

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

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 理論

2.1 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．一つの目的は，全体の特徴もしくは規則性を捉えるために用いられます．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のでしょうか．偏差値やセンター試験の平均点は統計の結果得られた立派な量です．これらの数値は，大学の合格者の手段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．

2.2 データ

N 個のデータがあるとするとそのデータのセット X は次のように表されます．

$$X = \{x_1, x_2, \dots, x_N\} \quad (1.1)$$

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

2.3 代表値

例えば，表 1.1 に示すデータセットがあったとします．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．データセットの特徴を 1 つの値で表したいときに用いるのが代表値です．今回用いる代表値は，平均，中央値，分散，標準偏差です．

■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代表値です．平均は次の式で表されます．

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad (1.2)$$

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．例のデータの数に奇数でしたので，順番としての中央を決定できましたが，データの数に偶数の場合はどうしたら良いのでしょうか．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．中央値は，平均に対して利点があります．後のほどう演習で確かめましょう．

図 1.1 度数分布表

階級	度数
0 以上 9 以下	0
10 以上 19 以下	0
20 以上 29 以下	0
30 以上 39 以下	1
40 以上 49 未満	6
50 以上 59 未満	22
60 以上 69 未満	43
70 以上 79 未満	20
80 以上 89 未満	8
90 以上 99 未満	0
100 以上	0
合計	100

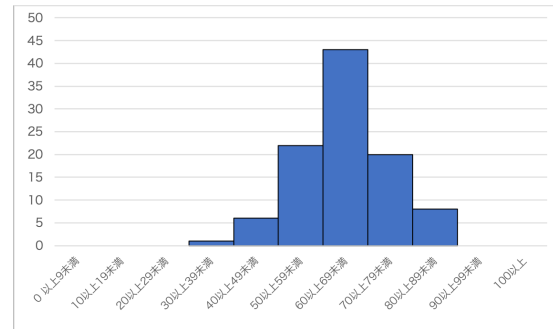


図 1.2 ヒストグラム

■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。

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

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。

2.4 度数分布表とヒストグラム

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、それぞれ階級にあるデータの数 (度数) を表したものです。表 1.1 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多さを表すことになります。ヒストグラムに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。

3 Excel 実験

3.1 データ形式

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 xlsx ファイルを用います。csv ファイルは、Comma-Separated Vales 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。xlsx 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が、演習結果の保存は xlsx ファイルにしましょう。

3.2 総和、平均、中央値、分散

Excel での総和の計算は次の手順で行います。

1. 総和を入力したいセルを選択する。
2. “=sum(” と入力する。
3. 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る (図 1.3)。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。^{*1}
4. “)” 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 “MEAN” に変えます。代表値と excel の関数の関係を表 1.2 に示します^{*2}。

^{*1} 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。また、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。

^{*2} 分散や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。

SUM		✖	✔	fx	=sum(A1:A5)
	A	B	C		
1	1				
2	2				
3	3				
4	4				
5	5				
6	=sum(A1:A5)				

図 1.3 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	MEAN
分散	VAR.P
標準偏差	STDEV.P
最大値	MAX
最小値	MIN

■演習 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 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.

3.3 度数分布表

Excel を用い、データ (表 1.1) から図 1.4 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 1.5 の A 列のように Excel にデータを入力します.
2. 次に階級を設定します. 今回は、0 から 9, 10 から 19, 20 から 29, 30 から 39, 40 から 49, 50 から 59, 60 から 69, 70 から 79, 80 から 89, 90 から 99, 100 以上という階級にします.
3. 度数分布表を図 1.5) に作成します. ここではまだ度数は空欄になっています.
4. 図 1.6 のように度数を入力したいセルを選択します.
5. 関数を入力するフォームに “=FREQUENCY(” と入力します.*3
6. データを選択します. そうすると、図 1.7 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、各階級の大きい数値が入力されたセルを選択します. そうす

*3 “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.

ると、図 1.8 のようにセル番号が入力されます.

8. “)”を入力する。
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 1.5 のような度数分布表が出来上がります。

階級 (区間)	内容	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 1.4 Excel で作成した度数分布表

	A	B	C	D	E
1	82		階級 (区間)	内容	度数
2	69		9	0から9	
3	74		19	10から19	
4	87		29	20から29	
5	83		39	30から39	
6	55		49	40から49	
7	74		59	50から59	
8	63		69	60から69	
9	63		79	70から79	
10	69		89	80から89	
11	66		99	90から99	
12	79			100以上	
13	72			合計	

図 1.5 Excel で作成した度数分布表

C	D	E
階級 (区間)	内容	度数
9	0から9	
19	10から19	
29	20から29	
39	30から39	
49	40から49	
59	50から59	
69	60から69	
79	70から79	
89	80から89	
99	90から99	
	100以上	
	合計	

図 1.6 度数を入力するセルを選択した状態




SUM				=FREQUENCY(A1:A100
	A	B	FREQUENCY([data_array], [bins_array])	
1	82		階級 (区間)	内容
2	69		9	0から9
3	74		19	10から19
4	87		29	20から29
5	83		39	30から39
6	55		49	40から49
7	74		59	50から59
8	63		69	60から69
9	63		79	70から79
10	69		89	80から89
11	66		99	90から99
12	79			100以上
13	72			合計

図 1.7 データを選択した状態

■演習 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

SUM				<i>fx</i>	=FREQUENCY(A1:A100,C2:C11)		
				=FREQUENCY([data_array], [bins_array])			
	A	B					
1	82			階級 (区間)	内容	度数	
2	69			9	0から9	C2:C11	
3	74			19	10から19		
4	87			29	20から29		
5	83			39	30から39		
6	55			49	40から49		
7	74			59	50から59		
8	63			69	60から69		
9	63			79	70から79		
10	69			89	80から89		
11	66			99	90から99		
12	79				10R x 1C		
13	72				合計		

図 1.8 階級を選択した状態

しなさい*4.

3.4 ヒストグラム

3.3 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。*5

1. リボンの Insert (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 1.9 のようにセルを選択します。
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 2.6 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。
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 1.11 のように選択し、図 1.12 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。

■演習 ??節の演習で作成した度数分布表に基づきヒストグラムをかけ。

*4 <http://www.qmss.jp/databank/medcomed/default.htm> の血糖値データを用いました。

*5 ヒストグラムもかき方が多々あるので各自の好みで。

C	D	E
階級 (区間)	内容	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 1.9 セルを選択した状態

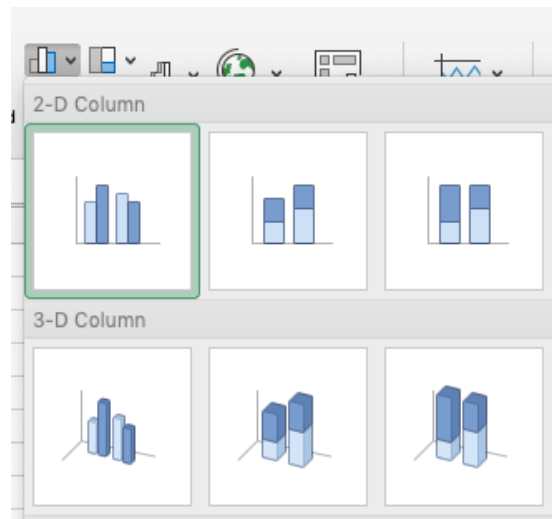


図 1.10 棒グラフを選択する

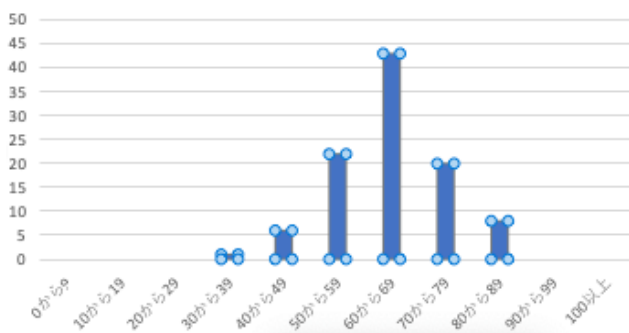


図 1.11 棒を選択した状態

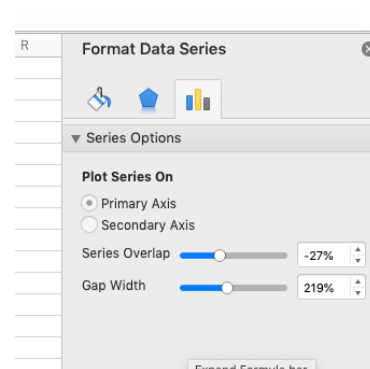


図 1.12 棒の太さを選ぶ画面

3.5 レポート提出

レポートテンプレートを参考に演習の結果にまとめ、kazuhsa.fujita@komatsu-u.ac.jp へ電子データとして提出しなさい。データ形式は pdf とする。レポートの提出期限は次回の講義日までとします

4 おまけ

ファイル形式

確率分布

ガウス分布
不偏分散

第 2 回

Excel による統計処理 2

1 目的

データを解析において重要なデータの可視化 (グラフ化) の方法を学ぶ.

2 原理

2.1 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, 折れ線グラフ, 円グラフ, 散布図を取り扱います.

■棒グラフ 棒グラフは大きさの比較に用いられます. 時間変化を把握するのに用いてはいけません.

■折れ線グラフ 値の時間変化の把握に用いられます. 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.

■円グラフ 割合をは把握するために用いられます. 扇型の角度は全体に対する割合を表します. 円グラフは合計で 100% にならなければなりません.

例えば, 表 1.1

■散布図 データの値同士の関係性を知りたいときがあります. 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 2.1) を示します. 散布図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，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
がありそうなのが分かります．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．

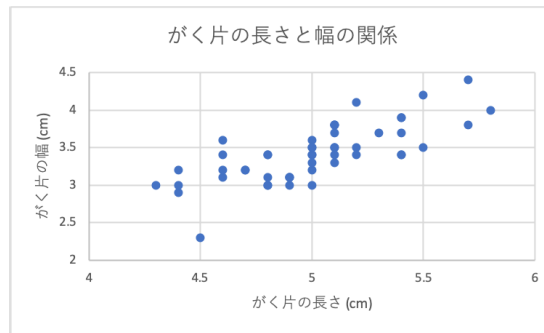


図 2.1 散布図

2.2 相関係数

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 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．

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

相関係数は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．

2.3 散布図と回帰直線

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 2.2 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。^{*1}

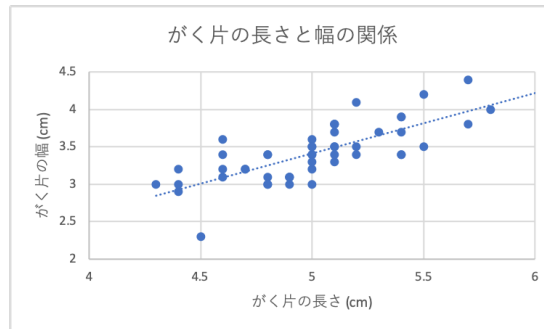


図 2.2 散布図と回帰直線

3 演習

3.1

3.2 折れ線グラフ

手順は前回のヒストグラムと同じです。

■演習 csv ファイルから折れ線グラフを作りなさい。ただし、横軸は 縦軸は にしなさい。

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

図 2.3

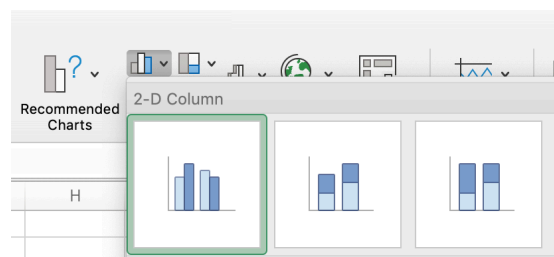


図 2.4

^{*1} 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。

3.3 円グラフ

■演習

3.4 折れ線グラフ

表に示すデータを折れ線グラフにします。横軸を、縦軸をとします。よく軸のデータと縦軸のデータを選びます。グラフの散布図の折れ線グラフを選びます。

■演習 csv ファイルから折れ線グラフを作りなさい。ただし、横軸は 縦軸は にしなさい。

3.5 散布図

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. 散布図にしたいデータを選びます (図 2.5).
2. 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??に示す散布図ボタンを押します。そうすると図 2.7 のような散布図が出来上がります。
3. 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 2.8) を選びます。
 - (b) ウィンドウの右端に出てくる軸の書式設定 (図 2.9) 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 2.10 のようになります。
4. 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 2.11) 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。

	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 2.5 セルを選択した状態

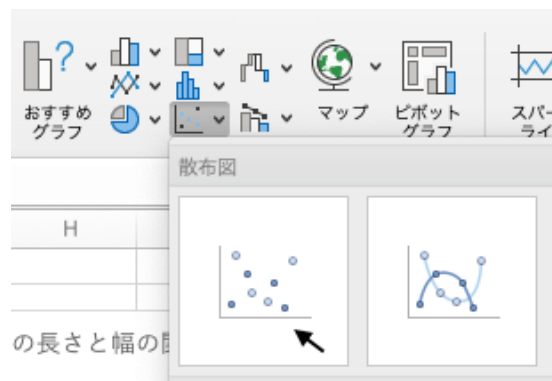


図 2.6 散布図を選択する

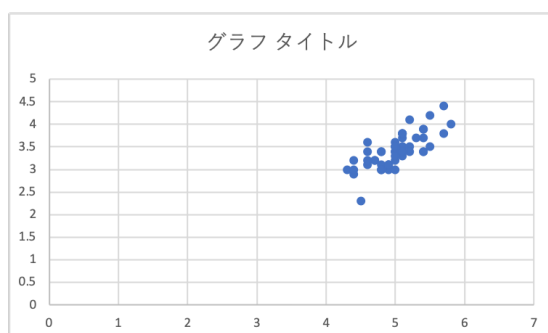


図 2.7 未完成な散布図

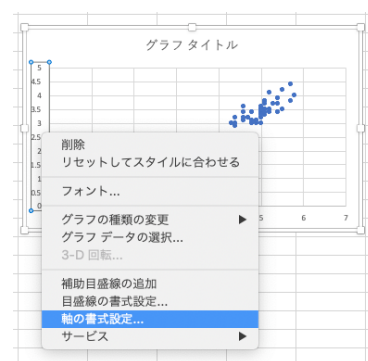


図 2.8 軸のメニュー

(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“がく片の幅 (cm)” にしましょう*2。同様の手順で横軸を“がく片の長さ (cm)” しましょう。そうすると、図 2.12 のようなグラフができます。

5. 必ずしもグラフにタイトル付ける必要はありませんが、せっかくなので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 2.1 のようなグラフになります。

■演習 csv ファイルから散布図をかきなさい。ただし、横軸は 縦軸は にしなさい。

*2 軸には単位を付けましょう。



図 2.9 軸の範囲の設定

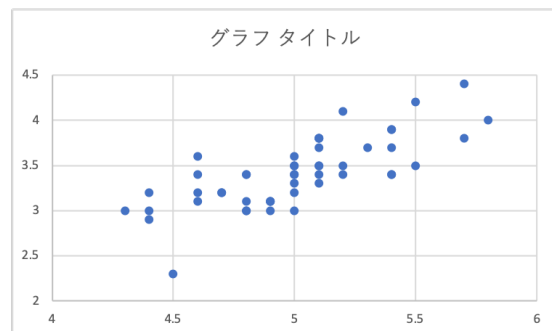


図 2.10 軸の範囲を変えた散布図

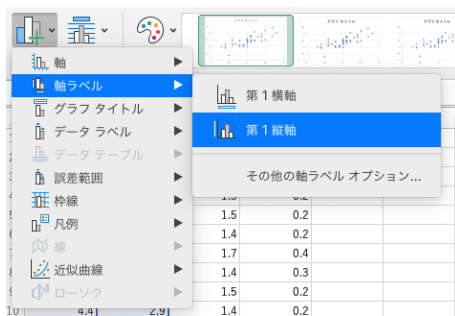


図 2.11 軸の範囲の設定

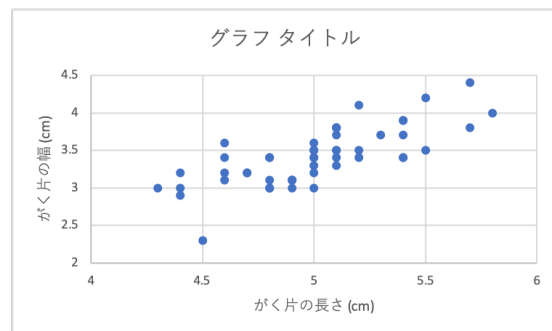


図 2.12 軸のラベルを付けた散布図

3.6 相関係数

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. 相関係数を入れるセルを選択します。
2. “=CORREL(” と入力します。
3. 相関係数を求めたいデータの列を 2 つ選びます。
4. “)” と入力します。そうすると、相関係数が表示されます。

■演習 csv データから相関係数を求めなさい。

3.7 回帰直線

先程作成した散布図に回帰直線を追加します。

1. リボンの“グラフのデザイン”を選びます。
2. 図 2.13 のように、近似曲線の線形予測を選びます。そうすると、図 2.2 のような回帰直線がかかります。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。

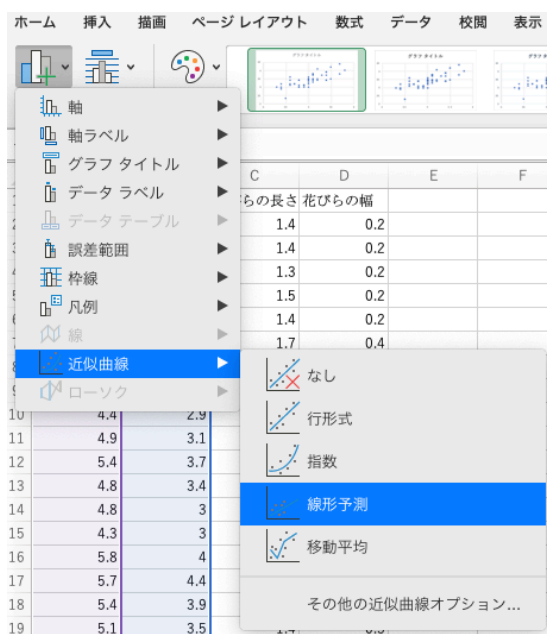


図 2.13 散布図

■演習 csv データから回帰直線のかかれた散布図を作成しなさい。

3.8 レポート提出

レポートは kazuhisa.fujita@komatsu-u.ac.jp へ電子データとして提出する。データ形式は pdf とする。レポートの提出期限は次回の講義日までとする。

4 おまけ

R, python

プロット R, matplotlib, gnuplot

有料なら matlab がある.