プログラミング演習
-Excel による統計処理-

公立小松大学臨床工学科
藤田 一寿

公開場所
https://github.com/KazuhisaFujita/BasicStatisticsUsingExcel

第 1 回

Excel による統計処理 1

1　目的

Excel の操作を通し，データ処理の基礎とグラフの作成の仕方を学ぶ.

2　統計とは

統計や確率は様々な場面で用いられています．それはなぜでしょうか．一つの理由は，全体の特徴もしくは規則性を捉えるためです．皆さんは，大学受験で大学の偏差値や合格者のセンター試験の平均点などを調べたのではないでしょうか．偏差値やセンター試験の平均点は立派な統計によって得られた量です．これらの数値は，大学の合格者の集団の特性を表しています．今後，卒業研究はもちろん就職してからも大量のデータを扱い，集団の特性や規則性を調べることになると思います．この演習を通じ，統計の基礎と Excel による簡単な統計処理を身につけてください．しかし，**統計やそれに関わる機械学習，人工知能を本気で使うのであれば R や Python といったプログラミング言語を学ぶ必要があります.**

3　データ

3.1　データセット

N 個の値からなるデータセットがあるとするとそのデータセット X は次のように表されます.

$$X = \{x_1, x_2, ..., x_N\}. \tag{1.1}$$

3.2　データの保存形式

データを保存するためのデータ形式は様々ありますが，この演習では csv ファイルと xslx ファイルを用います．csv ファイルは，Comma-Separated Values ファイルの略です．その名の通り，csv ファイルの中身は，データがカンマで区切ってあるだけのテキストファイルです．単なるテキストファイルですので，どのようなコンピュータを使っている人でも必ず見ることができます．csv ファイルは Excel など表計算ソフトが入っていれば，表計算ソフトに関連付けされています．ただし，csv ファイルはあくまでも単なるテキストファイルですので，表計算ソフトによる関数による自動計算，作成したグラフ，文字の装飾，表の罫線などは保存されません．xslx ファイルは Excel2007 以降で標準的に用いられるファイルです．データだけではなく，関数による自動計算，作成したグラフ，文字の装飾，表の罫線なども保存されます．この演習では，データは csv ファイルで提供します．**演習結果の保存は必ず xslx ファイルにしましょう．**

4　統計量

4.1　様々な統計量

例えば，表 1.1 に示すデータがあったとします．このデータの特徴は何でしょうか．といきなり言われても困りますよね．そこで, データの特徴を表すために統計学の力を借りることが多くあります．データの特徴を表す統計学的な方法により求められた量のことを統計量と言います．今回用いる統計量は，平均，中央値，分散，標準偏差です[*1]．

■平均　平均 μ は集団の値の重心を表す最も頻繁に用いられる統計量です．平均は次の式で表されます．

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{1.2}$$

皆さんおなじみの平均です．表 1.1 のデータの平均は 65.09 です．

■中央値　中央値はデータを大きさ順で並べ替えて，順番としてちょうど真ん中にあたる値のことです．例えば，$\{1, 4, 5, 7, 8\}$ のようなデータがあったとすると，中央値は順番的に真ん中の 5 となります．例のデータの数は奇数でしたので，順番としての中央

[*1] 特に，これらの統計量は基本統計量と呼ばれます．

表 1.1　データ

82	66	39	66	54	56	58	72	53	60
69	79	71	68	50	68	61	66	74	77
74	72	73	56	47	59	56	76	69	67
87	66	57	45	84	53	47	52	49	74
83	69	87	61	59	64	66	69	79	68
55	68	50	66	60	69	60	58	83	72
74	79	65	77	52	65	48	56	76	65
63	62	63	77	72	68	69	59	63	82
63	68	80	61	48	58	55	61	54	66
69	56	79	61	62	61	65	65	75	69

を決定できましたが，データの数が偶数の場合はどうしたら良いでしょうか．例えば，$\{1, 3, 4, 5, 7, 9\}$ のようなデータがあったとすると，ちょうど真ん中の値はありません．このような場合は，真ん中の 2 つの値，4 と 5 を足して 2 で割った値，すなわち 4.5 が中央値となります．表 1.1 のデータの中央値は 65.5 です．表 1.1 のデータの場合は，平均と中央値にあまり差がないので，中央値がある必要性が分からないかもしれませんが，中央値は，平均に対して利点があります．後のほど行う演習で確かめましょう．

■分散と標準偏差　分散はデータのばらつき具合を表します．分散は次の式で表されます．

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2. \tag{1.3}$$

分散の平方根を標準偏差といいます．表 1.1 のデータの分散は約 102 で標準偏差は約 10.1 です．

4.2　Excel による統計量の計算

統計量の一つである総和を Excel で計算してみます．手順は次のとおりです．

1. データを入力します．今回は，図 1.1 のように A 列に 1, 2, 3, 4, 5 と入力してください．
2. 総和を入力したいセルを選択する．
3. “=SUM(” と入力する．

4. 総和を計算したいセルをマウスで選択する．そうすると “(” の後ろにセル番号が入力されます (図 1.1)．また，データのセル番号を直接入力することで，マウスで選択する操作と同様のことができます．例の場合では，セル A1 から A5 までの総和なので “=SUM(A1:A5)” と書きます[*2][*3]．
5. “)” を入力する．

同じやり方で，平均，分散，標準偏差が計算できます．平均なら “SUM” の部分を “AVERAGE” に変えます．統計量と Excel の関数の関係を表 1.2 に示します[*4]．

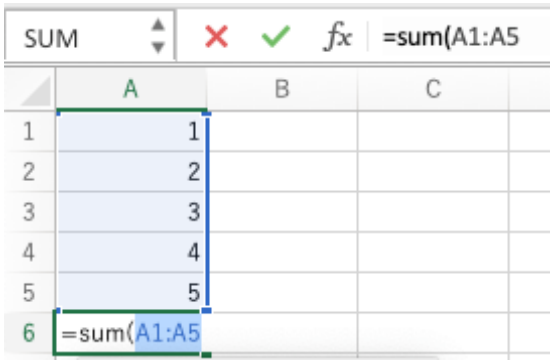


図 1.1　総和を計算したいセルを選択した状態．

表 1.2　代表値と Excel の関数

総和	SUM
平均	AVERAGE
中央値	MEDIAN
分散	VAR.P
標準偏差	STDEV.P
最大値	MAX
最小値	MIN

5　度数分布表とヒストグラム

5.1　度数分布表，ヒストグラムとは

生データを並べただけでは，それが持つ特徴を直感的に理解することは難しいです．そこで，データの整理の方法の一つに，度数分布表があります．度数分布表は，データの取りうる範囲をいくつかの階級に分け，それぞれの階級にあるデータの数 (度数) を表したものです．表 1.3 に度数分布表の例を示します．

この度数分布表を図で表したものがヒストグラム (図 1.2) です．ヒストグラムの横軸は階級を表し，縦軸が度数を表します．つまり，棒の長さ (面積) が階級に占めるデータの

[*2] データを選択する方法はいくつかあります．皆さんの慣れた方法でやってください．
[*3] Excel も自動でいろいろやってくれるので，閉じカッコが保管される場合もあります．臨機応変に対応してください．
[*4] 分散と標準偏差に関連する関数は複数あります．それぞれ意味が違います．それぞれの意味はこの講義で扱う範疇を超えているので説明しません．興味がある人は調べてみましょう．

表 1.3　度数分布表

階級	度数
0 から 10	0
11 から 20	0
21 から 30	0
31 から 40	1
41 から 50	8
51 から 60	23
61 から 70	40
71 から 80	21
81 から 90	7
91 から 100	0
100 以上	0
合計	100

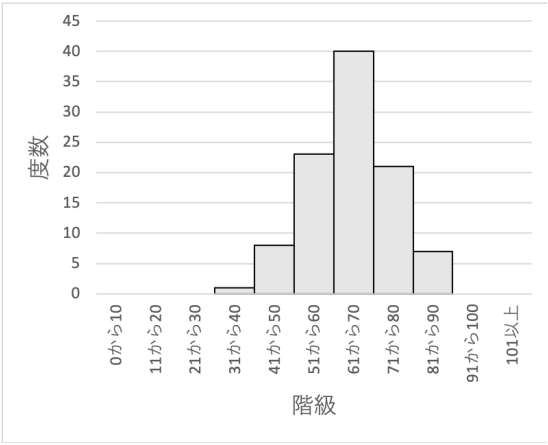


図 1.2　ヒストグラム

多さを表すことになります．ヒストグラムの形状を分布といいます．分布の形状はデータの特徴として非常に重要です．

5.2　度数分布表の作成

Excel を用い，データ (表 1.1) から図 1.7 のような度数分布表を作るためには次の手順が必要です．

1. まず，図 1.3 の A 列のように Excel にデータを入力します．表 1.1 のデータは freq.csv に入っていますので，それを開くと A 列にデータが表示されます．
2. 次に階級を設定します．今回は，0 から 10，11 から 20，21 から 30，31 から 40，41 から 50，51 から 60，61 から 70，71 から 80，81 から 90，91 から 100 という階級にします．
3. 度数分布表を図 1.3 に作成します．ここではまだ度数は空欄になっています．Excel の度数分布作成の機能の都合上，図 1.3 のように階級の上限を階級の区切りとして入力しておきます．
4. 図 1.4 のように度数を入力したいセルを選択します．

5. 数式を入力するフォーム (数式バー) に “=FREQUENCY(” と入力します.*5
6. データを選択します．そうすると，図 1.5 のようにデータがあるセル番号が入力されます.
7. “,” を入力したあと，各階級の上限が入力されたセル (今回は階級の区切の列) を選択します．そうすると，図 1.6 のようにセル番号が入力されます.
8. “)” を入力します.
9. ここが重要です．Ctrl + Shift + Enter を押します.
10. 合計は先程用いた “sum” を用いて計算します．そうすると，図 1.7 のような度数分布表が出来上がります.

	A	B	C	D	E
1	82		階級の区切	階級	度数
2	69		10	0から10	
3	74		20	11から20	
4	87		30	21から30	
5	83		40	31から40	
6	55		50	41から50	
7	74		60	51から60	
8	63		70	61から70	
9	63		80	71から80	
10	69		90	81から90	
11	66		100	91から100	
12	79			101以上	
13	72			合計	
14	66				

図 1.3　Excel で作成した度数分布表

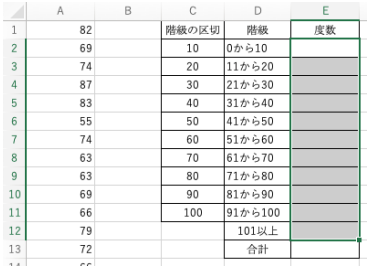


図 1.4　度数を入力するセルを選択した状態

5.3　ヒストグラムの作成

5.2 節で作成した度数分布表から図 1.2 に示すヒストグラムを作ります．横軸は内容，縦軸は度数とします．作る手順は次のとおりです.*6

1. リボンの挿入を選びます．そうすると，グラフや図などの挿入ボタンがたくさん出てきます.
2. 図 1.8 のようにグラフにしたい数値の入ったセルを選択します.
3. リボンにある棒グラフアイコンを押し，図 1.9 に示す棒グラフボタンを押します．そうするとヒストグラムが出来上がります.

*5 “COUNTIFS” という関数を使っても度数分布表を作ることが可能です.
*6 ヒストグラムもかき方が多々あるので各自の好みで.

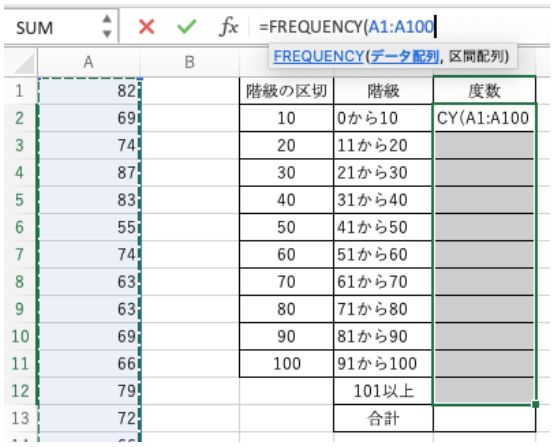


図 1.5　データを選択した状態

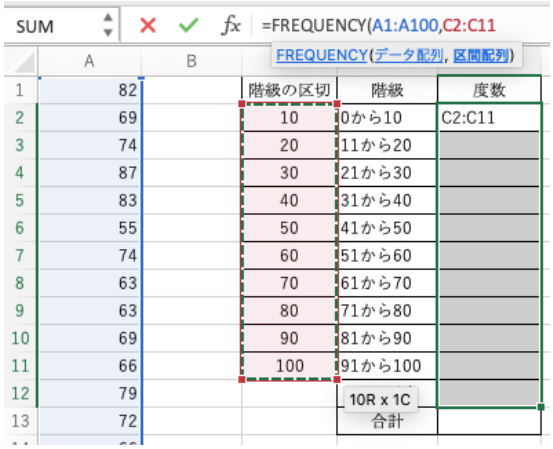


図 1.6　階級の区切りを選択した状態

階級の区切	階級	度数
10	0から10	0
20	11から20	0
30	21から30	0
40	31から40	1
50	41から50	8
60	51から60	23
70	61から70	40
80	71から80	21
90	81から90	7
100	91から100	0
	101以上	0
	合計	100

図 1.7　完成した度数分布表

4. ヒストグラムの棒の幅を変えたい場合は，ヒストグラムの棒を図 1.10 のように選択し，図 1.11 の Gap Width (要素の間隔) を 0 にします．そうすると図 1.2 のようなヒストグラムが完成します．

5.4　COUNTIF，COUNTIFS を用いた集計処理

データの中から，ある条件に当てはまるものの数を数えたい場合があると思います．例えば, 表 2.3 のような名簿から男女の数を求める場合です．このような場合は, COUNTIF という関数を使います．それではまず，表 2.3 から，石川県出身者の人数を計算してみましょう．

1. まず表??に示すデータを図 1.12 のように入力します．今回は，A 列を番号，B 列に出身県を入力します．

	C	D	E
	階級の区切	階級	度数
	9	0から9	0
	19	10から19	0
	29	20から29	0
	39	30から39	1
	49	40から49	6
	59	50から59	22
	69	60から69	43
	79	70から79	20
	89	80から89	8
	99	90から99	0
		100以上	0
		合計	100

図 1.8　セルを選択した状態

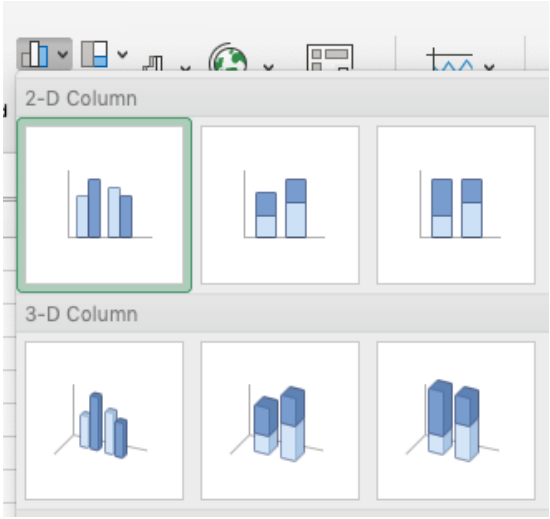


図 1.9　棒グラフを選択する

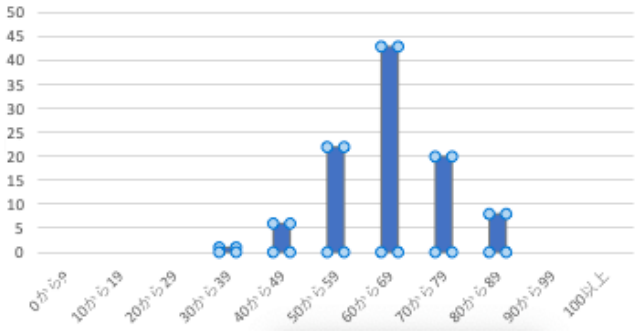


図 1.10　棒を選択した状態

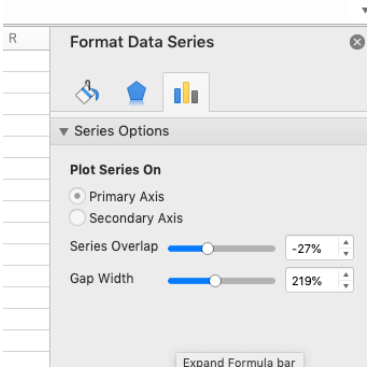


図 1.11　棒の太さを選ぶ画面

2. 図 1.12 のように人数を入れる表を作ります．
3. まず石川県出身の人を数えたいので，人数を入れるセルに “COUNTIF(” と入力します．
4. 出身が書かれたセルをマウスで選択します．そうすると “(” の後ろにセル番号が入力されます (図 1.13)．また，データのセル番号を直接入力することで，マウスで選択する操作と同様のことができます．この場合では，セル B2 から B6 までなので “=COUNTIF(B2:B6” と書きます．
5. カンマ “,” を入力します．
6. 図 1.14 のように条件を入力します．今回は石川県出身者の人数を数えるので，出身県のセルの内容が石川と等しいという条件になります．石川と等しいという条件

表 1.4　データ

番号	出身県
1	石川
2	福井
3	富山
4	石川
5	岐阜
6	福井
7	石川
8	大阪
9	石川
10	富山

表 1.5　データ

演算子	条件
=	同じ
<>	異なる
>	より大きい
<	より小さい
>=	以上
<=	以下

は Excel では，“ ”=石川” ” と書かれます．= は演算子と呼ばれ，条件を表す記号です．演算子を変えることで，等しい以外に異なるなどの条件をつけることが可能です．使用できる演算子は表 1.5 に示します．

7. 条件を入力し終えると人数が表示されます．

うまくできたら，富山県出身者と福井県出身者の人数も計算してみましょう．

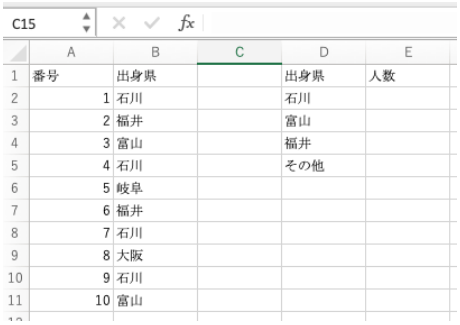


図 1.12　エクセルで作成した表

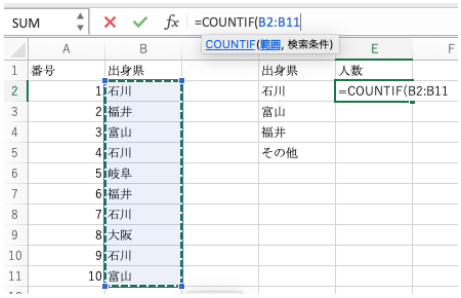


図 1.13　COUNTIF を入力しセルを選んだ状態

先の方法で，石川県出身者の人数を計算できました．では，石川県，富山県，福井県以外をその他として人数を計算した場合はどうすれば良いでしょうか?この場合，3 県以外なので，それぞれの県に対して異なるという条件が必要になります．つまり，条件が 3 つ必要です．しかし，COUNTIF 関数は数える条件を 1 つしか設定できませんでした．そこ

図 1.14　条件を入力した状態

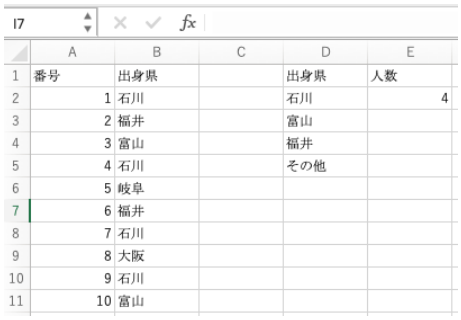


図 1.15　出力結果

で，COUNTIFS を用います．COUNTIF と COUNTIFS の違いは，条件が一つか，複数かの違いです．

1. その他の県出身者の人数を入れるセルに “COUNTIFS(” と入力します．
2. 出身が書かれたセルをマウスで選択します．そうすると “(” の後ろにセル番号が入力されます．また，データのセル番号を直接入力することで，マウスで選択する操作と同様のことができます．この場合では，セル B2 から B6 までなので “=COUNTIFS(B2:B6” と書きます．
3. “,” を入力します．
4. 石川以外という条件 “ ”<> 石川” ” を入力し，“,” を入力します．
5. 富山以外という条件 “ ”<> 富山” ” を入力し，“,” を入力します．
6. 福井以外という条件 “ ”<> 福井” ” を入力し，“)” を入力します．
7. そうすると，入力した 3 条件をすべて満たすセルの数，すなわち，3 県出身者以外の人数が計算されます．

“COUNTIFS” は，すべての条件を満たす，すなわち条件の論理積を計算しています．複雑な条件の場合，COUNTIF を複数使わなければならなくなります．

6　演習

演習 1　アヤメデータ iris.csv 内にある，がく片の長さ，がく片の幅，花びらの長さ，花びらの幅についてそれぞれ平均，中央値，分散，標準偏差を求めなさい．

演習 2　gaussian.csv に書かれている数値の平均と中央値を求めなさい．そして，平均値と中央値の差について考察しなさい．ただし，gaussian.csv は，表 1.1 の最後のデータを 500 に置き換えたものです．(ヒント:中央値の利点を調べましょう．)

演習 3 glucoselevel.csv データから度数分布表を書きなさい．ただし，階級は，60 以下，60 より大きく 80 以下，80 より大きく 100 以下，100 より大きく 120 以下，120 より大きく 140 以下，140 より大きく 160 以下，160 より大きく 180 以下，180 より大きい としなさい[*7]．

演習 4 演習 3 で作成した度数分布表に基づきヒストグラムをかきなさい．

演習 5 第節でその他の県出身者の人数を COUNTIFS で計算しました．実は，COUNTIF でも可能です．COUNTIF と四則演算だけでその他の県出身者の人数を計算できる数式を答えなさい．

ヒント: 10−COUNTIF(B2:B6, ”=石川”) は 10 人中石川県出身者以外の人数になります．

7 レポート提出

レポートは report1.docx の空欄を埋める形で作成し，kazuhisa.fujita@komatsu-u.ac.jp へ電子データで提出してください．データ形式は docx もしくは pdf とします．レポートの提出期限は次回の講義日までとします．

8 おまけ

Excel のユーザーインターフェースを構成する各部分の名称は図 1.16 のようになります．

[*7] http://www.qmss.jp/databank/medcomed/default.htm の血糖値データを用いました．

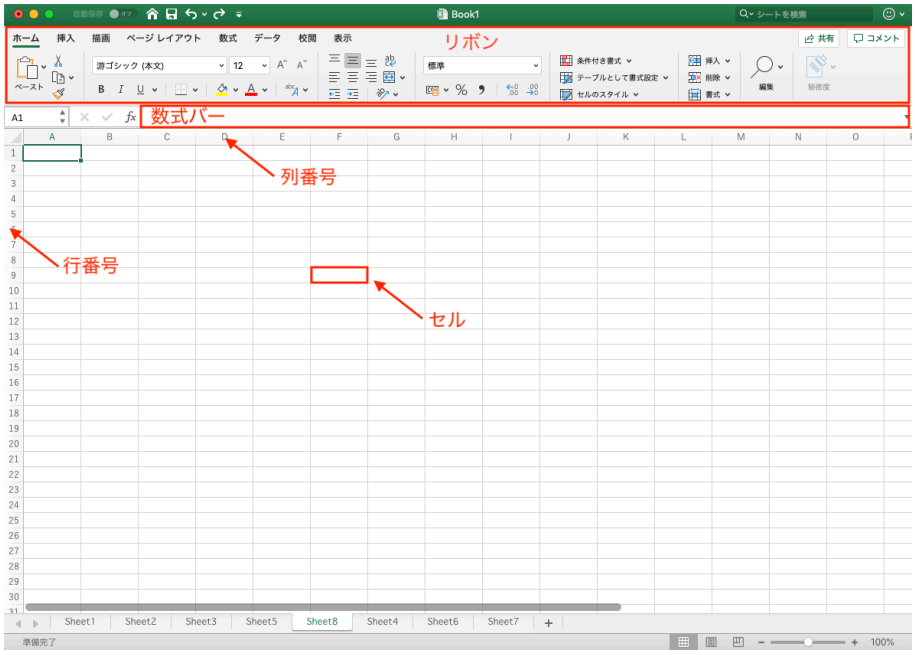


図 1.16　ユーザーインターフェース

第 2 回

Excel による統計処理 2

1　目的

データ解析おいて重要なデータの可視化 (グラフ化) の方法を学ぶ.

2　グラフ

データの特徴を視覚的に把握するためにグラフを用います．代表的なグラフの一つが前回取り扱ったヒストグラムです．ヒストグラム以外にも様々なグラフが存在します．今回は，棒グラフ，折れ線グラフ，散布図を取り扱います.

2.1　棒グラフ

■棒グラフとは　棒グラフは値の大きさの比較に用いられます．例えば，表 2.1 に示す商品の売上高をグラフ化すると図 2.1 のようになります．グラフにしたことで，それぞれの商品の売上の大小が視覚的に分かりやすくなります．しかし，棒グラフは基本的に時間変化を把握するためには用いません.

■棒グラフの作成　手順は前回のヒストグラムと同じです．表 2.1 に示すデータを棒グラフにしましょう.

1. データを入力します．今回は図 2.2 のように入力してください.
2. 棒グラフにしたいデータを図 2.2 のように選択します.
3. リボンの挿入を押します．そして，リボンにある棒グラフアイコンを押し，図 1.9 に示す棒グラフボタンを押します．そうすると，図 2.4 のような棒グラフができ

表 2.1

4 月の売上高	
商品	売上高 (万円)
商品 A	350
商品 B	120
商品 C	50
商品 D	250

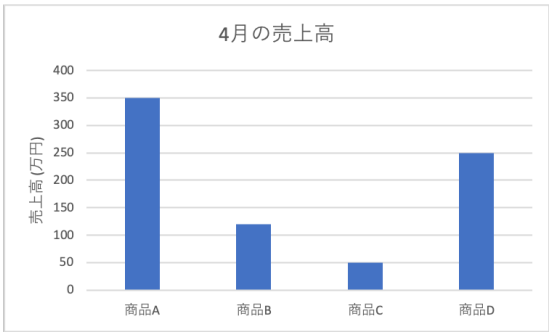


図 2.1　棒グラフ

図 2.2　データを選択した状態

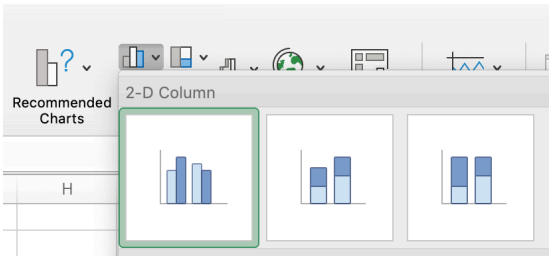


図 2.3　棒グラフを選択する

ます.

4. 出来上がったグラフには軸のラベルがありません．軸が何をあらわすか分かるようにラベルを付けます.
   (a) 軸にラベルを追加するには，軸を加えたいグラフをクリックし (グラフが選択された状態で)，リボンのグラフのデザインをクリックします.
   (b) そして，リボンのグラフ要素を追加をクリックします.
   (c) 出てきたメニュー (図 2.5) の軸ラベル第 1 縦軸を選びます．そうすると，縦軸に “軸ラベル” というラベルが追加されます.
   (d) “軸ラベル” をクリックし選んだあと，更にクリックすると編集できるようになります．縦軸は “売上金額 (万円)” にしましょう[*1].
   (e) 必ずしもグラフにタイトル付ける必要はありませんが，今回はタイトルを付けます．“グラフ タイトル” をクリックし選んだあと，更にクリックすると編集できるようになります．グラフタイトルを “4 月の売上金額” にしましょう．そうすると図 2.1 のようなグラフができます.

[*1] 軸には単位を付けましょう.

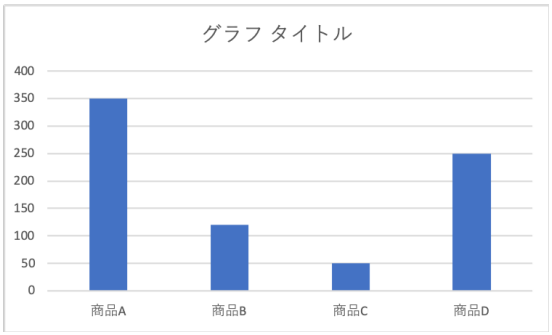


図 2.4　作成したグラフ

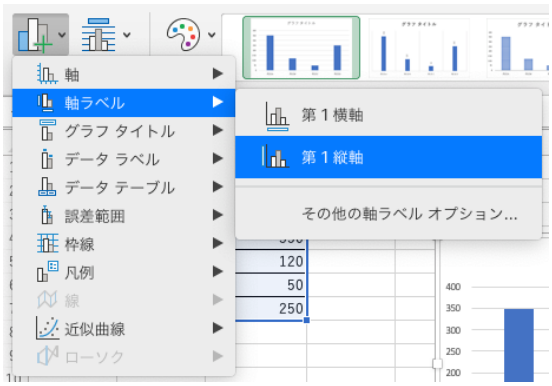


図 2.5　軸ラベルを追加する

2.2　折れ線グラフ

■折れ線グラフとは　折れ線グラフは値の時間変化の把握のために用いられます．例えば，表 2.2 に示す商品の売上高の推移をグラフ化すると，図 2.6 のようになります．商品の売上が時間とともに変化していることが，グラフ化することで視覚的に分かりやすくなります．特に，商品 B と商品 C の売上がいつ，どのように逆転したのかグラフ化することでよく分かります．ただし，大きさを比べる目的で折れ線グラフを使わないようにしましょう．

表 2.2

売上高 (万円)		
月	商品 B	商品 C
4	120	50
5	120	55
6	110	75
7	90	80
8	100	110
9	80	120

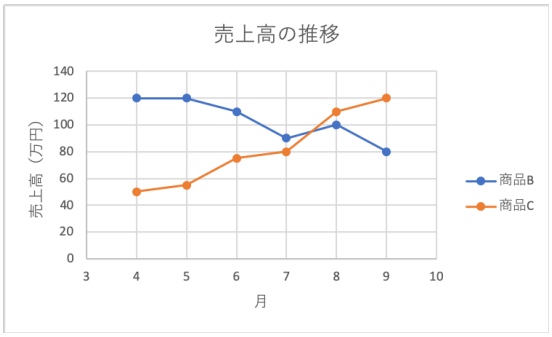


図 2.6　折れ線グラフ

■折れ線グラフの作成　表 2.2 に示すデータを折れ線グラフにします．横軸を月，縦軸を売上金額とします．

1. データを入力します．今回は図 2.7 のように入力してください．
2. 図 2.7 のように横軸のデータと縦軸のデータを選びます．値だけではなく商品名まで選択しました．そうするとグラフの線の説明 (凡例) まで自動で入力されます[*2]．
3. グラフの散布図の直線とマーカーを選びます[*3]．そうすると図 2.9 のようなグラフがかけます．
4. 凡例が下だと邪魔ですので右に移動させます．グラフを選択し，リボンのグラフのデザインをクリックします．グラフ要素の追加をクリックし，凡例にカーソルを移動させると，図 2.10 のように凡例の場所を選ぶことができます．そうすると，図 2.11 のようになります．
5. 図 2.11 を見れば分かりますが，線がグラフ全体に広がらず右上に偏っています．軸の値の範囲を変更して，線を全体に広げます．
   （a）横軸をクリックし (選び), 右クリックを押して出てくる軸の書式設定 (図 2.12) を選びます．
   （b）ウインドウの右端に出てくる軸の書式設定 (図 2.13) の最小値，最大値を変更します．今回は最小値を 3.0，最大値を 10.0 とします．同様のやり方で，縦軸の最小値を 40.0，最大値を 130.0 とします．そうすると図 2.14 のようになります．
6. 軸のラベルを追加します
   （a）軸にラベルを追加するには，まず，軸を加えるグラフを選び，リボンのグラフのデザインをクリックします．
   （b）リボンのグラフ要素を追加をクリックします．
   （c）出てきたメニュー (図 2.15) の軸ラベル第 1 縦軸を選びます．そうすると，縦軸に “軸ラベル” というラベルが追加されます．
   （d）“軸ラベル” をクリックし選んだあと，更にクリックすると編集できるようになります．縦軸は “売上高 (万円)” にしましょう[*4]．同様の手順で横軸を “月” にしましょう．そうすると，図 2.16 のようなグラフができます．
7. 必ずしもグラフにタイトル付ける必要はありませんが，今回はタイトルを付けます．“グラフ タイトル” をクリックし選んだあと，更にクリックすると編集できる

[*2] いつもうまくいくとは限りません．うまく凡例がつかなかったときは，値だけ選択してグラフをかきます．グラフを選択し，リボンのグラフのデザインを選び，グラフデータの選択をクリックします．そうするとウインドウが出てくるので，そのウインドウにある名前の欄に凡例を書き込みます．

[*3] 折れ線グラフを選んでもグラフはかけますが，軸の範囲が固定など面倒なことになります．今回は散布図にします．

[*4] 軸には単位を付けましょう．

売上高の推移（万円）		
月	商品B	商品C
4	120	50
5	120	55
6	110	75
7	90	80
8	100	110
9	80	120

図 2.7　データの選択

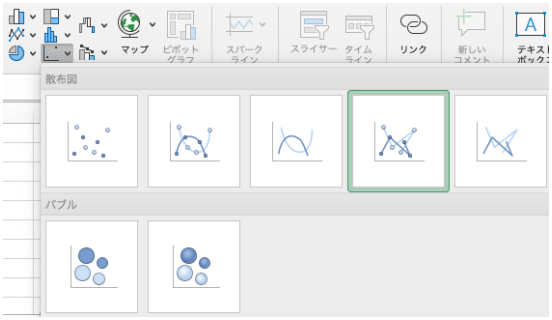


図 2.8　散布図の直線とマーカーを選択する

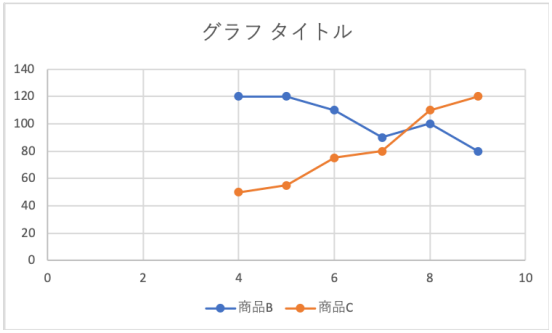


図 2.9　作成されたグラフ 1

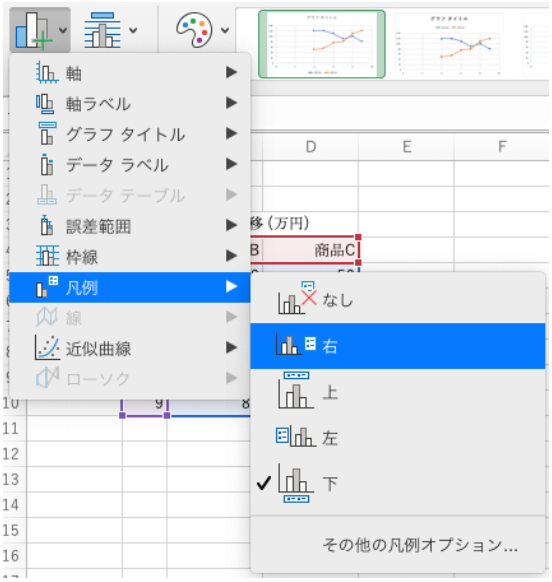


図 2.10　凡例の位置

ようになります．その状態で，“売上高の推移” に変更しましょう．以上の操作で，図 2.6 のようなグラフになります．

2.3　円グラフ

前回，表 2.3 から出身県ごとの人数を計算しました．では，各出身県出身者の人数の全体の割合がどの様になっているのか視覚的に分かりやすくするにはどうすればよいでしょうか．このようなときは，円グラフを用います．円グラフは，円を扇型に分割し，その扇

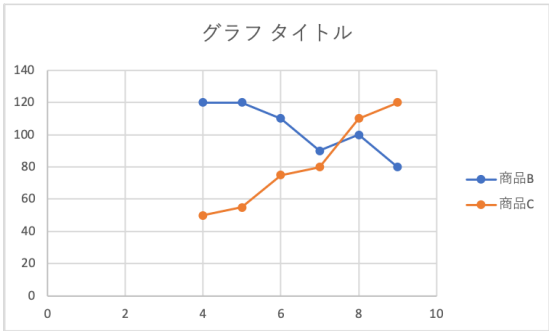


図 2.11　作成されたグラフ 2

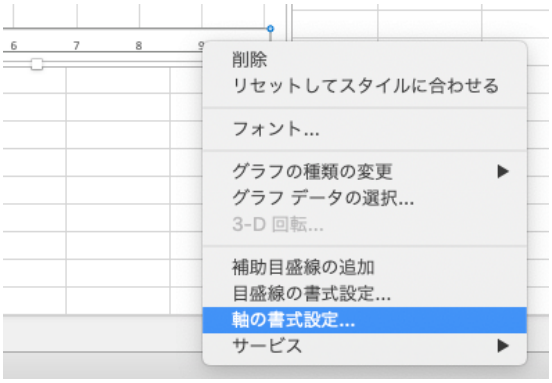


図 2.12　軸の書式設定クリックする

図 2.13　軸の範囲を変更

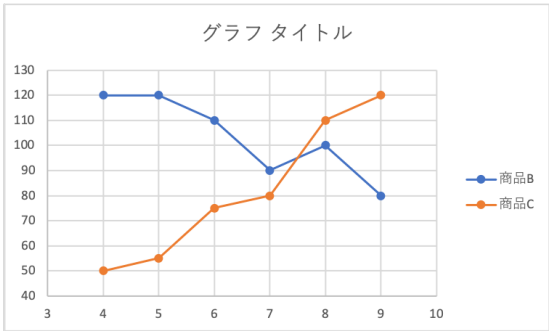


図 2.14　作成されたグラフ 3

型に割当てられた割合の大きさに応じ扇型の角度を決めたものです．扇型に割り当てられた割合をすべて足すと 100% になります．逆に言えば，円一周で 100% にならなければなりません．図 2.17 は表 2.3 から作成した出身県の人数の割合を表す円グラフです．

■円グラフの作成　表 2.3 に示すデータから出身券の割合を円グラフにします．

1. 前回行った COUNTIF，COUNTIFS の演習で作成したエクセルファイルを開きます．そのファイルにはすでに図 2.18 のように各県の出身者が計算されているはずです．
2. 図 2.19 のように円グラフにしたい系列と値を選びます．
3. リボンの挿入を押します．そしてグラフアイコンの中で円グラフを押すと図 2.20 のように表示されます．その中で 2D 円の左端のものをクリックします．そうすると，図 2.21 のようなグラフができます．

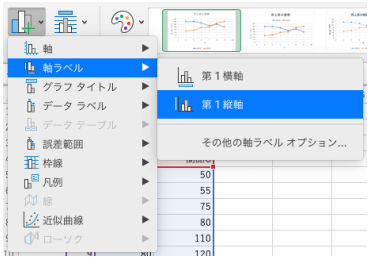


図 2.15　軸のラベルを追加

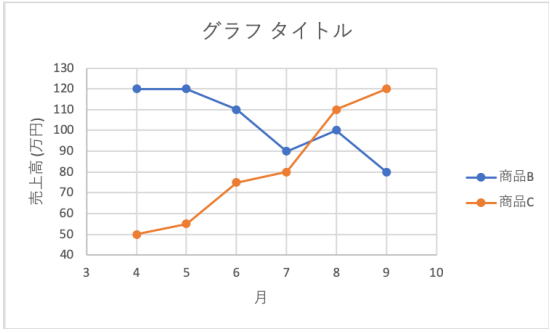


図 2.16　作成されたグラフ 4

表 2.3　データ

番号	出身県
1	石川
2	福井
3	富山
4	石川
5	岐阜
6	福井
7	石川
8	大阪
9	石川
10	富山

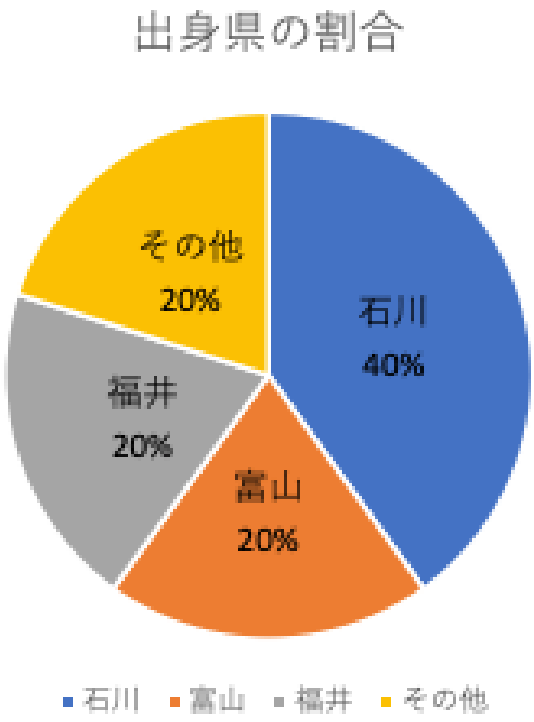
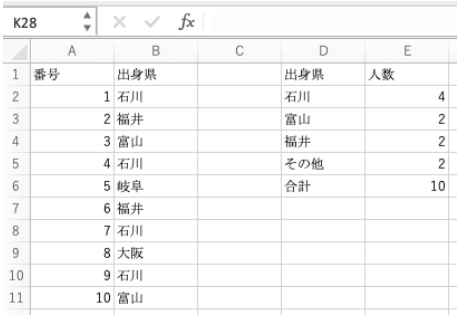


図 2.17　円グラフ

4. この状態では，グラフに割合が入っていないなど味気ないですね．そこで次に，グラフに割合を表示させましょう．まず，グラフをクリックし，リボンのグラフのデザインをクリックします．そして，グラフ要素を追加をクリックし，マウスカーソルをデータラベルに移動させると，図 2.22 のように表示されます．ここでは中央を選びます．そうすると，グラフが図 2.23 のようにグラフに人数が表示されます．
5. 人数ではなく割合 (パーセンテージ) に表示を変えます．変えるには，人数を選び右クリックをします．そうすると，図 2.24 のように表示されるので，それのデー



	A	B	C	D	E
1	番号	出身県		出身県	人数
2	1	石川		石川	4
3	2	福井		富山	2
4	3	富山		福井	2
5	4	石川		その他	2
6	5	岐阜		合計	10
7	6	福井			
8	7	石川			
9	8	大阪			
10	9	石川			
11	10	富山			

図 2.18　ファイルを開いた状態

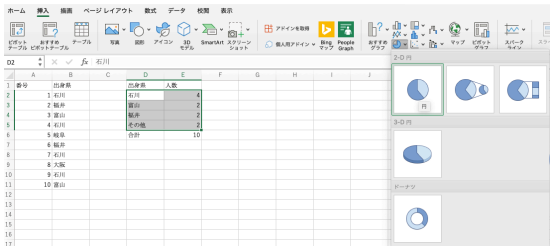


図 2.19　データを選択した状態

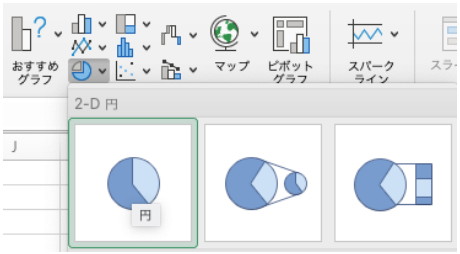


図 2.20　リボン内の円グラフのアイコン

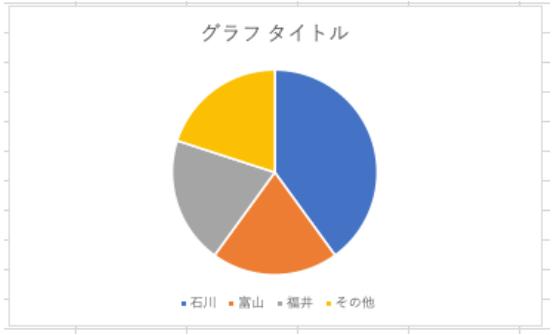


図 2.21　できた円グラフ

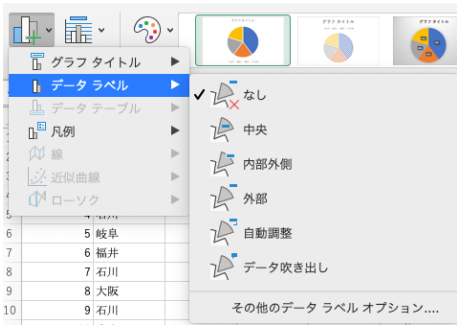


図 2.22　リボン内のラベルメニュー

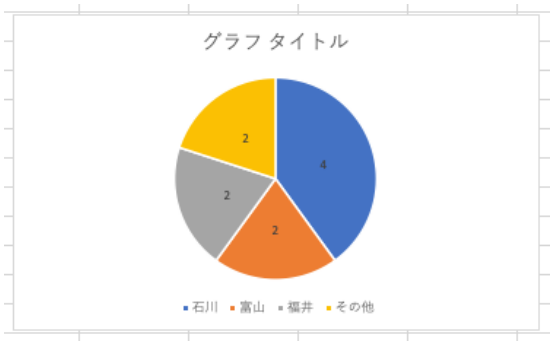


図 2.23　人数が記入された円グラフ

タラベルの書式設定をクリックします．そうすると，ウインドウの右端に図 2.25 が表示されます．図のように，分類名，パーセンテージにチェックを入れます．今回は，引き出し線はチェックしてもしなくても良いです．

6. 最後に，タイトルを編集します．グラフ タイトルをダブルクリックし，出身県に変えます．そうすると，図 2.17 が完成します．

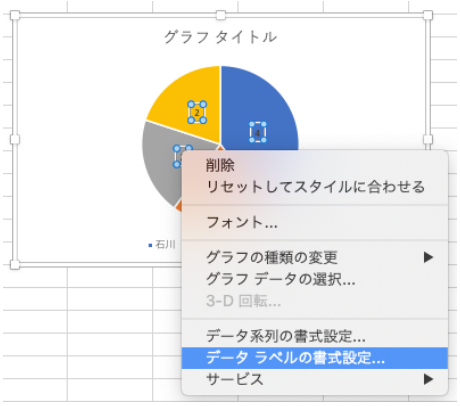


図 2.24　ラベルの書式設定を選ぶ

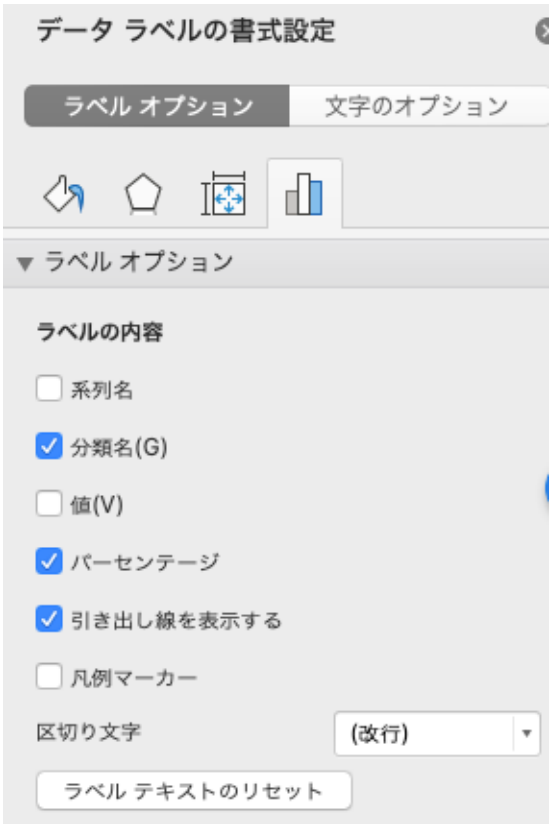


図 2.25　ラベルの書式設定

2.4　散布図

■散布図とは　データの値同士の関係性を知りたいときがあります．散布図は値の関係性を見るために用いられます．例として，前回の演習で用いたアヤメのがく片の長さと幅 (iris.csv) の散布図 (図 2.26) を示します．散布図では関係を見るデータ，例ではアヤメのがく片の長さと幅をそれぞれ軸とします．そして，データの組み合わせ，例ではアヤメのがく片の長さと幅の組み合わせを点で表します．例の散布図を見ると，がく片の長さが長いほど幅が太くなる傾向がありそうなことが分かります．データそれぞれ無関係な場合は，点は垂直，水平，円に分布します．

■散布図の作成　散布図も他のグラフの書き方と同じですが，データの列を 2 つ選ぶ点が異なります．

1. データを開きます．今回は，アヤメデータ iris.csv を用います．

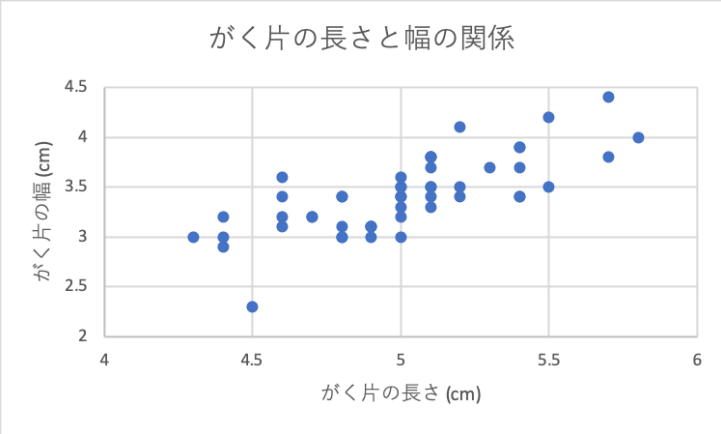
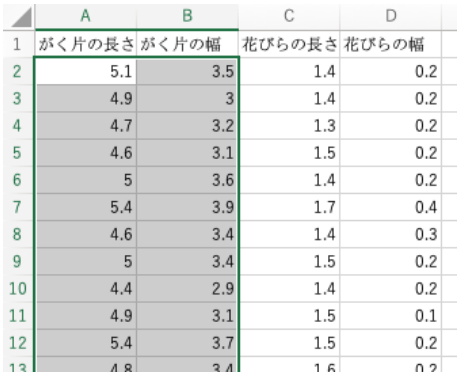


図 2.26　散布図

2. 散布図にしたいデータを選びます (図 2.27).
3. リボンの挿入を押します．そして，リボンにある散布図アイコンを押し，図 2.28 に示す散布図ボタンを押します．そうすると図 2.29 のような散布図が出来上がります.
4. 作成した散布図はグラフの右上にデータ点が偏っています．グラフ全体にデータ点を広げるために，軸の値を調整します．はじめに縦軸を変更します (横軸から変更しても良いです).
   (a) 縦軸をクリックし (選び)，右クリックを押して出てくる軸の書式 (図 2.30) を選びます.
   (b) ウインドウの右端に出てくる軸の書式設定 (図 2.31) の最小値，最大値を変更します．今回は最小値を 2.0，最大値を 4.5 とします．同様に，横軸の最小値を 4.0，最大値を 6.0 とします．そうすると図 2.32 のようになります.
5. 以上の操作で散布図は作れますが，グラフとして少し格好が悪いです．グラフには軸が何をあらわすか分かるようにラベルを付けます.
   (a) 軸にラベルを追加するには，リボンのグラフのデザインをクリックします.
   (b) そして，軸を追加したいグラフを選び，リボンのグラフ要素を追加をクリックします.
   (c) 出てきたメニュー (図 2.33) の軸ラベル第 1 縦軸を選びます．そうすると，縦軸に “軸ラベル” というラベルが追加されます.
   (d) “軸ラベル” をクリックし選んだあと，更にクリックすると編集できるように



	A	B	C	D
1	がく片の長さ	がく片の幅	花びらの長さ	花びらの幅
2	5.1	3.5	1.4	0.2
3	4.9	3	1.4	0.2
4	4.7	3.2	1.3	0.2
5	4.6	3.1	1.5	0.2
6	5	3.6	1.4	0.2
7	5.4	3.9	1.7	0.4
8	4.6	3.4	1.4	0.3
9	5	3.4	1.5	0.2
10	4.4	2.9	1.4	0.2
11	4.9	3.1	1.5	0.1
12	5.4	3.7	1.5	0.2
13	4.8	3.4	1.6	0.2

図 2.27　セルを選択した状態

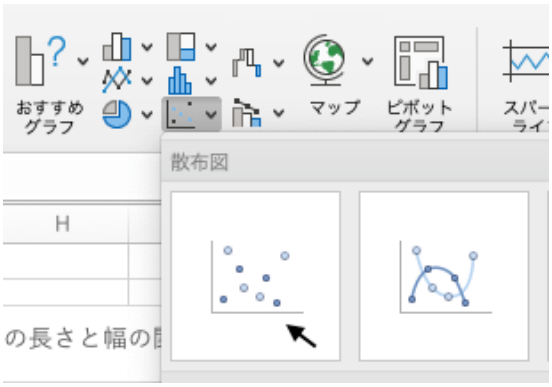


図 2.28　散布図を選択する

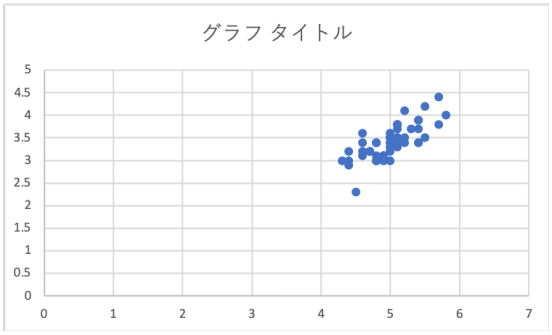


図 2.29　未完成な散布図

図 2.30　軸のメニュー

なります．縦軸は “がく片の幅 (cm)” にしましょう[*5]．同様の手順でよく軸を “がく片の長さ (cm)” しましょう．そうすると，図 2.34 のようなグラフができます．

6. 必ずしもグラフにタイトル付ける必要はありませんが，今回はタイトルを付けます．“グラフ タイトル” をクリックし選んだあと，更にクリックすると編集できるようになります．その状態で，“がく片の長さと幅の関係” に変更しましょう．以上の操作で，図 2.26 のようなグラフになります．

3　相関係数

■相関係数とは　散布図では，視覚的に 2 つのデータの関係性を見ることができます．しかし，図ではそのデータがどの程度関係しているのか定量的に分かりません．定量的に 2

[*5] 軸には単位を付けましょう．

図 2.31　軸の範囲の設定

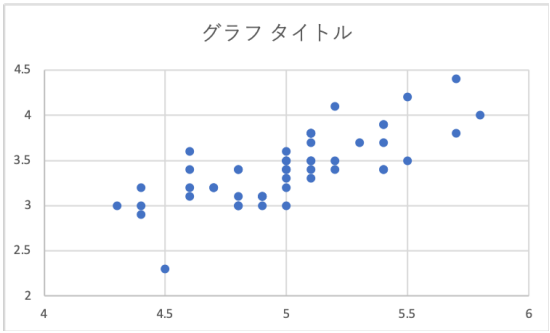


図 2.32　軸の範囲を変えた散布図

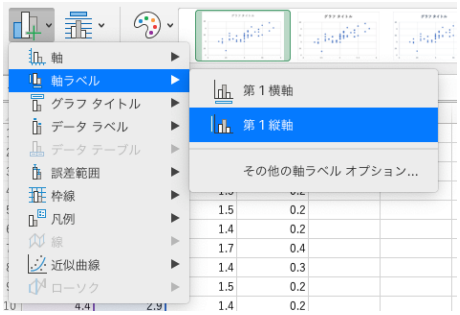


図 2.33　軸の範囲の設定

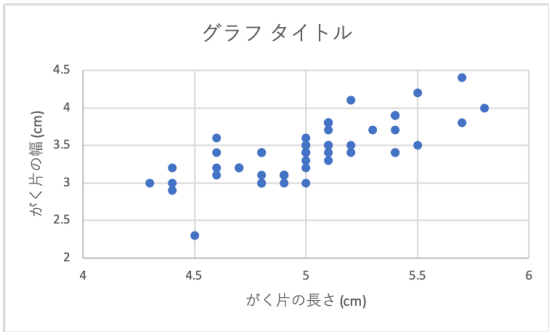


図 2.34　軸のラベルを付けた散布図

つのデータの関係を見る場合は相関係数を用います．相関係数は次の式で計算されます．

$$r = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{2.1}$$

相関係数はデータ同士が直線的な関係かどうかをみることができます．相関係数が正の場合は正の相関，負の場合を負の相関と言います．正の相関の場合，一方のデータの値が増えれば，もう一方のデータの値も増えるという関係があることが分かります．負の相関の場合，一方のデータの値が増えれば，もう一方のデータの値は減るという関係があることが分かります．しかし，U 字型であったり S 字型の関係であった場合は相関係数ではその関係性を定量化することはできません．

■相関係数の計算　Excel で相関係数を求めるのは簡単です．前回の総和を求めたのと同じ要領で行います．しかし，今回は 2 列データがありますので，そこだけ注意しましょう．

1. データを開きます．今回も先程用いたアヤメデータ iris.csv を用います．
2. 相関係数を入れるセルを選択します．
3. “=CORREL(” と入力します．
4. 相関係数を求めたいデータの列を 2 つ選びます．
5. “)” と入力します．そうすると，相関係数が表示されます．

4　回帰直線

■回帰直線とは　2 つデータの値同士の関係を知るために散布図を用いる事ができることを先程述べました．図 2.35 のように，その関係性を直線で表したものを回帰直線と言います．回帰直線は，データの分布を直線に近似したものです．もう少し簡単に言えば，2 つのデータの関係を直線の式 $y = ax + b$ に強制的に表したものです[*6]．

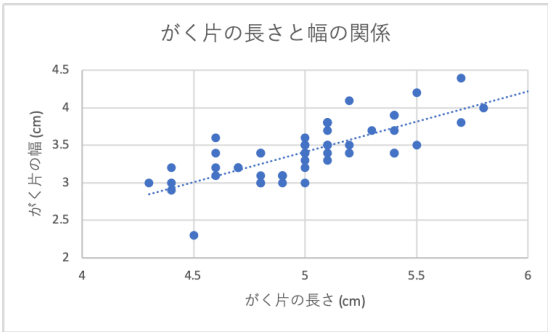


図 2.35　散布図と回帰直線

■回帰直線の作成　先程作成した散布図に回帰直線を追加します．

1. 回帰直線を追加したい散布図を選びます．今回は，先ほど作成したアヤメデータの散布図に回帰直線を追加します．
2. リボンの “グラフのデザイン” を選びます．
3. 図 2.36 のように，近似曲線の線形予測を選びます．そうすると，図 2.35 のような回帰直線がかかれます．軸の範囲が変わった場合は，見やすいように変更しましょう．

[*6] 直線近似の方法は，この講義の範疇を超えているので説明しません．興味がある人は調べてみましょう．

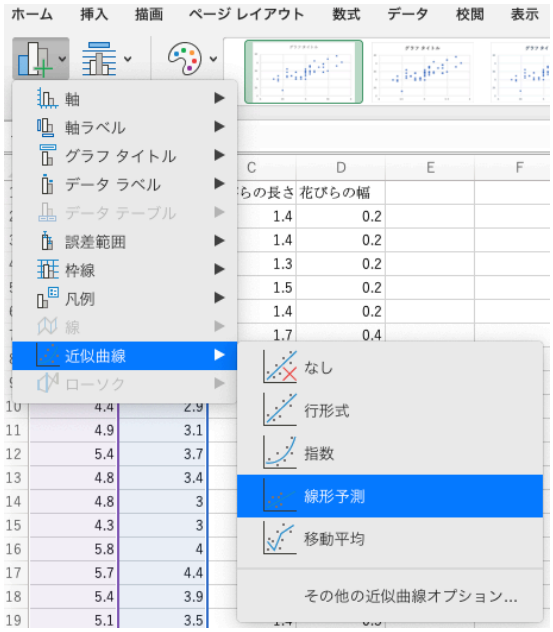


図 2.36　散布図

5　演習

演習 1　temp_city.csv ファイルから棒グラフを作りなさい．ただし，横軸は都市，縦軸は気温 (度) にしなさい．

演習 2　2019 年 4 月 8 日の小松市の気温データ temp.csv から折れ線グラフを作りなさい．ただし，横軸は時間 (時)，縦軸は気温 (度) にしなさい．

演習 3　2020 年 4 月 21 日までの石川県内のコロナウイルス感染者情報 COVID-19_Ishikawa.csv[*7]から，感染者の男女の割合を円グラフで表しなさい．

演習 4　私鉄の資本金と従業員数のデータ shitetsu.csv から散布図をかきなさい．ただし，横軸は資本金，縦軸は従業員数にしなさい[*8]．

演習 5　私鉄の資本金と従業員数のデータ shitetsu.csv 内の資本金と従業員数の相関係数を求めなさい．

演習 6　演習 4 で作成した散布図に回帰直線加えたものを作成しなさい．

[*7] SIGNATE の COVID-19 Challenge で作成されているデータを用いました．https://signate.jp/competitions/261

[*8] http://www.qmss-matsubara.sakura.ne.jp/databank/index.html の私鉄データを用いました．

6　レポート提出

レポートは report2.docx の空欄を埋める形で作成し，kazuhisa.fujita@komatsu-u.ac.jp へ電子データで提出してください．データ形式は docx もしくは pdf でお願いします．レポートの提出期限は次回の講義日までとします．