衛生学
公衆衛生学
実習

［編集］

松久保隆　中垣晴男　山中すみへ

［執筆］

東京歯科大学講師
杉原直樹

鶴見大学歯学部教授
鶴本明久

愛知学院大学歯学部教授
中垣晴男

鶴見大学歯学部助手
福島眞貴子

東京歯科大学教授
松久保隆

愛知学院大学歯学部講師
村上多恵子

愛知学院大学歯学部講師
森田一三

東京歯科大学客員教授
山中すみへ

［五十音順］

医歯薬出版株式会社

第7章 統計的手法の選択と統計分析の実際

最近の統計分析は，電卓などで計算することはなくなり，パソコンの統計ソフトを利用して行っている．これまで，一部の研究者のみが行うことができた統計分析が誰にでもできるようになったため，統計処理はブラックボックス化している．このことは，たとえ利用者の統計手法に対する理解が不十分でも，簡単に結果が出てしまうことを意味している．

以下に，パソコンを用いた統計処理を行う際に注意するべき点をあげる．

①データの入力

②データを俯瞰的に眺め，その分布の特徴を把握し，入力ミスおよび外れ値などをよくみる

③適切な解析方法を選択する

④解析ソフトの解析結果をよく読みとる

なお，本章では，Microsoft® 社の Excel を用いた統計手法を主体に記載し，あわせて市販の統計ソフト（SAS）を使用した処理結果についても掲載し説明している．

① データの入力

Excel を用いた統計処理を行う場合，あらかじめ分析ツールをインストールする必要がある．方法は，Excel を起動後，ツールバーの「ツール」から「アドイン」を選択し，現れたウィンドウ内にある「分析ツール」のボックスをチェックして，「OK」ボタンをクリックすると，エクセルに「分析ツール」がインストールされる．

データの入力の多くは Excel を用いることが多いので，Excel を利用した場合の注意事項をあげる．基本的には入力ミスをできるかぎり少なくする工夫と統計ソフトで解析することを考慮に入れて行う．

①調査票と同じ順序で，記録された文字を入力する．

入力ミスがないようにするためである．とくに入力時に計算しながら入力しない．また，調査票は入力時のことを考えて作成する必要がある．

②データは，半角，英数字で入力する．

③データに記号（=，/，*，−，+，$，¥など）を用いない．

統計処理ソフトは，計算時に全角文字や記号があると計算不可能になることがある．

④欠測値（データがない，あるいは回答がないなどの場合）は「.（ピリオド）」とする

表 7-1　データの種類

質的データ	0-1 データ			はい (1)，いいえ (0)
	カテゴリーが 3 以上	順序のないもの		職業，人種など
		順序のあるもの	計数データ	良好，普通，不良 －，＋，＋＋，＋＋＋など
数量データ	離散データ			出産回数，齲蝕経験歯数など
	連続データ		計量データ	身長，体重，検査値など

（宮原英夫，丹後敏郎：医学統計学ハンドブック．朝倉書店，1995 より）

⑤数量化あるいは符号化できる値は，できるかぎり数値として入力する．

たとえば「男性」は「1」，「女性」は「2」，「はい」は「1」，「いいえ」は「0」のように入力する．

⑥複数回答のような場合は，文字として「01100011」のように入力する．

これは，8 つの回答肢のうち，対象者の回答が 2，3，7，8 番目の回答肢に○をつけている場合である．そして，Excel で関数を用いて新しく列を作成する．また，このようなデータがある場合はあらかじめワークシート全体を文字に設定しておく．

⑦データの削除およびソート時の注意

データを修正する場合，たとえば「削除」を行うとデータがずれてしまうことに注意する．また，ソート時においても範囲を拡大して行わないとソート列のみがソートされてしまうことに注意する．

② データの種類

データは質的データ（qualitative data）と数量データ（quantitative data）とに分けられる（**表 7-1**）．質的データは，カテゴリーに分類されたものである．このなかでカテゴリーが 2 つのもの，たとえば質問紙調査での「いいえ」あるいは「はい」などのようにデータが 0 あるいは 1 で示されるものをとくに 0-1 データという．カテゴリーが 3 つ以上ある場合は，それに順序がある場合とない場合に分けられる．順序がない例としては職業分類，人種などであり，順序がある場合は，「良好」，「普通」，「不良」や唾液細菌検査の「－」，「＋」，「＋＋」，「＋＋＋」などである．順序のない質的データは名義尺度（nominal scale）の水準といい，順序のある質的データは順序尺度（ordinal scale）の水準という．

数量データは数で表されたデータのことである．数量データは距離をもつが意味のある 0 点をもたないものを間隔尺度（interval scale）の水準といい，距離と意味のある 0 点をもつものを比例尺度（ratio scale）という．数量データで連続的な値を示すものを連続データ（continuous data），離散的な値しかとらないものを離散データ（discrete data）という．離散データと質的データは異なる尺度であるが，解析方法は共通することが多いので両者をまとめて計数データという．

質的データを統計的解析をするために数量データに変えることを数量化（quantification）という．これに対して職業分類などのデータに数字をつけることがある．この場合の数字は単なる符号であ

	A	B	C	D	E	F	G	H	I	J	K
1	NO	性別	年齢	年齢群	最高血圧	最低血圧	総コレステロール	中性脂肪	血糖値	GOT	GPT
2	1	0	39	30	118	67	207	275	90	33	67
3	2	0	39	30	132	83	193	156	81	27	32
4	3	0	39	30	117	73	209	129	91	17	25
5	4	0	39	30	121	76	220	55	85	15	14
6	5	0	38	30	120	76	153	60	85	17	20
7	6	0	39	30	105	68	180	88	84	23	33
8	7	0	39	30	126	70	170	76	89	18	18
9	8	0	39	30	130	75	203	92	92	17	23
10	9	0	39	30	116	73	200	87	84	20	14
11	10	0	38	30	124	88	187	117	89	55	84
12	11	0	38	30	136	83	214	70	93	26	29
13	12	0	39	30	139	87	171	52	95	16	14
14	13	0	39	30	140	80	210	57	98	16	16
15	14	0	39	30	133	83	233	349	85	21	28
16	15	0	38	30	128	60	176	63	95	25	31
17	16	0	38	30	113	72	187	111	89	22	21
18	17	0	39	30	139	82	228	124	87	33	24
19	18	0	39	30	125	71	219	150	81	18	19
20	19	0	38	30	120	74	157	62	81	19	26
21	20	0	39	30	124	70	190	326	83	23	40
22	21	0	39	30	126	84	231	56	87	16	14
23	22	0	39	30	98	68	138	72	88	15	15
24	23	0	37	30	122	70	167	80	97	16	17
25	24	0	37	30	107	65	182	117	74	18	21
26	25	0	38	30	118	76	238	107	87	15	25
27	26	0	39	30	90	64	166	69	90	17	7
28	27	0	39	30	110	74	188	92	96	32	32
29	28	0	37	30	120	70	279	817	89	24	47
30	29	0	38	30	100	63	186	94	92	15	11
31	30	0	39	30	97	64	125	143	91	11	7
32	31	0	38	30	135	82	164	86	85	37	57
33	32	0	37	30	152	94	220	119	174	21	35
34	33	0	36	30	114	62	196	93	93	12	6
35	34	0	37	30	115	58	197	82	85	18	19
36	35	0	37	30	130	80	176	114	90	28	42
37	36	0	36	30	127	75	234	128	93	19	27
38	37	0	37	30	108	71	171	65	85	21	12
39	38	0	37	30	127	84	207	99	87	21	36
40	39	0	37	30	121	81	246	234	84	33	44
41	40	0	37	30	118	74	185	93	82	24	29
42	41	0	37	30	142	78	187	78	82	20	20
43	42	0	38	30	127	74	194	56	84	18	21
44	43	0	37	30	124	73	313	342	74	25	37

検査値

図 7-1　定期健康診断の結果

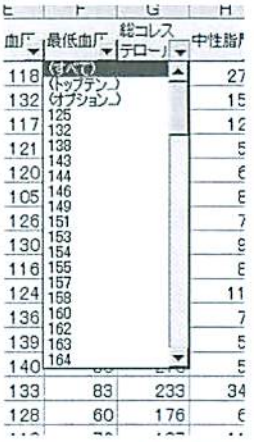


図 7-2　フィルターで最大値・最小値を確認する

るので符号化（coding）とよんでいる．

③ データの表示

A　単独変量の場合

1）データの分布

(1) 頻度分布（frequency distribution）

データが名義変数（nominal data）や順序データ（ordinal data）では，その頻度分布はそれぞれの区分に対応する数値からなる．データが離散データや連続データでは，ヒストグラムの幅やクラスの数（階級数）を決めることになるが，基本的にはデータ数を n として $\sqrt{n}+1$ 程度を目安にする．

(2) 相対頻度（relative frequency）

データの数が異なるデータセットを比較する際に有用である．

(3) 幹葉表示（stem-and-leaf display）

統計ソフトによっては観測値を使用して頻度分布を幹葉表示できるものがあり，それによってお

およその分布を知ることができる（図 7-9 参照）.

(4) 箱ひげ図（ボックスプロット）(box plot)

データの分布をボックスと上下に数を出して表示させてみるものである（図 7-9 参照）.

2）異常値（discordant value），外れ値（outlier）

頻度分布，幹葉表示や箱ひげ図に，分布からはずれるデータが存在していたら，注意する必要がある．これは，データの入力ミスや異なった母集団のサンプルを入力した可能性がある.

外れ値が 1 つや 2 つ程度の場合は，除去することもできる．しかしながら，検査値などのデータはほとんどの場合に著しい異常値があるので，統計的には外れ値であったとしても，その値を除去するかどうかの判断は慎重にすべきである.

3）分布の対称性

ヒストグラムを見る際に重要なことは，分布が対称的かどうかである．とくに医学データは，対称でないことが多く，大きいほうの異常値はいくら大きくても存在する.

統計解析は対称的な分布を仮定して行うことが多い．計量データの統計分析では正規分布を仮定する手法（t 検定など）が多い．分布が対称でない場合は，データの変換（対数変数，平方根変換，べき乗変換，指数変換など）を行うことによって正規分布型に直すことができる.

例 7-1　Excel でのデータ分布の処理 I（ヒストグラム，度数分布表）

(1) 使用するデータ

某企業の定期健康診断の結果（180 名，35 歳～ 59 歳）（図 7-1）．このデータを用いて　総コレステロール値のヒストグラムを作成する.

(2) Excel での操作

①データの最大値と最小値をみる.

データでは各項目にフィルターを設定し▼をクリックすることで最大値と最小値をみることができる．図 7-2 に示す項目である総コレステロール値の最小値 125，最大値 315 である.

②階級数を決める

階級数はデータ数が 180 であるので$\sqrt{180}+1$で約 14 となるが，区切りのいいところで階級数を 20 とし，119，129，139，……309，319 と 10 きざみにワークシートの右に作成する（図 7-3）.

③ヒストグラム，度数分布表の作成

a．ツールバーの「ツール」から「分析ツール」を選択し，現れた「データ分析」ウィンドウより「ヒストグラム」を選択して，「OK」ボタンをクリックする.

b．現れた「ヒストグラム」ウィンドウ中の「入力範囲」ボックスの横にあるボタンをクリックし，総コレステロールの値全体を選ぶ（図 7-4）.

c．つづいて，カーソルを「データ区間」ボックスに移動し，ボックスの横にあるボタン

図 7-3　階級の値を入力する

図 7-4　ヒストグラムデータ選択

入力元の入力範囲にはヒストグラムを作成したいデータの範囲を選択する．項目（ラベル）である「総コレステロール」を選んだときは，ウィンドウの「入力元」の「ラベル」をチェックし，「出力オプション」では「累計度数分布の表示」と「グラフ作成」とをチェックする．

をクリックして，Excel 上に入力した階級数の値（119，129，139，・・・・・・・・309，319）を選択する（図 7-4）．

d．項目のラベルを選んだ場合には「ラベル」のボックスをチェックし，出力オプションでは「累積度数分布の表示」と「グラフ作成」のボックスをチェックしてから，「OK」ボタンをクリックする（図 7-4）．

e．この操作によってヒストグラム，度数分布表，累積度数分布表がワークシートに新しく作成される（図 7-5）．

g．作成されたヒストグラムの変更は，作成された棒グラフの上にカーソルをあわせ右クリックして現れたメニューより「データ系列の書式設定」を選択し（図 7-6），「棒の間隔」を 0 にしてから，「OK」ボタンをクリックする（図 7-7）．

h．図 7-8 に作成されたヒストグラムを示す．ほぼ左右対称の分布であるが，300 を超

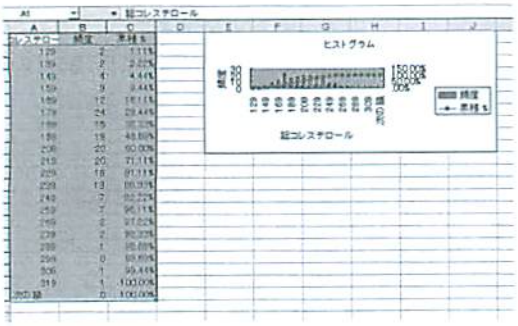


図 7-5　ヒストグラム作成結果
度数分布表ヒストグラムが作成される

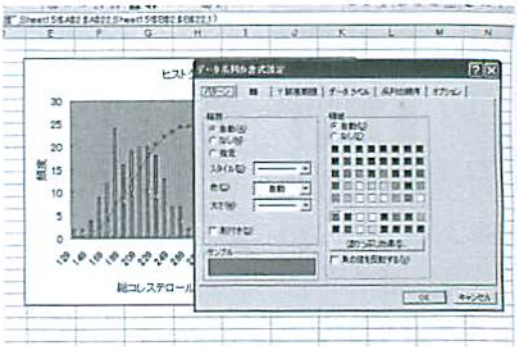


図 7-6　ヒストグラム作成結果

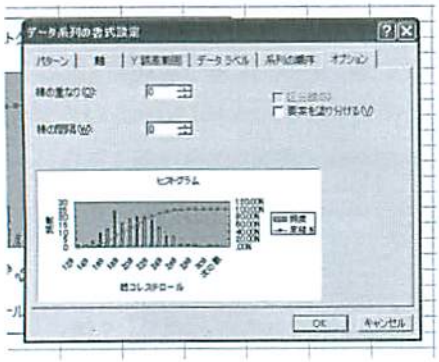


図 7-7　ヒストグラム作成結果
「棒の間隔」を 0 にする

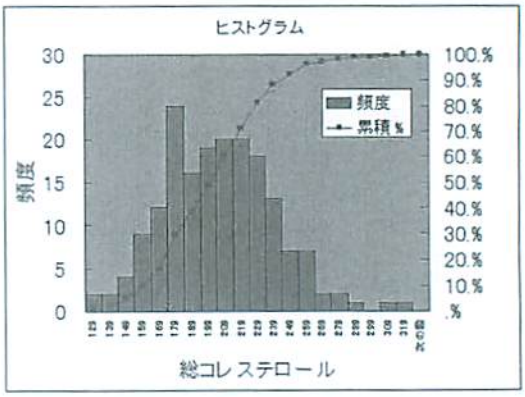


図 7-8　ヒストグラム作成結果

える値のものが 2 名いることがわかる．

i．同様の手順で血糖値および中性脂肪の度数分布表ヒストグラムを作成する．

(3) SAS による出力例

①幹葉表示（stem-and-leaf display）（図 7-9 左）

幹葉表示は測定値を幹と葉という 2 つの部分に分けて表示する．たとえば測定値 220 では 22 を幹，0 を葉と考える．この場合，22 の幹の右に 0 が 5 つ並んでいるので，測定値 220 のものが 5 名いることを表している．幹葉表示は，ヒストグラムと同じような形ができあがるが，元のデータの情報をより多く含んでいるのでヒストグラムより優れている．

②箱ひげ図（box plot）（図 7-9 中央）

箱中央の下端，中央，上端の点線で示した水平線は，それぞれ，第 1 四分位数，中央値，第 3 四分位数を表し，中央に示した「+」は平均値を示す．この場合は，中央値（201）と平均値（201.37）がほぼ同じ（図からはわからない）であることがわかる．

第 3 四分位数と第 1 四分位数との差を四分位偏差といい，箱の上端と下端から 1.5 四分位偏差以内で最も中央値から離れた点まで「ひげ」とよばれる垂直線がひかれる．垂直線

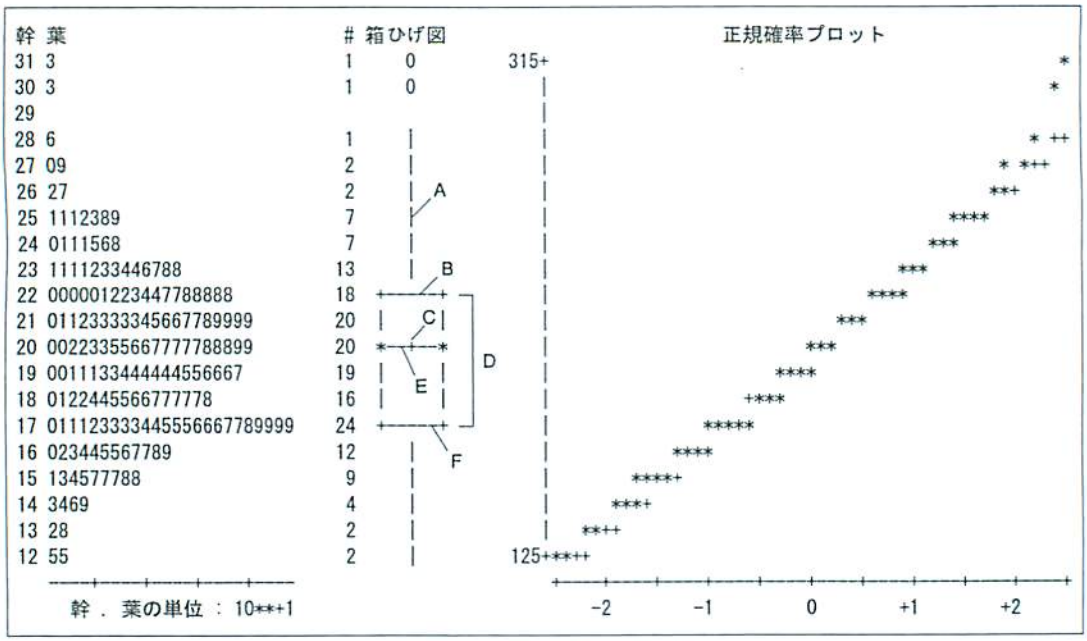


図 7-9　SAS による出力例

A：ひげ　B：第３四分位数　C：平均値　D：四分位偏差　E：中央値　F：第１四分位数

の端からさらに離れた測定値については箱から第３四分偏差までのものを外側値といい，丸（○），もしくはゼロ（0）で表し，それ以上離れた測定値は極外値といい，黒丸（●），あるいはアスタリスク（＊）で表される．外側値と極外値が外れ値の候補である．

箱ひげ図は統計量と分布の関係や外れ値の存在などを見るのに適している．この場合は，観測値の 303 と 313 が 0 で表されている．

③正規確率プロット（normal probability plot）（図 7-9 右）

正規分布に従っているかどうかを調べる場合のプロットを正規確率プロットとよぶ．データを大きさの順に並びかえ，標準正規分布のパーセント点を横軸に，実際のデータを縦軸にとってプロットしたときのグラフである．データが正規分布に従うときは，このプロットは直線になる．図中のアスタリスク（＊）は実際のデータを用いてプロットされたもので，「＋」はそのデータと同じ標本平均と標本標準偏差をもつ正規分布からプロットされた基準線である．データが正規分布からずれるほどこの基準線から離れることになる．図 7-9 では，高い値にズレがあることがわかる（分布の正規性の検定は 168 頁を参照）．

	A	B	C	D	E
1		学生	母	父	
2	1	155	155	173	
3	2	158	155	170	
4	3	152	154	168	
5	4	158	155	170	
6	5	159	160	175	
7	6	165	155	173	
8	7	160	155	175	
9	8	160	156	166	
10	9	159	154	174	
11	10	164	154	175	
12	11	163	160	165	
13	12	161	156	175	
14	13	156	153	170	
15	14	160	155	162	
16	15	155	153	160	
17	16	162	156	172	
18	17	157	157	162	

図 7-10　身長データ

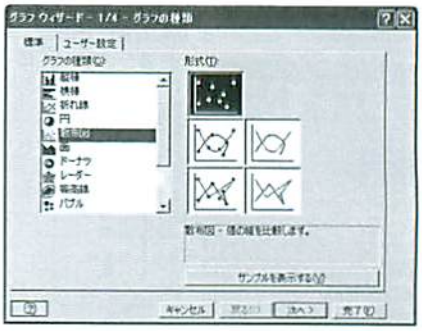


図 7-11　グラフウィザードグラフの種類

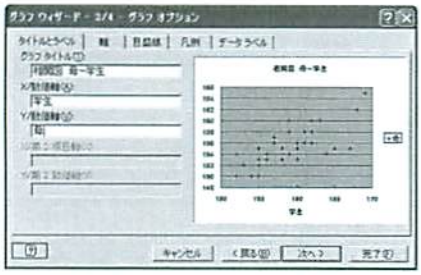


図 7-12　グラフウィザード記入

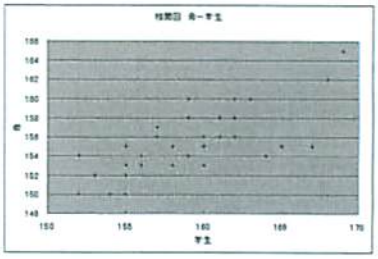


図 7-13　女子学生と母親の相関図

B　2 変量以上の場合

1）散布図（scatter diagram）から相関を判断する

変量が 2 つ以上の場合，解析対象とする 2 つを組み合わせて散布図を描き，相関がどの程度かを判断する．一般的に，2 変量の分析は直線的な相関を前提としている．したがって，相関が曲線的である場合は，尺度を変換するか，またはノンパラメトリックな統計解析（171 頁参照）を用いる．

例 7-2　Excel でのデータ分布の処理 Ⅱ（散布図，相関係数，回帰曲線）

(1) 使用するデータ

女子学生 39 名の身長とその両親の身長（図 7-10）．

(2) Excel での操作

①相関を調べたいデータを選択する（ここでは女子学生と母親の身長を選ぶ）．

②ツールバーの「挿入」メニューにある「グラフ」を選び，現れたウィンドウの「グラフの種類」から「散布図」を選択し，「次へ」をクリックする（図 7-11）．つぎのグラフウィザードで図を確認し，「次へ」をクリックする．次のウィザードでグラフタイトルと X 軸，Y 軸の値を入力し（図 7-12），「完了」ボタンをクリックすると（図 7-13），散布図が作成される．同様の方法で学生－父親，父親－母親の散布図を作成してみる．

身長との相関は，図から判断して，父親より母親とのほうが高いことがわかる（図

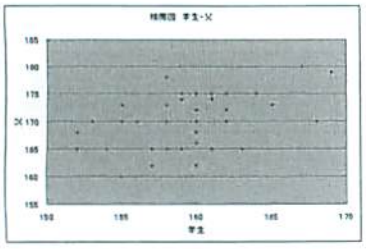


図 7-14　女子学生と父親の相関図

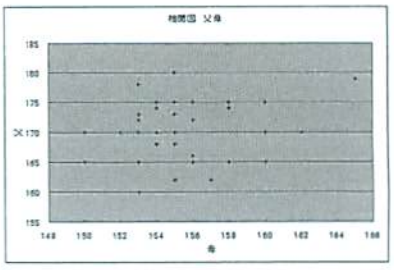


図 7-15　父親と母親の相関図

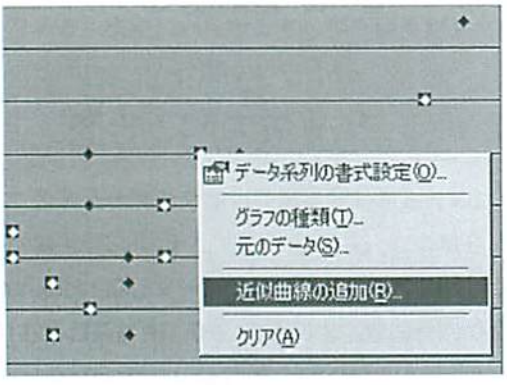


図 7-16　回帰直線の求め方

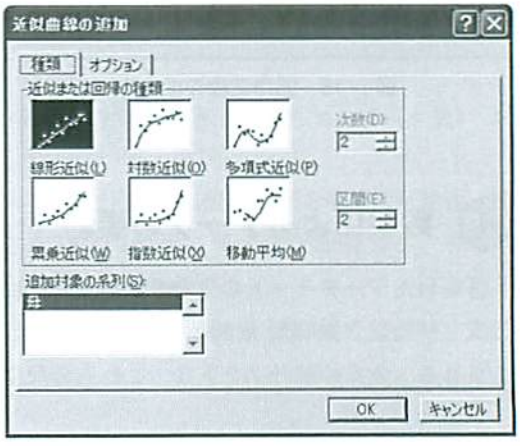


図 7-17　回帰直線の求め方

7-13, 14)．一方，父親と母親の関係はほとんどないように考えられる（図 7-15）．この関係を数値で示すのが相関係数であり，一般に r で示す．

③作成した散布図より相関係数および回帰直線を求める．

a．描かれた散布図（相関図）の任意のデータポイントを右クリックし，メニューより「近似曲線の追加」を選択する（図 7-16）．

b．「近似曲線の追加」ウィンドウより「線形近似」を選択し，つぎに「オプション」タブをクリックする（図 7-17）．

c．「オプション」内にある「グラフに数式を表示する」と「グラフに R-2 乗値を表示する」のボックスをチェックして（図 7-18），「OK」ボタンをクリックすると，グラフ中に，回帰直線，回帰式，R-2 乗値が表示される（図 7-19）．計算された R^2 を展開した値が r 値（相関係数）である．

④同様の方法で父親－学生，父親－母親の図中にも表示させる．

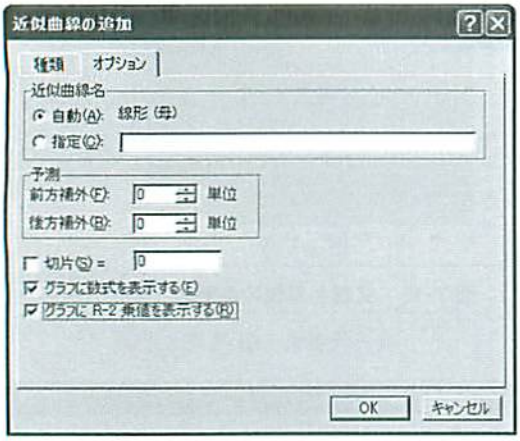


図 7-18　回帰直線の求め方

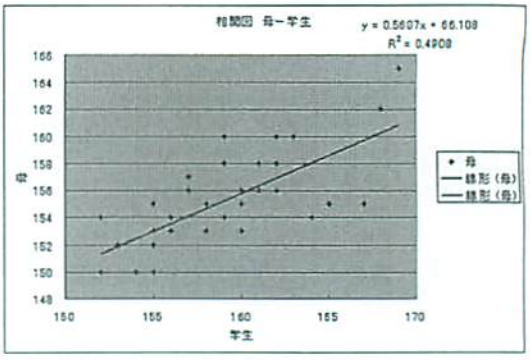


図 7-19　回帰直線完成

④ 数字によるデータの要約

得られたデータセットの分布の状態ならびに変数間の相関を把握した後に，データ分布の中心の尺度（平均値，中央値，最頻値），ばらつきの尺度（範囲，四分偏差，標準偏差，分散，変動係数など）を求める．多くの統計ソフトは，これらの尺度を容易に計算することができる．

A　分布の中心の尺度とばらつきの尺度

1）分布の中心の尺度

(1) 平均値（mean）

最もよく使われる分布の中心の尺度である．データのすべての観察値を合計し，測定度数で割ることによって計算される．質的データや順序データには用いない．平均値は外れ値に大きく影響を受ける．

(2) 中央値（mediam）

測定値の 50 パーセント点である．各測定値にそれほど左右されない．すなわち，異常値に左右されない．中央値は順序データ，離散または連続データに用いることができる．

(3) 最頻値（mode）

頻度が最も多い観察値である．すべてのタイプのデータに対する要約尺度として用いることができる．分布の中央を示す最もよい尺度は，値の分布の状態に依存する．左右ほぼ対象の分布であれば，平均値，中央値，最頻値はほぼ同じ値となる．分布が左右対称であっても二峰性の分布である場合は，平均値と中央値はおおよそ同じになるが，この場合は 2 つの最頻値を代表値とするか，2 つを別々に取り扱ったほうがよい．データの分布が左または右に偏っている場合は，平均値のそれぞれ左あるいは右になる．この場合は中央値が最もよい中央の尺度であることが多い

2） ばらつきの尺度

(1) 範　囲 (range)

観察値の最大値と最小値の差である．計算が容易であるが，外れ値に非常に左右されるのでその有用性はかぎられている．

(2) 四分偏差 (interquartile range)

分布の 75 パーセント点のデータから 25 パーセント点の差である．これは観察値の中心 50%を含むことになる．

(3) 標準偏差 (SD；standard deviation)

データセットのばらつきの尺度として用いられるのが分散であり，標準偏差は分散の正の平方根である．実際には分散よりも標準偏差がよく用いられる．

$$\text{分散}\ (s^2) = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2,\ s = \sqrt{s^2}\ (\bar{x}：\text{平均値}\quad s：\text{標準偏差}\quad n：\text{データ数})$$

値の全体的な分布の特徴を要約するためにデータセットの平均値と標準偏差を用いて平均値±標準偏差 (SD) として用いられる．この場合，平均値 ± 1SD の範囲にデータの約 67%が，平均値 ± 2SD の範囲にデータの約 95%が，平均値 ± 3SD にほとんどすべてのデータが入る．標準偏差は測定値と同じ単位であるので測定単位が異なる 2 つのデータグループの標準偏差を比較することは意味がない．

3） その他の尺度

(1) 分　散 (variance)

上記の「(3) 標準偏差」を参照

(2) 変動係数 (CV；coefficient of variation)

2 つ以上のデータセットの変動を比較するための尺度である．

$$\text{変動係数}\ (CV) = \frac{s}{\bar{x}} \times 100$$

標準偏差 (s) の平均値 ($\bar{x}$) に対する割合で，相対的な変動の尺度である．

(3) 標準誤差 (SE；standard error)

標準偏差はデータのバラツキの大きさを表す指標であるが，標準誤差は標準偏差を$\sqrt{n}$で割ったもので平均の推定精度を表すものである．

$$\text{標準誤差}\ (SE) = \frac{s}{\sqrt{n}}$$

標準誤差は標準偏差よりも値が小さくなることからデータのバラツキを表す指標とする誤用が見られるが，この 2 つ統計量の使い方は明確に区別する必要がある．

表 7-1　歪度と尖度の特徴

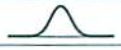
	分　布		
歪度	右にすそを引いている	左にすそを引いている	左右対称
	正 	負 	0
尖度	長くすそを引く	すそが切れている	正規分布
	3より小さい 	3より大きい 	3

図 7-20　基本統計量

B　分布の型

(1) 歪度（skewness）および尖度（尖り，kurtosis）（表 7-1）

分布が左右対称であれば，歪度は 0 になる．歪度および尖度は外れ値や異常値が含まれていると大きくなる．いずれもデータ数が 30 以上の場合に有効である．

(2) 分布の正規性

① Shapiro-Wilk の検定（W 値）

測定数（n）が 2,000 以下のときには，分布の正規性の判定に Shapiro-Wilk の検定（W 値）を用いる．Shapiro-Wilk の検定統計量（W）は，0 と 1 の間に値をとり，0 に近いほど正規分布からのずれが大きいことを示す．P 値（Pr < W）が 0.05 以下であれば 5%有意水準で正規性が棄却される（P 値については 170 頁図 7-23 を参照）．

② Kolmogorov-Smirnov の検定

測定数（n）が 2,000 以上のときには，分布の正規性の判定に Kolmogorov-Smirnov の検定を用いる．Kolmogorov-Smirnov の検定統計量（D）は，値が大きいほど正規分布からのずれが大きいことを示す．

例 7-3　Excel でのデータ分布の処理 III

(1) 使用するデータ

例 7-1 で使用の検査値

(2) Excel での操作法のながれ

①ツールバーの「ツール」より「分析ツール」を選択し，現れた「データ分析」ウィンドウ中で「基本統計量」を選び，「OK」ボタンをクリックする（図 7-20）．

②「基本統計量」ウィンドウの「入力範囲」ボックスに要約したい変数を選択する（図

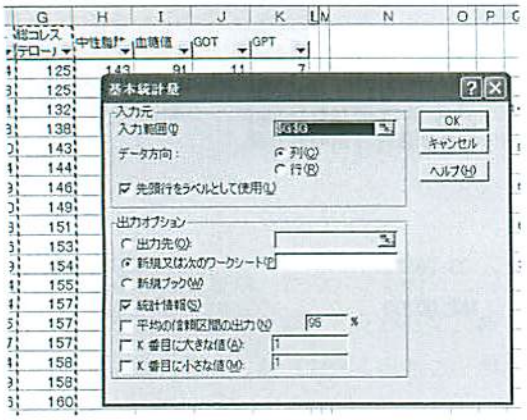


図 7-21　基本統計量

	A	B
1	総コレステロール	
2		
3	平均	201.37
4	標準誤差	2.48
5	中央値（メジアン	201.00
6	最頻値（モード）	194.00
7	標準偏差	33.27
8	分散	1106.78
9	尖度	0.36
10	歪度	0.37
11	範囲	188.00
12	最小	125.00
13	最大	313.00
14	合計	36247.00
15	標本数	180.00
16		

図 7-22　基本統計量

7-21 では総コレステロールを選択している）．先頭行を選択した場合は，「先頭行をラベルとして使用」のボックスをチェックする．出力オプションでは，「統計情報」を選択し，その他にも必要な項目があればボックスをチェックして「OK」ボタンをクリックすると（図 7-21），「基本統計量」が作成される（図 7-22 では「出力オプション」で「新規又は次のワークシート」を選択した）．

③ Excel の分析ツールでは四分偏差や変動係数が求められない．求める必要がある場合は，QUARTILE 関数を用いて求める．

a．ツールバーの「挿入」メニューにある「関数」をクリックし，現れた「関数の貼り付け」ウィンドウの左右より「統計」，「QUARTILE」を選んで，「OK」ボタンをクリックする．

b．現れたウィンドウの「配列」ボックスの横にあるボタンで対象とするデータの範囲を指定し，「戻り値」のボックスに 0 から 4 までの整数を入れて，「OK」ボタンをクリックする．最大値，最小値と 3 つの四分位点をそれぞれ求めることができる．戻り値は 0：データの最小値，1：第 1 四分位数，2：中央値，3：第 3 四分位数，4：最大値である（四分位範囲＝第 3 四分位数－第 1 四分位数）．

(3) SAS での出力例（図 7-23）

このデータセットは n が 180 であるので，正規性の検定は Shapiro-Wilk の統計量を用いる．P 値が 0.1727 であるので，このデータセットの正規性は棄却されない．（図 7-8，9 までに，総コレステロール値は歪度と尖度が正の値であることから，分布は右にすそをひいた形であり，P の正規確率プロットから正規分布がずれた値が高い値にあることが示されていた．）

⑤ 検定の基礎 －統計的仮説検定－

統計的仮説検定（statistical hypothesis test）では，「差がある」という仮説の逆の「差がない」という仮説を立てる．これを帰無仮説（null hypothesis）とよび記号 Ho で表す．データを評価し

標本数	180	重み変数の合計	180
平均値	201.372222	合計	36247
標準偏差	33.2683403	分散	1106.78246
歪度	0.36940343	尖度	0.35694531
無修正平方和	7497253	修正平方和	198114.061
変動係数	16.520819	平均の標準誤差	2.47967568

基本統計量

位置		ばらつき	
平均値	201.3722	標準偏差	33.26834
中央値	201.0000	分散	1107
最頻値	194.0000	範囲	188.00000
四分位範囲	46.00000		

位置の検定 : μ 0=0

検定	---統計量---		------p 値------	
Student の t 統計量	t	81.2091	Pr > t	<.0001
符号検定	M	90	Pr >= M	<.0001
符号付順位検定	S	8145	Pr >= S	<.0001

正規性の検定

検定	----統計量----		------p 値------	
Shapiro-Wilk	W	0.988906	Pr < W	0.1727
Kolmogorov-Smirnov	D	0.047476	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.056733	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.383359	Pr > A-Sq	>0.2500

パーセント点	推定値
100% 最大値	313.0
99%	303.0
95%	255.5
90%	241.0
75% Q3	222.0
50% 中央値	201.0
25% Q1	176.0
10%	161.0
5%	152.0
1%	125.0
0% 最小値	125.0

極値のオブザベーション

---最小値から-		---最大値から-	
値	Obs	値	Obs
125	139	270	22
125	119	279	117
132	162	286	133
138	111	303	71
143	50	313	132

図 7-23　SAS での総コレステロールの出力例

て，偶然では説明できないほどのデータがこの Ho からずれていれば，「差がない」という仮説は棄却されることとなり「差がある」とする．Ho からのずれが偶然の範囲であれば結論を保留する．この偶然の範囲を調べるために帰無仮説の下で生じる確率である P 値 (P value) を計算する．この P 値があらかじめ決めた有意水準 (significance level) (0.05 または 0.01 に設定することが多い)

よりも小さいときに有意な差があるとし,「5%（1%）の有意水準で有意である」という．また,5%の水準で有意であれば＊を記し，1%であれば＊＊で記すことがよくあるが，Excel や統計ソフトでは P 値は計算されるので，P 値をそのまま記載することが望ましい．

検定法には，両側検定法と片側検定法とがある．

1）検　定

平均値の差の検定は，差の大きさだけを対象としており，2つの標本の一方の値が大きいとか小さいことを示してはいない．したがって，多くの検定では両側検定を用いるべきである．事前に2つの標本の間のどちらかが大きくなることがわかっているような場合にのみ片側検定を用いる．多くの場合，両側検定の P 値は片側検定の2倍になる．よって，両側検定より片側検定のほうが差は有意になりやすいが，安易に片側検定を用いてはならない．

2）有意水準

有意水準5%は数学的にまったく根拠がなく，10%や1%の有意水準で検定を行うこともあり，経験的に用いられている数字である．たとえば，同じ強さだと仮定した野球チームが試合をして，どちらかが3連敗する確率は，$0.5 \times 0.5 \times 0.5 = 0.125$ である．3連敗は偶然と考えてもおかしくないが4連敗することは偶然とは考えられず，2つのチームの強さが同じとは考えにくい．このときの確率は $0.5 \times 0.5 \times 0.5 \times 0.5 = 0.0625$ であり，0.05 に近いのでこれを有意水準としている．

3）検定の選択（ノンパラメトリックとパラメトリック）

検定の方法の選択は，得られたデータの性質を把握したうえで決める．得られたデータが正規分布が仮定できる計量データの場合はパラメトリック手法を選択し，外れ値などがあって正規分布を仮定できない場合はデータを順位に変換して検定する方法であるノンパラメトリック手法を選択する．前者の代表が t 検定であり，後者は Wilcoxon 検定である．

A　2つの平均の比較

1）正規性がある2群の標本データ

計量データで正規性がある2群の標本データでは，対応がある場合とない場合で扱いが異なる．

①対応のない両群の分散が等しい場合の平均値の差の検定（Student の t 検定）

②対応のない両群の分散が等しくない場合の平均値の差の検定（Welch の t 検定）

③対応がある場合の平均値の差の検定（paired　t 検定）

例 7-4　対応のない両群の分散が等しい場合の平均値の差の検定（Student の t 検定）

(1) 使用するデータ

例 7-2 で使用した女子学生とその両親の身長のデータ

図 7-24　*F* 検定

	A	B	C
	F-検定：2 標本を使った分散の検定		
		学生	母
	平均	159.1795	155.3590
	分散	16.3090	10.4467
	観測数	39	39
	自由度	38	38
	観測された分散比	1.5612	
	P(F<=f) 両側	0.0872	
	F 境界値 両側	1.7167	

図 7-25　*F* 検定結果

	A	数式バー	C
1	t-検定：等分散を仮定した2標本による検定		
2			
3		学生	母
4	平均	159.18	155.36
5	分散	16.31	10.45
6	観測数	39	39
7	プールされた分散	13.38	
8	仮説平均との差異	0	
9	自由度	76	
10	t	4.61	
11	P(T<=t) 片側	0.0000079	
12	t 境界値 片側	1.67	
13	P(T<=t) 両側	0.0000158	
14	t 境界値 両側	1.99	

図 7-26　*t* 検定結果

	A	B	C
	F-検定：2 標本を使った分散の検定		
		父	母
	平均	170.23	155.3(
	分散	23.34	10.4!
	観測数	39	3!
	自由度	38	3!
	観測された分散比	2.23	
	P(F<=f) 両側	0.00756	
0	F 境界値 両側	1.71669	

図 7-27　*F* 検定結果

(2) Excel での操作

①まず，等分散の *F* 検定，すなわち両群での分散が等しいかどうか検定する．

②ツールバーの「ツール」より「分析ツール」を選択し，現れた「データ分析」ウィンドウのボックスより「*F* 検定:2 標本を使った分散の検定」を選択して,「OK」をクリックする．

③*F* 検定のウィンドウの「変数 1 の入力範囲」および「変数 2 の入力範囲」に比較する 2 変数を選択する（図 7-24 では，学生と母の身長を選択している）．先頭行を選択した場合は「ラベル」のボックスをチェックし，「OK」をクリックすると，「*F* 検定」が作成される（図 7-25）．

④図 7-25 によると，分散比が 1.5612 であり，*P*（*F* <= f）両側の値が 0.05 よりも大きいので，両変数の分散に差が認められないことがわかる．

⑤分散が等しい（等分散）場合の *t* 検定を行う．②と同様の手順で「*t* 検定：等分散を仮定した 2 標本による検定」を選択し，「OK」ボタンをクリックする．

⑥現れた *t* 検定のウィンドウで「変数 1 の入力範囲」および「変数 2 の入力範囲」に比較する 2 変数を選択して,「OK」ボタンをクリックすると, *t* 検定が作成される（図 7-26 では，学生と母の身長を選択している）．

⑦図 7-26 によると，*P* 値は 0.001 以下であり，両群の間の平均値に有意な差が認められる．

例 7-5　対応のない両群の分散が等しくない場合の平均値の差の t 検定（Welch の t 検定）

(1) 使用するデータ

例 7-2 で使用した女子学生とその両親の身長のデータ

(2) Excel での操作

①等分散の F 検定を父親と母親のデータから例 7-4 と同じ方法で行う（図 7-27）.

②分散比が 2.23 であり，P（$F <= f$）両側の値が 0.05 よりも小さいので両変数の分散に差が認められる.

③分散が等しくない場合の t 検定を例 7-4 と同様の方法で行う.

④ F 検定ウィザードで，比較する 2 変数を「変数 1 の入力範囲」および「変数 2 の入力範囲」で選択する．先頭行を選択した場合は先頭行を「ラベル」として使用のボックスをクリックし，OK をクリックする（ここでは，学生と母の身長を選択している）.

⑤ P 値は，0.001 以下であるので両群の間の平均値に有意な差が認められる.

2）　正規性のない 2 群の標本データ（ノンパラメトリック法）

平均値の差の検定では，両群の標本データの分布が変換を含めて正規分布でない場合，t 検定は使用できない．この場合，先に述べた分布に依存しない方法であるノンパラメトリック法の Wilcoxon の検定が適用される.

(1) Wilcoxon の順位和検定（Wilcoxon rank sum test）

①両群とも数が 20 以下の場合：U 表を使用する

②少なくとも大きいほうの群が 20 以上の場合：正規近似を使用する

(2) Wilcoxon の符号付き順位和検定（Wilcoxon signed-ranks test）

①差が 0 でないペア数が 20 以下の場合：T 表を仕様する

②異なったデータを示すペア数が 20 以上の場合：正規近似を使用する

Excel の標準の分析ツールにはこれらの手法は含まれていないので，Excel に対応したアドインソフトを追加する必要がある.

B　χ^2 検定

質的データや離散データを項目別に集計し，さまざまな要因に対して結果となる度数に差があることを検出する手法のうちで最も基礎となるのは，2 × 2 分割表（two-by-two frequency table, contingency table）である．この 2 × 2 の分割表の度数に，要因 - 結果の関連性が統計的に有意に出現しているかどうか，すなわち 2 群間で割合の出現の差を検定する手法が χ^2（かいじじょう）検定である．ただし，1 つのセルの度数が 5 以上でない場合はより精密な Fisher の直接確率法（Fisher's exact test）に従う必要がある.

2 × 2 の分割表におけるある要因が 2 群間の度数に差があるどうかは，2 群間で割合が等しいとする帰無仮説と等しくないとする対立仮説を，有意水準 5 %（0.05）に設定して行う.

検定を行う際には，はじめに帰無仮説が正しいとする仮定のもとに，分割表の各セルの期待度数

を計算する.

標本数（n）に対する観測度数の2×2の分割表を示すと次のようになる.

変数A	変数B		
	有	無	計
有	a	b	a + b
無	c	d	c + d
計	a + c	b + d	n

これに対する期待度数は次のようである.

変数A	変数B		
	有	無	計
有	E_1 ((a+b) (a+c) / n)	E_2 ((a+b) (b+d) / n)	a + b
無	E_3 ((c+d) (a+c) / n)	E_4 ((c+d) (b+d) / n)	c + d
計	a + c	b + d	n

χ^2検定は，観察度数（実際のデータ）と期待度数の差が偶然による違いより大きくないかどうかを確かめるために用いられる.

観察度数と期待度数とのずれを示す指標であるχ^2値は次のようになる.

$$\chi^2 = \frac{(a - E_1)^2}{E_1} + \frac{(b - E_2)^2}{E_2} + \frac{(c - E_3)^2}{E_3} + \frac{(d - E_4)^2}{E_4}$$

この値がχ^2分布し，$n \times m$の分割表の自由度は，$(n-1)(m-1)$で求められるので，2×2分割表は，$(2-1)(2-1)=1$となる．この値が，3.84以上であれば観察度数と期待度数とのずれが大きく差があるといえる.

例7-6　Excelでのクロス集計

(1) 使用するデータ

例7-1で使用の検査値を用いる．1つのセルの観察値と期待値が5以上であるので，χ^2検定を行うことができる.

(2) Excelでの操作

①コレステロール値の基準以上の者が男性に多いかどうかを検定する.

②ツールバーの「データ」から「ピボットテーブルとピボットテーブルグラフレポート」を選択する.

③はじめのウィザードは，そのまま「次へ」をクリックし，次のウィザードでデータ全部を選択して（図7-28），最後のウィザードで作成先を選択したら「OK」ボタンをクリックすると，ピボットテーブルが作成される.

⑤ピボットテーブルの項目より集計したい項目を選ぶ．図7-29では左の枠に性別を，上の枠にコレステロール判定をドラッグする.

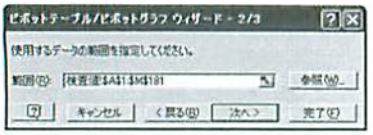


図 7-28　ピボット

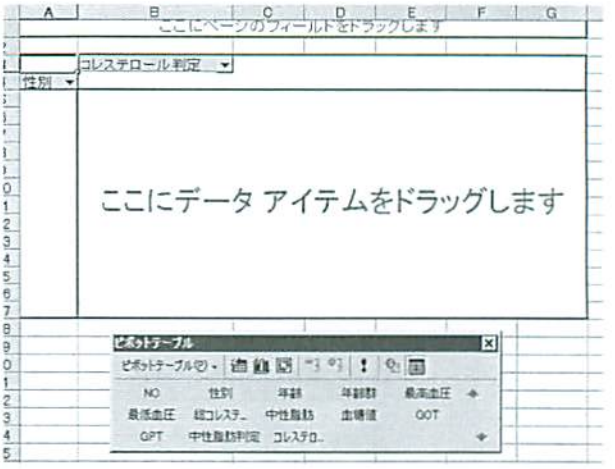


図 7-29　ピボット

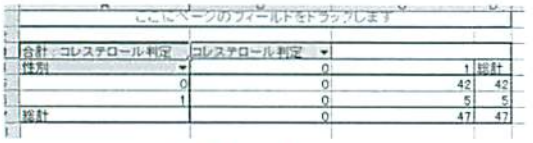


図 7-30　ピボット

図 7-31　ピボット

⑥コレステロール判定を表にドラッグすると図 7-30 のようになる．表に示される値は合計値であるのでこれをデータの個数に変える必要がある．

⑦クロス集計された値にカーソルを移動し，右クリックメニューより「フィールドの設定」を選択する（図 7-31）．データの個数を選択する（図 7-32）．

⑧性別とコレステロール判定によるクロス集計結果が示される．「データの個数：コレステロール判定」の上にカーソルを移動し，クリックすると，表全体が選択される．「編集」から「コピー」を選択し集計結果の表の下に「形式を選択して貼り付け」を選び，新しく表を作成する（図 7-34）．

⑨χ^2 検定のための期待度数とカイ 2 乗値を求める（図 7-35）．

関数（CHITEST）を選択する（図 7-36）．

実測地と期待値の範囲を選択する（図 7-37）．

a．期待度数の求め方

期待度数とは各行（または列）の合計度数を全体の行（または列）比率に応じて書くセルに配分した値である．したがって，男性の基準値未満の期待度数は，男性の合計度数

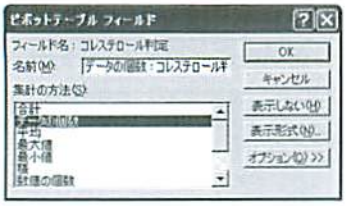


図 7-32　ピボット

データの個数：コレステロール判定	コレステロール判定		
性別	0	1 (空白)	総計
0	96	42	138
1	37	5	42
(空白)			
総計	133	47	180

図 7-33　ピボット

	A	B	C	D
1				
2				
3	データの個数：コレステロール判定	コレステロール判定		
4	性別	0	1	総計
5	0	96	42	138
6	1	37	5	42
7	総計	133	47	180
8				
9				
10	データの個数：コレステロール判定	コレステロール判定		
11	性別	基準値未満	基準値以上	総計
12	男性	96	42	138
13	女性	37	5	42
14	総計	133	47	180
15				

図 7-34　ピボット

	A	B	C	D	E	F	G	H
1								
2								
3	データの個数：コレステロール	コレステロール判定						
4	性別	0	1	総計				
5	0	96	42	138				
6	1	37	5	42				
7	総計	133	47	180				
8								
9								
10	データの個数：ココレステロール判定							
11	性別	基準値未満	基準値以上	総計				
12	男性	96	42	138				
13	女性	37	5	42				
14	総計	133	47	180				
15								
16								
17		期待度数				期待度数		
18		基準値未満	基準値以上	総計			基準値未満	基準値以上
19	男性	101.967	36.033	138		男性	=D12*B14/D14	=D12*C14/D14
20	女性	31.033	10.967	42		女性	=D13*B14/D14	=D13*C14/D14
21	総計	133	47	180		総計	=D14*B14/D14	=D14*C14/D14
22								
23		カイ2乗値				カイ2乗値		
24		基準値未満	基準値以上	総計			基準値未満	基準値以上
25	男性	0.3491	0.9880			男性	=(B12-B19)^2/B19	=(C12-C19)^2/C19
26	女性	1.1472	3.2463			女性	=(B13-B20)^2/B20	=(C13-C20)^2/C20
27	総計					総計	=(B14-B21)^2/B21	=(C14-C21)^2/C21
28								
29		カイ2乗値(z)	5.7306			カイ2乗値(z)	=SUM(B25:C26)	
30		p値	0.0167			p値	=CHITEST(B12:C13,B19:C20)	

図 7-35　ピボット

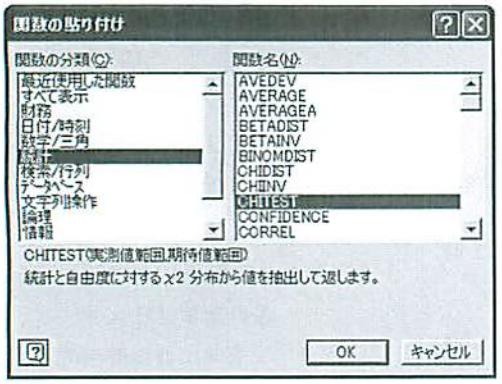


図 7-36　カイ 2 乗

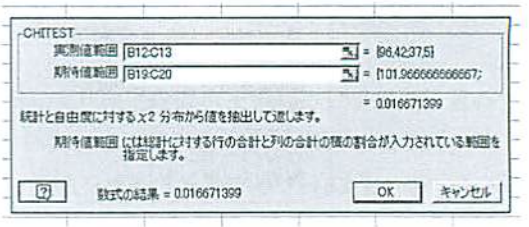


図 7-37　カイ 2 乗

138 を 133：180 の割合で配分すると，138 × 133 / 180　となり，101.967 である．他のセルについても同様に求める（図 7-35）．

b．カイ 2 乗値の求め方

カイ 2 乗値は，各セルの（実測地－期待度数）^{2} を期待度数で序した値である．

⑫ P 値を求める

P 値の求め方は図 7-36 に示すように関数の CHITEST を選択し，実測値と期待値の範囲を図 7-37 に示したウィザードで選択し求める．この場合 P 値が 0.0167 であり，2 つの変数間に関連性が認められる．

(3) SAS での X^2 検定（図 7-38）

⑥ 相　関

2 つの計量データ間の関連性を調べる場合は，まず散布図を作成してデータのおおよその関係を観察することが重要である．その後，相関の程度を記述する相関係数を計算する．

① 2 変数が正規分布し，直線的な関係があるときは Pearson 積率相関係数（Pearson product-moment correlation coefficient）を用いる．

② 2 変数間に曲線的な関連がみられ，2 変量正規分布が仮定できない場合には，2 つの変数間の相関の大きさを順位データの相関として測る尺度として Spearman の順位相関係数（Spearman's rank correlation coefficient）を用いる．

例 7-7　Excel での母相関係数の検定

(1) 使用するデータ

例 7-2 に同じ

(2) Excel での操作

①ツールバーの「分析ツール」を選択し，現れたウィンドウより「相関」を選択して，「入力元」

表 : CHOCLASS * GENDER

CHOCLASS 度数 パーセント 行のパーセント 列のパーセント	GENDER 0	1	合計
0	96 53.33 72.18 69.57	37 20.56 27.82 88.10	133 73.89
1	42 23.33 89.36 30.43	5 2.78 10.64 11.90	47 26.11
合計	138 76.67	42 23.33	180 100.00

CHOCLASS と GENDER の統計量

統計量	自由度	値	p 値
χ 2 乗値	1	5.7306	0.0167
尤度比 χ 2 乗値	1	6.4529	0.0111
連続性補正 χ 2 乗値	1	4.8104	0.0283
Mantel-Haenszel の χ 2 乗値	1	5.6988	0.0170
ϕ 係数		-0.1784	
不確実性係数		0.1757	
Cramer の V 統計量		-0.1784	

Fisher の正確検定

セル (1,1) 度数 (F)	96
左側 Pr <= F	0.0111
右側 Pr >= F	0.9970
表の確率 (P)	0.0081
両側 Pr <= P	0.0164

サンプルサイズ = 18

図 7-38　SAS での χ^2 検定の結果

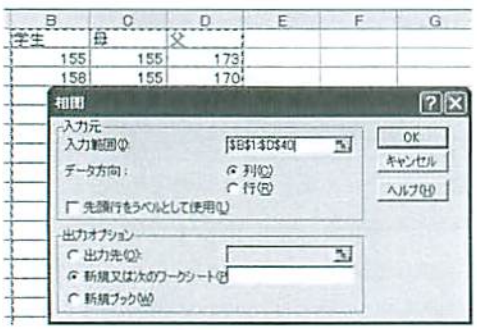


図 7-39　相関行列

	A	B	C	D	F
1		学生	母	父	
2	学生	1			
3	母	0.70057	1		
4	父	0.45237	0.215329	1	
5					
6					
7		検定統計量(t値)			
8		学生	母	父	
9	学生				
10	母	5.97183			
11	父	3.08541	1.34126		
12					
13		有意水準5%での棄却値		2.02619	=TINV(0.05,39-2)
14		有意水準1%での棄却値		2.715406	=TINV(0.01,39-2)
15		有意水準0.5%での棄却値		2.985253	=TINV(0.005,39-2)
16		有意水準0.1%での棄却値		3.573659	=TINV(0.001,39-2)
17					

図 7-40　相関行列

の「入力範囲」に求める2度数の範囲を選択すると（図7-39），相関行列が作成される（図7-40）．

②母親と学生との相関係数（0.70057）と n（39）より検定統計量を求める．まず，任意のセルを選択し，つぎに数式バーのボックスに「=B3 / SQRT(39 - 2) /SQRT (1 - B3)」とすべて半角で入力して，エンターキーを押して数値を求める．値は5.97183となる．

④ Excel関数のTINVを用いて有意水準0.1～5%での棄却値を求める．有意水準5%の棄却値は，任意のセルを選択してから，数式バーのボックスに「=TINV (0.05, 39 - 2)」と入力して求める．

⑤相関係数の検定

父親と母親には相関が認められないが，学生と母親間には0.1%以下の有意水準で，学生と父親間には0.5%以下の有意水準で相関係数が0でないと判断できる．

A　回帰分析

回帰分析とは，ある変数の値を他の変数の線形方程式（1次式）によって予測しようとする手法である．

1）回帰方程式

一般に変数 y を x の一次式で予測する場合，モデルで説明されないバラツキ（誤差）ε は，以下のようになる．

$$y = \alpha + \beta + \varepsilon$$

このとき，予測される変数 y を従属変数（dependent variable）あるいは目的変数，予測に用いる変数 x を独立変数（independent variable）あるいは説明変数（explanatory variable）という．定数 α は，$x = 0$ のときの y の値を示し切片（intercept）といい，β は x が1増加するときの y の増加分を表し，回帰係数（regression coefficient）という．

方程式で求められた値 y を「y の予測値（predicted value）」といい，実測地と予測値の差は残差（redidual）とよぶ．

2）予測値と残差

従属変数（y）の分散をV（y），回帰方程式によってもとめた y の予測値の分散を Var（$\hat{y}$），残差の分散をV（e）とすると，以下の関係が成立する．

$$\mathrm{Var}\ (y) = \mathrm{Var}\ (y) + \mathrm{Var}\ (e)$$

予測値のもつ分散を「モデルによって説明される分散」という．また，元の変数のもつ分散 Var(y) に対する Var($\hat{y}$) の比率を，寄与率あるいは分散説明率ということがある．また決定係数(coefficient of determination）ともいう．説明変数が1つの場合は，目的変数と説明変数の相関係数の2乗と等しい．また，決定係数はこの回帰方程式によって説明される部分の割合を説明することができる．

図 7-41　回帰分析 2

	A	B	C	D	E	F	G	H	I
1	概要								
2									
3	回帰統計								
4	重相関 R	0.70057							
5	重決定 R2	0.4907863							
6	補正 R2	0.4770361							
7	標準誤差	2.920452							
8	観測数	39							
9									
10	分散分析表								
11		自由度	変動	分散	観測された分散比	有意 F			
12	回帰	1	304.1691	304.1691	35.6627619	6.655E-07			
13	残差	37	315.5745	8.52904					
14	合計	38	619.7436						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	23.18770	22.77701	1.01803	0.31527416	-22.96286	69.33827	-22.96286	69.33827
18	母	0.87534	0.14658	5.97163	0.00000059	0.57834	1.17233	0.57834	1.17233
19									
20	残差出力								
21									
22	観測値	予測値：学生	残差	標準残差					
23	1	158.8653	-3.8653	-1.3413					
24	2	158.8653	-0.8653	-0.3003					
25	3	157.9699	-5.9699	-2.0786					
26	4	158.8653	-0.8653	-0.3003					
27	5	163.2420	-4.2420	-1.4720					
28	6	158.8653	6.1347	2.1288					
29	7	158.8653	1.1347	0.3938					
30	8	159.7406	0.2594	0.0900					
31	9	157.9899	1.0101	0.3505					
32	10	157.9899	6.0101	2.0856					
33	11	163.2420	-0.2420	-0.0840					
34	12	159.7406	1.2594	0.4370					
35	13	157.1146	-1.1146	-0.3868					
36	14	158.8653	1.1347	0.3938					
37	15	157.1146	-2.1146	-0.7338					
38	16	159.7406	2.2594	0.7840					
39	17	160.6159	-3.6159	-1.2548					
40	18	154.4886	-0.4886	-0.1695					
41	19	161.4913	-2.4913	-0.8645					
42	20	159.7406	-2.7406	-0.9510					

Sheet14 / Sheet15 / Sheet16 / Sheet17 / Sheet18 / Sheet19 / Sheet20 / Sheet21 \ Sheet22 / Sheet23 / Sheet1 /

図 7-42　回帰分析 3

回帰分析の結果（目的変数：学生の身長，説明変数：母親の身長）

例 7-8　Excel での回帰分析

(1) 使用するデータ

例 7-2 で使用の女子学生と両親の身長

(2) Excel での操作

①ツールバーの「ツール」より「分析ツール」を選択し，現れた「データ分析」ウィンドウ中で「回帰分析」を選択する．

	A	B	C	D	E	F	G	H	I
1	概要								
2									
3	回帰統計								
4	重相関 R	0.45237036							
5	重決定 R2	0.20463894							
6	補正 R2	0.18314269							
7	標準誤差	3.64995342							
8	観測数	39							
9									
10	分散分析表								
11		自由度	変動	分散	観測された分散比	有意 F			
12	回帰	1	126.82367	126.82367	9.51975287	0.003836			
13	残差	37	492.91992	13.32216					
14	合計	38	619.74359						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	94.8077479	20.871456	4.5424597	0.0000574	52.5182	137.0973	52.5182	137.0973
18	父	0.37814397	0.1225588	3.0854097	0.0038358	0.129817	0.626471	0.129817	0.626471
19									
20	残差出力								
21									
22	観測値	予測値：学生	残差	標準残差					
23	1	160.2267	-5.2267	-1.4512					
24	2	159.0922	-1.0922	-0.3033					
25	3	158.3359	-6.3359	-1.7592					
26	4	159.0922	-1.0922	-0.3033					
27	5	160.9829	-1.9829	-0.5506					
28	6	160.2267	4.7733	1.3253					
29	7	160.9829	-0.9829	-0.2729					
30	8	157.5796	2.4204	0.6720					
31	9	160.6048	-1.6048	-0.4456					
32	10	160.9829	3.0171	0.8377					
33	11	157.2015	5.7985	1.6100					
34	12	160.9829	0.0171	0.0047					
35	13	159.0922	-3.0922	-0.8586					
36	14	156.0671	3.9329	1.0920					
37	15	155.3108	-0.3108	-0.0863					
38	16	159.8485	2.1515	0.5974					
39	17	156.0671	0.9329	0.2590					
40	18	157.2015	-3.2015	-0.8889					
41	19	160.9829	-1.9829	-0.5506					

図 7-43　回帰分析 4

回帰分析の結果（目的変数：学生の身長，説明変数：父親の身長）

②回帰分析ウィンドウの「入力 Y 範囲」（目的変数）と「入力 X 範囲」（説明変数）にそれぞれの範囲を選択して入力して，「OK」ボタンをクリックすると概要が作成される（図 7-41 ～ 43）．（図 7-42 は，目的変数に「女子学生の身長」，説明変数に「母親の身長」を選択し，図 7-43 は説明変数に「父親の身長」を選択した概要である）

③女子学生の身長と母親の身長の回帰式は，「女子学生の身長＝ 23.18770 ＋ 0.87534 ×母親の身長」であり，P 値は 0.0001 以下であるので学生の身長と母親の身長の間には相関関係があることがわかる．さらに決定係数が 0.49 であるので学生の身長の 49%は説明できることになる．一方，女子学生と父親との回帰式は，「学生＝ 94.807 ＋ 0.378 ×父親の身長」で P 値は 0.0038 となり，決定係数は 0.205 となって，学生の身長の約 20%は説明できることになる．したがって，父親と女子学生の身長との間には，母親ほどの相関がないことがわかる．

B　重回帰分析

1 つの説明変数 x で目的変数 y を予測する分析を単回分析といい（前述），2 つ以上の説明変数

図 7-44　重回帰

図 7-45　　重回帰

重回帰分析の結果（目的変数：学生の身長，説明変数：母親と父親の身長）

x_1，x_2，x_3，…… x_p を同時に使って y を予測する分析方法を重回帰分析（multiple regression analysis）という．一般に重回帰式は

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots \beta_p x_p + \varepsilon$$

と示すことができる．

各独立変数にかかる係数 β_1，β_2 …… β_p を偏回帰係数（partial regression coefficient），α を切片（intercept）という．

重相関分析において，目的変数と予測値との相関係数を重相関係数（multiple correlation coefficient）といい，この値の２乗を寄与率あるいは重決定係数としてあらわしている．

一般に説明変数を追加することにより，重相関係数は増加する．

例 7-9　Excel による重回帰分析

(1) 使用するデータ

例 7-2 で使用の女子学生と両親の身長

(2) Excel での操作

学生の身長を目的変数とし，母親と父親の身長を説明変数としたときの重回帰分析（図 7-44）．

①ツールバーの「ツール」より「分析ツール」を選択し，現れた「データ分析」ウィンドウ中で「回帰分析」を選択する．

②回帰分析ウィンドウの「入力Ｙ範囲」（目的変数）と「入力Ｘ範囲」（説明変数）にそれ

ぞれの範囲を選択して入力して，「OK」ボタンをクリックすると概要が作成される（図7-44では目的変数に「女子学生の身長」，説明変数には「母親と父親の身長」を選択した）．

③作成された概要（図7-45）から，「女子学生の身長」を目的変数とし，「母親と父親の身長」を説明変数としたときの重回帰方程式は，

学生 = 0.79 ×母親（x_1）+ 0.264 ×父親（x_2）− 8.59

となり，寄与率（重決定係数）は0.586である．「母親の身長」を説明変数としたときの寄与率が0.49であったのに対して「母親と父親の身長」を説明変数とすることによって寄与率は増加し，半分以上がこの式によって予測が可能であることが示されている．

C　回帰分析の評価

回帰分析が成功したかどうかは寄与率で判断できるが，予測する対象によって回帰式が十分な精度をもつかどうかは残差の標準偏差から判断する必要がある．

すなわち「女子学生の身長」を目的係数，母親の身長を説明係数に用いた場合の残差の標準偏差は2.92（図7-42のB7のセル）であり，この回帰式で予測した女子学生の身長と実際の身長の大部分（約95%）は2.92の2倍（5.84 cm）以内であるということが近似的にいえる．また，父親の場合では3.65（図7-43のB7のセル）の2倍（7.3 cm）であり，母親と父親の身長を用いた場合は2.67の2倍(5.3 cm)である．これが十分な精度かどうかが予測式としての有効性の評価となる．

⑦ 分散分析

t 検定およびWilcoxon検定は2つのグループの平均値の差の検定であった．ここでは，3つ以上の正規母集団のグループについての平均値の差の検定を行う．

A　一元配置分散分析

分散分析（analysis of variances）は，変動をいくつかの因子に分解する手法であり，以下の2条件を満たす必要がある

①各群が正規分布に従う母集団から抽出された標本である（正規性）

②母集団のなかでは各群の分散は等しい（等分散性）

B　多重比較（multiple comparison）

一元配置の分散分析では，グループ間で平均値に差があるかないかしか検定できない．差があるという結果が得られた場合，どの群とどの群の間に平均値の差があるのかを検定する必要がある．その方法には，① Fisher のLSD法，② Bonferroni の方法，③ Tukey の方法，④ Dunnett の方法，⑤ Scheffe の方法がある．

繰返しのある2元配置

	A	B	C
a	10	9	10
a	12	11	9
a	11	10	7
b	11	11	10
b	10	10	8
b	11	8	8
c	9	10	10
c	10	9	9
c	12	10	9

図 7-46　分散分析

図 7-47　分散分析

C　正規性・等分散が成り立たない場合

母集団が正規性をもたないときは，全データに順位をつけて行う Kruskal-Wallis の検定を使う．

例 7-10　Excel の分散分析

(1) 使用するデータ

ある化合物の収量に影響を与える要因を評価する目的で，酵素の種類（要因 A ～ C）と温度（要因 a ～ c）を変えて繰り返して 3 回ずつ実験を行った結果（繰返しのある二元配置）である（図 7-46）．

(2) 一元配置の分散分析

二元配置のデータにおいて，温度の影響を無視して同じ温度で酵素の種類ごとに 9 回実験したとして酵素による効果のみで分析することにする．

①ツールバーの「ツール」より「分析ツール」を選択し，現れた「データ分析」ウィンドウ中で「分散分析：一元配置」を選択する．

②「分散分析：一元配置」のウィンドウの「入力範囲」でラベル A ～ C も含めてデータ全体を選択する（要因 a ～ c は選択しない）（図 7-47）．「先頭行をラベルとして使用」をチェックし，「OK」ボタンをクリックすると，「分散分析：一元配置」のワークシートが作成される（図 7-48）．

④ワークシートは，概要として計算された酵素の表と分散分析表からなる．分散分析表から F 値に対する P 値は 0.004 で 0.4％の有意水準では有意であり，酵素の種類によって収量が影響を受けていることが確認できる．

(3) 二元配置の分散分析

二元配置のデータで酵素と温度の 2 要因の効果で分析する．基本操作は 1 元配置と同じである．

①ツールバーの「ツール」より「分析ツール」を選択し，現れた「データ分析」ウィンドウ

	B	C	D	E	F	G
分散分析：一元配置						
概要						
グループ	標本数	合計	平均	分散		
A	9	96	10.66667	1		
B	9	88	9.777778	0.944444		
C	9	80	8.888889	1.111111		
分散分析表						
変動要因	変動	自由度	分散	観測された分散	P-値	F 境界値
グループ間	14.22222	2	7.111111	6.981818	0.004075	3.402832
グループ内	24.44444	24	1.018519			
合計	38.66667	26				

図 7-48　分散分析

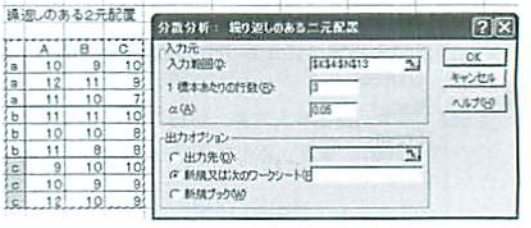


図 7-49　分散分析

	A	B	C	D	E	F	G
1	分散分析：繰り返しのある二元配置						
2							
3	概要	A	B	C	合計		
4	a						
5	標本数	3	3	3	9		
6	合計	33	30	26	89		
7	平均	11	10	8.666667	9.888889		
8	分散	1	1	2.333333	2.111111		
9							
10	b						
11	標本数	3	3	3	9		
12	合計	32	29	26	87		
13	平均	10.66667	9.666667	8.666667	9.666667		
14	分散	0.333333	2.333333	1.333333	1.75		
15							
16	c						
17	標本数	3	3	3	9		
18	合計	31	29	28	88		
19	平均	10.33333	9.666667	9.333333	9.777778		
20	分散	2.333333	0.333333	0.333333	0.944444		
21							
22	合計						
23	標本数	9	9	9			
24	合計	96	88	80			
25	平均	10.66667	9.777778	8.888889			
26	分散	1	0.944444	1.111111			
27							
28							
29	分散分析表						
30	変動要因	変動	自由度	分散	観測された分散比	P-値	F 境界値
31	標本	0.222222	2	0.111111	0.088235	0.915939	3.554561
32	列	14.22222	2	7.111111	5.647059	0.012486	3.554561
33	交互作用	1.555556	4	0.388889	0.308824	0.86829	2.927749
34	繰り返し誤	22.66667	18	1.259259			
35							
36	合計	38.66667	26				

図 7-50　分散分析（繰り返しのある二元配置）

中で「分散分析：繰り返しのある二元配置」を選択する．

②「分散分析：繰り返しのある二元配置」のウィンドウの「入力範囲」でラベルA～Cおよびa～cを含めてデータ全体を選択する(図7-49)．「先頭行をラベルとして使用」のボックスをチェックし，「OK」ボタンをクリックすると，「分散分析:繰り返しのある二元配置」のワークシートが作成される（図 7-50）．

④作成されたワークシート内の分散分析表の列すなわち酵素の F 値に対する P 値は 0.0125（図 7-50 の F32 のセル）で 1.25%の有意水準では有意である．一方，標本で表されてい

```
Dependent Variable: atai

                                    Sum of
Source                      DF      Squares     Mean Square    F Value    Pr > F
Model                        2   18.07407407     9.03703704       7.18    0.0036
Error                       24   30.22222222     1.25925926
Corrected Total             26   48.29629630

R-Square     Coeff Var      Root MSE     atai Mean
0.374233      11.65327      1.122167      9.629630

Source                      DF        Type I SS     Mean Square    F Value    Pr > F
class                        2   18.07407407     9.03703704       7.18    0.0036
Source                      DF       Type III SS     Mean Square    F Value    Pr > F
class                        2   18.07407407     9.03703704       7.18    0.0036

Tukey's Studentized Range (HSD) Test for atai

Alpha                                          0.05
Error Degrees of Freedom                         24
Error Mean Square                          1.259259
Critical Value of Studentized Range        3.53170
Minimum Significant Difference              1.3211

Means with the same letter are not significantly different.

Tukey
Grouping        Mean        N    class

      A        10.6667       9    A
      A
B     A         9.5556       9    B
B
B               8.6667       9    C
```

図 7-51　SAS での一元配置分析結果（Tukey の方法）

る温度の F 値に対する P 値は 0.9159（図 7-50 の F31 のセル）であるので有意水準に達していない．また，2 つの要因の交互作用の F 値に対する P 値は 0.868（図 7-50 の F33 のセル）であり，これも有意ではないので，このデータは一元配置分散分析が適切である．

(4) SAS での一元配置分析結果（Turkey の方法）

Excel の分析では各群間の有意性は求められていないが SAS では分散分析の多重比較に Turkey などの方法がある．図 7-51 に Turkey の方法による分析結果を示したが図下に grouping として群平均の左側に AAA と BBB が記されている．この意味は AAA で囲まれている A と B 群の群平均，BBB でかこまれている B と C 群の群平均の間には有意差がないことを示している．この例では A と C 間は同じ文字で囲まれていないので有意差があることがわかる．

参考文献

1）竹内　啓監修：SASで学ぶ統計的データ解析1 SASによるデータ解析入門 第2版．東京大学出版，東京，1993.
2）竹内　啓監修：SASで学ぶ統計的データ解析5 SASによる実験データの解析．東京大学出版，東京，1989.
3）竹内　啓監修：SASで学ぶ統計的データ解析6 SASによる回帰分析．東京大学出版，東京，1996.
4）涌井良幸，涌井貞美：Excelで学ぶ統計解析．ナツメ社，東京，2003.
5）上田太一郎，苅田正雄：実践ワークショップ Excel徹底活用 多変量解析．秀和システム，東京，2003.
6）渡辺美智子，小山　斉：実践ワークショップ Excel徹底活用 統計データ分析．秀和システム，東京，2003.
7）浜田知久馬：学会・論文発表のための統計学，統計パッケージを誤用しないために．真興交易医書出版部，東京，1999.
8）竹内正弘監訳：ハーバード大学講義テキスト 生物統計学入門（Pangono, M. and Gauvreau, K.：Pronciples of Biostatics 2nd ed）．丸善，東京，2003.
9）森田茂穂監訳：医学統計データを読む（Dawson-Saunders, B., and Trap, R. G.：Basic & Clinical Biostatics）．メディカル・サイエンス・インターナショナル，東京，1997.
10）宮原英夫，丹後敏郎：医学統計学ハンドブック．朝倉書店，東京，1995.
11）石村貞夫，デズモンド・アレン：すぐわかる統計用語．東京図書，東京，1997.
12）真野喜洋編：スタンダード公衆衛生学．文光堂，東京，2003.