統計解析のはなし　大平　平

バスの発車時刻 8 時 45 分，牛肉 100g，ウエスト 79cm，
制限速度 60km のように公然とかかげている数値を**公称値**というのですが，公称値の中には当てにならないものが多いのです．
・・・
　各人の責任に帰することのできない理由によって誤差が発生し，またその程度の誤差はがまんができるとき，**公称値には誤差が含まれものが許されている**という趣旨をくだくだと書いてきました．

　統計学は，だいぶ前までは記述統計学と推測統計学とに二大別して説明されるのがふつうでした．記述統計学は大量の事実を観測して大量のデータを集め，それを集計して見やすく使いやすい形の表やグラフにして表現するための学問で，大量の数字の効果的な集め方，そのじょうずな整理のしかた，結果を表現するさまざまな方法などの理屈や手法で構成されています．いっぽう，**推測統計学**，略して**推計学**は，全体の中から取り出された一部の見本の性質を調べ，その結果から全体の性質を確率的に推しはかるための学問で，推定と検定とを 2 本の柱とし，その応用としての抜取検査法や実験計画法なども含めて取り扱うのがふつうでした．

抜取検査
抜き取った標本も性質から全体の性質の良否を判定する

正規分布の加法性

さて，ここに 2 つの集団があり，それぞれ

$N(\mu_1,\ \sigma_1^2)$

$N(\mu_2,\ \sigma_2^2)$

の正規分布をしているとしましょう．このとき，両方の正規分布から 1 つずつの値を取り出して

$N(\mu_1,\ \sigma_1^2)$から取り出された値

$+N(\mu_2,\ \sigma_2^2)$から取り出された値

という新しい値を作ることをくり返すと，この新しい値は

$N(\mu_1+\mu_2,\ \sigma_1^2+\sigma_2^2)$

という正規分布をする性質があり，これを正規分布の加法性*というのでした．

標本から計算した標準偏差は，

母集団のほんとうの標準偏差よりも小さい値になる傾向があり，標本の数が少ないほどその傾向が著しいからです・

その理由は，つぎのとおりです．いまかりに，平均値も標準偏差もわからない正規分布から，２つの標本

$$1,\quad 5$$

が得られていたとします．この２つの標本の平均値は

$$\bar{x}=\frac{1+5}{2}=3$$

ですから，標準偏差は

$$s=\sqrt{\frac{(1-3)^2+(5-3)^2}{2}}=2$$

です．平均値を３として計算するとこうなるのですが，けれども，母集団のほんとうの平均値が３であるという保証はどこにもありません．前にも書いたように，標本から求めた平均値は母集団の平均値の不偏推定値ではありますが，それは，母集団の平均値が３より大きい確率と３より小さい確率が等しいといっているだけであり，母集団の平均値がぴったり３であると保証しているわけではないのです．実際には，母集団の平均値は３よりもいくらか大きかったり小さかったりする可能性が大です．

もしも，ほんとうの平均値が３ではなくて４であったらどうでしょうか．そのときには，母集団の標準偏差は

$$\sqrt{\frac{(1-4)^2+(5-4)^2}{2}}\fallingdotseq 2.24$$

とみなすのが公平なところです．また，ほんとうの平均値が３ではなく１であったとすれば，母集団の標準偏差は

$$\sqrt{\frac{(1-1)^2+(5-1)^2}{2}}\fallingdotseq 2.83$$

とみなさなければなりません．いずれにせよ，ほんとうの平均値が３でないとすれば，平均値を３として計算した

$$s=2$$

よりは大きな値になってしまいます．いいかえれば，$s=2$ は母集団の標準偏差を表わす値としては小さく見積もりすぎていることは明らかです．なにしろ，s を計算していく過程を見ていただくとわかるように，２つの値との差の合計がもっとも小さくなるように，いいかえれば２つの値との差を２乗した値の合計がもっとも小さくなるように，標本が自分たちだけで架空の平均値を決めてしまっているのですから……．

記号

母平均	μ
母分散	σ^2
母標準偏差	σ
母平均の不偏推定値	$\hat{\mu}$
母分散の不偏推定値	$\hat{\sigma}^2$
母標準偏差の不偏推定値	$\hat{\sigma}$
標本平均	$\bar{x}$
標本分散	s^2
標本標準偏差	s
標本の値	$x_1, x_2, \cdots, x_i, \cdots$

「標本から求めた平均値は母集団の平均値の不偏推定値である」

母集団の平均値や標準偏差は神様にしかわからない値なのでギリシア文字で表わすと前に書きましたが，これに対して，標本の平均値や標準偏差は計算すれば容易に知ることができる値なので，ありふれたローマ字で表わすのがしきたりです．

結論を先に書くと，n 個の標本がある場合には

$$\hat{\sigma}^2=\frac{n}{n-1}s^2 \qquad (2.1)$$

であることがわかっています．

2つの標本は

1，5

であったのですが，2つの標本の平均を3と決めてしまうと，いっぽうの標本が1なら他方は自動的に5に決まってしまうし，いっぽうが5であれば他方は1でなければならないので，2つあるように見える標本も実は1つしかないのと同じことになってしまうのです．このように標本から算出した平均値を使うことによって，標本の数が1つだけ減少したのと同じ効果が発生するのですから，s を求める過程で，標本の数 n で割って

$$s^2=\frac{\Sigma(x_i-\bar{x})^2}{n} \qquad (2.2)$$

とするのではなく，標本の数が $n-1$ であるとみなして

$$s^2=\frac{\Sigma(x_i-\bar{x})^2}{n-1} \qquad (2.3)$$

とすれば，勝手に平均値を作り出して使ったために s が小さくなりすぎる欠点を除去できようというものです．こういうわけなので，式(2.3)の s^2 を σ^2 の代用として，つまり $\hat{\sigma}^2$ として使うことにすると

$$\hat{\sigma}^2=\frac{\Sigma(x_i-\bar{x})^2}{n-1}$$

この本では、(2.2)を使う。

自由度：標本から作り出している使用している平均値の数を標本の数から差し引いた値

$\phi = n-1$

ϕ は f に相当するギリシャ文字。 f は degree of freedom（自由度）。

ϕ は標本由来なのに、文字はギリシャ文字なのは、統計数学では f は frequency（頻度）を使うため。

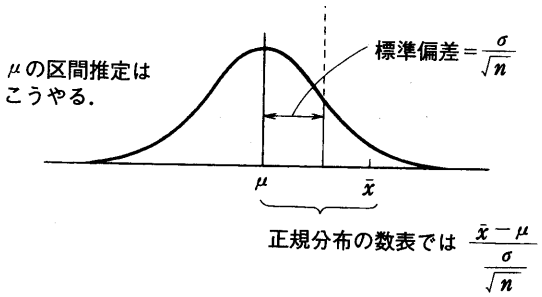


図 2.4　規準化された正規分布表を使うために

x バーが μ の周りに $\sigma/\sqrt{n}$ の標準偏差で正規分布する。

σ がわからない場合、 σ の代わりに σ ハットを使う。しかし、s は正規分布しない。（なぜ？）

$$\hat{\sigma}=\sqrt{\frac{n}{n-1}}s \qquad (2.5)$$

を使えばよいことに気がつきました*．そこで，式(※)の σ の代わりに式(2.5)の $\hat{\sigma}$ を代入すれば

$$\frac{\bar{x}-\mu}{\sqrt{\frac{n}{n-1}}\frac{s}{\sqrt{n}}}=\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n-1}}}$$

s^2 と s の分布とでは形が異なる。

＊　母分散の不偏推定値 $\hat{\sigma}^2$ は式(2.1)のように $\hat{\sigma}^2=\frac{n}{n-1}s^2$ です．この式の両辺を平方に開くと $\hat{\sigma}=\sqrt{\frac{n}{n-1}}s$ となりますが，これは数式の形だけのことであり，σ の不偏推定値 σ は $\sqrt{\frac{n}{n-1}}s$ ではありません．s^2 の分布と，それを平方に開いた s の分布とでは形が異なるからです．詳しくは，『統計のはなし』(改訂版) 109 ページを見てください．

母分散未知は、 t 分布を使う。

$$t=\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n-1}}}$$

s の分布を調べるのには√がついているぶん面倒。

s ではなく s^2 の分布を調べる。

σ^2 ハットと s^2 には(2.1)式の関係（比例）がある。

s^2 と S（偏差平方和）は比例する。

よって、σ²ハットとS（偏差平方和）は比例する。
σ²ハットは、σの不偏推定値なので、σ²とS（偏差平方和）は比例する傾向があるはず。

$$\chi^2=\frac{S}{\sigma^2} \qquad (3.2)$$

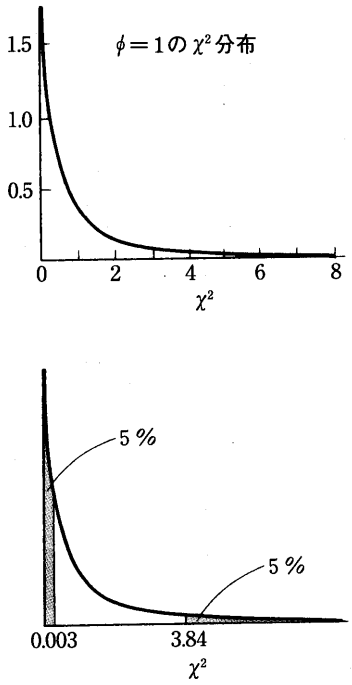


図3.1　怪　な　分　布

不偏分散 V_1、V_2

$$\left.\begin{aligned}\frac{n_1}{n_1-1}s_1^2=V_1\\ \frac{n_2}{n_2-1}s_2^2=V_2\end{aligned}\right\} \qquad (3.6)$$

略記することにします．V_1 と $\hat{\sigma}_1^2$，V_2 と $\hat{\sigma}_2^2$ とは同じ値ですが，偏りのない分散，つまり**不偏分散**という感じで使うときには V_1 とか V_2 とか書き，分散の偏りのない推定値というように「推定値」であることの思い入れを強調したい折には $\hat{\sigma}_1^2$ や $\hat{\sigma}_2^2$ と書くのが統計数学のしきたりです．

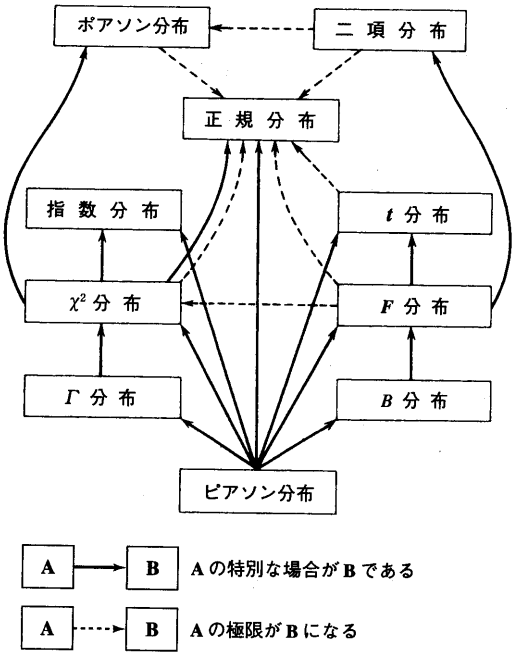
F 分布

$$F=\frac{V_2/\sigma_2^2}{V_1/\sigma_1^2} \qquad (3.7)$$

二項分布で n を無限大にした極限が正規分布
F 分布の特殊な場合が t 分布
χ^2 分布の特殊な場合が t 分布

付録2　分布間の相互関係

この本には，正規分布，t 分布，χ^2 分布，F 分布など，たくさんの分布が現れましたが，これらは下図のような親戚関係で結ばれています．



分布間の相互関係

（大村　平，今田直孝：『推測統計の FORTRAN』，オーム社，1972，より）

標本に占める割合から母集団の中の割合を区間推定する。→グラフ。

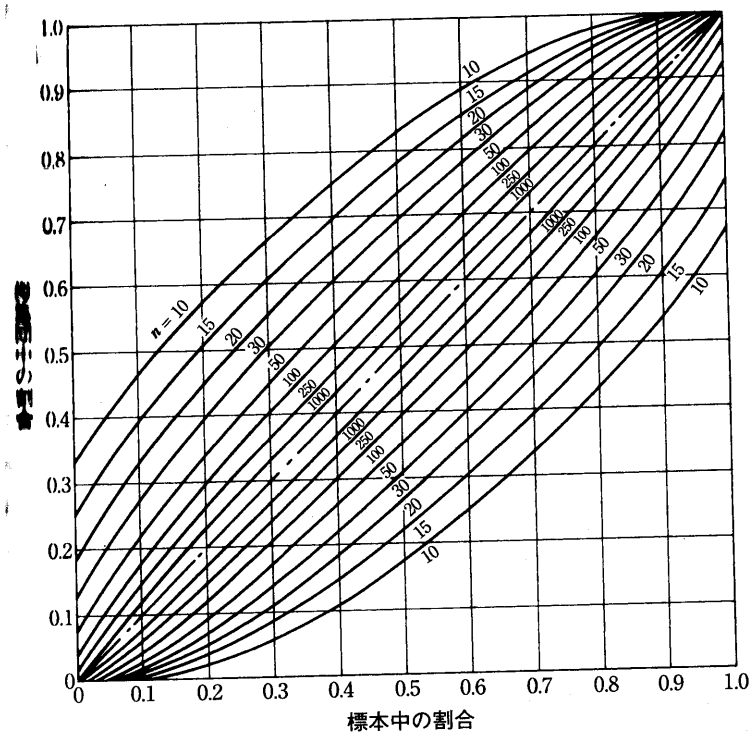


図 3.6　割合の区間推定のグラフ

こういう考え方，つまり，偶然によってはめったに起こらないような事象が起こったならは，偶然に起こったのではなく，その事象

を起こすような原因が存在したと信じようという考え方に従って，
ある現象や効果が何らかの原因から必然的に生じているかどうかを
統計的に調べる手法を**検定**といいます．そして，検定は推定と並ん
で統計解析のもっとも基本的な手法です．

したがって，この検定がまちが
った結論を出している確率が 3.125%だけあることは明らかです．
　こういうわけですから，5%以下の確率は小さい確率とみなすと
いう私たちの約束に従って検定を行なえは，その検定がまちがった
答えを出す確率が 5%はあることになります．このように検定がまち
がってしまう確率を**危険率**といいます．
　危険率は検定がまちがった答を出す確率ですから，こんなものは
小さいほうがいいに決まっています．

したがって，検定の危険率を
小さくすることは「石橋を叩いて渡る」ことに相当し，危険率を小
さくしすぎると「石橋を叩いても渡らない」ことになってしまいま
す．で，検定がまちがってもたいして困らないような場合には危険
率を 5%にし，**まちがうと被害甚大なとき**には 1%とか 0.1%とか
の危険率をとるのがふつうです．

ある事実を否定するような仮定を設けてみると思いがけない矛盾が発生するの
で，これは仮定がまちがっているにちがいないと思い返して事実を肯定するよう
な証明法を**背理法**といいます．

無に帰することを期待している仮説という意味で**帰無仮説**と名付けている。

食い違いの大きさを下式で表す。

$$\Sigma \frac{(\text{実現値} - \text{期待値})^2}{\text{期待値}}$$

分母が期待値の 2 乗ではない理由：偏差平方和の平均値は、データの数の 2 乗にではなく、データの数に正比例するので、データの数の影響を除去するためには期待値で割るのが正しい。

データが少なすぎると χ^2 検定に誤差が生じる。

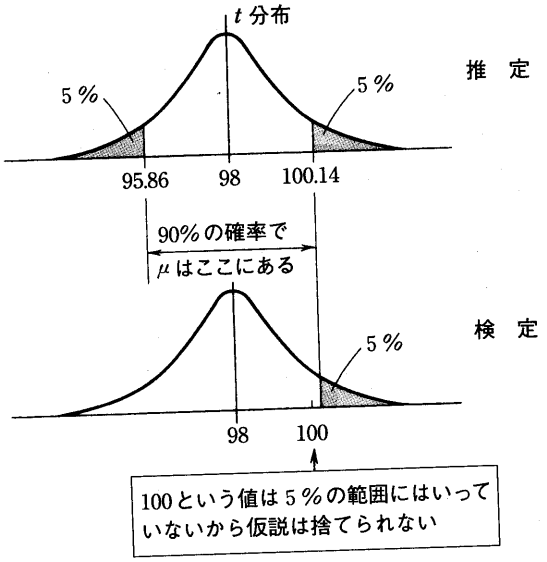


図 4.3　推定と検定は兄弟どうし

はげ
しい、ランニングの途中で心臓発作を起こして落命したり，マイホームを購入したためのローンに追われてみじめな生活を送ったり……，これらは，いずれも**手段を目的と取り違えたところに起きる悲劇**です．病気の治療は健康を回復するという目的を達成するための手段なのに，病気の治療そのものを目的にしてしまったり，体力を推持する手段としてのランニングなのに，走ることだけを目的にしてしまい，体力どころか命まで失ったり，楽しい生活を送るのが目的で購入するマイホームなのに，マイホームを入手することだけが人生の目的であるかのように錯覚してしまったり……手段を目的と取り違えることのないよう，お互いに注意したいものです．

ポアソン分布

n 個の標本を取り出すと，その標本の中に r 個だけ不良品が含まれている確率は

$$P(r)=\frac{(np)^r}{r!}e^{-np} \qquad (5.1)$$

で表わされます*.

＊　これは，母集団が大きく，p が小さく（0．1 以下くらい），n が大きく（50 以上ぐらい），そして期待値 np が 0～10 くらいのときに使用できる近似式でポアソン分布といいます.

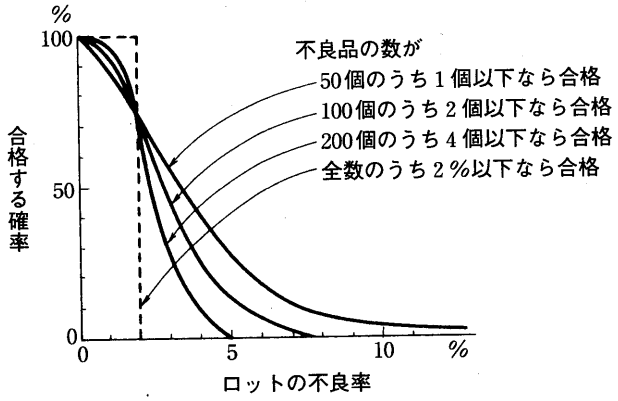


図 5.4　抜取個数によって OC 曲線はこう変わる

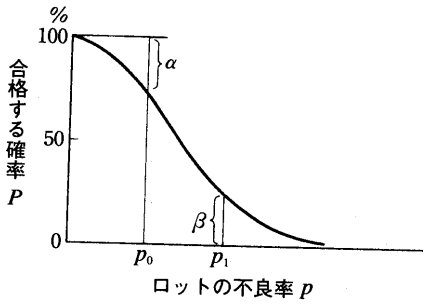


図 5.5　リスクを分担しあう

生産者リスク α、消費者リスク β

2 回抜取検査

では，「50 個中の不良品がゼロなら合格，不良品が 2 個以上なら不合格，不良品が 1 個ならばさらに 50 個を検査してその中の不良品がゼロなら合格」という 2 回抜取検査で不良率 p のロットが合格する確率はどうかと調べていきます．ごめんどうでも 138 ページをめくって式(5.2)を思い出していただきます．

$$P(r)=\frac{m^r}{r!}e^{-m}$$　　(5.2)と同じ

この式は，抜取個数が n であるとき

$$np=m$$

とおいて，不良品がちょうど r 個であるような確率 $P(r)$ を求める式でした．つまり，

不良品がゼロの確率　$P(0)=\frac{m^0}{0!}e^{-m}=e^{-m}$

不良品が 1 個の確率　$P(1)=\frac{m^1}{1!}e^{-m}=me^{-m}$

ということになります．そうすると，私たちの 2 回抜取検査に合格する確率 P は

$$P=\underset{\substack{\uparrow\\ \text{1回めの抜き取りが不良品ゼロである確率}}}{e^{-m}}+\underset{\substack{\uparrow\\ \text{1回めの抜き取りが不良品1個である確率}}}{me^{-m}}\times\underset{\substack{\uparrow\\ \text{2回めの抜き取りが不良品ゼロである確率}}}{e^{-m}}$$　　(5.3)

と表わされることになります．

一例として，$p=0.01$，つまり，1% であるとしてみましょうか．

$m=np=50\times0.01=0.5$

$e^{-m}=e^{-0.5}\fallingdotseq0.6065$　（表 5.1 から）

ですから，不良率 1% のロットが私たちの 2 回抜取検査に合格する確率は

$P=0.6065+0.5\times0.6065\times0.6065\fallingdotseq0.790$

自由度：標本から作り出している平均値の数を標本の数から差し引いたもの。
平均値が５つある場合、４つの平均値を決めれば他の１つは自動的に決まってしまう。この場合は、５ではなく４を引かれる数にする。

自由度＝（行の数－1）（列の数－1）

3 因子（分散分析）

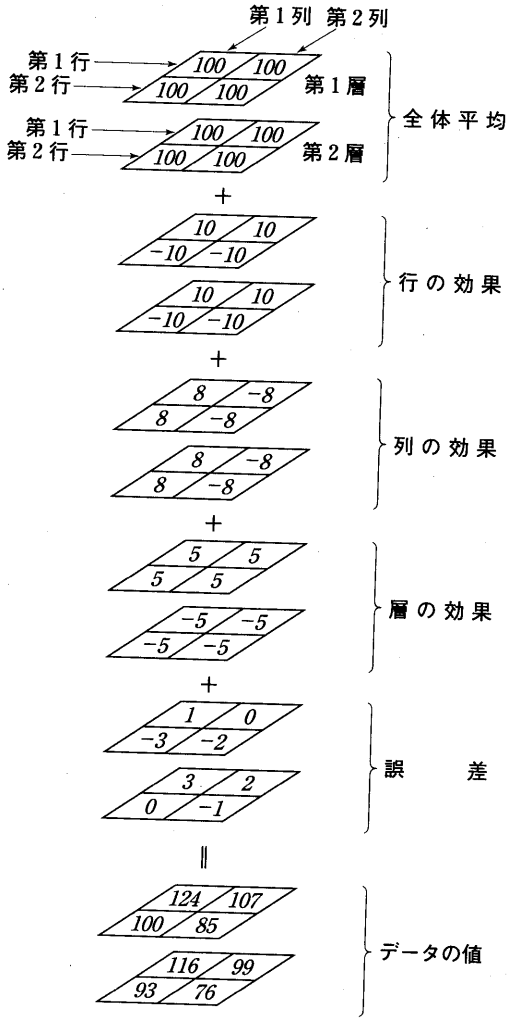
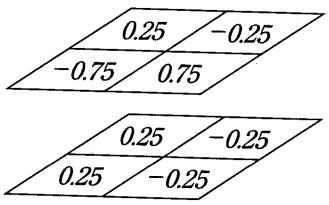


図 6.1　3 因子の場合の種作り

表 6.11　こうして誤差を分離する



では，行や列や層の効果に有意差があるかどうかを検定してみましょう．まず，誤差の不偏分散 V_1 は

$$V_1=\frac{0.25^2+(-0.25)^2+\cdots\cdots+(-0.25)^2}{4}=0.375\ *$$

です．

つぎに，行の効果を不偏分散 V_2 を求めると

$$V_2=\frac{11.5^2\times4+(-11.5)^2\times4}{2-1}=1058$$

です．そうすると，このときの F の値は

$$F=\frac{V_2}{V_1}=\frac{1058}{0.375}\fallingdotseq2821$$

であるのに

$$F_1^1(0.05)=161$$

ですから，ゆうゆうと行の効果は‘あり’と判定されます．

＊　分母の 4 は誤差の自由度です．8 つの誤差を求めるのに使われた平均値は，全体平均が 1 つ，行と列と層の平均が 2 つずつの計 7 つですが，行平均，列平均，層平均の各合計が全体平均の 2 倍と同じという 3 つの拘束条件があるので，自由度を減らす平均値は 4 つです．で，自由度は 8−4=4

——自由度の難しさ

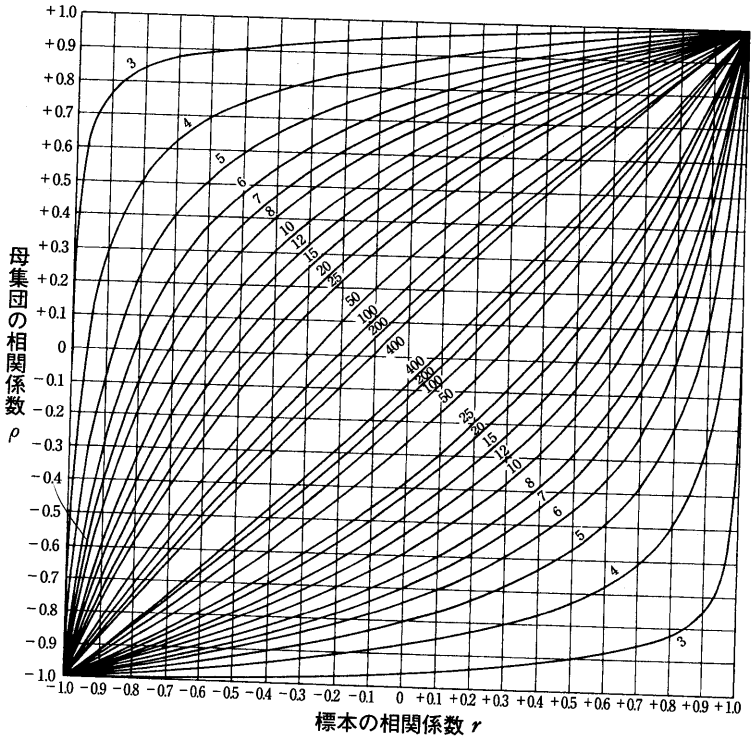


図 7.5　母集団の相関係数を区間推定する（95％信頼区間）

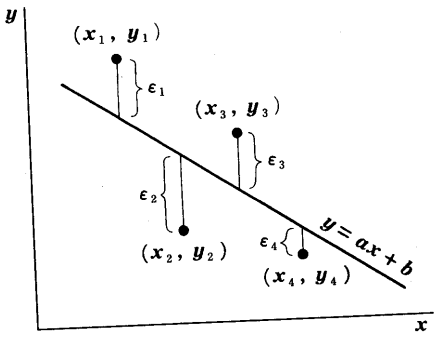


図 7.8　代表する直線を決める方法

＊　ε はイプシロンと読むギリシア文字で，ローマ字の e に相当します．error の頭文字に相当するので，誤差を表わす記号としてよく使われます．

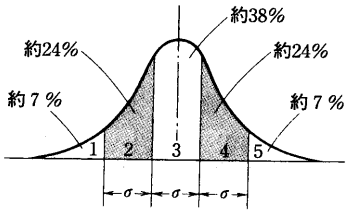


図 9.1　５段階評価のパターン

外的な基準が数値で与えられているときには**数量化Ｉ類**の手法が，基準が分類で示されているなら**数量化 II 類**の手法が使えるのでした．

外　的な基準がまったく与えられていないとき，２つの変量の相関をもっとも強くするように変量の順序を入れ換えることによって変量を数量化するこのような方法を**数量化Ⅲ類**の方法と呼ぶことがあります．

この例のように，外的な基準はもとより，力を貸してくれる他の変量もなく，身内どうしの比較だけを手がかりにして行なう数量化は，数**量化Ⅳ類と**呼ばれることがあり，多くの事象を似たものどうしに分類する場合などに利用される手法です．

数量化の技術は，まだ開発の途上にあります．ここでご紹介してきた数量化Ｉ類からⅣ類までの手法から放したり改善されたりした手法が，これからもどんどん開発され実用化されてくることでしょう．