1.2 量的データの集計　【動画】

1.2.1 量的データの集計

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。

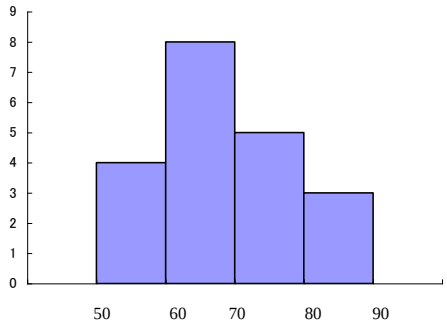

図 1.2.1 ヒストグラム

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。

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。

標準偏差＝$\sqrt{分散}$ ＝2.098

標準偏差＝$\sqrt{不偏分散}$ ＝2.345

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。サンプルは College Analysis の本体 CAnalysis.exe のあるフォルダーの下の Samples フォルダーに保存してあります。

ファイルを選択すると、質的データの集計でみた画面になります。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。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以下のデータが現れます。

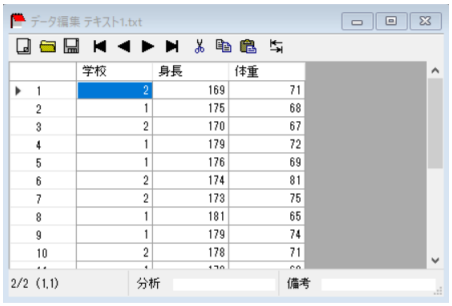

図 1.2.2　ファイル読込画面

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、以下のような量的データの集計実行画面を表示します。

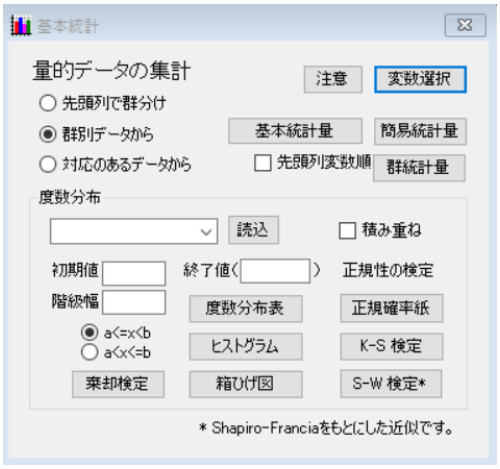

図 1.2.3　量的データの集計実行画面

「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。

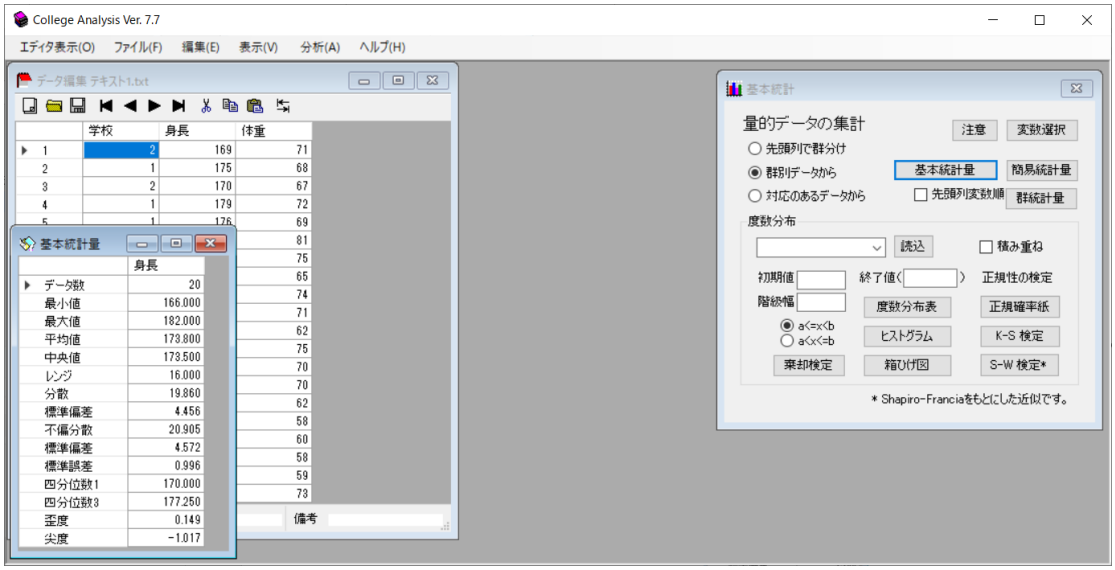

図 1.2.4　基本統計量表示画面

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、どちらでもよいのですが、我々は下側の不偏分散と標準偏差を使うことにします。

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。

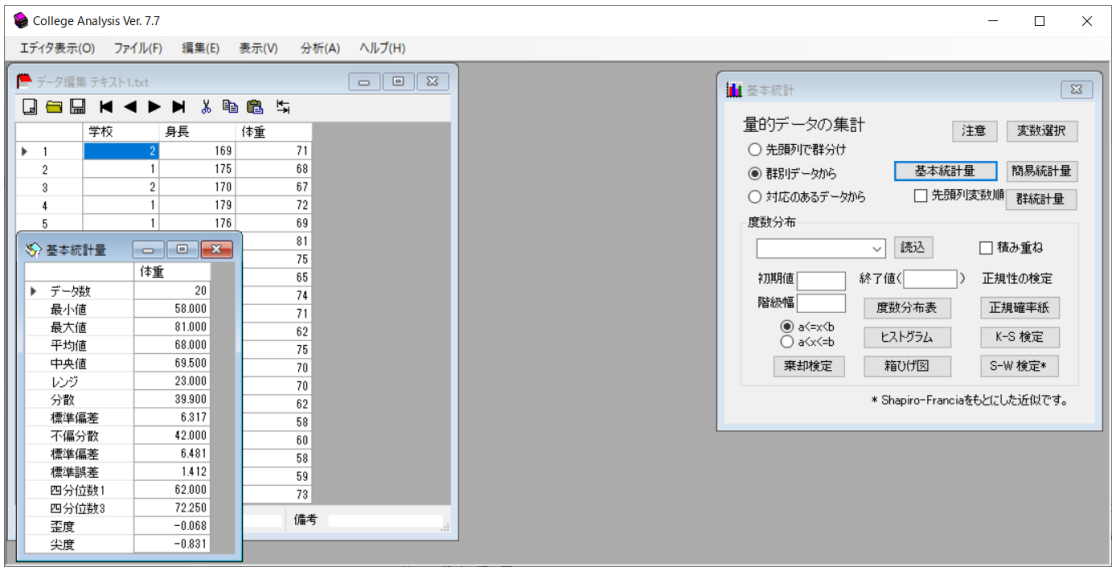

図 1.2.5　体重の基本統計量出力画面

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。

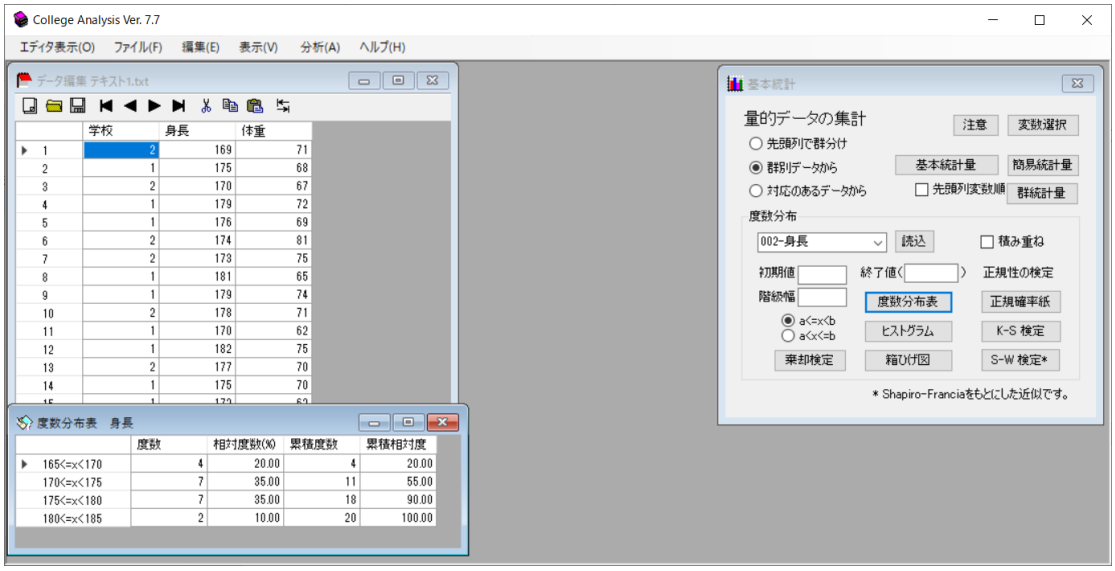

図 1.2.6　身長の度数分布表表示画面

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。

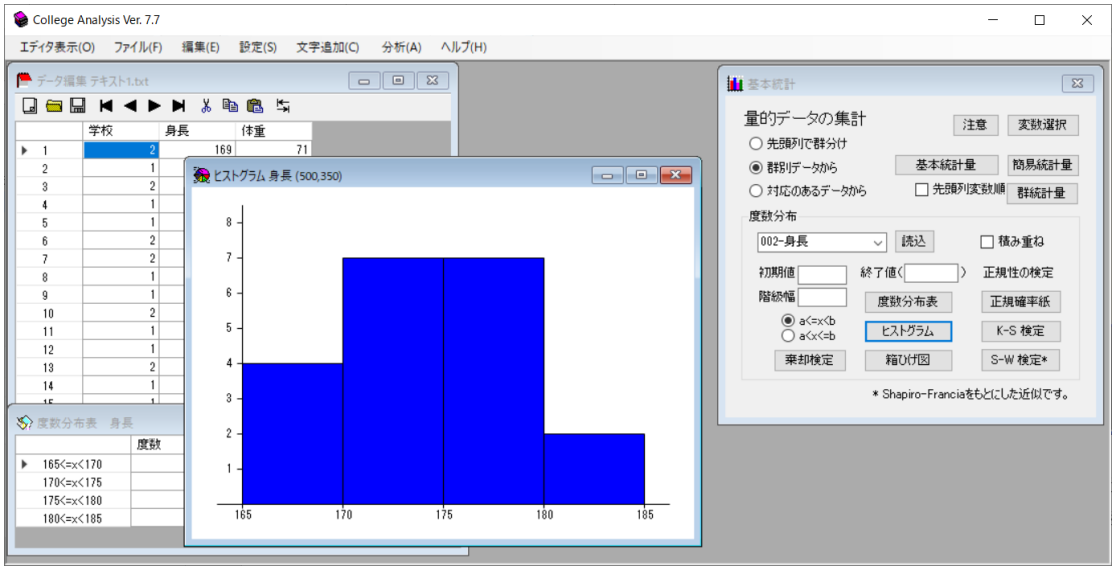

図 1.2.7　身長のヒストグラム表示画面

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。

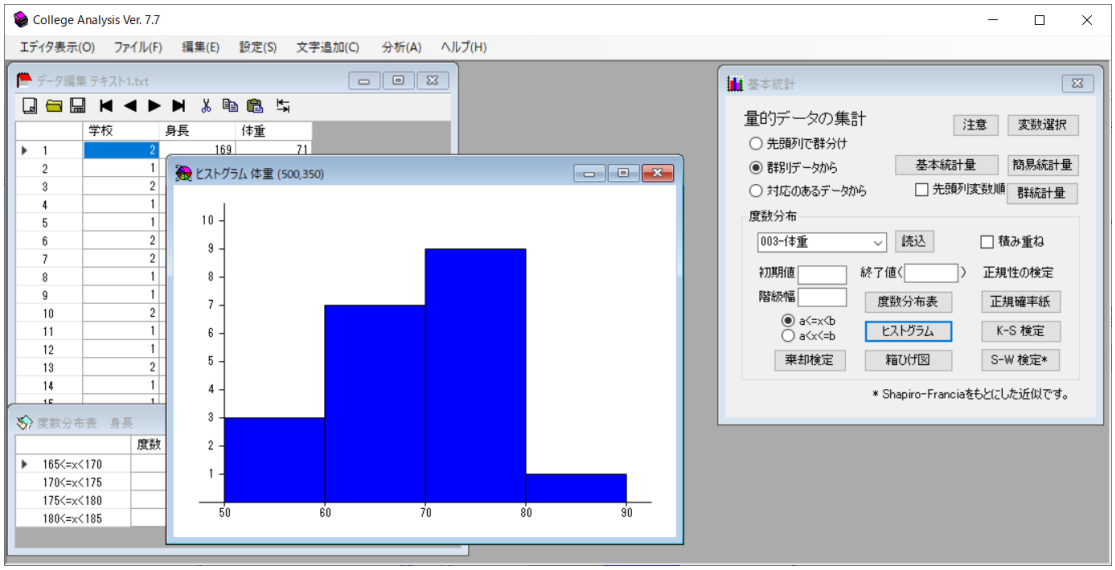

図 1.2.8　体重のヒストグラム表示画面

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。

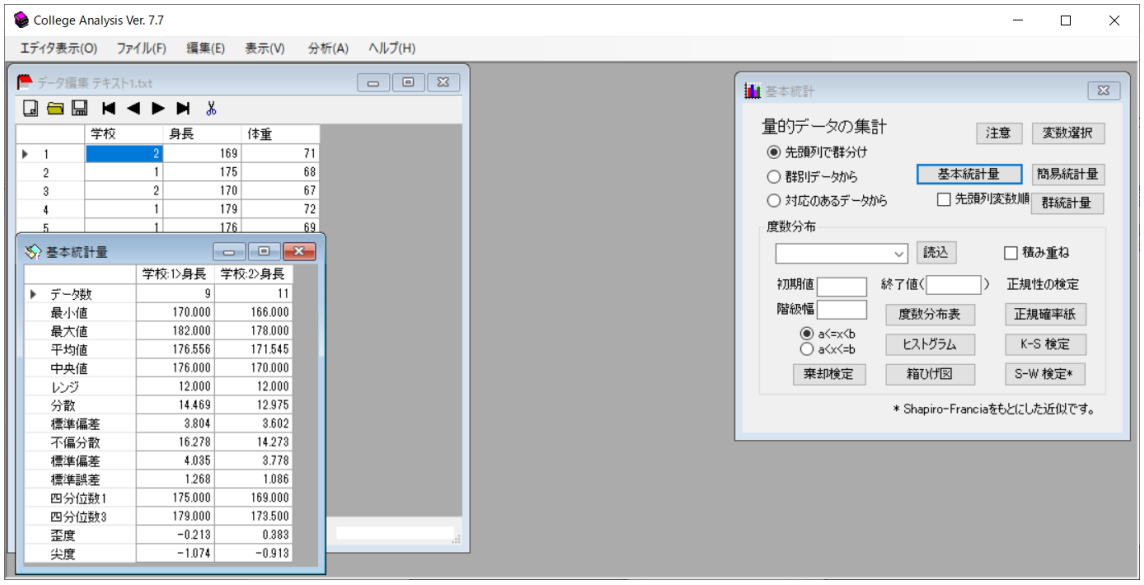

図 1.2.9　学校別に群分けされた身長の基本統計量

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。

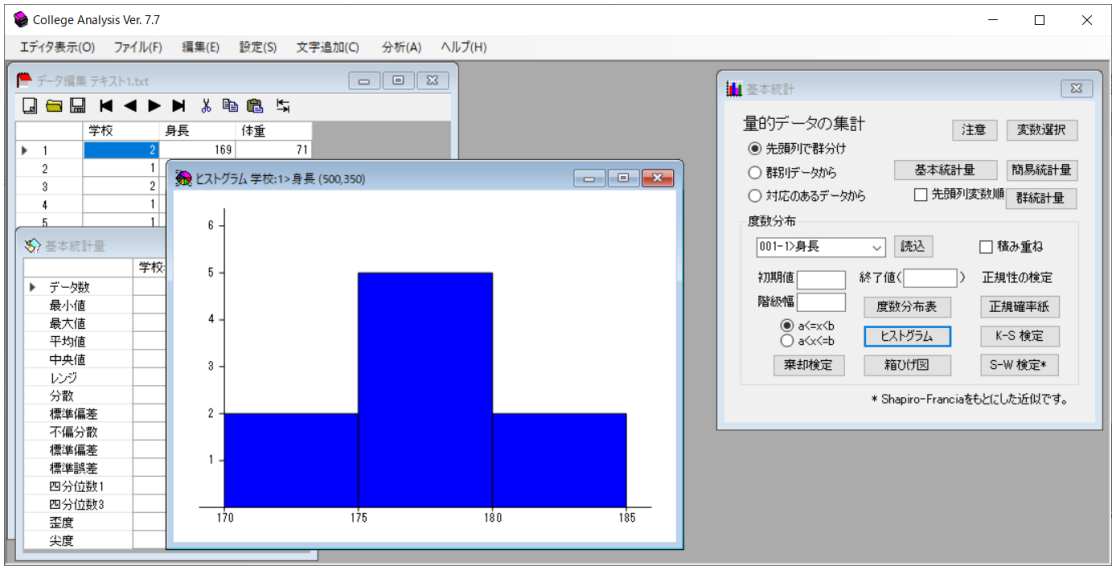

図 1.2.10　学校 1 の身長のヒストグラム