1.2 量的データの集計

1.2.1 分布とヒストグラム

量的なデータの集計では、まずデータの分布を見ることが大切です。どの範囲にどれだけの数のデータがあるのかを示すのが度数分布表です。度数分布表の階級がデータを分類する範囲で、度数がどれだけのデータがその範囲に入っているかを表します。相対度数は、その度数の全体から見た割合です。また、それに加えて累積度数と累積相対度数を加える場合もあります。累積度数はその階級以前の度数の合計、累積相対度数はその全体から見た割合です。

表 1.2.1 度数分布表

階級	度数	相対度数(%)	累積度数	累積相対度数(%)
50 <= x < 60	4	20	4	20
60 <= x < 70	8	40	12	60
70 <= x < 80	5	25	17	85
80 <= x < 90	3	15	20	100
計	20	100		

度数分布表の度数の部分を棒グラフのように表示したグラフをヒストグラムといいます。階級が等間隔の場合、ヒストグラムは高さが度数になっています。

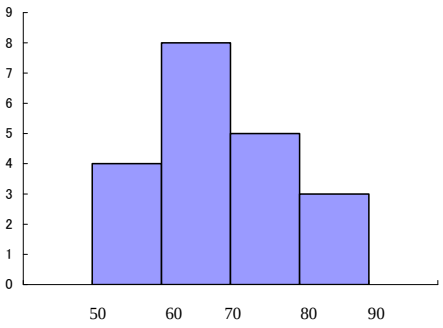


図 1.2.1 ヒストグラム

ヒストグラムが富士山型に近い形をしている場合、データの分布は正規分布であるといいます。正規分布についての詳しい話は、基礎からの統計学 7 章を参照して下さい。特に後に述べる検定などを利用する場合、正規分布かどうかは非常に大切です。

1.2.3 基本統計量

度数分布やヒストグラムはデータの特徴を最も良く表すものですが、人に一言で情

報を伝えたい場合には不便です。そのために我々は統計量と呼ばれる分布の特徴を要約した数値を使います。分布の特徴としては中心がどこなのか、広がりはどの程度かといったことが重要になります。正確には分布の中心を表す統計量のことを基本統計量と呼ぶようですが、ここでは他の統計ソフトなどと同様、総称して基本統計量と呼ぶことにします。まず分布の中心を与える統計量について説明します。今簡単のため、「3,3,4,2,8」というデータを考えてみます。

分布の中心を表す統計量

よく知られている分布の中心を表す統計量は平均値です。平均値はデータの合計をその個数で割ったものです。実際に上のデータについて求めてみましょう。

$$\text{平均値} = \frac{1}{5}(3+3+4+2+8) = 4$$

次に検定と呼ばれる処理でよく使われる分布の中心を表す統計量は中央値です。中央値はメジアンとも呼ばれ、文字どおりデータを小さい順に並べて真ん中の値です。このデータの場合は以下のようになります。

中央値 = 3　　2, 3, 3, 4, 8 のとき

またデータが偶数の場合は真ん中の 2 つのデータの平均になります。

中央値 = (3+4)/2=3.5　2, 3, 3, 4, 6, 8 のとき

分布の拡がりを表わす統計量

次は分布の拡がりを表す統計量についてです。最も簡単な指標はレンジ（範囲）と呼ばれる指標です。これはデータの最大値から最小値を引いたものです。

レンジ = 8 – 2 = 6

しかしこの指標には欠点があります。例えば極端に大きなデータが 1 つあった場合、そのデータの影響でレンジが極端に大きくなってしまいます。この欠点を取り除いたものが分散です。分散には通常の分散と不偏分散と呼ばれる量があります。分散は各データの平均値からのずれの 2 乗をデータ数で割ったもので、データを使って以下で与えられます。

$$\text{分散} = \frac{1}{5}\left[(2-4)^2 + (3-4)^2 + (3-4)^2 + (4-4)^2 + (8-4)^2\right] = 4.4$$

不偏分散は各データの平均値からのずれの 2 乗を（データ数－1）で割ったもので、分散より大きな値になります。

$$\text{不偏分散} = \frac{1}{4}\left[(2-4)^2 + (3-4)^2 + (3-4)^2 + (4-4)^2 + (8-4)^2\right] = 5.5$$

通常 Excel の関数で与えられる var(範囲) は不偏分散を表しており、アンケート調査の結果などで用いられるのは不偏分散の方です。2 つの分散の違いは標本調査のところで話をします。これらの分散では大きなずれは平均化されるのでレンジのときのような影響はありません。

これで問題解決と行きたいところですが、これらの分散にも問題があります。それは分散の定義にずれの 2 乗を使っているため、データの単位と異なることです。即ち例えばデータの単位が cm ならば、分散の単位は cm^2 になってしまい、拡がりを横軸上に表示できません。これを訂正するためには分散の平方根を取って単位を元に戻す方法が考えられます。このようにしてできたのが標準偏差で、通常の分散から求められるものと不偏分散から求められるものの 2 種類があります。名前に区別がありませんので、注意する必要があります。

$$\text{標準偏差} = \sqrt{\text{分散}} = 2.098$$

$$\text{標準偏差} = \sqrt{\text{不偏分散}} = 2.345$$

今までは 1 つの変数についてだけ見てきましたが、2 つの変数の関係を見る場合はどうでしょうか。これには通常散布図と呼ばれるグラフが使われます。例えば身長と体重の関係を見ようと思えば、身長を横軸に体重を縦軸に取って、各データをそこにプロットしていきます。しかし、それでは特徴を端的に言えないので、通常相関係数という統計量を用いて 2 つの変数の関係を表します。

今散布図の中に一本の直線を描くことを考えます。直線はできるだけ点のならびに近いように引きます。そして点がどれだけ直線に近いかで相関係数 r を以下のように決めます。

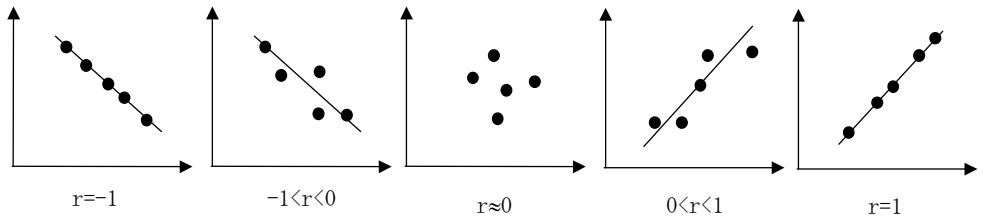


図 1.2.2 散布図と相関係数

右端と左端は 2 つの変数の関係がきっちり決まっていますので、相関係数の絶対値が 1 に近いほど相関が高いと言われます。相関係数がほとんど 0 だと無相関です。また、

相関係数がプラスだと正の相関、マイナスだと負の相関と言います。また、データ点に近くなるように引いた直線を回帰直線、この回帰直線の式を使って２つの量の関係を調べる分析を回帰分析といいます。それでは例題を用いて具体的な集計方法を見て行きましょう。

例

以下のデータ（Samples¥テキスト 1.txt）を用いて次の問いに答え、結果は文書にまとめよ。

学校	身長(cm)	体重(kg)	学校	身長(cm)	体重(kg)
2	169	71	1	170	62
1	175	68	1	182	75
2	170	67	2	177	70
1	179	72	1	175	70
1	176	69	1	172	62
2	174	81	2	166	58
2	173	75	2	168	60
1	181	65	2	173	58
1	179	74	2	169	59
2	178	71	2	170	73

この例題では身長と体重が量的データで、学校が２つのデータを分類する質的データです。まず、エディターの［ファイルー開く］メニューで、表示されるファイルから、テキスト 1.txt を選びます。通常サンプルは College Analysis の本体 CAnalysis.exe のあるフォルダーの下に Samples フォルダーを作って保存しておくと良いでしょう。

ファイルを選択すると、以下のような画面になります。

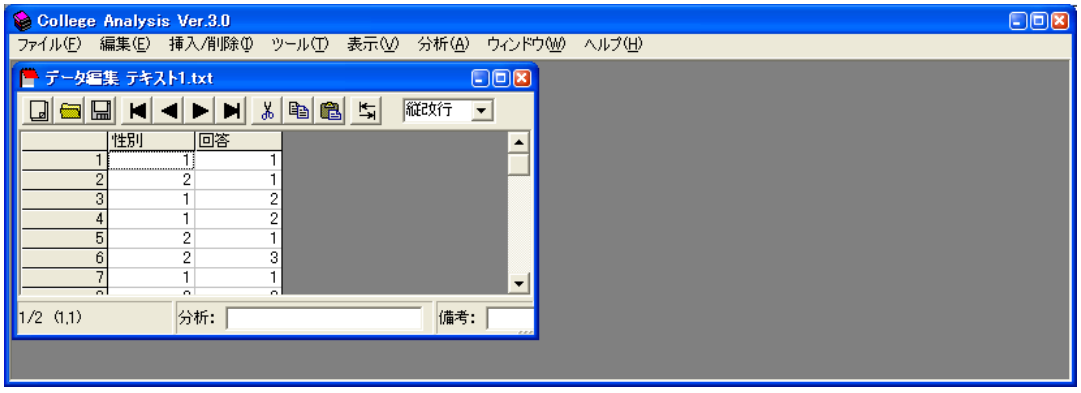


図 1.2.3 ファイル読込画面

この画面の右下の「1/2」の表示から、このファイルは 2 ページからなり、今 1 ページ目を表示していることが分かります。例題のデータは章ごとに「テキスト＊.txt」としてファイルに入っています。今回使うデータは 2 ページ目にあるので、ツールバーにある▶マークをクリックするとデータが現れます。

ページ切り替えには ⏮、◀、▶、⏭ の記号を使いますが、左から、先頭のページへ、前のページへ、次のページへ、最後のページへという意味です。

それでは、問題に答えて行きましょう。

1）身長についての基本統計量を求めよ。

まずメニュー［分析－基本統計－量的データの集計］を選び、図 1.2.1 のような量的データの集計メニューを表示します。

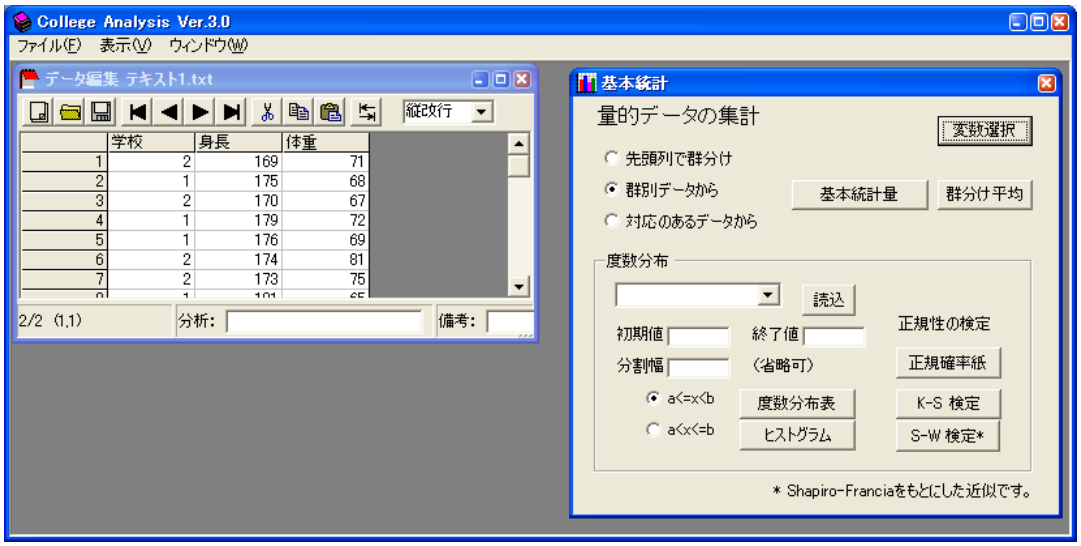


図 1.2.3 量的データの集計メニュー

「変数選択」ボタンで身長を選択します。その後「基本統計量」ボタンをクリックすると以下の結果が表示されます。但し、これは出力された 2 つの Window を横に並べて表示したものです。

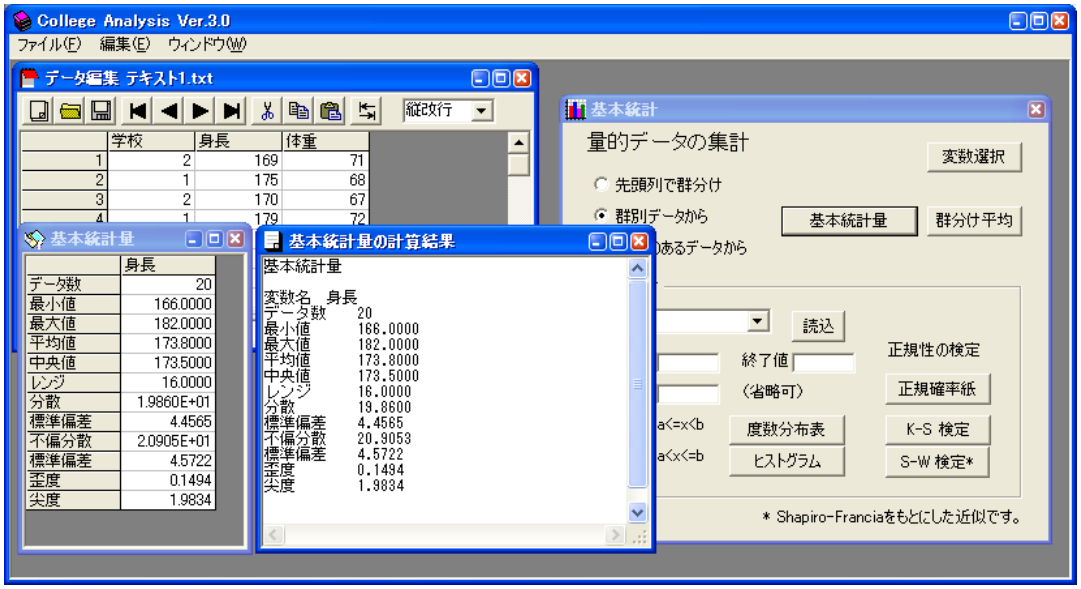


図 1.2.4 基本統計量表示画面

これを見ると基本統計量がテキストとグリッドで表示されています。これは用途によってコピーし易くするためで、内容は同じです。グリッド表示の分散に 1.9860E+01 という表示がありますが、これは浮動小数点表示と言って、$1.9860\times10^{1}=19.860$ を表します。この表記は Excel でもときどき使われます。また標準偏差については分散の下と不偏分散の下にありますが、これらはそれぞれ上の分散と不偏分散の平方根を取ったもので、通常我々は下側の不偏分散と標準偏差を使います。

2）体重についての基本統計量を求めよ。

今度は「変数選択」で体重を選択し、「基本統計量」ボタンをクリックすると以下の結果が出力されます。

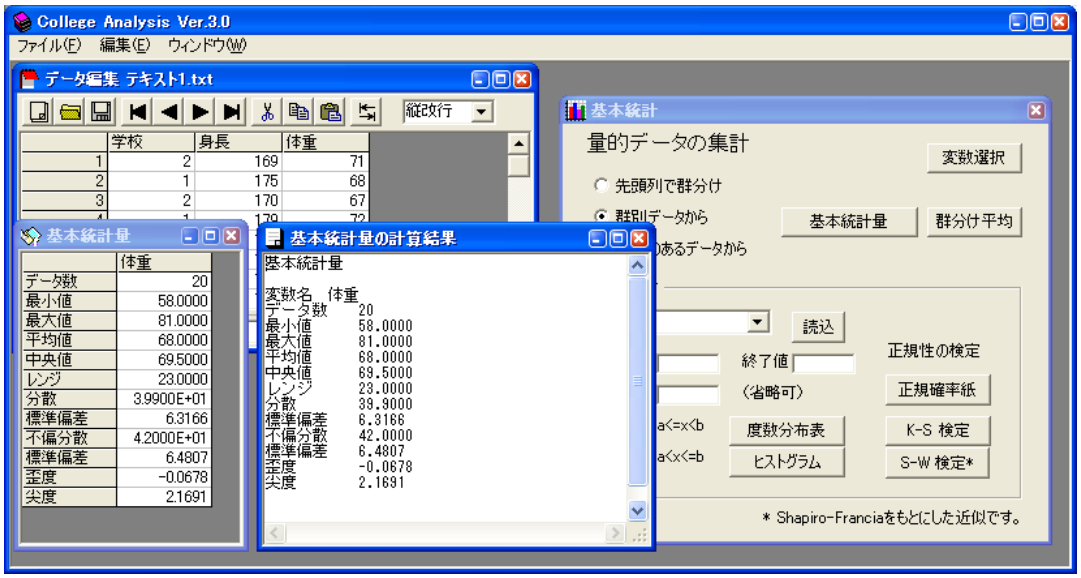


図 1.2.5 体重の基本統計量出力画面

3）身長について 5cm 毎の度数分布表を描け。

「変数選択」で身長を選び、度数分布のグループボックス内の「読込」ボタンで左のコンボボックスに身長を表示させ（1 変数を選択した場合は省略可)、「度数分布表」ボタンをクリックします。

図 1.2.6 身長の度数分布表表示画面

ここで今回は何の設定もなくうまく表示されましたが、初期値や分割幅（階級の幅）、場合によっては終了値（省略可です）を入力してから「度数分布表」をクリックすると好みの分割が作れます。

４）身長について 5cm 毎のヒストグラムを描け。

上の状態で「ヒストグラム」ボタンをクリックすると以下のようなヒストグラムが描けます。

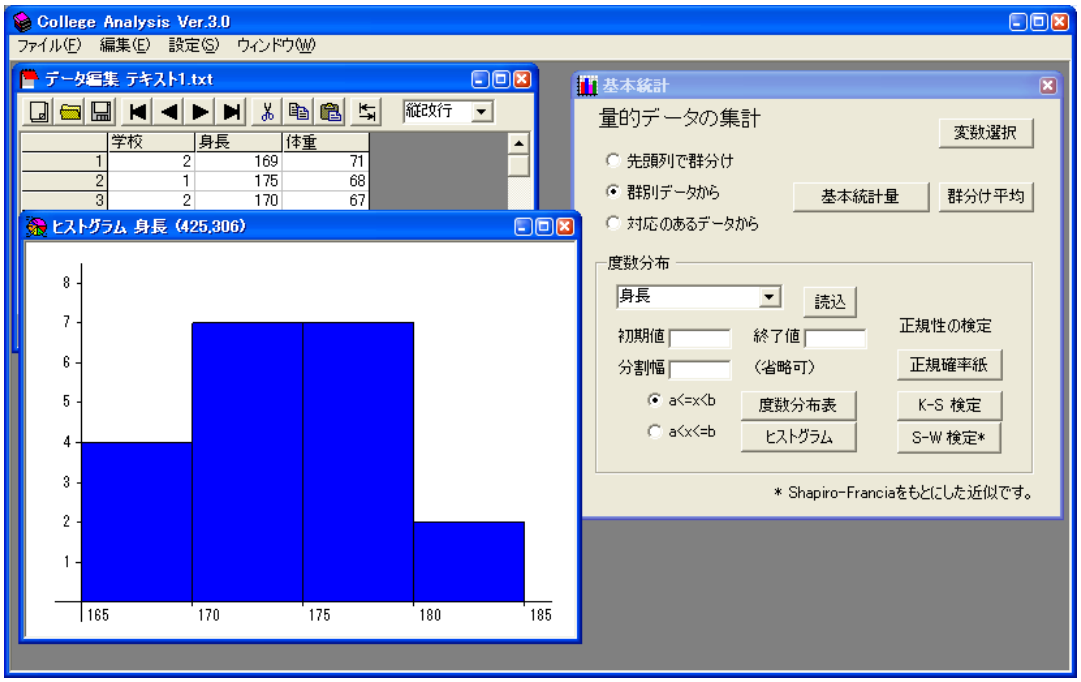


図 1.2.7 身長のヒストグラム表示画面

グラフの縦横を伸ばしたり縮めたりするとそれに合わせてグラフが変化します。その際文字の大きさは一定ですので、文字が重なるようなら、横に伸ばして拡げることもできます。

５）体重について 10kg 毎のヒストグラムを描け。

これも同様に「変数選択」ボタンで体重を選び、「読込」ボタンで左のコンボボックスに体重を表示し、「ヒストグラム」ボタンをクリックします。

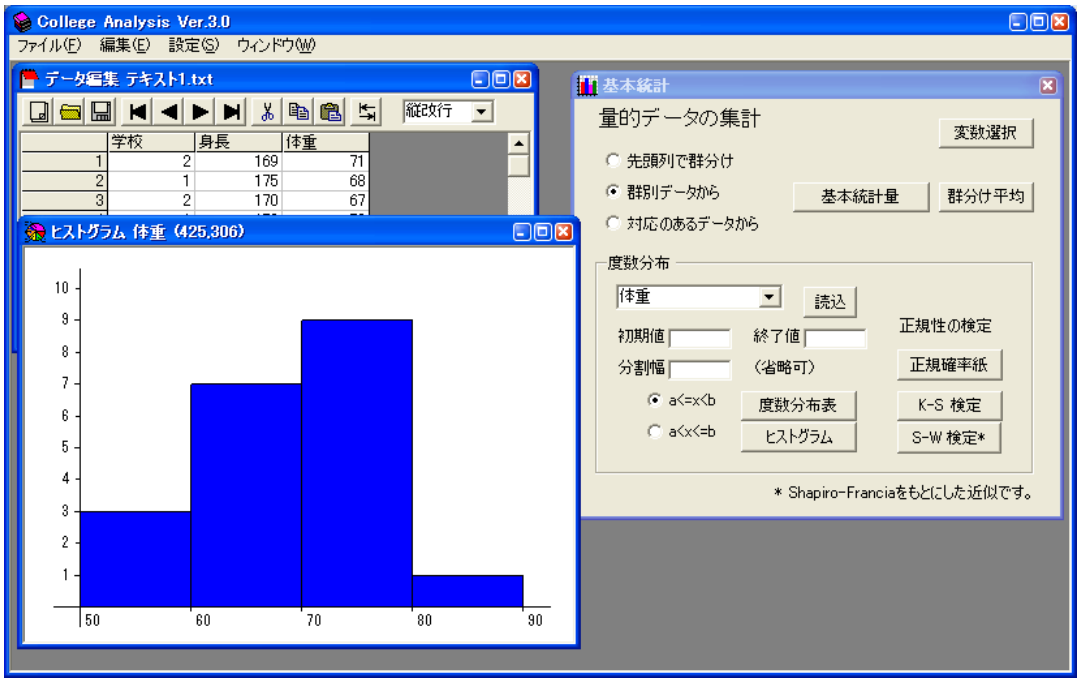


図 1.2.8 体重のヒストグラム表示画面

6）学校別に身長についての基本統計量を求めよ。

これは身長を分類しながら集計する問題です。まず「変数選択」で学校と身長を 2 つ選びます。ここで重要なのは分類する方の変数を先に選ぶことです。次にメニュー左上のラジオボタンで「先頭列で群分け」を選びます。これは最初に選んだ変数を群分けに利用せよという意味です。その後「基本統計量」をクリックすると以下のように分類された結果が表示されます。

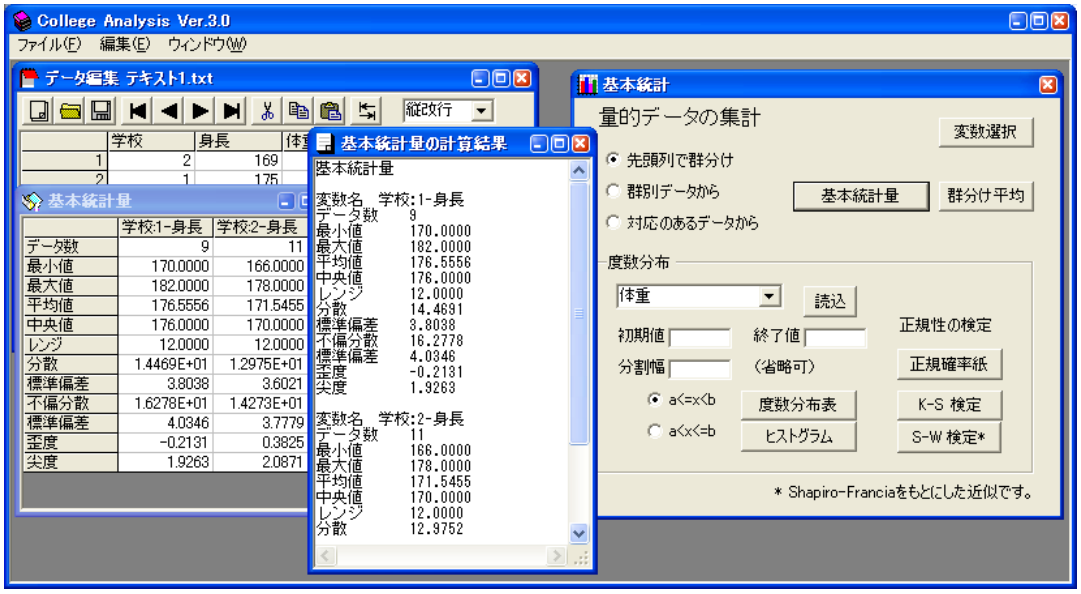


図 1.2.9 学校別に群分けされた身長の基本統計量

7）学校 1 について、身長のヒストグラムを描け。

これも「変数選択」で学校と身長、ラジオボックスは「先頭列で群分け」とします。「度数分布」グループボックス内の「読込」をクリックすると左のコンボボックスには「学校:1-身長」、「学校:2-身長」、「すべて」が設定され、先頭の要素が表示されます。このコンボボックスで表示したいものを選んで、「ヒストグラム」ボタンをクリックすると以下のようなヒストグラムになります。

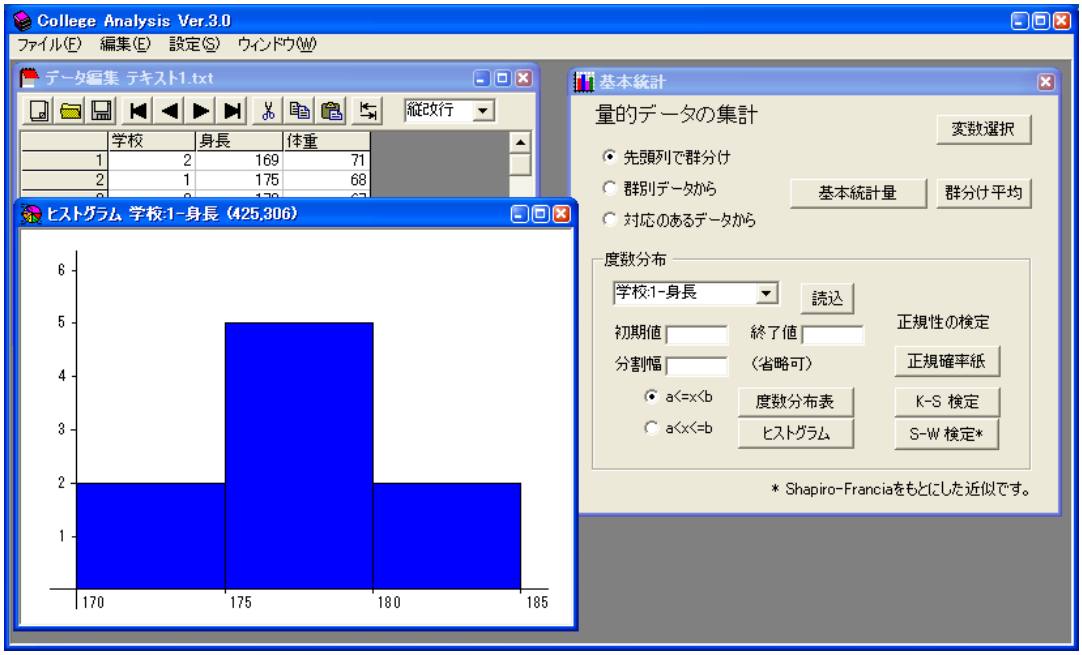


図 1.2.10 学校 1 の身長のヒストグラム

8）身長と体重に関する散布図を描け（体重を縦軸）。

2 つの量的データ間の関係を見るには、メニュー［分析－基本統計－相関と回帰分析］を選択します。

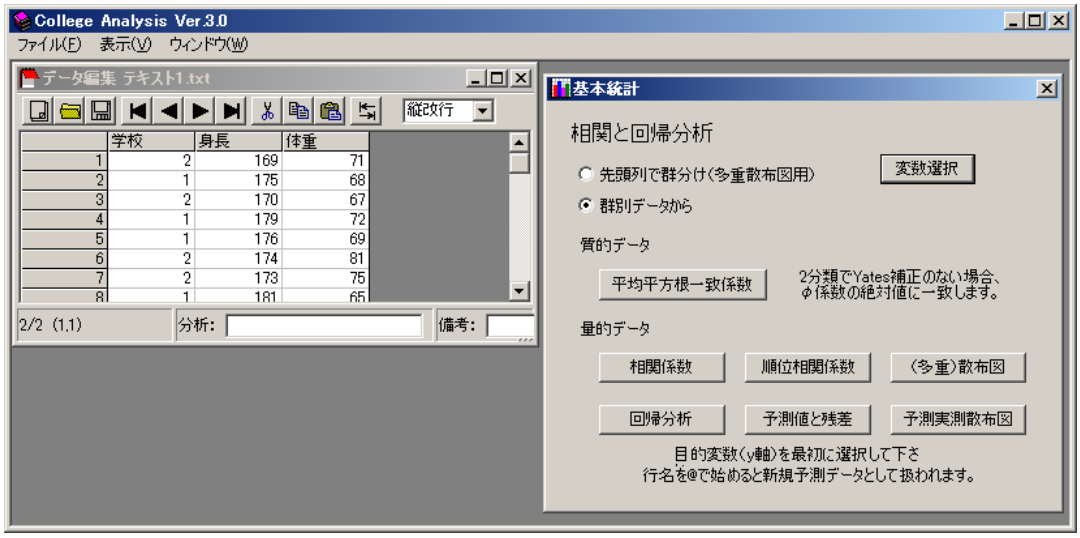


図 1.2.11 相関と回帰分析メニュー画面

「変数選択」ボタンで身長と体重を選びますが、縦軸にするものを最初に選択します。ここでは体重を縦軸にしますから、体重、身長の順で変数を選びます。「(多重) 散布図」ボタンをクリックすると以下の結果が表示されます。

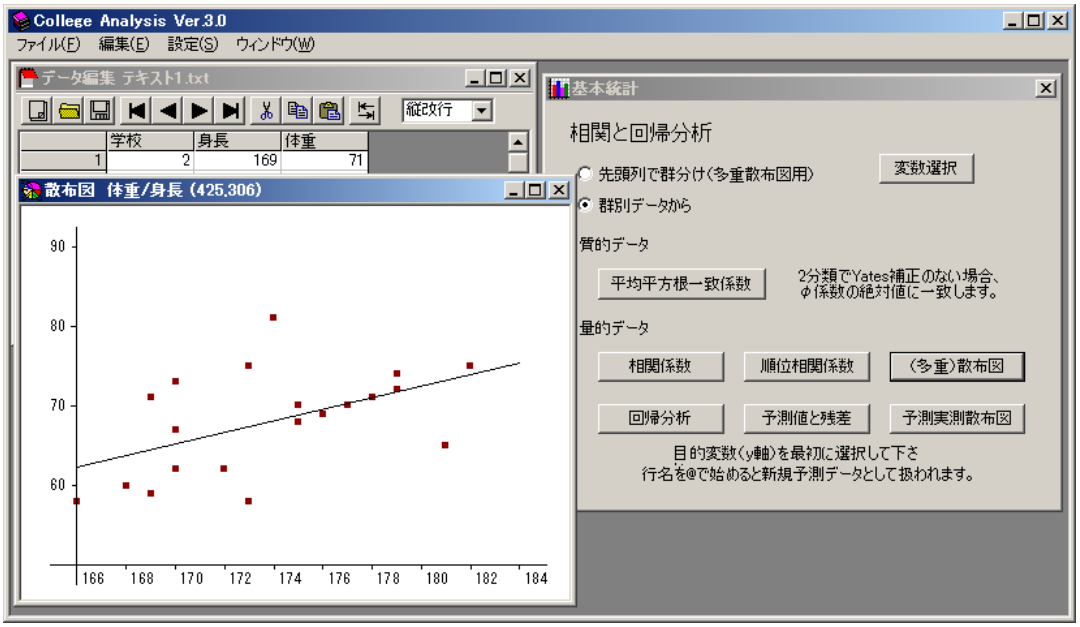


図 1.2.12 散布図表示画面

グラフメニューには「編集」と「設定」がありますが、これらの中のサブメニューにはグラフによって使えるものと使えないものがありますので、試してみて下さい。

9）身長と体重の相関係数を求めよ。

「変数選択」はこのままで「相関係数」ボタンをクリックします。相関係数の値を含む以下のような画面が表示されます。

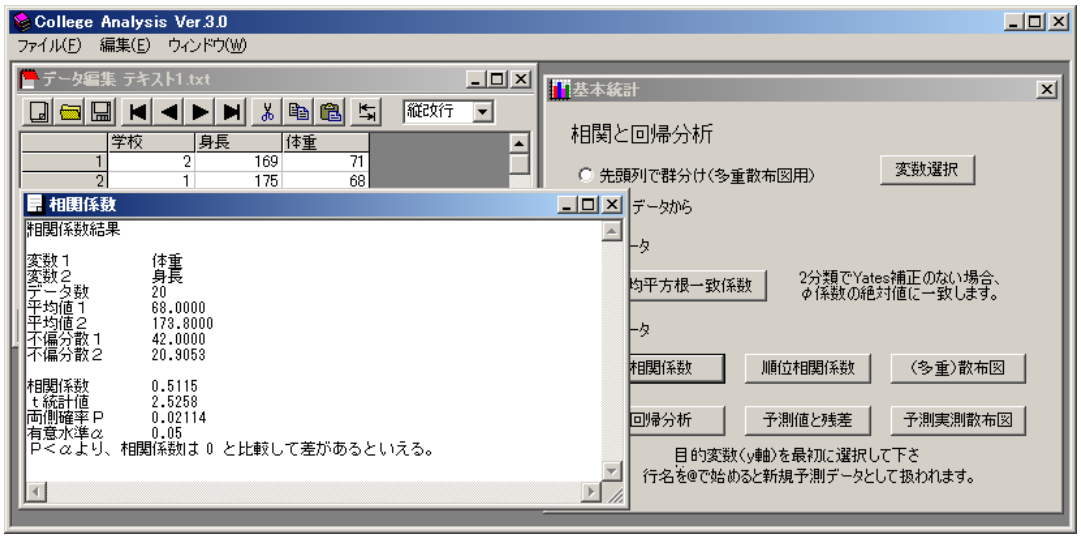


図 1.2.13 相関係数の表示画面

相関係数は中ほどで、そこから下は後に述べる相関係数の検定についての結果です。

10）身長で体重を予測する回帰式を求めよ。

最後に散布図のところで表示された直線についてです。これは回帰直線といい、身長で体重を予測する際の最も確からしい直線です。この直線の方程式は、「回帰分析」ボタンをクリックすることによって以下のように求められます。

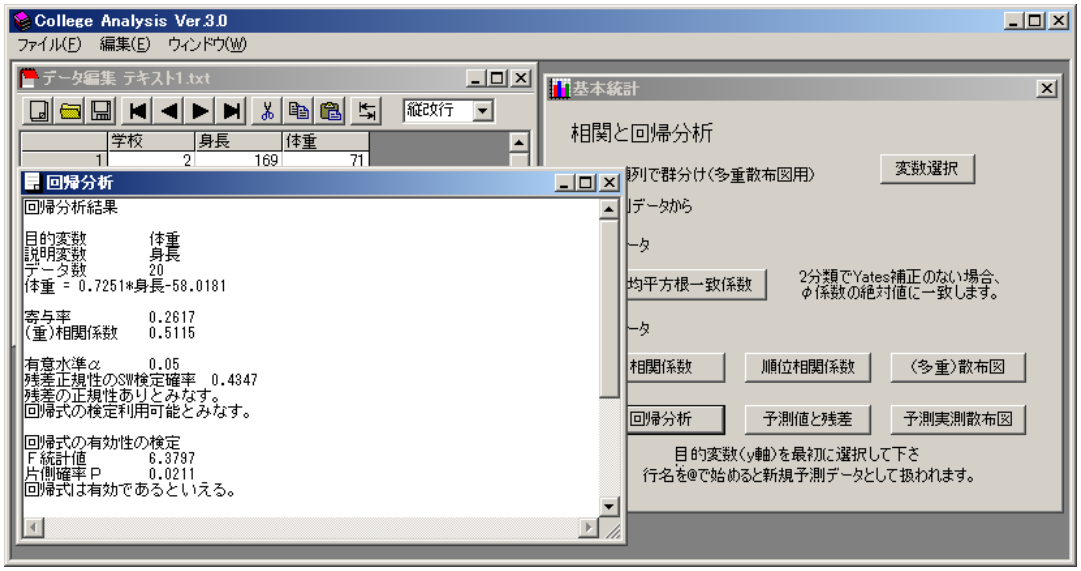


図 1.2.13 回帰分析結果表示画面

回帰式は上のほうに出ています。その下は後に述べる検定の結果です。

以上で例題は終わりです。以下の問題を解いてみて下さい。

問題２

Samples¥テキスト 9.txt を用いて以下の問いに答え、結果は文書にまとめよ。但し、地域について 1：市街、2：郊外とする。

１）年収に関する基本統計量を求めよ。

データ数	最小値	最大値	平均値	中央値	不偏分散	標準偏差

２）地域別の年収に関する基本統計量を求めよ。

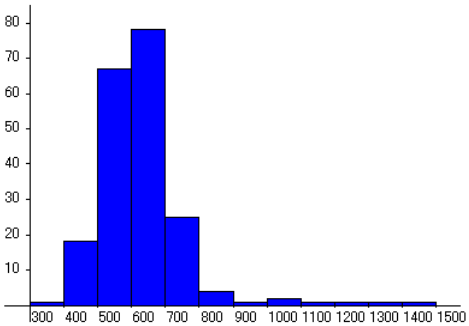
	データ数	最小値	最大値	平均値	中央値	不偏分散	標準偏差
市街							
郊外							

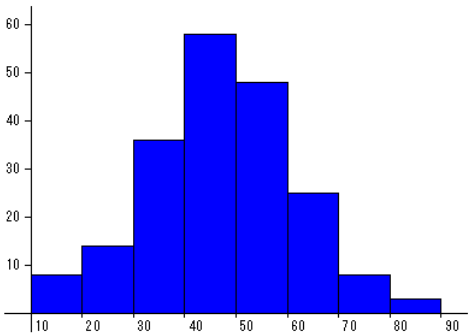
３）年収に関する度数分布表（累積度数・累積相対度数は省略）を描け。

年収	度数	相対度数(%)
300<=x<400		
400<=x<500		
500<=x<600		
600<=x<700		
700<=x<800		
800<=x<900		
900<=x<1000		
1000<=x<1100		
1100<=x<1200		
1200<=x<1300		
1300<=x<1400		
1400<=x<1500		

４）年収に関するヒストグラムを描け。（下図左）

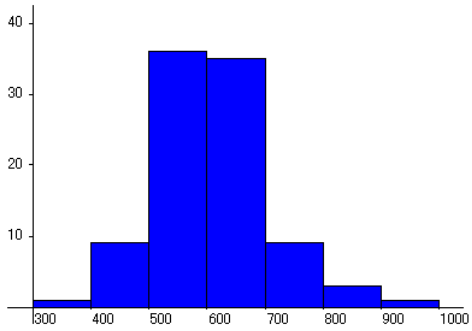
５）支出に関するヒストグラムを描け。（下図右）

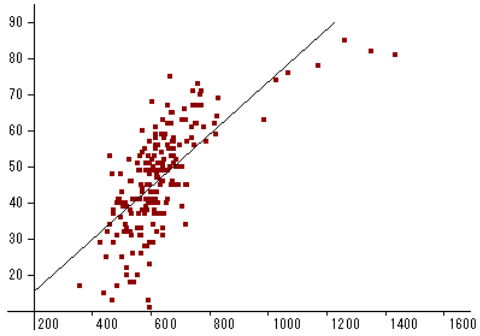




６）地域:1 の年収に関するヒストグラムを描け。（下図左）

７）年収と支出に関する散布図を描け（支出を縦軸，下図右)。





８）年収と支出に関する相関係数を求めよ。

相関係数［　　］

９）支出を目的変数に年収を説明変数としたときの回帰式を求めよ。

支出＝［　　］×年収＋［　　］

1.3 欠損値の除去

アンケート調査などを実施すると、問題の中には回答されていないものが出てきます。これを欠損値と言い、統計処理では取り除いて処理する必要があります。ここではこの欠損値の取り除き方について見てみます。

以下の例について見て下さい。これにはあるテストを行った場合のデータで、受けた生徒の番号、学校（番号表示)、受験科目の点数が出ています。このデータに対していくつかの処理を考えてみましょう。

例

番号	学校	国語	数学
1	1	76	82
2	2		63
3	1	62	58
4		73	74
5	2	81	
6	2	73	65
7	1		46

まず国語の平均を考えてみます。この処理で利用されるのは何番の人でしょうか。通常考えられるのは 1, 3, 4, 5, 6 です。この処理では他の科目の点数が何であれ、国語を受験している人の点数が問題です。このように、他の項目を見ずにある項目だけで

取り除く人を決める方法をデータ単位の欠損値の除去と言います。我々は後の問題用にこれを①番の方法としましょう。

次に学校別に国語の平均を求める場合は、何番の人を利用するでしょうか。これには国語の点数の他に分類に使う学校の番号も必要です。そこで利用される人は 1, 3, 5, 6 になります。国語についてはデータ単位の除去ですが、分類変数が必要ですので、このような除去の方法を（分類変数を除いて）データ単位の除去と呼びましょう。我々はこれを②番の方法とします。

最後に国語と数学についての散布図を描く場合はどうでしょうか。この場合は国語と数学両方のデータが必要ですので、利用される人は 1, 3, 4, 6 となります。このように選んだ変数すべてのデータがそろっていない場合に除去する方法を（選択変数について）レコード単位の除去と言います。ここにレコードとはデータベースで使う言葉で、横 1 列のデータ、即ちここでは個人のデータのことです。我々はこの方法を③番の方法とします。場合によっては②番の方法と③番の方法が重なる場合もありますが、これについては問題の中で話します。

さて教科書の中には、各変数ごとに欠損値を取り除く方法をデータ単位の除去、1 つのレコードに 1 つでも欠損値があれば取り除く方法をレコード単位の除去とあっさりと書いてある場合がありますが、ここではもう少し現実的に分類しています。

今の話をまとめると以下の表のようになります。

表 1.3.1 集計内容と欠損値の除去方法

集計内容	利用データ	欠損値の除去方法
国語の平均	1,3,4,5,6	①データ単位の除去
学校ごとの国語の平均	1,3,5,6	②（分類変数を除いて）データ単位の除去
国語の数学の散布図	1,3,4,6	③（選択変数について）レコード単位の除去

次の問題は集計の復習ですが、欠損値のある場合の一般的な処理の方法も選択するようになっています。College Analysis では変数選択の Window の中で下図のようにこの処理の方法が選べるようになっていますが、通常は自動にしておけば上のような処理をしてくれます。

図 1.3.1 欠損値の除去

問題３

欠損値を含む Samples¥テキスト 9b.txt を用いて、以下の問いに答え、よく使われる欠損値の除去方法について、上の①、②、③のどれに一番近いか右の［　］に答えよ。

１）意見１に関する１次元分割表を描け。　［　　］

意見１:1	意見１:2	合計

２）意見１と意見２に関する２次元分割表を描け。　［　　］

	意見２:1	意見２:2	意見２:3	合計
意見１:1				
意見１:2				
合計				

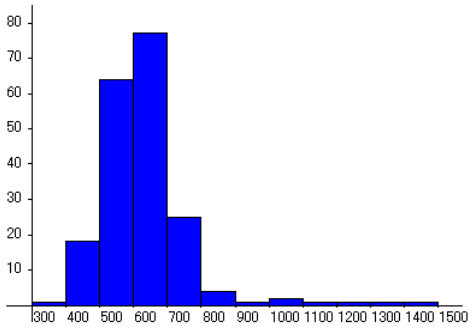
３）年収と支出に関する以下の基本統計量を求めよ。　［　　］

	最小値	最大値	平均値	中央値	標準偏差
年収					
支出					

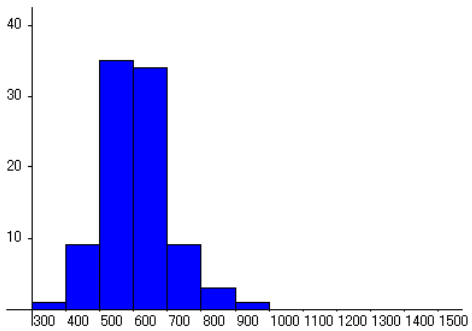
４）地域別の年収に関する基本統計量を求めよ。　［　　］

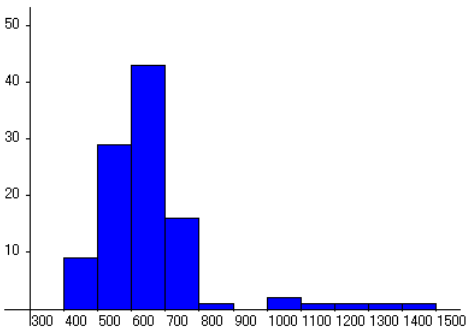
	最小値	最大値	平均値	中央値	標準偏差
地域:1					
地域:2					

5）年収に関するヒストグラムを描け。　　［　　］

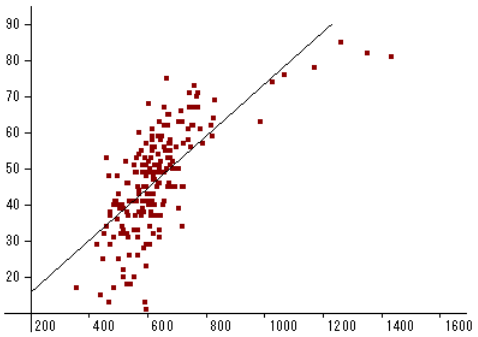


6）地域 1, 2 の年収に関するヒストグラム　　［　　］





7）年収と支出に関する散布図を描け。　　［　　］



8）年収と支出に関する相関係数を求めよ。　　［　　］

相関係数＝［　　］

9）支出を年収で予測する回帰式を求めよ。　　［　　］

支出＝［　　］×年収＋［　　］