高知工科大学 経済・マネジメント学群

統計学 2

3. 記述統計とデータの可視化

矢内 勇生（やない ゆうき）

https://yukiyanai.github.io

yanai.yuki@kochi-tech.ac.jp

このトピックの目標

- 記述統計をRで計算する
  - 基本的な統計量を思い出す
  - Rでそれらを計算する方法を身につける
- データを可視化して要約する
  - データの可視化で注意すべきポイントを学ぶ

記述統計

統計学における文字の使い分け

- ギリシャ文字：見えない真実（パラメタ; parameters）
  - https://ja.wikipedia.org/wiki/ギリシア文字
- ギリシャ文字に修飾：推定値
- アルファベットに修飾：統計量（データから計算されるもの）
- アルファベット：データ

$$x \to \bar{x} \to \hat{\mu}_x \to \mu_x$$

データの要約に使われる主な統計量

- データの中心的傾向を表す統計量
  - 算術平均、加算平均 (mean)
  - 中央値、中位値 (median)
  - 最頻値 (mode)
- データのばらつきを表す統計量
  - 範囲 (range)
  - 四分位範囲 (interquartile range; IQR)
  - 分散（variance）
  - 標準偏差 (standard deviation)

算術平均 (mean)

- 観測個体を識別する添字： i $(i = 1,2,\ldots,N)$
- データ: x_i
- x の平均値を $\bar{x}$ と表記し、「エックスバー」と読む

$$\bar{x} = \frac{1}{N}(x_1 + x_2 + \cdots + x_N)$$

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

- Rで変数 x の平均値を求める：mean(x)

算術平均の弱点

- **外れ値** (outlier) 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)

- x_i を小さい順（または大きい順）に並べ替えたとき、「ちょうど真ん中」の位置にある x_i を x の中央値と呼ぶ
  - 「ちょうど真ん中」が2つあるときは、その2つの値の平均値が中央値
  - より正確には、$P(x_i \leq x_k) \geq 1/2$ かつ $P(x_i \geq x_k) \geq 1/2$ となる x_k が中央値
- Rで変数 x の中央値を求める：`median(x)`

中央値の欠点

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

年収	
A社	B社
850	350
900	900
1000	1000
1100	1100
1150	1150

単位：万円

最頻値 (mode)

- x_i のなかで最も頻繁に現れる値を x の最頻値と呼ぶ
  - より正確には x のうち、確率（離散型確率変数の場合）または確率密度（連続型確率変数の場合）が最大のものが x の最頻値
- 最頻値は複数存在することがある（短所）
- Rで変数 x の最頻値を求める：一発で求める関数はないので省略（必要になったときに説明する）

平均値、中央値、最頻値の関係

- 完全に左右対称の分布：三者が一致
- 右に歪んだ分布：

最頻値 ＜ 中央値 ＜ 平均値

Mode ＜ Median ＜ Mean （辞書に出てくる順番が早いほど大きい）

範囲 (range)

- x_i のうち最小のものを x_m , 最大のものを x_M としたとき、$x_M - x_m$ を x の範囲と呼ぶ
  - 例：$x_m = 12, x_M = 50$ なら、範囲は $50 - 12 = 38$
  - Rで求める：`max(x) - min(x)`
- 文脈によっては最小値と最大値の間の区間 $[x_m, x_M]$ を範囲と呼ぶことがあるので注意
  - 例：[12, 50]
  - Rで求める：`range(x)`

範囲の弱点

- 範囲は、外れ値の影響を受け易い
  - C組の試験得点の範囲：14
  - D組の試験得点の範囲：51
- D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う

試験の得点	
C組	D組
68	30
70	70
75	75
78	80
82	81

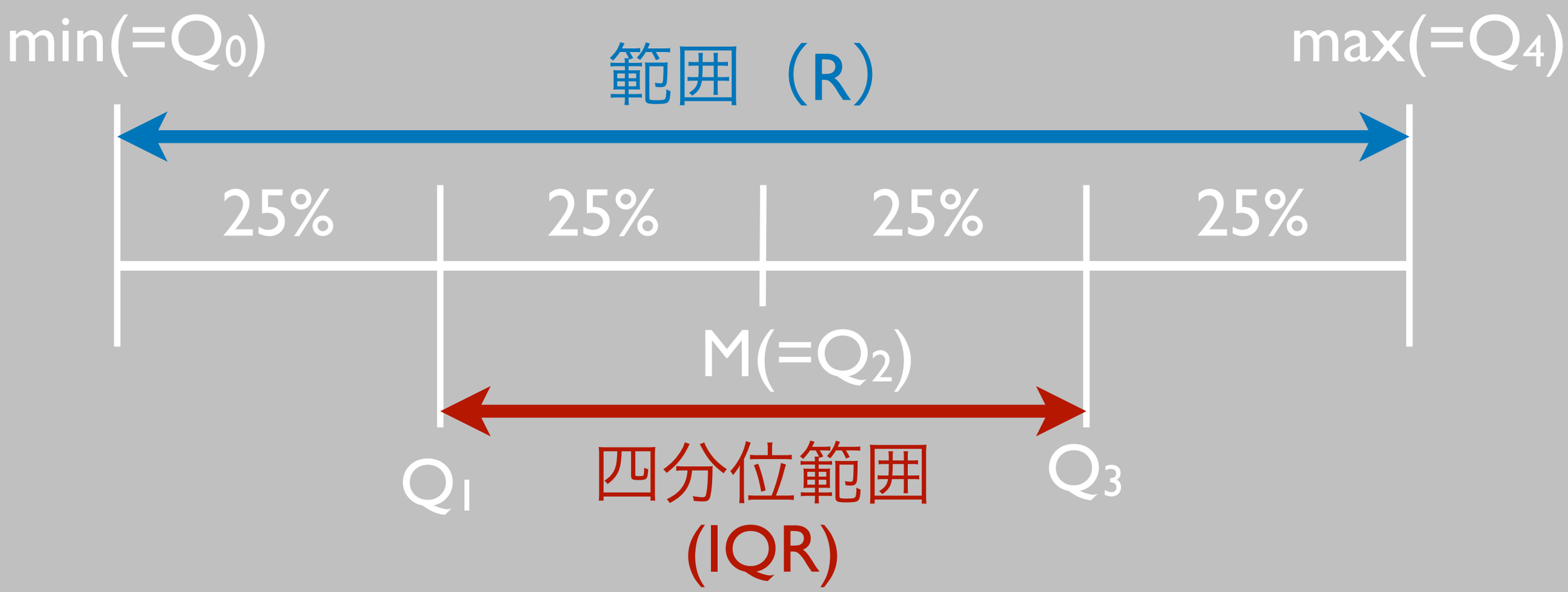
範囲の弱点：極端な例

- C組では99人が100点、1人が90点を取った
- D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 [(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

四分位数の求め方（1）

- 5つの境界線のうち、3つは簡単
  - （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1四分位数と第3四分位数の求め方

四分位数の求め方（2）

1.中央値を見つける

2.第1四分位数：データ全体の中央値より小さい値の中の中央値

3.第3四分位数：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1四分位数：77より小さい値の中の中央値 → (68 + 70) / 2 = 69
- 第3四分位数 ：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）ほうの半分に中央値を含まない

- 中央値 = 76
- 第1四分位数：68
- 第3四分位数 = 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

Rで分位数を求める

- 四分位範囲: `IQR(x)`
- パーセンタイルを求める例
  - 25パーセンタイル（第1四分位数）：

```
quantile(x, prob = 0.25)
```

  - 75パーセンタイル（第3四分位数）：

```
quantile(x, prob = 0.75)
```

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

試験の得点

E組	F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
- Rで求める（どちらでも同じ）：
  - `quantile(x, prob = c(0, 0.25, 0.5, 0.75, 1))`
  - `fivenum(x)`

五数要約の例

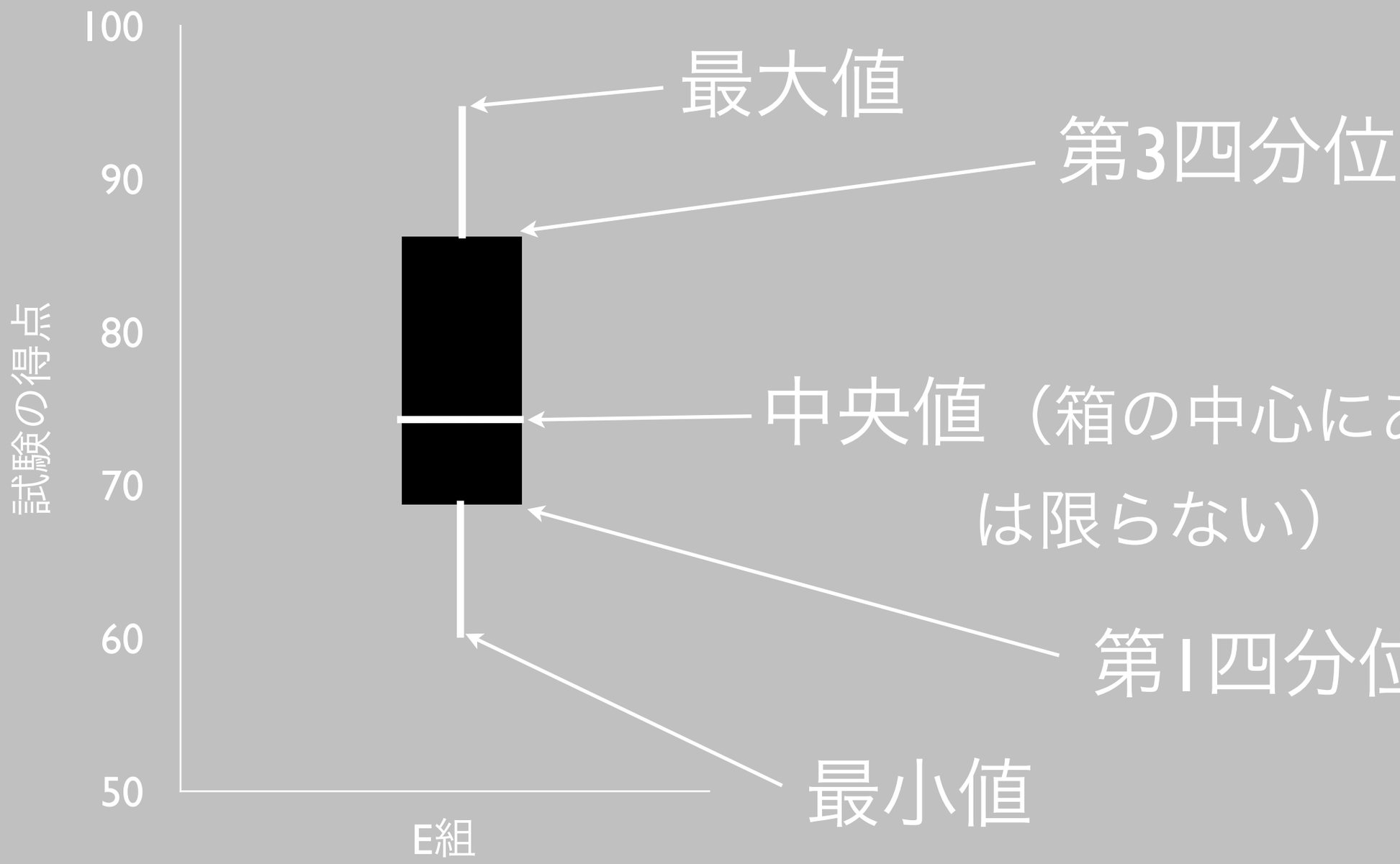
表：E組とF組の得点の五数要約

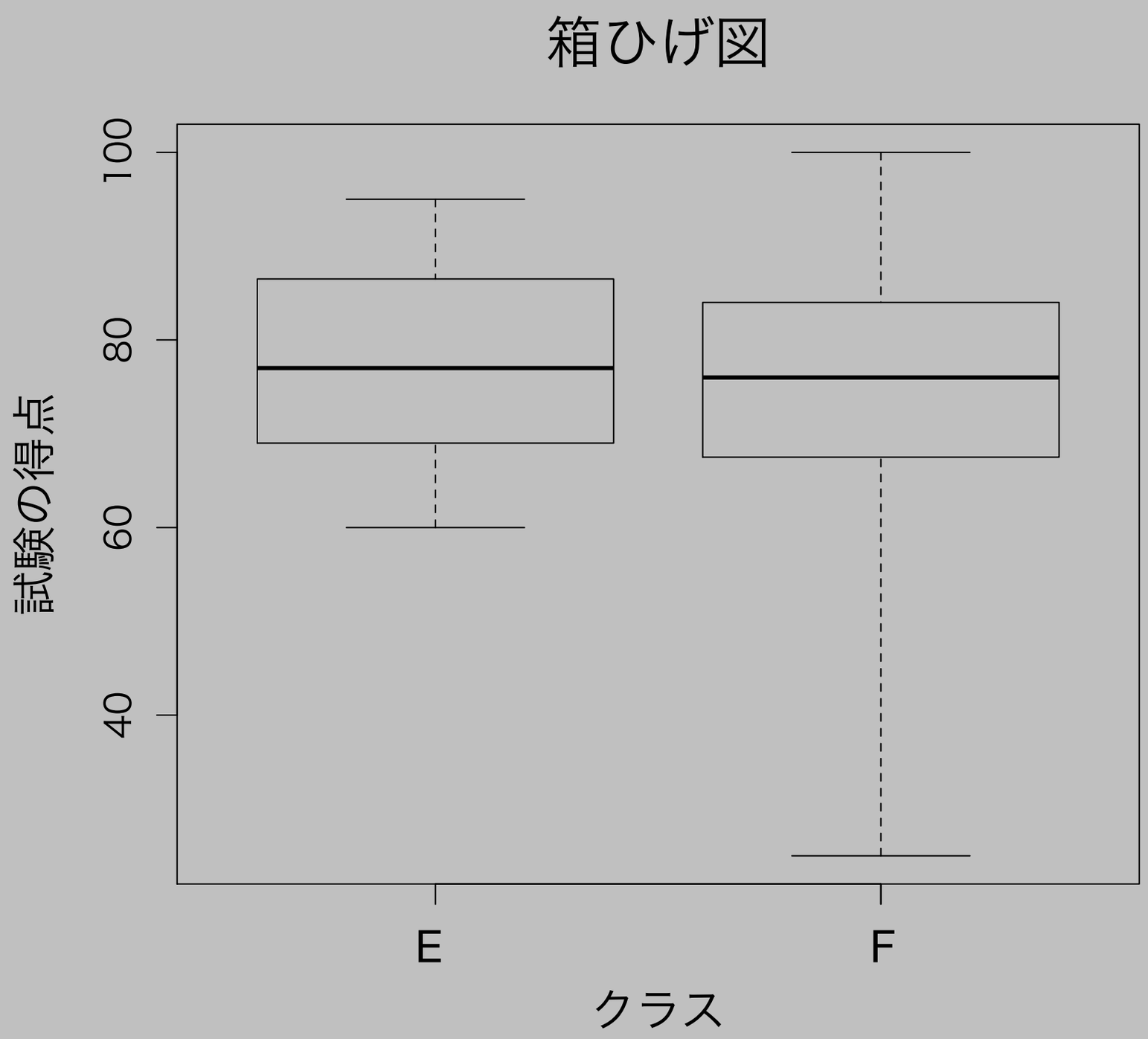
	最小値	第1四分位	中央値	第3四分位	最大値
E組	60	69	77	86.5	95
F組	25	67.5	76	84	100

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





分散 (variance)

- x の分散 $\mathrm{Var}(x)$ は x のばらつきを表す統計量：

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

- つまり、分散は「偏差の二乗（平方）」の平均値
  - [平均まわりの] 偏差 (deviation [from the mean]) ：$x_i - \bar{x}$
  - 偏差平方和：$\sum_{i=1}^{N}(x_i - \bar{x})^2$
- **統計学で最も重要な統計量**
- R で分散を求める：`var(x)`

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（長さ）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

➡ 長さのデータのばらつきを面積で表現されても意味がつかみにくい

✦意味のわからない単位になってしまうことも（例：年収を円で測定 → 円の二乗??）

標準偏差 (standard deviation)

- x の標準偏差 $\mathrm{SD}(x)$ は x のばらつきを表す統計量：

$$\mathrm{SD}(x) = \sqrt{\mathrm{Var}(x)} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- つまり、標準偏差は分散の平方根
- 分散とは異なり、元の変数の単位でばらつきを捉える
- R で標準偏差を求める：sd(x)

標準偏差の特徴

- 平均まわりのばらつきを測る

➡ 平均が代表値としてふさわしくない場合、標準偏差も使うべきではない（例：双峰型の分布）

- すべてのx の値が同じときのみ0となり、それ以外は正の値をとる
- 散らばりが大きいほど標準偏差が大きくなる
- 外れ値の影響を受ける

標準偏差でわかること

- それぞれの値が平均値から標準偏差で何個分離れているかを見る

➡ それぞれの値が「普通」か「特殊」かがある程度見分けられる

- 大まかな目安：標準偏差1個分以内は「普通」、2個分以上は「特殊」

統計量をRで求める方法についての

詳細は web資料を参照！

デモンストレーション

- RStudio のプロジェクト機能を使う！

データの可視化

データの可視化

- データ（分析結果）を**正確に、分かり易く**示す
  - データを要約する
  - 直感に訴える（誤解を生じさせないよう要注意）

- よくある**間違い**
  - データの一部を強調して相手を「説得する」
    - 科学の作法ではない：統計学の「誤用」「悪用」の一種

例：身長のデータ

- 20代の女性40人のデータ（架空）
- データの特徴は？

  データを見ただけではわかりにくい

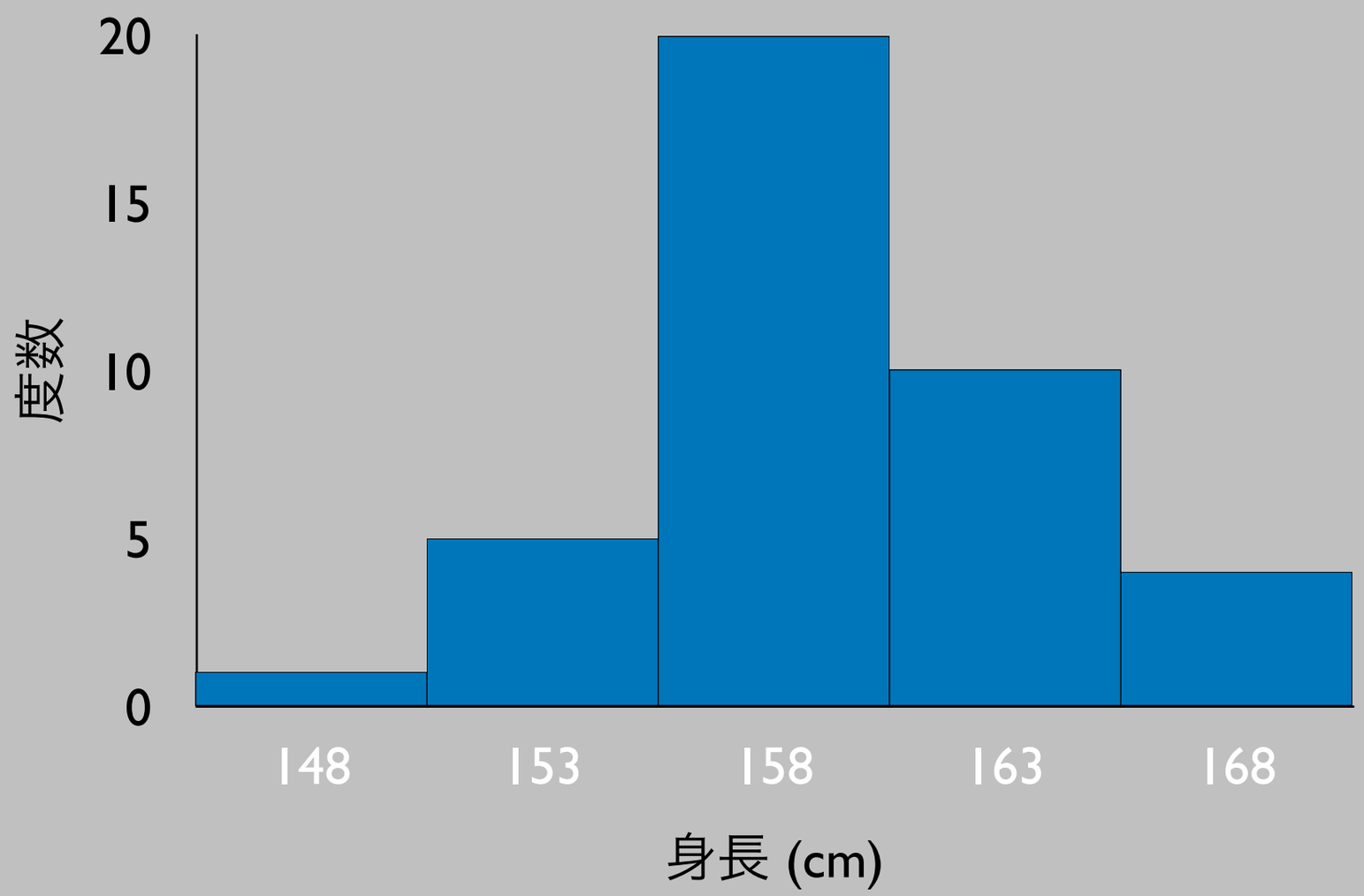
166	155	156.9	155.4
161.9	157.3	161.4	156.7
156.9	153.4	168.8	157.7
158	166.7	155.8	160.7
157.6	155.8	162	161.9
156.6	156	147.8	156.5
155.6	167.1	162.4	160.4
163.9	159.2	163.8	157.5
156.4	161.9	160.5	160
150.7	154.5	159.8	159.1

単位：cm

度数分布表とヒストグラム

- 各階級値の上に、対応する度数の高さの棒を立てる
  （棒の幅 = 階級の幅）

階級値	度数
148	1
153	5
158	20
163	10
168	4



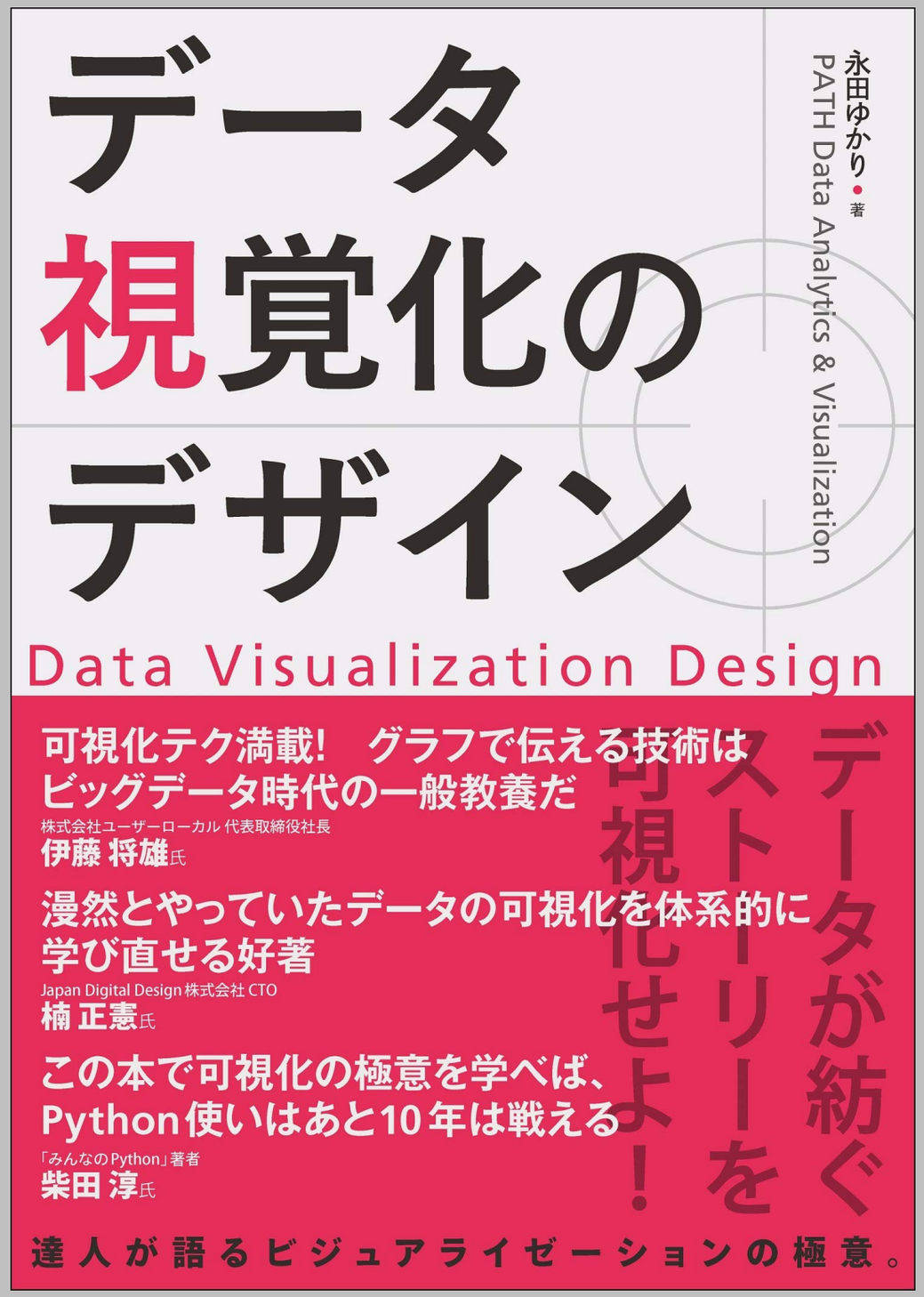
とにかく

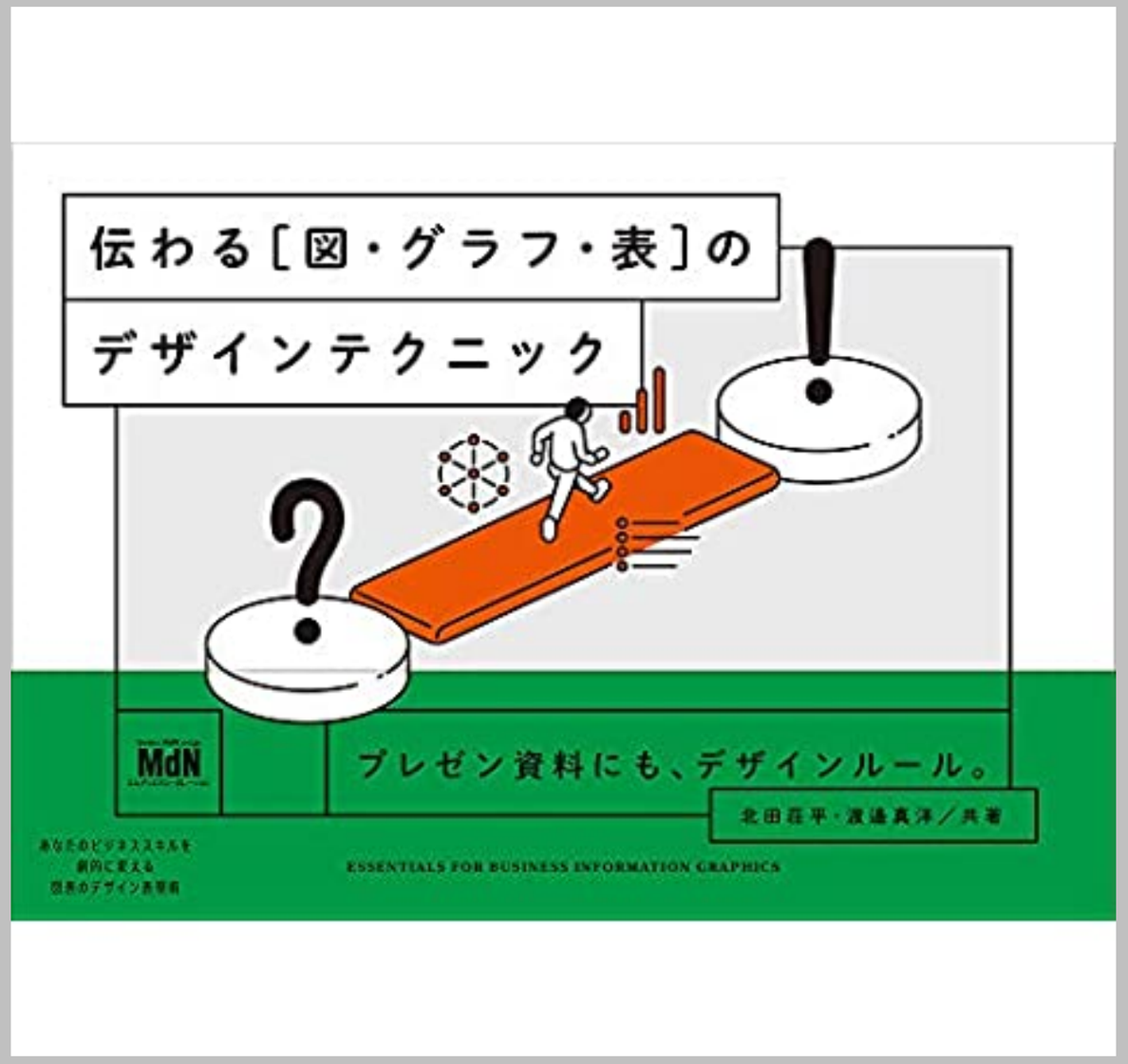
可視化すればいい？？？

スライド40枚をカット

受講生はKUTLMS にアップロードされた完全版を見るように！

参考文献


永田ゆかり・著
PATH Data Analytics & Visualization
データ
視覚化の
デザイン
Data Visualization Design
データが紡ぐストーリーを可視化せよ！
可視化テク満載！　グラフで伝える技術は
ビッグデータ時代の一般教養だ
株式会社ユーザーローカル 代表取締役社長
伊藤 将雄氏
漫然とやっていたデータの可視化を体系的に
学び直せる好著
Japan Digital Design株式会社 CTO
楠 正憲氏
この本で可視化の極意を学べば、
Python使いはあと10年は戦える
「みんなのPython」著者
柴田 淳氏
達人が語るビジュアライゼーションの極意。


伝わる［図・グラフ・表］の
デザインテクニック
MdN
プレゼン資料にも、デザインルール。
北田暁平・渡邉真洋／共著
ESSENTIALS FOR BUSINESS INFORMATION GRAPHICS

Topic 3の課題

- 提出が必要な課題はなし
- 以下のページの内容を自分で実行し、内容を理解する
  - http://yukiyanai.github.io/jp/classes/stat2/contents/R/introduction-to-RStudio.html

次回予告

4. R Markdown によるレポート作成