

研究を始めるための 統計のはなし

兵庫医科大学
地域包括ケア学
関 進

統計とは

- ある対象の集団の性質を数量的に捉える
 - 個々の対象のデータは数量だけでなく
 - 平均身長165.9cm
 - 平均体重59.8kg
 - カテゴリーでも
 - 男性25人
 - 女性20人
- 数量的に捉えることが可能

尺度(変数)

- 個々の対象の性質によって3種類に分類
 - 連続尺度
 - 数値的に扱える
 - 足し算, 引き算ができる
 - 例: 身長, 体重, 年齢
 - 名義尺度
 - カテゴリーとして分類できる
 - どのカテゴリーに属するかがはっきりわかる
 - 例: 男女, 居住地域
 - 順序尺度
 - 名義尺度の性質に加えてカテゴリーに順位がある
 - 例: 年少(0-14歳)・生産年齢(15-64歳)・老年(65歳-)

集団としての連続尺度

- 集団を特徴づけるいろいろな量が提案
特に重要なのは
 - 集団の中心
 - 平均
 - 中央値
 - 集団のバラツキ
 - 分散(標準偏差)
 - 四分位数, 最小値, 最大値

値をもとにした中心

- 平均

得られたデータ

$x_1, \dots, x_N$

の平均は

$$M = \frac{1}{N} \sum_{i=1}^N x_i = \frac{x_1 + x_2 + \dots + x_N}{N}$$

値をもとにしたバラつき

- 分散

- 平均からのずれの平均

$$S^2 = \frac{1}{N} \sum_{n=1}^N (x_i - M)^2$$

- 標準偏差

- 分散の根号: もとの量に合わせる

$$S = \sqrt{S^2}$$

平均・分散・標準偏差の例

- データ

160.2, 164.3, 170.8, 171.5, 171.2, 168.5, 167.6, 170.7, 181.9
9個

の平均

$$\begin{aligned}\bar{M} &= \frac{160.2 + 164.3 + 170.8 + 171.5 + 171.2 + 168.5 + 167.6 + 170.7 + 181.9}{9} \\ &= 169.6\end{aligned}$$

分散

$$\begin{aligned}\bar{S}^2 &= ((160.2-169.6)^2 + (164.3-169.6)^2 + (170.8-169.6)^2 + (171.5-169.6)^2 \\ &\quad + (171.2-169.6)^2 + (168.5-169.6)^2 + (167.6-169.6)^2 + (170.7- \\ &\quad 169.6)^2 + (181.9-169.6)^2) / 9 = 31.3\end{aligned}$$

標準偏差

$$\bar{S} = \sqrt{31.3} = 5.6$$

順序をもとにした中心と バラつき

- 中央値
 - 順序をもとにした中心の値
 - データを小さいものから順番に並べた $1/2$ の位置の値
- 四分位数
 - 順序をもとにしたばらつき
 - データを小さいものから順番に並べた $1/4$, $3/4$ の位置の値

中央値・四分位数の例

- データ

160.2, 164.3, 170.8, 171.5, 171.2, 168.5, 167.6, 170.7, 181.9

9個

並べ替え

160.2, 164.3, 167.6, 168.5, 170.7, 170.8, 171.2, 171.5, 181.9

最小値

1/4

1/2

3/4

最大値

中央値

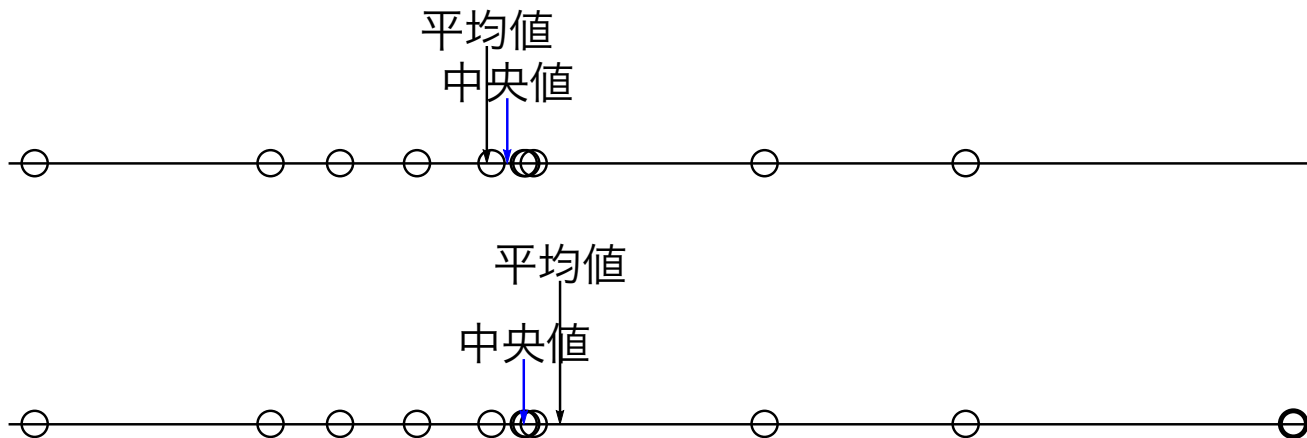
$$\text{Med} = 170.7$$

四分位数

$$\text{IQR} = (166.0 - 171.4)$$

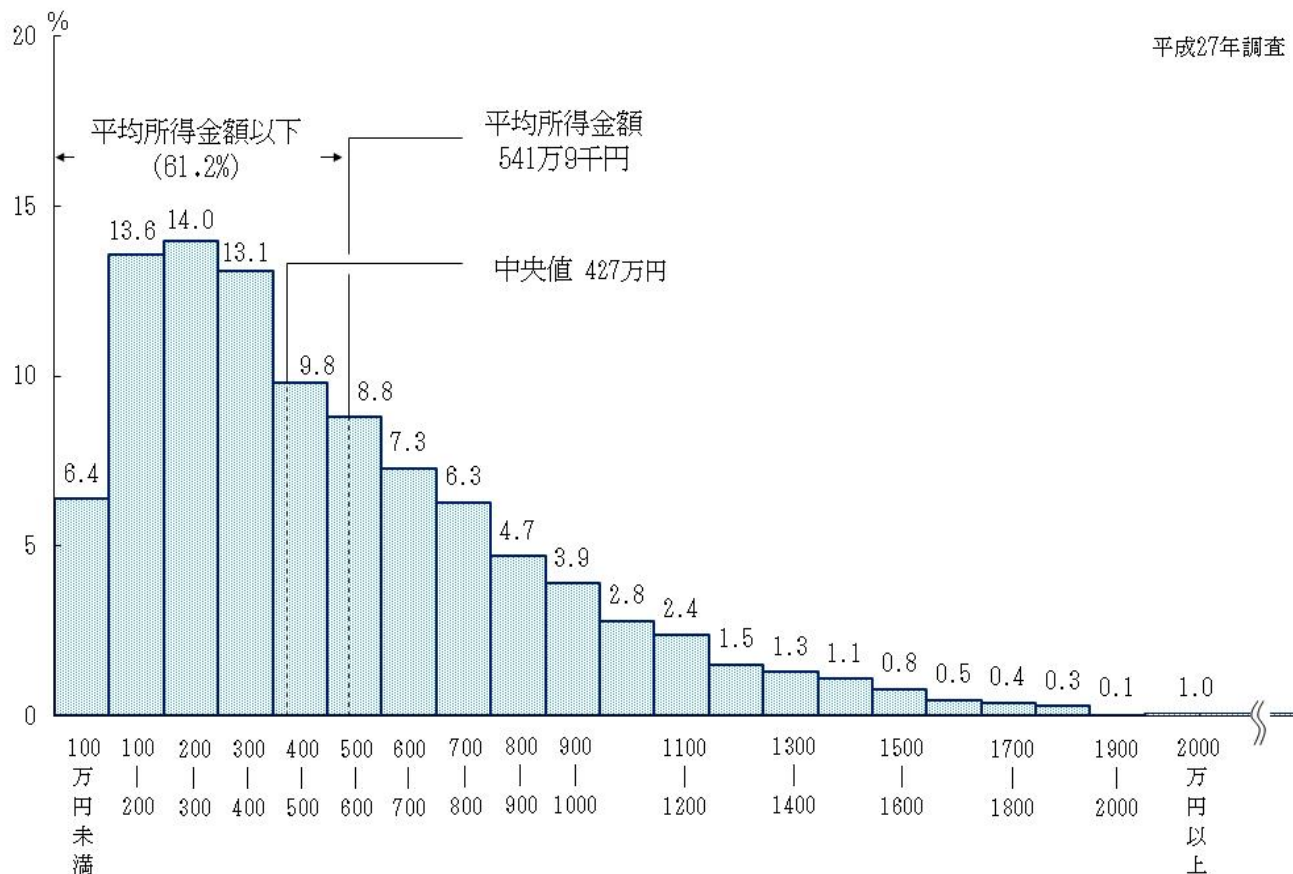
平均と中央値の比較

- 平均
 - メリット: 理論的な扱いが容易
 - デメリット: 極端な測定値の影響を受けやすい
- 中央値
 - メリット: 測定に対して安定
 - デメリット: 理論的な扱いが難しい



平均と中央値の例

- 平均は中央値に比べ高額所得者の影響大



厚生労働省平成27年国民生活基礎調査の概要

<http://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa15/index.html>

統計の対象とする集団

- 統計で分析する集団
 1. 分析する集団そのものが対象の場合
 - 例) 学校のクラスで担任が子供たちの全体としての様子を把握したい
 2. 分析する集団の背後に普遍的な対象がある場合
 - 例) ある治療法が効果的かどうかを調べたい
研究で実際に治療される患者さんの背後に、
この治療法を受ける可能性のある患者さん全体
- ここでは2について述べる

母集団

- 母集団

- 調査対象として想定している全体

- 例) 有権者全体

- 選挙では開票が進む前に結果(当選確実)が報道
 - 統計により推測. しかし有権者全体を調べてはいない
 - 有権者全体は有限→有限母集団

- 例) 実験

- コインを投げたとき表の出る確率を求めたい
 - コインは何回でも(無限回)投げられる→無限母集団

標本と母集団

- 標本

- 母集団から実際にデータを収集する対象

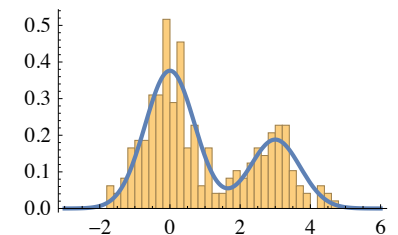
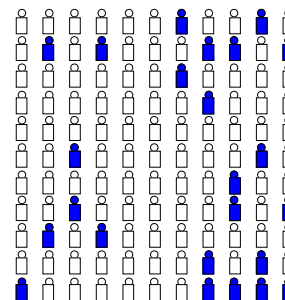
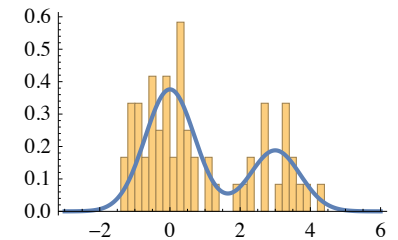
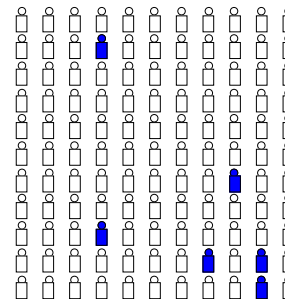
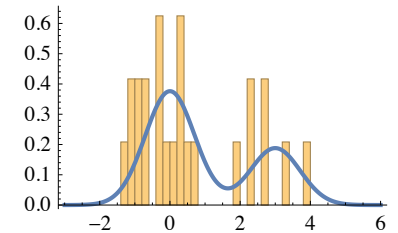
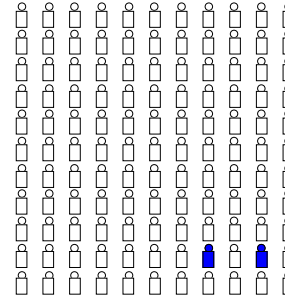
統計では標本のデータから母集団の性質を推測することが重要

- 例) 治療法の効果

- 例) 候補者の得票率

標本サイズ

- 標本サイズ
 - 実際に測定した数
 - 標本サイズの増加に伴ない母集団の性質が明確化



母集団のモデル

- 母集団は確率的現象として記述
 - 例)サイコロを投げて1が出る確率: $1/6$
 - 実際に目の前のサイコロの1の目が出るのが $1/6$ かどうかは実験をしなければわからない!
 - 6回投げたら $6/6=1$ 回1の目が出るはず.
 - 1回も出なかった→サイコロは歪んでる?
 - 600回投げたら $600/6=100$ 回1の目が出るはず.
 - 80回しか出なかった→サイコロは歪んでる?

母数

- 母数
 - 母集団の分布を特徴づけるパラメータ
 - 母平均 μ
 - 母集団の平均
 - 母分散 σ^2
 - 母集団の分散

点推定

- 点推定
 - 標本から母集団の母数をピンポイントで推定
 - 点推定とは別に母数を範囲で推定する区間推定もある(ここでは述べない)

母平均と標本平均

- 母平均

- 母平均の点推定は標本平均

$$M = \frac{1}{N} \sum_{n=1}^N x_i \longrightarrow \mu$$

- 大数の法則

- N が大きくなれば M は μ に限りなく近づく

母分散と標本分散

- 母分散
 - 母分散の点推定は

$$\cancel{s^2 = \frac{1}{N} \sum_{n=1}^N (x_i - M)^2} \quad s^2 = \frac{1}{N-1} \sum_{n=1}^N (x_i - M)^2 \longrightarrow \sigma^2$$

μ≠Mではばらつきが小さめに出るための補正
不偏推定量

仮説検定

- 統計手法を用いて母集団の性質について仮説を立てて、それが正しいかどうかを調べる
- 医療分野の統計研究で非常にしばしば用いられる
 - － 例) 治療法の有効性
 - － 例) 患者集団の違いによる差

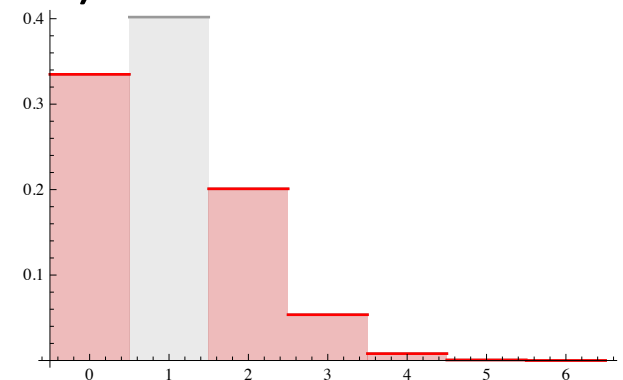
仮説検定の考え方

- 仮説検定は以下の手順に従う.
 1. 対立仮説, 帰無仮説をたてる.
 2. 観察または実験を行う. 結果について確率計算を行い, あらかじめ決めておいた確率 (危険率, 有意水準) α (通常は5%や1%) と比較する. α 未満のときは帰無仮説を棄却. 対立仮説が危険率 α で有意.

仮説検定の例

(サイコロの1の目の出方, その1)

- サイコロは歪んでいるかを $\alpha=0.05$ で検定
 - 帰無仮説: サイコロの1の目の出る確率 $=1/6$.
対立仮説: サイコロの1の目の出る確率 $\neq 1/6$
 - 6回投げたら1の目が0回
 - 確率は2項分布で計算可能
 - $P(1の目が0回) + P(1の目が2回以上)$
 $= 0.60 > 0.05$
- 帰無仮説は棄却できない

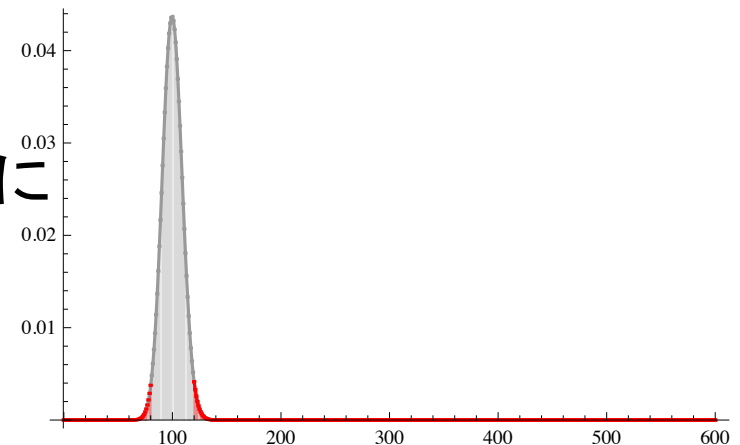


仮説検定の例

(サイコロの1の目の出方, その2)

- サイコロは歪んでいるかを $\alpha=0.05$ で検定
 - 帰無仮説: サイコロの1の目の出る確率 $=1/6$.
対立仮説: サイコロの1の目の出る確率 $\neq 1/6$
 - 600回投げたら1の目が80回
 - $P(1の目が80回以下)+P(1の目が120回以上)$
 $=0.032 < 0.05$

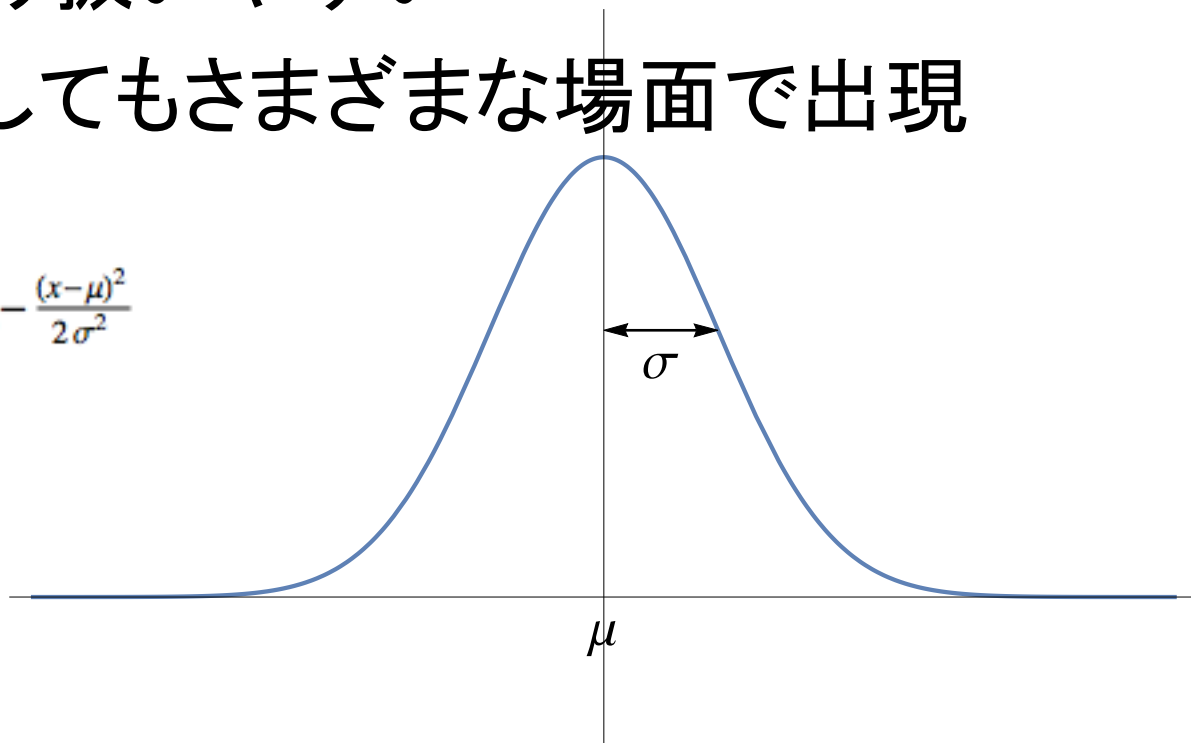
帰無仮説は棄却される
対立仮説が危険率5%で有意に
成り立つ



正規分布

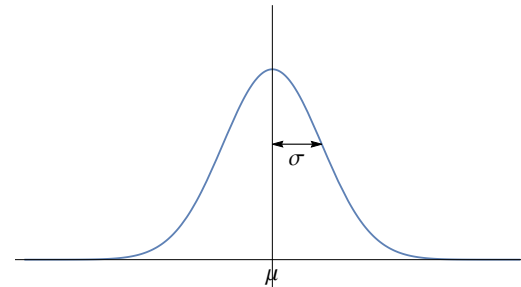
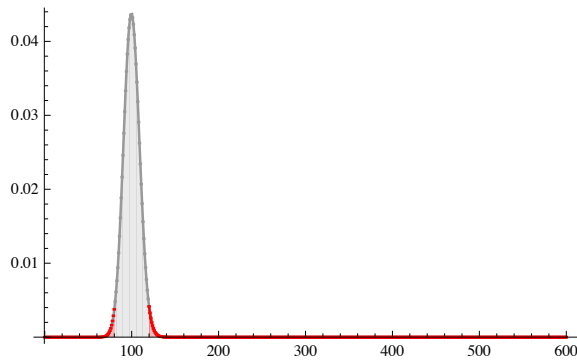
- 統計では母分布が正規分布であるときにさまざまな分析方法が確立
- 理論的に取り扱いやすい
- 自然現象としてもさまざまな場面で出現

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



中心極限定理

- サイコロの目の検定に出てきた2項分布の形は正規分布に似ている



一般には母平均 μ , 母標準偏差 σ の母分布の標本で以下の式は平均0, 分散1の正規分布に近づく

$$\frac{1}{\sqrt{N} \sigma} \sum_{n=1}^N (x_i - \mu)$$

正規分布を用いた母平均の検定

- 仮定
 - 2種類のグループの各グループの母分布が同分散の正規分布
- 仮説
 - 帰無仮説: 2グループの母平均は同じ
 - 対立仮説: 2グループの母平均は同じではない
- 検定
 - 有意水準 α を設定し確率計算
 - t検定

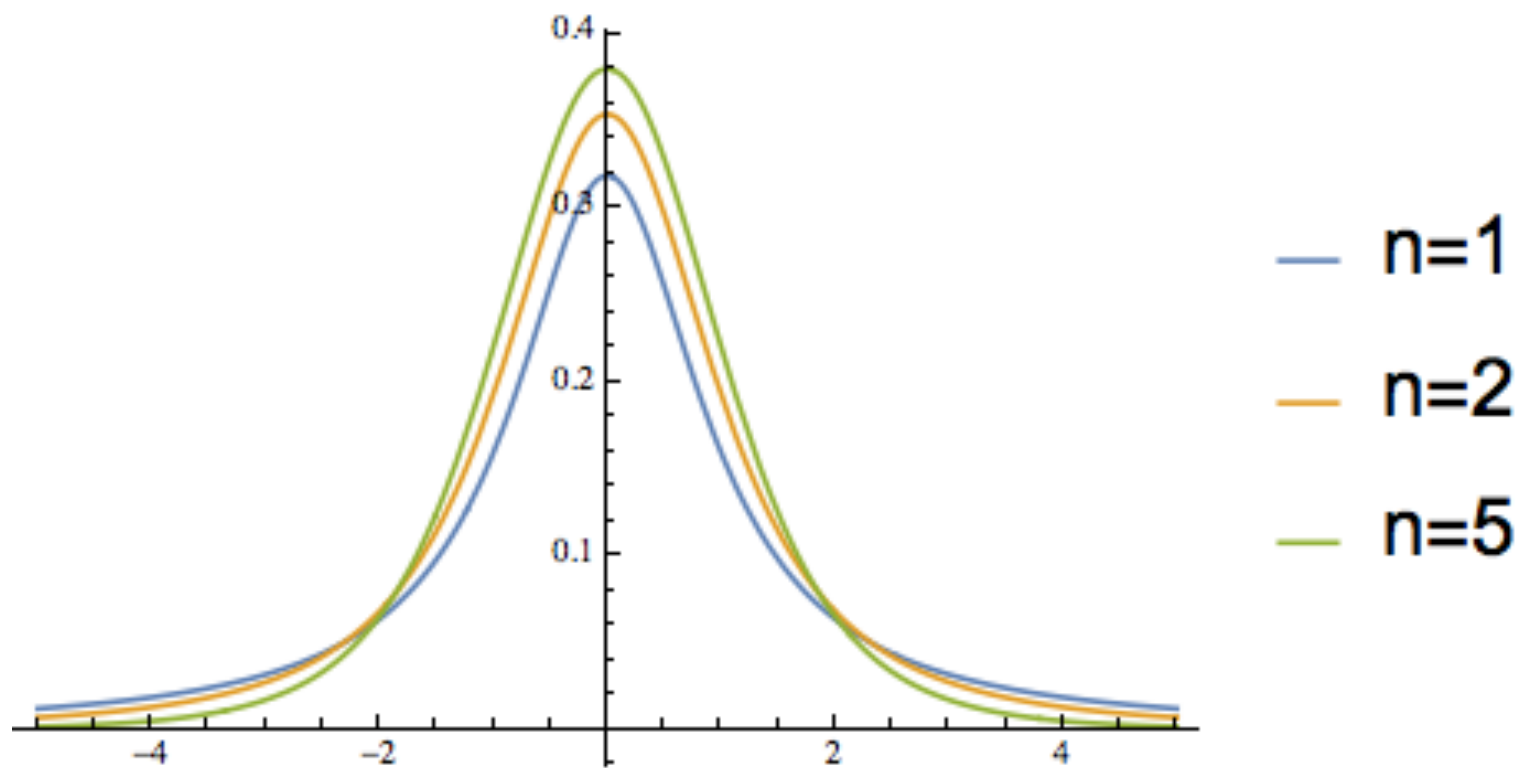
t検定

- 2グループが仮定を満たすとき
 - それぞれの標本サイズ N_1, N_2
 - 標本平均 M_1, M_2
 - 標本標準偏差(不偏推定) s_1, s_2

$$t = \frac{\sqrt{N_1 N_2 (N_1 + N_2 - 2)} (M_1 - M_2)}{\sqrt{(N_1 + N_2) ((N_1 - 1) s_1^2 + (N_2 - 1) s_2^2)}}$$

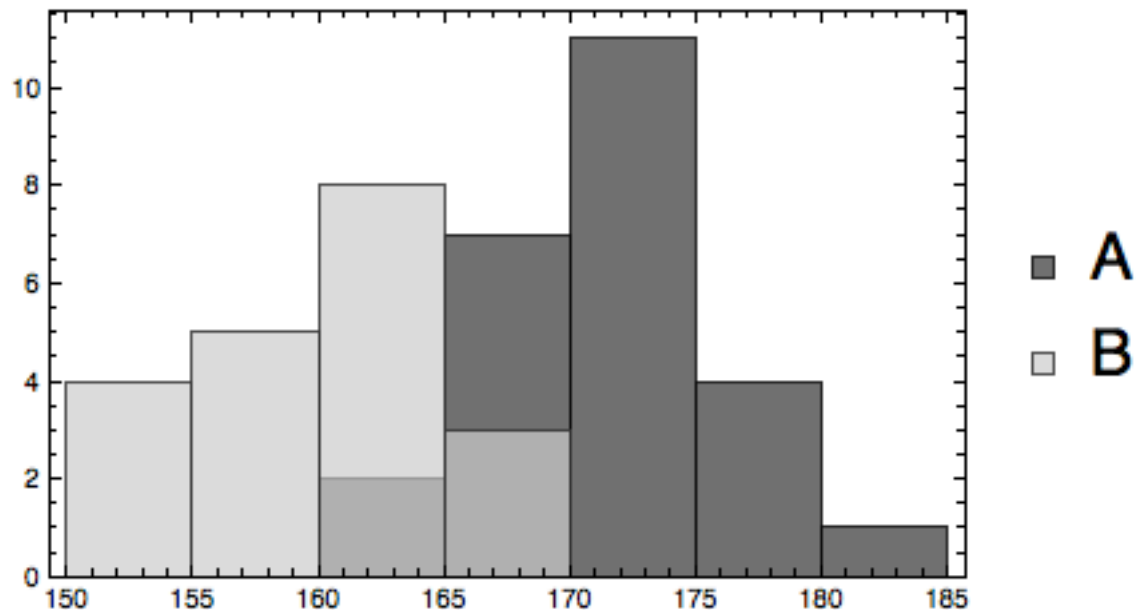
は自由度 $N_1 + N_2 - 2$ のt分布に従う.

t分布



t検定の例(1)

- グループA, Bで以下のようなデータが得られた. このときのA, Bで母平均に差があるかどうかを調べたい.



t検定の例(2)

- 仮定
 - 2種類のグループの各グループの母分布が同分散の正規分布
- 仮説
 - 帰無仮説: 2グループの母平均は同じ
 - 対立仮説: 2グループの母平均は同じではない
- 検定
 - 有意水準 $\alpha=0.05$ で検定

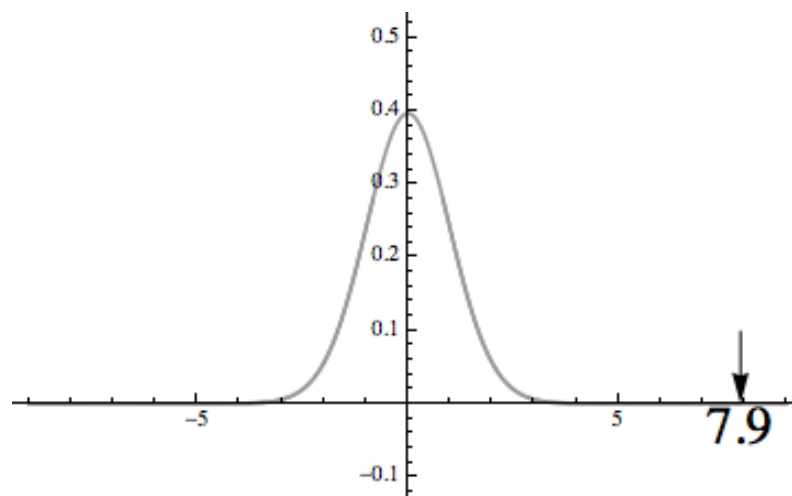
t検定の例(3)

• tの計算

$$t = \frac{\sqrt{N_1 N_2 (N_1 + N_2 - 2)} (M_1 - M_2)}{\sqrt{(N_1 + N_2) ((N_1 - 1) s_1^2 + (N_2 - 1) s_2^2)}}$$
$$= \frac{\sqrt{25 \times 20 (25 + 20 - 2)} (170.6 - 160.0)}{\sqrt{(25 + 20) ((25 - 1) 19.2 + (20 - 1) 21.1)}}$$
$$= 7.9$$

$$p < 0.001$$

帰無仮説は棄却される
対立仮説が危険率5%で
有意に成り立つ



今後の統計の勉強のために

- 「マンガでわかる統計学」

- 高橋信, トレンド・プロ

- オーム社

統計の初心者向け. わかりやすく考え方を説明

- 「確率・統計入門」

- 小針アキ(日見)宏

- 岩波書店

数学好き大学1,2年レベル. 考え方の面白さを記載