実験計画学
2006 年 10 月 10 日

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　5 品種のイネの 1 穂穎花数について

①　ぱっとデータを見て，何かいえるか

②　大きさの順に並べてみる　何かいえるか

③　分布を書いてみる

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

→書き方にはいくつかの方法がある．ホームページを参照のこと

④　代表値を考える

（　　），（　　），（　　）

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$x_1, x_2, x_3, x_4, \cdots, x_n$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 $x_1, x_2, x_3, x_4, \cdots\cdots, x_{n-1}, x_n$ と表すとき，平均 $\bar{x}$ は

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

（　　）

（　　）

② ばらつきを表す値

偏差

（　　）

$$V = \frac{S}{n-1}$$

平方和

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

（　　）

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

（　　）

（　　）

（　　）

★　別のデータについて，もう一度データの要約をやってみよう．

針に糸を通す時間									
8	26	11	16	21	15	13	39	18	10
8	25	5	14	35	5	6	43	22	21
49	12	4	40	19	10	15	35	7	33
9	6	20	6	21	1	14	15	7	16
19	7	6	32	15	9	27	13	23	18
19	17	12	6	30	5	6	15	23	9
13	8	10	31	24	8	16	11	15	2
13	47	13	28	8	6	8	7	40	13

①　上のデータを入力する．
②　大きさの順に並び替える
③　度数分布とヒストグラムを書く
④　統計量を求める．

平均　　メジアン
分散　　標準偏差
最大値　　最小値　　レンジ
第 1 四分位量，第 3 四分位量　両者の差が四分位範囲
変動係数

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			
3		平均	6.6
5		分散	9.3
6		標準偏差	3.04959
8		メジアン	6
11		レンジ	8
		変動係数	46.20591

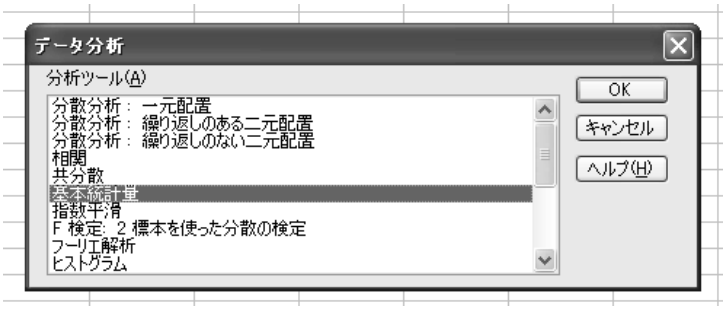
```
=AVERAGE(C4:C8)
=VAR(C4:C8)
=STDEV(C4:C8)
=MEDIAN(C4:C8)
=MAX(C4:C8)-MIN(C4:C8)
=STDEV(C4:C8)/AVERAGE(C4:C8)*100
```

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

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．

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

３．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．

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るべきである．

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．

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（　　）的傾向を示すような値を求める

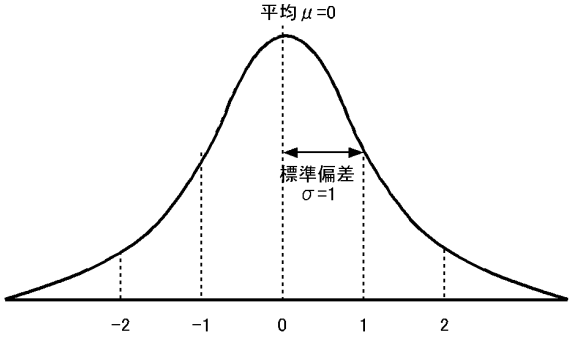
４）　（　　）を評価する値を求める

D．要約されたデータから何を読みとるか？

１．　それぞれの統計量は何を意味するのか？どういう利用価値があるのか？

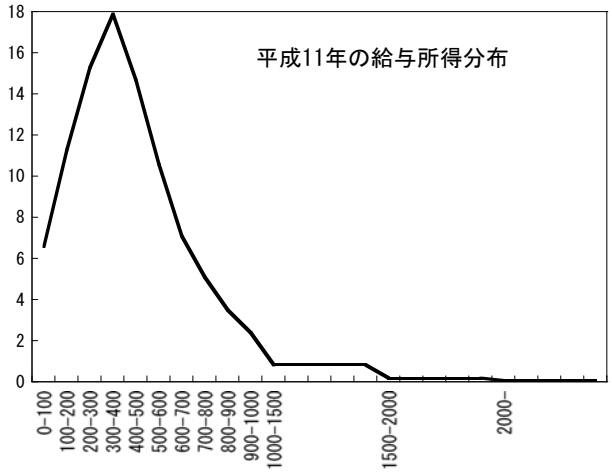
①　中心的傾向を表す代表値として

平均

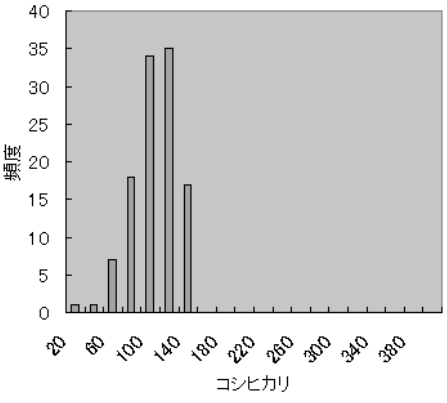


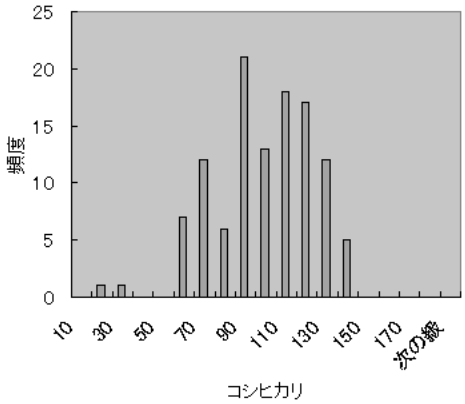
メジアン

モード



	コシヒカリ	日本晴	New PlantIR72		タカナリ
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





② ばらつきを評価する指標として

分散・標準偏差

レンジ

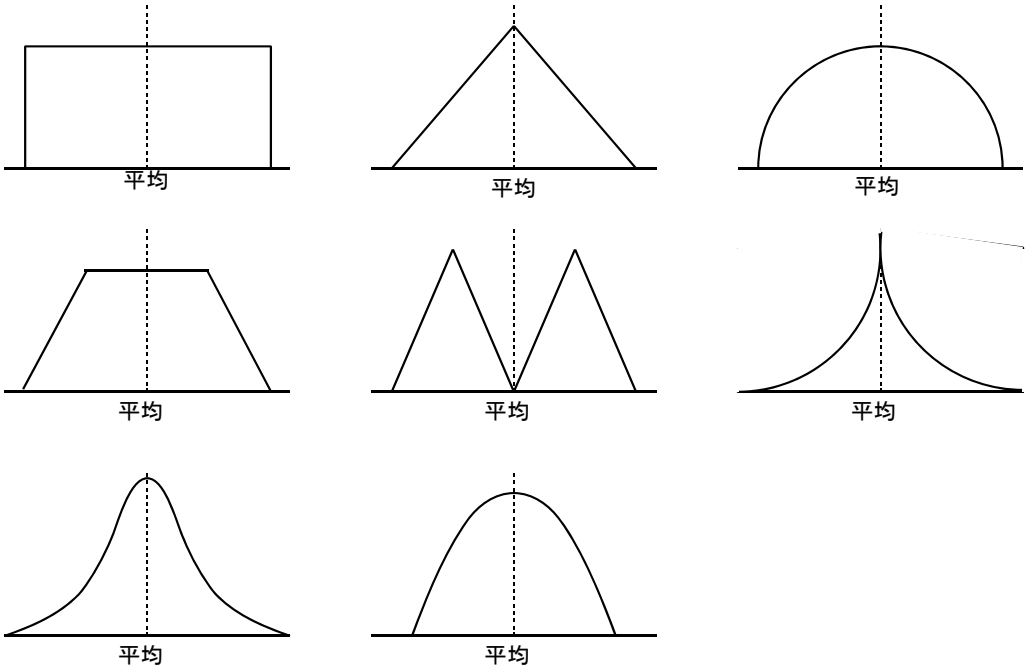
四分位範囲

２．要約する過程でどのくらいの情報が失われたのか？

では，逆に要約した情報からもとの分布が再構築できないだろうか？

３．種々の統計量から分布を再構築してみよう

★何らかのデータについてどのような分布が想定できるか.



D．宿題

１．前回の宿題２．について，今回学んだデータの要約の仕方を利用して，エクセルの２種類の計算方法を用いて，平均，メジアン，分散，標準偏差，レンジ，変動係数を求めよ．さらに前回の宿題で作成したヒストグラム 3 種類それぞれについてモードを求めよ．有効数字にも注意すること．

２．データの分布の形にはいろいろ考えられる．次回以降の授業では二項分布，ポアソン分布，正規分布について学ぶ．予備的に次の調査を行う．

① 二項分布に従うと考えられる以下の現象のうち，一つを選んで自分で実際に最低でも 25 回以上実験・調査せよ．その結果を度数分布になおし，ヒストグラム（階級範囲は１つずつとし，0 と 1 をいっしょの階級にしない）を書け．さらに平均，分散を計算せよ．集めたデータからわかったことを箇条書きで書け．

★ さいころを 10 回振って，そのうち 1 が何回出るか？

★ トランプから 1 枚のカードを抜く．元に戻して，再びよく混ぜる．10 回カードを引いて，ハートが何枚出るか．

★ 麻雀牌から 1 枚の牌を抜く．元に戻して，よく混ぜる．10 回牌を引いて，字牌が何枚出るか．

★ 10 本の鉛筆を用意し，1 本だけ赤鉛筆を入れる．よく混ぜて 1 本を取り出す．取り出した後，その鉛筆は元に戻す．これを 10 回行い，赤鉛筆を何回引いたか．

② めったに起こらないことはポアソン分布に従うと考えられる．以下の現象のうち，一つを選んで自分で実際に実験・調査せよ．その結果を度数分布になおし，ヒストグラム（階級範囲は１つずつとし，0 と 1 をいっしょの階級にしない）を書け．さらに平均，分散を計算せよ．集めたデータからわかったことを箇条書きで書け．

★ ここ 100 年間日本において火山の噴火した回数（理科年表でしらべることができる）．

★ ここ 100 年間において，世界で大地震の起こった回数（理科年表でしらべることができる）．

★ 100 個入りのお菓子に不良品に入っている数を 100 袋について調べる．

★ 島根県住民がここ 100 回の宝くじの特等に当たった人数．

★ 100 日間における島根県での交通事故の発生件数
警察庁のホームページ（http://www.npa.go.jp）から統計→交通事故発生状況→１．死者日報
全国で見ると毎日，交通事故死があるから，県レベルでみた方がよい．

提出締め切りは１０月１６日（月）午後１時までに生物資源科学部 2 号館 204 室に提出のこと．

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