統計学

講義資料その1
1変数データの記述と要約

兵庫県立大学 社会情報科学部

山本 岳洋

t.yamamoto@sis.u-hyogo.ac.jp

2021年度前期・火曜3限　神戸商科キャンパス 全学共通科目

この資料の内容

- **変数の種類**
  - ここは 数学I にはない新しいトピック
- **平均，中央値，最頻値・外れ値**
- **度数分布とヒストグラム**
- **四分位数・箱ひげ図**
- **分散・標準偏差**

この資料の確認ポイント

- **定義もわかるし計算もできる**
  - 平均，中央値，最頻値，四分位数，分散，標準偏差
- **変数の種類が分かる**
  - 名義尺度，順序尺度，間隔尺度，比例尺度
- **外れ値に対する頑健性の概念が分かる**
  - 平均と中央値はどちらが頑健性がある？
  - 範囲と四分位範囲は？

この資料の確認ポイント

- **標準化されたデータ $z_1, z_2, \ldots, z_n$ は平均 $\bar{z} = 0$, 標準偏差 $S_z = 1$ となることが導出できる**
  - 同様に，偏差値に変換されたデータの平均が 50 になることも導出できる

この資料の内容

- **教科書「確率統計」**
  - 2章
- **参考書: 小波先生「統計学入門」**
  - 1章
- **参考書 東大出版会「統計学入門」**
  - 2章1節 – 2章2節　くらいまで

本講義「統計学」で扱う内容の概観

（5月12日の講義で改めて説明します）

記述統計

データの整理

- 1変数データ
- 2変数データ

推測統計

母集団と標本

母集団の推定

仮説検定

確率の基礎

集合・確率

確率分布

記述統計学

- **収集したデータを正しく，効率的に把握するための手法**
  - 例: 受講生の学部の内訳は？
  - 例: 受講生の通勤時間の分布は？

	所属学部	PC保有状況	通勤時間
学生1	社会情報科学部	有り	40分
学生2	国際商経学部	有り	60分
学生3	社会情報科学部	無し	120分
学生4	社会情報科学部	有り	30分

1変数データの記述と要約

- **この資料では変数が１種類のときを扱う**
  - 例: 受講生の**通勤時間**の分布を知りたい
    - 平均や分散，箱ひげ図など
- **次回の資料は変数が2つある場合を扱う**
  - 例: 受講生の**数学**と**国語**の得点の関係は？
    - 散布図，相関，回帰など

変数の種類

変数の種類

- **変数: いろいろな値をとりうるもの**
  - 所属学部，　PC保有状況，通勤時間　など
  - **変量** ということもある
- **変数には取り得る値が数値であるもの（量的変数）や記号であるもの（質的変数）がある**

	所属学部	PC保有状況	通勤時間
学生1	社会情報科学部	有	40分
学生2	国際商経学部	有	60分
学生3	社会情報科学部	無	120分
学生4	社会情報科学部	有	30分

質的変数と量的変数

- **所属学部**
  - 取り得る値: 社会情報科学部，国際商経学部など
  - 取り得る値が記号なので，所属学部は質的変数
- **PC保有状況**
  - 取り得る値: 有，無
  - 取り得る値が記号なので，PC保有状況は質的変数
- **通勤時間**
  - 取り得る値: 40分，60分，…
  - 取り得る値が数値なので，通勤時間は量的変数

	所属学部	PC保有状況	通勤時間
学生1	社会情報科学部	有	40分
学生2	国際商経学部	有	60分
学生3	社会情報科学部	無	120分
学生4	社会情報科学部	有	30分

変数の種類

- **質的変数と量的変数はさらに2種類に分類される**
- **質的変数**
  - 名義尺度
  - 順序尺度
- **量的変数**
  - 間隔尺度
  - 比例尺度

名義尺度と順序尺度

- **名義尺度: 単なるラベル．順序付けられないデータ**
  - **学部**（値の例:「社会情報科学部」「国際商経学部」）
  - **性別**（値の例:「男性」「女性」）
- **順序尺度: 順序だけに意味があるデータ**
  - **システムの使いやすさ**
    （値の例:「悪い」「どちらとも言えない」「良い」）
    「悪い」＜「どちらとも言えない」＜「良い」
    という順序関係がある（が，間隔に意味はない）

- **両者とも足し算に意味がないようなデータ**
  - 「悪い」＋「良い」という演算は×

間隔尺度と比例尺度

- **間隔尺度: 数の間隔に意味があるデータ**
  - 温度（摂氏）（値の例: 「40°C」「100°C」）
  - 西暦（値の例: 「1000年」「2020年」）
- **比例尺度: 数の比にも意味があるデータ**
  - 通勤時間（値の例： 「40分」「130分」）
  - 身長（値の例: 「150cm」, 「170cm」）

- ○○倍大きい（小さい）と言えたらそれは比例尺度
- あるいは，0（原点）が本当に「0」であれば比例尺度
  - 温度（摂氏）の0°Cは便宜上そこを0°Cとしているだけ

なぜ変数の種類が重要か

- **間隔尺度と比例尺度が**
  ごっちゃになっている例（以下は誤り）
  - 40°Cのお湯は20°Cのお湯より2倍熱い
  - 西暦2000年は西暦1000年より2倍歴史がある

- **できない演算をしてしまう**
  - 「男性」＋「女性」は定義できないのに
    計算してしまう
    - 今はピンとこないと思いますが，プログラミングをするようになると気づかない内にしてしまうことも

代表値

1変数データ（量的データ）の要約

1変数データ ＝ 扱うデータの種類が1種類のデータ

学生50人の数学の成績

67	100	72	53	75	74	60	90	80	100
68	100	78	98	76	72	73	71	73	86
82	65	80	75	70	84	70	96	56	86
91	85	87	79	51	82	53	71	79	100
91	72	100	84	79	63	94	51	64	96

代表値

- **収集したデータの特徴的な値を知りたい**
  - 「今回のテストの**平均点**は・・・」
  - データの特徴を端的に表す値
- **代表的な代表値**
  - 平均（mean）
  - 中央値（median，メディアン）
  - 最頻値（mode，モード）
  - 四分位数

記号の説明

- **データ: $x_1, x_2, \cdots, x_n$**
  - n: データの大きさ（個数）
    - 今回の例だと $n = 50$
  - x_i : 測定した個々のデータの値
    （**観測値**ともいう）
    - $x_1 = 67$点, $x_2 = 100$点, $\cdots$, $x_n = 96$点

平均（mean）

- **いわゆる我々が知っている平均**
  - 厳密にいうと，算術平均（相加平均）
    - 平均には他にも相乗平均や幾何平均がある
- **すべてのデータの値の合計を**
  データの個数 n で割ったもの

平均

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

$$= \frac{1}{n}\sum_{i=1}^{n} x_i$$

（ $\bar{x}$ はエックスバーと読む）

平均: 具体例

50名の成績の平均 $\bar{x}$ は,

$$\bar{x} = \frac{1}{50}(67 + 100 + \cdots + 96)$$

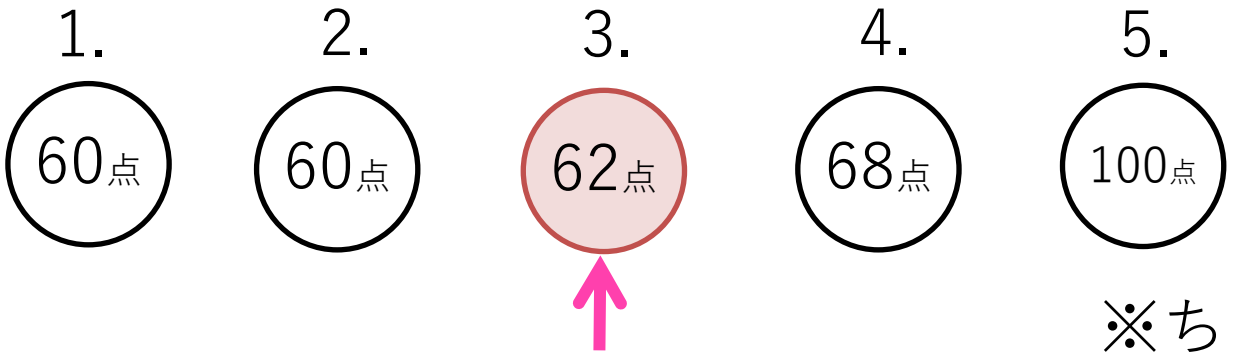
$$\fallingdotseq 78.0 \text{ 点}$$

中央値 (median，メディアン)

- **データを小さい方から順番に並べたときの，真ん中に位置するデータの値**
  - 平均とは異なる概念

- **簡単な具体例:**

学生5人の英語成績を低い順に並べたもの



※ちなみにこの例の平均は70点

中央値

中央値

データの個数が n 個のとき,

- n が奇数: 小さい方から $\dfrac{n+1}{2}$ 番目の値
- n が偶数: 小さい方から $\dfrac{n}{2}$ と $\dfrac{n}{2}+1$ 番目の値の平均

学生50人の数学の成績を
値の小さい順に並べ替えた
（ソートした）もの

51	51	53	53	56	60	63	64	65	67
68	70	70	71	71	72	72	72	73	73
74	75	75	76	78	79	79	79	80	80
82	82	84	84	85	86	86	87	90	91
91	94	96	96	98	100	100	100	100	100

中央値: 具体例

- 50名の成績の例だと，$n = 50$（偶数）なので，25番目の値（78点）と26番目の値（79点）の平均が中央値となる．
- したがって，中央値は $\frac{78+79}{2} = 78.5$ 点

最頻値（mode, モード）

- **データの中で出現回数が最も多い値**

- **50名の成績の例だと**
  - 100点をとった学生が5人と最も多いので,
    最頻値は100点

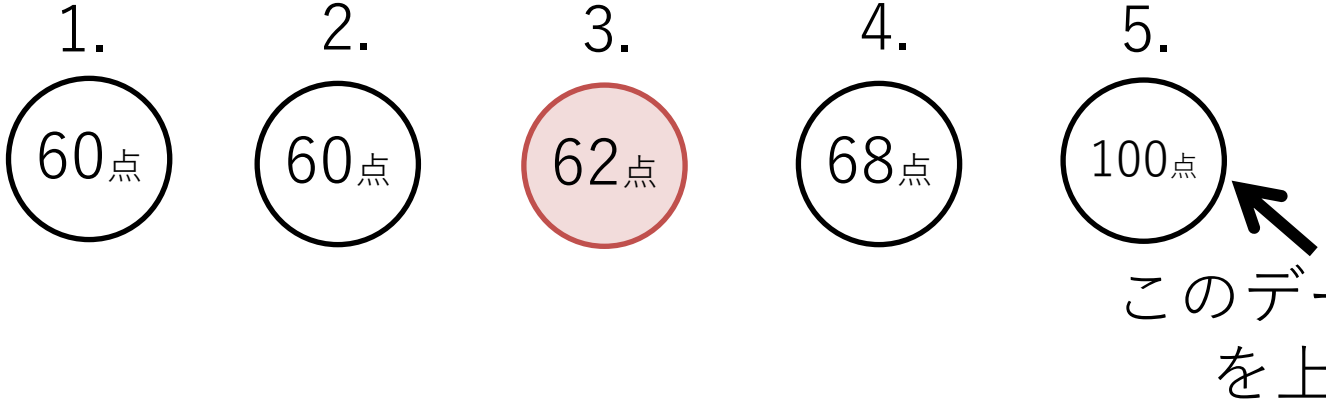
平均・中央値・最頻値の直感的理解

- **平均**
  - すべての学生の成績を合計して
    学生数で割った値
- **中央値**
  - ちょうど「真ん中」の人が取った成績
- **最頻値**
  - 最も多くの学生が取った成績

- **これら3つの値は一致するとは限らない**

外れ値

- 平均は，極端な値をとるデータに大きく影響をうける



- 他の値から極端に小さかったり大きかったりする値のことを外れ値（outlier，異常値）と言う

頑健性（ロバスト）

- **たとえば，前ページの例で100点の学生が仮に75点だったとすると**
  - 平均: 70点 → 65点 に変化
  - 中央値: 62点　→ 62点 （変わらず）

- **中央値は，少々の外れ値に対してはあまり影響を受けない．このような性質を頑健（robust，ロバスト）である，と呼ぶ**

平均と中央値が異なる例

- どのような種類のデータだと，
  平均と中央値が大きく異なるか？

- データの種類は？
- そのときの分布の形は？

度数分布表・ヒストグラム

1変数データ（量的データ）の要約

学生50人の数学の試験成績

67	100	72	53	75	74	60	90	80	100
68	100	78	98	76	72	73	71	73	86
82	65	80	75	70	84	70	96	56	86
91	85	87	79	51	82	53	71	79	100
91	72	100	84	79	63	94	51	64	96

分布を把握したい

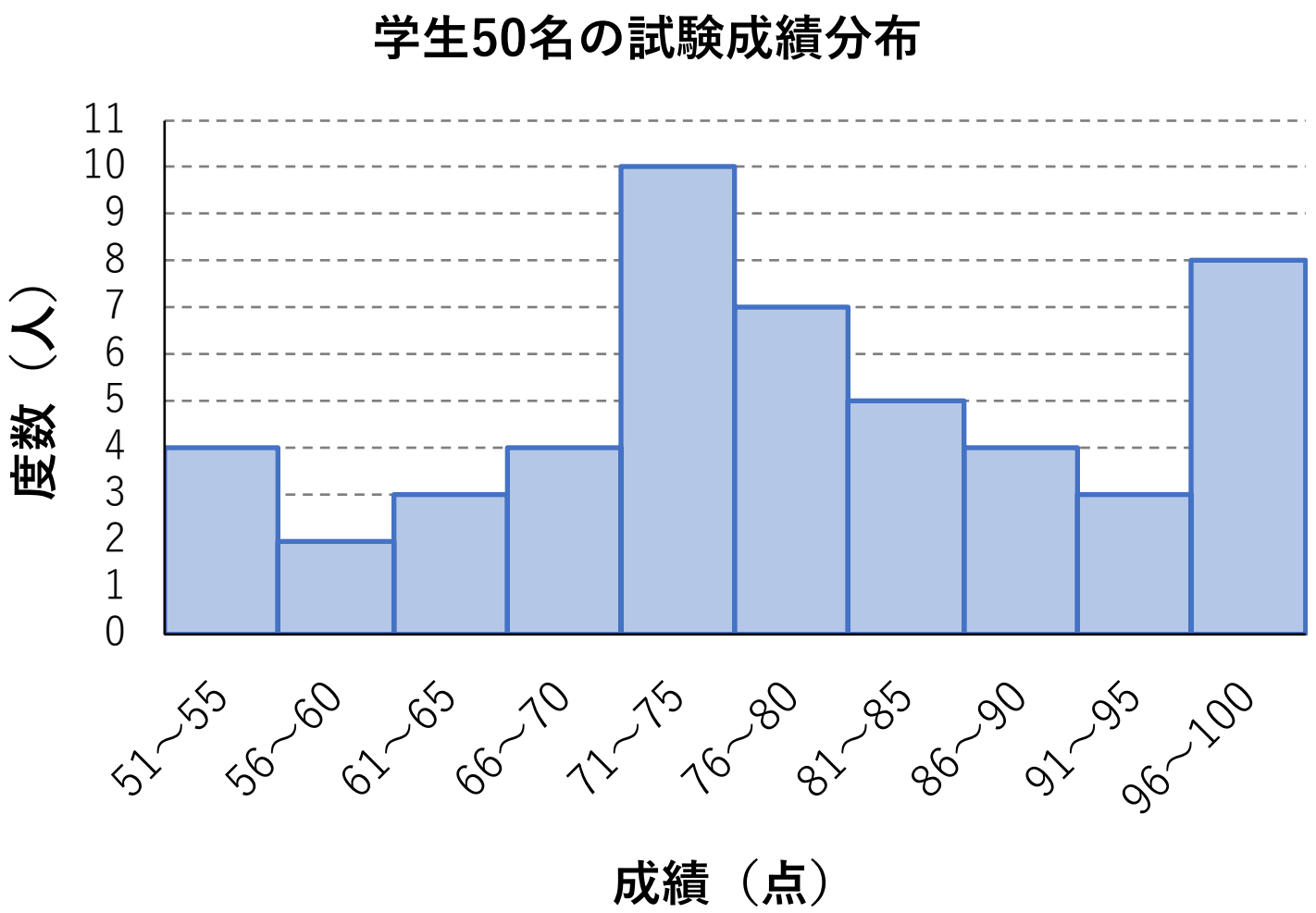
度数分布表

階級	階級値	度数
51点～55点	53点	4
56点～60点	58点	2
61点～65点	62点	3
66点～70点	68点	4
71点～75点	73点	10
76点～80点	78点	7
81点 ～85点	83点	5
86点～90点	88点	4
91点～95点	93点	3
96点～100点	98点	8
合計		**50**

- **階級**
  - データを分ける区間
- **階級値**
  - 階級の代表的な値（後述）
  - 度数分布表には記載しないこともある
- **度数（frequency）**
  - その階級に属するデータの個数

ヒストグラム

度数分布表をグラフで表したもの



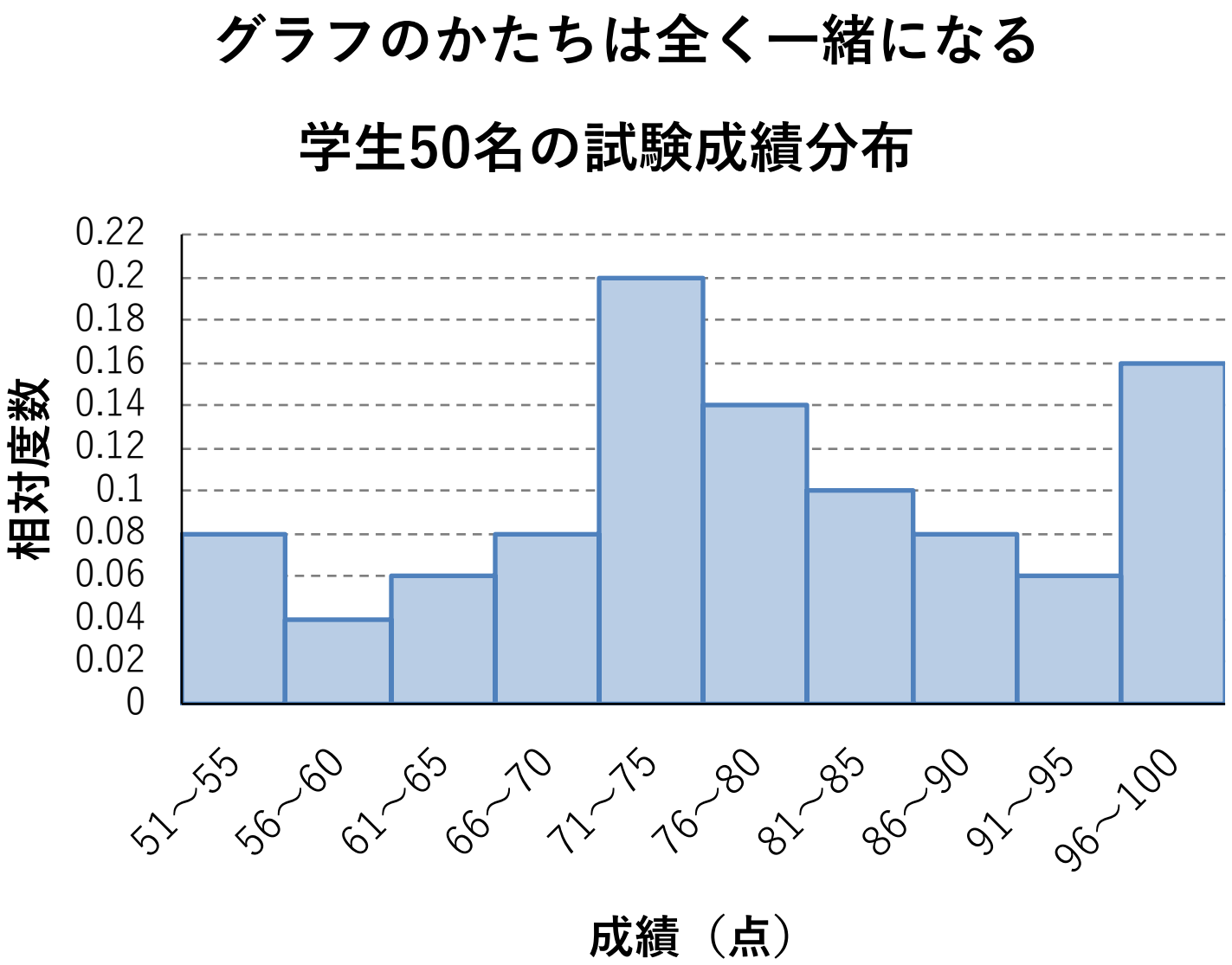
相対度数

階級	階級値	度数	相対度数
51点～55点	53点	4	0.08
56点～60点	58点	2	0.04
61点～65点	62点	3	0.06
66点～70点	68点	4	0.08
71点～75点	73点	10	0.20
76点～80点	78点	7	0.14
81点 ～85点	83点	5	0.10
86点～90点	88点	4	0.08
91点～95点	93点	3	0.06
96点～100点	98点	8	0.16
合計		**50**	**1.00**

- **相対度数**
  - 度数をデータの個数で割ったもの
  - 相対度数の合計は必ず **1** になる
- **相対度数 ＝ その階級に属するデータの割合**
  - 個数が異なる他のデータとの比較時に便利
  - A中学（50名）とB中学（300名）の数学の成績を比較

ヒストグラム（相対度数）

グラフのかたちは全く一緒になる



累積度数

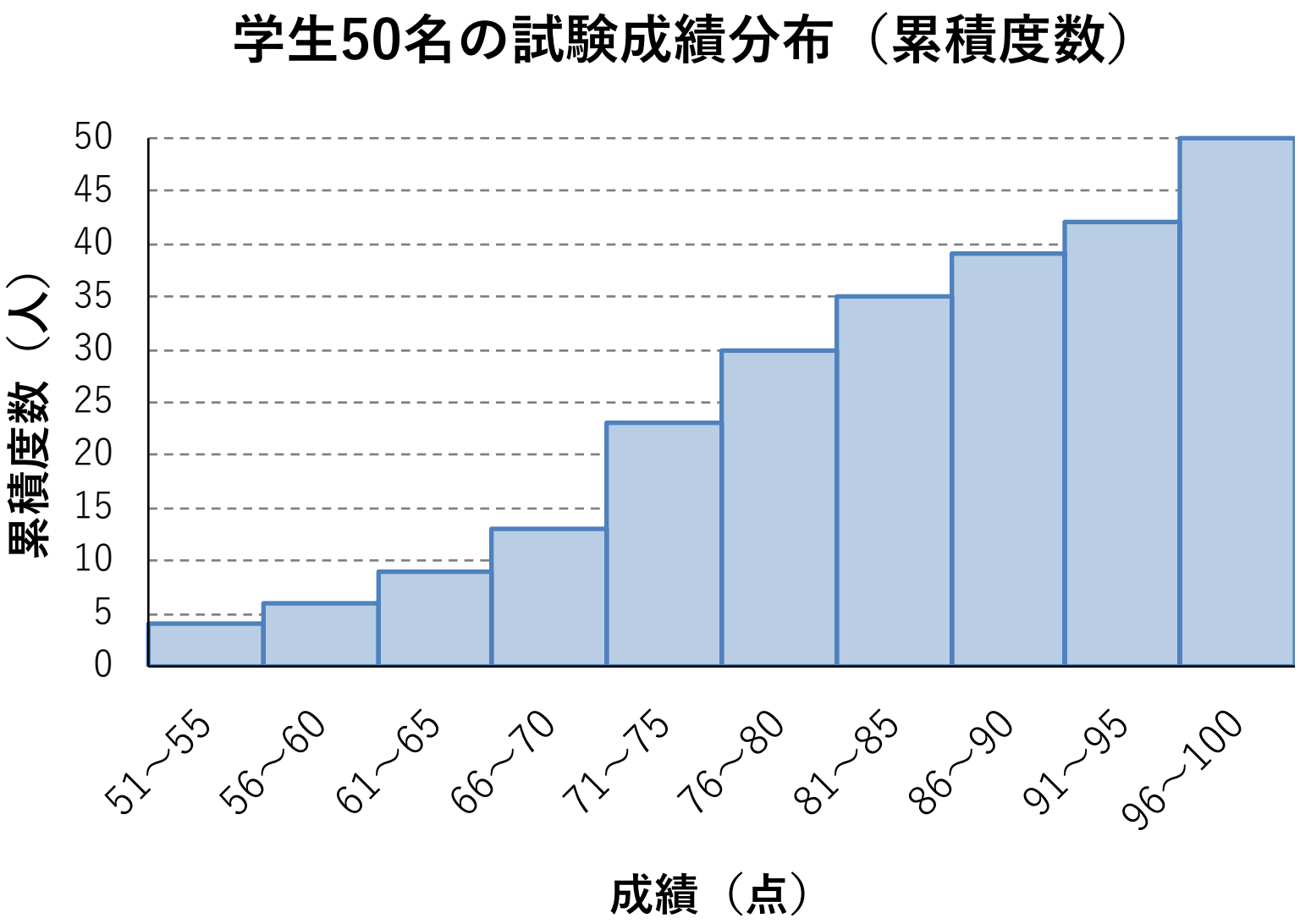
階級	階級値	度数	累積度数
51点～55点	53点	4	4
56点～60点	58点	2	6
61点～65点	62点	3	9
66点～70点	68点	4	13
71点～75点	73点	10	23
76点～80点	78点	7	30
81点 ～85点	83点	5	35
86点～90点	88点	4	39
91点～95点	93点	3	42
96点～100点	98点	8	50
合計		**50**	

- **累積度数**
  - 度数を，下の階級から順に積み上げたときの累積和

- **その階級までに属するデータの数が分かる**
  - 試験が75点以下の学生数は23人

- **同様に，相対度数に対する累積和は累積相対度数**

累積度数グラフ


学生50名の試験成績分布（累積度数）
累積度数（人）
50
45
40
35
30
25
20
15
10
5
0
51～55
56～60
61～65
66～70
71～75
76～80
81～85
86～90
91～95
96～100
成績（点）

度数分布表と平均

階級	階級値	度数
51点～55点	53点	4
56点～60点	58点	2
61点～65点	62点	3
66点～70点	68点	4
71点～75点	73点	10
76点～80点	78点	7
81点 ～85点	83点	5
86点～90点	88点	4
91点～95点	93点	3
96点～100点	98点	8
合計		**50**

- **度数分布表だけからでも，（近似としての）平均を求めることができる**

- **階級値**
  - その階級内にデータがまんべんなく分布（**一様に分布**）していると仮定したときの，データの平均
  - その階級の下限値と上限値の平均

度数分布表と平均

$$\frac{1}{50}(53 \times 4 + 58 \times 2 + \cdots + 98 \times 8)$$

$$= 77.9\text{点}$$

本当の平均（≒78.0点）と近い値が得られる

- **つまり，階級値とは**
  - 個々のデータの値は分からないので，とりあえず中間の値にしておけば平均を求めるときに誤差は少ないだろう，という考え

度数分布表と中央値

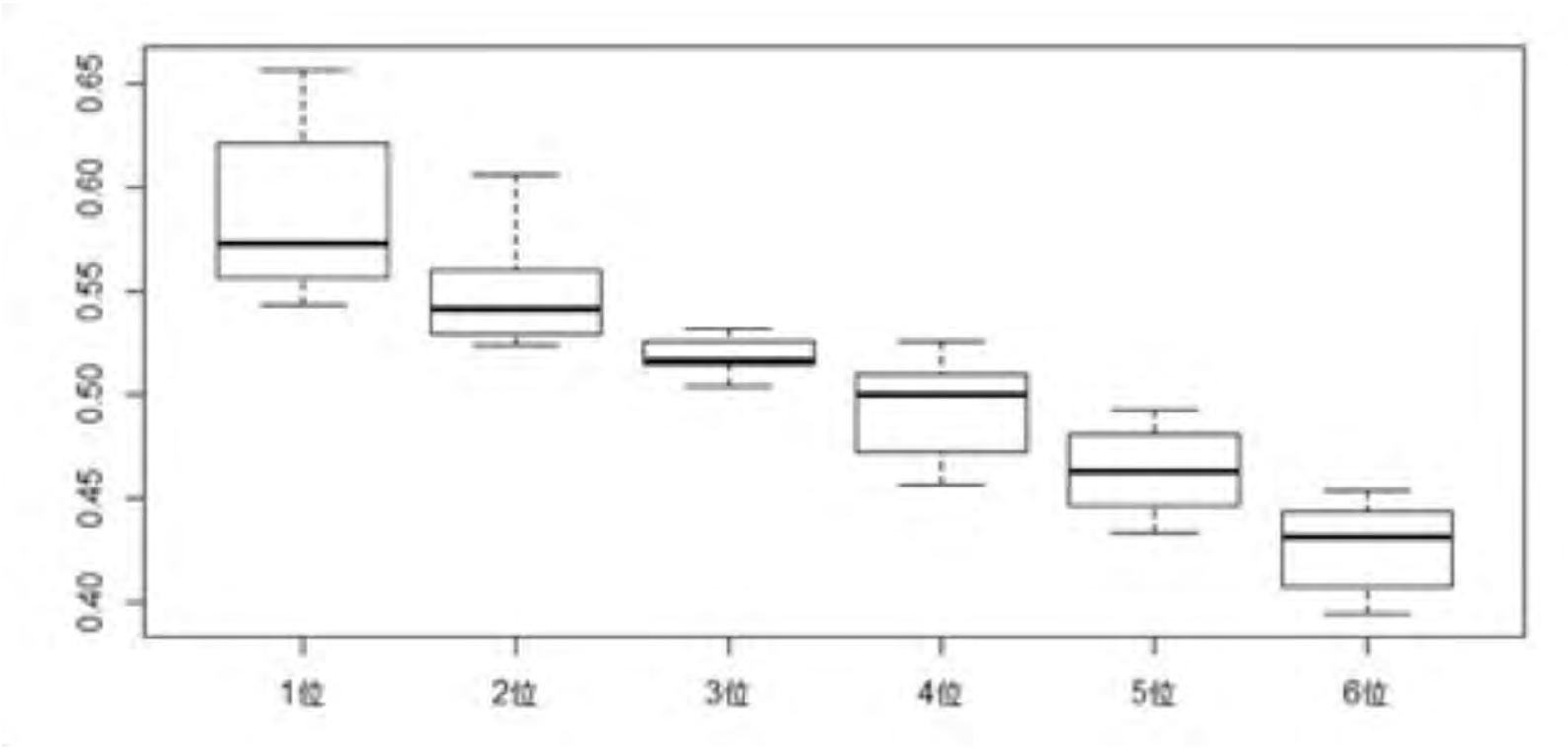
階級	階級値	度数	累積度数
51点～55点	53点	4	4
56点～60点	58点	2	6
61点～65点	62点	3	9
66点～70点	68点	4	13
71点～75点	73点	10	23
76点～80点	78点	7	30
81点 ～85点	83点	5	35
86点～90点	88点	4	39
91点～95点	93点	3	42
96点～100点	98点	8	50
合計		**50**	

- **累積度数を求めれば，中央値がどの階級に属するかが分かる**
- **中央値（24番目と25番目の平均）**
  - 度数分布表を見れば，76点 – 80点の階級にあることが分かる
    - 本当の中央値 = 78.5

四分位数・箱ひげ図

箱ひげ図の例

2006～2015年の10年間のプロ野球パ･リーグの順位毎の勝率



ferret, 箱ひげ図をマスターしよう！誰が見ても一瞬で伝わるレポート資料の作り方 より引用, https://ferret-plus.com/8234

四分位数（quartile）

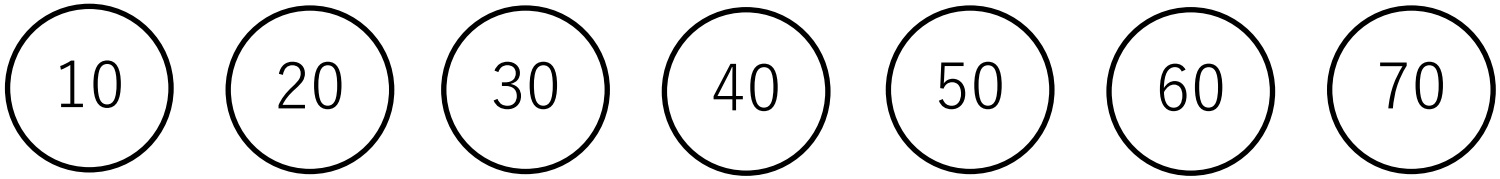
- **データを小さい順に並べたときに，**
  データを4等分したときの各区切りの値

- **小さい方から順に:**
  - **第1四分位数**（Q_1）：25%に位置する値
  - **第2四分位数**（Q_2）：50%に位置する値
    - つまり，中央値
  - **第3四分位数**（Q_3）：75%に位置する値

四分位数の求め方
（いろいろ定義あり）

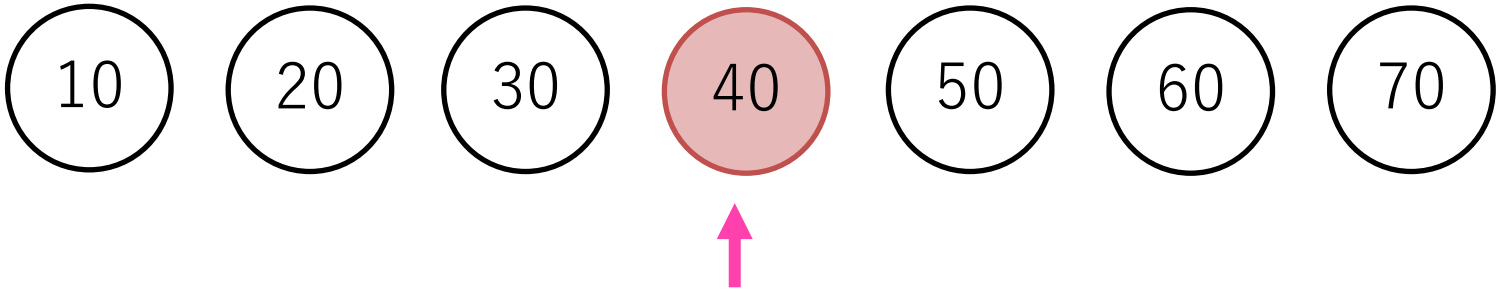
1. **第2四分位数 Q_2（中央値）を求める**
2. **データを中央値より小さなグループ，大きなグループの２種類に分割する**
   - n が奇数のとき，各グループに中央値を含める方法と含めない方法の2通りがある
   - 本資料は含める方法を採用
3. **小さなグループ内，大きなグループ内でそれぞれ中央値を求める**
   - 小さいグループの中央値を 第1四分位数 Q_1
     大きいグループの中央値を 第3四分位数 Q_3 とする

具体例： 四分位数の求め方



具体例： 四分位数の求め方

1. 第2四分位数（中央値）を求める

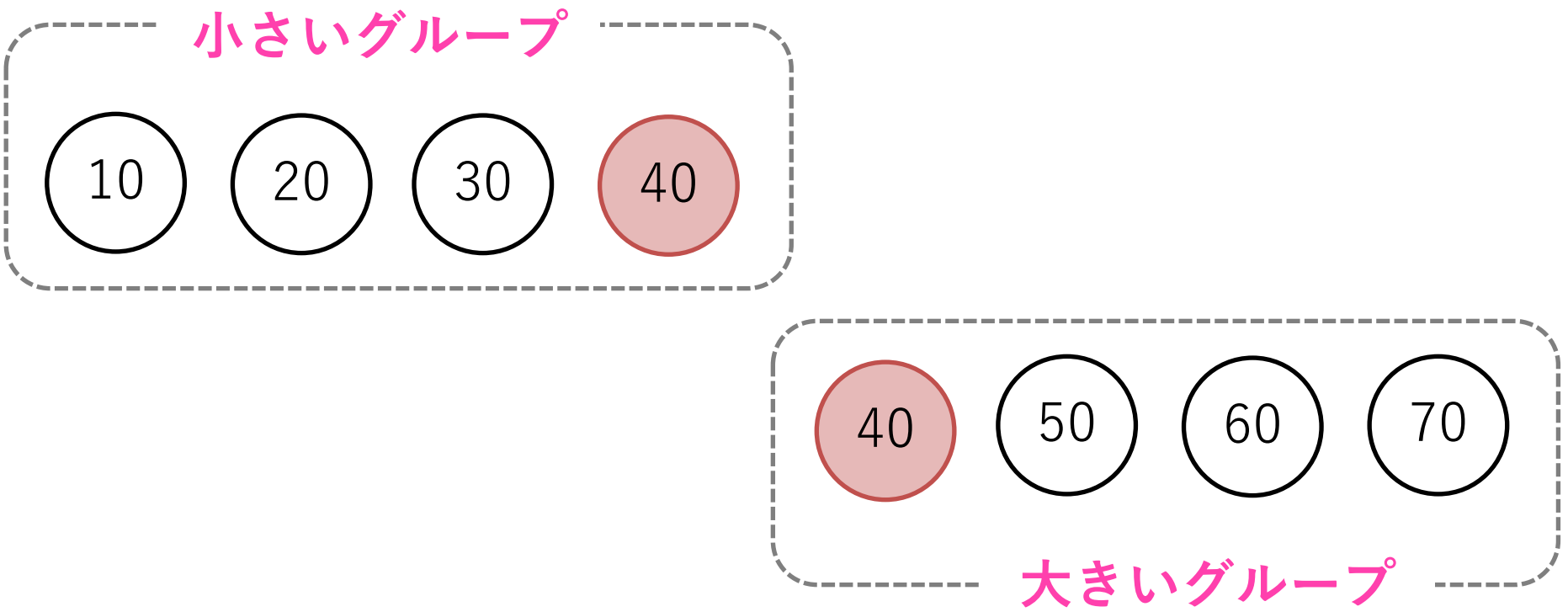


中央値 = 40

（第2四分位数 Q_2）

具体例： 四分位数の求め方

2. データを中央値より小さいグループ，
   大きいグループの２種類に分割する



具体例： 四分位数の求め方

3. 小さいグループ，大きいグループ
   それぞれで中央値を求める

小さいグループ

10　20　30　40

中央値 = 25

（第1四分位数 Q_1）

（第3四分位数 Q_3）

中央値 = 55

40　50　60　70

大きいグループ

参考: パーセンタイル数 （percentile）

- データを**100等分**したときの，小さい方から $\boldsymbol{p}$ **%**のところにある値を
  p**パーセンタイル数**という

- 25パーセンタイル数 $= Q_1$
- 50パーセンタイル数 $= Q_2$
- 75パーセンタイル数 $= Q_3$

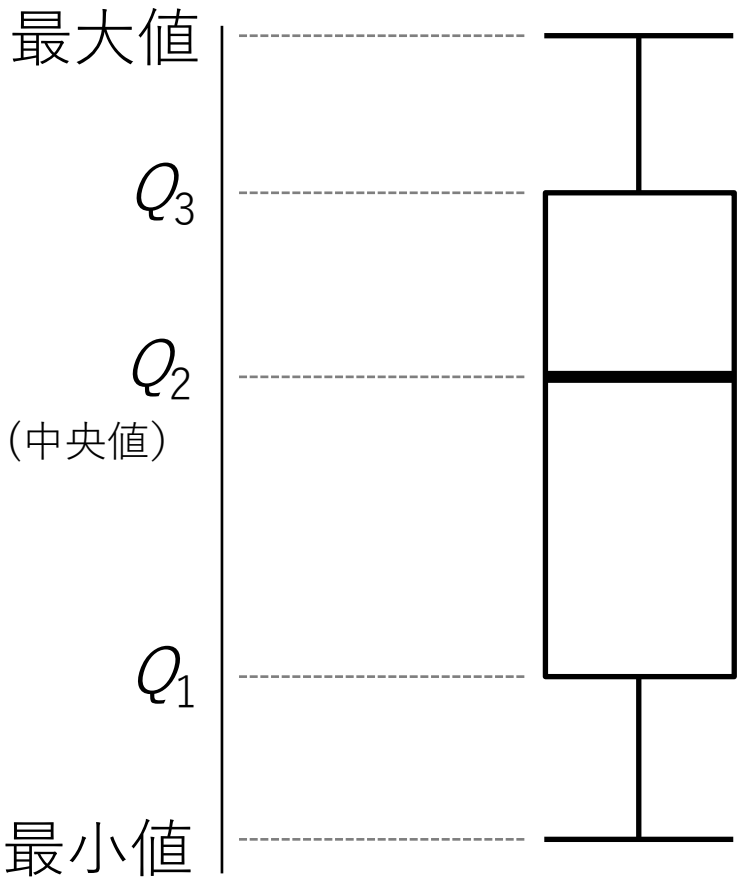
五数要約

- データの最小値
- 第1四分位数（Q_1）
- 第2四分位数（Q_2）
- 第3四分位数（Q_3）
- データの最大値

をまとめて五数要約と呼ぶ

箱ひげ図（box plot）

五数要約をグラフで表したもの



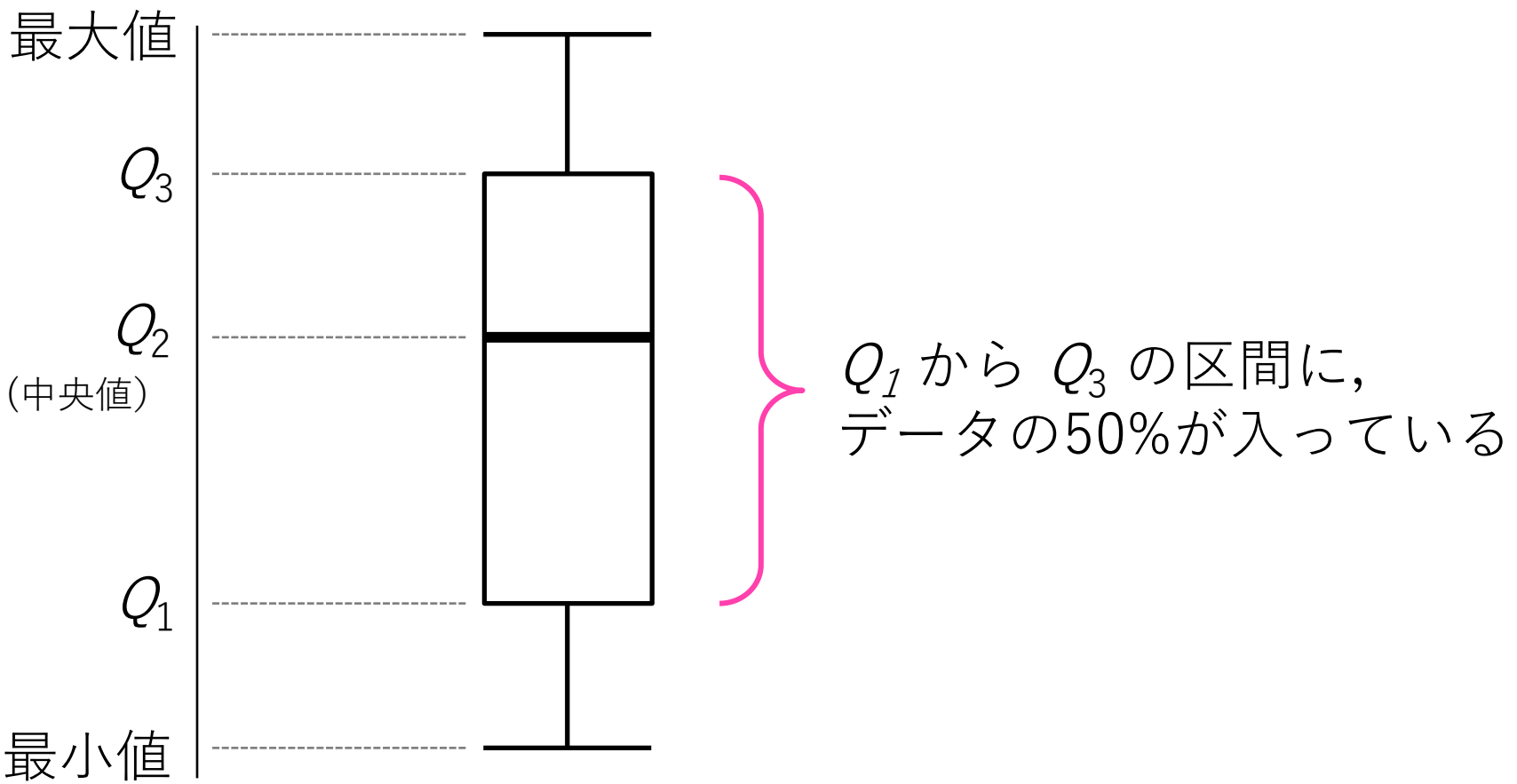
- **データの取りうる範囲や分布の形をある程度把握できる**
  - 平均だけを記載した棒グラフよりも多くのことがわかる

※外れ値の扱い方や，平均をグラフ中に記入するかどうかなど，箱ひげ図の作成方法はいろいろとやり方がある

この資料では最も単純な，外れ値を考慮しない方法について説明

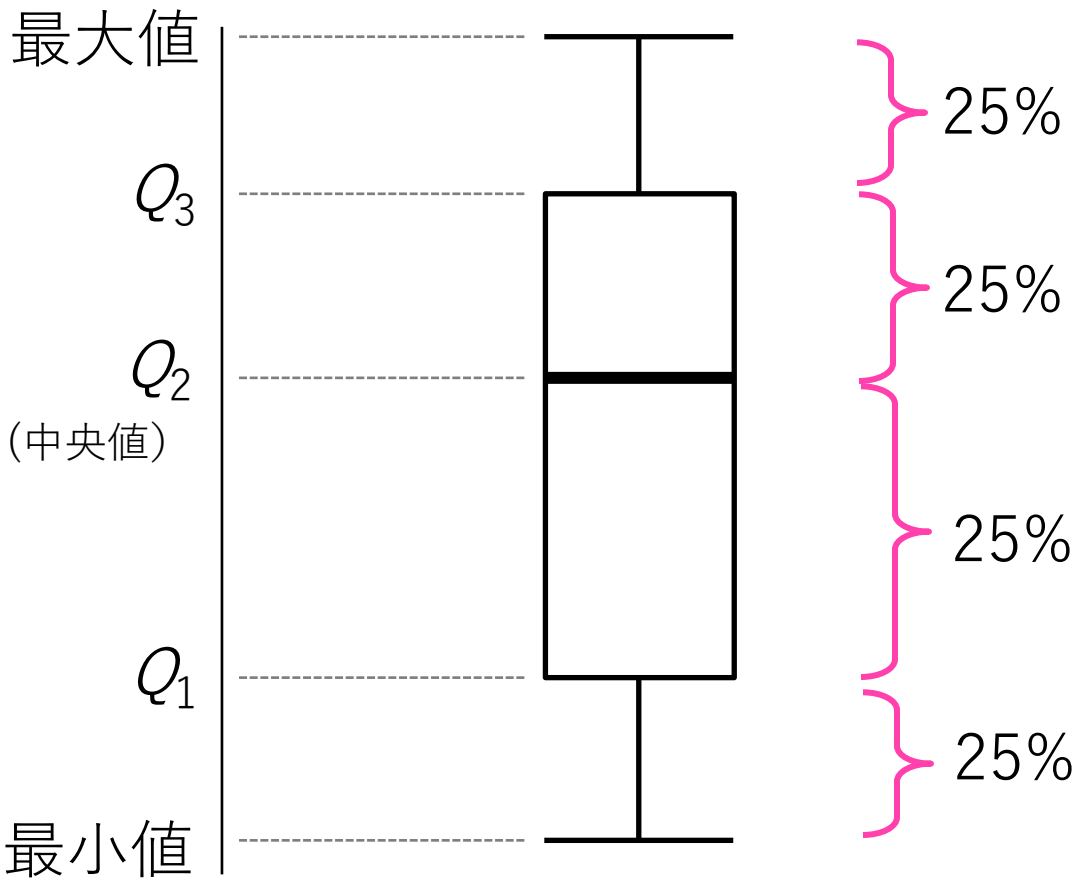
箱ひげ図（box plot）

五数要約をグラフで表したもの

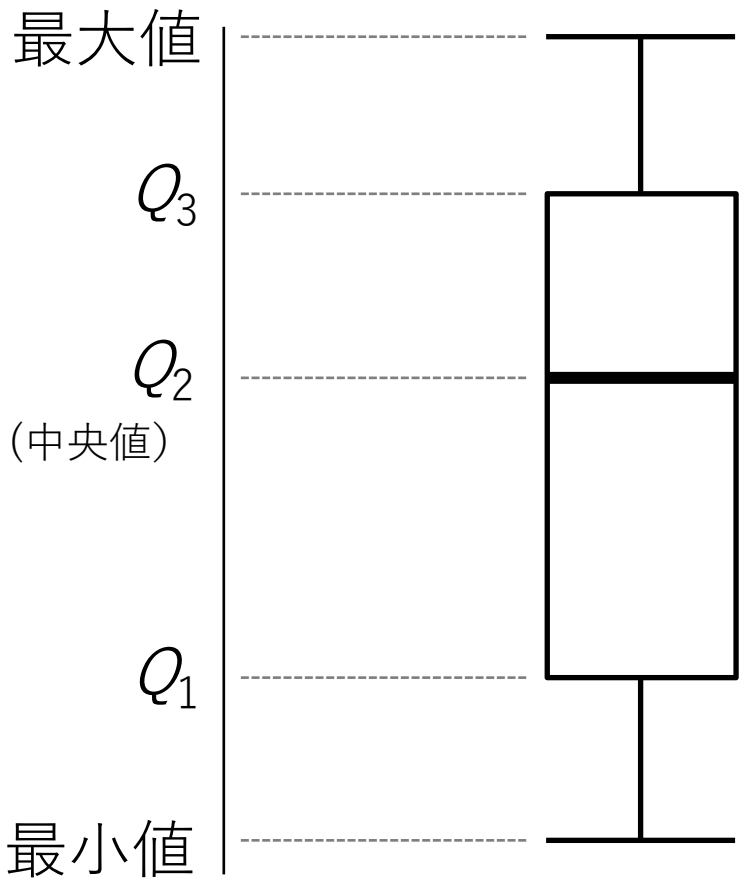


箱ひげ図（box plot）

五数要約をグラフで表したもの

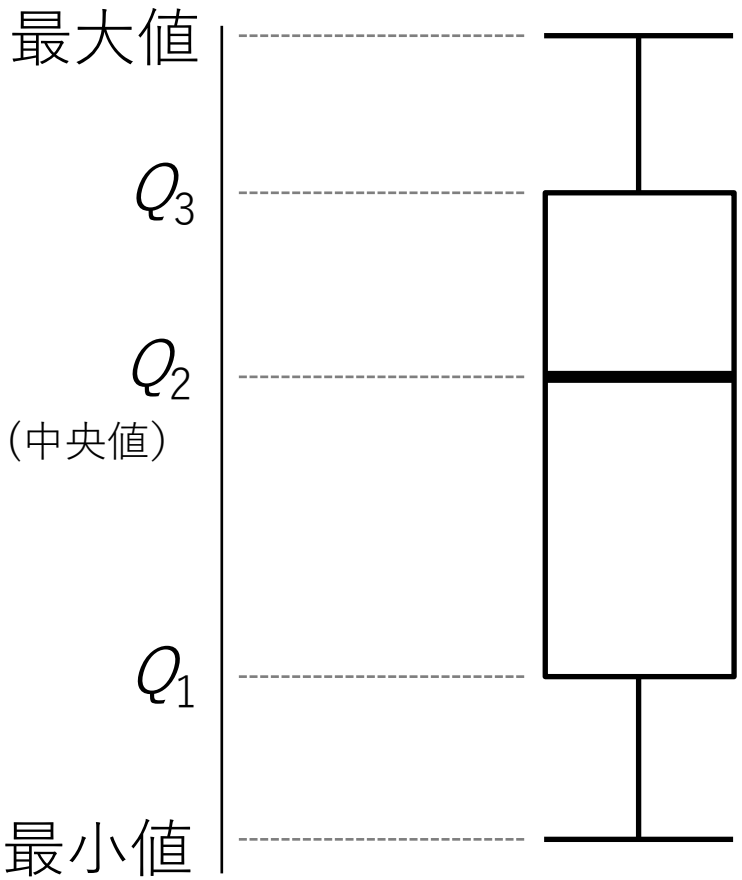


範囲と四分位範囲



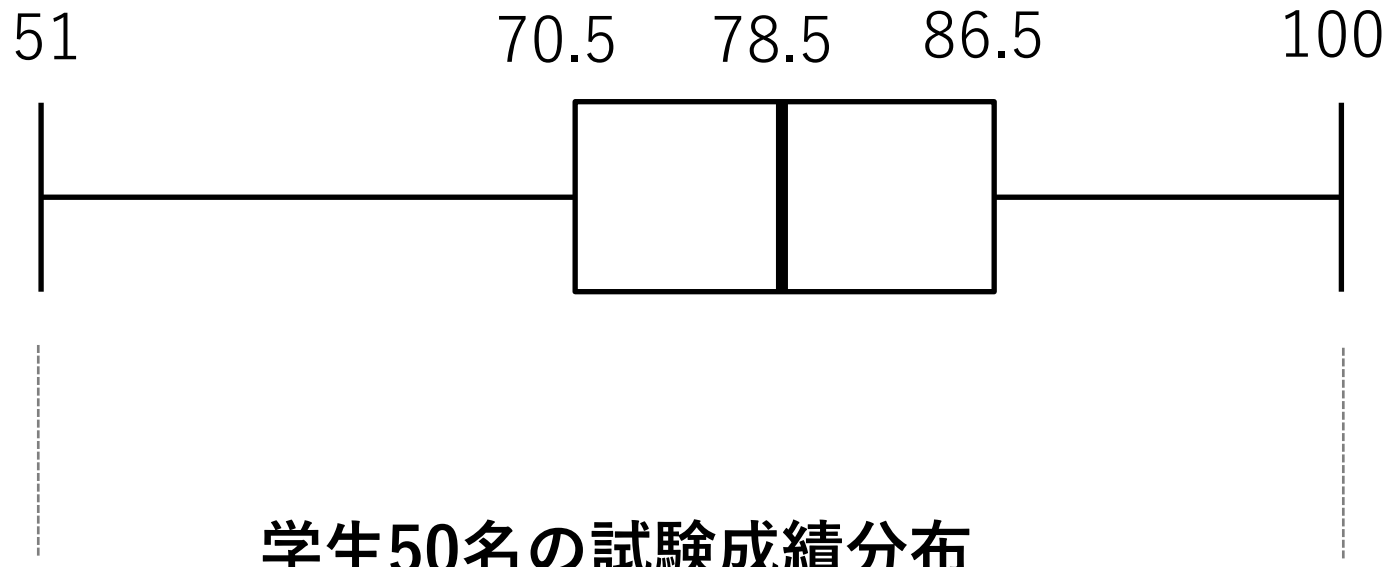
- **範囲（range, レンジ）**
  - 最大値 – 最小値

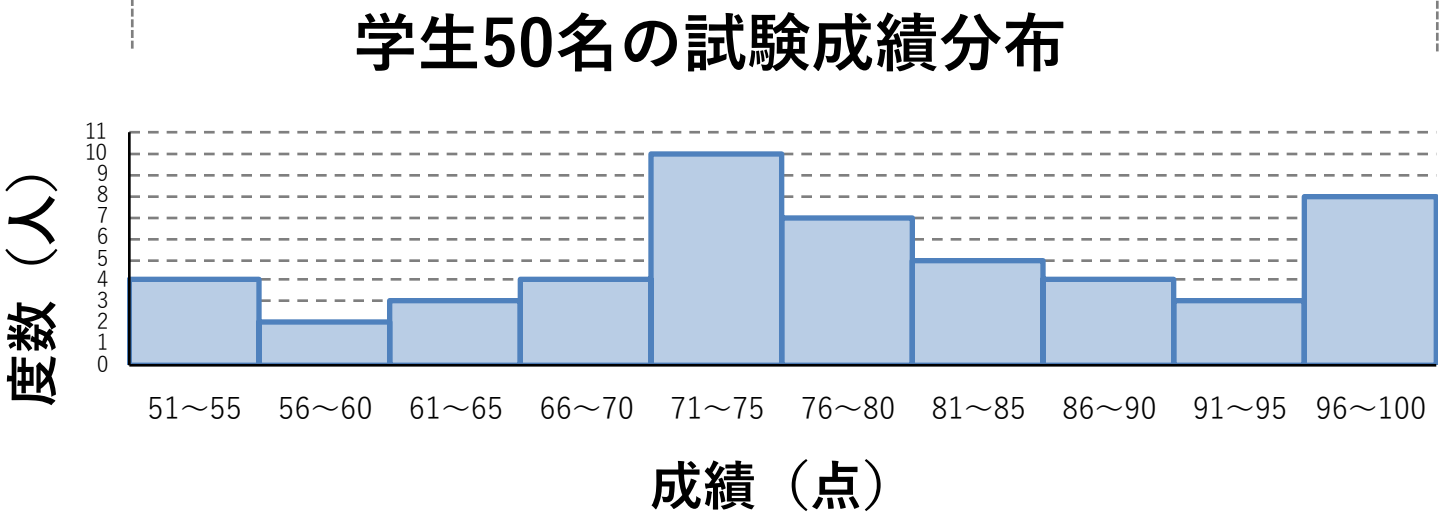
範囲と四分位範囲



- **四分位範囲**
  - $Q_3 - Q_1$
  - 範囲よりも**ロバスト**

箱ひげ図の具体例


51
70.5
78.5
86.5
100


学生50名の試験成績分布
度数（人）
0
1
2
3
4
5
6
7
8
9
10
11
51～55
56～60
61～65
66～70
71～75
76～80
81～85
86～90
91～95
96～100
成績（点）

分散・標準偏差

データのばらつきの表現

- **平均値や中央値だけでは**
  データの分布が分からない
  - 平均値や中央値が同じだとしても
    分布が一致するとは限らない

分散（variance）

分散

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

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

- 分散 S^2 が大きい → データが散らばっている
  分散 S^2 が小さい → データが平均値付近に集まっている
- 分散の式に出現している $x_i - \bar{x}$ を偏差と呼ぶ
  - 偏差: 平均からの差

分散の計算

- 学生50人の数学の成績を例に取ると，
  分散 S^2 は

$$S^2 = \frac{1}{50}\{(67-78.0)^2 + (100-78.0)^2 + \cdots + (96-78.0)^2\}$$

$$\fallingdotseq 186.6 \text{ 点}^2$$

分散の別表現
（こちらの方がよく使います）

分散

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

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

$$= \overline{x^2} - \bar{x}^2 \quad \text{ただし，} \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

分散 ＝ 二乗の平均 － 平均の二乗

標準偏差（standard deviation, SD）

- 分散 S^2 の正の平方根を**標準偏差**と呼ぶ

標準偏差

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

- 学生50名の試験成績の標準偏差 S は

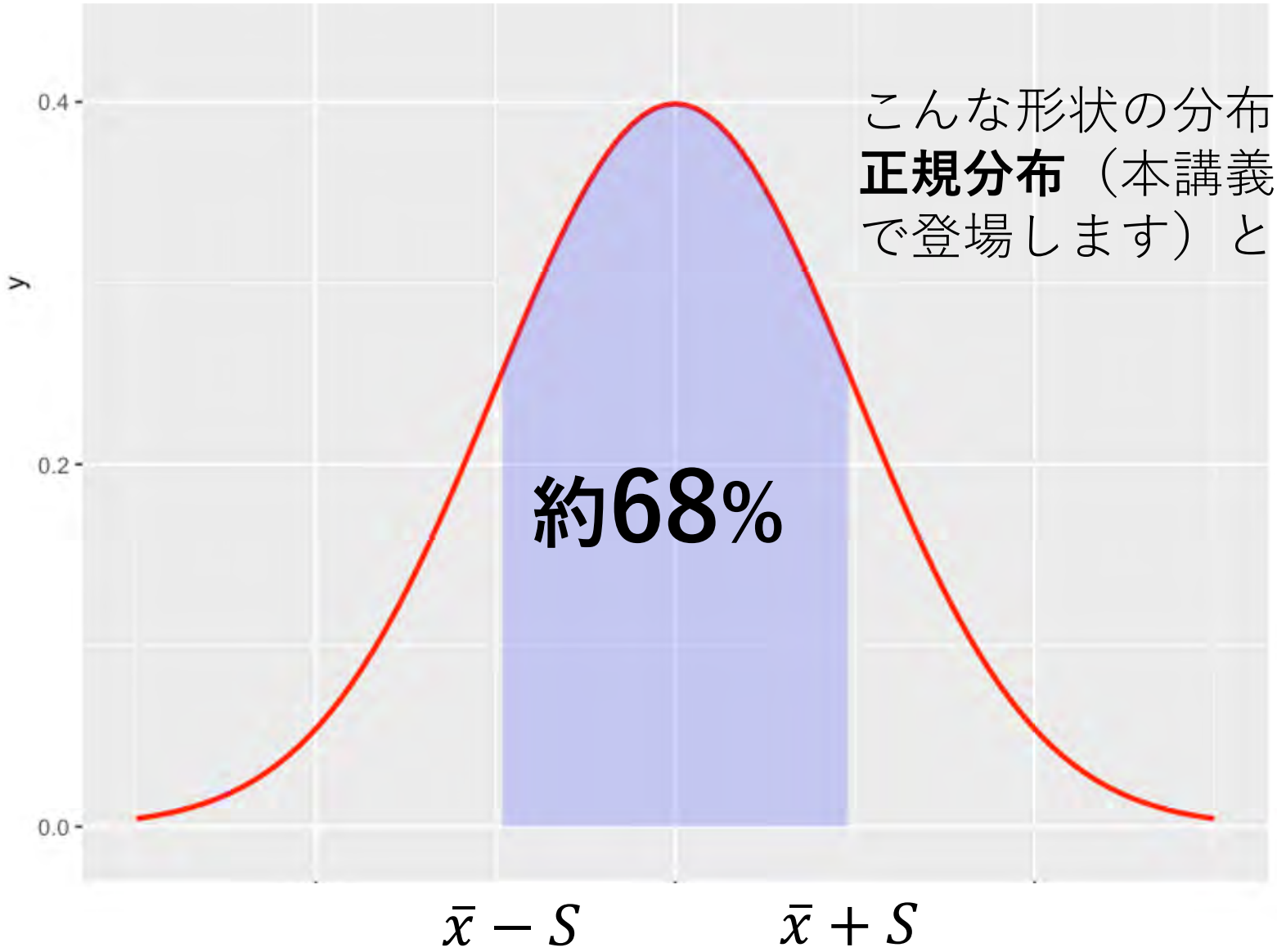
$$S = \sqrt{S^2} \fallingdotseq 13.7$$

– ちなみに，このときの単位は **点**

参考: 標準偏差の意味

- **データが（本講義後半で扱う）正規分布に従っているとすると，全データの:**
  - 約68%は $\bar{x} - S$ から $\bar{x} + S$ の区間に入る
  - 約95%は $\bar{x} - 2S$ から $\bar{x} + 2S$ の区間に入る

参考: 標準偏差の意味



こんな形状の分布を**正規分布**（本講義後半で登場します）という

標準化

- データ $x_1, x_2, \ldots, x_n$ に対して，
  以下の変換を行うことを考える

$$z_i = \frac{x_i - \bar{x}}{S}$$

ただし，S は
$x_1, x_2, \ldots, x_n$の標準偏差

- このような変換を**標準化**と呼び，
  z_iを**標準得点**と呼ぶ

標準化

- **標準化されたデータ $z_1, z_2, \ldots, z_n$ は平均 $\bar{z} = 0$, 標準偏差 $S_z = 1$ となる**

- **異なる種類のデータを比較する際などに使用**
  - 例：国語のテストにおける75点と
    数学のテストにおける75点のどちらが
    「良い」成績か

偏差値

- **標準化したデータにさらに以下の変換を行ったものを偏差値という**

$$T_i = 10z_i + 50$$

- 変換後のデータ $T_1, T_2, \ldots, T_n$ **は平均** $\bar{T} = 50$**, 標準偏差** $S_T = 10$ **となる**
  - （もしデータが正規分布に従っているとすると）偏差値30～70の間に95%の学生が入る

参考: 基本統計量

- **データ代表値やばらつきの尺度など，データの統計的性質を表す値を基本統計量とよぶこともある**
- **基本統計量の例**
  - 平均，中央値，最頻値，四分位数
  - 最大値，最小値
  - 分散，標準偏差
  - 他にも，歪度（わいど），尖度（せんど）など

この資料のまとめ

- 変数の種類
- 平均，中央値，最頻値・外れ値
- 度数分布とヒストグラム
- 四分位数・箱ひげ図
- 分散・標準偏差