2024年度春学期

統計学

第5回
分布をまとめる
— 記述統計量(平均・分散など)

関西大学総合情報学部
浅野 晃

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」で経典を翻すだけで「読む」ように,
データも一目見るだけで内容がわかればいいけれど…

※大般若会(だいはんにゃえ)とは, 600巻に及ぶ「大般若経」を,
僧侶が翻すことで「読む」という儀式です。
(講義ウェブサイトにあるリンク先を参照してください)

代表値とは

現実には, 人間は, 数字をざっと眺める
だけで一瞬で理解できるほど, 賢くありません

そこで
- 図示する(ヒストグラム)
- ひとつの数にまとめる ⟶ [代表値]
  数字で表されていれば, 計算ができる

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）

↑

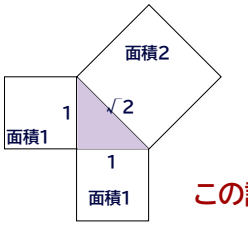
”／”でも同じ意味（÷よりもよく用います）

算術？

算術平均以外にも，「幾何平均」「調和平均」というものもあります

算術平均 ＝（データの総和）÷（数値の個数）

幾何平均 ＝（データ全部の積）の，（数値の個数）乗根



√（ルート，2乗根）が出てくるのは幾何の問題です

この講義では，以後「平均」といえば算術平均のことです。

算術平均を式で書くと

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

※「エックスバー」と読んでください。「バー」は平均を表すのによく用います。

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

※Σは「合計」を表す記号です。

和

※もし日本人がΣ記号を発明していたら，きっと「和」と書いていたことでしょう。

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つの数値をさすのではない）

データの中に含まれる数値の個数をデータの大きさ（サイズ）という

※データの数とはいいません。

家族（family）という言葉に似ている

※人数の多い家族は「大家族」といい，「多家族」とはいいません。

度数分布から平均を求める

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%)

度数分布とは，こんなやつでした

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)

ひとつの階級に入っている数値は，
みな「階級値と同じ」とみなすから，

ひとつの階級には，「階級値」と同じ数値が，
度数(個)あるとみなされる

よって，ひとつの階級の数値の合計は，
「階級値×度数」で表される

=((階級値×度数)の合計)/(データサイズ)

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)

かけ算(×)と割り算(/)について，
カッコ"( )"の位置をかえる

= [階級値 × (度数 / データサイズ)]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%)

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 / データサイズ)]の合計

(度数 / データサイズ)のことを
「相対度数」という

=[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%)

度数分布から平均を求める

階級	階級値	相対度数
0 〜 9（点）	5	0.04
10 〜 19	15	0.16
20 〜 29	25	0.08
30 〜 39	35	0.12
40 〜 49	45	0.10
50 〜 59	55	0.10
60 〜 69	65	0.12
70 〜 79	75	0.08
80 〜 89	85	0.18
90 〜 100	95	0.02
合計		1.0

テキストに載っている別の例で，
計算してみましょう

度数分布から平均を求める

階級	階級値	相対度数	階級値×相対度数
0 〜 9（点）	5	0.04	5×0.04 = 0.2
10 〜 19	15	0.16	15×0.16 = 2.4
20 〜 29	25	0.08	25×0.08 = 2.0
30 〜 39	35	0.12	35×0.12 = 4.2
40 〜 49	45	0.10	45×0.10 = 4.5
50 〜 59	55	0.10	55×0.10 = 5.5
60 〜 69	65	0.12	65×0.12 = 7.8
70 〜 79	75	0.08	75×0.08 = 6.0
80 〜 89	85	0.18	85×0.18 = 15.3
90 〜 100	95	0.02	95×0.02 = 1.9
合計		1.0	合計 49.8

各階級で，
[階級値×相対度数]を求めて
合計する

これが平均

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

では，どのくらいばらばらかを，数字で表そう

A, B, Cは，いずれも10個の数値からなるデータです。

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

平均はどれも5

では，どう違う？

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，

Aの青線部とBの赤線部を比べると

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

※AもBも平均は5ですから，各数値と5との差を書いていきます。

-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でしょうか？

偏差の平均？

だめです🤦 偏差を平均したらゼロになるからです。

-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

そこで，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，2乗してから平均する

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

［分散］＝(偏差)²の平均　　この「分散」を，ばらつきの指標とする

25 4 4 0 0 0 0 4 4 25　　平均 6.6 ＝ Aの分散

-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 ＝ Bの分散

Bのほうが分散が大きい。Bのほうがよりばらついている。

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

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

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

n 個たして
n で割る

分散の平方根(√)を［標準偏差］という

分散を求める計算の途中で数値を2乗しているので，平方根を求めてもとにもどす

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

一方，分散＝(偏差)²の平均　　※ここを置き換える

だから，分散 ＝ ［(偏差)²×相対度数］の合計

分散 ＝ ［(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%)

度数分布から分散を求める

テキストに載っている例で，計算してみましょう

階級	階級値	相対度数	階級値×相対度数	偏差	(偏差)²	(偏差)² × 相対度数
0 ～ 9 (点)	5	0.04	0.2	5 - 49.8 = -44.8	(-44.8)²=2007.4	2007.4×0.04 =80.28
10 ～ 19	15	0.16	2.4			
20 ～ 29	25	0.08	2.0			
30 ～ 39	35	0.12	4.2			
40 ～ 49	45	0.10	4.5			
50 ～ 59	55	0.10	5.5			
60 ～ 69	65	0.12	7.8			
70 ～ 79	75	0.08	6.0			
80 ～ 89	85	0.18	15.3			
90 ～ 100	95	0.02	1.9			
合計		1.0	49.8 (=平均)			

度数分布から分散を求める

テキストに載っている例で，計算してみましょう

階級	階級値	相対度数	階級値×相対度数	偏差	$(\text{偏差})^2$	$(\text{偏差})^2 \times$ 相対度数
0 ～ 9（点）	5	0.04	5×0.04 $= 0.2$	$5 - 49.8$ $= -44.8$	$(-44.8)^2$ $= 2007.04$	2007.04×0.04 $= 80.28$
10 ～ 19	15	0.16	2.4	−34.8	1211.04	193.77
20 ～ 29	25	0.08	2.0	−24.8	615.04	49.20
30 ～ 39	35	0.12	4.2	−14.8	219.04	26.28
40 ～ 49	45	0.10	4.5	−4.8	23.04	2.304
50 ～ 59	55	0.10	5.5	5.2	27.04	2.704
60 ～ 69	65	0.12	7.8	15.2	231.04	27.72
70 ～ 79	75	0.08	6.0	25.2	635.04	50.80
80 ～ 89	85	0.18	15.3	35.2	1239.04	223.03
90 ～ 100	95	0.02	1.9	45.2	2043.04	40.86
合計		1.0	49.8 （=平均）			696.96（=分散） $\sqrt{696.96} = 26.4$ （=標準偏差）

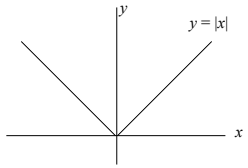
分散

分散の平方根が標準偏差

ところで，どうして2乗するの？

偏差の「マイナス」を「プラス」にしたいのなら，
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



2乗を表す関数のグラフ（放物線）には折り目はない　だから2乗を用います。

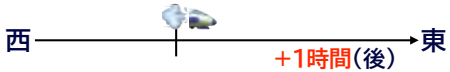
なぜマイナスかけるマイナス＝プラス？

プラスとマイナスは，「向きが反対」と考えましょう。

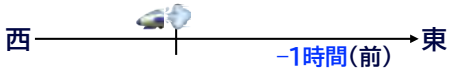
東に1km進むのが「+1km」なら
西に1km進むのは「−1km」

1時間後が「+1時間」なら
1時間前は「−1時間」

+50km/h（東向き）　　+50km/h × +1時間 = +50km（東にいる）



−50km/h（西向き）　　−50km/h × −1時間 = +50km（東にいた）



標準得点🤔

「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~~　…？

分散(標準偏差)も考えないと，答えられない

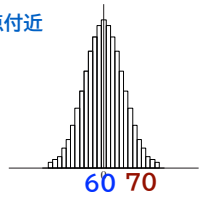
「試験で70点」は優れているのか

一緒に試験を受けた人たちが

※平均は高いがばらつきが小さく，たいていの人が60点付近

平均60点で標準偏差5点なら

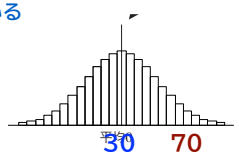
70点の人は，平均を
標準偏差の2倍上回っている



※平均は低いがばらつきが大きく，70点付近の人もいる

平均30点で標準偏差20点

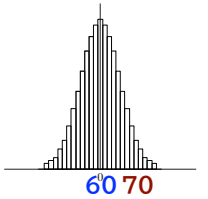
70点の人は，やはり平均を
標準偏差の2倍上回っている



70点の「地位」，
つまり受験者の中で
どのくらい優れている
かは，どちらも同じ

標準得点

平均を標準偏差の2倍上回っている



これを，[標準得点]が+2点 であるという

※標準得点は，Zスコア，Z値ともよばれます。

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

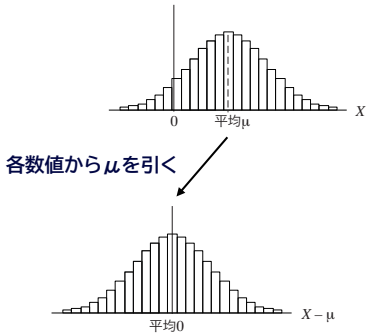
分布そのものを，
平均0，標準偏差1に「変換」したら？

その数値の変換後の値が，そのまま標準得点になる
ある人の点数を変換して「+2」になったら，
+2は標準偏差(=1)の2倍だから「標準得点2点」

分布の変換

分布中の各数値から，平均を引く

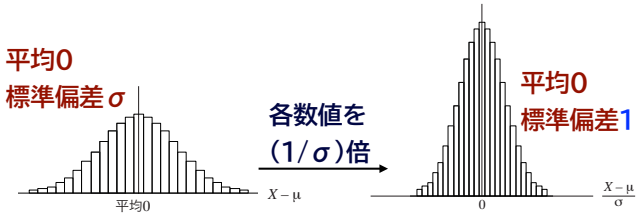
平均μ
標準偏差σ



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/σ)倍
分散は(偏差)^{2}の平均　(1/σ)^{2}倍
標準偏差は分散の平方根　(1/σ)倍



式で書くと

分布そのものをXとすると，　※分布そのものを，
数のようにひとつの文字で表す

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

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]

※わかりやすい数値で書いている。
「華氏温度」を思い出してください。

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