CC BY NC ND

2017年度日本疫学会スライドコンテスト受賞作品

基本的な統計
ーデータの分布とばらつきー

藤田保健衛生大学医学部
公衆衛生学
柿崎　真沙子

CC BY NC ND

基本的な統計

① 母集団と標本
② 誤差・ばらつきを表す要約値
③ データの分布
④ 標準化と偏差値

CC BY NC ND

母集団と標本

- **母集団**：調査対象とする（推測結果を適用したい）個体の集まり、集団
  - 母集団は一般的に全数調査が困難
    - 例：日本人の平均身長を知りたい！
- **標本**：無作為に抽出された個体
    - 例：電話帳から100名無作為に抽出
- Ａ県在住の地域住民（**＝母集団**）の生活習慣を調査したいのに、特定の疾患を持つ患者だけを調査しても偏りが生じてしまう（**＝誤差**）
- 母集団から偏りなく標本を集めることが重要

CC BY NC ND

例）2群の平均を比較する

- 某アイドルグループ全体を母集団とした場合、チームAは標本として適切か？
- チームAとチームBの平均年齢に違いは有るか？

どうやって検討したらいいだろうか？

（その前段階として分布とばらつきを検討）

CC BY NC ND

まずは誤差とばらつき

- 誤差・ばらつき
- 誤差・ばらつきを表す要約値
  - 平均
  - 偏差
  - 偏差平方和
  - （不偏）分散
  - 標準偏差(S)
  - 標準誤差

CC BY NC ND

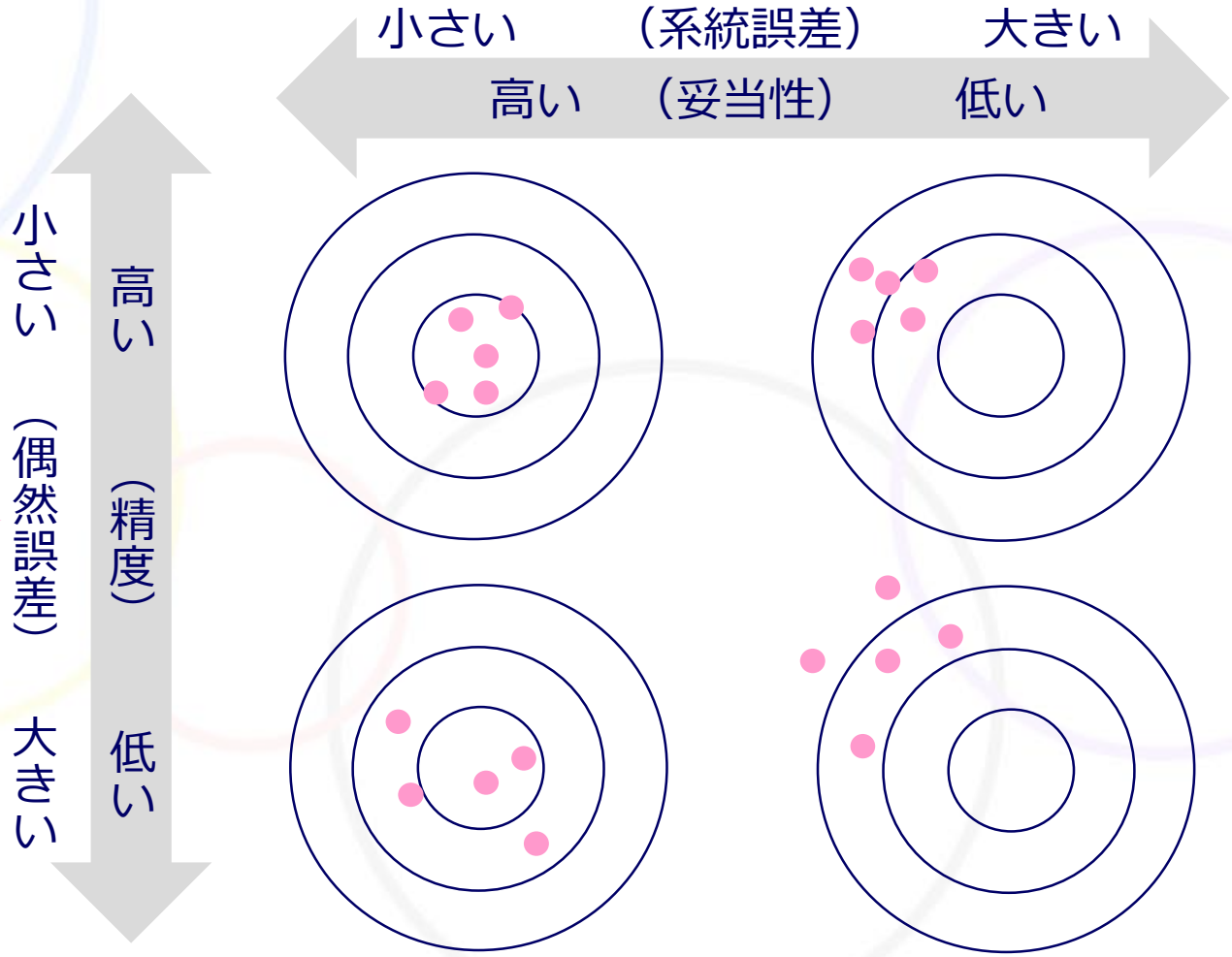
誤差・ばらつき

- **偶然誤差（Random error）**
  - 偶然によってのみ生ずるばらつき
  - 真の値を中心にばらつきが分布
  - 偶然を支配する法則（統計学）により対処可能

- **系統誤差（Systematic error）**
  - 中心が真の値から偏っている
  - バイアス（Bias）とも言う
  - 統計学では対処不能

CC BY NC ND

誤差・ばらつきの多い少ない



統計で
対処可能

CC BY NC ND

どうやって誤差・ばらつきを表すか？

- 平均
- 分散（Variance）
- 標準偏差（Standard deviation：SD）
- 標準誤差（Standard error：SE）
- 分位点
- パーセンタイル
- 範囲（Range）

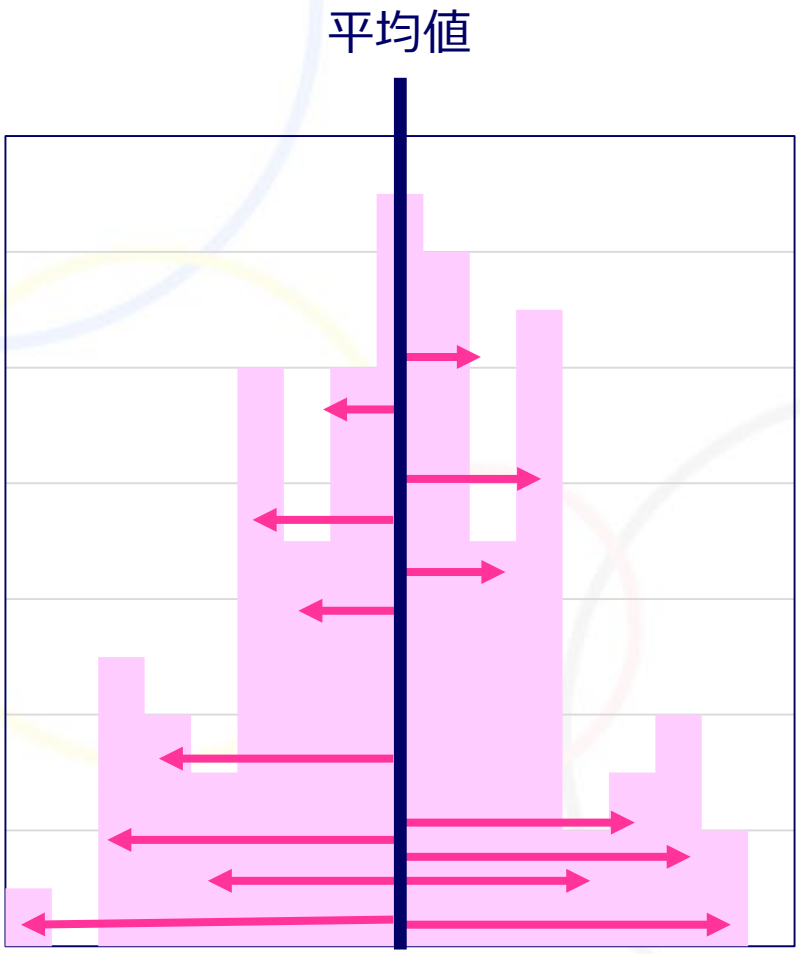
ばらつきを表す要約値

N個のデータをそれぞれ、X1, X2, X3・・・Xnとする

- **平均 ($\bar{X}$)**：(X1+X2+X3+・・・Xn)/n
- **偏差**：（データ－平均）
  - (X1- $\bar{X}$)、(X2- $\bar{X}$)、(X3- $\bar{X}$)、・・・(Xn- $\bar{X}$)
- **偏差平方和**：偏差の二乗を全てのデータで合計
  - $(X1-\bar{X})^2+(X2-\bar{X})^2+(X3-\bar{X})^2+\cdots(Xn-\bar{X})^2$

CC BY NC ND

分　散



- 平均から個々のデータがどのくらい離れているか？
- 「**平均からの離れ具合の平均値**」が分散
- 偏差平方和をn-1で割ったもの

CC BY NC ND

標準偏差

- **分散の平方根**
- 標準偏差はデータそのものの分布の広がり幅 (ばらつき) をみる尺度
  - 散らばり具合を見る
- 平均値と標準偏差：この2つがわかれば、データの分布 がある程度明らかに

CC BY NC ND

標準誤差

- **真の平均値と標本の平均値のズレ**が標準誤差
- 標準誤差の値が大きいほど推定精度が低い

- 日本人全員の平均身長を知りたい！
  - 真の平均値
- 電話帳からランダムに100名選択
  - 標本の平均値

ズレ

CC BY NC ND

例）チームAとチームBの年齢

チームA	チームB
20	24
25	28
23	24
20	22
28	21
23	25
22	26
24	27
25	27
18	23
27	27
20	22
23	28
29	23
17	25

※架空のデータです

CC BY NC ND

では計算してみよう！

		チームA	チームB
平均年齢	すべて足してNで割る		
偏差平方和	(データ－平均) ^{2}を全てのデータで合計		
（不偏）分散	偏差平方和をn-1で割ったもの		
標準偏差	分散の平方根		
標準誤差	標準偏差を√nでわる		

CC BY NC ND

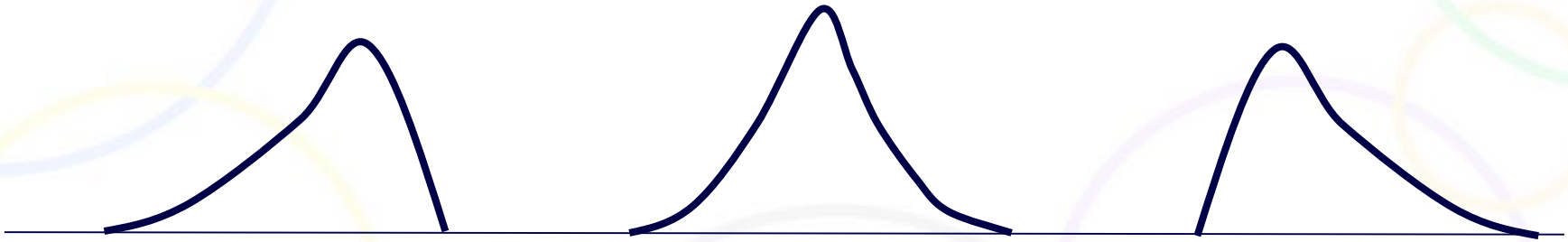
では計算してみよう！

		チームA	チームB
平均年齢	すべて足してNで割る	344/15 =22.93	372/15=24.80
偏差平方和	(データ－平均) ^{2}を全てのデータで合計	$(20-22.93)^2$+・・・・ =174.933…	$(24-24)^2$+・・・・ =74.4
（不偏）分散	偏差平方和をn-1で割ったもの	174.933/14＝12.495	74.4/14＝5.314
標準偏差	分散の平方根	√12.495・・＝3.535	√5.31・・＝2.305
標準誤差	標準偏差を√nでわる	√3.534/√15＝0.9127	√2.227/√15＝0.5952

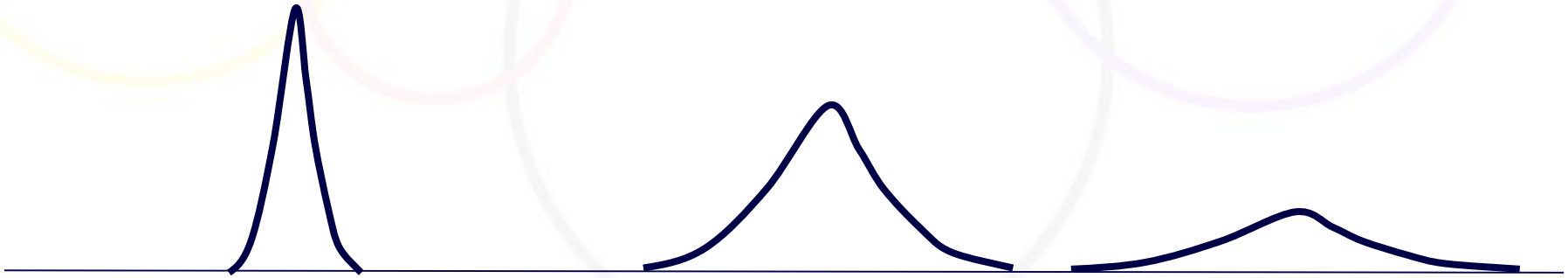
CC BY NC ND

データの分布を表す指標

歪度 (ひずみ、skewness)：分布の左右対称度



尖度 (とがり、kutosis):分布の裾の長さ



CC BY NC ND

✓「正規分布」と言っても様々な分布の形がある

分布の形が異なると比較がしにくい

↓

形をそろえて比較しやすくする

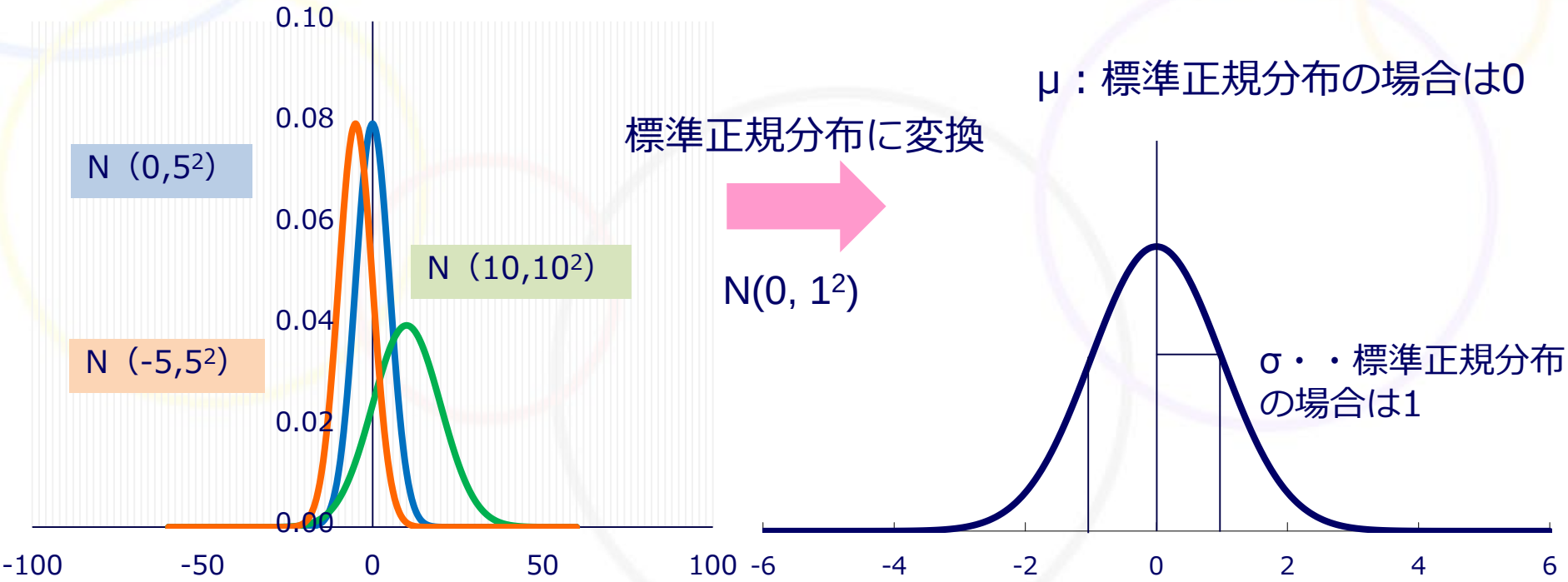
↓

標準化（z変換）

CC BY NC ND

標本分布を標準化（Ｚ変換）

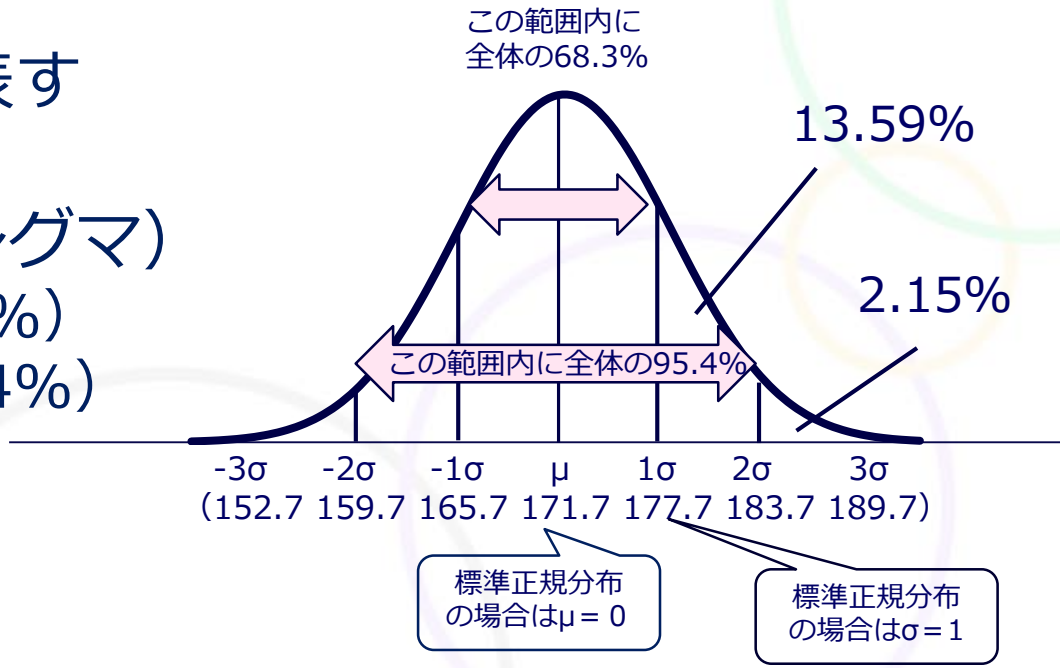
- N (μ, σ^2) を N(0, 1^2) に変換できる
- 全てのデータを標準化した時、変換後の値を標準得点（ｚ値）と呼ぶ
- Z値＝「どのくらい平均から離れているか」を示す。



CC BY NC ND

（おさらい）正規分布とは

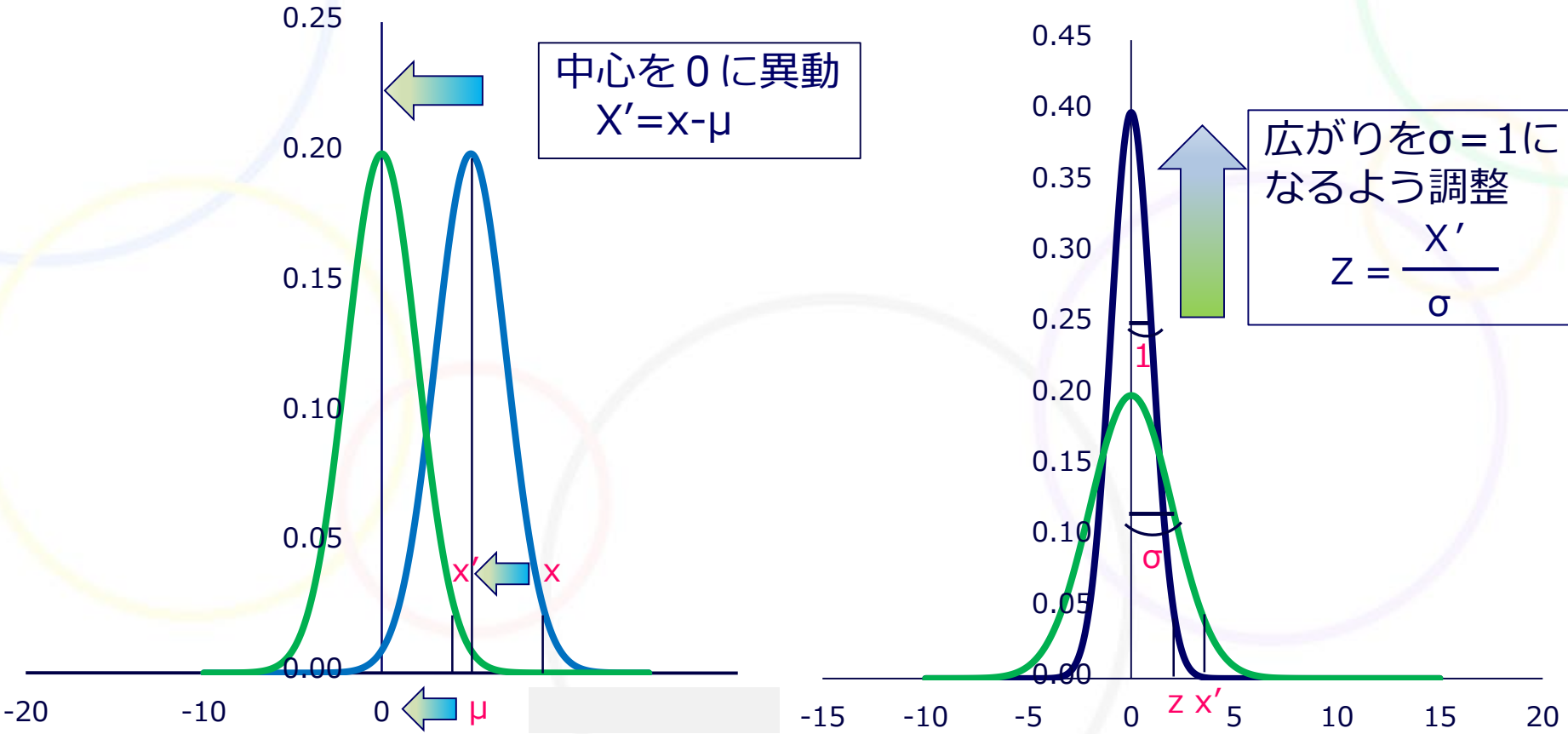
- 正規分布N (μ, σ^2)と表す
- 平均値：μ（ミュー）
- 標準偏差 (SD)：σ（シグマ）
  - ±1SDに約2/3（68.3%）
  - ±2SDに約95%（95.4%）
  - ±1.96SDに95%
- 標準正規分布N (1, 0)



- 例：平成27年国民健康・栄養調査より26-29歳の男性の身長は平均171.7cm、標準偏差（σ）6.0cmである。この度数分布が正規分布に従うとすると図のように示される。

CC BY NC ND

標準化の手順（イメージ）



CC BY NC ND

標準化の公式

- すべてのデータを以下の公式を使って変換

$$Z = \frac{X - \mu}{\sigma}$$

①平均をひいて

②標準偏差で割る

① それぞれのデータから平均を引く
② それを標準偏差で割る

- Zスコアはある値xiが分布の中でどの辺りに位置するかを平均値0、標準偏差1の標準正規分布に置き換え、表したもの

CC BY NC ND

なぜ標準化？

- 定期テスト（英語90点、国語80点）
  - 英語の方がよくできたのか？？
- 平均点が異なるため一概に比較できない
- 標準化を行い（いわゆる）偏差値を出す
- 偏差値＝平均値50、標準偏差10に標準化

$$偏差値T = \frac{X - \mu}{\sigma} \times 10 + 50$$
$$= z \times 10 + 50$$

CC BY NC ND

偏差値を計算してみよう！

科目	平均値	標準偏差	得点	偏差値
英語	60	12	70	
国語	65	20	65	
数学	50	18	80	
理科	55	10	80	
社会	52	15	50	
5教科	282	75	345	

※架空のデータです

CC BY NC ND

偏差値を計算してみよう！

科目	平均値	標準偏差	得点	偏差値
英語	60	12	70	58.3
国語	65	20	65	50.0
数学	50	18	80	66.7
理科	55	10	80	75.0
社会	52	15	50	48.7
5教科	282	75	345	58.4

※架空のデータです