統計学

第2回
1変数データの記述と要約1

兵庫県立大学 社会情報科学部

山本 岳洋

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

平成31年度前期・火曜2限　神戸商科キャンパス 全学共通科目

この資料の内容

- **1変数データの記述と要約**
  - **データとは**
  - **種々の代表値**
    - 平均，中央値，最頻値
  - **度数分布とヒストグラム**

この資料の内容

- **教科書「確率統計」**
  - 2章1節
- **参考書「統計学入門」**
  - 2章1節 – 2章2節　くらいまで

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？

B	C	O
学部について教えてくだ	ノートPCをすでに持って	大学までの通学時間（分）
社会情報科学部	すでに持っている	20
社会情報科学部	学部でBYODが指定されて	60
社会情報科学部	すでに持っている	100
社会情報科学部	すでに持っている	100
社会情報科学部	学部でBYODが指定されて	90
社会情報科学部	すでに持っている	20
社会情報科学部	すでに持っている	80
社会情報科学部	すでに持っている	20
社会情報科学部	学部でBYODが指定されて	30
社会情報科学部	学部でBYODが指定されて	30
社会情報科学部	すでに持っている	60
社会情報科学部	すでに持っている	60
社会情報科学部	すでに持っている	90
社会情報科学部	すでに持っている	120
社会情報科学部	すでに持っている	20

データの種類

- **データは大きく2種類に分類される**

- **質的データ**
  - 名義尺度
  - 順序尺度

- **量的データ**
  - 間隔尺度
  - 比例尺度

質的データ

- **名義尺度: 単なるラベル．順序付けられないデータ**
  - 学部:「社会情報科学部」「国際商経学部」
  - 性別:「男性」「女性」
- **順序尺度: 順序だけに意味があるデータ**
  - 「知らない」＜「知っている」＜「とても知っている」
  - 「悪い」＜「どちらとも言えない」＜「良い」

- **両者とも足し算に意味がないようなデータ**
  - 「知らない」＋「知っている」という演算は×

量的データ

- **間隔尺度: 数の間隔に意味があるデータ**
  - 温度（摂氏）：「40°C」「100°C」
  - 時刻: 「午後1時15分」「午後2時30分」
- **比例尺度: 数の比にも意味があるデータ**
  - 通勤時間 「40分」「130分」
  - 身長，体重など

- **○○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**

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）
- **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,

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

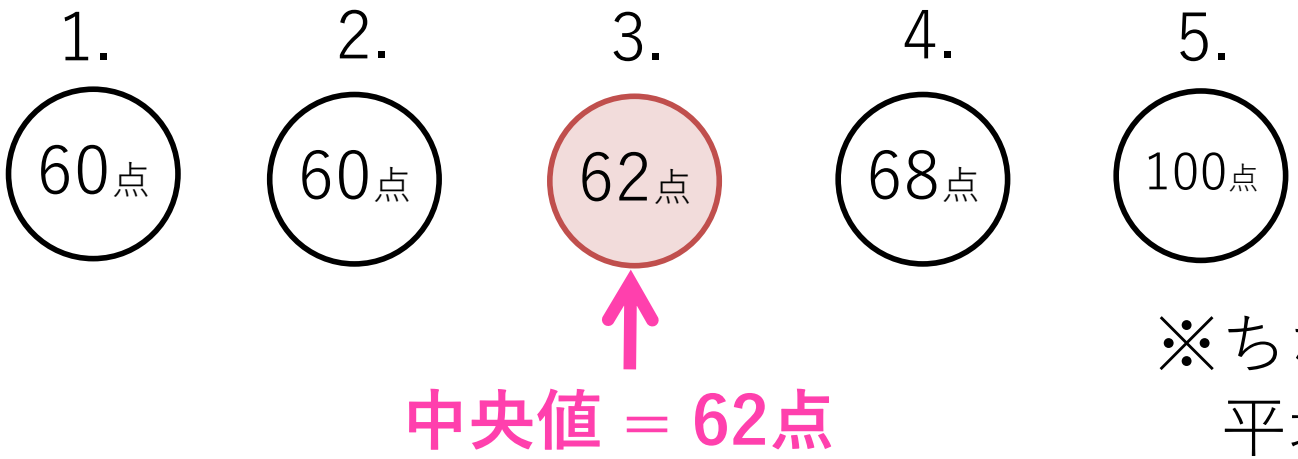
$$\fallingdotseq 78.0 \text{ 点}$$

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

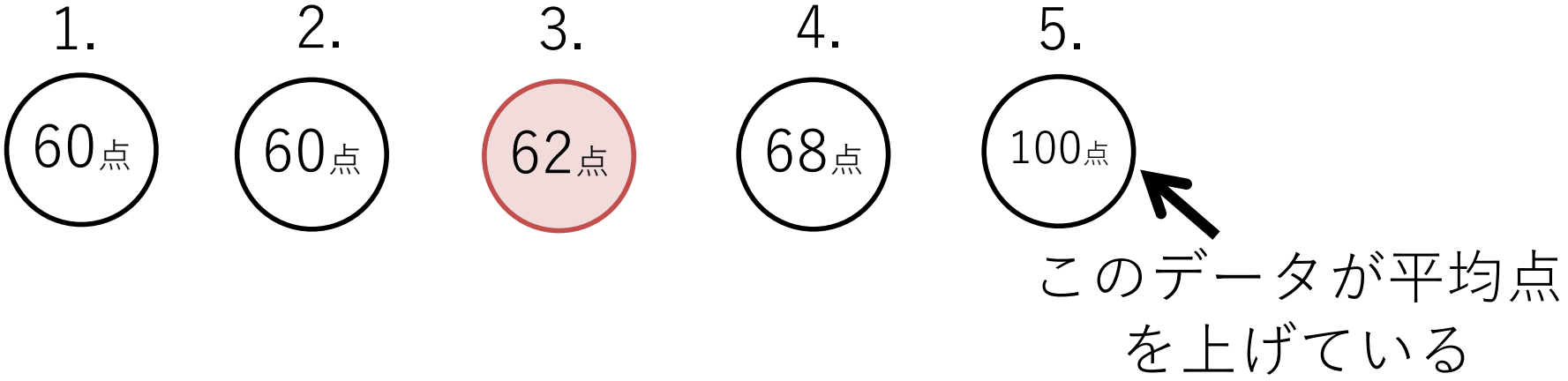
平均・中央値・最頻値の直感的理解

- **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**

平均と中央値

- 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

平均と中央値

- **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？

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

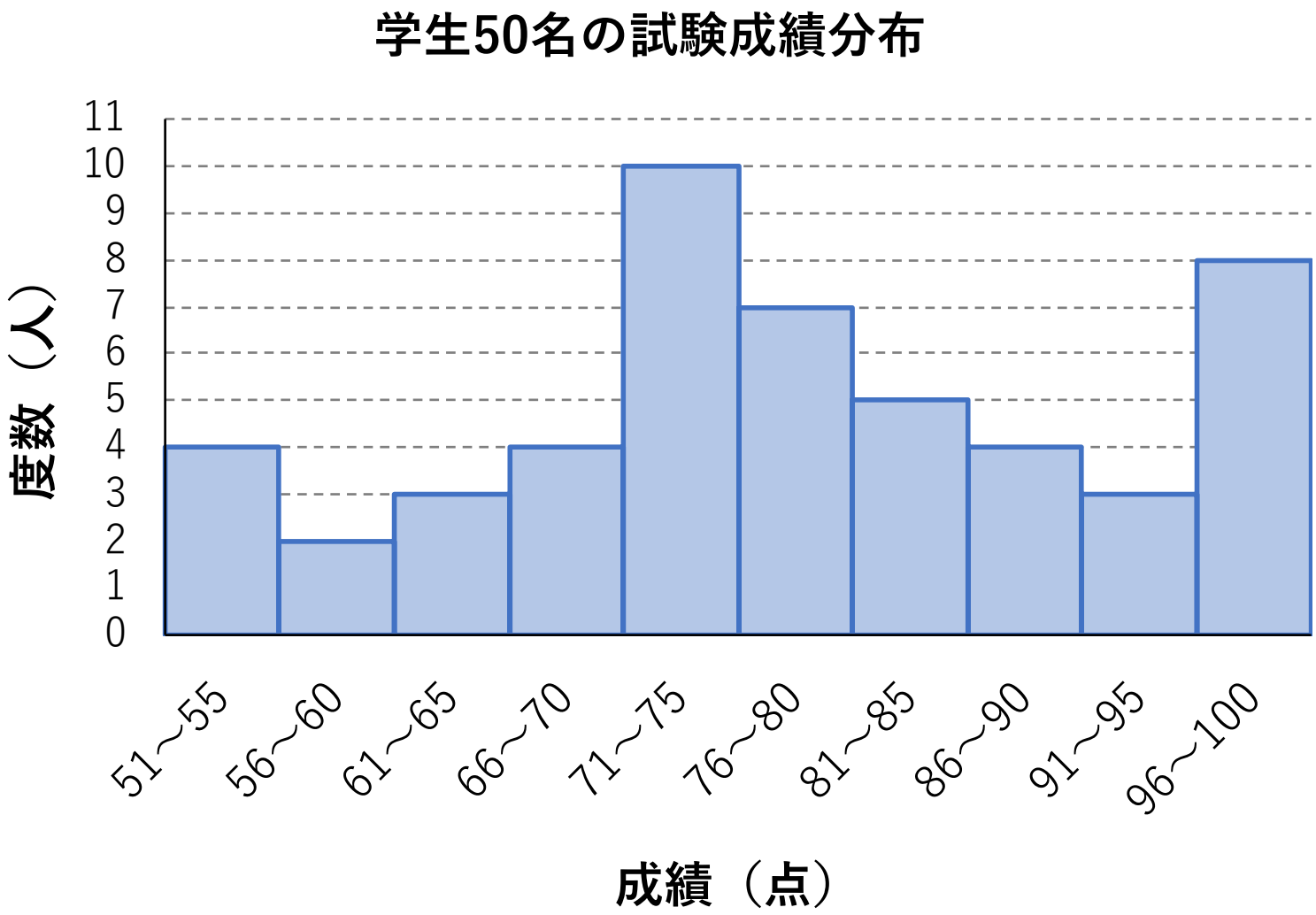
度数分布表

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



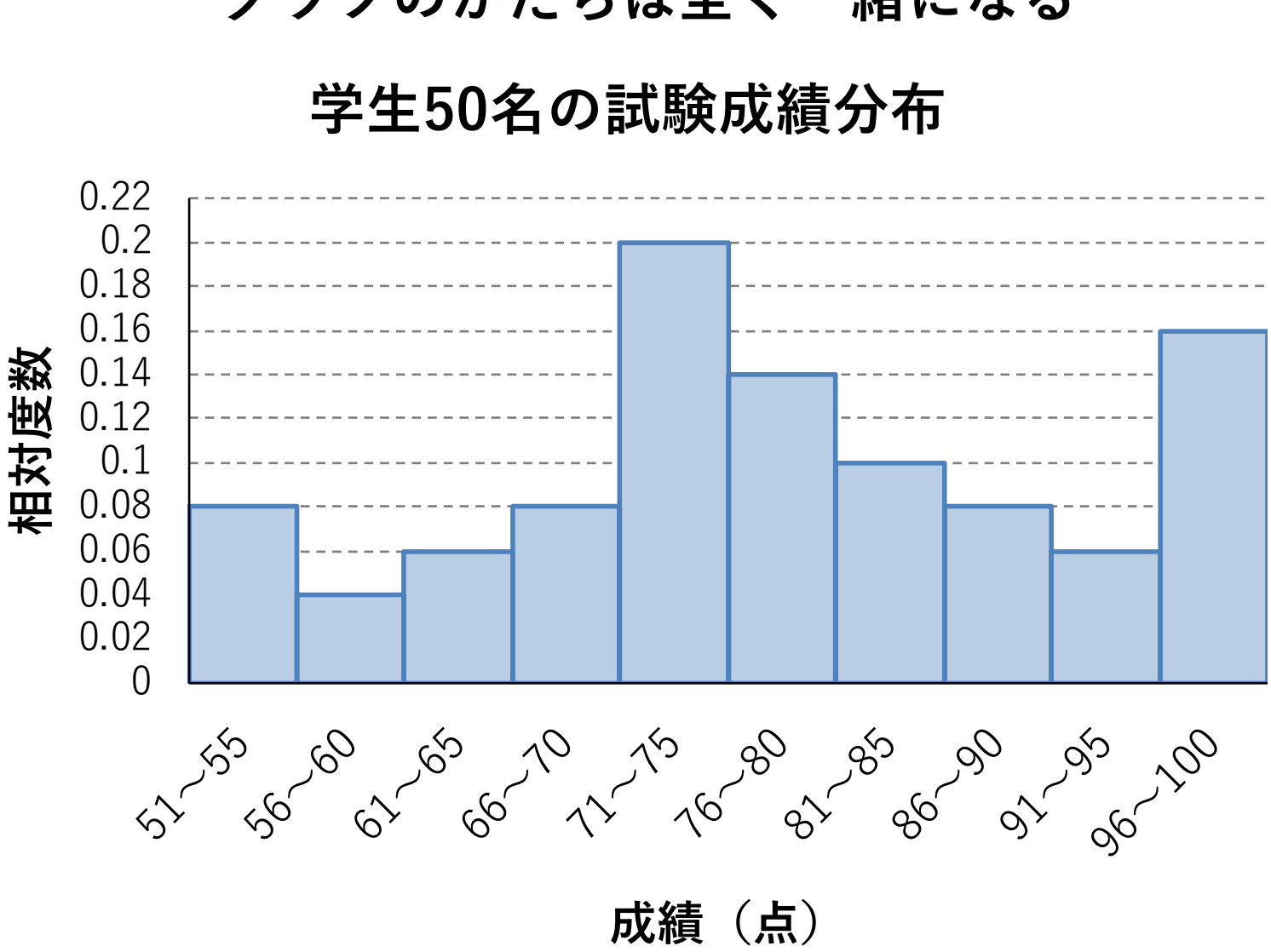
相対度数

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

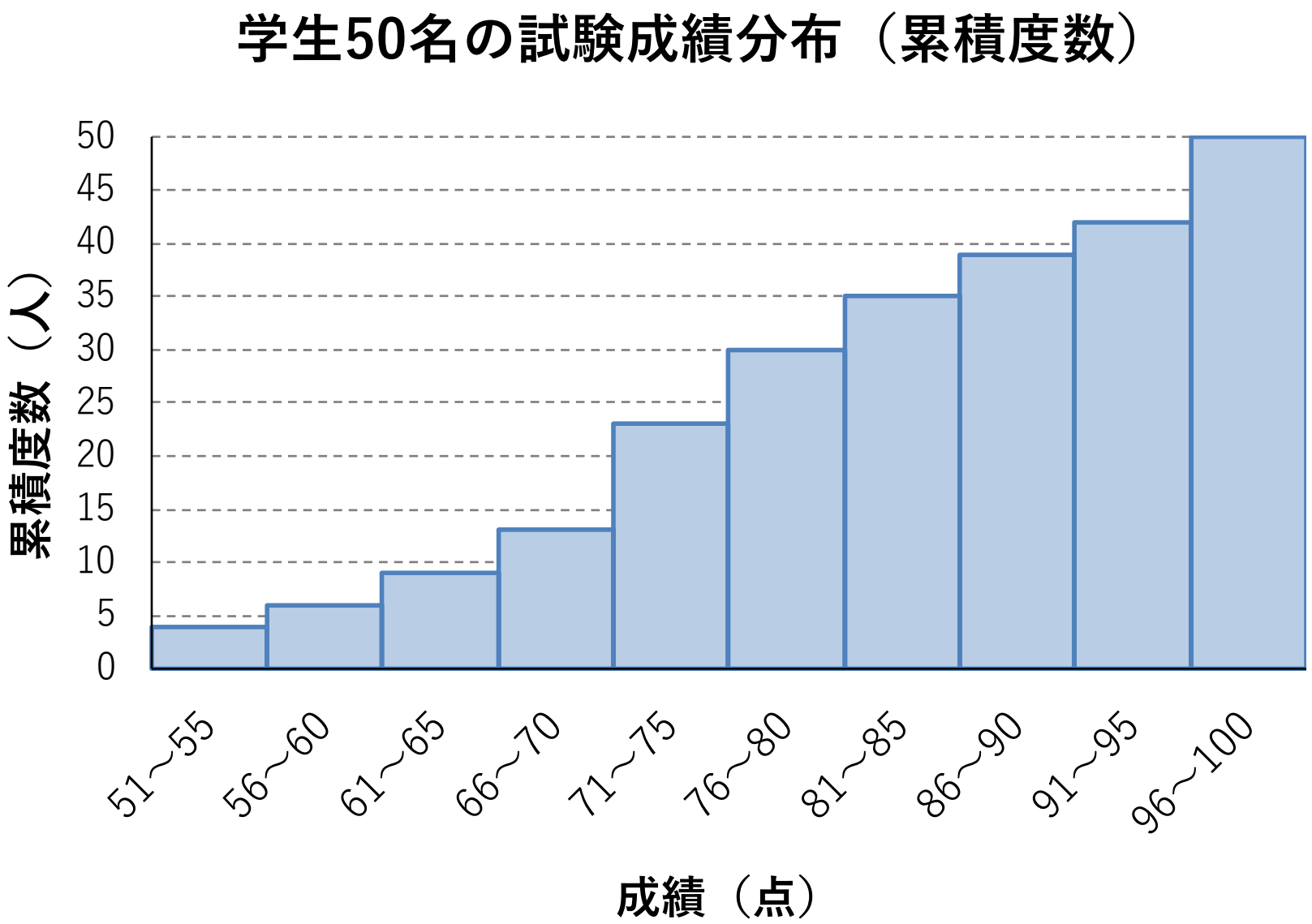
階級	階級値	度数	累積度数
51点～55点	53点	4	4
56点～60点	58点	2	6
61点～65点	62点	3	9
66点～70点	68点	4	13
71点～75点	73点	10	23
76点～80点	78点	7	30
81点 ～85点	83点	5	35
86点～90点	88点	4	39
91点～95点	93点	3	42
96点～100点	98点	8	50
合計		**50**	

- **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

この資料のまとめ

- **データの種類**
  - 質的データと量的データ
    - 平均を求めることができるのは･･･？
- **代表値**
  - 平均，中央値，最頻値
  - データの特性を客観的に記述
- **度数分布表とヒストグラム**
  - データの分布を素早く把握するのに便利
  - **データを収集したら，まずは分布を確認**

レポート課題その1 について

- **締切: 4月23日（火）講義開始時（厳守）**
- **手書き，パソコンで作成し印刷**
  どちらでもOK
- **配布した紙に直接手書きでもOK**
  - スペースが十分ではないかもしれないので，
    その場合は紙を適宜追加してください
  - なお，5月以降は紙では配付しない
    予定です

レポート課題その1 について

- 紙が複数枚になる場合は,
  左上をホチキス留めしてください
- 表紙は必要ありません
- 学生番号・氏名を忘れないように

レポート課題その1 について

- **余裕があるひとは，表やグラフなどを Excelで作成してみるとよいでしょう**
  - もっと余裕があるひとは，R について調べ，R で作図してみても良いでしょう

注意事項

- **他人のレポートをコピーすることは剽窃（ひょうせつ）にあたります**
  - 大学の規定に従い厳しく処分します
- **友だちで教え合うのはOKです**
  - むしろ推奨します
  - つまり，自分のことば，データでレポートを書いてください
- **参考にした書籍やURLがあれば記載**

次回

- **4月23日（火）**
- **講義開始時にレポートを回収します**
- **講義中に簡単な小テストを実施します**
  - 講義資料を見ながら回答してもらってよいので，ノートPCやスマートフォンを用意，あるいは講義資料を事前に印刷するなど，資料が見られるようにしておいてください