

第 2 回 平均と分散・データの要約

A. データの要約

1. 100 個以上のデータを要約する

架空の 5 品種のミニトマトの着果数について（生物統計学__授業用データ集の中から「第 2 回データの要約」を参照）

（数字は例なので，トマトの着果数をニワトリの産卵数などに置き換えて考えてみてもよい）

① ぱっとデータを見て，何かいえるか
どの品種が一番多いか？

5 つの品種の順番は？

② 大きさの順に並べてみる 何かいえるか
どの品種が一番多いか？

5 つの品種の順番は？

③ 度数分布やヒストグラムを書いてみる

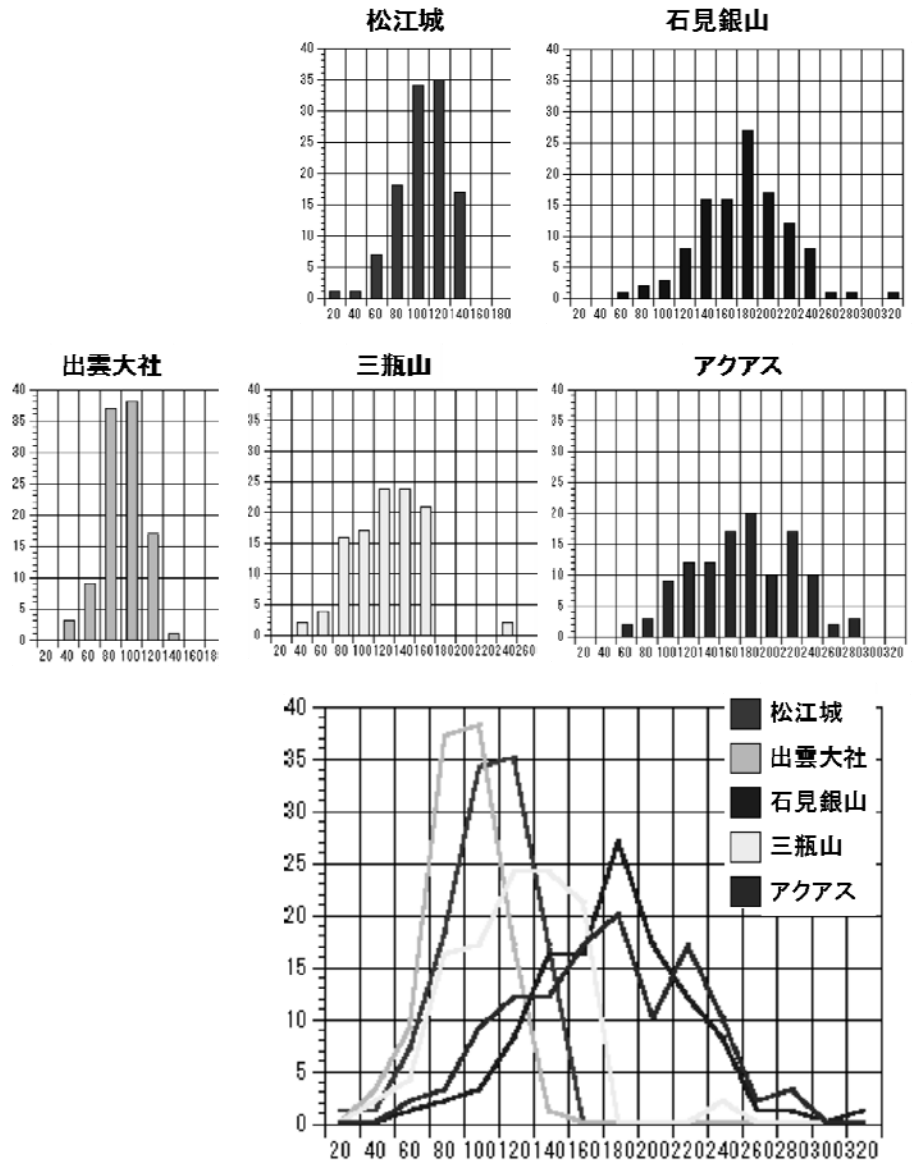
度数分布とヒストグラムの書き方（いくつに分級するべきか，区間の幅をどう決めるか）にはいくつかの方法がある．詳細はホームページを参照のこと．

★どのような傾向が読み取れるか？

度数分布

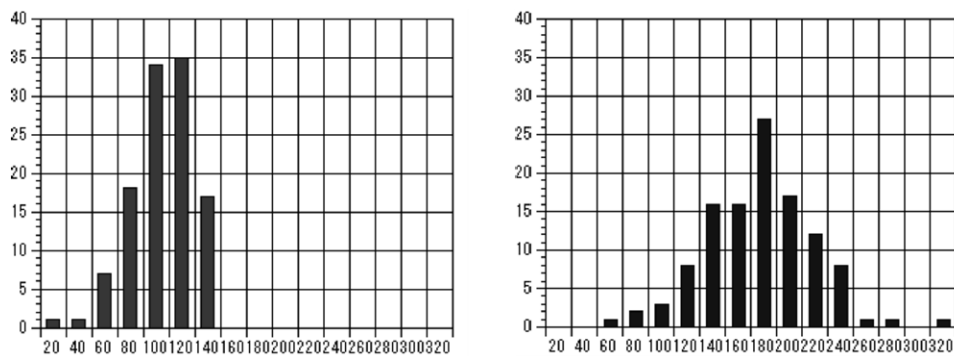
	松江城	出雲大社	石見銀山	三瓶山	アクアス
20	1	0	0	0	0
40	1	3	0	2	0
60	7	9	1	4	2
80	18	37	2	16	3
100	34	38	3	17	9
120	35	17	8	24	12
140	17	1	16	24	12
160	0	0	16	21	17
180	0	0	27	0	20
200	0	0	17	0	10
220	0	0	12	0	17
240	0	0	8	2	10
260	0	0	1	0	2
280	0	0	1	0	3
300	0	0	0	0	0
320	0	0	1	0	0
合計	113	105	113	110	117

ヒストグラム



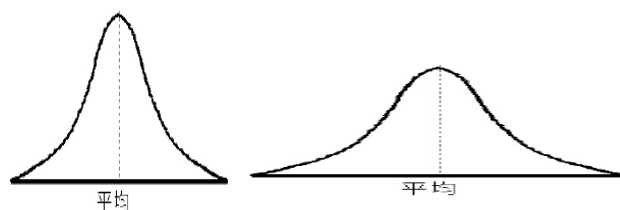
④ 代表値を考える

(), (), (), ()



	()	()	()
松江城 (左)	95	97	84
石見銀山 (右)	167.1	170	177

- ⑤ () を考える
(), (), (), ()
()



2. さまざまな統計量の定義

サンプル数を n , データを $x_1, x_2, x_3, x_4, \dots, x_n$ とする.

- ① 代表値 中心を表す値
()

観測数が n であり, すべての観測値を $x_1, x_2, x_3, x_4, \dots, x_{n-1}, x_n$ と表すとき, 平均 $\bar{x}$ は

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + \dots + x_{n-1} + x_n}{n} = \sum_{k=1}^n x_k \div n = \frac{\sum_{k=1}^n x_k}{n}$$

- () データを大きさの順に並べて
データが奇数個なら中央に位置するデータの値,
データが偶数個なら中央に位置する 2 つのデータの平均である

例 1 1 1, 1 4, 1 5, 1 7, 2 0 例 2 2, 4, 8, 1 0, 1 2, 1 6

- () 集めたデータの中で, 最も多く現れた値あるいは階級のこと.
名目データではモードを代表値とする

- ② ばらつきを表す値
偏差

()

$$V = \frac{S}{n-1} \quad \text{単位は元の値の 2 乗となる. } S \text{ は平方和である}$$

平方和 平均からの偏差の二乗の和

$$S = \sum_{i=1}^n (x_i - \bar{x})^2$$

()

$$S.D. = \sqrt{V}$$

()

最大値と最小値の差

もっとも両極端な値を使っているので信頼は低い

両方とも異常値であることもある.

四分位範囲

第 3 四分位点と第 1 四分位点の差である.

第 1 四分位点はデータを大きさの順に並べて, 最小値から数えて, 全体の 4 分の 1 番目に当たるデータの値である. 第 3 四分位点は同じく全体の 4 分の 3 番目に当たるデータの値である. メジアンは第 2 四分位点と一致する.

()

標準偏差を平均で割ったものでばらつきの大きさを相対的に評価する時に使う.

おおまかに 5%以下だと小さいばらつき, 6-14%はふつうのばらつき, 15%以上は大きいばらつきと考える

B. 代表値の計算の実際

次のデータから平均, 分散, 標準偏差, メジアン, レンジ, 変動係数 (%) を計算せよ.

例題 山を調査して発見したツキノワグマの頭数 3, 5, 6, 8, 11 頭

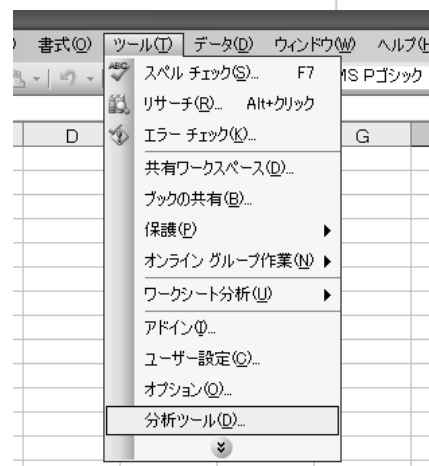
1. エクセルの関数を使う計算方法

データ				=AVERAGE(C4:C8)
3	平均	6.6		=VAR(C4:C8)
5	分散	9.3		=STDEV(C4:C8)
6	標準偏差	3.04959		=MEDIAN(C4:C8)
8	メジアン	6		=MAX(C4:C8)-MIN(C4:C8)
11	レンジ	8		=STDEV(C4:C8)/AVERAGE(C4:C8)*100
	変動係数	46.20591		

2. エクセルの分析ツールを使う計算方法

① ツール→分析ツール

(エクセル 2007 ではデータ→データ分析)



- ② 基本統計量を選ぶ
- ③ 入力範囲にデータの範囲を入れる。データの最初にラベル（データ名）があるときはチェックを入れる。出力先を指定する。

- ④ 指定した出力先に計算結果が出る。

The image shows two Excel dialog boxes and a summary table. The 'データ分析' (Data Analysis) dialog box has '基本統計量' (Basic Statistics) selected. The '基本統計量' (Basic Statistics) dialog box shows the input range as '\$B\$3:\$E\$28', data in columns, and output to '\$G\$4'. The 'データ' (Data) table summarizes the statistics for the range B3:E28.

データ	
平均	13
標準誤差	1.47196
中央値（メジアン）	13
最頻値（モード）	#N/A
標準偏差	7.359801
分散	54.16667
尖度	-1.2
歪度	0
範囲	24
最小	1
最大	25
合計	325
標本数	25

次のデータから平均、分散、標準偏差、メジアン、レンジ、変動係数（%）を計算せよ。

- ① 農園で収穫したミカンの数 11, 13, 15, 17, 22, 27, 36, 41, 55 個

- ② 根の長さ 1.4, 2.7, 5.9, 6.3, 10.5, 13.1, 15.0, 18.2, 22.7cm

★ 分散の単位は cm² のように 2 乗がつくことが多い。エクセルで 2 乗を示すのは以下のようにすればよい。

The image shows an Excel context menu with 'セルの書式設定(F)...' (Format Cells) selected. The 'セルの書式設定' (Format Cells) dialog box is open, showing the 'Number' tab. The 'Fraction' section has '上付き' (Superscript) checked, which is circled with a red line.

4. データの要約について

- 1) データの全体的な傾向を表す表にまとめる 大きさの順に並べる, 度数分布
- 2) () などの図を書く
- 3) 平均など () 化の傾向を示すような値を求める
- 4) () を評価する値を求める

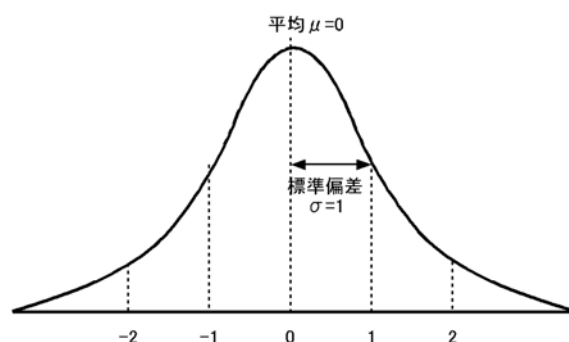
D. 要約されたデータから何を読みとるか？

それぞれの統計量は何を意味するのか？ どのような利用価値があるのか？

1. 中心化の傾向を表す代表値として

① 平均

平均は分布に () のないとき
() にもっとも有用な統計量である.
とくに母集団が () を示すときに非常に有用である.



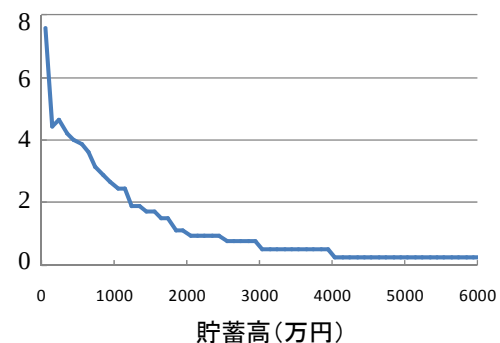
分布がゆがんでいるとき, 平均の利用には注意が必要である

② メジアン

右の給与分布の場合, 平均値とメジアンにはずれがある. 一般的な年収の人を考えると, 年収が平均の人とメジアンの人を選べば, それぞれ別のタイプの人になる.

メジアンは正規性のないデータあるいは標本全体を同時に調査できないようなデータ (性成熟日数, 生存日数, 電球の寿命のような打ち切り標本, 以上・以下を含む標本など) などの形質に用いる. データに異常値があるときのように, 極端に離れた値を含む標本では平均よりメジアンを用いる方が無難である.

(%) 日本の世帯別貯蓄高の分布 (2007年)



③ モード

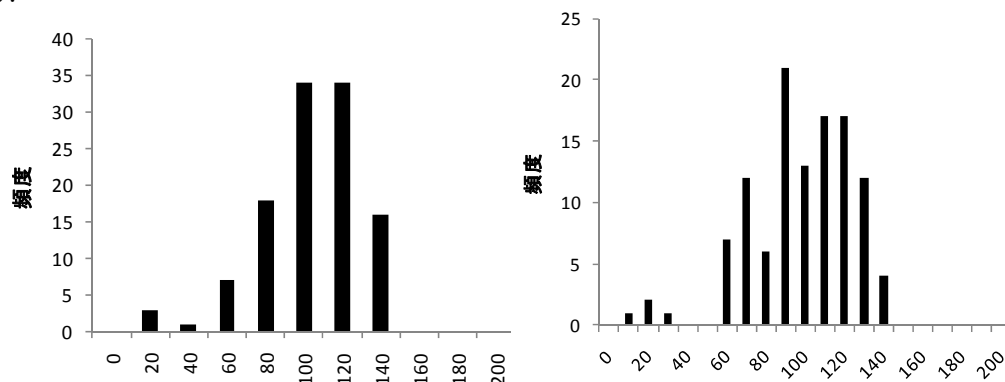
色などの名目 (質的) データならモードを使うしかない.

右の表はある交差点を通過した車の色を調べた結果である. この場合, モードは白となる.

白	47
黒	10
黄	8
赤	7
緑	2
青	1
その他	6

量的データでは度数分布の階級のとりかたによってモードが変わる。

下は最初に使った仮想のミニトマト品種松江城を階級のとり方の異なる2つのヒストグラムにしたものである。



左のヒストグラムではモードは 81~100 と 101~120 の2つある。右のヒストグラムではモードは 81~90 である。このように量的データではデータのまとめ方次第でモードが変わるので、モードの使用には注意が必要となる。

2. ばらつきを評価する指標として

分散・標準偏差

たいていの場合、平均と分散（標準偏差）の2つをばらつきの指標としたらよい

レンジ

計算が簡単なので、標準偏差の代用とすることもある。異常値に左右されやすいので、信頼性は低い

四分位範囲

あまり使わないが、メジアンといっしょに分布のゆがんだデータで利用される。

D. 宿題

宿題はホームページ <http://www.ipc.shimane-u.ac.jp/food/kobayasi/biometrysyukudai.html> を見てください。

参考 数字の丸め方：四捨五入の注意点

一般的な四捨五入では、丸めるべき桁の数字が 0～4 ならば切り捨て、5～9 ならば切り上げてきた。しかし、末尾が 5 で終わるデータのとき、末尾の数字をいつも切り上げると、厳密には下の表からわかるように四捨五入では、ほんのわずかに大きい方に数字が偏る。

元の数	5.00	5.01	5.02	5.03	5.04	5.05	5.06	5.07	5.08	5.09	5.045
四捨五入 すると	5.0	5.0	5.0	5.0	5.0	5.1	5.1	5.1	5.1	5.1	5.05
差	0	-0.01	-0.02	-0.03	-0.04	0.05	0.04	0.03	0.02	0.01	平均

そこで末尾が 5 で終わり、その数字を丸める場合には丸めた数字の末尾が偶数になるように丸める JIS, ISO 式四捨五入を使う方がよい。この方法なら丸めることによる誤差は最小となる。

元の数	5.00	5.01	5.02	5.03	5.04	5.05	5.06	5.07	5.08	5.09	5.045
四捨五入 すると	5.0	5.0	5.0	5.0	5.0	5.0	5.1	5.1	5.1	5.1	5.04
差	0	-0.01	-0.02	-0.03	-0.04	-0.05	0.04	0.03	0.02	0.01	平均

元の数	5.10	5.11	5.12	5.13	5.14	5.15	5.16	5.17	5.18	5.19	5.145
四捨五入 すると	5.1	5.1	5.1	5.1	5.1	5.2	5.2	5.2	5.2	5.2	5.15
差	0	-0.01	-0.02	-0.03	-0.04	0.05	0.04	0.03	0.02	0.01	平均

上のように、5.05 の場合は切り下げ、5.15 の場合は切り上げるのでこの 2 つで偏りを相殺できる。

JIS, ISO 式四捨五入の規則は以下の 3 つである。

- ① 一番近い丸め結果候補が 1 つだけなら、その数に丸める。
- ② 一番近い丸め結果候補が 2 つある場合は、末尾が偶数のものに丸める。
- ③ 丸め処理は 1 段階で行わなければならない。

この方法で以下に例示したいくつかの数字を整数第 1 位に丸める。

1000.4999→1000, 1000.5→1000, 1000.5001→1001, 1001.5→1002

練習 次の数値を JIS, ISO 式四捨五入に基づいて、小数第 1 位に丸めよ。

3.589, 100.03, 0.045, 8.85

答え

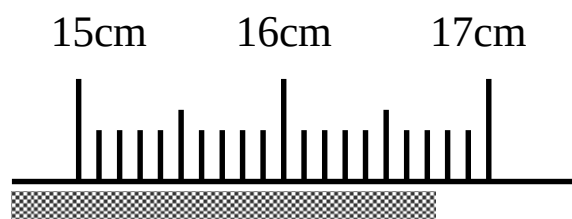
3.6, 100.0, 0.0, 8.8

参考．データ計算における注意：計算結果を何桁まで表示すべきか？ 有効数字の概念

計算結果を何桁まで表示できるかは有効数字によって決まる．ただ有効数字すべてを記載すると桁数が多くて煩雑なときは実用的な桁数に丸めてもよいこともある．

確実な数字とさらにいくぶん不確実な数字をその下に一桁付け加えて、それらの数字を**有効数字**と呼ぶ．たとえば 1mm 目盛りの物差しで長さを測ると、最小目盛りの 10 分の 1 まで測定できる．10.5mm というように測定できる．この場合、最後の桁の数字である 5 にはいくらか不確実である．有効数字で表示された測定値、身長 160cm, 160.0cm, 160.00cm はそれぞれ同じものではない．160cm よりも 160.00cm の方がより正確である．160cm という表記では 159.5～160.5 までの範囲あるいはもっとそれより大きい範囲内であることを示すが、160.00cm ならばそれよりずっと狭い範囲内であることを示す．

左の棒を物差しで測った場合、目盛りは 1mm 単位までであるが、目分量で 0.1mm 単位まで測ることができる．16.7mm とするより、16.72mm と測る方が正確である．



有効数字どうしの計算の規則は 2 つある．加減算のときはいちばん粗い精度となる桁に揃える．乗除算のときは有効数字の数がいちばん少ないデータに有効数字の数を合わせる．この規則より厳密な方法もあるが煩雑なのでここでは紹介しない（ホームページ参照）．

ここでは平均や標準偏差を計算するので、有効数字はデータの位に合わせるとよい．小数第 1 位のデータについて平均を出すときは小数第 2 位を四捨五入し、小数第 1 位までを最終的に表記する．分散は有効数字の個数を合わせるとよい．

ただし途中の計算ではできるだけ多くの桁数までとり（パソコン上の計算では何もしなければよい）、最後に四捨五入すること．エクセルで何段回かにかけて計算するときは参照座標を使って、途中の結果を四捨五入しないで代入すべきである．

例 カキの葉の長さ 16.3, 17.3, 17.6, 18.5, 18.9, 19.3, 20.7cm

平均、標準偏差のときは元のデータの桁にあわせる

$$\text{平均} = \frac{128.6}{7} = 18.37143 \dots \approx 18.4 \text{ cm}$$

分散は元のデータの有効数字の数にあわせる

$$\text{分散} = 2.102381 \dots \approx 2.10 \text{ cm}^2$$

なお人数のような離散データの平均ではデータより多く桁を取る方がよい場合がある．子供の数は整数で示されるが、ある都市の世帯平均の子供の数を 3.56 人と表記するのは問題がない．子供 1 人という場合、小数第 1 位以下が不確実な数字ではなく、1.000...と見なせるからである．しかし、3 世帯からしか調べなかったら、小数第 1 位までしか表記してはいけない．1.666...という割り切れない数字が得られても 1.7 とする．なぜなら、3 世帯のデータで子供 1 人が増減すると、0.333...だけ数値が増減する．このことから小数第 2 位以下の数字は意味がないことがわかる．