４章 母集団と指定値との量的データの検定

4.1 検定手順

今までは質的データの検定の方法を学んで来ましたが、これからは量的データについてよく利用される方法を説明します。量的データでは、データの分布が正規分布か否かで検定の方法が著しく異なります。この章ではまずデータの分布の正規性を調べる方法を述べ、次にデータの平均値または中央値がある指定された値と違うかどうかの検定方法を説明します。以下の図 4.1.1 を見て下さい。

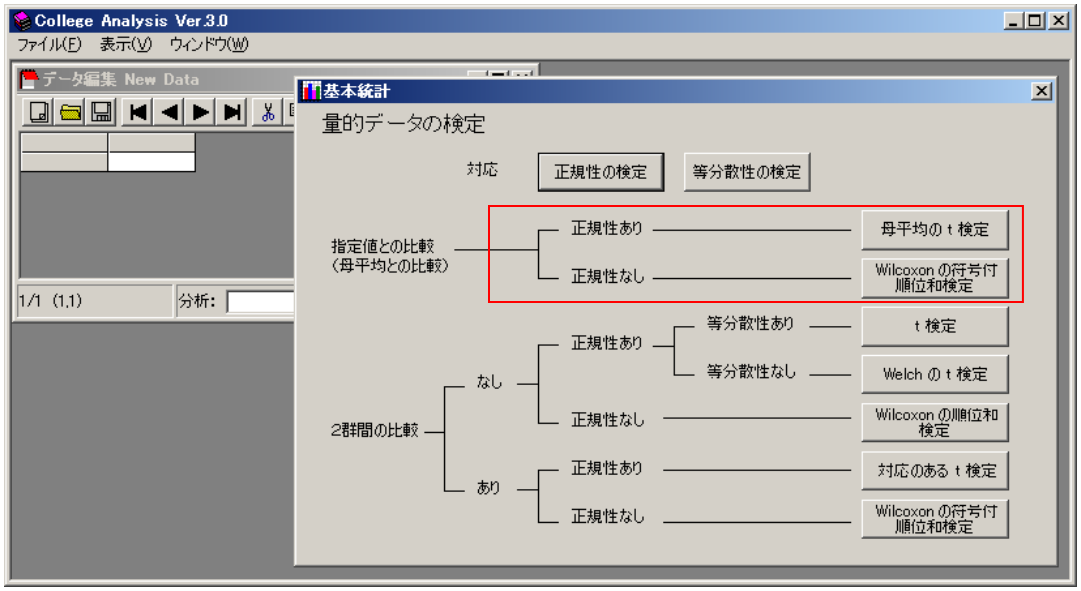


図 4.1.1 母集団の統計量と指定値との量的データの比較検定

これは前に示した量的データの検定選択ツリーですが、この章では赤い四角で囲まれた部分について利用法を学習します。最終的な検定方法の名前は母平均の t 検定及び Wilcoxon の符号付順位和検定といい、どちらを利用するかは正規性の有無によって決められます。

4.2 正規性の検定

最初は正規性を調べる方法についてです。これまではヒストグラムがきれいな富士山型をしている場合は正規分布と言ってきましたが、正規分布はデータ数が多くなければきれいな形になりません。データが少ない場合はどうするのでしょうか。また、

きれいな形といっても個人が感じることですから、個人差があります。この差をなくすにはどうするのでしょうか。

正規性を調べる方法には大きく分けて視覚的な方法と数値的な方法の 2 種類があります。視覚的な方法では、データ数が多い場合にはヒストグラムを用いる方法と正規確率紙による方法があります。正規確率紙は古くから売られていたグラフ用紙で、ある手順に沿ってデータをプロットして行くと、正規分布と思われる場合はその点が直線に近く並ぶというものです。直線に並ぶ場合は曲線と違ってかなりはっきりと直線からのずれを認識することができます。この方法はデータ数が多くても使えますし、かなり優れた方法です。

しかし、やはり人間の個人的な感性の違いから人によって異なった結果になる可能性が残っています。これをできるだけなくすために、数値的な方法も考えられています。よく利用されるのは Kolmogorov-Smirnov 検定（略して K-S 検定）と Shapiro-Wilk 検定（略して S-W 検定）ですが、一般に S-W 検定の方が正規分布との違いを見つけ出し易くよく利用されているようです。しかし K-S 検定はデータ数が数千を超える場合は良い結果を与えるとされています。私のソフトにはきちんとした S-W 検定がなく、Shapiro-Francia 検定を元にした近似計算になっています。実用上問題はないと思いますが、利用する場合はその点だけ念頭に置いて下さい。それでは具体的に視覚的方法と数値的方法の 2 つの視点から、正規性を調べてみましょう。

例

以下のデータの正規性を調べよ。

2.5，2.1，3.4，2.8，4.6，3.2，3.8，4.8，4.0

まずファイル Samples¥テキスト 1 .txt を開きます。メニュー［分析－基本統計－量的データの集計］を選択し、変数選択でデータ 1 を選んだ画面が以下の図です。

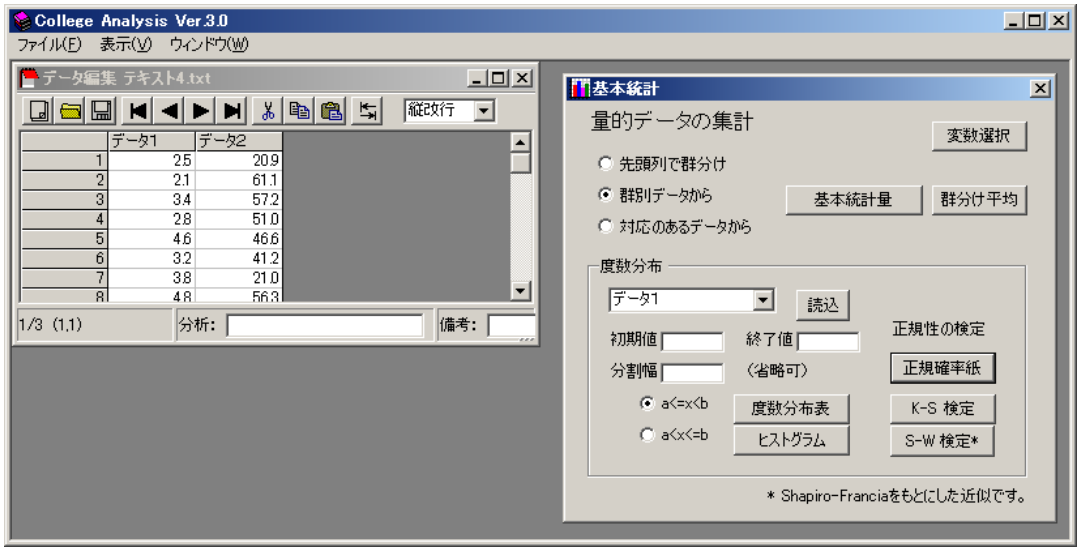


図 4.2.1 正規性の検定メニュー

これは一度利用した分析メニューですが、ヒストグラムの他に正規確率紙や S-W 検定などの処理も含まれています。このデータは数が少なくヒストグラムで正規性を示すことができませんので、「読込」ボタンを押した後「正規確率紙」ボタンをクリックします。結果は以下のように示されます。

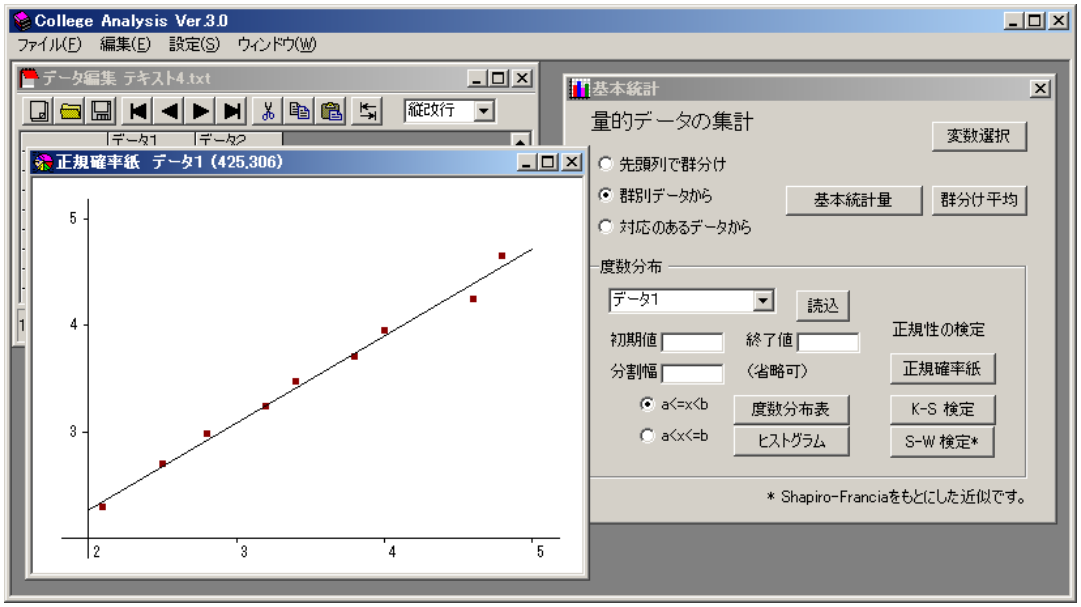


図 4.2.2 正規確率紙表示画面

これを見るとプロットがほぼ直線状に並んでいますので、データは正規分布してい

るものと考えられます。また「S-W 検定」ボタンをクリックすると以下の結果が表示されます。

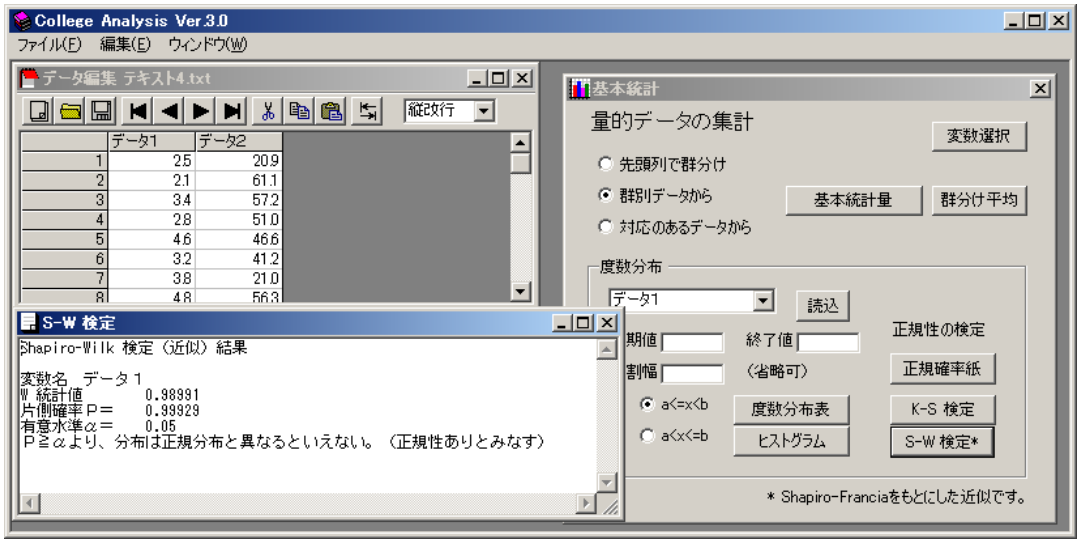


図 4.2.3 Shapiro-Wilk 検定（近似）表示画面

これは帰無仮説としてデータに正規性がある、対立仮説としてデータに正規性がないとする検定で、この結果によると帰無仮説が採択されます。

正規性の検定では、正規分布でないということは言えますが、正規分布であることは「正規分布でないといえない」という弱い言い方しかできません。それはデータ数が増えると差を見出し易くなって「正規分布でない」という結論になってしまうかも知れないからです。ただ実際の検定の場面では、これを正規分布と考えて処理を行うこともあるようで、我々のソフトではこういったニュアンスを込めて、「正規性ありとみなす」という表現にしています。

問題１

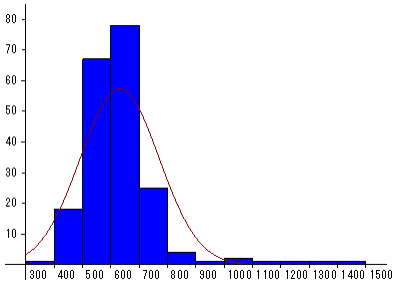
以下のデータの正規性を調べよ。

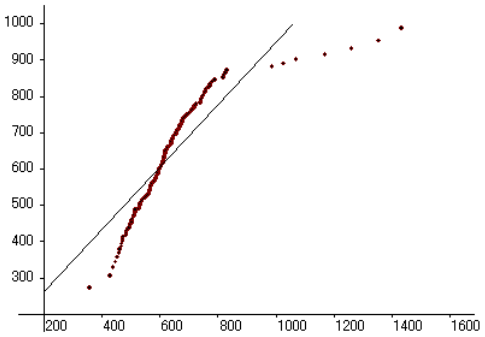
20.9, 61.1, 57.2, 51.0, 46.6, 41.2, 21.0, 56.3, 49.5, 49.3, 22.4, 23.5

問題２

Samples¥テキスト 9.txt のデータについて以下の問いに答えよ。但し、ヒストグラムについては、密度関数から得た理想的な正規分布の形を加えること。

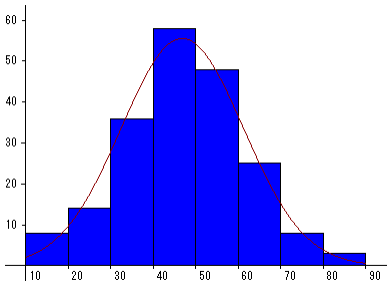
１）年収のデータの正規性をヒストグラム、正規確率紙、S-W 検定で調べよ。

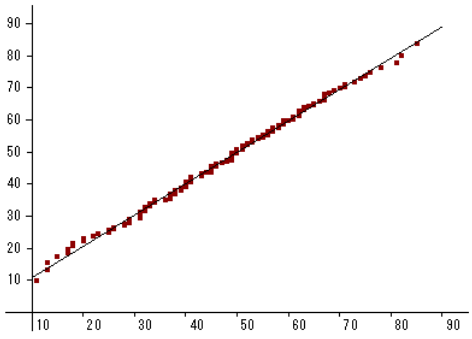




S-W 検定　　確率 [　　] 判定　正規分布と［みなす・いえない・判定困難］。

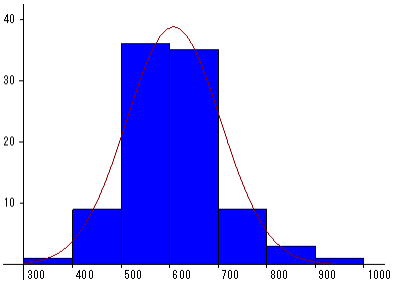
２）支出のデータの正規性をヒストグラム、正規確率紙、S-W 検定で調べよ。

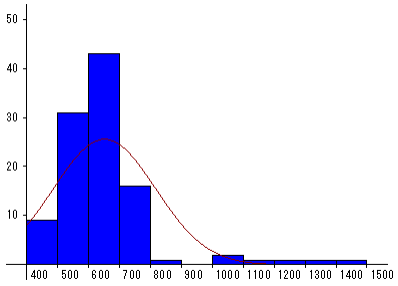




S-W 検定　　確率［　　］判定　正規分布と［みなす・いえない・判定困難］。

３）地域別に年収のデータの正規性を調べよ。





地域１　　確率［　　］判定　正規分布と［みなす・いえない・判定困難］。

地域２　　確率［　　］判定　正規分布と［みなす・いえない・判定困難］。

4.3 母集団の平均値と指定値との比較（正規性あり）

正規性を調べる方法が分かりましたので、次はデータに正規性があった場合の具体的な検定の方法についてです。以下の例を見て下さい。

例

ある地域のある規模の会社 9 社について 1 人当り売上高のデータを集めたら、正規分布し、平均 2410（万円）、不偏分散から求めた標準偏差 150（万円）であった。この地域の会社の 1 人当り売上高は日本の同じ規模の会社の 1 人当り平均売上高 2260（万円）に比べて差があるといえるか？有意水準 5%で判定せよ。

この問題は量的データの検定問題ですから、メニューから［分析－基本統計－量的

データの検定－量的データ検定メニュー］を選択します。

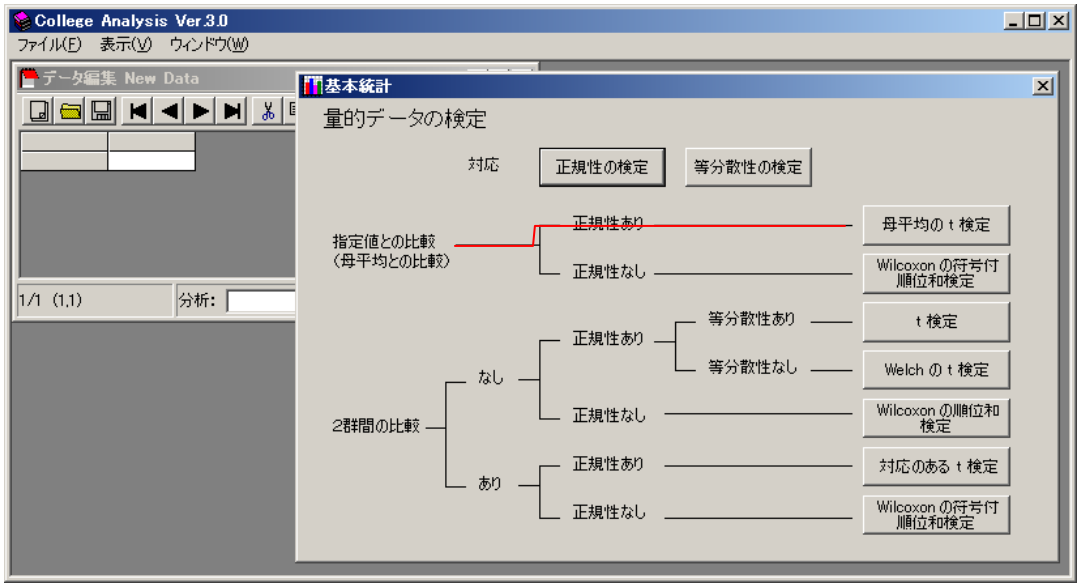


図　量的データの検定メニュー

この例題では標本を取り出した母集団の母平均（中央値）と指定値 2260（万円）との比較ですから、指定値との比較のラインをたどることになります。正規性の検定では、問題の仮定より正規分布することが分かっているので、検定名は母平均の t 検定となります。これをクリックすると以下のようなメニューが現れます。

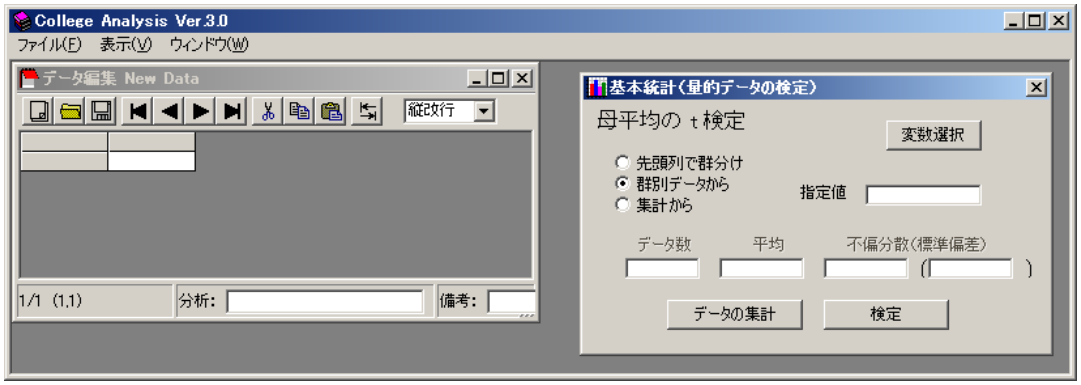


図　母平均の t 検定画面

この問題は集計結果が分かっているので、ラジオボックスは「集計から」を選択し、

集計されたデータを以下の図のように入力します。

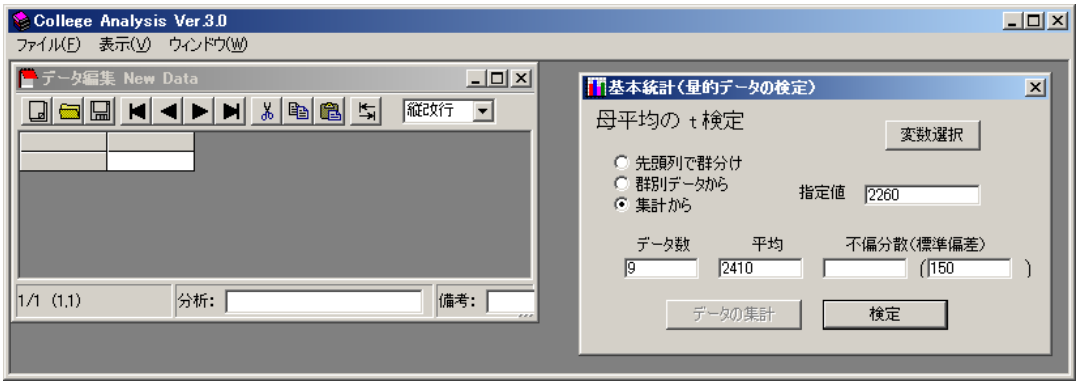


図 母平均の t 検定集計入力画面

ここで「検定」ボタンを押すと以下のような出力結果を得ます。

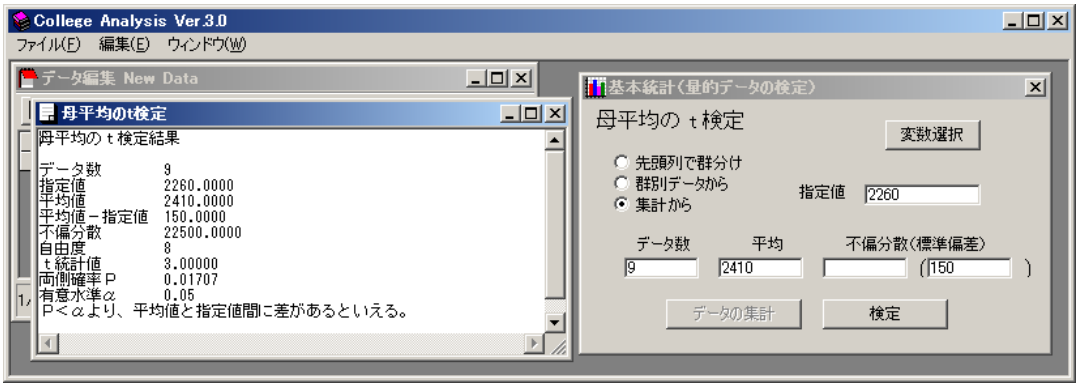


図 母平均の t 検定結果表示画面

この結果から検定確率 $p = 0.0171 < 0.05$ となり、平均値と指定値間に差があるといえるということになります。

最後にこの計算の基になっている理論式を示しておきましょう。

理論　母平均の t 検定

指定値と比べて平均に差がないとして、

$$t = \frac{\sqrt{n}(\bar{x} - \mu)}{u} \sim t_{n-1} \text{分布}$$

4.4 母集団の中央値と指定値との比較（正規性なし）

例

ある地域のある規模の会社の 1 人当り売上高（万円）は以下の通りである。これらの会社は同じ規模の会社の中央値 2260（万円）に比べて売上高に差があるといえるか。有意水準 5%で判定せよ。

2060, 2350, 1550, 1720, 1800, 1990, 1510, 1720, 2910, 1820, 2600

この問題は正規性を仮定しない問題です。ファイル Samples¥テキスト 4.txt を開き、前に述べた量的データの検定メニューから、以下の道筋を通ります。

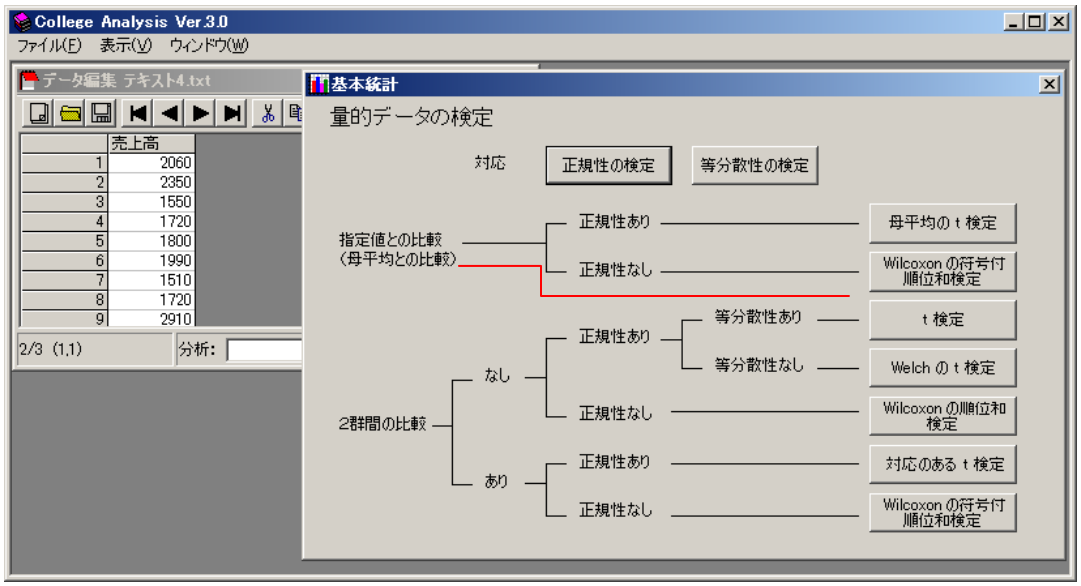


図 量的データの検定メニュー

これに従って、「Wilcoxon の符号付順位和検定」ボタンをクリックすると以下のような検定画面が表示されます。

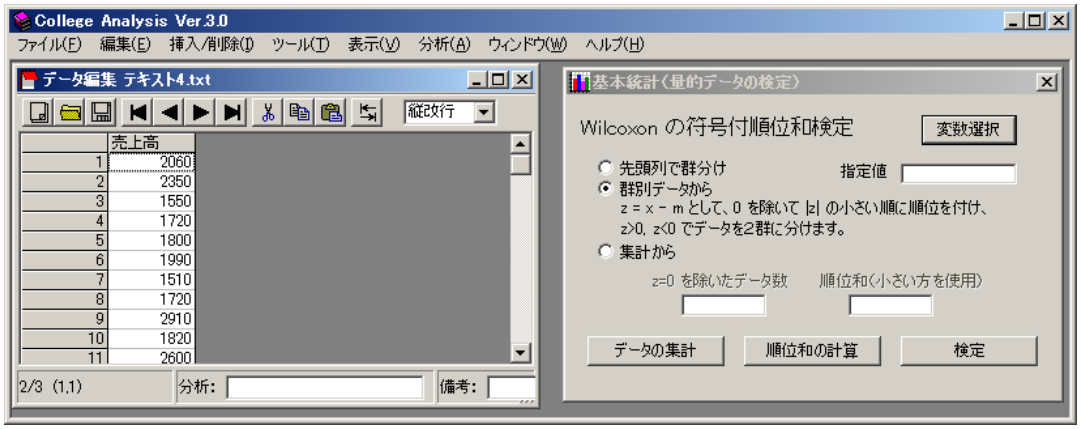


図 Wilcoxon の符号付順位和検定画面

ここで、変数選択で「売上高」を選択し、ラジオボタンは変数が 1 つなので「群別データから」指定値に 2260 を入れて「検定」ボタンをクリックします。結果は以下のように差は見出されませんでした。

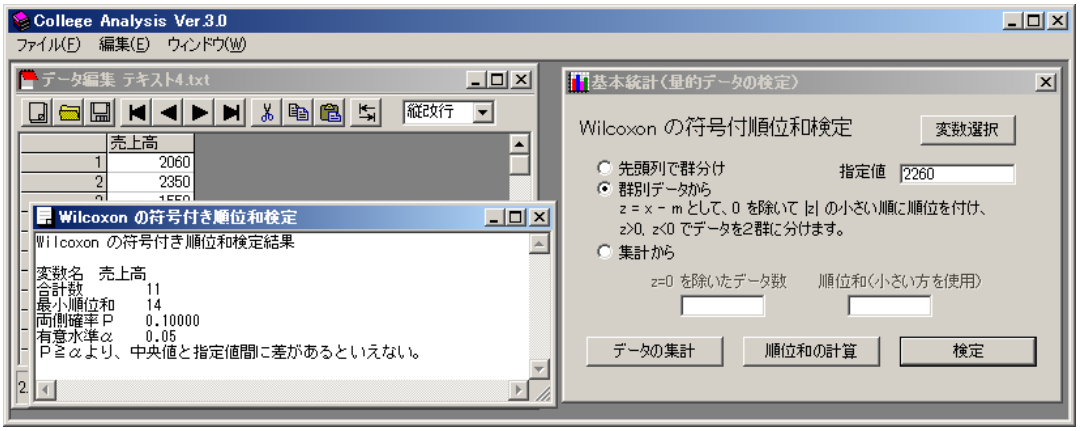


図 Wilcoxon の符号付き順位和検定結果画面

さて、Wilcoxon の符号付き順位和検定とはどんなことを使って検定をしているのでしょうか。今、8 個ずつ 3 種類のデータを用意し、データの値－指定値を横軸にして、3 本のライン上に各データをプロットしてみます。図を見て下さい。

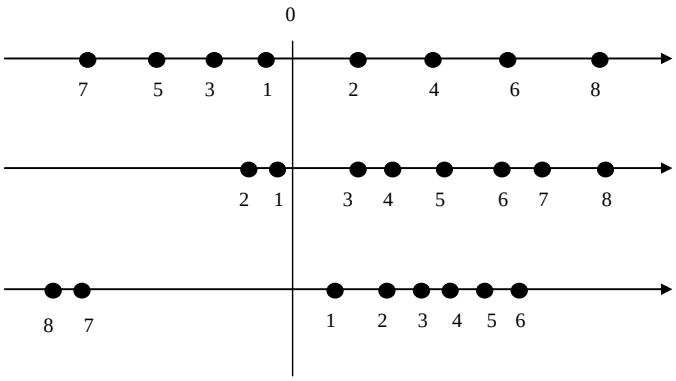


図　検定概念図

この 3 種類のデータのうち、平均から最もずれているのはどれでしょうか。一番上

のデータは平均から左右均等に散らばっていますので、これは違います。しかし真ん中のデータは極端に右に偏っていますので、これだと分かります。視覚的にはすぐに分かりますが、数値的には何を使ってずれを判定するのでしょうか。一番下のデータはあまりずれていないように感じますが、平均から右にずれている個数は 2 番目と同じなので、左右の個数ではありません。

このデータに対して左右に関係なく 0 に近いところから順番に番号を付けてやることにします。それが上の図に付いた番号です。この番号を 0 以上と 0 未満のところで合計します。上は、20 と 16、真ん中は 33 と 3、下は 21 と 15 です。真ん中のデータは合計が極端に違います。この番号（0 に近い順位）の和によってデータの偏りをみる検定が Wilcoxon の符号付き順位和検定です。

実際に利用する式は以下です。

理論

左右の順位和を求め、その小さい方を R とする。

標本数が多いとき

$$z=\frac{|R-n(n+1)/4|-1/2}{\sqrt{n(n+1)(2n+1)/24}} \quad \sim \quad N(0,1)$$ 分布（正の部分）（Yates の連続補正）

標本が少ないとき

数表を利用

問題３

以下のデータの平均値（中央値）は 5.5 と比べて差があるといえるか。検定を選んで有意水準 5%で判定せよ。

8.4, 4.6, 5.2, 6.3, 7.2, 5.8, 6.0, 5.4, 4.9, 6.9

正規性の判定

ヒストグラムにはデータ不足。正規確率紙を描く。S-W 検定　確率［　　］

判定　正規分布と［みなす・いえない］。

検定名［　　］　確率［　　］

判定　5.5 と比べて差があると［いえる・いえない］。

４章問題４

Samples¥テキスト 9.txt のデータを用いて以下の問いに答えよ。

1）年収の平均値（中央値）は 610 万円より多いといえるか。分析を選んで有意水準 5%で判定せよ。

正規性の判定

ヒストグラムを描く。正規確率紙を描く。S-W 検定　確率［　　］

判定　正規分布と［みなす・いえない］。

検定名［　　］　確率［　　］

判定　610 万円より多いと［いえる・いえない］。

2）支出の平均値（中央値）は 44 万円より多いといえるか。分析を選んで有意水準 5%で判定せよ。

正規性の判定

ヒストグラムを描く。正規確率紙を描く。S-W 検定　確率［　　］

判定　正規分布と［みなす・いえない］。

検定名［　　］　確率［　　］

判定　44 万円より多いと［いえる・いえない］。