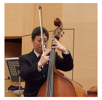
2016年度秋学期　統計学　　第５回

分布をまとめる—平均・分散

浅野　晃
関西大学総合情報学部



代表値

代表値とは

統計学が相手にするのは,
「分布」しているデータ

（大般若会の写真）

データをこんな
ふうに読めれば
いいけれど…

http://www3.ic-net.or.jp/~yaguchi/houwa/daihannya.htm

代表値とは

こんなことはできないので,

- 図示する（ヒストグラム）
- ひとつの数にまとめる

（大般若会の写真）

[代表値]
数字で表されていれば,
計算ができる

http://www3.ic-net.or.jp/~yaguchi/houwa/daihannya.htm

平均

とくに［算術平均］は
代表的な代表値

（算術）平均
＝（データの総和）÷（数値の個数）

平均

データ $x_1, x_2, \ldots, x_n$ のとき
数値の個数（データサイズ） n

平均 $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{1}{n} \sum_{i=1}^{n} x_i$

和

データサイズ？

「データ」という言葉は，
数値の集まりをさす
（１つ１つの数値ではない）

データの中に含まれる
数値の個数を大きさ（サイズ）という

家族(family)という言葉に似ている

度数分布から平均を求める

度数分布とは，これでした

以上	未満	階級値	度数	相対度数
15	25	20	4	0.08 (8%)
25	35	30	3	0.06 (6%)
35	45	40	3	0.06 (6%)
45	55	50	8	0.16 (16%)
55	65	60	12	0.24 (24%)
65	75	70	8	0.16 (16%)
75	85	80	9	0.18 (18%)
85	95	90	3	0.06 (6%)
x	x	x	計 50	計 1 (100%)

度数分布から平均を求める

平均=(データの合計)/(データサイズ)
=([階級値×度数]の合計)/(データサイズ)
=[階級値×(度数/データサイズ)]の合計
=[階級値×相対度数]の合計

以上	未満	階級値	度数	相対度数
15	25	20	4	0.08 (8%)
25	35	30	3	0.06 (6%)
35	45	40	3	0.06 (6%)
45	55	50	8	0.16 (16%)
55	65	60	12	0.24 (24%)
65	75	70	8	0.16 (16%)
75	85	80	9	0.18 (18%)
85	95	90	3	0.06 (6%)
x	x	x	計 50	計 1 (100%)

分散と標準偏差

「ばらつき」を数字で

分布は，大小ばらばらな数値からなるデータ

どのくらいばらばらかを，
数字で表そう

A: 0, 3, 3, 5, 5, 5, 5, 7, 7, 10　どう違う？
B: 0, 1, 2, 3, 5, 5, 7, 8, 9, 10　平均は
C: 3, 4, 4, 5, 5, 5, 5, 6, 6, 7　どれも5

レンジとばらつき

A: 0, 3, 3, 5, 5, 5, 5, 7, 7, 10
B: 0, 1, 2, 3, 5, 5, 7, 8, 9, 10
C: 3, 4, 4, 5, 5, 5, 5, 6, 6, 7

Cは，最大と最小の差［レンジ］が違う

A, Bはレンジは同じだが，
Bのほうがばらついている
ように見える

偏差

各数値と平均との差を［偏差］という

-5 -2 -2 0 0 0 0 +2 +2 +5

$A: 0, 3, 3, 5, 5, 5, 5, 7, 7, 10$

$B: 0, 1, 2, 3, 5, 5, 7, 8, 9, 10$

-5 -4 -3 -2 0 0 +2 +3 +4 +5

偏差を平均したら，AとBのばらつきの違いが表せる？

偏差の平均？

だめ。平均したらゼロ

-5 -2 -2 0 0 0 0 +2 +2 +5

$A: 0, 3, 3, 5, 5, 5, 5, 7, 7, 10$

$B: 0, 1, 2, 3, 5, 5, 7, 8, 9, 10$

-5 -4 -3 -2 0 0 +2 +3 +4 +5

偏差を２乗する

偏差を２乗したら，全部正の数になるから，それから平均する

25 4 4 0 0 0 0 4 4 25

-5 -2 -2 0 0 0 0 +2 +2 +5

$A: 0, 3, 3, 5, 5, 5, 5, 7, 7, 10$

$B: 0, 1, 2, 3, 5, 5, 7, 8, 9, 10$

-5 -4 -3 -2 0 0 +2 +3 +4 +5

25 16 9 4 0 0 4 9 16 25

分散

平均 *6.6*

25 4 4 0 0 0 0 4 4 25

-5 -2 -2 0 0 0 0 +2 +2 +5

$A: 0, 3, 3, 5, 5, 5, 5, 7, 7, 10$

$B: 0, 1, 2, 3, 5, 5, 7, 8, 9, 10$

-5 -4 -3 -2 0 0 +2 +3 +4 +5

25 16 9 4 0 0 4 9 16 25

平均 *10.8*

［分散］＝(偏差)^{2}の平均

分散と標準偏差

［分散］＝(偏差)^{2}の平均　式で書くと

1番の数値　データの平均

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

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

n個たして

nで割る

分散の平方根を［標準偏差］という

A. Asano, Kansai Univ.

度数分布から分散を求める

データの平均＝[階級値×相対度数]の合計

分散＝(偏差)^{2}の平均

＝[(偏差)^2×相対度数]の合計

＝[(階級値−データの平均)^2×相対度数]の合計

以上	未満	階級値	度数	相対度数
15	25	20	4	0.08 (8%)
25	35	30	3	0.06 (6%)
35	45	40	3	0.06 (6%)
45	55	50	8	0.16 (16%)
55	65	60	12	0.24 (24%)
65	75	70	8	0.16 (16%)
75	85	80	9	0.18 (18%)
85	95	90	3	0.06 (6%)
x	x	x	計 50	計 1 (100%)

A. Asano, Kansai Univ.

なぜ２乗？

偏差の２乗ではなく，

偏差の「絶対値」ではいけないの？

絶対値の関数は，途中に折れ目があってむずかしい

放物線には折り目はない

A. Asano, Kansai Univ.

標準得点

A. Asano, Kansai Univ.

「試験で70点」は優れているのか

試験で70点をとった。
まわりより優れているのか？

一緒に受けた人たちの平均点が

50点なら　優れている

80点なら　劣っている

「試験で70点」は優れているのか

試験で70点をとった。
まわりよりとても優れているのか？

一緒に受けた人たちの平均点が

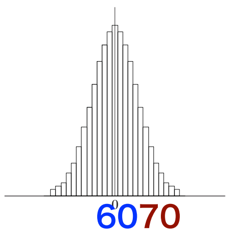
~~50点なら　まあ優れている~~

~~30点なら　とても優れている~~　…？

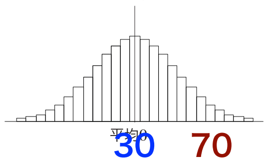
「試験で70点」は優れているのか

一緒に受けた人たちが

平均60点で
標準偏差5点



平均30点で
標準偏差20点

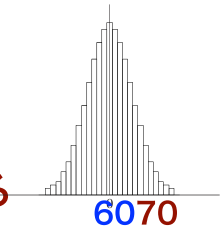


70点の
「地位」
は同じ。

「地位」を数字で表す

一緒に受けた人たちが

平均60点で標準偏差5点なら
70点の人は，平均を
標準偏差の2倍上回っている

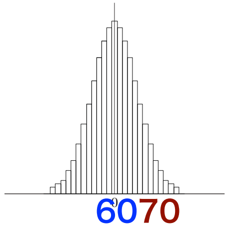


平均30点で標準偏差20点なら
70点の人は，やはり平均を
標準偏差の2倍上回っている

70点の「地位」は同じ

標準得点

平均を
標準偏差の2倍上回っている



［標準得点］が2点

平均を標準偏差の2倍
下回っているなら

標準得点が-2点

標準得点への換算

標準得点＝
　分布中のある数値が，
　平均を標準偏差の何倍
　上回って／下回っているか

分布そのものを
平均0，標準偏差1に「変換」したら？

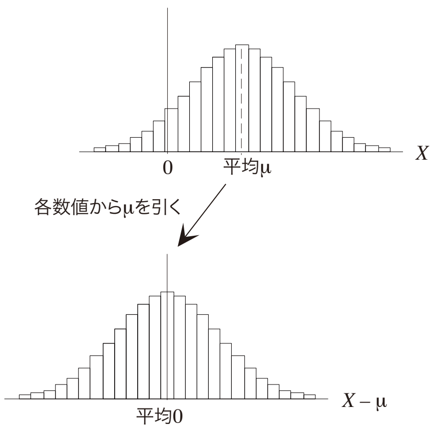
その数値の変換後の値が，
そのまま標準得点になる

分布の変換

分布中の各数値から，平均を引く

平均μ
標準偏差σ

平均0
標準偏差σ



分布の変換（続き）

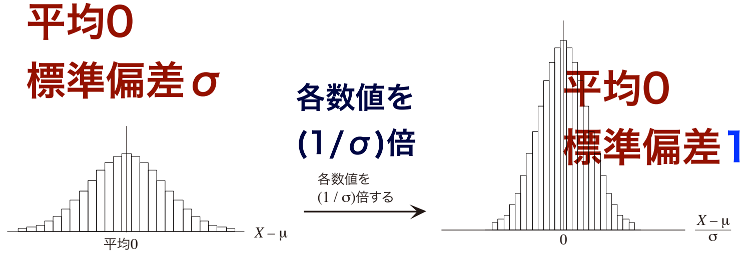
分布中の各数値から，平均を引いて
標準偏差で割る

各数値の偏差は	$(1/\sigma)$倍
分散は(偏差)^{2}の平均	$(1/\sigma)^2$倍
標準偏差は分散の平方根	$(1/\sigma)$倍

平均0
標準偏差σ

各数値を
$(1/\sigma)$倍

平均0
標準偏差1



式で書くと

分布そのものをXとすると

$Z = (X - \mu) / \sigma$

と変換すると，Zは平均0，標準偏差1

受験産業でいう「偏差値」

平均0，標準偏差1の分布Zを，さらに

$W = 10Z + 50$

と変換すると，Wは平均50，標準偏差10

これが［偏差値］

偏差値70	平均よりも，標準偏差の2倍上回っている
偏差値40	平均よりも，標準偏差の1倍下回っている