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.

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

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.

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

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．

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．表に度数分布表の例を示します．

表 1.2: 度数分布表

階級	度数
0 以上 9 未満	0
10 以上 19 未満	0
20 以上 29 未満	0
30 以上 39 未満	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 109 未満	0
合計	100

この度数分布表を図で表したものがヒストグラム (図 1.1) 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．

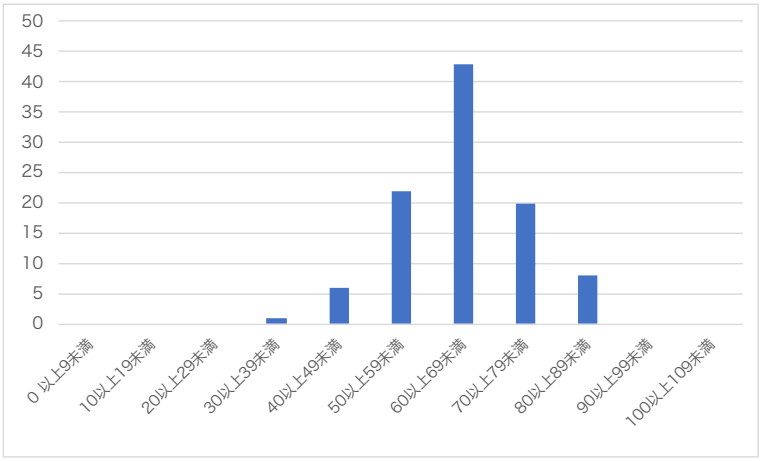

図 1.1: ヒストグラム

3　Excel 実験

3.1　総和，平均，中央値，分散

エクセル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選択する．[1]そうすると “(” の後ろにセル番号が入力される (図 1.2)．
4.) を入力する．

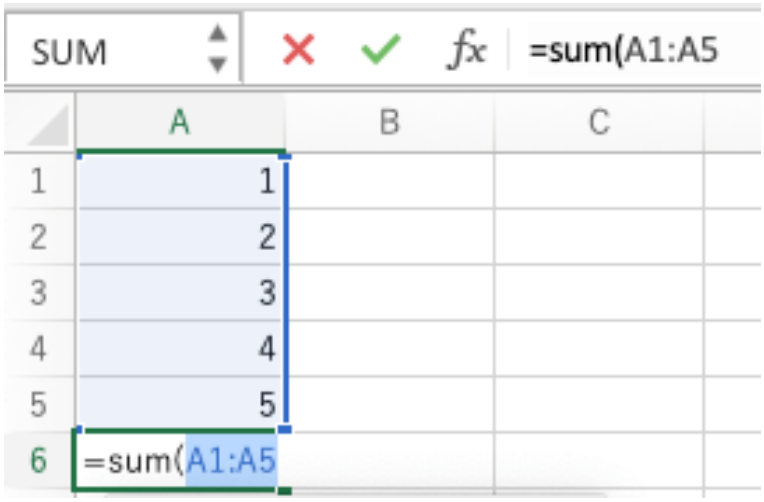

図 1.2: 総和を計算したいセルを選択した状態．

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**??**に示します[2]．

演習　csv の平均，分散，標準偏差を求めなさい．

演習　csv の平均，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．

[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．

表 1.3: 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

3.2　度数分布表とヒストグラム

図のような度数分布を作るにはどうすればよいでしょうか．まず階級を設定します．

演習　csv データから度数分布表を書け．

演習　度数分布表に基づきヒストグラムをかけ．

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

エクセルにはデータを解析するための関数が多く用意されている．今回はその一部を用い，データ解析の初歩を学ぶ．

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．

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．

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

2.3 散布図と回帰直線

平均に戻る．回帰.

3 実験

3.1 折れ線グラフ

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.

3.2 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.

3.3 相関係数

演習 csv データから相関係数を求めなさい.

3.4 共分散

演習 csv データから相関係数を求めよ.

3.5 演習

演習 csv データから散布図と回帰直線を書きなさい.

4 おまけ

R，python
プロット R，matplotlib，gnuplot
有料なら matlab がある.

5　レポートの出し方

提出は電子データで送る．データ形式は pdf もしくは docx とする．