

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

(実験計画学授業用データ集を参照)

(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よい)

アジア(データ数113)

56	125	130	98	27	119	65	129	140	113	112	106
52	60	65	84	62	124	59	107	131	88	109	64
64	108	74	111	101	87	83	81	105	109	97	89
60	85	70	118	108	134	62	119	92	92	84	
17	83	66	111	93	95	83	122	95	100	117	
98	97	114	107	72	91	88	104	68	114	60	
122	84	103	73	85	67	59	131	118	118	103	
101	108	113	133	97	127	71	91	110	87	75	
128	120	84	116	102	103	81	87	130	129	86	
85	88	81	73	113	64	62	130	107	127	114	

アフリカ(データ数105)

80	96	22	78	73	67	122	65	89	44	93	
84	111	81	73	96	84	107	47	67	78	71	
64	78	65	36	55	82	102	77	82	93	78	
71	80	71	75	95	89	108	106	84	74	88	
81	80	55	57	100	103	95	83	83	83	43	
93	94	89	81	74	104	107	93	63	81		
107	85	73	40	104	85	99	61	61	98		
104	92	47	104	107	77	69	82	94	69		
77	91	75	111	102	83	115	88	79	79		
60	76	78	86	47	71	109	94	62	76		

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?

アジア(データ数113)

17	62	70	83	86	92	100	107	111	118	127	133
27	62	71	83	87	92	101	107	112	118	127	134
52	64	72	83	87	93	101	107	113	118	128	140
56	64	73	84	87	95	102	108	113	119	129	
59	64	73	84	88	95	103	108	113	119	129	
59	65	74	84	88	97	103	108	114	120	130	
60	65	75	84	88	97	103	109	114	122	130	
60	66	81	85	89	97	104	109	114	122	130	
60	67	81	85	91	98	105	110	116	124	131	
62	68	81	85	91	98	106	111	117	125	131	

アフリカ(データ数105)

22	57	67	74	78	81	84	89	95	104	109	
36	60	69	74	78	81	84	91	95	104	111	
40	61	69	75	78	81	84	92	96	104	111	
43	61	71	75	78	82	85	93	96	104	115	
44	62	71	76	79	82	85	93	98	106	122	
47	63	71	76	79	82	86	93	99	107		
47	64	71	77	80	83	88	93	100	107		
47	65	73	77	80	83	88	94	102	107		
55	65	73	77	80	83	89	94	102	107		
55	67	73	78	81	83	89	94	103	108		

オーストラリア(データ数113)

188	177	111	302	201	222	151	163	104	187	212	142
169	158	217	214	176	201	147	138	178	171	124	129
131	127	123	197	150	182	192	132	176	152	268	79
233	162	149	190	152	144	196	127	228	111	236	
186	117	203	177	178	114	138	169	159	232	188	
183	144	170	211	151	48	126	222	170	128	222	
167	165	175	202	78	165	155	190	178	178	166	
135	98	152	173	98	128	224	175	161	116	205	
148	112	95	177	189	172	166	129	196	259	115	
126	196	194	128	194	194	214	177	208	208	143	

南アメリカ(データ数110)

131	160	56	117	97	138	111	145	139	158	140	
129	115	135	121	67	96	142	129	110	129	134	
106	113	134	72	152	234	147	136	106	78	118	
95	71	114	120	155	108	143	99	144	65	93	
152	76	150	96	118	134	117	82	102	110	102	
152	101	134	58	224	111	93	78	124	134	69	
145	76	134	83	130	105	120	91	144	103	84	
144	68	130	116	147	122	118	73	94	102	73	
153	37	94	91	143	89	137	61	138	130	73	
83	94	145	51	74	40	145	61	139	142	51	

北アメリカ(データ数117)

177	96	175	220	169	178	262	241	193	219	191	162
207	152	78	154	209	100	206	231	96	197	134	126
179	162	84	218	191	101	104	208	96	157	163	112
173	111	208	83	159	222	119	202	233	207	127	147
109	139	116	210	115	271	160	129	152	184	90	88
151	154	141	77	211	225	188	107	119	191	146	82
143	132	205	144	138	225	224	118	148	128	131	116
162	214	165	238	213	213	242	176	179	162	163	
123	174	215	189	54	136	273	142	132	226	167	
188	163	154	224	48	180	154	222	80	195	178	

オーストラリア(データ数113)

48	114	128	142	152	165	173	178	190	201	214	259
78	115	128	143	152	166	175	178	190	201	217	268
79	116	128	144	152	166	175	178	192	202	222	302
95	117	129	144	155	167	176	182	194	203	222	
98	123	129	147	158	169	176	183	194	205	222	
98	124	131	148	159	169	177	186	194	208	224	
104	126	132	149	161	170	177	187	196	208	228	
111	126	135	150	162	170	177	188	196	211	232	
111	127	138	151	163	171	177	188	196	212	233	
112	127	138	151	165	172	178	189	197	214	236	

南アメリカ(データ数110)

37	68	78	93	102	111	118	130	136	143	150	
40	69	78	94	102	111	120	130	137	144	152	
51	71	82	94	102	113	120	131	138	144	152	
51	72	83	94	103	114	121	134	138	144	152	
56	73	83	95	105	115	122	134	139	145	153	
58	73	84	96	106	116	124	134	139	145	155	
61	73	89	96	106	117	129	134	140	145	158	
61	74	91	97	108	117	129	134	142	145	160	
65	76	91	99	110	118	129	134	142	147	224	
67	76	93	101	110	118	130	135	143	147	234	

北アメリカ(データ数117)

48	96	115	129	143	154	163	178	191	208	219	233
54	96	116	131	144	154	163	178	191	208	220	238
77	96	116	132	146	157	165	179	193	209	222	241
78	100	118	132	147	159	167	179	195	210	222	242
80	101	119	134	148	160	169	180	197	211	224	262
82	104	119	136	151	162	173	184	202	213	224	271
83	107	123	138	152	162	174	188	205	213	225	273
84	109	126	139	152	162	175	188	206	214	225	
88	111	127	141	154	162	176	189	207	215	226	
90	112	128	142	154	163	177	191	207	218	231	

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．

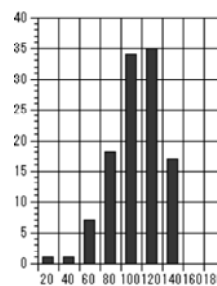
★どのような傾向が読み取れるか？

度数分布

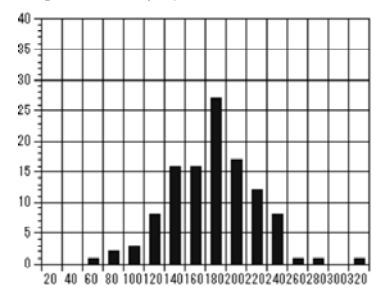
	アジア	アフリカ	オーストラリア	南アメリカ	北アメリカ
20	1	0	0	0	0
40	1	3	0	2	0
60	7	9	1	4	2
80	18	37	2	16	3
100	34	38	3	17	9
120	35	17	8	24	12
140	17	1	16	24	12
160	0	0	16	21	17
180	0	0	27	0	20
200	0	0	17	0	10
220	0	0	12	0	17
240	0	0	8	2	10
260	0	0	1	0	2
280	0	0	1	0	3
300	0	0	0	0	0
320	0	0	1	0	0
合計	113	105	113	110	117

ヒストグラム

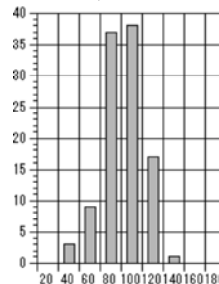
アジア



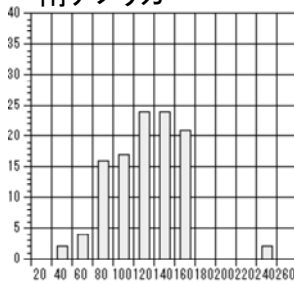
オーストラリア



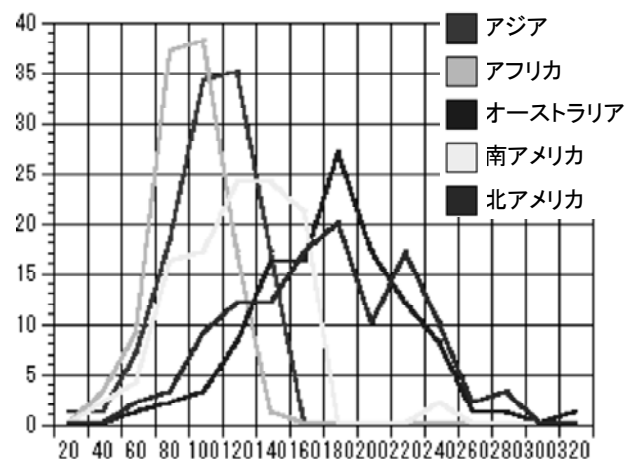
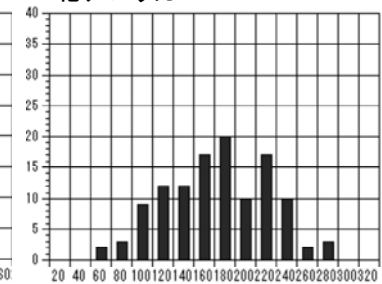
アフリカ



南アメリカ

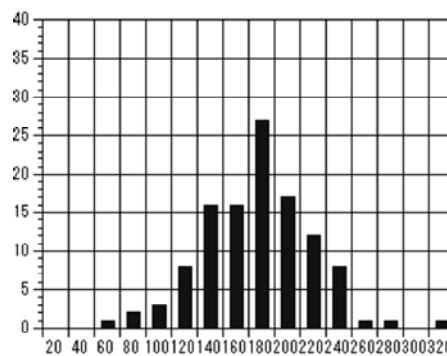
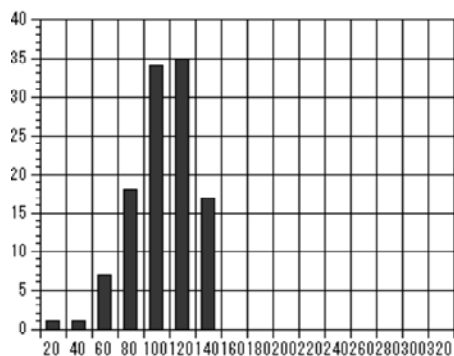


北アメリカ



④ 代表値を考える

(), (), ()



	()	()	()
アジア (左)	95	97	84
オーストラリア (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 11, 14, 15, 17, 20 例2 2, 4, 8, 10, 12, 16

() 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

()

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

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

★ 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.

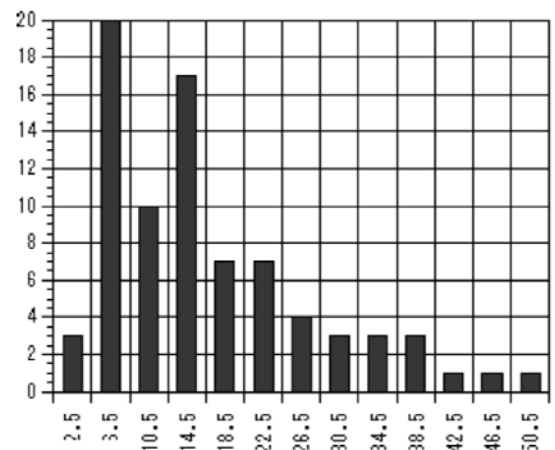
針に糸を通す時間										
	8	26	11	16	21	15	13	39	18	10
	8	25	5	14	35	5	6	43	22	21
	49	12	4	40	19	10	15	35	7	33
	9	6	20	6	21	1	14	15	7	16
	19	7	6	32	15	9	27	13	23	18
	19	17	12	6	30	5	6	15	23	9
	13	8	10	31	24	8	16	11	15	2
	13	47	13	28	8	6	8	7	40	13

- ① 上のデータを入力する.
- ② 大きさの順に並び替える

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

- ③ 度数分布とヒストグラムを書く

階級	階級値	度数
-4.5	2.5	3
4.5-8.5	6.5	20
8.5-12.5	10.5	10
12.5-16.5	14.5	17
16.5-20.5	18.5	7
20.5-24.5	22.5	7
24.5-28.5	26.5	4
28.5-32.5	30.5	3
32.5-36.5	34.5	3
36.5-40.5	38.5	3
40.5-44.5	42.5	1
44.5-48.5	46.5	1
48.5-	50.5	1



- 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

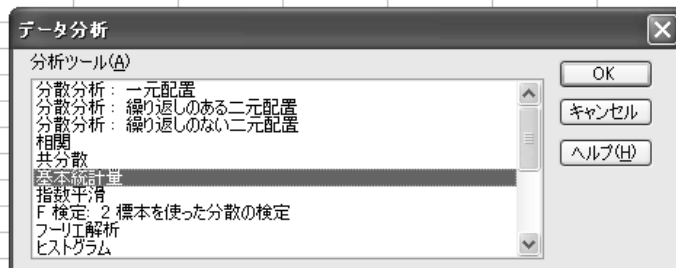
① ツール→分析ツール

（エクセル 2007 ではデータ→データ分析）

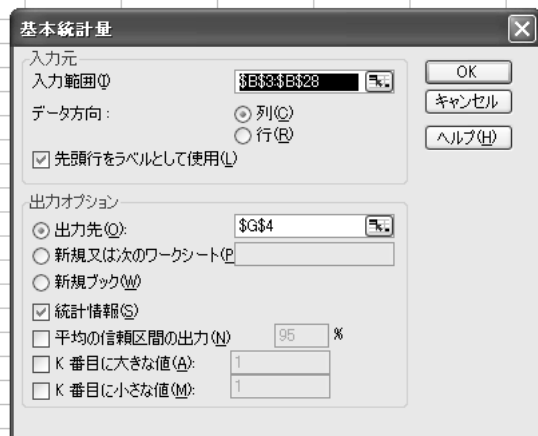


② 基本統計量を選ぶ

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。



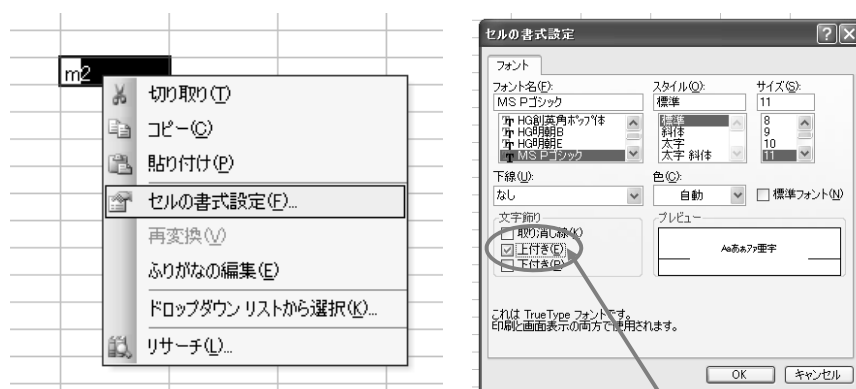
データ	
平均	13
標準誤差	1.47196
中央値（メジアン）	13
最頻値（モード）	#N/A
標準偏差	7.359801
分散	54.16667
尖度	-1.2
歪度	0
範囲	24
最小	1
最大	25
合計	325
標本数	25

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 cm^2 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。

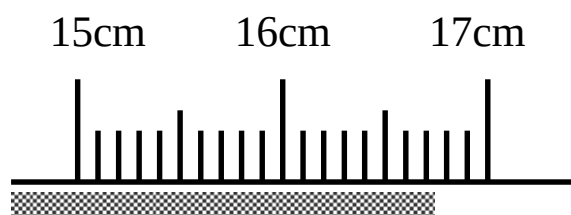


3. データ計算における注意：計算結果を何桁まで表示すべきか？ 有効数字の概念

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。

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てとよい. 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.

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.

例 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

平均, 標準偏差のときは小数第 1 位まで

$$\text{平均} = 6.6$$

$$\text{標準偏差} = 3.04959 \dots \approx 3.0$$

分散は小数第 1 位まででよい

$$\text{分散} = 9.3$$

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 大ききの順に並べる, 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？

1. 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？

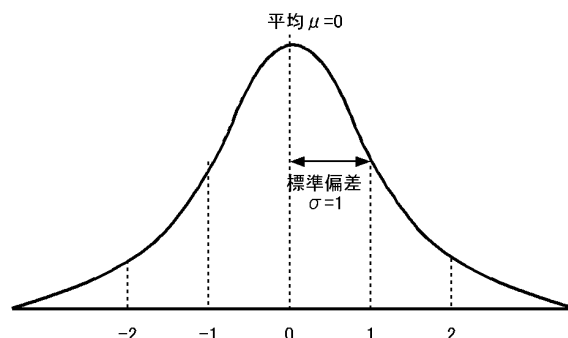
① 中心化の傾向を表す代表値として

平均

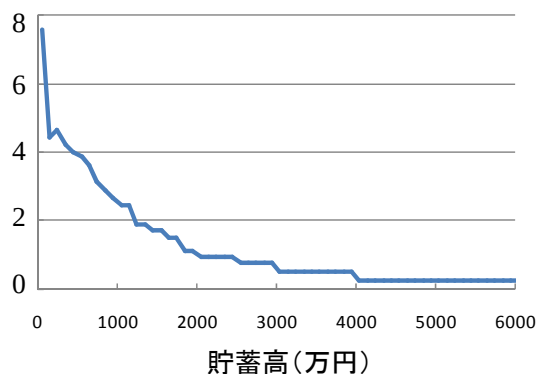
平均は分布に（ ）のないとき
（ ）にもっとも有用な統計量である。
とくに母集団が（ ）を示すときに非
常に有用である。

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

右の場合、目的によってはメジアンを用いる方がよい



(%) 日本の世帯別貯蓄高の分布(2007年)



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。

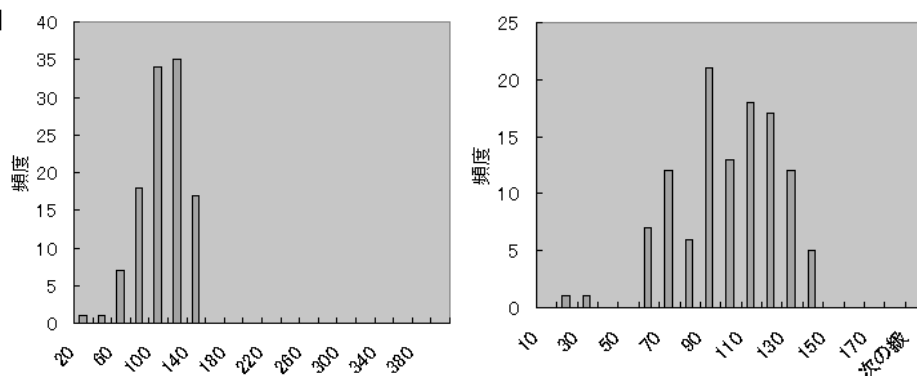
データに異常値があるときのように、極端に離れた値を含む標本では平均よりメジアンを用いる方が無難である。

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。

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アジア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。



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。

2. 要約する過程でどのくらいの情報が失われたのか？

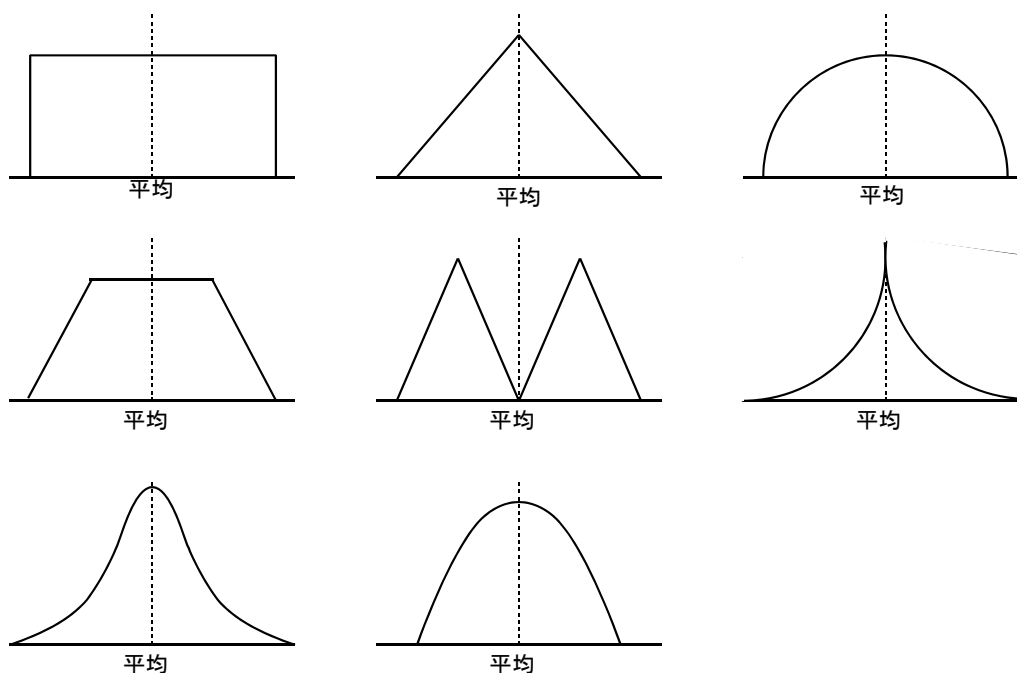
平均・分散・標準偏差の計算にはすべてのデータが関係する。

メジアンでは真ん中にある値1つか2つしか、計算に関係しない。

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 1. 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、分散を計算し、小数第 2 位までで示せ。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？
- ★ ジョーカーを含まない 52 枚のトランプから 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、分散を計算し、小数第 2 位までで示せ。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→4. 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。

提出するときには元となったデータもきちんと印刷すること。

提出締め切りは 10 月 26 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。
（以降のレポートも原則として、月曜日午後 1 時までに提出すること）

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看、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