データを科学する
～統計学的な見方•考え方～

田中 勝人(Tanaka Katsuto)
学習院大学経済学部教授，一橋大学名誉教授

松本深志高校にて

2019年10月12日

講 義 の 内 容

1. 統計学・・・文理融合型の学問
2. データの縮約・・・「平均のウソ」と「中央値の魔術」
3. 帰納的推論・・・「母集団」に思いを巡らす
4. 「中心極限定理」・・・その不思議さと美しさ
5. ビッグデータの時代・・・データは21世紀の石油

1．統計学・・・文理融合型の学問

「データ分析のための方法論的科学」

（1）データの収集（標本調査，実験計画）

（2）データの整理（記述統計：データそのものの分析）

（3）推論（推測統計：母集団特性を確率的に推論）

（4）決定と予測（今後の行動や意思の決定，将来予測）

統計学は，文理融合的，あるいは文理の枠を超えた学問

文理融合的な他の学問の例：

数理科学，情報科学（数学を応用した学問分野）

生命科学（生命を取り巻く関連諸科学）

行動科学（人間行動の科学的研究と法則性の解明）

統計学は，各分野で使われる

- 数理統計学　• 統計数学　• 医学統計
- 計量生物学　• 統計物理学　• 統計力学
- 情報理論　• 制御理論　• 信号処理
- 経済統計学　• 計量経済学　• 人口統計学
- 計量政治学　• 計量言語学　• 計量国語学
- 数量経済史　• 心理統計学　• スポーツ統計学
- • • • • •

これらを総称して，「統計科学」ということもある．

記 述 統 計

● もともとのデータは数字や文字の羅列

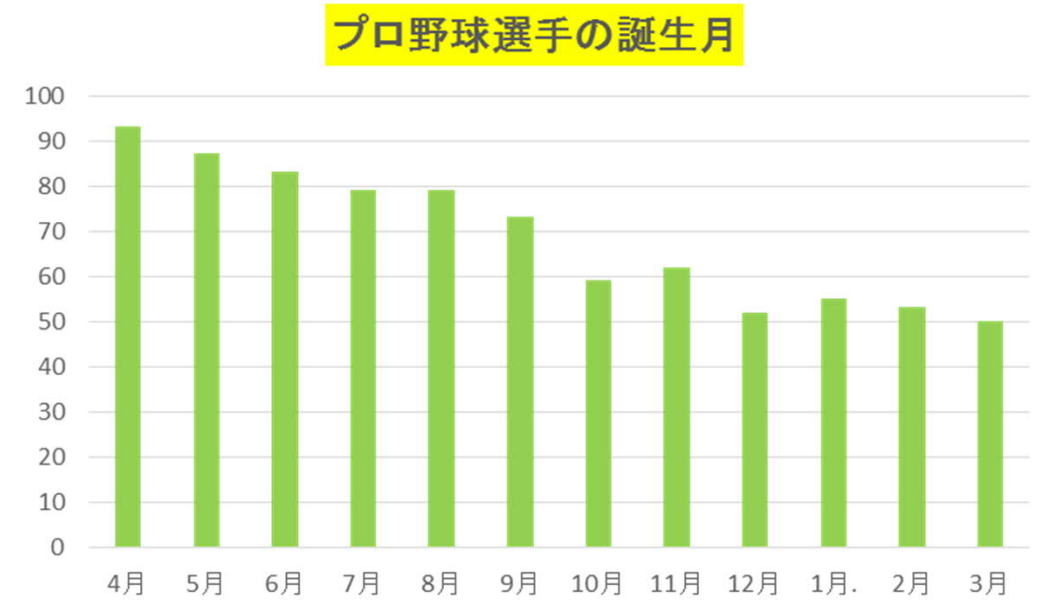
（GDP統計，労働力調査，貿易統計，消費者物価指数，
身長･体重のデータ，成績のデータ，生年月日のデータ，
ゲノムデータ，臨床データ，気象データ，経済データ，
選挙データ，文学作品，・・・）

ビッグデータ：非定型データ（従来の統計学の対象外）

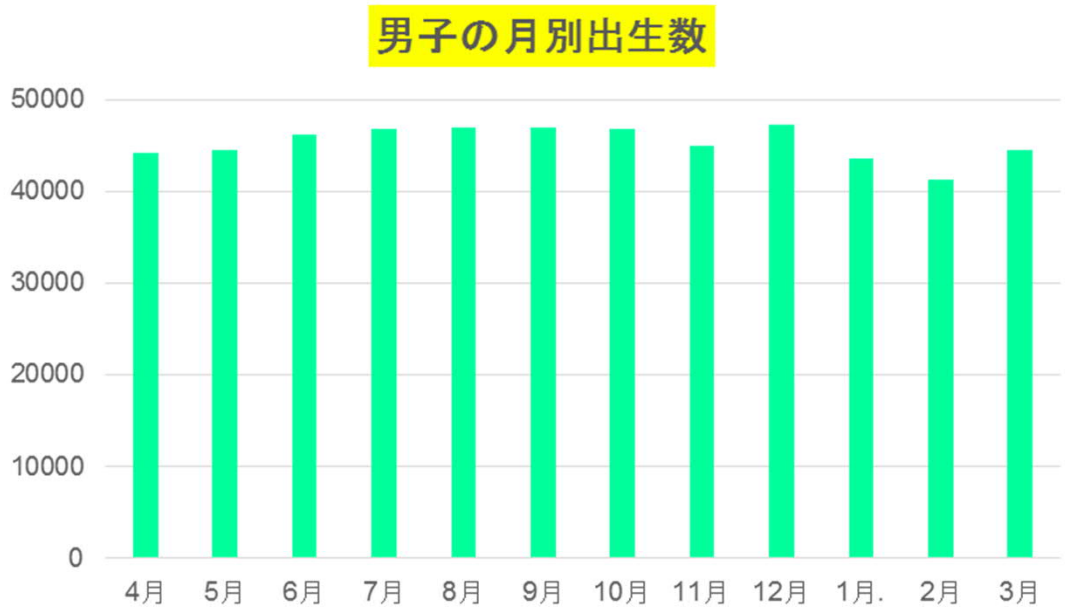
● データの特徴を視覚的•数量的に把握する

ヒストグラム，度数分布，箱ひげ図，散布図，・・・
平均，標準偏差，変動係数，相関係数，・・・

運動神経は誕生月と関係？（相対年齢効果）

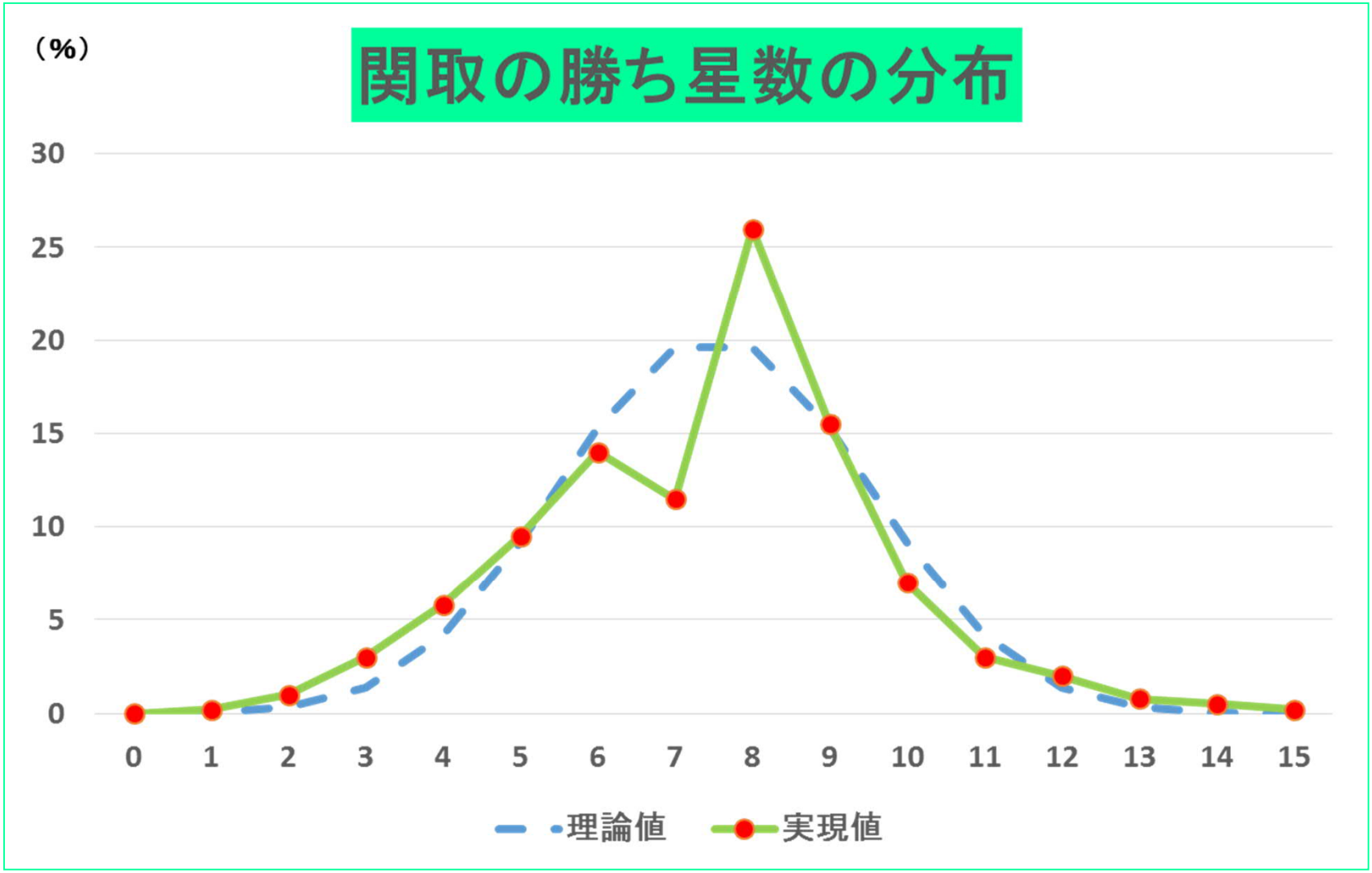


プロ野球日本人選手 825 人の
誕生月データ (2019年)



日本人男子 54 万人 の
月別出生数
（人口動態調査：2016年)

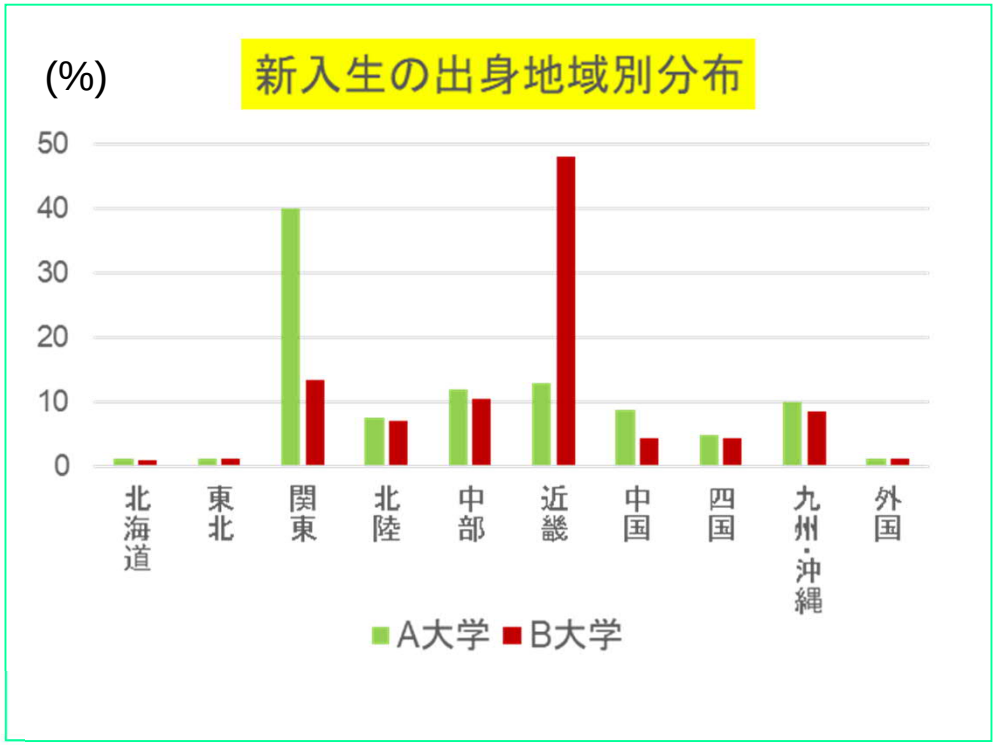
大 相 撲:八 百 長？


（%）
関取の勝ち星数の分布
30
25
20
15
10
5
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
理論値
実現値

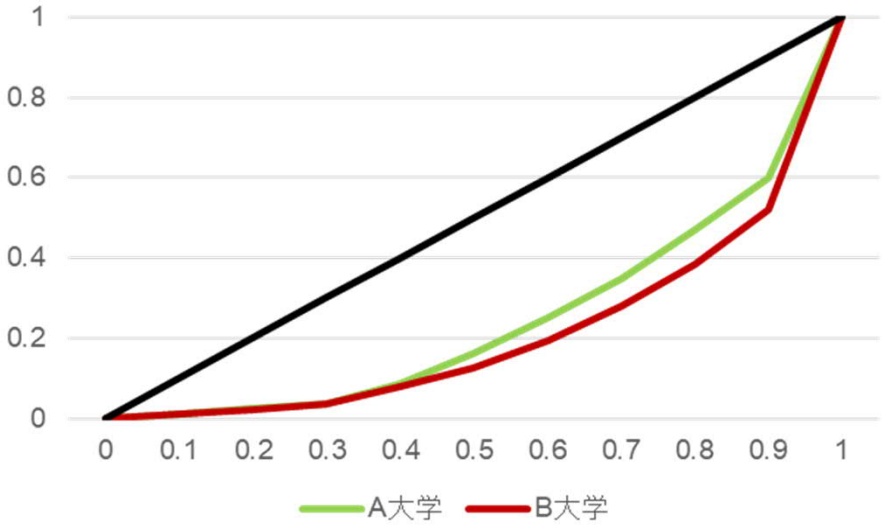
どちらの大学が，より全国型か？

大学新入生の出身地域別分布

	A 大学	B大学
北海道	50	100
東北	50	120
関東	1600	1350
北陸	300	700
中部	480	1050
近畿	520	4800
中国	350	450
四国	200	450
九州・沖縄	400	850
外国	50	130
合計	4000	10000



ローレンツ曲線

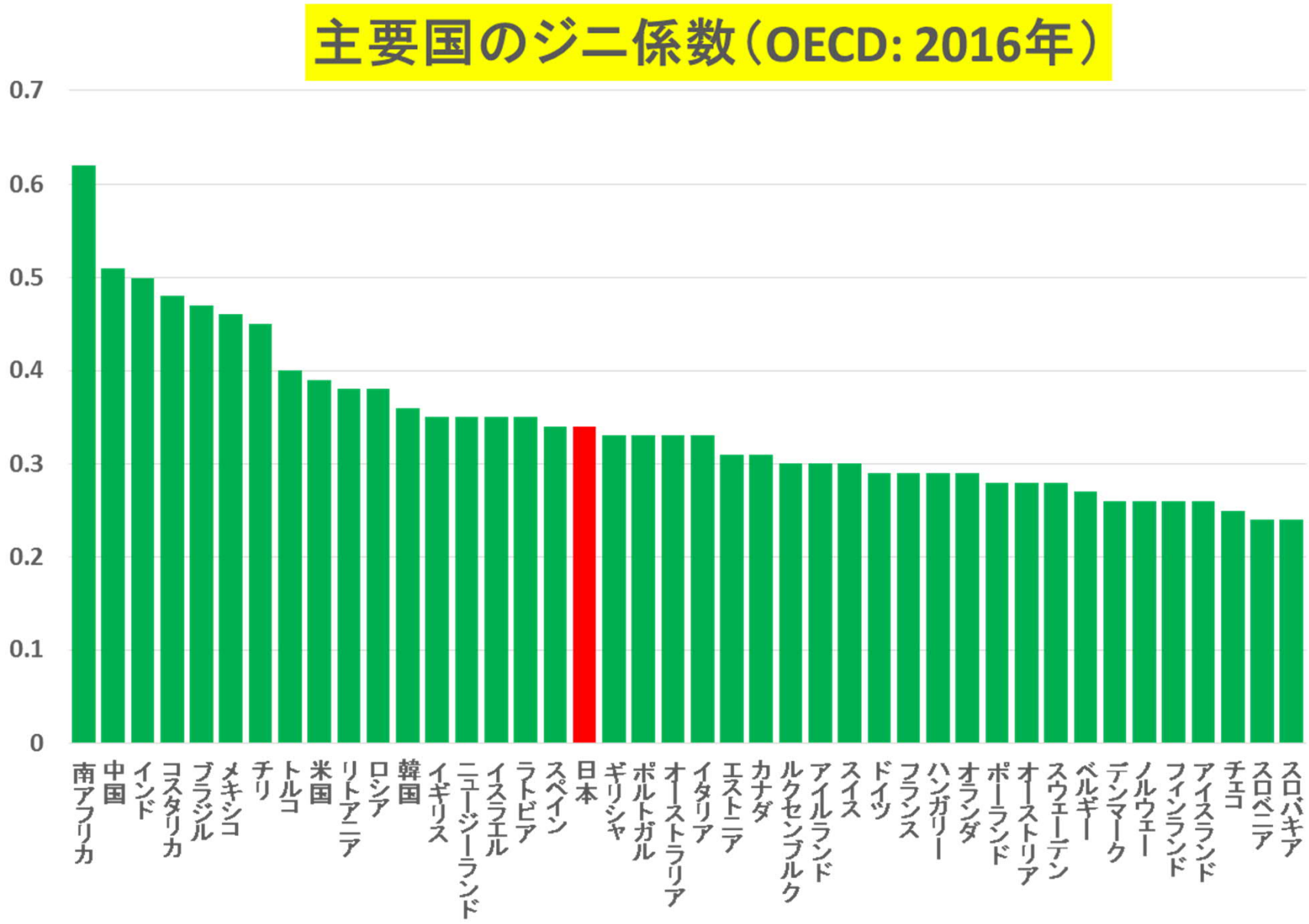


ジニ係数：45度線とローレンツ曲線で囲まれた弓型の面積の2倍
（0以上1以下の値．大きいほど，集中度が大きくなる）

A大学：0.50

B大学：0.57

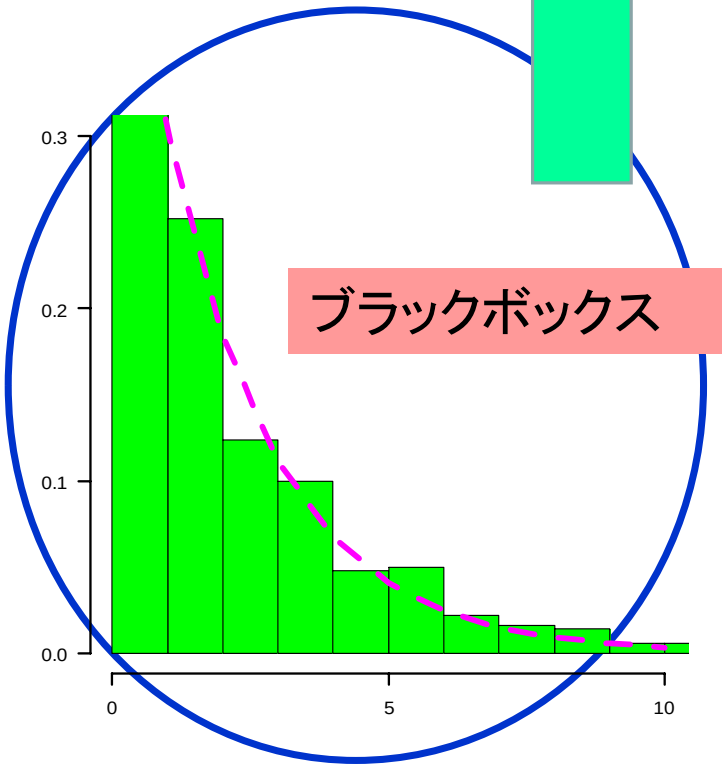
所 得 格 差 の 比 較


主要国のジニ係数（OECD: 2016年）
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
南アフリカ
中国
インド
コスタリカ
ブラジル
メキシコ
チリ
トルコ
米国
リトアニア
ロシア
韓国
イギリス
ニュージーランド
イスラエル
ラトビア
スペイン
日本
ギリシャ
ポルトガル
オーストラリア
イタリア
エストニア
カナダ
ルクセンブルク
アイルランド
スイス
ドイツ
フランス
ハンガリー
オランダ
ポーランド
オーストリア
スウェーデン
ベルギー
デンマーク
ノルウェー
フィンランド
アイスランド
チェコ
スロベニア
スロバキア

推 測 統 計

データ（母集団の一部分）から，母集団の特性を推論する
（推論の正しさを確率的に評価する）

母集団



$X_1, X_2, \cdots, X_n$

に基づいて，背後の母集団のなかみを推測する．

- 誤差を確率的に評価
- 仮説検定，因果関係の推論
- データの生成過程を確率モデルで表現

2. データの縮約・・・
「平均のウソ」と「中央値」の魔術

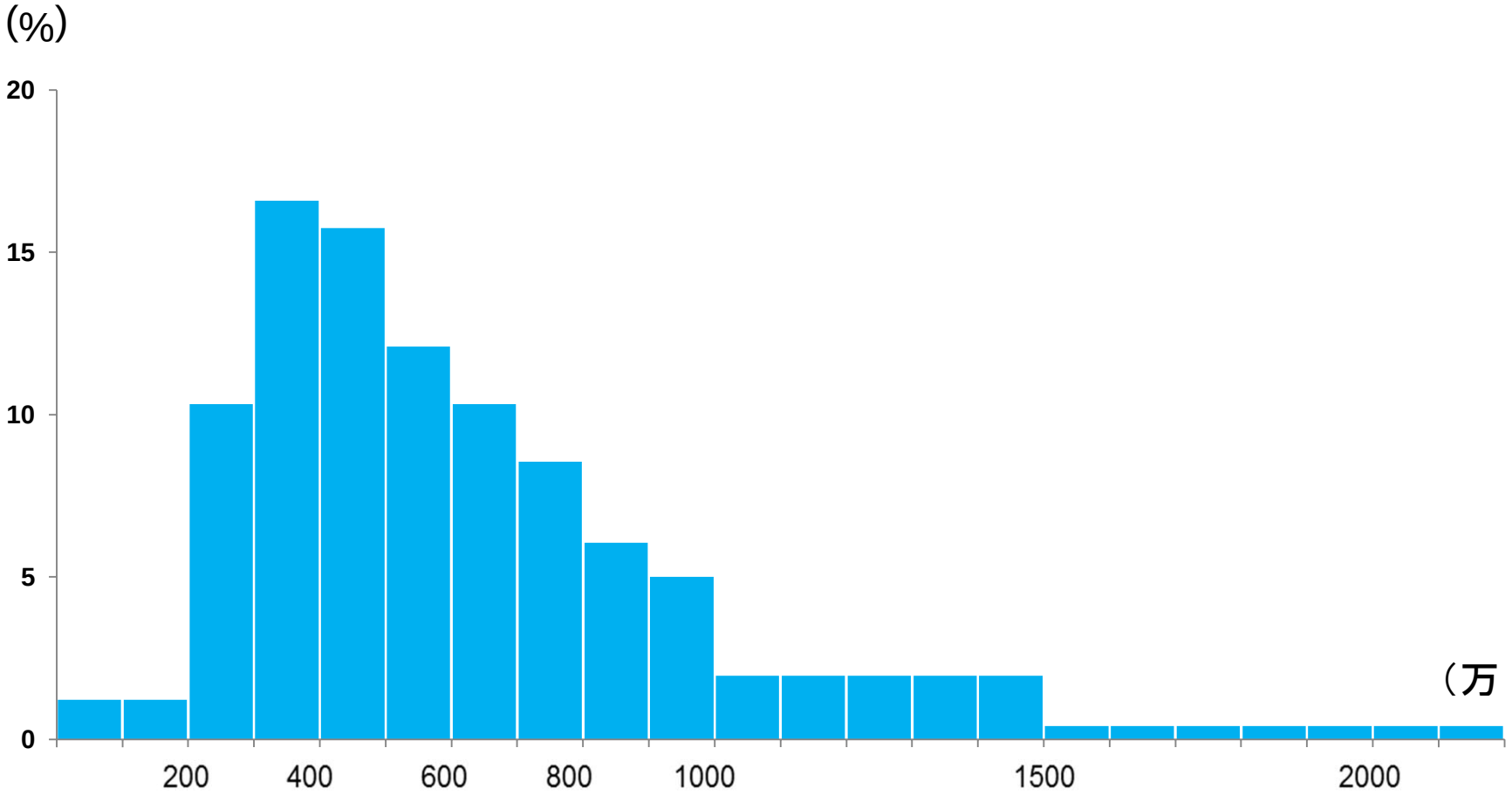
- 縮約: データをまとめる(次元を落とす)こと
- n 個のデータ $x_1, x_2, \cdots, x_n$ の平均は,

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

である.

問題：下のヒストグラムは，8000世帯の年間収入の分布を表している．平均収入は？

家計調査（総務省統計局）



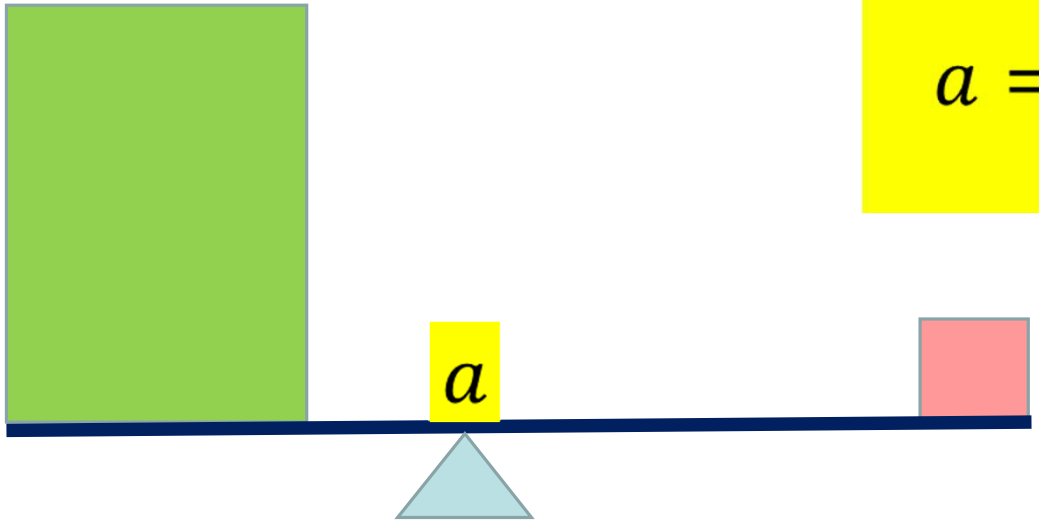
平均とは重心のことである.

$\sum_{i=1}^{n}(x_i - a) = 0$ をみたす a は,

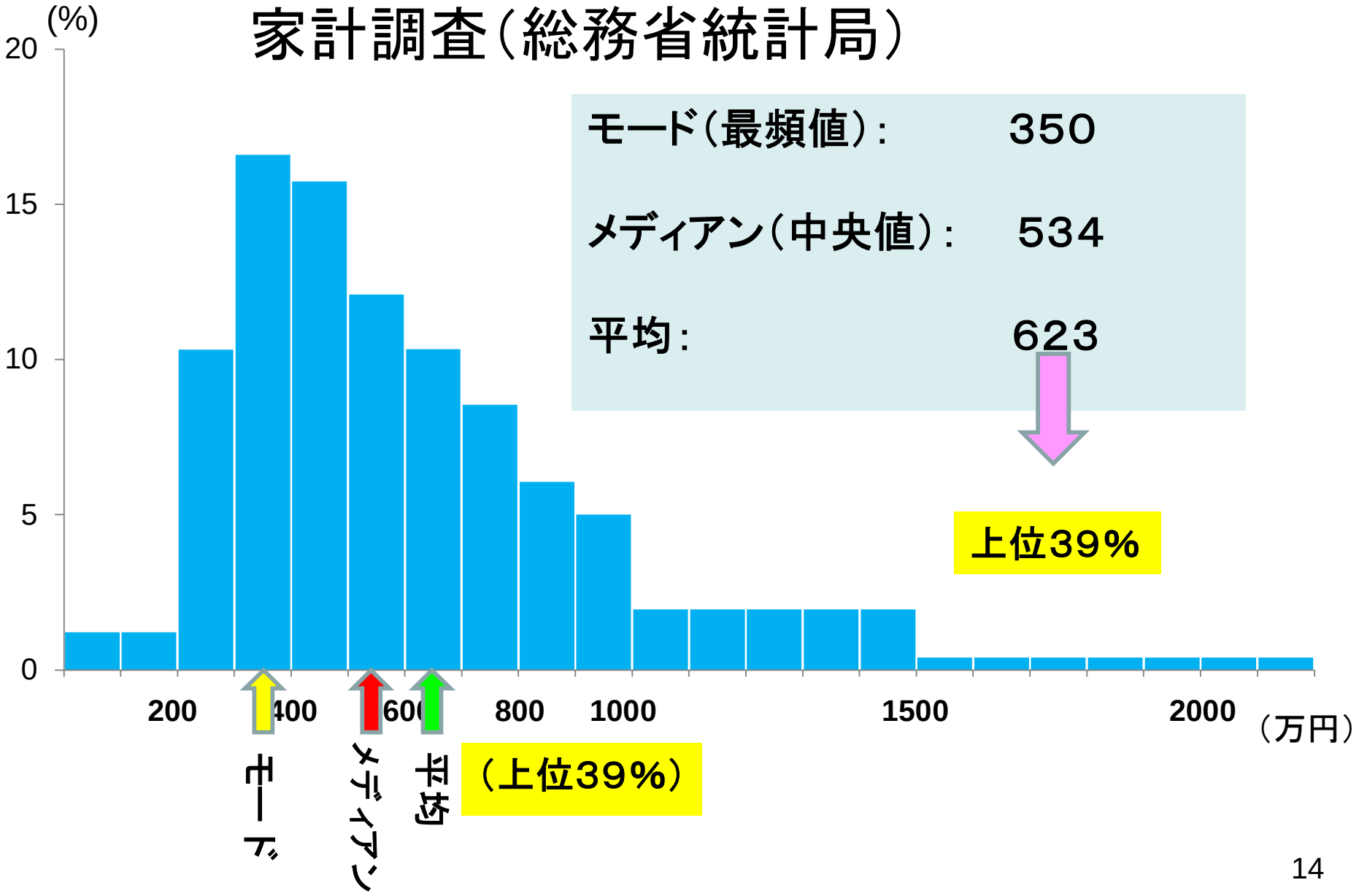
データ　$x_1, x_2, \cdots, x_n$ の平均であると同時に,

重心である.

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



世帯の年間収入の分布


家計調査（総務省統計局）
(%)
20
15
10
5
0
モード（最頻値）: 350
メディアン（中央値）: 534
平均: 623
上位39%
200
400
600
800
1000
1500
2000
（万円）
モード
メディアン
平均
（上位39%）

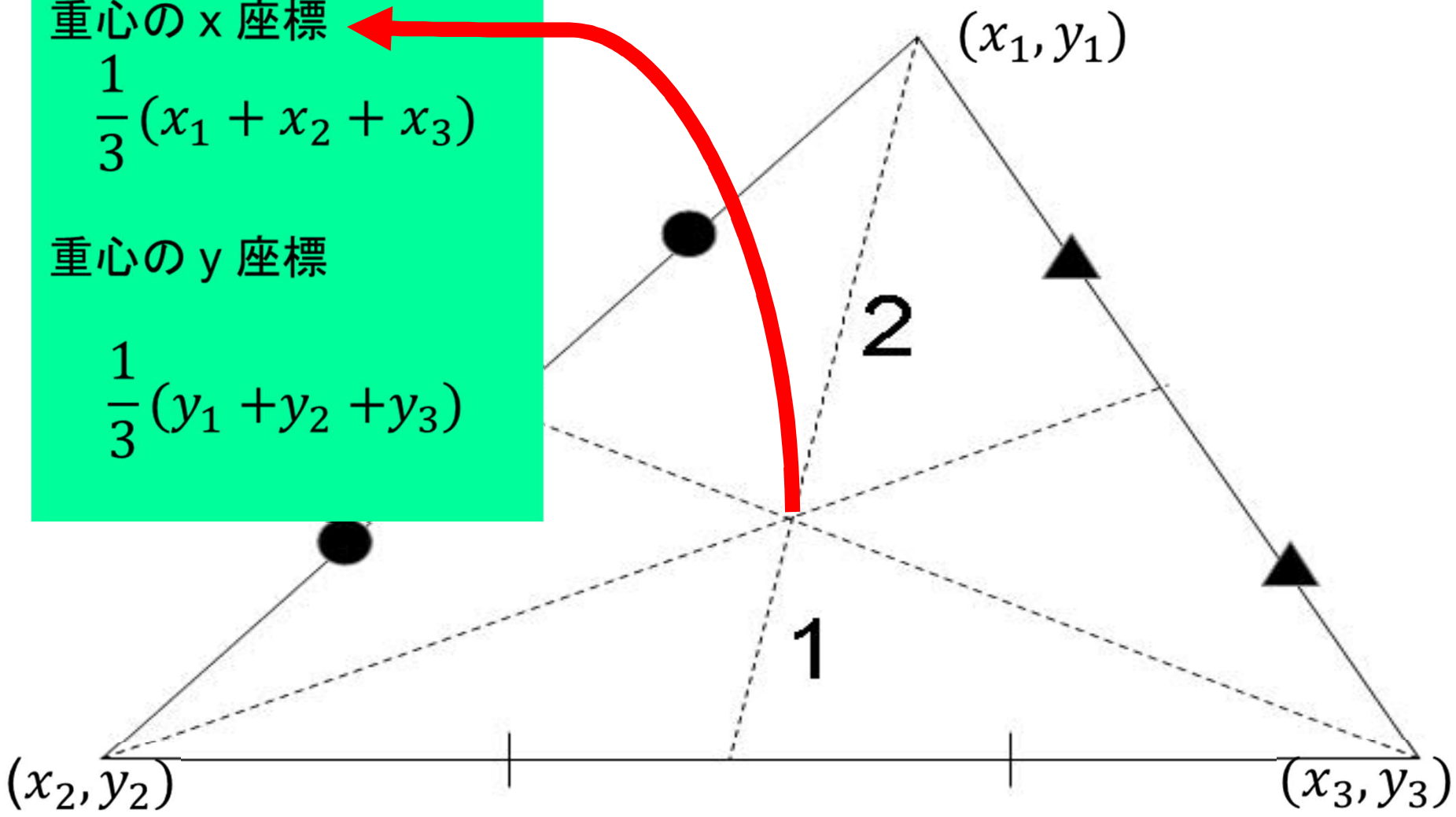
三 角 形 の 重 心

重心の x 座標

$$\frac{1}{3}(x_1 + x_2 + x_3)$$

重心の y 座標

$$\frac{1}{3}(y_1 + y_2 + y_3)$$



分 布 の 形 状

- 右にゆがんだ分布

（右スソが長い分布）

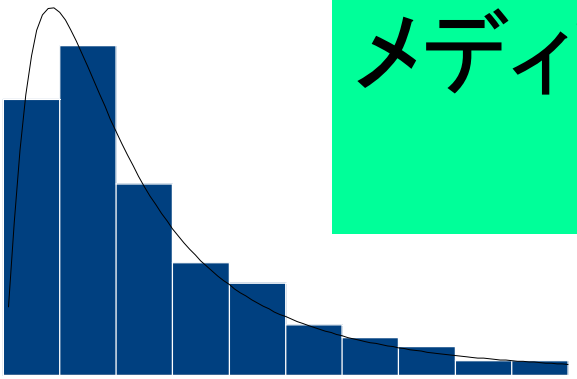
所得，待ち時間，入客数，
事故件数，新生児の体重，
むずかしい試験の点数，
家の広さ（日本？）

- 左にゆがんだ分布

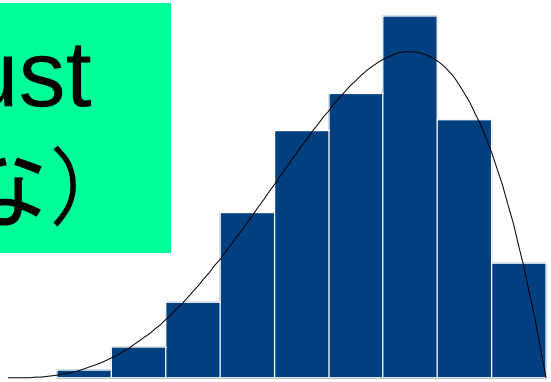
（左スソが長い分布）

機械の寿命，年齢（日本の場合），
やさしい試験の点数，
家の広さ（米国？）

メディアンは，robust
（頑健な）



モード＜メディアン＜平均



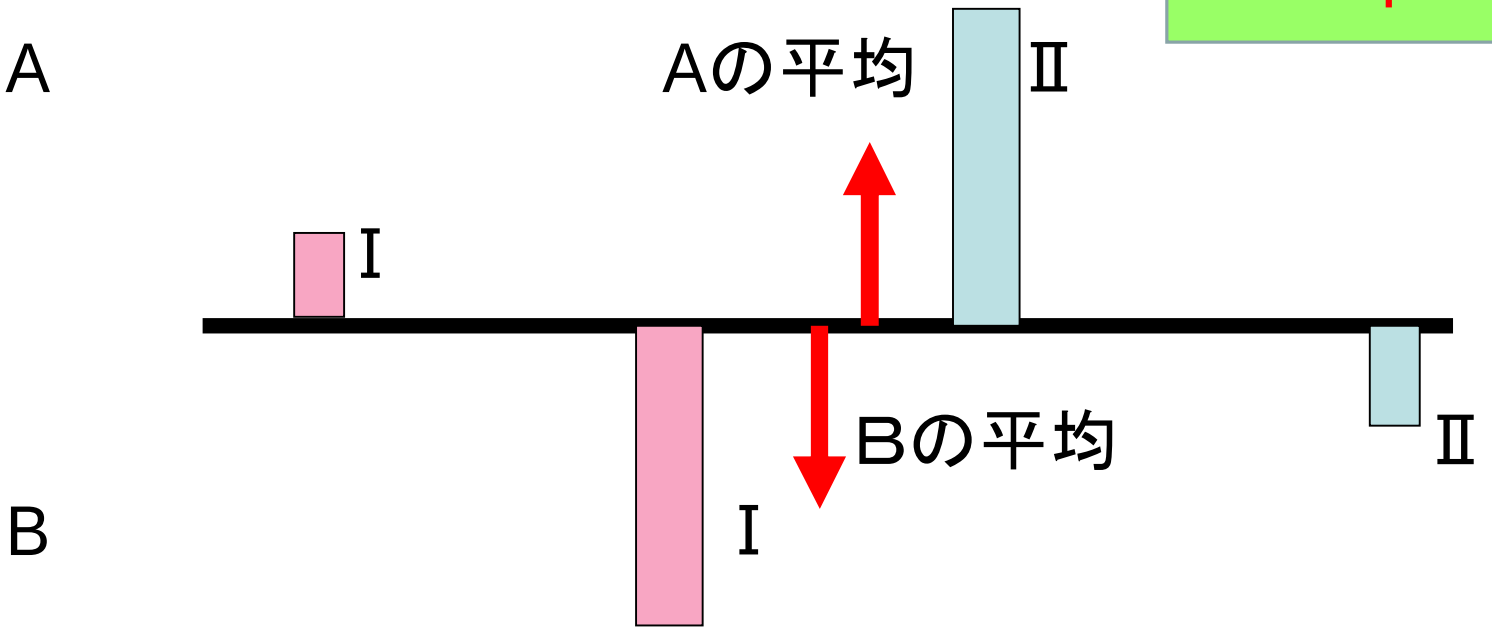
平均＜メディアン＜モード

平 均 の ウ ソ

		会社 A 人数	A 給与	B 人数	B 給与
職種	Ⅰ	20	20	80	25
	Ⅱ	80	30	20	35
	計	100	28	100	27

職種ごとには，
Bの給与が高い．
しかし，全体では，
Aの方が高い．

Simpson’s paradox



「中央値（メディアン）」の魔術

ボーダーライン上に成績順に並んだ4名

A＞B＞C＞D

の合否を21名の審査員が可否投票して，次の結果を得た.

4回の可否投票

	A	B	C	D
○	20	19	11	5
×	1	2	10	16

多数決原理により，A，B，C の 3 名が合格.

この結果を 1 回で決めることができる方式がある.

ボーダーラインの合否決定

21名の審査員が，上位何名を合格とするか，その人数を投票

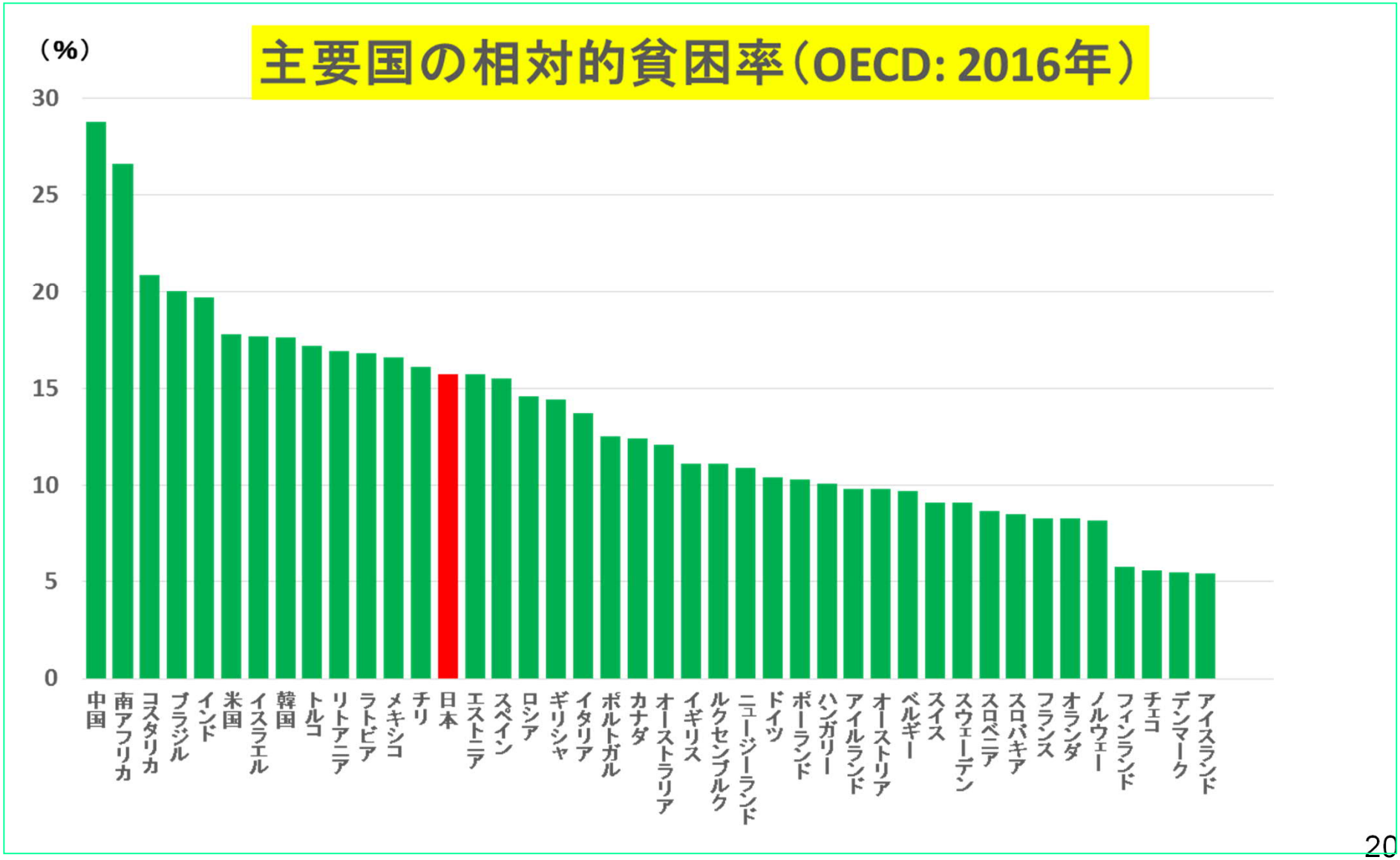
投票番号	0	1	2	3	4
票　　数	1	1	8	6	5
累積票数	1	2	10	16	21

上記の1回の投票と下記の4回の可否投票は同値

	A	B	C	D
○	20	19	11	5
×	1	2	10	16

1回の投票において，投票番号をデータとみなしたとき，
メディアンが合格人数の解となる.
0, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4

相対的貧困率
（＝メディアン収入の半分以下の人の割合）



平 均 とメディアン

平均

$$\sum_{i=1}^{n}(x_i - a) = 0 \quad \text{の解 } a$$

$$\sum_{i=1}^{n}(x_i - a)^2 \quad \text{を最小にする } a$$

メディアン

$$\sum_{i=1}^{n}|x_i - a| \quad \text{を最小にする } a$$

3. 帰納的推論・・・
「母集団」に思いを巡らす

◆現実のデータは，標本調査や実験の結果であり，全体の一部にすぎない.

◆本当に知りたいことは，背後にある母集団の情報である.

◆しかし，母集団を完全に知ることは不可能.

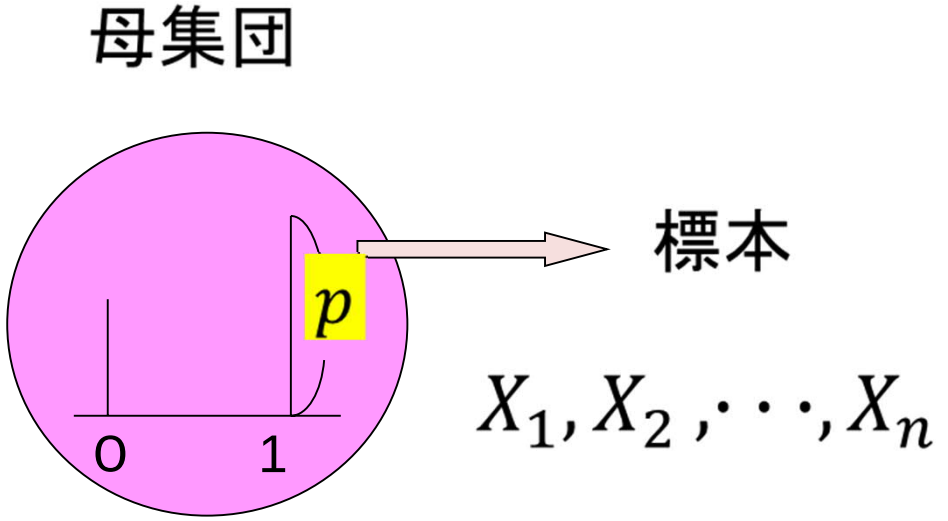
◆ただし，母集団の特性値（平均，標準偏差などのパラメータ）を確率的に推測することは可能である.

統 計 的 推 測 の 必 要 性

- 実験回数の制約
- 全数調査が困難

⇒ 一部から全体を推し測る必要性（帰納的推論）

- ➢ 母集団の構造を推測
- ➢ 誤差を確率的に評価
- ➢ 因果関係の特定化
- ➢ モデルの構築



$\hat{p} = f(X_1, X_2, \cdots, X_n)$

（p の推定量）

$P(|\hat{p} - p| > a)$

（推定量の誤差評価）

母集団失業率 p の推定

- 知りたいのは，母集団（約7,000万人）における失業率
- 実際には，10万人程度を調査（総務省統計局）

$\hat{p} = 0.022$ ：母集団失業率 p の推定値

- 誤差を確率的に評価（中心極限定理による）
- 95%の信頼度で，母集団失業率は，

$0.0211 < p < 0.0229$ の範囲にある．

（誤差は，± 0.0009 程度）

- 100万人を調査した結果とみなすと，

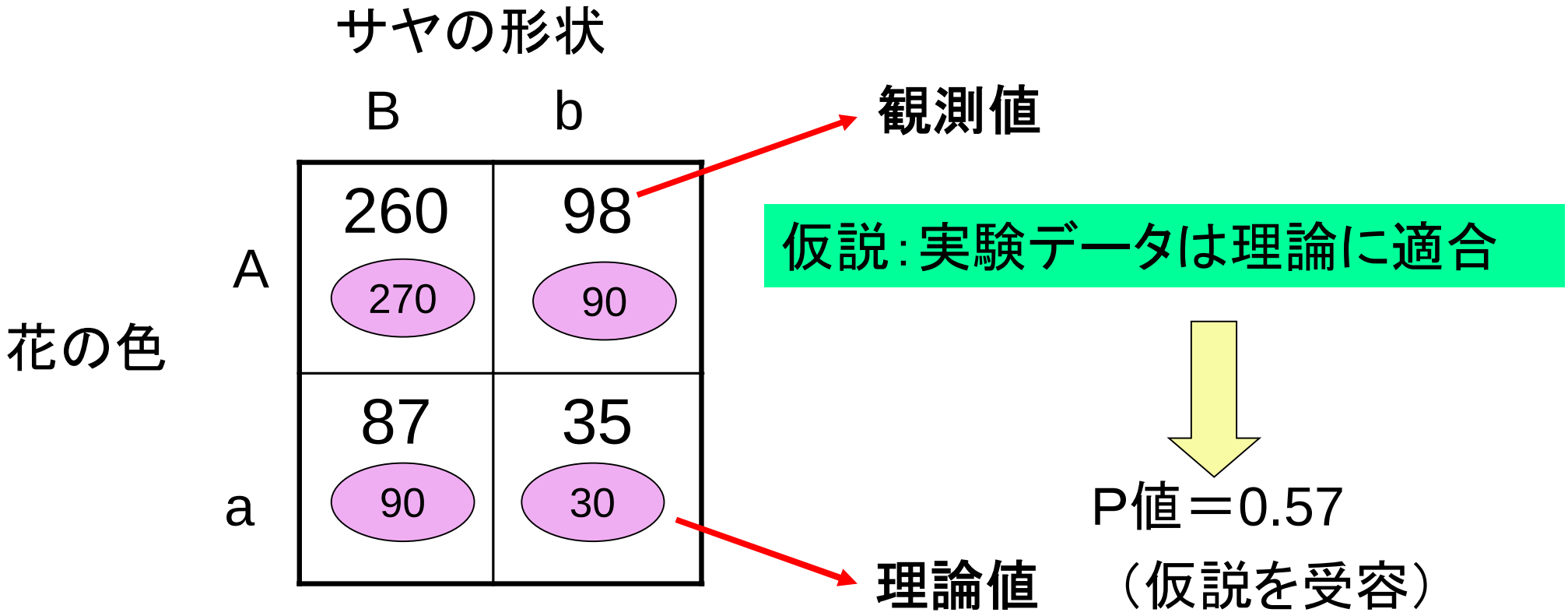
$0.0217 < p < 0.0223$ の範囲にある．

（誤差は，± 0.0003 程度）

理 論 仮 説 の 検 定

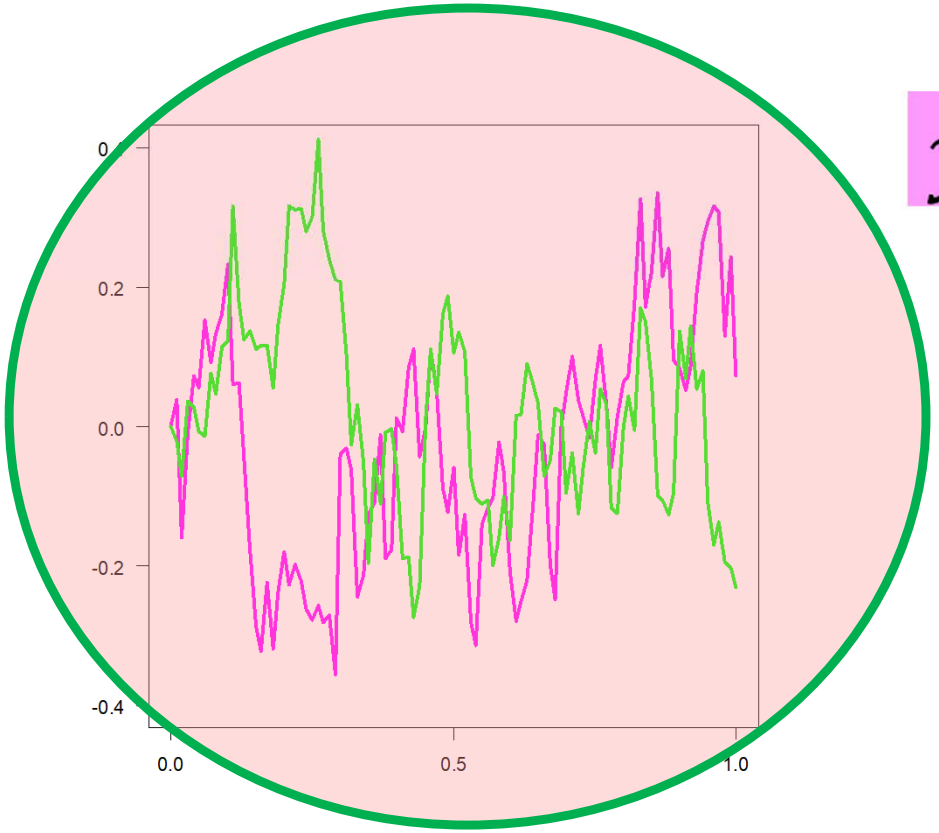
- メンデルの法則(分離の法則：第2世代の交配)

実験（480例）



確率モデルを使った推測

◆　データの生成プロセスが複雑な場合（神のみぞ知る）は，
確率モデルを作って推測する．



$$y_t = f(y_{t-1}, x_t, w_t; \theta) + u_t$$

さまざまな確率モデル：
単純確率モデル，回帰モデル，
ロジットモデル，パネルモデル
同時方程式モデル
時系列モデル，長期記憶モデル
点過程，状態空間モデル
ブラウン運動，拡散過程，•••

データ分析の例

- ゲノムデータによる遺伝子機能の推定（生物情報科学）
- 臨床データによる新薬の効能の検査

- 経済予測・需要予測
- 失業率の推定
- 金融市場のリスク分析

（金融工学，数理ファイナンス）

- 選挙当確速報

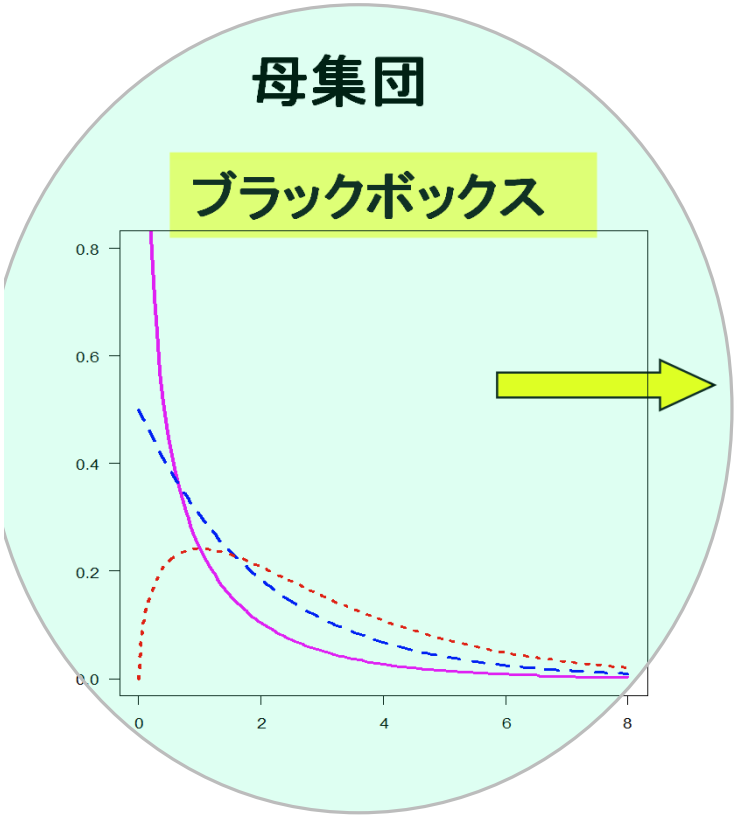
- 気象予報
- 地球環境データの分析
- 地震データの解析
- 生息する生物の個体数の推定

- 文学作品の作者の真贋問題（「源氏物語」の宇治十帖？）
- 大学入試センター試験の信頼性・妥当性の分析
- ビッグデータに基づく市場分析

27

4. 中心極限定理・・・その不思議さと美しさ

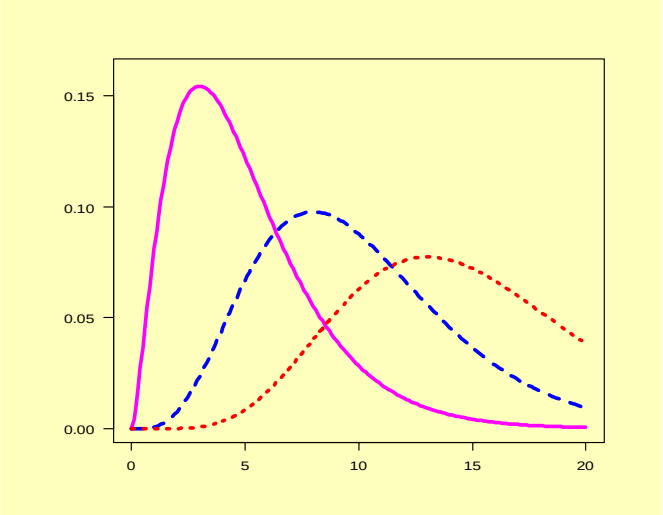
未知の母集団分布の平均などの推論
（誤差を確率的に評価）

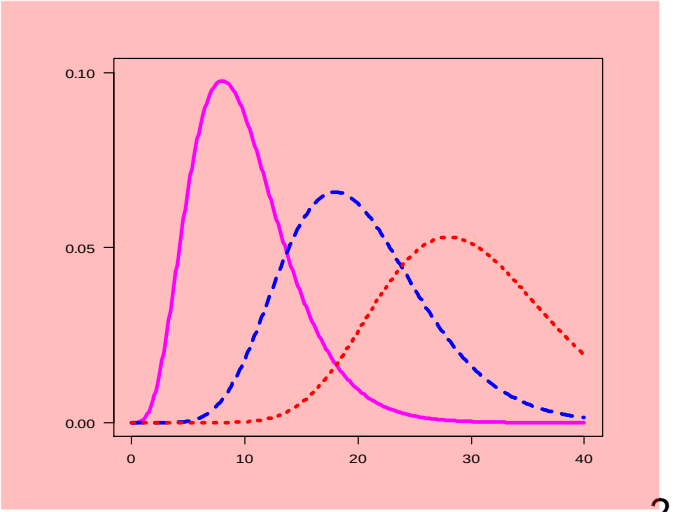


$$X_1 + \cdots + X_5$$

$$X_1, X_2, \cdots, X_n$$

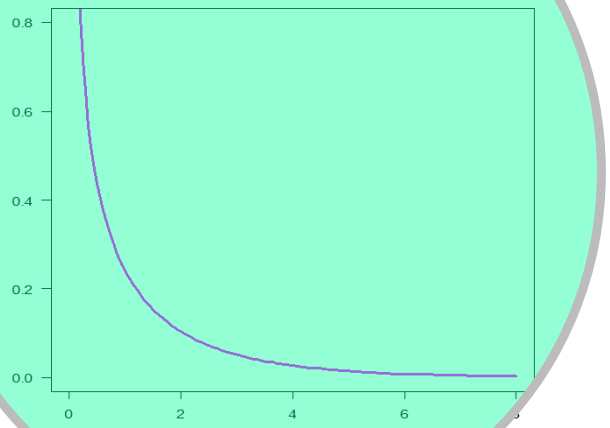
$$X_1 + \cdots + X_{10}$$



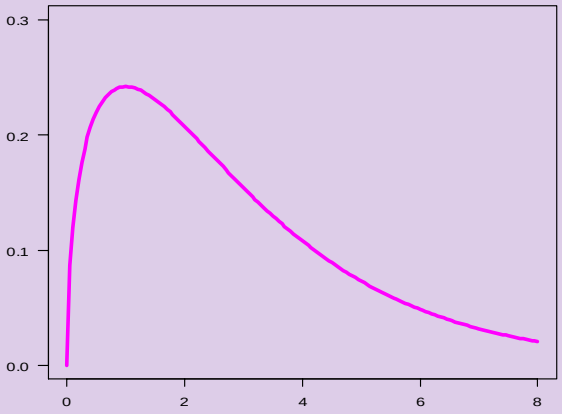


中心極限定理（Central Limit Theorem）

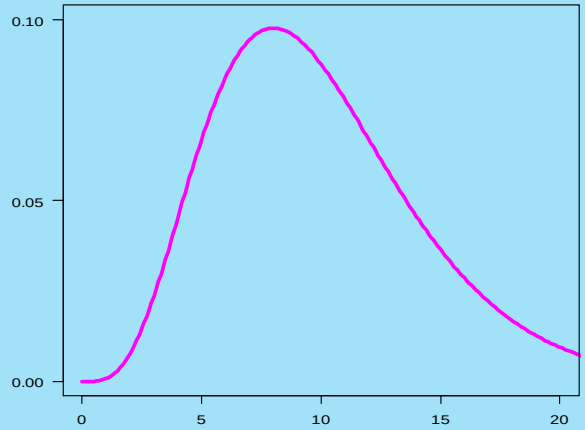
母集団

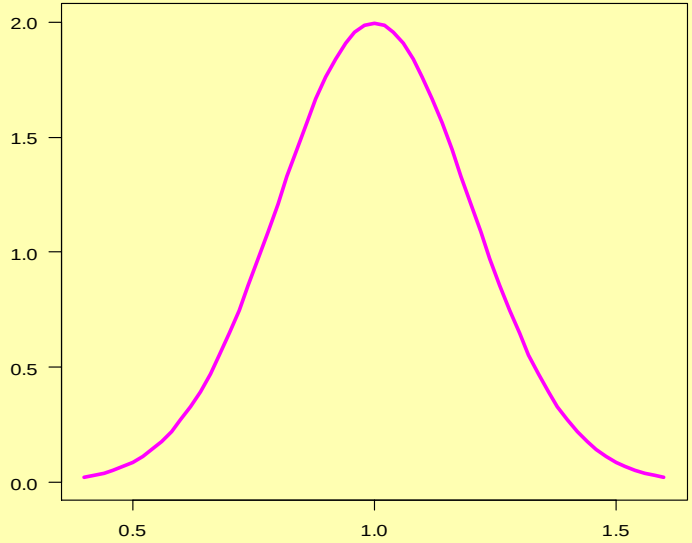


$X_1 + X_2 + X_3$



$X_1 + X_2 + \cdots + X_{10}$



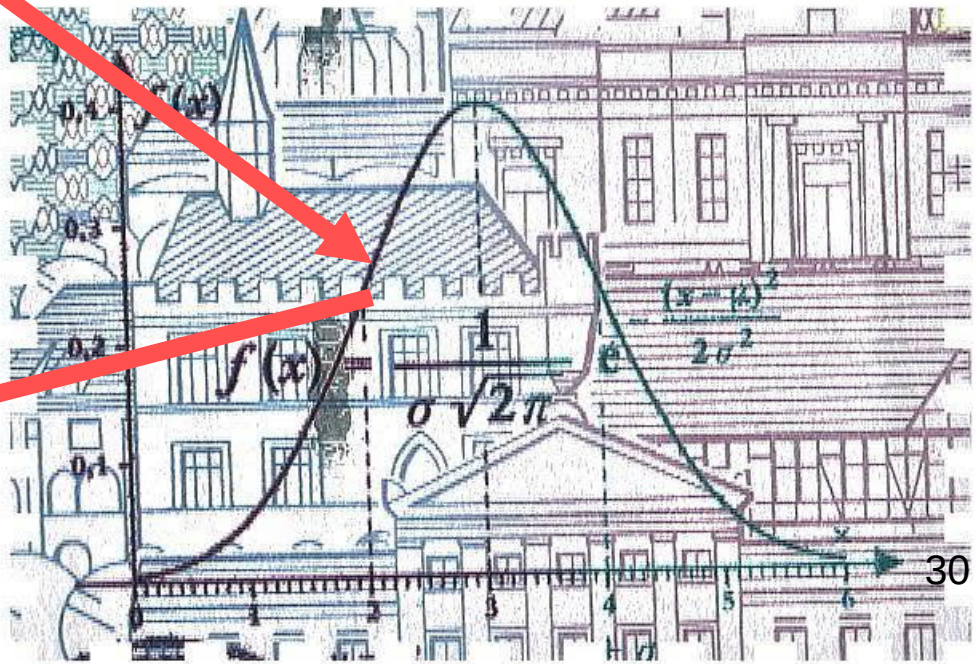


行き着く先は，正規分布

正規分布(Normal Distribution)=Gauss 分布

Carl Friedrich Gauss (1777-1855)

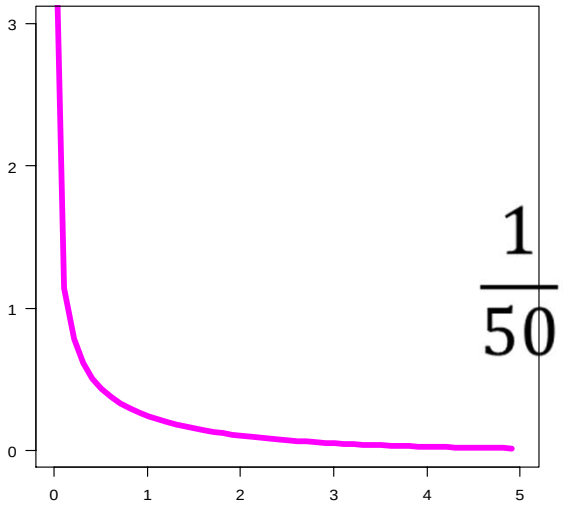
10マルク紙幣（西ドイツ）



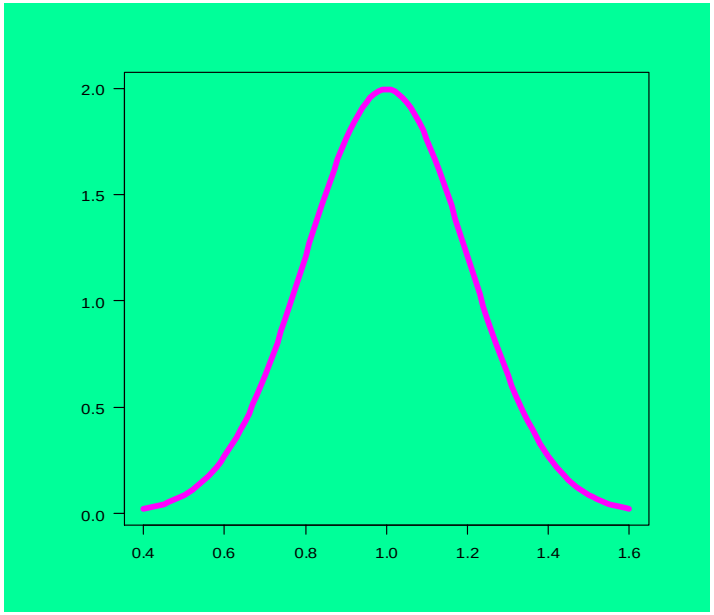
正規分布の密度関数

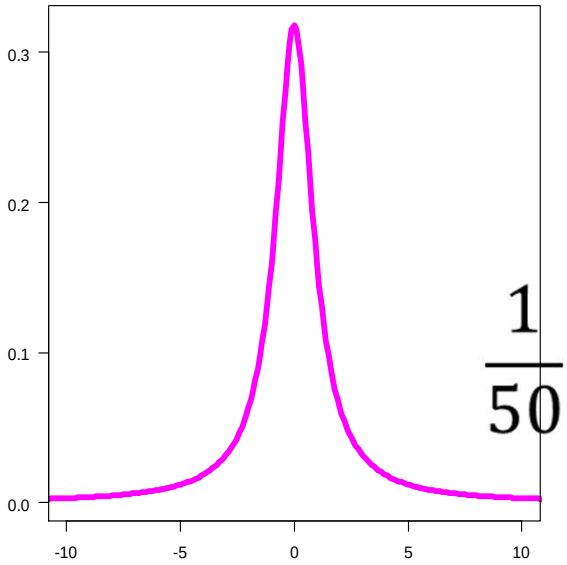
$$f(x) = \frac{1}{\sigma\sqrt{2\,\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

There is no rule without exceptions.

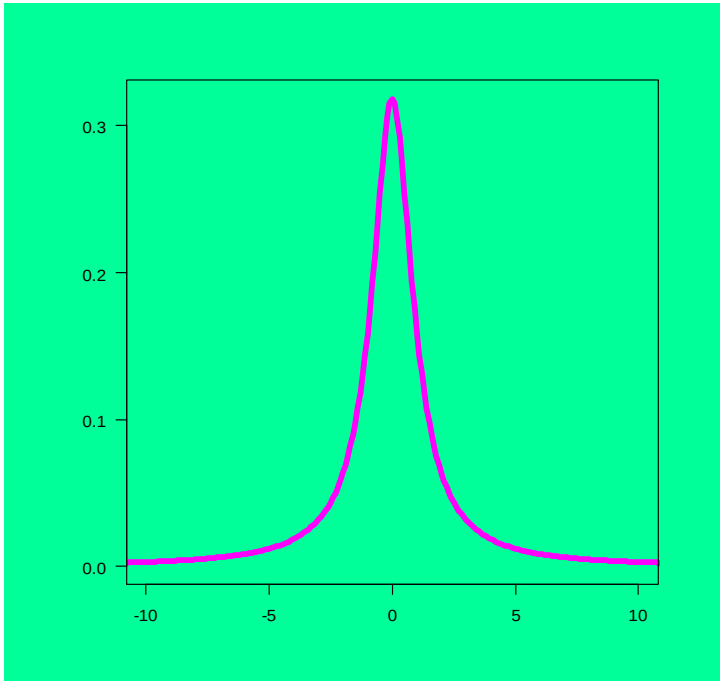


$$\frac{1}{50}(X_1 + \cdots + X_{50})$$





$$\frac{1}{50}(Y_1 + \cdots + Y_{50})$$

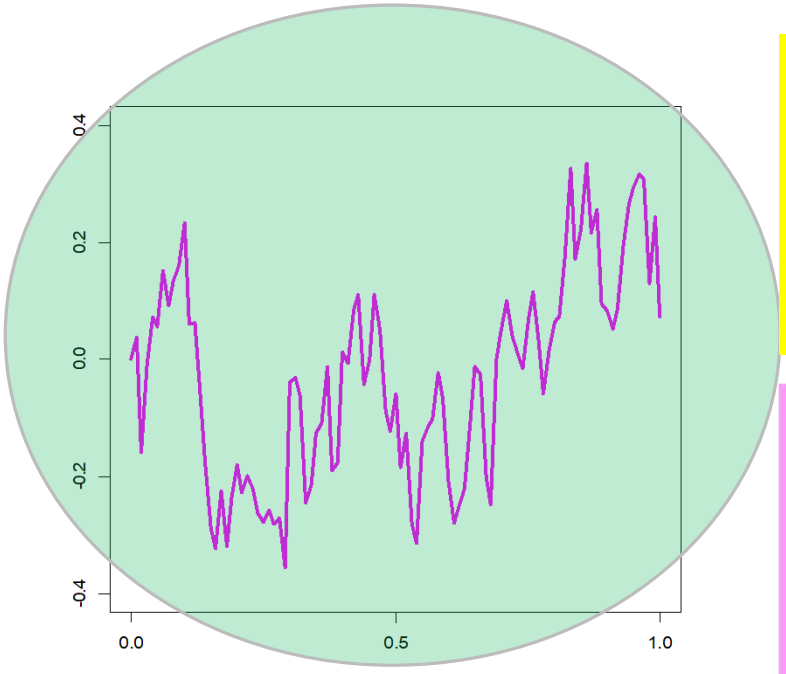


非正規分布への収束

連続時間確率過程 {Y(t)} の確率モデル

$$dY(t) = \alpha Y(t)dt + dB_H(t), \qquad 0 \le t \le T$$

$B_H(t)$： フラクショナル・ブラウン運動



左の母集団から得られるデータの生成プロセスとして，上の確率モデル（確率微分方程式）を考える．

このとき，統計的問題は，モデルに含まれるパラメータ α の推定問題に帰着する．

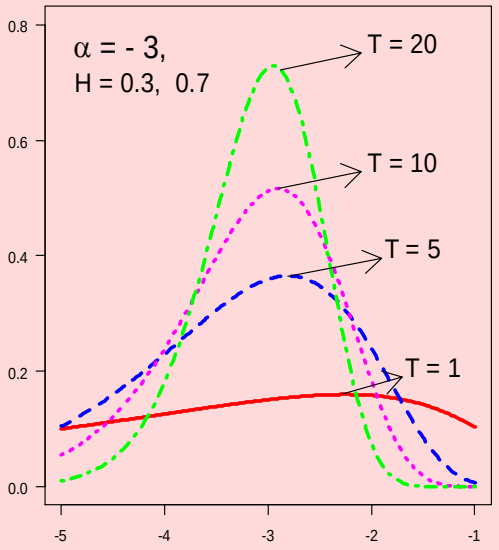
非正規分布への収束

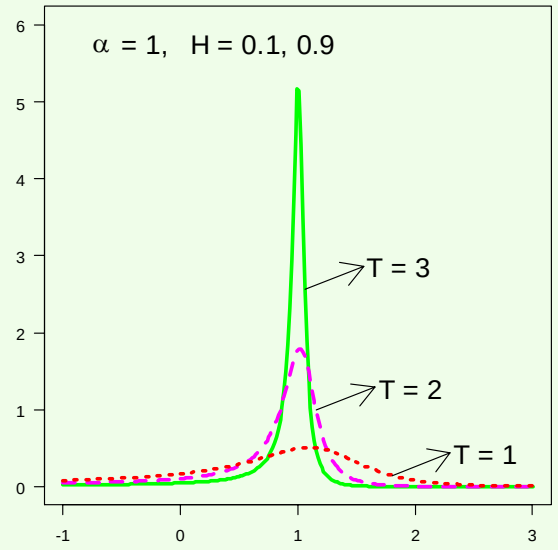
連続時間確率過程 $\{Y(t)\}$ の確率モデル

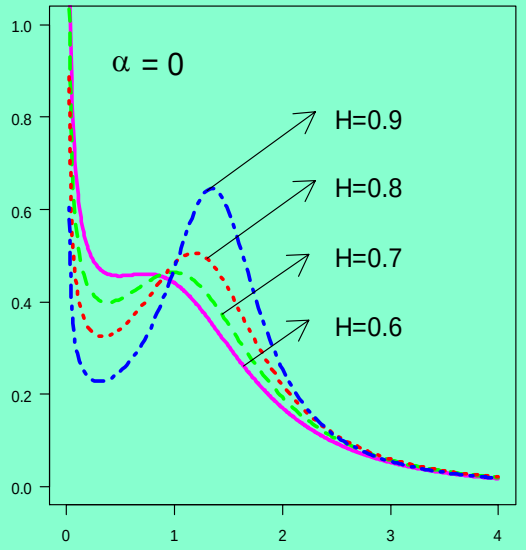
$$dY(t) = \alpha Y(t)dt + dB_H(t), \qquad 0 \le t \le T$$

$B_H(t)$： フラクショナル・ブラウン運動

パラメータ α の推定量の分布は，推定方法，および，
α の符号に依存して異なる．



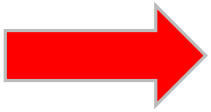




5. ビッグデータの時代・・・
データは21世紀の石油

- ビッグデータ：従来の統計データとは異なる
  （非定型的，非構造的）

＝スマートフォンの普及とIT技術の進化によって生まれた，
大容量で多様な高頻度データ（3つの V）
（Volume, Variety, Velocity)
動画データ，ツイッター，GPSデータ，
インターネット閲覧・購入履歴，ポイント・カードの
購入履歴，クレジッドカード履歴，・・・



ビッグデータの活用が，企業に「ビジネスチャンス」をもたらす．

第四の科学「データサイエンス」の出現

ビッグデータから，新たな知見を発見したり，価値を創造するための科学

- データサイエンスは，第四期の科学

第一期：天動説のような自然観測に基づく科学

第二期：理論と実験の繰り返しによる方法

第三期：コンピュータによる計算科学

第四期：ビッグデータから新たな知見を見出す

（データ駆動型科学）

「データサイエンティスト」の台頭

企業だけでなく，官庁，自治体の意思決定において，
ビッグデータ分析の重要性が高まっている．

データサイエンスのスペシャリスト：求められる3つのスキル

理系的スキル

(1) データ処理：「情報学」，「コンピュータサイエンス」

(2) データ分析：「統計学」（高度な理論は不要）

文系的スキル

(3) 価値創造：「ビジネスや問題解決に繋げる提案力」，
「専門家とのコミュニケーション能力」

上記 3 つのスキルを同時に備えた人材は希少．
「データサイエンス学部」の創設（社会的要請）．