# ヒストグラムと等高線図

- データ解析の一つの目的に、データ源の確率分布を求めることがある
- ヒストグラムは１次元確率分布を推定する一番わかりやすい方法
- ヒストグラムで重要なのは区切りの幅。これ次第で結果が変わる
- Excel では標準アドインソフト（分析ツール）を使うと簡単にヒストグラムが作成できる

## 1. 分析ツールを使えるようにする

Excel では、「分析ツール」とよばれる追加ソフトを組み込まないとヒストグラム作成を行えない[1]。分析ツールが組み込まれているかどうかの確認および組み込み方法は以下の通り。

### 1.1 分析ツールが組み込まれているかどうかの確認

分析ツールは、[ツール]メニューから呼び出すことができるが、Excel のインストール直後の状態では利用可能になっていない。利用可能になっていない時は[ツール]メニューを開いた時には図 1 のような[2]メニューが表示される。このようなメニューが開く状態では分析ツールを使えないため、1.2 の手続きに従って分析ツールを有効にする必要がある。

---

[1] この作業は、通常の環境なら 1 回やればよい。ただし、情報処理センターの実習用環境では定期的にシステムが初期化されてしまうため必要に応じて繰り返すこと。

[2] 「のような」というのは、Excel のバージョン違いや他のアドインをインストールするとまた違う項目が表示されたりするから。図１は Excel97 のインストール直後でのメニュー表示。

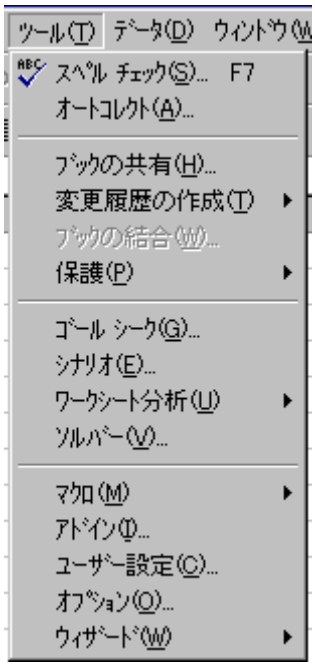



**図 1　「分析ツール」が使えない状態の[ツール]メニュー**

## 1.2 アドインメニューから分析ツールを有効にする

分析ツールは Excel のアドイン(add-in)ソフトウェアとして Excel の一部として提供されているが、初期状態では有効になっていない。これを有効にするためには、まず[ツール]-[アドイン]メニューを選ぶ（図 2)。


**図 2 [ツール]-[アドイン]**

すると、図 3 のようなアドイン管理メニューが開く。チェックマークが付いているアドインは Excel に自動的に読み込まれて利用可能になる。初期状態では図 3 のように、分析ツールと分析ツール-VBA 関数にはチェックが入っていないため、Excelから利用できるようにはなっていない。

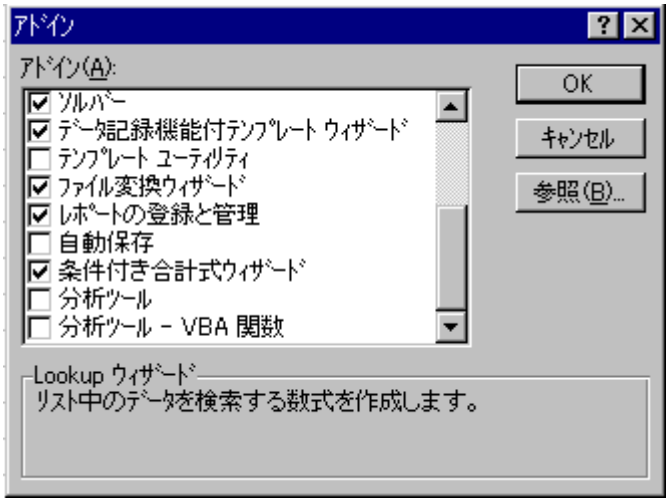



**図 3　アドイン管理メニュー**

そこで､図 4 のようにこの２つにチェックを入れて､[OK]ボタンを押してExcelにもどる。これでアドインが読み込まれているはずである。

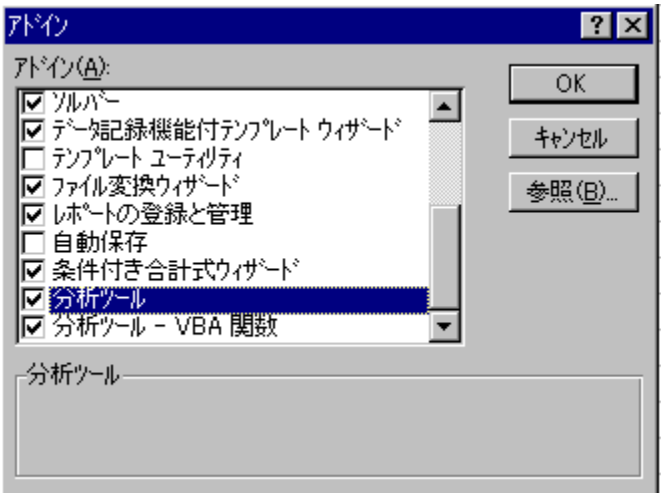



**図 4　分析ツールを有効にする**

本当に「分析ツール」が利用可能になっているかどうか、もう一度[ツール]メニューを開いてみると、確かに「分析ツール」がメニューに追加されている(図 5)。


**図 5　分析ツールが入った[ツール]メニュー**

## 2.　ヒストグラム作成ツールの使い方

### 2.1 ヒストグラム作成ツールの起動方法

ヒストグラム作成ツールを起動するためには、[ツール]-[分析ツール]で表示されるデータ分析メニュー（図 6）から「ヒストグラム」を選択して OK を押せばよい。すると、図 7 のようなメニューが開きヒストグラムが作成できる。メニュー内の項目の意味については、ヘルプ参照。


**図 6　データ分析メニュー**

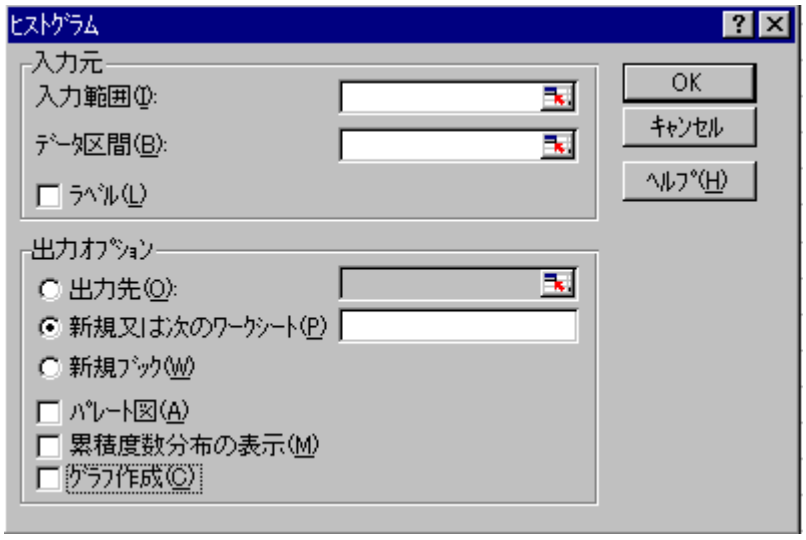



**図 7　ヒストグラム作成メニュー**

## 2.2　　ヒストグラムを作ってみる

Excel のヒストグラムツールでは、データ区間として指定した値は区間の上限（区間に含まれる）を表している。グラフ表示をするとあたかも階級値のように見えるが騙されてはいけない。

そのため、データ区間を$\{k_1, k_2, k_3, k_4, \ldots, k_n\}$と与えた場合には、$k_1$ とラベルが付いている区間は$\#\{ X_i \leqq k_1\}$を表し、$k_2$ ラベルが付いている区間は $\#\{k_1<X_i \leqq k_2\}$を表している。区切りが n 個であるから、全体は n+1 個の区間に分けられるが、最後の区間は「次の級」とラ

ベルがつき、#{ $X_i$ ＞ $k_n$}を示している。この例を 2002-06-17.xls の「ヒストグラム」シートに示す（図 8)。ただし #{ }は条件を満たすデータの個数を表す。

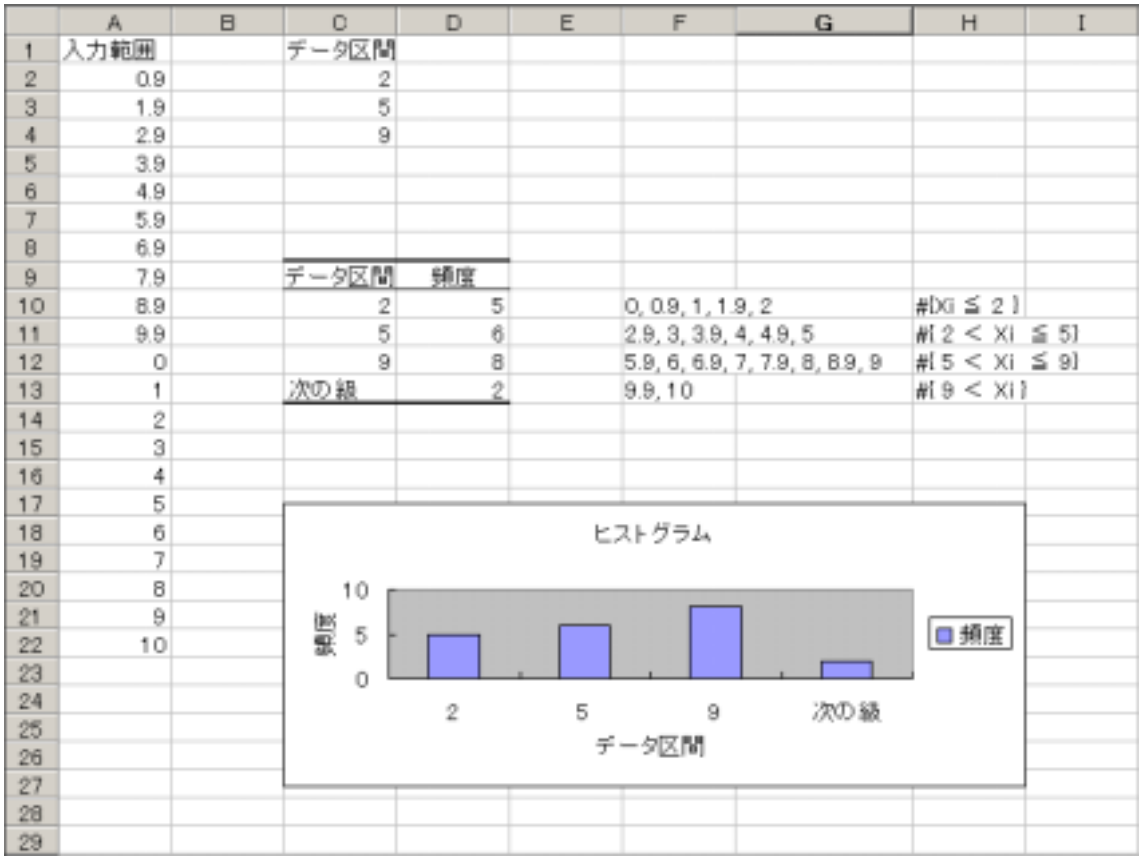



**図 8 ヒストグラムの作成例**

### 2.2.1 区間幅を Excel お任せでヒストグラムを作る

図 8 の例では人間が明示的にデータ区間を指定してヒストグラムを作成したが、図 7 のヒストグラム作成メニューで、入力範囲のみ指定してデータ区間を指定しないと Excel が適当に（区間生成方法はヘルプに書いてないためよくわからない）等間隔の区間を作成してヒストグラムを作成してくれる。

これは一見楽そうなのだが、ヒストグラムを用いて密度関数を近似するときには区間幅（区間数）をどのように決定するかが本質的な問題であるため、実はあまりお勧めできない。

## 3. ヒストグラムにおける区間幅（区間数）の意味

2002-06-17.xls の「所定内賃金」データは、賃金が対数正規分布すると仮定して乱数を用いて生成した賃金列である。ただし、平均と分散の異なる 2 系統のデータを 100 個ずつ使って全サンプルの 200 個にしている。ヒストグラムの区間数を変えていくことによって、2

系統を識別できるかどうかを試してみよう。下の図は、区間数をいろいろ変えてヒストグラムを描いたものである。図を見れば明らかであるが、区間数をどのように設定するかによって 2 系統のピークが正しく認識できるかどうかが変わってくる。実際の２系統の分布がどうなっているかは、図 9 に示した通りである。これがヒストグラムから見えてくればよい。

区間数が 5 つ程度まででは、そもそもピークが一つしか見えないため、単調減少しているかのように見える。もちろん、対数正規分布を 2 系統混ぜているのであるから単調減少に見えるのはグラフの作り方が悪い。

区間数が 7 から 11 個程度になると、高い方の分布のピークが徐々に見えてくる。しかしながら低い方の分布のピークは依然として一番下の階級に埋もれたままで、対数正規分布が 2 系統混じっているようには見えてこない。

区間数が 13～17 になると、低い方のピークと一番下の階級が分離されてきてやっと対数正規分布が 2 つ混ざっている感じになってくる。

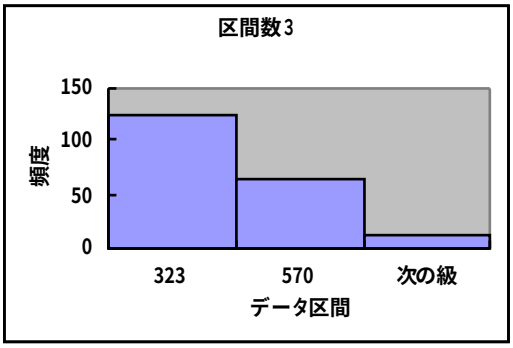
このように、ヒストグラムを作成する際には区間をどのように設定するかが最大の意思決定項目となる。

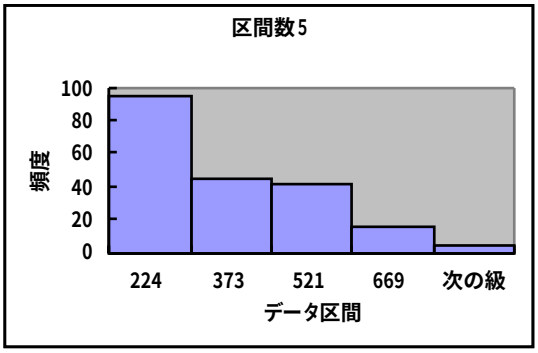
### 3.1 よく使われる区分基準

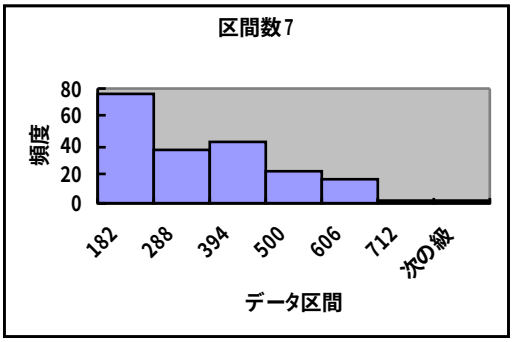
ヒストグラムの区間数（幅）を決定するために使われることのあるスタージェスの公式は、データのサンプル数から区切り個数を決定するものであり、区間数=$1+\log_2 n$ で区間数を決める。今回のデータ(n=200)でこの公式から区間数を求めると 8.64 となる。上述の通り、区間数が 9 個程度では今回のデータには少なすぎる。

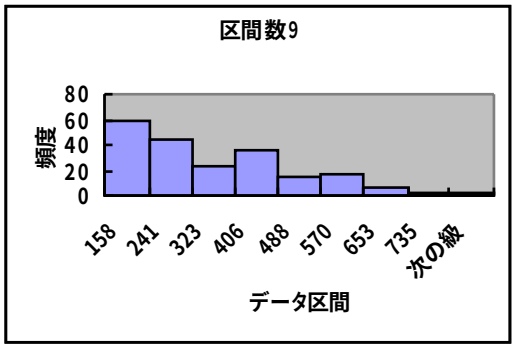
元々の分布が正規分布であると仮定した場合に最適の区間幅を求める公式もある。この公式は、区間幅=$3.491\,\sigma\,n^{-1/3}$ で区間幅を求める。正規分布を仮定しての話なので、ピークが複数あるようなデータに適用した場合はσが過大評価され、結果的に区間幅が大きくなりすぎる傾向を持つ公式である。今回のケースでも区間幅が 93 程度になってしまい、区間数にするとほぼ 8 の場合に該当し、予想通り区間幅は過大である。

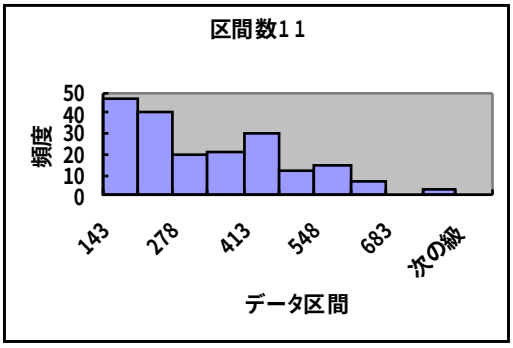
Excel は内部的に何をやっているのか不明であるが、Excel おまかせで区間幅を決めるとなかなか良い感じのグラフになっている。

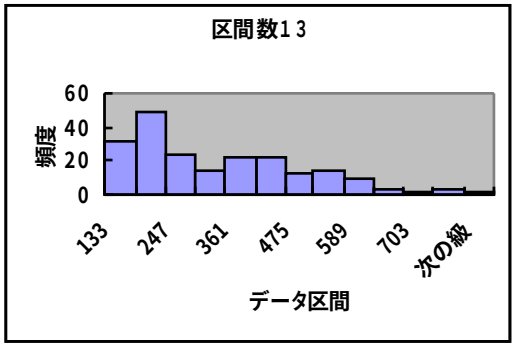

区間数3
150
100
50
0
頻度
323
570
次の級
データ区間

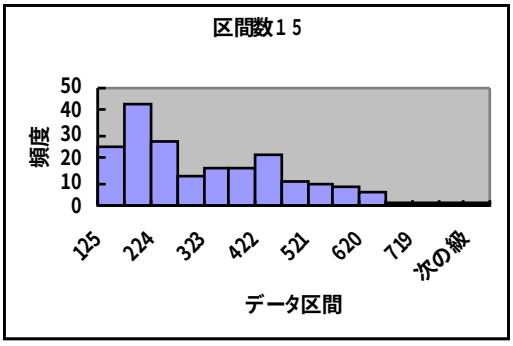

区間数5
100
80
60
40
20
0
頻度
224
373
521
669
次の級
データ区間

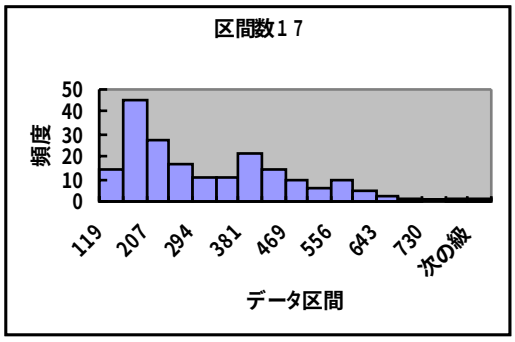

区間数7
80
60
40
20
0
頻度
182
288
394
500
606
712
次の級
データ区間


区間数9
80
60
40
20
0
頻度
158
241
323
406
488
570
653
735
次の級
データ区間


区間数11
50
40
30
20
10
0
頻度
143
278
413
548
683
次の級
データ区間


区間数13
60
40
20
0
頻度
133
247
361
475
589
703
次の級
データ区間


区間数15
50
40
30
20
10
0
頻度
125
224
323
422
521
620
719
次の級
データ区間


区間数17
50
40
30
20
10
0
頻度
119
207
294
381
469
556
643
730
次の級
データ区間

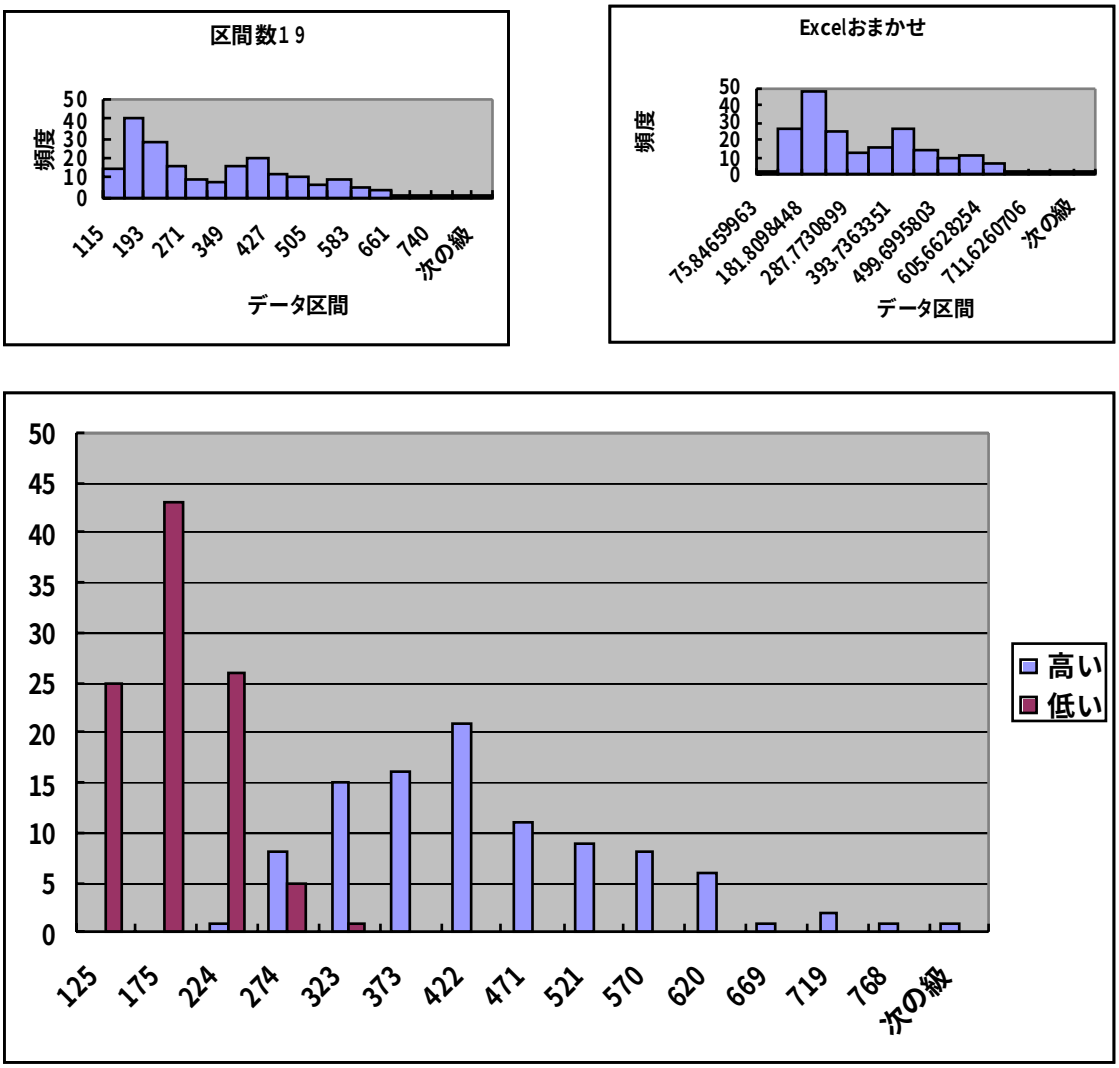



図 9　2つの分布の中身

## 3.2 端点の選択

よりマイナーな問題であるが、ヒストグラムの見栄えに大きく影響する可能性がある要素として端点をどのような値に設定するかが挙げられる。例えば、上のヒストグラム群はExcel おまかせの物を除いてデータの最小値+区間幅を一番小さい区切り値として利用しているため、常に左端の区分にある程度のデータが入っている。一方、Excelおまかせの区分だと最小値を一番小さい区切り値として利用しているため、左端の区分は常にデータ数が１である。

Excel のような方法をとると、データ分布の最小値がヒストグラムの中に明示的に表示されるという利点がある。ただし、一番下の階級がつねに頻度１になっているのは階級構成上

の都合であることを意識している必要はある。

### 練習問題 1

2002-06-17.xls の「練習問題１」シートに入っているデータは、複数の異なる正規分布に従うデータの混合物である。ヒストグラムを適切に使って、(1)いくつの正規分布から得られたデータと考えられるか(2)それぞれの正規分布の平均はどの程度の値になるか、の 2 つを推測しなさい。余力があれば、それぞれの正規分布の標準偏差も推定してみなさい。

## 4. 多変量(２変数)での密度推定

- 散布図は 2 変量の関係を知るためには適しているが、プロットが重なってしまうようなケースでは無力
- Excel では、このようなデータを扱うために、散布図を拡張したバブルグラフと、ヒストグラムを 3 次元に拡張した 3D 縦棒グラフ、さらに等高線グラフが使える

### 4.1 散布図の限界

散布図は非常に有用なツールであるが、データの密度推定には哀しいほど役に立たない。図 10 は 2002-06-17.xls の「2 変量」シートのデータを散布図にプロットしたものであるが、この図からは「X と Y には大した相関がない」といった程度のことしかわからない。これは、点が重なってしまうと見分けがつかなくなるからである（もっとも、図 10 ではデータ点の完全な重なりはほとんどない）。

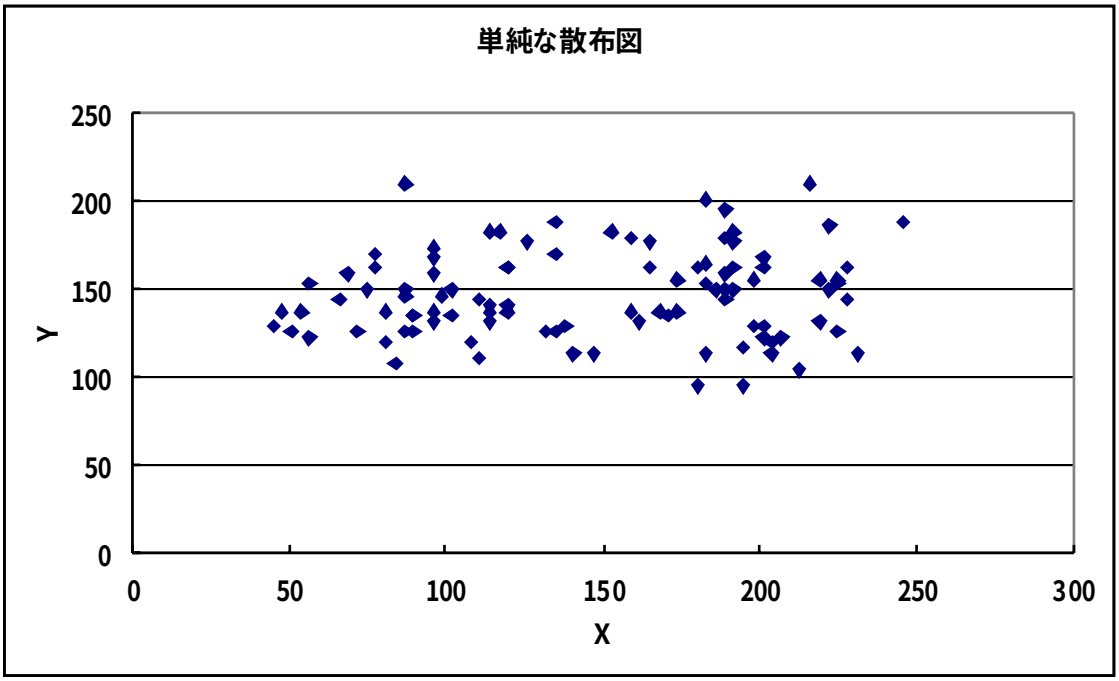


**図 10　データ密度を散布図でみる**

次に、Excel のバブルグラフを使ってみる。これはマーカーサイズを点ごとに指定できる散布図だと思えばよい。元のデータそのままだと厳密な意味での重なりはほとんどないため、区間幅を 20 で区切って、その区切りに入るデータ件数をマーカーサイズにして表示したものが図 11 である。散布図よりは、データの集まっているポイントがなんとなく見えてきて、このデータには山が 2 つありそうに見えてきた。

## 4.2 3D 縦棒グラフを使ってみる

Excel には縦横の集計表形式のデータを 3 次元のグラフにする機能がある。この機能を使って、図 11 と同じデータをグラフ化したものが図 12 である。2 つのピークは確かに確認できるが、それ以外の部分はよくわからない。3D グラフはよほどうまく使わないとピークに他のデータが隠れてしまい役に立たないが、このケースでもあまりわかりやすいとは言えない。

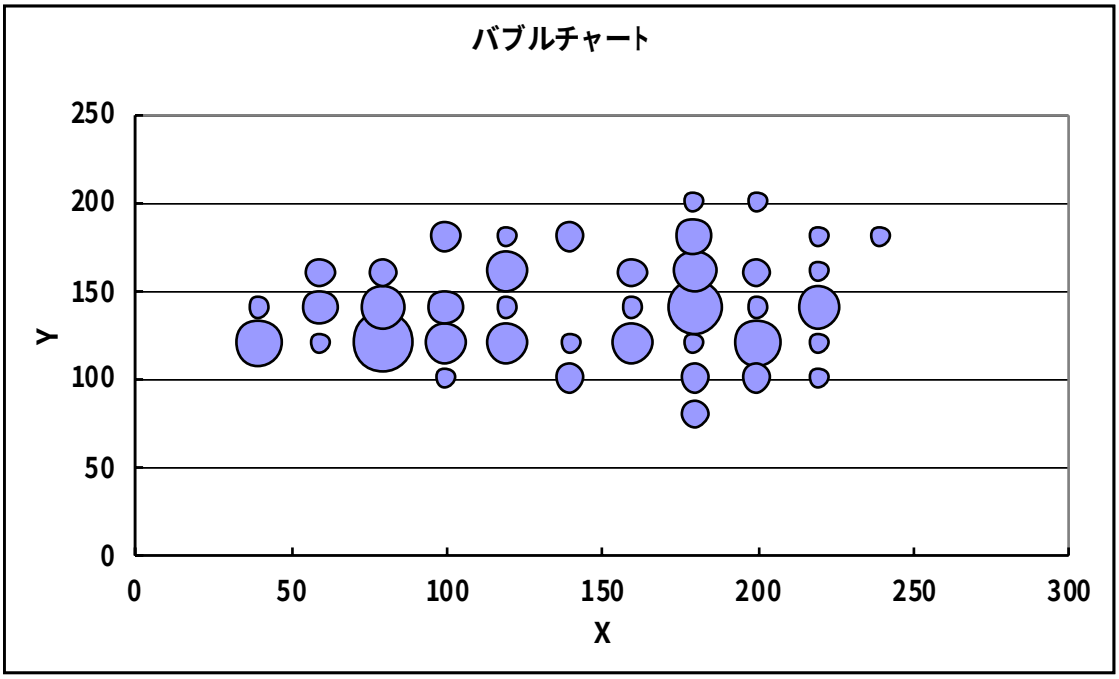



**図 11　データの密度をバブルチャートでみる**

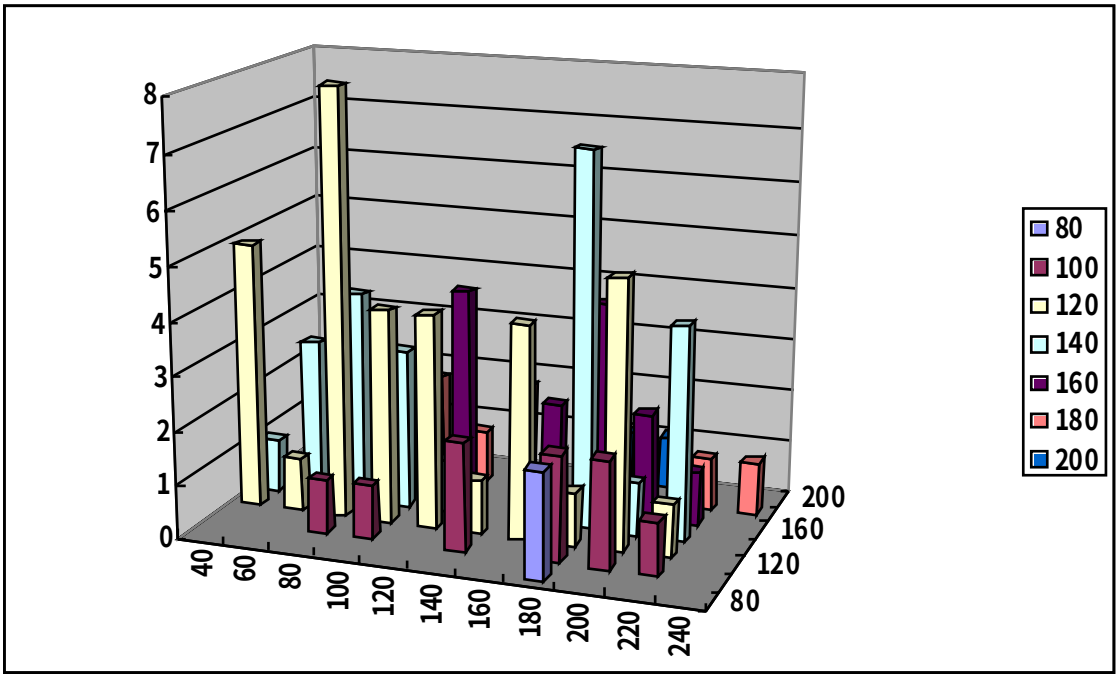



**図 12　データの密度を 3D 棒グラフでみる**

## 4.3 等高線グラフを使ってみる

3D 棒グラフにできるデータは、等高線グラフにすることができる。等高線グラフを作成した例が図 13 である。今度は 2 つのピークの位置が明らかにわかり、しかもピークの周りになだらかに密度が下がっていくようすもわかりやすい（もっとも、データ数が少ないため若干でこぼこしているが)。

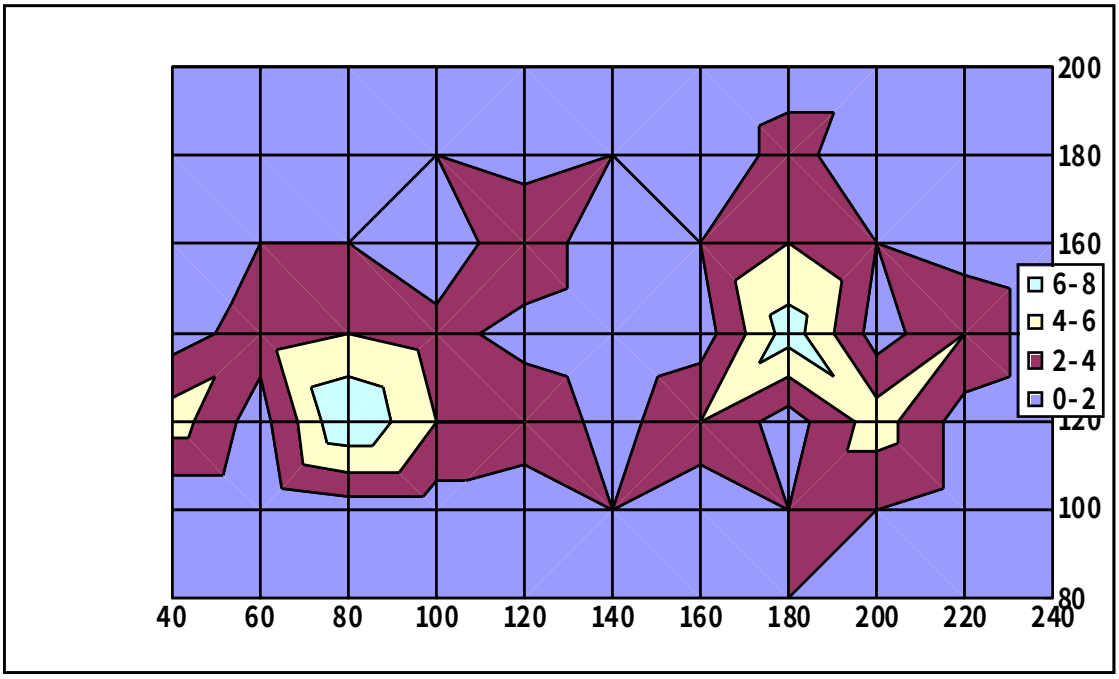



**図 13　等高線グラフで密度をみる**

## 4.4 3 変量以上の密度を視覚化する

2 次元までは等高線を使うことによって比較的わかりやすい視覚化が可能であったが、3 つ以上の変量がある場合はどうなるだろうか？　実はこの場合うまい方法はない。変量が増えていくことによって生じる問題には以下のようなものがある。

- 区切り方法を決めなければいけない変数が増える。変数が増えると最適な区切りを求めるのがどんどん困難になる
- 視覚化が難しくなる。等高線グラフは 2 次元上に密度をプロットしていくものであるから、3 変量以上のデータを簡単に扱うことはできない。このような場合は、もともとの多次元空間を平面で切断した切断面の密度を等高線として何枚も提示するくらいしか手がない（例えば、太陽の内部密度分布を平面に表示することを考えて見よ）
- 多次元化していくと、データが散らばれる範囲がどんどん広がってくるため意味のある密度推定をするためにはより多くのデータが必要になる。また局所的な密度の集中が減るため、簡単には密度推定が行えなくなる

社会科学の領域では、必要なだけデータを取れるとは限らないためあまりに高次元のモデルを作成するとロクなことにはならないことが多い。

## 練習問題 2

2002-06-17.xls の「練習問題 2」シートのデータを使って、散布図、バブルチャート、3D 棒グラフ、等高線図をそれぞれ作成してみなさい。また、その結果からこのデータはどのような構造をもっていそうか推定しなさい。

なお、等高線グラフでの等高線階級の幅は凡例のプロパティで設定できる。