計量経済学

8. 測定 II

矢内　勇生

高知工科大学　経済・マネジメント学群

2018年10月30日

今日の目標

- 測定したデータを確認し、その内容を把握する方法を身につける
  - データの可視化 (visualization)
    - ヒストグラム
    - 箱ひげ図
    - 散布図
  - 記述統計による要約

データの可視化

- データを読み込んだら、まず可視化する！
  - ヒストグラム
  - 箱ひげ図
  - 散布図

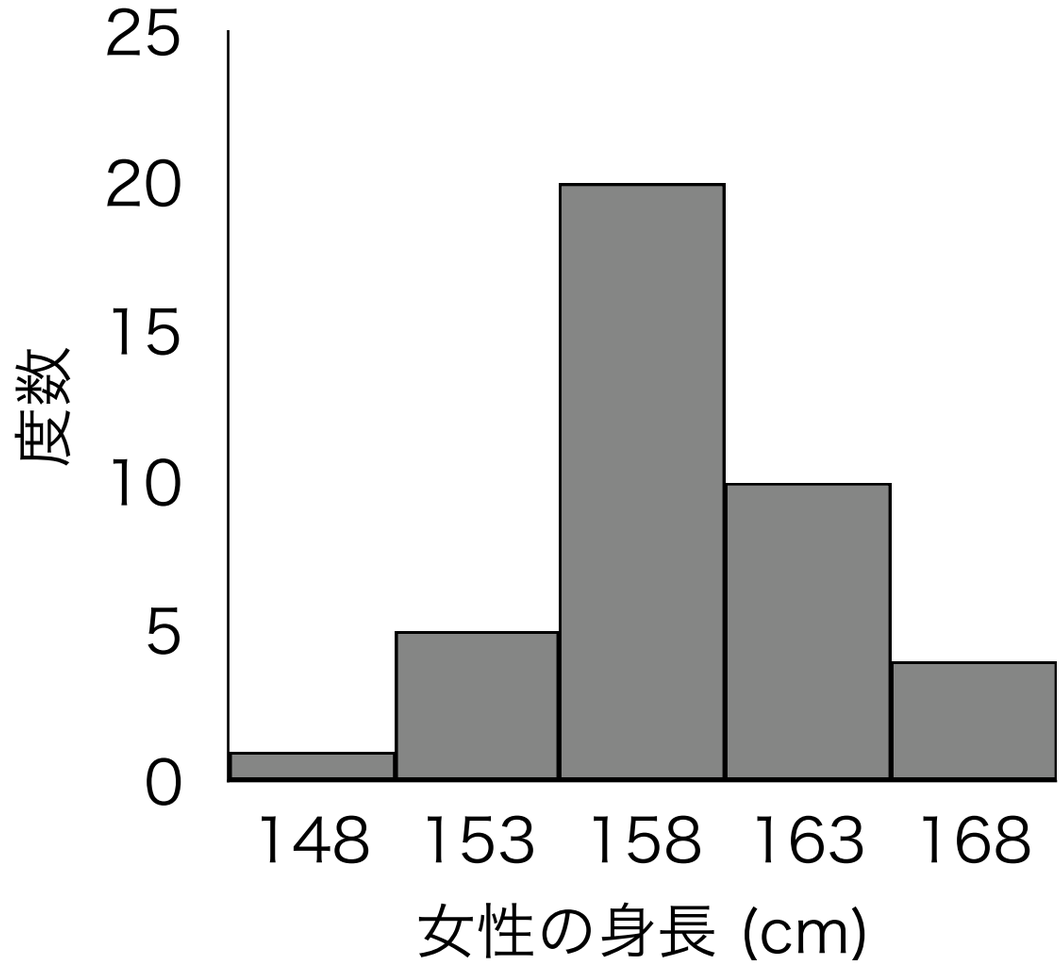
ヒストグラム (histogram)

- 注目するポイント

1. どこにデータが集まっているか（棒が高いのはどこか）

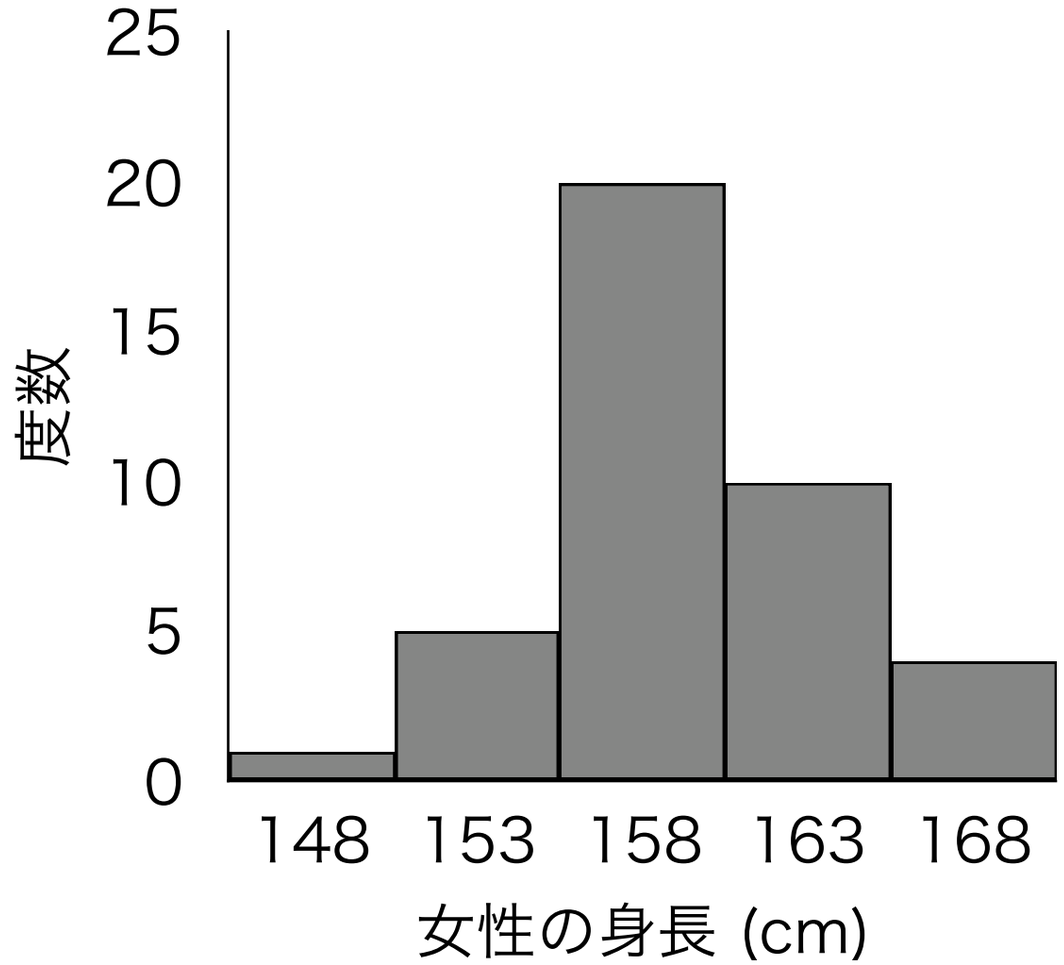
2. データが分布している範囲は？

3. 全体の形状は？



ポイント1：どこにデータが集まっているか

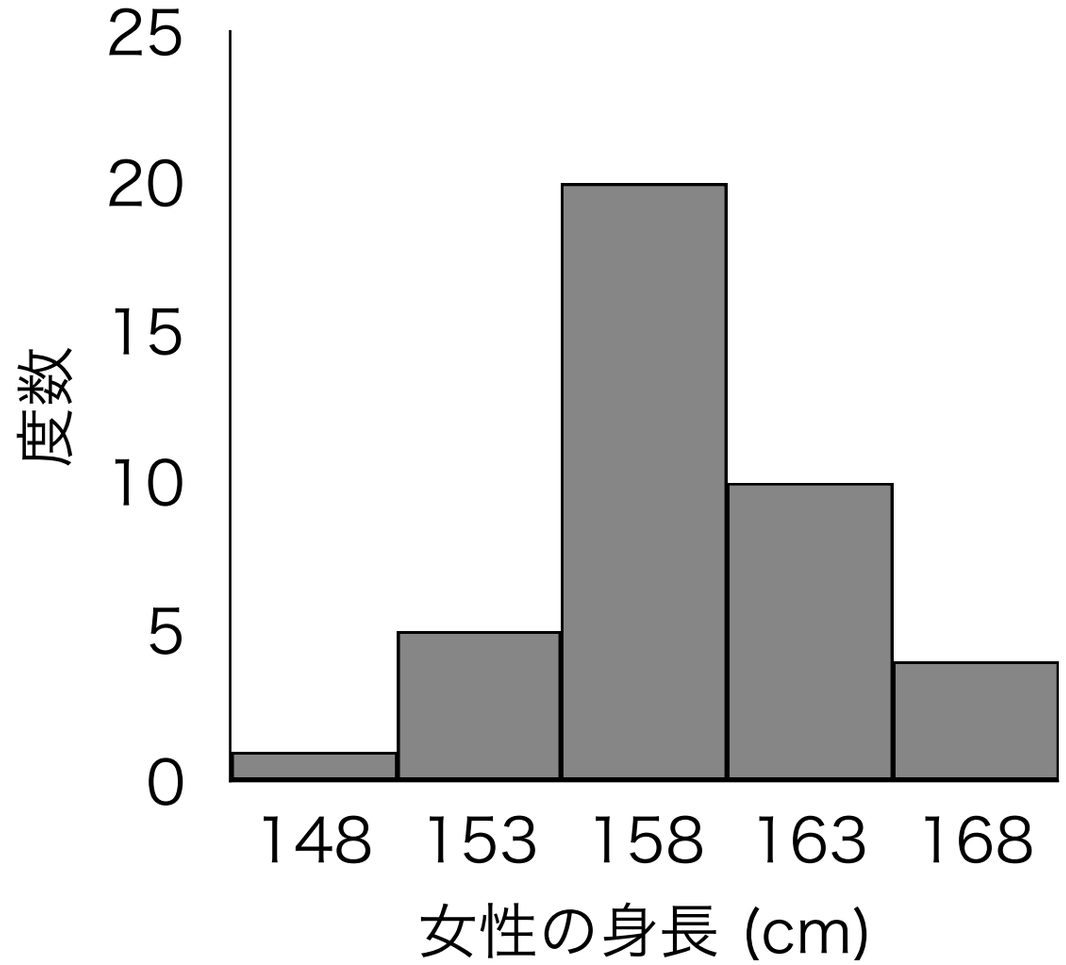
- 棒が高いところにデータが集まっている
- 高い棒と周りの棒との差 ＝ データの集中度
  - 身長が158cm ほどの女性が多い



ポイント2：データの分布している範囲は？

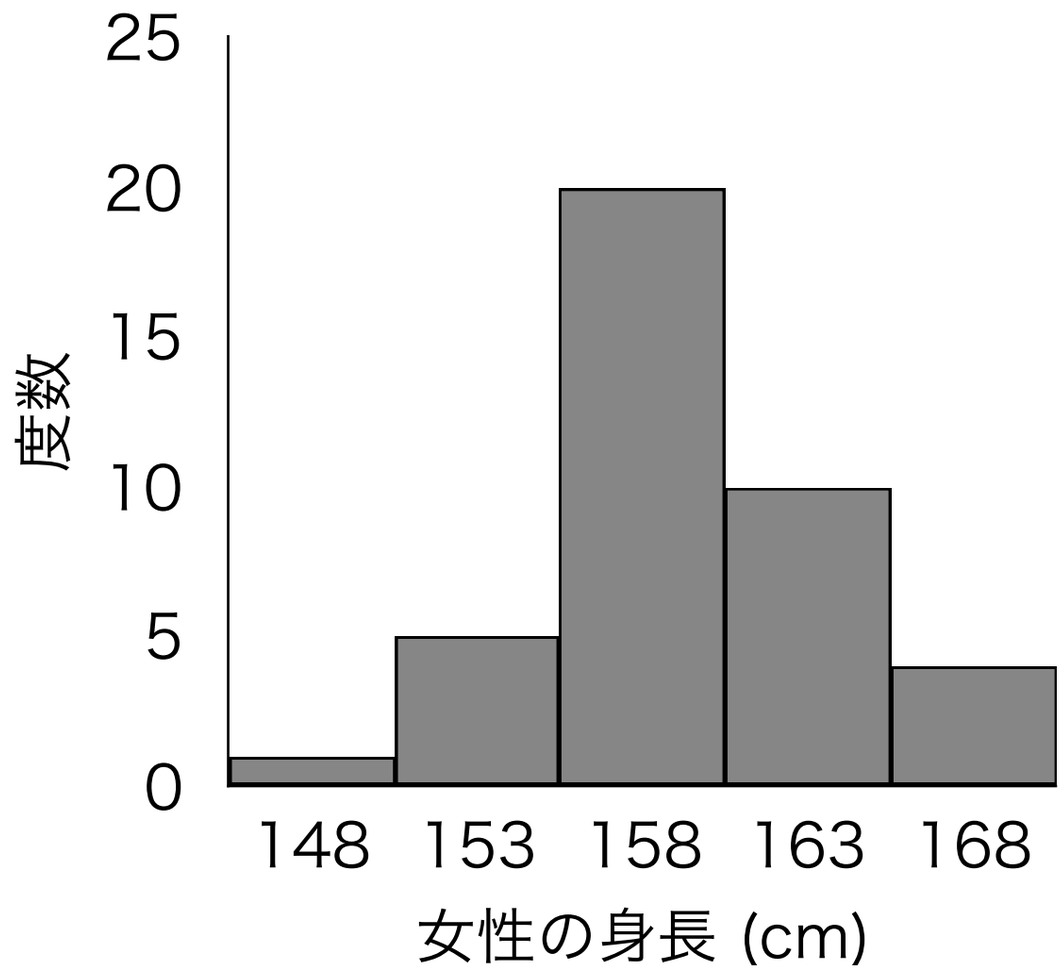
- データがある場所と
  ない場所がある

- 145cm 以下や
  171cm 以上の女性
  はいない（注：デー
  タをとった40人の中
  にいないだけ！）



ポイント3：全体の形状は？

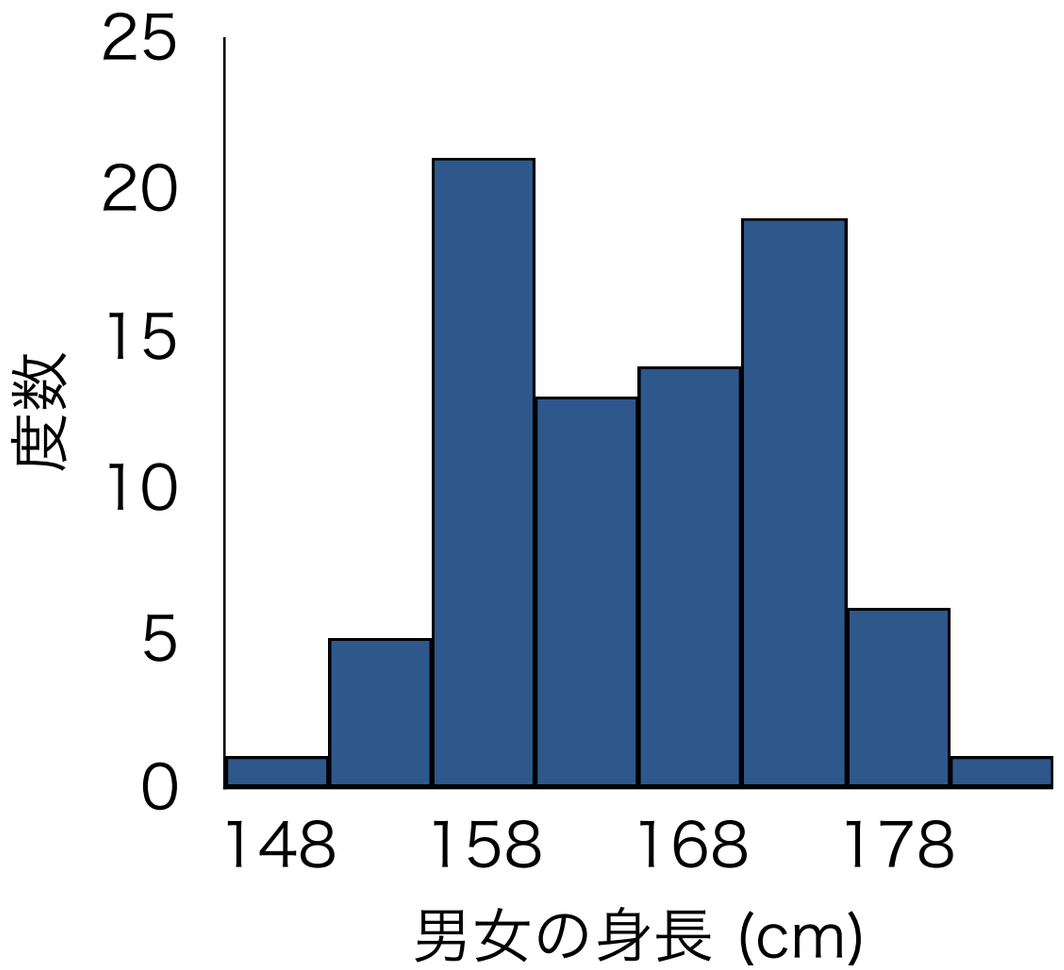
- 山はいくつある？
  - 山は1つ = 単峰型分布
- 左右対称？
  - ほぼ左右対称



ポイント3（続）：山の数

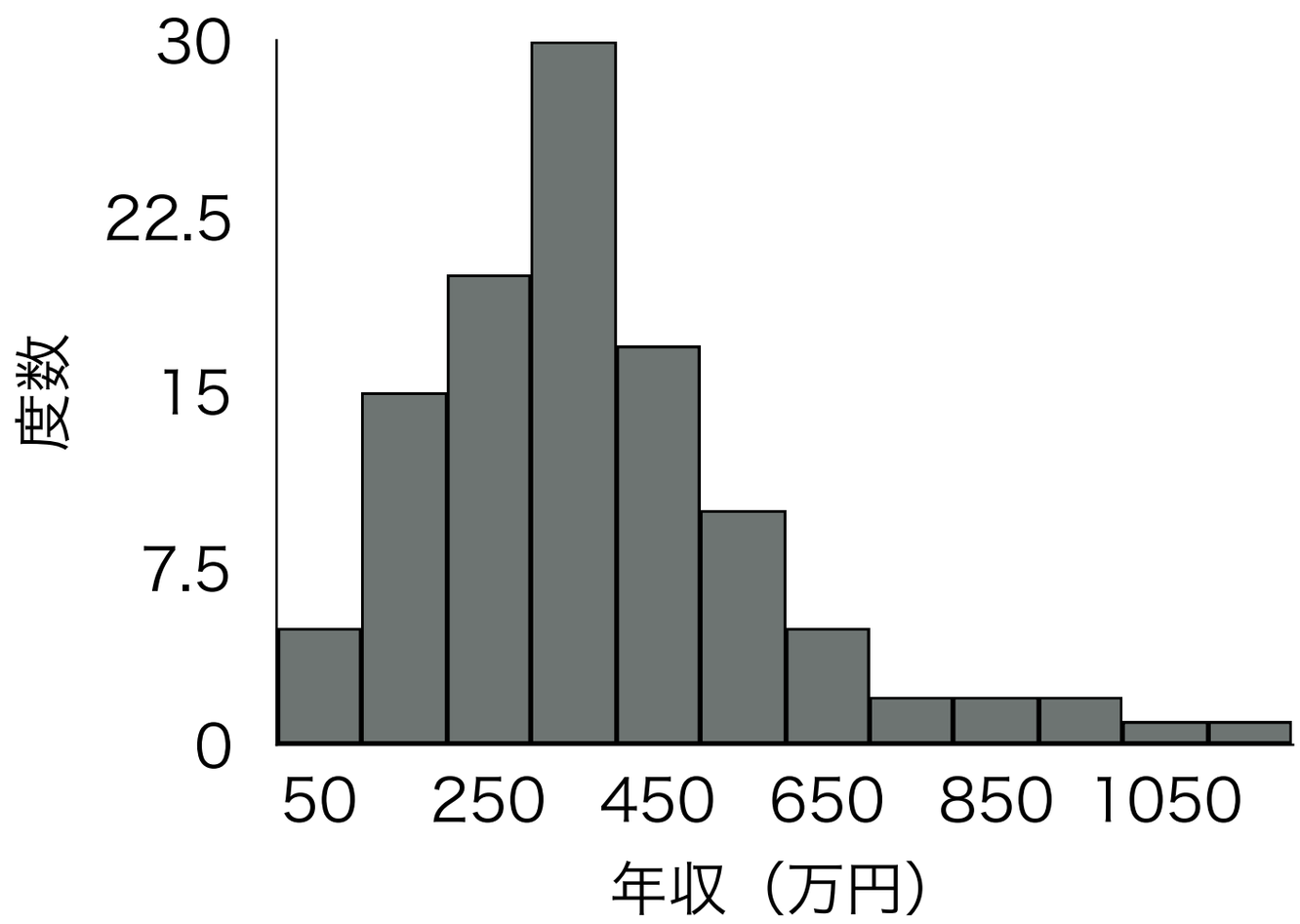
- 右のヒストグラムの山は
  2つ ＝ 双峰型分布
  - 異質なグループをひとつにまとめると、双峰型になりやすい

- 山が3つ以上の場合は多峰型と呼ぶ



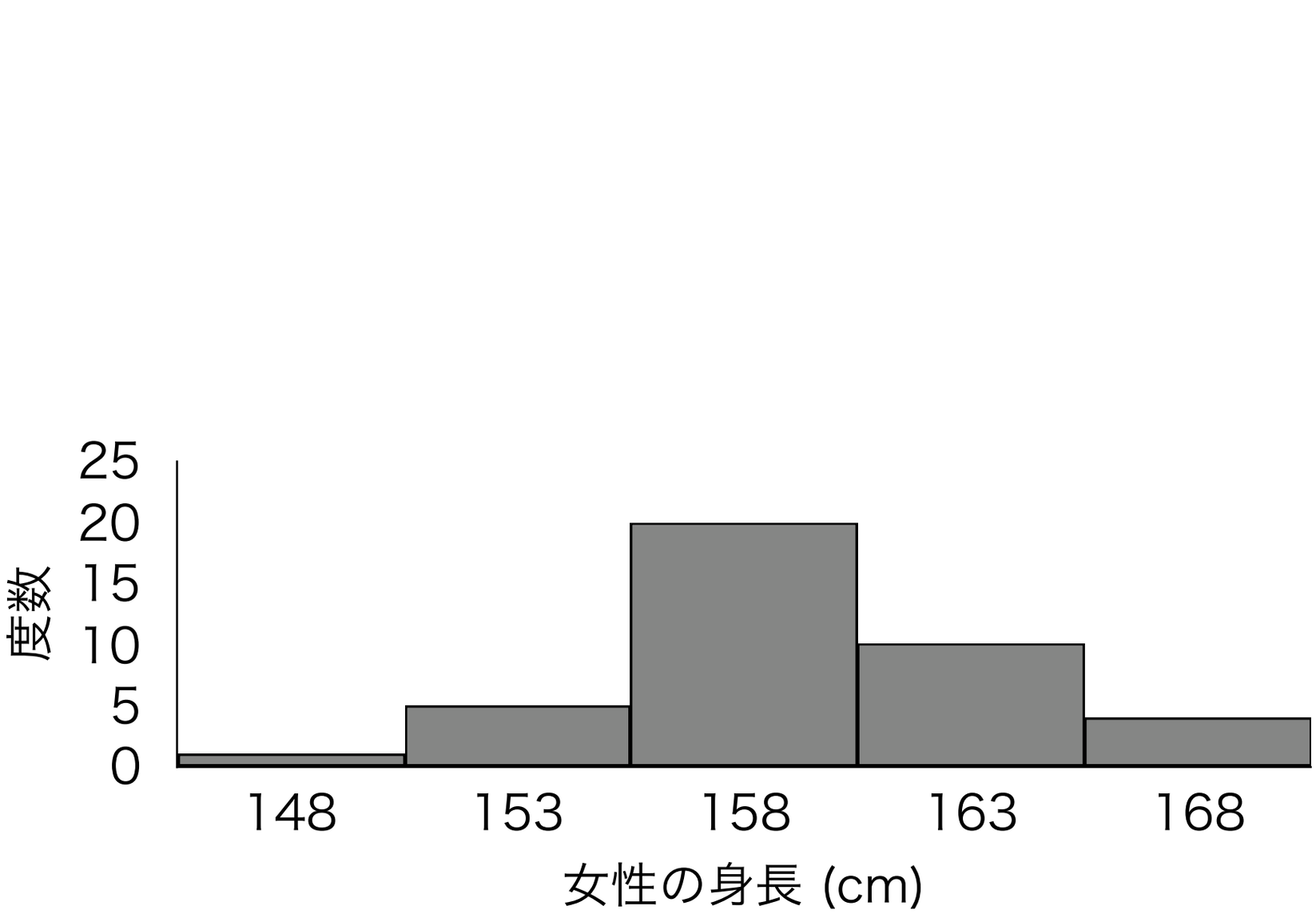
ポイント3（続）：対称性

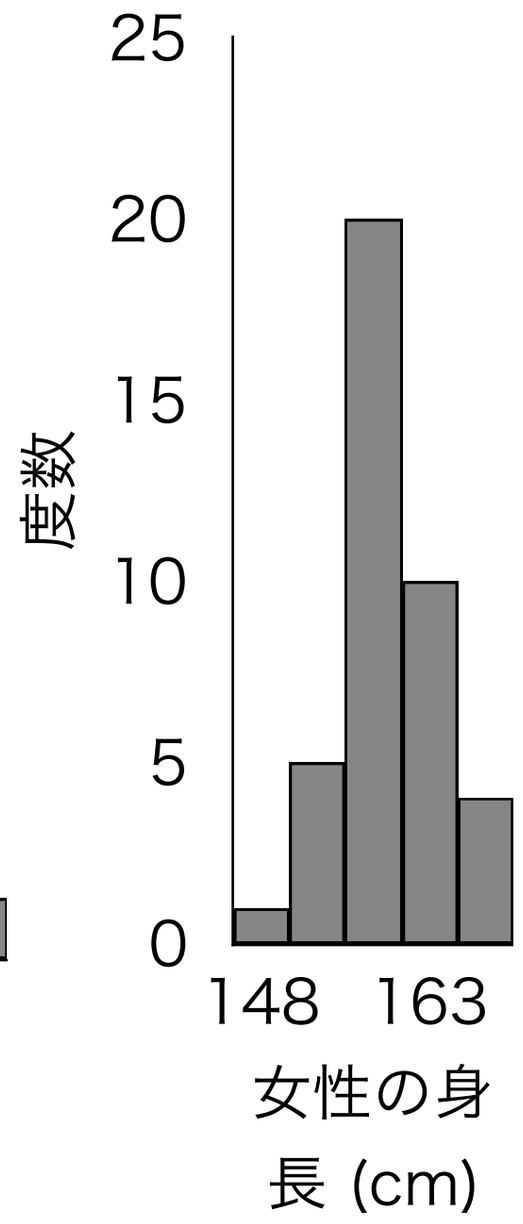
- 山より右に離れた位置に大きな値が少数存在する = 右に歪んだ分布
- 右に歪んだ分布は社会のデータに多い
- 左右対称は、自然のデータに多い



ヒストグラムの読み方は主観的

- まったく同じヒストグラムなのに、見た目の印象が違う！





データの中心とばらつきを調べる（統計学の復習）

- データの中心的傾向を表す統計量
  - 平均値、中央値、最頻値
- データのばらつきを表す統計量
  - 分散、標準偏差（範囲、四分位範囲）

統計量（statistic）

- 統計量とは
  - データの**ある特徴**を表す数字
  - 統計学で決められた方法を使うことによって得られる
- 様々なstatistic について研究するのがstatistic**s** （統計学）

代表値

- データの中心的傾向を表す統計量を「代表値」とよぶ
- 代表値の例
  - 平均値 (mean)
  - 中央値 (median)
  - 最頻値 (mode)

平均値（mean）

- 平均にはいくつかの種類がある
  - 算術平均、相加平均（arithmetic mean）
    単に「平均」と言ったらこれのこと
  - 幾何平均、相乗平均（geometric mean）
  - 調和平均（harmonic mean）
- 目的に合わせて適切なものを選ぶ

算術平均

- 算術平均 = 値の合計 ÷ n
- 例）5人の年収の平均を求める

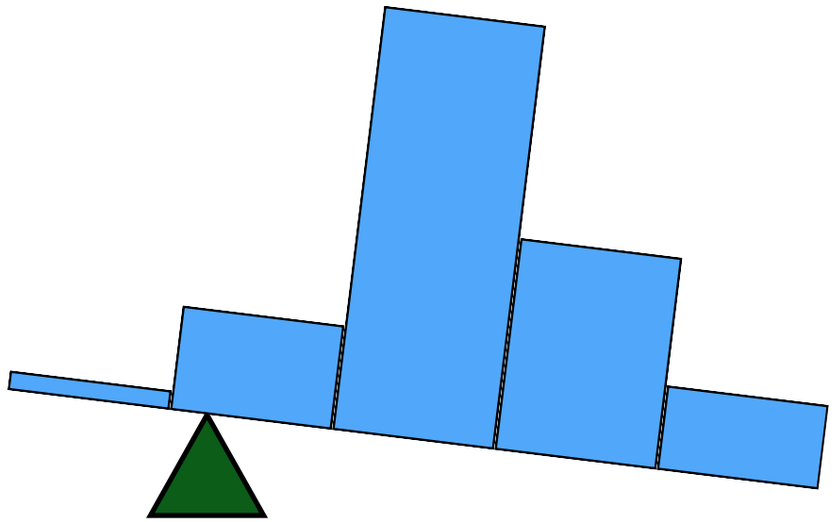
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{5} x_i}{5}$$
$$= \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$
$$= \frac{350 + 450 + 500 + 600 + 800}{5}$$
$$= \frac{2700}{5} = 540$$

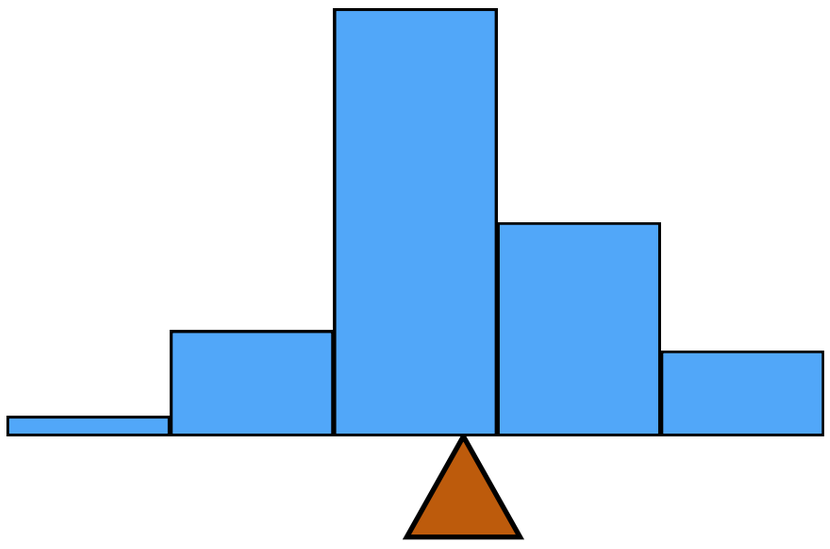
年収
350
450
500
600
800

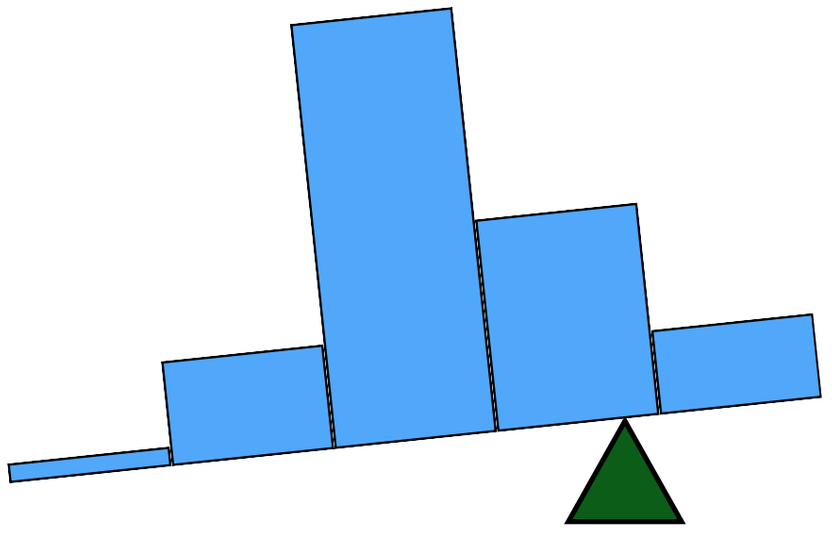
単位：万円

算術平均とヒストグラム

- 算術平均はヒストグラムのバランスをとる支点（やじろべえの支点、重心）







平均値の弱点

- 外れ値の影響を受けやすい
  - 「外れ値」とは、データの中の他の値に比べ、飛び抜けて大きい（小さい）値

➡外れ値に強い統計量は？

中央値（median）

- データの中央にある値（中位値ともいう）

- 中央値の求め方

1. データを小さいものから順番に並べる

2. ちょうど真ん中にあるものが中央値

3. 真ん中がない（2つある）場合、2つの値の算術平均が中央値

中央値の例

- C社：600万円
  - たまたま平均と同じ値
- D社：400万円
- E社：真ん中が2つ

➡ (450+510) / 2 = 480万円

年収		
C社	D社	E社
550	350	400
550	350	420
600	**400**	**450**
650	450	**510**
650	1450	550
		600

単位：万円

中央値の長所

- 外れ値の影響を受けにくい

  例：ある会社の給料の変化（2017年から2018年）

  - 平均値：500万円→780万円
  - 中央値：500万円のまま

従業員	2017	2018
1	400	400
2	450	450
3	500	500
4	550	550
5	600	2000

中央値の欠点

- 与えられた情報をすべて使っていない
  - （例）F社もG社も中央値は1000万円
  - しかし、分布の中身は違う

年収	
F社	G社
850	350
900	900
1000	**1000**
1100	1100
1150	1150

単位：万円

どの代表値を使う？

- 一致するときはどれでもよい
- 目的に応じて使い分けることが必要
- 統計を読むときは、どの代表値が使われているか意識することが大事

代表値だけに頼らない！

- 2つのグループの代表値（平均、中央値）が同じだからといって、グループ同士が似ているとは限らない

➡範囲を確かめてみる

範囲は代表値とセットで

- 範囲だけを見てもあまり意味がない
  - C社の年収の範囲とD社の年収の範囲は同じ（300万円）だが・・・
  - 平均は？

年収	
C社	D社
350	600
400	650
500	750
600	850
650	900

単位：万円

範囲の弱点

- 範囲は、外れ値の影響を受け易い
  - E組の試験得点の範囲：14
  - F組の試験得点の範囲：51
- F組は1人の得点がきわめて悪かったため、範囲が大きくなってしまった

試験の得点	
E組	F組
68	30
70	70
75	75
78	80
82	81

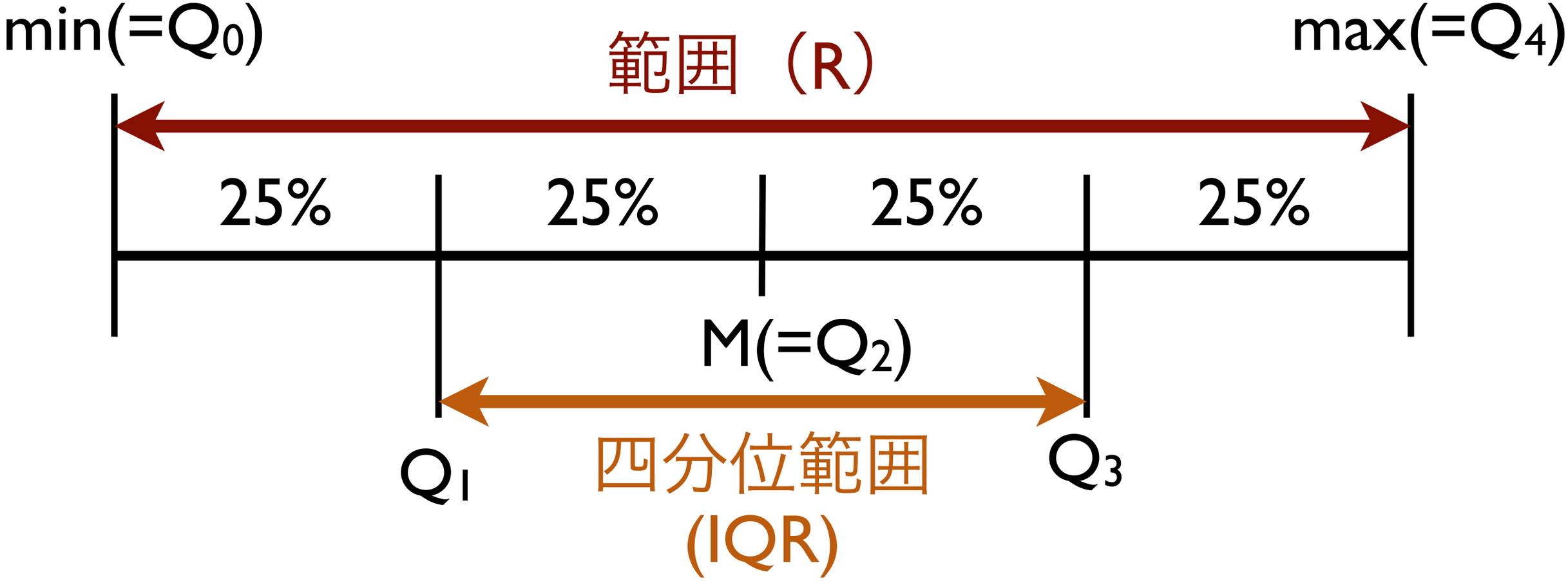
範囲の弱点：極端な例

- E組では99人が100点、1人が90点を取った
- F組では99人が100点、1人が10点を取った
  - それぞれの範囲はどうなる？
  - 範囲の値が大きく異なるからといって、2つのグループがまったく異質だといえる？

四分位数 (quartile)

- データを4等分する区切り（境界線）の値
- 4等分すると境界線は5つできる
  - 最小値 [(Q_0 =) min]
  - 第1四分位 [Q_1]
  - 第2四分位 = 中央値（中位値）[(Q_2 =) M]
  - 第3四分位 [Q_3]
  - 第4四分位 = 最大値 [(Q_4 =) max]

データを小さい順に並べ替え、4等分する



四分位範囲
(interquartile range)

- 略してIQR

- $IQR = Q_3 - Q_1$

- 小さい方から25%のデータと大きい方から25% のデータを省いているので、外れ値の影響を受けにくい

注意：4等分にするのはデータの値の「個数」

- データの範囲を4等分にするのではない
- 例：データ = {0, 1, 2, 3, 4, 8, 9, 10}

×範囲を4等分する：2.5, 5.0, 7.5 を区切りにして
{0, 1, 2}, {3, 4}, { }, {8,9,10}の4グループに分ける
（注：3つ目のグループは空集合）

○個数を4等分する：{0, 1}, {2, 3}, {4, 8}, {9, 10} の
4グループに分ける

四分位の求め方（1）

- 5つの境界線のうち、おなじみの統計量
  - （第0四分位＝）最小値
  - （第4四分位＝）最大値
  - 第2四分位　＝　中央値

➡問題は、第1四分位と第3四分位の求め方

四分位の求め方（2）

1.中央値を見つける

2.第1四分位：データ全体の中央値より小さい値の中の中央値

3.第3四分位：データ全体の中央値より大きい値の中の中央値

四分位の求め方：例1

- 中央値 = (76 + 78) / 2 = 77

- **第1四分位**：77より小さい値の中の
  中央値 → (68 + 70) / 2 = 69

- **第3四分位** ：77より大きい値の中の
  中央値 → (85 + 88) / 2 = 86.5

試験の得点

60	78
62	81
68	85
70	88
75	90
76	95

n=12

四分位の求め方：例2

★ n が奇数のとき

➡小さい（大きい）方の半分に中央値を含まない

- 中央値 = 76
- **第1四分位** = 68
- **第3四分位** = 85
  - 注：中央値と同じ値であっても、中央値そのものでなければ除外しない

試験の得点

60	76
62	81
68	85
70	88
76	90
76	

n=11

m分位数

- 四分位数はデータを4つに分ける（m=4）が、他にも様々な分け方が考えられる
- 他によく使われる分位数
  - m = 10：十分位数（decile）
  - m =100：百分位数（percentile）

百分位数

- 「パーセンタイル（percentile）」
- データを100等分したときの境界線
  - 25パーセンタイル = 第1四分位
  - 50パーセンタイル = 第2四分位 = 中央値
  - 75パーセンタイル = 第3四分位

注：四分位の求め方は色々ある

（この頁は興味がある者のみ読むこと）

- 厳密には、その値以下の値の数が25%（75%）になるような値を第1四分位（第3四分位）という

- 授業で解説した方法では、上の定義とずれることがある（多くの場合、ズレはわずか）

- 授業で解説した方法で求めたものをヒンジ（hinges）と呼び、四分位とは別のものとして扱う場合もある

  - 授業で求めた第1四分位 ：下側ヒンジ

  - 授業で求めた第3四分位：上側ヒンジ

範囲と四分位範囲

- 中央値：77（G組）> 76（H組）
  - 中央値はほとんど同じ
- 範囲 ：35（G）< 75（H）
- 四分位範囲 ：17.5（G）> 16.5（H）
  - 範囲はH組の方が大きいが、四分位範囲はG組のほうが大きい

試験の得点	
G組	H組
60	25
62	65
68	67
70	68
75	73
76	76
78	76
81	80
85	84
88	84
90	87
95	100

五数要約
(five-number summary)

- 最小値、第1四分位、中央値、第3四分位、最大値の5つの数字でデータの特徴を表すこと

- メリット：データの中心的傾向とともに範囲、四分位範囲という散らばりの傾向もわかる

五数要約の例

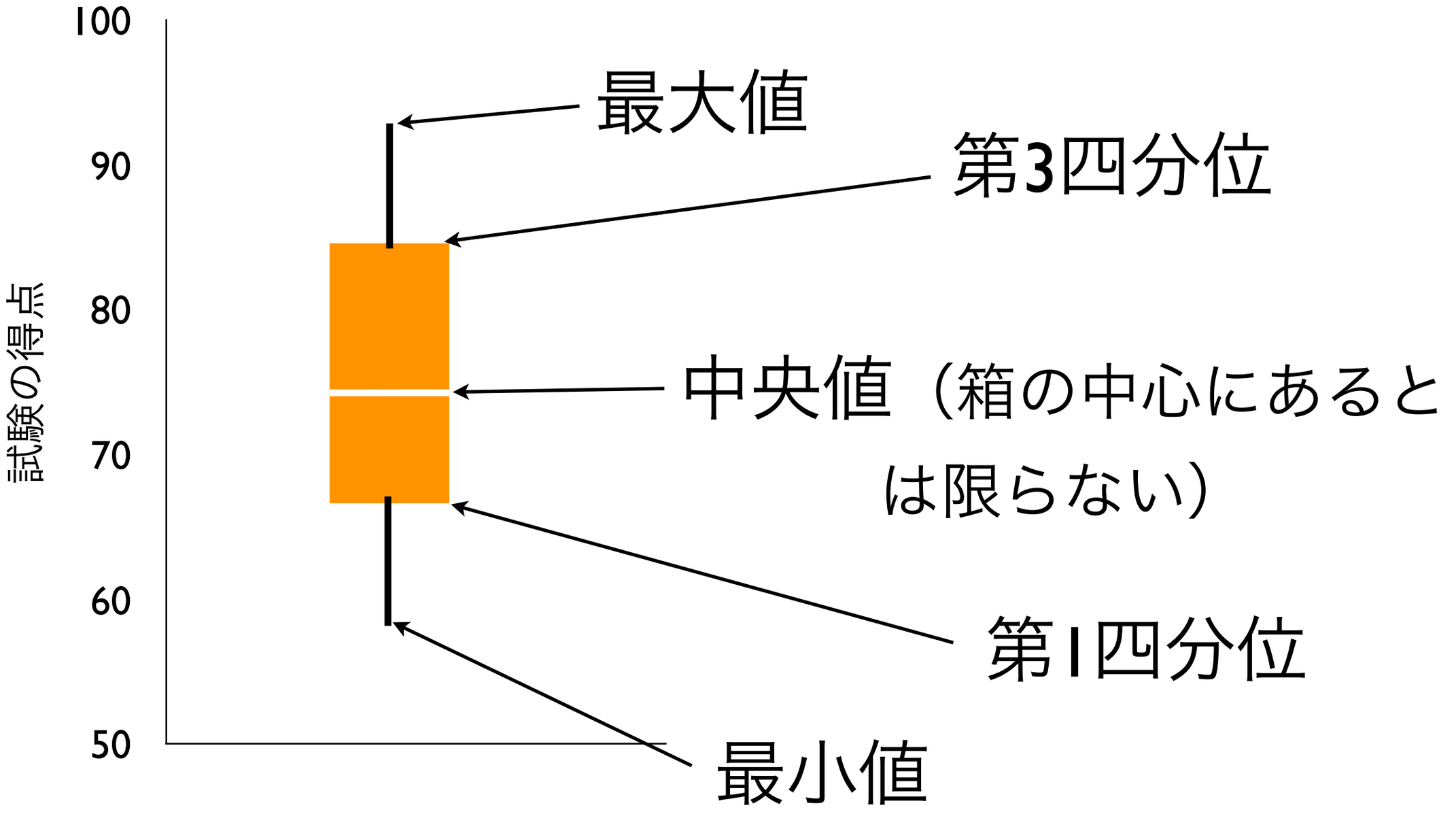
表：G組とH組の得点の五数要約

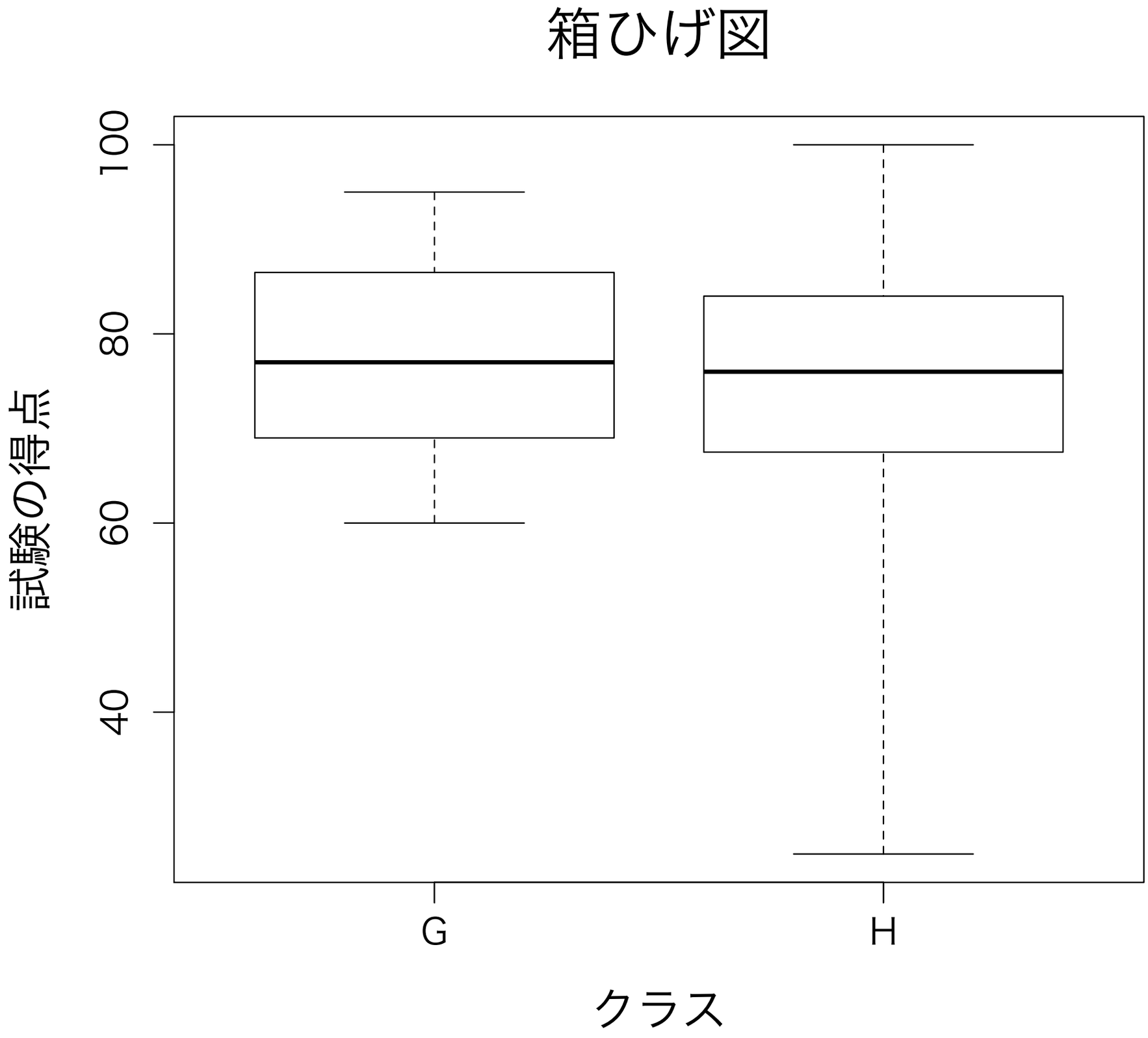
	最小値	第1四分位	中央値	第3四分位	最大値
G組	60	69	77	86.5	95
H組	25	67.5	76	84	100

五数要約を図示する

- 箱ひげ図（box-and-whisker plot）
  - 箱で四分位範囲を表す
  - ひげで四分位外の範囲を表す
  - 箱の中の線で中央値を表す

箱ひげ図


100
90
80
70
60
50
試験の得点
最大値
第3四分位
中央値（箱の中心にあるとは限らない）
第1四分位
最小値


箱ひげ図
試験の得点
100
80
60
40
G
H
クラス

IQR で外れ値を見つける

- 外れ値を見分けるためにIQR を利用する
- 第1四分位から1.5×IQR より小さい値は「外れ値の疑いがある」と考える
- 第3四分位から1.5×IQR より大きい値は「外れ値の疑いがある」と考える

1.5×IQR ルールの適用例（1）

- G組
  - IQR = 17.5 → 1.5IQR = 26.25
  - 第1四分位は69：69-26.25=42.75 より小さい値は「外れ値の疑い」
  - 第3四分位は86.5：86.5+26.25=112.75 より大きい値は存在しない

➡G組の得点に外れ値はない

試験の得点

G組	H組
60	25
62	65
68	67
70	68
75	73
76	76
78	76
81	80
85	84
88	84
90	87
95	100

1.5×IQR ルールの適用例（2）

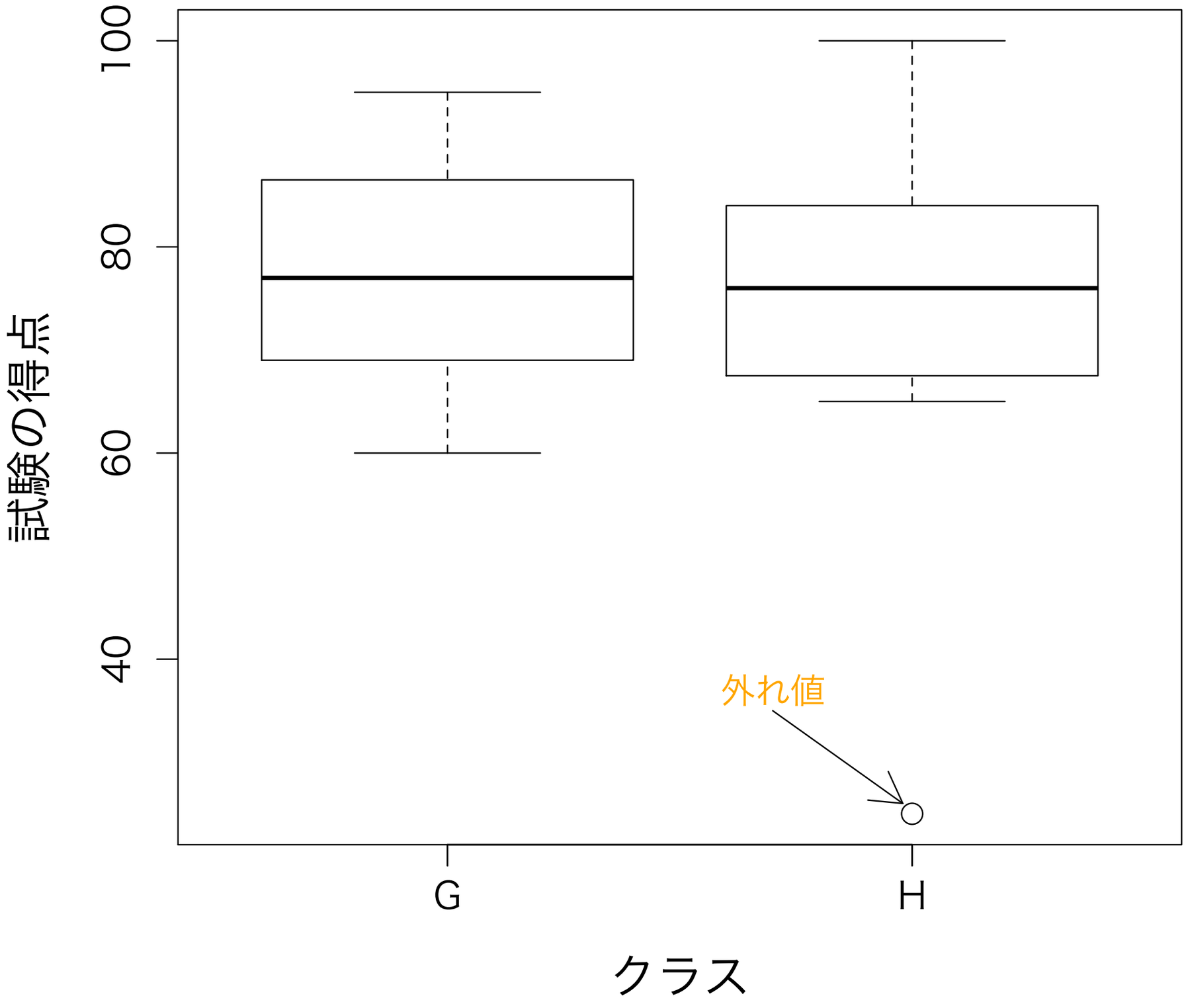
- H組
  - IQR = 16.5 → 1.5IQR = 24.75
  - 第1四分位は67.5：67.5-24.75=42.75
    より小さい値は「外れ値の疑い」
  - 第3四分位は84：84+24.75=108.75
    より大きい値は存在しない

➡H組の25点は外れ値

試験の得点

G組	H組
60	**25**
62	65
68	67
70	68
75	73
76	76
78	76
81	80
85	84
88	84
90	87
95	100

外れ値を考慮した箱ひげ図



外れ値を探す理由

- データの間違いによる外れ値ではないか確認する
  - 入力・記入ミス
  - 異質なデータ（例：国語の点数の中にひとつだけ数学の点数、成人の身長の中に小学生の身長など）

  ➡データを修正する必要がある

- 例外的な値だからといって、何も考えずに分析から除外していいわけではない

範囲、四分位範囲の問題点

- 範囲や四分位範囲はすべての情報を利用していない
- 「全体的な」ばらつき（散らばり具合）がわからない

➡すべての情報を利用して全体的なばらつきを考えよう！

データの全体的なばらつきを調べる

- 中心的傾向が同じでも、似たようなデータとは限らない
  - 平均値も中央値も一緒だが・・・

試験の得点

A組	B組
40	10
45	30
50	50
55	80
60	100

分散 (variance)

- データのばらつきを表す統計量
- **統計学で最も重要な統計量**
- s^2 という記号で表す
- 分散 s^2 ＝ 「偏差の二乗」の平均値
- 標本で計算するときは、不偏分散を使う

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

不偏分散の例

- 右のデータの分散を求めてみる

x	偏差	偏差^2
1	-4	16
4	-1	1
5	0	0
7	2	4
8	3	9

$$s_x^2 = \frac{16 + 1 + 0 + 4 + 9}{5 - 1}$$

$$= 7.5$$

分散の問題点

- 値を二乗するので、単位が変わってしまう
  - 例：身長をcm（距離）で測ったデータを二乗すると、単位がcm^2（面積）に変わってしまう

➡距離データの散らばり具合を面積で表現されても意味がつかみにくい

標準偏差 (standard deviation)

- 略して SD　あるいは sd

- 単位を元に戻すために、分散の平方根をとる

- 標準偏差 s は、不偏分散s^2の平方根

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

Rで記述統計を求める

1.平均値： mean()

2. 中央値： median()

3. 五数要約：fivenum()

4. パーセンタイル：quantile()

5. 不偏分散：var()

6. 不偏分散の平方根（標準偏差）：sd()

Rで図を作る

- ggplot2パッケージを使う!
  - ヒストグラム：geom_histogram()
  - 箱ひげ図：geom_boxplot()
  - 散布図：geom_point()
  - etc.