{\sqrt{\phi+3}}$



$$P_r = 2\int_r^1 \frac{\left(1-X^2\right)^{(\phi/2)-1} dX}{B\left(\frac{\phi}{2},\frac{1}{2}\right)}$$

上の表から，大まかに相関係数が 0.5 以上あれば，有意な相関が認められます。
データ数で有意になる相関係数は異なります。データ数が多ければ r ＝0.2 程度でも有意となります。
また，近似式も使用できます。

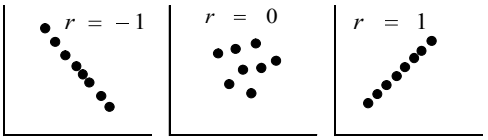
注１）検定については，「第１２章　有意差検定の解釈の誤りと検定に必要なデータ数」で説明します。

参考１　**相関係数**について

相関係数は

$$r \equiv \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})/n}{\sqrt{\sum (x_i - \bar{x})^2/n \sqrt{\sum (y_i - \bar{y})^2/n}}} = \frac{s_{xy}^2}{\sqrt{s_x^2}\sqrt{s_y^2}} = \frac{s_{xy}^2}{s_x s_y}$$



で計算します。少し詳しく説明します。

相関係数は２変量正規分布を仮定し，直線性の強さのを表しています。

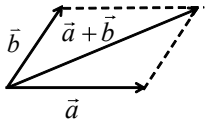
相関係数の表現では，高い，低いは不適切です。負の相関があるために，相関は**「強い」「弱い」**で表現します。

母相関係数は通常 ρ で表し，

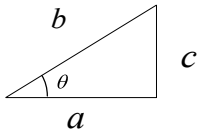
$$\rho = \frac{\sigma_{xy}^2}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

となります。

共分散がベクトルの内積であることを示し，相関係数が $\cos\theta$ であることを，これから示します。



ベクトル　$\vec{a}+\vec{b}$



$$\sin\theta = \frac{c}{b} \quad \cos\theta = \frac{a}{b} \quad \tan\theta = \frac{c}{a}$$

ベクトル a と b の積である内積 a・b は高校で習いますが，この概念は解り難いものです。

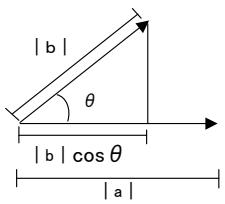
内積は

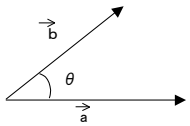
$$a \cdot b \equiv |a||b|\cos\theta$$

$$a \cdot b \equiv a_1 b_1 + a_2 b_2 + \ldots + a_n b_n$$

と，通常は定義します。（$|a|$ はノルムで，内積はベクトルではなく，スカラーになります。）

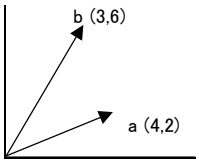
$$a \cdot b \equiv |a||b|\cos\theta \qquad \cos\theta = \frac{a \cdot b}{|a||b|}$$





例えば，右の図のベクトルの内積は $a \cdot b \equiv a_1 b_1 + a_2 b_2 + \ldots + a_n b_n$ から，

4×3＋2×6＝24 になります。



第 6 章　相関係数の 2 乗とは何か

x（x 1，x 2，…x n）と **y**（y 1，y 2，…y n）とし，平均偏差である

$$x_R = \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ x_3 - \bar{x} \\ \cdot \\ \cdot \\ \cdot \\ x_n - \bar{x} \end{pmatrix} \quad y_R = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ y_3 - \bar{y} \\ \cdot \\ \cdot \\ \cdot \\ y_n - \bar{y} \end{pmatrix}$$

を考えます。

このベクトルの長さの 2 乗の和（偏差平方和）を n で割ると，その平方根は標準偏差になります。

$$s_x = \sqrt{\frac{x_R . x_R}{n}} = \sqrt{\frac{|x_R|^2}{n}} = \frac{1}{\sqrt{n}}|x_R|$$

$$s_y = \sqrt{\frac{y_R . y_R}{n}} = \sqrt{\frac{|y_R|^2}{n}} = \frac{1}{\sqrt{n}}|y_R|$$

x R と y R のベクトルの積である，内積を n で割ったものは共分散になります。

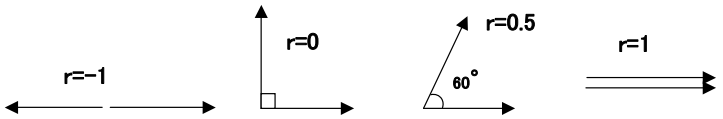
$$s_{xy}^2 = \frac{x_R \cdot y_R}{n}$$

内積は $\frac{a \cdot b}{|a||b|} = \cos\theta$ でしたが，このことから相関係数は

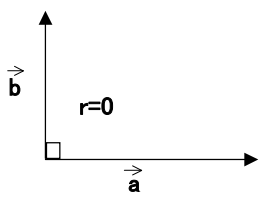
$$r = \frac{s_{xy}^2}{\sqrt{s_x^2}\sqrt{s_y^2}} = \frac{s_{xy}^2}{s_x s_y} = \frac{\frac{x_R . y_R}{n}}{\frac{|x_R||y_R|}{n}} = \frac{x_R . y_R}{|x_R||y_R|} = \cos\theta$$

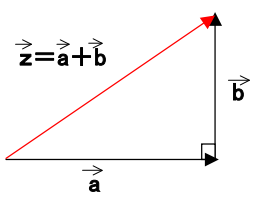
となります。相関係数 r は cos θ で表現できます。

つまり，x R と y R のベクトルを考えると，**標準偏差**の s x と s y は**ベクトルの大きさ**と考えられ，**共分散** s 2 x y は**ベクトルの内積**で，**相関係数**は x R と y R の余弦 **cos θ** です。相関係数が 1 とはベクトルの方向が同じで，相関係数 0 は x R と y R が 90°（直交）となっています。内積の概念を理解していれば，ベクトルにより相関係数のイメージがより明確になります。



統計量	ベクトル
標準偏差	ベクトルの長さ
共分散	内積
相関係数	c os θ





参考２ **相関の検定で t 検定による方法**も参考として示します。

$\rho = 0$ の２次元正規分布に従う母集団から，標本相関係数は

$$f(r) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left((n-1)/2\right)}{\Gamma\left((n-2)/2\right)} \left(1-r^2\right)^{(n-4)/2}$$

の分布に従い，

$$T = \sqrt{\frac{(n-2)r^2}{1-r^2}}$$

は t 分布に従います。

このことから t 検定をします。自由度は $\phi = n - 2$ です。

n＝20 で相関係数 r ＝0.5 ならば

$$T = \sqrt{\frac{(n-2)r^2}{1-r^2}} = \sqrt{\frac{(20-2)\times 0.5^2}{1-0.5^2}} = 2.45$$

エクセルで t 分布の確率を計算すると，

有意水準 0.05 で ϕ ＝20－2＝18

＝TINV(確率,自由度)=TINV(0.05,18)=2.101

で 2.101 となります。

t ＝2.45 で 2.101 より大きいので， n ＝20 で相関係数 r ＝0.5 ならば有意に相関があります。

参考；薩摩 順吉：「確率・統計」岩波書店（1989 年）

参考 3　Spearman の順位相関係数（ノンパラメトリック法）

通常相関係数と言う場合 Pearson の相関係数 r です。

先ほど述べたように，Pearson の相関係数 r は 2 変量正規分布で直線性の強さを示しています。

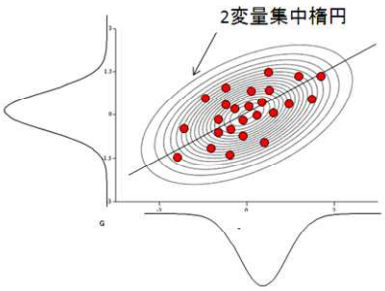
この Pearson の相関係数 r のノンパラメトリック法が Spearman の順位相関係数です。

Spearman の順位相関係数は Pearson の相関係数を順位に変えたもので，2 変量正規分布や直線性を要求しないノンパラメトリック法です。

つまり，分布に偏りがある場合とかで，Pearson の相関係数 r が適用できない時に一般的に使用されます。

濃度分析では 2 変量正規分布でない場合が多く

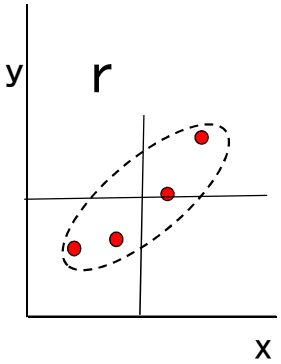
Spearman の順位相関係数が頻繁に使用されます。



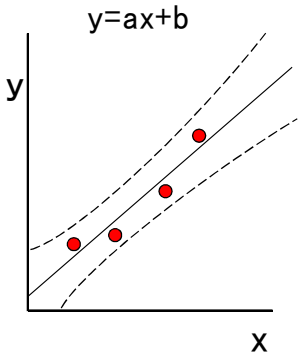
Pearson の相関係数

参考 4　相関分析と回帰分析

相関分析と回帰分析を区別しておきます。



相関分析



回帰分析

相関分析と回帰分析は異なる統計手法です。このことは理解しておくべきことです。

第７章　小標本のランダム化の注意点

統計的検定の厳密性を保障するのには，検定に影響する確率的変動を０にする必要があり，実験や調査のランダム化が行われます。実験する前は，ランダム化は確率の偏りがないように見えます。しかし，事後的にはランダム化は必ず偏りが生じます。

よく知られた例として円周率πは乱数と見なせるが，同じ数字が繰り返される部分が多くあります。大標本法ならランダム化であらゆる方向の偏りがなくなりますが，小標本ではランダム化であらゆる方向に平等に偏ることになります。

少し前の統計の本には必ず乱数サイコロの写真が載せてあり，ランダム化し，人為的な部分を除き，確率の偏りを無くすためにランダム化しなければならないと記載されていますが，小標本でランダム化すると，事後的に必ず偏りが生じます。これはよく経験することで，偏りがひどいと再度ランダム化をしたりします。

例えば下記のようなことが起こります。

均等な割付(サンドイッチ法)[注 1)]

ABBA　BAAB　　ABBA　BAAB ...

ランダムな割付

ABBA　BBAA　BBAA　　BABA　...

化学実験データの多くは小標本ですから，ランダム化は止めて，均等に割り振る方が安全です。実験データではランダム化の前に偏りの無いデータを集めます。このためには，データに影響すると思われる要因を洗い出してその要因の影響を打ち消せる，また把握できるようにすることです。

つまり，「偏り」がないことが大切で，さらに，ランダム化により「客観性」を持たせます。

ランダム化を第一に考えている人がいますが，偏りを生じている場合がありますので注意すべきです。

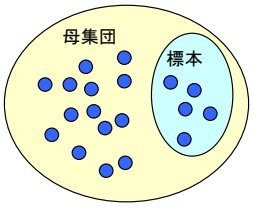
注 1)　竹内 啓：「統計解析におけるランダム化の問題」日本計量生物学会，応用統計学会 1999 年合同年次合同大会

竹内 啓：「偶然とは何か――その積極的意味」岩波新書　2010 年

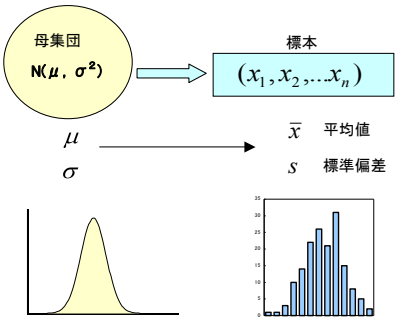
第 8 章　正規分布について

8.1　正規分布

母集団とは，ある集合の全体で，その一部を標本として取り出していると考え，母集団と標本の関係は下記の図のようになります。



一部を測定している



測定値（標本）を増やすと釣鐘状の「**正規分布**」に近似していくことはよく経験することです。濃度測定では多く場合，母集団として誤差は正規分布を仮定します。

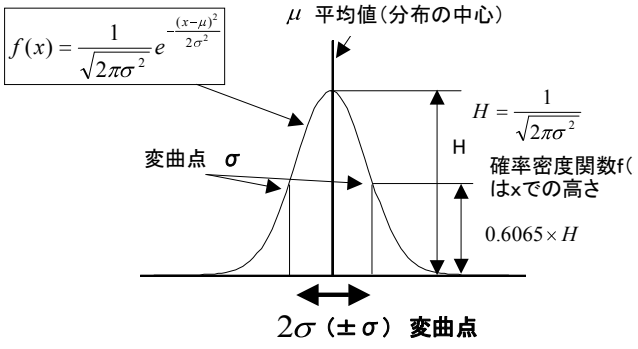
正規分布は「ガウス分布」とも呼ばれますが，ガウスが詳細な検討をしましたが，発見しのはド・モアブルのようです。最近ではガウス分布ではなく正規分布を一般的に使用します。

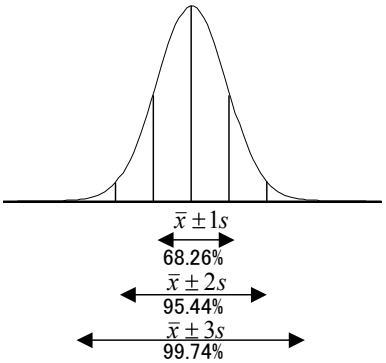
正規分布の**確率密度関数**は

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

です。μは平均，σ²は分散，πは円周率 3.14…で，e はネイピアで 2.718…です。

正規分布では標準偏差の± 2 s.d.で約 95%範囲になります。

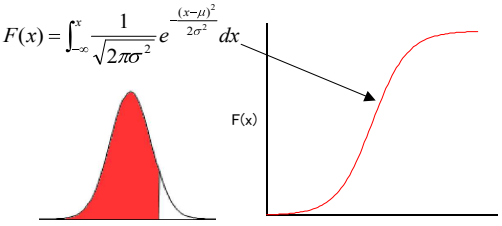




また，確率密度関数を積分したものは**累積分布関数**（分布関数）と呼び，

$$F(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

になります。



正規分布は平均と分散のみで決まるので，平均μ，分散σ^2の場合**N(μ，σ^2)**と記載します。

変数（データ）xは次式の**標準化変換**（**Z変換**）で規格化されて，平均が0，分散が1に従う**標準正規分布N（0，1^2）**になります。

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

このZ変換による値は**Z-スコア**（または**標準得点，ＳＤＩ**）と呼ばれ，

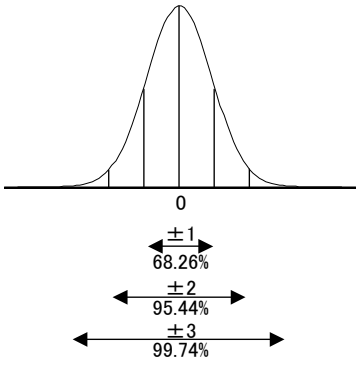
右の図のように

±1 の範囲で 68.26%

±2 の範囲で 95.44%

±3 の範囲で 99.74%

となります。



Z-スコアで測定精度を評価する場合がありますが，無次元化（単位を持たない）されているので，そのデータ数，平均値，標準偏差なども明記しないと意味をなさないことがあります。

確率分布として，積分すれば1になります。[注1)]

$$\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\left(\frac{z^2}{2}\right)}=1$$

標準正規分布N(0，1^2)の密度関数を**ϕ（.）**で表し，その累積率分布関数を**Φ（.）**で表す習慣です。

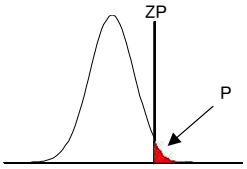
注1）数式処理ソフトがあれば面倒な数式を解かなくても，積分すれば1になることが簡単に確認できます。本文の「付録2　無料パソコンソフトの利用」に数式処理ソフトMaximaの説明があります。

さらに，N(0，1^2)の確率変数ｕに対して，下記の式を満たす

Ｚｐを N(0，1^2)の上限Ｐ％点と呼びます。

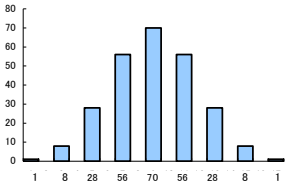
$\Pr(u \geq Zp) = P = 1 - \Phi(Zp)$

の関係があります。



ところで，パスカルの三角形は正規分布に近づきます。少し不思議に感じますが２項係数，２項分布は正規分布に近似していきます。（ド・ モアブル - ラプラスの定理）

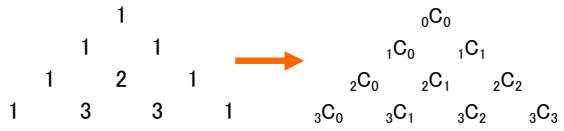
1
1　1
1　2　1
1　3　3　1
1　4　6　4　1
1　5　10　10　5　1
1　6　15　20　15　6　1
1　7　21　35　35　21　7　1
1　8　28　56　70　56　28　8　1



参考 **２項分布**：２項分布は離散分布です。

２項定理は

$$(a+b)^n = \sum_{x=0}^{n} {}_nC_x a^x b^{n-x}$$



です。${}_nC_x$（２項係数）は「ｎ個の中からｘ個を選ぶ組み合わせの数」です。ｎ！は階乗です。

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

例えば，

$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$

となります。この係数は上のパスカルの三角形の４行目と同じです。1，3，3，1となっています。

２項分布Ｂｉ（ｎ，ｐ）：ｎ回の試行で確率がｐで，ｘ回起きる確率は

$p(x) = {}_nC_x p^x (1-p)^{n-x}$

です。

先ほどの２項定理の式と見比べて下さい。

簡単な例を示します。

例：サイコロの１の目の出る確率は1/6であると考えられます。２回投げて，２回とも１である確率は，式に当てはめるならば，（１－ｐ）は１でない確率ですから

$$p(2) = {}_2C_2 \left(\frac{1}{6}\right)^2 \left(1-\frac{1}{6}\right)^0 = 0.028$$

となります。

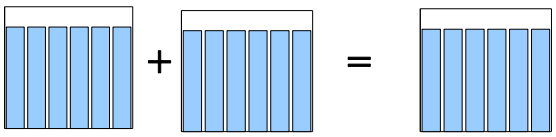
8.2　中心極限定理

さて，次の問題を考えてみて下さい。

一様分布（矩形分布）などもよく誤差の計算で使用します。例えば，サイコロの目の出方は１から６まで同じ確率で出るので一様分布であると考えられます。

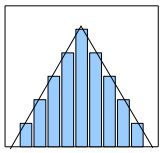
２つの測定誤差が共に一様分布に従う場合，２つの測定知値を加えるとどの様な誤差分布が得られるでしょうか。誤差の範囲は等しいとします。

一様分布と一様分布を合わせると，そのまま一様分布のような気もします。



しかし，実際は，一様分布と一様分布を加えると三角分布になります。

例えば，サイコロの目の出る数は一様分布ですが，２つのサイコロの目の和を考えると，２になるのは１＋１しかありませんが，６になるのは１＋５，２＋４，３＋３，２＋４，５＋１の５通りがあります。全て数え上げて図にすると三角分布になります。



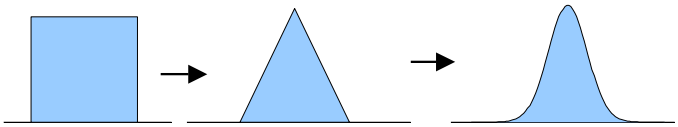
このことはあたりまえのようですが，感覚的には少し変な気もします。

実際に現場の問題として，確率密度関数の合成が必要な場合があります。

２つの正規分布 N（μ_1，σ_1^2）と N（μ_2，σ_2^2）の和は，N（$\mu_1+\mu_2$，$\sigma_1^2+\sigma_2^2$）になります。再度正規分布になるので，「**正規分布の再生の定理**」と呼びます。

また，どのような分布でも，その和 $X_1+X_2+\ldots+X_n$ の確率分布は漸近的に正規分布に従うことは大切なことです。これは「**中心極限定理**」と呼ばれています。[注1）]

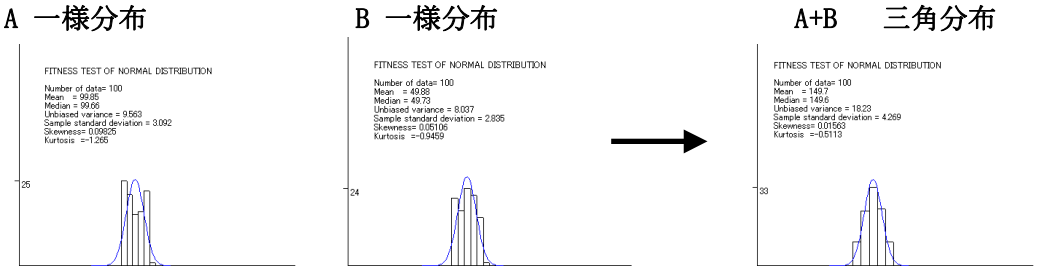
先ほどの一様分布も $Z=X_1+X_2$ が三角分布になり，さらに加えていけば中心極限定理に従い正規分布に近づきます。



注１）中心極限定理は，後で述べる母関数を使用して証明できます。

母関数は本文「第１４章　母関数の魅了」を参照して下さい。

一様分布の乱数ｎ＝100 の A と B を加えたシミュレーションの結果を示します。A+B は三角分布になっています。



8.3　ランダムなものを数式として扱えるようにする正規分布の重要性

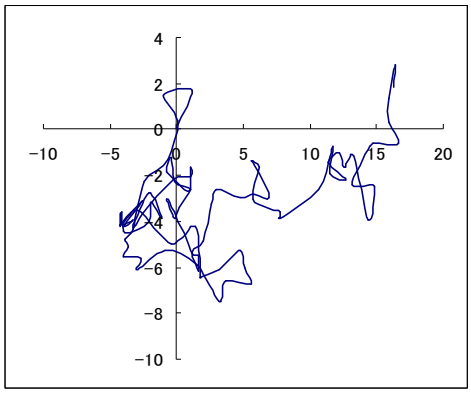
正規分布の式は

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

ですが，濃度測定データの誤差のほとんどが正規分布として仮定でき，濃度測定では最も身近にある分布です。

重要なのは無秩序なものを数式として捕らえることが可能であることです。無秩序から正規分布が現れてきます。

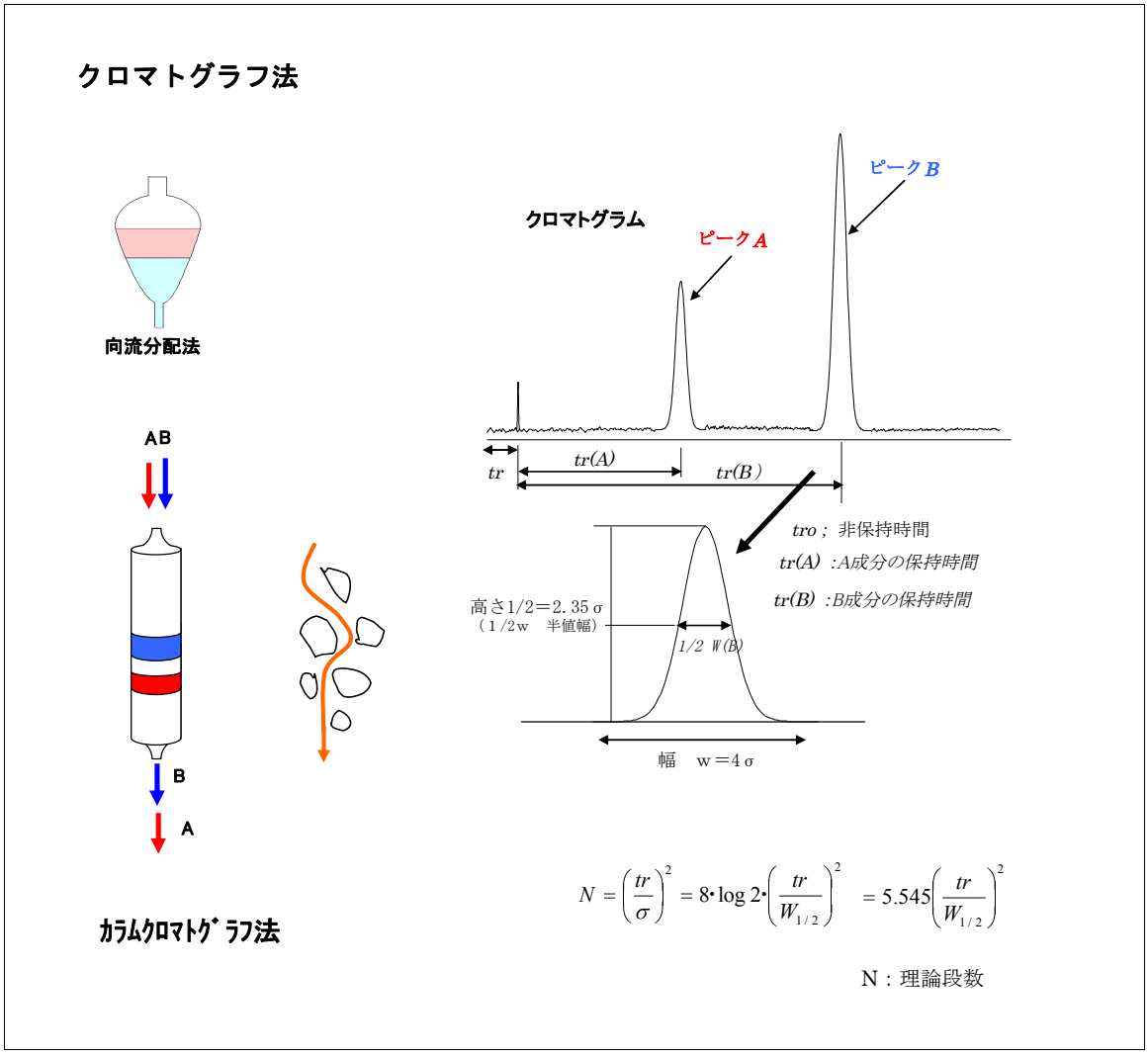
例えば，ランダム・ウォークはランダムですから規則性がありません。しかし，多くのランダム・ウォークのヒストグラムを作成すれば正規分布が現れます。



分離分析のクロマトグラフ法のピークも正規分布と見なすことが可能であることは，ピークの形状からも想像できます。

分液ロートを使用した向流分配法が２項分布になり，正規分布に近似するように，段理論のクロマトのピークを正規分布と考えて，理論段数などが算出できます。段理論では，向流分配を連続的に行っていると見なすことができます。

クロマトグラフ法では，段理論，速度論などから分離を調べますが，段理論，速度論（拡散）共に正規分布が現れます。[注１）]



注１）津田 考雄：「クロマトグフィー　ー分離のしくみと応用ー」丸善（1995 年）

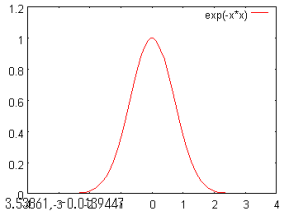
参考１　**正規分布の密度関数**がどのように導き出せるのかを，参考として簡単に紹介します。

ド・モアブルにより発見された関数

$f(x)=e^{-x^2}$

は，パソコンでも簡単に確認できますが，右の図のように

釣鐘状の形になり，誤差分布としての仮定を満たします。



この分布が「確率の公理」を満たすためには，積分して１になる

必要があります。積分して１になるように係数を求めていくと

$$\int_{-\infty}^{\infty} e^{-x^2} = \sqrt{\pi}$$

から，標準正規分布

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{x^2}{2}\right)} = 1$$

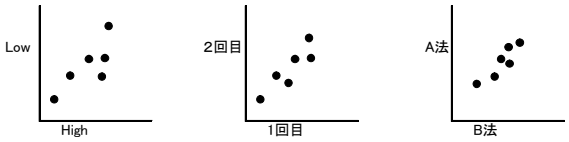
が求まります。つまり，正規分布関数の $\frac{1}{\sqrt{2\pi}}$ などの係数は，積分して１にするためです。

参考：小杉　肇：「ｅの数学」恒星社厚生閣（1986 年）

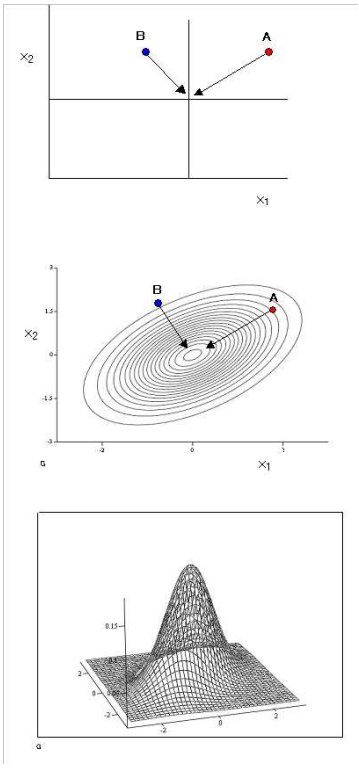
参考２　**２変量正規分布**について

２変量の図はよく使用します。

例えば，臨床検査での新法と旧法の比較などです。また，コントロール検体が２濃度あれば散布図にすることができ，精度管理の解析にも利用できます。精度管理など解析では，２変量正規乱数を発生させてシミュレーションしてみることも有効です。



相関係数と正規分布について述べましたので，２変量正規分布について述べます。



右の上の散布図でＡとＢの中心までの距離は，Ｂの方が近いと思われす。

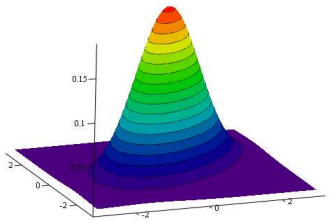
しかし，等確率長円を描くとＡの方が確率では中心に近いことが解ります。

この確率の距離を「**マハラノビスの距離**」と呼びます。

データを基準化し

$$u_1 = (x_1 - \bar{x}_1)/s_1$$
$$u_2 = (x_2 - \bar{x}_2)/s_2$$

s ：標準偏差



とするとマハラノビスの距離 D^2 は

$$D^2 = \frac{u_1^2 - 2\rho u_1 u_2 + u_2^2}{1-\rho^2}$$

ρ ：相関係数

で計算できます。

２変量正規分布は D^2 を含む下記の式です。

$$f(u_1,u_2;\rho) = \frac{1}{2\pi(1-\rho^2)}\exp\left[-\frac{1}{2(1-\rho^2)}(u_1^2 - 2u_1u_2 + u_2^2)\right] = \frac{1}{2\pi(1-\rho^2)}\exp\left(-\frac{D^2}{2}\right)$$

このとき平均０ 分散１の正規分布 N(0,0;1,1;ρ)に従います。

２変量正規分布の分布関数は

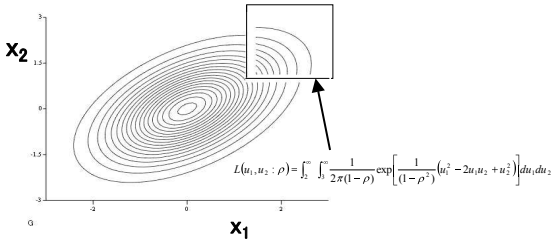
積分なので

$$L(u_1,u_2:\rho) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{2\pi(1-\rho^2)}\exp\left[\frac{1}{(1-\rho^2)}\left(u_1^2 - 2u_1u_2 + u_2^2\right)\right]du_1du_2$$

です。

例えば,

相関係数 ρ =0.8 で u1>2　u2>3 となる確率の計算式は



$$L(u_1,u_2:\rho) = \int_{2}^{\infty}\int_{3}^{\infty}\frac{1}{2\pi(1-\rho^2)}\exp\left[\frac{1}{(1-\rho^2)}\left(u_1^2 - 2u_1u_2 + u_2^2\right)\right]du_1du_2$$

$$= 1.1314\times10^{-3}$$

となります。

つまり，相関係数 ρ =0.8 での u1>2 ， u2>3 の確率は 0.0011314 です。

この計算は２変量の精度管理などに利用されます。

実際の計算は，数学処理ソフトで行えます。本文「付録２　無料パソコンソフトの利用」を参照して下さい。

参考 3　**変動係数 c. v. %とこれまでの説明で間違い易い計算**について

変動係数（c.v.または RSD）で，例えば c.v.%=5 は同時再現性で良好なのか。

$$c.v.\% = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}}{\bar{x}} \times 100 = \frac{s.d.}{\bar{x}} \times 100$$

または，c.v.%が何%なら良好と言えるのかなどの質問をよく受けます。

注意として述べましたが,0 や負の数を含む c.v.%は考えません。

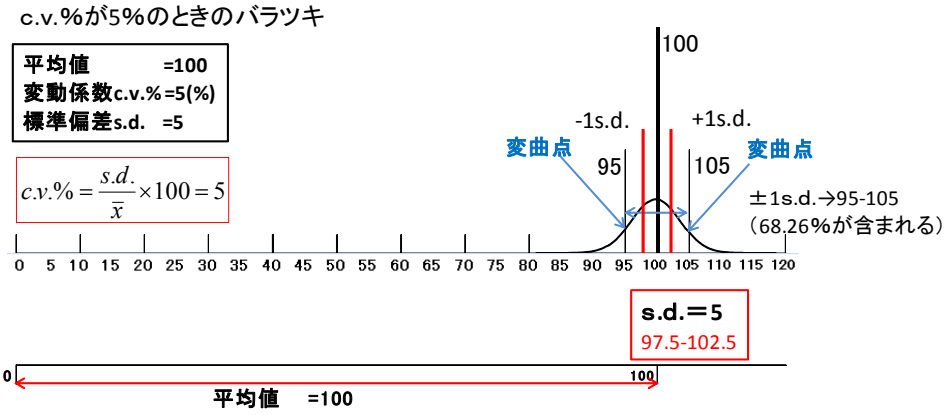
また，通常の濃度測定では，原点付近（低濃度）では分母が小さくなるために，c.v.%が大きくなります。

s.d.は各データ x_i と $\bar{x}$ の差の平均で，c.v.%は s.d.を平均 $\bar{x}$ で割って%にした割合です。

変動係数 c.v.%は特によく使用するので，理解を深めるために簡単な図を付け加えておきます。平均から 1s.d.離れた点は正規分布ならば**変曲点**になっています。

例として，平均値 100 で c.v.%＝5 の時のバラツキを図にすると下記のようになります。

（注）高さなど図は模式図ですが，図にするとバラツキを把握できます。



その他の確認

平均：対数正規分布の場合は幾何平均を使用する。(「第 2 章　平均値の計算方法を考える」で説明しました。

相関係数：相関関係は「強い」「弱い」で表現し，2 変量正規分布を仮定しています。

このため,原点回帰の相関係数は原則使用しない。

ノンパラメトリック法として Spearman の順位相関係数などを使用する。

相関係数がどの程度なら良いのかの質問を受けますが，検査項目，条件，目的などで変わります。

第９章　検査スケジュールの最適化

検査を始める前に各検査員の配置，検査項目，分析機器の測定予定などスケジュールを決める必要があります。スケジュールを立てなければ効率的な検査はできません。

簡単なスケジュールを組む問題を考えてみます。

食品検査で，検査材料はスイカＡ，リンゴＢ，みかんＣ，玄米Ｃがあり，前処理Ｍ１（粉砕，精製など）の後に測定Ｍ２（分析機器）を行い，前処理Ｍ１を済ませないと，測定Ｍ２には移れないとします。下記の作業時間が必要であるとします。

処理時間　　単位：時間

	スイカ A	リンゴ B	みかん C	米 D
前処理　M1	5	1	3	3
測定　　M2	1	4	2	5

何も考えずに，スイカＡ，リンゴＢ，みかんＣ，玄米Ｄの順に処理すると

時間	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
前処理M1	A					B	C			D							
測定　M2						A	B				C		D				

となります。上の**ガントチャート**(Gantt chart)から 17 時間が必要です。

通常スケジュールは**互換操作**を繰り返して，最適と思うスケジュールを組んでいるのではないでしょうか。この例でも，組合わせパターンは 24 種あります。

よく上の図（ガントチャート）を眺めると，前処理（Ｍ１）の短いリンゴ（Ｂ）を先にして，測定時間（Ｍ２）の短いスイカ（Ａ）を最後にして，みかん（Ｃ）と米（Ｄ）も同様に考えて配置すれば良いことに気が付きます。

この考えで配置してみます。

時間	1	2	3	4	5	6	7	8	9	10	11	12	13
前処理M1	B	D			C			A					
測定　M2		B				D					C		A

１７－１３＝４で４時間の短縮ができます。

いま示した作戦は「ジョンソンの定理」と呼ばれるものです。
ジョンソンの定理を先ほどの例で再度説明します。

処理時間　　単位:時間

	スイカ A	リンゴ B	みかん C	米 D
前処理　M1	5	1	3	3
測定　　M2	1	4	2	5

ジョンソンの定理

ステップ１

全体で一番小さいものを探し，Ｍ１にあれば一番はじめに実行し，Ｍ２にあれば最後に実行する。

（Ｍ１のリンゴＢ＝１なので最初に実行し，次にＭ２のスイカＡ＝１が見つかり最後に行う。）

ステップ２

選択したものは除く。

（リンゴとスイカを除く。）

ステップ３

残りのものでステップ１を行う。残っていない場合は作業を終える。

（みかんＣと米Ｄでは，Ｍ２のみかんＣ＝２が小さいのいで，みかんＣを後に行う。）

時間	1	2	3	4	5	6	7	8	9	10	11	12	13
前処理M1	B	D			C			A					
測定　M2		B				D					C		A

ここで述べたものは**２機械フローショップ問題（ジョンソンの問題）**と呼ばれるものです。ＡからＤの４つでも 24 種ものパターンがあり，全てのパターンでの作業時間を計算するのは面倒です。このため，ジョンソンの定理は優れています。

日常の検査では，その日のスケジュールを仕事の前に決める必要があり，大きなプロジェクトの日程計画や，新規項目の検討，立ち上げなどでもスケジュールを決める必要があります。

実際は，ここで示したような単純な問題ではないかも知れません。

しかし，「**最適スケジューリング**」の方法を理解している方が，効率的なスケジュールを組むのに有利です。

「オペレーション・リサーチ」「組合わせ理論」「離散数学」は企業経営，生産管理，プロジェクト，ネットワーク計画などに幅広く関係し，有効な戦略を得るための手法として大切です。

参考 **最適化スケジューリング**について

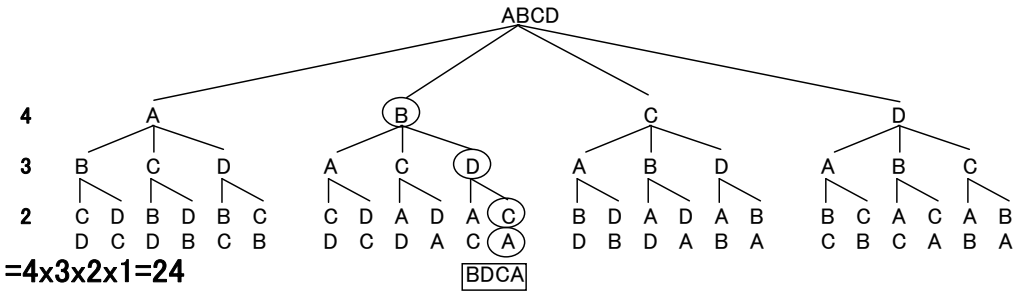
効率的な仕事や最適な在庫管理などは「オペレーション・リサーチ」や「離散数学」の分野です。
「組合せ最適化」,「グラフ理論」,「ネットワークフロー最適化」,「需要予測」,「線形計画問題」,有名な「4 色問題」,「巡回セールスマン問題」「ＮＰ困難」などにつながって行きます。
簡単な例で「最適スケジューリング」を紹介しました。
また，営業で幾つかの会社を廻る場合，効率の良い廻り方を考えるのも同じような問題です。
仕事の効率化,在庫管理などを行うためには，「オペレーション・リサーチ」や「離散数学」を知り利用すのが効果的です。
仕事の「割当て問題」も含めて，順列，組み合わせを考えているので，先ほどの「最適スケジューリング」の問題は，n 個のものからｎ個の順列を作り，最小となるものを選び出す問題となります。
n 個のものから n 個を取った順列は

$${}_n p_n = n! = n(n-1)(n-2)...3 \cdot 2 \cdot 1$$

です。先ほどのＡからＤの４個の問題では ABCD,ABDC,ACBD,....となり

$$n! = 4 \times 3 \times 2 \times 1 = 24$$

24 個のパターンがあるはずです。24 個すべてのパターンの計算値を求めれば正確に最適な順列が求まります。（**完全数え上げ法**）
少ない数での順列を，間違いなく作成するには「**木**」の作成が便利です。



最適解

n	n!
2	2
3	6
4	24
5	120
6	720
7	5040
8	40320
9	362880
10	3628800
11	39916800
12	4.79E+08

しかし，もしも２０個の問題なら

$$n! = 20 \times 19 \times ... \times 4 \times 3 \times 2 \times 1 = 2.43^{18}$$

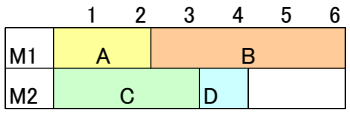
で，１秒に１パターンを計算すると 77146816596 年かかることになり，高々２０個の割り当てでも大変な時間が必要です。
ｎ！はｎが少し増えると急激に増加します。
このため「**分枝限定法**」などの手法が考案されています。

第 9 章　検査スケジュールの最適化

スケジューリングについて，さらに簡単な問題を少し考えてみます。

2 人（M1，M2）で順番は関係なくＡからＤの作業が下記の表のように与えられたる場合，2 人で行う最適なスケジューリングを考えてみます。M1，M2 の 2 人の能力差は無いとします。

A	2
B	4
C	3
D	1



順番の制約がないので組合わせの問題になります。

何も考えずに並べると，M2 に 2 時間の空白があります。

４つから２つを取る組合わせなので，

$$ {}_4C_2 = \frac{4!}{2!(4-2)!} = 6 $$

で 6 パターンですが，M1 とM2 は区別がないので 6/2＝3 で 3 パターンが考えられます。

まず，2 人の分担は作業時間の長いものを先にして，短いもので調整するのが楽だと考えられます。

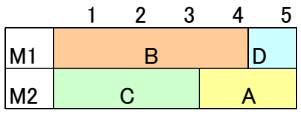
さらに，全体の作業時間を考えます。

$$ 2+4+3+1=10 $$

作業時間は 10 時間で 2 名なので 10/2＝5　から，可能ならば各 5 時間にするのが最適なスケジュールとなるはずです。

①　作業時間の長いものを先に入れる。

②　全体の時間を計算して，人数で割り，均等な時間になるように組合わせる。



同様な問題では，さらに人数，仕事の種類が増えた場合も同じ様にスケジューリングします。

「オペレーション・リサーチ」「組合わせ理論」「離散数学」については多くの本が出版されています。

仙波　一郎：「組合わせ数学」　コロナ社（1999 年）

斉藤，西澤，千葉：「離散数学」　朝倉書店（1998 年）

中川，三道：「オペレーションズ・リサーチ」日刊工業新聞社（2005 年）　など

読んで少し楽しい本も紹介しておきます。

野崎　昭弘：『離散数学「数え上げ理論」：「おみやげの配り方」から「Ｎクイーン問題」まで』

ブルーバックス　（2008 年）

参考　**九去法について**

整数論の本を読んでいて，九去法を知り，面白かった手品を紹介します。

（読み飛ばしたてもかまいません。）この手品の証明は整数論の本を読めば自分で簡単に出来ます。

現象

１）演者は予言の数を紙に書いて伏せて置く。

２）計算方法を簡単に説明する。

３）好きな３桁の数字を書いてもらう　　（例　１３５　）

４）３桁の数字を自由に並び変える　　（　　５１３　）

５）大きい方から小さい方の数を引く　　（　５１３－１３５＝３７８　）

６）１桁になるまで各桁を足す　　（３＋７＋８　＝　１８　　１＋８＝９　　）

７）「自由に書いた数字だから３）から６）の計算で３とか５でもよいのですが」偶然９になったことを述べる。

８）予言の紙を見ると９と書いてある。

説明

手品の本では種として，「必ず９になる」としか書かれていません。

不思議に思ったのは，何桁でも好きな数を書いて，順番を入れ替えて，大きき方から小さい方を引き，各桁を加えて行くと，必ず９になること自体です。なんで必ず９になるのだろうか？

合同式を使用すと簡単に説明できます。

不思議なことは，どんな数でも各桁を足すのを繰り返し１桁にした数は，９で割った余りに一致します。

１）好きな数を書いて，各桁を足して１桁にすると９の余りになる。

２）数の順番を入れ替えて，大きい方から小さい方を引き，その各桁の数を足して１桁にすると必ず９になる。

１）と２）について説明します。

算数でも説明できますが，合同式を使用すと簡単に説明できます。

　合同式を参考に示すと

　自然数ｎが整数 ab に対して，a-b がｎの倍数であるとき a と b はｎを法として互いに合同であるといいます。

a≡b（mod n）　　mod は modulus

９を法としたとき

１０≡１　（mod ９）

あたりまえで，１０－１は９です。また，１０を９で割った余りは１で，１を９で割った余りも１で，同じである。

ここである数を１０進数でＡとして

A＝an...a1a0　　０≦ai＜10

Ａ＝１０^{n}　an＋...１０a1＋a0

これも当然です。２２０は　２×１００＋２×１０　と同じなのですから

各桁の数の和をＳ（A)とするなら

Ｓ（Ａ）＝an＋..＋.a1＋a0

です。

先ほどの

１０≡１　（mod ９）

をｋ乗すると

１０^{k}＝１^{k}　（mod ９）

１０^{k}＝１　　（mod 9)

となります。

両辺をak倍してｋ＝0,1,,,,nについて足すと

Ａ≡Ｓ（A)　（mod ９）

つまり

１０^{n}　an＋...１０a1＋a0≡１０^{n}　an＋...１０a1＋a0　（mod ９）

で９で割った余りに一致します。

これで１）の疑問であった，各桁の和は９の余りになることが説明できました。

次に，２）の順番を入れ替えた数を引くとどうなるのかを考えてみます。

１桁にした時は９の余りになる，数字の順番を替えても，加える順番がかわるだけで

和は同じですから，余りは同じです。

数Ａの順番を入れ替えた数をＢとすると　ＡとＢの各桁を足した和は同じですから

Ｓ（Ａ）＝Ｓ（Ｂ）

から

Ｓ（Ａ）－Ｓ（Ｂ）＝０

９≡０　（ mod ９）

必ず９に成らざる得ません。

もしも，順序を入れ替えなければ０です。

（５４６－５４６＝０）

入れ替えると９になるます。

（５４６－４５６＝９０　９０は９で割りきれます。

９＋０＝９　９は９で割りきれ，余りは０です。）

第Ⅱ部　統計処理

適切なデータを得ることは大切ですが，正しいデータ解析を行わないと判断を誤る危険性があります。
このため，分析現場ではデータ解析は重要です。

第１０章　誤差伝播の法則－不確かさの計算－

「不確かさ」の国際ルールとして GUM(Guide to the expression of uncertainty in measurement)が出されています。GUM で示された「不確かさ」の計算の元となる「誤差伝播の法則」について説明します。

GUM では「不確かさの伝播則」となっていて，誤差と不確かさを明確に区別していますが，ここでは誤差の伝播則として説明します。不確かさと読み替えても計算式は成り立ちます。分析データは誤差を伴うものですから，分析技術者として，ここで述べる「誤差の伝播則」は理解すべき事柄です。

10.1　測定値が和，差，積，商の場合を考える

１）和と差の場合を考える

測定値が和になる場合をはじめに考えます。

A と B の誤差として，標準偏差 σ_a と σ_b とします。

Z＝A＋B で各平均を m_z ，m_a，m_b とすると，m_z の期待値は $m_z = m_a + m_b$ ですが，

誤差を考慮すると，

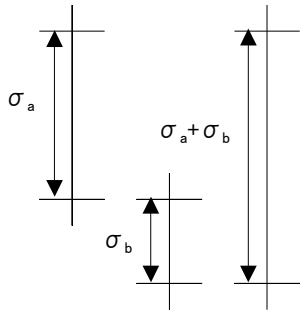
最小値 $Z_{min} = m_z - (\sigma_a + \sigma_b)$，最大値 $Z_{max} = m_z + (\sigma_a + \sigma_b)$

と考えられます。

誤差の最大は右の図からも $\sigma_a + \sigma_b$ で，

$\sigma_z \approx \sigma_a + \sigma_b$

となります。



Z＝A－B の差の場合も，誤差は加えればよいのではないかと考えられます。

明確にイメージするために，簡単な例で考えてみます。

誤差として標準偏差を使用するのならば，

A の平均値 $\bar{X}_a = 10$　標準偏差 $s_a = 2$ なら

A の測定値＝$\bar{X}_a \pm s_a = 10 \pm 2$

B の平均値 $\bar{X}_b = 20$　標準偏差 $s_b = 5$ なら

B の測定値＝$\bar{X}_b \pm s_b = 20 \pm 5$

と考えられます。

Z＝A＋B は

$\bar{X}_a+\bar{X}_b\pm$（s_a+s_b）

＝10＋20±（2＋5）

＝30±7

となります。

実際は，Z＝A＋B と Z＝A－B の場合の，

$\sigma_z \approx \sigma_a+\sigma_b$

では大きく見積もりすぎていて，「8.2　中心極限定理」で述べ，この後の「誤差の伝播則」や「第 14 章　母関数の魅了」で示すモーメント母関数からも

$$\sigma_z=\sqrt{\sigma_a^2+\sigma_b^2}$$

が良いことが解ります。[注 1)]

直角三角形になっていて，ピタゴラスの定理です。

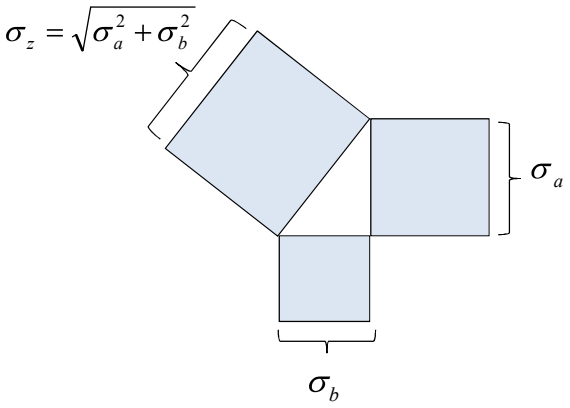
先ほどの例では，Z の測定値は

$\bar{x}_a+\bar{x}_b\pm\sqrt{s_a^2+s_b^2}$

$=10+20\pm\sqrt{2^2+5^2}$

$=30\pm5.4$

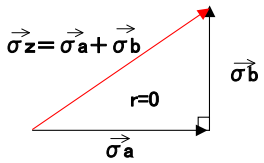
となります。



注 1）本文「第６章　相関係数の２乗とは何か」の参考の相関係数のところで，標準偏差はベクトルの大きさで，相関係数が 0 とはベクトルが直交することを示しました。

Z＝A+B で，A とＢの相関ｒ＝0 とは直交することで，右下の図からも，AとＢの標準偏差σａのベクトルとσｂのベクトルを加えたものがσ ｚのベクトルになっています。つまり，A と B に相関がなければ，標準偏差がベクトルの大きさでしたので，下記の式になります。

$\sigma_z=\sqrt{\sigma_a^2+\sigma_b^2}$



相関がある場合の計算式は，この章の最後のに参考として示しました。

２）積と商の場合を考える

Z＝A×B を考えますが，

平均はmz, ma，mb で，相対標準偏差 RSD（RSD と変動係数 c.v.は同じです・）

$$RSD_z = c.v._z = \frac{\sigma_z}{m_z}$$

を使用します。

測定値は相対標準偏差を使用すると，A の場合は$m_a\left(1 \pm \frac{\sigma_a}{m_a}\right)$と考えられます。

このことから Z＝A×B の最大値は

$$m_a m_b\left(1+\frac{\sigma_a}{m_a}\right)\left(1+\frac{\sigma_b}{m_b}\right) = m_a m_b\left(1+\frac{\sigma_a}{m_a}+\frac{\sigma_b}{m_b}+\frac{\sigma_a}{m_a}\frac{\sigma_b}{m_b}\right)$$

となります。

積の$\frac{\sigma_a}{m_a}\times\frac{\sigma_b}{m_b}$は一般に小さくなるので，省略します。

$$m_a m_b\left(1+\frac{\sigma_a}{m_a}+\frac{\sigma_b}{m_b}+\frac{\sigma_a}{m_a}\frac{\sigma_b}{m_b}\right) \approx m_a m_b\left(1+\frac{\sigma_a}{m_a}+\frac{\sigma_b}{m_b}\right)$$

となります。負の場合も同様に計算し，±で示すと，

Z の測定値＝$m_a m_b\left(1 \pm \left(\frac{\sigma_a}{m_a}+\frac{\sigma_b}{m_b}\right)\right)$

となります。

Z の測定値＝$m_z\left(1 \pm \frac{\sigma_z}{m_z}\right)$

と比べて，Z の期待値は$m_z = m_a m_b$になり，誤差の計算は

$$\frac{\sigma_z}{m_z} \approx \frac{\sigma_a}{m_a}+\frac{\sigma_b}{m_b}$$

が求まります。

Z＝A×B の誤差は相対標準偏差 RSD の和として求められることが解ります。

和と差のときと同様に，大きく見積もっているので，

$$\left(\frac{\sigma_z}{m_z}\right)^2 = \left(\frac{\sigma_a}{m_a}\right)^2 + \left(\frac{\sigma_b}{m_b}\right)^2$$

$$c.v._z^2 = c.v._a^2 + c.v._b^2$$

となります。

積の場合を，簡単な例で計算してみます。

Z＝A×B の

誤差として標準偏差を使用するのならば，

A の平均値$\bar{X}_a$＝10　標準偏差 s_a=2 なら

A の RSD＝$\frac{2}{10} = 0.2$

B の平均値$\bar{X}_b$＝20　標準偏差 s_b=5 なら

B の RSD＝$\frac{5}{20} = 0.25$

から，

Z の測定値＝

$$m_a m_b\left(1 \pm \sqrt{\left(\frac{\sigma_a}{m_a}\right)^2 + \left(\frac{\sigma_b}{m_b}\right)^2}\right) = 10 \times 20 \times \left(1 \pm \sqrt{\left(\frac{2}{10}\right)^2 + \left(\frac{5}{20}\right)^2}\right)$$

$$= 200(1 \pm 0.32)$$

$$= 200 \pm 64$$

となります。

商の場合 Z＝$\frac{A}{B}$ は少し面倒です。

テイラー展開から近似的に[注1)]

$$\frac{1+\sigma_a}{1-\sigma_b} \approx 1+\sigma_a+\sigma_b$$

が導けます。このことから，最大 Z_{max} と最小 Z_{min} は

$$Z_{max}\text{の測定値} = \frac{m_a}{m_b}\left(1+\frac{\sigma_a}{m_a}\right)\Big/\left(1-\frac{\sigma_b}{m_b}\right) \approx \frac{m_a}{m_b}\left(1+\frac{\sigma_a}{m_a}+\frac{\sigma_b}{m_b}\right)$$

$$Z_{min}\text{の測定値} = \frac{m_a}{m_b}\left(1-\frac{\sigma_a}{m_a}\right)\Big/\left(1+\frac{\sigma_b}{m_b}\right) \approx \frac{m_a}{m_b}\left(1-\frac{\sigma_a}{m_a}-\frac{\sigma_b}{m_b}\right)$$

となります。

参考注 1）**テイラー展開**について

微分可能な関数 f （x）を微分して

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots$$

と有限項まで展開するとき，テイラー展開と呼びます。このときの係数は

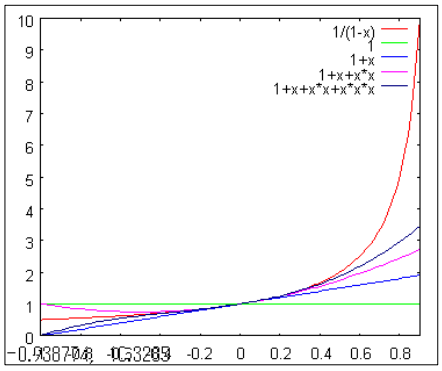
$$a_n = \frac{f^{(n)}(x)}{n!}.$$ です。

f $^{(n)}$ は n 階の微分で，n ！は階乗です。

右の図でテイラー展開について説明すると $|x|<1$ で，

$f(x) = \frac{1}{1-x}$ ならテイラー展開で，$1+x$，$1+x+x^2$，$1+x+x^2+x^3$

となり，$f(x) = \frac{1}{1-x}$ に近似していくことが図からも理解できます。



1 階の微分までで

$\frac{1}{1-x} \approx 1+x$ で，$\frac{1}{1+x} \approx 1-x$ も導けます。x を σ に変えて，

テイラー展開　$\frac{1}{1-x} = 1+x+x^2+x^3+\ldots$

$x<1$

$\frac{1}{1-\sigma_a} \approx 1+\sigma_a$ と $\frac{1}{1+\sigma_a} \approx 1-\sigma_a$ になります。このことから

$\frac{1+\sigma_a}{1-\sigma_b} \approx (1+\sigma_a)(1+\sigma_b) = 1+\sigma_a+\sigma_b+\sigma_a\sigma_b \approx 1+\sigma_a+\sigma_b$ になります。

つまり,

Z の測定値＝$\frac{m_a}{m_b}\left(1\pm\left(\frac{\sigma_a}{m_a}+\frac{\sigma_b}{m_b}\right)\right)$

になります。このことと

Z の測定値＝$m_z\left(1\pm\frac{\sigma_z}{m_z}\right)$

を比較して

$\frac{\sigma_z}{m_z}\approx\frac{\sigma_a}{m_a}+\frac{\sigma_b}{m_{b\ b}}$

であることを示すことができます。これも大きく見積もっているので

$$\left(\frac{\sigma_z}{m_z}\right)^2=\left(\frac{\sigma_a}{m_a}\right)^2+\left(\frac{\sigma_b}{m_b}\right)^2$$

$$c.v._z^2=c.v._a^2+c.v._b^2$$

となります。c.v.(RSD)の 2 乗を加えればよいことになります。

商の場合の簡単な例も示します。

$$Z=\frac{A}{B}$$

誤差として標準偏差を使用するのならば,

A の平均値$\bar{X}_a$＝10　標準偏差 s_a=2 なら

A の RSD＝$\frac{2}{10}=0.2$

B の平均値$\bar{X}_b$＝20　標準偏差 s_b=5 なら

B の RSD＝$\frac{5}{20}=0.25$

から,

Z の測定値＝

$$\frac{m_a}{m_b}\left(1\pm\sqrt{\left(\frac{\sigma_a}{m_a}\right)^2+\left(\frac{\sigma_b}{m_b}\right)^2}\right)=\frac{10}{20}\times\left(1\pm\sqrt{\left(\frac{2}{10}\right)^2+\left(\frac{5}{20}\right)^2}\right)$$

$$=0.5(1\pm0.32)$$

$$=0.5\pm0.16$$

となります。

10.2　シミュレーションでの計算方法の確認

Z＝A±B の場合は分散から

$$\sigma_z = \sqrt{\sigma_a^2 + \sigma_b^2}$$

で計算します。

Z＝A×B または Z＝$\frac{A}{B}$ の場合は，相対標準偏差 RSD から

$$RSD_z = \sqrt{RSD_a^2 + RSD_b^2}$$

で計算します。

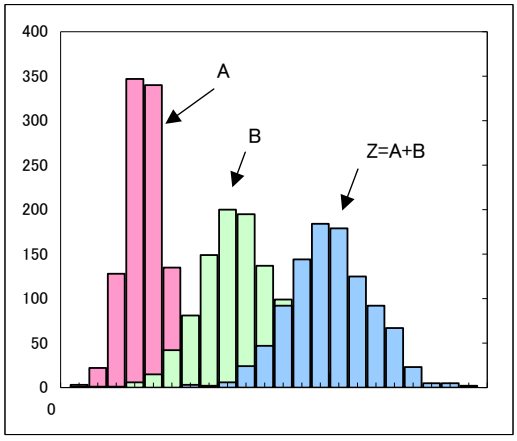
Z＝A＋B＋…＋Y や Z＝A×B×…×Y は上記と同様に計算します。

ここで，和と差の場合と積と商の場合で計算が異なることに注意して下さい。

この法則が実際に成り立つのか，エクセルの正規乱数でシミュレーションしてみます。

正規分布に従う A の測定値が N(10,2^2)と，B の測定値が N(20,4^2)に従う A＋B や A/B などの誤差を考えてみます。N（10,2^2）とは平均 10 で分散 4 の正規分布に従う測定値のことです。

各正規乱数 1000 個を発生させてヒストグラムを作成しました。



Z＝A＋B のヒストグラム

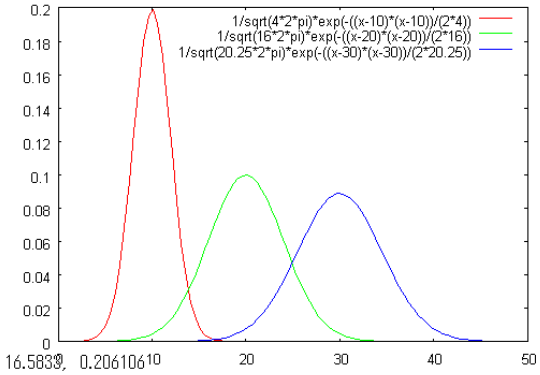
A，B 各 n＝1000 の正規乱数のシミュレーション結果

シミュレーション	A	B	A+B	A×B	A/B
平均	10	20	30	201	0.53
S.D.	2.0	4.1	4.5	57.7	0.19
RSD(C.V.)	0.20	0.20	0.15	0.29	0.36

A＋B の RSD が 0.15 と加えることによって低くなっていることは注意して下さい。

これは A＋B で平均が 30 になることにより見かけ上 RSD が低くなりますが，s.d.は増えています。

実際に先ほどの式で上手く計算できているのか調べてみます。



設定値から先ほどの式で算出　　A，B，A＋B の分布

計算値	A(設定値)	B(設定値)	A+B	A×B	A/B
平均	10	20	30	200	0.5
S.D.	2.0	4.0	4.5	56.6	0.14
RSD(C.V.)	0.20	0.20	0.15	0.28	0.28

A，B 各 n＝1000 での正規乱数のシミュレーション結果

シミュレーション	A	B	A+B	A×B	A/B
平均	10	20	30	201	0.53
S.D.	2.0	4.1	4.5	57.7	0.19
RSD(C.V.)	0.20	0.20	0.15	0.29	0.36

A＋B の誤差は計算値が一致しています。

A×B は相対標準偏差で使用可能な範囲で一致しています。

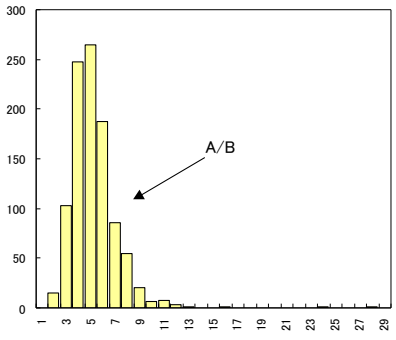
しかし，A／B が正規乱数のシミュレーションでは RSD＝0.36

$RSD_z^2 = RSD_A^2 + RSD_B^2$

では RSD＝0.28 と少し離れています。

A／B がどのような分布になっているのかシミュレーションのヒストグラムを調べてみます。

右の図になり，左に傾いていて，飛び離れた値もあり，正規分布でないことは解ります。このことは後で述べます。



概ね，

$\sigma_z = \sqrt{\sigma_a^2 + \sigma_b^2}$ と $RSD_z = \sqrt{RSD_a^2 + RSD_b^2}$ が成り立つことは確認できます。

ここで，これまでの計算方法を**まとめ**ておきます。

測定値 Z が A と B の和，または差である場合

Z＝A±B

の合成誤差は A と B の分散を σ_a^2 と σ_b^2 として

$$\sigma_z = \sqrt{\sigma_a^2 + \sigma_b^2}$$

で計算します。

測定値 Z が A と B の積，商の場合

Z＝A×B　　$Z = \frac{A}{B}$

は相対標準偏差 RSD(c.v.)を使用して

$$RSD_z = \sqrt{RSD_a^2 + RSD_b^2}$$

で計算します。

和，差，積，商が混ざる時は遂次的に計算していきます。

10.3　誤差の伝播則

「誤差の伝播則」について述べます。「誤差の伝播則」はテイラー展開から導けます。

誤差の伝播則を下記に記載します。

測定量をA，B，…，Yとして
Z＝ｆ（A，B…Y）
であるとすると
A，B，…，Yの平均をma，ｍｂ…myの近傍でテイラー展開され，その１次の項で近似されるなら，Zの分散σ$_z^2$は分布の形に関わらず

$$\sigma_z^2 = \left(\frac{\partial f}{\partial a}\right)^2 \sigma_a^2 + \left(\frac{\partial f}{\partial b}\right)^2 \sigma_b^2 + \ldots + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2$$

で近似することができる。$\left(\frac{\partial f}{\partial a}\right)^2 \sigma_a^2$はA＝maにおける微分係数を表す。

これまで述べた式と同じ式が得られますが，関数を偏微分するだけで誤差を推定できることを示します。
この誤差の伝播則でZ＝A＋B，Z＝A×B，Z＝A／Bを再度計算してみます。

Z＝A＋B

ｆ（Z）として，Zの平均をｍｚとすると測定値$f(z)=a+b$は偏微分すると
$f(z)$をaで微分して

$$\frac{\partial f(z)}{\partial a} = 1$$

となり，
ｂで微分して

$$\frac{\partial f(z)}{\partial b} = 1$$

となりますので，誤差の伝播則から

$$\sigma_z^2 = \sigma_a^2 + \sigma_b^2$$

$$\sigma_z = \sqrt{\sigma_a^2 + \sigma_b^2}$$

となります。
以下同様に計算します。

Z＝A×B

$$f(z)=a\times b$$

$$\frac{\partial f(z)}{\partial a}=b$$

$$\frac{\partial f(z)}{\partial b}=a$$

$$\sigma_z^{\,2}=b^2\sigma_a^{\,2}+a^2\sigma_b^{\,2}=(ab)^2\left(\frac{\sigma_a^{\,2}}{a^2}+\frac{\sigma_{\mathrm{b}}^{\,2}}{\mathrm{b}^2}\right)$$

$$\frac{\sigma_z^{\,2}}{(ab)^2}=\frac{\sigma_a^{\,2}}{a^2}+\frac{\sigma_{\mathrm{b}}^{\,2}}{\mathrm{b}^2}$$

$$\frac{\sigma_z}{m_z}=\sqrt{\frac{\sigma_a^{\,2}}{m_a^{\,2}}+\frac{\sigma_{\mathrm{b}}^{\,2}}{m_{\mathrm{b}}^{\,2}}}$$

Z＝A／B

$$f(z)=\frac{a}{b}$$

$$\frac{\partial f(z)}{\partial a}=\frac{1}{b}=\left(\frac{a}{b}\right)\frac{1}{a}$$

$$\frac{\partial f(z)}{\partial b}=-\frac{a}{b^2}=-\left(\frac{a}{b}\right)\frac{1}{b}$$

$$\sigma_z^2=\left(\left(\frac{a}{b}\right)\frac{1}{a}\right)^2\sigma_a^2+\left(-\left(\frac{a}{b}\right)\frac{1}{b}\right)^2\sigma_b^2=\left(\frac{a}{b}\right)\left(\frac{\sigma_a^2}{a^2}+\frac{\sigma_{\mathrm{b}}^2}{\mathrm{b}^2}\right)$$

$$\frac{\sigma_z^2}{\left(\frac{a}{b}\right)}=\frac{\sigma_a^2}{a^2}+\frac{\sigma_{\mathrm{b}}^2}{\mathrm{b}^2}$$

$$\frac{\sigma_z}{m_z}=\sqrt{\frac{\sigma_a^2}{m_a^{\,2}}+\frac{\sigma_{\mathrm{b}}^2}{m_{\mathrm{b}}^{\,2}}}$$

となります。

このことより，Z＝A＋B，Z＝A－B は分散を加え，Z＝A×B，Z＝$\frac{A}{B}$ は相対標準偏差を加えることにより求めることができます。

本当は，正規分布でもZ＝A×Bはベータ分布となり，$Z=\frac{A}{B}$はコーシー分布（平均，分散が存在しない分布）と呼ばれる分布になりますから，シミュレーションした例では，平均値，分散の正確な値を推定できていません。

特に商の場合は，よい推定値が得られない場合がありますが，通常はシミュレーションで示した例よりも標準偏差は十分小さいく，平均値も 0 から離れていると思われるので，推定値として問題のないレベルになります。

10.4　実際の計算例

それでは，濃度計算でよく見かける式の誤差を計算してみます。

$$Z = F \times A \times \frac{1}{B}$$

Aは秤量値で10gの標準偏差 s_a＝0.01，RSD＝0.1

Bは100mLのフラスコで標準偏差 s_b＝0.1，RSD＝0.1

F＝1000は単位変換係数でgをmgにする。

Z＝$100_{mg/mL}$が得られるとします。この測定誤差を計算してみます。

a，b で偏微分して

$$\frac{\partial f(z)}{\partial a} = \frac{F}{b} = \frac{1000}{100} = 10$$

$$\frac{\partial f(z)}{\partial b} = F\left(-\frac{a}{b^2}\right) = 1000\left(\frac{-10}{100^2}\right) = -1$$

となり，誤差の伝播則の式に代入すると

$$s_Z = \sqrt{10^2 \times 0.01^2 + (-1)^2 \times 0.1^2} = 0.14$$

になります。

また簡単に式は積と商の形なのでRSD(またはc.v.)を使用して

$$RSD_Z = \sqrt{0.1^2 + 0.1^2} = 0.14$$

となります。

さらに，**包含係数**を２とすると，[注1)]

Zの測定値＝100±２×0.14

となります。

注 1）**包含係数**は，拡張不確かさを求めるために合成標準不確かさに乗じる数として用いられる数値係数と定義されていますが，２s.d.の２と同じ意味です。（JIS Z8103）

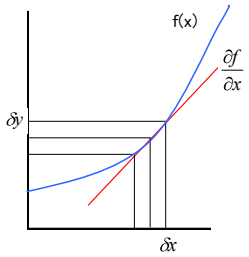
参考１　**「誤差の伝播則」**がどのように導き出せるのかを紹介します。

誤差の伝播則はテイラー展開から誘導できます。

１変量の場合は，誤差はｘ$_0$の期待値での微分から

$$\delta y = \frac{\partial f}{\partial x}\delta x$$

となります。



このことは右の図からも理解できます。

関数ｆ（x）のｘ$_0$での微分は傾きになり，ｘの誤差δｘは，ｙではδｙとなります。このことを多変数関数に拡張すれば誤差の伝播則が得られると想像できます。

テイラー展開から微分可能な関数ｆ（ｘ$_0$+δ$_x$）を微分して

$$f(x_0+\delta x) = f(x_0) + \frac{f^1(x_0)}{1!}\delta x + \frac{f^2(x_0)}{2!}\delta x^2 + \ldots + \frac{f^n(x_0)}{n!}\delta x^n$$

で１次で打ち切ると

$$f(x_0+\delta x) \approx f(x_0) + f^1(x_0)\delta x$$

となります。

２変数のｚ＝ｆ（a，b）で，期待値 E[a]＝μ_aのまわりの１次テイラー展開をおこなうと，ｚのμ_zのまわりの微小偏差を，ａのμ_aのまわりの微小偏差で表すと，

$$z-\mu_z = \frac{\partial f}{\partial a}(a-\mu_a) + \frac{\partial f}{\partial b}(b-\mu_b)$$

となります。高次項は全て無視すると，その平方は

$$(z-\mu_z)^2 = \left[\frac{\partial f}{\partial a}(a-\mu_a) + \frac{\partial f}{\partial b}(b-\mu_b)\right]^2$$

で（ｚ－μ_z）^{2}の期待値はｚの分散で，σ_zで

$$\sigma_z^2 = \left(\frac{\partial f}{\partial a}\right)^2\sigma_a^2 + \left(\frac{\partial f}{\partial b}\right)^2\sigma_b^2 + 2\frac{\partial f\partial f}{\partial a\partial b}\sigma_{ab}^2$$

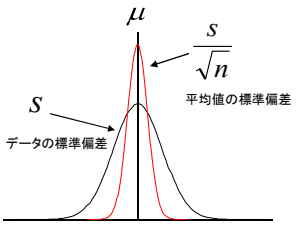
となります。ａとｂに相関がなければ（σ_{ab}^2は共分散で，相関が無ければ$\sigma_{ab}^2=0$です。）

$$\sigma_z^2 = \left(\frac{\partial f}{\partial a}\right)^2\sigma_a^2 + \left(\frac{\partial f}{\partial b}\right)^2\sigma_b^2$$

となります。

参考：今井 秀孝（訳）・飯塚 幸三（監修）：「ISO 国際文章　計測における不確かさの表現のガイド」：
日本規格協会

参考２　**標準誤差**：平均値の標準偏差



誤差の伝播則を利用して，平均値の標準偏差を簡単に導くことができます。

この標準誤差は

$$S.E.=\frac{s}{\sqrt{n}}$$

と，計算します。

平均値の分布はＮ（μ，σ^2／ｎ）に従い，平均値の標準偏差は$\sigma/\sqrt{n}$で計算します。誤差の伝播則からは下記のように導けます。

平均は$\bar{x}=\frac{x_1+\ldots+x_n}{n}$で，各ｘは正規分布すると考え，その各誤差は，誤差伝播の法則から各偏微分で，分散は同じになることから，

$$\sigma_{\bar{x}}=\sqrt{\left(\frac{\partial\bar{x}}{\partial x_1}\sigma_{x_1}\right)^2+\ldots+\left(\frac{\partial\bar{x}}{x_N}\sigma_{x_n}\right)^2}=\sqrt{\left(\frac{\sigma_{x_1}}{n}\right)^2+\ldots+\left(\frac{\sigma_{x_n}}{n}\right)^2}$$

$$=\sqrt{n\left(\frac{\sigma_{x_1}}{n}\right)^2}=\frac{\sigma_x}{\sqrt{n}}$$

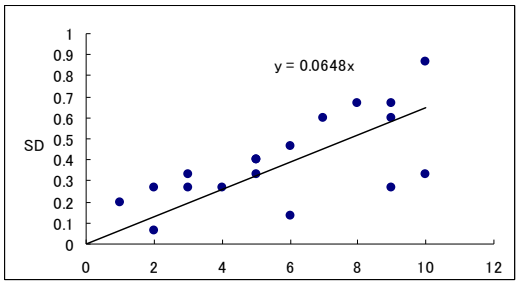
となります。このことは後で述べる最尤法，モーメント母関数からも導けます。

第１１章　RER と PP 図で誤差を解析する

Rodbaed，Ekins らにより考案された RER（Response Error Relationship）や PP（Precision Profile）を呼ばれる誤差を調べる方法があります。RIA（Radio immune assay）の WHO の精度管理方法として使用されていましたが，現在はあまり広く使用されてはいないようです。[注1）] しかし，その方法は簡単で，測定誤差の解析に非常に有効です。

11.1　RIA の RER と PP 図

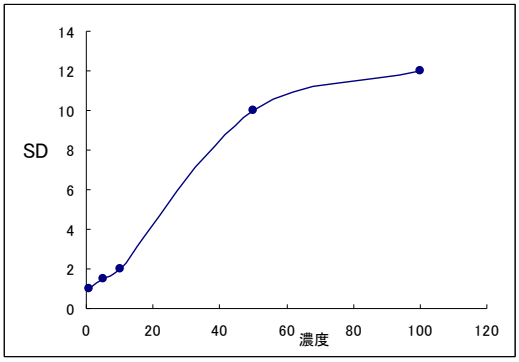
RER は横軸に測定値を取り，縦軸にその標準偏差 s.d.を取ります。RIA では２重測定の各 s.d.をプロットし，原点回帰の傾き（勾配）slope で精度管理をします。



PP は検量線の勾配と RER の勾配を使用して

$$s.d. = \frac{RERからのs.d.}{検量線の勾配}$$

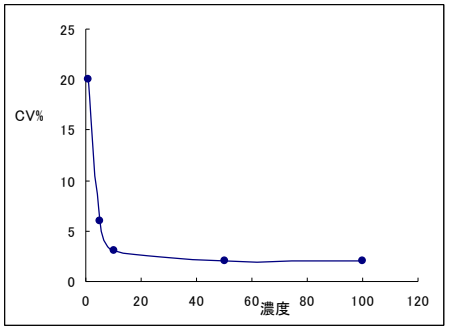
を縦軸にして，横軸に濃度を取ります。



注１）「Radioimmunoassay の精度管理の実際と解説　–WHO 方式ー」ダイナボット・ラジオアイソトープ研究所

このときの縦軸との切片は最小感度となります。

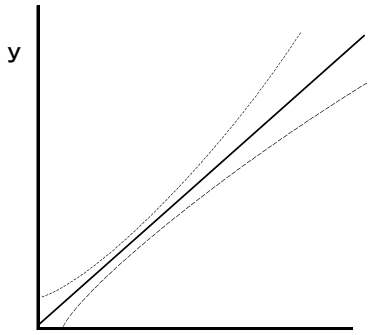
また，c.v.%＝$\frac{s.d.}{濃度}\times 100$ にすると下の図のようになります。



全濃度域での精度（精密度）を知ることができます。

11.2　通常の濃度測定での PP 図の利用

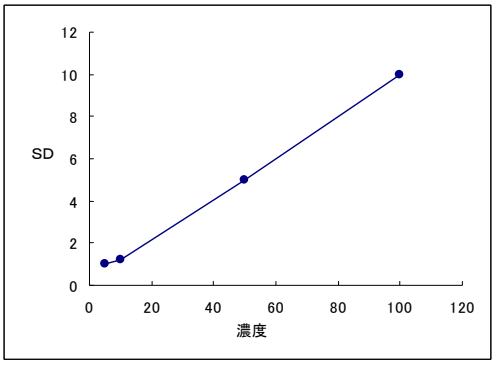
RER や PP 図が有効なのは，測定の全濃度域での精度を図で示すことができ，管理できる点です。

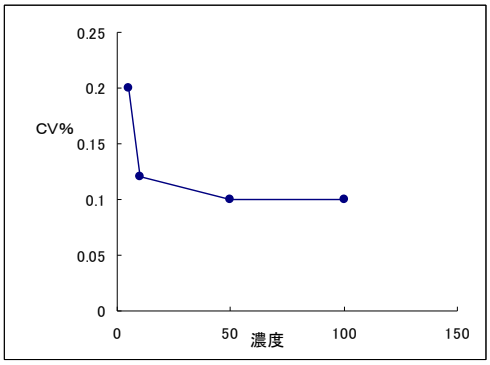


濃度測定で検量線作成し，各濃度での標準偏差を求めた場合，全濃度域で標準偏差が一定であることはほとんどなく，定量範囲を広げると上記の図のようになる場合がほとんどです。

つまり，測定精度を調べるには，１濃度ではなく全濃度域の精度を調べる必要があります。数濃度の精度を調べたとしても，その中間にある測定値の精度は何らの推測や補間が必要となります。

例えば，濃度と標準偏差が比例していなければ，ある検査項目の精度は c.v.%=5.6 であると言うような表現では不十分で，全濃度域での精度を示すべきです。このときに役立つのが PP 図です。各濃度で精度を調べて PP 図が得られます。



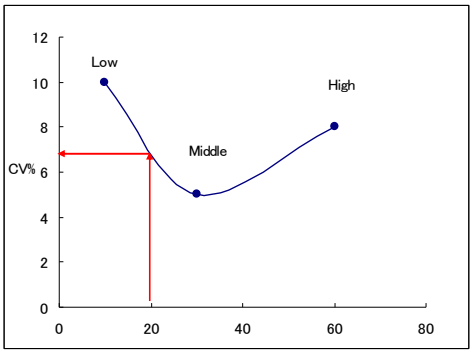


検量線の濃度に合わせて PP 図を作成すると，検量線の重み付けなどを考慮するときにも有効です。

検量線の重み付については「第１６章　検量線の重み付きと回帰式の選択方法」で述べます。

また，日常の精度管理検体の Low，Middle，High の３濃度を上手く取れば，測定値全体の精度を捉えることも可能です。

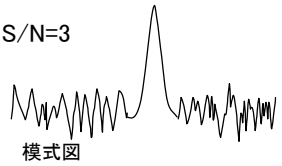
	濃度	C.V.%
Low	10	10
Middle	30	5
High	60	8



例えば右の図のように，調べていない濃度 20 での c.v.%を 7%程度と推定できます。

11.3　検出限界を求める

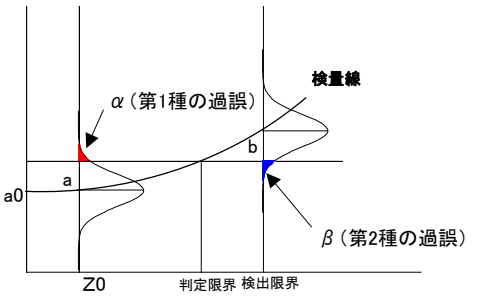
検出限界の求め方には，クロマトグラフ法では S／N の３倍とか，EIA（enzyme immunoassay）では c.v.%が 30%（または 20%）や低濃度域での t 検定などが使用されます。



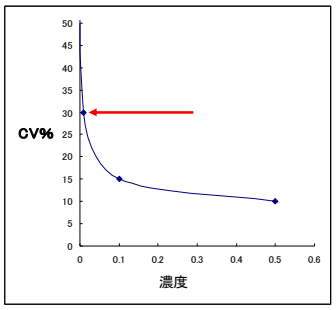
また，右の図に示す 0 濃度の標準偏差の 3 倍の値（IUPC では 3 を採用）

$b = a + 3\sigma$

を検出限界とする方法もあります。[注 1）]



PP 図により，検出限界を求めることもできます。c.v.を縦軸にとり，通常 30%のところの濃度を読み取り検出限界とします。



さらに，PP 図で縦軸に標準偏差 s.d.をとり検出限界を求める方法を紹介します。[注 2)]

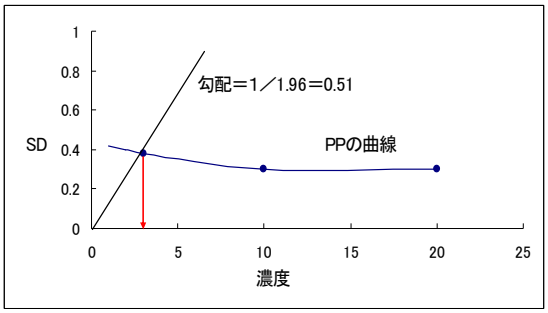
勾配は

$$勾配 = \frac{\sqrt{r}}{t} = \frac{\sqrt{1}}{1.96} = 0.51$$

で求めます。

r は測定回数，t は t 値で，通常の 1 回の測定で t 値＝1.96 とすると，勾配は 0.51 になります。

この勾配の直線を引き，PP との交点の濃度を求め検出限界とします。



検出限界や定量限界については，さらに多くの考えや方法がありす。

注 1）図の「第 1 種の過誤」「第 2 種の過誤」は，本文の「12.4　第 1 種の過誤，第 2 種の過誤」で述べる内容を参考にして下さい。

注 2）杉山 慶彦：「ラジオイムノアッセイの精度管理」RADIOISOTOPES,33,502-509(1984)

第１２章　有意差検定の解釈の誤りと検定に必要なデータ数

差の検定を行うのに，どの程度のデータ数を集めたら良いか判らないとしたら，検定する以前の問題として，データを集めることすら出来ません。

また,信頼区間のみで良いという意見もあります。実際に有意差検定がなくてもよいし，誤って使用するぐらいなら使用しない方が良いのかも知れません。有意差検定の手順は簡単ですが，実際の解釈は解り難いものです。

仮説検定は数学の論理学の「背理法」と同じですが，例えば，平均値の差の検定でＡとＢの有意差が認められないときに，ＡとＢは等しいと結論するのは誤りです。棄却したい仮説（帰無仮説）を立て，その仮説を否定，棄却することを目的とします。

つまり，平均値の差の検定では，差があることを示すために，まず差が無いという仮説を立て，この仮説を否定できるなら差があると結論します。差がないことを示す検定ではありませんので注意して下さい。

また，データが少ないと有意になり難く，逆にデータ数が多いと少しの差でも有意になります。つまり，検定結果はデータ数に左右されます。このため，検定結果をある程度自由に操作することが可能です。だからと言って検定がいい加減であることとは違います。正しく理解すれば検定は有効です。

検定について理解し，誤用しないためには「第１種の過誤」,「第２種の過誤」,「検出力」などについて理解する必要があります。検定の結果が簡単に得られるとしても，これらについて理解しないで検定を行うのは危険です。一見，これから述べる「第１種の過誤」,「第２種の過誤」の話はつまらない内容のように感じられ,「検出力」は難しい話のように思われるかも知れません。有意差検定の結果を理解するには,「第１種の過誤」,「第２種の過誤」,「検出力」などの概要だけでも理解する必要があります。

それでは，よく使用する２標本の平均値の差の検定を例として考えてみます。

注１）柴田義貞：「閑却されたフィシャーの遺産－有意差検定・推定確率－」
日本計量生物学会・応用統計学会 1997 年合同年次大会

参考「平均値の差の検定」では等分散性の「F 検定」をしてから，等分散でない場合はウェルチの t 検定（Welch's t tes）をすることが教科書ではよく記載されています。
しかし，分散に差がありそうな場合は「F 検定」をしないで，ウェルチの t 検定をはじめから使用した方が検定の多重性の問題が生じないとする意見もあります。
多くの場合，ウェルチの t 検定で有意ならば，通常の t 検定でも有意であり，濃度分析ではウェルチの t 検定で十分である場合があります。

12.1　検定方法

平均値の差の検定で，分散は未知であるが同じであると仮定して，平均μ_0と平均μ_1の差を検定する場合を考えてみます。

否定したい仮説$\mu_0=\mu_1$を立てます。仮説を立てたときに，これと対立する仮説を立てます。これを，対立仮説と呼びます。

帰無仮説　H_0　（棄却することを目的として立てる仮説ですので，帰無仮説と呼びます。）

（注）帰無仮説は$\mu_0-\mu_1=0$で差がないことを仮定するためH_0の記号を通常使用します。

H_0：　$\mu_0=\mu_1$

対立仮説　H_1

H_1：　$\mu_0\neq\mu_1$　　$\mu_0>\mu_1$　$\mu_0<\mu_1$

母分散が未知の場合は検定するために検定統計量t_0を計算します。各平均値を$\bar{x}_1$，$\bar{x}_2$とし，各データ数をn_1，n_2とし，分散をＶとすると，下記の検定統計量t_0はｔ分布に従います。[注1）]

$$t_0=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{V(1/n_1+1/n_2)}}$$

分母はデータ数ｎに左右され，データ数が大きければt_0も大きくなり有意になり易くなります。t_0が有意水準（危険率）$\alpha=0.01$や$\alpha=0.05$で有意かを判断します。

有意水準は$\alpha=0.05$ならば５％の確率で検定が間違っている可能性があるが，有意差があることを示しています。$\alpha=0.01$で有意差があれば，より明確に差があると言えます。

注１）ｔ分布は標準正規分布とχ^2分布から導き出すことができます。

$$t_0=\frac{z}{\sqrt{\frac{\chi^2}{\phi}}}=\frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}}$$

Ｚ：標準正規分布に従う変数

上式から導き出せるｔ分布の確率密度関数は本文「12.6　検出力」で示します。

ｔ検定そのものは，本文の「第１３章　最尤法について」で述べる尤度比検定と「ネイマン・ピアソンの補題」などの知識から，有意水準αの尤度比検定はｔ検定を誘導します。

尚，平均値の差の検定で分散が概知の場合は，検定統計量は標準正規分布Ｎ（0，1^2）に従います。

$$u_0=\frac{\bar{x}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$$

検定統計量は「第８章　正規分布について」で述べた規格化になっています。

さらに，ｔ検定などは$\mu_0-\mu_1=0$で表せる線形仮説の検定です。

有意差があれば，H_0を棄却（否定）します。もしも，仮説 H_0が棄却できないときは，仮説 H_0を採択しますが，仮説を積極的に支持したわけではありませんので，「保留」したと表現する場合もあります。

12.2　数値例（データ数の違い）

検定方法を示しましたが，実際に簡単な例で平均の差の検定をしてみます。計算はエクセルの「分析ツール」を使用しました。

A が平均 μ_a=10　標準偏差 σ_a=2

B が平均 μ_b＝８　標準偏差 σ_b=2

に従う正規乱数のデータをエクセルで各 5 個発生してみました。

	A	B
1	9	9
2	8	9
3	11	7
4	10	5
5	6	4

５個でも，平均値の差 10－8＝2 の有意差が認められるのか確認してみます。

帰無仮説 H_0　：$\mu_a=\mu_b$

対立仮説 H_1　：$\mu_a\neq\mu_b$

A，B の平均値をエクセルの分析ツールの「t 検定 ： 等分散を仮定した２標本による検定」で検定してみます。すると下記の結果が出力されます。[注1)]

t-検定 : 等分散を仮定した2標本による検定

	変数 1	変数 2
平均	8.8	6.8
分散	3.7	5.2
観測数	5	5
プールされた分散	4.45	
仮説平均との差異	0	
自由度	8	
t	1.499	
P(T<=t) 片側	0.08612	
t 境界値 片側	1.860	
P(T<=t) 両側	0.1722	
t 境界値 両側	2.306	

少しこの出力されて結果の見方を説明します。

ｔ＝1.499 となっていますが，計算した検定統計量が

$$t_0=\frac{\bar{x}_a-\bar{x}_b}{\sqrt{V(1/n_a+1/n_b)}}=1.499$$

であることを示しています。

注１）Microsoft Excel の分析ツールの「t 検定 ： 等分散を仮定した２標本による検定」は Excel の機能として備わっています。

本文中有意水準を α としますが，エクセルでは有意確率を P が表示されます。

このときの両側検定の有意確率は P＝0.1722 となっています。有意水準 0.05 と比較します。P＝0.1722 は 0.05 よりも大きいので，有意水準 0.05 でも H_0は棄却できません。つまり，有意差を認めることができないと言えます。差がないと積極的に言っているのではない点に注意して下さい。もし，P＜0.05 ならば危険率 0.05 で有意差があり，P＜0.01 ならば危険率 0.01 で有意差があると判断します。

ところで，データ数が 5 と少ないから有意差がなかったと思われます。
そこで
A が平均μ_a=10　標準偏差σ_a=2
B が平均μ_b＝８　標準偏差σ_b=2
として，エクセルの正規乱数でデータ数を 20 程度に増やして検定してみます。

	A	B
1	9	9
2	8	9
3	11	7
4	10	5
5	6	4
6	9	7
7	9	9
8	7	7
9	9	5
10	13	8
11	11	8
12	8	4
13	9	8
14	12	12
15	10	8
16	10	7
17	10	8
18	10	7
19	12	6
20	10	6

t-検定：等分散を仮定した2標本による検定

	変数 1	変数 2
平均	9.65	7.2
分散	2.871	3.642
観測数	20	20
プールされた分散	3.257	
仮説平均との差異	0	
自由度	38	
t	4.293	
P(T<=t) 片側	5.86E-05	
t 境界値 片側	1.686	
P(T<=t) 両側	0.0001172	
t 境界値 両側	2.024	

結果は P＝0.0001172 で, P＜0.01　でも有意差が認められます。データ数を増やせば有意差を検出できることを示しています。

10－8＝２の差は，データ数を 20 にすれば検出できることが解ります。もし差が無いのならば，データ数を 20 に増やしても有意差が検出されないのは当然です。

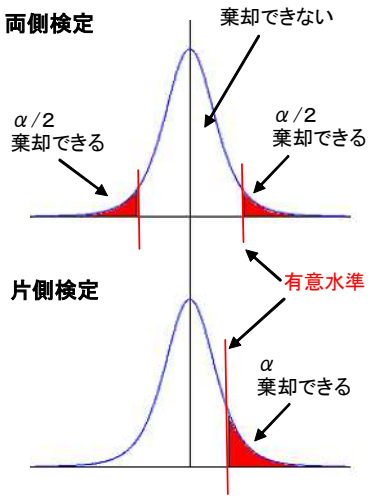
12.3　有意水準α（片側検定と両側検定）

検定の例を示しましたので，検定についてさらに考えてみます。
検定するときに検定統計量（ｔ値など）を計算し，確率分布（ｔ分布など）と比較します。
このとき判断基準の有意水準をα=0.05 とかα=0.01 にして検定します。
模式図で示すと下の図のようになります。
片側検定は２標本の検定では，一方が低いとか高いとかの予測がすでに得られている場合です。両側検定は，どちらが低いとか高いとかの予測が得られていない場合です。

この有意水準α＝0.01 や 0.05 は，人の経験にるもので，数学的な意味付けがあるわけではありません。0.013 や 0.04 でもよいのですが，99%や 95%を判断基準にする方が感覚的に理解し易いためです。

「第１３章　最尤法について」で述べる AIC(赤池の情報量基準)の関係から，α＝0.2 以下ならば有意になる可能性があるので，0.2 を注意すべき目安とする考えもあります。

通常は，この有意水準をα＝0.01 や 0.05 のように小さい値にするのは，後で述べる「第１種の過誤」を小さくしたいためです。
「棄却できる」とはαの過誤の確率があるが，有意差を認めることができることを意味しています。



12.4　第１種の過誤，第２種の過誤

それでは，検定ではどの様な過誤があるか考えてみます。
有意水準αの意味は，先ほどの図から考えると「正しい仮説 H_0 を誤って棄却する確率」です。つまり，先ほどの例では，平均値に差が無いのに差があると見なす確率がαです。
これは**第１種の過誤**と呼ばれる確率です。
有意水準 0.05 で有意とは，「平均値に差が無いのに差があると誤って判断する確率が 0.05（５％）である」と言っています。
逆に「正しくない仮説 H_0 を誤って採択する確率」を**第２種の過誤**の確率と呼びます。
この関係は下の表のようになります。

	H_0 帰無仮説が正しい **ｔ分布**	**H_1 対立仮説が正しい** **非心ｔ分布**
H_0 を選択した場合	**１－α** 仮説（H_0）が正しく，H_0 を選択	**β** **第２種の過誤**
H_1 を選択した場合	**α** **第１種の過誤**	**１－β（検出力）** H_1 が正しく，H_1 を選択

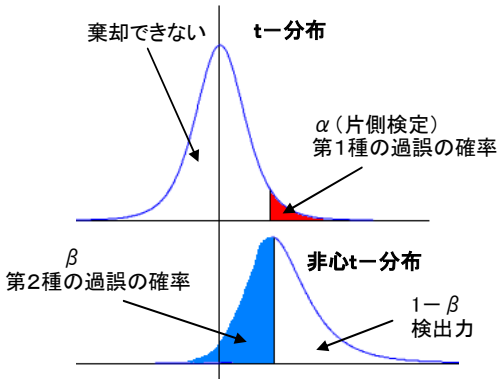
データ数が一定の時に第１種の過誤と第２種の過誤を同時に小さくできません。
そこで，仮説検定の目的からして，第１種の過誤の確率を少なくするために，α=0.05 とかα=0.01 にします。
当然，有意水準 0.01 で有意ならば，0.05 よりも誤りの危険が少ない，または，明確に差があると判断できます。

第１種の過誤の確率を **α** とします。

第２種の過誤の確率は **β** とします。

第２種の過誤βを犯さない確率 **1－β** を「**検出力**」**Power** と言います。この検出力は検定の良し悪しを判断する基準となり，1－β＝0.8 程度を確保するように実験を計画します。ここで示した「検出力」は計算できます。計算例を後で示しますが，このことにより，検定に必要なデータ数も算出できます。また，検定が複雑ならばシミュレーションで求めることもできます。

第１種の過誤の確率αとか第２種の過誤の確率βとか少し分かり難いものです。次の図をよく見て下さい。例として平均値の差の検定で使用するｔ検定での片側検定でのαとβの関係を模式図で示しています。Ho でのαはｔ分布から計算できますが，対立仮説 H_1 下で確率であるβは，非心ｔ分布と呼ばれる分布に従います。



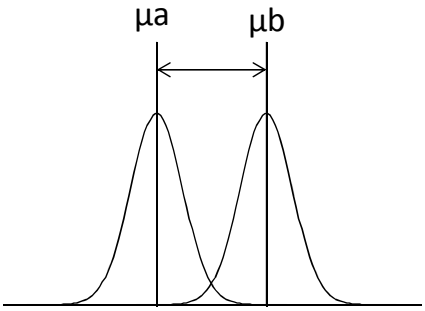
第２種の過誤の確率βは対立仮説（H_1）下の分布で，中心からズレた非心ｔ分布に従います。

上の図から有意水準αの 0.05 か 0.01 で判断すると決めると，そのときの検出力は非心分布から１－βを求めます。

12.5　有意差ｄ

さらに，次の用語を付け加えます。

有意差ｄ：仮説 H_0は，一般に差がない，下記式のｄ＝0 を仮定します。どのくらいの差ｄを有意差と考えるかの，検出したい有意差の量的表現です。



例えば，平均値の差の検定でμ_aとμ_bの差の検定では，

$$d = \frac{|\mu_a - \mu_b|}{\sigma}$$

です。

ｄは標準偏差の何倍の差があるのかを示しています。

標準偏差σが小さいか，μ_aとμ_bの差が大きければ有意差ｄは大きくなります。ｄが大きければ検出し易くなります。

２標本の平均値の差の検定では Cohen(1988 年)が提案した

１）小さな差を検出する場合ｄ＝0.2

２）中位な差を検出する場合ｄ＝0.5

３）大きな差を検出する場合ｄ＝0.8

を目安にすることもできます。[注1)]

この有意差ｄにより，検出力，検定に必要な検体数を算出できます。

注 1）古川 俊之，丹後 俊郎：「医学への統計学」朝倉書店（1993 年）

12.6　検出力

先ほど説明したように，検出力は下記のようになります。

検出力１－β：第２種の過誤を犯さない確率です。

仮説が正しくないときに，仮説を棄却する確率，または，正しい対立仮説を正しく採択する確率です。

検出力を求めるための数式を少し示しますが，詳細は省略します。

実際の計算は式を知らなくても，後で述べるパソコンの利用などで計算結果を得ることができます。

ｔ検定に使用するｔ値は，自由度φの下記のｔ分布の確率密度関数に従うとして，検定するものです。

$$f(x:\phi)=\frac{1}{\sqrt{\phi}B(1/2,\phi/2)}\left(1+\frac{x^2}{\phi}\right)^{-(\phi+1)/2} \qquad (-\infty<x<\infty)$$

B（１／2，φ／2）はベータ分布です。

実際はエクセルでもｔ分布やベータ分布の値は関数として計算できます。

検出力の計算には非心ｔ分布の値を知る必要があります。

先ほどの有意差ｄから

$$d=\frac{|\mu_A-\mu_B|}{\sigma}$$

$$\lambda=\sqrt{n}\left(\frac{|\mu_A-\mu_B|}{\sigma}\right)=d\sqrt{n}$$

でλは非心度で，非心ｔ分布は

$$f(x:\phi,\lambda)=\frac{\exp\left(-\frac{\lambda^2}{2}\right)}{\sqrt{\phi\pi}\Gamma\left(\frac{\phi}{2}\right)}\sum_{j=0}^{\infty}\Gamma\left(\frac{\phi+j+1}{2}\right)\frac{(\phi x)^j}{j!}\left(\frac{2}{\phi}\right)^{\frac{j}{2}}\left(1+\frac{x^2}{\phi}\right)^{-\left(\frac{\phi+j+1}{2}\right)}$$

となります。

Γ（.）はガマン関数です。

自由度φ，　ｄ＝μ／σ，　ｄ＝０なら通常のｔ分布です。

この非心ｔ分布をから，ｔ$_0$＝λとなる検出力（Power）は両側検定で

$$Power = 1-\beta = P_r\{t_0 \le -t(\phi,\alpha)\} + P_r\{t_0 \ge t(\phi,\alpha)\}$$

で計算します。
尚，片側検定では

$$Power = 1-\beta = P_r\{t_0 \le -t(\phi,2\alpha)\}$$

または

$$Power = 1-\beta = P_r\{t_0 \ge t(\phi,2\alpha)\}$$

です。

これまでに説明した用語を理解すれば，JIS Z 9041 には例があり，その通りの手順に従えば，検出力，必要なデータ数が計算できます。永田靖：「サンプルサイズの決め方」朝倉書店に計算例があり，「統計数値表」JSA-1972　日本規格協会にも非心分布の表やプログラムが記載されています。(JSA-1972　日本規格協会のプログラムはバグがあるようですので，注意が必要です。)

また，本文の「付録３」にプログラムを記載しました。エクセルの VBA の標準モジュールに関数として貼り付ければ，エクセルの関数として

関数　＝pow2(データ数 A,データ数 B,有意差 D,有意水準 A)

で計算できます。[注 1)]

さらに，無料の統計処理ソフト「**R**」でも検出力は計算できます。
本文の「付録２　無料パソコンソフトの利用」に，Ｒによる検出力の計算方法を示しました。[注２)]
とにかく，検出力はデータ数ｎ，有意差ｄ，有意水準αの値が解れば計算できます。

注 1）Microsoft Excel のマクロのＶＢＡ（Visual Basic for Application）で動きます。
（永田靖：「サンプルサイズの決め方」朝倉書店に従い，プログラムにしたものです。）

注２）検出力を求めるのが困難な場合はシミュレーションで求めます。
予測される乱数を発生させて検定し，検出力を求めます。（モンテカルロシミュレーション）
参考：山田，杉澤，村井：「「R」によるやさしい統計学」オーム社（2008 年）

12.7　検出力の計算例

用語の説明をしたので，検出力と必要なデータ数について，先ほどのデータで考えてみます。

A が平均 μ_a=10　標準偏差 σ_a=2

B が平均 μ_b＝８　標準偏差 σ_b=2

に従う正規乱数の５個のデータでは有意差が認められませんでしたが，20 個では有意差が認められました。

	A	B
1	9	9
2	8	9
3	11	7
4	10	5
5	6	4

t-検定：等分散を仮定した2標本による検定

	変数 1	変数 2
平均	8.8	6.8
分散	3.7	5.2
観測数	5	5
プールされた分散	4.45	
仮説平均との差異	0	
自由度	8	
t	1.499	
P(T<=t) 片側	0.08612	
t 境界値 片側	1.860	
P(T<=t) 両側	0.1722	
t 境界値 両側	2.306	

有意差　無し

	A	B
1	9	9
2	8	9
3	11	7
4	10	5
5	6	4
6	9	7
7	9	9
8	7	7
9	9	5
10	13	8
11	11	8
12	8	4
13	9	8
14	12	12
15	10	8
16	10	7
17	10	8
18	10	7
19	12	6
20	10	6

t-検定：等分散を仮定した2標本による検定

	変数 1	変数 2
平均	9.65	7.2
分散	2.871	3.642
観測数	20	20
プールされた分散	3.257	
仮説平均との差異	0	
自由度	38	
t	4.293	
P(T<=t) 片側	5.86E-05	
t 境界値 片側	1.686	
P(T<=t) 両側	0.0001172	
t 境界値 両側	2.024	

有意差　有り

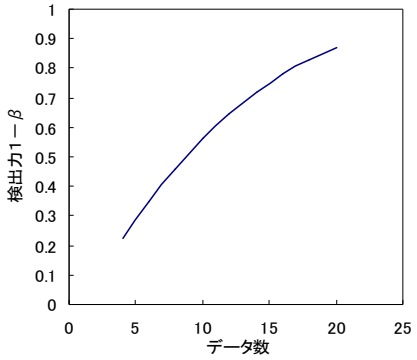
A と B の平均値の差は 10－8＝2 で，標準偏差は共に 2 です。有意水準 α ＝0.05 で検出力を調べてみます。

有意差は先ほど示したように

$$d = \frac{\mu_A - \mu_B}{\sigma} = \frac{2}{2} = 1$$

です。
検出力は非心 t 分布に従います。

計算は先ほどの「12.6 検出力」を参考にして行うと，
各 5 個では
$1 - \beta$ ＝0.3
各 20 個では
$1 - \beta$ ＝0.9
になります。
データ数を各 20 にすれば検出力は $1 - \beta$ ＝0.9 になり，ほぼ確実に平均値の差 10－8＝2 を検出できることになります。

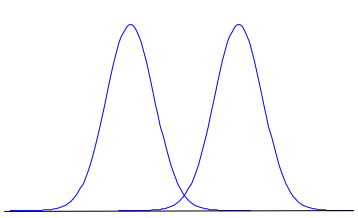


この例では，検出力から，データ数は 20 程度集めなければ，有意差は検出でいないことになります。

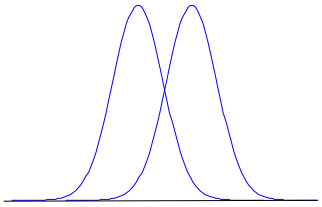
12.7.1　有意差とデータ数

データ数が少ないと有意になり難く，増やすと少しの差でも有意になります。

データが多く無意味な有意差である場合は，技術的に意味のない差と結論します。逆にデータ数が少なくて検出できない場合は，必要な検出力 1－β が得られるデータを集めます。



n =5
差が大きくないと検出できない



n ＝1000
差が小さくても検出できる

12.7.2　有意差 d からのデータ数（局外母数の問題とデータ数）

検出力１－β を求めるにはデータ数 n，有意水準 α，有意差 d （平均値の差と標準偏差 s），の値が必要です。

データ数 n，有意水準は 0.05 として決めても，有意差 d に未知の標準偏差（平均値の検定では**局外母数**となる）が含まれます。

しかし，実際の現場では，技術的にある程度の標準偏差が推定できれば，検出したい意味のある差からデータ数と検出力の計算ができます。

さらに，何の情報がなくても有意差 d で述べたように，２標本の平均値の差の検定では Cohen が提案した

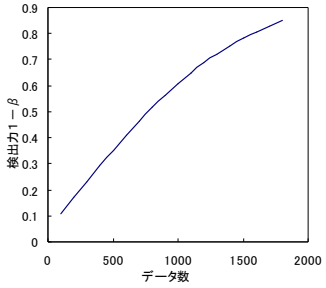
１）小さな差を検出する場合 d ＝0.2
２）中位な差を検出する場合 d ＝0.5
３）大きな差を検出する場合 d ＝0.8

を目安にすることもできます。

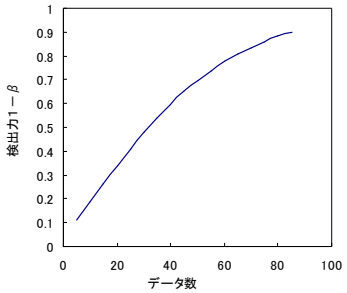
d ＝0.１， d ＝0.5， d ＝１.0， d ＝2.0 の時の有意水準 0.05，検出力 0.8 で図を作成してすると，次のようになります。

有意差ｄからのデータ数

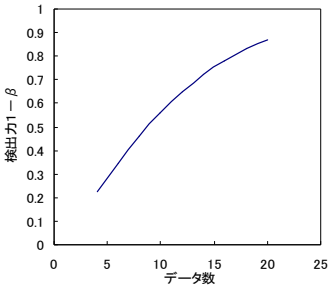
$$d = \frac{|\mu_A - \mu_B|}{\sigma}$$



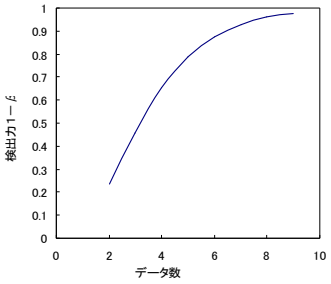
ｄ＝0.1



ｄ＝0.5



ｄ＝1.0



ｄ＝2.0

上の図から以下のことが読み取れます。

２標本の平均値の差が標準偏差の 1／10 程度，つまり，有意差ｄ＝0.1 では 1500 個程度のデータが必要となります。
２標本の平均値の差が標準偏差と同じ程度（ｄ＝１）ならば，20 個程度のデータは欲しいところです。
明らな差，つまり２標本の平均値の差が標準偏差の２倍程度の差（ｄ＝2.0）を検出したいのならば，５個程度のデータ数で検出できます。

ここに示した図を参考にある程度データ数を決めることが可能です。

本文では，平均値の差の検定を例として示しましたが，他の場合は永田　靖：「サンプルサイズの決め方」朝倉書などを参考にして下さい。

また，簡単に検出力，データ数を求められない場合も少なくありません。その場合はシミュレーションで求めます。[注２)]

注１)統計処理ソフト「R」にで必要なデータを求める方法は「付録２　無料パソコンソフトの利用」の R の使用方法で記載しました。

注２)山田，杉澤，村井：「R」によるやさしい統計学」オーム社（2008 年）

12.8　検定結果について

平均値の差の検定で，A と B の有意差が認められないときに，A と B は等しいと結論するのは誤りであることを述べました。さらに，データ数が少ないと有意になり難く，逆にデータ数が多いと少しの差でも有意になります。

有意意にならない場合は検出力が低い可能性があります。また，有意になった場合は，意味のある差であるか検討する必要があります。

有意差を認めた場合で重要なのは，**統計的有意**と**実質的有意**と言う判断です。

つまり，統計的有意となっても，技術的に意味のない有意差は問題として扱う必要はありません。技術的判断が優先されます。

（1）有意差が無い場合，同等であると積極的に言うことは出来ない。

（2）データ数を増やすと差は検出し易くなる。

（3）実質的に意味のある差であるかを検討する必要がある。

データに差がなければ，データ数を増やしても有意とはならないと思われますが，実際は完全に測定条件が同じになることはないので，データを相当数増やせば無意味な少しの差でも有意となります。

重要なのは，どの程度の差を問題として検出したいのかを明確にし，そのために必要な検出力が得られるデータ数を集め，有意差検定を行うことです。

12.9　検定は繰り返してはいけない（多重性の問題）

平均値の差の検定を例にして有意差検定について述べましたが，検定を繰り返すことが出来ないことを述べます。

これは「**検定の多重性の問題**」と呼ばれるものです。

検定で第 1 種の過誤の有意水準を 0.05 にしても，同じデータで検定を繰り返すと $\alpha = 0.05$ よりも大きくなってしまいます。

Welch の t 検定について

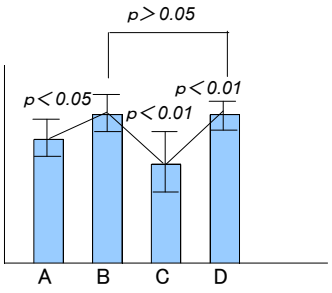
平均値の検定でスチューデントの t 検定を例に述べましたが，教科書的には F 検定で 2 標本が等分散ならスチューデントの t 検定を行い，分散が異なる時は Welch の t 検定を行うと述べてあるものがあります。

しかし，明らかに分散が異なるのなら，多重性の問題から F 検定を行わないで Welch の t 検定を直接行うべできであるとする意見もあいます。この意見が支持されてきています。

第１種の過誤

右の図のような，一元配置分散分析で
t 検定を繰り返すことは出来ません。
$\mu_A=\mu_B=\mu_C=\mu_D$ を検定するのに，
$\mu_A=\mu_B$，$\mu_A=\mu_C$，$\mu_A=\mu_D$，$\mu_B=\mu_C$，$\mu_B=\mu_D$，$\mu_C=\mu_D$ と分けて検定すれば，
${}_4C_2=6$ 通りあるので，第１種の過誤が大きくなることが想像できます。

α＝0.05 ではなくなり，有意になり易くなります。



６回の検定で１つでも第１種の過誤を犯している確率は，繰り返し数を k とすると下記のように計算できます。（Bonferroni 法）

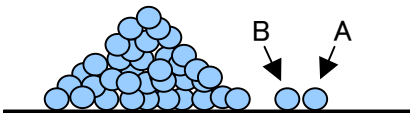
$$1-(1-\alpha)^k \approx k\alpha$$

より，有意水準は 0.05 が $6\times0.05=0.3$ と大きくなります。[注 1)]
Bonferroni 法は有意水準を α / k にする方法です。
第１種の過誤を 0.05 に維持するために，多重比較法として Tukey 法，Dunnett 法，Scheff 法などがあります。[注 2)]

平均値の差の検定だけではなく，検定は基本的に繰り返すことは誤用と見なされます。
これは，第１種の過誤の確率を維持出来ないためです。

例えば，飛び離れたデータを棄却するのに Gurubbs-Smirnov 検定がありますが，１個のデータの棄却検定法で，複数回棄却検定を繰り返すことは出来ません。

２個のデータを棄却したい場合，保守的になりますが，下の図のような場合はＢを検定して，有意ならば，Ａも棄却することは可能です。



注 1）シミュレーションでも α＝0.05 より大きくなることは確認できます。
山田，杉澤，村井：「「R」によるやさしい統計学」オーム社（2008 年）

注 2）計算手法は下記の本などを読んで下さい。
永田靖，吉田道弘：「統計的多重比較法の基礎」サイエンティスト社（1997 年）

注 3）多重比較は未解決な問題もあり，臨床検査や環境検査では薬物分析などとは異なり，Bonferroni 法で有意差を示せば十分な場合が多くあります。

第１３章　最尤法について

13.1　尤度関数とは何か

推定量を求める方法として**最尤法**（さいゆうほう）があります。尤度は理解し難い概念の一つですが，基本的な統計手法です。

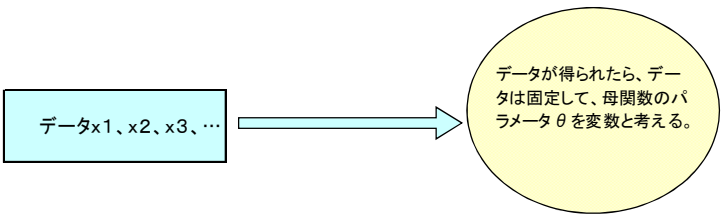
最尤法は一見当たり前のことを，見方を変えただけに過ぎないように思えます。正規分布を例として説明します。[注1)]

通常，私たちは母集団の一部としてデータｘを得たと考えます。分析での母集団は通常無限のデータを考えていますから，その一部としてのデータから母集団の母平均値や母分散を推定しています。

しかし，最尤法ではデータｘが一度得られたら，このデータに尤もらしい平均や分散は何かを考えます。「データｘは，確率最大のものが実現した」と仮定します。（**最尤原理**）

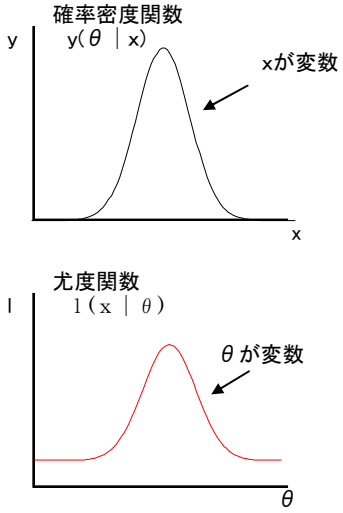
確率密度関数ｙ(θ｜ｘ)に従うとし，θはパラメータ（母数）で正規分布なら平均と分散がθで，ｘはデータです。測定データx_1，x_2，…x_nが得たときにx_1，x_2，…x_nは得られ固定されたと考えて，ｙ(θ｜ｘ)のｘの関数ではなく，逆にθの関数として見方を変え l(ｘ｜θ)を考えます。この l(ｘ｜θ)を**尤度関数**と呼び，その値が**尤度**です。

尤度関数はもはや確率である必要はありません。
つまり，確率ではないので，尤度関数を積分して１になる必要もありません。[注2)]



注１）本文「第４章　ベイズの定理」で述べたように，ベイズの定理から最尤法を誘導することができます。「第４章　ベイズの定理」の例で示した，赤球が得られたとして，それがＡの箱かＢの箱かを考えたのと似ていることは，感覚的にも理解できます。つまり，「第４章　ベイズの定理」で示した$p(x|\theta)$は尤度関数となります。

注２）正規分布ならば，$F(x)=\int_{-\infty}^{\infty} f(x)dx=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx=1$で１になります。



尤度関数 l (x ｜ θ)を最大にする，尤もらしい θ の推定値 $\hat{\theta}$ を**最尤推定値**と呼びます。

それでは，簡単な数値例で考えてみます。
正規分布の確率密度関数は

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

です。
例えば，分散 σ ²= 1 の分析データ 10，８，９，11 を得たとします。

データ	10	8	9	11

このデータを得る確率は積の形になるので，積記号Πを使用して

$$f(x_1,x_2,..x_n)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}\times\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_2-\mu)^2}{2\sigma^2}}\times\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_3-\mu)^2}{2\sigma^2}}\times...\times\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

$$=\prod_{i=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right\}$$

$$=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

になります。

このとき得たデータは固定して，平均を変数θと考えると，尤度関数は

$$l(\theta)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2}$$

となります。

この尤度関数で最大となるものが，尤もらしいθと考えられます。尤度関数Ｌ（θ）では大きさのみに関心が向きます。大きさのみならば対数にした方が，積が和に変わり微分するのが簡単で，知りたい最大となる位置は変わりません。（対数微分法）

対数にすると先ほどの式は

$$L(\theta)=\log l(\theta)=-\frac{n}{2}\log(2\pi)-n\log\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2$$

となります。

はじめに，分散σ²＝１のデータ 10，９，8，11 を得たとしますと述べました。

関心は大きさのみなのでθに関する部分のみで十分です。

$$L(\theta)=-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2$$

この式の最大値が最尤推定値となります。

式を展開して，x_iにデータ 10，９，8，11 を入れると

$$L(\theta)=-2\theta^2+38\theta-183$$

となります。実際にθに適当に数値を入れて計算すると

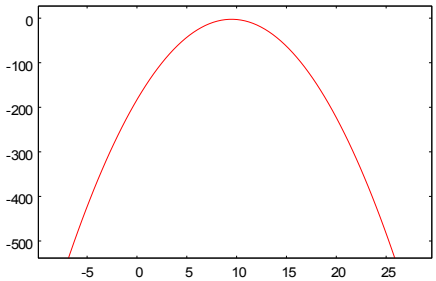
θがもし８ならＬ（θ）＝－７

θがもし 9.5 ならＬ（θ）＝－2.5

θがもし 10 ならＬ（θ）＝－3

で，9.5 付近に期待値として尤もらしい最尤推定値があると考えらます。

先ほどのθの２次関数$L(\theta)=-2\theta^2+38\theta-183$を図にすると下記のようになります。



この図の最大となる θ の値を求めるには,

$$L(\theta) = -2\theta^2 + 38\theta - 183$$

を微分して 0 とおけばよいので

$$\frac{\partial L(\theta)}{\partial \theta} = -4\theta + 38 = 0$$

$$\theta = \frac{38}{4} = 9.5$$

になります。これは平均値に他なりません。同様なことは「第２章　平均値の計算方法を考える」でも述べました。

$$\frac{10+9+8+11}{4} = 9.5$$

正規分布すると仮定した場合，データ 10，９，8，11 が得られときの期待値の最尤推定量は平均値で 9.5 となります。

先ほど示したように，下記の正規分布の対数尤度関数

$$L(\mu,\sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

から，母平均 μ の最尤推定量は微分して 0 とおくことにより

$$\frac{\partial L(\mu,\sigma))}{\partial \mu} = \frac{1}{\sigma^2}\sum(x_i - \mu) = 0$$

より

$$\hat{\mu} = \frac{\sum x_i}{n}$$

で平均値の式になります。

また，母分散 σ ２の最尤推定値も微分して 0 とおくことにより

$$\frac{\partial L(\mu,\sigma))}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum(x_i - \mu)^2 = 0$$

から

$$\hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \hat{\mu})^2$$

となり，分散の最尤推定量が求まります。これは不偏推定量のｎ―1 で割る不偏分散ではありませんが，最尤推定量です。

何か当たり前のことを難しくしたような気がしますし，また何をやったのか解り難い話です。ここまででやったことは，データが得られたら，データはもはや得られた時点で，母数のパラメータを変数とみて尤度関数を考える。尤度の対数をとり対数尤度の最大となるものを，微分して０とおいて尤もらしい最尤値を求めました。[注1)]

検定では尤度の比で検定します。[注2)]

一般尤度比は

$$\lambda(x)=\frac{\sup L(\theta_0;x)}{\sup L(\theta_1;x)}$$

sup:最大のもの

です。この一般尤度比から各種の検定方法を導き出すことができます。

－２×対数尤度比は

$$-2\times\log\frac{\sup L(\theta_0;x)}{\sup L(\theta_1;x)}$$

自由度ｐ（パラメータ数）のχ^2分布に従います。χ^2検定できます。

ベイズ統計，統計モデルの選択，尤度比検定など，尤関数を使用します。

よく使用する「第 12 章　有意差検定の解釈の誤りと検定に必要なデータ数」で説明したｔ検定も，有意水準αでの一般尤度比から導けます。[注3)]

最尤法は統計学の基本です。

注1）　尤度方程式が簡単な例を示しましたが，連立微分方程式になっています。

$$\frac{\partial L(\theta)}{\partial \theta_\mu}=\frac{\partial L(\theta)}{\partial \theta_\sigma}=0$$

尤度関数が複雑であるとか，非線形である場合などは数値解析的に求めることがよくあります。

一般に尤度関数Ｌ（θ）の最尤推定値は上に凸の関数となり，偏微分して０となる

$$\frac{\partial L(\theta)}{\partial \theta_1}=\frac{\partial L(\theta)}{\partial \theta_2}=\ldots=\frac{\partial L(\theta)}{\partial \theta_m}=0$$

を求める必要があります。この連立微分方程式を解くにはニュートン・ラフソン法などの反復収束法を使用します。

注 2）竹本康彦，有薗育生：「２つの正規分布の同等性性に関する尤度比検定の検出力特性に関する考察」応用統計学　Vol.31.No.2.(2002).141-162

注 3）**ネイマン・ピアソンの補題**：

尤度比による検定が，仮説検定では最も検出力が高い。

野田 一雄，宮岡 悦良：「入門・演習　数理統計」共立出版

13.2　エクセルで最尤法を理解する

13.1 で対数微分して平均と分散を算出しました。

しかし, 最尤値を求めることは, 対数微分をして解を求めることは本質ではありません。エクセルで計算すると最尤法の計算の本質が理解できます。

「尤度関数 l (x ｜ θ)を最大にする，尤もらしい θ の推定値が最尤推定値です。」

対数を取り，積を和に変えます。

$$L(\theta) = \ln(\theta) = \sum \ln\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x_i-\mu)^2}{2\sigma^2}}\right]$$

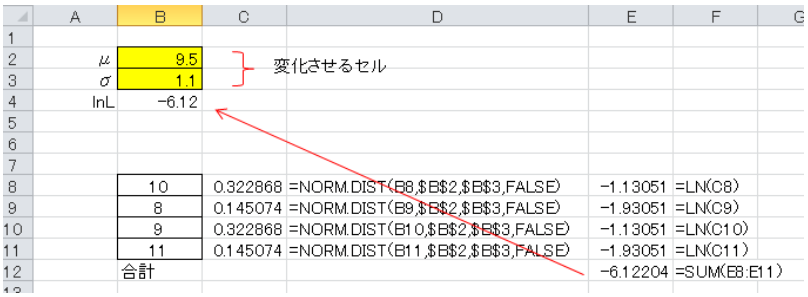
先ほどの例，分析データ 10，８，９，11 を得たとします。

データ	10	8	9	11

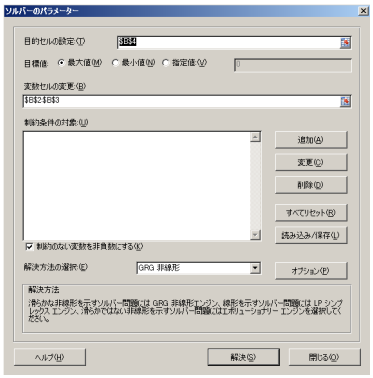
この平均（期待値）μ と標準偏差は σ はソルバーで同時に求めることが出来ます。

下記の図のエクセルのセルに 10，8，9，11 を入れて，上のデータに対する確率密度関数の値を求めるには NORMDIST 関数（x,平均,標準偏差,FALSE）

を使用して，ソルバーで尤度関数が最大になる値を求めればよいのです。



	A	B	C	D	E	F	G
1							
2	μ	9.5		変化させるセル			
3	σ	1.1					
4	lnL	-6.12					
5							
6							
7							
8		10	0.322868	=NORM.DIST(B8,$B$2,$B$3,FALSE)	-1.13051	=LN(C8)	
9		8	0.145074	=NORM.DIST(B9,$B$2,$B$3,FALSE)	-1.93051	=LN(C9)	
10		9	0.322868	=NORM.DIST(B10,$B$2,$B$3,FALSE)	-1.13051	=LN(C10)	
11		11	0.145074	=NORM.DIST(B11,$B$2,$B$3,FALSE)	-1.93051	=LN(C11)	
12		合計			-6.12204	=SUM(E8:E11)	



$$\hat{\mu} = 9.5$$
$$\hat{\sigma} = 1.1$$

対数尤度関数で最大となる値として, 最尤値が求まります。

データ 10，8，9，11 の標準偏差 σ は，n で割った自由度を考慮しない値になっていています。

最尤法でも自由度を考慮すれば，通常の標本標準偏差 s.d.=1.3 が求まります。

13.3　AIC とは何か

最尤法を説明しましたので，よく使用する **AIC**（赤池の情報量規準 Akaike's Information Criterion）による統計モデルの選択方法を説明します。[注１）]

AIC は

$$AIC \equiv -2\log(L) + 2p$$

で L は最大尤度で，ｐはパラメータ数で，パラメータ数をペナルティとして AIC の小さいモデルを最適なモデルとして選択します。

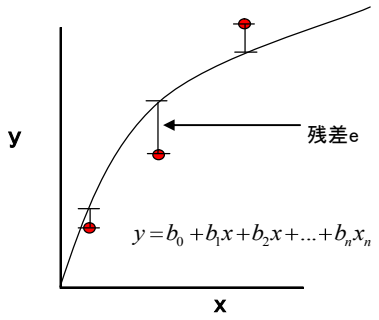
検量線で $y = ax + b$ の１次回帰式（直線回帰）か２次回帰などの高次回帰式かの選択を AIC で行うことを考えてみます。

ｍ次回帰式の尤度関数は，回帰モデルを一般化すると

$$y_i = b_0 + b_{i1}x + b_{i2}x^2 + \ldots + b_{ik}x^m + \varepsilon_i$$

より，回帰式からの残差は正規分布すると考えると，

$$l = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - b_0 - \sum_{j-1}^{m} b_j x_{ij})^2}$$



となります。先ほどの正規分布の尤度関数と見比べて下さい。

１を最大にする回帰係数ｂ０ｂ１…ｂｍの最尤推定値は，最小２乗推定値と一致します。分散の推定値で，正規分布で示した式は

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

でしたが，

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{m}\hat{b}_j x_{ij})^2$$

$$= \frac{1}{n}S_e(m)$$

となります。

Se（ｍ）はｍ次回帰での残差平方和です。

注１）赤池 弘次，中川 東一郎：「ダイナミックシステムの統計的解析と制御」
サイエンス社（1972 年）

これを AIC の式に代入すると

$$AIC(m) = n\log\frac{S_e(m)}{n} + 2(m+1) + n(\log 2\pi + 1)$$

です。

mとは関係のない部分を除くと，

$$AIC(m) = n\log\frac{S_e(m)}{n} + 2(m+1)$$

になります。この AIC の値が最小となる統計モデル（回帰式）が，最も尤もらしい統計モデルです。

13.4　AIC による回帰モデルの選択例

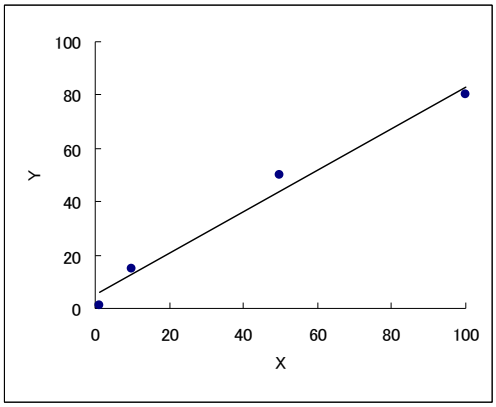
それでは，実際に簡単な例で計算してみます。

X	Y
1	1
10	15
50	50
100	80

このデータの回帰式をエクセルで 1 次回帰してみます。

エクセルの回帰分析から下記の結果が得られます。

y=0.7801x+5.1029



R=0.9904

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	3724.356	3724.356	102.5368	0.009612
残差	2	72.64428	36.32214		
合計	3	3797			

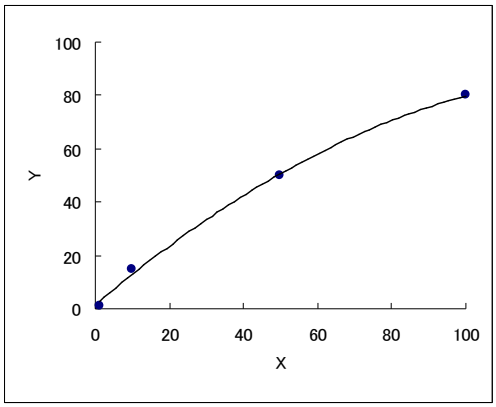
回帰次数 1 と上の表の残差の変動（平方和）72.64428 を AIC の式に代入すると

$$AIC(1) = 4\log\frac{72.64428}{4} + 2(1+1)$$
$$= 15.60$$

この１次回帰でも R＝0.9904 で良いように思えますが，２次回帰を試してみましょう。

$y=-0.004x^2+1.1878x+1.3962$

$$AIC(2) = 4\log\frac{7.548479}{4} + 2(2+1)$$
$$= 8.54$$



AIC は 15.60 と 8.54 から

$AIC(1) > AIC(2)$

で，２次回帰式 $y=-0.004x^2+1.1878x+1.3962$ を採用した方が良いことになります。

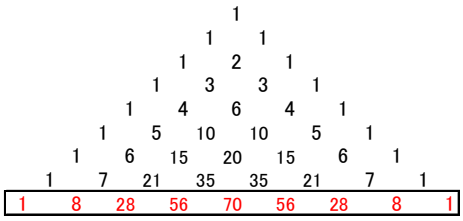
第１４章　母関数の魅了

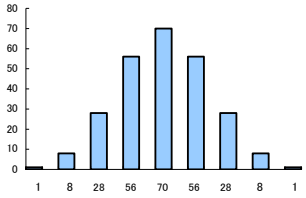
統計学のモーメント母関数は１次のモーメントとして平均値，２次のモーメントとして分散，３次のモーメントとして歪度（わいど），４次のモーメントとして尖度（せんど）を生み出す関数です。歪度と尖度は正規性の検定や確認のためによく使用します。母関数，モーメント，モーメント母関数の順で説明します。

この母関数の概念について，パスカルの三角形，フィボナッチ数列，黄金比などを例として説明します。母関数は非常に強力な数学手法です。

14.1　母関数

よく知られているようにパスカルの三角形は下記のものです。





この数列は統計学では２項分布として表れます。正規分布に近づくことは「第８章　正規分布について」でも述べました。この数列を生み出す関数である，母関数があります。

中学校で $(1+x)^2 = 1+2x+x^2$ を習いますが，

$(1+x) = 1+x$

$(1+x)^2 = 1+2x+x^2$

$(1+x)^3 = 1+3x^2+3x^3+x^4$

…

から $(1+x)^n$ の係数がパスカルの三角形の横の数列に一致していることに気が付きます。

$(1+x)^n$ はパスカルの三角形の横の数列を生成する「**母関数**」です。つまり，パスカルの三角形の何段目の数列でも $(1+x)^n$ から計算（生み出せる）できます。

$$(1+x)^n = 1 + nx + \frac{1}{2}n(n-1)x^2 + \ldots + x^n$$

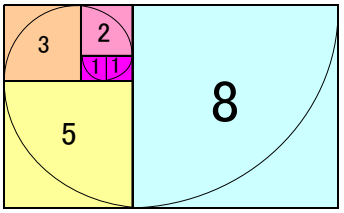
よく見るとｘのベキ乗の多項式になっている点を注意してください。

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots$$

さて，フィボナッチ数列は１，１，２，３，５，８，１６，…です。
１つ前と２つ前の数を加えることで生成します。（F_1＝１とする。）

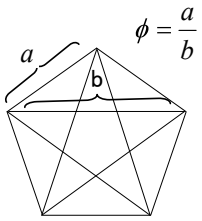
F_1　　＝１
F_2＝０＋１＝１
F_3＝１＋１＝２
F_4＝１＋２＝３
F_5＝２＋３＝５
F_6＝３＋５＝８
…

で $F_n = F_{n-1} + F_{n-2}$ です。



この数列は自然界でよく現れます。例えば，ひまわりの種の配列，オオムガイの殻などが有名です。また，フィボナッチ数列の隣り合う数の比

$\frac{1}{1}, \frac{1}{2}, \frac{2}{3}, \frac{3}{5}, \frac{5}{8}, \frac{8}{13}, \ldots$ は，黄金比の $\phi = \frac{\sqrt{5}-1}{2} \approx 0.618$ になっていきます。黄金比は芸術的に美しいとされる比です。



五芒星の対角線と１辺の比も黄金比です。

黄金比 ϕ は $x^2 - x - 1 = 0$ の解であることより，連分数表示にすると

$$\phi = \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1} \ldots}}}} \approx \frac{1}{0.618}$$

であり，

$$\phi = \sqrt{1 + \sqrt{1 + \sqrt{1 + \sqrt{1 + \ldots}}}} \approx \frac{1}{0.618}$$

の表示も有名で，黄金比は単数１が並び，0.618…になる少し不思議で綺麗な形になります。

1	1	2	3	5	8
1	1	1	1		
1	2	3	4		
1	3	6	10		
1	4	10	20		
1	5	15	35		
1	6	21			

ところで，フィボナッチ数列は，パスカルの三角形の中にも斜めに現れます。

パスカルの三角形で母関数は $(1+x)^n$ であることは示しました。このフィボナッチ数列の母関数はどのようなものであるかの話をします。

パスカルの三角形の母関数はベキ展開でした。

$$f(x)=a_0+a_1x+a_2x^2+a_3x^3+\ldots$$

フィボナッチ数列もベキ展開できないかを考えます。この母関数の考えにより，

$$\frac{1}{1-x-x^2}=1+1x+2x^2+3x^3+5x^4+\ldots+F_nx^n+\ldots$$

となります。

母関数はベキ級数表示であり

$$f(x)=a_0+a_1x+a_2x^2+a_3x^3+\ldots$$

です。

母関数は組合わせ理論においては，各係数の組合わせ的解釈をすることにより，計算を容易にします。簡単な例を示してみますが，式の理解には離散数学の本などを参考にして下さい。[注２）]

例えば，10 円，100 円，500 円貨幣が多量にあり，10 円を a，100 円を b，500 円を c とし，各貨幣は０個か１個を選べるとします。この母関数は

$f(x)=(1+ax)(1+bx)(1+cx)$　　となり，展開すると

$f(x)=1+(a+b+c)x+(ab+bc+ac)x^2+abcx^3$

となります。

係数の部分を見ると，$(a+b+c)x$ から１個選ぶ選び方は 10 円か 100 円か 500 円かの３通り，$(ab+bc+ac)x^2$ から２個選ぶ選び方は３通り，$abcx^3$ から各１個ずつ選ぶ選び方は１通りであることが見て取れます。

注１）　佐藤修一「自然にひそむ数学―自然と数学の不思議な関係」講談社〈ブルーバックス〉，1998 年

注２）　大山達雄「パワーアップ　離散数学」共立出版，1997 年

Joseph H.Silverman(鈴木治郎　訳)「はじめての数論」ピアソン・エデューション，2001 年

「第 10 章　誤差伝播の法則」の注）のテイラー展開で述べた $|x|<1$ で

$$\frac{1}{1-x}=1+x^2+x^3+\ldots$$

が成り立ちますが，自然数 0,1,2,3,4,...の母関数は

$$\frac{x}{(1-x)^2}=x+2x^2+3x^3+\ldots$$　　で，係数が自然数です。

数学用語として，「組み合わせ」でなく「組合わせ」が使用されますが，どちらでも良い。

14.2　統計学のｎ次のモーメントとは

それでは，統計学のモーメント母関数について述べていきます。考え方，概念は先ほどの母関数の話と同じです。つまり，統計学のモーメントを生成します。
推定法と使用される**モーメント法**のｋ次のモーメントとは何かから述べます。モーメントは物理学のモーメントの概念からきています。
平均（ｘの期待値）は[注1)]

$$E[x] \equiv \int_{-\infty}^{\infty} xf(x)dx$$

で，分散は

$$V[x] \equiv E\left[(x-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$$

です。
平均

$$E[x]$$

のｘの指数部分に注目して１次なので，原点のまわりの１次のモーメントと呼びます。

同様に$E\left[x^2\right]$は原点のまわりの２次のモーメントで，

$E\left[x^k\right]$

は，**原点のまわりのｋ次のモーメント**と定義します。
分散は

$$V[x] = E\left[(x-\mu)^2\right] = E\left[x^2\right] - E[x]^2$$

ですから，平均μのまわりの２次のモーメントと呼びます。
同様に

$E\left[(x-\mu)^k\right]$

から，**平均のまわりのｋ次のモーメント**を定義します。

このようにして３次のモーメント，４次のモーメントが計算できることになります。

注１）期待値についてついては本文「第２章　平均値の計算方法を考える」で説明しました。

データが正規分布ならば,原点のまわりの１次のモーメントである平均は位置に関係し，平均のまわりの２次のモーメントである分散は幅，広がりに関係し，平均のまわりの３次のモーメントは分布の形の歪み方に関係し，平均のまわりの４次のモーメントは分布の形の尖り方（実際は裾の広がり度）に関係します。

平均のまわりのｋ次のモーメントは

$$\mu_k' = \int_{-\infty}^{\infty} (x-\mu)^k f(x)dx$$

で計算します。

そこで，歪み方を示す平均のまわりの３次のモーメントは平均に関して対称ならば$\mu_3 = 0$です。標準偏差の３乗σ^3で割って規格して，**歪度**として計算します。

$$\beta_1 = \frac{\mu_3}{\sigma^3}$$

尖度は平均のまわりの４次のモーメントを使用して

$$\beta_2 = \frac{\mu_4}{\sigma^4} - 3$$

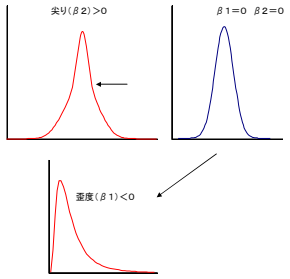
で計算します。

－３してあるのは，正規分布ならば

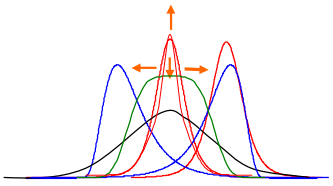
$$\frac{\mu_4}{\sigma^4} = 3$$

になり，指標として０にするためで，－３しない流儀もあります。

まとめると下記の表のようになり，確率密度関数の分布形状に関係するものが，統計学のモーメントです。



１次のモーメント	平均μ	位置
２次のモーメント	分散σ^2	幅
３次のモーメント	歪度β_1	歪みの程度
４次のモーメント	尖度β_2	尖りの程度 （裾の広がり度）



尖度，歪度の実際の計算例は「第１８章　正規分布に変換する方法」で示します。

14.3　モーメント母関数

モーメントの説明をしましたので，モーメント母関数を説明します。

再度，目標とする母関数の多項式を示します。

$$f(x)=a_0+a_1x+a_2x^2+a_3x^3+\ldots$$

で，係数はテイラー展開では[注1)]

$$a_n=\frac{f^{(n)}(x)}{n!}.$$

です。この関数を求めます。

e^x のベキ展開（テイラー展開）[注2)] は上記式より

$$e^x=1+\frac{x}{1}+\frac{x^2}{2!}+\frac{x^3}{3!}+\ldots$$

が求まります。この展開式はよく知られています。

指数型母関数である統計学のモーメント母関数を M（θ）として，

モーメント母関数を

$$M(\theta)\equiv E\left[e^{\theta x}\right]$$

と定義します。

$e^{\theta x}$ をベキ展開すると e^x と同様に，指数部分の x を θ x に変えただけなので，

$$e^{\theta X}=1+\frac{\theta x}{1}+\frac{(\theta x)^2}{2!}+\frac{(\theta x)^3}{3!}+\ldots..$$

となることは上の式より解ります。

このことよりモーメント母関数は

$$M(\theta)=E\left[e^{\theta x}\right]=E\left[1+\frac{\theta}{1}x+\frac{\theta^2}{2!}x^2+\frac{\theta^3}{3!}x^3+\ldots..\right]$$

$$=1+\frac{\theta}{1}E[x]+\frac{\theta^2}{2!}E\left[x^2\right]+\frac{\theta^3}{3!}E\left[x^3\right]+\ldots$$

となります。

平均は E[x]です。

注 1）テイラー展開については「第 10 章　誤差伝播の法則」の参考**テイラー展開**を読んでください。

注 2）x ＝ 0 の場合のベキ指数展開をマクローリン展開と呼ぶ流儀もあります。

また，分散は

$$V[x]=E[(x-\mu)^2]=E[x^2]-E[x]^2$$

ですから，原点のまわりの２次のモーメント E[ｘ^2]から E[ｘ]^{2}を引いたものです。
それでは，このモーメント母関数から平均μと分散σ^{2}を抽出してみます。

$$M(\theta)=1+\frac{\theta}{1}E[x]+\frac{\theta^2}{2!}E[x^2]+\frac{\theta^3}{3!}E[x^3]+\ldots$$

から，E[X]を抽出するには M（θ）をθで微分して

$$M'(\theta)=E[x]+\frac{\theta}{1!}E[x^2]+\frac{\theta^2}{2!}E[x^3]+\ldots$$

で，θ＝０を代入すると

$$M'(0)=E[x]$$

になります。これで E[ｘ]が抽出できました。
また，分散は２階微分してθ＝０を代入すると

$$M''(0)=E[x^2]$$

で，分散が抽出されます。
以下同様にｋ回微分して，いつでも原点のまわりのｋ次のモーメントが得られます。

$$\mu_k'=\frac{d^k}{d\theta^k}M(\theta)\Big|_{\theta=0}=E[x^k]$$

それでは，実際に正規分布で平均と分散を求めてみます。
正規分布の確率密度関数は

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

ですから，正規分布 N(μ, σ^2)のモーメント母関数は

$$M(\theta)=E[e^{\theta x}]=\int_{\infty}^{-\infty}e^{\theta x}f(x)dx$$

$$=\frac{1}{\sqrt{2\pi\sigma}}\int_{\infty}^{-\infty}e^{\theta x}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx$$

$$=e^{\mu\theta+\frac{1}{2}\sigma^2\theta^2}$$

と求められます。

$$M(\theta) = E\left[e^{\theta x}\right] = e^{\mu\theta + \frac{1}{2}\sigma^2\theta^2}$$

からＭ（θ）を１回微分してθ＝０を代入すると

$$M'(0) = E[x] = \mu$$

１次のモーメントが平均で，さらに微分して

$$M''(0) - M'(0)^2 = E\left[x^2\right] - E[x]^2 = V[x] = \sigma^2$$

で平均のまわりの２次のモーメントの分散が求まります。

モーメント母関数と確率密度関数は１：１で対応していますので，モーメント母関数が同じならば，確率密度関数も同じです。

正規分布のベキ展開であるモーメント母関数を説明しまいたが，確率密度関数の合成なども直接たたみこみ積分などで求めるよりも，モーメント母関数で求めた方が簡単です。正規分布に従う N(μ_1, σ_1^2) と N(μ_2, σ_2^2)の X＋Y の分布を求める場合を示します。

$$M_X(\theta) = e^{\mu_1\theta + \frac{1}{2}\sigma_1^2\theta^2}$$

$$M_Y(\theta) = e^{\mu_2\theta + \frac{1}{2}\sigma_2^2\theta^2}$$

から

$$M_{X+Y}(\theta) = M_X(\theta)M_Y(\theta) = e^{\mu_1\theta + \frac{1}{2}\sigma_1^2\theta^2} e^{\mu_2\theta + \frac{1}{2}\sigma_2^2\theta^2}$$

$$= e^{(\mu_1+\mu_2)\theta + \frac{\theta^2}{2}(\sigma_1^2+\sigma_2^2)}$$

より，N($\mu_1 + \mu_2$，$\sigma_1^2 + \sigma_2^2$)の正規分布に従うことが簡単に求まります。

つまり，X＋Y の平均は X の平均と Y の平均を加え$\mu_1 + \mu_2$で求まり，分散は X の分散と Y の分散を加え$\sigma_1^2 + \sigma_2^2$で求められます。

このことは「8.2　中心極限定理」や「第 10 章　誤差伝播の法則－不確かさの計算－」でも述べました。

参考　**特性母関数とキュムラント**

モーメント母関数と同様なものに**特性関数**（characteristic function）があります。

オイラーの公式

$e^{i\theta x} = \cos\theta x + i\sin\theta x$　　$\cos\theta x \le 1, \sin\theta x \le 1$

$i = \sqrt{-1}$

から

$$E(\cos\theta x) = \int_{-\infty}^{\infty} \cos\theta x f(x)dx$$

$$E(\sin\theta x) = \int_{-\infty}^{\infty} \sin\theta x f(x)dx$$

$$E(e^{i\theta x}) = E(\cos\theta x) + iE(\sin\theta x)$$

$$\varphi(\theta) = E(e^{i\theta x}) = \int_{-\infty}^{\infty} e^{i\theta x} f(x)dx = \int_{-\infty}^{\infty} (\cos\theta x) f(x) + i\int_{-\infty}^{\infty} (\sin\theta x) f(x)dx$$

この$\varphi(\theta)$を特性関数と呼びます。この特性関数はモーメント母関数のθを$i\theta$に変えたものです。

確率密度関数のフーリエ変換になっています。

$$\varphi(\theta) = \int_{-\infty}^{\infty} e^{i\theta x} f(x)dx$$

フーリエ変換なのでフーリエ逆変換して元の関数を求めることができます。

フーリエ逆変換は

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\theta x}\varphi(\theta)dx$$

です。

モーメント母関数と同様にテイラー展開できて

$$\varphi(\theta) = 1 + E[x]\frac{i\theta}{1} + E[x^2]\frac{i\theta^2}{2!} + E[x^3]\frac{i\theta^3}{3!} + \ldots$$

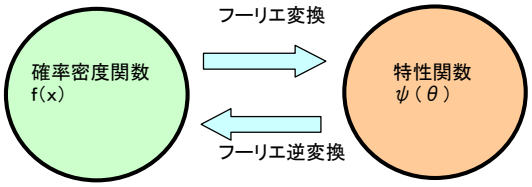
です。

正規分布の特性関数は

$$\varphi(\theta) = e^{i\mu\theta + \frac{1}{2}\sigma^2\theta^2}$$

となります。

確率分布の特性関数は必ず存在しますが，モーメント母関数は存在するとはかぎりません。この意味において特性関数の方が優れています。



さらに，特性関数の対数のテイラー展開した各項の係数を**キュムラント**(cumulant)といいます。

$$\log\varphi(\theta)=\kappa_1\frac{i\theta}{1}+\kappa_2\frac{i\theta^2}{2!}+\kappa_3\frac{i\theta^3}{3!}+\ldots$$

$$\kappa_n=\left[\frac{(-i)^n d^n\log\varphi(u)}{du^n}\right]_{u=0}$$

n 次のモーメントをm_nとすると，キュムラントとモーメントと関係は

$\kappa_1=m_1=\mu$
$\kappa_2=m_2$
$\kappa_3=m_3$
$\kappa_4=m_4-3m_2^2$

です。正規分布の場合，３次以上のキュムラントは全て０になります。

母関数からモーメント，モーメント母関数，特性関数，キュムラントまでの概要を説明しました。モーメント母関数，特性関数から微分するだけでいつでも平均や分散が求まり，確率密度関数とモーメント母関数，特性関数は１：１に対応しますので，モーメント母関数や特性関数で計算した方が簡単になることが多くあります。

第１５章　測定精度の推定方法（併行，日間精度）

併行精度（同時再現性）や日間精度の算出で不適切な発表や論文がよくあります。

ISO15189 や厚生労働省の残留農薬，環境検査では「試験の妥当性」で枝分かれ分散分析法が採用されています。枝分かれ分散分析法や REML 法（制限付き最尤法）により精度を計算すべきです。本文の理解には分散分析法の基礎知識が必要で，REML 法の理解には最尤法や一般化線形モデルの知識が必要です。

分析精度を推定するためのデータ解析手法ですから，濃度測定を行う技術者として使いこなせる必要があります。

15.1　不適切な計算例

毎日 2 回，5 日間測定して下記のデータを得たとします。[注1)]

測定日	1	2	平均
1日目	0.0485	0.0436	0.0461
2日目	0.0512	0.0564	0.0538
3日目	0.0559	0.0587	0.0573
4日目	0.0391	0.0385	0.0388
5日目	0.0468	0.0446	0.0457

平均の平均	0.0483
s.d.	0.0073
c.v.%	15.1

各平均の値から日間精度は c.v.%＝15.1 と報告すると，不適切です。

もし 1 日 2 回ではなく，1 日 100 回の平均値を 5 日間取れば日間精度の c.v.%は小さくなると推定できます。

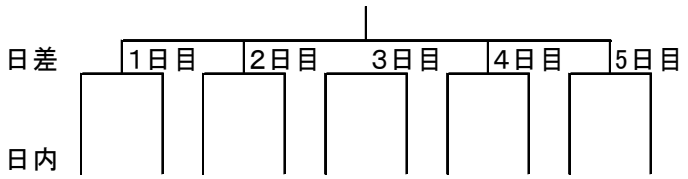
ここで計算した精度は「1 日 2 回 5 日間の平均値の精度」で，適切な日間精度ではありません。

次に述べる分散分析法や REML 法から計算します。

注１）厚生労働省　食安発 1115001 号
「食品中に残留する農薬等に関する試験の妥当性評価ガイドラインについて」で記載されている例題のデータです。

15.2　枝別れ分散分析からの計算例

「枝分かれ」とは同一の検体を小分けして数日間測定する場合は，下記のようになっています。



一元配置分散分析はエクセルでも行うことができ，下記の計算ができますが，「分散の期待値」に注意して下さい。

変動要因	平方和（変動）	自由度	分散	分散比	分散の期待値
日差	S_{RW}	j−1	V_{RW}	$F=V_{RW}/Ve$	$\sigma e^2+n\sigma_{RW}^2$
併行（日内、同時再現性）	Se	j(n−1)	Ve		σe^2
	S_T	jn−1			

j：測定した日数　n：日内の測定数

実際の数値で計算方法を確認します。

先ほど示した表と同じ数値です。

測定日	1	2	平均
1日目	0.0485	0.0436	0.0461
2日目	0.0512	0.0564	0.0538
3日目	0.0559	0.0587	0.0573
4日目	0.0391	0.0385	0.0388
5日目	0.0468	0.0446	0.0457

分散分析の知識が無くても，上記のデータをエクセルの分析ツールの一元配置分散分析で計算すると下記の結果が得られます。

分散分析表						
変動要因	変動	自由度	分散	観測された分散比	P−値	F 境界値
日差	0.000427	4	0.000107	16.64206584	0.00429	5.192168
併行	3.2E−05	5	6.41E−06			
合計	0.000459	9				

日差には先ほどの「分散の期待値」を見ると併行精度が混ざっています。

$\sigma_e^2 + n\sigma_{RW}^2$

そこで，「分散の期待値」の式から併行精度と日間差は次のように計算します。[注1)]

併行再現性（同時再現性，日内再現性）：σｅ（併行標準偏差）

$$\hat{\sigma}_e = \sqrt{V_e} = \sqrt{6.41E-06} = 0.00253$$

日間差再現性：σ RW（日間標準偏差）

$$\hat{\sigma}_{RW} = \sqrt{V_{RW} - V_e / n} = \sqrt{(0.000107 - 6.41E-06)/n} = 0.00708$$

先ほど不適切な日間精度は c.v.%=15.1 でしたが，分散分析からの推定値は

$$c.v.\% = \frac{\sigma_{RW}}{\bar{x}} = \frac{0.00708}{0.0483} \times 100 = 14.6$$

となります

室内再現性：（室内標準偏差，総合精度）

同一施設内において，試験日，試験実施者，器具，機器等を変えて測定する場合の精度のことです。「第 10 章　誤差伝播の法則－不確かさの計算－」から

$$\sqrt{\hat{\sigma}_e^2 + \hat{\sigma}_{RW}^2} = 0.00752$$

となり，測定施設の精度を表す値は室内再現性です。

つまり，1 検体のデータに対して，日を変えてもどの程度の精度で測定できているのかを室内精度（総合精度）が表しています。

$$c.v.\% = \frac{\sqrt{\hat{\sigma}_e^2 + \hat{\sigma}_{RW}^2}}{\bar{x}} = \frac{0.00752}{0.0483} \times 100 = 15.0$$

この場合，室内精度は c. v. %=15. 0 です。

注１）厚生労働省　食安発 1115001 号

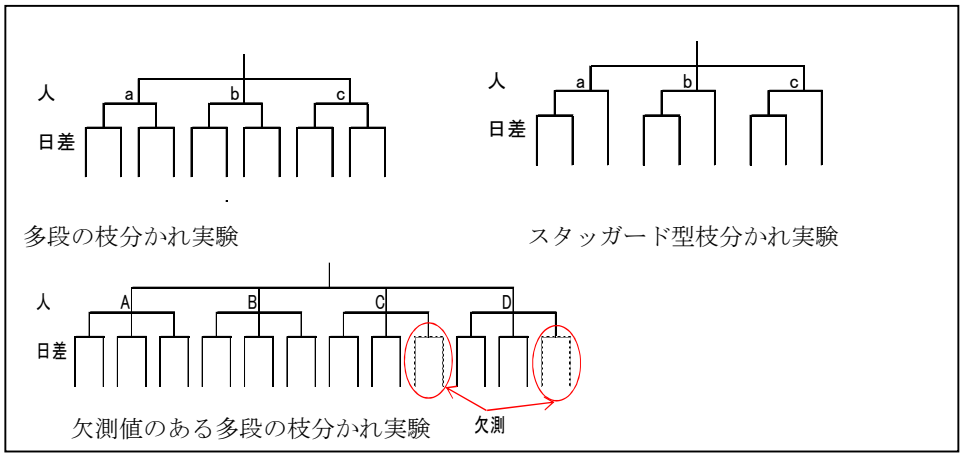
「食品中に残留する農薬等に関する試験の妥当性評価ガイドラインについて」

に記載されている分散分析法です

15.3　制限付き最尤法（ＲＥＭＬ法）での計算

分析現場のデータは，欠測値などで非釣り合い型データ(unbalanced data)や，効率の良いスタッガード型実験や，分析機器や人の差などの誤差要因の解析のために多段の枝分かれ実験になることがあります。このようなデータでは，厚生労働省のガイドラインに示されている古典的な分散分析法では計算が複雑です。

そこで，制限付き最尤法（REML 法. restricted maximum likelihood）での解析が有効です。



REML 法で計算出来る実験モデルの例

REML 法による計算は統計ソフト JMP や R で行えます。

また，下記のインターネットに，高橋行雄のエクセルでの詳細な説明が記載されていますので，エクセルや数式処理ソフトでも計算可能です。

http://www.sascom.jp/download/pdf/usergroups11_A-12.pdf#search=%27REML%E6%B3%95%27

方法の解説は下記の参考 1 文献 1）にあります。

また，実際の例は下記の参考 1 文献 2）を読んで下さい。

また，参考 2 に簡単な基本となる式と実際のデータでの解析例を示します。

参考 1　**参考文献**

1）南美穂子：「制限付き最尤推定法（REML 推定法）」応用統計学 Vol.25，Ｎｏ.2，73-78，1996

2）秋山功　他：「残留農薬分析の制限付き最尤法(REML 法)による妥当性の評価」第 108 回　日本食品衛学会，2014

参考２**制限付き最尤法（REML 法）**とは

制限付き最尤法（REML 法）は，線形混合モデルの分散の推定に使用されます。

$$y = X\beta + Zu + e$$

y はデータのベクトル，X と Z は計画行列，β は固定効果ベクトル，u は変量効果ベクトルとして，u と e は互いに独立で平均 0 の多変量正規分布に従うとする。データ y に関する尤度 f(y)と固定効果の尤度 g(β)の比を最大化する尤度比から各分散を推定します。

$$L = \frac{f(y)}{g(\beta)}$$

対数尤度は下記の式になります。

$$-2\ln L = -2\times\left(\ln f(y) - \ln g(\beta)\right)$$

1　完備型データでの分散分析法と REML 法の結果の比較

ホルムアルデヒドで，厚生労働省のガイドラインに従い，5 日間 2 回のデータから計算した分散分析法と REML 法の結果を示します。

ホルムアルデヒド
超純水に8.0ppb添加　単位：ppb

1日目	8.69	9.00
2日目	7.83	7.99
3日目	7.82	7.91
4日目	7.84	7.92
5日目	8.00	8.12

分散分析法での解析結果

分散分析表

要因	平方和	自由度	平均平方	F比	p値
日差	1.39126	4	0.3478	23	0.002
併行	0.0753	5	0.0151		
全体	1.46656	9	0.1630		

	s.d.	RSD(%)
室内精度	0.426	5.3
日間差精度	0.408	5.0
併行精度		
（同時再現性）	0.123	1.5
全体の精度	0.404	5.0

REML 法での解析結果

REML法による分散成分推定値

変量効果	分散比	分散成分	標準偏差
日間差精度	11.05	0.16638	0.408
併行精度		0.01506	0.123
全体		0.18144	0.426
-2対数尤度=		2.64058	

	s.d.	RSD(%)
室内精度	0.426	5.3
日間差精度	0.408	5.0
併行精度		
(同時再現性)	0.123	1.5
全体の精度	0.404	5.0

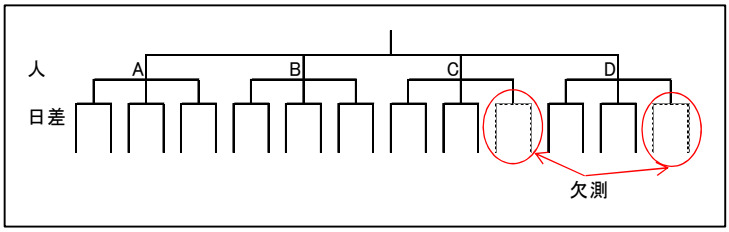
以上の結果から，完備型であるデータからの推定精度は，古典的な分散分析法と REML 法の推定精度は同じ結果が得られます。このため，古典的な分散分析法から REML 法に移行するのに問題はありません。

2　多段の非釣り合い型データでの REML 法での推定の例

単位 : ppb

	A	B	C	D
1日目	15.0	14.8	14.6	14.6
	14.9	15.0	14.2	14.5
2日目	14.8	14.3	14.9	14.2
	14.4	14.3	14.3	14.0
3日目	14.9	15.3	-	-
	15.2	15.2	-	-

ホルムアルデヒド　　天然水に15ppb添加
4人で3日間　　GC/MSで測定



REML法による分散成分推定値

変量効果	分散比	分散成分	標準偏差s.d.	標準誤差s.e.	百分率%	変動係数RSD%
人	0.36	0.01603	0.127	0.0617	10.3	0.86
日間[人]	2.16	0.09512	0.308	0.0694	61.3	2.10
併行精度		0.04400	0.210	0.0197	28.4	1.43
室内精度		0.15514			100	2.68
-2対数尤度=		13.48661				

欠測値があっても，REML 法では適切な推定値が得られます。

参考文献

秋山功　他：「水質検査の制限付き最尤法（REML 法）による妥当性の評価」埼玉県環境計量協議会
第 34 回研究発表会，2016 年

第１６章　検量線の重み付きと回帰式の選択方法

濃度測定などにおける検量線では，重み付き回帰なども使用されます。検量線を直線にするのか曲線にするのか，重み付け変数をどのようにするのかなどについて迷うことがあります。採用した検量線により精確さが左右されます。濃度測定においては精確さに影響するのであるから，なぜその検量線の回帰モデルを選択したかの理由を明確にする必要があります。

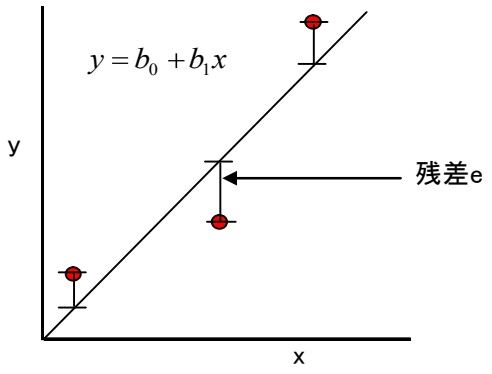
測定現場で日常的に使用する重み付き回帰式は，入門的な統計学の本ではあまり記載されていませんので，重み付き回帰の簡単な説明と，重み付けの選択と直線回帰なのか曲線回帰なのかの選択方法，回帰式の信頼区間などについて触れます。

式が多く，電卓で計算できるレベルではありませんが，検量線は分析機器では自動で計算してくれます。数式は多くなりますが，自動で計算してくれる内容を理解することは，分析機器のデータを正しく理解し，使いこなすことになり，正しいデータを報告することになります。

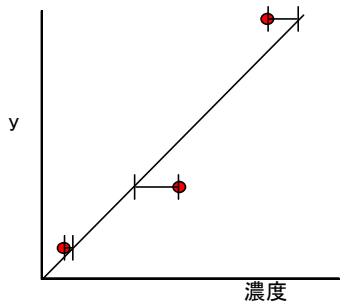
16.1　回帰式について

横軸（ｘ）は独立変数として濃度などバラツキが無いものにして，縦軸（ｙ）は従属変数としてバラツキのある測定値などにして，データをプロットします。慣例としてｘを横軸，ｙを縦軸にします。

回帰式は下記の図に示すように，プロットした点と回帰直線（または曲線）の縦方向の差（残差）を最小にする線です。

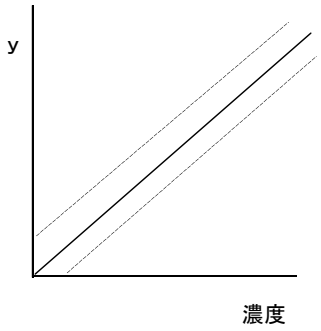


ｘとｙを入れ替えて回帰式を計算するとｘに誤差があるとして回帰するので，下記の図のように残差の向きが変わり，別の回帰式になりますので注意して下さい。[注1）]

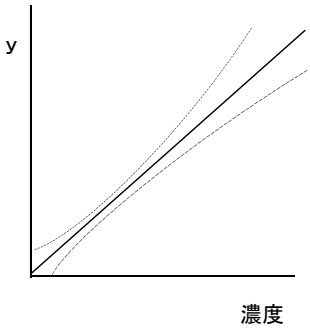


エクセルなどでは縦軸にバラツキがあるとして回帰式を算出していますので，検量線などでは横軸は濃度で縦軸は測定値にする必要があります。

濃度測定の検量線にエクセルを使用すると回帰式は，下記の左の図のように全濃度で誤差が一定として回帰式を算出しています。



全濃度域で誤差が一定

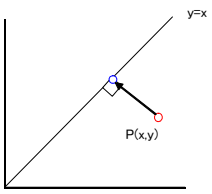


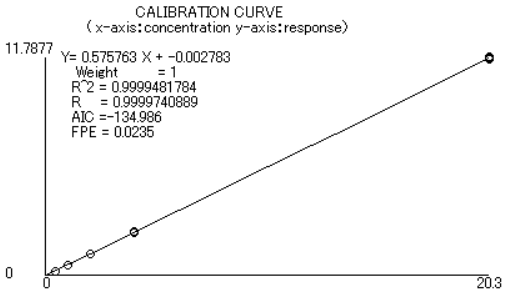
濃度で誤差が異なる

しかし，通常の測定はバラツキを調べると，右上の図のように高濃度になるほどバラツキが大きくなります。このためバラツキの変化を考慮した「重み付き回帰式」による検量線を使用する場合があります。

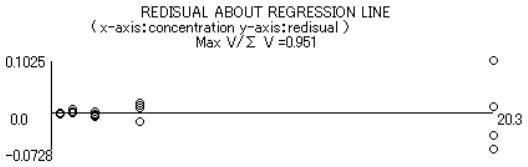
注１）平均値のところ述べたｙ＝ｘへの垂線を求める方法も最小２乗法ですが，このようにして求める場合は線形関係式（直交回帰法）と呼びます。

私たちが通常の検量線で使用する方法とは異なります。多くの検量線はｙに誤差があるとして，残差の２乗の和を最小にする線です。





検量線



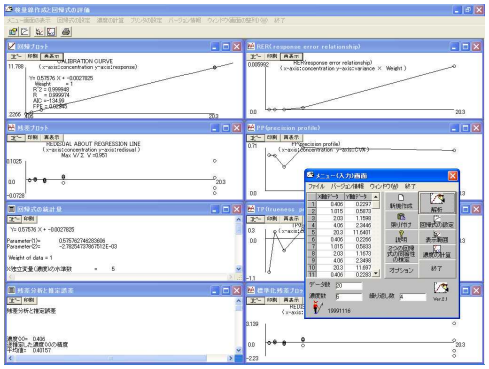
残差プロット[注1)]

重み付き回帰式は重みを一定にすれば，表計算ソフトなどで計算できる通常の回帰式と同じになります。つまり重み付き回帰式は，重みの無い回帰式を含みます。重み付き回帰は通常の回帰を一般化した形になっています。

実際の重み付き回帰は，通常分析機器のデータ処理ソフトで計算出来ます。[注2)]

注１）一般の残差プロットとは異なります。残差プロットは本来，横軸に回帰推定値 y_i をとり，縦軸に基準化した残差 d_i を取ったのもを言います。濃度 x_i に対する重み付きを考察するために横軸を x_i にしてあります。

注２）本文中の回帰式の作成，解析は Microsoft Visual Basic　6.0　で自作したプログラムを使用しました。



16.2　最小２乗法と最尤法について

さて，重み付き最小２乗法の計算式から述べます。
最小２乗法は先ほど説明した残差の２乗和を最小とするものです。

つまり，直線回帰（1 次回帰）なら $f(x) = b_0 + b_1 x$ ですし，

高次回帰ならば，$f(x) = b_0 + b_1 x + b_2 x^2 + \ldots + b_n x^n$ ですが，残差 e の２乗和を S_e とすれば

$$S_e = \sum e_i^2 = \sum (y_i - f(x_i))^2$$

となり，この S_e を最小にするのが最小２乗法です。

各 y_i の信頼性，精度などが異なる場合は重みを付けます。重みを wi として，重み付けると，

$$S_e = \sum w_i (y_i - f(x_i))^2$$

となります。
これはデータを

$$\sqrt{w_i}\, y_i = \sqrt{w_i}\, f(x_i) + \sqrt{w_i}\, e_i$$

として回帰するのと同じです。

「第 13 章　最尤法について」で説明した最尤法から，最適な重みの掛け方を考えてみます。

はじめに求める回帰式は直線回帰式 $f(x)=b_0+b_1x$ ですが，高次回帰でも同じです。

まず,測定誤差は正規分布すると考えられます。正規分布の確率密度関数は以下の式です。

$$P(y)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(y-\mu)^2/2\sigma^2}$$

μ は真値で y は観測値， σ は標準偏差， π は円周率の 3.14…で，e は自然対数の底で 2.718...です。

観測値の $y_{1,...,}$ y_n の誤差は正規分布に従うと仮定します。

正規分布の確率密度関数に，求める回帰式を $f(x)=b_0+b_1x$ を代入すると

$$P(y_i)=\frac{1}{\sigma_i\sqrt{2\pi}}e^{-(y_i-b_0-b_1x_i)^2/2\sigma_i^2}$$

になります。

観測値の $y_1,\ldots, y_n$ の確率は積として求められるので,

$$l=P(y_1).P(y_2)...P(y_n)=\prod_{i=1}^{n}\frac{1}{\sigma_i\sqrt{2\pi}}e^{-(y_i-b_0-b_1x_i)^2/2\sigma_i^2}$$

となり，最尤法で説明したように，対数をとると

$$L=\log l=-\frac{n}{2}\log(2\pi)-n\log\sigma-\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}(y_i-b_0-b_1x_i)^2$$

となります。

余分な部分は取り除くと

$$Sew=\sum_{i=1}^{n}\frac{1}{\sigma_i^2}(y_i-b_0-b_1x_i)^2$$

となります。これは χ^2 分布に従います。[注1)] 重みを w_i とすると，最適な重みは上記式より

$$w_i=\frac{1}{\sigma_i^2}$$

となり，各濃度の y の分散の逆数で重み付ければ良いことになります。

注 1） χ^2 分布

$$\begin{cases} f_v(\chi^2)=\dfrac{1}{2^{v/2}\Gamma\left(\dfrac{v}{2}\right)}\left(\chi^2\right)^{v/2-1}e^{-\chi^2/2} & \chi^2>0 \\ f_v(\chi^2)=0 & \chi^2\le 0 \end{cases}$$

先ほどの Sew の式の１／σ i^2 を wi として展開すると

$$Sew = \sum w_i y_i^{\,2} + \sum w_i b_0^{\,2} + (2\sum w_i x_i) b_0 b_1 + (\sum w_i x_i^{\,2}) b_1^{\,2} - (2\sum w_i y_i) b_0 - (\sum w_i y_i x_i) b_1$$

となります。回帰係数の b 0，b 1 を求めるのは，偏微分して，結果を 0 とおけばよいので

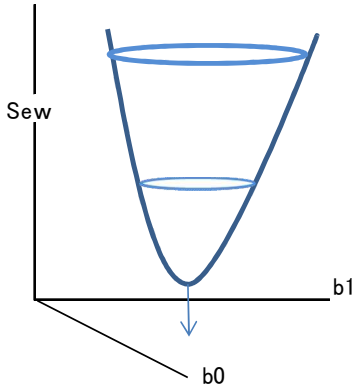
$$\frac{\partial L}{\partial b_0} = 2\sum w_i b_0 + 2\sum w_i x_i b_1 - 2\sum w_i y_i$$

$$\frac{\partial L}{\partial b_1} = 2\sum w_i b_0 + 2\sum w_i x_i^2 b_1 - 2\sum w_i y_i x_i$$

$$b_0 \sum w_i + b_1 \sum w_i x_i - \sum w_i y_i = 0$$

$$b_0 \sum w_i x_i + b_1 \sum w_i x_i^2 - \sum w_i y_i x_i = 0$$

となります。



$$b_0 \sum w_i + b_1 \sum w_i x_i = \sum w_i y_i$$

$$b_0 \sum w_i x_i + b_1 \sum w_i x_i^2 = \sum w_i y_i x_i$$

となりますが，これは正規方程式と呼ばれます。

これを解くと

$$b_1 = \frac{\sum w_i x_i y_i - (\sum w_i x_i)(\sum w_i y_i) / \sum w_i}{\sum w_i x_i^{\,2} - (\sum w_i x_i)^2 / \sum w_i}$$

$$b_o = \frac{\sum w_i y_i}{\sum w_i} - b_1 \frac{\sum w_i x_i}{\sum w_i} = \bar{y} - b_1 \bar{x}$$

となり，回帰式

$$y = b_0 + b_1 x$$

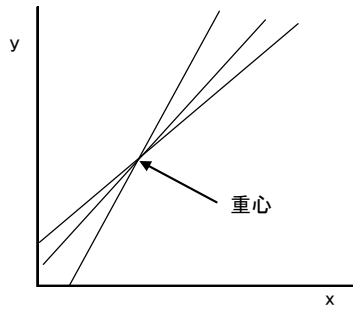
を求めることができます。

重心である平均（重み付き平均）は

$$\bar{y} = \frac{\sum w_i y_i}{\sum w_i}$$

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

で，この点を回帰式は通ります。[注1)]



重みは

$$w_i = \frac{1}{\sigma_i^{\,2}}$$

が良いことを述べましたが，正確な各濃度のｙの分散を知る必要がありますが，分散の比が判っていれば，比を重み付けに利用できます。

注１）w_iを一定にすると通常の重み無しの回帰式になります。

尚，$w_i = 1$である必要はなく，定数など一定であれば回帰式は同じになります。

重心と平均については「重み付き平均」でも述べました。

wi=1 すれば $\bar{y} = \frac{\sum y_i}{n}$ と $\bar{x} = \frac{\sum x_i}{n}$ になり，通常の平均値となります。

参考　**行列**について

和と差

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} a+e & b+f \\ c+g & d+h \end{pmatrix} \qquad \begin{pmatrix} a & b \\ c & d \end{pmatrix} - \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} a-e & b-f \\ c-g & d-h \end{pmatrix}$$

積

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \times \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}$$

連立方程式

$$\begin{cases} ax1 + bx2 = y1 \\ cx1 + dx2 = y2 \end{cases} \quad \text{は} \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \times \begin{pmatrix} x1 \\ x2 \end{pmatrix} = \begin{pmatrix} y1 \\ y2 \end{pmatrix} \quad \text{になります。}$$

行列の積の計算

A，B 行列で，一般に $AB \neq BA$ であることから，

$$(A+B)^2 = A^2 + AB + BA + B^2$$

$$(A-B)^2 = A^2 - AB - BA + B^2$$

となります。

16.3　m次回帰式の求め方

重み付き直線回帰（１次回帰）について述べたので，重み付き高次回帰式についても述べます。

回帰式

$$y = b_0 + b_1 x + b_2 x^2 + \ldots + b_m x^m$$

とします。行列で表示すると

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \qquad B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ 1 & x_3 & x_3^2 & \cdots & x_3^m \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix}$$

で，重みを

$$W = \begin{bmatrix} w_1 & & & O \\ & w_2 & & \\ & & \ddots & \\ O & & & w_n \end{bmatrix}$$

とすると，重み付き残差平方和 S は

$$S = (Y - XB)' W (Y - XB)$$

となります。

これを微分して 0 とおくと

$$\frac{\partial S}{\partial B} = -2X'WY + 2X'WXB$$

$$X'WXB = X'WY$$

となります。この正規方程式を解けば係数が求まります。

$X^{'}WXB = X^{'}WY$
を詳細に示します。

左辺は

$$X^{'}WXB = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_n^2 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_1^m & x_2^m & x_3^m & \cdots & x_n^m \end{bmatrix} \begin{bmatrix} w_1 & & & & \\ & w_2 & & 0 & \\ & & w_3 & & \\ & 0 & & \ddots & \\ & & & & w_n \end{bmatrix} \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ 1 & x_3 & x_3^2 & \cdots & x_3^m \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

$$= \begin{bmatrix} \sum w_i & \sum w_i x_i & \sum w_i x_i^2 & \cdots & \sum w_i x_i^m \\ \sum w_i x_i & \sum w_i x_i^2 & \sum w_i x_i^3 & \cdots & \sum w_i x_i^{(m+1)} \\ \sum w_i x_i^2 & \sum w_i x_i^3 & \sum w_i x_i^4 & \cdots & \sum w_i x_i^{(m+2)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum w_i x_i^m & \sum w_i x_i^{(m-1)} & \sum w_i x_i^{(m-2)} & \cdots & \sum w_i x_i^{(2m)} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

右辺は

$$X^{'}WY = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_n^2 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_1^m & x_2^m & x_3^m & \cdots & x_n^m \end{bmatrix} \begin{bmatrix} w_1 & & & & \\ & w_2 & & 0 & \\ & & w_3 & & \\ & 0 & & \ddots & \\ & & & & w_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \begin{bmatrix} \sum w_i y_i \\ \sum w_i x_i y_i \\ \sum w_i x_i^2 y_i \\ \vdots \\ \sum w_i x_i^m y_i \end{bmatrix}$$

となりますから

$$X^{'}WXB = X^{'}WY$$

は

$$\begin{bmatrix} \sum w_i & \sum w_i x_i & \sum w_i x_i^2 & \cdots & \sum w_i x_i^m \\ \sum w_i x_i & \sum w_i x_i^2 & \sum w_i x_i^3 & \cdots & \sum w_i x_i^{(m+1)} \\ \sum w_i x_i^2 & \sum w_i x_i^3 & \sum w_i x_i^4 & \cdots & \sum w_i x_i^{(m+2)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \sum w_i x_i^m & \sum w_i x_i^{(m-1)} & \sum w_i x_i^{(m-2)} & \cdots & \sum w_i x_i^{(2m)} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} \sum w_i y_i \\ \sum w_i x_i y_i \\ \sum w_i x_i^2 y_i \\ \vdots \\ \sum w_i x_i^m y_i \end{bmatrix}$$

となります。

この正規方程式からガウス消去法などで，係数のｂ$_0$からｂ$_m$を求めることができ，回帰式

$$y = b_0 + b_1 x + b_2 x^2 + \ldots + b_m x^m$$

が求まります。[注1）]

注１） $X^{'}WXB = X^{'}WY$ は $B = (X^{'}WX)^{-1} X^{'}WY$ となります。

計算のプログラムは，市販の多くの数値解析の本に記載されています。

参考：　小田 政明：「やさしい　数値計算法（２次方程式から有限要素法まで)」日刊工業新聞社（1994 年）

16.4　エクセルのソルバーで重み付き最小２乗法の原理を理解する

「16.2　最小２乗法と最尤法について」を読み返して下さい。

正規分布の確率密度関数に，求める回帰式 $f(x)=b_0+b_1x_i$ を代入すると

$$P(y_i)=\frac{1}{\sigma_i\sqrt{2\pi}}e^{-(y_i-b_0-b_1x_i)^2/2\sigma_i^2}$$

になり，この最尤値を求めることです。

結果として正規分布関数の指数部分の下記の残差の２乗を最小にすることです。

$$S_{ew}=\sum_{i=i}^{n}\frac{1}{\sigma_i^2}(y_i-b_0-b_1x_i)^2\approx\sum_{i=1}^{n}\frac{1}{x_i^2}(y_i-b_0-b_1x_i)^2$$

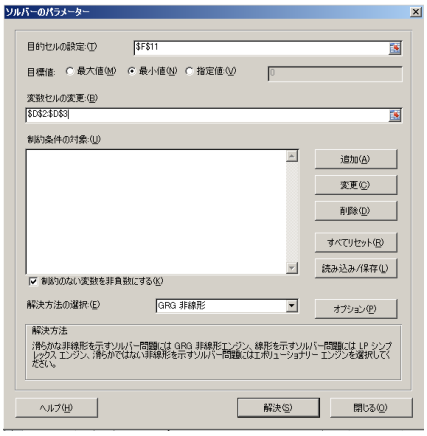
簡単に重みを w＝1/ｘ２としてエクセルのソルバーで回帰係数を求めてみます。

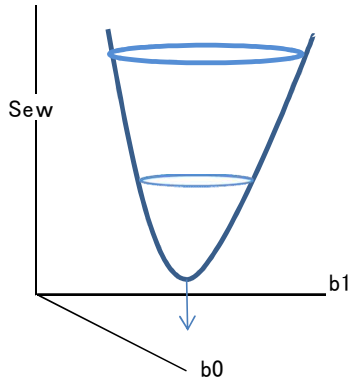
その例を下記に示します。回帰式を変えれば，エクセルや数式処理ソフトで高次回帰，非線形回帰も簡単に求められます。

さらに言えば，ソルバーが無くても残差の２乗の和が最小となる値を探せばよいので，試行錯誤で求めても良いのです。

	C	D	E	F	H
1				変化させるセル	
2	b0=	0.0859			=1/(C5*C5)
3	b1=	0.8156			
4	x	y	w=1/x^2	残差	式
5	2	1	0.2500	0.12853	=E5*(D5-($D$2+$D$3*C5))^2
6	3	3	0.1111	0.02427	=E6*(D6-($D$2+$D$3*C6))^2
7	3	4	0.1111	0.23925	=E7*(D7-($D$2+$D$3*C7))^2
8	7	6	0.0204	0.00086	=E8*(D8-($D$2+$D$3*C8))^2
9	8	4	0.0156	0.10648	=E9*(D9-($D$2+$D$3*C9))^2
10	12	10	0.0069	0.00011	=E10*(D10-($D$2+$D$3*C10))^2
11			合計	0.49950	=SUM(F5:F10)
12	y=0.0859+0.8156x				
13				最小にするセル	

ソルバー





16.5　回帰式の選択方法

１次回帰でなく曲線回帰が必要であるかは，赤池の情報量基準 AIC で調べることができます。

血中薬物濃度測定などでは定量限界付近，また，食品中残留農薬測定では基準値付近（一律基準では 0.01ppm）の精確さが大切です。高濃度の場合も希釈再検しなで測定したいものです。このため，測定可能な範囲を広げる方が有利です。

検量線の範囲を広げると，検量線は曲線になることがあります。検量線は直線回帰が一般的であり，直線と判断できる範囲で検量線を作成すべきであると考える人もいますが，自然界の現象の多くは指数関数的な変化を示すものが多く，許容できる精度が得られるならば濃度範囲を広げ曲線にした方が自然です。

AIC による統計モデルの選択は「第 13 章　最尤法について」で記載しましたが，AIC は下記の式で計算できます。

$$AIC = -2\log(L) + 2p$$

L は最大尤度で，p はパラメータ数で，パラメータ数をペナルティとして AIC の小さいモデルを最適なモデルとして選択します。

高次回帰の AIC は，回帰次数m，データ数n，残差平方和 $Se_{(m)}$ とすると，

最大対数尤度 MLL(最大となる対数尤度)は

$$MLL = -\frac{n}{2}(\log Se_{(m)} / n)$$

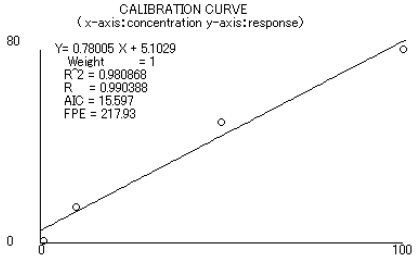
であることから

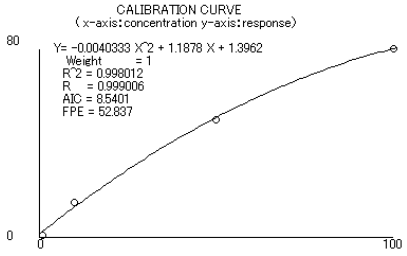
$$AIC = -2MLL + 2(m+1)$$

$$= n\log\frac{Se_{(m)}}{n} + 2(m+1)$$

となります。「第 13 章　最尤法について」のところで記載したデータと同じですが，再度記載しておきます。

X	Y
1	1
10	15
50	50
100	80





１次回帰式の AIC＝15.6 と２次回帰式の AIC＝8.5 から２次曲線が適合しています。

他にも回帰式の適合を調べる方法は考えられます。例えば，回帰式で各プロットの位置が何個連続して上または下に位置したかを調べ，検定する方法などもあります。

16.6　重み付けの選択方法

16.6.1　等分散性

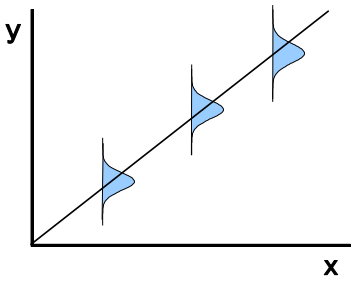
最小２乗法の回帰式では誤差が

１）等分散性　$V(\varepsilon_i)=\sigma$　　誤差ε_iの母分散は全て等しい。

２）不偏性　　$E(\varepsilon_i)=0$　　誤差ε_iの期待値は０である。

３）無相関性　誤差ε_iは互いに無相関である。

４）正規性　　誤差ε_iは正規分布に従う。



の仮定を満たす必要がありますが，検量線では１）の等分散性が満たされないことがほとんどです。そこで，重み付けにより等分散にします。

最尤法から$w_i=\dfrac{1}{\sigma_i^{\,2}}$で，各濃度のｙの分散$\sigma_i^2$の逆数を重み付けるべきであると述べました。

しかし，各濃度の分散ではなく，多くの分析機器では１／ｘまたは１／ｘ^{2}がよく使用されます。各濃度のｙの分散の比が濃度の１／ｘまたは１／ｘ^{2}に比例していれば，重みとして使用できます。低濃度に重みがかかり，１／ｘ^{2}の方が重み付けが大きいといえます。[注1)]

重み付きをどの様にするのか迷うところです。

注１）曲線回帰であるとか,対数変換した場合も残差が上記 1）から 4）を満たすかを調べる必要があります。

参考；竹内 啓，芳賀 敏郎，野澤 昌弘，岸本 淳司：「SAS による回帰分析」東京大学出版会（1996 年）

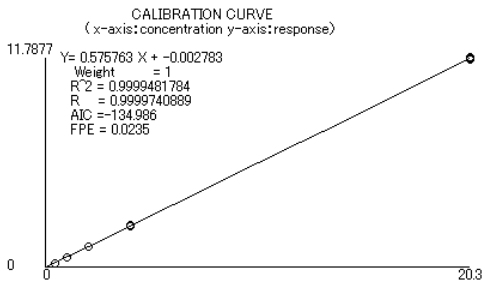
16.6.2　残差ﾌﾟﾛｯﾄと等分散性

それでは,実際のデータで検討してみます。ある血中薬物濃度の検量線用のデータです。
測定は HPLC 法（高速液体クロマトグラフィー）です。
５濃度を４検体測定したものです。検量線用の検体はブランク血清に薬剤を添加して処理して測定したものですから，各濃度の精度を示しているといえます。

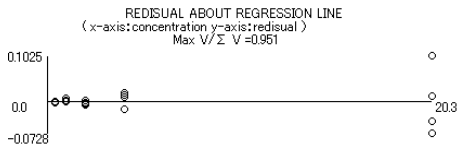
	X(濃度)	Y(IS比)4回測定				各分散
		1	2	3	4	
1	0.406	0.2297	0.2266	0.2283	0.2291	1.81E-06
2	1.015	0.5873	0.5833	0.5802	0.5802	1.13E-05
3	2.03	1.1598	1.1673	1.1566	1.1631	2.10E-05
4	4.06	2.3446	2.3498	2.3173	2.3541	0.0002743
5	20.3	11.6401	11.697	11.6124	11.7877	0.00599203

分散の合計	0.00630046

重み無しで検量線を作成すると，下記のようになります。



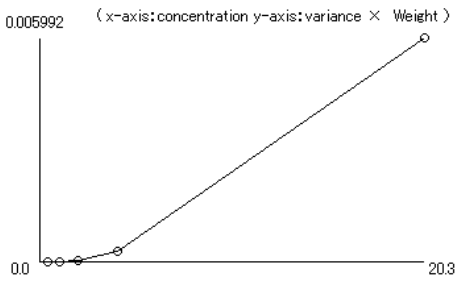
検量線



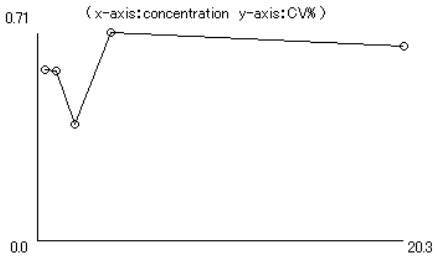
残差プロット

検量線の残差プロットから重みの必要性を判断することができます。
標準偏差の最小と最大が３倍以上ならば重み付けは必要です。
しかし，重みが必要であるかは，等分散性を調べることです。

PP 図からは，どのような重みが必要なのか読み取れます。PP 図を下記に示します。
PP 図については，「第 11 章　RER と PP 図で誤差を解析する」で述べました。



縦軸分散



縦軸 C.V.%

縦軸を分散にすると濃度で分散が大きくなっていくことが読み取れます。
さらに，縦軸を各濃度での c.v.%にすると

$$c.v.\%_i = \frac{\sigma_i}{\bar{x}_i} \times 100$$

が一定になっていることが解ります。
c.v.%が一定とは，どのような重み付けが必要なのかを，次に説明します。

参考 **y での重み付**けについて

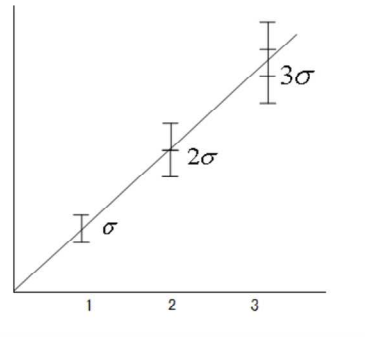
統計学の本などでは重み付けに１／yi，１／yi^2などを使用しているものがあります。[文献 1)]
統計学上は，本来 y の分散 σ^2の逆数で重み付けるべきなので，yi を使用すべきあるとする意見もあります。１／yi，１／yi^2の場合は，遂次計算し，収束した近似推定係数を求める必要があります。yi で重み付けるとその回帰式からの各推定値 Yi が求まります。そこで，その推定値 Yi で今度は重み付けます。これを繰り返し，重み付け回帰式を求める必要があります。（η による重み付け）
機器分析では，x 軸側は標準物質の濃度ですから，x 側の誤差は 0 として扱います。
このため，１／xi，１／xi^2を重みに使用すれば収束した近似推定値を求める必要がありません。
機器分析では x 側を重み付けに使用しているものが多くあります。
もし，x で重み付けても，y の分散が直線的でない場合は，濃度に対する y の分散の変化を曲線で近似して重み付けを行うことも可能です。
また，y で重み付ける場合でも，η による重み付け（推定値 Yi での重み付け）でない，単に yi での重み付けを使用している分析機器もあります。

文献 1）竹内 啓，芳賀 敏郎，野澤 昌弘，岸本 淳司：「SAS による回帰分析」東京大学出版会

16.6.3　１／x と １／x^2 の重み付けの選択

各濃度の y の分散（または標準偏差）を調べればよいのですが，分析機器のデータ処理ソフトに組み込まれいる１／x^2 と １／x の重みとは何かを考えます。

１／x^2 とは



「濃度 x_i に y の標準偏差が比例する」と仮定し，残差は平均０で分散（$x_i\sigma$）^{2} の正規分布に従うとします。右の図のような濃度と σ が比例する場合。

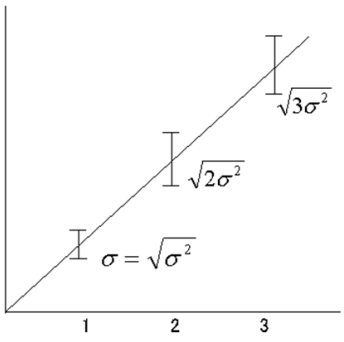
$$\varepsilon_i \sim N(0,(x_i\sigma)^2)$$

分散は

$$(x_i\sigma)^2 = x_i^2\sigma^2$$

となり，分散 $x_i^2\sigma^2$ を一定にするための重みは $w_i = \frac{1}{x_i^2}$ とになります。

１／x とは



「濃度 x_i に y の分散が比例する」と仮定すると

$$\varepsilon_i \sim N(0,x_i\sigma^2)$$

となり，分散 $x_i\sigma^2$ を一定にするための重みは $w_i = \frac{1}{x_i}$ とになります。

右の図のような関係の誤差がある場合になります。

まとめ

重み付けは下記のようになります。

$w_i = \frac{1}{x_i^2}$　は濃度 x_i に y の標準偏差が比例すると仮定した場合に使用する。

$$\sigma_i \propto x_i$$

$w_i = \frac{1}{x_i}$　は濃度 x_i に y の分散が比例すると仮定した場合に使用する。

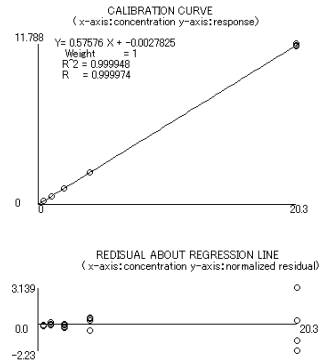
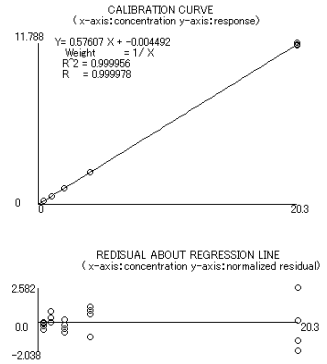
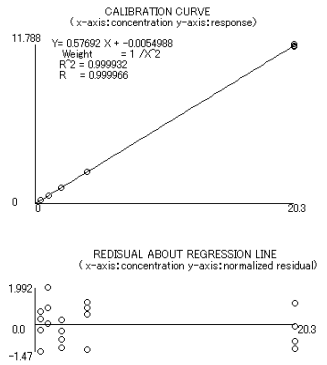
$$\sigma_i^2 \propto x_i$$

PP 図で c.v.が一定とは，濃度と標準偏差の比が一定であることですから $w_i = \frac{1}{x_i^2}$ が仮定できます。

重み無しと１／ｘ，１／ｘ^{2}の重みの結果を示します。

	X(濃度)	Y(IS比)4回測定				各分散
		1	2	3	4	
1	0.406	0.2297	0.2266	0.2283	0.2291	1.81E-06
2	1.015	0.5873	0.5833	0.5802	0.5802	1.13E-05
3	2.03	1.1598	1.1673	1.1566	1.1631	2.10E-05
4	4.06	2.3446	2.3498	2.3173	2.3541	0.0002743
5	20.3	11.6401	11.697	11.6124	11.7877	0.00599203

分散の合計	0.00630046

重み無し	**重み＝１／ｘ**	**重み＝１／ｘ^2**
		
Y= 0.57576 X + -0.0027825 Weight of data = 1 Hartley の等分散性の検定 Max V/Min V =3312.041 5% 点 =50.9 等分散性の仮説は危険率 5%で棄却された Cochran の等分散性の検定 Max V/Σ V =0.951 5% 点 =0.5981 等分散性の仮説は危険率 5%で棄却された	Y= 0.57607 X + -0.004492 Weight of data = 1/ X Hartley の等分散性の検定 Max V/Min V =66.241 5% 点 =50.9 等分散性の仮説は危険率5%で棄却された Cochran の等分散性の検定 Max V/Σ V =0.7594 5% 点 =0.5981 等分散性の仮説は危険率5%で棄却された	Y= 0.57692 X + -0.0054988 Weight of data = 1 /X^2 Hartley の等分散性の検定 Max V/Min V =3.269 5% 点 =50.9 **等分散性の仮説は危険率 5%で棄却出来ない** **Cochran の等分散性の検定** **Max V/Σ V =0.2857** **5% 点 =0.5981** **等分散性の仮説は危険率 5%で棄却出来ない**

このデータでは，等分散性の仮説が棄却されることがない重み＝１／ｘ^{2}が良いことが解ります。

実際の検量線の重みの選択は，重み付けたときの残差プロット，検定結果，全領域での打ち返し濃度，どの濃度域が重要かなどを総合的に判断し重み付けを決めます。

16.7　濃度の計算方法

回帰式は測定値 y がバラツキ，濃度 x はバラツキがないとして回帰式を求めています。濃度 x から測定値 y を求めることには問題はありません。

しかし，濃度の算出は，測定値 y が得られ濃度 x を求めます。この逆推定には問題があります。

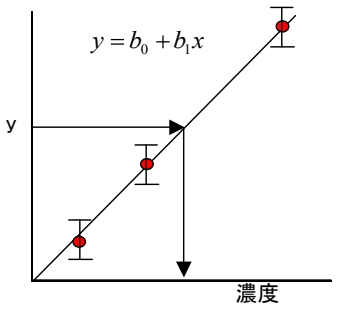
回帰式が

$$y = b_0 + b_1 x$$

であると，逆推定は

$$x = \frac{y - b_0}{b_1}$$

となります。



b_1 も誤差があることから正規分布と正規分布の分数となっていて厄介です。

このことは「第 10 章　誤差伝播の法則　10.2」でも示しました。

各データが検量線に上手くフィットしていることが必要です。各濃度の点が検量線から外れていると，相当精確さのない濃度を算出している可能性があります。

高次回帰式を考えると，逆推定は 2 次回帰なら 2 次方程式の解を求めることなので「解の公式」から，

$$y = ax^2 + bx + c$$

の逆推定は

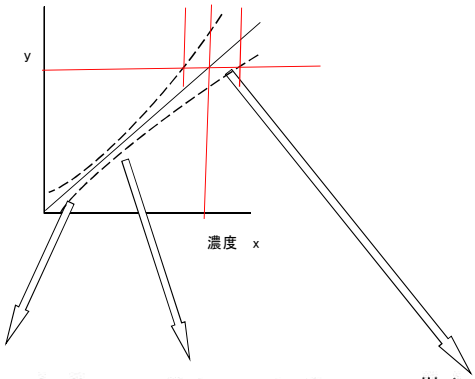
$$\frac{-b \pm \sqrt{b^2 - 4a(c - y)}}{2a} = x$$

になります。2 次方程式なので解は 2 つありますが，検量線の範囲内の値を採用します。3 次は「カルノダ法」，4 次は「フェラリ法」がありますが，5 次以降の「解法」は数学上存在しません。

しかし，実際は何次回帰式でも，さらに非線形回帰式の場合も，解の範囲を定めれば，解は数値的に反復収束させて求めることが出来ます。ニートン・ラプソン法，はさみ打ち法，遂次代入法などがあります。[注 1)]

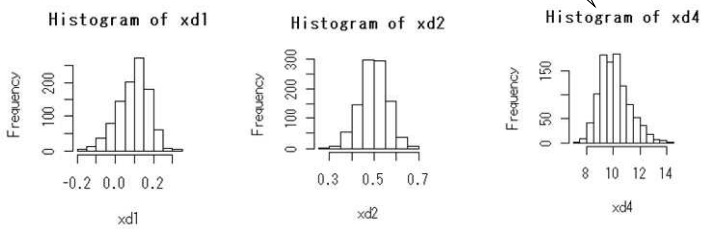
注 1）荻原 国宏：「BASIC による　実用数値計算　2 次方程式から有限要素法まで」山海堂（1987 年）

逆推定の信頼区間も知りたいところです。[注1，2)]
通常ｙの誤差を調べ，その時逆推定から得られる濃度の信頼区を知ることが出来なけば，ISO17025 などの信頼出来る不確かさを推定するこは出来ません。
機器分析では，秤量や定容の誤差が無視できるほどで，前処理の誤差もさほど大きくはあいません。環境検査，臨床検査，食品検査，薬物濃度測定などの濃度分析の分析機器は質量分析に移行してきています。
調べてみると，分析機器の測定誤差が大きいことがよくあります。
この場合，どうしても分析機器での逆推定の信頼区間を知る必要が出てきます。
しかし，重み付き回帰の逆推定の信頼区間を計算出来る市販の統計ソフトがありません。
この問題に答えているインンターネット上の記載があります。
下記の高橋行雄の
第３回 応用回帰分析１ －各種の重み付き回帰における逆推定－
http://www.yukms.com/biostat/takahasi2/rec/003.htm
に逆推定の信頼区間に関して詳細に記載されています。
エクセルでの例が記載されていますが，数式処理ソフトならば簡単に推定値を求めることがでます。
直線回帰の場合，低濃度では右に傾いた分布になり，高濃度では左に傾いた分布になります。



$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + x^n + \varepsilon_i \quad \varepsilon_i \sim N(0, x_i^{\delta}\sigma_i^2)$$

ｘ軸方向の９５％信頼区間の推定は下記の式から求めることができます。



$$(\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + ... + \hat{\beta} x_i^n) \pm t_{0.05}\sqrt{Var(\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + ... + \hat{\beta} x_i^n) + x_{95\%h}^{\delta}\sigma_e^2 / m} = y_0$$

$$(\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + ... + \hat{\beta} x_i^n) \pm t_{0.05}\sqrt{Var(\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + ... + \hat{\beta} x_i^n) + x_{95\%h}^{\delta}\sigma_e^2 / m} - y_0 = 0$$

注１）秋山功，富山茂巳，高橋行雄：「残留農薬分析における検量線の重み付けの選択と信頼区間の推定」第 106 回　日本食品衛生学会，ｐ101，2013
注２）秋山功：「等分散性が成り立たない場合の回帰分析 －検量線を事例として－」第７回定例会　医薬安全研定例会，2010

16.8　分散分析表，回帰係数の誤差

さらに，重み付き回帰式の誤差について検討する必要がある場合も考えられます。そこで，分散分析表についても記載しておきます。

重み付き回帰の分散分析表

重み付き回帰

$$f(x) = b_0 + b_1 x$$

の残差 e の残差平方和は

$$s_e = \sum w_i (y_i - f(x_i))^2$$

であることはすでに示しました。
この誤差分散は

$$V_e = \frac{\sum w_i (y_i - f(x_i))^2}{n-2}$$

となります。
分散分析表は下記のようになります。

要因	平方和 S	自由度 ϕ	分散 V
全体	$S_T = \sum_{i=1}^{n} w_i (y_i - \bar{y})^2$		
回帰	$S_R = \sum_{i=1}^{n} w_i (f(x_i) - \bar{y})^2$		
残差	$S_e = \sum_{i=1}^{n} w_i (y_i - f(x_i))^2$	n －p	$V_e = \frac{\sum_{i=1}^{n} w_i (y_i - f(x_i))^2}{n-p}$

重み付き平均 $\bar{y}$ ： $\bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$　　n：データ数　　p ：パラメータ数

決定係数は $R^2 = \frac{S_R}{S_T} = 1 - \frac{S_e}{S_T}$ です。

16.9　関数の当てはめと微分方程式

自然現象の多くは微分方程式で記述できます。難しい面倒なことは考えなくてもパソコンで微分方程式を解くことも可能な時代になりました。[注1)]

よく使用する検量線の濃度による変化は，微分方程式の解として理解することが出来ます。

どの濃度でも変化の割合αが一定であると仮定すると，微分方程式は

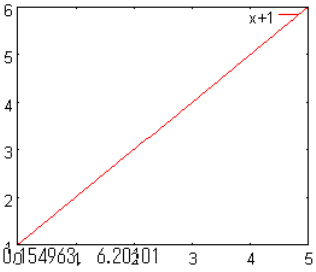
$$\frac{dy}{dx} = \alpha$$

になります。

直接積分すれば，この一般解は普段よく使用する直線回帰

$y = \alpha x + \beta$

になります。



濃度に比例して一定の割合αでレスポンスが増加するモデルを考えると，その微分方程式は

$$\frac{dy}{dx} = \alpha x + \beta$$

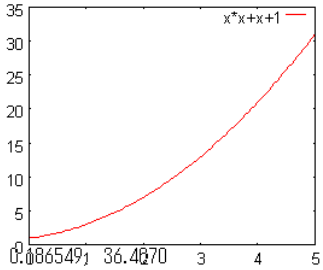
となります。この解は２次回帰式

$$y = \frac{\alpha}{2}x^2 + \beta x + \gamma$$

です。

$y = ax^2 + bx + c$

の形になります。



変化速度が一定であるとして，変化速度をβとすると

$$\frac{dy}{dx} = \beta y$$

$$\frac{dy}{dx} \cdot \frac{1}{y} = \beta$$

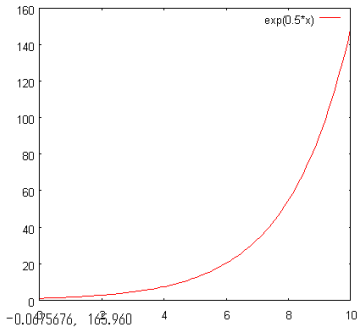
注１）本文「付録２　無料パソコンソフトの利用」を参照して下さい。

この一般解が

$$y = \alpha e^{\beta x}$$

です。

つまり，指数関数的に変化するとは，変化速度 β が一定であるとする微分方程式の解となっています。



$$y = \alpha e^{\beta x}$$

は，曲線となりますが，対数変換することにより

$$\log y = \log \alpha + x\beta$$

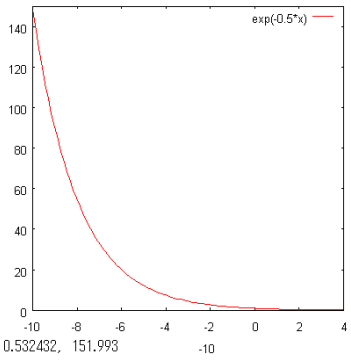
になり，

$$\log y = Y$$

$$\log \alpha = A$$

とすれば

$$Y = A + x\beta$$

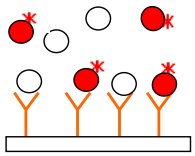


ですから，データ y を対数にして Y とすれば，直線回帰にすることができます。

しかし，対数変換で低濃度に重みが掛ります。

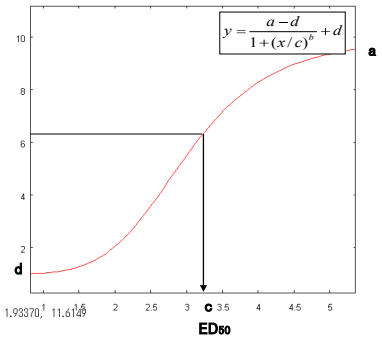
指数関数的な変化は生体内代謝，薬物動態，刺激に対する反応，化学物質の残留率（残留農薬）など多くの現象に見られます。

RIA や EIA(enzyme immunoassay)の検量線として log-log 3 次，4 係数 logit，3 次スプラインなど多くの方法が利用されますが，抗原抗体反応（測定系）を考えると，指数関数や 2 次曲線，S 字曲線になることが理解できます。



よく知られている Verhulst の人口モデル, 細菌の増殖なども S 字曲線になり, 自然界の多くの現象は S 字曲線になり, ロジスティック方程式と呼ばれる微分方程式で記述できます。

ロジスティック方程式は自然科学のみならず，社会科学においても広範囲に重要な方程式です。



ロジスティック方程式は

$$\frac{dy}{dx} = (\alpha - \beta y)y$$

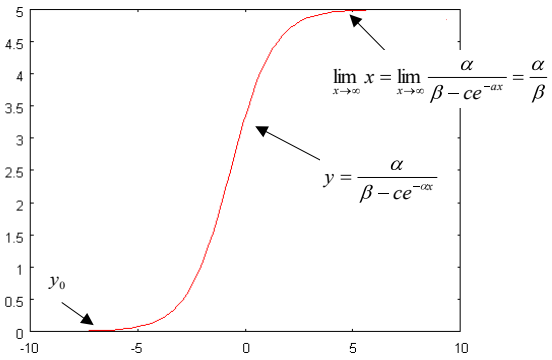
です。

この一般解は $x = 0$ の時の値を y_0 とすれば

$$c = \frac{\alpha - \beta y_0}{y_o}$$

$$y = \frac{\alpha}{\beta - ce^{-\alpha x}}$$

となります。



ロジスティック曲線は非線形ですが，非線形最小２乗法は数学ソフトやエクセルでも計算できます。[注１）]

注１）本文「付録２　無料パソコンソフトの利用」を参照して下さい。

gnuplot：グラフ作成ソフト（フリーソフト）http://www.gnuplot.info/

Mathcad：Parametric Technology Corporation (PTC)数値計算ソフト

R：統計解析ソフト（フリーソフト）http://www.r-project.org/

参考１ロジスティック方程式

$$\frac{dy}{dx} = (\alpha - \beta y)y$$

は

$$\int \frac{1}{(\alpha - \beta y)y} dy = \int dx$$

$$\frac{1}{\alpha}\int \left(\frac{\beta}{\alpha - \beta y} + \frac{1}{y} \right) dy = \int dx$$

$$\frac{1}{\alpha}\{-\log(\alpha - \beta y) + \log y\} = x + C$$

$$\log \frac{y}{\alpha - \beta y} = \alpha x + \alpha C$$

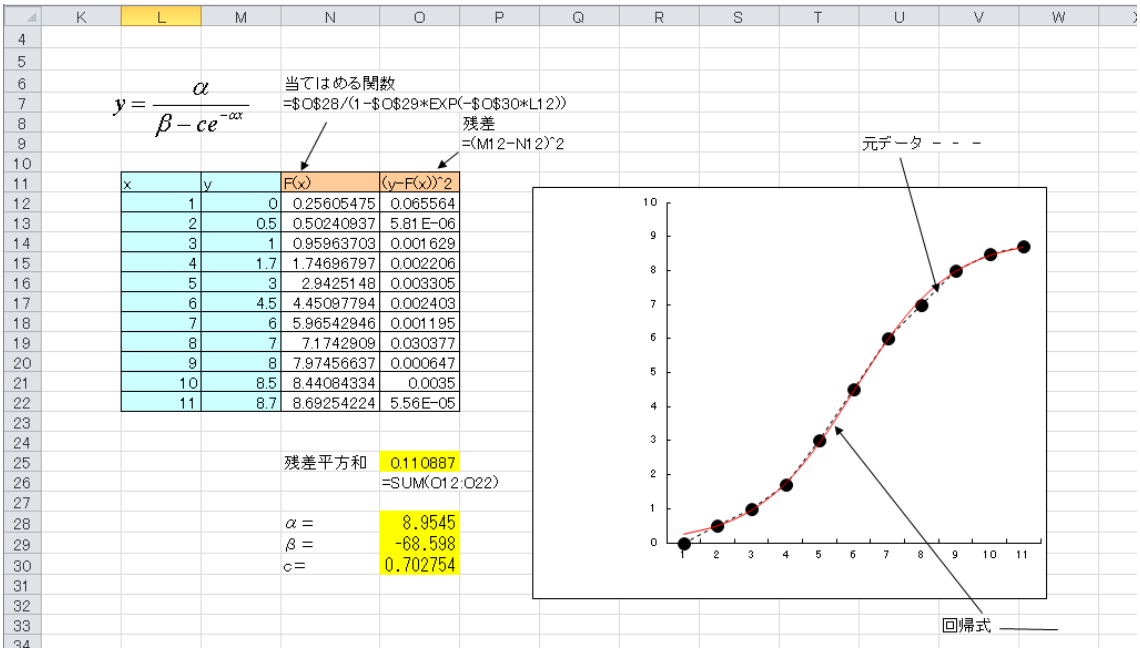
$$\frac{y}{\alpha - \beta y} = e^{\alpha x} e^{\alpha C}$$

$e^{\alpha C}$ をあらためて c とおいて

$$y = \frac{\alpha}{\beta - ce^{-\alpha x}}$$ となります。

尚，「付録２　無料パソコンソフトの利用」の Maima を利用することも可能です。

エクセルのソルバーによるロジスティック回帰式の求め方を参考として下記に示します。

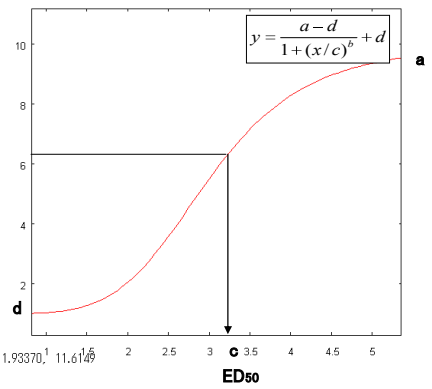


「16.4　エクセルのソルバーで重み付き最小２乗法の原理を理解する」と同じです。

エクセルのソルバーと同じものは，統計ソフトや数式処理ソフト，グラフソフトなどにも組み込まれていますので，ソルバーを使用してロジスティック回帰式を求めることができます。

参考２ 重み付き４パラメータロジスティック検量線

抗原抗体反応を利用した EIA などでは S 字曲線になり，下記の４パラメータロジスティック検量線の式がよく使用されます。この場合も各濃度に対応する重き付けを行うことがあります。(yi の分散 σi^2 を重みにする)a が最大値で d が最小値になり，c が中点になりますので，各係数に意味があります。



$$y = \frac{a-d}{1+(x/c)^b} + d$$

重み付けは本文で述べたように，回帰式の等分散性を確保するために行います。

検量線はその他にも，いろいろな種類のものがありますが，式の意味を理解して使用しないと危険です。

とにかく，重み付けは「等分散性」を確保するためです。

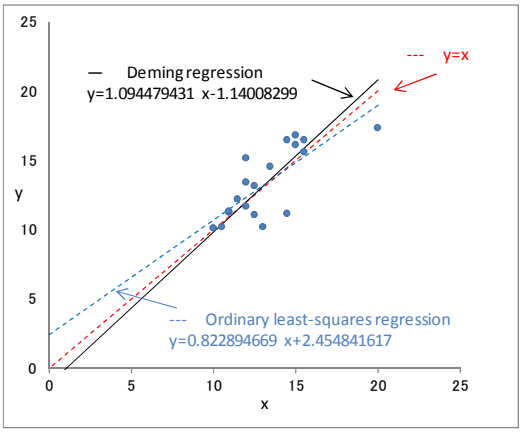
また，重み付けにより回帰式は安定します。

第１７章　測定値の比較検討（重み付き Deming 法）

濃度分析での方法間や試薬 Lot.間差の比較では，ｙ軸側のみに誤差があるとするエクセルの通常の回帰は不適切であると，多くの論文でも指摘されていて，日本臨床化学学会などでもｘ軸ｙ軸側共に誤差があるとする線形関係式の使用を推奨しています。
線形関係式には標準主軸法や Deming 法がなどがあります。
統計ソフトの多くが重み付き Deming 法や Bland-Altman プロットが実装されてきていいます。本文で示す重み付き Deming 回帰法の計算値や図は，本文で示す数式から自作したもので計算した結果です。
また，フリーソフトＲや日本臨床化学学会のソフトでも計算可能です。
十分に相関が強ければ，通常のｙ軸側のみに誤差があるとする回帰式を使用しても大きな差はないので問題ないとする考えもありますが，便宜的な統計手法を使用することになります。

17.1　実データでの線形関係式の必要性

下記の図は臨床の方法間比較の実データでの回帰式です。
通常の回帰では
y=0.823x+2.45
となり，傾きは 0.823 で，切片+2.45 でｙ＝ｘとは考えられません。
つまり，通常のｙ軸側のみに誤差があるとする回帰では 2 方法で差が出ると判断することになります。
しかし，ｘ軸ｙ軸側共に誤差があるとする Deming 法で回帰すると
y=1.09x－1.14
となります。
傾き 1.09 で，切片－1.14 となりｙ＝ｘに近い値になります。

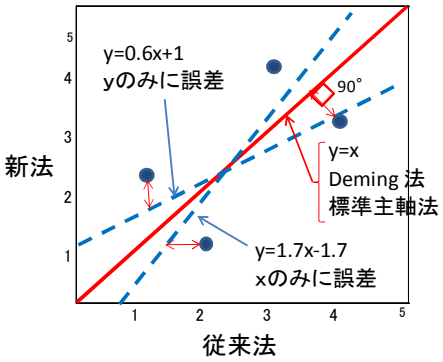


ここで示した例からも，方法間比較ではｘ軸ｙ軸側共に誤差があるとする Deming 法などの線形関係式で検討する必要があります。
簡単に計算できるｙ軸側のみに誤差があるとするエクセルの回帰式を使用すべきではありません。

17.2　回帰式と線形関係式

通常の回帰ではｙ軸のみに誤差があるとして計算しています。通常のエクセルの回帰分析が比較試験では不適切であるのかを，簡単な数値例と図で再度考えてみます。
下記のデータが得られたとすると，平均=2.5 で回帰式は y=x が妥当と考えられます。

	x従来法	y新法
1	1	2
2	2	1
3	3	4
4	4	3
平均	2.5	2.5



上のデータから，右に各種の回帰直線を計算してみました。

1）ｙのみに誤差があるとした回帰式（通常の回帰式）
2）ｘのみに誤差があるとした回帰式
3）標準主軸法
4）Deming 法

1）の通常の回帰はｙ軸方向のみに誤差があるとする回帰式でｙ=0.6ｘ＋1 となり，低い傾きとなっています。逆に 2）のｘ軸方向のみに誤差があるとした場合は，y=1.7ｘ－1.7 で高い傾きとなり，この式も妥当とはいえません。

「測定値の比較」ではｘ，ｙ共に誤差があるのでｘ，ｙ共に誤差があるとする線形関係式である，標準主軸法，Deming 法では共に y=x と一致した回帰式が得られます。

この例からも，方法間比較や Lot.間比較では線形関係式を使う必要性を示しています。

x 軸，y 軸共に誤差がある場合に適用できる Deming 回帰では，分散の比を知る必要があります。

$$\lambda = \frac{x\text{の分散}}{y\text{の分散}}$$

この分散比λが不明なときは，λ=1 とすることが，海外の臨床化学では推奨する文献が多いようです。

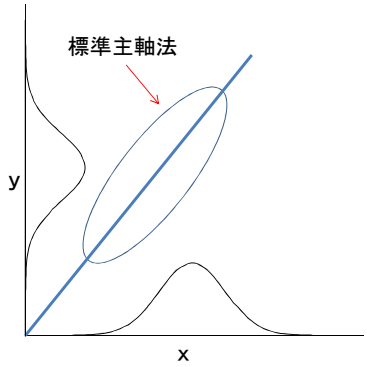
しかし，集中楕円を考えるとその長軸に一致する標準主軸法が考えられます。

λを偏差平方和 Syy，Sxx と各分散 ex2，ey2 が比例していると仮定し

$$\lambda = \frac{e_x^2}{e_y^2} \approx \frac{Sxx}{Syy}$$

とすると，標準主軸法になります。（17.3 に示す偏差平方和の式を見て下さい。）

標準主軸法は散布図に集中楕円を描くとその長軸に一致する利点があります。



標準主軸法では，２変量正規分布でない場合は変数変換して正規分布にする必要があります。

２変量正規分布でない場合や重み付けが必要な場合は，分割して標準主軸法を使用する考えもあるようですが，便宜的な方法です。

Deming 回帰のλを変えれば標準主軸法にりますので，Deming 回帰に標準主軸法が含まれます。

17.3　線形関係式（重み付き Deming 回帰）

通常の回帰は

$y_i = \alpha + \beta x_i + \varepsilon_i$

で ε は y のみ誤差があるとして，x の誤差は 0 として回帰式を求めています。

Deming 回帰では

$x_i = X_i + \varepsilon_i$

$y_i = Y_i + \delta_i$

として x と y に誤差があるとして回帰式を求めています。

通常の回帰モデルと同様に誤差は下記の仮定をします。

$$\varepsilon_i \sim \text{i.i.d. } N(0,\ \sigma_\varepsilon^2)$$

$$\delta_i \sim \text{i.i.d. } N(0,\ \sigma_\delta^2)$$

標準主軸法と重み付き Deming 法の式を以下に示します。

標準主軸法は簡単に

$$Slope = \sqrt{\frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{Syy}{Sxx}}$$

$$Intercept = \bar{y} - Slope \times \bar{x}$$

で求めることもできます。

重み付き Deming 法について述べます。

w＝1 にすれば重み無しの Deming 法になり。

λ を $\lambda = \dfrac{Sxx}{Syy}$　にすれば標準主軸法になります。

つまり，「重み付き Deming 法」は「無しの Deming 法」や「標準主軸法」を含みます。

通常の回帰式（y 軸側にのみ誤差がある場合）では，下記の y 軸方向の残差平方和 Se を最小にする係数を求めます。

$$Se = \sum (y_i - \hat{y}_i)^2$$

同様に，重み付き Deming 法では，下記の x 軸，y 軸両方向の重み付き残差平方和 Sew を最小にする係数を求めます。

$$Sew = \sum w_i \left\{ (x_i - \hat{x}_i)^2 + \lambda (y_i - \hat{y}_i)^2 \right\}$$

下記の分散比λを決める必要があります。

$$\lambda = \frac{\sigma_x^2}{\sigma_y^2}$$

不明な場合はλ=1 とします。λ=1 にすれば，回帰直線と 90°の残差を最小にするので，理解し易いです。
回帰式で等分散にする理論上の重み w は分散の逆数を使用することです。
分散比λが判っている場合は Deming 回帰を使用すべきであり，分散比λはある程度推定できる場合が多くあります。

精度管理図や x 軸側と y 軸側の 2 重測定からも推定できます。

$$\lambda = \frac{x\text{の分散}}{y\text{の分散}} = \frac{s.d._x^2}{s.d._y^2} = \frac{c.v.\%_x^2}{c.v.\%_y^2}$$

分散比λが判っていて，濃度でバラツキが変化する場合は「重み付き Deming 回帰」を使用します。
「16. 6. 3　１／x と１／x^2 の重み付けの選択」で述べたように，濃度分析では通常は濃度に比例してバラツキが大きくなる傾向にあります。

濃度とバラツキが標準偏差 s. d. に比例している場合は $\lambda \approx 1$ として

$$\sigma_i \propto (x_i + y_i)/2$$

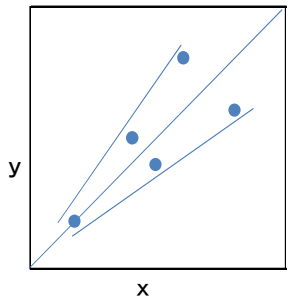
が成り立つとして

$$\varepsilon_i \sim N(0,((x_i + y_i)/2 \times \sigma)^2)$$

から

$$w_i = \frac{1}{\left[\frac{\hat{x}_i + \hat{y}_i}{2}\right]^2}$$

の重みが考えられます。



濃度に分散 σ^2 が比例している場合は下記の重みになります。

$$w_i = \frac{1}{\left[\frac{\hat{x}_i + \hat{y}_i}{2}\right]}$$

各濃度の x 軸側と y 軸側のバラツキが調べられていればその値を重み付けます。

さらに，分散比 λ を考慮した重み付けもできます。本章の参考文献)

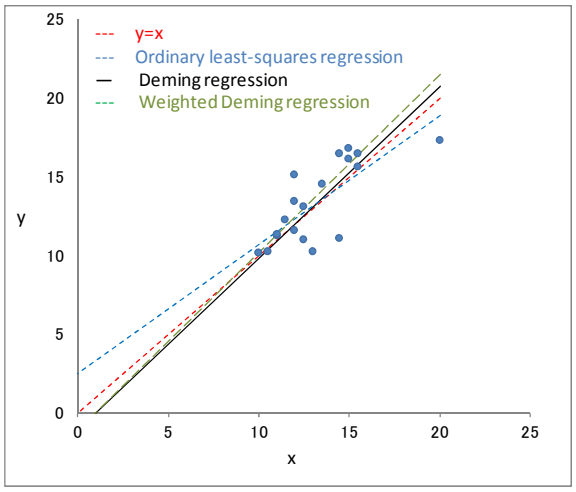
$$w_{i\lambda} = \frac{1}{\left[\frac{\hat{x}_i + \lambda\hat{y}_i}{1+\lambda}\right]^2}$$

Kristian Linnet[5) 6)] はこの式を採用しています。重み付けにより，各濃度での等分散性を確保できるようにします。

重み付き平均は

$$x_{wm} = \frac{\sum w_i x_i}{\sum w_i}$$

$$y_{wm} = \frac{\sum w_i y_i}{\sum w_i}$$



となります。通常の重み付き平均と同じです。

重み付き偏差平方和は

$$Sxx_w = \sum w_i (x_i - x_{wm})^2$$

$$Syy_w = \sum w_i (y_i - y_{wm})^2$$

$$Sxy_w = \sum w_i (x_i - x_{wm})(y_i - y_{wm})$$

となり

重み付き Deming 回帰式は下記の式で求めることができます。[5) 6)]

$$Slope = \frac{(\lambda Syy_w - Sxx_w) + \sqrt{(Sxx_w - \lambda Syy_w)^2 + 4\lambda Sxy_w^2}}{2Sxy_w}$$

$$Intercept = y_{wm} - Slope \times x_{wm}$$

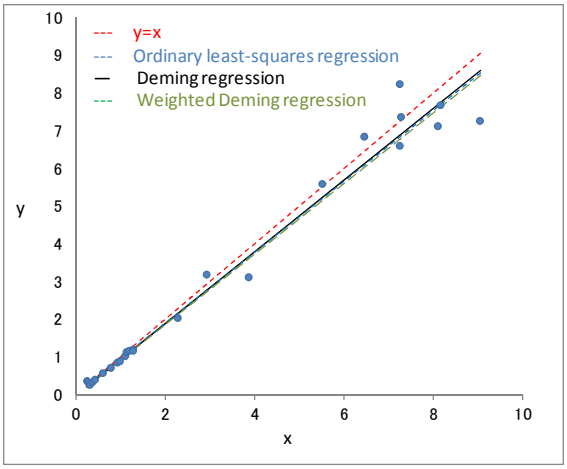
上記式で，w＝1 とすれば重み無しの Deming 回帰式になり，λ=Sxx/Syy にすれば標準主軸回帰式になります。

よくパラメータの 95%区間推定にエフェロンの考案した Bootstrap method を使用しているものがありますが，検定を「測定値の比較検討」で行うことは適切でない場合があるの注意すべきです。有意差検定で述べたように，技術的判断や臨床的判断が重要です。

17.4　Bland-Altman プロット（偏差プロット）

測定方法間の比較で，特によくある Lot.間比較でｙ＝ｘであるかが問題になる場合は Bland-Altman プロットは有効な手法です。
横軸に（ｙ＋ｘ）／２に平均を取り，ｙ－ｘの差を縦軸に取った図です。
下記の図は，通常ｘ－ｙですが，ｘを基準とすることが多いので，見にくいのでｙ－ｘにしました。差の標準偏差 s.d.を計算し，95%区間を計算しています。
この区間に０を含まない場合は明らかに差があることになります。

1　　Regression Analysis



重み付きDeming回帰　　(Weighted Deming regression)

λ ＝ 1

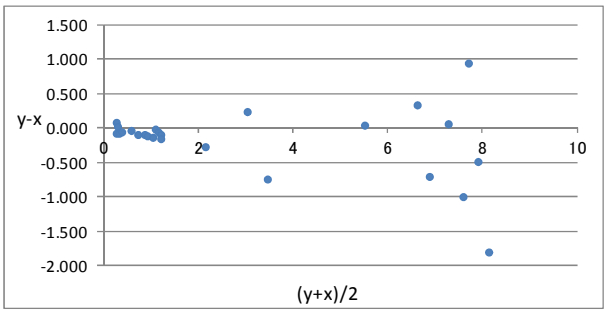
Slope= 0.934157379　　r = 0.9863533

Intercept= −0.017493364　　R^2= 0.9728928

	x	y
1	0.27	0.33
2	0.31	0.31
3	0.34	0.24
4	0.39	0.3
5	0.45	0.38
6	0.63	0.57
7	0.81	0.69
8	0.94	0.82
9	1	0.87
10	1.13	0.99
11	1.15	1.11
12	1.22	1.14
13	1.29	1.18
14	1.3	1.14
15	2.31	2.02
16	2.94	3.16
17	3.87	3.11
18	5.54	5.56
19	6.48	6.81
20	7.27	6.56
21	7.27	8.2
22	7.29	7.34
23	8.11	7.1
24	8.17	7.67
25	9.06	7.25

濃度に比例してバラツキが大きくなっているので，重み付き Deming 回帰を行った。

2　　Bland-Altman Analysis
Difference plot



上限(95%)
0.8123

平均
−0.18760

下限(95%)
−1.18747

平均＝ 3.088
差の平均= −0.1876
差の標準偏差Std Dev= 0.4999

参考１ソフトと参考文献

下記の文献にはプログラムも記載されています。

丹後 俊郎：「測定誤差を考慮に入れた線形関係式-測定法の比較のための統計学的方法-」

線形関係式に関する参考文献を下記に示します。

Ｒや日本臨床化学学会のエクセルソフトなどでも計算可能です。

１）丹後俊郎：「測定誤差を考慮に入れた線形関係式」.臨床病理.XXXVI. 9:1101-1108.1988

２）丹後俊郎：「測定誤差のある線形モデル」.統計モデル入門.朝倉書店. 61-74.2000

３）市原清志：「臨床化学検査の分析能の比較評価」臨床化学. 27：21‐49.1998

４）市原清志:「臨床検査の方法間比較」 臨床検査増刊号 臨床検査のための情報処理技術の進歩 49, 12, : 1315-1326, 2005

５）Kristian Linnet: 「Performance of Deming regression analysis in case of misspecified analytical error ration in method comparison studies」 Clinical Chemistry 44:5.1024-1031.1998

６) Kristian Linnet:「Evaluation of Regression Procedures for Methods Comparison Studies」 Clinical Chemistry.39.3:424-432.1993

７）P.Joanne Cornbleet and Nathan Gochman :「ncorrect Least-Squares Regression Coefficients in Method Comparison Analysis」 Clinical Chemistry.25.3:432-438.1979

８）R:Package’mcr’ .version 1.21:Febrary20,2 : https://cran.r-project.org/web/packages/mcr/mcr.pdf (統計ソフトＲではパッケイジ’mcr’が使用できる：Deming 回帰，重み付き Deming 回帰，および Passing-Bablok 回帰を提供している。)

９）M A Pollock, S G Jefferson, J W Kane, K Lomax, G MacKinnon and C Bwinnard :「Method comparison-a different approach」Ann Clin Biochem. 29: 556-560.1992

１０）Katy Dewitte, Colette Fierens, Dietmar Stöckl, Linda M. Thienpont :「Application of the Bland‐Altman Plot for Interpretation of Method-Comparison StudiesA Critical Investigation of Its Practice」. Clinical Chemistry .48, 5, 799-801. 2002

１１）Validation-Support/Excel. 日本臨床化学学会 http://www.jscc-jp.gr.jp/?page_id=1145

１２）Deming Regression on the NCSS statistical software. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Deming_Regression.pdf

１３）日本臨床衛生検査技師会：「臨床検査精度保証教本」正確さの評価・管理方法.57-61.2010

本文の計算や図はエクセルで自作したもので，結果は統計ソフト JMP で確認しました。

注意）λは本文の説明と逆の（ｙの分散）/（ｘの分散）となっている文献やソフトもありますので，注意して下さい。

参考 2Q&A

問い

Deming 法

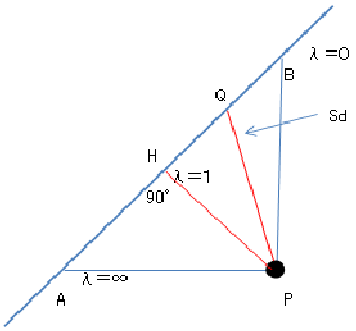
「分散比が不明なときでもλ=1 にすれば，回帰直線と 90°の残差を最小にするので，理解し易い。」

理解しやすいですが

λ＝0.5 はグラフにするとどのような意味になるのでしょうか？

角度が変わるのなか？

答え



λ＝(x 軸側の誤差 ex)／(y 軸側の誤差 ey)として

λ＝１の場合は，ＰＨは主成分分析の第一主成分Ｙ＝α＋βＸと垂直になります。

λ＝０の場合は，ｘの誤差がｙの誤差に比べ無視できるくらい小さい場合で，点Ｑは点Ｂに一致して通常の回帰式になります。

λ＝0.5 は図のλ＝１とλ＝０の中間の点Ｑになります。

Sd の２乗の総和を最小にする関数を求めますが，角度については，下記の関係が文献に記載されています。図から

$$\tan(\angle BAP) = Slope$$

$$\tan(\angle QPB) = \lambda \times Slope$$

傾きが１で垂直に交わるのは∠ＡＰＢが 90°なので，点Ｑを動かして∠ＱＰＢ＝∠ＨＰＢの時で，半分の 45°になるのは，λ＝1 の時であることは図から理解できます。

もしも傾きが１でλ＝0.5 なら

$$\angle QPB \cong 26.6\,\mathrm{deg}$$

になります。

第１８章　正規分布に変換する方法
－臨床基準値の算出－

データが正規分布にならない場合があります。例えば臨床検査の基準値（正常値）は対数正規分布を示すことがあります。しかし，多くの統計解析の推定や検定では正規分布を仮定しています。このため，データを変換して正規分布に近づけて解析する方法が有効です。

18.1　修正ベキ変換（ボックス・コックス変換）

下記のλによるベキ変換を**ボックス・コックス変換**と言います。

$$x^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda}-1}{\lambda} + \lambda & \lambda. \neq 0 \\ \log x & \lambda = 0 \end{cases}$$

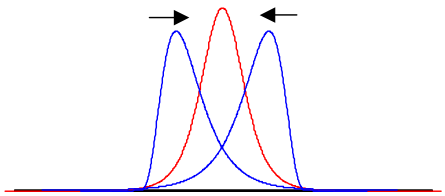
この変換で正規分布に近似させます。

ここに示した式は，Box-Cox1964 年を大瀧慈・後藤昌司 1999 年が報告した修正ベキ変換です。[注１）]

ベキ乗のλは 0 でない場合は

$$x^{(\lambda)} = \frac{x^{\lambda}-1}{\lambda} + \lambda$$

で，0 の場合は対数変換します。



この変換は無変換，対数変換，指数変換を含みます。最適な変換式は最尤法を利用しを求めます。[注２）]

λは最尤法よりデータ　$x_1 \ldots x_n$ で

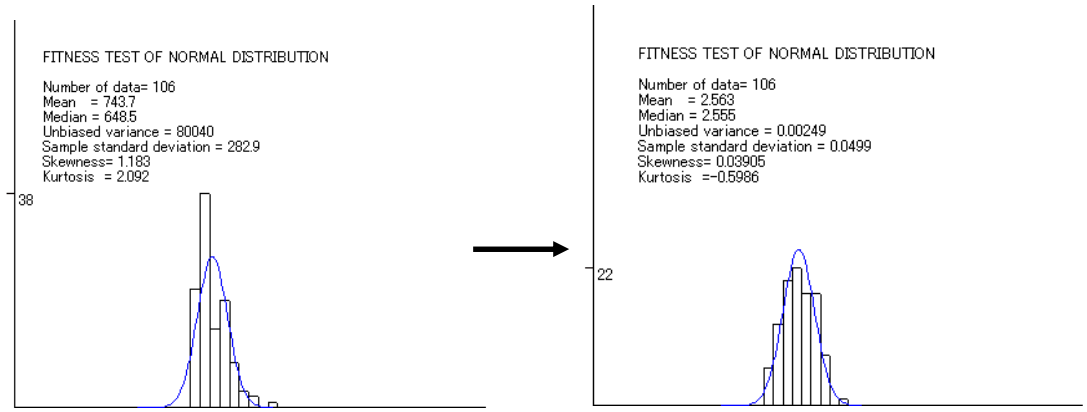
$$MLL(\lambda) = -\frac{N \log \sum_{i=1}^{n} (x_i^{(\lambda)} - \bar{x}^{-(\lambda)})^2}{2} + (\lambda - 1)\sum_{i=1}^{N} \log x_i$$

MLL が最大になる変換式を採用します。

注１）濱崎 俊光　他：「ベキ変換とその変型」応用統計学，Vol.２８,No3 179-190 ,1999
大瀧 慈：「ブートストラップ法を用いた外れ値の検出」応用統計学会，日本計量生物学会合同年次大会　139-144,2001

注２）最尤法は本文中「第 13 章　最尤法について」で説明しました。

パソコン用にプログラムを作製して，ボックス･コックス変換した例を下記に示します。
計算は先ほどの式を使用し，ヒストグラムと正規分布の比較もできるようにしました。



左に傾いた分布を正規分布に近づけたものです。
なぜベキ変換を単純に $x^{(\lambda)} = x^{\lambda}$ としないのかを説明します。
まず，0 乗は $x^0 = 1$ で全て 1 なるので 0 のときは log にします。
これは，

$$\lim_{\lambda \to 0} \frac{x^{\lambda} - 1}{\lambda} + \lambda = \log x$$

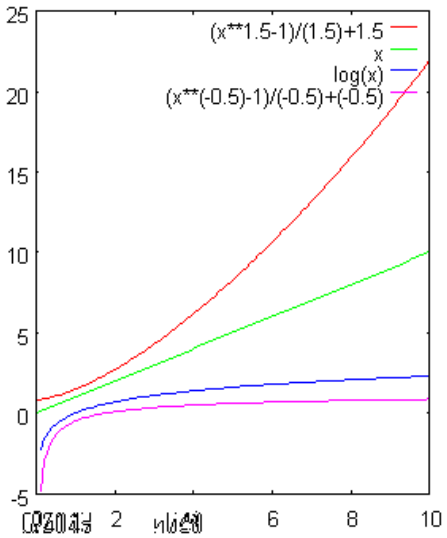
となるためです。
また，１乗のときも変換無しと同じにするために，このような式になっています。

$$\frac{x^1 - 1}{1} + 1 = x$$

つまり，一貫した変換ができる，

$$x^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} + \lambda & \lambda. \neq 0 \\ \log x & \lambda = 0 \end{cases}$$

の式を使用します。



臨床検査のデータは正規分布しないことが多いので，ベキ変換して正規分布に近づけ95%範囲の基準値（正常値）を計算することがあります。

また，外部精度管理のデータも同様に正規分布しない場合はベキ変換し，Z－スコアを計算し評価した方が正確な管理範囲を示せます。

ところで

$$x^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda}-1}{\lambda}+\lambda & \lambda \neq 0 \\ \log x & \lambda = 0 \end{cases}$$

で変換し，解析した後で値を元に戻す式を示すと，

log(x)＝ x $^{(\lambda)}$ の時は簡単に $e^{x^{(\lambda)}} = x$ で元の値に戻ります。

$\lambda \neq 0$ のときは

$$x^{(\lambda)} = \frac{x^{\lambda}-1}{\lambda}+\lambda$$

なので,

$$(\lambda x^{(\lambda)} + \lambda^{\lambda} + 1)^{\frac{1}{\lambda}} = x$$

で元の値に戻せます。

再度ここで述べた式を示しておきます。

データ x_i が得られたら，λ をいろいろ変えて下記の式に入れます。

$$x_i^{(\lambda)} = \frac{x_i^{\lambda}-1}{\lambda}+\lambda$$

最尤値は下記の式で計算します。

$$MLL(\lambda) = -\frac{N\log\sum_{i=1}^{n}(x_i^{(\lambda)} - \bar{x}^{-(\lambda)})^2}{2} + (\lambda-1)\sum_{i=1}^{N}\log x_i$$

MLL（λ）が最大になるモデルを選択します。

18.2　ジョンソン分布

ジョンソン分布を紹介します。

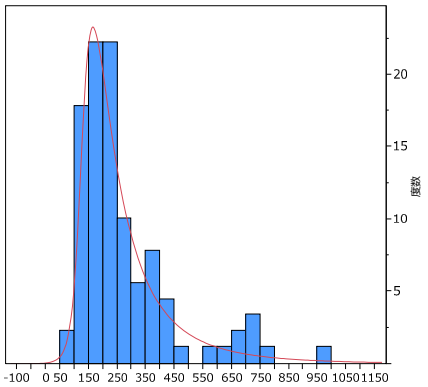
Johnson Su 分布は 3 次のモーメントである歪度と 4 次のモーメントである尖度の最適な変換を最尤法から求める方法です。

統計ソフト JMP や R で計算が可能です。

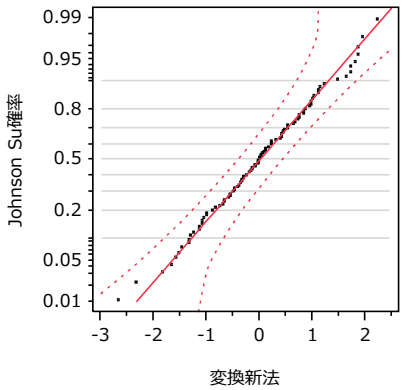
R は SuppDists パッケイジで計算できます。対称性と裾の広さの最適な分布を求めることが可能です。

「第１４章　母関数の魅力」で説明した 4 つのパラメータ，1 次から 4 次のモーメントのパラメータを変化させます。

実際に JMP で計算した例を以下に示します。



Johnson Suのあてはめ		
パラメータ推定値		
種類	パラメータ	推定値
形状	γ	-1.56522
形状	δ	1.051073
位置	θ	118.3377
尺度	σ	48.35704
(-2)*対数尤度 = 1155.89855724093		



適合度検定		
Shapiro-WilkのW検定		
W		p値(Prob<W)
0.992942		0.9077

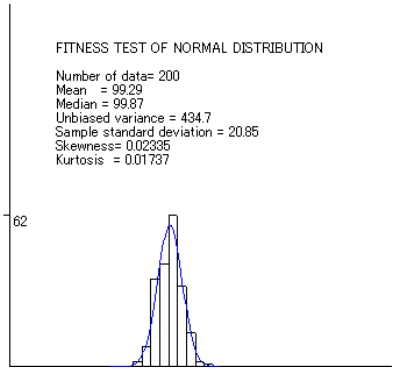
この例では，対数変換でも適合しませんが，ジョンソン分布ではよい適合が得られました。

18.3　正規性の確認

正規性の確認方法の幾つかを紹介します。

１）正規分布の確率密度関数とヒストグラムの比較

一番簡単で解り易いのは，データのヒストグラムと正規分布を比較することです。



２）χ^2適合度検定による正規性の検定

χ^2分布を利用した適合度検定により，正規性を検定しようとする方法です。

３）正規確率紙の利用　確率のプロット

正規確率紙にプロットすると，正規分布なら直線になります。

直線からはずれているかで正規性を判断します。パソコンでプログラムを組むことも可能ですが，正規確率紙も市販されています。

多くの統計ソフトでは確率プロットとして図が出力されます。

４）コルモゴロフ＝スミルニフ検定（標本分布関数の検定）

「統計数値表」JSA-1972　日本規格協会（1972 年）を参照して下さい。

５）歪度と尖度による方法

歪度と尖度は「第 14 章　母関数の魅了」で 3 次と 4 次のモーメントとして説明しましたので，計算例と検定方法を示します。

エクセルで下記のデータを計算してみます。

データ	6	3	4	2

歪度（正規分布なら分布は左右対称なので０となる。）は

ｓは標本標準偏差で

$$b_1 = \frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i - \bar{x}}{s}\right)^3$$

と，エクセルでは定義されています。

上記データのエクセルの関数での計算結果を示します。

```
=SKEW(6,3,4,2)
0.752837199
```

尖度（尖り度または広がり度を表す。）は

$$b_2 = \left\{\frac{n(n+1)}{(n-1)(n-3)(n-3)}\sum\left(\frac{x_i - \bar{x}}{s}\right)^4\right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

と，エクセルでは定義されています。

エクセルの関数での計算結果を示します。

```
=KURT(6,3,4,2)
0.342857143
```

このの値から下記のように判断します。

b_1　＞０　なら右に裾を引く

b_1　＜０　なら左に裾を引く

b_2　＞０　なら裾が長い

b_2　＜０　なら裾が短い

検定は，尖度と歪度の標準化でＮ（0,1）に変換して検定する方法を以下に示します。

歪度と尖度による正規性の検定[注 1)]

歪度ｂ$_1$，尖度ｂ$_2$，データ数をｎとして

$$Z_1 = \frac{|b_1|}{\sqrt{\frac{6(n-1)}{(n+1)(n+2)}}}$$

$$Z_2 = \frac{|b_2|}{\sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}}$$

を計算します。

Ｚ$_1$，Ｚ$_2$ともに

1.96（α=0.05)

2.58（α=0.0１）

3.29（α=0.001)

と比較し，これよりも大きければ有意差を認めます。（正規分布ではない。）

注 1）丹後 俊郎：「臨床検査への統計学」　朝倉書店

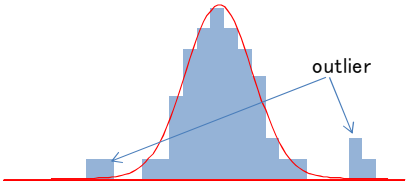
第１９章　切断した正規分布と打ち切りのある正規分布の平均と標準偏差

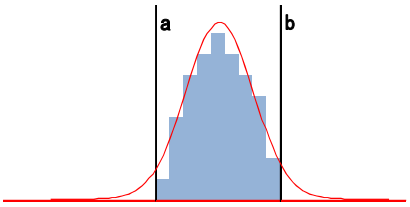
濃度分析では定量限界（limit of determination）がある場合が多くあります。
このため，ここで述べることは分析現場では頻繁に遭遇する問題です。
例えば，臨床検査で LT（以下），GT（以上）がありパラメトリック法で基準値を設定するとか，環境検査で定量限界があるデータの平均値や標準偏差を推定する必要がある場合などです。

データ解析では，定量下限(minimum limit of determination)は「左側打ち切り」で定量上限（maximum limit of determination）は「右側打ち切り」です。
この打ち切りがある正規分布（censored normal distribution）と切断された正規分布（truncated normal distribution）は意味が異なります。「切断」（truncated）と「打ち切り」（censored）を明確に区別する必要があります。「打ち切り」は「生存分析」と同様にデータはあるが，正確な値が不明な場合であり，切断は一定以上または以下を切断して推定します。切断は安易な 2s.d.や 3s.d.切り捨てることではありません。一部が切断された確率分布として推定する方法です。
この打ち切りのある正規分布と切断した正規分布との平均と標準偏差の計算方法を述べます。

19.1　切断した正規分布（truncated normal distribution）

下記のヒストグラムのように飛び離れたデータ(outlier)がある場合があります。この時，飛び離れたデータを除き平均と標準偏差を推定したいと考えます。
Grubbs-Smirnov の棄却検定は基本的には繰り返せません。
そこで，安易に 2s.d.や 3s.d.を計算して，外れた値を取り除き，再度平均と標準偏差を計算することが，臨床基準値の計算や外部精度管理のスクリーニングとして行われるが，それで良いのかを考えると，統計上不適切であることは明白です。
データ数が減り必ず標準偏差は小さくなります。その値を推定値とするのは不適切です。
ではどのようにすれば良いのでしょうか。
切断正規分布として平均と標準偏差を計算するのが適切であると考えます。





truncated normal distribution

丹後俊郎は「臨床検査への統計学」朝倉書店で患者データを含む場合，3s.d.や 2s.d.で切断することは「統計学的に信頼のおけない方法」であると述べていて，トランケート分布に基づく方法を示しています。計算方法を導く偏微分方程式と，その為のプログラムも記載しています。切断された正規分布の分布関数は，蓑谷千凰彦「統計分布ハンドブック」朝倉書店（2003）に示されているように，下側切断点を xL，上側切断点を xU としたときに，正規分布の下側確率の差 F(xU)－F(xL)で確率密度関数を除した下記の最尤値を求めればよいのです。[注 1)]

$$f(x_i)=\frac{\frac{1}{\sqrt{2\pi\tilde{\sigma}^2}}\exp\left[-\frac{(x_i-\tilde{\mu})^2}{2\tilde{\sigma}^2}\right]}{\int_{-\infty}^{x_U}\frac{1}{\sqrt{2\pi\tilde{\sigma}^2}}\exp\left[-\frac{(x_U-\tilde{\mu})^2}{2\tilde{\sigma}^2}\right]-\int_{-\infty}^{x_L}\frac{1}{\sqrt{2\pi\tilde{\sigma}^2}}\exp\left[-\frac{(x_L-\tilde{\mu})^2}{2\tilde{\sigma}^2}\right]}$$

つまり， f (xi)の対数を取り，対数尤度の合計の lnL が最大となる μ （平均）と σ （標準偏差）を求めれば良いので，エクセルにソルバーと NORMDIST()関数でも求めることが出来ます。
エクセルでの例を次に示しますので，セルの式を参考にして下さい。

注 1）秋山功，富山茂巳，高橋行雄：「定量下限未満を含むデータの要約統計量と各種の統計解析」医薬安全性研究会，2016（公表準備中.）

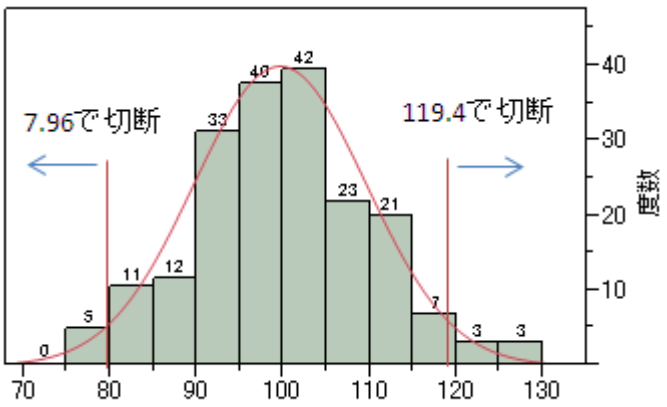
	A	B	C	D	E	F	G	H
1			下限切断	上限切断				
2		トランケート=	79.6	119.4				
3		μ=	99.2					
4		σ=	9.4	lnL=	−667.7851	←	lnLの合計	
5	No.	データ	密度	密度/分布	ln Li			
6	1	104	0.0372	0.0305	−3.2561	←	=NORMDIST(B6,$C$3,$C$4,FALSE)	
7	2	79	0.0043	0.0044	0.0000	←	=C7/(NORMDIST($D$2,$C$3,$C$4,TRUE)−NORMDIST($C$2,$C$3,$C$4,TRUE))	
8	3	101	0.0415	0.0430	−3.1454	←	=IF(AND($D$2>=B8,$C$2<=B8),LN(D8),0)	
9	4	92	0.0316	0.0327	−3.4202			
10	5	106	0.0327	0.0338	−3.3861			
11	6	97	0.0411	0.0426	−3.1551			

使用方法
①データを入力する。
②切断する下限と上限の位置を決める。
③μ、σの近い値を入れておく。
③ソルバーを立ち上げ、目的セルに「lnL」のセルを指定。(オレンジ色のセル)
④変数セルの変更に「μとσ」のセルを指定する。(黄色のセル)
⑤ソルバーの目標値は「最大値」を選択し「解決」をクリックする。
⑥ソルバーの「解決」で解を求める。

注)ソルバーはアドインで入れておく。

⑥ソルバーでμとσが計算される。
切断する範囲を多少変えても
同じような値が得られる。

例題を考えてみます。正規乱数で n=200　平均=100.0　s.d.= 10.0 (正規乱数)
でヒストグラムを作成してみました。
±2s.d.で 79.6 と 119.4 になる。この残りのデータで標準偏差を計算すると当然データが中心に集まり，標準偏差は小さくなり s.d.=8.6 が得られます。
79.6 と 119.4 で切断したと考え，トランケート分布として計算すると標準偏差は s.d.= 9.4 になり，データが全てある場合の 10.0 に近い値が得られます。

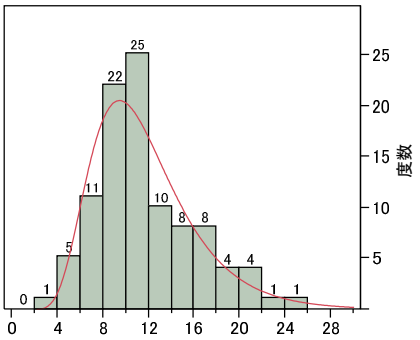


正規乱数 n=200 平均＝100 s.d.=10	理論値
	10

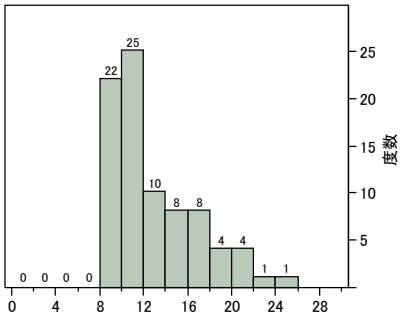
方法	推定したs.d.
±2s.d.で切り捨てて計算	8.6
±2s.d.で切断された**正規分布として計算**	9.4

19.2　打ち切りのある正規分布（censored normal distribution）

濃度分析では LOQ，LOD などので打ち切りのあるデータがあります。
下記の図は臨床検査の健常 100 人の検査データのヒストグラムです。
対数正規分布しています。
8 で打ち切りがあるとして，どのようにしたら全データがある場合の平均値と標準偏差を再現できるのかを考えてみます。

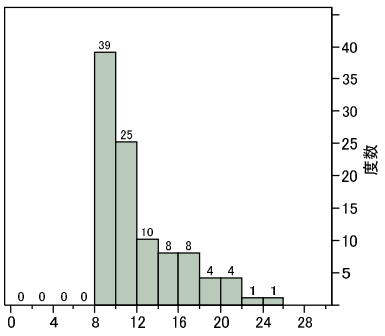


全データのヒストグラム

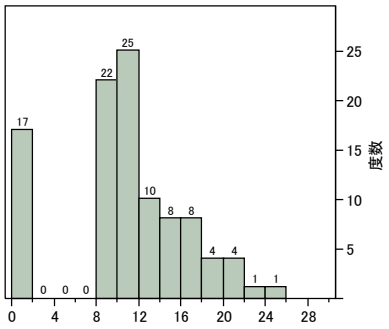


8 以下を無いとして除外するヒストグラム

分析技術者として打ち切りのあるデータを右上のように「無い」として扱ってはならないことです。[2)]
データを捨て去ることになり，データ解析上問題があり，倫理上も問題が生じることがあります。
そこで，下記のようなヒストグラムも考えられます。



8 以下を 8 として扱うヒストグラム



8 以下を 0 としたヒストグラム

打ち切りがある場合の平均と標準偏差をどのように計算すべきか迷うところです。
海外では定量限界の 1/2 を使用する方法も使用されます。

打ち切りのあるデータではどのようにすべきかを，例題のデータでどのようなになるのかを計算した表を下記に示します。

全データ	平均	標準偏差
	11.75	4.66

			平均	標準偏差
17%打ち切り	最尤法	JMP　寿命1変量	11.78	4.34
		R　NADA	11.78	4.34
		Excel	11.78	4.34
	便宜的方法	1/2で補完	11.51	5.99
		＜を除く	12.78	3.69
		LOQで補完	11.94	3.77
		0で補完	10.64	6.04

全データに近い，最適なのは最尤法による方法です。

1/2 で補完　:「打ち切り」では最尤法での推定が優れていますが，最低 1/2 で補完すべきです。

＜を除く　　：平均は高値に傾き，標準偏差は小さく推定されます。
LOQ で補完　：平均は高値に傾き，標準偏差は小さく推定されます。
0 で補完　　：平均は低値に傾き，標準偏差は大きく推定されます。

最尤法からの推定は統計ソフト JMP や R で行うことができ，最尤法を理解していれば，エクセルでも行うことができます。
JMP では寿命分析を利用できますが，R では library(NADA)が使用できます。
http://practicalstats.com
また，エクセルの Normdist() 関数を使用して計算することもできます。
方法の説明は下記を参考にして下さい。

1）秋山功，富山茂巳，高橋行雄 :「定量下限未満を含むデータの要約統計量と各種の統計解析 - 最尤法についての補足 - 」，医薬安全性研究会，2016（公表準備中）
2）Dennis R. Helse :「Statistics for Censored Environmental Data Using Minitab® and R, Second Edition」Wiley（2012 年）。

参考**臨床検査の基準値（正常値）**の計算について

患者データからの基準値の推定法として，反復切断補正法(Hoffmann 法，臼井法)や丹後の方法などがあります。反復切断補正法の 2s.d.や 3s.d.の標準偏差での反復切断は，経験的な係数の使用や方法そのものも不自然です。反復切断はやめて，丹後が示した確率密度関数として切断正規分布（truncated normal distribution）または打ち切りがある正規分布（censored normal distribution）を考えた方が自然です。
参考：丹後 俊郎：「臨床検査への統計学」朝倉書店（1986 年）

基準値は基本的に健常人のデータから推定し，「**健常人 95%の範囲**」と定義します。
しかし，健常人からの推定では，健常人の定義，年齢，性別，採取時間，地域，データに LT（以下）や GT（以上）など打ち切りがある場合などを考慮すると，臨床検査の基準値の設定は困難な多くの問題を含みます。(LT，GT での打ち切りがある場合は打ち切りがある分布になります。)
基準なのだから可能な限り質の良い多くのデータを集め，定義から考えて分布に影響されないノンパラメトリックな手法（パーセント点）が利用できれば理想的です。パーセント点とはデータの小さい順に並べて，この順序から 95%の範囲を求める方法です。
しかし，多くのデータを確保できない場合もあります。そこで，パラメトリックな手法（正規分布，対数正規分布などを仮定する）を使用します。一般にデータの 2s.d.の範囲を 95%の範囲として使用します。

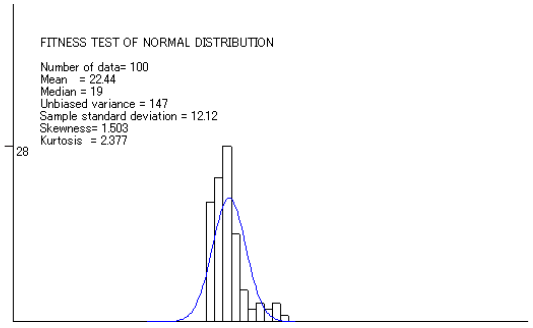
また，飛び離れたデータ(outlier)を取り除くことは，安易に Grubbs-Smirnov の棄却検定などですべきではありません。飛び離れたデータの人について調べ直し，医学的根拠から取り除くべきです。飛び離れたデータが重要な意味を持つ場合もあります。反復切断補正法の問題点，Grubbs-Smirnov の棄却検定は基本的には繰り返せないことなどをよく考慮して，飛び離れたデータを取り除きます。
それでは下記の表のデータで，先ほど述べた「第１８章　正規分布に変換する方法」の修正ベキ変換で臨床検査の基準値を算出してみます。
全て健常人で，各人の年齢，性別，健康診断データも得られているものとします。

データ

n=100

28	39	15	14	14	6	12	17	27	10
17	15	38	12	10	19	26	14	19	11
19	14	20	22	24	23	28	31	23	26
15	22	19	12	19	58	13	24	24	21
19	18	45	53	27	9	30	27	21	20
7	18	24	19	15	9	28	57	18	22
19	44	25	9	17	6	18	23	10	8
10	20	18	17	23	36	9	12	58	18
22	19	28	32	17	66	33	26	45	13
26	50	19	17	22	17	29	11	34	12

ヒストグラムを作成してみす。



データ数=	100	
平均値 =	22.4	
中央値 =	19	
最小値 =	6	
最大値 =	66	
分散 =	147	
標準偏差=	12.1	
2 SD 区間	: −1.3	<--- 46.2
ノンパラメトリック法	95%区間 6.53	<--- 58

正規性の検定

歪度の検定

歪度 =	1.5
Z1 =	6.26

Z1=6.259 > Z=3.29　P<0.001で正規性は棄却される。

尖度の検定

尖度 =	2.38
Z2 =	5.23

Z2=5.227 > Z=3.29　P<0.001で正規性は棄却される。

左に傾いていますので正規分布ではないようです。

正規分布の検定でも，正規性は棄却されました。

修正ベキ変換して，最尤法から変換式を探ってみます。

下記の式のλを変化させてMLLを計算すると

$$x_i^{(\lambda)} = \frac{x_i^{\lambda} - 1}{\lambda} + \lambda$$

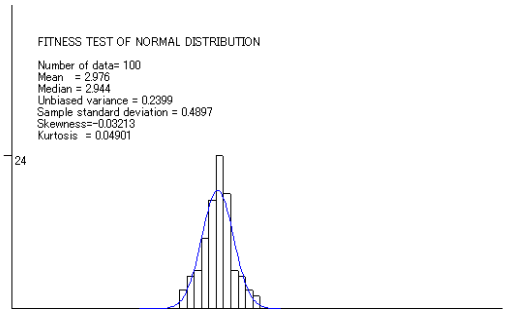
$$MLL(\lambda) = -\frac{N \log \sum_{i=1}^{n} (x_i^{(\lambda)} - \bar{x}^{-(\lambda)})^2}{2} + (\lambda - 1)\sum_{i=1}^{N} \log x_i$$

から

Power Transfomation	
λ	MML
−0.8	−239.987
−0.6	−235.143
−0.4	−231.713
−0.2	−229.749
0	−229.283
0.2	−230.328
0.4	−232.87
0.6	−236.873
0.8	−242.277
1	−249.004
2	−299.224
3	−369.109

MML＝－229.28 でλ＝0 が最大となり，対数変換が良いようです。対数変換してみます。

対数変換したデータで解析してみます。

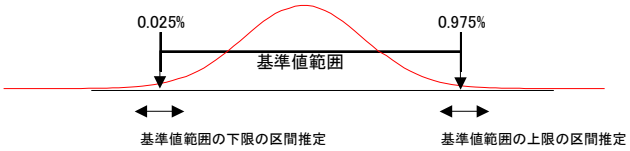


データ数=		100	
平均値 =		2.976	19.6
中央値 =		2.944	19
分散 =		0.2399	1.27
標準偏差=		0.4897	1.63
2SD 区間	:	1.996	<---> 3.955

正規性の検定	
歪度の検定	
歪度 =	-0.03213
Z1 =	0.1338
P>0.05で正規性を棄却することができない。	
尖度の検定	
尖度 =	0.04901
Z2 =	0.1078
P>0.05で正規性を棄却することができない。	

対数変換で，正規分布と見なせるようです。

推定した基準値（２ＳＤ区間）は点推定ですから誤差を伴います。その区間推定は下記の図のようになり，



下限，上限での区間推定は $(\bar{x} \pm 2s) \pm 2s\sqrt{\frac{3}{n}}$ で求められます。

$$xL = (\bar{x} - 2s) \pm 2s\sqrt{\frac{3}{n}} = 1.996 \pm 0.4897$$

$$xH = (\bar{x} + 2s) \pm 2s\sqrt{\frac{3}{n}} = 3.955 \pm 0.4897$$

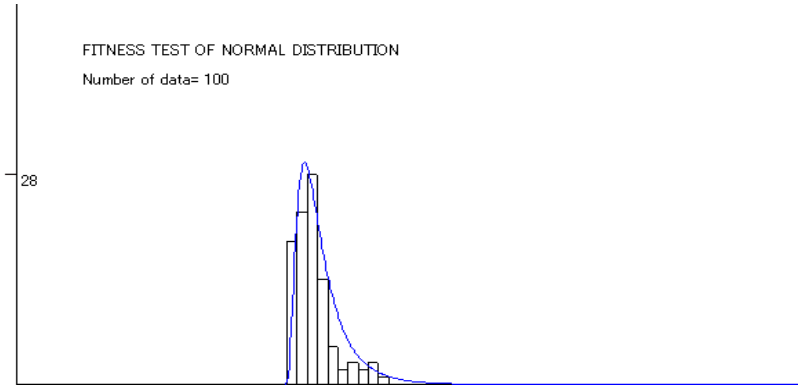
から

下限： 1. 826 から 2. 166

上限： 3. 785 から 4. 124

対数変換をしたので元に戻す必要があります。

元に戻すと下記のようになります。対数正規分布に重なり合います。

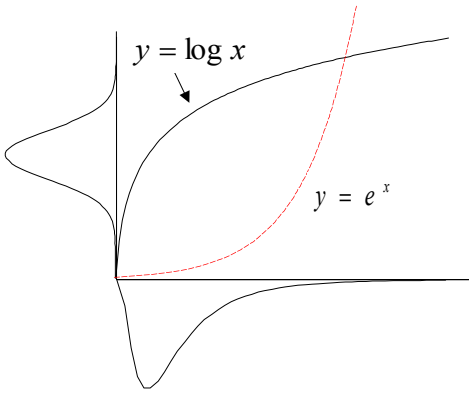


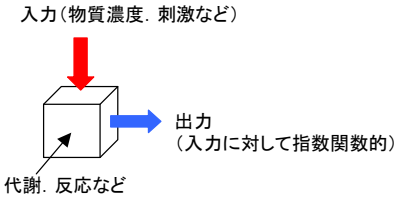
第１９章　切断した正規分布と打ち切りのある正規分布の平均と標準偏差

基準値	
下限 7.4　（6.2 から 8.7）	
上限 52.2　（44.1 から 61.8）	（）は基準値の 95%信頼区間
	幾何平均　19.6

基準値は 7 から 52 になりますが，95%信頼区間を考慮すると $\pm 2s\sqrt{\frac{3}{n}}$ の幅があり，6 から 62 になる可能性もあります。

生体内の反応，代謝は指数関数的な変化をすることが多く，指数関数的な変化は「2.4　幾何平均」「16.9　関数の当てはめと微分方程式」で述べたように，対数正規分布で近似できると推測されます。





修正ベキ変換で，最尤法から変換式を探しましたが，単に統計処理をして範囲を求めたにすぎません。

変換式には明確な医学的な意味はありません。何らかの代謝などを考慮した生体モデルを仮定した訳ではない点は注意すべきです。95%範囲のみに意味があります。

基準値は一度設定すると多くの患者データと比較し，多くの人の生命に関係するのであるから，質の良いデータを可能な限り多く集める必要があります。

参考：丹後 俊郎：「医学への統計学」朝倉書店　（1993 年）
養谷千鳳彦：「統計分布ハンドブック」 朝倉書店（2003 年）

第２０章　その他の統計学の利用について

統計学は広範囲ですが，最後に有効であると思われる統計手法を紹介します。
医学関係で使用される，オッズや生存分析などの説明は省略します。

20.1　直交配列実験など（分散分析的手法）

実験計画法の分散分析の一元配置と二元配置はエクセルの「分析ツール」でも計算できます。

データを d11 から d34 とすると配置は下記のようになります。

一元配置分散分析

要因 A	データ			
A1	d12	d12	d13	d14
A2	d21	d22	d23	d24
A3	d31	d32	d33	d34

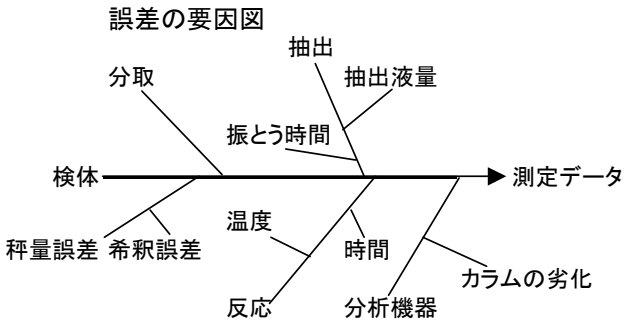
二元配置分散分析

要因 A, B	B1	B2	B3	B4
A1	d12	d12	d13	d14
A2	d21	d22	d23	d24
A3	d31	d32	d33	d34

要因が多くなると取る必要があるデータ数は膨大になります。そこで，一部実験法である直交配列実験が有効です。

濃度測定の前処理である秤量，定容，抽出，分液，分取，反応，測定などで測定誤差が生じます。この誤差の要因図を作成したときに，各要因によるデータに及ぼす影響を調べるのに直交配列実験は有効です。直交配列実験で，特に小規模な L8，L9 程度の実験を繰り返すことにより，日々の検査の中でも負担なく簡単に誤差の解析ができます。直交配列実験が日常的に広く利用されれば誤差要因は明確になっていきます。

交互作用がないことが大切ですが，サンプリングから測定までの誤差の要因を幾つか拾い出して検討していきます。



Ｌ８直交配列

	1	2	3	4	5	6	7	データ
1	1	1	1	1	1	1	1	d1
2	1	1	1	2	2	2	2	d2
3	1	2	2	1	1	2	2	d3
4	1	2	2	2	2	1	1	d4
5	2	1	2	1	2	1	2	d5
6	2	1	2	2	1	2	1	d6
7	2	2	1	1	2	2	1	d7
8	2	2	1	2	1	1	1	d8

例えば，要因（温度，ｐＨなど）が６あり，２水準（10℃，20℃など）であると，2^6で６４個のデータを取る必要があります。L_8（2^7）で交互作用が無視でき，７列目を誤差に割り振れば８個のデータでよく，$\frac{8}{64}=\frac{1}{8}$で８分の１の実験ですみ効率的です。[注１）]

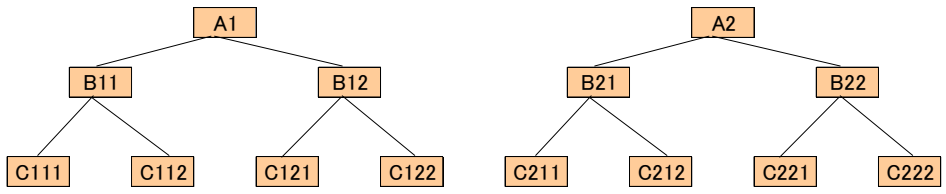
分散分析法である「枝分かれ配置」は分割法で，従来から使用されてきましたが，ISO の不確かさの推定としても使用されます。

測定データの誤差の構造が枝分かれ配置になっていることが多く，濃度測定では使用頻度が高い統計手法です。

例えば測定で，１日目と２日目を A，２台の分析機器を B，２重測定を C とすれば，８個のデータは，下記のような構造から得られたことになります。

実験計画，枝分かれ実験については，「第１５章　測定精度の推定方法」で述べました。

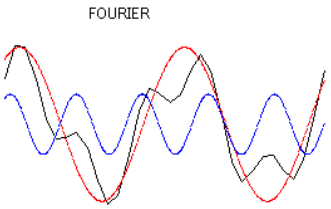
枝分かれ実験（nested dsign）



注１）中里，川崎，平栗，大滝：「品質管理のための実験計画法テキスト」日科技連
近藤，船阪編：「技術者のための　統計的方法」共立出版
奥野 忠一，芳賀 敏郎：「実験計画法」 培風館　　など

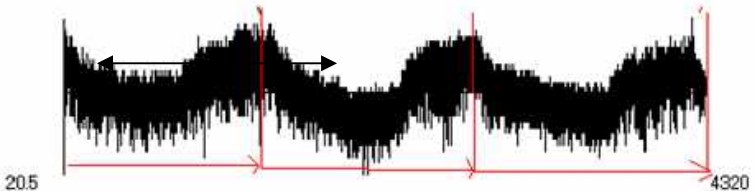
20.2　スペクトル解析

ルーチン的な仕事では毎日同じ検査項目を測定しますが，それが時系列データになりスペクトル解析の利用を可能にします。測定値の他にも検査では多くの時系列データがあります。注１）
下記の本を参考にしてパソコン用のプログラムを作製し，解析した図を示します。

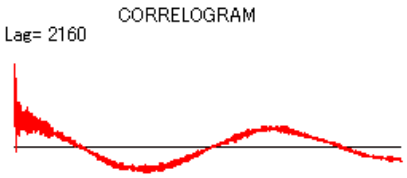


日野 幹雄：「スペクトル解析」朝倉書店（1977 年）

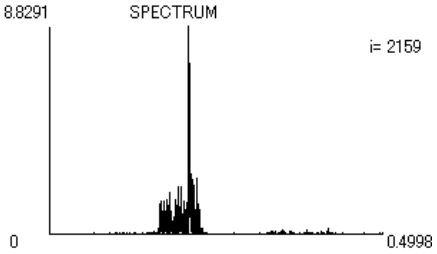
元データ



自己相関係数



スペクトル解析 MEM（最大エントロピー法）



注１）**時系列分析**について

ある程度の臨床検査センターでの検体数（病院の患者数）の予測などは，時系列分析により驚くほど一致します。

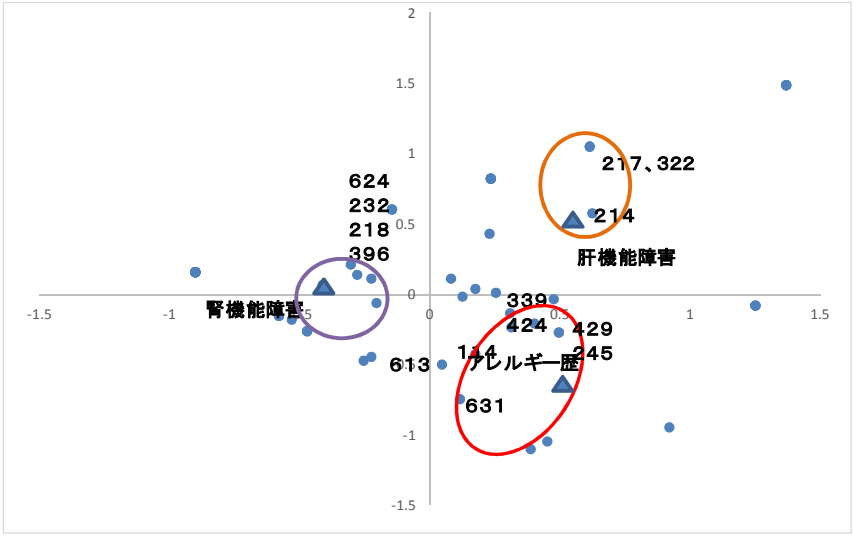
時系列分析として，TCSI 法は T(Trebd)傾向，C(Cycle)循環，S(Seasonality)季節，I(Irregularity)不規則に分解した後，それを合成します。ＡＲＭＡモデル（自己回帰移動平均モデル）なども将来的な予測に利用できます。

参考）尾崎タイヨ：「計量モデル分析と数値計算法」ホルト・サウンダース（1985 年）
溝口 敏行，刈屋 武昭：「経済時系列分析入門」日本経済新聞社（1983 年）　　など

20.3　多変量解析 multivariate analysis

下記の本は理解し易く一読する価値があります。
奥野 忠一，久米 均，芳賀 敏郎，吉澤 正：「多変量解析法」日科技連（1971 年）
この本以外にも多くの多変量解析の本が出版されています。多変量解析には色々な解析方法があります。
パソコンの普及により，多変量解析が誰でも簡単に使用できる時代になったので，現場にあるいろいろなデータについて解析してみることを勧めます。環境検査や臨床検査などでは毎日多くのデータが出力されます。
解析すれば，新たな発見があるかも知れません。

対応分析（数量化Ⅲ類）の実例



おわりに

濃度分析技術者として基礎となる部分で，データ解析の入り口にすぎないものですが，現場に即したものを説明しました。

説明不足で理解できない部分もあると思いますが，単なる手順書や数式での説明ではない，納得できるような記述を試みたつもりです。

手順を知ることではなく，少しでも原理を納得して使用することが大切なようです。

何故なら，実際のデータ解析では入門的な統計学の手順書では対応する手法が見つからないなど，簡単なデータ解析では処理できない場合も少なくありません。

ここで述べた内容が多少でも役立てばと思います。

付録１　本の紹介

本文を読んで，さらに疑問に感じることがあるかも知れません。

また，本文では難解な統計手法を紹介している部分もあります。

そこで，本を少し紹介しておきます。（注）下記以外に，本文中でも本を紹介しています

１）　遠山 啓：「数学入門」岩波書店（1960 年）

２）　松坂 和夫：「数学読本」全 6 巻 岩波書店（1990 年）

３）　高木 貞治：「解析概論」岩波書店（1983 年）

４）　日本数学会　編集：「数学辞典 第 4 版」岩波書店（2008 年）

５）　原田 耕一郎：「群の発見 (数学，この大きな流れ)」岩波書店 (2001 年)

６）　David Burghes，Morag Borrie （垣田 高夫，大町比佐栄　訳）：
「微分方程式で数学モデルを作ろう」日本評論社（1990 年）

７）　佐藤 總夫：「自然の数理と社会の数理」日本評論社（1984 年）

８）　「統計数値表」JSA-1972　日本規格協会（1972 年）

９）　橋本 洋志，石井 千春，山浦 富雄，大山 恭弘：
「微分方程式＋モデルデザイン教本」オーム社（2003 年）

１０）鈴木 七緒，安岡 善則，志村 利雄：
「詳解　確率と統計演習」共立出版（1979 年）

１１）小針 晛宏：「確率・統計入門」岩波書店（1973 年）

１２）山田 剛史，杉澤 武俊，村井 潤一郎：
「「R」によるやさしい統計学」オーム社（2008 年）

１３）P.G.ホーエル（浅井，村上　訳）：「入門数理統計学」培風館（1978 年）

１４）守谷 栄一：「詳解.演習　数理統計」日本理工出版会（1998 年）

１５）永田 靖：「サンプルサイズの決め方」朝倉書店（2000 年）

１６）吉澤 康和：「新しい誤差論」共立出版（1989 年）

１７）佐和 隆光：「回帰分析」朝倉書店（1979 年）

１８）渡辺 洋：「ベイズ統計学入門」福村出版（1999 年）

１９）奥野 忠一，久米 均，芳賀 敏郎，吉澤 正：「多変量解析法」日科技連（1971 年）

２０）日野 幹雄：「スペクトル解析」朝倉書店（1977 年）

２１）市原 清志：「バイオサイエンスの統計学」南江堂（1990 年）

２２）丹後 俊郎：「臨床検査への統計学」朝倉書店（1986 年）

２３）Dennis R. Helse：「Statistics for Censored Environmental Data Using Minitab® and R, Second Edition」Wiley（2012 年）

２４）芳賀敏郎：「-医薬品開発のための-　統計解析」全 3 巻，サイエンティスト社，(2016 年)

付録２　無料パソコンソフトの利用

データ処理のソフトとしてエクセルを使用することが多いのではないでしょうか。

グラフの作成, 数式処理, 統計解析のソフトで, 利用価値のあると思う gnuplot, Maxima, R を紹介します。これらのソフトは無料ですが，データ解析に対して十分な能力を持っています。世界中の多くの研究者の献身的な努力により，さらに強力で利用し易いソフトになって行くと思われます。すばらしいソフトですので，使用してみることを勧めます。

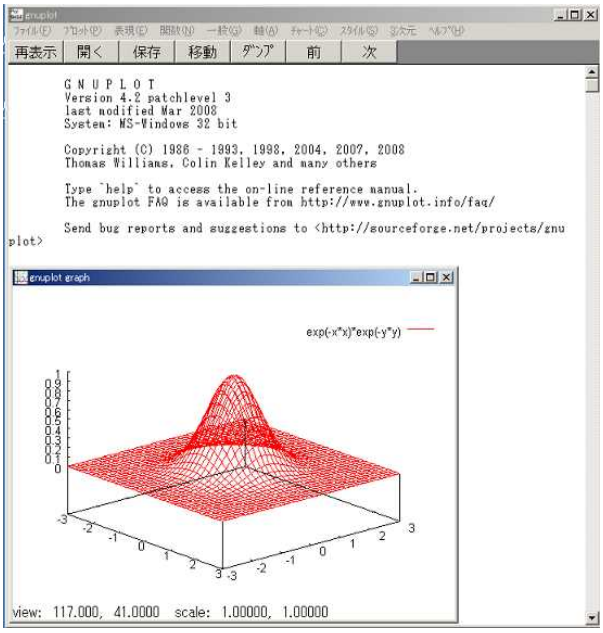
gnuplot

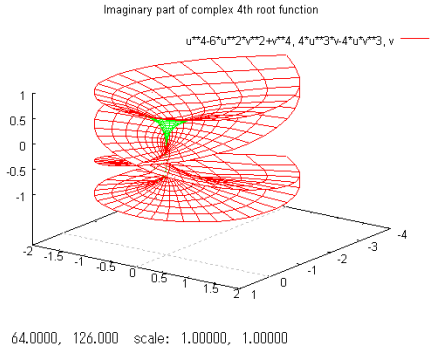
gnuplot は，下記のように数式を入力すれば簡単にグラフを作成できます。強力なグラフ作成ソフトです。

ダウンロードは下記ホームページから行えます。

http://www.gnuplot.info/

gnuplot の fit 関数で Levenberg-Marquadt 法による非線形最小 2 乗法を解くこともできます。





http://gnuplot.sourceforge.net/demol

_4.2/singulr.htmgnuplot demo script: singulr.dem

Maxima

Maxima は数式処理ソフトです。微積分の数式から一般解を求めたり，確率分布の計算もできます。有名な Mathematica などは高価であり，個人での購入は困難です。

無料の Maxima は高価な Mathematica に迫る処理能力を持っています。

ダウンロードは下記ホームページから行えます。

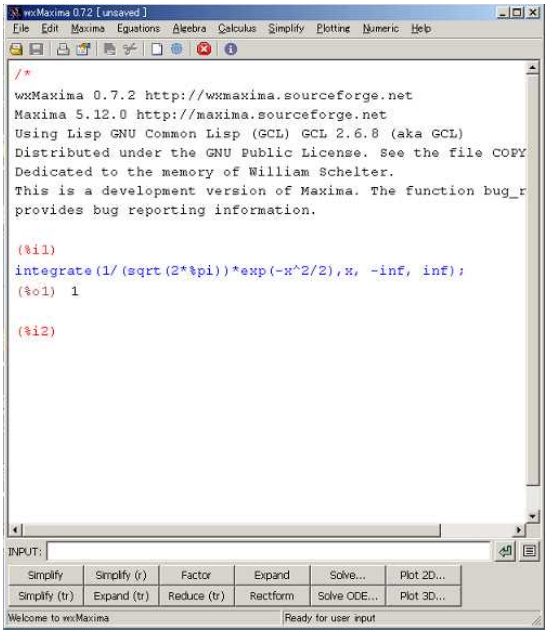
http://wxmaxima.sourceforge.net/

例えば，標準正規分布の式が積分して 1 になることを確認してみます。

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{x^2}{2}\right)} = 1$$

下記のように入力すると結果 1 が出力されます。

```
integrate(1/(sqrt(2*%pi))*exp(-x^2/2),x, -inf, inf);
1
```



注 1）井上 勝：「定義域を伴った関数の数式処理」応用統計学　Vol.37.No.1.17-35(2008 年)

Maxima の使用説明として下記のものが理解しやすい。

中川 義行　Maxima 入門ノート　1.2.1　（2005 年）

http://www.eonet.ne.jp/~kyo-ju/maxima.pdf

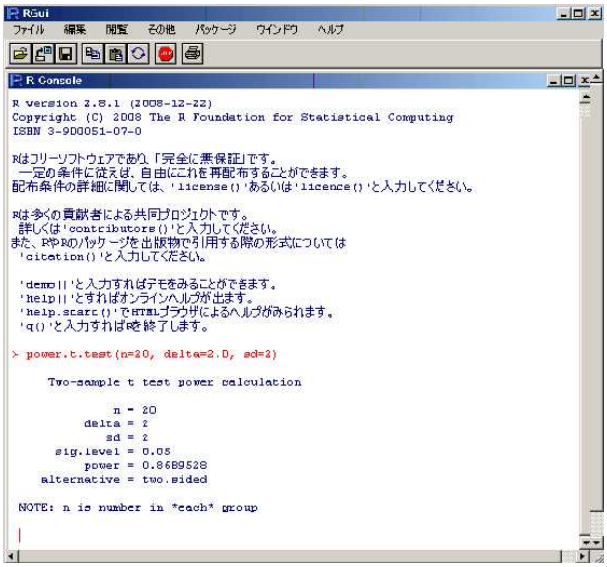
R

Rは統計解析ソフトです。

エクセルは統計ソフトではありませんが，R は本格的な統計処理が可能なソフトで，信頼性も確保できます。[注1）]

ダウンロードは下記ホームページから行えます。

http://www.r-project.org/



注1）山田 剛史，杉澤 武俊，村井 潤一郎：「「R」によるやさしい統計学」オーム社（2008 年）

R はオープンソースですからプログラムを見ることが可能です。

検出力の計算例を示します。
統計解析ソフトRの関数 **power.t.test()**を使用します。

１）検出力を求める

本文の例：データ数(n)，平均値の差(delta)，s.d.=2 を＞の後に，下記のように入力します。

```
> power.t.test(n=20, delta=2.0, sd=2)

     Two-sample t test power calculation

              n = 20
          delta = 2
             sd = 2
      sig.level = 0.05
          power = 0.8689528
    alternative = two.sided

 NOTE: n is number in *each* group
```

検出力(power)=0.8689528 が出力されます。

２）必要なデータ数を求める

本文の例：平均値の差 10-8＝2，s.d.＝2，検出力 0.8 を＞の後に，下記のように入力します。

```
> power.t.test(n=NULL, delta=2.0, sd=2,power=0.8)

     Two-sample t test power calculation

              n = 16.71477
          delta = 2
             sd = 2
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

 NOTE: n is number in *each* group
```

データ数はｎ=16.71477，つまり，データ数は 17 個必要であることを示しています。

R には一元配置分散分析 power.anova.test()，比率の検定 power.prop.test()の検出力を算出する関数もあります。

付録３　プログラム（検出力の計算）

片側検定

Microsoft エクセル

関数　＝pow１(データ数 NA,データ数 NB,有意差 D,有意水準 A)

```
Function POW1(NA, NB, D, A) As Double
'
Dim df As Integer
Dim RA, TD, PR2, PRA As Double

df = NA + NB - 2
A = 2 * A
'RA = Sqr(NA * NB / (NA + NB)) * D
TD = Application.WorksheetFunction.TInv(A, df)
PR2 = TD * (1 - (1 / (4 * df))) - RA
PRA = 1 + TD * TD / (2 * df)
PRA = Sqr(PRA)
PR2 = PR2 / PRA
PR2 = Application.WorksheetFunction.NormSDist(PR2)
 POW1 = 1 - PR2

End Function
```

両側検定

Microsoft エクセル

関数　＝pow２(データ数 NA,データ数 NB,有意差 D,有意水準 A)

```
Function POW2(NA, NB, D, A) As Double

Dim df As Integer
Dim RA, TD, PR1, PR2, PRA As Double

df = NA + NB - 2
RA = Sqr(NA * NB / (NA + NB)) * D
'TD = Application.WorksheetFunction.TInv(A, df)
PR1 = -TD * (1 - (1 / (4 * df))) - RA
PRA = 1 + TD * TD / (2 * df)
PRA = Sqr(PRA)
PR1 = PR1 / PRA
PR1 = Application.WorksheetFunction.NormSDist(PR1)
PR2 = TD * (1 - (1 / (4 * df))) - RA
PR2 = PR2 / PRA
PR2 = Application.WorksheetFunction.NormSDist(PR2)
  POW2 = PR1 + 1 - PR2

End Function
```