データの分析と処理
科学研究に必要な統計学

青森大学ソフトウェア情報学部
角田均

科学研究と統計学

実験データと統計処理

統計学の重要性

実験・研究の正しさを保証する

（例）10回の測定値と平均値

Aの測定			
10.27	10.32	10.28	10.28
10.04	10.25	10.34	10.15
10.18	10.12	10.16	10.40
10.24	10.19	10.22	10.24
10.29	10.25	10.24	10.04
10.25	10.13	10.34	10.40
10.19	10.21	10.04	10.20
10.38	10.27	10.25	10.29
10.27	10.32	10.34	10.27
10.28	10.26	10.24	10.16
10.24	10.23	10.25	10.24

Bの測定			
10.57	7.82	11.09	12.49
9.72	12.26	10.19	12.91
10.87	8.41	8.62	11.18
9.45	7.26	9.22	9.74
7.11	9.49	10.47	8.66
10.16	11.00	9.68	13.25
7.55	6.40	7.60	9.36
14.00	11.88	7.59	8.36
12.09	6.50	12.41	13.28
10.88	11.64	11.83	11.91
10.24	9.27	9.87	11.11

測定値

平均値

データの「確からしさ」を統計学で説明・記述する

統計とは

- 「統計」 ＝ データを整理して分析する

測定や調査で
データを集める

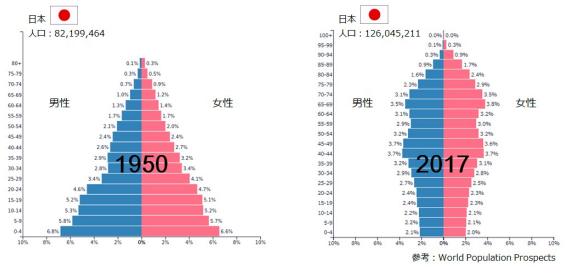
表やグラフに
整理・まとめ

分析、検証、予測

統計学の分類

- 記述統計学
  - グラフや表で大量の事象（データ）を記述する
- 数理統計学
  - 推測統計学
    - 少ないサンプルから全体を予測する
  - 多変量解析
    - 複雑に絡み合った要因を分解、因果関係を明らかにする

人口ピラミッド

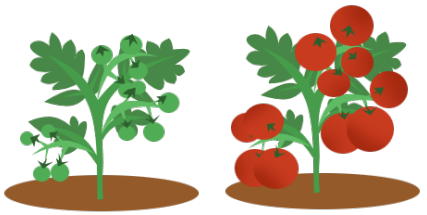


記述統計学

アンケート調査

推測統計学

水?　肥料?　温度?

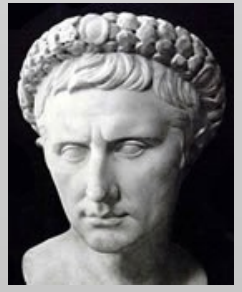


多変量解析

統計学の３つの源流

Adolphe Quetelet
(1796-1874)

近代統計学の祖



Augustus
(BC63-14)

William Petty
(1632-87)

国の実態をとらえる

Edmond Halley
(1656-1742)

大量の事象をとらえる

Blaise Pascal
(1623-62)

Pierre de Fermat
(16??-1665)

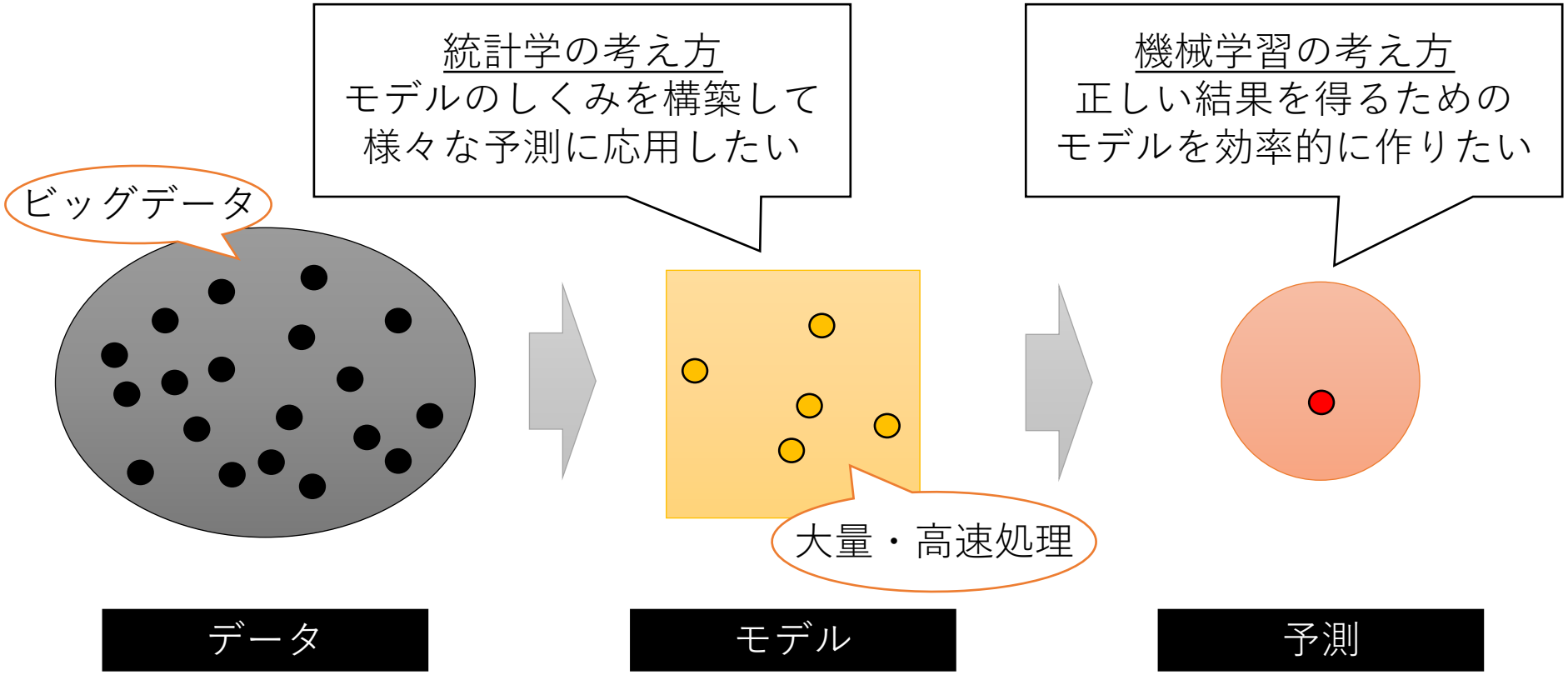
確率的な事象をとらえる

統計学の過去、現在、未来

- 統計学は難しい？
  - 学校で習わない
  - 処理、計算が大変
- 統計学は最強？
  - コンピュータの普及で変わった世界
  - ビジネスの世界に浸透
- これからの統計学
  - 「ビッグデータ」で変わる統計学
  - 「データサイエンス」「機械学習」「人工知能」

人工知能と統計学

- 機械学習は応用統計学？

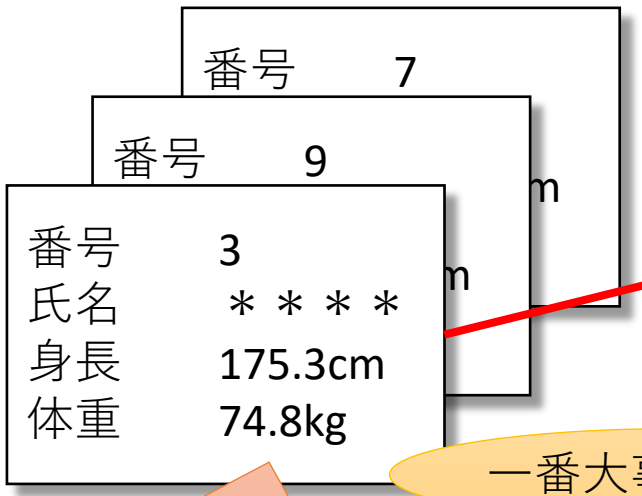


データの整理と俯瞰

個票とヒストグラム

データを整理する

- 「個票」を作る



生データ
（未加工データ）

一番大事なデータ

個票
（一次データ）

番号	身長(cm)	体重(kg)
1	169.6	63.3
2	172.2	62.8
3	175.3	74.8
4	181.2	
5	1	
6	168.9	
7	183.5	69.7
8	169.9	59.0
9	158.6	46.1
10	164.2	70.3

個体（レコード）
… 1組のデータ

「変量」

データを整理・集計

データ全体を視覚化する

- 度数分布表とヒストグラム

クラス1の得点分布

得点データ

平均値はどちらも3.0

クラス2の得点分布

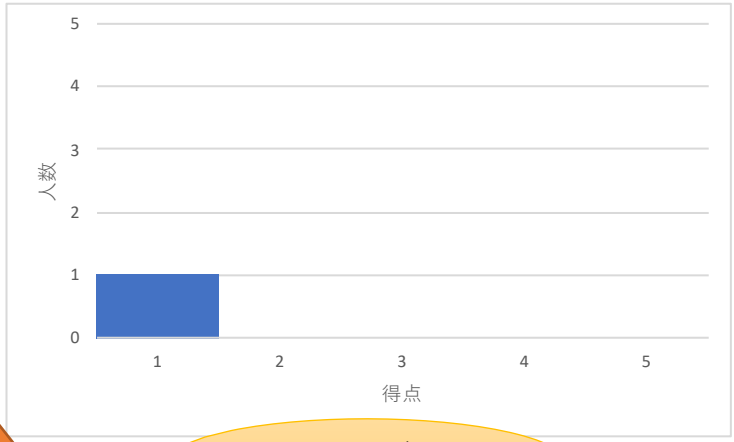
番号	得点
A	2
B	5
C	2
D	3
E	4
F	1
G	3
H	3
I	4
J	3

番号	得点
A	1
B	1
C	4
D	5
E	5
F	3
G	5
H	1
I	4
J	1

得点区間	人数
1	1
2	
3	
4	
5	

度数分布表

得点区間	人数
1	
2	
3	
4	
5	



ヒストグラム

データの代表値

平均値、中央値、最頻値

データの代表値

- 主な代表値
  - 平均値
  - 中央値（メジアン）
  - 最頻値（モード）
  - 最大値・最小値
  - レンジ

データ全体の性質を示す

平均値

- データを平らに均(なら)した値

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$ （平均値の公式）

例）5人分のテストの成績

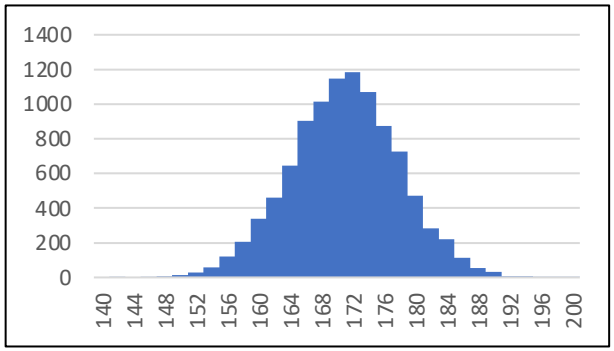
番号	得点
A	7
B	5
C	3
D	9
E	6

$$\frac{7+5+3+9+6}{5} = \frac{30}{5} = 6$$

平均値は最も重要な代表値

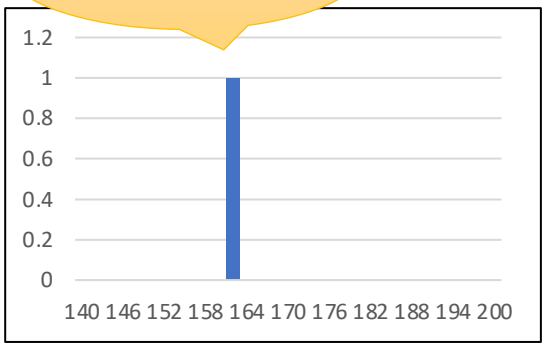
大数の法則

- サンプルの平均値は、サンプル数を増やすと真の平均値に近づく

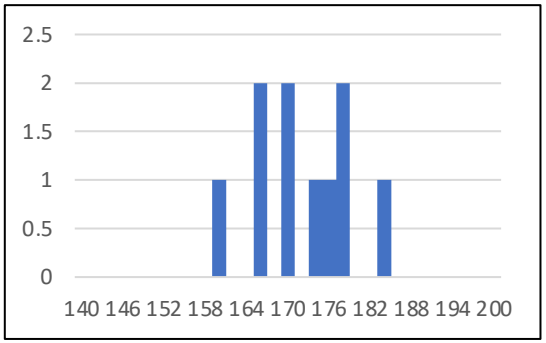


（母集団10000個）平均170.0

偶然、異常値を得る可能性もある

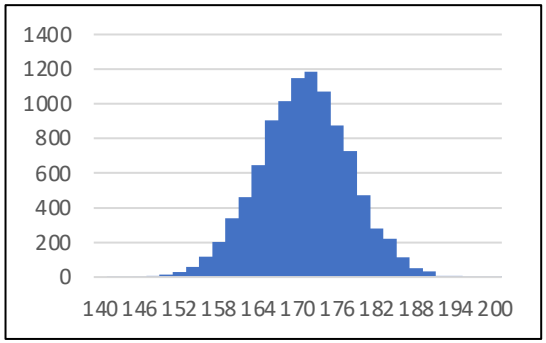


（サンプル1個）平均162.6



（サンプル10個）平均171.4

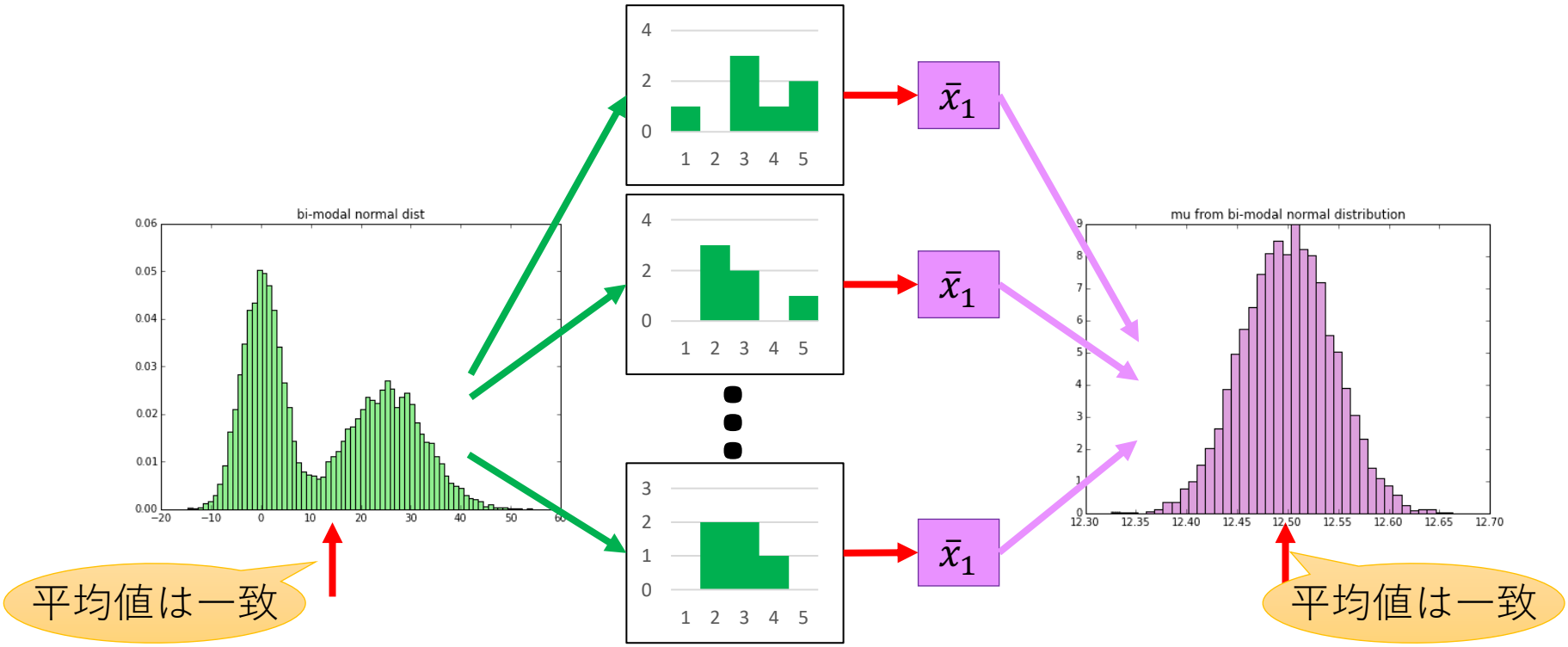
…



（サンプル10000個）平均170.0

中心極限定理

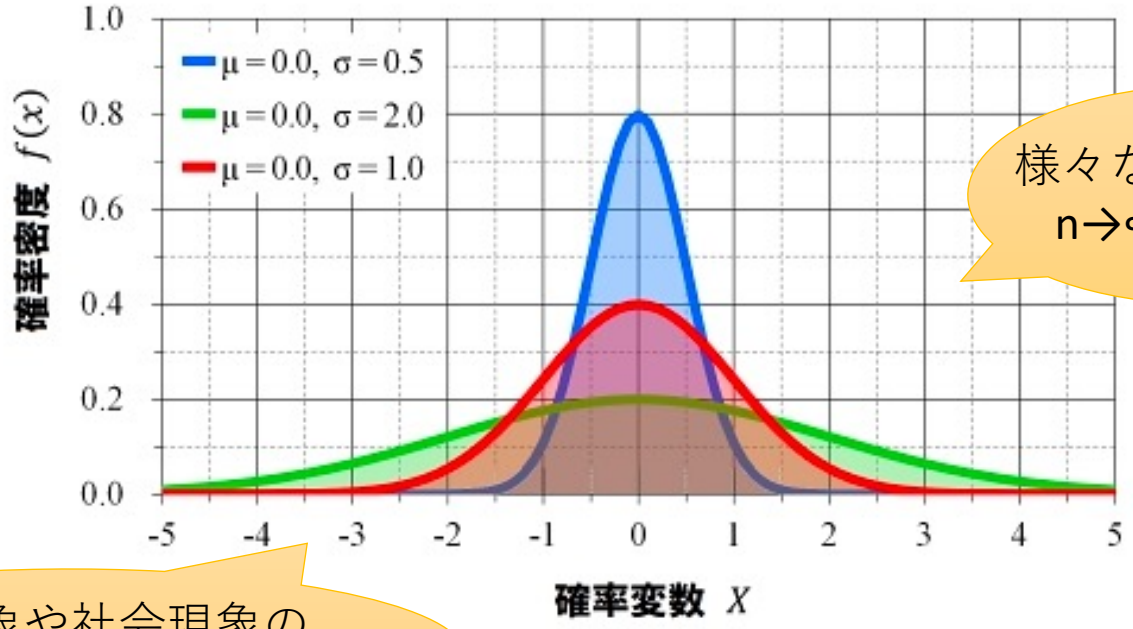
- 元がどんな分布のデータでも、サンプルの平均値は母集団の平均値の周辺に正規分布する



正規分布

- 定義

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

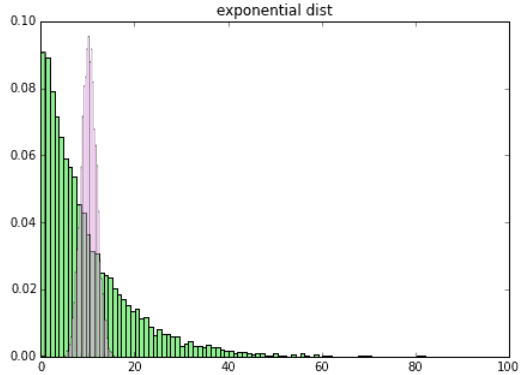


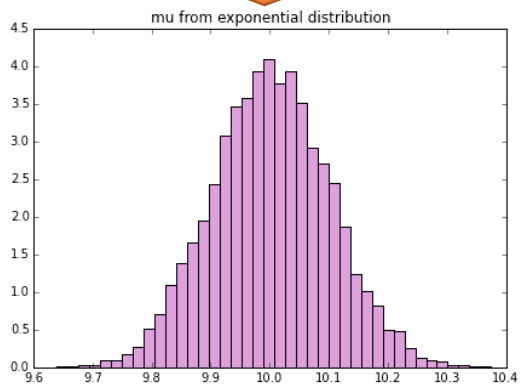
様々な分布関数の
n→∞の極限値

自然現象や社会現象の
多くがこの分布で説明できる

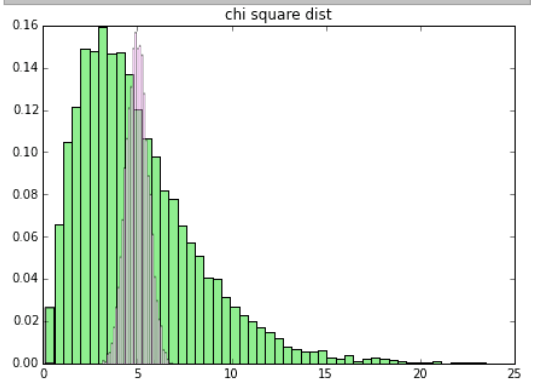
実際にやってみる

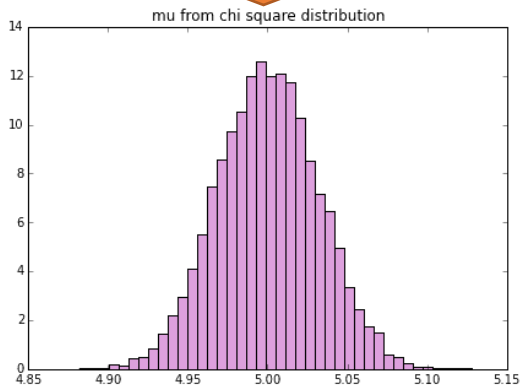
指数分布



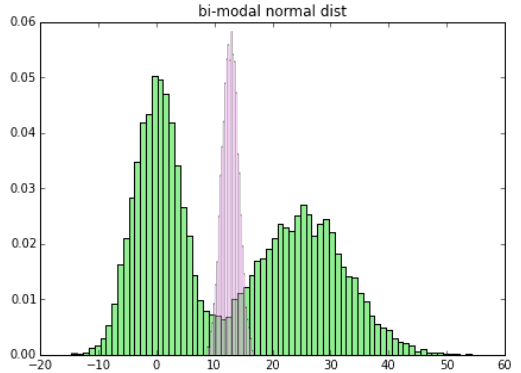


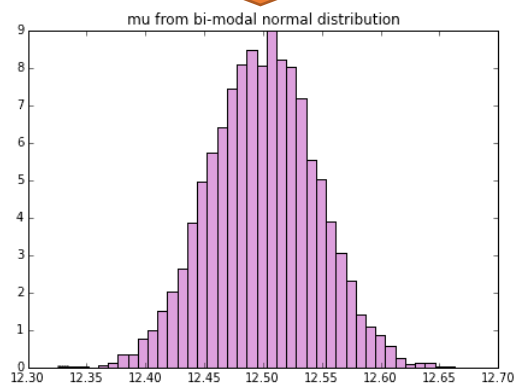
カイ２乗分布





双峰正規分布

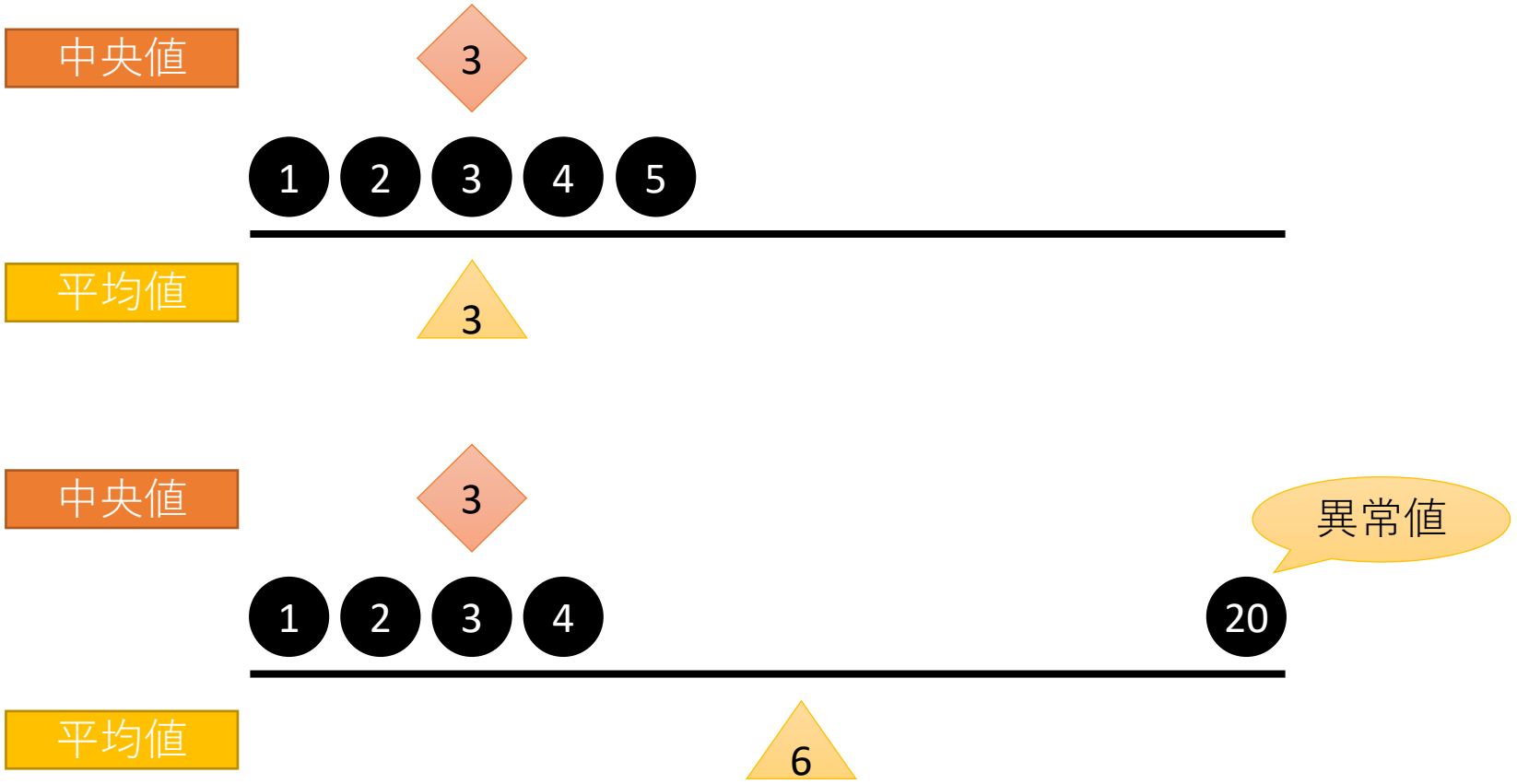




(参考) http://qiita.com/kenmatsu4/items/351284ef430bcfd2c8ed

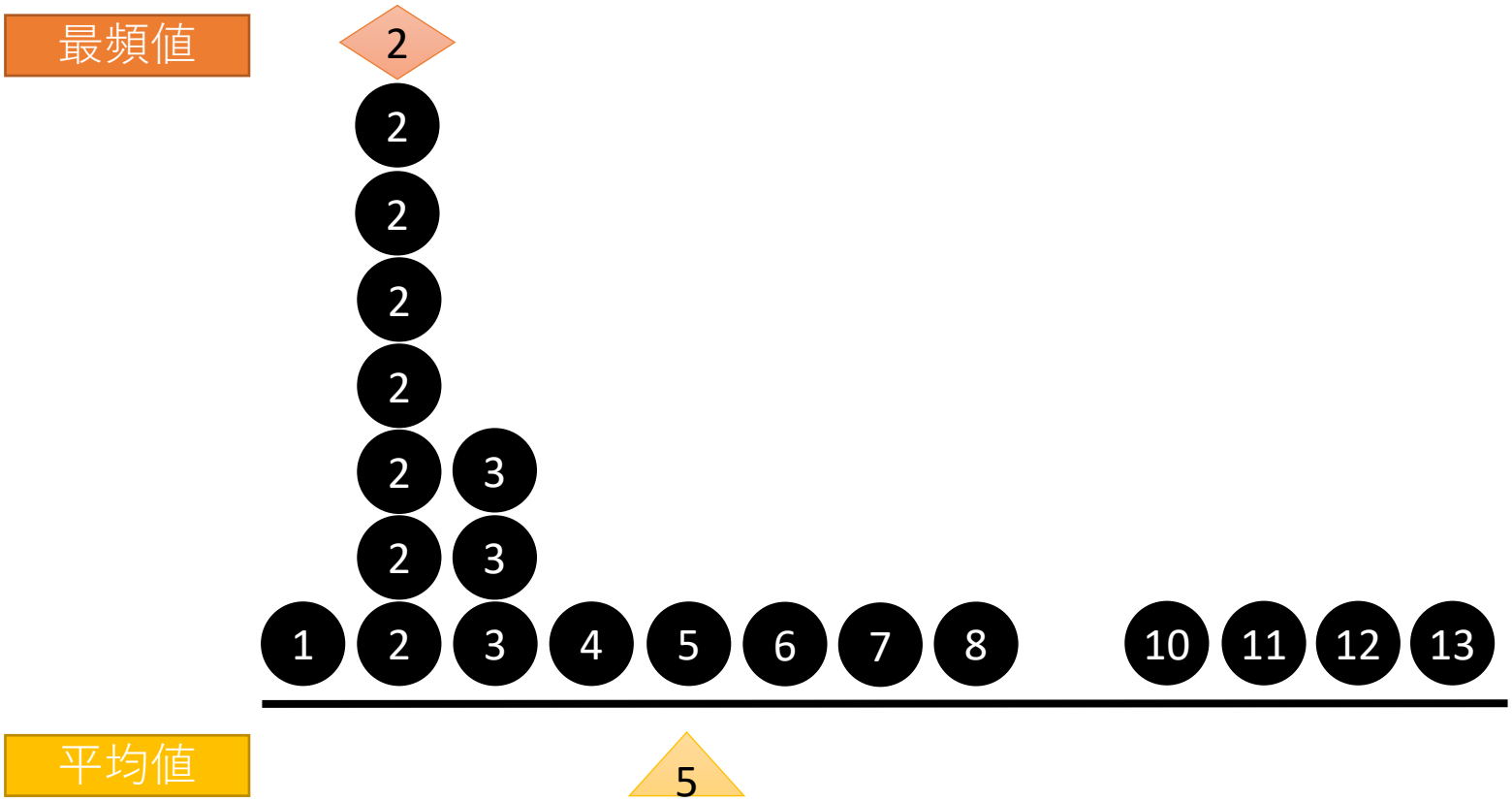
中央値（メジアン）

- 中央値 = 変量の真ん中の値



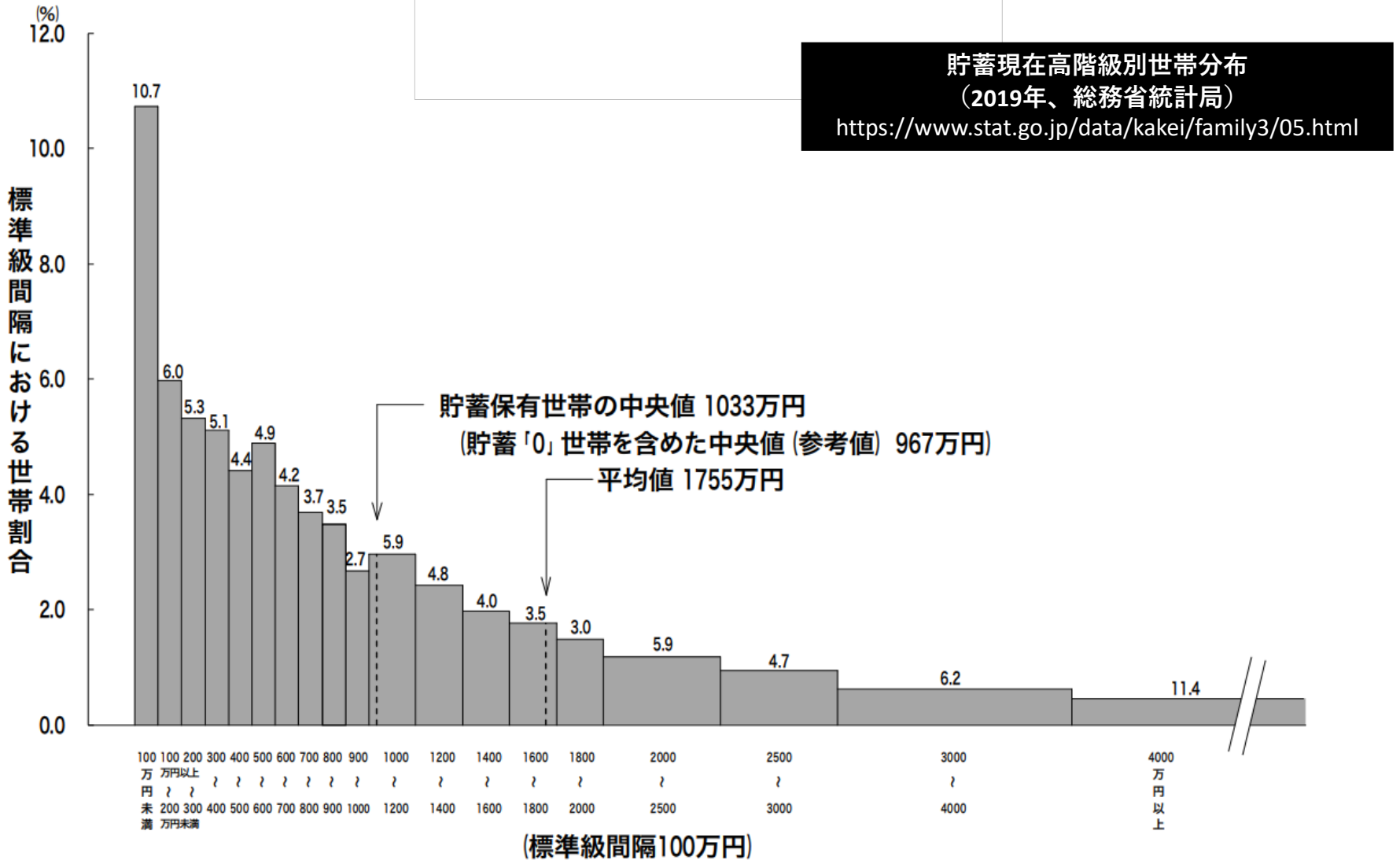
最頻値（モード）

- 最頻値 = 最も頻繁に現れる値



一致するデータの数が一番多い

どれが「代表値」？

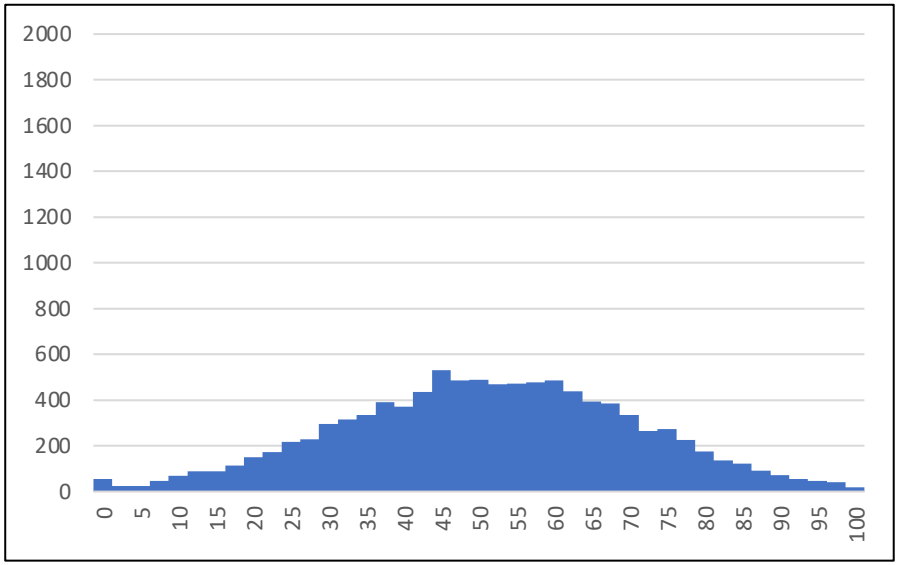

貯蓄現在高階級別世帯分布
（2019年、総務省統計局）
https://www.stat.go.jp/data/kakei/family3/05.html
(%)
標準級間隔における世帯割合
12.0
10.0
8.0
6.0
4.0
2.0
0.0
10.7
6.0
5.3
5.1
4.4
4.9
4.2
3.7
3.5
2.7
5.9
4.8
4.0
3.5
3.0
5.9
4.7
6.2
11.4
貯蓄保有世帯の中央値 1033万円
(貯蓄「0」世帯を含めた中央値（参考値） 967万円)
平均値 1755万円
100万円未満
100万円以上200万円未満
200～300
300～400
400～500
500～600
600～700
700～800
800～900
900～1000
1000～1200
1200～1400
1400～1600
1600～1800
1800～2000
2000～2500
2500～3000
3000～4000
4000万円以上
(標準級間隔100万円)

データのばらつき

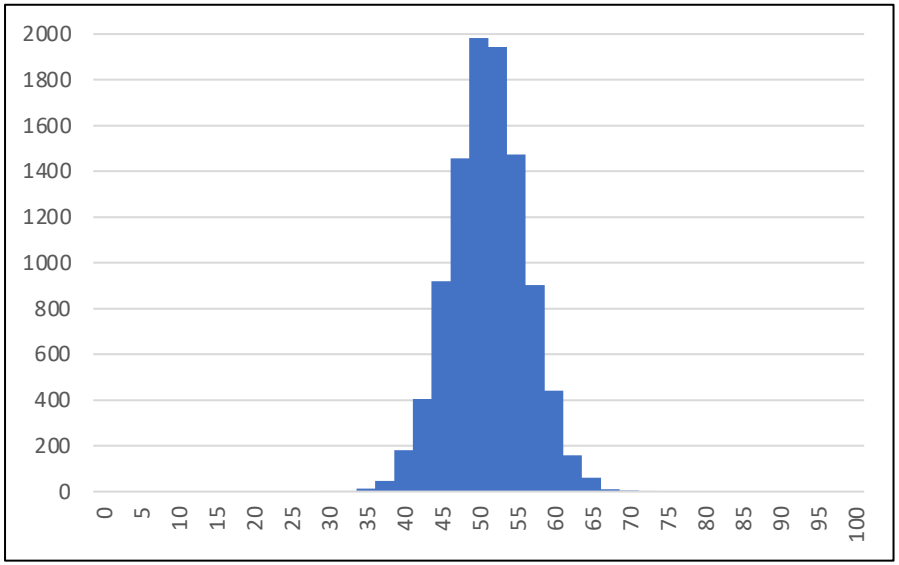
標準偏差

ばらつき度合いの表現

- 同じ平均値でも、ばらつきの違うデータ



データ数10000、平均値50.0



データ数10000、平均値50.0

ばらつき度合いを表す指標

- 「偏差」 … 平均値との差

$$\text{偏差} = x - \bar{x}$$

- 「分散」 … 偏差の2乗の平均

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

- 「標準偏差」 … 分散の平方根

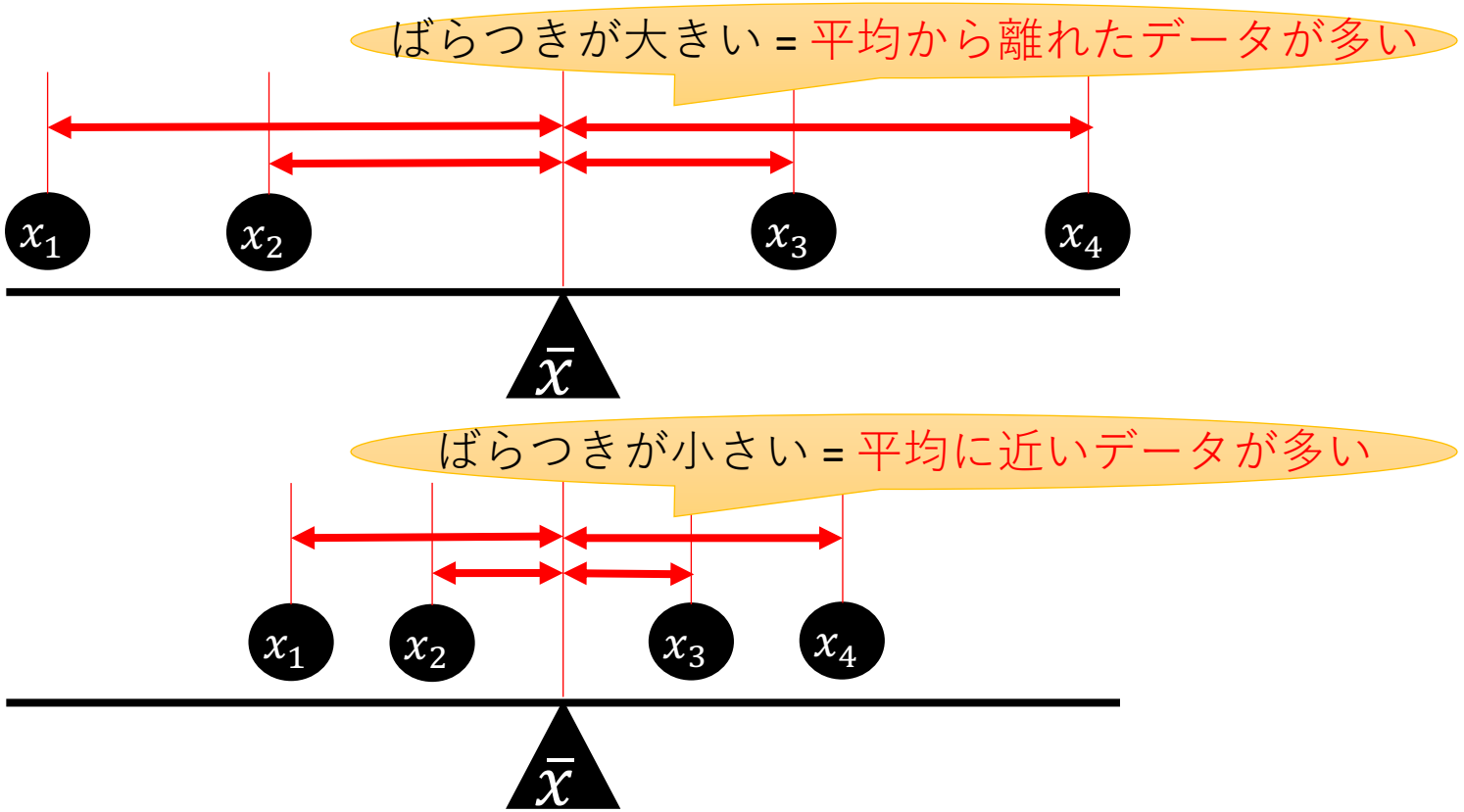
$$\sigma = \sqrt{\sigma^2}$$

偏差

- 偏差 … 平均値との差

個別のデータごとの指標

偏差$= x - \bar{x}$

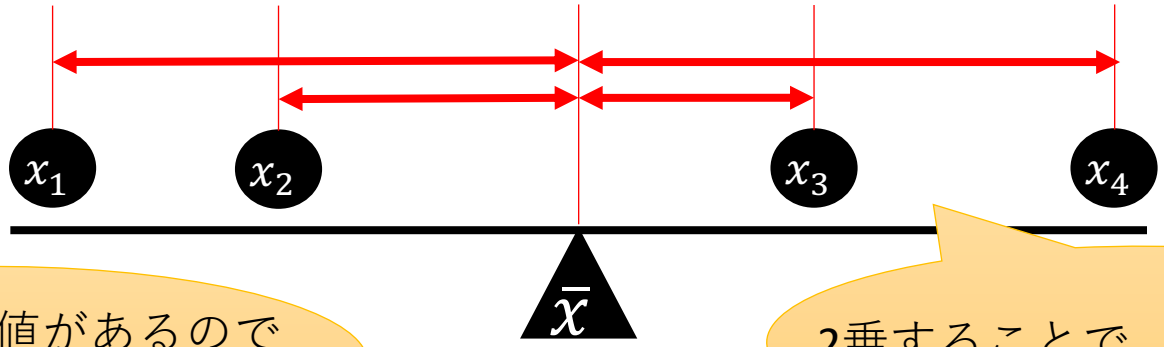


分散

データ全体の指標

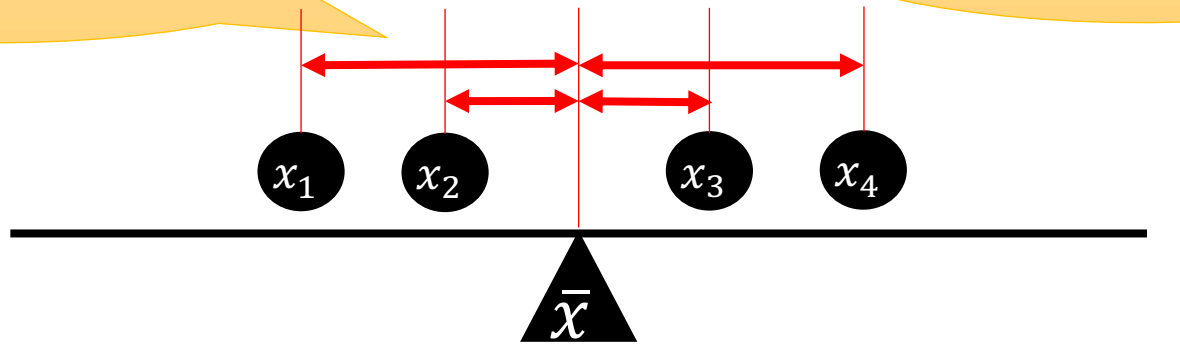
- 分散 … 偏差の2乗の平均

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



偏差は正負の値があるので
偏差のまま合計すると0になる

2乗することで、差を強調する



標準偏差

- 標準偏差 … 分散の平方根

$$\sigma = \sqrt{\sigma^2}$$

偏差$= x - \bar{x}$ — 元のデータと同じ単位

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}$$ — 元のデータの単位の2乗

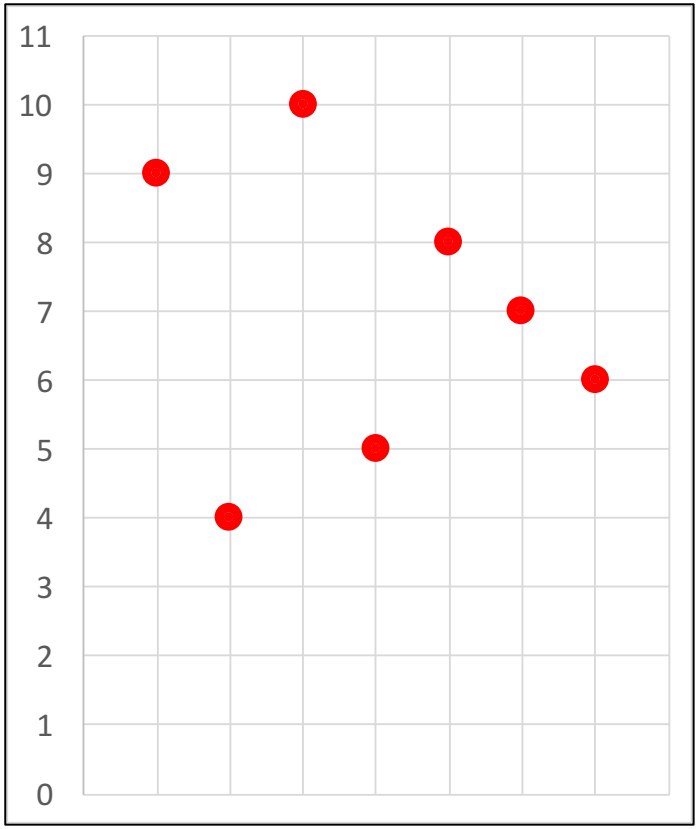
$$\sigma = \sqrt{\sigma^2}$$ — 元のデータと同じ単位

標準偏差は元のデータと直接比較できる

標準偏差の意味

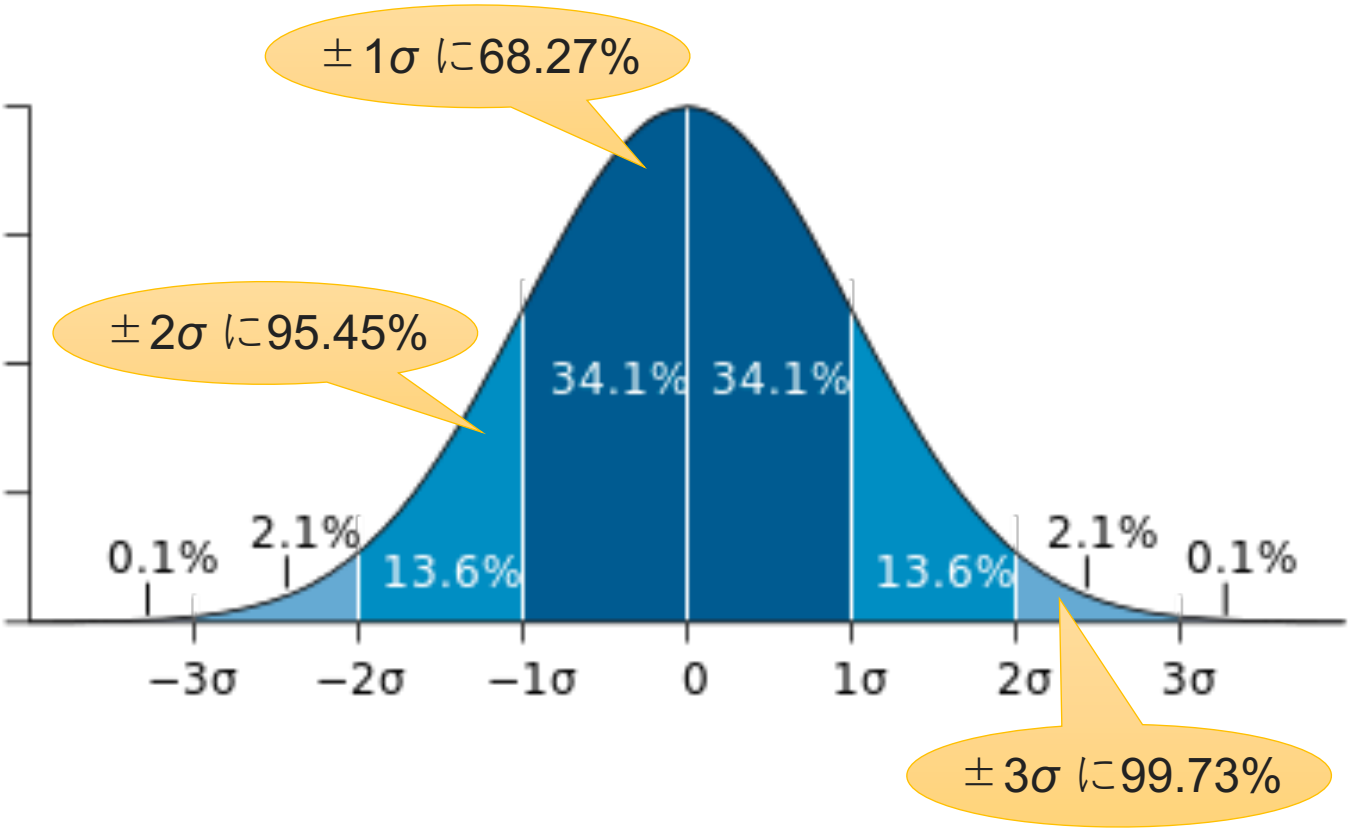
（例）7人分のテスト

番号	得点	偏差	偏差の2乗
A	9		
B	4		
C	10		
D	5		
E	8		
F	7		
G	6		



正規分布と標準偏差

- 平均値の周りにデータが含まれる確率



標本標準偏差と母標準偏差

- 標本の場合はnをn-1に置き換える。
- 理由
  - 標本の平均値を使うと、標本の偏差が最小になる（標本の平均値がそのように定義されている）
  - この、標本の平均値によって拘束された１次元分の自由度の減少を反映させる
- Excelの場合
  - STDEV.S … 標本標準偏差
  - STDEV.P … 母標準偏差

標準誤差

- 標準誤差 … 推定値（平均値）の標準偏差

$$\text{標準誤差} = \frac{\sigma}{\sqrt{n}}$$

2種類の測定データ

①多くのバラついた値を測定する場合
　集団全体の特性を記述・分析する
　例）40人のクラスのボール投げの距離を集計した

平均値 + 標準偏差

②唯一の正しい値を測定する場合
　誤差を含む測定から、正しい値を推定する
　例）学校から自宅までの距離を10回測った

平均値 + 標準誤差

標準誤差の意味

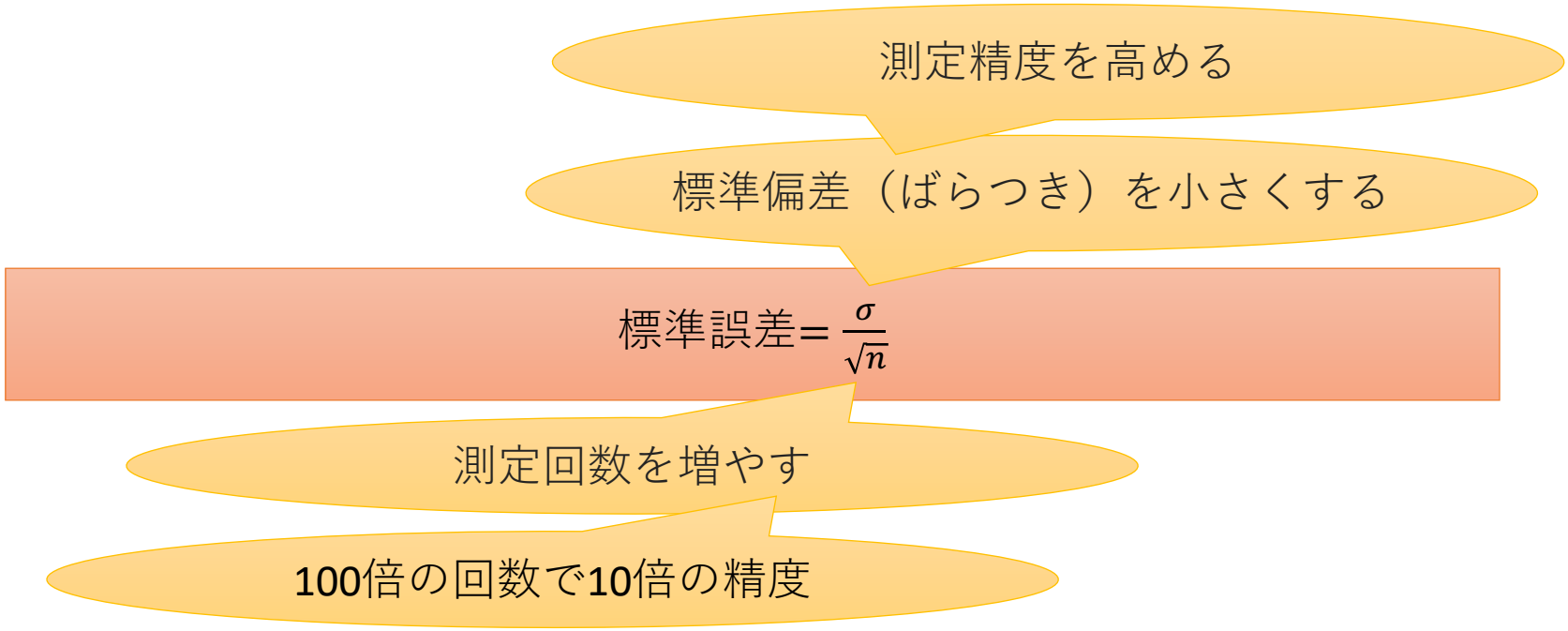
（例）10回の測定値と平均値と標準誤差

Aの測定			
10.57	7.82	11.09	12.49
9.72	12.26	10.19	12.91
10.87	8.41	8.62	11.18
9.45	7.26	9.22	9.74
7.11	9.49	10.47	8.66
10.16	11.00	9.68	13.25
7.55	6.40	7.60	9.36
14.00	11.88	7.59	8.36
12.09	6.50	12.41	13.28
10.88	11.64	11.83	11.91
10.24	9.27	9.87	11.11

Bの測定			
10.27	10.32	10.28	10.28
10.04	10.25	10.34	10.15
10.18	10.12	10.16	10.40
10.24	10.19	10.22	10.24
10.29	10.25	10.24	10.04
10.25	10.13	10.34	10.40
10.19	10.21	10.04	10.20
10.38	10.27	10.25	10.29
10.27	10.32	10.34	10.27
10.28	10.26	10.24	10.16
10.24	10.23	10.25	10.24

測定の信頼度を上げる

- 標準誤差を小さくするために

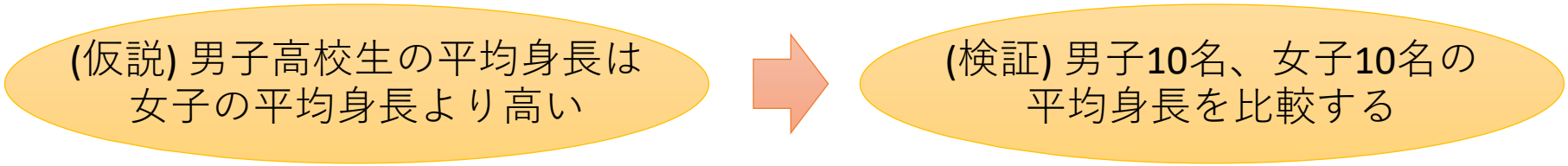


仮説の検証

統計的仮説検定

仮説と検定

- 仮説
  - 母集団の性質(平均値、分散、比率など)を説明するために立てた仮定(命題)

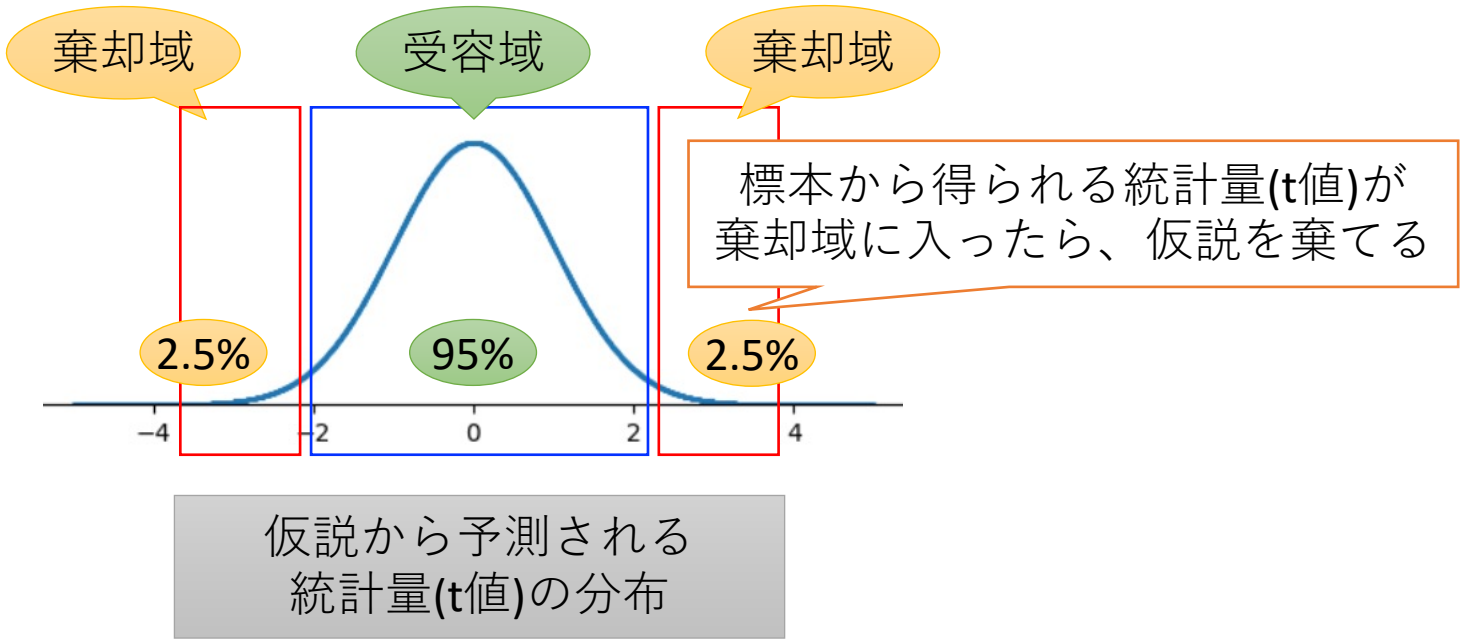


- 検定
  - 標本(サンプル)から仮説を検証する
    - 仮説から導かれる結果が得られているか
    - 仮説が正しく測定結果を説明しているか
    - 偶然ではないのか

仮説検定の基本

- 基本的な考え方

起こりづらいことが起きた → **仮説が間違っている**



仮説から予測される
統計量(t値)の分布

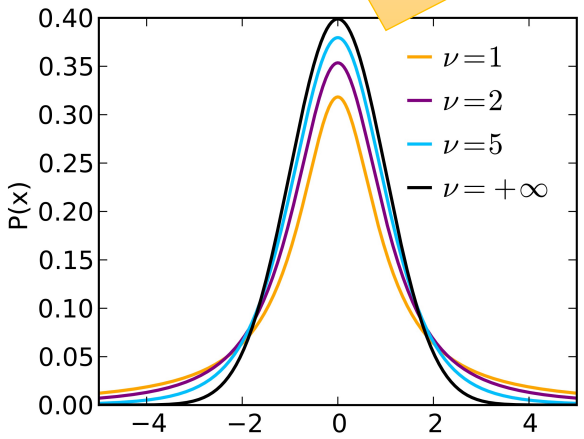
仮説検定の基本

背理法
帰無仮説の方を検証→棄却することで
対立仮説の正しさを主張する

- 仮説の設定
  - 帰無仮説 … 支持したい仮説の逆の命題を設定
  - 対立仮説 … 支持したい仮説
- t検定
  - 母集団の平均値の違いを検定
    - 母集団が正規分布
    - 標本数が少ない場合
  - 標本のt値はt分布に従う

t値　$T = \dfrac{\bar{X} - \mu}{s/\sqrt{n}}$

スチューデントのt分布



自由度νのt分布
(Wikipediaより)

仮説検定の例

表. 投薬前後の体温測定

患者	投薬前	投薬後	低下体温
1	35.4	35.3	0.1
2	38.8	36.2	2.6
3	37.0	34.8	2.2
4	38.2	38.1	0.1
5	38.8	36.2	2.6
6	37.9	36.6	1.3
7	38.5	36.2	2.3
8	37.6	36.0	1.6
9	38.7	36.6	2.1
10	39.4	36.5	2.9

- 問題設定: 解熱剤の効果
  - 10人の患者に投薬
  - 投薬前と投薬後の体温を測定
  - 解熱剤は効いたのか？
- 仮説
  - 「解熱剤によって体温が下がった」
- 検定手順
  - 帰無仮説: 投薬前と投薬後の平均体温に差はない
  - 対立仮説: 投薬前と投薬後の平均体温に差がある
  - 帰無仮説の矛盾を示す(棄却する)
- 利用するExcelの関数
  - 平均 … average(データ範囲)
  - 標準偏差 … stdev.s(データ範囲)
  - 平方根 … sqrt(値)
  - t検定 … t.test(系列1, 系列2, 検定の種類, データの種類)

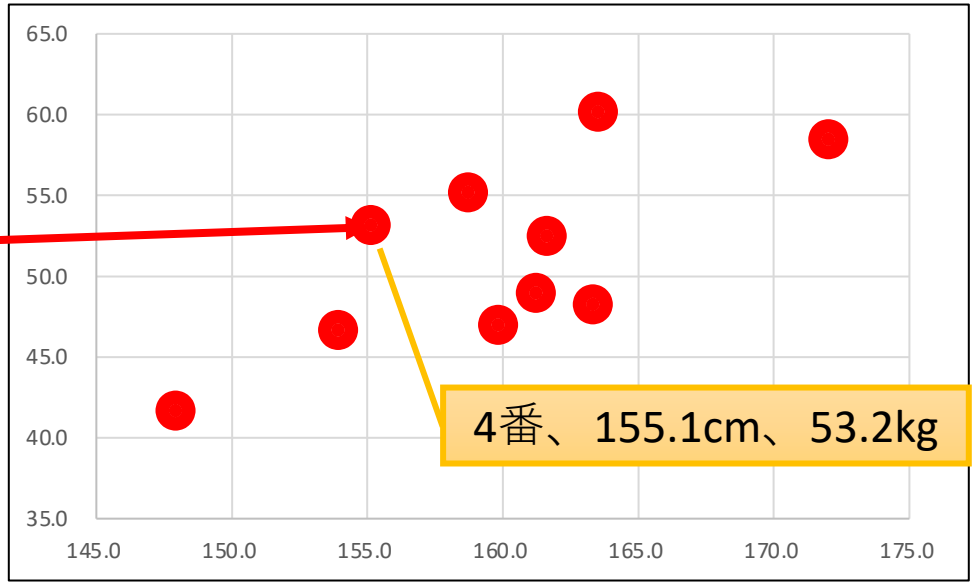
変量同士の関係

回帰分析と相関係数

散布図と相関

（例）身長と体重を同時にプロット

番号	身長	体重
1	147.9	41.7
2	163.5	60.2
3	159.8	47.0
4	155.1	53.2
5	163.3	48.3
6	158.7	55.2
7	172.0	58.5
8	161.2	49.0
9	153.9	46.7
10	161.6	52.5

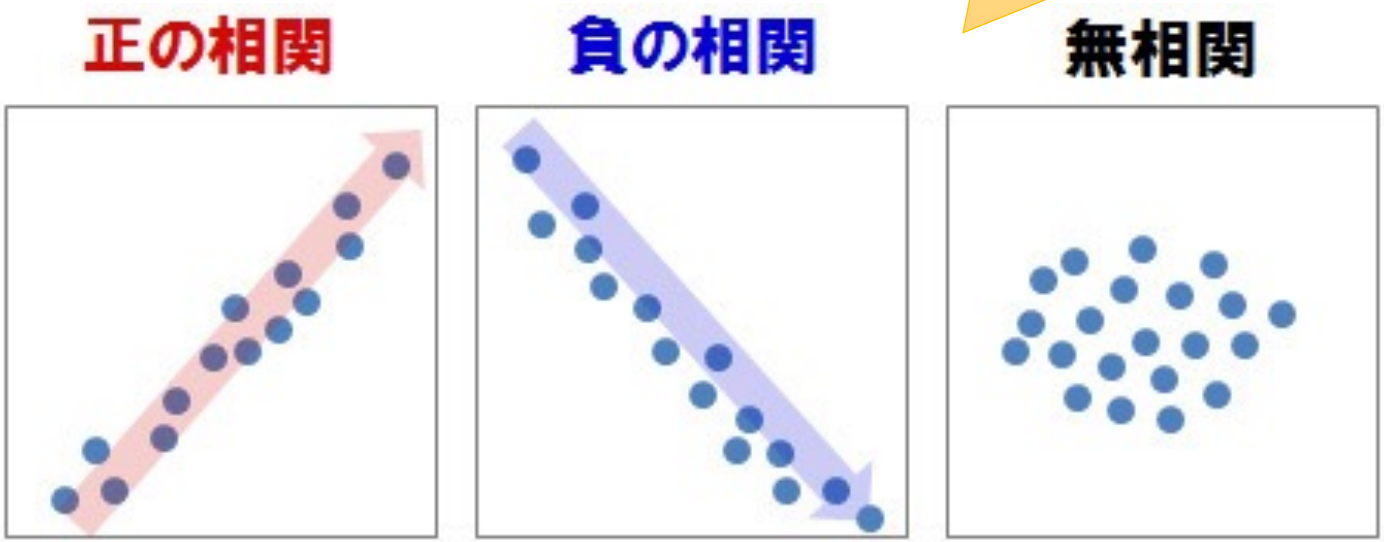


2つの変量による分布に関連性(「相関」)は見られるか？

相関

- 変量間の関係

一方の変量の変化に対する
他方の変量の変化の傾向



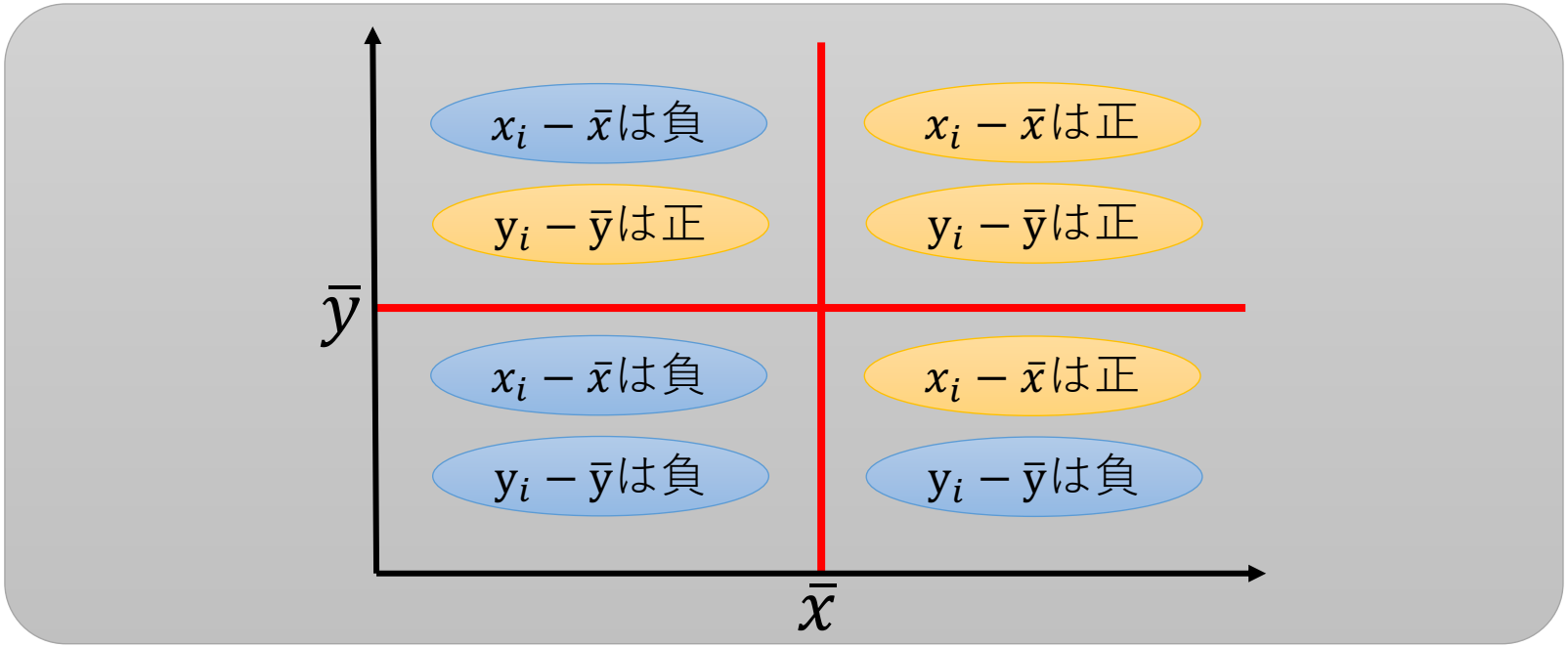
一方が増えれば 他方も増える	一方が増えれば 他方が減る	どちらでもない

共分散

- 定義

$$s_{xy} = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

$s_{xy} > 0$ならば正の相関　　$s_{xy} \fallingdotseq 0$ならば無相関　　$s_{xy} < 0$ならば負の相関

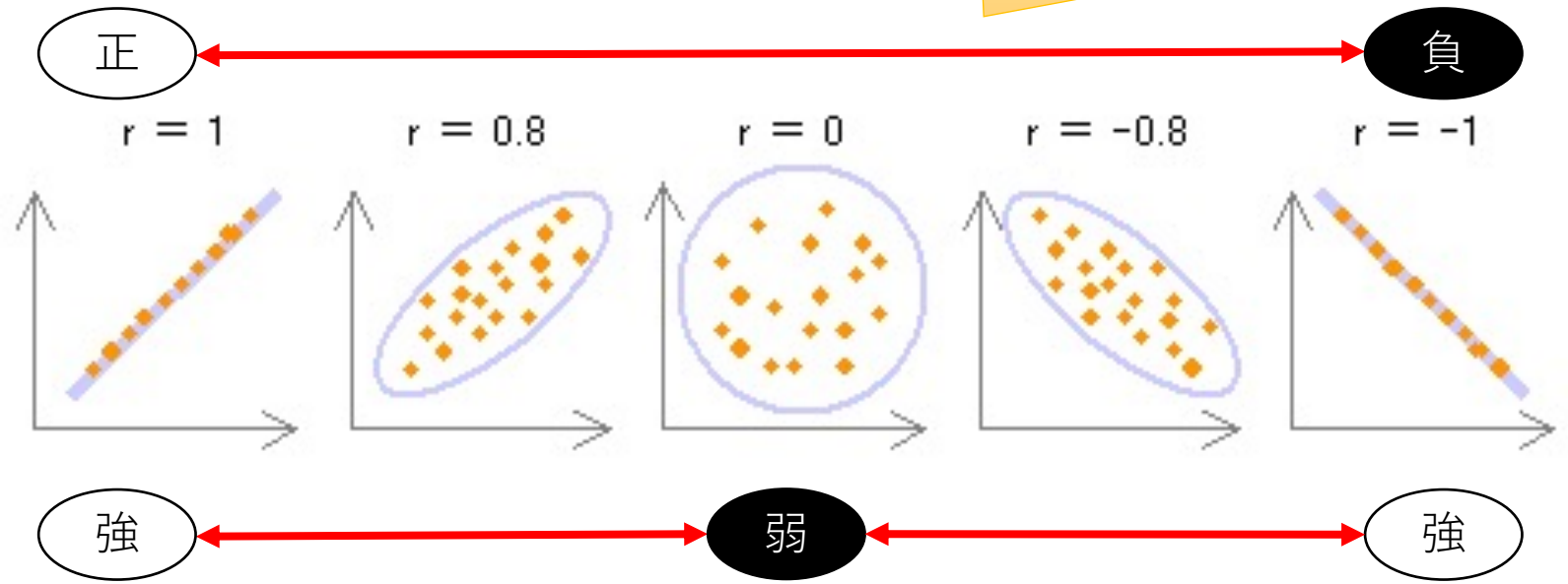


相関係数

- 定義

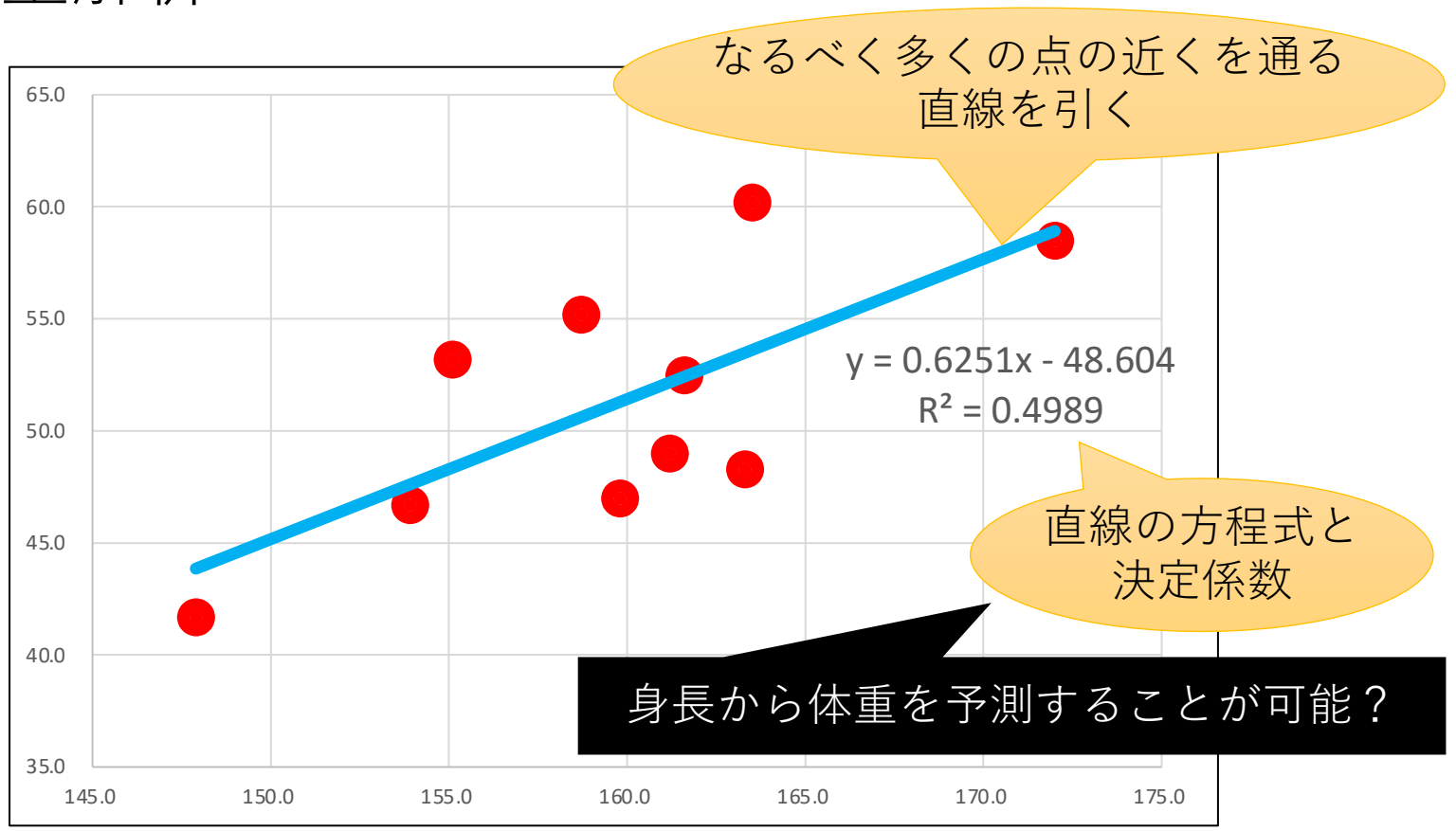
$$r_{xy} = \frac{s_{xy}}{\sigma_x \sigma_y}$$

相関係数は−1〜1の範囲で
相関の正負と強弱を示す



回帰直線

- 多変量解析


なるべく多くの点の近くを通る
直線を引く
y = 0.6251x - 48.604
R² = 0.4989
直線の方程式と
決定係数
身長から体重を予測することが可能？
65.0
60.0
55.0
50.0
45.0
40.0
35.0
145.0
150.0
155.0
160.0
165.0
170.0
175.0

最後に

実験科学と統計学

統計学の位置づけ

- 実験科学の流れ