記述統計の「なぜ？」を統計的推測により理解する
～四分位数を例として～

酒折文武，早崎久登
2019 年 3 月 2 日
中央大学理工学部数学科，中央大学大学院理工学研究科数学専攻

Contents

はじめに

記述統計と統計的推測の接点

四分位数の定義

標本四分位数の評価

まとめ

はじめに

「四分位数」等の扱い

中等教育における四分位数，四分位範囲，箱ひげ図の扱い

- 現行の学習指導要領：高等学校 数学 I
- 新学習指導要領：中学 2 年

現場での混乱

- 四分位数の計算結果が Excel と異なる！
- コンピュータの利用が推進されているのに…どうしたらよいのか？

「四分位数」等の扱い

中等教育における四分位数，四分位範囲，箱ひげ図の扱い

- 現行の学習指導要領：高等学校 数学 I
- 新学習指導要領：中学 2 年

現場での混乱

- 四分位数の計算結果が Excel と異なる！
- コンピュータの利用が推進されているのに…どうしたらよいのか？

さまざまな指摘・意見

何森 (2013)

- 現場での混乱を避けるために固定の定義を押し付けるべきではない
- 入試などに四分位数は適さない

⇒ 四分位数や四分位範囲を読み取っての判断は可能であり，あくまでも四分位数の計算を主に問うべきではないというだけ

小林 (2013)

- 四分位数や箱ひげ図を扱わないほうがよい

⇒ 学習指導要領の意図を理解しようとしない暴論

新学習指導要領では

新学習指導要領解説 (2017)

- 四分位数を求める方法として幾つかの方法が提案されているが，ここでは四分位数の意味を把握しやすい方法を用いる

西仲 (2018)

- 中学での導入に際して，再度この点に警鐘を鳴らしつつ，解説の記述の不十分さを指摘

奥村 (2019)

- Tukey の定義，Excel で使える 2 通りの定義，統計ソフト R で使える 9 通りの定義のどれでもない，おそらく文科省独自のもの
  ⇒ 最後の点は後述の通り誤り
- 一つを選ぶのであれば Tukey の定義にすべきであった

新学習指導要領では

新学習指導要領解説 (2017)

- 四分位数を求める方法として幾つかの方法が提案されているが，ここでは四分位数の意味を把握しやすい方法を用いる

西仲 (2018)

- 中学での導入に際して，再度この点に警鐘を鳴らしつつ，解説の記述の不十分さを指摘

奥村 (2019)

- Tukey の定義，Excel で使える 2 通りの定義，統計ソフト R で使える 9 通りの定義のどれでもない，おそらく文科省独自のもの
  ⇒ 最後の点は後述の通り誤り
- 一つを選ぶのであれば Tukey の定義にすべきであった

本発表の目的

現場の混乱が起きないよう，内容を整理しておきたい！ サーベイ

- なぜ四分位数の定義が複数あるのか
- 中学校・高等学校で扱っているもの以外にどのようなものがあるか，どのような違いがあるか

安心して四分位数を使ってほしい！ 研究

- 他の四分位数と比べて計算しやすいメリットはあるが，他にメリット，デメリットはないのか
- その他気をつけるべき点はないか

記述統計と統計的推測の接点

記述統計と統計的推測

記述統計

- 得られたデータを要約する方法
  ex. 平均，中央値，最頻値，標準偏差，分散，ヒストグラム，散布図，相関係数，...

統計的推測（推測統計）

- 標本から母集団を推測する方法
  ex. 点推定，区間推定，仮説検定

一般的には 記述統計 ⇒ 統計的推測 と学ぶが，
記述統計を深く理解するためには統計的推測の考えかたが重要！

平均，分散

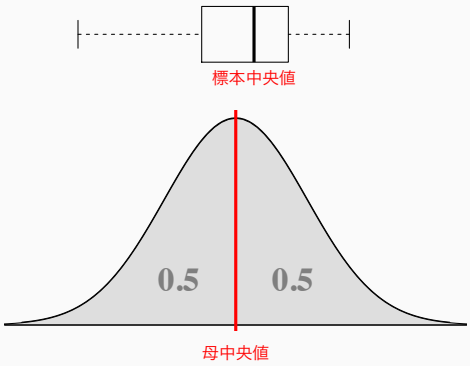
平均（標本平均） $\bar{X} = \frac{1}{n} X_i$

- 母集団の平均（母平均）μ を推定するもの（推定量）
- 言い換えれば，母平均を推定するために標本平均を求める

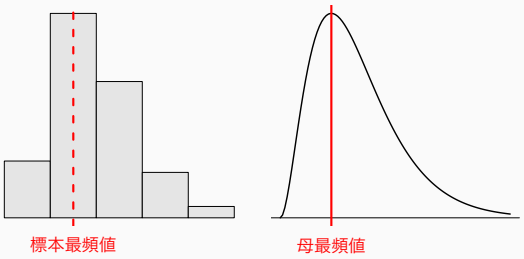
分散（標本分散） $S^2 = \frac{1}{n}(X_i - \bar{X})^2$

- 母集団の分散（母分散）σ^2 を推定するもの（推定量）
- 言い換えれば，母分散を推定するために標本分散を求める

母中央値と標本中央値



母最頻値と標本最頻値



そもそもヒストグラム自体も母集団分布を推測したもの！

数理統計学では

- あるパラメータを推定するための推定量は無数にある
- そのうちどれを用いるとよいかを様々な指標で評価する

母平均 $\mu = E(X)$ の推定量の代表例

- 標本平均 $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ 不偏性 一致性 有効性
- トリム平均 ロバスト性

母分散 $\sigma^2 = V(X) = E[(X-\mu)^2]$ の推定量の代表例

- 標本分散 $S^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 一致性
- 不偏分散 $U^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 不偏性 一致性 UMVU

以上を踏まえて

中央値，最頻値，四分位数なども母集団での値を推定するもの

- 推定量が複数存在するのは自然
- どれを用いるべきか，数理統計学的に優劣を判断できる
- 中等教育ではとくに使いやすさという観点も重要

⇒ 中学 2 年の段階では母集団と標本について未学習

- 誤差を伴うことの理解は難しい
- 高等学校・数学 I の仮説検定での理解が大事

1. 「この計算式が四分位数である」というような固定の定義を押し付けるやり方は避けるべき
2. 「集団を四等分する点を四分位数と呼ぶ．それをデータから計算する方法はいくつかあるが，ここでは簡易的なこれを用いる．こんなことが読み取れるよ」という指導をすべき

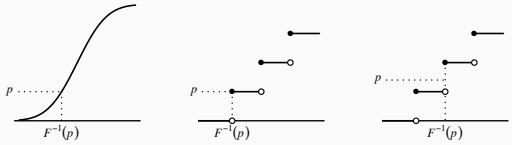
四分位数の定義

母分位数

X: 累積分布関数 $F(x) = P(X \leq x)$ をもつ確率変数

p 分位数（下側 $100p$ パーセント点）

$$Q(p) = F^{-1}(p) = \inf\{x : F(x) \geq p\}, \quad 0 \leq p \leq 1$$



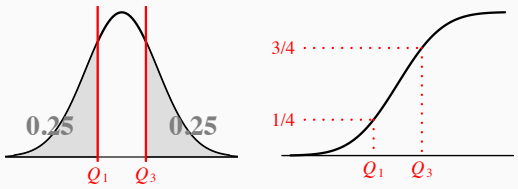
正整数 m に対し，$i = 1, \ldots, m-1$ に対する i/m 分位数を，
m を漢数字で表記して第 i m 分位数ともいう

母四分位数

四分位数

- 第 1 四分位数 $Q_1 = Q(1/4)$
- 第 2 四分位数 $Q_2 = Q(2/4) = Q(1/2) =$ 中央値
- 第 3 四分位数 $Q_3 = Q(3/4)$

母集団分布 $F(x)$ によって一意に定まる



なぜ四分位数を考えるのか

パラメトリックな分析

- 母集団分布に正規分布など特定の分布を仮定
- 標本から，平均，分散などのパラメータを推測

ノンパラメトリックな分析

- 母集団分布に仮定を置かない
- 分布の形や特性を様々な方法で知る
  - 中央値：分布の中央の位置
  - 最頻値：分布のピーク
  - 分位数，四分位数：分布の形やボリュームゾーン
  - 密度推定：分布の形

標本分位数

- $X_1, \ldots, X_n$: $F(x)$ からの無作為標本

$\Rightarrow$ $X_{(1)}, \ldots, X_{(n)}$: $X_1, \ldots, X_n$ の**順序統計量**
$(X_{(1)} \leq \cdots \leq X_{(n)})$

標本 p 分位数

- 主に順序統計量を用いて母分位数 $Q(p)$ を推定
- 多くのものは以下の形

$$\hat{Q}(p) = (1-\varepsilon)X_{(j)} + \epsilon X_{(j+1)}$$

隣接する 2 つの順序統計量の線形補間

Hyndman & Fan (1996): 9 種類の推定量を 6 つの観点から比較

さまざまな標本四分位数

Langford (2006): 15 種類の推定量

- 標本 p 分位数の $p = 1/4,\ 3/4$ の場合を用いることが可能

累積分布関数に基づく方法

経験分布関数を累積分布関数と思って，母四分位数の要領で標本四分位数標本四分位数 $\hat{Q}_1, \hat{Q}_3$ を定める方法.
$n = 4j + r,\ r = 0, 1, 2, 3$ のとき

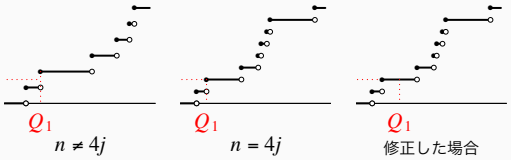
$$\hat{Q}_1 = X_{(\lceil n/4 \rceil)} = \begin{cases} X_{(j)} & r = 0 \\ X_{(j+1)} & r = 1, 2, 3 \end{cases}$$
$$\hat{Q}_3 = X_{(\lceil 3n/4 \rceil)} = X_{(3j+r)}$$

さまざまな標本四分位数

累積分布関数に基づく方法の改良（CDF に基づく推定量）

$n = 4j$ のときに中央に修正.

$$\hat{Q}_1 = \begin{cases} \frac{1}{2}X_{(j)} + \frac{1}{2}X_{(j+1)} & r = 0 \\ X_{(j+1)} & r = 1, 2, 3 \end{cases}$$
$$\hat{Q}_3 = \begin{cases} \frac{1}{2}X_{(3j)} + \frac{1}{2}X_{(3j+1)} & r = 0 \\ X_{(3j+r)} & r = 1, 2, 3 \end{cases}$$



さまざまな標本四分位数

Mathematica の推定量

$p = 1/4,\ 3/4$ に対して「$np + 0.5$ 番目」の順序統計量とする方法.

【例】データが $\{1, 2, 3, 4, 5\}$ のとき，$5 \times 1/4 + 0.5 = 1.75$ より

$$\hat{Q}_1 = \frac{1}{4}X_{(1)} + \frac{3}{4}X_{(2)} = 1.75$$

Excel の `QUARTILE.EXC` 関数や MINITAB の推定量

「$(n+1)p$ 番目」の順序統計量とする方法.

Excel の `QUARTILE.INC` 関数や R のデフォルトの推定量

「$(n-1)p + 1$ 番目」の順序統計量とする方法.

さまざまな標本四分位数

Excel の QUARTILE.EXC 関数や MINITAB の推定量

$n = 4j + r,\ r = 0, 1, 2, 3$ のとき

$$\begin{aligned}\hat{Q}_1 &= (1-\varepsilon)X_{(\lceil (n+1)/4 \rceil)} + \varepsilon X_{(\lceil (n+1)/4 \rceil + 1)} \\ &= \begin{cases} (1-\varepsilon)X_{(j)} + \varepsilon X_{(j+1)} & r = 0, 1, 2 \\ (1-\varepsilon)X_{(j+1)} + \varepsilon X_{(j+2)} & r = 3 \end{cases} \\ \hat{Q}_3 &= (1-\delta)X_{(\lceil 3(n+1)/4 \rceil)} + \delta X_{(\lceil 3(n+1)/4 \rceil + 1)} \\ &= (1-\delta)X_{(3j+r)} + \delta X_{(3j+r+1)}\end{aligned}$$

ここで ε, δ は $\lceil\ \rceil$ 内の数の小数部:

$$\varepsilon = \frac{n+1}{4} - \left\lceil \frac{n+1}{4} \right\rceil,$$
$$\delta = \frac{3(n+1)}{4} - \left\lceil \frac{3(n+1)}{4} \right\rceil$$

さまざまな標本四分位数

Tukey のヒンジ

- 標本中央値を含む前半のデータの中央値，後半のデータの中央値
- 母四分位数の推定量として用いることもある（inclusive な推定量）

$X_{(1)}\ X_{(2)}\ X_{(3)}\ X_{(4)}\ X_{(5)}\ X_{(6)}\ X_{(7)}\ X_{(8)}\ X_{(9)}\ X_{(10)}\ X_{(11)}\ X_{(12)}\ X_{(13)}$

さまざまな標本四分位数

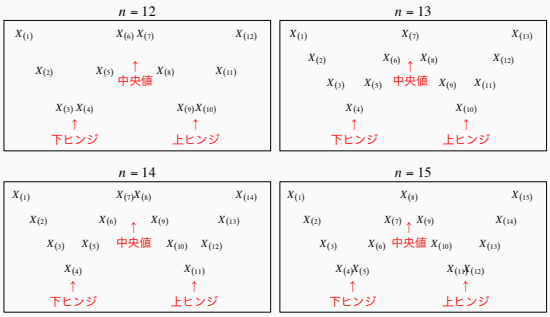
Tukey のヒンジ

- 標本中央値を含む前半のデータの中央値，後半のデータの中央値
- 母四分位数の推定量として用いることもある（inclusive な推定量）

$X_{(1)}\ X_{(2)}\ X_{(3)}\ X_{(4)}\ X_{(5)}\ X_{(6)}\ X_{(7)}\ X_{(8)}\ X_{(9)}\ X_{(10)}\ X_{(11)}\ X_{(12)}\ X_{(13)}$

$X_{(1)}$ $X_{(7)}$ $X_{(13)}$
$X_{(2)}$ $X_{(6)}$ $X_{(8)}$ $X_{(12)}$
$X_{(3)}$ $X_{(5)}$ 中央値 $X_{(9)}$ $X_{(11)}$
$X_{(4)}$ $X_{(10)}$
↑ ↑
下ヒンジ 上ヒンジ

さまざまな標本四分位数



さまざまな標本四分位数

高等学校の四分位数

- 標本中央値を含まない前半のデータの中央値，後半のデータの中央値
- exclusive な推定量

$n = 4j + r,\ r = 0, 1, 2, 3$ に対して

$$\hat{Q}_1 = \begin{cases} \frac{1}{2}X_{(j)} + \frac{1}{2}X_{(j+1)} & r = 0, 1 \\ X_{(j+1)} & r = 2, 3 \end{cases}$$
$$\hat{Q}_3 = \begin{cases} \frac{1}{2}X_{(3j+r)} + \frac{1}{2}X_{(3j+r+1)} & r = 0, 1 \\ X_{(3j+r)} & r = 2, 3 \end{cases}$$

さまざまな標本四分位数

	各データセットの $(\hat{Q}_1,\ \hat{Q}_3)$			
推定量	$\{1, 2, 3, 4\}$	$\{1, 2, 3, 4, 5\}$	$\{1, 2, \ldots, 6\}$	$\{1, 2, \ldots, 7\}$
inclusive	$(1.5,\ 3.5)$	$(2,\ 4)$	$(2,\ 5)$	$(2.5,\ 5.5)$
exclusive	$(1.5,\ 3.5)$	$(1.5,\ 4.5)$	$(2,\ 5)$	$(2,\ 6)$
CDF	$(1.5,\ 3.5)$	$(2,\ 4)$	$(2,\ 5)$	$(2,\ 6)$
Mathematica	$(1.5,\ 3.5)$	$(1.75,\ 4.25)$	$(2,\ 5)$	$(2.25,\ 5.75)$
QUARTILE.EXC	$(1.25,\ 3.75)$	$(1.5,\ 4.5)$	$(1.75,\ 5.25)$	$(2,\ 6)$
R のデフォルト	$(1.75,\ 3.25)$	$(2,\ 4)$	$(2.25,\ 4.75)$	$(2.5,\ 5.5)$

標本四分位数の評価

数値実験 1

- 正規分布 $N(50, 10^2)$ から大きさ $n = 8, 9, \ldots, 20, 30, 40$ のデータセットを発生（1,000,000 回）
- 母四分位数 $Q_1 \approx 43.2551, Q_3 \approx 56.7449$ からの標本四分位数の平均的なズレ（バイアス）を計算

$$\mathrm{Bias}(\hat{Q}_j) = E(\hat{Q}_j) - Q_j$$

- （バイアスがない推定量を不偏推定量という）

数値実験 1 の結果（抜粋）

第 1 四分位数

n	8	9	10	11	12	13	14	15	16
inclusive	0.12	1.02	0.19	0.80	0.10	0.72	0.13	0.60	0.08
exclusive	0.12	−0.78	0.19	−0.54	0.10	−0.52	0.13	−0.41	0.08
CDF	0.12	1.03	0.19	−0.54	0.10	0.72	0.13	−0.41	0.08
Mathematica	0.12	0.12	0.19	0.13	0.10	0.13	0.09	0.08	0.05
QUARTILE.EXC	−0.83	−0.78	−0.68	−0.54	−0.54	−0.52	−0.47	−0.41	−0.41
R	1.07	1.03	0.88	0.80	0.74	0.72	0.64	0.60	0.56

- n が大きくなるに従ってバイアスは減少傾向
- inclusive，Mathematica，R は第 1 四分位数を過大推定
- `QUARTILE.EXC` は，第 1 四分位数を過小推定
- exclusive，CDF は n の値によってバイアスの方向が反転する
- inclusive，exclusive は n が奇数のときのバイアスが大きい
- Mathematica はバイアスがかなり小さいが，$n = 4j + 2$ のときにバイアスが若干大きい

数値実験 1 の結果（抜粋）

第 3 四分位数

n	8	9	10	11	12	13	14	15	16
inclusive	−0.12	−1.02	−0.18	−0.79	−0.09	−0.71	−0.12	−0.59	−0.07
exclusive	−0.12	0.78	−0.18	0.55	−0.09	0.52	−0.12	0.41	−0.07
CDF	−0.12	−1.02	−0.18	0.55	−0.09	−0.71	−0.12	0.41	−0.07
Mathematica	−0.12	−0.12	−0.18	−0.12	−0.09	−0.10	−0.12	−0.09	−0.07
QUARTILE.EXC	0.83	0.78	0.68	0.55	0.55	0.52	0.47	0.41	0.41
R	−1.07	−1.02	−0.88	−0.79	−0.73	−0.71	−0.64	−0.59	−0.56

- 全体的な傾向は第 1 四分位数とほぼ同じ
- バイアスの符号がすべて反転
- n が大きくなるに従ってバイアスは減少傾向

数値実験 2

- 正規分布 $N(50, 10^2)$ および一様分布 $U(0, 100)$ から大きさ $n = 10$ のデータセットを発生（1,000,000 回）
- 母四分位数 $Q_1 \approx 43.2551$ からの標本四分位数のバイアス，標準誤差，平方根平均二乗誤差を計算

	$N(50, 10^2)$			$U(0, 100)$		
	Bias	SE	RMSE	Bias	SE	RMSE
inclusive	0.188	4.185	4.189	2.258	12.847	13.040
exclusive	0.188	4.185	4.189	2.258	12.847	13.040
CDF	0.188	4.185	4.189	2.258	12.847	13.040
Mathematica	0.188	4.185	4.189	2.258	12.847	13.040
QUARTILE.EXC	−0.677	4.085	4.141	−0.015	11.906	11.906
R	0.889	3.982	4.080	4.534	12.605	13.395

- 標準誤差や RMSE はどの方法でもあまり変わりがない
- 正規分布と一様分布で全く異なる傾向

漸近的な評価

Mudholkar and Huston (1997): `QUARTILE.EXC` の漸近バイアス等の導出

漸近バイアス

$$\mathrm{Bias}(\hat{Q}_1) = \frac{3Q_1''}{32n} + \frac{B_1(\varepsilon)}{2048n^2} + O(n^{-3}),$$
$$\mathrm{Bias}(\hat{Q}_3) = \frac{3Q_3''}{32n} + \frac{B_3(\delta)}{2048n^2} + O(n^{-3})$$

ここで

$$B_1(\varepsilon) = 128\left[8\varepsilon(1-\varepsilon) - 3\right]Q_1'' + 64Q_1''' + 9Q_1^{(4)},$$
$$B_3(\delta) = 128\left[8\delta(1-\delta) - 3\right]Q_3'' - 64Q_3''' + 9Q_3^{(4)}$$

- $\frac{1}{n}$ の項は Q'' のみに依存する
- 正規分布のとき $Q_1'' < 0, Q_3'' > 0$ よりバイアスもそれぞれ負，正
- 一様分布のとき，$Q_1'' = Q_3'' = 0$ より漸近バイアスは n^{-2} のオーダー

漸近的な評価

酒折・早崎（投稿準備中）：exclusive な推定量の漸近バイアス

漸近バイアス（$\hat{Q}_1$）

$$\mathrm{Bias}(\hat{Q}_1) = \begin{cases} \frac{8Q_1'+3Q_1''}{32n} + \frac{-512Q_1'+64Q_1''+112Q_1'''+9Q_1^{(4)}}{2048n^2} + O(n^{-3}) & n=4j \\ \frac{3Q_1''}{32n} + \frac{-128Q_1''+64Q_1'''+9Q_1^{(4)}}{2048n^2} + O(n^{-3}) & n=4j+1 \\ \frac{8Q_1'+3Q_1''}{32n} + \frac{-512Q_1'-192Q_1''+112Q_1'''+9Q_1^{(4)}}{2048n^2} + O(n^{-3}) & n=4j+2 \\ \frac{3Q_1''}{32n} + \frac{-384Q_1''+64Q_1'''+9Q_1^{(4)}}{2048n^2} + O(n^{-3}) & n=4j+3 \end{cases}$$

- $\frac{1}{n}$ の項は Q' と Q'' に依存する
- 正規分布のとき，$Q_1' > 0, Q_1'' < 0$ であり，n が奇数のときバイアスは消し合う方向
- 一様分布のとき，$Q_1' > 0, Q_1'' = 0$ より，Q' によりバイアスが増える

漸近的な評価

漸近バイアス（$\hat{Q}_3$）

$$\mathrm{Bias}(\hat{Q}_3) = \begin{cases} \frac{-8Q_3'+3Q_3''}{32n} + \frac{512Q_3'+64Q_3''-112Q_3'''+9Q_3^{(4)}}{2048n^2} + O(n^{-3}) & n=4j \\ \frac{3Q_3''}{32n} + \frac{-128Q_3''-64Q_3'''+9Q_3^{(4)}}{2048n^2} + O(n^{-3}) & n=4j+1 \\ \frac{-8Q_3'+3Q_3''}{32n} + \frac{512Q_3'-192Q_3''-112Q_3'''+9Q_3^{(4)}}{2048n^2} + O(n^{-3}) & n=4j+2 \\ \frac{3Q_3''}{32n} + \frac{-384Q_3''-64Q_3'''+9Q_3^{(4)}}{2048n^2} + O(n^{-3}) & n=4j+3 \end{cases}$$

第 1 四分位数と同様の傾向

ここからわかること

QUARTILE.EXC の推定量について

- 正規分布のとき Q_1 を過小推定，Q_3 を過大推定
  - ⇒ 四分位範囲を広く推定
- 一様分布のときバイアスは非常に小さい

exclusive な推定量について

- 正規分布のとき，n 偶数ならば，Q_1 を過大推定，Q_3 を過小推定
  - ⇒ 四分位範囲を狭く推定
  - QUARTILE.EXC よりバイアスの絶対値は小さい
- 正規分布のとき，n 奇数ならば，Q_1 を過小推定，Q_3 を過大推定
  - ⇒ 四分位範囲を広く推定
  - QUARTILE.EXC と全く同じ推定結果
- 一様分布のときバイアスは QUARTILE.EXC より大きい

まとめ

本研究のまとめ

現場の混乱が起きないよう，内容を整理しておきたい！ サーベイ

- なぜ四分位数の定義が複数あるのか
  - ⇒ 母四分位数の推定量がいくつも提案されているから．n で割る分散と $n-1$ で割る分散があるのと同じ．
- 中学校・高等学校で扱っているもの以外にどのようなものがあるか，どのような違いがあるか
  - ⇒ ヒンジや Excel の方法など紹介した．

安心して四分位数を使ってほしい！ 研究

- 他の四分位数と比べて計算しやすいメリットはあるが，他にメリット，デメリットはないのか
  - ⇒ バイアスの絶対値自体は Excel のものより小さいが，正負が交互に来ることには注意．
- その他気をつけるべき点はないか
  - ⇒ 四分位範囲を n が異なる群で比較するときに多少注意が必要．

とはいえ

- 大事なことは四分位数の計算ではない
- 大まかに分布の形状やボリュームゾーンなどを読み取るために使うわけであり，細かな数字の違いをどうこういうべきものではない

今後について

- 四分位数については一定の結論を得たと感じている（inclusive な方法などについても漸近バイアスは得ておきたいが）
- 他にも，理論的に整理しておくべき内容があるのではないか？ 統計家はもっとしっかりと関わるべきでは？
  - すでにいくつかは他の院生，学部生たちとチャレンジ中
  - 何か疑問に感じること，数学的に整理しておきたいことなどあればご意見ください！