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.

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

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．

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

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．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．

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

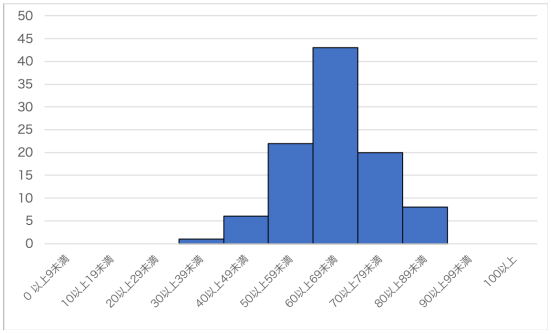

図 1.2　ヒストグラム

■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．

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

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．

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 xslx 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．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が，演習結果の保存は xslx 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. 総和を表示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．

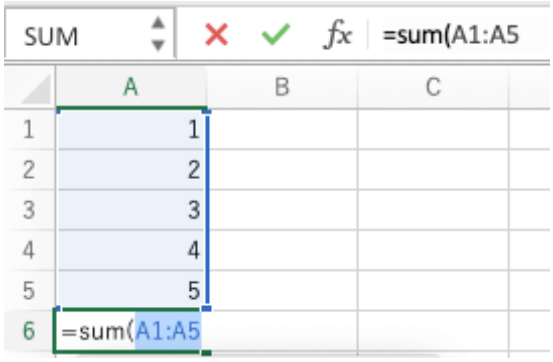

図 1.3　総和を計算したいセルを選択した状態.

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

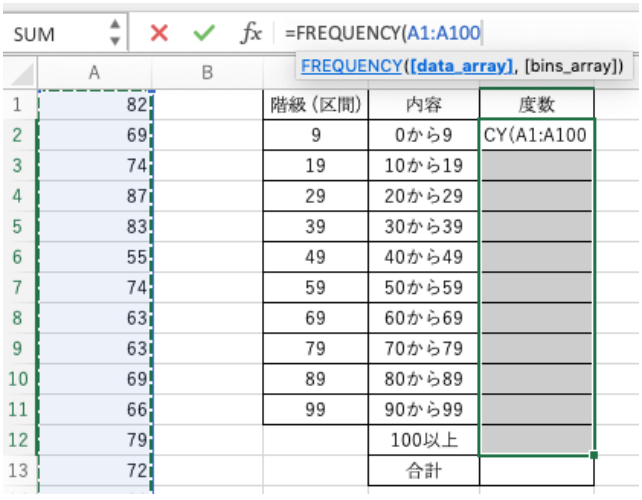

図 1.7　データを選択した状態

■**演習**　csv データから度数分布表を書きなさい．ただし，階級は としなさい．

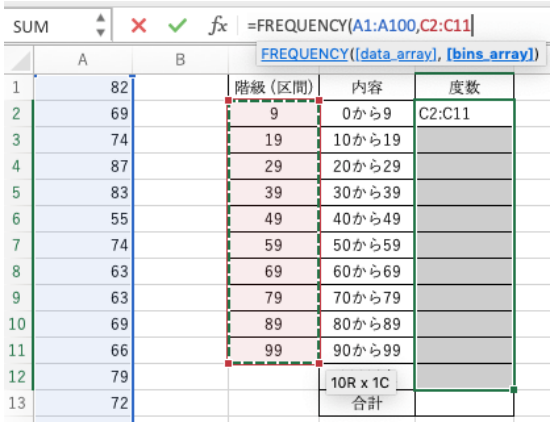

図 1.8　階級を選択した状態

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．[*4]

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. リボンにある棒グラフアイコンの横の “への字の逆” のようなボタンを押し，図 1.10 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　3.3 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] 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

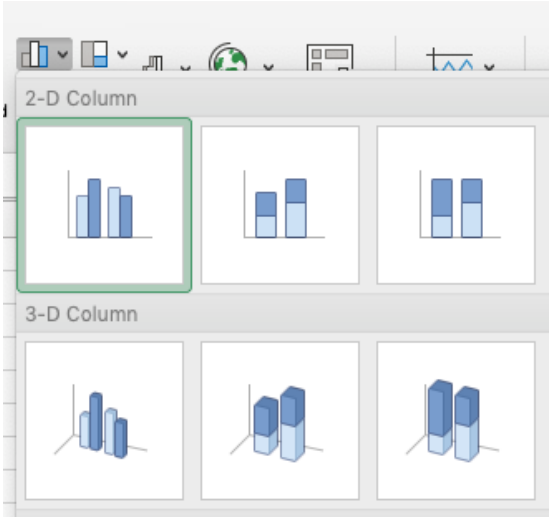

図 1.10　棒グラフを選択する

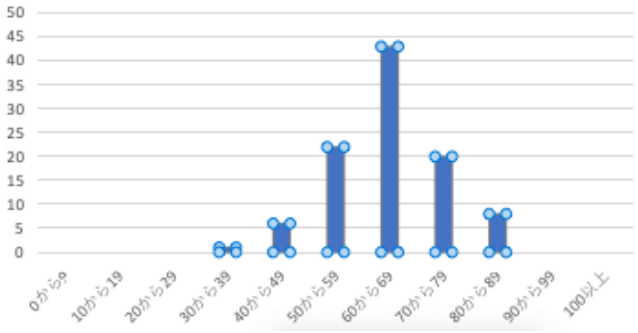

図 1.11　棒を選択した状態

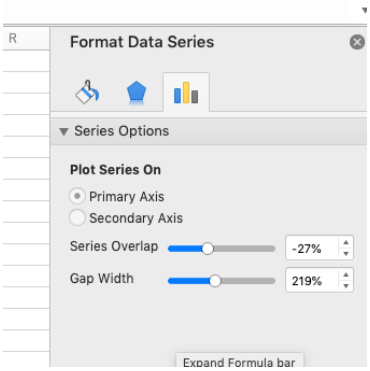

図 1.12　棒の太さを選ぶ画面

3.5　レポート提出

レポートテンプレートを参考にに演習の結果にまとめ，kazuhisa.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．例えば，身長と体重の関係です.

2.2　相関係数

散布図では，視覚的に値同士の関係性を見ることができます．しかし，定量的にどの程度関係しているか知りたいときには相関係数を使います．相関係数は次の式で計算されます．

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

相関係数は直線的な関係かどうかをみることができます．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みることができません．

正の相関負の相関

2.3　散布図と回帰直線

平均に戻る．回帰．

3　演習

3.1

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

図 2.1

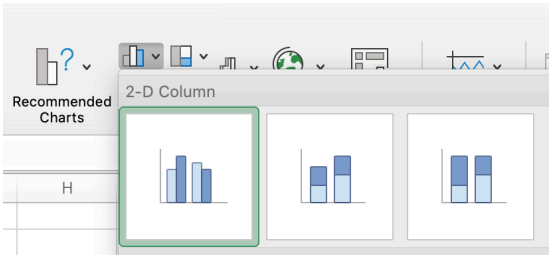

図 2.2

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．

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

■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．

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．

CORREL

■演習　csv データから相関係数を求めなさい．

3.7　回帰直線

■演習　csv データから散布図と回帰直線をかきなさい．

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．