生命科学の統計学

第１部

Katsumi Wakabayashi, Ph. D.
Prof. Emer. Gunma University
Technical consultant, Shibayagi Co. Ltd.

記述統計学と推測統計学

記述統計学

　情報を読み取りやすいようにデータを加工・整理
　（グラフ化など）する

推測統計学

　標本となるデータから全体の情報を推測する
　（バラツキを伴う情報を客観的に分析･評価する）

母集団（population）と層化抽出標本

何を調べようとしているのか？
　調べようとする対象
　＝あるいくつかの条件（性質）を共有する集合の全体＝母集団
　例えば：小学一年生、選挙権のある成人、2型糖尿病の患者、など

母集団の何を調べるか？
　ひとつの母集団のある特徴を調べる
　性質の異なるいくつかの母集団の特徴を比べる、など

全体（全数）調査と標本調査
　全体調査はしばしば困難である　⇒　標本調査
　標本は母集団の代表的意味を持たねばならない
　母集団の構成員から代表的なものを抽出する　⇒　標本(Sample)
　抽出法：無作為抽出（母集団からランダムに選ぶ）
　層化抽出: 標本の構成は様々な性質に関して母集団の構成と同一
　　　　　であるべき　⇒　層化(Stratification)

層化の例

ヒトの母集団から標本を選ぶ際
　母集団において、ある条件を充たすヒトの割合を考えて、
　標本でもそのように揃える
　例えば、
　　○男女の割合を揃える
　　○年齢の分布を揃える
　　○年収の分布を揃える
　　○学歴の分布を揃える
　　○職業の分布を揃える
　　○既婚、未婚、離婚者の分布を揃える
比較を行う場合には、他の条件を同じようにして１つの条件を変えた時にどうなるかを観察する

母集団の構成

構成要素の分布

分布の一例：　二項定理（パスカルの三角形）

２ｎ＝（ａ＋ｂ）ｎ　仮にａを硬貨の表、ｂを裏と考え、２回トスするとする．可能性としてはa2+2ab+b2　で2回とも表の出る可能性は１/４、表と裏の出る可能性は２/４、2回とも裏のである可能性は１/４である

1　1

1　2　1

1　3　3　1

1　4　6　4　1

1　5　10　10　5　1

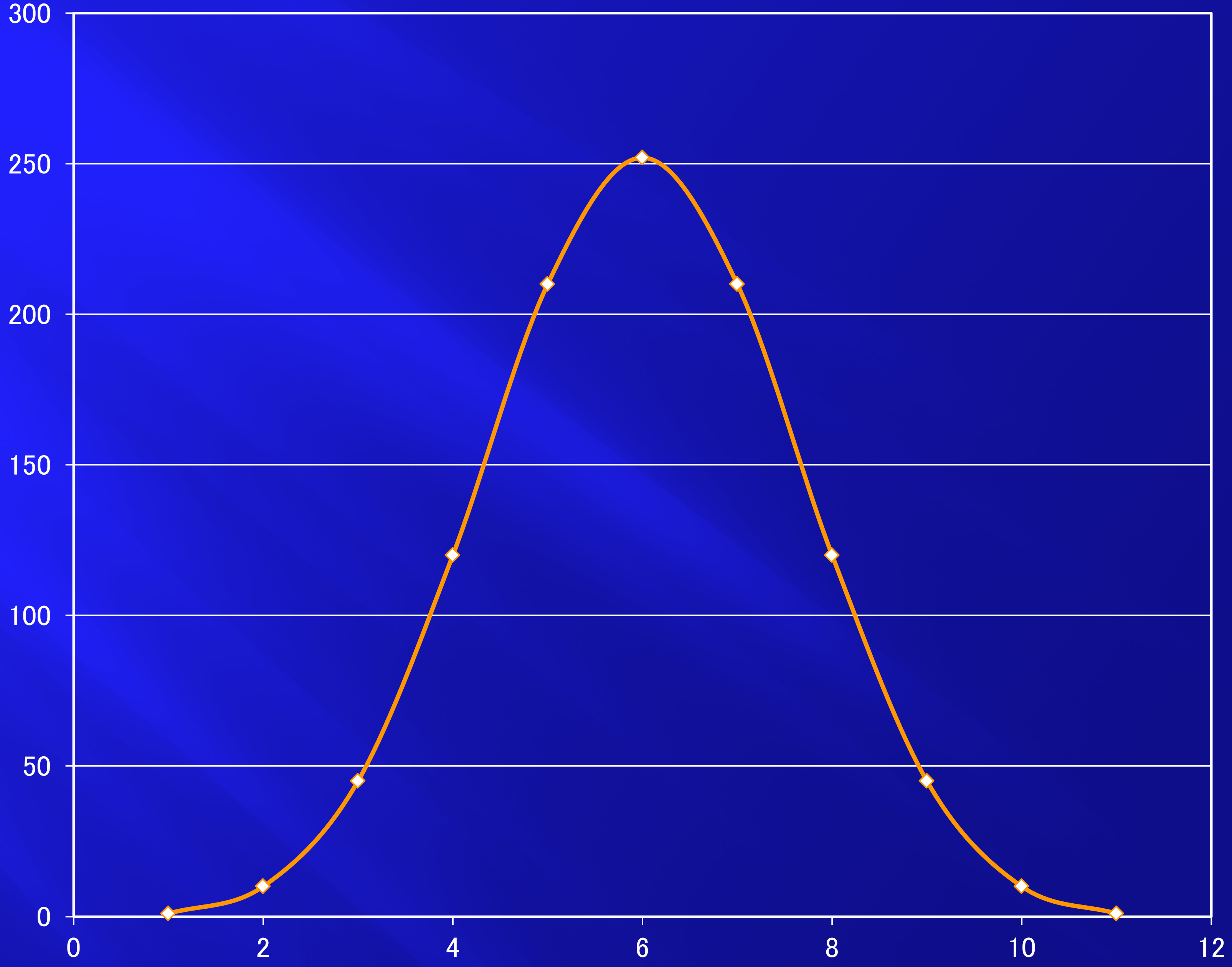
1　6　15　20　15　6　1

1　7　21　35　35　21　7　1

1　8　28　56　70　56　28　8　1

1　9　36　84　126　126　84　36　9　1

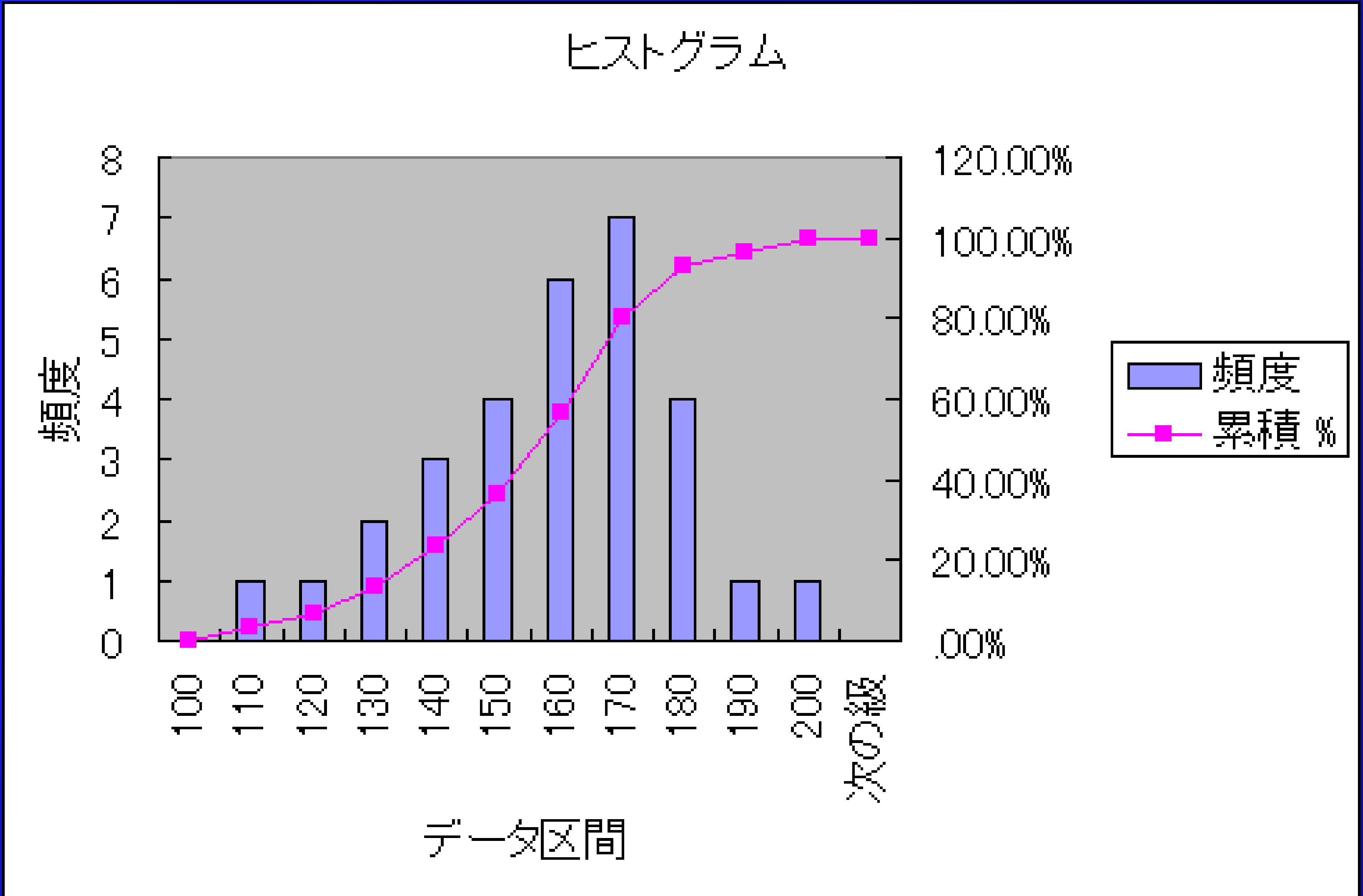
1　10　45　120　210　252　210　120　45　10　1


300
250
200
150
100
50
0
0
2
4
6
8
10
12

ヒストグラム

統計でもっとも基本的な２つのツール、
「基本統計量ツール」と「ヒストグラム」ツールの使用を解説します
表に示すような30個のデータがあるとします

No.	DATA	No.	DATA	No.	DATA
1	102	11	163	21	138
2	123	12	158	22	169
3	145	13	155	23	186
4	120	14	134	24	149
5	156	15	148	25	151
6	145	16	178	26	128
7	178	17	180	27	190
8	180	18	196	28	165
9	176	19	158	29	171
10	167	20	152	30	133



データ区間については「～迄」と理解してください。

基本統計量の表	
平均	154.6667
標準誤差	3.811457
中央値 （メジアン）	157
最頻値 （モード）	145
標準偏差	20.87621
分散	435.8161
尖度	0.3428
歪度	-0.42821
範囲	94
最小	102
最大	196
合計	4640
標本数	30
信頼区間(95.0%)	7.795309

基本統計量解説（１）

平均値（mean, average）

：全データの和をデータ個数で割ったもの

中央値（median）

：データを小さいものから大きいものへと順に並べたとき、その中央に位置する数値．データの個数が偶数である場合は、中央に位置する2個の数値の平均値

最頻値（mode）

：最も頻度の多い数値

基本統計量解説（２）

分散（variance）

：標準偏差と同様母集団の広がりを示す．平均値とデータの個々の数値との差の平方和．Σ(個々の数値ー平均値)２を自由度（データの個数ー1）で割った値

標準偏差（standard deviation, SD）

：母集団の広がりを示す．平均値±SDの範囲は、母集団全体の約67%が含まれる区間．平均値±２SDの範囲には母集団全体の約95%が含まれる．標準偏差は、分散の平方根として求められる

相対的標準偏差（RSD，relative standard deviation）＝変動係数（CV, coefficient of variation）

：平均値に対する標準偏差の割合

標準誤差（standard error, SE, SEM）

：推定平均値の広がりを示す（平均値の標準偏差）
標準偏差をデータ個数の平方根で割った値

基本統計量解説（3）

尖度（kurtosis）
：データの分布が正規分布からどの程度逸脱しているかを示す統計量
　尖度＞０　　　裾広がりの強い分布
　尖度＝０　　　正規分布
　尖度＜０　　　裾が途切れた分布（一様分布、ドーム型分布など）
　　註）尖度の表現には2種類あるので注意．正規分布を３とする方法もある

歪度（skewness）
：同じくデータの分布が正規分布からどの程度逸脱しているかを示す統計量
　歪度＞０　　　右裾広がりの分布
　歪度＝０　　　正規分布
　歪度＜０　　　左裾広がりの分布

範囲（range）：データの最大値ー最小値

最小（minimun）：データの最小値

最大（maximum）：データの最大値

合計（sum）：データの総和

標本数：データの個数

信頼区間（95.0%）(confidence limit)

：平均値の95%信頼限界．即ち5%の危険率で示される平均値の範囲
母集団の全数検査で無い限り、サンプリングにより平均値が変わってくる可能性がある
そのため，母集団の推定平均値の範囲を95%の信頼区間として求めておくことが必要である
この範囲は、t表で自由度と危険率αから tn-1(α)を求め（nはデータの個数）、平均値 ± tn-1(α) x SEの範囲として求められる

危険率αとして0.05、即ち5を採ると、この例ではn=30なので、
t29(0.05)値は2.045である
したがって標準誤差3.8114×2.045=7.7953が信頼区間として示される

計算式

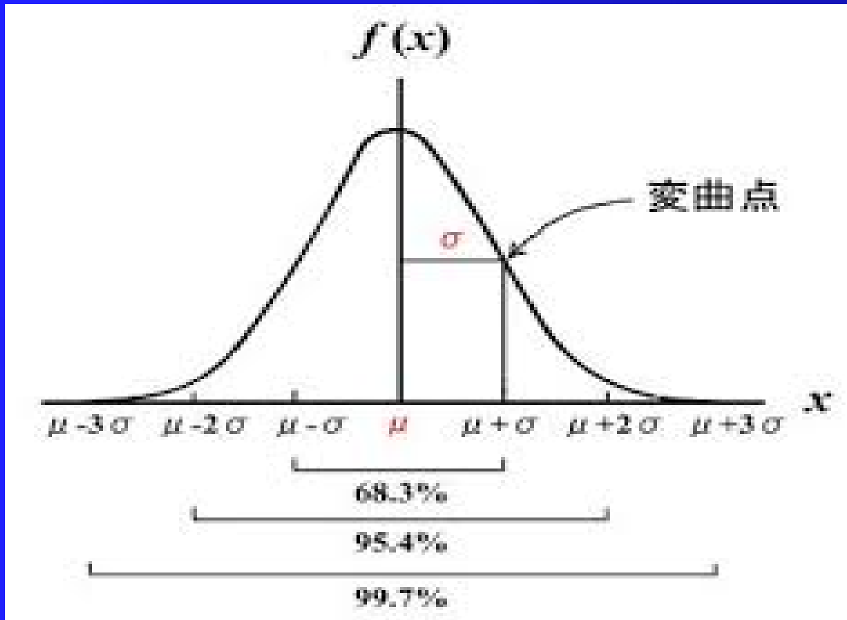
平均値Mean(Average)	M=x=1/n Σ1_(i=1)^n▒xi
偏差	xi-x ‾‾
偏差平方和Sum of square （変動，variation, Sx）	SS=Σ1▒〖(xi−x)^2 〗
（不偏）分散	V=s^2=ss/(n−1)
自由度　Degree of freedom	n-1
標準偏差 Standard deviation	SD=√V
標準誤差 Standard error または (Standard error of Mean, SEM)	SE=√(V/n)

正規分布(y:確率)Normal distribution

N(μ,σ^2)

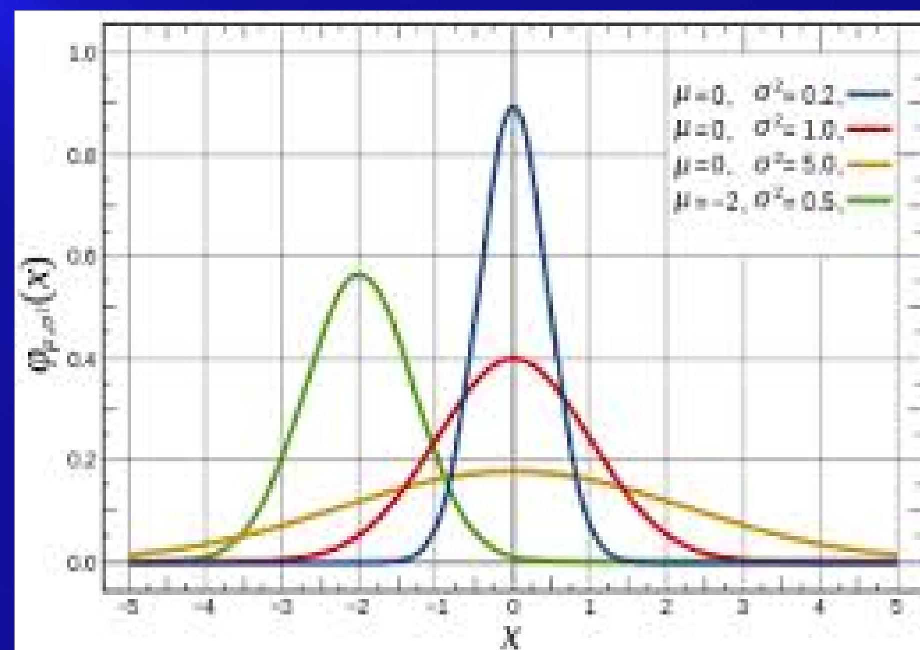
y=1/√ 2$\pi\sigma$ e^(〚-(x-μ)〛 ^2/ 〚2σ〛 ^2)

μ ：平均値，σ２：分散



正規分布をグラフにしてみると
平均値 μ から $\pm\sigma$ のところが
変曲点となる
$\mu\pm\sigma$の範囲が68.3%
　$\pm 2\sigma$の範囲が95.4%
　$\pm 3\sigma$の範囲が99.7%
をカバーする

平均値 μ と標準偏差 σ の
大きさが変わると、曲線は
いろいろに変わることが分かる



母集団に関するいくつかの解析

●平均値と標準偏差の関係（標準偏差が大きいか小さいか）
相対標準偏差(Relative standard deviation, RSD)
あるいは変動係数(Coefficient of variation, SD)
＝SD／mean x100 (%)

●偏差（母集団に属するある数値が平均値からどのくらい離れているか）
＝xi－$\bar{x}$

偏差値（ｓ-スコア）＝(xi－x)／SD x１０＋５０
xｉ＝ x の時、偏差値は５０となる

●ｔ 分布
偏差を標準誤差 (SD／$\sqrt{N}$）で割った値
（xi-$\bar{x}$）／SE の確率分布 ⇒ｔ表
この方法は対応の有る２群の比較にも用いられる
即ちデータｘiとyiとが対応している時、
例えばｎ匹のラットに或る薬物を投与して、投与前と
投与後t時間でのあるマーカーＫの血中濃度をラット
ごとに測定したようなデータ
（投与前値：xi, 投与後値：yi）

yi－xi＝di として、di の平均値を d の SE で割った値を
検討することで di の平均値が０から有意に離れているか
否かを判定する

誤差

- 判定上の誤差
- 第一種の誤差
- 第二種の誤差
- 正確さの誤差
- 確率誤差
- 系統誤差
- グロスエラー

この項　終わり