情報科学

統計解析(1)

島根大学医学部医療情報学講座
安田 晃

今日の内容

- 統計とは何だ？
- 統計は勘や当てずっぽうか？
- 統計を使って何がわかる？
- 統計で何が科学的にわかるのか？
- 統計のはじめの一歩，平均値．
- 統計のはじめの二歩，分散，標準偏差．
- 平均値と標準偏差を使って変な差をつけたがる？偏差値を知る．

統計とは何だ？

統計・・・集団における個々の要素の分布を調べ、その集団の傾向・性質などを数量的に統一的に明らかにすること。また、その結果として得られた数値。

解析・・・物事をこまかく解き開き、理論に基づいて研究すること。

理論・・・個々の事実や認識を統一的に説明することのできる普遍性をもつ体系的知識。

統計とは何だ？

私たちの周囲に統計は溢れている．

例えば，（2003年度飲食業ランキング 日本経済新聞より）

順位	社名（主な店名）	2003年度売上高（百万円）	前年度比伸び率(%)	売上高内訳		店舗数合計
				直営	FC	
1(1)	日本マクドナルド	386,688	▲4.0	270,785	115,903	3,773
2(2)	すかいらーく	280,873	▲1.5	280,873	0	2,450
3(3)	ほっかほっか亭総本部	197,362	2.6	0	197,362	3,519
4(6)	日清医療食品	136,600	8.8	136,600	0	3,357
5(5)	モンテローザ（居楽屋白木屋、のみくい処魚民、居楽屋笑笑）	132,435	▲3.6	132,435	0	1,106
6(9)	ダスキン（ミスタードーナツ）	130,000	8.5	8,384	121,616	1,319
7(4)	日本ケンタッキー・フライド・チキン（ケンタッキーフライドチキン、ピザハット）	127,358	▲7.6	44,296	83,062	1,494
8(7)	ロイヤル	121,974	▲2.4	74,681	47,293	501
9(8)	本家かまどや	120,850	▲0.6	20,017	100,833	2,570
10(10)	モスフードサービス（モスバーガー）	107,500	0.2	13,500	94,000	1,476

統計とは何だ？

「東北農政局秋田統計・情報センターは、03年の県漁業・養殖業統計を発表した。生産量は1万1470トンで前年比0・9％（101トン）増加した。ハタハタの資源回復を反映したもので、3年連続の増加となった」

「朝、すっきりと目覚められた人は、高校生では7人に1人にとどまり、親の世代の4人に1人よりかなり低いことが18日までに、静岡県が実施した子どもの生活実態調査で分かった。親の方が快眠できているとも受け取れる皮肉な結果で、県は詳しく分析し、問題点を洗い出したいとしている」

このような内容も「統計」として扱われる．

統計は当てずっぽうか？

例えば

- 基準と比較して本当に異なっているのか？
  - –ここ1週間の平均最高気温22℃．今日は24℃だった．今日は暑い日だった．
- 同じ2つのものを比較して本当に異なっているのだろうか？
  - –LソンとPプラの弁当では，押しなべて見ればLソンが安いだろう（かもしれない）．
- 集団の傾向を的確に表わしているか？
  - –身長が高ければ体重も大きい．大相撲の世界でも？
- 複数の基準を比較して，異なっている？
  - –徒歩と自転車と自動車とスーパーやくもで速いのは？

統計は当てずっぽうか？

これらのことを，単なる当てずっぽうや，そんな気がする，で終わらせてほしくありません．

看護学で得られたデータは，客観的な方法で集計したり，計算してほしいからです．

それらを行うことにより，データの背後にある何かが見えるかもしれません．そのことは今まで誰も発見していないことかもしれません．

それらは当てずっぽうではできません．統計学は看護学同様，科学的に議論しましょう．

統計を使って何がわかる？

統計を使えば，

- Play Typeがどのような分布をしているのか．
- 通学時間と成績は関係あり？
- バイトで収入があったら，その日からしばらくは食生活が豊かだろうか．
- Mドナルドのひとりあたりの購買量のムラは，Mスタードーナツのそれと比べ大きいか．

などを具体的な数字で示してくれる．

統計を使って何がわかる？

今までのことを一般化すれば，

論理的な考えから，得られたデータの主成分をとったり，2つの変量の関係を見たり，変量の代表値を比較したりすることによって，現象の原因と結果を科学的に見るひとつの手段として統計を学ぼう．

統計で何が科学的にわかるのか？

例えば，102人に聞きました．Mドナルドを何と呼んでいますか？

呼び名	マクド	マック	Mドナルド そのまんま	合計
人数	40	28	34	102

マクドと呼ぶ人が多そうだ．
フルネームのMドナルドも1/3ある．
マックは少数派か？
‘‘‘．
この呼び名の区分に対して差異はあるのだろうか？
こんなとき，

統計で何が科学的にわかるのか？

このように考える．

呼び名	マクド	マック	Mドナルド	合計
人数	40	28	34	102

から，102人が3つのグループに均等に分かれたら，

呼び名	マクド	マック	Mドナルド	合計
人数	34	34	34	102

となる．これを理論値と呼ぼう．実際の値とこの理論値の離れ具合をK. Pearsonは次のように計算して，その値を優位水準というものに置き換えた．その計算は，

統計で何が科学的にわかるのか？

$$\frac{(\text{マクド}-\text{マクド理論値})^2}{\text{マクド理論値}}+\frac{(\text{マック}-\text{マック理論値})^2}{\text{マック理論値}}+\frac{(\text{Mドナルド}-\text{M理論値})^2}{\text{Mドナルド理論値}}$$
$$=\frac{(40-34)^2}{34}+\frac{(28-34)^2}{34}+\frac{(34-34)^2}{34}$$
$$=0.471+0.471+0=0.942$$

エクセルにもあるχ2乗分布表から，0.942を上回る優位水準（確率）は0.624．これが0.05より小さかったらデータは均一ではないと言える．この場合は，マクドが多そうに見えるが，102名のデータからでは特に多いとは言えないようだ．

統計のはじめの一歩，平均値

平均･･･広辞苑第4版から見てみれば，

不揃いのないようにすること。ならすこと。また、不揃いのないこと。

つり合いがとれていること。平衡。

多くの量または数の中間的な値。また、それを求める演算。

統計ではこのように考えてみよう･･･．

統計のはじめの一歩，平均値

今，10，7，16の平均は，

$$\frac{1}{3}\times(10+7+16)=11$$

なんでこのような計算をするのだろう．

ここで，10，7，16を代表とする値は，10，7，16それぞれの値からの誤差の2乗和がもっとも小さくなるよう考える．
代表する値を a とおいて，2乗和を S とすれば，

$$S=(10-a)^2+(7-a)^2+(16-a)^2$$

が最小となる a を求める問題となった．

統計のはじめの一歩，平均値

S が最小となる a は，どのような数か？

$a=8$ では，

$$S=(10-8)^2+(7-8)^2+(16-8)^2=4+1+64=69$$

$a=12.76$ では，

$$S=(10-12.76)^2+(7-12.76)^2+(16-12.76)^2$$
$$=7.6176+33.1776+10.4876=51.2328$$

⋮

何か良い方法はないか．･･･，ある．

統計のはじめの一歩，平均値

ここで，高校生のとき勉強したことを基礎にして，次のように考える．

S が a で最小となるためには，

$\frac{dS}{da}=0$ を満たす a があるはずである．よって，

$$\frac{dS}{da}=\frac{d}{da}(10-a)^2+\frac{d}{da}(7-a)^2+\frac{d}{da}(16-a)^2=0$$

となって，

$$2(10-a)+2(7-a)+2(16-a)=0$$

を得る．更に，

統計のはじめの一歩，平均値

$$(10-a)+(7-a)+(16-a)=0$$
$$10+7+16=3a$$
$$a=\frac{1}{3}(10+7+16)=11$$

誤差の2乗和を最小とする値は，私たちが考えている平均値であった．

実際に，

a	10	7	16	2乗和
7	9	0	81	90
8	4	1	64	69
9	1	4	49	54
10	0	9	36	45
11	1	16	25	42
12	4	25	16	45
13	9	36	9	54
14	16	49	4	69

となる．

統計のはじめの一歩，平均値

今までのことを一般化して，

n 個のデータ $x_1, x_2, \cdots, x_n$ があるとき，

平均値 $\bar{x}$ は，

$$\bar{x}=\frac{1}{n}(x_1+x_2+\cdots+x_n)=\frac{1}{n}\sum_{i=1}^{n}x_i$$

であった．この平均を算術平均と呼ぶことがある．

統計のはじめの一歩, 平均値

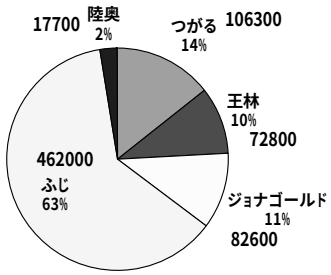
このような考え方もある.
りんごを3種類を以下の値段で買った. りんごの平均値は？

りんごの種類	買った個数	1個の値段	合計金額
ふじ	5	128	640
ジョナゴールド	3	158	474
陸奥	10	198	1980

$$\bar{x}_{りんご} = \frac{5\times128+3\times158+10\times198}{5+3+10}$$
$$= \frac{640+474+1980}{18} = \frac{3094}{18} = 171.89$$

統計のはじめの一歩, 平均値

2003年度の全国りんご生産高　（単位トン）



農林水産省2003年産りんごの収穫量及び出荷量より

統計のはじめの一歩, 平均値

一般化して,

データ　x	重み（係数）　w
x_1	w_1
x_2	w_2
⋮	⋮
x_n	w_n

としたとき, 平均値 $\bar{x}$ は,

$$\bar{x} = \frac{w_1x_1+w_2x_2+\cdots+w_nx_n}{w_1+w_2+\cdots+w_n} = \frac{\sum_{i=1}^{n} w_ix_i}{\sum_{i=1}^{n} w_i}$$

である.　このような平均を特に加重平均という.

統計のはじめの一歩, 平均値

この他に

幾何平均

$$\bar{x}_G = \sqrt[n]{x_1x_2\cdots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

調和平均

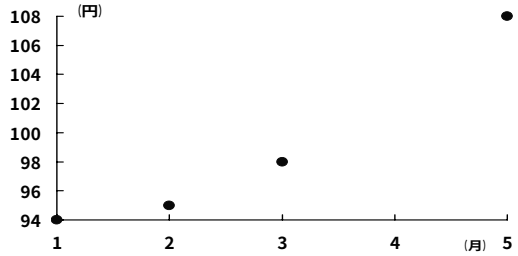
$$\bar{x}_H = \frac{n}{\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

がある.

統計のはじめの一歩, 平均値

幾何平均の例:

ガソリンの小売値は2004年に入って以下のようであった.
4月に観測を忘れたが, どの程度だろうか.

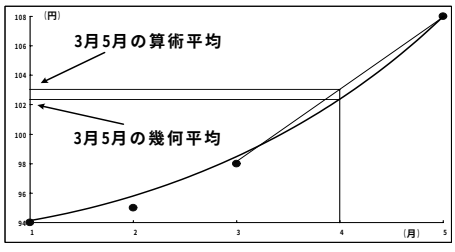


統計のはじめの一歩, 平均値

4月の小売値を x_4 とおけば,

$$x_4 = \frac{1}{2}(98+108) = 103$$

しかし,



となっていると考えたほうがよさそうだ.

統計のはじめの一歩，平均値

グラフのように変化していると思われるので，(算術) 平均

では大きく見積もられるだろう．そこで，

$$x_4 = \sqrt{98 \times 108} = 102.879$$

としたほうが，日常的な値となるだろう．

統計のはじめの一歩，平均値

調和平均は，例えば，出雲-新大阪間401kmを，行きは平

均スピード114.57km/h，帰りは同じく98.204km/hで往復した．

平均のスピードは？

出雲-新大阪間往復に要した時間は，

行き　$\frac{401}{114.57}$時間　　帰り　$\frac{401}{98.204}$時間

統計のはじめの一歩，平均値

往復の距離は802kmだから

$$\frac{802}{\frac{401}{114.57}+\frac{401}{98.204}} = \frac{802}{401\left(\frac{1}{114.57}+\frac{1}{98.204}\right)}$$

$$= \frac{2}{\frac{1}{114.57}+\frac{1}{98.204}} = 105.76$$

決して，

$$\frac{1}{2}(114.57+98.204) = 106.39$$

ではない．

統計のはじめの二歩，分散，標準偏差

例えば，Mドーナツにあるケーキ･ドーナツとイースト･ドーナツの値段の散らばりが気になった．

ケーキ･ドーナツ	値段
ホームカット	94
シナモン	94
ココナツ	105
バタークランチ	105
プレーンクルーラー	94
シュガークルーラー	94
シナモンクルーラー	94

平均＝97.14円

イースト･ドーナツ	値段
ハニーディップ	94
シュガーレイズド	94
チョコリング	105
ホワイトチョコリング	105
ストロベリーリング	105
エンゼルクリーム	115
スタードクリーム	115
チョコカスタード	115
カリーパン	147
コーヒーロール	126
ツイスト	126

平均＝113.4円

統計のはじめの二歩，分散，標準偏差

一見して，イースト･ドーナツのほうが散らばっているように見える．散らばりを，代表する値を平均値とするとき，平均値の周りにどの程度散らばっているかを量的に見たい．その時，このように考える．

①各ドーナツと平均値までの差を1辺とする正方形を考える．

②その正方形の面積をすべて足し算する．

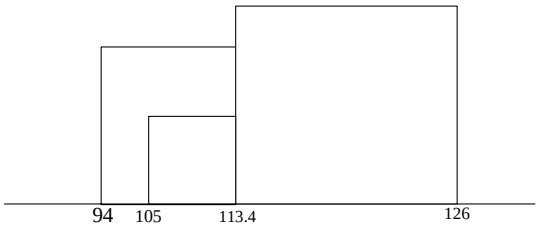
③正方形の平均を考える．

④正方形の面積はこの場合，円^2．これは馴染み難い．

⑤そこで平方根をとると円となる．

統計のはじめの二歩，分散，標準偏差

今までを図にしてみれば，



統計のはじめの二歩, 分散, 標準偏差

各値段から平均を引き算し, 総和しただけでは,

$$(94-97.14)+\cdots+(94-97.14)=0$$

となって, うまくゆかない. イースト･ドーナツの場合も同様. そこで, 2乗和を考える. つまりイースト･ドーナツの平均113.4円をそれぞれから引き算して2乗し, ハニーディップからツイストまで足し算してみる.

$$(94-113.4)^2+\cdots+(126-113.4)^2=2418.56$$

この式の各項は, 各値段から平均値までを1辺とする正方形の面積である. その面積をすべて足し算したものである.

次に, この2乗和の平均をとってみる.

統計のはじめの二歩, 分散, 標準偏差

即ち,

$$\frac{1}{11}\times\left\{(94-113.4)^2+\cdots+(126-113.4)^2\right\}=219.86$$

である. これで平均的な正方形の面積が分かった. ここで

はこの面積の単位は 円^2 である. そこで, 平方根をとって,

$$\sqrt{\frac{1}{11}\times\left\{(94-113.4)^2+\cdots+(126-113.4)^2\right\}}=14.828$$

を得る. これがデータのばらつきを示している.

統計のはじめの二歩, 分散, 標準偏差

ケーキ･ドーナツの場合も同様に

$$(94-97.14)^2+\cdots+(94-97.14)^2=172.86$$

$$\frac{1}{7}\times\left\{(94-97.14)^2+\cdots+(94-97.14)^2\right\}=24.694$$

$$\sqrt{\frac{1}{7}\times\left\{(94-97.14)^2+\cdots+(94-97.14)^2\right\}}=4.969$$

となる.

統計のはじめの二歩, 分散, 標準偏差

一般化して, n 個のデータ $x_1, x_2, \cdots, x_n$ があるとき,

それぞれのデータから平均値までの距離の2乗したもの総

和を偏差平方和 (S) という. 即ち,

$$S=(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2=\sum_{i=1}^{n}(x_i-\bar{x})^2$$

統計のはじめの二歩, 分散, 標準偏差

次に, 平均的な乖離を見るには, S を n で割ればいい.

この結果を分散 (V) という.

$$V=\frac{1}{n}\left\{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2\right\}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

統計のはじめの二歩, 分散, 標準偏差

次いで, V の単位では一般的ではないし, ばらつき程度を

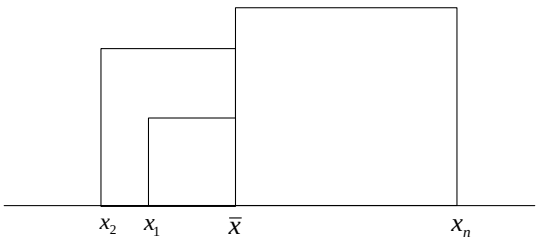
知るには適当ではない. そこで, V の平方根をとって標準

偏差 (s) と呼ぼう. 即ち,

$$s=\sqrt{\frac{1}{n}\left\{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2\right\}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

統計のはじめの二歩，分散，標準偏差

今までを図にしてみれば，



平均と標準偏差を使って

平均値と標準偏差を使って，

$$\frac{\text{それぞれの値}-\text{平均値}}{\text{標準偏差}}$$

のような計算を行ってみよう．今まで使った記号で示せば，

n 個のデータ $x_1, x_2, \cdots, x_n$ において，平均を $\bar{x}$ ，

標準偏差を s とすれば，i 番目のデータは，

$$\frac{x_i - \bar{x}}{s}$$

と書ける．例えば，

平均と標準偏差を使って

Mドナルドのイートス･ドーナツデータでは，

イースト･ドーナツ	値段	$\frac{x_i - \bar{x}}{s}$
ハニーディップ	94	-1.306
シュガーレイズド	94	-1.306
チョコリング	105	-0.564
ホワイトチョコリング	105	-0.564
ストロベリーリング	105	-0.564
エンゼルクリーム	115	0.110
スタードクリーム	115	0.110
チョコカスタード	115	0.110
カリーパン	147	2.268
コーヒーロール	126	0.852
ツイスト	126	0.852

具体的には

$$\frac{147-113.4}{14.83}$$

この11個の平均は0，
標準偏差は1となった．

平均=113.4円，標準偏差=14.83円

平均と標準偏差を使って

ケーキドーナツではどうだろう．

ケーキ･ドーナツ	値段	$\frac{x_i - \bar{x}}{s}$
ホームカット	94	-0.632
シナモン	94	-0.632
ココナツ	105	1.581
バタークランチ	105	1.581
プレーンクルーラー	94	-0.632
シュガークルーラー	94	-0.632
シナモンクルーラー	94	-0.632

この7個の平均は0，
標準偏差は1となった．

平均=97.143，標準偏差=4.969

平均と標準偏差を使って

同じ値段のチョコリングとバタークランチ，前者は符号がマイナスに，後者はプラスとなっている．このような計算をするには何か裏がありそうだ．

･･･裏はないが，このような計算は皆さんが高校時代によく計算している．次にそれを示そう．

平均と標準偏差を使って

先ほどの計算，

$$\frac{\text{それぞれの値}-\text{平均値}}{\text{標準偏差}}$$

あるいは，

$$\frac{x_i - \bar{x}}{s}$$

を統計の世界ではデータの標準化という．このように計算すれば，どのようなデータの集団だって平均は0，標準偏差（あるいは分散）は1となる．

平均と標準偏差を使って

例えば，12個のデータ

1000	2	0.3	40	5	600	700	0.08	900	10	11	0.12

$\bar{x} = 272.38,\ s = 384.23$

標準化すれば，

1.894	-0.704	-0.708	-0.605	-0.696	0.853	1.113	-0.709	1.633	-0.983	-0.680	-0.719

$$\frac{1}{12}(1.894 + \cdots - 0.719) = 0$$

$$\sqrt{\frac{1}{12}\left\{(1.894-0)^2 + (-0.719-0)^2\right\}} = \sqrt{\frac{1}{12}\times 12} = \sqrt{1} = 1$$

何故，平均は0，標準偏差は1になるのか，一般化したデータで証明してほしい．

平均と標準偏差を使って

フィクション・・・

S大学では最終合否判定となった5名の取り扱いに困っていた．

それは，

	A	B	C	D	E	平均	標準偏差
数学	50	80	20	40	60	50	20
化学	35	50	45	55	65	50	10
合計点	85	130	65	95	125	100	24.5
合計点の順位	4	1	5	3	2		

から1名を合格させることであった．

平均と標準偏差を使って

ここで，当然最高点をとったBを選ぶところだが，偏差値という問題が浮上してきた．
偏差値は，

$$50 + 10\times\frac{\text{それぞれの値-平均値}}{\text{標準偏差}}$$

$$50 + 10\frac{x_i - \bar{x}}{s}$$

で示される．先ほどの表から偏差値を計算してみよう．

平均と標準偏差を使って

先ほどの表は，

	A	B	C	D	E
数学	$50+10\times\frac{50-50}{20}=50$	65	35	45	55
化学	$50+10\times\frac{35-50}{10}=35$	50	45	55	65
偏差値の平均	$\frac{1}{2}(50+35)=42.5$	**57.5**	40	50	60
合計点の偏差値	$50+10\times\frac{85-100}{24.5}=44$	62	36	48	60

この表から矛盾が起きている．合計点ではBが合格しているが，偏差値の平均から見てみれば，Eのほうが成績が良くなっている．S大学は困り果て・・・という措置をとった．

平均と標準偏差を使って

偏差値は，平均が50，標準偏差が10である．

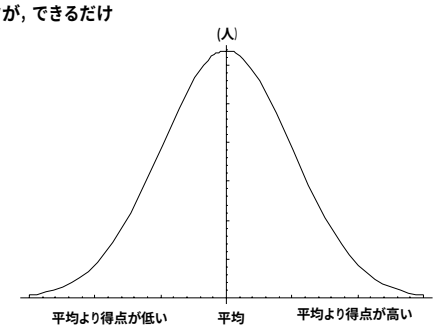
						平均	標準偏差
A県	100	0	50	10	90	**50**	**40.497**
B県	55	50	60	55	60	**56**	**3.741**
C府	100	90	95	95	95	**95**	**3.163**

A県	$50+10\times\frac{100-50}{40.497}$ 62	38	50	40	60	**50**	**10**
B県	47	34	61	47	61	**50**	**10**
C府	66	34	50	50	50	**50**	**10**

このデータから何が言えるか考えてほしい．

平均と標準偏差を使って

注意：偏差値あるいはデータの標準化を行うときには，データが，できるだけ



のようになっていることが必要である．

注 意

1. 今日話したことはあくまで統計の基礎の基礎に過ぎません．
2. 今後統計が必要になったときには，文献や成書を参考に勉強願います．
3. データがあれば，データそのままを人に見てもらうことも可能でしょう．しかし，それでは冗長すぎます．そのデータの必要なところをうまくまとめてください．
4. 今日話したことは，平均や分散，標準偏差が計算できるデータが対象です．アンケートの回答などには適用しないほうがいいでしょう．

注 意

4.の具体化．例えば，「当店の食事はいかがでしたか？」に

番号	選択肢	人数
1	とても満足	20
2	満足	32
3	普通	14
4	不満	21
5	とても不満	5

と集計し，

加重平均をとって，

$$\frac{1\times 20+2\times 32+3\times 14+4\times 21+5\times 5}{20+32+14+21+5}=2.554\cdots$$

ということは，92名の対象者はこの店の食事に満足しているが，かなり普通と感じている？ ———— 平均を計算することはナンセンスである．

2.554に相当する選択肢はどんな感じだろう．

再来週は，

- 2つの変量の関係
  - この関係を簡単な式で表わしたい．
  - この関係を具体的な数値で示したい．
- ひとつの変量とp個の変量の関係
  - この関係を簡単な式で表わしたい．
  - この関係を具体的な数値で示したい．
- アンケートを行う，その前に．