生物統計学入門

筑波大学体育科学系
高井省三

1.0　はじめに

1.1　生物統計学とは

生物学的問題を解決するための，一つ一つのではない量的情報の科学的研究方法の応用。

ここではなるべく数式を使わずに，Macintoshの統計ソフトーJMP4.0ーを使って生物学的な題材で数理統計学の入門といこう。

コンピュータソフトを使うため従来のように統計解析（すなわち計算）に大きなエネルギーを割く必要はなくなった。この空いた時間を研究に関してどの様に使ったらよいのだろうか？　(1) データの収集，(2) データ整理，(3) 考察，といったところだろう。(1)，(3)はいうまでもないだろう。(2) のデータ整理は探索的データ解析法（Exploratory Data Analysis, EDA）でじっくりやってみよう。これは生物統計学と相補う手法である。

1.2　歴史（数理統計学）

Quetelet, Adolphe (1796-1874)
　ベルギーの天文学者。確率論を取り入れ数学的方法としてひろめる。

Galton Francis（1822-1911)
　相関・回帰の概念，midparent形質の定義。

Pearson, Karl (1857-1936)
　英国人。Biometrika創刊，現在の記述統計学的方法の確立。

Gosset, William Sealy (Student, 1876-1937)
　英国人。小標本理論。t 分布。

Fisher, Ronald Aylmer (1890-1962)
　英国人。F 分布，z 変換。推測統計学の確立。

Neyman, Jerzy (1894-1981)
　ポーランド人。推測統計学の精密化。

2.0　生物学のデータ

2.1　標本と母集団

標本：ある一定の方法で選ばれた個体の集合体

母集団：全地球上に存在する，あるいは，少なくとも空間的・時間的に限定された標本空間に存在する、我々がそれらについて推測しようとするすべて。

2.2　生物学での変数

2.2.1　測量変数

四則演算ができる。実数で表される連続変数（計量データ）と自然数で表される離散変数（計数データ）に分ける。

A　連続変数

(1) 間隔尺度
　ほとんど全ての統計的操作が適用できるが，変動係数は適用できない。
　距離は定義できるが，「零点」には意味がない尺。例えば，相対温度，℃は零度を任意に（水の凍る温度）決めたものだ。したがって，気温20度の日は10度の日より2倍暖かいということにはならない。

(2) 比例尺度
　絶対温度，° Kのように絶対零点がある。

B　離散変数

歯の数，個体数などの度数。（20xx年の日本人夫婦の子供の数の平均は1.75人とはどういうことだろうか？）

2.2.2 順序変数（順序尺度）

等間隔性は保証されていない。四則演算ができないが中央値，パーセンタイル，順位相関を求めることが出来る

2.2.3 属性（名義尺度）

数は全く名目的に使用されるに過ぎない。群やカテゴリの命名（第１学群，第２学群など）

頻度，定性的相関を求めることができる

2.3 誘導変数（計算変数）

示数，比，率の弱点

(1) 相対的な不正確さ

1.2/1.8の比0.6666...は，２つの数字の真の値の範囲（1.2：1.15～1.25；1.8：1.75～1.85）から計算されるので1.15/1.85～1.25/1.75の範囲，つまり0.622～0.714の値をとりうる。

(2) 正規分布をしないことが多い

逆正弦（arcsin）変換を考えなければならない時もある

(3) もとの２つの変数の関係の情報が失われる

2.4 データの変換（再表現）

2.4.1 効用

(1) 変数間の関係を単純化する

(2) 分布の広がりを一定にして集団を比較し易くする

(3) 正規分布や他の標準的分布への近似

2.4.2 指針

(1) ベキ乗変換

（最大値／最小値）≦20のとき

(2) 対数変換

データが全て正であるとき，計数データであるとき

(3) 平方根変換

度数分布のならし

(4) 折り重ね対数変換

比率データに対して，$\ln\sqrt{p} - \ln\sqrt{(1-p)}$で両極を引き伸ばす

(5) (Y - a) / (b - a)変換

データYに上限 (a)と下限 (b)があるとき

(6) 順位→比率

データ数nの集団のi番目データの累積比率を，p = i / (n + 1)，p = (i - 1/2) / (n + 1/2)，p = (i - 1/3) / (n + 1/3)などと変換する

3.0 基本統計量

3.1 頻度（度数）分布

データがどの様に分布しているかを概観することはデータ解析の第一歩だ。分布のかたちから(1)歪み，(2) はずれ値，(3) ギャップ，(4) 多峰性，を吟味して初めて集団の比較に平均値を利用すべきか，中央値を使うべきかが決る。

3.1.1 生物学的意味

(1) 非対称分布

(2) 分布の一方の側に分布する生物への淘汰（指向性選択）

(3) 測定に使った尺度のため
(4) 二峰性分布（bimodal distribution）
(5) 異なった種，系統が混在
(6) 両性の混在
(7) 異なった年齢集団の混在
(8) 分布の中間の生物が不利な選択（分断選択）

3.1.2 階級の数

生物学では階級数（L）は経験的に15〜20。ただし，例数（n）によって異なる：40例以下では12未満だろうし，数千例では20以上のこともある。

$L = 1 + \log_2 n$（Sturgesの公式）または，$L = 10 \log_{10} n$（探索的データ解析における公式）がある。後述するJMP（Macintoshの対話的統計プログラム）で試行錯誤してみよ。

3.1.3 Macintosh／Windowsのソフトー ＪＭＰ ーで度数分布グラフをつくる

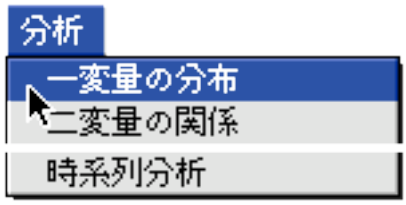
《問題》例題フォルダのテキストファイルsomatometry.textをインポートしてJMPファイルをつくれ。

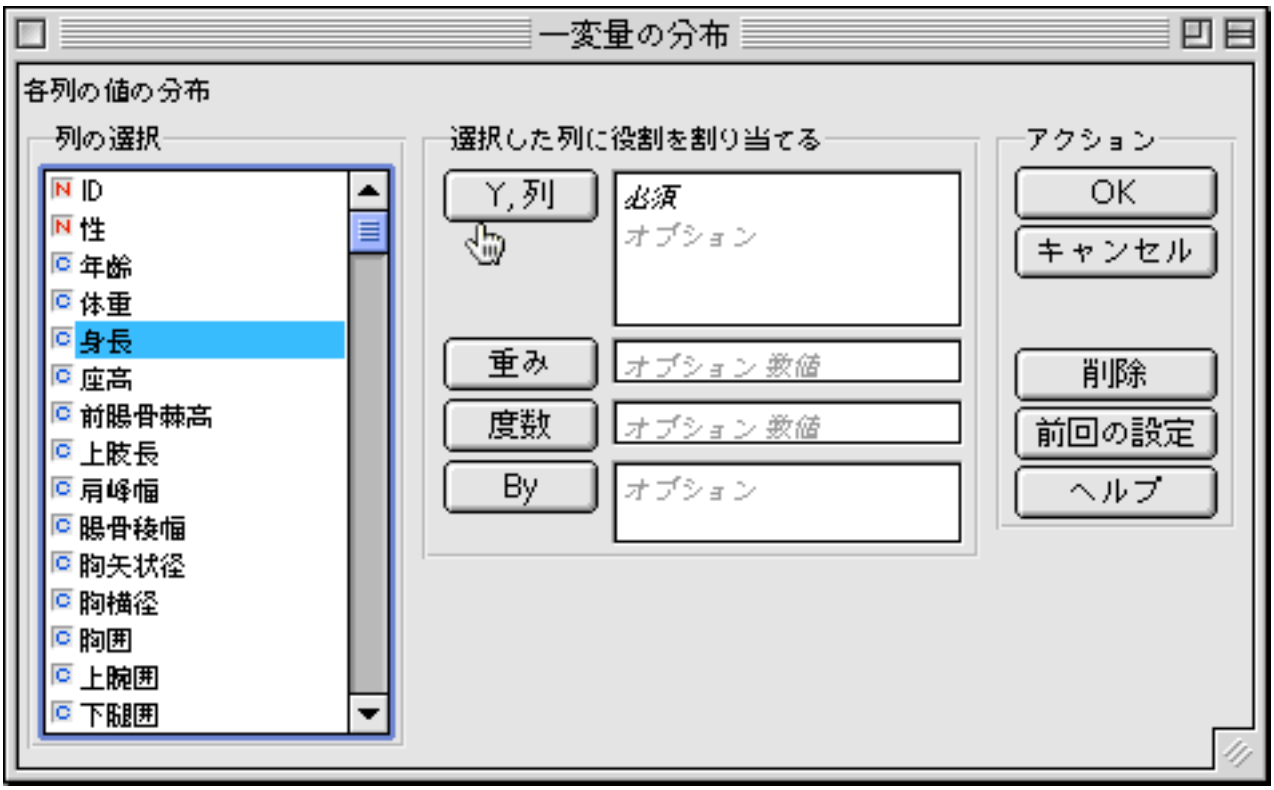
変数は，(1) ID番号，(2) 性，(3) 年齢，(4) 体重，(5) 身長，(6) 座高，(7) 前腸骨棘高（下肢長），(8) 上肢長，(9) 肩峰幅（肩幅），(10) 腸骨稜幅（腰幅），(11) 胸郭矢状径（胸厚），(12) 胸郭横径，(13) 胸囲，(14) 上腕囲，(15) 下腿囲，(16) 三頭筋皮脂厚，(17) 肩甲下皮脂厚，(18) 前腕末端幅，(19) 大腿上顆幅，(20) 頭長，(21) 頭幅，(22) 下顎角幅の順にならんでいる。

《問題》皮脂厚（Y）に，$100 \times \log_{10}(10 \times Y mm - 18)$という変換式（Edwards et al.，1955）がある。ナマの値と変換値の分布を比べてみよ。

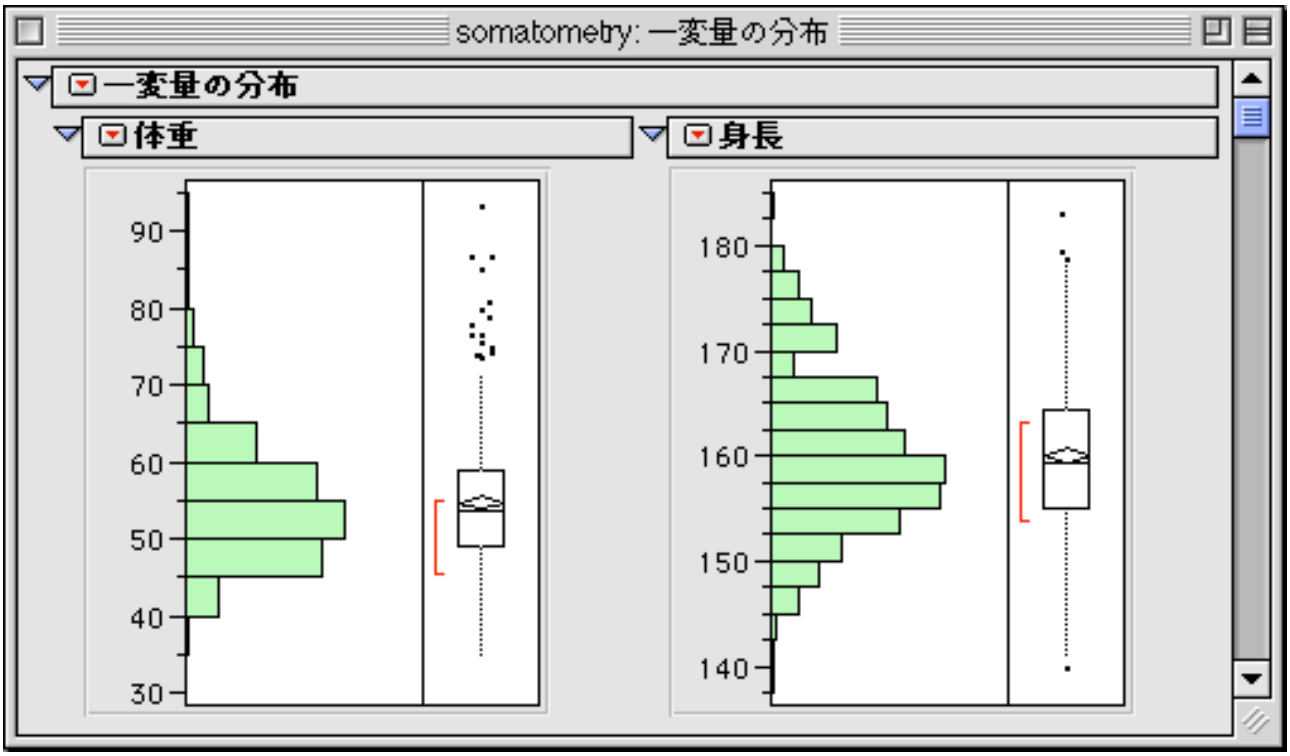
(1) 分布のグラフ化

分析メニューから**一変量の分布**を実行する




一変量の分布
各列の値の分布
列の選択
ID
性
年齢
体重
身長
座高
前腸骨棘高
上肢長
肩峰幅
腸骨稜幅
胸矢状径
胸横径
胸囲
上腕囲
下腿囲
選択した列に役割を割り当てる
Y, 列
必須
オプション
重み
オプション 数値
度数
オプション 数値
By
オプション
アクション
OK
キャンセル
削除
前回の設定
ヘルプ

(2) ヒストグラム

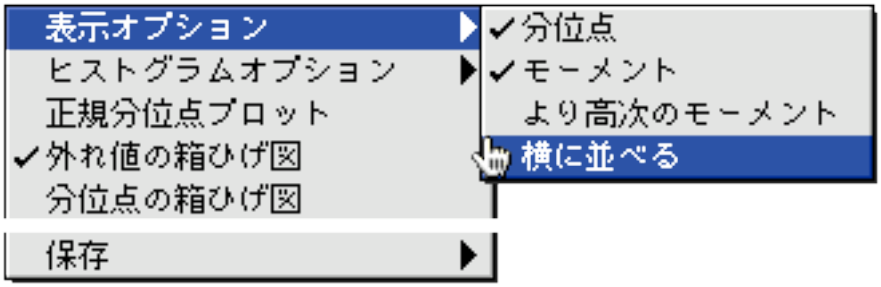


(3) 階級数（階級幅）の変更

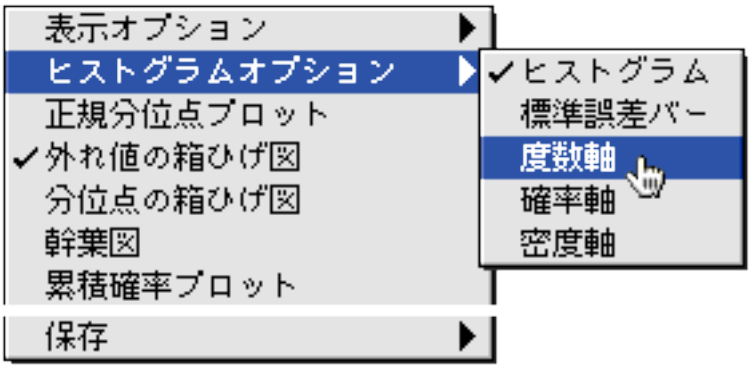
ツールメニューから手のアイコンを選ぶ⇒ヒストグラムにのせて，左に動かすとドラッグすると階級幅は大きくなる（ヒストグラムのバーの本数は減る），右にドラッグすると階級幅は小さくなる（バーの数は増える）。

(4) ヒストグラムの表示オプション

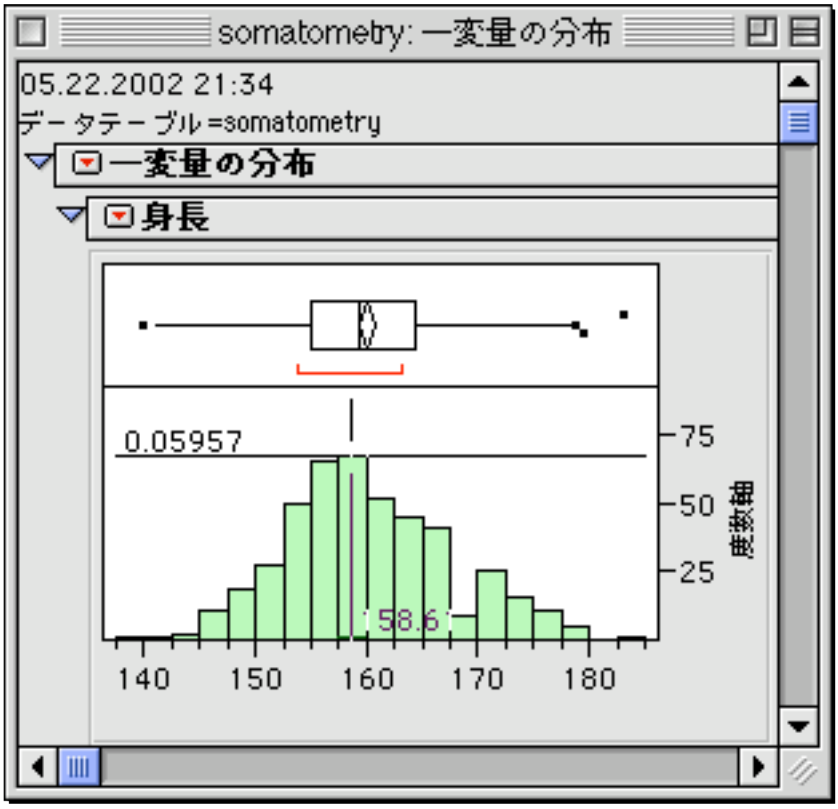
身長タイトルバーの左コーナーの「チェックマーク▼」をクリックし，**表示オプション⇒横に並べる**を選んでみよう。見なれた水平配置のヒストグラムが表示される。



同様にしてチェックマーク▼をクリックしヒストグラムの高さ目盛り（Count Axis）等を表示させることができる。



ツールメニューからクロスヘアー（十字）を選び，度数軸目盛りにあてて，ヒストグラムの高さを読んでみよ。



《問題》somatometryファイルで適当な変数をいくつかとりあげ，男女のヒストグラムを比べよ。性的二型（sexual dimorphism）が見られる変数はあるか？

《問題》頭示数を求め，長頭，中頭，短頭に分け，ヒストグラムをつくれ。

3.2 代表値

3.2.1 平均値（算術平均）

$$\bar{Y}=\left(\sum_{i=1}^{N} Y_i\right)/N$$

ヒストグラムの重心に相当する。飛び離れたとっぴょうしもない値（はずれ値）に影響される！

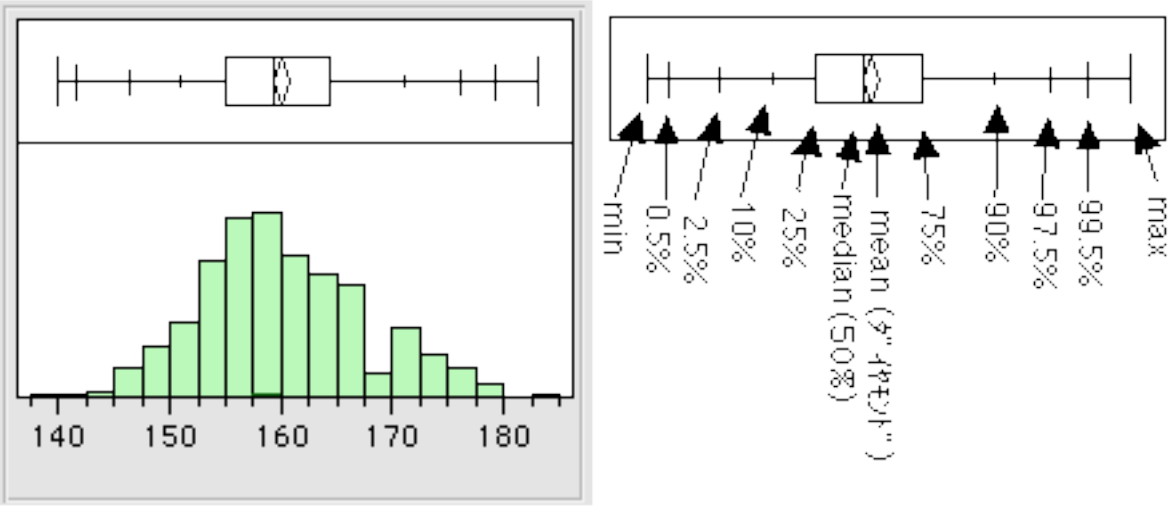
3.2.2 メディアン

観測値をその値の大きさの順に並べたとき，その両側に同数の観測値がくるような値：LD_{50}（中央致死量），ED_{50}（中央有効投与量）

分布が正規でないときのノンパラメトリック統計法で使う。

nが偶数の時は，n/2と(n/2)+1番目の中間点；タイ(同値）があるときの計算は面倒だがJMPにまかせよう。

JMPではヒストグラムの脇の箱ひげ図（box plot）に注目しよう（左図）。ここには平均値（ダイヤモンド），四分位範囲（箱），最大・最小値，各パーセンタイル値が図示されている。



3.3 バラツキ度

3.3.1 標準偏差

(1) 偏差

$y=Y-\bar{Y}$ で表す。

《問題》集団のバラツキの代表値としてこの偏差yの平均を使うとするとどうなるか？JMPで新変数yをつくって試してみよ。

(2) 偏差平方和

平方和（SS, sum of squares）

$$\sum y^2=\sum (Y-\bar{Y})^2$$

(3) 分散（平均平方）

分散（v, variance），平均平方（mean square）

$v = \Sigma y^2 / (n-1)$

(4) 標準偏差

s，SD (standard deviation)

$s = \pm \sqrt{v}$

3.3.2 変動係数

母集団の平均値が大きく異なっているときは分散や標準偏差そのものの比較はあまり有用ではない。平均値が大きい生物では，そのバラツキは平均値の小さい生物よりも，大きいのが普通である。

$CV=(s \times 100)/\bar{Y}$

で母集団のバラツキの相対的程度をあらわす。これは，測定値の単位とは無関係であり，パーセントで表される。体重（kg）と皮脂厚（mm）の比較などに有効。

《問題》身長，体重の標準偏差と変動係数を比較せよ。

3.4　基本統計量をＪＭＰで求める

《問題》基本統計量と度数分布図をじっくり比べてみよ。数字（基本統計量）からヒストグラムが浮かんでくるか？

《問題》男女別にいくつかの変数について，基本統計量を求めてみよ。

《問題》１つの変数の全ての観測値に(1) 定数3.3を加えたとき，(2) ２を掛けたとき，(3) 3.3を加えたのちに２を掛けたとき，(4)２を掛けたのちに3.3を加えたとき，基本統計量はどうなるか？

ヒストグラムの下（横）には分位点，モーメントが表示されている。

分位点			モーメント	
100.0%	最大値	183.10	平均	159.99912
99.5%		179.36	標準偏差	7.4721518
97.5%		176.06	平均の標準誤差	0.3510723
90.0%		171.00	平均の上側95%信頼限界	160.68905
75.0%	4分位点	164.45	平均の下側 95%信頼限界	159.30918
50.0%	中央値（メディアン）	159.30	N	453
25.0%	4分位点	155.00		
10.0%		151.00		
2.5%		146.30		
0.5%		141.66		
0.0%	最小値	139.80		

3.5　基本統計量の有効数字

標準偏差を３で割り，商の最初の０でない有効桁に注目。平均値はこの桁まで，標準偏差はもう１桁余分に。

《問題》2.354±0.363，2.354±0.243と計算されたときに，記載すべき有効数字を求めよ。

3.6　母数

標本統計量から母集団統計量（母数）を推定する

不偏推定値：標本の大きさにかかわらず，ある母集団から抽出された標本統計量が平均的に見て，母数に等しくなるような推定値

標本平均は不偏推定値であるが，標本分散は不偏ではない。一般に標本分散は母分散を小さめに推定するので，平方和をn（標本数）で割らずに自由度（この場合はn-1）で割る。

4.0　確率分布入門

4.1　確率分布の初歩

標本から母集団を帰納推理するのが近代の統計学（推測統計学）であることは先に述べたが，このような帰納推理を形式化するにあたり最も重要なものが確率論であろう。Quetletに始まる近代統計学はまさにこの確率論の上に築かれているのである。確率の概念によって，統計的結論の不確実性の程度（すなわち信頼性の程度）を表現するのが推計学である。

4.1.1　確率，無作為標本，仮設検定

1992.1.1現在の筑波大学の学生を，男の学群生（MU，6,084人），女の学群生（FU，2,791人），男の

院生（MG，2,168人），女の院生（FG，636人）に分類しよう。ある実験で学生を抽出する状況を想定しよう。

まず，**確率**の定義をしよう。ある形質（男，女，学群生，院生）を持った標本を抽出する確率は，標本の大きさをどんどん大きくしていったときの全標本数に対する当該標本数の比率の極限として定義される。形質（属性）FGをもつ確率を，P[FG]＝0.02，と書こう（PはprobabilityのP）。

標本を抽出するには偏らないように，無作為にしなければならない。**無作為標本**（random sample）では，母集団の中にある全ての個体が同じ確率で標本の中にふくまれる。小さな母集団では次々に抽出実験をするときは，抽出の度に標本を元に戻さないと，確率の定義に反するようになる。5本のあみだくじで１人の鬼を選ぶケースを考えてみよ。初めは1/5の確率であるが，１番目の抽出で鬼がはずれると1/4となり，２番目の抽出でもはずれると1/3となり確率は変ってしまう。生物集団は有限ではあるが，非常に大きいことが多いので，標本抽出に際しては，抽出した標本を元に戻さなくても（置換しなくても），無限の母集団と考えてよい。

《**問題**》どの様な抽出計画を立てたら無作為に（at random）に学生を抽出できるだろうか？

ある実験でのすべての確率を含んだ集合を標本空間とよぶ。集合｛MU, FU, MG, FG｝が標本空間だ。１回の実験でただ１つの結果しか起こらないような要素を単事象という。｛MU｝，｛FU｝，｛MG｝，｛FG｝はそれぞれ単事象だ。間違いやすいが，**事象**とは標本空間のいかなる部分集合をも表す。だから，｛FG｝，｛MU, FU｝，｛MG, FG｝，｛MU, MG｝，｛FU, FG｝，...はそれぞれ事象である。｛MU, FU, FG｝は学群生かあるいは女子学生という場合の事象だ。

事象AをA={MU, FU}，事象BをB={FU, FG}とすると，事象AとBの積事象はA∩B={FU}である。事象AとBの和事象はA∪B={MU, FU, FG}である。

ここで，単事象のそれぞれにある数字p $(0 \leqq p \leqq 1)$ をあたえ，$\Sigma p=1$とすると，その空間は確率空間となる。つまり，

｛MU , FU ，MG , FG ｝の標本空間は，

{0.52, 0.24, 0.19, 0.05}の確率空間に対応する。

4.1.2 事象の独立

ある一つの事象のおこる確率が，その事象がすでに起こったかどうかによっては影響されないとき，各々の事象は**独立**であるという。記号で書くと，P[A∩B]=P[A]・P[B]ならば事象AとBは独立。**事象A，Bが独立であるならば，その確率はそれぞれの確率の積として表すことが出来る。**

［例１］

女子学生がむれがちであるならば，1番目の抽出学生がならば，2番目の抽出学生もFUである確率は0.24より高いかもしれない。このようなときは事象は従属している（独立でない）といい，その確率を条件付き確率という。このような事象の背景にはなんらかの因果関係を想定できよう。

もし，2人の学生を抽出した結果が，両方ともFG（女子の院生）であったなら，その確率はP[FG]・P[FG]=0.05×0.05=0.0025となる。これは，0.25%という偶然だけでは得ることの出来ない希なことである。P[FG]=0.05は事実だから，標本抽出が無作為でなかったか，事象が独立でなかったか，あるいは両方であると疑わざるをえない。

［例２］

ストレスを体験すると病気に対する抵抗力が増すかどうかを23頭のサルを使って調べた実験がある。ストレス群はある作業を怠けると電気ショックが与えられるという状況におかれた。全てのサルにポリオウィルスを接種した。ストレス群は12頭中7頭が生き残った。対照群は12頭中ただ1頭しか生き残らなかった。このような違いが偶然に起こる確率はどれほどだろうか？

	生存数	死亡数	合計
ストレス群	7	4	11
対照群	1	11	12
合計	8	15	23

全てのサルの中では23頭のなかで8頭が生き残ったわけだが，この結果起こる場合は23C8 =490314のケースがある。ストレス群では11頭のうち7頭が，対照群では12頭のうち1頭が生き残った。 こうなるには11C7=330，12C1=12のケースがある。ストレス群，対照群で生き延びたケースは，330×12=3960である。これが偶然に起こる確率は，3960/490314=0.00808で約0.81%だ。さらに，偶然が極端な結果を引き起こすケースを考えよう。つまり，対照群は全て死に，生き残り8頭は全てストレス群の場合である。

12C0=1，11C8=165であるから，この極端なケースは1×165=165通りありその確率は165/490314=0.000337で約0.03%だ。先のケースと足して0.81＋0.3=0.84%がこのような実験結果を生じさせる確率である。つまり，100回に1回の割りよりも小さい。このことは，ストレスと抵抗力という２つの事象は独立であるという仮説は支持できそうもないことを意味している。すなわち，ストレスと病気に対する抵抗力には因果関係があるという結論を99%の信頼性で下せるのである。

以上は**Fisherの独立性の精密検定**と呼ばれるものである。標本数nが20より小さいか，または20≦n≦40で期待数の最小が5より小さい場合はこの方法が適用できる。

	事象Aが起こる	事象Aが起こらない
事象Bが起こる	a	b
事象Bが起こらない	c	d

とすると，

p = {a+bCa c+dCc} / a+b+c+dCa+c

= {(a+b)! (c+d)! (a+c)! (b+d)!} / {a! b! c! d! (a+b+c+d)!}

である。

4.2 二項分布

離散的な確率分布の例として二項分布をとりあげよう。これは遺伝学のような分野で重要である。**二項分布**は起こり得る結果（事象）が２通りあるような実験をn回試行したときの，いずれか一方の生起回数の確率を求めたりするときに利用できる。たとえば，母不良率が分かっている製品のロットからn個を取り出したときの不良品の個数kの分布にもあてはまる。

以下では話しを簡単にするために，学群生，院生の分類は無視して，標本空間を二つの要素，男と女{M, F}，で成り立つものとする。そして，男子学生である確率をp=P[M]，女子学生である確率をq=P[F]（q=1-p）として，確率空間を{p, q}であらわすことにする。

２人の学生を抽出したときの標本空間は，

{MM, MF, FF}で，その確率空間は，

$\{p^2 , 2pq , q^2\}$である。

事象MFの確率が２倍になっているのは，Mの後でFを抽出したときと，Fの後でMを抽出したときの２通りの抽出があるからだ。

３人の学生を独立に抽出するとその標本空間は，

{MMM, MMF, MFF, FFF}となり確率空間は，

$\{p^3 , 3p^2q , 3pq^2, q^3 \}$となる。

このような結果は二項展開の方法でまとめることが出来る。二項展開の方法は，事象が２つあり，それ

ぞれが独立で起こる場合に応用できる。例えば，MN型の血液型のように 1 対の対立遺伝子が表現型を決定している場合などがそうである。これは二項式$(p + q)^n$を展開して計算できる。ここで，nは抽出する標本の大きさを表す。

二項展開の係数（二項係数）を得るには，**パスカルの三角形**を使えばよい。

n
0　　1
1　　1　1
2　　1　2　1
3　　1　3　3　1
4　　1　4　6　4　1
5　　1　5　10　10　5　1

また，二項展開の一般項は

nCk $p^k q^{n-k}$

で表され，このうち二項係数はnCkでそれは，

nCk = n! / [k! (n-k)!]

で計算できる（！は階乗を示す）。学生の性比のケースで3人を独立に抽出したとき，{MMF}の事象の確率は$3p^2q$で表され，k=2は男子学生、q=1は女子学生の数を表している。

確率分布をつぎのようなケースを材料に考えてみよう。

1991年の人口動態統計によると日本人の出生時の性比は，男児：女児＝M:F=627：595であった。いまこの母集団から 6 人の子供を抽出したとき，子供の性の構成の期待頻度の分布をみてみよう。

男児誕生の確率は，p=627／1222=0.513，女児のそれはq=1-.513=0.487となる。期待される比率は$(p+q)^n$ $=(0.513+0.487)^6$の二項式の展開となる。確率空間は，パスカルの三角形または二項展開の一般項から，

$\{p^6, 6p^5q, 15p^4q^2, 20p^3q^3, 15p^2q^4, 6pq^5, q^6\}$

となる。標本空間は，

{MMMMMM, MMMMMF, MMMMFF, MMMFFF, MMFFFF, MFFFFF, FFFFFF}

p，qに実際の数値を入れて計算すると，それぞれの項は理論的な頻度分布すなわち確率分布を示す。この二項式展開型の確率分布が二項分布である。

実際に度数分布（絶対期待頻度の分布）を求めるには，この相対期待頻度に標本（母集団）の大きさを掛けてやればよい。

《**問題**》上の男女の性比のケースで標本を10,000回抽出したときの，相対期待頻度，絶対期待頻度を求めて，ヒストグラムを描け。Microsoft Excelで計算し，JMPに読み込むとよいだろう。

ここで，もしあらかじめそれぞれの事象の観測頻度があるならば，絶対期待頻度と観測頻度の一致／不一致を調べることが出来る。不一致の原因は偶然によるものであるかも知れない。あるいは，(1) 出生性比はM:F=627：595である，(2)標本抽出は無作為である，(3)事象が独立である，という仮定の 1 つ以上を棄却したほうがよいかもしれない。これは統計学における仮設検定の考え方である。

《**問題**》上の男女の性比のケースで，**二項分布の平均値と標準偏差**をそれぞれ絶対および相対期待頻度のそれについて求めよ。

絶対期待頻度では，μ=np，　$\sigma=\sqrt{npq}$

相対期待頻度では，μ=p，　$\sigma=\sqrt{pq/n}$

で与えられることがわかっている。

4.3　正規分布

4.3.1 二項分布との関係

n=20の二項分布を展開すると，あるいはnpq≧3のような二項分布を展開すると**正規分布**で十分近似できることが知られている。実際，正規分布の式を発表したde Moivre (1733) は二項式に対する近似を求めるさなかに発見したそうだ。遺伝学の材料で，離散分布の二項分布から連続分布の正規分布への近似を考えてみよう。

1909年にスウェーデンのNilsson-Ehleはコムギの種皮の色（赤色の濃淡）がメンデル遺伝する2つの対立遺伝子で説明できることを明かにした。赤色と白色の両親の遺伝子型R1R1R2R2，r1r1r2r2からのF1はR1r1R2r2の遺伝子型をもつ。このF1同士を交配したF2の表現型を見てみよう。Rを4つもつ暗赤色，Rを3つもつ濃赤色，Rを2つもつ赤色，Rを1つもつ淡赤色，Rを1つももたない白色の比率は1:4:6:4:1になる。

《問題》クロス表をつくって，このことを確かめてみよ。

これはn=4の二項展開の係数である。このようにある形質について同じ効果をもつ遺伝子が複数の遺伝子座に見られるとき，これらを同義遺伝子（multiple gene）という。

ここで遺伝子座が10個でそれぞれに１対の対立遺伝子があるケースを考えると，種皮の色は表現型の間で識別困難になり，遺伝子座がもっと増えると最終的に個体間の変異は連続的になり二項分布から正規分布に近付くことが期待できる。このような，ある形質を表現させる遺伝子の数が多く，１つ１つの影響は小さいが相加的に働くような同義遺伝子をポリジーン（polygene，量的遺伝子）という。身長の遺伝もこのようなものとされている。

4.3.2 正規分布の性質

正規分布の確率密度関数（**正規確率密度関数**）は次のようである，

$$Z = \{1 / (\sigma\sqrt{2\pi})\} \exp[-1/2\{(Y-\mu)/\sigma\}^2 \quad \text{あるいは} \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

つまり，正規分布は母平均μと母標準偏差σの２つの母数で決る。

正規曲線（正規確率密度関数）は平均値を中心として左右対称であるから，平均，中央値，最頻値は一致する。

正規分布ではある範囲に観測個体がふくまれる割合は：

$\mu \pm \sigma$　　は全観測数の68.27%をふくむ

$\mu \pm 2\sigma$　　は全観測数の95.45%をふくむ

$\mu \pm 3\sigma$　　は全観測数の99.73%をふくむ

これらは「正規分布の上側確率」の表（日本規格協会「簡約統計数値表」）で求められる（ただし1-2Q(u)の値として読む）。

または，

全観測数の50%は$\mu \pm 0.674\sigma$の間にはいる

全観測数の95%は$\mu \pm 1.960\sigma$の間にはいる

全観測数の99%は$\mu \pm 2.576\sigma$の間にはいる

これらは「正規分布のパーセント点」の表（日本規格協会「簡約統計数値表」）で求められる（ただし，(1-Q)/2として表を引く）。

これらの数表の数値は縦軸，横軸を相対的な数値で表現している。正規分布の上側確率の表のuは平均からの距離をσの単位で表した基準正規偏差値，$u = (Y-\mu) / \sigma$，を使っている。

《問題》受験で問題の偏差値とはどのようなものか？

4.3.3 正規分布の応用

1990（平成2）年乳幼児身体発育調査結果によると，標本数6,014人の日本人男児の出生時体重の平均値は3.19kg，その標準偏差は0.43kgであった。出生時体重が正規分布していると仮定して，この母集団から無作為に標本抽出したとき，出生時体重が1.8kg以下の子供を抽出する確率はどれくらいだろう？

まず正規分布の上側確率の表に合わせるために，基準正規偏差値を計算する。(1.8-3.19)/0.43=-3.233

これは1.8kgの出生時体重は平均値よりも3.233標準偏差単位小さいことを意味する。正規分布の上側確

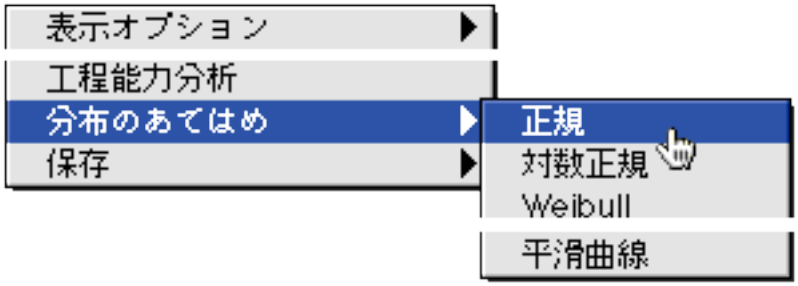
率の表でu=3.23のQ(u)は0.0006であることがわかる。つまり，0.06%の確率だ。もし，母数の分からない母集団から無作為に抽出した１人の新生児がこのような体重を持っていたのなら，考えている母集団から本当ににきたものか疑問を持ってもよいだろう。つまり，日本人の男児であるのかどうかを。

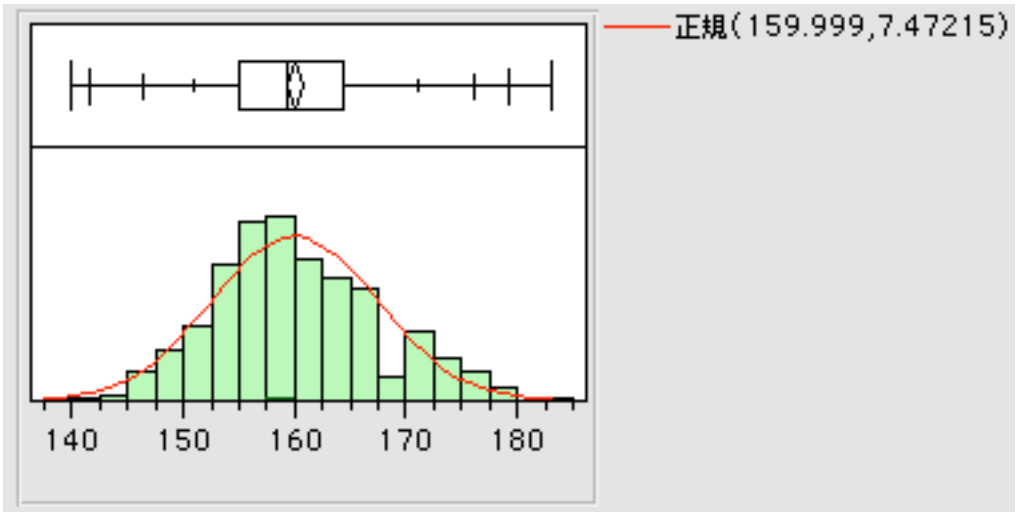
4.3.4　正規分布のチェックと検定

もし，ある分布が正規分布に従うと仮定できるならば，これに基づいて**予測，仮説の検定**ができる。これから学ぶ，分散分析，回帰分析などはいづれも正規分布理論に基づいている。

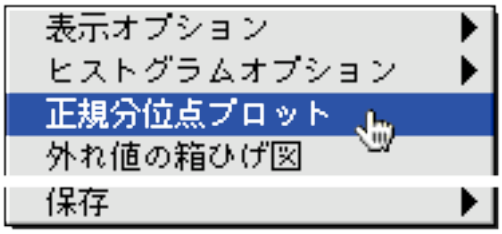
上での日本人男児の出生時体重の仮説を検定する前に，まず出生時体重が正規分布するかどうかを確認しなければならない。Shapiro-Wilksの検定（n≦50）またはKolmogorovの検定（n≧51）を援用することになる。

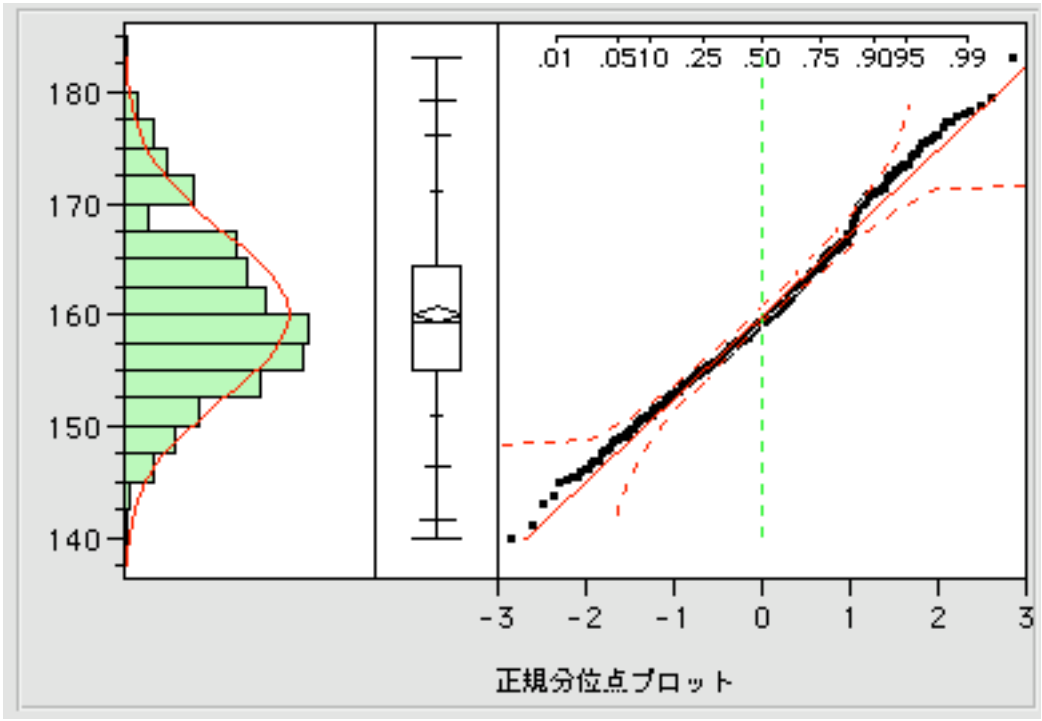
分析⇒一変量の分布メニューで分布図が描かれたら，タイトルバー左の▼マークをクリックし**分布のあてはめ⇒正規**をクリックする。すると正規曲線を分布図に重ねて描く。





正規分位点プロットをチェックすると正規累積百分率をプロットする。後者は変数が正規分布していれば対角直線になる（正規確率紙をつかったものと同じ）。下の図は女子の身長の分布。

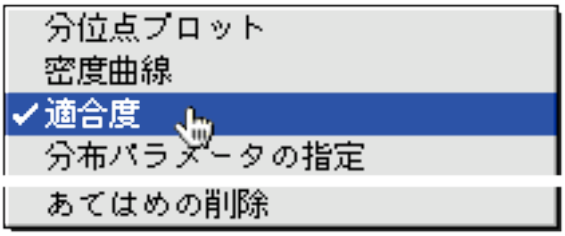


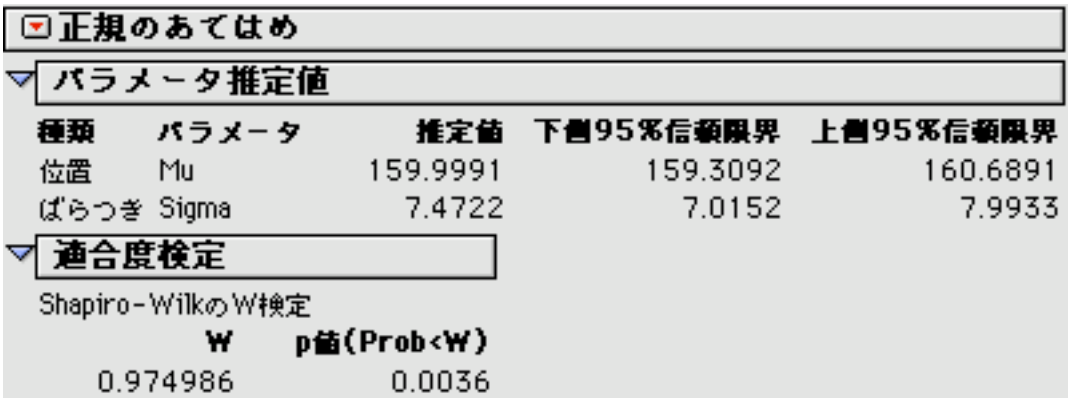


正規分布の有意性検定について，JMPでは標本数が2,000以下のときはShapiro-WilkのW検定が，2,000以上の標本に対してはKSL（Kormogorov-Smirnov-Lelliefors）検定が使われる。

分析⇒一変量の分布メニューを選択する。項目タイトルバー左端にある▼をクリックして**分布のあてはめ⇒正規**をクリックするとヒストグラムの下の方に正規のあてはめというタイトルでリポートが表示される。

正規のあてはめタイトルバーの▼をクリックして適**合度**を選ぶ。





上の結果は女子の腸骨稜幅の分布の例だ。標本数は2,000以下なので，Shapiro-WilkのW検定の値が計算されている。これはこのような実測値が正規分布をするデータから取り出されたものであるという仮説を検定するものだ。0.7777という数字はこの仮説が起こる確率が80%くらいあるという意味だ。８割というのは

非常に良くあることだから，この腸骨稜幅データは正規分布する母集団からのサンプルである，すなわちこの腸骨稜幅データは正規分布することだ。どれくらいの割合なら仮説を棄てる（正規分布を否定する）かというと0.05より小さい数字をとるのが一般的だ。

《問題》男子の腸骨稜幅についても正規性を検定せよ。体重，皮脂厚でも試してみよ。また，変数変換（対数など）をしたらどうなるかを検証せよ。

5.0 推定と仮説検定

生物学で出会う二つの基本的な統計学的な問題，(1) 結果はどれほど信用できるものなのか？(2) 観測結果と仮説から期待される差が偶然に起こる確率はどれくらいか？を考えよう。(1)は信頼限界について，(2)は仮説検定の問題である。

5.1 平均値の分布と分散

まず，正規分布に関する重要な定理，中心極限定理を頭に入れよう。

（定理１）正規母集団からの標本平均値（$\overline{Y}$）は，標本数にかかわらず，それ自身正規分布をする。

（定理２）標本数が増えるにつれて，どんな分布を示す母集団からとられた標本平均も，正規分布に近づく。

さらに，　　標本平均の分散の期待値をみてみよう。期待値とは，無限回数同じような標本抽出を繰り返して得た平均の値を意味する。

標本平均の分散の期待値は

$$\sigma^2_{\overline{Y}} = \sigma^2 / n$$

だから，平均値$\overline{Y}$の標準偏差の期待値は

$$\sigma_{\overline{Y}} = \sigma / \sqrt{n}$$

平均値$\overline{Y}$の標準偏差の推定値（標準誤差）

$$s_{\overline{Y}} = s / n$$

5.2 信頼限界入門

標本平均mは母平均μの不偏推定量であることは述べたが，$\overline{Y}$はμとしてどれほど信頼できる値であろうか？たいてい，母統計量の真値は分からないが，標本統計量に信頼限界をもうけてそれを推定する。まず，母標準偏差が分かっている正規母集団について考えてみよう。ついで，母標準偏差が分からず，σの推定値のｓとスチューデントのｔ分布をつかった信頼区間を考えてみる。

定理１より，正規分布の性質－全観測数の95%は$\mu \pm 1.960\,\sigma$の間に入る－が応用できる。平均値$\overline{Y}$の標準偏差の期待値は $\sigma_{\overline{Y}} = \sigma / \sqrt{n}$ だから，

大きさｎの標本からの標本平均の95%は，$\mu \pm 1.960\,\sigma / \sqrt{n}$の間に入る。

式で書けば，

$$\mu - 1.960\,\sigma / \sqrt{n} \leqq \overline{Y} \leqq \mu + 1.960\,\sigma / \sqrt{n}$$

左と右の不等式を移項して，それぞれμについての式で表し，両式を一緒にすれば，

$$\overline{Y} - 1.960\,\sigma / \sqrt{n} \leqq \mu \leqq \overline{Y} + 1.960\,\sigma / \sqrt{n}$$

これが平均値の95%信頼区間である。

同様にして，平均値の99%信頼区間は，

$$\overline{Y} - 2.576\,\sigma / \sqrt{n} \leqq \mu \leqq \overline{Y} + 2.576\,\sigma / \sqrt{n}$$

で与えられる。

ここで，統計量の信頼限界の意味は，下限と上限で設定された区間が母平均を含む確率が0.95（0.99）

であることを注意しておこう。

《問題》 1990（平成2）年乳幼児身体発育調査結果によると，標本数6,014人の日本人男児の出生時体重の平均値は3.19kg，その標準偏差は0.43kgであった。母平均の95%信頼区間を求めよ。

信頼区間の幅を小さくするには，標本数nを大きくするしかない。工業製品の母集団などではσを小さくすることが技術的に可能だろうが，自然界の母集団ではσを小さくする方法はない。

さて，この考えは標本が十分大きいとき（e.g.，$n \geqq 100$）は応用できるが，少数例では標本標準偏差のバラツキを考えなくてはならない。そこで，登場したのがWilliam S. Gossetである。彼はStudentのペンネームで知られている。

5.3 Studentのt分布

母標準偏差が既知のときの平均値の95%信頼区間の不等式を変形してみよう。

$$-1.960 \leqq \{ (\overline{Y}-\mu) / (\sigma/\sqrt{n}) \} \leqq +1.960$$

中かっこ {} の式は基準正規偏差値，$u=(Y-\mu)/\sigma$に対応する。従って，$(\overline{Y}-\mu)/(\sigma/\sqrt{n})$は正規分布に従う（変数に定数を足したり，掛けたりしたときの基本統計量の変化の観察でこれは確認したことを思い起こせ！）。

母標準偏差が未知のときはσの代りにsを使って上のuに相当するtを考えてみよう。

$$t = (\overline{Y}-\mu) / (s/\sqrt{n})$$

標本平均値mは母平均値μの不偏推定量であったが，sはσの不偏推定量ではなかった。だから，tはもはや正規分布には従わない。このtの分布が**t分布**であり，Gosset(Student)が1908年に発見し，Fisherが1926年に完成した。t分布は正規分布と同様に左右対称であり$-\infty$から$+\infty$にわたる。大標本の場合t分布は$\mu=0$，$\sigma=1$の正規分布とほとんど違わない。標本数が30くらいでは両者の区別は困難である。t分布の特徴は自由度－ここでは$(n-1)$－によって分布のかたちが変ることである。

5.4 標本統計量に基づく信頼限界

それぞれの自由度についてのtの確率分布は，数表のかたちで用意されている（例えば，日本規格協会「t分布のパーセント点」）。

母標準偏差がわからない正規母集団からの標本平均の95%信頼限界（区間）を求めるには，母標準偏差が既知のときのσを標本標準偏差sで置き換え，1.960を「t分布のパーセント点」の$t\,2\alpha(\nu)$の値と置き換えるだけだ。95%信頼限界を求めるとき「t分布のパーセント点」表は片側確率を与えてあるので$\alpha=0.025$ではなく，両側確率の$2\alpha=0.05$で表を引く。自由度ν（ニュー，ローマ字のnに相当）は$n-1$を示す。

$$\overline{Y} - t\,2\alpha(\nu) \cdot (s/\sqrt{n}) \leqq \mu \leqq \overline{Y} + t\,2\alpha(\nu) \cdot (s/\sqrt{n})$$

《問題》 高校3年生の男子50名の身長について，$m=169.9$，$s=5.14$を得た。母平均μの95%信頼区間を求めよ。

《問題》 JMPの例題（somatomery data）から，女子の身長について母平均の95%信頼範囲を求めよ。

5.5 カイ二乗（χ^2）分布

まず，つぎのようにしてχ^2分布を導いてみる。

1つの箱に赤玉と白玉を同数ずつ，ただし非常にたくさん入れておく。この玉を十分によくかき混ぜて，その中から目をつむって50個の玉を取り出して，赤玉，白玉の個数を数える。そしたら，赤玉が27個，白玉が23個であったとしよう。

ところで，この実験での赤玉と白玉の期待値はともに25個である。この観測値と期待値の食い違いの測度を，

Σ {（観測値－期待値）^{2}／期待値}

であらわそう。今の場合は食い違いの測度は0.32となる。この玉を箱にもどし，この抽出実験を何回も繰り返してみる。その結果を，横軸に食い違いの測度を，縦軸に度数をとったグラフで表わそう。この分布が自由度１のχ^2分布である。同様な実験を赤玉，白玉，黒玉の３つの玉で行うと（このとき玉の比率はどの様でもよい），自由度２のχ^2分布が得られる。４種類の玉での実験は自由度３のχ^2分布を生じさせる。

このような考えは，前述のFisherの独立性の精密検定の場合に応用出来ることを後にしめそう。

今度は標準正規分布を示す変数，$u = (Y-\mu) / \sigma$，を考えてみよう。ここで，平均値μ，標準偏差σをもつ正規母集団からｎ個の変量ｕで成る標本を何回も繰り返して抽出実験をしてみる。それぞれの抽出標本でもとの変量Yをuに変換し，Σu^2を計算する。上と同様に横軸にΣu^2を，縦軸に出現頻度をとると，自由度ｎのχ^2分布が求まる。Σu^2を書き直してみよう。

$$\Sigma u^2 = \Sigma \{ (Y-\mu)^2 / \sigma^2 \} = (1/\sigma^2) \Sigma (Y-\mu)^2$$

ここで，母平均μを標本平均$\overline{Y}$に置き換えると，

$$(1/\sigma^2) \Sigma (Y-\overline{Y})^2$$

さて，$s^2 = \Sigma y^2 / (n-1) = \Sigma (Y-\overline{Y})^2 / (n-1)$であったことを思い出そう。上の式にあてはめると，

$$(n-1)\, s^2 / \sigma^2$$

もし正規母集団から大きさｎの標本を，究極的に無限回抽出すると，上の式の値は自由度ｎ－１のχ^2分布を作り出す。

χ^2分布はその値が０から＋∞に広がる確率密度関数だ。

5.6 分散の信頼限界

それぞれの自由度についてのχ^2の確率分布は，数表のかたちで用意されている（例えば，日本規格協会「χ^2分布のパーセント点」）。

比$(n-1)\, s^2 / \sigma^2$について次の式をつくることが出来る：

$$P \{ \chi^2(1-\alpha/2)(\nu) \leqq (n-1)\, s^2 / \sigma^2 \leqq \chi^2(\alpha/2)(\nu) \} = 1-\alpha$$

左の不等式にσ^2を掛け，$\chi^2(1-\alpha/2)(\nu)$で割って，$\sigma^2 \leqq (n-1)\, s^2 / \chi^2(1-\alpha/2)(\nu)$

右の不等式にσ^2を掛け，$\chi^2(1-\alpha/2)(\nu)$で割って，$\sigma^2 \geqq (n-1)\, s^2 / \chi^2(\alpha/2)(\nu)$

一緒にまとめて，

$$P \{ (n-1)\, s^2 / \chi^2(\alpha/2)(\nu) \leqq \sigma^2 \leqq (n-1)\, s^2 / \chi^2(1-\alpha/2)(\nu) \} = 1-\alpha$$

ふたたび$s^2 = \Sigma y^2 / (n-1) = \Sigma (Y-\overline{Y})^2 / (n-1)$から，$(n-1)\, s^2 = \Sigma y^2$を使って，

$$P \{ \Sigma y^2 / \chi^2(\alpha/2)(\nu) \leqq \sigma^2 \leqq \Sigma y^2 / \chi^2(1-\alpha/2)(\nu) \} = 1-\alpha$$

を得る。

《例》高校３年生の男子40名の身長について標本平均$\overline{Y}$=169.6，標本標準偏差ｓ＝5.2748を得た。これから母分散σ^2の95%信頼限界を求めてみよう。

標本の平方和は，

$$\Sigma y^2 = 39 \times 5.2748^2 = 1085.1171$$

95%の信頼限界を求めるのだから，χ^2曲線の95%の面積を囲むパーセント点は，$\chi^2 0.025(39) = 58.1201$と$\chi^2 0.975(39) = 23.6543$である。したがって，母分散$\sigma^2$の95%信頼限界は，

$1085.1171 / 58.1201 \leqq \sigma^2 \leqq 1085.1171 / 23.6543$，すなわち

$18.6703 \leqq \sigma^2 \leqq 45.8740$，あるいは母標準偏差で表せば，$4.32 \leqq \sigma \leqq 6.77$

《問題》JMPの例題（somatomery data）から，女子の身長について母分散の95%信頼範囲を求めよ。

5.7 仮説検定入門

科学的な研究は，(1) 仮説を立てる。この仮説は，ある結果を見せられてもそれは偶然の所産だと考える，というものである（帰無仮説），(2)実験する，(3)仮説を検定する，という方法と順序をもって進むだろう。

仮説検定の方法とは，(1)仮説が正しいとしたときに，実験で得られた結果と，それよりも仮説が捨てられやすい結果をあわせて，これらの結果の起こる確率を計算する，(3)あらかじめ決めておいた小さな（滅多に起こりそうもない）確率と比べ，その確率以下なら，偶然起こったと考えるよりも仮説を捨てて反対の仮説をとる，ということで表わせる。

5.7.1 帰無仮説と２つの誤り

ある製薬会社が”新薬”を開発したので，従来のものと比べてよく効くかどうかを調べることになった。従来の薬効に関する特性値が平均100であることが分かっているとしよう。ここで，帰無仮説（null hypothesis, H_0）というのを，H_0：μ=100と書こう。ということは，新薬が従来のものと効き目が変らない，というものだ。会社にとっては，開発費を投じてつくった薬が従来品と変らないでは”歓迎すべからざる”仮説ということになる。願わくば，H_0が捨てられ，対立仮説H_1：μ>100が真と判断されるのが望ましい。つまり，H_0の仮説が当然捨てられるように設定されていたほうがよいということになる。これが，無に帰する意図でたてる仮説という意味で，帰無仮説という。

帰無仮説とは性悪説にたった仮説といえよう。なぜ性悪説にたったほうが都合がよいかというと，Aという人が２種類のビールの目隠しテストで銘柄を当てらるかどうかという官能テストを考えてみよう。銘柄当てが出来ないということは，当るも八卦当らぬも八卦で５分５分ということだ。つまり，1/2の確率ということ。超能力何回やっても当るときは確率は１だ。つまり，銘柄当てが出来る確率Pは，1/2＜P≦1ということだ。さて，Aは銘柄当てが出来るという性善説にたって，これを帰無仮説にするとなると，このPのすべての値について検定をする必要がある。ところが性悪説では，P=1/2という一義的な値を考えればよい。

「標本を抽出した，母集団に関する真の値と，仮説の値には差がない」。この仮説が真であるならば，得られた”差がある”という結果は偶然だけの所産である。しかしその確率は非常に小さく滅多に起こりそうもない。このようなとき，仮説の真／否と仮説の受け入れ／棄却の組み合せは４種類生ずる。真を受け入れる，否を捨てることはともに正しい結論となる。しかし，真に正しい帰仮説を捨てる誤り（第一種の過誤）と間違った仮説を受け入れる誤り（第二種の過誤，その確率はβで表わす）がある。第二種の過誤の小ささは検定力として表される。検定力を増加させるには，(1)標本の大きさを大きくする，(2)検定の性質を変化させること，の方法がある。(2)はノンパラメトリック検定法の援用が考えられる。

ほとんど全ての統計的仮説検定は第一種の過誤の確率（αと表す）α=0.05またはα=0.01（5%または1%の有意水準ともいう）で帰無仮説の棄却を決定する。

ここで，次のことは重要である。すなわち，たとえα=0.05で帰無仮説が棄却できなかったとしても，帰無仮説を採択すること，仮説が正しいといことにはならない。ただ，”帰無仮説が正しくないとはいえない”ということを意味するに過ぎない。だから，この実験だけでは対立仮説を積極的に認めることができないということを意味するだけだ。

5.7.2 片側検定と両側検定

二項分布の復習をしてみよう。17人の学生をランダムに抽出したら，14人の男子学生と3人の女子学生であった。もし母集団の性比が男子：女子＝１：１（p=q=0.5）ならば，このような標本を抽出する確率は0.005188である。さらにすべての極端な結果（15：2, 16：1, 17：0）を含むとその確率は0.006363となる。すなわち，このような結果が偶然によってだけ得られたということはありそうもないという結論だ。

じつはこの計算は片側検定の考えに基づいている。つまり，うえでは性比１：１の期待値から男子学生が多くなるようなズレの方向にだけ関心をもって計算したのである。もし１：１の期待値から逆の方向へのズレも考えるならば，男子：女子＝3：14とそれより極端なケースについての標本を得る確率をも計算しなければならない。これが両側検定である。

後にでてくる，２集団での２つの平均値の検定のケースでは，片側検定では帰無仮説$\mu_1 > \mu_2$または$\mu_1 < \mu_2$を検定するのが片側検定で，帰無仮説$\mu_1 \neq \mu_2$を検定するのが両側検定である。

5.8 t 分布を使った簡単な仮説検定

《例》ウィスキーのシングルは30mlである。ところが，あるバーテンはワン・フィンガーなどといって，ジガー分量器を使わずに計る。そして，分量器をつかわずとも正確であるとワンフィンガーの腕前を誇っている。そこで，こっそり10杯の出されたシングルを正確に計ったら，平均27ml，標準偏差7mlであった。さて，このバーテンの腕前は彼が誇るほどであろうか？

帰無仮説はH0：$\mu = \mu 0$，対立仮説はH1：$\mu \neq \mu 0$となる。
標本からのｔ（ｔｓ)を計算すると，

$$t\ s = (\bar{Y} - \mu_0) / (s / \sqrt{n}) = (27-30) / (7 / \sqrt{10}) = -1.355$$

ｔ分布のパーセント点の表で$\nu = 9$の両側確率をみると，このような値は0.2と0.3の間にあることが分かる。すなわち，もし帰無仮説が本当に真ならば，3mlとそれ以上の偏差がでる確率は0.2と0.3の間のどこかにあることである。通常のα=0.05をとれば，標本平均27mlが基準30mlと有意に異なっているとはいいがたい。ゆえに，この観測からはバーテンダーのワン・フィンガーの腕前を疑えない。

《問題》somatometry　dataから男子の身長の統計量を求め，その平均値が母平均170cmの母集団と異なるかどうかを$\alpha = 0.05$で検定せよ。ただし，標本集団も母集団もその身長はともに正規分布すると仮定することにする。

生物学，医学では２つ操作（処理という）をするときに，しばしば『対になった標本』を対象にする。対を構成する成員は，同腹同性の２匹のマウス，同じ程度の能力を持つ２人の学生などでもよい。よく用いられるのは，１個体が２回の処理をうけたり，２回の測定をされるような『自己対合self-pairing』のケースだろう。例えば，練習の前と後で血圧を測定したり，同一個人の左と右の上腕囲を測るような場合がそうである。このような実験計画は，２つの処理の比較の精度を増大させる。もちろん，これらは独立の標本ではない。

このような対にした標本では，標本毎に両処理の差（Di）を求めると，その平均（$\bar{D}$）は0になることが期待される。ここで，偏差$D_i - \mu_D$は母集団平均値0をもち，正規かつ独立に分布するという仮定の元に，標本平均値差Ｄは標準偏差（すなわち標準誤差）$\sigma_D / \sqrt{n}$でμ_Dのまわりに正規分布する。だから，

$$t = (\bar{D} - \mu_D) / (s_D / \sqrt{n})$$

は自由度n-1のｔ分布に従う。となると，このｔ分布は，さきにウィスキーのｔ検定のと同じようにして，H0：$\mu_D = 0$の検定に用いられる。このケースではＹがＤに，$\mu 0$がμ_D（すなわち０）に対応するから，

$$t\ s = \bar{D} / (s_D / \sqrt{n})$$

を計算し，自由度n-1について$\alpha = 0.05$（または0.01）でｔ分布表の値とくらべればよい。

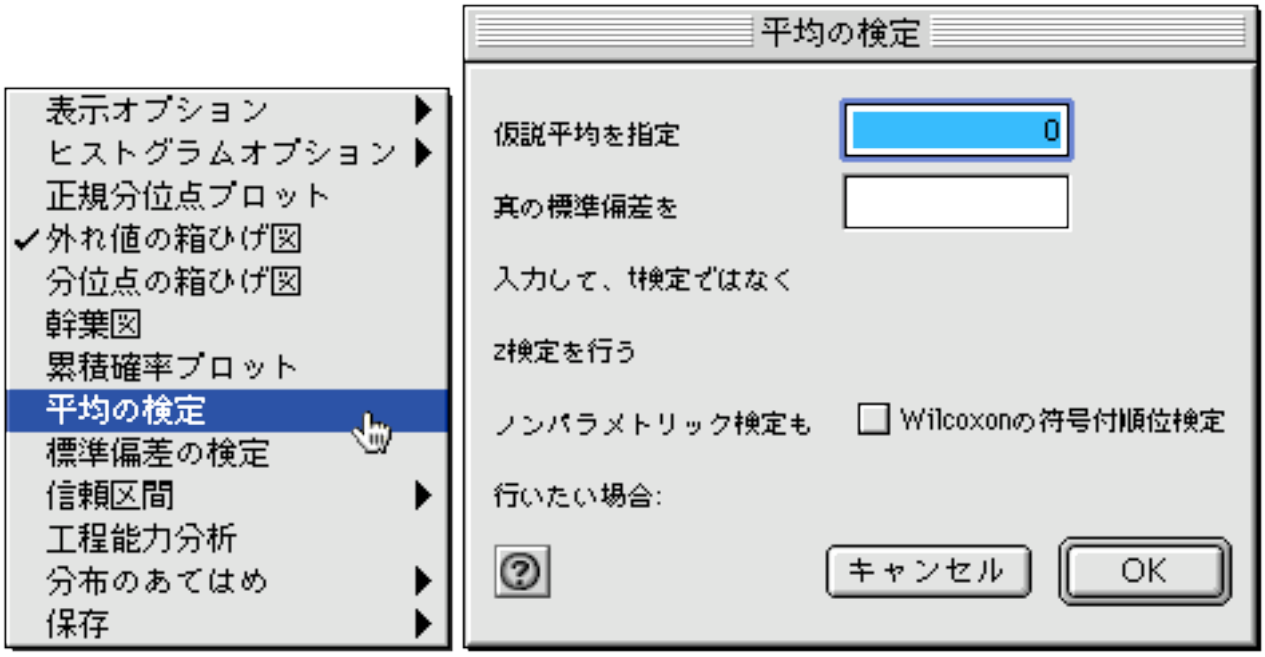
《例》つぎのデータはNewman & Meredith (1956)による米国の白人女児の下顎角幅（cm）の成長のデータである。15人の同一個人を５歳と６歳の２回計測した。

個体	5歳	6歳	個体	5歳	6歳
1	7.33	7.53	9	7.49	7.68
2	7.49	7.70	10	7.44	7.66
3	7.27	7.46	11	7.95	8.11
4	7.93	8.21	12	7.47	7.66
5	7.56	7.81	13	7.04	7.20
6	7.81	8.01	14	7.10	7.25
7	7.46	7.72	15	7.64	7.79
8	6.94	7.13			

このデータに基づいて，下顎角幅の年齢差を検定せよ。

データを新しいJMPファイルに用意して，差分データ（6歳－5歳）を**計算式...**を使い新たなカラムにつくる。

この差分変数を**分析⇒一変量の分布**メニューで分析する。タイトルバー（年齢差）の▼マークをクリックして，**平均の検定**サブメニューを選ぶ。



H_0：$\mu_D = 0$ を検定するのであるから，平均の値（Specity Hypothesized Mean）は0のままで良い。

OKボタンをクリックすると，H_0：μ_D＝の仮説を検定結果が表示される。

モーメント

平均	0.2
標準偏差	0.0392792
平均の標準誤差	0.0101419
平均の上側95%信頼限界	0.2217521
平均の下側 95%信頼限界	0.1782479
N	15

平均の検定

仮説値	0
実際の推定値	0.2
df	14
標準偏差	0.03928
t検定	
検定統計量	19.7203
p値(Prob>|t|	<.0001
p値(Prob>t)	<.0001
p値(Prob<t)	1.0000

年齢差の平均値は0.2，その標準誤差は0.01014であることが示されている。 t の値（0.2/0.01014）が19.72と求まる（検定統計量の値）。このような t 値のでる確率（ p 値（Prob>|t|）は0.0001以下（0.01%以下）であるから，年齢差は有意ということになる。

5.9 仮説H_0：$\sigma^2 = \sigma_0^2$の検定

分散においてはこの統計量が正規分布しないことをすでにみた。平方和とσ^2の比はχ^2分布に従うので，標本分散と母分散が異なっているかどうかはχ^2分布をつかって検定する。

《例》ある製品は規則で600gでなければならないとしよう。10個のこの製品からなる標本では，$\overline{Y}$=592.5， s=11.2であることをみた。ここでさらに規則は製品の分散は100を越えてはならないことをうたっているとしよう。さて，この標本分散は100より有意に大きいであろうか？標本からのχ^2（χ_s^2）を計算すると，

$\chi_s^2 = (n-1)\, s^2 / \sigma^2 = 9 \times 125.44 / 100 = 11.29$

ここでは片側検定をすればよいので，この11.29の値がとる確率は，帰無仮説が真のとき0.10よりは高い

が，0.50よりは低い。したがって，χ_s^2は5%水準では有意ではない。

《問題》somatometry dataから男子の身長の統計量を求め，その標本分散が母分散20の母集団と異なるかどうかを$\alpha=0.05$で検定せよ。

6.0　F分布

正規分布，t分布，χ^2分布がでてきたところで，ついでにF分布について学ぼう。F分布は後に学ぶ分散分析で必要になる分布だ。

母平均μと母分散σ^2をもった正規母集団から大きさn1とn2の標本をを抽出する（n1＝n2であってもよい）。あるいは，母平均は異なるが母分散は等しい２つの正規母集団からそれぞれ大きさn1とn2の標本を抽出しても同じである。これらの標本から標本分散s_1^2とs_2^2をもとめる。これから，

$$Fs=s_1^2/s_2^2$$

を計算する。s_1^2とs_2^2は母分散σ^2の推定値であるから，この値は1に近い値をとるだろう。この抽出実験を繰り返したときのFsの期待分布（横軸にFsを縦軸に出現頻度をとった頻度曲線）は，発表した彼R.A. Fisher(1924)の頭文字にちなんで，F分布と呼ばれる。F分布はt分布やχ^2分布とちがい，２つの自由度ν_1とν_2によって決定される。

F分布とχ^2分布の関係を見てみよう。まえに，正規母集団N（μ，σ^2）から大きさnの標本を，究極的に無限回抽出すると，$(n-1)s^2/\sigma^2$の値は自由度n－1のχ^2分布を作り出す，ことを学んだ。この式の分子を，(n - 1)で割ってやると，$Fs=s^2/\sigma^2$，すなわちF（n－1，∞）のF分布に従う。一般に，

$$\chi(\nu)^2/\nu=F(\nu,\infty),$$

となるからχ^2表をF表で代用できる。

6.1　帰無仮説$H_0：\sigma_1^2=\sigma_2^2$

F分布を利用して，２つの標本によって代表される２つの正規母集団が，同じ分散を持つという帰無仮説を検定できる。のちの，２つの平均値が同一母集団からのものかどうかを検定する（平均値の差の検定）方法は，母集団の等分散性を前提としている。体脂肪率を測定する２つの方法のどちらがバラツキが少ない方法かを知りたいときなどに，この検定が応用できよう。

Fsを計算するときの分子の分散（s_1^2）には大きい値をおいて求める（すなわち，Fs＞1である）。というのは，普通はF分布の表（例えば，日本規格協会「F分布のパーセント点」）では分布の右側の端についての値だけが載っているからだ。また，自由度ν_1と自由度ν_2はそれぞれ大きいほうの分散（s_1^2）を求めた自由度と，小さい方の分散を求めた自由度に対応している（実際の自由度の値の大小に関係ない！）

《問題》somatometry dataから男子と女子の身長の標本分散を求め，これらの標本がおなじ母分散σ^2をもつ正規母集団からの独立な無作為標本であるか同かを$\alpha=0.05$で検定せよ。

7.0　正規分布，t分布，χ^2分布，F分布の計算機シミュレーション

これまでに，正規分布をする母集団からさまざまなやり方で標本を抽出し，その統計量の期待分布がt分布，χ^2分布，F分布をすることを知った。ここでは，それらの分布の数理統計学的な検討はさておき，Macintosh上のJMPによって各分布のシミュレーションをやってみよう。

7.1　正規分布のシミュレーション

シュハートのノーマルチップスというものがある。これは998枚の円形のチップでそれには0から60までの番号が書いてある。その番号の枚数は，平均値30，標準偏差10の正規分布を構成するように決められている。ところで，計算機は自然数だけでなく実数を扱うことが出来るので，0から60の自然数にこだわらなくてもよい。だから，チップを丁度1,000枚にして，正規分布N（30，10）をJMPでつくってみよう。

(1) 新規のJMPファイルをつくる。

JMPアプリケーションを直接ダブルクリックすると新規ファイルが開かれる。すでにJMPが起動されているときは，**ファイル**メニュー⇒**新規**，で新規JMPファイルを準備する。ここでは列1に「Population」（母集団）と名前をつけよう。

(2) 1,000行分のセルを用意する。

Populationカラムを選択し，**行**メニュー⇒**行の追加...**，を選ぶ。**追加する行数：**のボックスに，1000を入力し，OKをクリックする。

(3) 正規データを作り出す。

Populationカラムを選択しておいて，**列**メニュー⇒**計算式**を選ぶ。電卓ウィンドウが開いたら，関数一覧ウィンドウの上欄（下図では関数（すべて）欄）より**乱数**⇒Random Normalを選ぶ。

電卓ウィンドウのキーパッドの×をクリックして10を入力してリターンを押す。ついで，電卓ウィンドウのキーパッドの＋をクリックして30を入力してリターンを押す。すると，計算式表示欄には，**Random　Normal() * 10+30**と表示されるはずだ。ここで，OKをクリックすると，Populationカラムに平均値30，標準偏差10の正規分布を示す1,000のデータが記入される。

(4) この正規分布データを保存する。

《問題》Populationカラムを分析メニュー⇒一変量の分布で分析せよ。統計量はどうであったか？

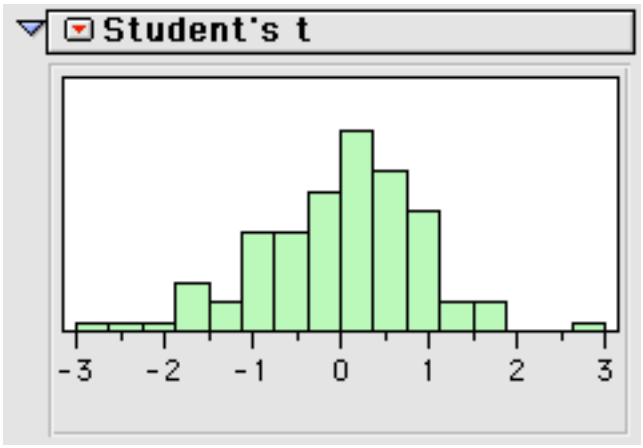
Population欄の▼をクリックし，**分布のあてはめ**⇒**正規**または**正規分位点プロット**オプションで正規性をチェックしてみよ。例数を1,000から5,000；10,000と増やしたらどうか？　Populationカラムを選択して，**行**メニュー⇒**行の追加**で4,000を入力すると自動的にデータを入力してくれ，合計5,000の正規分布のデータが得られる。この標本数（n）が増えるときに，標準誤差はどうかわるか？　95%信頼限界（**モーメント**欄の**平均の上側95%信頼限界**と**平均の下側95%信頼限界**の値で示されている。正規分布のパーセント点の代りにtを，σの代りにsを使っている。すなわち母集団を対象にではなく，標本からの信頼限界を与えていることに注意。）はどのように変わるか？

《問題》母平均30，母標準偏差10をもつ正規母集団から，40例の標本を100組抽出して統計量を求めよ。第2，第3の標本を作り出すには，計算式表示欄の式全体をドラッグして選択し。これをコピーする。

新たな変数（Sample 2, Sample 3, ...）をつくり，それぞれの計算式表示欄に先にコピーした式をペーストしてやればよい。一人では大変だろうから，何人かで手分けしてやるとよい。

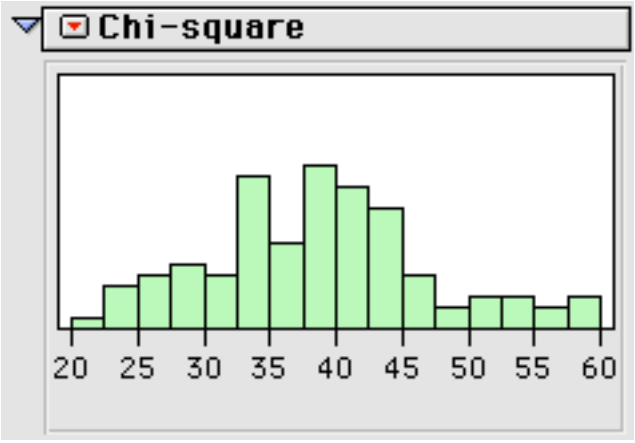
7.3 t 分布のシミュレーション

上の100例のデータから求めた t の分布はつぎのようになった。各々の標本は40例からなるから，これは自由度$\nu=39$の t 分布だ。



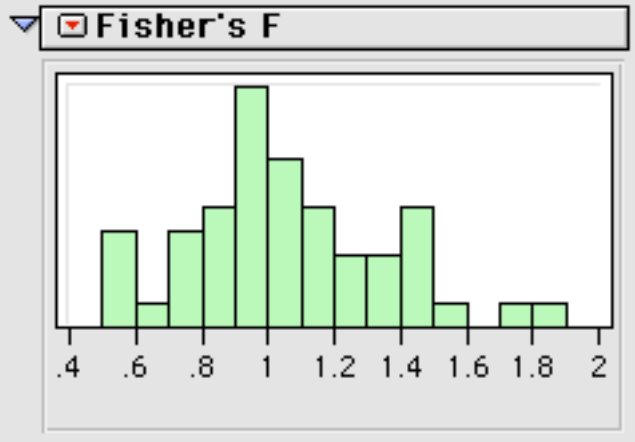
7.4 χ^2分布のシミュレーション

上の100例のデータから求めたχ^2の分布はつぎのようになった。各々の標本は40例からなるから，これは自由度$\nu=39$のχ^2分布だ。



7.5 F 分布のシミュレーション

100例のデータから第１行：第２行，第３行：第４行というように奇数行番号のデータと偶数行番号のデータとのペアーをつくった。この50例のデータから求めたＦ分布は次のようになった。これは，自由度$\nu_1=39$，$\nu_2=39$のＦ分布だ。



8.0　分散分析入門（一元分散分析）

分散分析（analysisofvariance,ANOVA）はR.A.Fisherによって1920年代に，農業試験を通じて展開された。分散分析の利用法の一つに，二つあるいはそれ以上の標本平均が同じ母平均をもつ母集団から得られたものかどうかを検定すること，がある。すなわち，帰無仮説H_0：$\mu_1=\mu_2=...=\mu_n$の検定法である。

8.1　分散分析の考え方

8.1.1　各種睡眠薬を飲んだときの睡眠時間はバラツク

ある製薬会社が４種類（A，B，C，D）の睡眠薬の薬効を比較しようとした。12人のボランティアをランダムに３人づつ４組にわけて，睡眠薬をあたえて睡眠時間を測ってみた。その結果は次のようになった。これから４種の睡眠薬の薬効に差があるかどうかを考えてみよう。

表8.1.1　観測値

繰り返し＼薬	A	B	C	D	
1	18	20	12	14	
2	17	19	15	13	
3	13	15	12	12	
平均	16	18	13	13	15

もしも，表8.1.1のデータが次のように（表8.1.2）きれいであったなら，薬効差は疑いの余地はないだろう。

表8.1.2　理想データ

繰り返し＼薬	A	B	C	D	
1	16	18	13	13	
2	16	18	13	13	
3	16	18	13	13	
平均	16	18	13	13	15

しかし，現実は体質的に睡眠時間の長いヒト，たまたま実験時に疲れていて睡眠時間が長くなったヒ

ト，神経が高ぶって眠れなかったヒトなどのように偶然誤差のために，表8.1.1のようになるのが普通だ。

8.1.2 バラツキはバラセ

さて，表8.1.1の総平均15時間はなにを意味するのであろうか？これは薬効差がなくて，12人の被験者の生理状態が同じで，さらに偶然によるもろもろの誤差がないときの睡眠時間だということになる。だから，観測値から総平均を引いたのこりは薬効差，個人差，偶然誤差が混じったものと考えるべきだ。このようにして表8.1.1のデータはバラツカナイ部分とバラツク部分とに分けられる。ではこれをバラシてみよう。

元のデータのバラツキ（観測値）					バラツカナイ部分（総平均）					バラツク部分			
18	20	12	14	=	15	15	15	15	+	+3	+5	-3	-1
17	19	15	13		15	15	15	15		+2	+4	0	-2
13	15	12	12		15	15	15	15		-2	0	-3	-3

バラツク部分は薬効差，個人差，偶然誤差からなっていて，このうち薬効差は，

薬効差によるバラツキ＝各睡眠薬のごとの平均－総平均

として求めることができる。しかし，被験者は１人が１回の実験しか受けてないので，個人差としてバラスことはできない。これは偶然誤差にふくめて扱わざるをえない。したがって，上式の右辺はつぎのようにさらにバラスことができる。

	バラツカナイ部分（総平均）					薬効差によるバラツキ					誤差によるバラツキ			
=	15	15	15	15	+	+1	+3	-2	-2	+	+2	+2	-1	+1
	15	15	15	15		+1	+3	-2	-2		+1	+1	+2	0
	15	15	15	15		+1	+3	-2	-2		-3	-3	-1	-1

8.1.3 どのようにして薬効検定をするか

さて，「睡眠薬の薬効に差があるか？」ということを上式と関連づけて，どのように表現したらよいだろう。もし，睡眠薬の効果に差があれば，上式の『薬効差によるバラツキ』部分のバラツキが非常に大きくなると予想できる。では，何に対して大きいとか小さいとかを評価したらよいだろう。ここでの比較相手は『誤差によるバラツキ』だ。これは，たとえ睡眠薬の効果に差があっても，個人差や偶然による誤差のバラツキ（誤差によるバラツキ）と同じくらいなら，実用上は薬効に差がないと考えても不都合はあるまい。

8.1.4 分散分析の実行

バラツキの大きさは**不偏分散**unbiased variance（分散，平均平方mean squareともいう。偏差平方和／自由度）で推定する。添字ｉはグループ（睡眠薬の種類）を，ｊはグループの繰り返しをあらわすことにする。グループ数をｋ（＝４），グループｉの繰り返し数をｎｉ（ここではｎ１＝ｎ２＝...＝ｎｋ＝３）とする。

バラツク部分全体の偏差平方和

$$= \Sigma \left(Y_{ij} - \overline{Y}.. \right)^2$$

$$= (+3)^2 + (+2)^2 + (-2)^2 + (+5)2 + \ldots + (-2)^2 + (-3)^2$$

$$= 90$$

薬効差の偏差平方和

$$= \Sigma \left\{ \left(\overline{Y}_{i}. - \overline{Y}.. \right)^2 \cdot n_i \right\}$$

$$= (+1)^2 \times 3 + (+3)^2 \times 3 + (-2)^2 \times 3 + (-2)^2 \times 3$$

$$= 54$$

バラツク部分全体を薬効差によるバラツキと誤差によるバラツキにバラシたのだから，

誤差によるバラツキ

＝バラツク部分全体－薬効差によるバラツキ

$=(Y_{ij}-\overline{Y}..)-(\overline{Y}_{i}.-\overline{Y}..)$

$=Y_{ij}-\overline{Y}_{i}.$

誤差の偏差平方和

$=\Sigma(Y_{ij}-\overline{Y}_{i}.)^{2}$

$=(+2)^2+(+1)^2+(-3)^2+(+2)^2+\ldots+(0)^2+(-1)^2$

＝36

ここで，上の数字から

バラツク部分の偏差平方和＝薬効差の偏差平方和＋誤差の偏差平方和

という関係が成り立つことがわかる。

さて，偏差平方和が計算できたので，分散の計算に必要な自由度を求めてみよう。自由度はさきに分散を計算するときはｎ－１であったことを思い出そう。これは，平均値を１つ使っているから，つまり平均値を１つ使うと自由度が１つ減ると考えられる。

バラツク部分全体の偏差平方和の自由度（データ総数－１）

$=(\Sigma n_i)-1$

＝3+3+3+3-1＝11

薬効差の偏差平方和の自由度（睡眠薬の数－１）

＝ｋ－１

＝3

誤差の偏差平方和の自由度（データの総数－睡眠薬の数）

$=(\Sigma n_i)-k$

＝8

ここで，上の数字から

バラツク部分全体の自由度＝薬効差の自由度＋誤差の自由度

という関係が成り立つことがわかる。

いよいよ，薬効差分散と誤差分散が計算できる準備が整い，分散比Ｆも計算できるようになった。つまり，先に述べたＦ分布表を使って薬効差があるかないかの検定ができるわけだ。

Ｆ＝(54 / 3) / (36 / 8)＝4.0

自由度$\nu 1=3$，$\nu 2=8$の$\alpha=0.05$の値は4.066であるから，帰無仮説H0：$\mu 1=\mu 2=\mu 3=\mu 4$は棄却できない。すなわち，薬効差があるとはいえない。

なお，薬効差の分散をクラス（グループ）間の分散という意味で**級間分散**，誤差の分散をクラスのなかの分散という意味で**級内分散**ともいう。

このように，標本グループがただ一つの基準で分類されている意味で，この分析方法を**一元分散分析（one-way ANOVA）**という。

8.2　ＪＭＰによる分散分析

上での計算をＪＭＰで実行してみよう。ここでは分析メニューの二変量の関係分析法を使う。

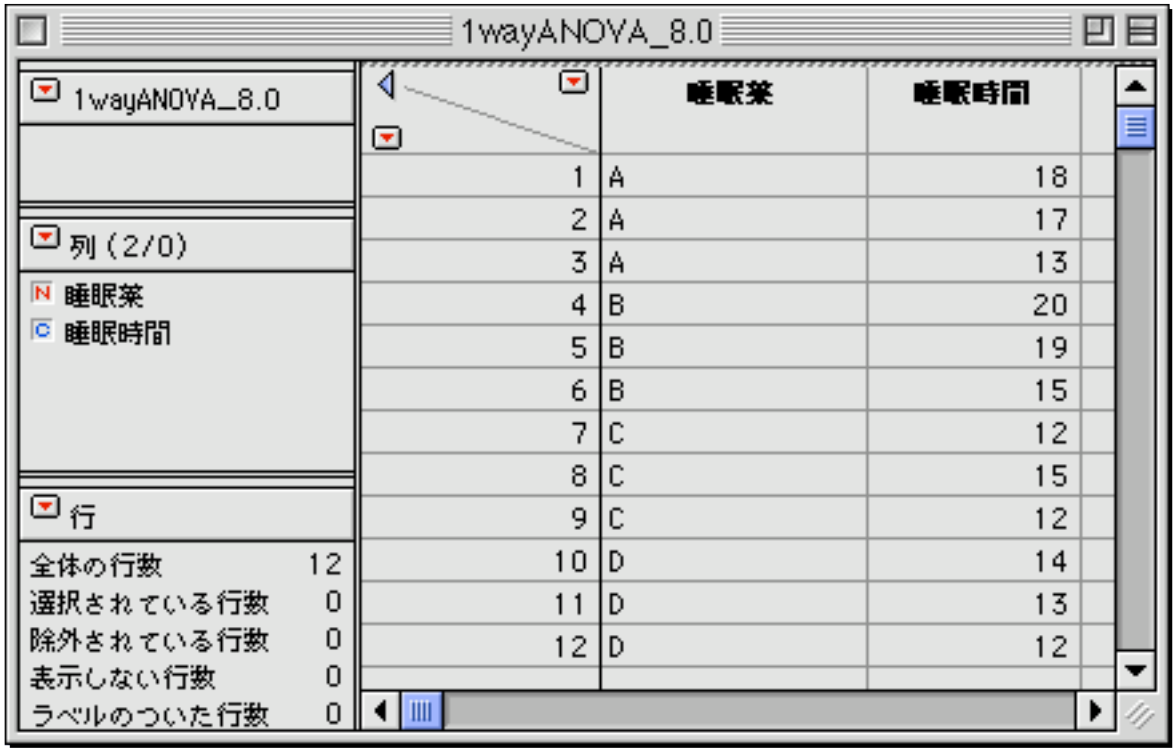
(1) データの準備

ファイル⇒新規のメニューを選び，新規データーシートをつくる。列1を選び，**列**メニュー⇒**列情報...**で睡眠薬のカラム（medicine）をつくる

ここで，データのタイプとして文字型（Character）をチェックするのを忘れないように。暗黙の型宣言では**数値**となっている。この**数値**では，A,B,C,Dの文字データが入力できない。ここで文字型を指定すると，データ表の変数の種類（尺度）は自動的に名義尺度となる。データ表で，始めに名義尺度を設定してもデータタイプは文字型にならないので，A,B,C,Dは入力できない。

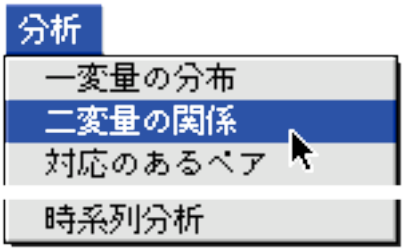
列メニュー⇒**列の新規作成...**で睡眠時間のカラム（sleep time）をつくる。

行メニュー⇒**行の追加**で12行のデータを準備する
データを入力する

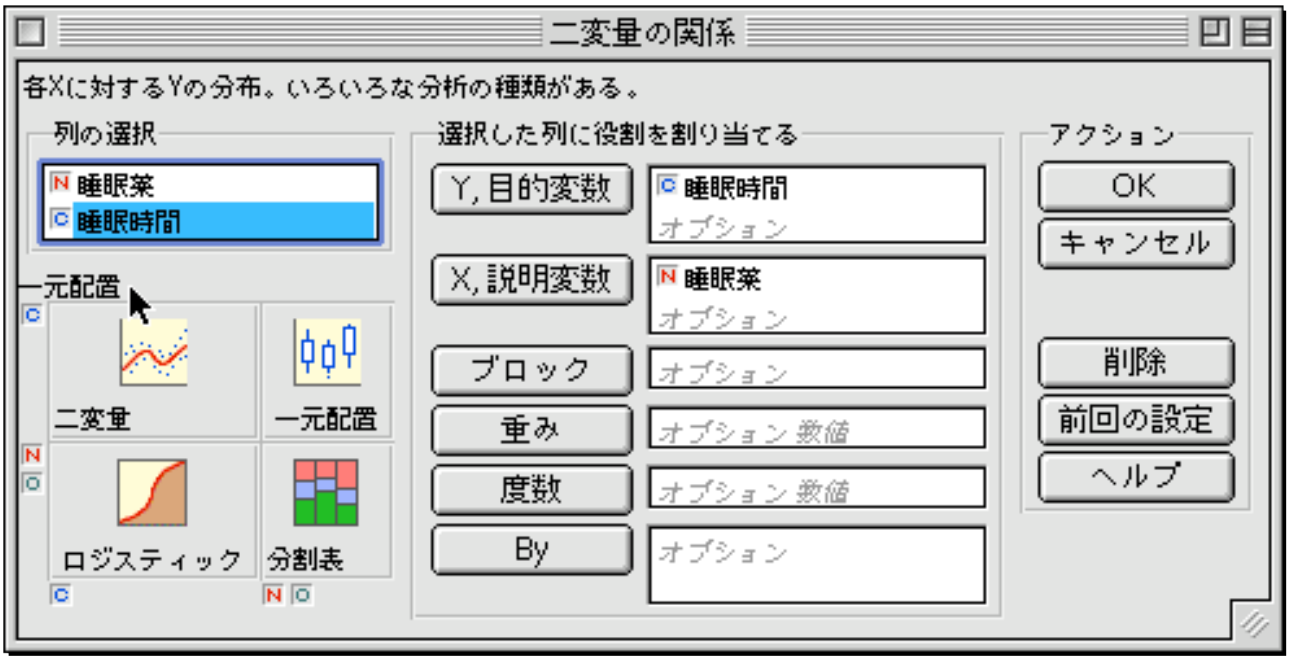


(2) 分散分析の実行

分析メニュー⇒**二変量の関係**を選択する



変数（列）の役割を設定する：睡眠薬はXに，睡眠時間はYに設定。



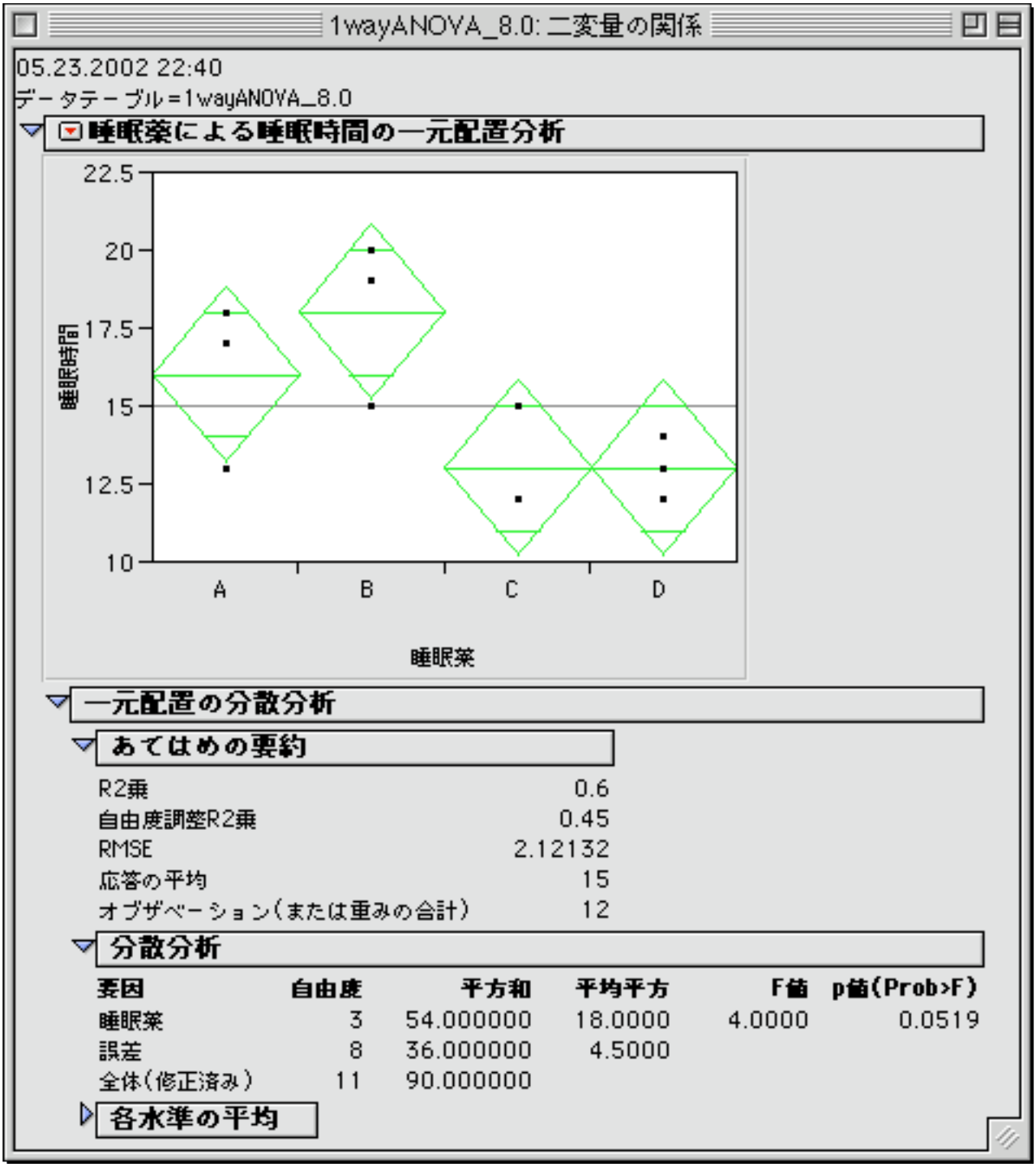
列の選択欄の下の分析方法を示すグラフのタイトル（矢印の先）が**一元配置**になっていることに注目

しよう。Y変数の尺度がC（間隔尺度），X変数の尺度がN（名義尺度）の組み合わせでは，二変量の関係分析を選ぶと自動的に一元分散分析法が選択される。

(3) 分散分析表を表示させる

タイトルバー（睡眠薬による睡眠時間の一元配置分析）の▼マークをクリックして，**平均/Anova/t 検定**を選ぶ。

グラフにダイヤモンドがつけ加わる。その上下の分割線は平均値を表し，上下の頂点は各睡眠薬での睡眠時間の平均値の95%信頼限界を表している。分散分析表（一元配置の分散分析欄）が得られる。F 値の確率（p値（Prob > F））は0.0519（5.2%）であることが分かる。すなわち，帰無仮説　H_0：$\mu_1 = \mu_2 = \ldots = \mu_n$は棄てることができないので，A, B, C, Dの薬の効き目には差はない（とりあえず）とする。



《問題》無作為に抽出した4匹のワタオウサギからダニ幼虫を採集した。その楯板（背の被い）の長さ（μm）を計測した。この楯板のサイズは宿主によって違いがあるだろうか？この結果は生物学的にどの様な解釈ができるだろうか（Sokal & Rohlf：生物統計学，203-207頁）。

hosts	length
A	380, 376, 360, 368, 372, 366, 374, 382
B	350, 356, 358, 376, 338, 342, 366, 350, 344, 364
C	354, 360, 362, 352, 366, 372, 362, 344, 342, 358, 351, 348, 348
D	376, 344, 342, 372, 374, 360

8.3　母数模型（Ｉ型）と変量模型（II型）

分散分析での因子の取り方で２つのモデルが考えられる。データのならべ方，計算方法は変らないが，有意性検定のときの分散比Ｆを構成する分母が異なってくる。

8.3.1　母数模型（fixed effect model）

母数模型（Ｉ型）の分散分析では，もしグループ（級）の平均値の間に差があるならば，それは実験者によって定められた処理効果に基づくものであると考える。目的とするところは，各グループ（グループ数は k）の平均値の間にある真の差を推定することだ。

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

ここで，i=1,2,...,k; j=1,2,...,n; $\varepsilon_{ij} \in N(0, \sigma^2)$。$\alpha$ は有意水準の α とは異なる！

母数模型の分散分析は，各グループの平均値の差を，処理による分散成分の存在を検定して調べる。もし，その様な成分があるならば，すべてのグループが同じ母集団からのものであるという帰無仮説（H0：$\mu_1 = \mu_2 = ... = \mu_k$）を棄て，これらのグループの少なくともいくつかは他のものと異なっているという対立仮説を受け入れる。（つぎには，どの α_i が違っているか調べたくなるだろう。）

このモデルでは，グループの平均値に差があればそれは実験者によって定められた処理効果によるものと考える。たとえば，一群の動物で薬効を調べる実験がこのタイプである。この処理は固定されたもので，実験者が決めたものである。この処理は必ずしも実験者が完全に理解し，そして制御可能なものである必要はない。それが固定されており，そして繰り返しができるものならば母数模型が通用する。

日本で生まれ育った日本人の成人身長と，アメリカで生まれ育った日本人の成人身長を比較するようなときは母数模型だ。処理効果は”日本の環境 vs. アメリカの環境”で，これは固定された処理で，繰り返しができる。つまり，これらの被験者の再調査が可能である。いろいろな人類集団でのいろいろな年齢階級にある体重の比較もこのモデルだ。多く場合この母数模型は，実験の結果や，実験者がその要因を意識的に操作した場合に使う。

8.3.2　変量模型（random effect model）

変量模型（II型）の構造は母数模型に似ている，

$$Y_{ij} = \mu + A_i + \varepsilon_{ij}$$

ここで，i=1,2,...,k；j=1,2,...,n；$A_i \in N(0, \sigma_A^2)$；$\varepsilon_{ij} \in N(0, \sigma^2)$で$A_i$と$\varepsilon_{ij}$は独立である。変量模型が母数模型と違うところは，固定された処理効果α_i　の代りに，偶然の効果A_i（グループ間で違っているが）を考えることだ。この効果は偶然であるからその大きさをあるグループについて推定したり，グループ間の違いを推定することはできないが，分散成分σ_A^2を推定することはできる。その存在を検定し，その大きさs_A^2を推定し，さらに変量模型の分散分析における全変動に対するこの変動の効果の割合を推定することになる。

ネズミを使って腓腹筋の単位面積当りの筋線維数を調べたいとしよう。ネズミの数は５匹（$k = 5$）で，それぞれのネズミから３つの標本（$n = 3$）を得たとする。この５匹のネズミを無作為標本と考える。一つの腓腹筋からの３標本の分散（σ^2）は技術的な違いや採取場所の違いによるものだ。５匹のネズミ間の分散（σ_A^2）があるなら，それは遺伝的，環境的な違いによるだろうが，その性質についてはなんら情報を得られない。３標本の分散σ^2に加えてネズミ間の分散σ_A^2があるだろうかを考えるのが変量模型の分散分析ということになる。もし，σ^2が小さくσ_A^2が比較的大きいなら，ｎを少なくしてｋを多くするべきだろう。また，この逆のことも成り立つだろう。このようにして，繰り返しレベルでの実験誤差の推定の問題を扱うことができる。

握力の変動を家族間の違いと一家族内での兄弟姉妹の違いを調べるとしよう。このようなときは家族内の分散をσ^2と考え家族間の分散成分σ_A^2検定する。家族間には遺伝的な差がありそれが握力を決定するならばσ_A^2の存在が期待できる。そして兄弟姉妹の間の分散はより小さいだろう。そこで，σ^2とσ_A^2相対

的割合から握力の遺伝についての知見が得られるかも知れない。

分散成分σ_A^2の推定

ワタオウサギからのダニ幼虫の楯板の長さの分散分析はこの変量模型である。

誤差分散σ^2の標本推定値s^2は分散分析表のMS Error（級内平均平方）で与えられるから，　分散成分σ_A^2の標本推定値s_A^2を推定してみよう。s_A^2は次の式で計算できる，

$$s_A^2 = (MS_{Model} - MS_{Error}) / n_0$$

$$n_0 = (\Sigma n_i - \Sigma n_i^2 / \Sigma n_i) / (k - 1),$$

ここに，iはグループ（級）を表し，kはグループ数を表す。

ダニの例では，$n_0 = \{37 - (8^2+10^2+13^2+6^2) / 37\} / (4 - 1) = 9.00901$，$s_A^2 = (602.576 - 114.485) / 9.00901 = 54.1781$となる。これを誤差分散（級内分散，$MS_{Error}$）と級間（宿主）の分散成分の総和（$s^2 + s_A^2 = 168.6631$）に対する比率で表すと，$s^2 : s_A^2 = 67.9\% : 32.1\%$となる。

上の例が，植物育種でのケースなら，級間の分散成分は遺伝分散ということになるだろう。握力の家族調査であるならばもう少し，解釈は複雑だ。

9.0　分散分析（2）：二つのグループの比較

統計学でしばしば行われる検定に，二つの平均値の差の有意性検定がある。これは，二つのグループについての母数模型（I型）の**分散分析**で実現できる。もう一つの方法は，二つの平均値の差のための**t検定**である。

9.1　二つの平均値の差のための t 検定

これは計算方法，理解の仕方においても実質的な利点はなく，分散分析と全く同じである。次の式は自由度$\nu = n1 + n2 - 2$のt分布に従う：

$$t_s = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\left[\{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\} / (n_1 + n_2 - 2)\right]\{(n_1 + n_2) / n_1 n_2\}}}$$

このような検定での帰無仮説は，$H_0 : \mu_1 = \mu_2$である。だから，$(\mu_1 - \mu_2)$は0であると期待される。また，ここでの検定は二つの母集団の分散が等しいことを仮定している。これは分散分析の仮定でもある。

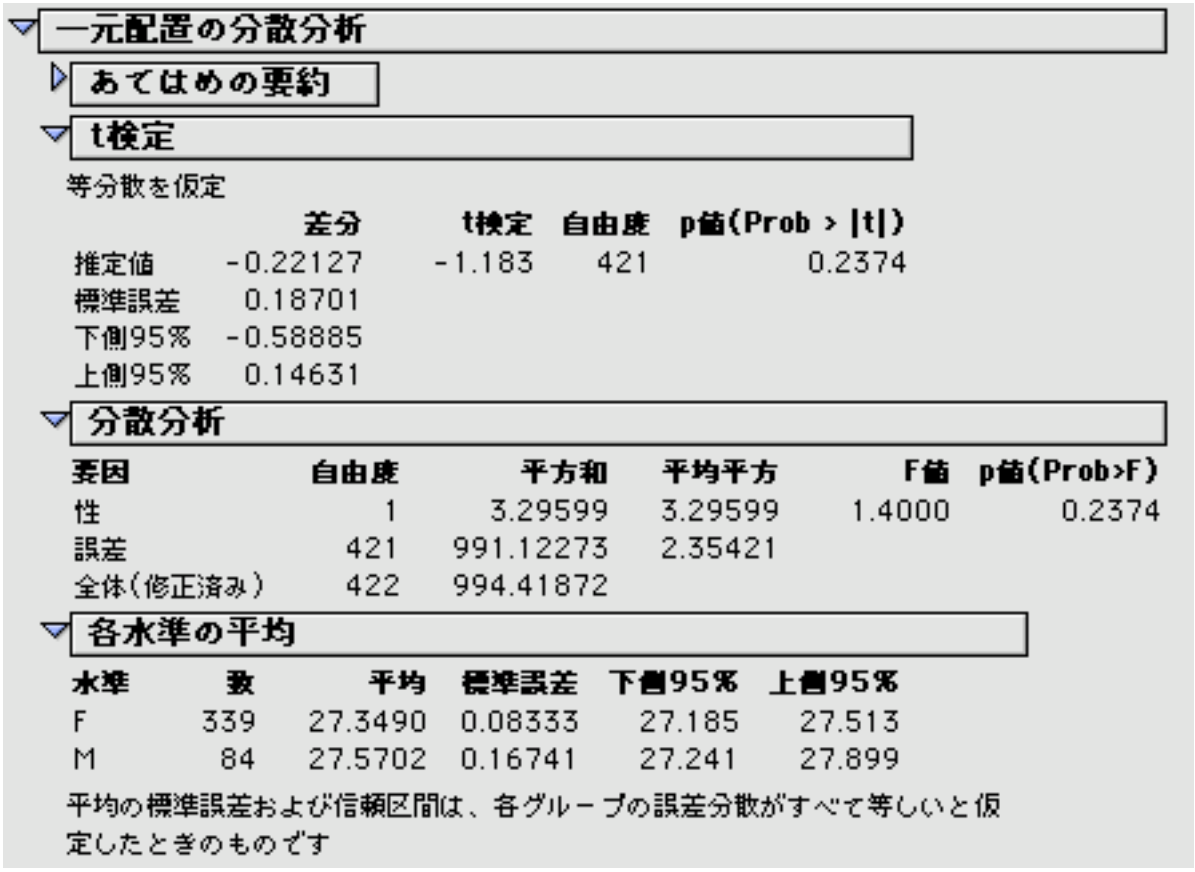
9.2　分散分析とt検定をJMPで比べてみる

《例》somatometry dataで腸骨稜幅の性差を検定してみる。

(1) 分析メニュー⇒二変量の関係を選ぶ。性をXに，腸骨稜幅をYに設定してOKボタンを押す。

(2) タイトルバーの▼マークをクリックして，平均/Anova/t 検定を選ぶ。

t 検定，分散分析表と男女の腸骨稜幅の平均値が表示される。



一元配置の分散分析

あてはめの要約

t検定

等分散を仮定

	差分	t検定	自由度	p値(Prob > |t|)
推定値	-0.22127	-1.183	421	0.2374
標準誤差	0.18701			
下側95%	-0.58885			
上側95%	0.14631			

分散分析

要因	自由度	平方和	平均平方	F値	p値(Prob>F)
性	1	3.29599	3.29599	1.4000	0.2374
誤差	421	991.12273	2.35421		
全体(修正済み)	422	994.41872			

各水準の平均

水準	数	平均	標準誤差	下側95%	上側95%
F	339	27.3490	0.08333	27.185	27.513
M	84	27.5702	0.16741	27.241	27.899

平均の標準誤差および信頼区間は、各グループの誤差分散がすべて等しいと仮定したときのものです

ここで，t 検定欄の推定値の値（t 検定，-1.183）と分散分析欄の性（要因）の値（F値，1.4000）の間には，$t s^2$［ν］＝F s［1，ν］という関係が成り立っていることに注意せよ。t の定義である $t = (\overline{Y} - \square) / s_{\overline{Y}}$ という式を２乗すると，$t^2 = |\overline{Y} - \mu|^2 / s_{\overline{Y}}^2$ となる。ところでこの式は分子，分母ともに分散を表している。これはF の定義である。適当な自由度をとって，t 分布表とF 分布表から両者の関係を確かめてみよ。

さて，この t 検定と分散分析から，5%の有意水準では帰無仮説は棄却できないことが分かった（偶然だけでこれ以上の t，F 値が得られる確率は0.2374すなわち23.7%）。腸骨稜幅は成人男女で違いがない数少ない形質である。

9.3 分散分析の仮定

上に述べた t 検定では $\overline{Y}_1$，$\overline{Y}_2$ は独立，正規に分布し，したがって，$\overline{Y}_1 - \overline{Y}_2$ もまた正規分布する（平均０の正規分布）することが仮定されている。さらに，２つの標本の分散は等しいことをも仮定している。分散分析でも同様である。ここで，分散分析の仮定をまとめてみよう。

(1) 無作為抽出

(2) 独立性

誤差項 ε は互いに独立である。たとえば，秤が十分に整備されてないとき，一連の過小な値に続いて一連の過大な値を生ずることがある。また，実験者がたびたび機器のゼロ点調節をするとき，周期的に過小，過大の値が生ずることがある。このようなときは ε は独立とはならない。この独立性欠如をカバーする簡単な方法はない。

(3) 分散の均一性

6.1でのべたH0：$\sigma_1^2 = \sigma_2^2$ をF検定してみる（このあとのJMPでの実行を参照）。等分散性の帰無仮説が棄却されたときは，この講義の始めにのべた，変数変換を試みる。それでもだめなときは，ノンパ

ラメトリック法を試してみる。ただし，多少の異分散性はそれほど深刻な問題ではないという（パラメトリック検定の頑強性の問題）。

(4) 正規性

Shapiro-WilkのWテスト，KSL（Kolmogorov-Smirnov-Lilliefors）テストでそれぞれの標本の正規性を検定する。正規性が保証されなかったときは，上の分散の均一性の記述と同様の対策を試みる。

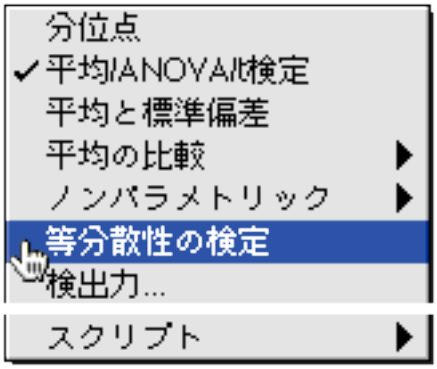
(5) 相加性

繰り返しのない二元分散分析の母数模型では，主効果の検定では交互作用がないことを仮定しなければならない。交互作用とは相乗効果，干渉効果にみられる現象と考えてよいだろう。この対策にはしばしば対数変換が有効である。

（変数変換の妥当性に関しては，Sokal & Rohlf「生物統計学」258～260頁を是非読んで欲しい！）

9.4 等分散性の検定

分析メニュー⇒二変量の関係を選ぶ。タイトルバー（性による腸骨稜幅の一元配置分析）の▼をクリックして等分散性の検定を選択する。JMPは４つのテストを提供している。もしも，このテストが等分散性を棄却したならば，Welchの分散分析を通常のANOVAに代えて使うべきだ。２変量のときのWelchの分散分析は**異分散性の元でのｔ検定**と等価である。



分散が等しいことを調べる検定

水準	度数	標準偏差	偏差の絶対値の平均	中央値からの偏差の絶対値の平均
F	339	1.553434	1.204262	1.203540
M	84	1.454016	1.192885	1.175000

検定	F値	分子の自由度	分母の自由度	p値(Prob>F)
O'Brien[.5]	0.4796	1	421	0.4890
Brown-Forsythe	0.0596	1	421	0.8073
Levene	0.0097	1	421	0.9218
Bartlett	0.5651	1	.	0.4522

Welchの分散分析は、標準偏差が等しくない場合に平均が等しいことを調べる検定。

F値	分子の自由度	分母の自由度	p値(Prob>F)
1.5164	1	133.96	0.2203

t検定
1.2314

（問題）上の男女の腸骨稜幅の等分散性を6.1のＦ検定の方法で検定せよ。

9.5 平均値のあらゆる対の差の検定（多重比較）

母数模型の分散分析の結果が有意と出たとき（有意のときだけ！），どれとどれのグループ間で差があるのかを調べたくなるだろう。この組み合わせは，k (k - 1)/2の数だけある。もっとも簡単な方法はFisherのLSD法（Least Significant Difference，最小有意差法）である。ただし，この方法は簡便なだけに問題を含む。LSD法の問題を調整している方法にTukerあるいはTukey-Kramer法がある。ふつうはこのTukey-Kramer法が使われる。JMPにはこのほかいくつかの検定法が備わっている。詳しくは，高橋行雄・大橋靖雄・芳賀敏郎「SASによる実験データの解析」（1989，東京大学出版会）のpp. 46-70，336-344を参照せよ。

《例》例題JMPファイル1-way-anova_ex-9.5はオーキシンを加えた組織培養液に，いろいろな糖類を加えて豆の切片の成長を調べたものである（Sokal & Rohlf「生物統計学」200頁）。母数模型の分散分析によって糖類の成長への影響を検定してみよう。

(1) 分散分析

分析メニュー⇒**二変量の関係**を選びTreatmentをXにGrowthをYに設定してOKボタンをクリックする。ついで，タイトルバーの▼マークをクリックして平均/Anova/t 検定を選ぶ。

一元配置の分散分析

あてはめの要約

R2乗	0.814412
自由度調整R2乗	0.797915
RMSE	2.335713
応答の平均	61.94
オブザベーション(または重みの合計)	50

分散分析

要因	自由度	平方和	平均平方	F値	p値(Prob>F)
Treatment	4	1077.3200	269.330	49.3680	<.0001
誤差	45	245.5000	5.456		
全体(修正済み)	49	1322.8200			

各水準の平均

水準	数	平均	標準誤差	下側95%	上側95%
Control	10	70.1000	0.73862	68.612	71.588
Fructose	10	58.2000	0.73862	56.712	59.688
Gl. + Fr.	10	58.0000	0.73862	56.512	59.488
Glucose	10	59.3000	0.73862	57.812	60.788
Sucrose	10	64.1000	0.73862	62.612	65.588

平均の標準誤差および信頼区間は、各グループの誤差分散がすべて等しいと仮定したときのものです

この結果から，組織培養液に加えた糖類の豆の切片に対する成長阻害の影響があることが分かった。ではどの糖類に差があるのかを調べてみよう。

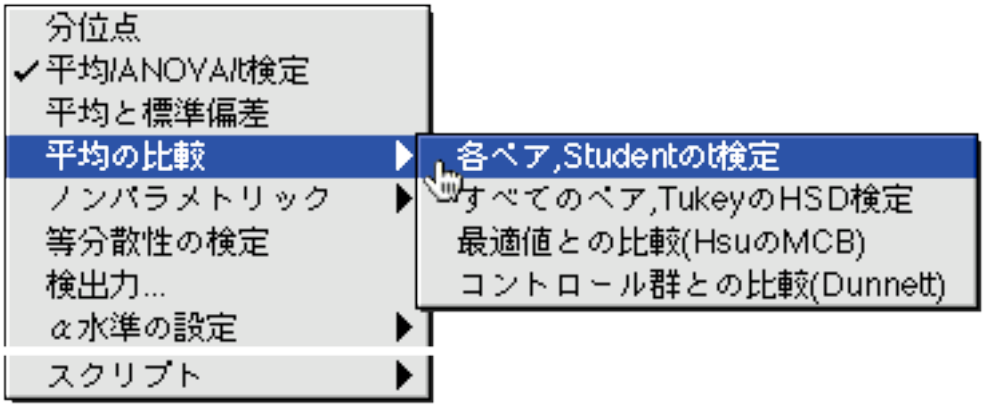
(2) 多重比較

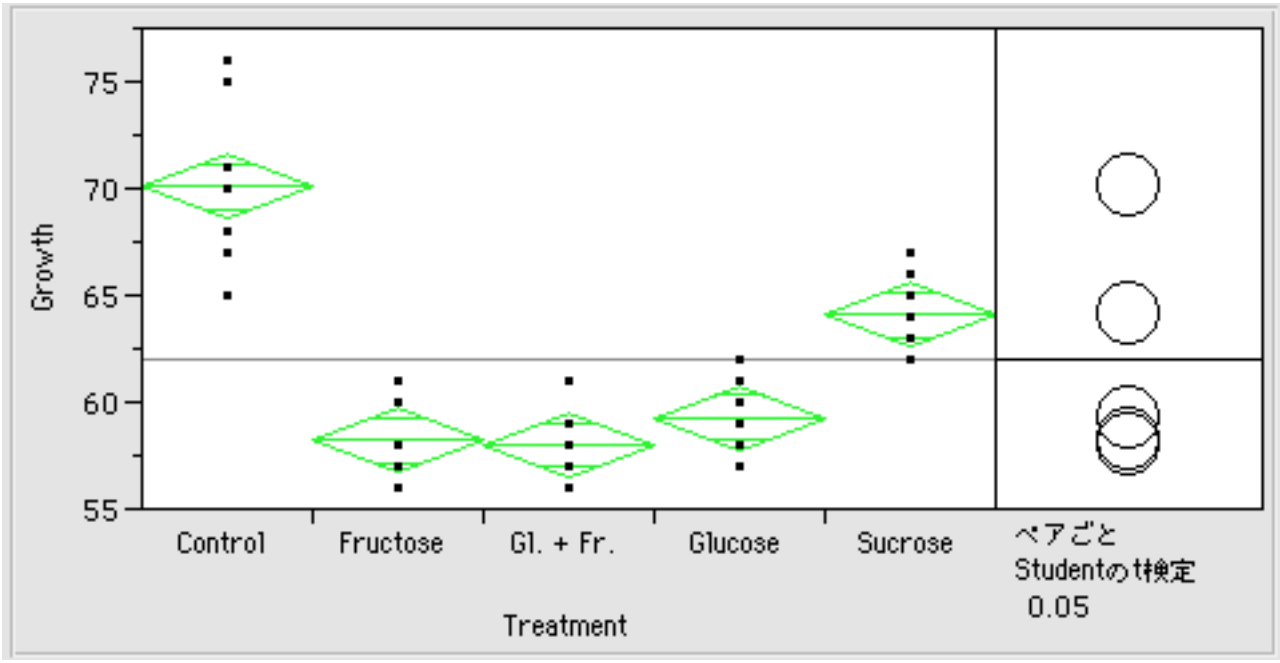
(2-1) LSD法

最小有意差（LSD）法は可能な組み合せの，平均値の差の t 検定と同等である。これは，それぞれを5%の有意水準で検定したときよりも，いずれかの検定が有意となる確率は5%よりもずっと大きくなるという点を調整してない。したがって，予備的な検討にだけ使うのがよいだろう。

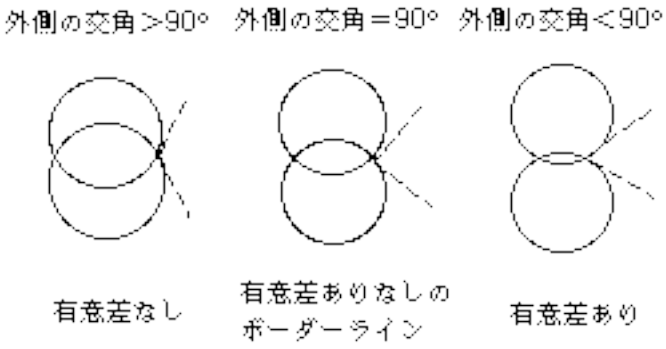
検定の有意水準はあらかじめ5%に設定されているが，変更もできる。

タイトルバーの▼マークから**平均の比較⇒各ペア，Studentのt検定**を選んでみよう。目でみて分かるような多重比較のグラフも描かれる。

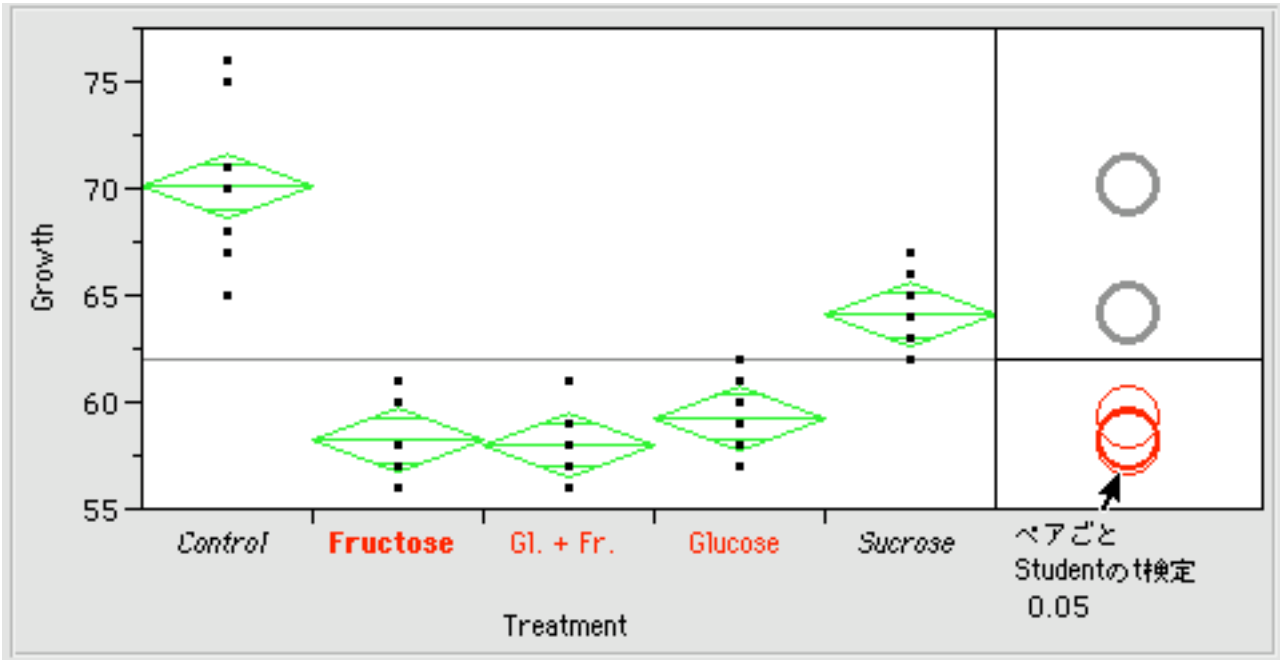




図の右の円の重なり具合が有意差の有無を表す。全く重なっていない時は有意差があることを示している。重なったときは，その角度によって次の様になっている。



１つの円をクリックしてみよ。現在対象になっている円（水準）は太く濃く表示される。同時に散布図のX軸のラベルも太字に表示される。この円（水準）と統計的に有意差がある円（水準）は太く薄い円で示される。同時にX軸のラベルはイタリック体で表示される。対象になっている円（水準）と統計的に有意差がない円（水準）は細字の円で示される。同時にX軸のラベルは普通の書体で表示される。



上のことをLSDの表でみてみよう。行と列で任意のペアーを選び，その交点の数字が正の値であれば，そのペアー間の平均値の差は有意である。

平均の比較

差=平均[i]-平均[j]

	Control	Sucrose	Glucose	Fructose	Gl. + Fr.
Control	0.000	6.000	10.800	11.900	12.100
Sucrose	-6.000	0.000	4.800	5.900	6.100
Glucose	-10.800	-4.800	0.000	1.100	1.300
Fructose	-11.900	-5.900	-1.100	0.000	0.200
Gl. + Fr.	-12.100	-6.100	-1.300	-0.200	0.000

Alpha= 0.05

Studentのt検定を使ったペアごとの比較

t

2.01410

Abs(Dif)-LSD

	Control	Sucrose	Glucose	Fructose	Gl. + Fr.
Control	-2.1039	3.8961	8.6961	9.7961	9.9961
Sucrose	3.8961	-2.1039	2.6961	3.7961	3.9961
Glucose	8.6961	2.6961	-2.1039	-1.0039	-0.8039
Fructose	9.7961	3.7961	-1.0039	-2.1039	-1.9039
Gl. + Fr.	9.9961	3.9961	-0.8039	-1.9039	-2.1039

値が正の場合、ペアになっている平均の間に有意差があることを示します。

これらの図表でみると，コントロールと2%ショ糖（Sucrose）は他のすべての処理に対して有意の差があることがわかる。しかし，2%ブドウ糖（Glucose），2%果糖（Fructose），1%ブドウ糖＋1%果糖の処理はそれぞれ他の処理に対して有意の差はないことも分かる。

しかし，前述のようにこれは，予備的に検討するときの検定法であるので，つぎに普通おこなわれる，Tukey-Kramer法の結果をみてみよう。

(2-2) Tukey-Kramer法

タイトルバーの▼マークから**平均の比較⇒すべてのペア，TukeyのHSD検定**を選んでみると，下のような表が現れる。

Tukey-KramerのHSD検定を使ったすべてのペアの比較

q^*

2.84145

Abs(Dif)-LSD

	Control	Sucrose	Glucose	Fructose	Gl. + Fr.
Control	-2.9681	3.0319	7.8319	8.9319	9.1319
Sucrose	3.0319	-2.9681	1.8319	2.9319	3.1319
Glucose	7.8319	1.8319	-2.9681	-1.8681	-1.6681
Fructose	8.9319	2.9319	-1.8681	-2.9681	-2.7681
Gl. + Fr.	9.1319	3.1319	-1.6681	-2.7681	-2.9681

値が正の場合、ペアになっている平均の間に有意差があることを示します。

表の見方は，先ほどのLSDと同じである。結果もLSDと矛盾しない結果が得られた。

10.0　分散分析（３）：繰返し(交互作用の項）のない二元分散分析

10.1　二元分散分析の考え方

8.1.1の４種類（A, B, C, D）の睡眠薬の薬効の比較実験をもういちど別の実験デザインで考えてみる。こんどは３人のボランティアの１人づつに，それぞれ４種類の睡眠薬をすべて一通り投与して睡眠時間を測ってみた。もちろん，各人が服用する薬の順序は無作為（ランダム）でなければならない（なぜかを考えてみよ！）。このような，観測値が２つの処理，要因（因子）の組み合わさったものから成っている実験計画を二元配置法とよび，その分析方法を二元分散分析（two-way　ANOVA）いう。以前の表8.1.1のような，観測値が１つの処理，要因（因子）から成っている実験計画は一元配置法とよび，その分析方法を一元分散分析（one-wayANOVA）とよんだことを思い出そう。こんどの実験では，薬効差によってタテ（列）の分類ができると同時に，被験者差によってヨコ（行）の分類ができる。これが一元配置法と二元配置法の違いだ。さて，その結果は次のようになった。

表10.1　観測値

薬 / 被験者	A	B	C	D	平均
1	18	20	12	14	16
2	17	19	15	13	16
3	13	15	12	12	13
平均	16	18	13	13	15

前の一元配置法では，

被験者１人あたりの睡眠時間＝総平均＋薬効差＋誤差，

と分解できたが，二元配置法では，

被験者１人あたりの睡眠時間＝総平均＋薬効差＋被験者差＋誤差，

と分解できる。**8.1.2**の誤差によるバラツキがさらに分解されたのが分かる。

元のデータのバラツキ　　バラツカナイ部分　　薬効差によるバラツキ

18	20	12	14
17	19	15	13
13	15	12	12

=

15	15	15	15
15	15	15	15
15	15	15	15

+

+1	+3	-2	-2
+1	+3	-2	-2
+1	+3	-2	-2

+

被験者差によるバラツキ　誤差によるバラツキ

+1	+1	+1	+1
+1	+1	+1	+1
-2	-2	-2	-2

+

+1	+1	-2	0
0	0	+1	-1
-1	-1	+1	+1

10.2 分散分析の実行

10.2.1 偏差平方和の計算

薬因子の水準を i で表し，その水準数を l （ここでは，l=4）とする。被験者因子の水準を j で表し，その水準数をm（ここでは，，m=3）とする。一元分散分析のときと同じようにして，

バラツク部分全体の偏差平方和

$$=\Sigma\ (Y_{ij}-\overline{Y}..)^2$$

$$=(18-15)^2+(17-15)^2+(13-15)^2+(20-15)^2+\ldots+(13-15)^2+(12-15)^2$$

$$=(+3)^2+(+2)^2+(-2)^2+(+5)^2+\ldots+(-2)^2+(-3)^2$$

$$=90$$

薬効差の偏差平方和

$$=m\cdot\Sigma\ (\overline{Y}_{i.}-\overline{Y}..)^2$$

$$=3\times\{(16-15)^2+(18-15)^2+(13-15)^2+(13-15)^2\}$$

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

$$=54$$

被験者差によるバラツキ

$$=l\cdot\Sigma\ (\overline{Y}_{.j}-\overline{Y}..)^2$$

$$=4\times\{(16-15)^2+(16-15)^2+(13-15)^2\}$$

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

$$=24$$

ここで，誤差によるバラツキは，

バラツク部分全体－薬効差によるバラツキ－被験者差によるバラツキ

であり，

$$(Y_{ij}-\overline{Y}..)-(\overline{Y}_{i.}-\overline{Y}..)-(\overline{Y}_{.j}-\overline{Y}..)$$

$$=Y_{ij}-\overline{Y}_{i.}-\overline{Y}_{.j}+\overline{Y}..$$

と表すことができるから，

誤差の偏差平方和

$$=\Sigma\ (Y_{ij}-\overline{Y}_{i.}-\overline{Y}_{.j}+\overline{Y}..)^2$$

$$=(18-16-16+15)^2+(17-16-16+15)^2+(13-16-13+15)^2+(20-18-16+15)^2+\ldots$$
$$+(13-13-16+15)^2+(12-13-13+15)^2$$

$$=(+1)^2+0^2+(-1)^2+(+1)^2+\ldots+(-1)^2+(+1)^2$$

$$=12$$

上の数字から

バラツク部分の偏差平方和＝薬効差の偏差平方和＋被験者差の偏差平方和
＋誤差の偏差平方和

という関係が成り立つことがわかる。

10.2.2 自由度の計算

さて，偏差平方和が計算できたので，分散の計算に必要な自由度を求めてみよう。バラツク部分全体の偏差平方和の自由度

＝ｌ×ｍ－１
＝12-1＝11

薬効差の偏差平方和の自由度

＝ｌ－１
＝3

被験者差の偏差平方和の自由度

＝ｍ－１
＝2

誤差の偏差平方和の自由度

＝（ｌ－１）×（ｍ－１）
＝6

ここで，上の数字から

全体の自由度＝薬効差の自由度＋被験者差の自由度＋誤差の自由度

という関係が成り立つことがわかる。

10.2.3 数学的モデル

ここで，モデルを数学的に整理しておこう。

睡眠時間をYで；薬剤の因子をA，その水準をA_i（i=1..4），その主効果をα_i；被験者の因子をB，その水準をB_j（j=1..3），その主効果をβ_j；誤差をεとすると，

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

上のモデルの制約条件は

$\Sigma\alpha_i = \Sigma\beta_j = 0$を満たし，$\varepsilon_{ij}$は互いに独立に正規分布$N$（0，$\sigma^2$）に従う，と仮定している。

これから検定すべき仮説は，

$H0_A$：水準A_i間に差はない（睡眠薬の違いによる睡眠時間に差はない）

$H0_B$：水準B_j間に差はない（被験者の違いによる睡眠時間に差はない），

ということになる。

10.2.4 分散比の計算

いよいよ，薬効差分散，被験者差分散，誤差分散が計算できる準備が整い，分散比Ｆも計算できるようになった。つまり，先に述べたＦ分布表を使って薬効差，被験者差があるかないかの検定ができるわけだ。

まず，薬効差と誤差の分散比は

Ｆ＝(54 / 3) / (12 / 6)＝9.0

自由度ν_1＝3，ν_2＝6のα＝0.05の値は4.757であるから，今回の２元配置法による実験では，睡眠時間には薬効差があるといえる。

つぎに，被験者差と誤差の分散比は

Ｆ＝(24 / 2) / (12 / 6)＝6.0

自由度ν_1＝2，ν_2＝6のα＝0.05の値は5.143であるから，被験者差も有意である。

ここで，前回の一元配置法による実験での誤差分散は4.5であったことを思い出そう。今回の誤差分散が

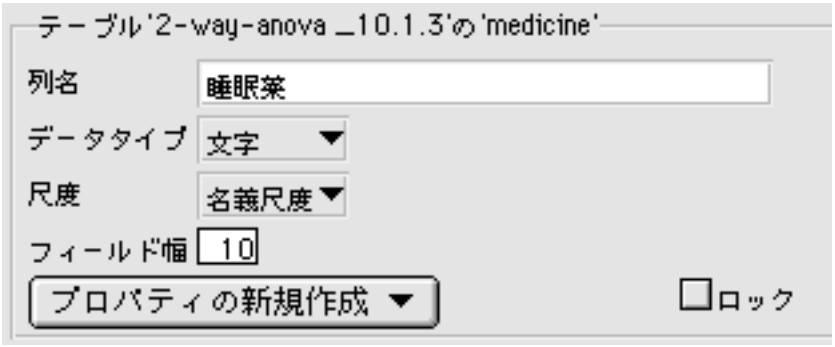
2.0と小さくなったのは，被験者差によるバラツキを誤差から分離できたからである。このように，二元配置法の有効性が分かる。

10.3　ＪＭＰによる二元分散分析

上での計算をＪＭＰで実行してみよう。ここでは**モデルのあてはめ**分析法を使う。

10.3.1　データの準備

列1を選び，**列**メニュー⇒**列情報...**で睡眠薬のカラム（medicine）をつくる



ここで，データのタイプとして文字型（Character）をチェックするのを忘れないように。

ついで，**列**メニュー⇒**列の新規作成...**で被験者のカラム（subject）をつくる。値には１，２，３を入力するが，これは背番号と同じ，すなわち，データタイプは文字型（Character）とする。

列メニュー⇒**列の新規作成...**で睡眠時間のカラム（sleep　time）をつくる。データタイプはそのままの数値（Numeric）でよい。

行メニュー⇒**行の追加**で12行のデータを準備する

データを入力する

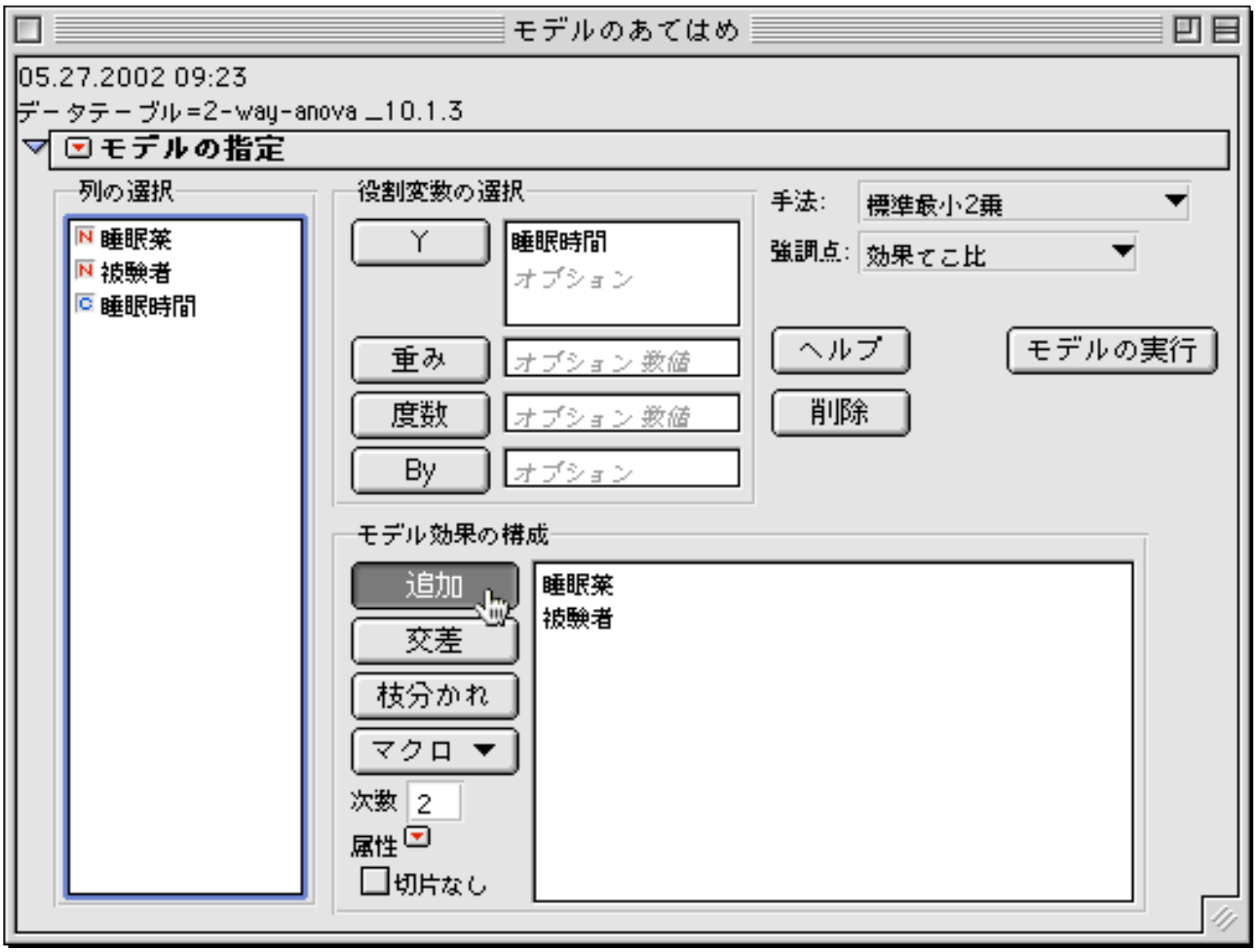
	睡眠薬	被験者	睡眠時間
1	A	1	18
2	A	2	17
3	A	3	13
4	B	1	20
5	B	2	19
6	B	3	15
7	C	1	12
8	C	2	15
9	C	3	12
10	D	1	14
11	D	2	13
12	D	3	12

10.3.2　分散分析の実行

分析メニュー⇒**モデルのあてはめ**を選択する。ここで，X, Yの変数を指定する：

睡眠時間を選んでYボタンをクリックする。睡眠薬と被験者を選んでそれぞれ**追加**ボタンをクリッ

クする。



モデルの実行ボタンをクリックして二元分散分析を実行する。

分散分析表は，効果の検定の欄に表示される：

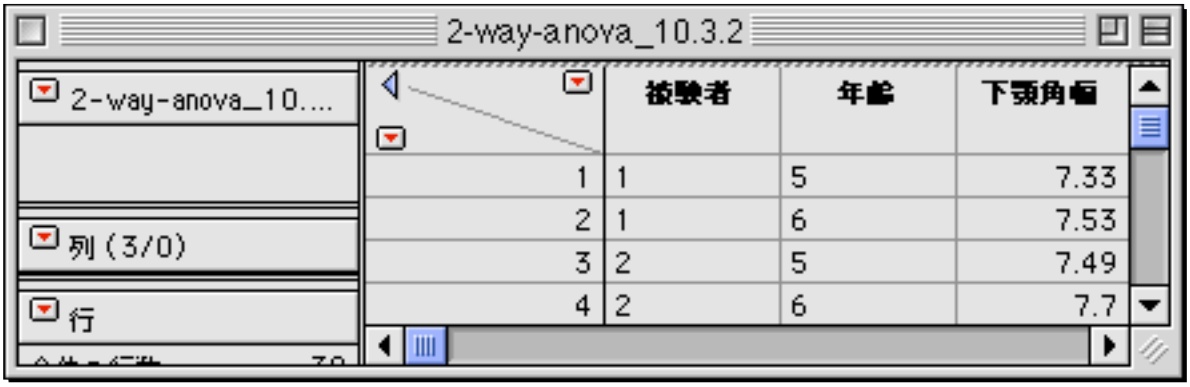
効果の検定

要因	パラメータ数	自由度	平方和	F値	p値(Prob>F)
睡眠薬	3	3	54.000000	9.0000	0.0122
被験者	2	2	24.000000	6.0000	0.0370

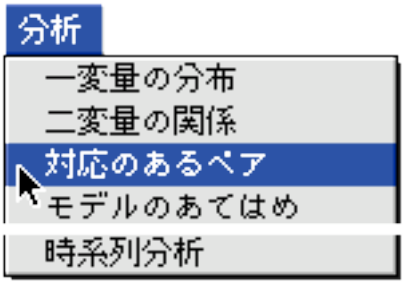
「睡眠薬の違いによる睡眠時間に差はない」という帰無仮説が真（true）である確率は1.2%，また，「被験者の違いによる睡眠時間に差はない」という帰無仮説が真である確率は3.7%であることが分かった。これらの値はあらかじめ決めた有意水準5%よりも小さいから，両方の帰無仮説を棄てる。すなわち，睡眠薬の違いは睡眠時間に効果があり，被験者の違いも睡眠時間に影響を及ぼす。

（問題）繰返しのない二元分散分析が適用される例として，**同一個体を何回も用いるケース**がある。すなわち，ある一定時間をおいて同じ個体のグループが何回も実験対象になる場合だ。もしも，２回の実験の対象になれば，それは対比較といわれる。このケースの t 検定については対応のある t 検定（paired t test）としてすでに5.8章で学んだ。このデータに基づいて，下顎角幅の年齢差を検定せよ。まず二元分散分析で解析し，ついで対応のある t 検定を試みよ。

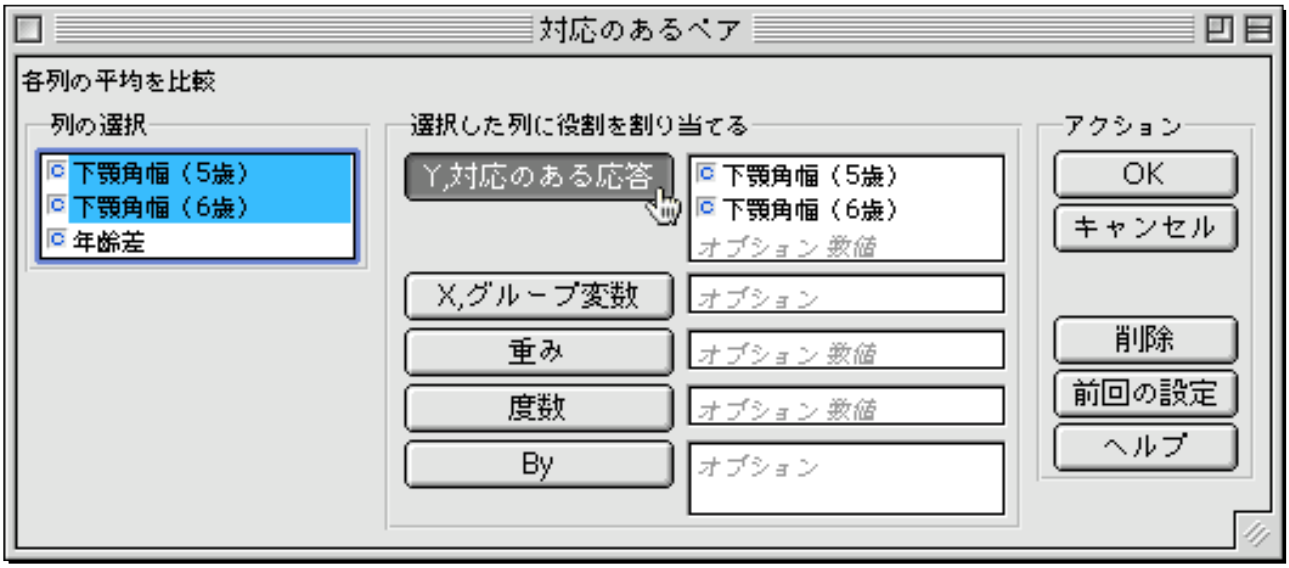
二元分散分析をするためのデータは次のようになる（一部を表示）。

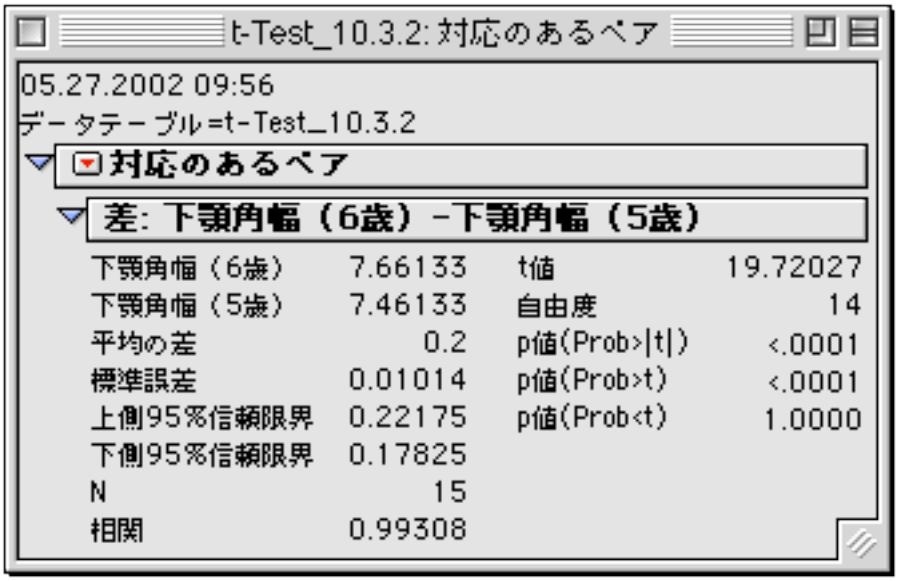


対応のあるt検定については，今度は，5.8章とは別の方法で試みよう。5.8章でのデータファイルを使って，5歳の値と6歳の値をX，Yに設定して分析メニューの**対応のあるペア**で分析する。



下顎角幅（5歳）と下顎角幅（6歳）を選び，**Y，　対応のある応答**ボタンを押してOKボタンで分析を実行する。





二元分散分析と対応のある t 検定の結果が等しいことは，どの数字に表れているか？その理由となる規則は何であったか？

繰返しのない二元分散分析では，因子Ａ（Ｂ）の影響が因子Ｂ（Ａ）のレベルによって変らない，ことを仮定している。すなわち，２つの因子の間に相乗効果，干渉効果といった交互作用（interaction）がないことを意味している。この交互作用が仮定できる場合は，因子の組み合せのサブセットに２例以上の観測値をおいた実験計画法に従い，「繰返し（交互作用）のある二元分散分析」で解析する必要がある。これについては次節で学ぶ。

11.0　分散分析（４）：繰返し(交互作用の項）のある二元分散分析

11.1　分析すべきこと

アフリカツメガエル（xenopus，発生学の実験でしばしば使われるカエル）のオタマジャクシの表皮の細胞分裂に関係する因子を実験的に観察した。分裂細胞に影響するある薬剤溶液に表皮を一定時間浸した。その後，その標本の細胞数に対して分裂している表皮細胞の割合（%）を観察した。薬剤の量は100μg，600μg，2400μgの３種類を；処理した時間は３時間，６時間，12時間，24時間の４種類である。この12種類の組み合せのそれぞれに３標本を使用した。結果は以下のようになった。

表11.1.1 アフリカツメガエルのオタマジャクシの表皮細胞の分裂の割合（%）

薬剤の量＼処理時間	3時間	6時間	12時間	24時間
100μg	13.2 15.7 11.9	22.8 25.7 18.5	21.8 26.3 32.1	25.7 28.8 29.5
600μg	16.1 15.7 15.1	24.5 21.2 24.2	26.9 31.3 28.3	30.1 33.8 29.6
2400μg	9.1 10.3 8.2	11.9 14.3 13.7	15.1 13.6 16.2	15.2 17.3 14.8

上のデータから次のことを分析したい。

(1) 薬剤の量と処理時間は互いに表皮細胞の分裂の割合に影響（相乗作用，相殺作用）を与えているか？

(2) 薬剤の量によって表皮細胞の分裂の割合に差があるのか？

(3) 薬剤で処理する時間によって表皮細胞の分裂の割合に差があるのか？

前回の二元配置法と異なる点は(1)である。二つの処理が同時に加わったときにその効果がそれぞれ一方の因子の効果からは予測できない時，この相互の影響を「交互作用（interaction）」とよぶ。このアフリカツメガエルの例では，高濃度の溶液で長時間処理すると細胞分裂は活性化される，逆に阻害される；あるい低濃度の溶液で長時間処理すると細胞分裂は活性化される，逆に阻害されるといった関係である。

このようなデータ構造を一般型で表してみる。

因子Aの水準数はlでA1，A2，...，Ai，...，Al

因子Bの水準数はmでB1，B2，...，Bj，...，Bm

各々のセルの繰返し数はnで1，2，...，k，...，nとする

表11.1.2　二元配置データの一般型

A＼B	B1	B2	.....	Bj	.....	Bm
A1	Y111 ..... Y11k Y11n	Y121 ..... Y12k Y12n	.....	Y1jk	.....	Y1m1 ..... Y1mk Y1mn
A2	Y211 ..... Y21k Y21n	Y221 ..... Y22k Y22n	.....	.....	.....	Y2m1 ..... Y2mk Y2mn
.....	.....	.....	.....	.....	.....	.....
Ai	Yi1k	Yi2k	.....	Yijk	.....	Yimk
.....	.....	.....	.....	.....	.....	.....
Al	Yl11 ..... Yl1k Yl1n	Yl21 ..... Yl2k Yl2n	.....	Yljk	·····	Ylm1 ..... Ylmk ..... Ylmn

11.2　数学的モデルといろいろな平均

一元配置法の母数（I 型）モデルは

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

繰返し（交互作用）のない二元配置法の母数（I 型）のモデルは

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

とかけた。

繰返し（交互作用）のある二元配置法の母数（I 型）モデルは

$$Y_{ijk} = \mu + \alpha i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

となる。α_i，β_j は主効果を，$(\alpha\beta)_{ij}$ は交互作用を表す。

ここで，$\Sigma\alpha_i = \Sigma\beta_j = 0$，$\Sigma_{i=1}(\alpha\beta)_{ij} = \Sigma_{j=1}(\alpha\beta)_{ij} = 0$，

ε_{ijk} は互いに独立に正規分布N（0，σ^2）に従う，と仮定する。

変量（II型）モデルでは，それぞれの期待値（E）が0であると仮定する。

$$E(\Sigma\alpha_i) = E(\Sigma\beta_j) = 0,$$

$$E(\Sigma_{i=1}(\alpha\beta)_{ij}) = E(\Sigma_{j=1}(\alpha\beta)_{ij}) = 0$$

データの変動は，観測値と平均の差の平方和であった。繰返しのある二元配置分散分析では，平均は4つ計算できる，

表11.2　いろいろな平均

	B1	B2	.....	Bj	.....	Bm	
A1	$\bar{Y}_{11}$	$\bar{Y}_{12k}$	.....	$\bar{Y}_{1j}$	.....	$\bar{Y}_{1m}$	$\bar{Y}_{1.}$
A2	$\bar{Y}_{21}$	$\bar{Y}_{22}$	.....	.....	.....	$\bar{Y}_{2m}$	$\bar{Y}_{2.}$
.....	.....	.....	.....	.....	.....	.....	
Ai	$\bar{Y}_{i1}$	$\bar{Y}_{i2}$	.....	$\bar{Y}_{ij}$	.....	$\bar{Y}_{im}$	$\bar{Y}_{i.}$
.....	.....	.....	.....	.....	.....	.....	
Al	$\bar{Y}_{l1}$	$\bar{Y}_{l2}$	.....	$\bar{Y}_{lj}$	·····	$\bar{Y}_{lm}$	$\bar{Y}_{l.}$
	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$		$\bar{Y}_{.j}$		$\bar{Y}_{.m}$	$\bar{Y}_{...}$

$\bar{Y}_{...}$は総平均を，$\bar{Y}_{i.}$は水準Aiの平均を，$\bar{Y}_{.j}$は水準Bjの平均を，
そして$\bar{Y}_{ij}$は（A_i，B_j）の平均を表す。
繰返しのある二元配置法ではデータY_{ijk}はこの平均を使って，

$$Y_{ijk}=\bar{Y}_{...}+(\bar{Y}_{i.}-\bar{Y}_{...})+(\bar{Y}_{.j}-\bar{Y}_{...})+(\bar{Y}_{ij}-\bar{Y}_{i.}-\bar{Y}_{.j}-\bar{Y}_{...})+(Y_{ijk}-\bar{Y}_{ij})$$

となる。$\bar{Y}_{...}$を移項して，

$$Y_{ijk}-\bar{Y}_{...}=(\bar{Y}_{i.}-\bar{Y}_{...})+(\bar{Y}_{.j}-\bar{Y}_{...})+(\bar{Y}_{ij}-\bar{Y}_{i.}-\bar{Y}_{.j}-\bar{Y}_{...})+(Y_{ijk}-\bar{Y}_{ij})$$

両辺の平方和をとると，一元配置，繰返しのない二元配置法のときと同様に，

$$\begin{aligned}\Sigma\Sigma\Sigma(Y_{ijk}-\bar{Y}_{...})^2 &= \Sigma_{i=1}^{m}\,l\,n\,(\bar{Y}_{i.}-\bar{Y}_{...})^2\\ &+\Sigma_{j=1}^{l}\,l\,n\,(\bar{Y}_{.j}-\bar{Y}_{...})^2\\ &+\Sigma_{i=1}\Sigma_{j=1}\,n\,(\bar{Y}_{ij}-\bar{Y}_{i.}-\bar{Y}_{.j}-\bar{Y}_{...})^2\\ &+\Sigma_{i=1}\Sigma_{j=1}\Sigma_{k=1}(Y_{ijk}-\bar{Y}_{ij})^2\end{aligned}$$

とバラツキがばらせる。すなわち次のようになる。
　総平方和＝要因Aの平方和＋要因Bの平方和＋交互作用A×Bの平方和＋誤差平方和
これを，

$$S=S_A+S_B+S_{A\times B}+S_E$$

と書こう。
そして，これに対する自由度も，

$$\begin{aligned}l\,m\,n-1=&(l-1)\\&+(m-1)\\&+(l-1)(m-1)\\&+l\,m(n-1)\end{aligned}$$

と分解できる。この自由度で割った分散（平均平方）を，

$V_A = S_A /（l - 1）$：要因Aの分散，$V_B = S_B /（m - 1）$：
要因Bの分散，$V_{A\times B} = S_{A\times B} /（l - 1）（m - 1）$：交互作用の分散
$V_E = S_E / l\,m（n - 1）$：誤差分散，級内分散
と書こう。さて，これで分散分析の準備が整った。

11.3 交互作用，主効果の検定

(1) 要因A，Bがともに母数モデルのとき

上の例のアフリカツメガエルの実験がこのモデルである。この例では，

要因	平方和	自由度	分散
A （薬剤の量）	889.52056	2	444.76028
B （処理時間）	798.20750	3	266.06917
A×B（交互作用）	89.88833	6	14.98139
E （誤差）	135.36670	24	5.64000
計	1912.98310	35	

となった。

まず，交互作用の分散と誤差分散の分散比を計算すると，

$F_{A\times B} = V_{A\times B} / V_E = 14.98139 / 5.64 = 2.6561$

$\nu_1 = 6$，$\nu_2 = 24$の自由度に対する有意水準５％のF値を分布表から引くと，

$F = 2.51$

である。よって，薬剤の量と処理時間の交互作用は細胞分裂の割合に有意に影響しているとみなせる。交互作用の分散が有意であったときは，通常の検定はこれで終りである（？！）。交互作用があるのだから，薬剤の量あるいは処理時間の要因を単独に考えても仕方ないからだ。

交互作用の分散が有意でなかったときは，繰返しのない二元分散分析のように，それぞれの要因の分散と誤差分散の分散比を計算する。（つぎの(2)で述べる，こみにした誤差分散に対する分散比で検定することもある）

$F_A = V_A / V_E$

$F_B = V_B / V_E$

(2) 要因Aが母数モデル，要因Bが変量モデルのとき（混合モデル）

要因Aには４種類の睡眠薬を，要因Bにはランダムに選んだ10人の被験者を，それぞれの組み合せを２回ずつ実験した（繰返し数２）ようなときは，

まず，交互作用の分散と誤差分散の分散比を計算する，

$F_{A\times B} = V_{A\times B} / V_E$

もし，この交互作用が有意であるならば，つぎに固定された要因A（母数モデル）の分散と交互作用の

分散との分散比を計算する。

$$F_A = V_A / V_{A\times B}$$

ランダムな要因B（変量モデル）の分散は誤差分散との比をとる，

$$F_B = V_B / V_E$$

この混合モデルでは固定された処理効果にたいする平均平方は交互作用の分散成分を同時に含む。一方，ランダムな要因に対する平均平方は誤差分散とそれ自身の分散成分だけを含み，交互作用の成分は含まない。

もし，この交互作用が有意でないならば，こみにした誤差分散に対して固定された要因A（母数モデル）の分散の分散比を計算することが普通である（？！）。

$$V_{E'} = (S_{A\times B} + S_E) / (\nu_{A\times B} + \nu_E)$$

$$F_A = V_A / V_{E'}$$

ランダムな要因B（変量モデル）の分散は誤差分散との分散比を計算する，

$$F_B = V_B / V_E$$

(3) 要因A，要因Bともに変量モデルのとき

まず，交互作用の分散と誤差分散の分散比を計算する，

$$F_{A\times B} = V_{A\times B} / V_E$$

もし，この交互作用が有意であるならば，つぎにランダムな要因A，Bの分散は交互作用の分散との分散比を計算する，

$$F_A = V_A / V_{A\times B}$$

$$F_B = V_B / V_{A\times B}$$

もし，この交互作用が有意でないならば，こみにした誤差分散に対して固定された要因A（母数モデル）の分散の分散比を計算することが普通である（？！）。

$$V_{E'} = (S_{A\times B} + S_E) / (\nu_{A\times B} + \nu_E)$$

$$F_A = V_A / V_{E'}$$

$$F_B = V_B / V_{E'}$$

11.4 ＪＭＰによる繰返し（交互作用）のある二元分散分析

上での計算をＪＭＰで実行してみよう。ここでは**モデルのあてはめ**を使う。

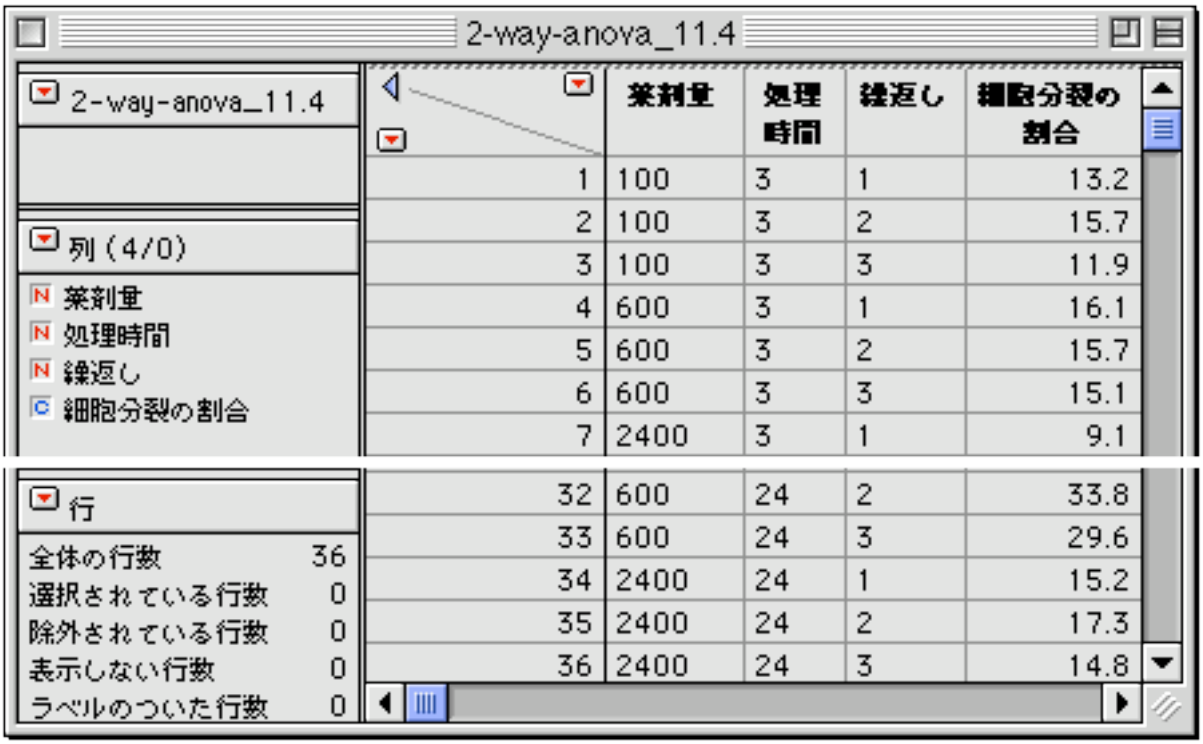
11.4.1 データの準備

ファイル⇒新規メニューで新規データシートをつくる。列1を選ぶ，**列**メニュー⇒**列情報...**，そして**列**メニュー⇒**列の新規作成...**で薬剤の量（amount），処理時間（time），繰返し（trial）のカラムをつくる。これらのデータのタイプとして文字型（Character）をチェックするのを忘れないように。

列メニュー⇒**列の新規作成...**で細胞分裂の割合（cell　div.）をつくる。Data　Typeはそのままの**数値**（Numeric）でよい。

行メニュー⇒**行の追加**で36行のデータを準備する。

データを入力する。



11.4.2 分散分析の実行

分析メニュー⇒**モデルのあてはめ**を指定すると，次の図のような変数の役割設定のダイアログがでる。 Yの変数として細胞分裂の割合（cell div.）を指定する：

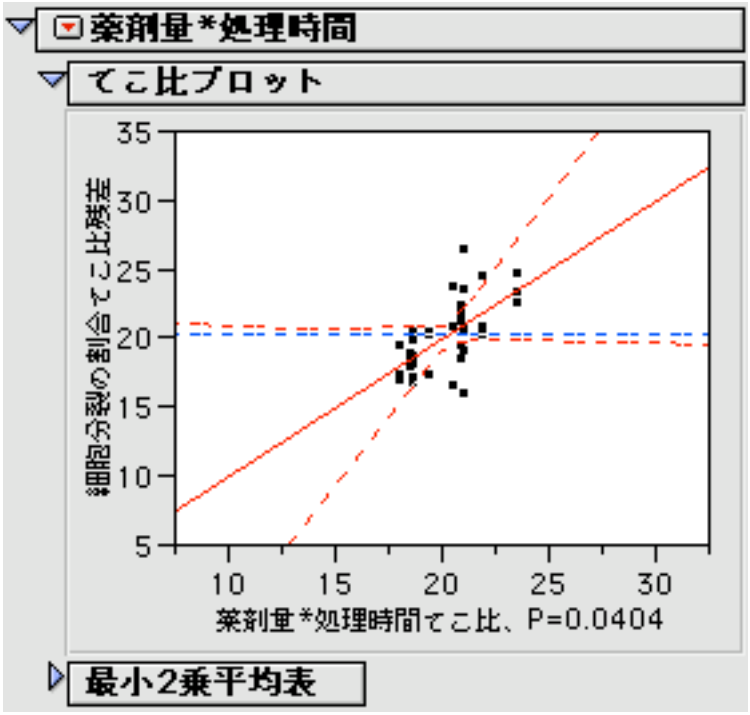
つぎに，X変数として，薬剤の量（amount）と処理時間（time）をコマンドキーを押しながら選択する。ついで，追加ボタンを押す。

次に，上と同様に薬剤の量（amount）と処理時間（time）をコマンドキーを押しながら選択して，こんどは 交差 ボタンをクリックする。

上の操作で，主効果と交互作用の項がつくられて，最終的に全部の変数が割り当てられる。

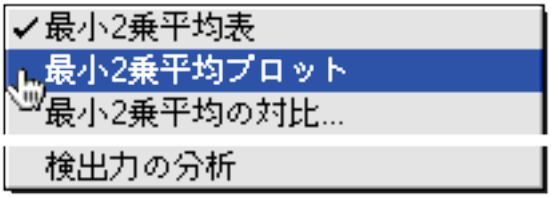
モデルの実行ボタンをクリックして，二元分散分析を実行する。

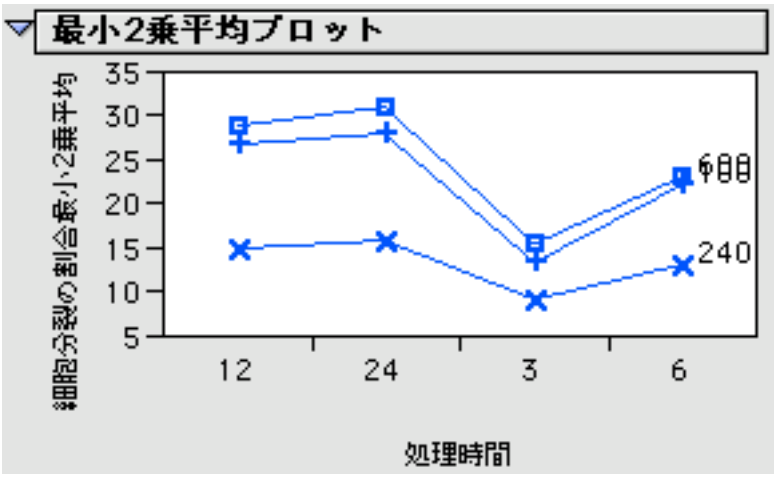
まず，モデル全体，薬剤量，処理時間，薬剤量＊処理時間のグラフをみてみよう。このグラフをてこ比プロットとＪＭＰでは呼んでいる。それぞれのグラフは回帰直線とその95%信頼限界が描かれている。グラフの水平線は全体の平均値である。この水平線と回帰直線・95%信頼限界曲線が交叉をしているとき，それぞれの効果は5%水準で有意である。



この交互作用は辛うじて95%信頼限界曲線が水平線と交わっているので，5%水準で有意であることが分かる。グラフ下を見ると，その確率は4.04%であることが示されている。

薬剤量＊処理時間タイトルの横の▼ボタンを押して，最小2乗平均プロットを選択してみよう。





薬剤600μgで24時間処理したとき，細胞分裂の割合が一番大きく；薬剤2400μgで3時間処理したとき，細胞分裂の割合が一番小さいことなどが分かる。これは，単純には薬剤の量，処理時間の効果は判断できないことを示している。

これらのグラフで分かる事柄を表にまとめたものが，画面の左端の **効果の検定** の欄に表示される：

効果の検定

要因	パラメータ数	自由度	平方和	F値	p値(Prob>F)
薬剤量	2	2	889.52056	78.8543	<.0001
処理時間	3	3	798.20750	47.1731	<.0001
薬剤量*処理時間	6	6	89.88833	2.6561	0.0404

Analysis of Variance に，誤差分散，全分散に関する情報がのっている。

分散分析

要因	自由度	平方和	平均平方	F値
モデル	11	1777.6164	161.601	28.6513
誤差	24	135.3667	5.640	p値(Prob>F)
全体(修正済み)	35	1912.9831		<.0001

このＪＭＰでの解析結果から，薬剤の量と処理時間の交互作用は有意であることが分かる（確率は4.04%で，これは5%よりも小さい）。であるから，ここでは，薬剤の量，処理時間の主効果は単独では判断できない（そのようなＦ値の確率はほとんど0%と小さいが）。

《問題》完全無作為化法によって，蛋白源，蛋白のレベルの２要因を組み合せた６種の飼料を与えた雄のラットの体重増加量を観察した。蛋白源の要因は牛肉，穀類，豚肉の３通りであり，蛋白のレベルは高水準，低水準の２通りである。

蛋白の水準	高蛋白			低蛋白		
蛋白源	牛肉	穀類	豚肉	牛肉	穀類	豚肉
繰返し	73	98	94	90	107	49
	102	74	79	76	95	82
	118	56	96	90	97	73
	104	111	98	64	80	86
	81	95	102	86	98	81
	107	88	102	51	74	97
	100	82	108	72	74	106
	87	77	91	90	67	70
	117	86	120	95	89	61
	111	92	105	78	58	82

このデータから，蛋白源，蛋白の水準およびその交互作用がラットの体重増加に影響を与えているかを検定せよ。

12.0 回帰

12.1 いとぐち

これからは二つの変数X，Yを同時に扱い，一つの変数Yの他の変数Xに対する依存性について学んでいく。数学ではYはXの「関数 function」であると表現するが，統計学では「回帰 regression」と表現する。このX，Yの関係を明かにすると，Xを与えたときのYを予測することができる。二元分散分析のときはXは質的変数として扱ったが，回帰ではXもYと同様に量的変数を用いる。ここではXとYの関係は線型（一次）であるときの直線回帰（単回帰 simple / linear regression）をみていこう。ここでは２変数，すなわちXが１個のケースを扱うが，Xが２個以上の回帰を重回帰（multiple regression）という。

関数　$Y = f(X)$ が一次式で表されるとき，その線型回帰は

$$Y = a + b X$$

と書ける。統計学ではこの関数を回帰式とよぶ。Yを従属変数（予測変数），Xを独立変数（説明変数）とよぼう。aはXが０のときのYの値であり，切片という。これは生物学ではあまり重要性をもたいことが多い（Xが０のケースはどれほどあるだろうか，Xが０のときYは存在し得るだろうか？）。bは数学的には微分係数（$= dY / dX$）であり直線の勾配だ。統計学では回帰係数という。これはXが１単位変化するときのYの変化量を表す。

$$y = \alpha + \beta x.$$

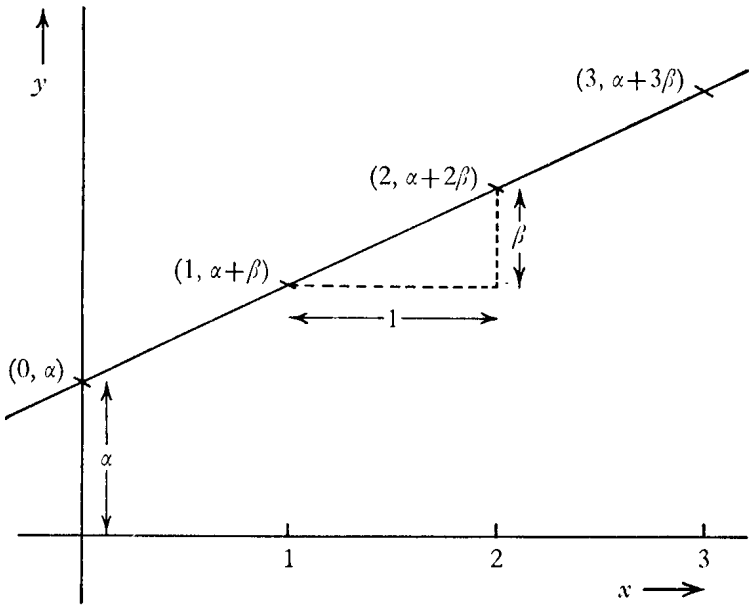


図12.1

12.2 回帰の統計モデル

これから学ぶ回帰については，次の仮定が置かれている。

(1) 独立変数**Xは固定**されているから測定誤差をもたない。これは分散分析での母数模型（Ⅰ型）と同じである。

(2) 選ばれたひとつのXに対するYの母集団は平均値μをもち，それは直線

$\mu = \alpha + \beta$（$X - \overline{X}$）$= \alpha + \beta\ x$ の上にある。α，βは母数であり，後述する。

(3) 選ばれた各々のX（X_i）についてYの正規分布がある（図12.2）。ここからYの標本が無作為に，独立に，必要ならば２つ以上のYが抽出される。

(4) 各母集団では，その平均値$\alpha + \beta\ x$のまわりのYの標準偏差は同じ値をもつ。それを$\sigma_{y \cdot x}$と書こう。これはXとは独立である。

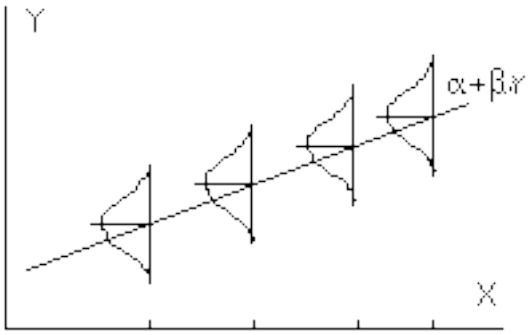


図12.2

すなわち，数学的モデルは

$Y = \alpha + \beta X + \varepsilon$ で表される。ここで，$\varepsilon \in N(0,\ \sigma_{y \cdot x})$。このモデルは，Yは確率的部分$\varepsilon$とXによる固定部分との和であることを表している。ここまで書くと，回帰の検定には分散分析が応用できそうだ！？

さて，ここで身長（X）に対する座高（Y）の回帰という問題を考えてみよう。このとき，身長を固定すること，身長を偶然誤差がなく与えることは可能だろうか？通常の調査ではまず不可能だろう。ということは，上の仮定を満たすことはできない。分散分析での変量模型（II型）の回帰分析を考える必要がある。しかし，このモデルの分析は統計学者の間でも統一見解がない。幸にも，上の(2)，(3)，(4)の仮定が満たさ

れるなら，これから述べる母数模型（Ｉ型）の回帰分析の手法がそのまま使える（Snedecor & Cochran, 1967）。

上のケースとは反対に，座高（X）に対する身長（Y）の回帰を考えても悪いはずはないだろう。このとき上のケースで計算された身長（X）に対する座高（Y）の回帰係数ｂの逆数（１／ｂ）を座高（X）に対する身長（Y）の回帰係数としてよいだろうか？後に回帰係数の求め方を示すが，ここでは回帰モデルの図から考えてみよう。（Ｉ型の）回帰分析では，こんどは座高（X）が与えられたときの身長（Y）の分布を考えなければならないはずである。だから，代数的に逆数をとった値と等しくなる保証はない。

12.3 回帰式の求め方

12.3.1 基本的な考え

年齢と血圧（収縮期血圧，最高血圧）の間にどの様な関係があるかを次の例で回帰分析をしてみよう。

表12.3.1

年齢 X	血圧 Y	平均からの偏差		偏差平方		積 xy
		x	y	x^2	y^2	
35	114	−20	−27	400	729	540
45	124	−10	−17	100	289	170
55	143	0	2	0	4	0
65	158	10	17	100	289	170
75	166	20	25	400	625	500
計 275	705	0	0	1000	1936	1380
平均 55	141					

回帰分析ではまず，年齢と血圧のグラフを描くことから始める。ここでの回帰分析は直線回帰（一次回帰）であるから，散布図の傾向線が曲線を予想させるなら，曲線回帰（non-linear regression）を考えるか，変数変換を考えなければならない。

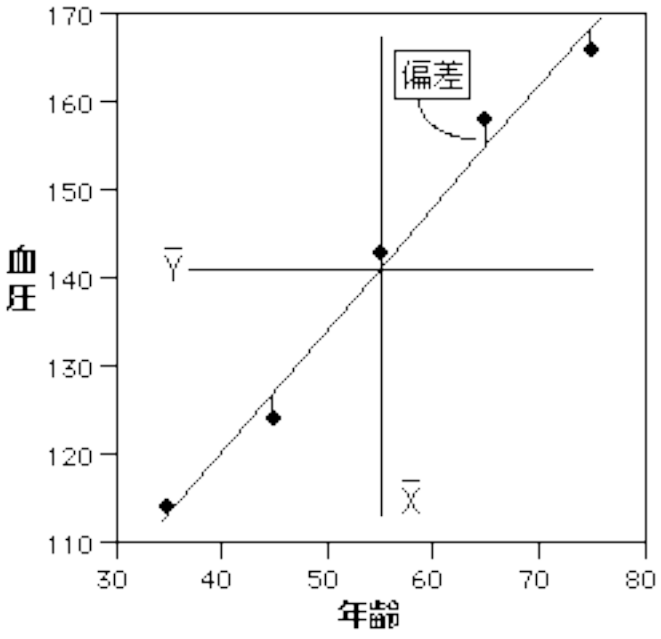


図12.3.1

この図12.3.1から血圧と年齢の間には，年齢が高くなるに連れて血圧は上がるという，正の関係があるように見える。観測点が一線に乗っていれば回帰直線は，数学的に，簡単に求めることができる。実際の観察ではこの図のように観測値は直線の両側にバラつく。順序が逆のようだが，では，この直線（回帰直線）はどの様にして求めるのだろうか？ここでは，最小自乗法（least sqaure method）をつかう。

では，何を最小にする直線をもとめるのだろうか。図12.3.1の斜めの直線（回帰直線）を，

$$\hat{Y} = a + b\,X$$

で表そう。ここで$\hat{Y}$（Yハット）は与えられたXに対して，まさに直線に乗っている（たぶん実際の観測値とは異なった）推定値である。だから，図12.3.1で見るように観測値Y_iと推定値Y_iの間にはズレ（回帰からの偏差）があるだろう。このズレは図12.3ではY軸に平行に描かれていることに注意しよう。この回帰からの偏差を，

$$d_{Y\cdot X} = (Y_i - \hat{Y}_i)$$

と書こう。さて，この回帰からの偏差の和は0となる（後で計算してみよう）。そこでこの二乗をとり，その和を考えよう。これは残差平方和とよばれ，

$$\Sigma\, d^2_{Y\cdot X} = \Sigma\,(Y_i - \hat{Y}_i)^2$$

と表す。最小自乗法とはこの残差平方和を最小にするような直線（回帰直線）を求める方法である。

残差平方和は観測値からX軸に下ろした垂線が回帰直線と交わる距離に基づいている（図12.3.1）。観測値から回帰直線に下ろした垂線の長さに基づくものではないことに注意しよう。このような偏差の二乗和を最小にするような方法で求めた直線は主成分分析（principal component analysis）の長軸に相当する。

12.3.2 回帰係数，切片の計算

最小自乗法による回帰式をもとめるには，微積分学が必要だ。ここではその過程は省略して，結果だけを示そう。回帰係数bは，

$$b_{Y\cdot X} = \Sigma\, x\, y \,/\, \Sigma\, x^2 = \Sigma\,(X-\overline{X})\;(Y-\overline{Y}) \,/\, \Sigma\,(X-\overline{X})^2$$

と計算できる。$\Sigma\, x\, y$は積和とよばれる。表12.3.1の値を代入すると，

$$b_{Y\cdot X} = 1380 / 1000 = 1.38$$

となる。切片aは

$$a = \overline{Y} - b\,\overline{X}$$

で与えられる。表12.3.1の値を代入すると，

$$a = 141 - 1.38 \times 55 = 65.1$$

となる。すなわち，求める回帰式は

$$Y = 65.1 + 1.38X$$

である。

回帰式，$\hat{Y} = a + b\,X$に切片aの式を代入してみると，

$$\begin{aligned}\hat{Y} &= a + b\,X \\ &= \overline{Y} - b\,\overline{X} + b\,X \\ &= \overline{Y} + b\,(X - \overline{X})\end{aligned}$$

と書けるから，$X = \overline{X}$を代入すると，$\hat{Y} = \overline{Y}$となる。これは，回帰式がかならず点（$\overline{X}$，$\overline{Y}$）を通ることを表している。

12.3.3 回帰式の検定

表12.3.1のX，Yにこの回帰式をあてはめ，推定血圧（Y），偏差（$d_{Y\cdot X}$），

偏差平方（$d^2_{Y\cdot X}$）を求めてみよう。

表12.3.3.1

年齢 X	血圧 Y	推定血圧 $\hat{Y}$	偏差 $d_{Y\cdot X}=Y-\hat{Y}$	偏差平方 $d^2_{Y\cdot X}$
35	114	113.4	0.6	0.36
45	124	127.2	-3.2	10.24
55	143	141.0	2.0	4.00
65	158	154.8	3.2	10.24
75	166	168.6	-2.6	6.76
計			$\Sigma d_{Y\cdot X}=$ 0.0	$\Sigma d^2_{Y\cdot X}=$ 31.60

ところで，個々の偏差はそれほど必要のないときが多いので，残差平方和は，

$$\Sigma d^2_{Y\cdot X}=\Sigma y^2-(\Sigma xy)^2/\Sigma x^2$$

の式で求めることができる。表12.3.1の値を代入すると，

$$\Sigma d^2_{Y\cdot X}=1936-1380^2/1000=31.60$$

となり，表13.2.3の偏差平方の和と一致した。

ここで，Yの偏差平方和から残差平方和を引いたものは回帰（による）平方和とよばれ，$\Sigma\hat{y}^2$と表す。これはXの違いによって生じた変動である。

$$\Sigma\hat{y}^2=\Sigma y^2-[\Sigma y^2-(\Sigma xy)^2/\Sigma x^2]$$
$$=(\Sigma xy)^2/\Sigma x^2$$

表12.3.1の結果から，$\Sigma\hat{y}^2=1380^2/1000=1904400/1000=1904.4$，となる。

さて，上に述べたこれらを別なかたちで書けば，

$$\Sigma y^2=\Sigma\hat{y}^2+\Sigma d^2_{Y\cdot X}$$

となる（図12.3.3）。

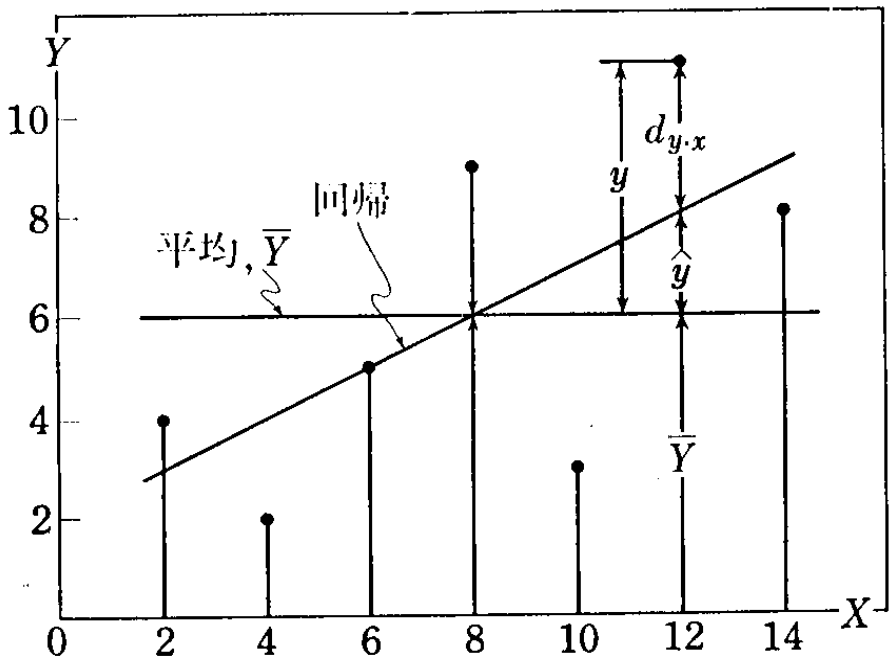


図12.3.3

バラツキがこのように，全変動＝回帰変動＋残差変動，と分解できたので，分散分析によって回帰式の検定ができそうだ。分散を計算するには自由度を求める必要がある。Ｙの変動に対する自由度は一変量のときと同じで，標本数－１だ。これは平均値を１個使って求めたからだ説明づけたことを思い出そう。では，残差平方和のほうはどうか。これを求めるために，αとβの２つのパラメータが使われたと考える。すると，標本数－２が残差平方和の自由度となる。回帰平方和に対する自由度は１である。したがって，

回帰分散 $s_{\hat{Y}}^2 = \sum \hat{y}^2 / 1$

残差分散 $s^2_{Y\cdot X} = \Sigma d^2_{Y\cdot X} / (n-2)$

検定すべきことは，帰無仮説H0：$\beta = 0$である。すなわち，標本統計量ｂは回帰係数の母統計量$\beta = 0$である母集団からのものかどうかを検定する。Ｆの値は残差に対する回帰の分散比，$Fs = S^2_{\hat{Y}} / s^2_{Y\cdot X}$として計算する。

表12.3.3.2

変動要因	自由度	平方和	分散	F
回帰	1	1904.4	1904.4	180.8
残差	3	31.6	10.53	
全体	4	1936		

$\nu 1=1$，$\nu 2=3$に対するFをF分布表で求めると，10.13である。したがって，帰無仮説を棄却する。

ここで，全変動Σy^2と回帰変動$\Sigma \hat{y}2$の比を決定係数とよびR^2で表す。

$$R^2=\Sigma y^2/\Sigma \hat{y}2$$

この値は1.0と0.0の間の値をとり，1.0に近いほど回帰直線の当てははまりが良いことを示している。

12.4 ＪＭＰによる回帰分析

回帰分析には**分析**メニューの**二変量の関係**を選択する。

(1) データ表の準備

ファイル⇒新規メニューで新規データシートを用意する。列1を選ぶ，**列**メニュー⇒**列情報...**，**列**メニュー⇒**列の新規作成...**で年齢（Age），血圧（B.P.：Blood Pressure）のカラムをつくる。

ここで，両データのタイプは数値型（Numeric）で，尺度（Measurement Level）の観点からはは間隔尺度（Interval）である。すなわち，初期状態のままでよい。

行メニュー⇒**行の追加**で5行のデータを準備する

データを入力する

図12.4.1

(2) 散布図を描く

分析メニューから**Fit Y byX**を選ぶ。年齢にX（独立変数），血圧にY（従属変数）の変数の役割を割り当てる。データタイプは数値，尺度は間隔尺度のままで良い。

列の選択欄の下の分析方法を示すグラフのタイトルが**二変量**になっていることに注目しよう。X，Y変数の尺度がContinuous（間隔尺度）同士の組み合わせでは，二変量の関係分析を選ぶと自動的に２変量分析（回帰と相関）が選択される。

OKボタンをクリックすると，散布図が出る。

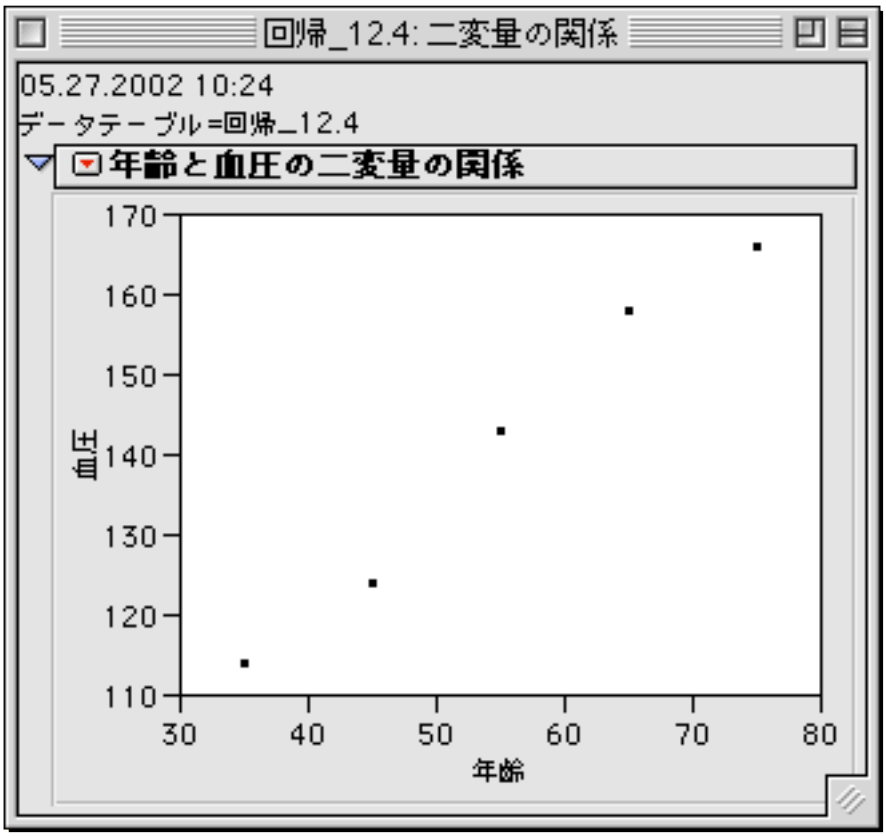


図12.4.2

散布図では，観測値は曲線上ではなく，ほぼ直線上にならんでいることことがわかる。したがって，このまま直線回帰で分析できそうだ。

(3) 分散分析による回帰式の検定

年齢と血圧の二変量の関係タイトルバー左端の▼マークをクリックして，**直線のあてはめ**を選ぶ。

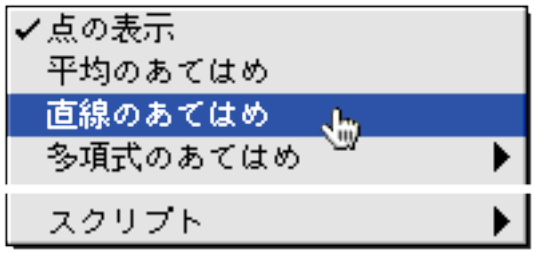


図12.4.3.1

分散分析表が表示される。 あてはめの要約欄の第１行にあるR2乗（決定係数）も0.984と１.0に近い値を示している。従って，分散分析の結果は有意に出ることが予想できる。

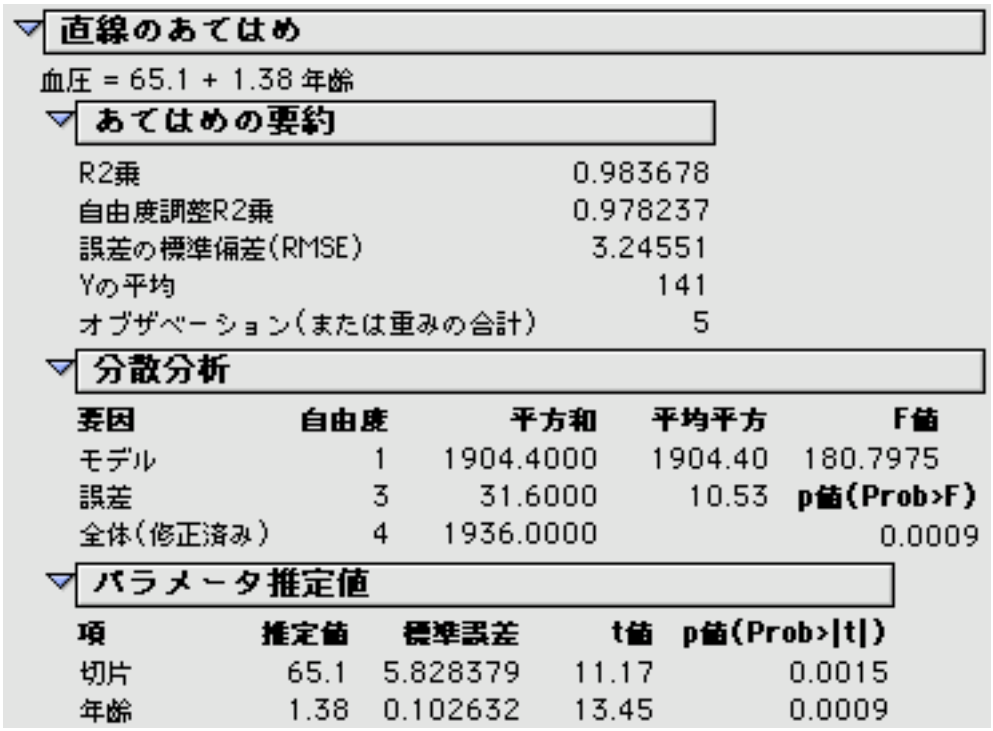


図12.4.3.2

パラメータ推定値の欄には回帰式，Ｙ＝ａ＋ｂＸのａ切片の行に，回帰係数は年齢の行に値が表示されている。この欄の右のカラムのp値（Prob>¦t¦）はH_0：$\alpha = 0$，H_0：$\beta = 0$の帰無仮説の検定での確率が表示されている。

(4) 回帰直線の信頼限界

散布図の下の**直線のあてはめ**欄の▼マークをクリックして，回帰の信頼区間メニューを選ぶ。

図12.4.4.1

すると散布図には回帰直線の95％信頼限界（信頼帯，信頼域）が描かれる。これは個々のＸに対する推

定値Ｙについて，一変量のときのように信頼限界を計算して（計算式は異なるが）線で結んだものだ。

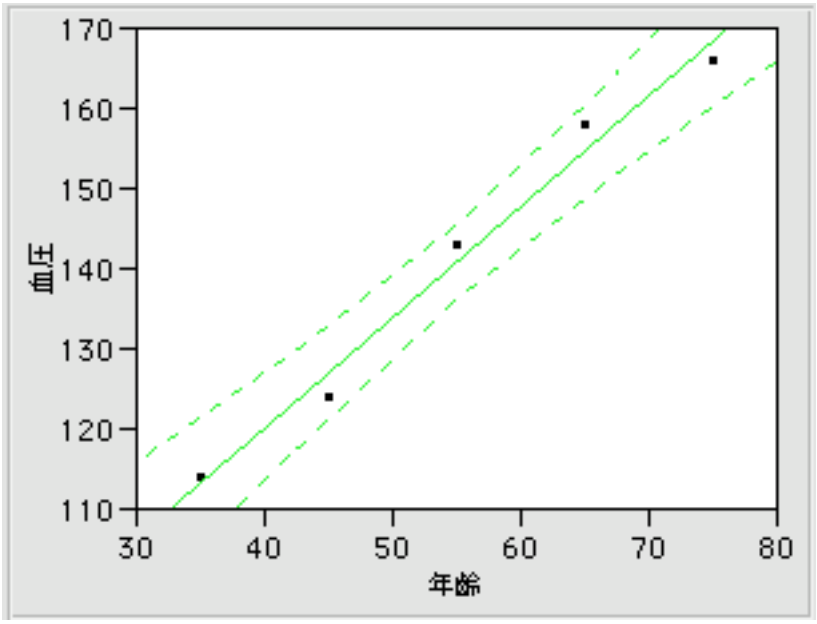


図12.4.5.2.

この図は$\mu = \alpha + \beta x$，すなわちＸが与えられたときのＹの母平均の信頼限界であることに注意しよう。

(6) 個々のＹの信頼限界

今度は，Ｘが与えられたときのＹの母平均ではなくて，推定値Ｙの信頼限界を見てみよう。これは推定によるバラツキのほかに観測値自身のバラツキも含まれるので，その信頼帯は(5)よりも大きくなる。Ｘが測定されている母集団のある新しいメンバについて，１個のＹを予測するとき適した信頼帯である。

図12.4.5.2の直線のあてはめ欄の▼マークをクリックして，**個別の値に対する信頼区間**メニューを選ぶ。これらの信頼限界のオプションスイッチはトグルになっているので，チェックマークが付いているオプションはもう一度クリックして，先のオプションをオフにしておくと図が見やすい。

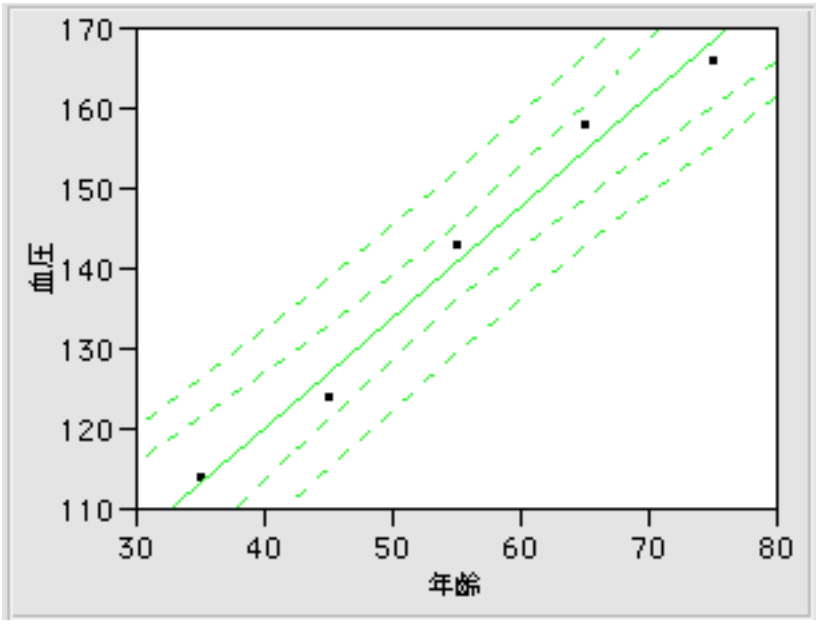


図12.4.6

回帰直線の信頼限界は，55歳の『平均』血圧を予測するときに使われるが，**個別の値に対する信頼区**

間はある新しい被験者の血圧を予測するときに使われる。両者を混同しないように！

《問題》

somatometryのデータを使って２変数を選び回帰分析を試みよ。選んだ２変数の関係を生物学的に考察できるか。回帰直線よりも他の可能性が考えられるか？

12.5 二つの回帰直線の傾きの検定

農薬の種類によって害虫の農薬量－死亡率曲線は異なるかもしれない。男女間で理想体重（身長に対する体重の回帰）は異なるかも知れない。これらの関数関係を表す回帰係数の比較を，一変量のときと同様に差の検定で行ってみよう。二つの回帰係数の差の検定方法はＦ検定として行える。

$$F_s = \frac{(b_1 - b_2)^2}{(\Sigma x_1^2 + \Sigma x_2^2)\,\overline{S}_{y\cdot x}} \Big/ \frac{}{(\Sigma x_1^2)\ (\Sigma x_2^2)}$$

ここで，$\overline{S}_{y\cdot x}$は二つのグループの重みづけ平均であり，

$$\overline{S}_{y\cdot x} = \frac{(\Sigma d^2_{y\cdot x})_1 + (\Sigma d^2_{y\cdot x})_2}{\nu 2}$$

である。$\nu 2$は$\nu 2 = n1 + n2 - 4$で与えられる。このＦｓをＦα［１，$\nu 2$］と比べる。

《問題》somatometry dataから男女それぞれの理想体重の回帰式を求めよ。この二つの回帰係数の差を５%有意水準で検定せよ。（ヒント：Σx^2は標準偏差（SD）から求めよ。

標準偏差＝$\sqrt{}$［Σx^2／（n - 1）］を利用せよ。$\Sigma d^2_{y\cdot x}$は分散分析表では誤差要因の平方和である。）

12.6 回帰分析の応用

これまでに学んだ回帰はどのように利用できるだろうか？

(1) 因果関係の研究

この問題はむずかしい。対象に関する洞察と常識が必要だろう。例えば，砂糖の消費量（Ｘ）と酒の消費量（Ｙ）の間に有意の回帰が成立したとする。では，酒屋は酒の売り上げを伸ばすために，砂糖の宣伝をして砂糖をたくさん売ろうとするだろうか？この両者の関係は偶然のものかも知れないし，両者に共通な影響を及ぼす別な原因があるかもしれない。ＹのＸへの有意な回帰によってＸの変化がＹのバラツキの原因であることは証明できない。しかし逆は真で，有意な回帰をしなければ，Ｘの変化はＹのバラツキに影響はおよぼさないといえよう。

(2) 科学的法則の記述と予測

回帰は，科学の目的の一つである「自然界における変数の間の関係を数学的に記述する」ということを可能にする。しかし，統計量ａ，ｂ（母統計量α，β）に生物学的な意味を見つけるのはむずかしい（e.g. アロメトリに対する批判）。だからといって，このような回帰直線に価値がないとはいえない。数学的に一番フィットしたものは予測という点では価値がある（e.g.体表面積の予測式）。

(3) 従属変数（Ｙ）の値の比較

例えば，最高血圧が年齢の関数であることが明らかになれば，中学生と大学教授の血圧を直接比較することは適当でない。このようなときは，独立変数Ｘの大きさを考慮して，**補正されたＹの値**を計算する必要がある。この値（Y_{adj}）は，

$$Y_{adj} = Y - b\,(X - \overline{X})$$

と表せる。

高コレステロール食と低コレステロール食をとっているさまざまな年齢構成の被験者の血圧のデータがあり，血圧は年齢で回帰できることが分かったときを考えよう。まず年齢で補正された平均値を求め，それぞれのコレステロール食の個体の血圧を比較することができる。こうして補正された平均値に対して食事の違いの影響を評価できる。

（問題）somatometry dataを使って，身長で補正された体重とナマの体重の双方を男女間で比較（度数分布，t 検定）せよ。（ヒント：身長による補正体重を求めるには，(1) 男女を込みにして，一変量の分布，二変量の関係の分析を実行し，共通の身長の平均値と回帰係数をメモしておく。(2) 列の新規作成...を実行し男女共通の補正体重を求める。計算式は下のようになるだろう。(3) 性をX，補正体重をYとして二変量の関係を実行し，t 検定を行う。）

$$体重-0.71022\bullet(身長-160)$$

12.7 残差の検査

回帰モデルの条件が成立しないとき，ＸとＹの関係（図12.7.1左）やＸと残差εの関係（図12.7.1右）は図12.7のようなグラフになるだろう。

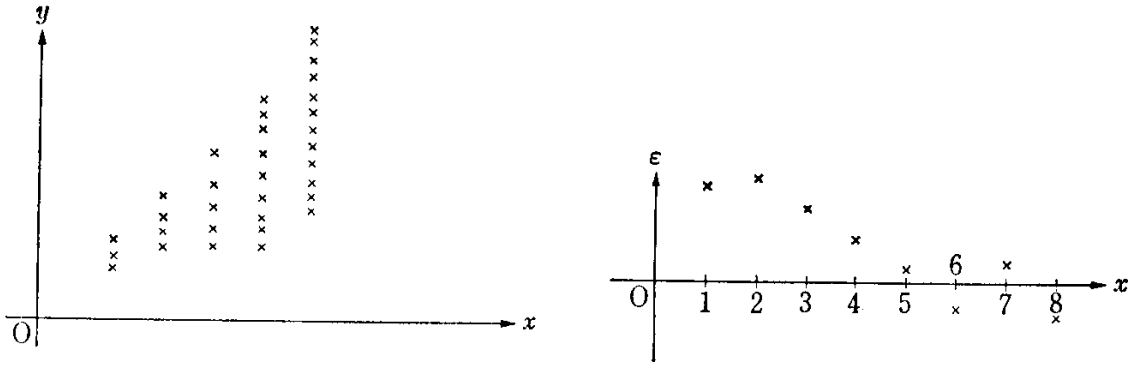


図12.7.1

ここでは，影響力係数（leverage coefficient；Hoaglin and Welsch, 1978）と規準化された残差を計算してみよう（できればDurbin-Watson testが望ましいが，現在のJMP 2.0ではまだサポートされてない）。

(1) 影響力係数

これを次の式で定義する，

$$hi = \frac{1}{n} + \frac{(Xi - \overline{X})^2}{\Sigma x^2}$$

hiは０から１まで変化し，特定のＹiの影響力の指標となる。つまり，Ｙiの推定値Ｙiへの影響力の尺度である。だから，Ｘの平均からずれたＸに対するＹは回帰直線の決定に際し，より大きな影響力を持つことになる。

(2) 基準化された残差

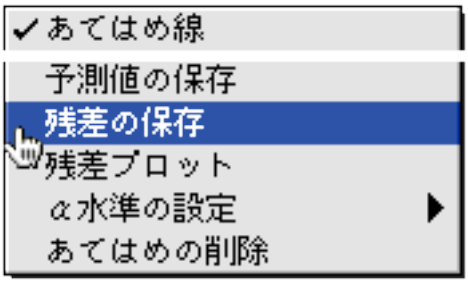
影響力係数だけでは不十分なので，基準化された残差を検討する。それは次式で与えられる，

$$\frac{dy \cdot x}{s} = \frac{dy \cdot x}{sy \cdot x \sqrt{(1 - hi)}}$$

高い値のhiでかつ基準化された残差が大きければ，回帰直線の傾きに過大の影響を及ぼしている可能性がある。高いhiと小さい残差，低いhiと大きい偏差の時は問題はないといってよい。高いhiとは，hi＞４／ｎの値である。大きい偏差は，ｔα［ｎ－２］の値と比較することで検定できる。このようにして見つかった過大の疑いのある観測値は除いて新しい回帰式を計算する。

(3) 残差の計算

さきに二変量の関係分析法で散布図を描かせ，あてはめオプションから**直線のあてはめ**を選び回帰直線を描かせた。この操作ののち，直線のあてはめ欄の▼マークをクリックして**残差の保存**オプションを選択すると，JMPデータ表に残差 Y-name（Y-nameは変数Ｙの名前，ここでは残差 血圧）という新しいカラムがつくられ，そこに（観測値－予測値）の残差が記入される。



年齢に対する血圧の回帰の例では下図のようになる（**残差プロット**のオプションを指定）。

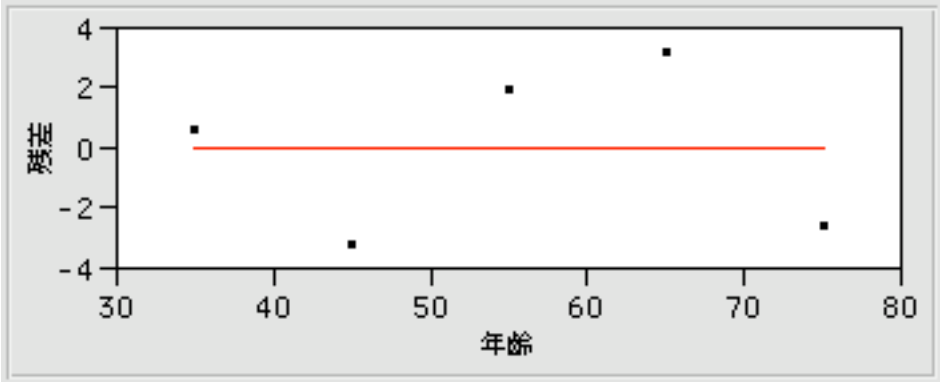
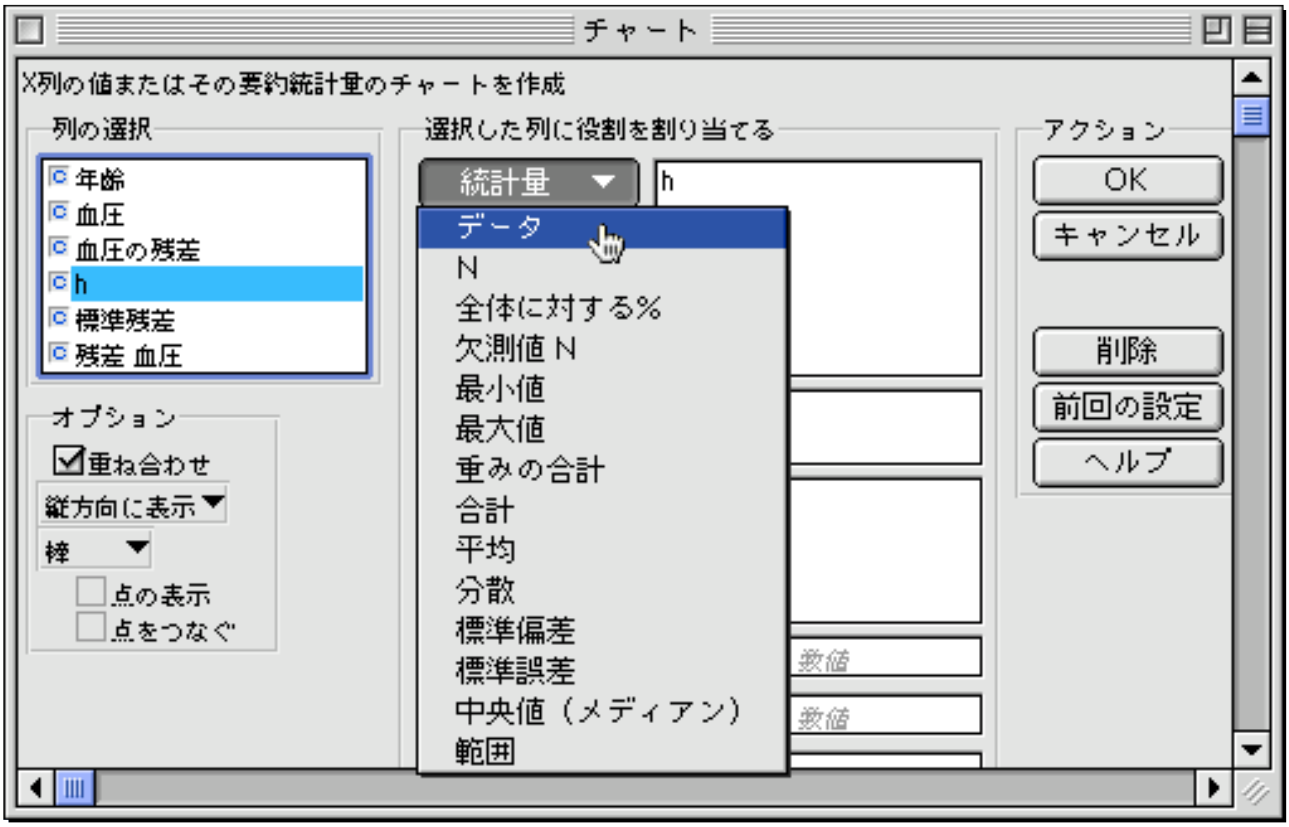
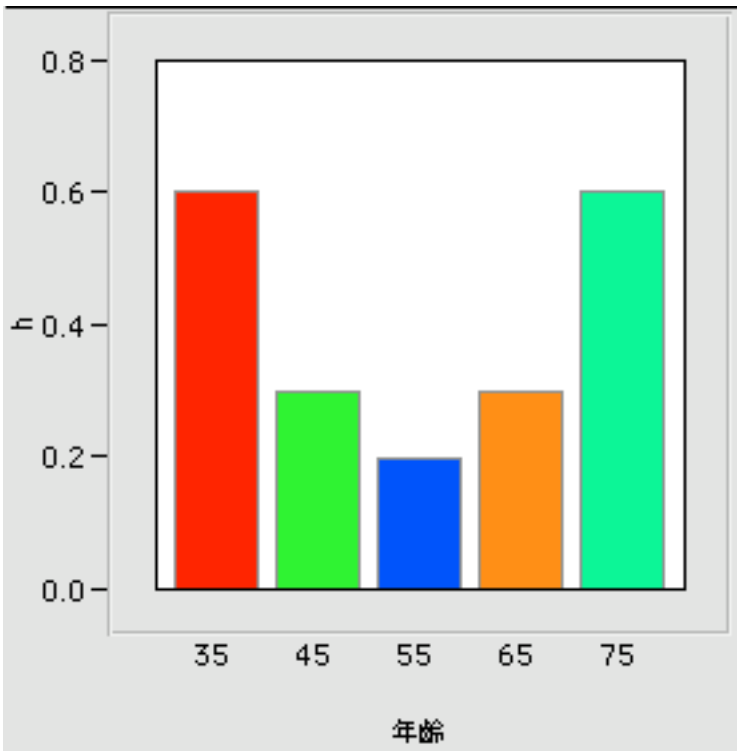


図12.7.2

影響力係数（h）をデータシートに追加してみよう。hのグラフを，**グラフ⇒チャート**メニューで描く。

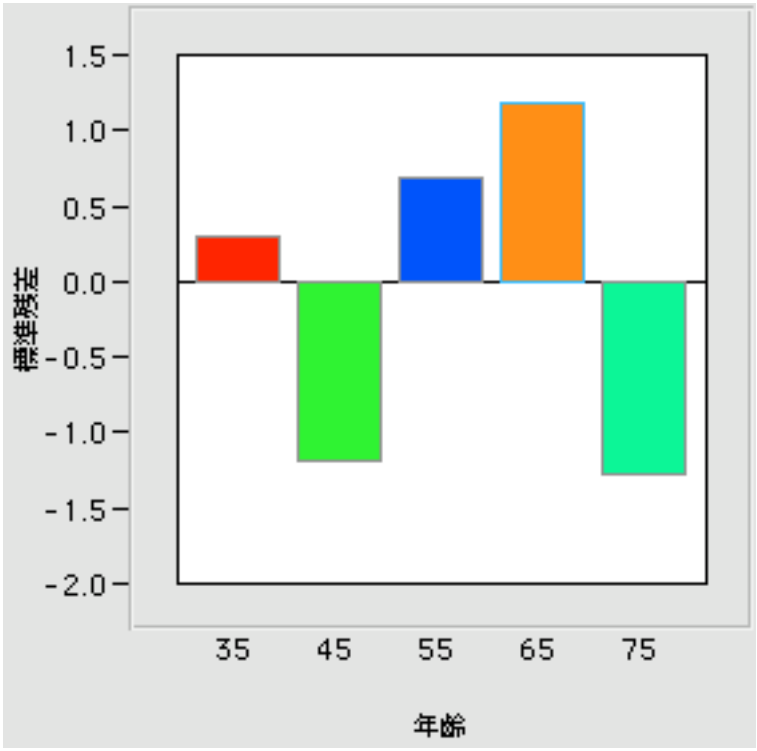


チャートウィンドウの**列の選択**欄でhを選択し，**統計量**ボタンの▼をクリックして**データ**を選ぶ。ついで，**列の選択**欄で年齢を選択し X，水準 を押す。OKボタンをクリックするとグラフが出る。



X（55）からはずれたXiに対するYiの影響力は大きくなっている。両端のX＝35、X=75に対するhはそれぞれ0.6である。この値は４／n=0.8よりも小さいから，大きなhiであるとはいえない。

基準化された残差（標準残差）をデータシートに追加してみよう。上のhと同様の手順で，標準残差のグラフを描く。



X＝45に対しては1.178，X＝65に対しても1.178，X＝75に対しては1.267であった。これらの値は，ｔ α［ｎ－２］＝ｔ0.05［3］＝3.182，に比べて小さい。したがって，外れ値はなく，年齢に対する血圧の関係は線型であるとみなしてよいだろう。

《問題》12.5の（問題）で求めた理想体重（身長に対する体重）を再検討せよ。このsomatometry dataに外れ値らしいものがあったときは，上の方法で除去し，回帰式を最計算してみよ。その新しい式は，一般に言われている「理想体重＝（身長(cm)－100）× 0.9」の式と比べてどうであるか？（ヒント：外れ値の行を行メニューの**除外する／除外しない**，**教示する／表示しない**のオプションを使って除外する）

この方法では，外れ値は相当大きく外れていないと棄却できないことが多い。元のデータを十分に吟味する必要がある（例えば，力士のデータが混じっていたとか）。

12.8 変数変換

回帰分析で一方または両方の変数を変換すると，曲線的な関係を直線的な関係に簡単化することができる場合がある。こうすると，観測値のバラツキが，回帰モデルの条件である正規性，等分散性になる傾向がある。ここでも散布図を描くことがの重要になる。

12.8.1 対数変換

(1) 従属変数Ｙの半対数（片対数）変換

ＸとＹの関係が指数関数

$$Y = a\ e^{b\,X} \quad \text{または} \quad Y = a\ b^{X} \qquad \text{(図12.8.1.1(a))}$$

と表されるとき，両辺の対数をとると，

$$\log Y = \log a + X \log b \qquad \text{(図12.8.1.1(b))}$$

となる。$y = \log Y$，$\alpha = \log a$，$\beta = \log b$ とおくと，

$$y = \alpha + \beta X$$

となる。すなわち，Ｘ－logＹ平面では線型式となる。回帰係数βの意味は，

$$\beta = \frac{dy}{dX} = \frac{d\log Y}{dX} = \frac{dY/Y}{dX}$$

であるから，**Ｘが１単位変化したときＹが何パーセント変化**するかを示す。ここでＸが年齢のような時間を表す変数ならば，このモデルの回帰係数はＹの**変化率**を表す。

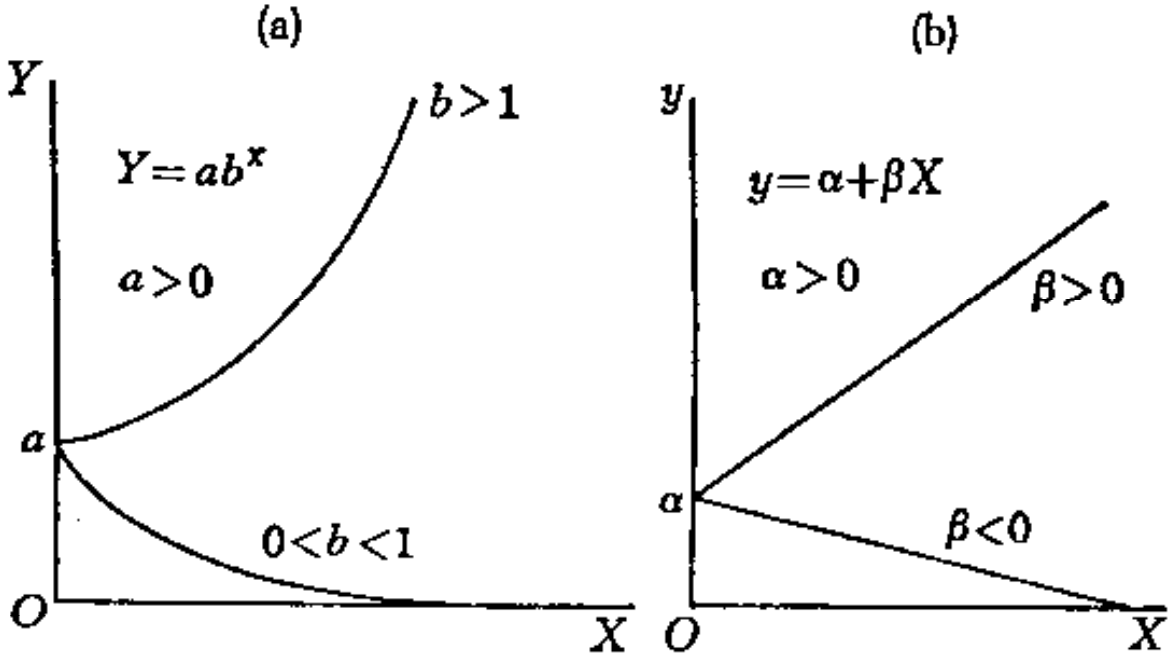


図12.8.1.1

(2) 独立変数Ｘの半対数変換

このモデルはXの水準が大きくなるに連れてYの変化は小さくなる。例えば，密度が増加するに連れて生物の体重が減少するようなケース，心理学でのWeber-Fechnerの法則（同量の反応増加を起こさせるためには刺激は同じ比率で増加されなければならない）があてはまる（図12.8.1.2）。

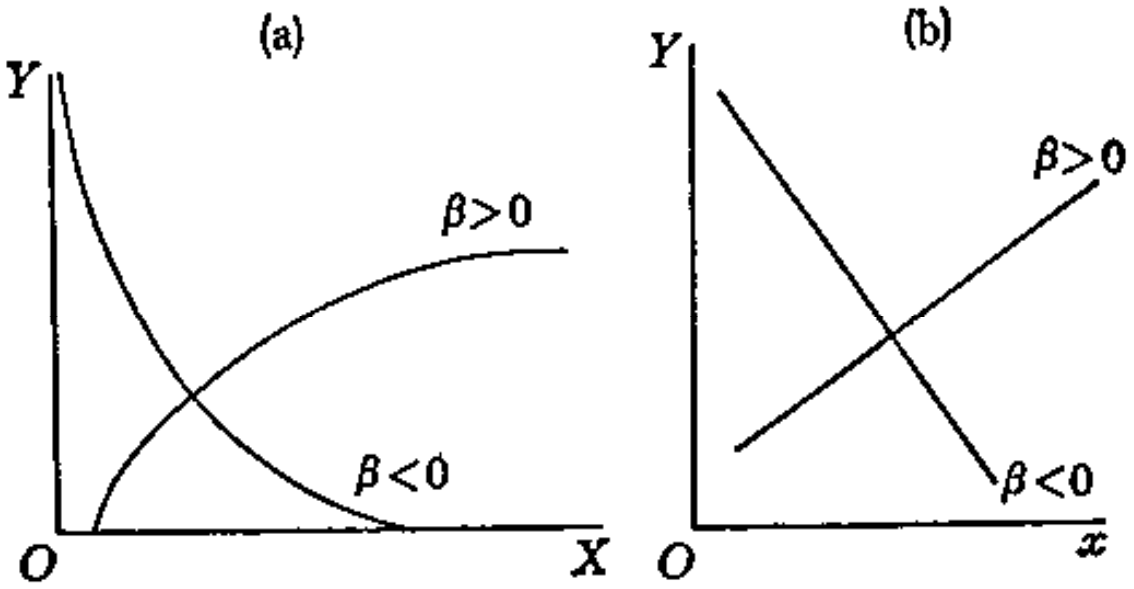


図12.8.1.2

βの意味は，ｄＹ＝β（ｄＸ／Ｘ）であるから，**Ｘの１パーセントの変化がＹを何単位変化**させるかを表す。

(3) 両対数変換

ＸとＹの関係が，

$$Y = a X^{b}$$

で表されるとき（図12.8.1.3(a)），両変数の対数をとると，

$$\log Y = \log a + b \log X$$

と線型化することができる（図12.8.1.3(b)）。

回帰係数βの意味は，

$$\beta = \frac{dy}{dx} = \frac{d\log Y}{d\log X} = \frac{dY/Y}{dX/X} = \frac{X}{Y}\frac{dY}{dX}$$

であるから，**Ｘの１パーセントの変化がＹの何パーセントを変化**させるかを表す。

Huxley（1932）のallometry式がこのケースである。

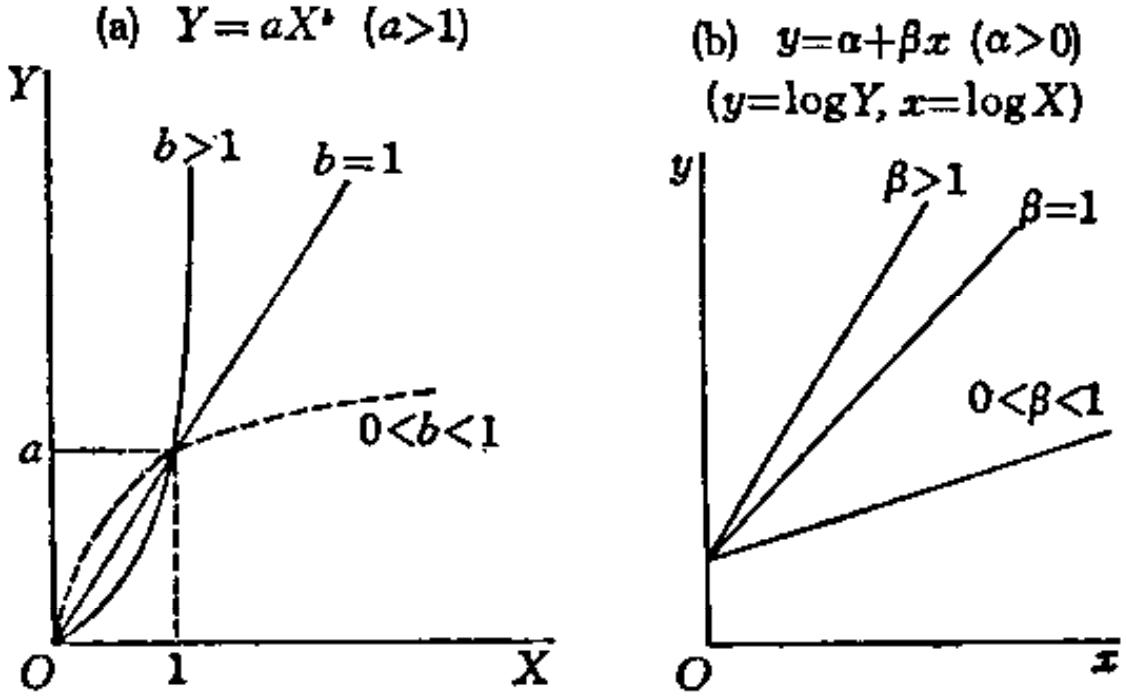


図12.8.1.3

allometryではｂ＝１のときはＸとＹのそれぞれの変化率の比が１であるから，両者のプロポーションは一定である（isometry，等成長）。このisometryを検定するには，回帰係数ｂの t 検定を使う。

$$t = \frac{b-\beta}{s_b}$$

ここで s bは，

$$s_b=\sqrt{[\,s^2_{y\cdot x}/\Sigma(X-\overline{X})^2]}$$

$\beta=0$ とすれば，Ｆ検定と同じである。isometryの検定であるから $\beta=1$ とすればよい。自由度は，$\nu=n-2$，である。

12.8.2　逆数変換

１秒当りの翅の振動数，１雌当りの産卵数などの**率**に関する現象は**双曲線**になる場合が多い。すなわち，

$$bX\cdot Y=1 \quad \text{あるいは} \quad (a+bX)Y=1$$

という関数である。この式から，

$$1/Y=bX \quad \text{あるいは} \quad 1/Y=a+bX$$

となる。したがって，従属変数Ｙの逆数をとることで回帰直線が得られる可能性がある。

13.0　相関

13.1　相関とはなにか

２つの変数の間の相互関係をあらわすもう一つの物差が相関（correlation）である。回帰と相関は数学的に非常に密接な関係にある。計算の大部分は共通だ。その基本的な量は積和［$\Sigma xy=\Sigma(X-\overline{X})(Y-\overline{Y})$］である。だから，回帰を計算するついでに簡単に相関も計算できる。

では，回帰と相関の違いはなんだろう？回帰は独立変数Ｘに対するＹの依存性（従属性）をみるのが目的であった。そして変数Ｘはバラツキを持たないことを仮定した。ところが，相関では二つの変数間に独

立，従属といった関係はない。相関は相互依存あるいは共変化の定量的な物差だ。たとえば，同じ両親から生まれた兄妹の間に，背の高い兄は背の高い妹をもつという傾向があるような，類似性の尺度といってもよい。

もう一つの違いは，回帰でのＸに相当する変数がバラツキをもつ，ということだ。変量模型（II型）の回帰モデルはその様に変数Ｘにバラツキを想定したが，２変数の間の依存性を見るという目的であった。相関はさきに述べたようにその様な依存性は目的としないので，もはや二つの変数をＸ，Ｙで表さずにＹ1，Ｙ2と表わそう。

回帰と同様に**相関も線形な関連性**（直線関係）をあつかうことに注意しよう。しかし，直線的傾向がないということだけで，必ずしも「関係がない」ことにはならない。

13.2 相関係数

ここではピアソン（KarlPearson）によって提唱された積率相関係数（Pearsoncorrelationcoefficient）について学ぼう。相関係数にはいくつかあるが，ふつう相関係数といえばこのピアソンの相関係数を指し，次の式で表す。

$$r_{12}=\frac{\Sigma y_1 y_2}{\sqrt{[\Sigma y_1^2 \Sigma y_2^2]}}=\frac{\Sigma (Y_1-\bar{Y}_1)(Y_2-\bar{Y}_2)}{\sqrt{[\Sigma (Y_1-\bar{Y}_1)^2 \ \Sigma (Y_2-\bar{Y}_2)^2]}}$$

相関係数ｒは完全な正の関連性をもつ＋１から完全な負の関連性をもつ－１までの値をとる。２変数間にまったく関連性のないときはｒ＝０である。このｒは単位あるいは次元を持たない数である。

表12.3.1の年齢と血圧の回帰のデータでは，

$$r_{12}=1380/\sqrt{(1000\times 1930)}=0.99181$$

となる。

この相関係数と先の回帰係数との関係は，

$$b_{2\cdot 1}=r_{12}[s_2/s_1]$$

$$b_{1\cdot 2}=r_{12}[s_1/s_2]$$

である。ここにｓ1，ｓ2は標準偏差（自由度ｎ－１をもつ）である。

表12.3.1の年齢と血圧の回帰のデータでは，

$$b_{2\cdot 1}=0.99181\times[\sqrt{(1936/4)}/\sqrt{(1000/4)}]=1.38$$

となり，回帰係数ｂy･xと一致する。

もし，変数が特定の分布，二変量正規分布に従うならば，相関係数ｒはこの分布の母統計量ρの推定値である。

13.3 二変量正規分布

ある母集団から標本を抽出し，各標本について２つの変数（Ｙ1，Ｙ2）を測定したときを考えよう。グラフ用紙を用意し，横軸にＹ1，縦軸にＹ2を目盛った直交座標系をつくる。これに２変数をプロットする。十分大きな標本を抽出すると同じ座標に数個の標本がプロットされるから，この座標の度数を縦軸に柱を立てるようにして更にプロットを続ける。すると，最も高い柱が中心のところにあり，そのまわりに短い柱が何本も立つようになるだろう。短い柱は上からみた円の周辺にむかってより短くなっていくだろう（図13.3.1）。

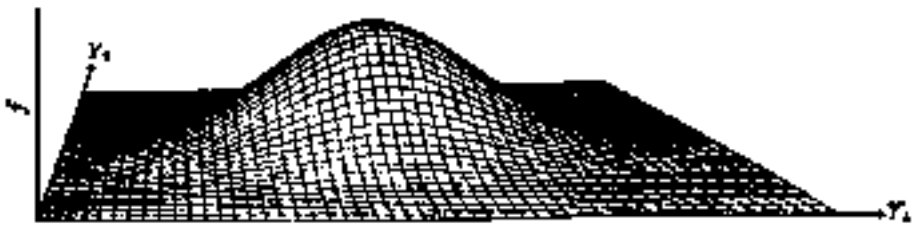


図13.3.1

図13.3.1は実は$\rho = 0$のケースである。したがって，上から見ると柱が林立してできるこの丘は長円ではなくて完全な円形である。このときＹ1，Ｙ2に関連性がない，つまりＹ1，Ｙ2は互いに独立である。

もし，Ｙ1，Ｙ2が独立でなく，ある程度の関連性（正の相関）があるならば，丘は長円形となろう。そして，中心の高い柱からは長円の主軸にそって尾根が伸びる（図13.3.2）。

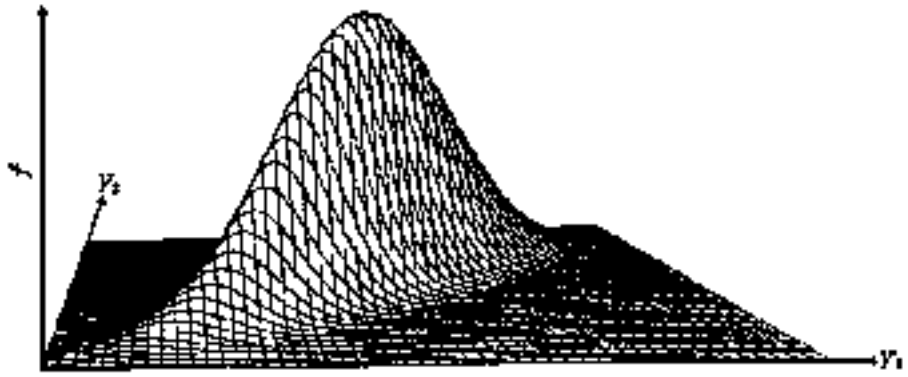


図13.3.2

図13.3.2は$\rho = 0.9$でのケースである。もし，$\rho = 1.0$ならば，このこの丘の厚みはなくなり，一枚の釣鐘型（じつは正規分布型）の屏風が立てられている状態となるだろう。このように立体散布図が示す丘のかたち（円形，長円形）の程度を母相関係数ρが表す。

この立体散布図，すなわち二変量正規分布の性質は次のとおりだ。

(1)　Ｙ1軸に垂直な各断面は正規分布をする（図12.2）。同様にＹ2軸に垂直な各断面も正規分布を示す。

(2) Ｙ1軸に垂直な度数分布はすべて同じ標準偏差$\sigma 2\cdot 1$をもち，その平均は回帰直線

$\mu 2\cdot 1 = \alpha 2 + \beta 2\cdot 1 Y 1$にのっている。

(3) Ｙ2軸に垂直な度数分布はすべて同じ標準偏差$\sigma 1\cdot 2$をもち，その平均は回帰直線

$\mu 1\cdot 2 = \alpha 1 + \beta 1\cdot 2 Y 2$にのっている。

(4) 周辺の度数分布はいずれも正規N（$\mu 1$，$\sigma 1$），N（$\mu 2$，$\sigma 2$）である。

２変数がこの二変量正規分布に従えば標本相関係数ｒからその信頼限界，検定が可能になる。図13.3.3は二変量正規分布から無作為に抽出した標本のバラツキを示す。散布図と母相関係数の値をよく観察せよ。

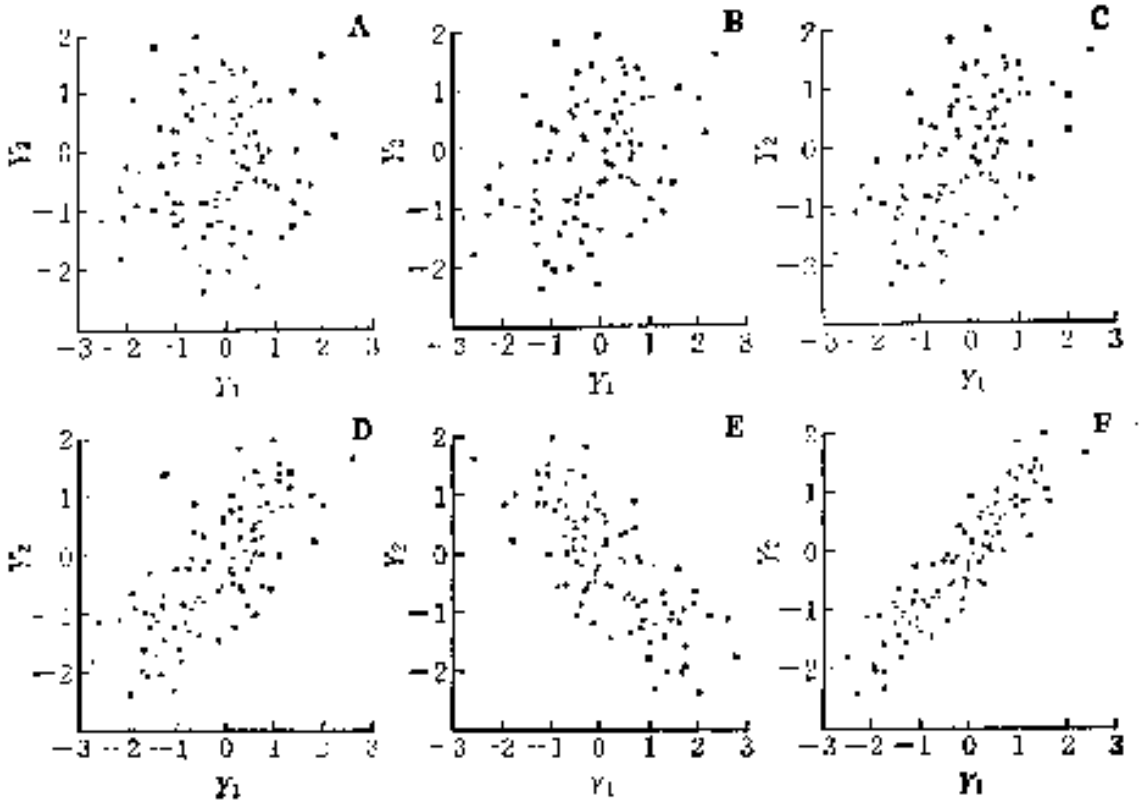


図13.3.3　いろいろなρの二変量正規分布から抽出した標本（n＝100）
A：ρ＝0.0，B：ρ＝0.3，C：ρ＝0.5，
D：ρ＝0.7，E：ρ＝-0.7，F：ρ＝0.9

13.4　相関の有意性検定

13.4.1　H_0：ρ＝0の検定

これは標本相関係数rが母相関係数ρ＝0の母集団からのものであるかどうかの検定だ。これには2つの方法がある。

(1)「相関係数の分布のパーセント点」表

この表（e.g.　　日本規格協会「簡約統計数値表」表6.1）はρ＝0のときのrの分布のパーセント点（5%，1%など）を提供する。表は自由度n－2を引数とする。rの符号は無視してよい。表12.3.1のデータでの標本相関係数r＝0.992を検定しよう。自由度3に対する有意水準（両側）5%の値は0.87834である。したがってH0：ρ＝0を棄却する。

(2) t検定

二変量正規分布をもちρ＝0の母集団からの標本の相関係数の標準誤差は，

$$s_r=\sqrt{[(1-r^2)/(n-2)]}$$

である。**5.0**推定と仮説検定でt統計量を，t＝（Y－μ）／（s／√n）と定義した。この式の分子は標準誤差であった。したがって，rとsrをつかって，

$$t_S = [r-0]/\sqrt{[(1-r^2)/(n-2)]}$$

$$= r\sqrt{[(n-2)/(1-r^2)]}$$

で検定できる。実はこの式は，Fisherが示したように，数学的には回帰係数の検定と同じである。従って，回帰係数の分散分析とも同じだ。

血圧と年齢の例では，

$$t_S=0.99181\times\sqrt{[3/(1-0.99181^2)]}=13.45$$

となる。自由度３，有意水準5%の t の値は3.182であるからH_0：$\rho = 0$ を棄却する。

13.4.2　$\rho \neq 0$ のときのH0：$\rho = \rho 1$の検定

13.4.1での数表や t 検定はH_0：$\rho = 0$ を検定するためだけに使われる。だから，H_0：ρ =0.5やH_0：$\rho 1 = \rho 2$といった帰無仮説の検定には別の方法を適用しなければならない。Fisherはこれについての解を与える z 統計量を考えた。

$$z = \ln \{(1 + r) / (1 - r)\} / 2$$

この r → z の変換表（e.g. 日本規格協会「簡約統計数値表」表10.4　z 変換）が用意されているので，それを使うこともできる。

この z の標準誤差は，

$$\sigma_z = 1 / \sqrt{\ [n - 3]}$$

である。したがって$\rho \neq 0$ のときのH0：$\rho = \rho 1$は

$$t_S = \frac{z - \zeta}{1 / \sqrt{\ [n - 3]}} = (z - \zeta) \sqrt{\ [n - 3]}$$

を使って検定できる。ここで z は r を，ζ はρ をそれぞれ z 変換したものである。しかし，**この式が（z 変換が）適用できるのは n がある程度大きいとき（n ≧50で十分，n ≧25でも十分耐えうる，n ≧10でもいいかも知れない）である。**

表12.3.1のデータでの標本相関係数 r =0.99181が母相関係数ρ =0.5からのものであるかどうかを検定してみよう。

$$r = 0.99181 \quad \text{では} \quad z = 2.74694$$
$$\rho = 0.5 \quad \text{では} \quad \zeta = 0.54931$$

したがって，

$$t_S = (2.74694 - 0.54931) \sqrt{2} = 3.10792$$

t 0.05［3］=3.182であるから5%有意水準ではH0：ρ =0.5を棄却できない。

13.4.3　二つの相関係数の差の検定

n が大きいとき，二つの相関係数についての帰無仮説H0：$\rho 1 = \rho 2$は，

$$t_S = \frac{(z1 - z2)}{\sqrt{\ [\{1 / (n1 - 3)\} + \{1 / (n2 - 3)\}\ ]}}$$

の値を **t α［∞］** と比較することで検定できる。

13.4.4　信頼限界

n が大きいとき（n ≧25，できれば n ≧50）は z 変換を使って r の信頼限界を計算できる。手順は，(1) 標本相関係数 r を z に変換し，この z の信頼限界を計算し，(2) その z の信頼限界を r に逆変換数する，というものだ。

下限 L 1は

$$L1 = z - t\ \alpha\ [\infty]\ \sigma\ z = z - [\ t\ \alpha\ [\infty] / \sqrt{\ (n - 3)\ }\ ]$$

で，上限 L 2は

$$L2 = z + t\ \alpha\ [\infty]\ \sigma\ z = z + [\ t\ \alpha\ [\infty] / \sqrt{\ (n - 3)\ }\ ]$$

で表される。

ためしに表12.3.1のデータで95%信頼限界を計算すると，r =0.99181では z =2.74694，t 0.05［∞］=1.960であるから，

$$z\text{ の } L1 = 2.74694 - [1.960 / 1.41421356] = 1.361$$
$$z\text{ の } L2 = 2.74694 + [1.960 / 1.41421356] = 4.133$$

となる。これを z 変換表を逆に引いて r に逆変換して，

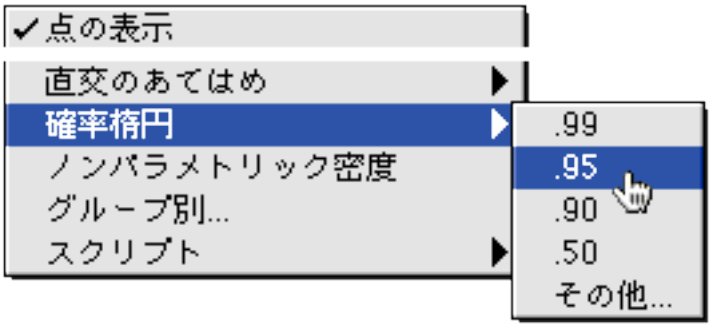
L 1≒0.876，および L 2＝1.000

が r ＝0.99181の95%信頼限界と求まった。

13.5　ＪＭＰによる相関の計算

(1)　分析メニューから**二変量の関係**を選ぶ。年齢にＸ，血圧にＹの変数の役割を割り当てる。相関分析では，回帰分析と違って，どちらの変数をＸにしてもＹにしても構わない。

(2)　　タイトルバー（年齢と血圧の二変量の関係）の左の▼マークをクリックして，**確率楕円**からさらに.95をえらぶ。散布図には95%確率楕円が描かれる。95%確率楕円とは95%の確率によって決められた標本を含む楕円で，二変量正規分布から計算される。



この例では95%確率楕円はあまりに細長くて画面からはみ出している。

(3) **相関**の欄の▼ボタンをクリックして相関係数を表示させる。

帰無仮説H_0：$\rho = 0$ の確率は0.09%であることが示されている。したがって，帰無仮説を棄却する。

《問題》somatometry dataから男女別に，身長と体重について次の計算をせよ。

(1) 相関の計算。

(2) H_0：$\rho 1 = 0$，H0：$\rho 2 = 0$ の検定。

(3) H_0：$\rho 1 = 0.5$，H0：$\rho 2 = 0.5$の検定。

(4) H_0：$\rho 1 = \rho 2$の検定。

(5) r 1，r 2の95%信頼限界。

14.0　属性（名義尺度）データの分析

これまでは主として連続変数である計測データをあつかってきた。ここではカテゴリーに対する反応の数（出現頻度）で表される属性データの分析を考える。すでに，この分野は４章でFisherの独立性の精密検定法として体験したことを思い出そう。ここでは尤度（ゆうど）比検定，χ^2検定による適合性検定と独立性検定を学ぶ。

14.1　適合性検定

14.1.1　対数尤度比検定

２つのカテゴリでの観測頻度を $f1$，$f2$；観測比率を p，q とする。期待頻度を$\hat{f}1$，$\hat{f}2$；期待比率を p，q とする。$f1 + f2 = \hat{f}1 + \hat{f}2 = n$，$p + q = \hat{p} + \hat{q} = 1$ である。二項分布での二項展開の一般項の式，

$$nCk\ p^k q^{n-k} \quad (nCk = n! / [k!\ (n-k)!])$$

を使って，母集団の統計量が観測された標本の比率と同じであるという仮定（仮説$\hat{p} = p$）のもとで，標本

の結果を得る確率を計算する。kに $f1$，n−kに $f2$を代入する。

$$\mathrm{n\,C}\ f1\ \ p^{f1}\,q^{f2}$$

ついで，帰無仮説を仮定したときに標本の結果を得る確率を計算する。

$$\mathrm{n\,C}\ f1\ \hat{p}^{f1}\hat{q}^{f2}$$

この両者の比を計算する。

$$\begin{aligned} L &= [\mathrm{n\,C}\ f1\ \ p^{f1}\,q^{f2}] \,/\, [\mathrm{n\,C}\ f1\ \hat{p}^{f1}\hat{q}^{f2}] \\ &= (\ p/\hat{p}\ )^{\ f1}\ (\ q/\hat{q}\ )^{\ f2} \end{aligned}$$

$f1=np$，$\hat{f}_1=n\hat{p}$，同様に $f2=nq$，$\hat{f}_2=n\hat{q}$であるから，

$$L = (\ f1/\hat{f}_1\ )^{\ f1}\ (\ f2/\hat{f}_2\ )^{\ f2}$$

Lの自然対数（ln）をとって，

$$\ln L = f1\ln(\ f1/\hat{f}_1\ ) + f2\ln(\ f2/\hat{f}_2\ )$$

ここで次の式は標本が十分大きいときはχ^2分布にしたがうことが知られている。

$$G=2\ln L$$

このGの値を［対数］尤度比（[log] likelihood ratio）といい，この検定法が［対数］尤度比検定である。

3つ以上（k）のカテゴリでのGの一般式は，

$$G=2\sum_{i=1}^{k} f_i\ \ln(\ f_i/\hat{f}_i\ )$$

また期待値が比率で与えられているときは，

$$G=2\ [\sum_{i=1}^{k} f_i\ \ln(\ f_i/\hat{p}_i\ )\ -n\ \ln n\ ]$$

で計算した方が簡単だ。

Gはχ^2分布に近似するが，少数例の場合はWilliamsの修正を施した方がよいだろう。$n<200$のときはこの修正をしたほうがよいという意見もある。修正項qは，

$$\mathrm{q}=1+(\ k^2-1\ )/6\ n\nu\quad(\text{このqは別のq})$$

で，kは水準数，nは標本数，νは$k-1$である。修正されたGは，

$$G_{\mathrm{adj}}=G/\mathrm{q}$$

で，これを$\chi^2[\alpha,\ \nu]$と比較する。

標本数nが20より大きく，また期待数の最小が5より大きい場合はこの方法を適用してよいだろう（そうでないときはFisherの精密法を使う）。

《例》1222人の新生児を無作為抽出したとき，男児が627人，女児が595人であった。これから，男女の出生の性比が1：1であるかどうかを検定してみよう。

まず，次の表を作成する。

表14.1.1

	観測頻度 f	観測比率 f/n	期待比率 $\hat{p}$, $\hat{q}$	期待頻度 $\hat{f}$	比 $f/\hat{f}$	$f\ln(f/\hat{f})$
男児	627	0.51309	0.5	611	1.02619	16.20978
女児	595	0.48691	0.5	611	0.97381	−15.78865
計	1222	1.0	1.0	1222		$\ln L$=0.42113

ついで,

$$G = 2\ln L = 2\times 0.42113 = 0.84226$$

を計算し，このGを$\chi^2[\alpha, 1]$とくらべて検定する。$\chi^2[0.05, 1] = 3.841$であるから，男女の出生の性比が１：１であるという帰無仮説は棄却できない。

Williamsの方法で修正されるGは,

$$q = 1 + (2^2 - 1)/(6\times 1222\times 1) = 1.00040917$$

$$G_{adj} = 0.84226/1.00040917 = 0.84191551$$

やはり，帰無仮説は棄却できない。

14.1.2 Pearsonのχ^2検定

厳密にはχ^2統計量を検定するのではないが，現在までにχ^2検定としてあまりにも有名なこの検定方法について知っておこう。一般にはここで求めるχ^2の値は尤度比検定のGの値に近いものになる。ここではχ^2の記号に代えてX^2を使おう。それは次の式で求められる,

$$X^2 = \sum_{i=1}^{k} [(f_i - \hat{f}_i)^2/\hat{f}_i]$$

この式は変形した次式の方が計算には便利だろう。

$$X^2 = [\sum_{i=1}^{k} (f_i^2/\hat{f}_i)] - n$$

この値を$\chi^2[\alpha, \nu]$と比較する（$\nu = n - 1$）。

先ほどの例を使おう。今度は次のような表をつくる。

表14.1.2

	観測頻度 f	期待頻度 $\hat{f}$	両者のズレ $f-\hat{f}$	ズレの２乗 $(f-\hat{f})^2$	$\frac{(f-\hat{f})^2}{\hat{f}}$
男児	627	611	16	256	0.4190
女児	595	611	−16	256	0.4190
計	1222	1222	0		X^2=0.838

$\chi^2[0.05, 1] = 3.841$であるから，男女の出生の性比が１：１であるという帰無仮説は棄却できない。

《**問題**》somatometry dataは骨成熟が完了した被験者のみからなっている標本である。今この標本で性比

を調べればそれは，男子：女子＝1：3に近い数字になることが分かるだろう。対数尤度比検定，ピアソンのχ^2検定によりこのことを確かめよ。

14.2 独立性の検定

2つの属性がそれぞれ2つの状態で起こりうるとき，その2つの属性はたがいに依存しているかどうかを検定してみよう。これは，2つの百分率の差の有意性検定をすることとも考えることができる。

《例》ある地区の保健所が，住民900人についてインフルエンザの罹患率を調査した。240人がインフルエンザにかかったことが分かり，さらにそのうちの60人はインフルエンザの予防注射をしていたことも分かった。この結果から，「予防注射をうけることと，インフルエンザにかかることは関係がない，すなわち，両者はたがいに依存せず独立である」という仮説を検定してみよう。全ての組み合せは次のような分割表（二元表）のかたちで表される結果であった。

表14.2.1　観測値

予防注射＼インフルエンザ	かかった	かからなかった	計
受けた	60	240	300
受けない	180	420	600
計	240	660	900

仮説が正しいとすると，予防注射を受けてインフルエンザにかからない人は，

$$900人\times\frac{300人}{900人}\times\frac{660人}{900人}=220人$$

であると期待される。同様に注射をしてインフルエンザにかかる人の期待値は，

$$900人\times\frac{300人}{900人}\times\frac{240人}{900人}=80人$$

すなわち，「分割表の各セルの期待頻度は，その交点を通る行と列の周辺頻度の積を標本数で割ってやる」ことで求めることができる。このようにして4つの期待値を求めて表にすると，

表14.2.2　期待値

予防注射＼インフルエンザ	かかった	かからなかった	計
受けた	80	220	300
受けない	160	440	600
計	240	660	900

対数尤度比の式

$$G=2\sum_{i=1}^{k} f_i \ln\left(f_i/\hat{f}_i\right)$$

に代入して，G＝10.570が求まる。

Gはχ^2［0.05，1］＝3.841の値よりも大きいので帰無仮説をすてることができる。すなわち，インフルエンザに対する予防注射の効果はある，といえる。

分割表をm行n列に拡張したm×n独立性検定の（簡便な）一般式は，

$$G = 2\times[(\Sigma\, f_c \ln f_c) - (\Sigma\, f_m \ln f_m) + n\ln n]$$

（f_cは各セルの頻度；f_mは各行・列の周辺頻度）

であり，自由度は，

$$\nu = (m-1)(n-1)$$

である。標本数が少ないときはWilliamsの修正をくわえるとよい。

独立性検定は連関性検定とも呼ばれる。独立と連関は表裏の関係だからだ。

《問題》ピアソンのχ^2検定によって上の例題を分析せよ。

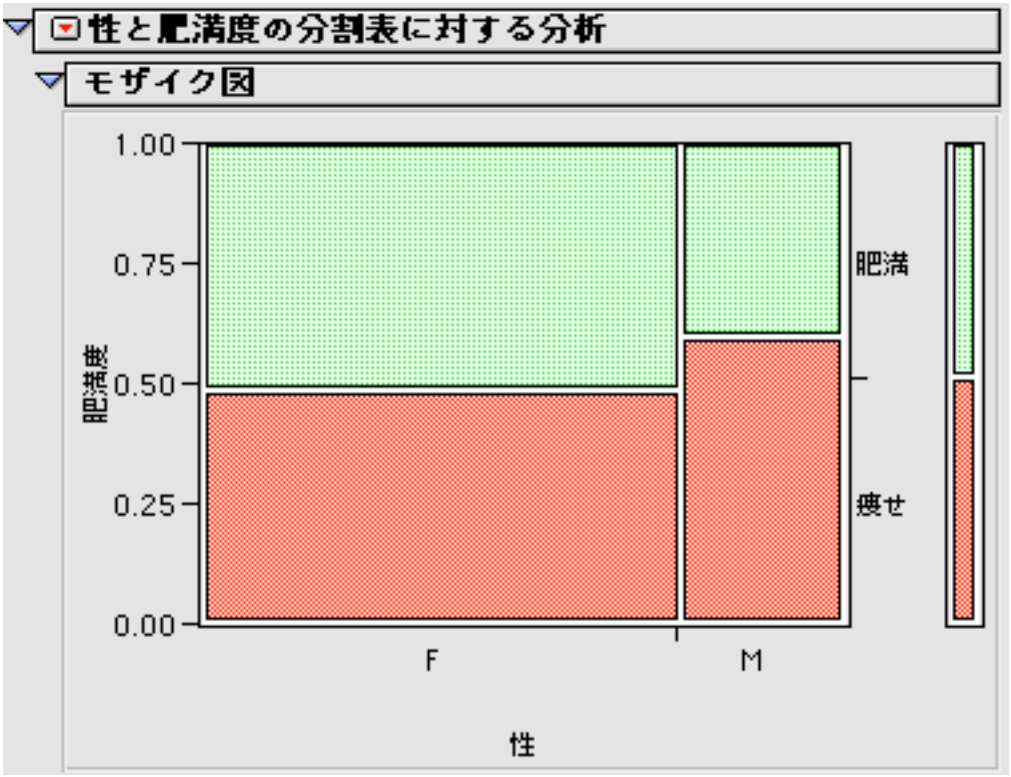
14.3　ＪＭＰによる独立性検定

somatometry　　dataを使って，性と肥満の連関性（独立性）を分析してみよう。ただし，肥満度（obesity）は理想体重（［身長cm－100］×0.9）を採用する。この値より体重が多ければ肥満（obese），少なければ（lean）とする。

(1) **列**メニュから**列の新規作成...**を選び，変数obesityをつくる。データタイプはもちろん文字型だ。

(2) **分析**メニュから**二変量の関係**分析法を選択する。性をXに，obesityをYに割り当てる。

列の選択欄の下の分析方法を示すグラフのタイトルが**分割表**になっていることに注目しよう。X，Y変数の尺度が名義尺度（Nominal）同士の組み合わせでは，二変量の関係分析を選ぶと自動的に分割表の分析方法が選択される。分割表がモザイク状にプロットされた図が表示される。



横軸に目盛られたF，M（棒の太さ）はそれぞれの標本数に比例している。縦軸の目盛りはX変数のそれぞれの水準の割合を表している。

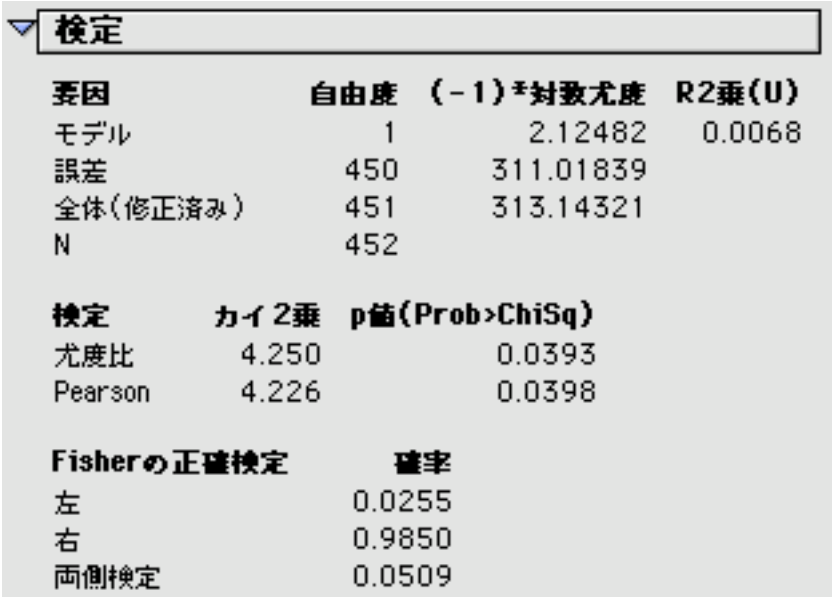
モザイクプロットの下には分割表が示されている。

分割表

		肥満度		
	度数 全体% 列% 行%	痩せ	肥満	
性	F	164 36.28 70.69 48.52	174 38.50 79.09 51.48	338 74.78
	M	68 15.04 29.31 59.65	46 10.18 20.91 40.35	114 25.22
		232 51.33	220 48.67	452

モザイクプロット，分割表から分かるように，女子と肥満，男子と痩せが連関しているようだ。すなわち，性と肥満度は独立ではないようだ。

分割表の下の**検定**欄を見ると，



検定

要因	自由度	(-1)*対数尤度	R2乗(U)
モデル	1	2.12482	0.0068
誤差	450	311.01839	
全体(修正済み)	451	313.14321	
N	452		

検定	カイ2乗	p値(Prob>ChiSq)
尤度比	4.250	0.0393
Pearson	4.226	0.0398

Fisherの正確検定	確率
左	0.0255
右	0.9850
両側検定	0.0509

性と肥満度が独立であるという帰無仮説の元でこのような結果を得る確率は，対数尤度比検定では4.2%であり，Peasonのχ^2検定では4.3%であることが示された。両者のχ^2値が近似していることに注意せよ。したがって，帰無仮説を棄却して性と肥満度には連関があると考える。しかし，Fisherの独立性の精密検定（4.1.2章，10ページ）では5%をわずかに越えている。したがって，性と肥満度には関連がないいう結論となる。さあ，どうしよう？？！！