

情報統計

第1回

2022年8月2日 神奈川工科大学



櫻井 望
国立遺伝学研究所
生命情報・DDBJセンター

スケジュール

	2日(火) データの 見える化	3日(水) 検定の これだけは	4日(木) 分散分析と多変 量解析の雰囲気	5日(金) データ準備 発表会
1限				13 補足 自習(課題、質問)
2限	1 ガイダンス PC環境準備、 データの見える化	5 区間推定、分布 との使い方	9 分布の仲間と、 分散分析	14 自習(課題、質 問)
3限	2 統計の基本と 用語	6 t検定	10 相関、主成分 分析	15 発表会
4限	3 プログラミング の基礎	7 検定で注意する こと	11 他の多変量解 析	
5限	4 自習(課題検討、 復習)	8 自習(課題検討、 復習)	12 自習(課題検討、 復習)	

授業で使うサイト

<https://github.com/nsaku/kait2022/wiki>

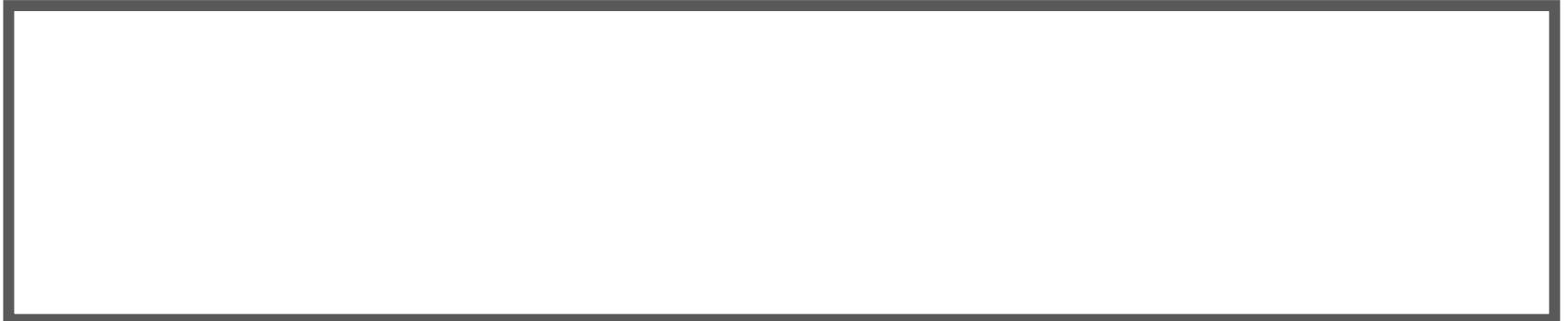
要パスワードサイト

ID:

PW:

ガイダンス

統計の知識は



大学で

- 実験の計画
- データの評価

仕事で

管理栄養士さん

- 調査研究等の設計
- データの評価
- 品質管理・問題解決

生活の中で

- 正しい情報を見抜く
- 話に説得力が出る

学習すること

- データの見える化
- 検定のこれだけは
- 分散分析
- 多変量解析の雰囲気

講義のメッセージ

統計情報を
うのみにしない

そのための力を身につけよう

スケジュール

	2日(火) データの 見える化	3日(水) 検定の これだけは	4日(木) 分散分析と多変 量解析の雰囲気	5日(金) データ準備 発表会
1限				13 補足 自習(課題、質問)
2限	1 ガイダンス PC環境準備、 データの見える化	5 区間推定、分布 との使い方	9 分布の仲間と、 分散分析	14 自習(課題、質 問)
3限	2 統計の基本と 用語	6 t検定	10 相関、主成分 分析	15 発表会
4限	3 プログラミング の基礎	7 検定で注意する こと	11 他の多変量解 析	
5限	4 自習(課題検討、 復習)	8 自習(課題検討、 復習)	12 自習(課題検討、 復習)	

課題のやり方

- 1班 1~5人くらい
- データを集めて解析
- 発表会

課題内容

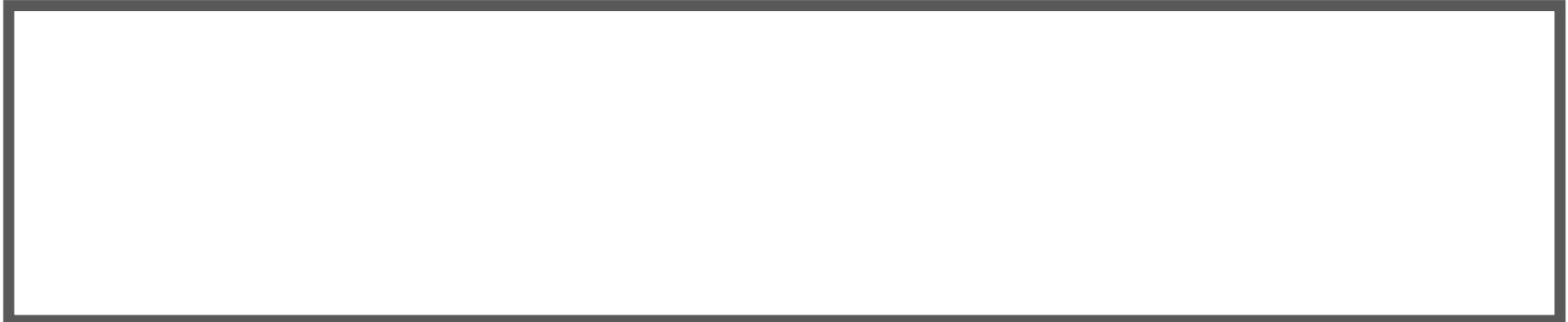
1. 統計データを公開しているサイト（厚生労働省など）から、データを取得して、
2. 統計結果を公開しているサイト（都道府県ランキングなど）から、考察されている情報を得て、さらに元データ入手して、
3. 独自のアンケートを作成して、データを収集し、
4. 1～3に代わるもので、

1～4のいずれか

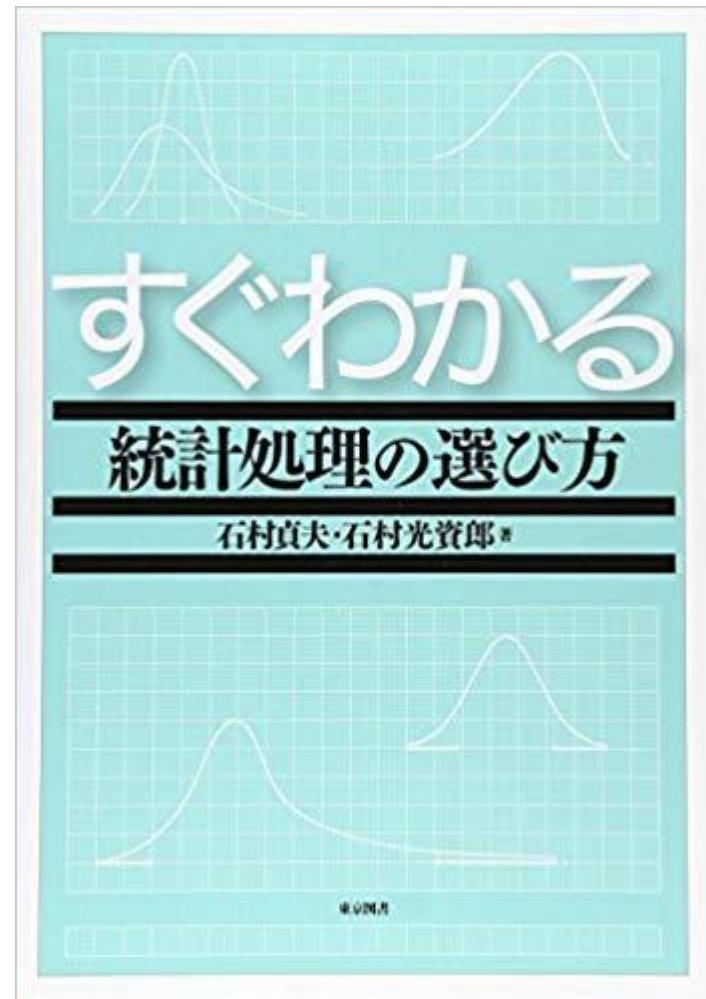
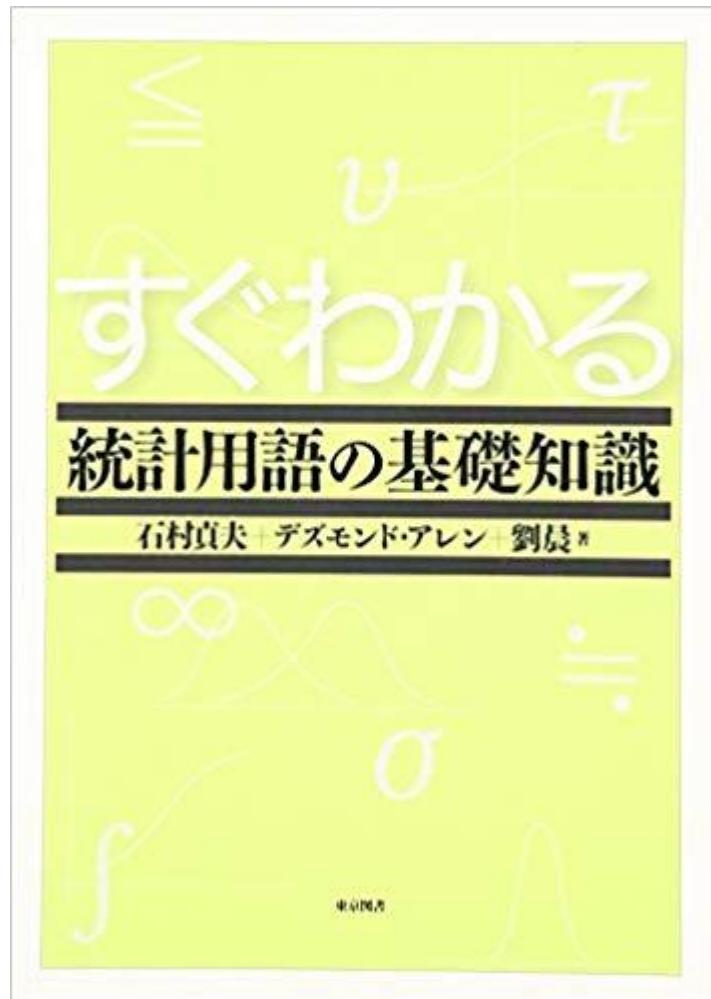


独自のグラフ等を作成し、統計的に解析して、結果・考察を発表する

参考図書は



参考図書は



第一回

データの見える化

学習目標

以下について確認します

- 統計学とは？
- 色々なグラフ

パレート図、ヒストグラム、折れ線グラフ、円グラフ、
帯グラフ、レーダーチャート、ガントチャート

- 統計サイト

厚生労働省、都道府県ランキング、アメリカ農務省、
WHO（世界保健機関）

統計



**統計学を学ぶこと
へのイメージ**

統計って？



出典 小学館 / デジタル大辞泉について 情報 | 凡例

百科事典マイペディアの解説

統計【とうけい】

多数の構成要素（統計単位）からなる**集団**において、各要素の観察によって得た**数値**（統計資料）を処理して集団の**性質・傾向**を明らかにすること。また統計資料をもいう。集団を一時点ととらえる静態統計と、一定期間でとらえる動態統計に分けられ、また統計調査の主体により**官庁統計**と民間統計がある。前者は**統計法**により規制され、特に重要なものは**指定統計**として扱われる。**→統計学**
→関連項目 **グラフ** | **大量観察法**

出典 株式会社平日社 / 百科事典 Wikipediaについて 簡易

大辞林 第三版の解説

とうけい【統計】

(名) スル [statistics]

集団現象を数量的に把握すること。一定集団について、調査すべき事項を定め、その集団の性質・傾向を数量的に表すこと。「-をとる」

出典 三省堂 / 大辞林 第三版について 情報

日本大百科全書(ニッポニカ)の解説

統計

とうけい

statistics英語

Statistik ドイツ語

statistiqueフランス語



統計とは、社会現象の量を反映する数字であり、とくに社会集団の状況を数字によって表現したものである。しかし、現代の統計学における統計的方法の急速な進歩とその普及に伴って、より一般的には、自然現象や抽象的な数値の集団をも含めて、いっさいの集団的現象を数字で表したものと統計とよんでいる。〔泉 俊術〕

統計の本質 [目次を見る](#)

統計の本質とは何かが問われるのは、主として狹義の意味における統計、つまり、社会的集団の状況を語る数字としての統計についてである。自然現象や單に抽象的な数値の集団にかかわる数字については、それらが統計としても意義がある性質は認められなくて問題にはならないからである。

統計の本質は、それがまず社会に実在する固有の事実と結び付き、同時に社会的存在としての集団についての数字データであることである。たとえば、ある人の賃金20万円、ある世帯の月収30万円などと、それが固有の事実に結び付き、また社会現象とみられるものであっても、それが単一の個体についての数字データであるとき、それはまだ統計とはよばない。それらが含まれた集団、つまり、労働者や世帯の具体的のある一定の集団についての数字データ、同種の事例（個体）を集めた集団についての数字が統計である。統計は、統計調査における統計集団の構成（単位、標識、特定の時点など）、時間的順序（時相）に依れば、アレムラムスの概念（統計集団をいかにも構成オブス、すれ目的に的）。

統計って？

集団の状況を
数値で表したもの



目的：集団の〇〇を知りたい

統計学

- データを集める
- 解析する
- 解釈する

ための方法論



結果：集団の〇〇がわかった！

データを集める準備

Google アカウント を作る

- この講義専用
- アンケートを作るのに使う
- プログラミングサイトpaiza.ioの登録用に使う

(注意) 授業が終わったら必ずログアウト

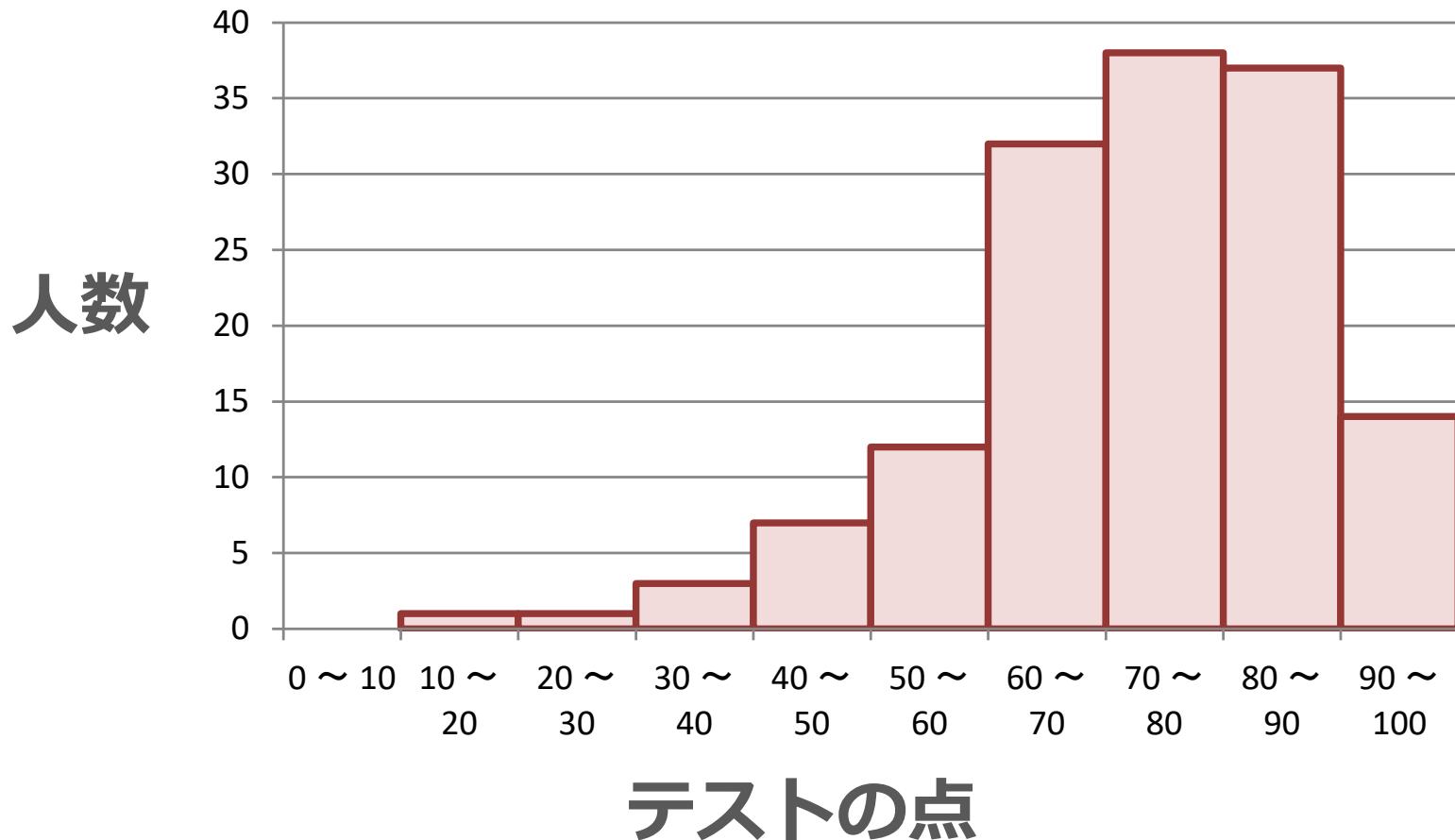
身長データを とってみる

授業のサイト

- リンク公開用（要PW）
- 作ったアンケートの
公開用Googleドキュメント



ヒストグラム (度数分布図)



ヒストグラムの描き方

1. 区間の数を決める

データ数の平方根が目安。四捨五入して整数にする

2. 区間の幅を決める

データの最大値-最小値（データの範囲）を、区間数で割る。四捨五入して、測定の刻み（最小単位）になるようにする

3. 区間の境界を決める

一つ目の区間の開始が、最小値-（測定の刻み÷2）になるように

4. それぞれの区間にに入った数を数える

目的に応じて、細かなアレンジは可

身長データの
ヒストグラムを
作ってみる



様々な見える化手法

- ヒストグラム、パレート図
- 折れ線グラフ
- 棒グラフ
- 円グラフ
- 帯グラフ
- レーダーチャート
- ガントチャート

厚生労働省 統計

で検索

テーマ別を探す

報道・広報

政策について

厚生労働省について

統計情報・白書

所管の法令等

申請・募集・情報公開

ホーム > 統計情報・白書 > 各種統計調査

各種統計調査

- ▼ [統計調査実施のお知らせ（直近10件まで）](#)
- ▼ [統計調査実施予定](#)
- ▼ [最近公表の統計資料](#)
- ▼ [年報等で公表・提供しているもの（直近5件まで）](#)
- ▼ [月報で公表・提供しているもの（直近10件まで）](#)
- ▼ [統計調査公表予定](#)
- ▼ [厚生労働統計一覧](#)
- ▼ [統計要覧一覧](#)
- ▼ [統計情報をご利用の方へ](#)
- ▼ [統計について学ぼう（統計学習サイトのリンク集）](#)
- ▼ [統計関連サイトリンク](#)

統計調査実施のお知らせ（直近10件まで）

▶ [実施のお知らせ一覧へ](#)

- | | |
|--------------|--|
| 2021年8月17日掲載 | ▶ 令和3年パートタイム・有期雇用労働者総合実態調査にご協力ください |
| 2021年8月6日掲載 | ▶ 令和3年度雇用均等基本調査のお願い |
| 2021年8月4日掲載 | ▶ 令和3年社会福祉施設等調査及び介護サービス施設・事業所調査へのご協力を
お願いします |

統計情報・白書

各種統計調査

- ▶ [統計調査実施のお知らせ](#)
- ▶ [統計調査実施予定](#)
- ▶ [最近公表の統計資料](#)
- ▶ [統計調査公表予定](#)

▶ [厚生労働統計一覧](#)

- ▶ [統計要覧一覧](#)
- ▶ [統計情報をご利用の方へ](#)
- ▶ [統計について学ぼう](#)

テーマ別に探す

報道・広報

政策について

厚生労働省について

統計情報・白書

所管の法令等

申請・募集・情報公開

▶ ホーム > 統計情報・白書 > 各種統計調査 > 厚生労働統計一覧

厚生労働統計一覧

- 1.人口・世帯**
- 2.保健衛生
- 3.社会福祉
- 4.介護・高齢者福祉
- 5.社会保障
- 6.社会保障等
- 7.雇用
- 8.賃金
- 9.労働時間
- 10.福利厚生
- 11.人材開発
- 12.労働災害・労働安全衛生・労働保険
- 13.労使関係
- 14.その他

厚生労働省で実施している主な統計調査や業務統計について、その調査内容、調査対象、調査周期、公表予定、実施担当部局及び集計結果表等の搭載場所等をみることができます。

- ▶ [厚生労働統計調査名英訳名称一覧](#)はこちら
- ▶ [厚生労働統計調査・業務統計等体系図（分野別・対象別一覧表）](#)はこちら
- ▶ [厚生労働統計調査・業務統計等体系図（ポイント）\[XSLX形式：217KB\]](#)はこちら

* 印は業務統計

1.人口・世帯

出生・死亡や人口の移動などによる人口変動や世帯の活動などに関するデータを提供しています

- 1.1.人口
- 1.2.世帯
- 1.3.縦断調査（パネル調査）

● 統計情報・白書

▼ 各種統計調査

▶ [統計調査実施のお知らせ](#)

▶ [統計調査実施予定](#)

▶ [最近公表の統計資料](#)

▶ [統計調査公表予定](#)

▼ 厚生労働統計一覧

▶ [地域児童福祉事業等調査](#)

● 統計要覧一覧

● [統計情報をご利用の方へ](#)

● [統計について学ぼう](#)

● [統計関連サイトリンク](#)

1.人口・世帯

出生・死亡や人口の移動などによる人口変動や世帯の活動などに関するデータを提供しています

[1.1.人口](#)

[1.2.世帯](#)

[1.3.総合調査（パネル調査）](#)

1.1.人口

統計・調査名	統計・調査内容
<p>▶ 人口動態調査 NEW 9月7日</p>	<p>出生・死亡・婚姻・離婚及び死産の人口動態事象を把握 本調査は、統計法に基づく基幹統計『人口動態統計』の作成を目的とする統計調査</p>
<p>▶ 人口動態職業・産業別統計</p>	<p>国勢調査年の4月1日から翌年3月31日までの1年間で発生した人口動態事象（出生・死亡・死産・婚姻・離婚）について職業（死亡については産業も含む）を調査し、人口動態事象と社会経済的属性との関連を明らかにする</p>
<p>▶ 人口動態統計特殊報告 NEW 7月30日</p>	<p>人口動態調査を基に、特定のテーマについてとりまとめたもの</p>
<p>▶ 生命表 NEW 7月30日</p>	<p>ある期間における死亡状況（年齢別死亡率）が今後変化しないと仮定したときに、各年齢の者が1年以内に死亡する確率や平均してあと何年生きられるかという期待値などを死亡率や平均余命などの指標（生命閾数）によって表したもの</p>

テーマ別に探す

報道・広報

政策について

厚生労働省について

統計情報・白書

所管の法令等

申請・募集・情報公開

ホーム > 統計情報・白書 > 各種統計調査 > 厚生労働統計一覧 > 人口動態調査

人口動態調査

お知らせ

人口動態統計の調査票について、平成16年、18年、21～29年に^①都道府県からの報告漏れがあることがわかりました。
再集計後の数値については^②「[人口動態統計\(確定数\)の概況](#)」をご覧ください。

- [平成29年人口動態調査で追加集計する統計表について\(募集結果\)](#)
- [人口動態調査における外国人統計に関する意見募集\(募集結果\)](#)

調査の概要

- [調査の目的](#)
- [調査の根拠法令](#)
- [抽出方法](#)
- [調査票](#)
- [調査の方法](#)
- [調査の沿革](#)
- [調査の対象](#)
- [調査事項](#)
- [調査の時期](#)

調査の結果

- [結果の概要 *New* 7月22日](#)

集計・推計方法

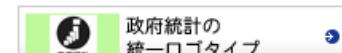
- [正誤情報](#)
- [利活用事例](#)

- [用語の解説](#)
- [利用上の注意](#)
- [統計表一覧](#)(政府統計の総合窓口e-Statホームページへ移動します)

公表予定

統計情報・白書

- [各種統計調査](#)
 - ▶ [統計調査実施のお知らせ](#)
 - ▶ [最近公表の統計資料](#)
 - ▶ [厚生労働統計一覧](#)
 - ▶ [統計要覧一覧](#)
 - ▶ [統計情報をご利用の方へ](#)
 - ▶ [統計について学ぼう](#)
 - ▶ [統計関連サイトリンク](#)
- [白書、年次報告書](#)



人口動態調査

■ 結果の概要

① 人口動態調査の結果

- 確定期 月報年計(概数) 月報(概数)
- 速報 報告書

② その他人口動態統計関連の公表物

- 我が国の人口動態 年間推計

■ 人口動態統計(確定数)の概況

● 月報年計(概数)に修正を加えた確定数です。毎年、調査年の翌年9月頃に公表しています。公表時期については [こちら](#)

● [都道府県からの報告漏れ](#)による再集計を反映した平成16～29年(2004～2017年)の確定数・保管統計表・保管統計表(都道府県編)の各統計表を [e-Stat](#) に掲載しました。

【注意】 概況の過去数値については、平成30年(2018年)以降の概況は再集計による過去数値の修正をおこなっていますが、平成16～29年(2004～2017年)の概況については、過去数値の修正をおこなっていません。口報告漏れによる再集計をおこなった過去数値を確認される場合は、平成30年以降の概況をご覧ください。

- 令和元年 平成30年 平成29年
- 平成27年 平成26年 平成25年 平成24年 平成23年
- 平成22年 平成21年 平成20年 平成19年 平成18年
- 平成17年 平成16年 平成15年 平成14年 平成13年
- 平成12年 平成11年 平成10年 平成9年 平成8年
- 平成7年

ページの一番下に…

■ 人口動態統計(報告書)

● 人口動態統計の報告書です。毎年、調査年の翌年3月に刊行しています。

- 令和元年 平成30年 平成29年

■ 我が国の人口動態

● 人口動態統計の主な内容をグラフ化したものです。

[平成30年我が国の人団動態\(平成28年までの動向\) \[1,522KB\]](#)

● グラフデータ及び統計表を.xls形式でダウンロードできます。

[人口・出生\(P6~14\) \[148KB\]](#) [死亡・乳児死亡\(P15~25\) \[221KB\]](#) [自然増減\(P26~27\) \[61KB\]](#)

[死産・周産期死亡\(P28~29\) \[48KB\]](#) [婚姻・離婚\(P30~36\) \[156KB\]](#) [特殊報告・平均寿命\(P37~39\) \[70KB\]](#)

[統計表\(P42~56\) \[226KB\]](#)

■ 人口動態統計の年間推計

● 月報(概数)と速報の公表数値を用いた推計です。

- 令和2年 令和元年 平成30年 平成29年 平成28年

ISSN 1345-5222



政府統計

平成30年

我が国の人口動態

Vital statistics in Japan

平成28年までの動向
Trends up to 2016



厚生労働省政策統括官(統計・情報政策担当)

DIRECTOR-GENERAL FOR STATISTICS AND INFORMATION POLICY,

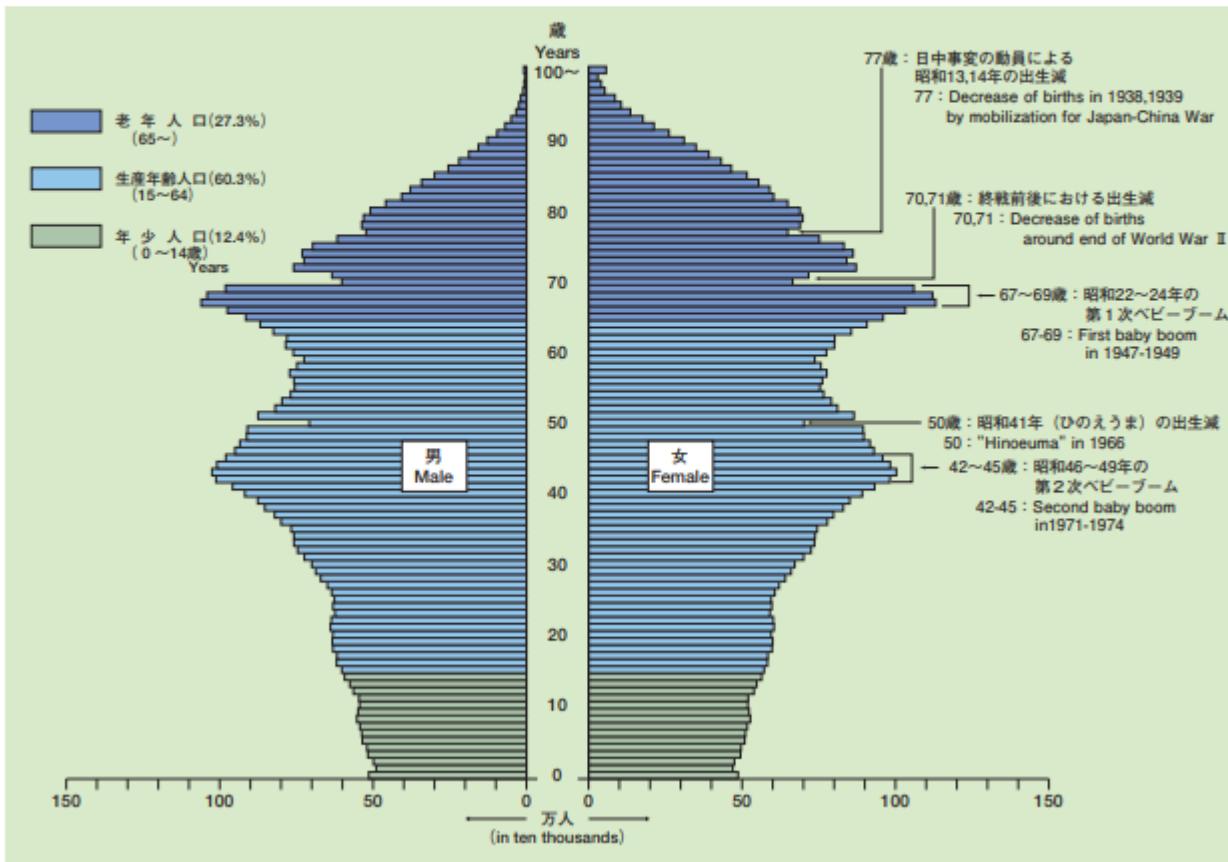
● ヒストグラム

人 口 Population

平成28年の総人口は1億2693万人 老年人口は27.3%

我が国の人団ピラミッド 平成28年10月1日現在

Population pyramid as of Oct.1, 2016

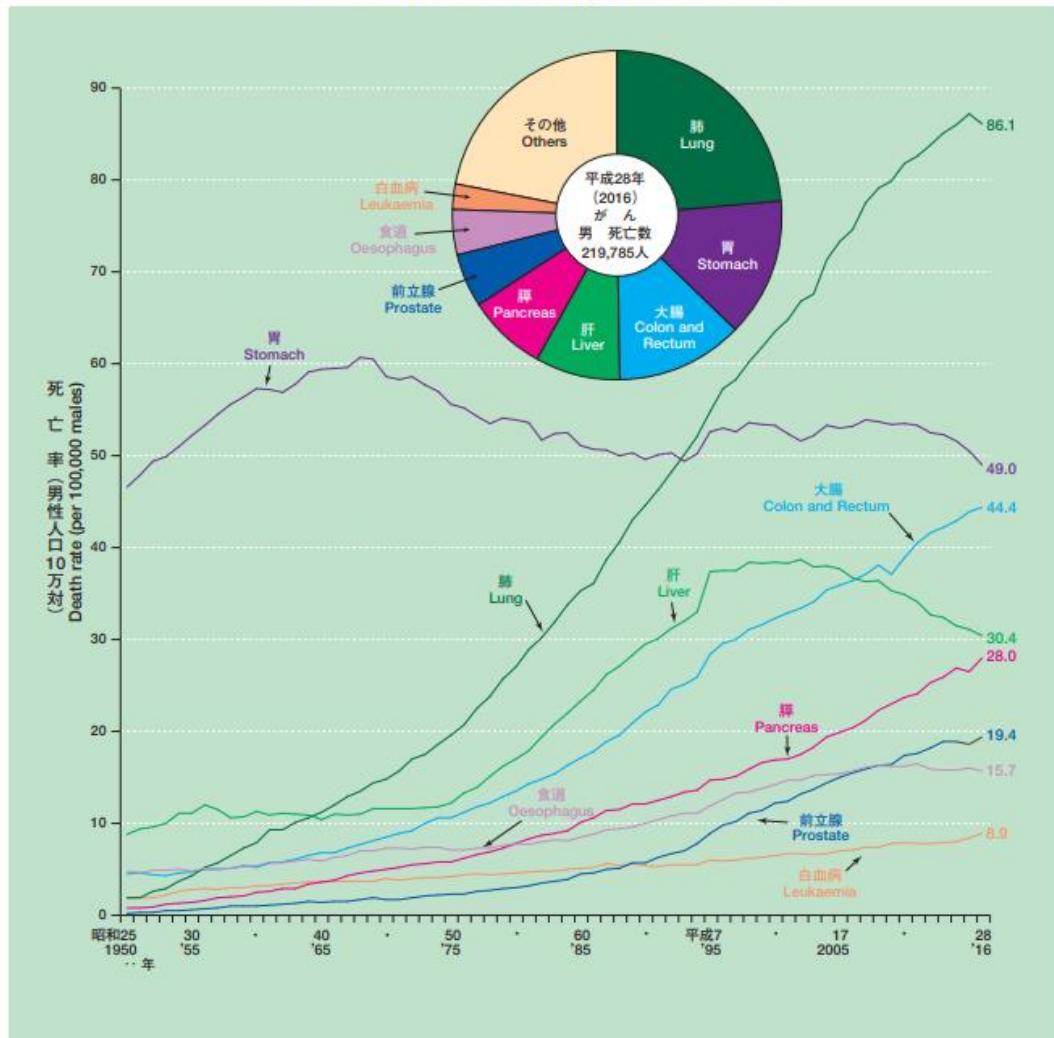


資料：総務省統計局 「人口推計（平成28年10月1日現在）」（総人口）

● 折れ線グラフ

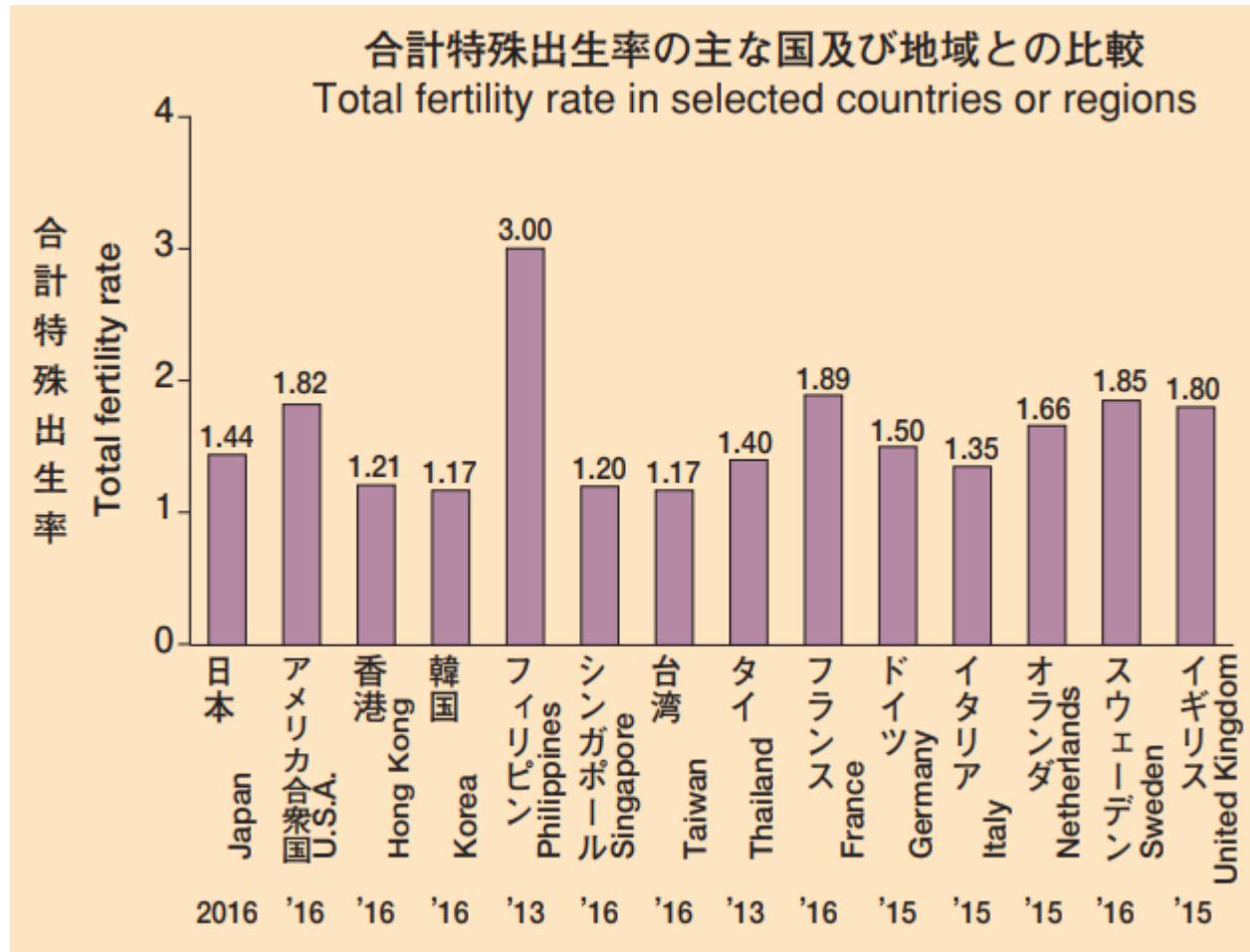
● 円グラフ

部位別にみたがんの死亡率の年次推移、男一昭和25～平成28年—
Trends in death rates for cancers by site, Male, 1950-2016



注：1) 大腸→結腸と直腸 S 状結腸移行部及び直腸（昭和42年まで直腸肛門部を含む。）Colon and Rectum→Colon and rectosigmoid junction and rectum
 2) 肝→肝及び肝内胆管（昭和32年まで胆のう及び肝外胆管を含む。）Liver→Liver and intrahepatic bile ducts
 3) 肺→気管、気管支及び肺 Lung→Trachea, bronchus and lung

● 棒グラフ



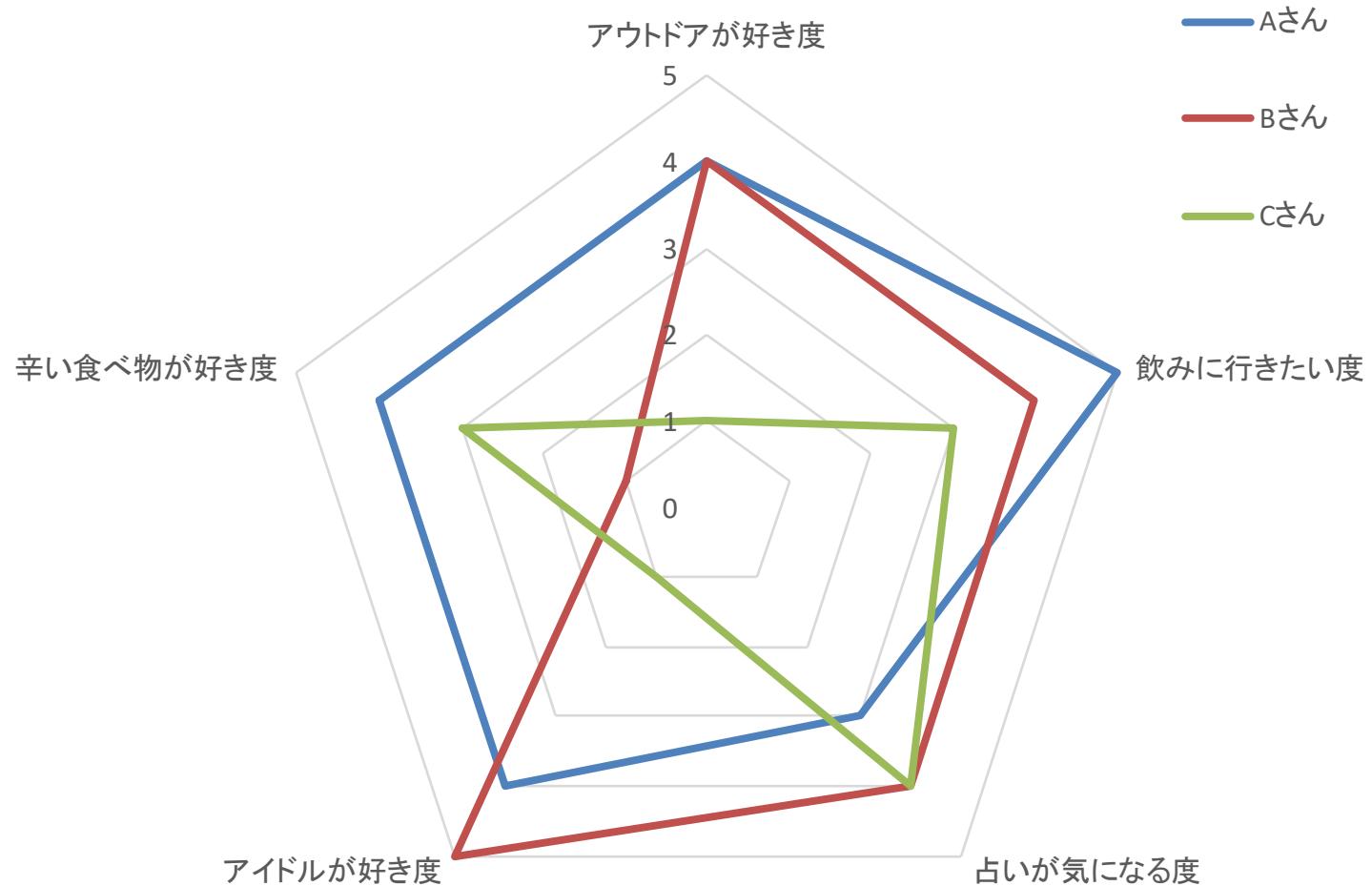
● 帯グラフ

都道府県別にみた年齢3区分別人口割合及び老人人口指数－平成28年－

Percent distribution of population by 3 age groups and aged dependency ratio, by prefecture, 2016

	千人 in thousands	年少人口 0~14歳 Years	生産年齢人口 15~64	老人人口 65~	老人人口指数 Aged dependency ratio
全 国 Total	126,933	12.4	60.3	27.3	45.2
北海道 Hokkaido	5,352	11.2	58.9	29.9	50.9
青 森 Aomori	1,293	11.2	57.8	31.0	53.7
岩 手 Iwate	1,268	11.6	57.2	31.1	54.4
宮 城 Miyagi	2,330	12.2	61.3	26.4	43.1
秋 田 Akita	1,010	10.3	55.0	34.7	63.1
山 形 Yamagata	1,113	11.9	56.5	31.6	55.8
福 島 Fukushima	1,901	11.9	58.7	29.5	50.2
茨 城 Ibaraki	2,905	12.4	60.0	27.6	46.0
栃 木 Tochigi	1,966	12.7	60.6	26.7	44.1
群 馬 Gunma	1,967	12.5	59.1	28.3	47.9
埼 玉 Saitama	7,289	12.4	62.1	25.5	41.0
千 葉 Chiba	6,236	12.2	61.2	26.6	43.4

レーダーチャート



ガントチャート

都道府県別の統計

<https://todo-ran.com/>

都道府県別統計とランキングで見る県民性

トップ 国土・インフラ 社会・政治 産業・経済 文化・くらし・健康 娯楽・スポーツ 店舗分布 その他



月額利用料
0円
振込手数料
無制限

①×

都道府県別統計を比較した都道府県ランキング。
1419 ランキング掲載中

都道府県
ベスト&ワースト

各都道府県の1位と47位だけを一覧表にまとめました。県民性が一目で分かります。

都道府県比較

東京vs大阪、埼玉vs千葉
vs神奈川など任意の都道府県の似たところ、似ていないところを一覧表にまとめました。

著者について

著者：久保哲朗
プロフィール
メール：odomon@gmail.com

Square

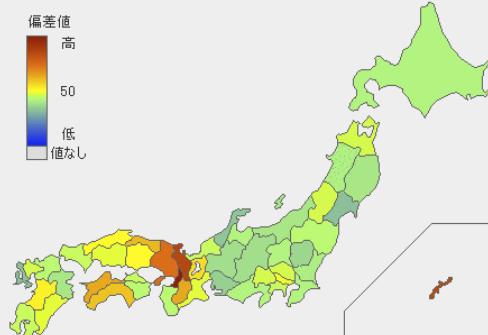
トップ

様々な都道府県別統計を比較したランキング。県民性をデータと都道府県ランキングで表します。
最終更新日:2021-9-1

最新ランキング

お笑い芸人出身地 [2021年 第一位 大阪府]

いいね！ ツイート ブックマーク



記事を探す

検索から探す (googleサイト内検索)

Google 提供
サイト内検索

カテゴリ別全記事一覧

新着順全記事一覧

テーマ別ランキング

- 東西対立型ランキング
東西で高低が分かれるランキング
- 都市地方型ランキング
都市と地方の格差が大きいランキング
- 東京突出型ランキング
東京が突出しているランキング
- ダントツ型ランキング
ダントツの都道府県があるランキング



Square請求書



月額利用料
0円

アメリカ農務省の統計

<http://www.fao.org/faostat/en/>

The screenshot shows the official website of the Food and Agriculture Organization (FAO) of the United Nations. At the top left is the FAO logo. To its right is the text "Food and Agriculture Organization of the United Nations". On the far right is a "Google Custom Search" bar with a magnifying glass icon. Below the header is a horizontal navigation menu with links: "About FAO", "In Action", "Countries", "Themes", "Media", "Publications", "Statistics", and "Partnerships". Underneath the menu are language links: العربية, 中文, English, Français, Русский, and Español. The main content area is titled "FAOSTAT". Below it is a navigation bar with icons for Home, Data, Selected Indicators, Compare Data, Definitions and Standards, and FAQ. A search bar on the right contains the placeholder "Search an Indicator or Commodity".

The screenshot shows the "Food and agriculture data" section of the FAOSTAT website. The background is a chalkboard covered in mathematical and scientific equations. In the center, the text "Food and agriculture data" is displayed above a blue button labeled "Explore Data". Below this, a subtext reads: "FAOSTAT provides free access to food and agriculture data for over 245 countries and territories and covers all FAO regional groupings from 1961 to the most recent year available." A pink rectangular box highlights the "Explore Data" button.

The screenshot shows several sections of the FAOSTAT website. On the left, there's a "Database Updates" section with a thumbnail of server racks and a blue circular button with a white edit icon. In the center, there's a photo of a person wearing a conical hat and carrying produce on their head, with a blue circular button with a white arrow icon below it. On the right, there's a "Bulk Download" section with a large blue download button and a "Database description" section with a link to XML and JSON files.

Data

[DOMAINS](#) [DOMAINS TABLE](#)

▶ Production

▶ Food Security and Nutrition

▶ Food Balance

▶ Trade

▶ Prices

▶ Land, Inputs and Sustainability

▼ Population and Employment

Annual population

Employment Indicators

▶ Investment

▶ Macro-Economic Indicators

▶ Climate Change

▶ Forestry

▶ Discontinued archives and data series



Annual population

[Back to domains](#)[DOWNLOAD DATA](#) [VISUALIZE DATA](#) [METADATA](#)

COUNTRIES **REGIONS** SPECIAL GROUPS

Q Filter results e.g. afghanistan

- World + (Total)
- World > (List)
- Africa + (Total)
- Africa > (List)
- Eastern Africa + (Total)
- Eastern Africa > (List)

Select All Clear All

World > (List)

ELEMENTS

Q Filter results e.g. total population - both sexes

- Total Population - Both sexes
- Total Population - Males
- Total Population - Female
- Rural population
- Urban population

Select All Clear All

Total Population - Both sexes

ITEMS

Q Filter results e.g. population - est. & proj.

- Population - Est. & Proj.

Select All Clear All

YEARS YEAR PROJECTIONS

Q Filter results e.g. 2018

- 2018
- 2017
- 2016
- 2015
- 2014
- 2013

Select All Clear All

1950	1951	1952	1953	1954
1955	1956	1957	1958	1959
1960	1961	1962	1963	1964
1965	1966	1967	1968	1969

Annual population

The FAOSTAT Population module contains time series data on population, by sex and urban/rural. The series consist of both estimates and projections... [Show More](#)

Food and Agriculture Organization of the United Nations

Bulk Downloads

All Data	1.31 MB
All Data Normalized	1.48 MB
All Area Groups	227 KB
Africa	323 KB
Americas	233 KB
Asia	280 KB
Europe	214 KB
Oceania	81 KB

Last Update
December 16, 2019

Related Documents

[Update history](#)

[Definitions and standa...](#)

[Metadata](#)

Annual population

[DOWNLOAD DATA](#)[VISUALIZE DATA](#)[METADATA](#)

Country/Region

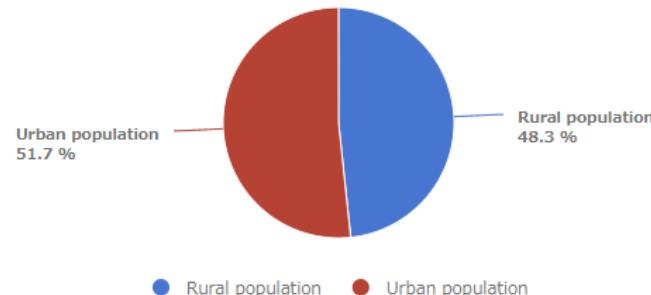
Year

World + (Total)

2010

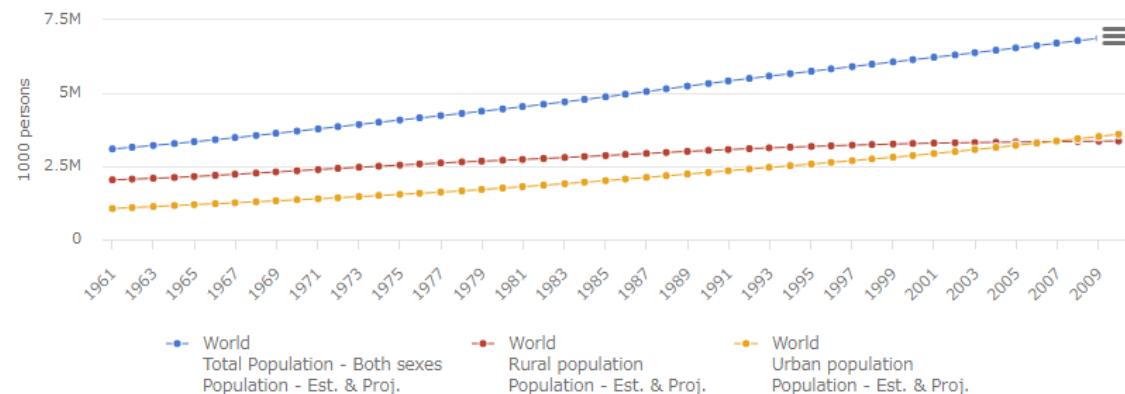
Population composition (area of residence)

World + (Total), 2010



Population dynamics

World + (Total), 1961 - 2010



WHOの統計

<https://www.who.int/>



The screenshot shows the official website of the World Health Organization. At the top, there is a navigation bar with several dropdown menus: "Health Topics", "Countries", "Newsroom", "Emergencies", "Data" (which is highlighted with a pink box), and "About WHO". Below the navigation bar, there are four main sections: "Data at WHO", "Dashboards", "Highlights", and "Reports". The "Dashboards" section contains links to "COVID-19 Dashboard" (which is also highlighted with a pink box) and "Malaria Dashboard". The "Reports" section contains links to "World Health Statistics 2022", "COVID excess deaths", and "DDI IN FOCUS: 2022". In the background, there is a photograph of people working, with the text "Our Work" overlaid. A "Learn more" button is visible in the bottom right corner of the image area.

Health Topics ▾ Countries ▾ Newsroom ▾ Emergencies ▾ Data ▾ About WHO ▾

Data at WHO »

- Global Health Estimates
- Health SDGs
- Mortality Database
- Data collections

Dashboards »

- COVID-19 Dashboard
- Malaria Dashboard
- Health Equity Monitor

Highlights »

- Global Health Observatory
- SCORE
- Insights and visualizations
- Data collection tools

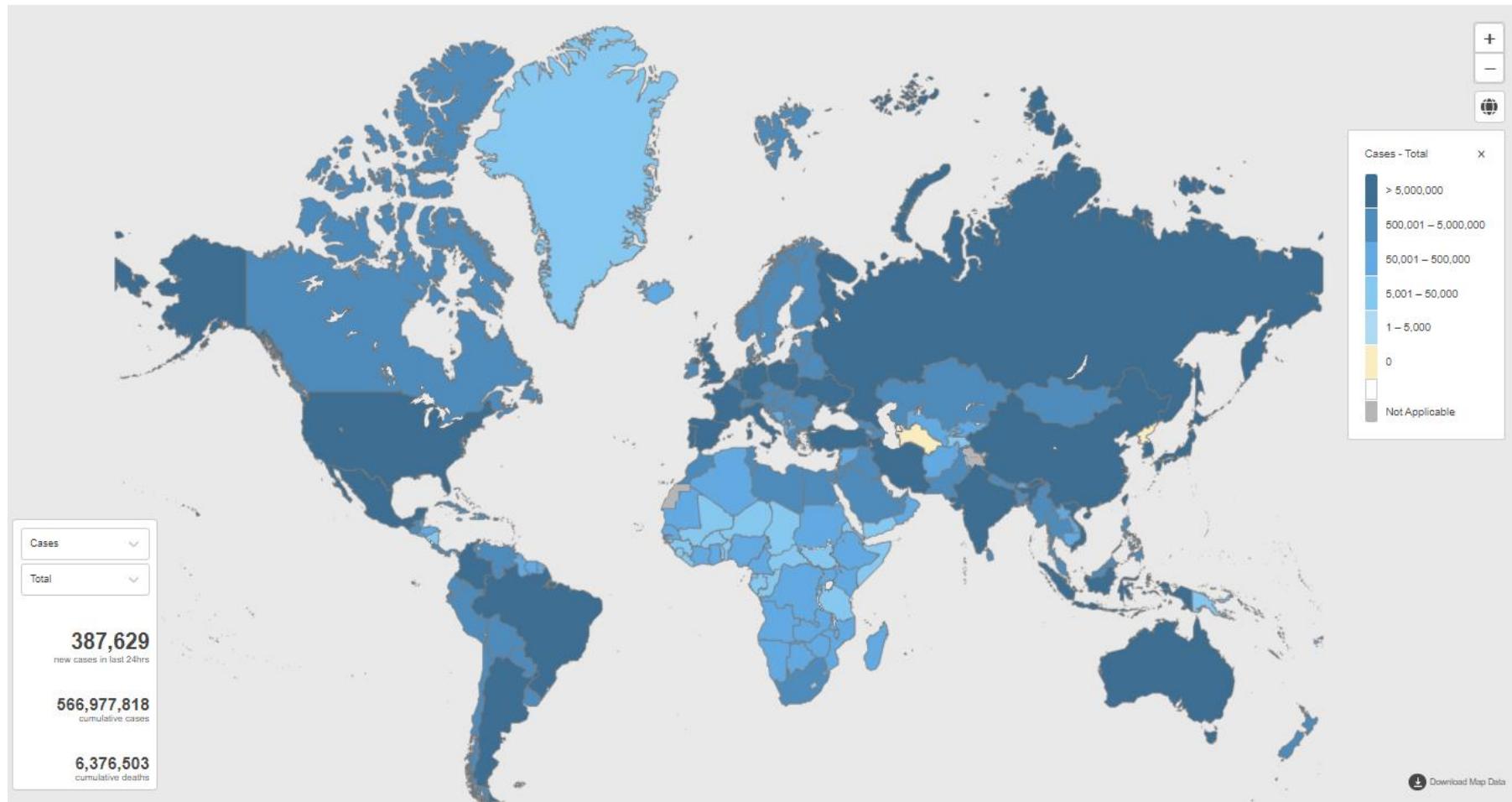
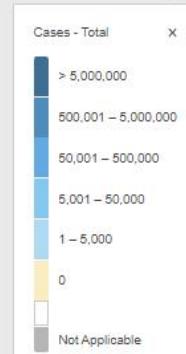
Reports »

- World Health Statistics 2022
- COVID excess deaths
- DDI IN FOCUS: 2022

Our Work

Learn more

Credits +



Globally, as of 5:12pm CEST, 25 July 2022, there have been 566,977,818 confirmed cases of COVID-19, including 6,376,503 deaths, reported to WHO. As of 18 July 2022, a total of 12,219,375,500 vaccine doses have been administered.

課題内容

1. 統計データを公開しているサイト（厚生労働省など）から、データを取得して、
2. 統計結果を公開しているサイト（都道府県ランキングなど）から、考察されている情報を得て、さらに元データ入手して、
3. 独自のアンケートを作成して、データを収集し、
4. 1～3に代わるもので、

1～4のいずれか



独自のグラフ等を作成し、統計的に解析して、結果・考察を発表する

自習

統計サイトの検索、閲覧など

情報統計 第2回

2022年8月2日 神奈川工科大学



櫻井 望
国立遺伝学研究所
生命情報・DDBJセンター

統計の基本と 用語

学習目標

以下の統計用語をマスターします

- 平均値、中央値
- 分散
- 標準偏差
- 母集団
- ランダムサンプリング
- 標本
- 統計的推定
- 母平均、母分散
- 標本平均、標本分散、不偏標本分散
- 分布
- 正規分布（ガウス分布）
- 標準誤差

統計って？

集団の状況を
数値で表したもの



目的：集団の〇〇を知りたい

統計学

- データを集める
- 解析する
- 解釈する

ための方法論



結果：集団の〇〇がわかった！

**第1回の
身長データを使って
解析してみる**

目的：このクラスの人の
身長はどのくらい？

データ



集団の状況を表す
代表的な値を計算

**平均値
中央値**

中心を表す値

**分散
標準偏差**

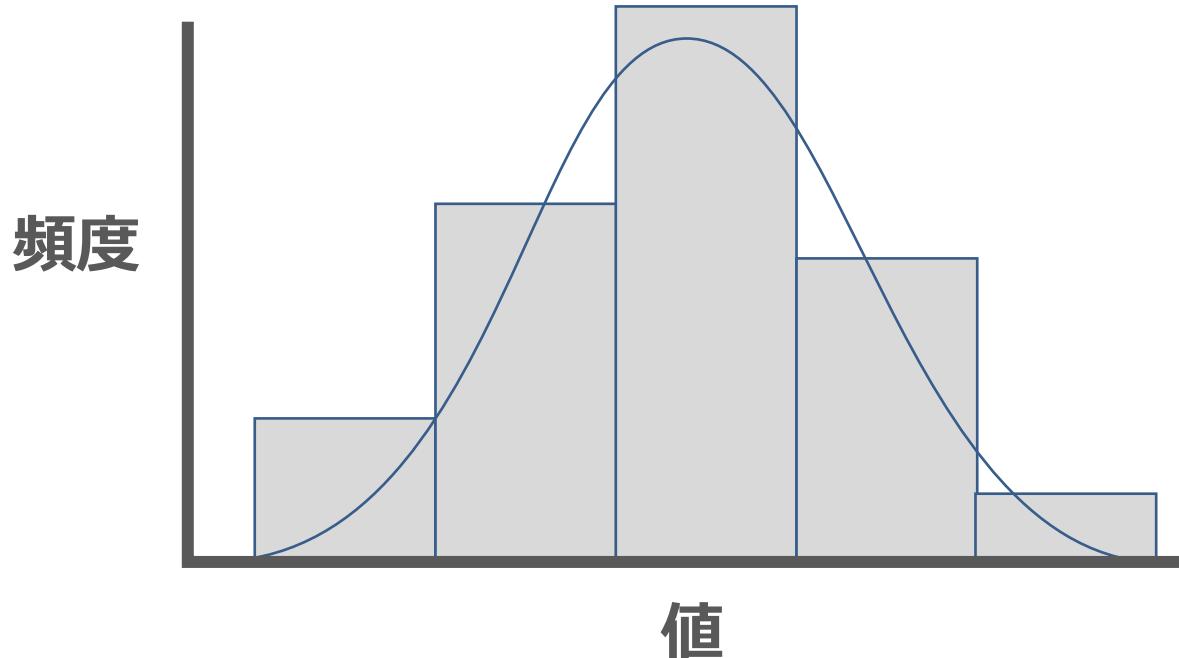
**ばらつきを
表す値**



(基本・基礎) 統計量

分布

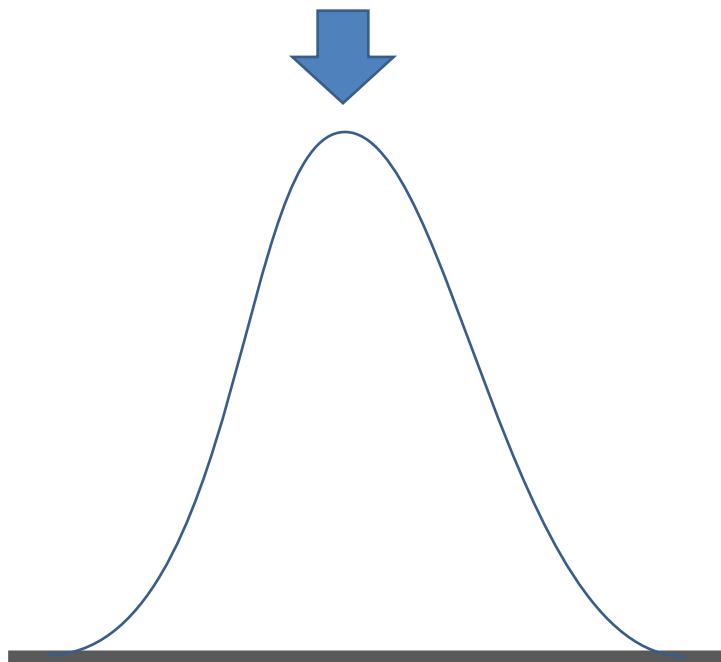
データの散らばり具合



ヒストグラム（頻度分布図）

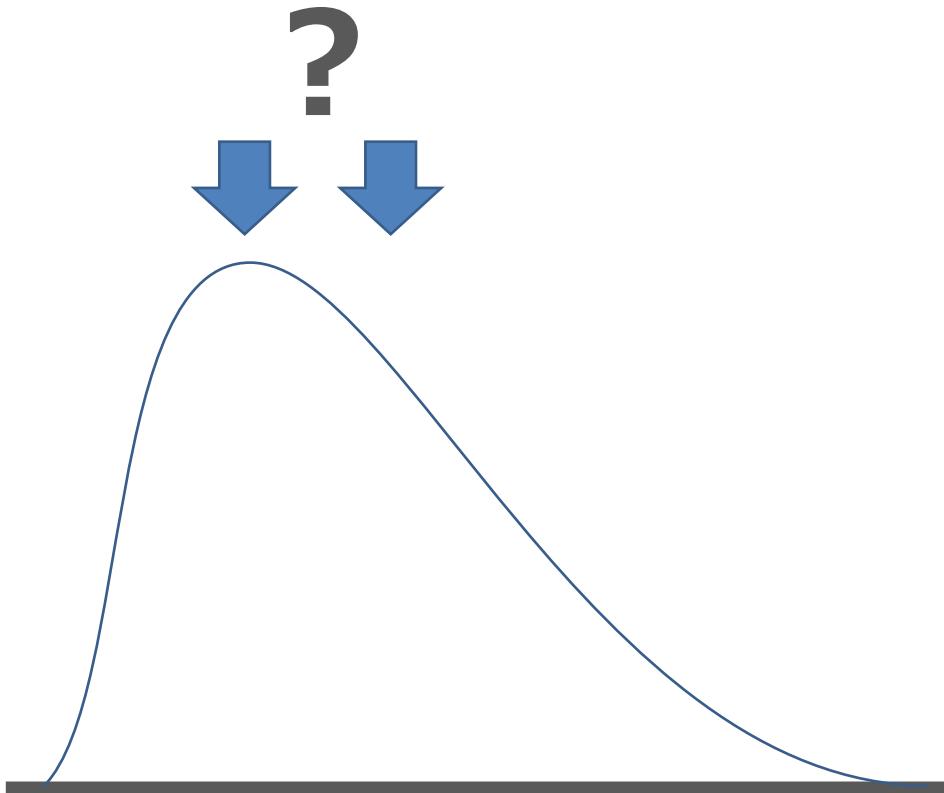
イメージ

データの中心



偏りのないデータ

身長の分布など

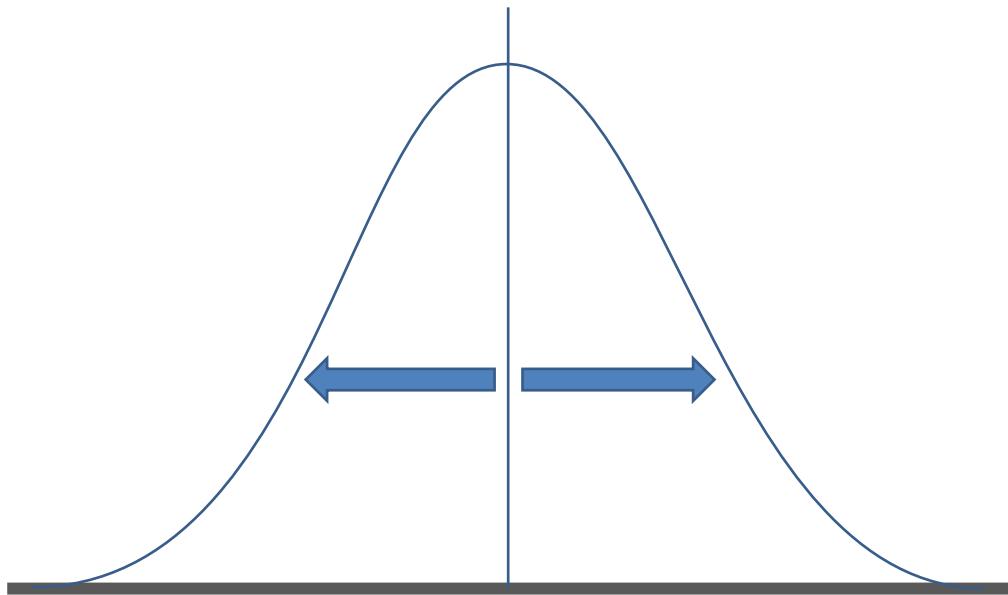


偏っているデータ

体重の分布など

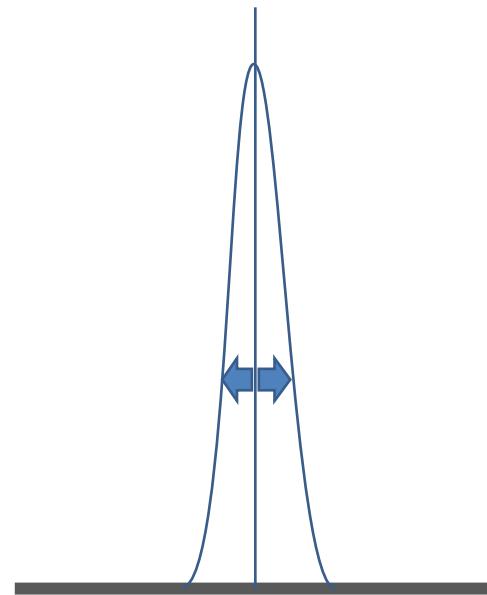
イメージ

ばらつき



ばらつき大きい

中心からの差が
全体的に大きい



ばらつき少ない

中心からの差が
全体的に小さい

平均値

- 合計を計算
- 要素数で割る



中央値

- 小さい順（大きい順）にならべる
- 要素が奇数の場合、真ん中の値を採用
- 要素が偶数の場合、中央の2要素の平均値を計算



ばらつきとは？

分散、標準偏差

平均値からのずれの大きさ

分散

- 平均値を計算
- 各要素-平均値を計算
- その値を2乗
- その平均値を計算



分散

②要素iと平均値の差

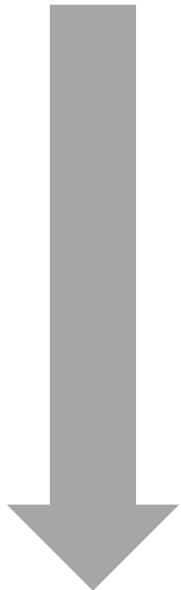
①平均値

⑤要素数nで
割って平均
にする

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{③その2乗}$$

④その全要素(iが1からnまで)の合計

分散…2乗された値



計測値と単位を
そろえるため

平方根を計算

標準偏差



目的：このクラスの人の 身長はどのくらい？

平均

男性

±

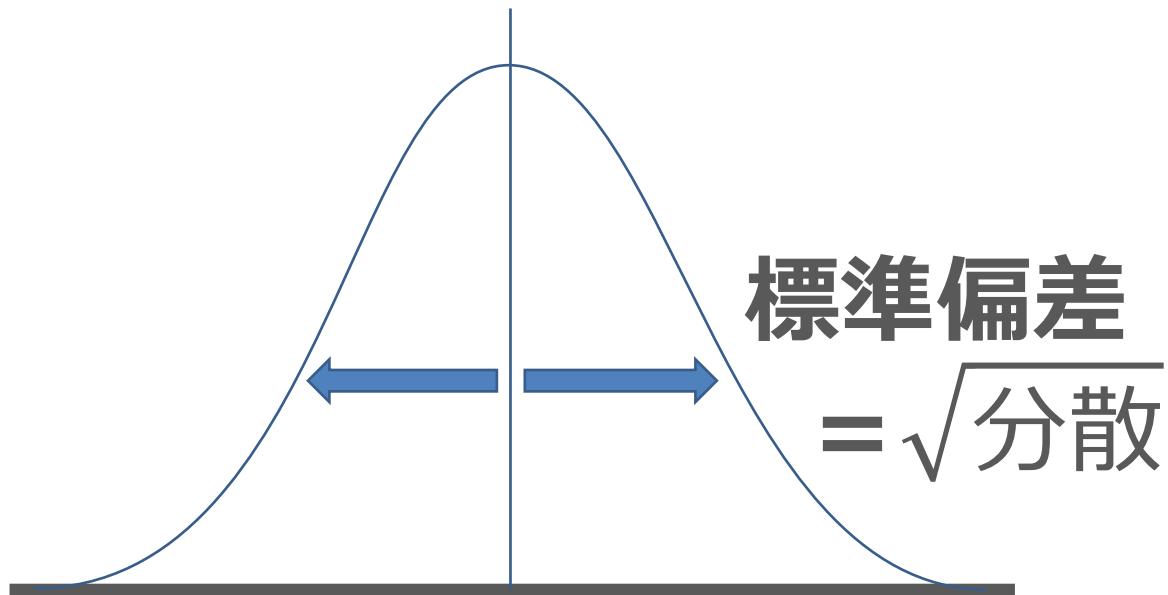
標準偏差

女性

±

イメージ

平均



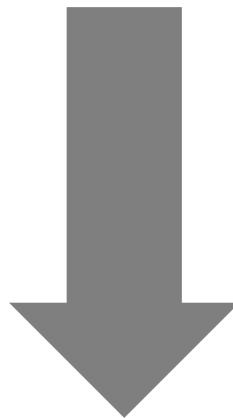
もっと広い
世界が知りたい

目的：このクラスの人の
身長はどのくらい？



目的：日本人の身長はど
のくらい？

全員の身長を測定して計算する



現実的ではない。
コストもかかる

何名かを抜き取り調査する



サンプリング（抽出）

サンプリング

偏りなくランダムに選ぶことが原則

↓
**ランダムサンプリング
(無作為抽出)**

サンプリングされた要素

↓
標本

(サンプル)

今回の目的の場合、

サンプリングされた人のこと

サンプリング前の要素全体



母集団 = 解析の対象

今回の目的の場合、日本人全員のこと

標本の数が多いほど、正確になる！

目的：日本人の身長はどのくらい？



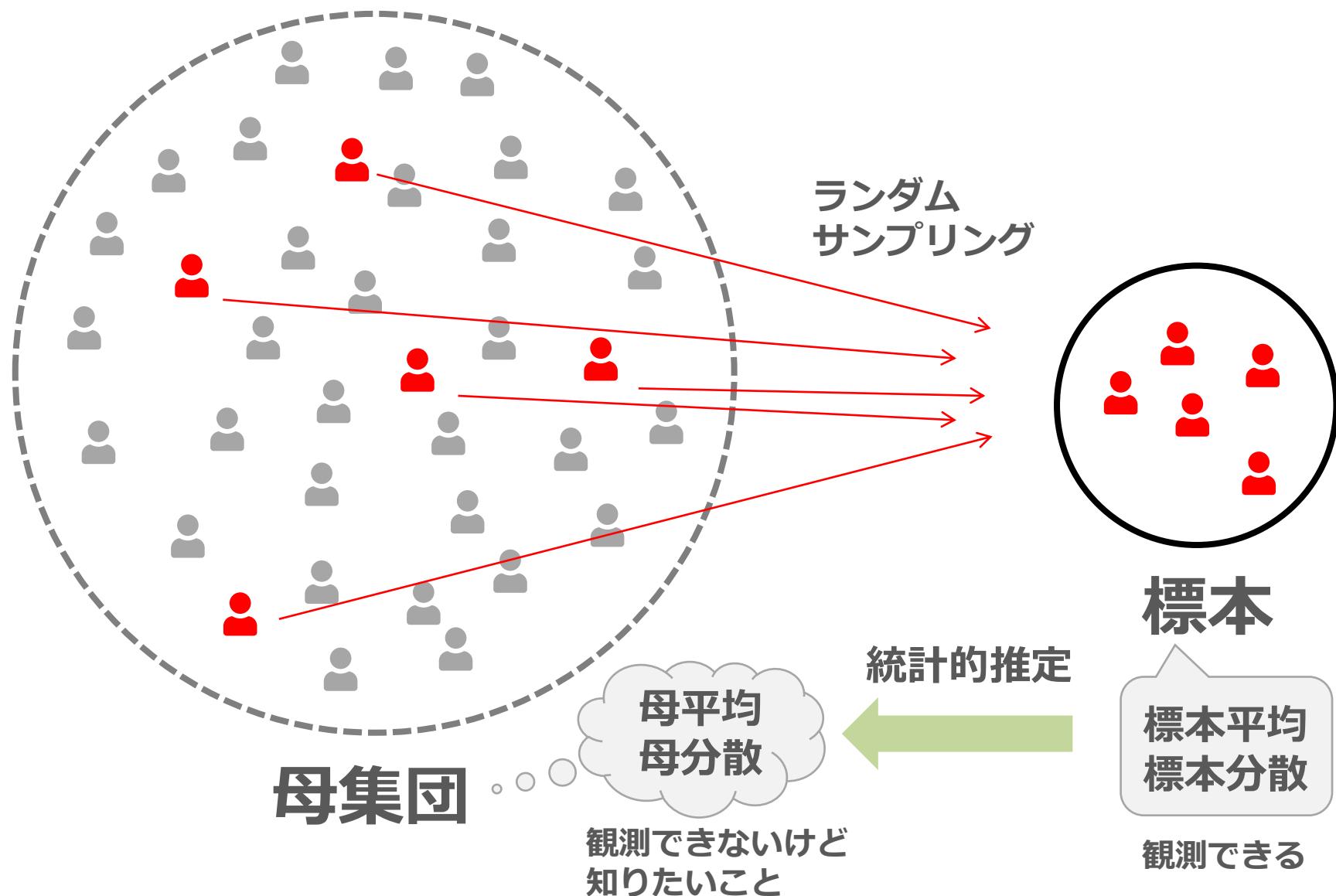
限られた標本から
母集団（日本人全体）の

- 推定の平均値や
- 推定のばらつき

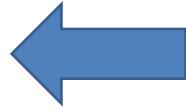
を計算する、という問題

統計的推定

母集団が大きい、あるいは無限で、直接観測できないとき、標本を観測することで、母集団の性質を調べる。



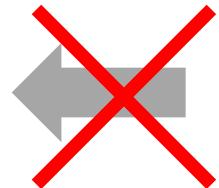
母平均 μ



標本平均 \bar{x}

一致が期待できる

母分散 σ^2



標本分散 s^2

母集団の全標本を観測できる場合は一致するが、
そうでない場合は、**実は一致が期待できない**



一致が期待できる

不偏(標本)分散 v^2

真の値から外れていないことを、
不偏性があると言うので。

標本分散

②要素*i*と平均値の差

①標本平均

⑤要素数nで
割って平均
にする

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

③その2乗

④その全要素(*i*が1からnまで)の合計

不偏(標本)分散

⑤n-1で割る

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

n-1で割る？

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

- 標本の数nが母集団の数N（大きな数）に近づくと、母分散に近くなる

→ 母分散の推定に使える

- 自由度を表している

自由度 = 互いに影響を与えない（独立した）値の数

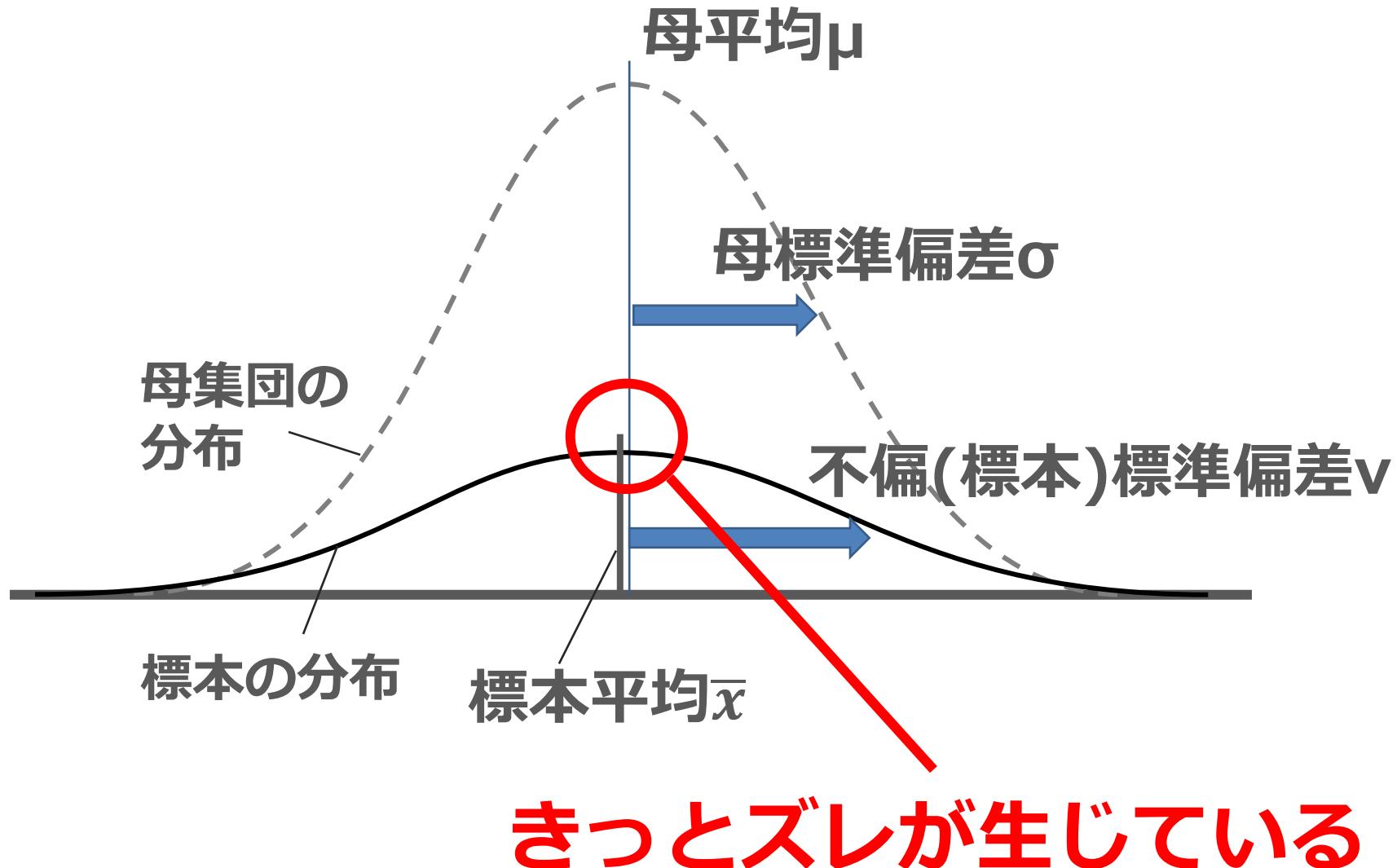
上の式で、一つの観測値 $x(i=a)$ は他と完全に独立ではなく、それ以外の(n-1)個の独立した観測値と平均値 \bar{x} によって求められる。

用語より、 $n-1$ で割っているか どうかに注目

書籍によって、標本分散 s^2 を不偏標本分散（不偏分散）のこととして記述しているものもあります。「（不偏）標本分散」と記述されることもあります。標本を考える時点で、そもそも母集団の推定を前提としていることが多いのです。

n で割っていたら、観測値の話
 $n-1$ で割っていたら、推定値の話 です

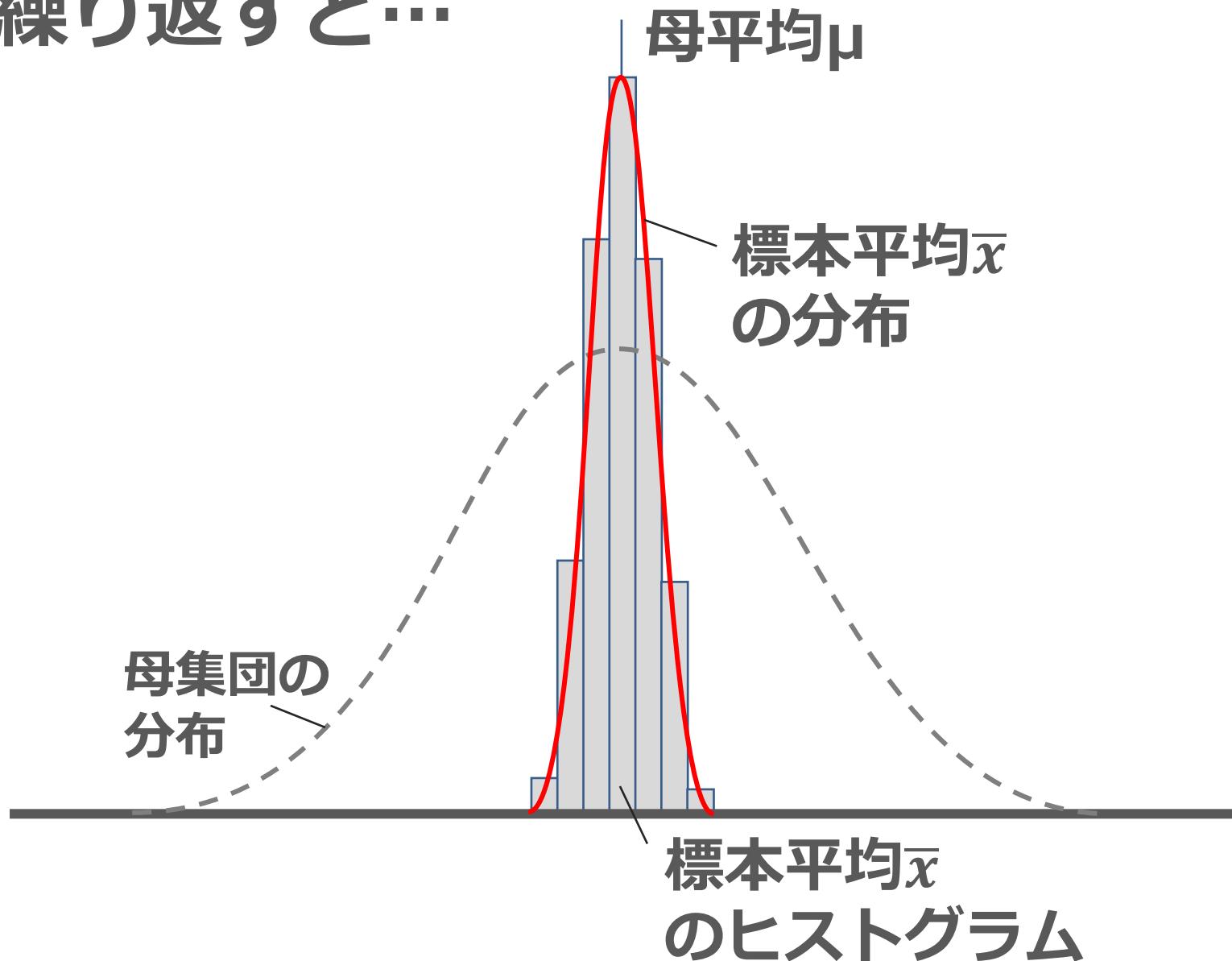
イメージ



誤差

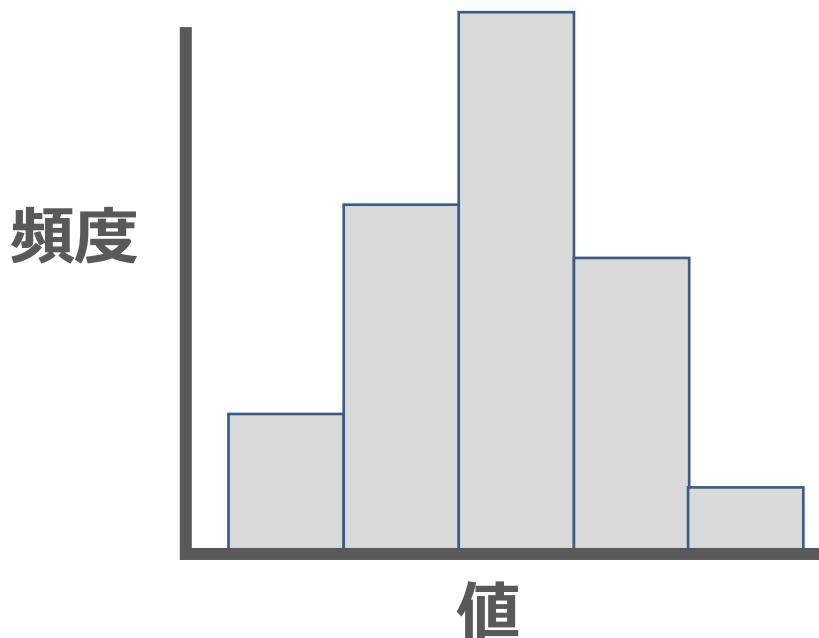
- サンプリング誤差
- 測定誤差

サンプリングして標本平均 \bar{x} を算出して、
を繰り返すと…



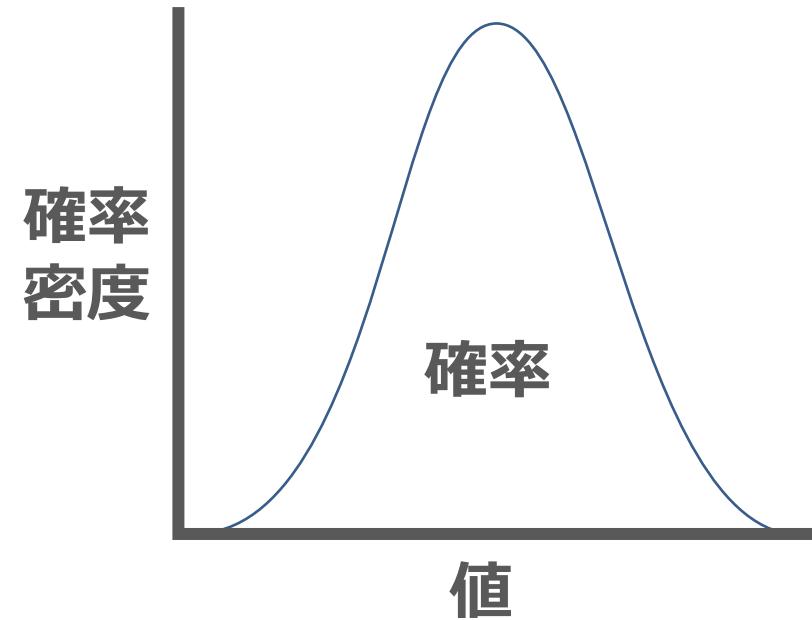
分布

データの散らばり具合



ヒストグラム

観測結果

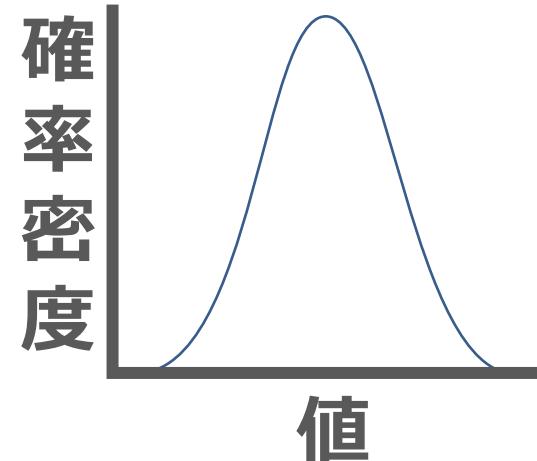


確率密度関数

事象の起こる確率
を表すモデル

正規分布（ガウス分布）

- 平均値を中心で、
- 平均値に近いものが多く、
- 左右に均等な釣り鐘状の分布



均等な確率で生じたばらつき
の場合にとる分布

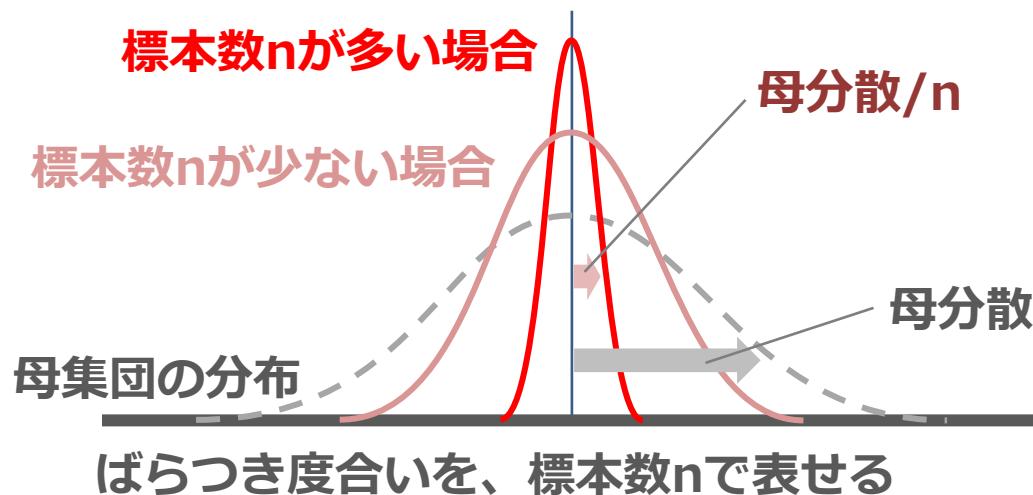
- ✓ 身長の分布
- ✓ 測定誤差の分布
- ✓ 自然界で起こるゆらぎ など

標本平均 \bar{x} の分布

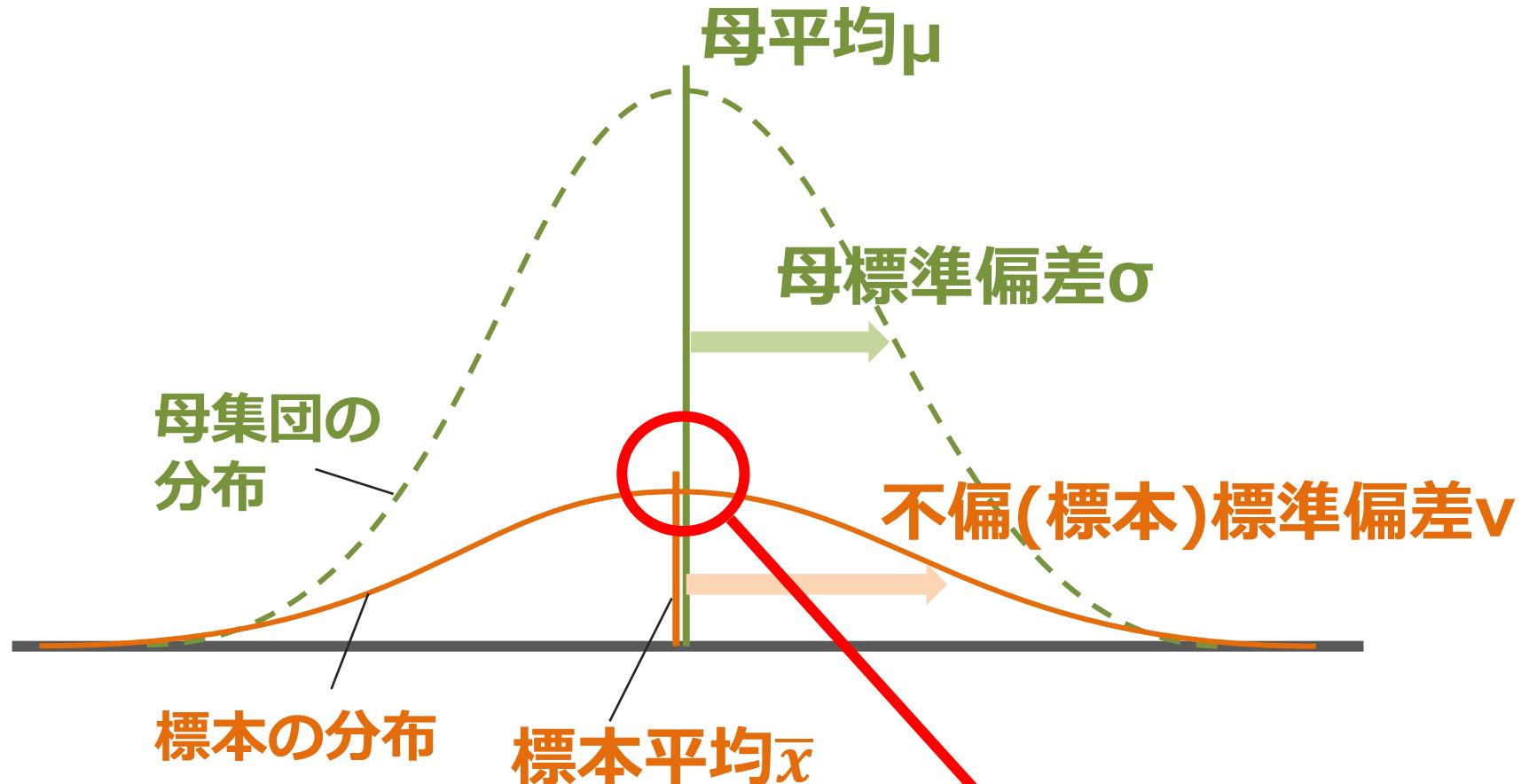
- 正規分布に従う
- 標本の数nが大きいほど、母平均 μ の推定確度は高まり、分散が小さくなる
- 分散は**母分散 σ^2 の $1/n$** になることが知られている

$n=$ 母集団数Nなら、全数検査なので、母平均 μ とのずれはゼロになる。

$n=1$ なら、母集団のうち一つずつを測定するのと同じなので、分散も同じ。



中心極限定理

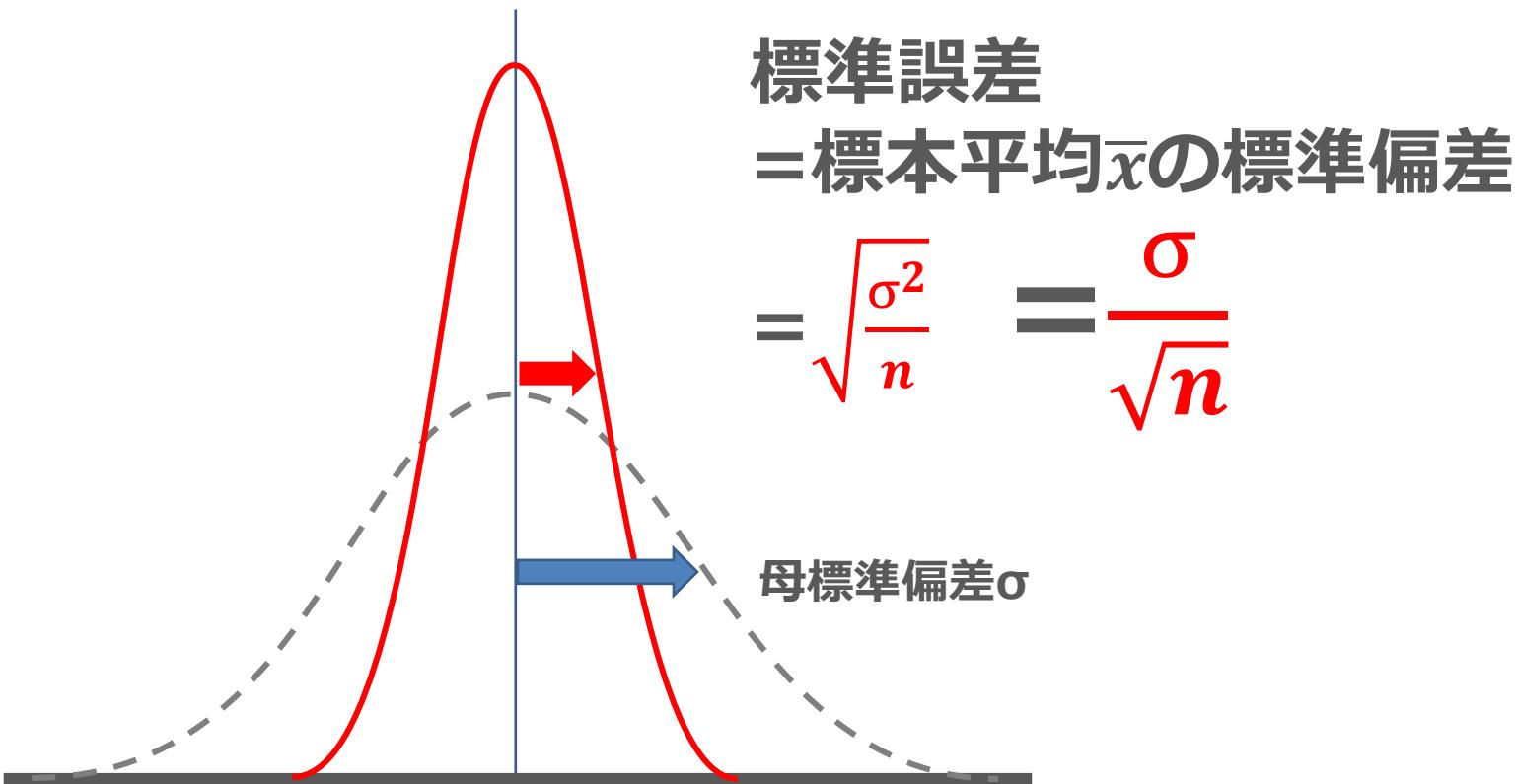


どれだけズってるの？

→ 標本数 n で示せる!!

標準誤差

- 標本平均 \bar{x} の分布の標準偏差のこと。
つまり、母平均 μ の推定値のばらつきを表す
- 母分散 σ^2 の $1/n$ の平方根



標準偏差と標準誤差

論文などでよく見る図

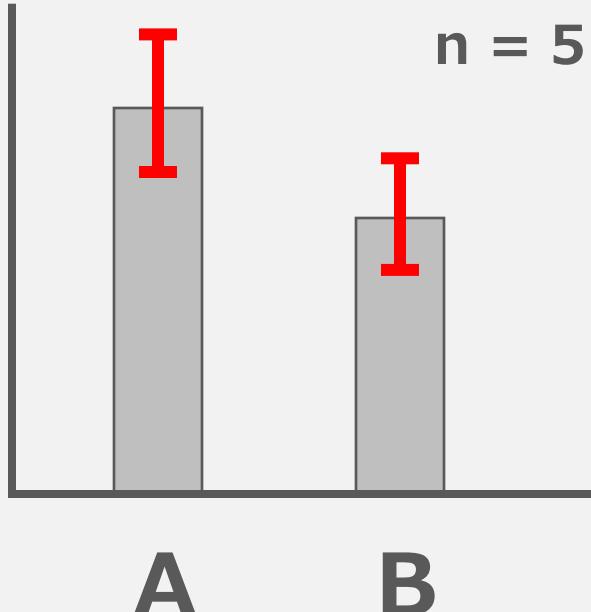


図1 A群とB群の**の違い
それぞれ5個体を測定した。
エラーバーは標準偏差を表す

エラーバーが**標準偏差**



測定した標本 자체の平均値を論じている

エラーバーが**標準誤差**



測定した標本から推定される母集団の平均値について論じている

標準誤差は標準偏差の $1/\sqrt{n}$ なので、エラーバーは短くなり、より明確な差がありそうな見栄えになります。標準誤差を示すことが適當なのかどうかを、正しく判断しながらデータを解釈しましょう。

計算してみよう

このクラスの身長データからいくつかのデータを抜き出し、クラスの身長の平均値を推定してみる



情報統計 第3回

2022年8月2日 神奈川工科大学



櫻井 望
国立遺伝学研究所
生命情報・DDBJセンター

プログラミング の基礎

学習目標

これからプログラミングを始めるときの
取り組み方のコツを学びます

平均値、標準偏差などを計算できるプロ
グラムを作つて動かし、授業に役立てま
す

プログラミング言語の種類

汎用

C, C++,
C#, Java,
Python,
Ruby, Perl

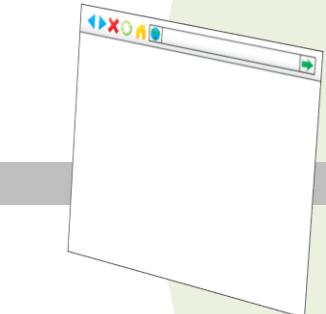
専用

R, Matlab (統計),
Unity (ゲーム)



PC、スマホ

JavaScript,
HTML,
CSS



ウェブブラウザ

インターネット

PHP,
Java,
Python,
Ruby,
Perl,
JavaScript,
Shell Script



ウェブ
サーバー

SQL



データベース
サーバー



学習するメリット

- スタンドアロンでもサーバーサイドでも、広く使える
- ライブラリが豊富
 - ✓ 機械学習、ディープラーニング
 - ✓ 数値計算
- 簡単（覚えやすい）
- はやっている（情報が多く困ったときに調べやすい）

若干のデメリット

- 書き方が、他の言語と少し違っていて独特（他の言語の学習時に少し苦労する、、、かも）
- オブジェクト指向プログラミング（Java, C#, C++などが得意）の習得にはあまり向いていない

プログラミングを
始めるときに
重要なこと

① いつ始めるか？

モチベーションや
必要性が出たとき

やりたい目標を持つことが必須。
「覚えなきや」と本読みから始め
ても、ほぼほぼ身につきません。

②今すぐ知っておく
とよいこと

学習のコツがここに
ある

プログラミングは



どんなことに役立つ？

可能性が広がる

「自分には到底できない」と思わず、つねに、「自分にもできるかも？」という発想になります。

一段上の仕事ができる

たとえば、大量のデータを扱う中で手作業のミスがないかを検証したりなど、仕事のクオリティーが上がります。

論理的な考え方ができる

目的を達成するにはどうすればよいか、工程を細かく分解して考える力がつきます。

学習のコツ

プログラミング言語に 共通する

- ✓ 5つのコア機能
- ✓ 3つの補助機能

をおさえる **これだけ！**

5つのコア機能

1. データを読み込む・書き出す **入出力**
2. データを覚えておく **変数**
3. データを処理する **演算子・命令**
4. データを比較する **比較演算子**
5. 処理の流れを変える **制御構造**

3つの補助機能

1. 一連の処理をひとまとめにして再利用する **関数・サブルーチン**
2. メモを書き込む **コメント**
3. どこにエラーがあるかを知る **デバッグ（バグとり）**

**プログラムの中身は、この8つの組み合
わせだけでほぼ100%できています！**

それぞれのプログラミング言語で、文法が多少違うだけ。

- 今自分が知りたいのは、どの機能のことか
- 本やネットを見ていて、どの機能の話をしているのか

これを意識するだけで、プログラミングは想像以上に短期間で修得できます。

プログラミング ハンズオン講習

使用するサイト



<https://paiza.io>

情報統計

第4回

2022年8月2日 神奈川工科大学



櫻井 望
国立遺伝学研究所
生命情報・DDBJセンター

自習

- 統計サイトのデータを見る・解析する
- プログラミングをする
- アンケートを作つてみる

などで、どんな課題発表にするか
考えてみましょう

情報統計 第5回

2022年8月3日 神奈川工科大学



櫻井 望
国立遺伝学研究所
生命情報・DDBJセンター

昨日

- 図で見える化
- 数値で見える化（統計の基礎）
平均、分散、標準偏差
母平均、母分散…
- アンケートでデータを作る
- Excelの基本操作
- Pythonやってみる

今日

- 見える化した数値や、そこから感じ取れる仮説が、どれだけ正しそうかを、客観的に評価する方法

検定

の考え方を学びます

**有意水準5%で
帰無仮説は棄却されました
よって、 ***です。**

検定

区間推定

分布とその使い方

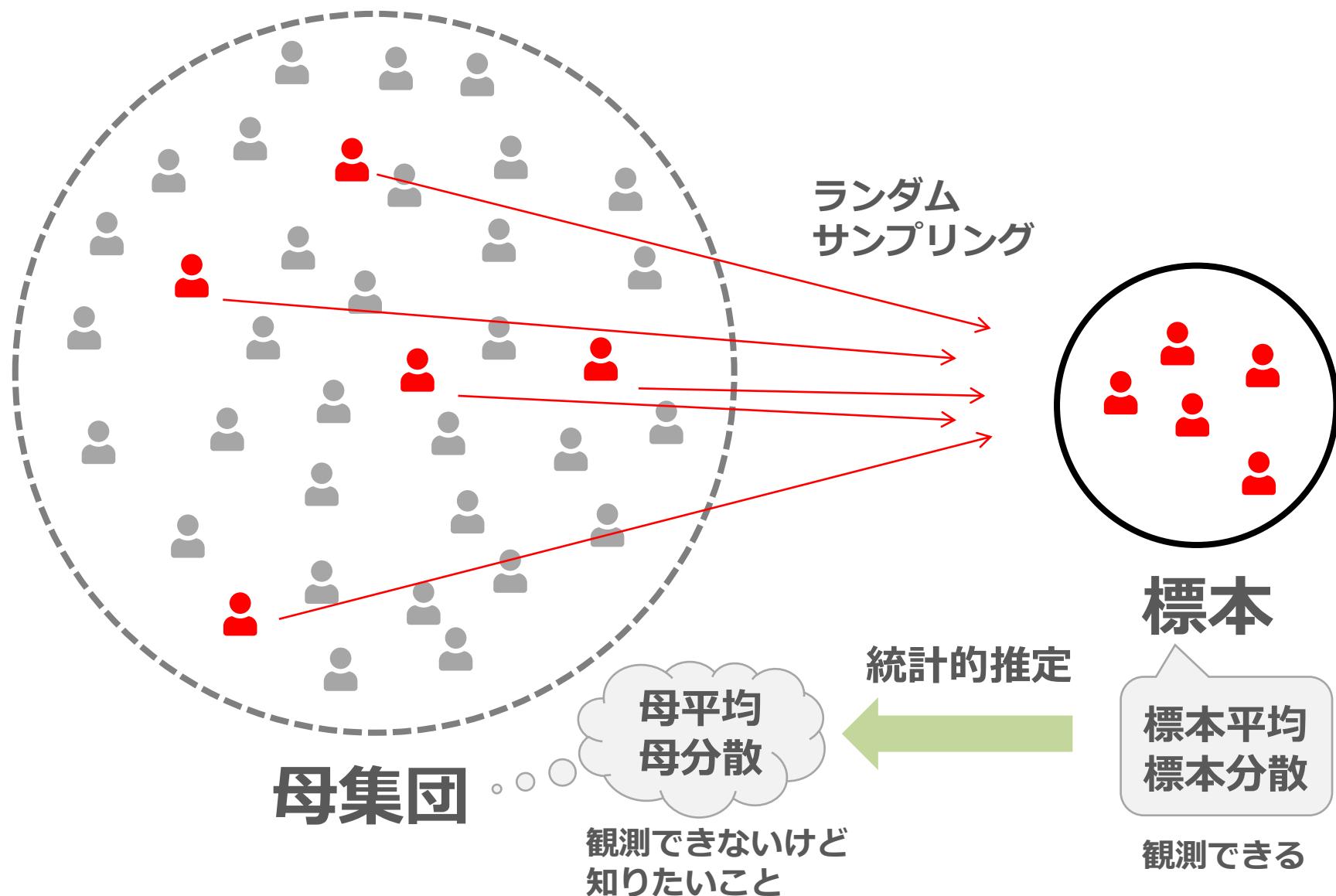
学習目標

区間推定を通じて、**検定などの基本となる分布**と、**その使い方**を身につけます

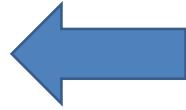
- ✓ 正規分布
- ✓ 標準正規分布
- ✓ t 分布

統計的推定

母集団が大きい、あるいは無限で、直接観測できないとき、標本を観測することで、母集団の性質を調べる。



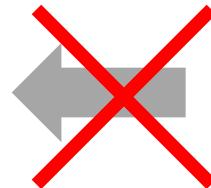
母平均 μ



標本平均 \bar{x}

一致が期待できる

母分散 σ^2



標本分散 s^2

母集団の全標本を観測できる場合は一致するが、
そうでない場合は、**実は一致が期待できない**



一致が期待できる

不偏(標本)分散 v^2

真の値から外れていないことを、
不偏性があると言うので。

点推定



「母平均 μ はこの値」、「母分散 σ^2 はこの値」のように、一つの代表値を決める方法

区間推定

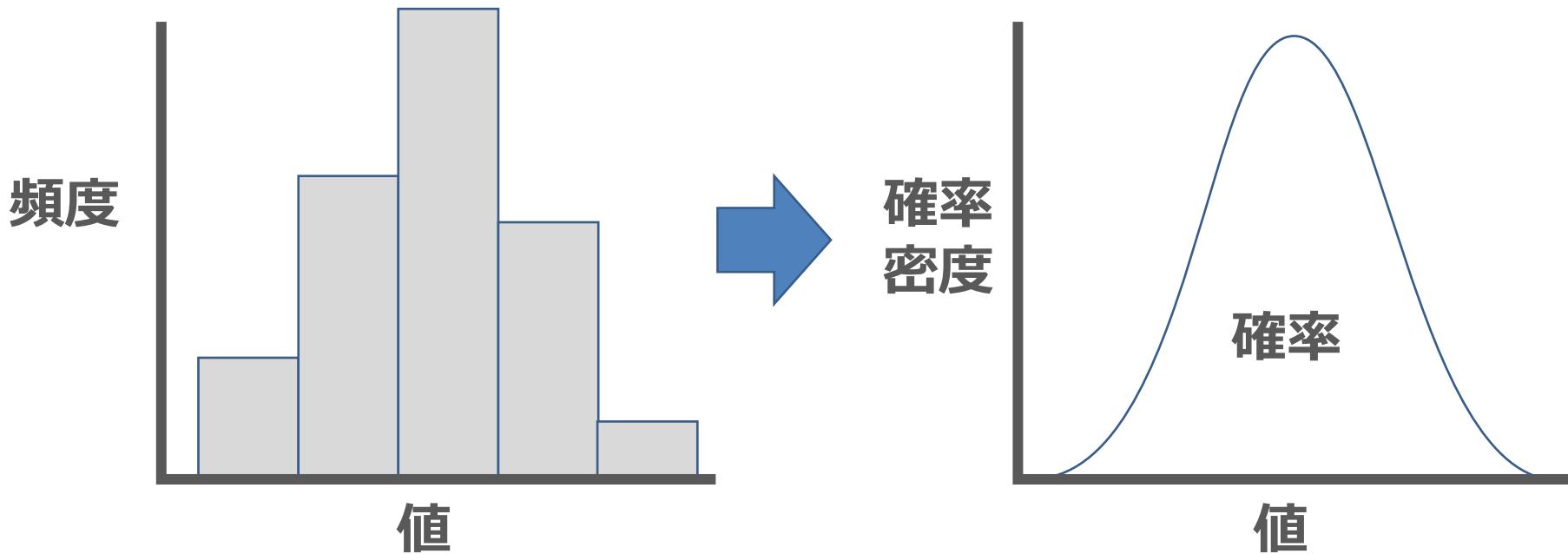


「神奈川県の男子の平均身長は、信頼係数95%で170.2 ~174.6 cmである」のように、幅を持たせて表現する方法

標準正規分布

分布

データの散らばり具合



ヒストグラム

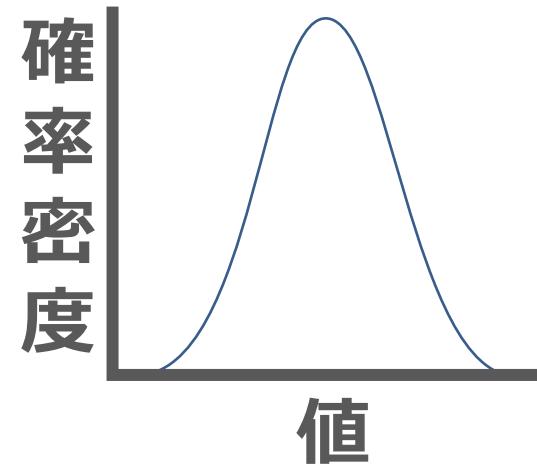
観測結果

確率密度関数

事象の起こる確率
を表すモデル

正規分布（ガウス分布）

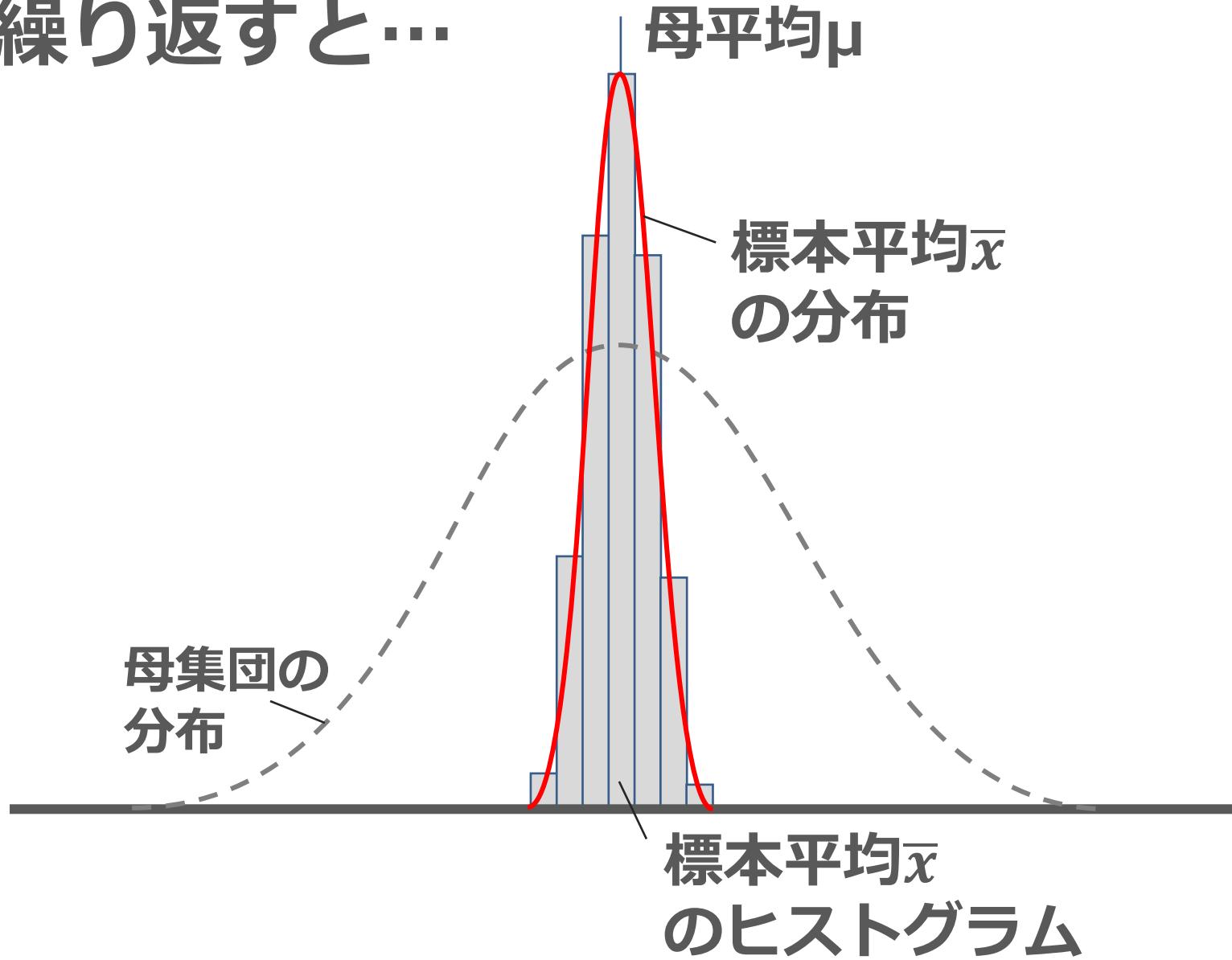
- 平均値を中心で、
- 平均値に近いものが多く、
- 左右に均等な釣り鐘状の分布



均等な確率で生じたばらつき
の場合にとる分布

- ✓ 身長の分布
- ✓ 測定誤差の分布
- ✓ 自然界で起こるゆらぎ など

サンプリングして標本平均 \bar{x} を算出して、
を繰り返すと…

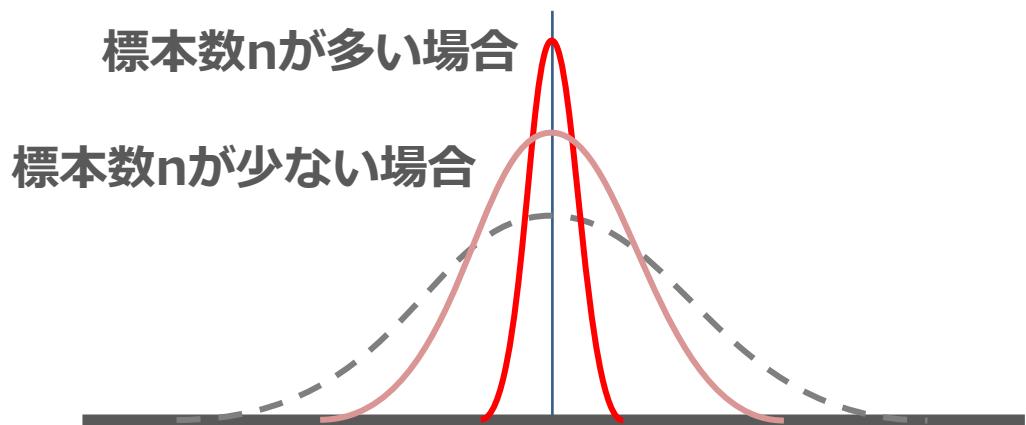


標本平均 \bar{x} の分布

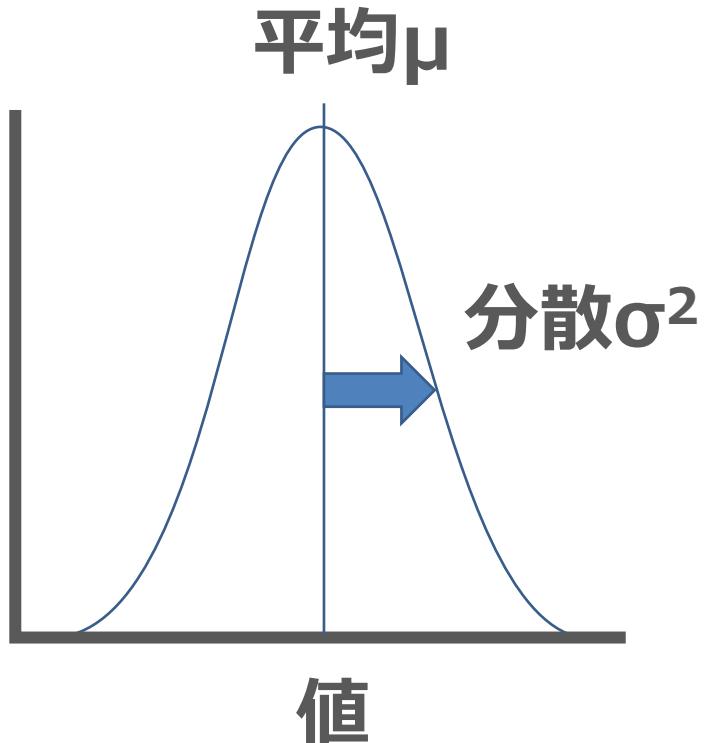
- 正規分布に従う
- 標本の数nが大きいほど、標本平均 \bar{x} の推定確度は高まり、分散が小さくなる
- 分散は母分散 σ^2 の $1/n$ になることが知られている

$n=1$ なら、母集団のうち一つずつを測定するのと同じなので、分散も同じ。
 $n=$ 母集団数Nなら、全数検査なので、母平均 μ とのずれはゼロになる。

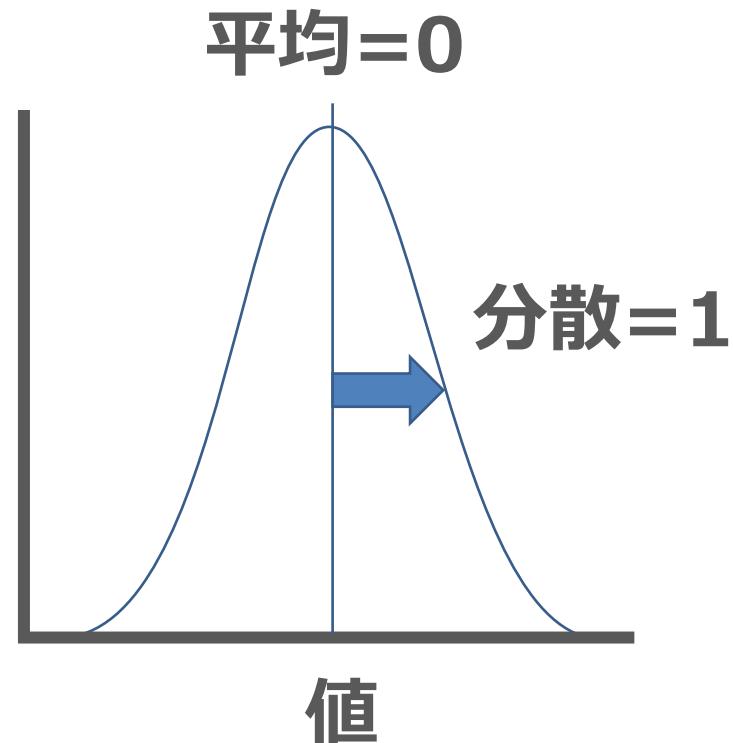
中心極限定理



正規分布



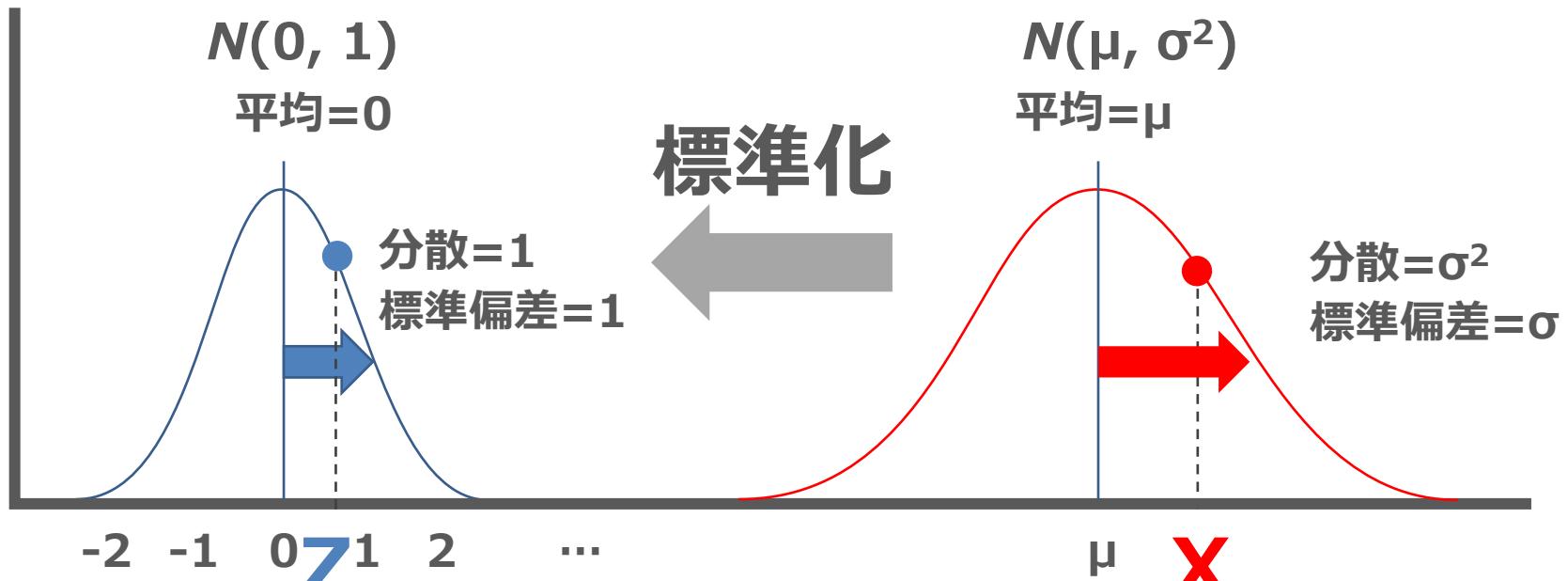
標準正規分布



標準化 (Z変換)

$N(\mu, \sigma^2)$ の正規分布に従う変数Xについて、

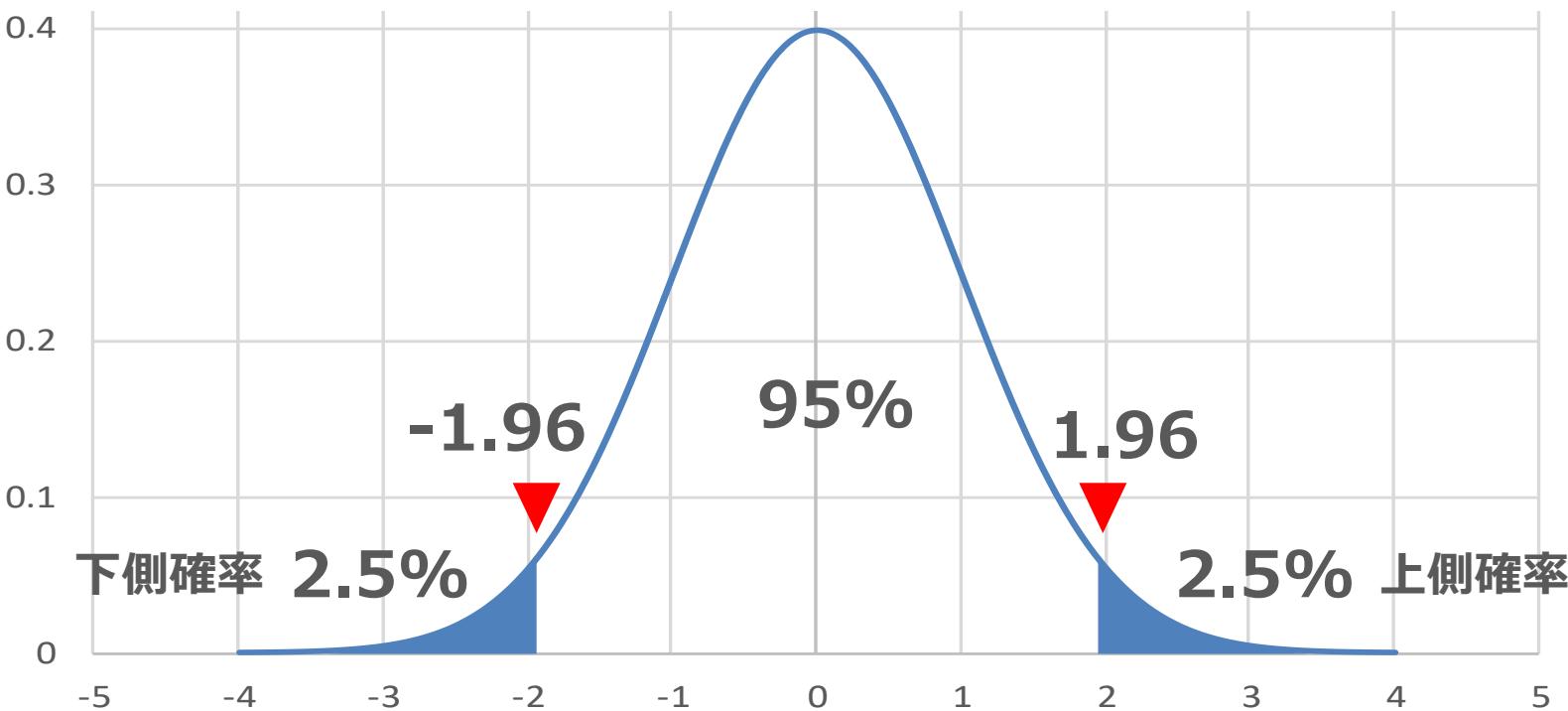
$$Z = \frac{X - \mu}{\sigma} \quad \text{と変換すると、標準正規分布になる。}$$



中央を μ ずらして、幅を1に合わせているだけ！

標準正規分布

- 形が一定なので、ある値より外側の面積が計算できる
例) 1.96以上なら2.5%
- 逆に言えば、外側がある面積（事象がおこる確率）となる境界値を求めることができる
- 左右対称。上側（下側）の面積を上側（下側）確率という



標準正規分布表

上側確率をあらかじめ
計算したもの

Excelでは、
NORM.S.DIST関数
NORM.S.INV関数
で求められる

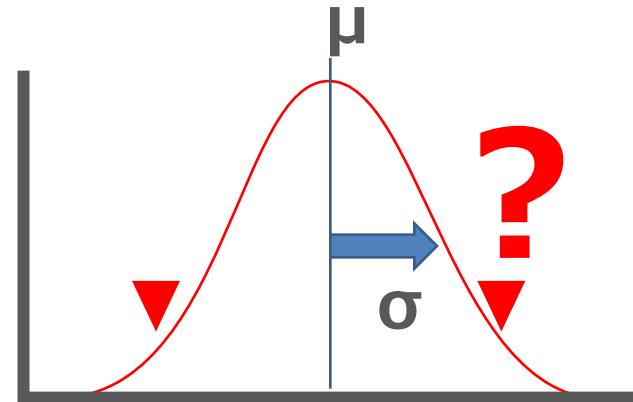
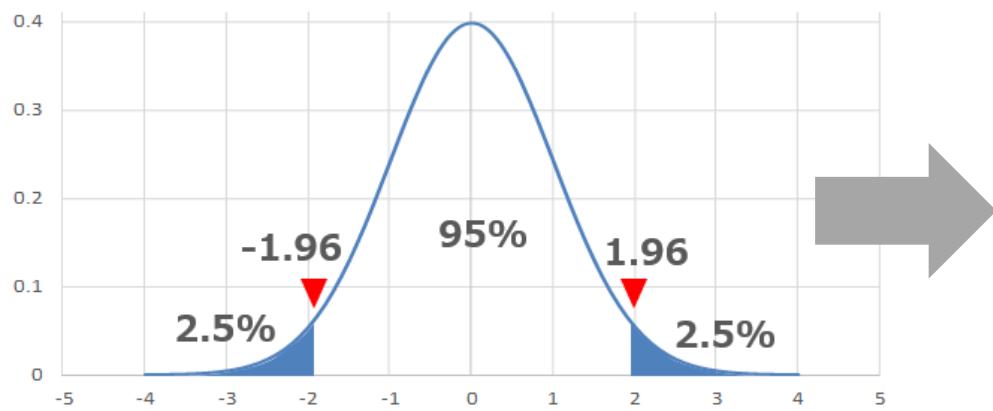
出典

<https://to-kei.net/distribution/normal-distribution/table/>

u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	
0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414	
0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465	
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591	
0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827	
0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207	
0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760	
0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510	
0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476	
0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673	
0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109	
1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786	
1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702	
1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853	
1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226	
1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811	
1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592	
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551	
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673	
1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938	
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330	
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831	
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426	
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101	
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842	
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639	
							0.00539	0.00523	0.00508	0.00494	0.00480
							0.00402	0.00391	0.00379	0.00368	0.00357

区間推定の考え方

- ある事象が正規分布に従っていることが分かっており、
- 平均 μ 、分散 σ^2 が分かっているなら、
- 標準正規分布における $a\%$ のときの境界値を用いて、その正規分布の境界値を求めればよい
- その境界値間を、 $a\%$ 信頼区間という



標準化

$$Z = \frac{X - \mu}{\sigma}$$

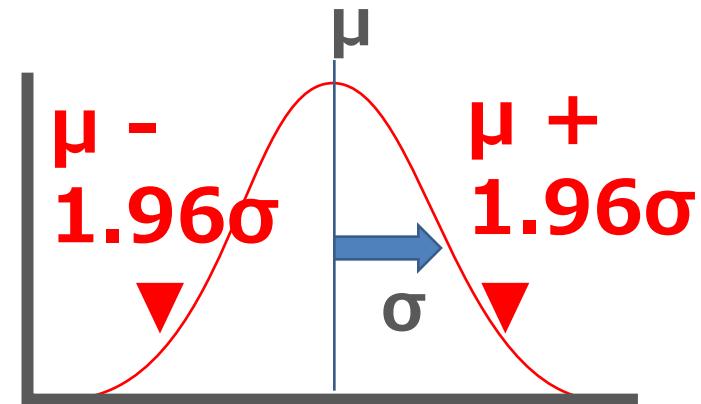
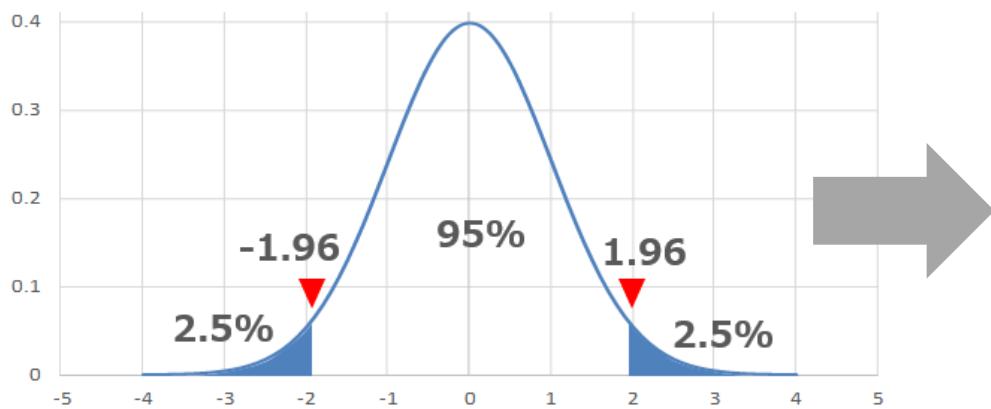


標準化の逆

$$X = \mu + Z\sigma$$

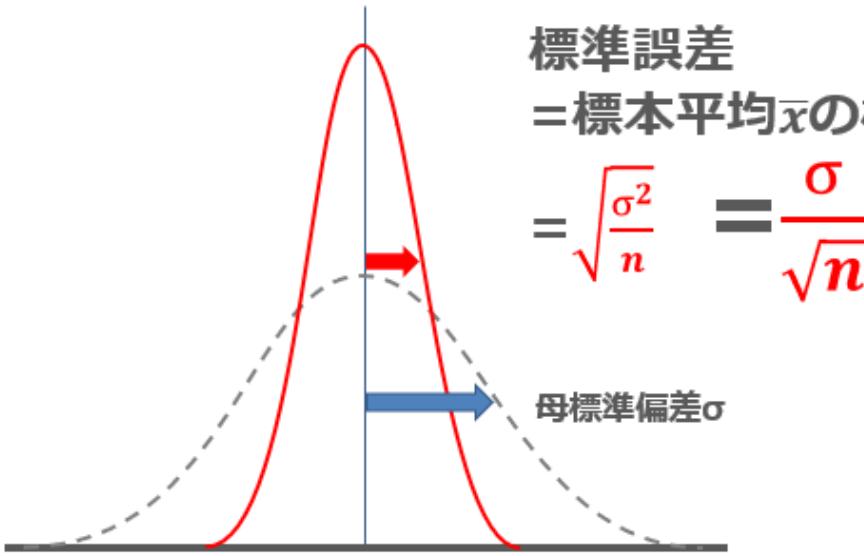
例) $Z = 1.96$ なら、

$$X = \mu + 1.96 \sigma$$



標準誤差

- 標本平均 \bar{x} の標準偏差のこと。
つまり、母平均 μ の推定値のばらつきを表す
- 母分散 σ^2 の $1/n$ の平方根



μ 推定値 : \bar{x}

標準偏差 : $\frac{\sigma}{\sqrt{n}}$

を当てはめる

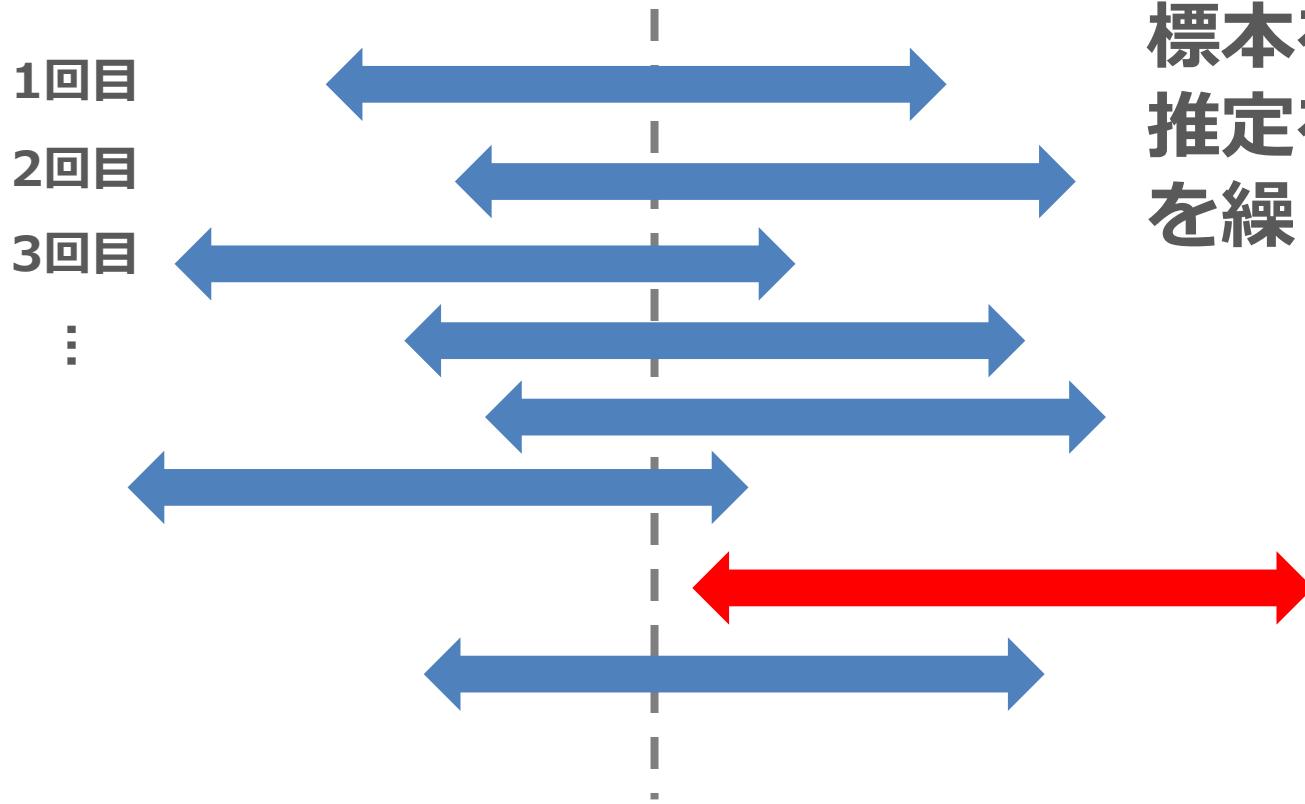
区間推定のまとめ

- 母平均 μ の推定値： 標本平均 \bar{x}
- 推定値の標準偏差： 標本平均の標準偏差 $\frac{\sigma}{\sqrt{n}}$
- の場合、95%信頼区間は、以下で求められる

$$\bar{x} - 1.96 * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 * \frac{\sigma}{\sqrt{n}}$$

意味：「母集団から標本を取り出して95%信頼区間を求めるという作業を100回やったとき、母平均がその区間に含まれるのが95回になる」

イメージ



真の母平均
(変わらない)

標本を取って区間
推定をする、
を繰り返したとき

95%の確率で正しいが、
5%は外している

一般化すると

区間推定 (分散既知の場合)

母平均 μ 、母分散 σ^2 の正規分布する母集団から抽出したn個の標本から求められる、a%信頼区間は以下となる。

$$\bar{x} - A * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + A * \frac{\sigma}{\sqrt{n}}$$

ここでAは、標準正規分布表から、

$$\alpha \text{ (信頼係数)} = (100-a)/2/100$$

で求められる境界値

ただし…

$$\bar{x} - 1.96 * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 * \frac{\sigma}{\sqrt{n}}$$

母分散 σ^2 は不明な場合がほとんど

母平均 μ が不明（推定したい）のに母分散 σ^2 だけ分かっているって、
どういうこと？ そんな状況はほとんどない！



母分散が不明な場合は、正規分布ではなく、*t*分布を用いて同様に考える

t 分布

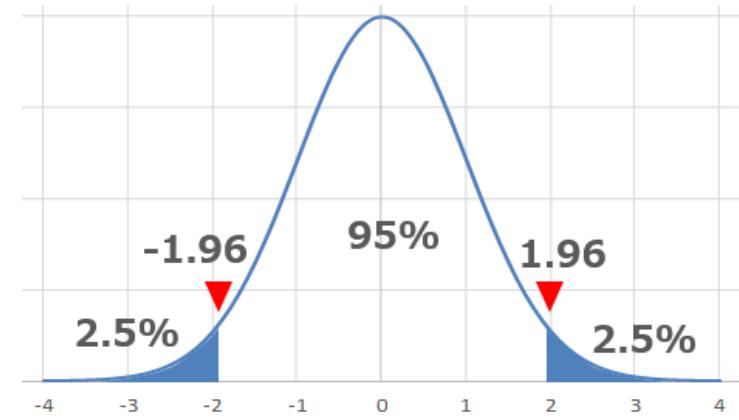
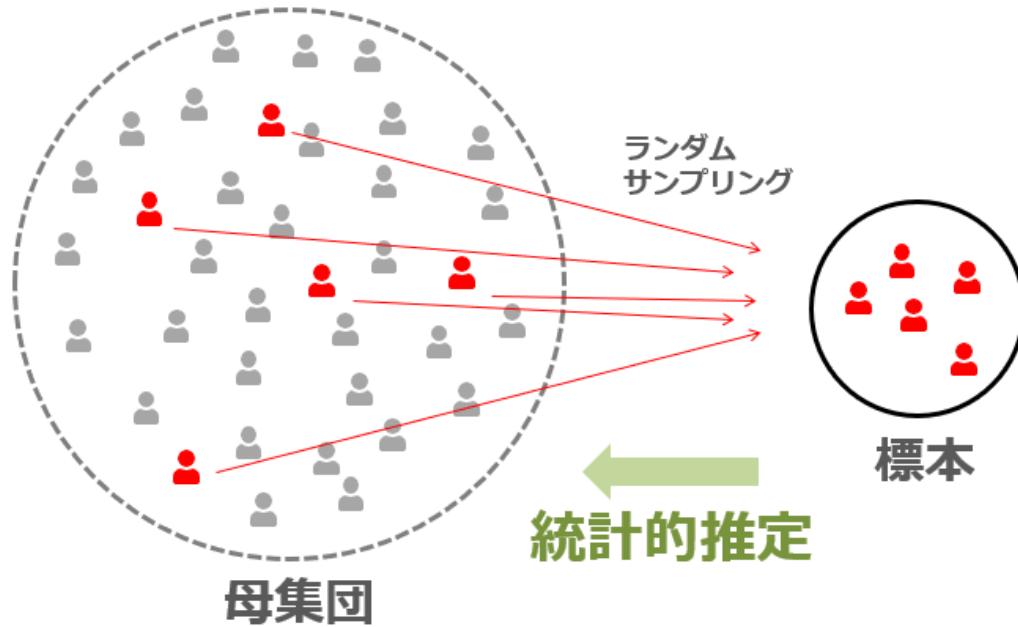
標準正規分布の、
標本数が少ない場合の
実用化バージョン by 櫻井

t分布

スチューデントのt分布

正規分布する母集団から標本をとり、母平均 μ を求めようとするとき、標本数が少ないと、標本側で起こる確率を、標準正規分布ではうまく表現しきれない。実際の実験などでは、標本数が少ないことがほとんど。そこで考え出された、**標準正規分布の、標本数を考慮した、実用化バージョン。**

by 櫻井



考えた人

ウィリアム・シーリー・ゴセット
William Sealy Gosset
イギリスの統計学者



出典：Wikipedia

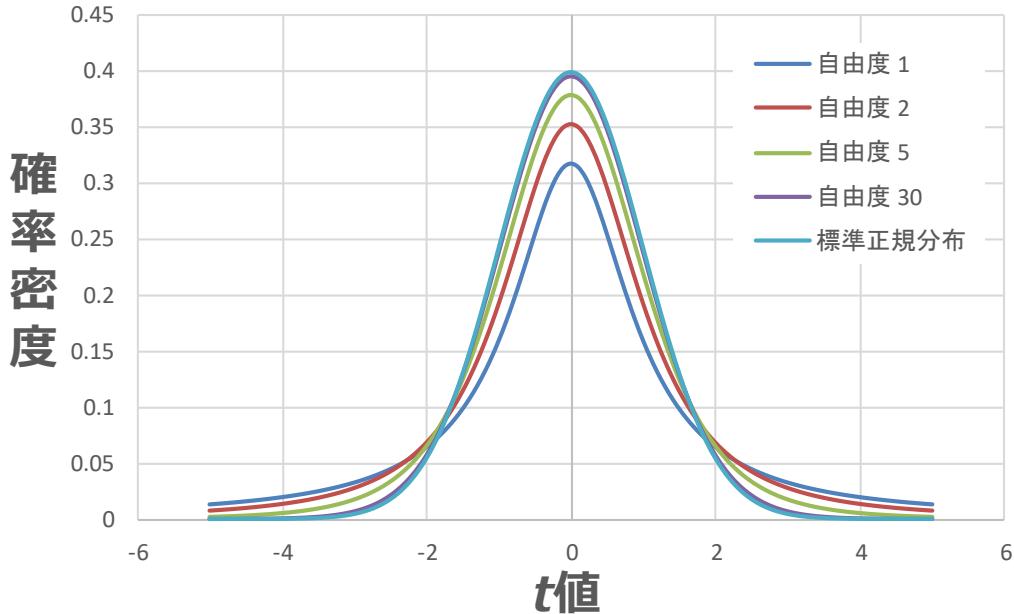


ギネスビール社で醸造とオオムギの品種改良の研究をするなかで t 分布を発見したが、ギネス社は社員の論文発表を禁じていたため、スチューデントというペンネームで論文発表した（1908年）。

出典：ギネス社HP

t 分布

t 分布表



自由度（標本-1）が小さいほど裾野が広がっており、自由度が高くなると標準正規分布に近づく

Excelでは、T.DIST, T.INV関数で計算できる

自由度 ν	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
出典	1.321	1.717	2.074	2.508	2.819

<https://to-kei.net/distribution/t-distribution/t-table/>

t 分布

性質：母平均 μ 、不偏分散 σ^2 の正規分布に従う母集団から抽出した n 個の標本を使って求めた次の統計量 t は、自由度($n-1$)の t 分布に従う。

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$z = \frac{X - \mu}{\sigma}$$

標準化 (z変換)

「標本平均 \bar{x} の分布を標準化した」と言える。
これまでと同様の考え方

区間推定 (母分散が不明な場合)

母平均 μ 、不偏分散 v^2 の母集団から抽出したn個の標本から求められる、a%信頼区間は以下となる。

$$\bar{x} - A * \frac{v}{\sqrt{n}} \leq \mu \leq \bar{x} + A * \frac{v}{\sqrt{n}}$$

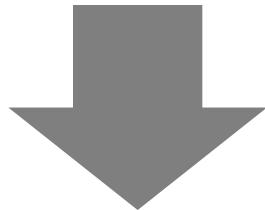
ここでAは、t分布表から、

- ✓ 自由度=n-1
- ✓ α (信頼計数) = $(100-a)/2/100$

で求められる境界値。

まとめ

分布（確率密度関数）



事象が起きる確率を推定できる！

描いてみよう

- 標準正規分布
- t 分布
- 補野の面積と境界値を計算

標準化してみよう



【参考】覚える必要はありません

正規分布の確率密度関数

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

標準正規分布の確率密度関数

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

【参考】覚える必要はありません

t 分布の確率密度関数

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi} \Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{t^2}{v}\right)^{-\left(\frac{v+1}{2}\right)}$$

v : 自由度

情報統計 第6回

2022年8月3日 神奈川工科大学



櫻井 望
国立遺伝学研究所
生命情報・DDBJセンター

*t*検定

学習目標

- 検定の考え方を学習し、
- 検定の基礎として、 t 検定を身につけます

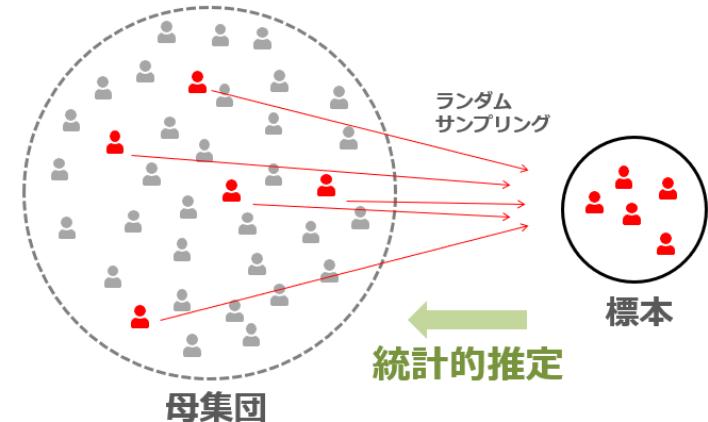
t 検定ツールの使い方を覚えるのではなく、Excelで自分で計算してみます

検定とは？

統計的仮説検定

- 統計的推定の手法のひとつ
- 母集団の性質や分布について立てた仮説を、標本を用いて、合理的・客観的に検証する方法
- 以下のステップをとる

- ① 仮説の設定
- ② 検定統計量の計算
- ③ 仮説採否の評価



例)

目標：カラオケ95点平均は本当？

- Aさんは、カラオケの平均点が95点くらいだと言っています。母平均 $\mu=95$ 点
- 実際の点数を、複数回にわたりこつそり記録した結果は以下でした。

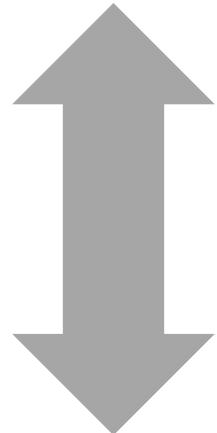
ランダムサンプリング

91, 90, 95, 88, 96, 89 標本

- 平均95点と言ってもよいでしょうか？

①仮説を立てる

Aさんのカラオケの平均は95点である



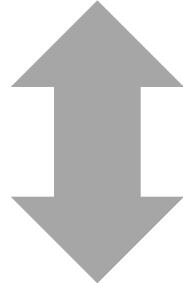
どちらでもよさそう
だが…

Aさんのカラオケの平均は95点ではない

帰無仮説と対立仮説

帰無仮説 H_0

Aさんのカラオケの平均は95点である



- 差異はみられない
- なんの関係もない

といった仮説を設定する

対立仮説 H_1

Aさんのカラオケの平均は95点ではない

帰無仮説が支持されない（棄却される）場合に採択される。検証したいことをこちらに持ってくる。

②検定統計量の計算

検定統計量

区間推定のときの境界値のように、分布に照らして確率を求めることができる数値のこと。

今回は、標本が6個なので、自由度5のt分布に従うと考え、t値を計算する。

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

②検定統計量の計算

標本平均

\bar{x}

不偏標本分散 s^2

母平均

μ

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$



③仮説採否の評価

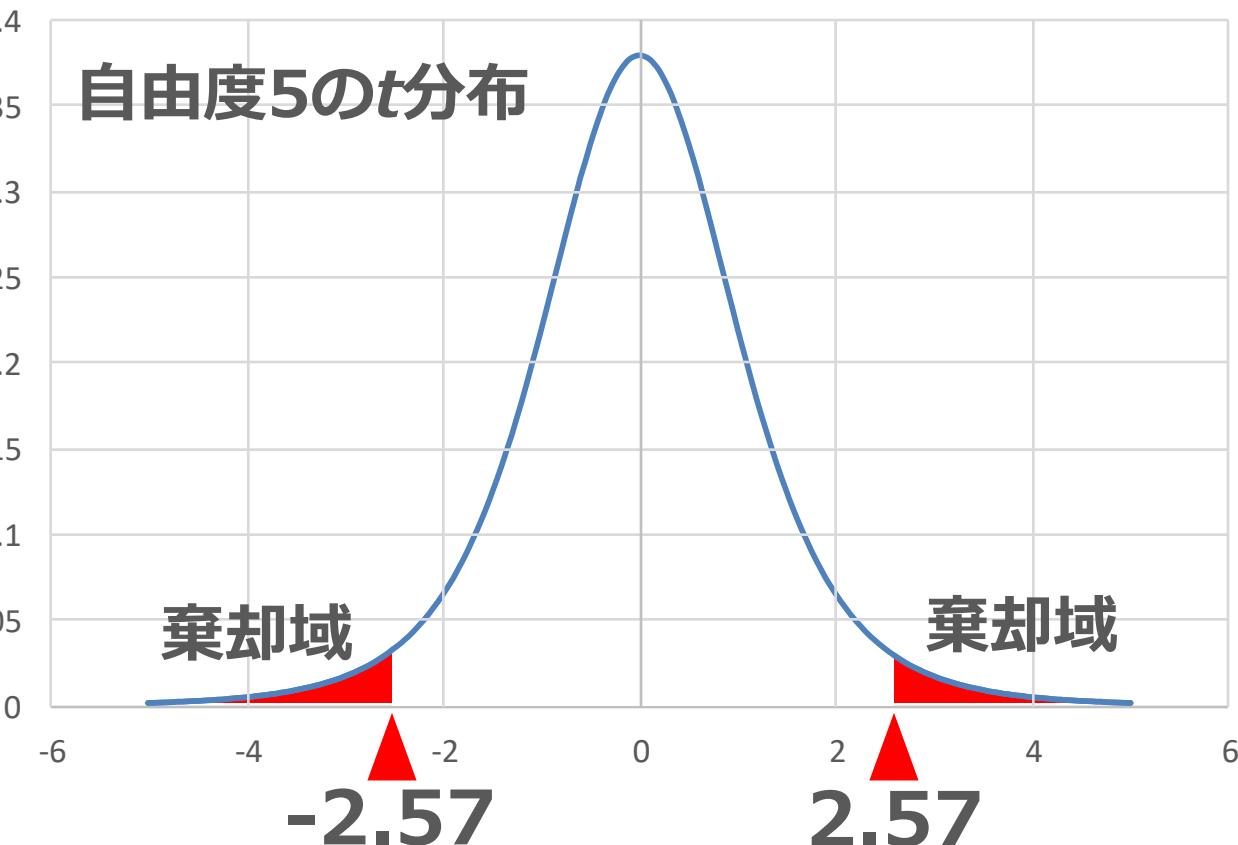
有意水準 α を0.05とする

有意水準 α

仮説を棄却するかどうかを決める基準の確率。これよりも小さい確率を持つ場合は、めったに起こらないことが起きていると考えられるため、帰無仮説（普通、変化がない）が棄却される。

③仮説採否の評価

t 分布表から、自由度5、 $\alpha = 0.05/2$
= 0.025の数値を読み取る



Excelで計算
してもよい



③仮説採否の評価

検定統計量が、棄却域に入ったかどうか
を確かめる

棄却域に入った！



結論

帰無仮説 H_0

Aさんのカラオケの平均は95点である

対立仮説 H_1

Aさんのカラオケの平均は95点ではない

有意水準0.05で帰無仮説は棄却された
ので、対立仮説を採択し、「Aさんのカ
ラオケの平均は95点ではない」とする。

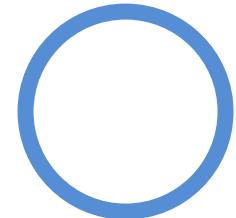
注意点

帰無仮説が棄却されないとき…

「帰無仮説が正しい」と安易に結論付けてはいけない。



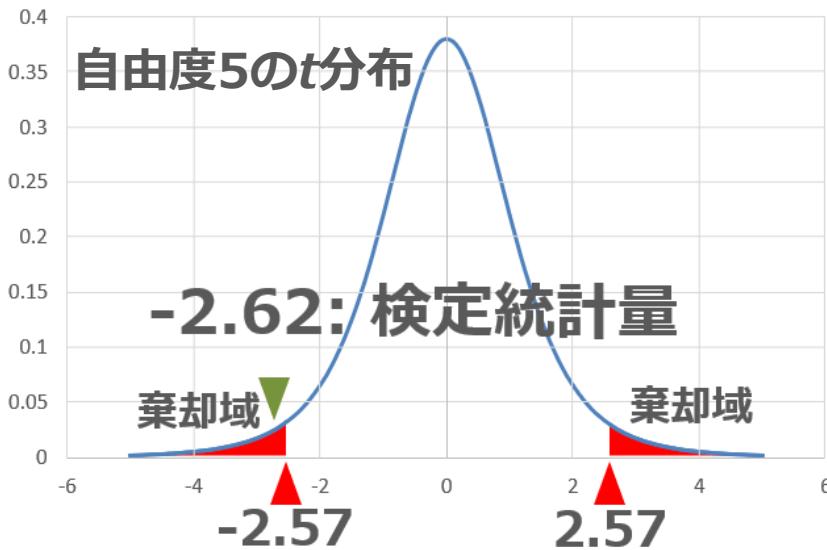
「帰無仮説が誤っているとは言えない」とは言える。



例えば今回では、帰無仮説が棄却されなくても、真の母平均は95点ではないかもしない。

p値（有意確率）

検定統計量と分布から計算される確率。
どれだけ例外的な事象が起きているかを表す。



境界値2.57は、自由度5、 $\alpha = 0.025$ の時に計算された値。 t 値2.62より外側の面積（p値）も、この分布から求めることができる。
0.025より小さい確率（より起こりにくい）を持っているはず。

※帰無仮説が正しい確率を示すのではない

有意と優位

検定を行った場合、「有意に＊＊だった」とか、「有意に＊＊とは言えない」のような表現をします。

検定では、確率的にまれに起こる事象かどうか、つまり「意味ありげ（有意）」かどうかを調べるからです。

一方、統計とは関係なく、数値の大小や傾向などを判断して、どちらかが優勢である状態を「優位」と表現します。

この違いに気を付けて正しく使い分けましょう。

エクセルで 計算してみよう

- 基本統計量
- 検定統計量
- 境界値
- p値
- 標本のカラオケ点を色々
変えて、結果がどうなる
かを見てみよう



情報統計

第7回

2022年8月3日 神奈川工科大学



櫻井 望
国立遺伝学研究所
生命情報・DDBJセンター

第6回の補足

検定で
注意すること

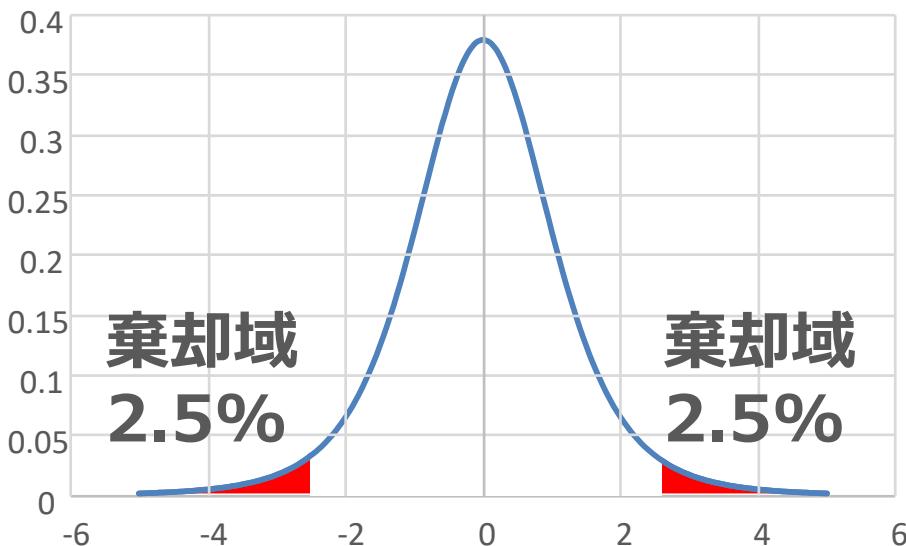
学習目標

- 第6回の補足
 - ✓ 両側検定と片側検定
 - ✓ t 検定のいろいろ
- 検定を考えるときに気をつけたいポイントをおさえる
 - ✓ 検定の間違い
 - ✓ p 値 < 0.05 にとらわれるな！
 - ✓ 多重性の問題、FDR（偽発見率）

第6回の補足

両側検定と片側検定

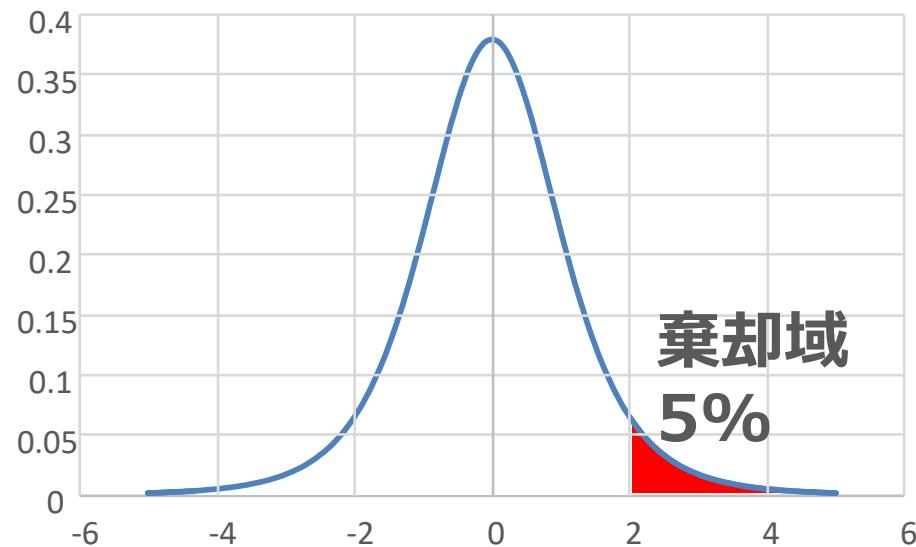
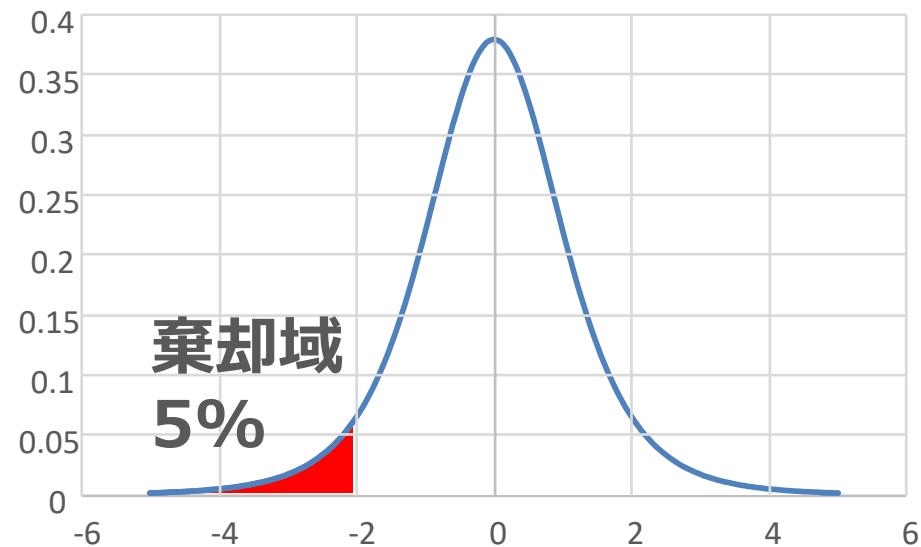
Aさんのカラオケ問題では、有意水準0.05を、その半分の0.025ずつに分け、 t 分布の両側に割り当てて考えました



これは
両側検定
と呼ばれます

片側検定

有意水準を、左右のどちらかにだけ重点配分することができ、これを**片側検定**と呼びます。



片側検定をするとき

明らかにどちらかに偏っている場合だけが問題になるような仮説検定をするときは、片側検定を行うことができます

- 例)
- 蛍光灯の寿命は仕様書にある＊＊時間よりも短いか？
 - 今年の給料は昨年の＊＊円よりも上がったか

ただ、有意水準の数字をいくつにするかだけの問題なので、通常は両側検定で問題ありません

色々なt検定

t検定には、実はいろいろあります。問題にしている群がひとつか二つか、2群の場合はさらに、対応関係があるかないかで分かれます。

- **1群のt検定**

- 母集団の平均値が特定の値であるかどうかの検定

- **2群のt検定**

- 2つの群の平均値に差があるかどうかの検定

- ✓ 対応のある2群の場合

- ✓ 独立した2群の場合

1群のt検定

母集団の平均値が、特定の値かどうかを検定します

Aさんのカラオケ平均点が95点かどうかで行ったのは、実は、1群のt検定です

他の例)

工場のラインで規格どおりに製品が製造されているかどうか？

2群のt検定（対応あり）

「対応がある」とは、例えば以下のような場合です。

介入試験をおこない、試験食の摂取前後で数値を測定した

被験者No.	摂取前	摂取後
1	120	122
2	108	107
3	115	118
4	123	130
5	111	119

被験者ごとに、摂取前（A群）と摂取後（B群）で対応関係があり、知りたいのは、摂取前後で差があるかどうかです。

2群のt検定（対応あり）

実はこの問題は、次の手順で、1群のt検定として処理できます

- 摂取前後の差をとる
- その平均値が0であることを帰無仮説として検定を行う

被験者No.	摂取前	摂取後	摂取前後の差
1	120	122	-2
2	108	107	1
3	115	118	-3
4	123	130	-7
5	111	119	-8

2群のt検定（独立2群）

実験科学の分野などでよく使われます

例)

- 介入試験で、試験食群とプラセボ群に差があるか？
- 二つのピーナッツ品種で、オレイン酸含量に差があるか？

2群間で、**分散が等しいか**どうかによって、二つのやり方があります。最近では、分散が等しいかどうかにかかわらず、等しくないことを仮定した**ウェルチの方法**が良く使われます。

2群のt検定（独立2群）

等分散の場合

1群目：標本数 n_1 , 不変標本分散 s_1^2 , 標本平均 \bar{x}_1

2群目：標本数 n_2 , 不変標本分散 s_2^2 , 標本平均 \bar{x}_2

プール分散
$$s^2 = \frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}$$

検定統計量
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

自由度： $n_1 + n_2 - 2$

帰無仮説： 2群の母集団の平均値は等しい

で、同様に検定できます

参考まで

2群のt検定（独立2群）

等分散が仮定できない場合 ウエルチの方法

1群目：標本数 n_1 , 不変標本分散 s_1^2 , 標本平均 \bar{x}_1

2群目：標本数 n_2 , 不変標本分散 s_2^2 , 標本平均 \bar{x}_2

検定統計量 $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

(近似)自由度 $v \approx \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}}$

帰無仮説：2群の母集団の平均値は等しい

で、同様に検定できます

参考まで

メッセージ

どんな検定でも

- 検定統計量
- 自由度
- 分布の計算方法

などさえ分かれば、身につけた
ステップで、**自分でできる！**

検定で
注意すること

①

検定の間違い

前提

検定では、
正しくない帰無仮説を棄却して、
対立仮説を採択することが、
主張したいこと（正しい姿）
とします。

検定の二つの間違え

第一種の過誤 偽陽性

本当は間違っていることを、正しいと判定してしまうこと。
[検定では、本当は帰無仮説が正しいのに、間違えだとして棄却してしまうこと]
この過誤を犯す確率は α で表され、実は、その値のことを**有意水準**と呼んでいる。

α :あーわてんぼうのお手つき率

第二種の過誤 偽陰性

本当は正しいことを、誤っていると判定してしまうこと。
[検定では、本当は帰無仮説が**間違え**なのに、正しいとして棄却しないこと]
この過誤を犯す確率は β で表され、(1- β 、つまりこの過誤を犯さない確率)を**検出力**と言う。第二種の過誤をなるべく犯さない (β が小さい) のが、**よい検定**とされる。

β :ぼーんやりものの見逃し率

		帰無仮説が本当は	
		間違え (正しい姿)	正しい
検定結果	棄却する (陽性)	$1 - \beta$ (検出力)	第一種の過誤 偽陽性 α
	棄却しない (陰性)	第二種の過誤 偽陰性 β	OK

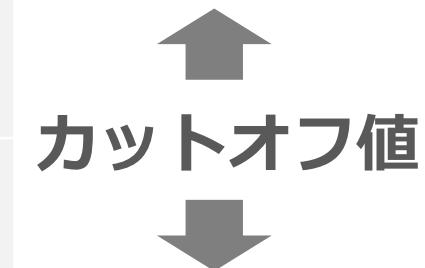


第一種の過誤を起こさないように α を下げて厳しく判定すると、 β が増えてしまい、検出力 ($1 - \beta$) が下がってしまう。
うまくバランスのとれた α を設定する必要がある。

少し脇道へ

スクリーニング検査

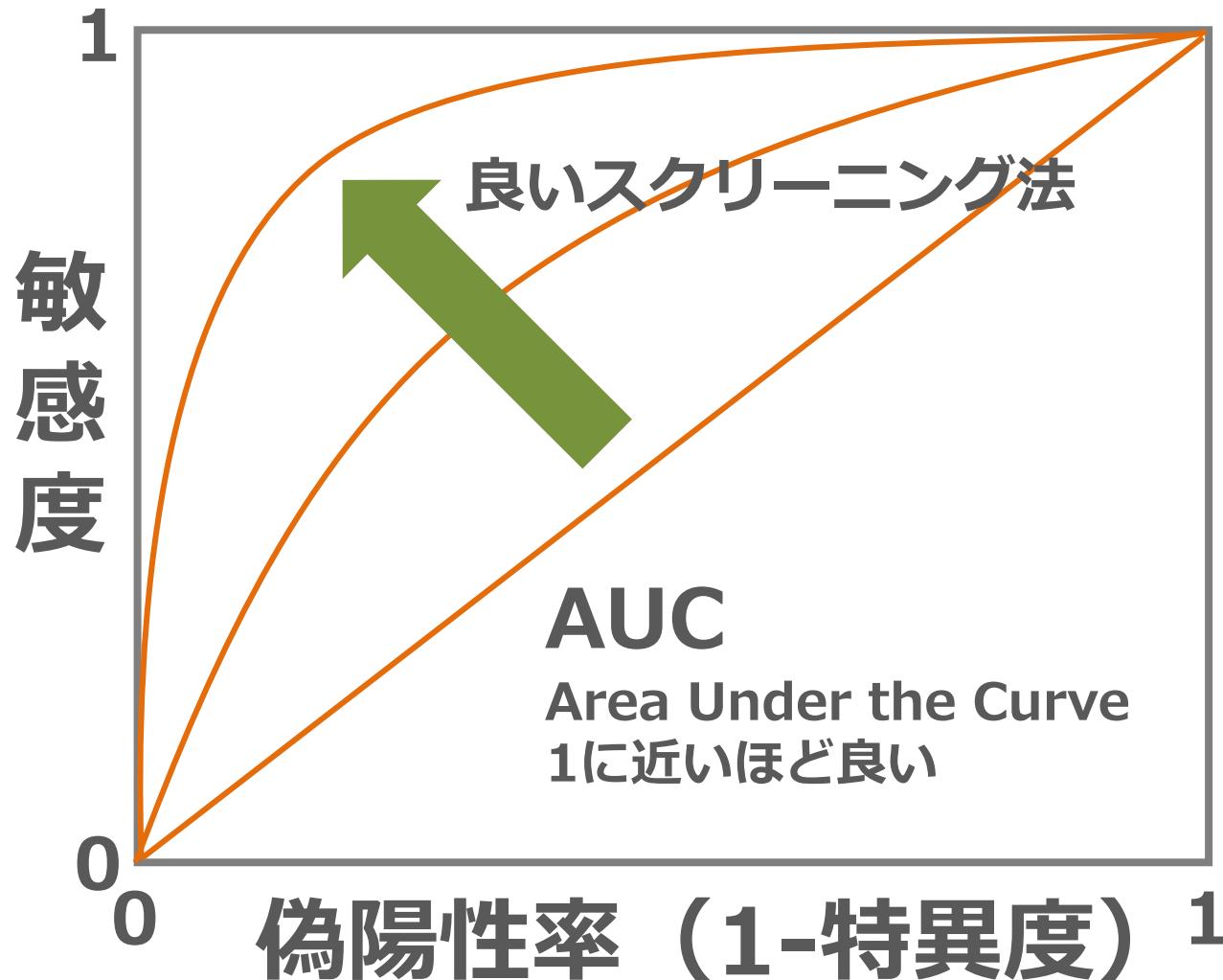
**の スクリーニング		本当は	
結果	陽性 (+)	病気	健康
		陰性 (-)	陰性
真陽性 True Positive 敏感度	偽陽性 False Positive 偽陽性度	偽陰性 False Negative 偽陰性度	真陰性 True Negative 特異度



敏感度を上げたり、偽陽性率を下げたりするためにカットオフ値を調整するのと似ています。
ただし、敏感度を上げるのに、カットオフ値を上げるか下げるかは、スクリーニング検査の方法に依存するので注意！

ROC曲線

Receiver Operating Characteristic curve



カットオフ値を変えたときの偽陽性率と敏感度をプロットしたもの

仮説検定では、何が真に正しいかがわからないため、ROC曲線が描けないことがほとんどです。

ただし、スクリーニング検査と同様に、診断システムの精度評価をする際などには多用されます。

データ解析ではとても重要な考え方です。

検定で
注意すること

②

多重性の問題

**検定は、
繰り返してはいけない**

検定を繰り返すと、誤りが大きくなる

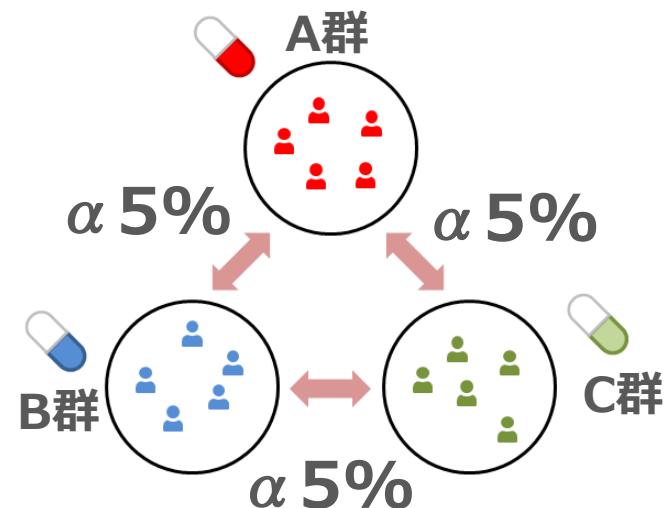
例)

3つの薬A, B, Cを与えた群で、差がなかったかどうかを、
A-B, B-C, C-A投与群間で α 5%で検定する。
3つの薬に差がないことを主張したい。

1回の検定で差がないという結果になる確率は0.95。

3回の検定でどれもが差がない結果となる確率は、0.95の3乗で、0.86。

どこかで有意な差が出てしまう確率は、 $1 - 0.86 = \textcolor{magenta}{0.14}$ 。



数打てば当たる状況！

多重比較のための 検定法を使う

Tukey (チューキー) の
多重比較検定など

Bonferroniの補正

有意水準 α を繰り返す検定の数で割り、それを有意水準として用いる

例)

$\alpha = 0.05$ で3回検定を繰り返す場合、

$$\alpha' = 0.05 \div 3 = 0.0167$$

を代わりに用いる

全体の α (お手つき率) が決して水準を超えないように、むりやり α を引き下げるのと、第二種の過誤の率 (見逃し率) β が上がってしまう恐れがある。

False Discovery Rate (FDR, 偽発見率)を調整する

ある程度 α が上がるのを許容しながら、 β を小さく抑える方法。

		帰無仮説が本当は		
		間違え(正しい姿)	正しい	計
検定結果	棄却する (陽)	s	v α 偽陽性	R
	棄却しない (陰)	t β 偽陰性	u	N-R
	計	N-n	n	N

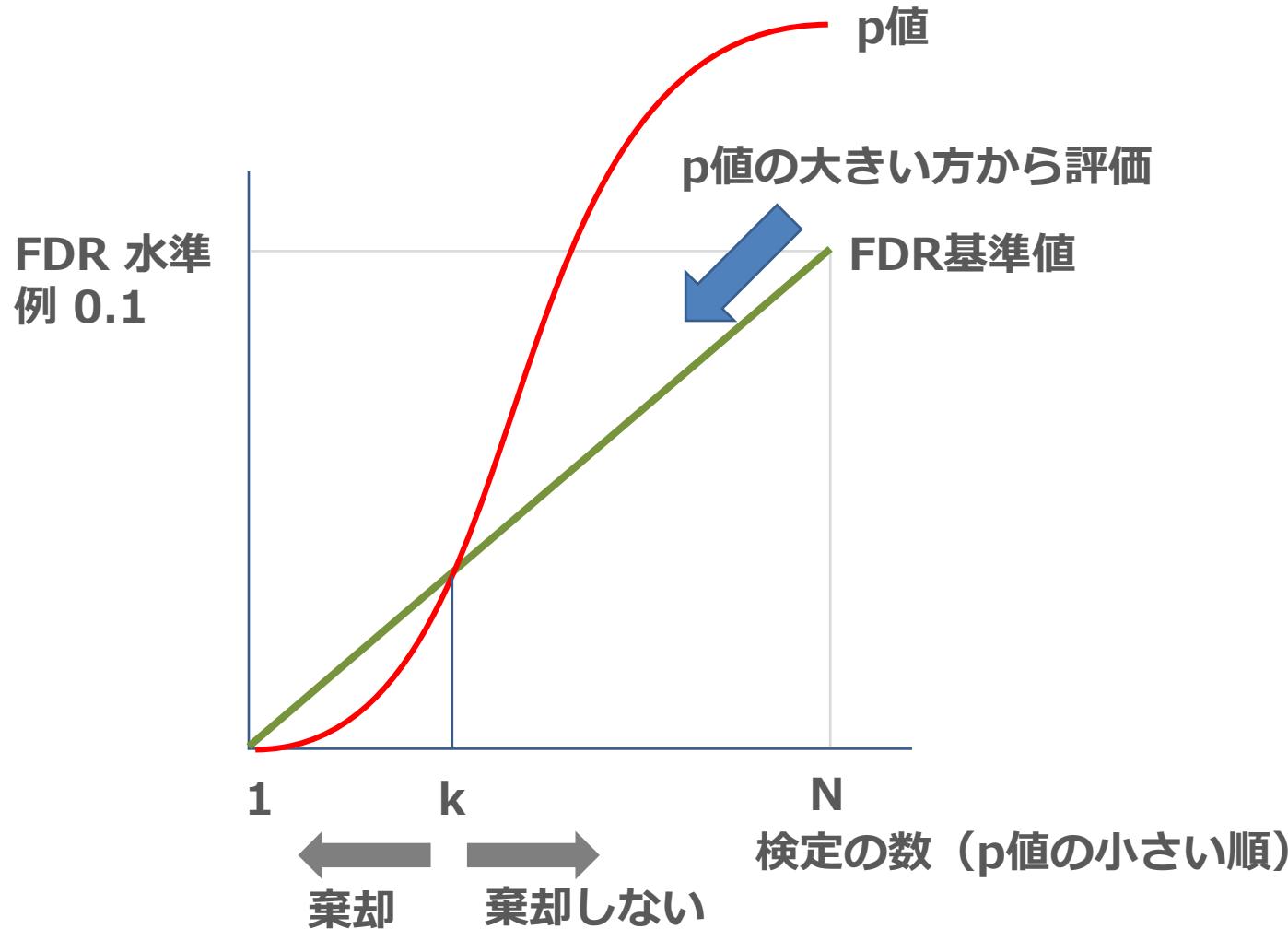
FDR $q = v/R$ 棄却したもののうち、偽陽性の率
これを、一定水準例えば0.05にする方法

FDR調整の手順

Benjamini & HochbergのFDR調整方法（BH法）（1995年に発表）
その後いろんな改良法が考案された。

- ① N個の検定結果について、p値の小さい順に並べる。
この時の順番を、 $i = 1$ 番目からN番目とする。
- ② $i = N$ （p値が一番大きいもの）とする。
- ③ $q \times i/N$ を計算する。
これが、もとのp値以上であれば、 $k = i$ として、④に進む。
もとのp値を下回れば、 $i = i - 1$ として、③を繰り返す。
 $i = 1$ に達したら、どの検定の帰無仮説も棄却しないものとする。
- ④ $i = 1$ からkまでの検定の帰無仮説を棄却する

FDRのイメージ



FDR調整のイメージ

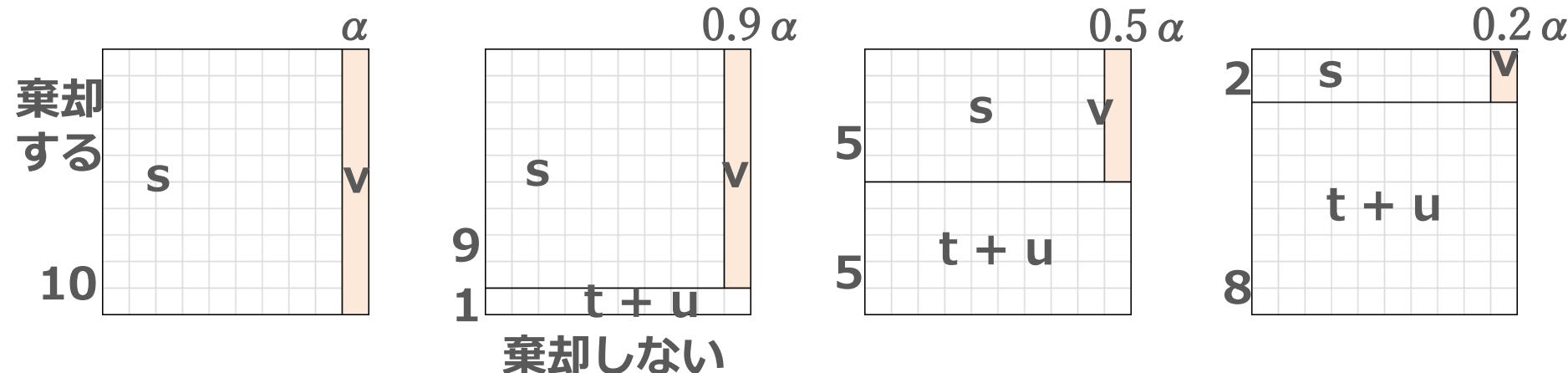
by 櫻井

p値は、検定を繰り返したときに誤る確率でもあるので、複数回検定を繰り返したときに、最大のp値が有意水準 α を下回っているなら、すべての検定が十分有意であると判断してもよいものとする（甘いが）。

10回検定し、FDRを0.1に制御したいとする。

10回を全部棄却したとき、FDRを0.1以下にするには、 α は0.1でよい。
1回分を棄却しないとすると、残り9回のFDRを0.1にするには、 α は $0.1 * 9 / 10$ に設定する必要がある。

以下同様、棄却する検定の数が減るほどに、 α を小さく調整する。



検定で
注意すること

③

p値 < 0.05
にとらわれるな！

**有意水準 α としてよく使われる0.05
という数字に、特に深い意味はない**

**起こりにくい確率のひとつの基準と
して使われているだけ**

アメリカ統計学会の声明

Wasserstein and Lazar 2016) The American Statistician 70: 129-133 Editorial

Wasserstein et al (2019) The American Statistician 73 (S1): 1-19 Editorial

- p値が特定の値以下だったことで「統計的に有意であった」と言ってはいけない
- それよりも、p値そのものを提示する
- p値は、仮説が正しい確率を測るものではない

など

2016年の声明の日本語訳が読める

<http://www.biometrics.gr.jp/>



一般社団法人
日本計量生物学会
The Biometric Society of Japan

HOME 学会について お知らせ ニュースレター 学会誌 [計量生物学の未来に向けて](#) 試験統計家認定制度

No.60～69

No. タイトル

61 研究不正と研究環境 井上永介(昭和大学)

60 計量生物学徒としてHTAに貢献する 萩原康博(東京大学大学院医学系研究科)

> トップページ

> 学会について

> お知らせ

> ニュースレター

> 学会誌

> [計量生物学の未来に向けて](#)

> 試験統計家認定制度

> 臨床研究に関する日本計量生物学会声明

> 統計家の行動基準

> 統計家の行動基準(英語版)

> [統計的有意性とP値に関するASA声明](#)

> メーリングリスト

> 当会へのお問合せ

No.50～59

No. タイトル

59 真実がわからない中で過去からの学びをどう活かすか 坂巻頸太郎(横浜市立大学)

58 計量生物学を理解したいと思って毎日挑戦しています 長島健悟(統計数理研究所)

57 これからの計量生物学の発展を担う生物統計家の育成 安藤宗司(東京理科大学)

56 一教員として貢献できること 高橋佳苗(大阪市立大学)

55 ベースラインハザードから思うこと 横田 熟(北海道大学)

54 放射線疫学と日本人のコホートを追跡する日米共同研究機関 三角宗近(放射線影響研究所)

53 実務の現場から:食品・栄養研究にも活用される生物統計学の専門性 高田理浩(味の素株式会社)

52 異分野、異文化の接点から 島津秀康(英国ラバーラ大学)

51 統計学を学んで 奥井 佑(九州大学)

50 教育・指導への感謝と未来への還元 井桁正堯(兵庫医科大学)

やってはいけない不正行為

- t 検定で有意にならなかつたので、**有意になる検定方法を試して、マン・ホイットニーのU検定を採用した**
- サンプルサイズを調整した



p値ハッキング

情報統計

第8回

2022年8月3日 神奈川工科大学



櫻井 望
国立遺伝学研究所
生命情報・DDBJセンター

そのほかの検定

明日の準備

自習（課題準備）

学習目標

ほかの検定方法を把握する

- ✓ パラメトリック検定と
ノンパラメトリック検定の違い
- ✓ χ^2 (カイ) 二乗検定

パラメトリック検定

- 分布を用いる
- 正規分布に従うとか、等分散性があるとか、何かしらの前提条件が必要

ノンパラメトリック検定

- 分布を用いない
- 前提条件が必要ない
- データを並び替えて検定する

例えば2群の差の検定

パラメトリック検定

対応ない場合

2群のt検定

対応ある場合

対応ある1群のt検定

ノンパラメトリック検定

対応ない場合

マンホイットニーのU検定

対応ある場合

ウィルコクソンの符号付き
順位和検定

分割表による検定

- カイニ乗検定
- フィッシャーの正確確率検定

	ゲームが好き	ゲームそれほどでもない	合計
朝食を食べる			
朝食を食べない			
合計			

など

情報統計

第9-12回

2022年8月4日 神奈川工科大学



櫻井 望
国立遺伝学研究所
生命情報・DDBJセンター

昨日

- 確率分布
 - t分布
 - 検定の手順
-
- Excelで分布を描く方法
 - t検定を手計算で行う

今日

- 分散分析(ANOVA)の概念を把握して、手で計算できることを確認する
- 相関
- 多変量解析（主成分分析）のイメージ

情報統計

第9回

2022年8月4日 神奈川工科大学



櫻井 望
国立遺伝学研究所
生命情報・DDBJセンター

学習の目標

- F検定（等分散性の検定）
- 分布の仲間
カイ二乗分布、F分布
- 分散分析ANOVA（F分布を使う）

2群のt検定（独立2群）

等分散の場合

1群目：標本数 n_1 , 不変標本分散 s_1^2 , 標本平均 \bar{x}_1

2群目：標本数 n_2 , 不変標本分散 s_2^2 , 標本平均 \bar{x}_2

プール分散
$$s^2 = \frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}$$

検定統計量
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

自由度： $n_1 + n_2 - 2$

帰無仮説： 2群の母集団の平均値は等しい

で、同様に検定できます

参考まで

2群のt検定（独立2群）

等分散が仮定できない場合 ウエルチの方法

1群目：標本数 n_1 , 不変標本分散 s_1^2 , 標本平均 \bar{x}_1

2群目：標本数 n_2 , 不変標本分散 s_2^2 , 標本平均 \bar{x}_2

検定統計量 $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

(近似)自由度 $v \approx \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}}$

帰無仮説：2群の母集団の平均値は等しい

で、同様に検定できます

参考まで

F検定

等分散性の検定

1群目：標本数 n_1 , 不変標本分散 v^2_1

2群目：標本数 n_2 , 不変標本分散 v^2_2

検定統計量：
$$F = \frac{v^2_a}{v^2_b}$$
 ※ v^2_a, v^2_b は、 v^2_1, v^2_2 のいずれか、分散の大きい方を分子にする。数値は1以上になる

自由度： $n_1 - 1, n_2 - 1$

※分子と分母に対応させて、二つ与える

帰無仮説：2群の分散は等しい

F分布を扱うExcel関数：F.DIST, F.DIST.RTなど

例) 身長データの場合

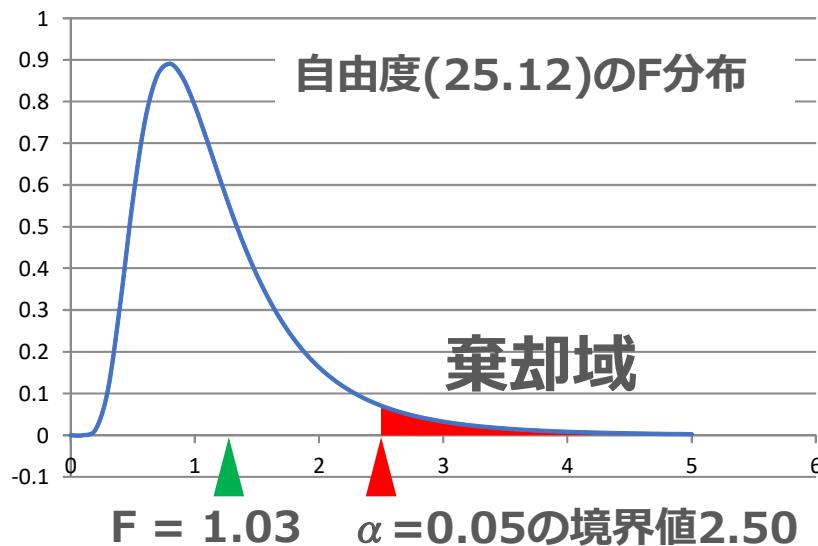
女性: $n_1 = 26, v^2_1 = 23.63$

男性: $n_2 = 13, v^2_2 = 23.02$

有意水準: 0.05とする

$$F = 23.63 \text{ (女性)} / 23.02 \text{ (男性)} = 1.03$$

自由度(25, 12)のF分布から、FDIST.RT関数を使って求めた右側確率pは、0.50



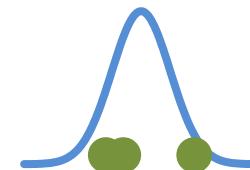
F値が棄却域の境界値より内側
($1.03 < 2.50, p=0.50 > \alpha$)
なので、帰無仮説は棄却できず、
「2群の分散に差があるとは言えない」と結論づけられた。

留意すべきこと

F検定で「分散に差がある」という結論を得たのち、2群の平均値に差があるかどうかをt検定すると、「**検定の多重性**」の問題にあたってしまう。

近年では、等分散かどうかに関係なく適用できるウェルチの検定を最初から行うことが望ましいという考えも出てきている。

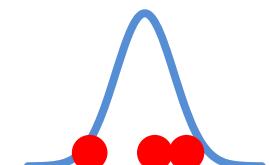
F分布



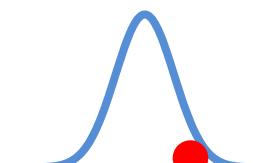
複数の変数を含む**2群**
(分散の比を考える)

カイ二乗分布

標準正規分布



複数の変数
(分散を考える)

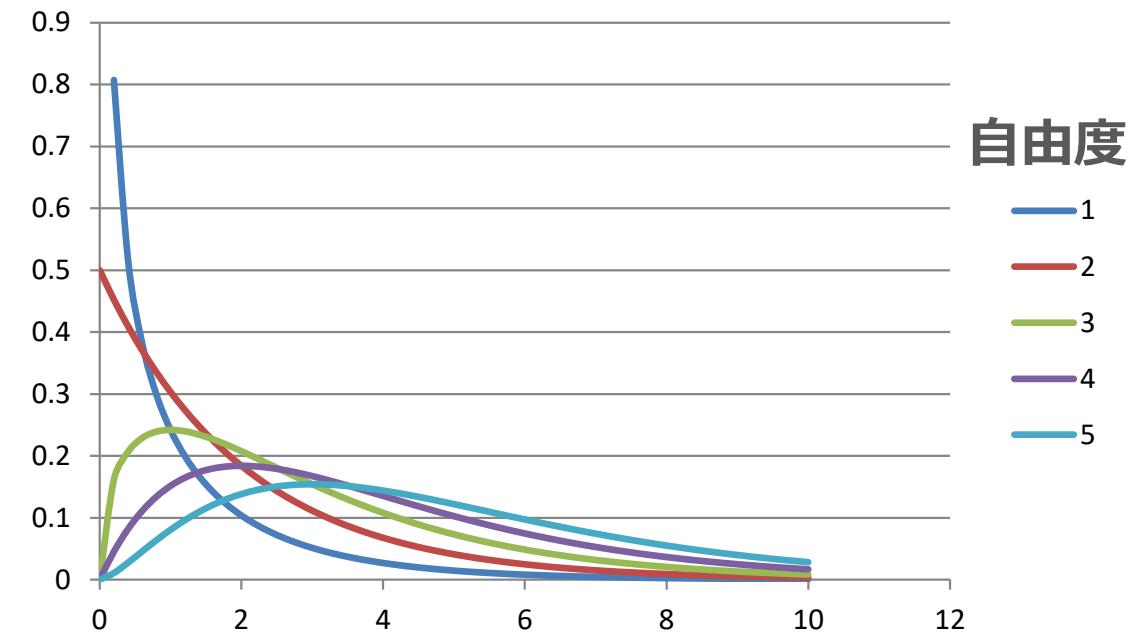
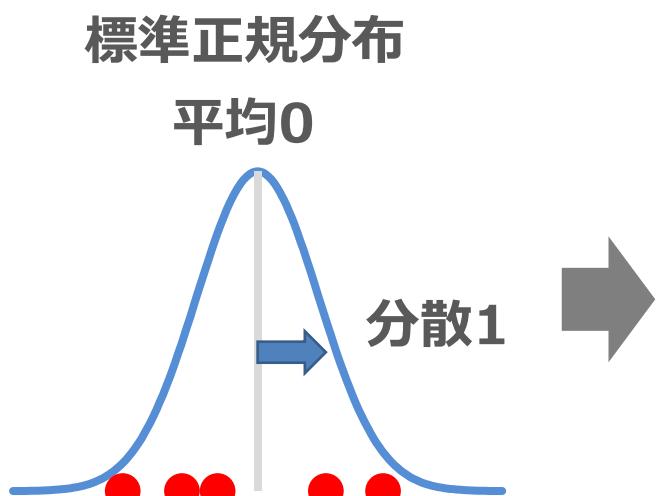


ひとつの変数

by 櫻井

カイ二乗分布

標準正規分布に従った独立した変数がいくつかあるとき、その二乗和が従う分布



カイ二乗分布の性質

正規分布 $N(\mu, \sigma^2)$ に従った k 個の変数 x_i について、偏差（平均からの差）の平方和と分散の比は、自由度 k のカイ二乗分布に従う

$$\chi^2 = \frac{\sum_{i=1}^k (x_i - \mu)^2}{\sigma^2} = \sum_{i=1}^k \left(\frac{x_i - \mu}{\sigma} \right)^2$$

カイ二乗検定

	ビール 好き	ビール あんまり
男性	23	12
女性	7	8

二つのカテゴリに関連があるかを調べたい

帰無仮説：

二つのカテゴリは独立である（関連がない）

有意水準：0.05

カイ二乗検定の手順

(1) 観測データから、カテゴリーごとに割合を出す

	ビール 好き	ビール あまり	合計
男性	69	36	105 70%
女性	21	24	45 30%
合計	90 60%	60 40%	150 100%

(2) 割合から、カテゴリーが独立な場合の度数（期待度数）を出す

	ビール 好き	ビール あまり	合計
男性	63	42	105 70%
女性	27	18	45 30%
合計	90 60%	60 40%	150 100%

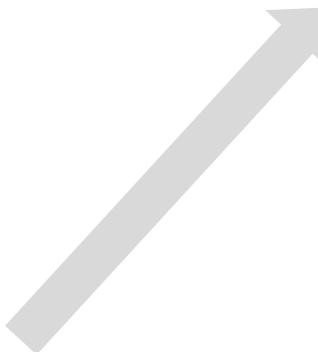
カイ二乗検定の手順

(3) 観測度数と期待度数の差を出す

	ビール 好き	ビール あんまり
男性	6	-6
女性	-6	6

(4) その二乗を出す

	ビール 好き	ビール あんまり
男性	36	36
女性	36	36



(5) 期待度数で割る

	ビール 好き	ビール あんまり
男性	$36/63 = 0.57$	$36/42 = 0.86$
女性	$36/27 = 1.33$	$36/18 = 2$

(6) その和を求める

$$x = 0.57 + 0.86 + 1.33 + 2 = 4.76$$

このように求めた値xは、カイ二乗分布に近似できる。

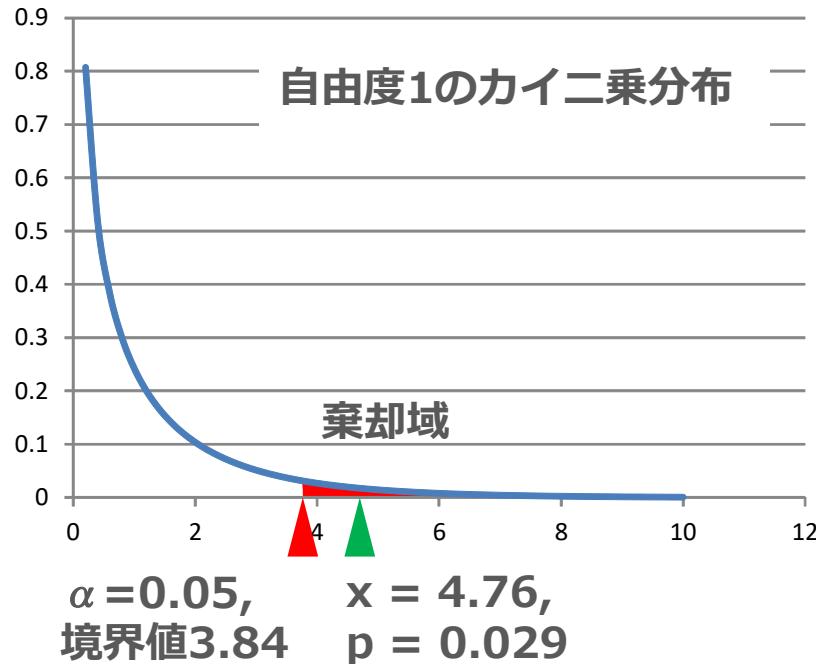
自由度は、各カテゴリ（性別、ビールの好み）の要素数をそれぞれ n_1 , n_2 とすると、 $(n_1-1)*(n_2-1)$ 。

この例の場合では、 $(2-1)*(2-1) = 1$

カイ二乗検定の手順

(7) 結論

x の値が棄却域の境界値の外側 ($3.84 < 4.76, p=0.029 < \alpha$) なので、帰無仮説は棄却され、「二つのカテゴリは独立ではない」と判断された。



よって、この母集団においては、「性別とビールの好みとの間に何かしらの関連性がある」と結論づけられた。

カイ二乗分布を扱うExcelの関数：
CHISQ.DIST, CHISQ.DIST.RT, CHISQ.INV.RTなど

カイ二乗検定の留意点

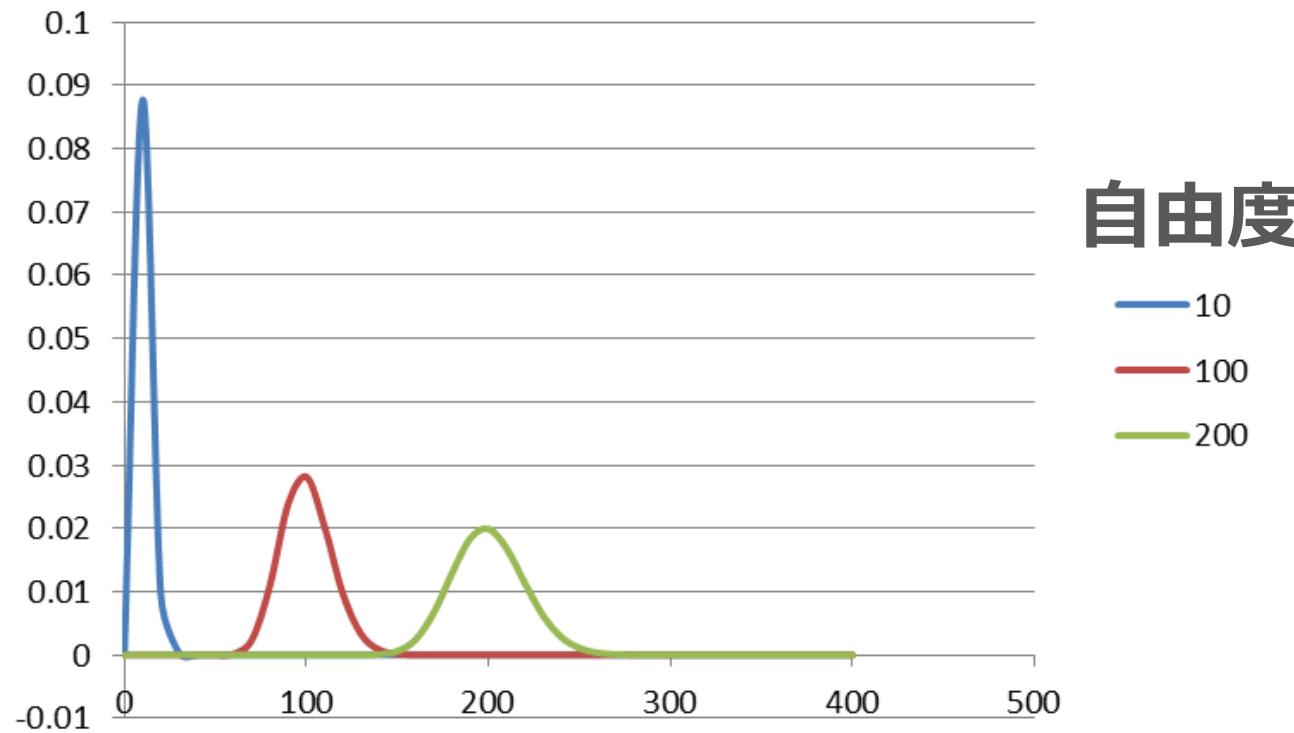
観測数が少ないとカイ二乗分布への近似ができないので、
その場合はフィッシャーの正確確率検定を行う。

目安：

期待度数が5未満のセルが、全セルの20%以上で存在
する場合、近似が不正確と考えられる
(コクラン・ルール)

期待度数が1未満のセルがあってはならない

カイ二乗分布の性質 その2



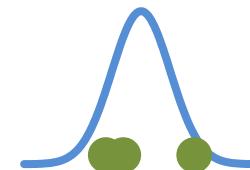
自由度 k が大きくなると、

平均値 : k

分散 : $2k$

の正規分布に近づいてゆく

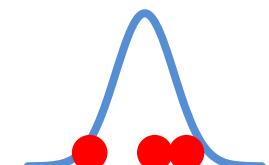
F分布



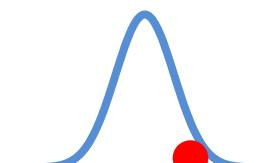
複数の変数を含む**2群**
(分散の比を考える)

カイ二乗分布

標準正規分布



複数の変数
(分散を考える)



ひとつの変数

by 櫻井

F分布とカイ二乗分布の関係

自由度 k_1 のカイ二乗分布 χ^2_1

自由度 k_2 のカイ二乗分布 χ^2_2

があるとき、次の値Fは、自由度(k_1, k_2)のF分布に従う

$$F = \frac{\chi^2_1/k_1}{\chi^2_2/k_2}$$

χ^2 分布が
ふたつ!!

F分布の活用

正規分布 $N(\mu_1, \sigma^2_1)$ に従った母集団から得た標本、
標本数 : n_1 、不偏標本分散 : v^2_1

正規分布 $N(\mu_2, \sigma^2_2)$ に従った母集団から得た標本、
標本数 : n_2 、不偏標本分散 : v^2_2

があるとき、

$$F = \frac{\chi^2_1/k_1}{\chi^2_2/k_2} = \frac{v^2_1/\sigma^2_1}{v^2_2/\sigma^2_2}$$

二つの母集団の分散 σ^2_1 と σ^2_2 が等しいと仮定できる場合は、

$$F = \frac{v^2_1}{v^2_2} \quad \text{←これをF検定で利用している！}$$

F分布の活用

カイ二乗分布の性質

$$\chi^2 = \frac{\sum_{i=1}^k (x_i - \mu)^2}{\sigma^2} \quad \text{自由度 } k$$

この式を変形すると、

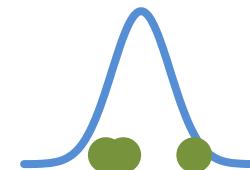
不偏標本分散 v^2 になっている！

$$\chi^2 = \frac{k \times \frac{\sum_{i=1}^k (x_i - \mu)^2}{k}}{\sigma^2} = \frac{k \times v^2}{\sigma^2}$$

したがって、

$$\frac{\chi^2}{k} = \frac{k \times v^2}{\sigma^2} \times \frac{1}{k} = \frac{v^2}{\sigma^2}$$

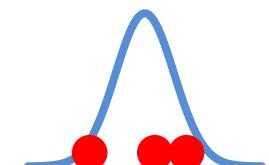
F分布



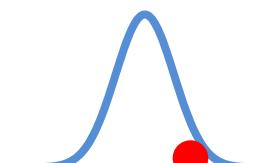
複数の変数を含む**2群**
(分散の比を考える)

カイ二乗分布

標準正規分布



複数の変数
(分散を考える)



ひとつの変数

by 櫻井

分散分析

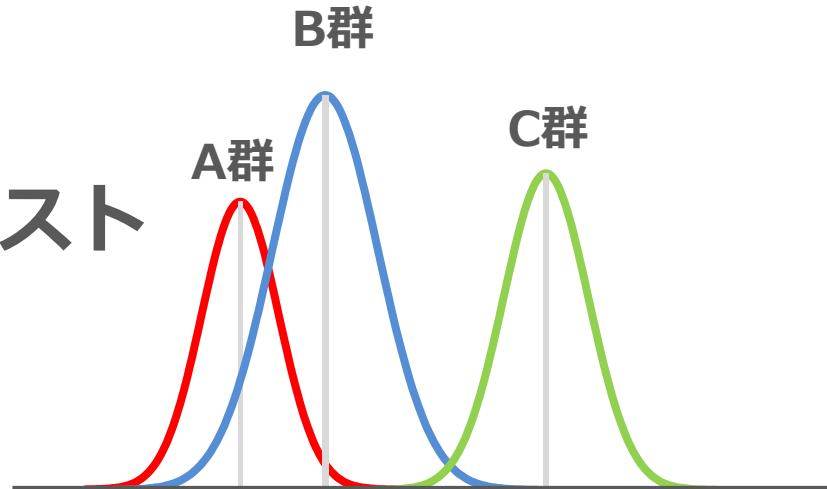
Analysis of Variance

ANOVA

- ✓ 3つ以上の群があるとき、
- ✓ 群の母平均に差があるかどうかを、
- ✓ 分散（F分布）を使って、

検定する方法

例) 1組、2組、3組で、テストの平均点に差があるか？



帰無仮説：

A群、B群、C群の母平均は等しい

対立仮説：

A群、B群、C群の母平均の中に、
異なる値がある

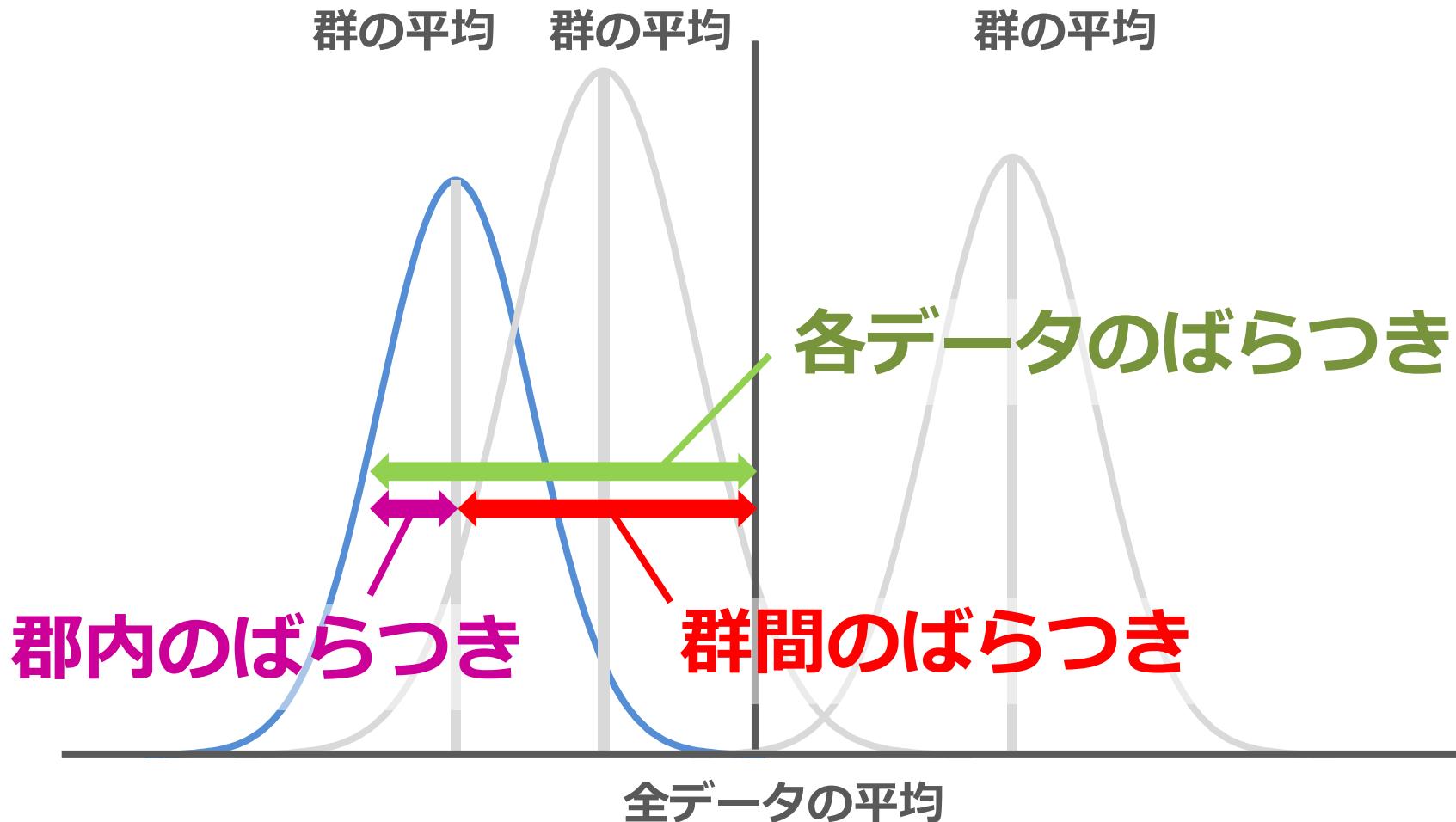


どれが異なるかまではわからない！

帰無仮説が棄却されたときは、解釈に注意が必要

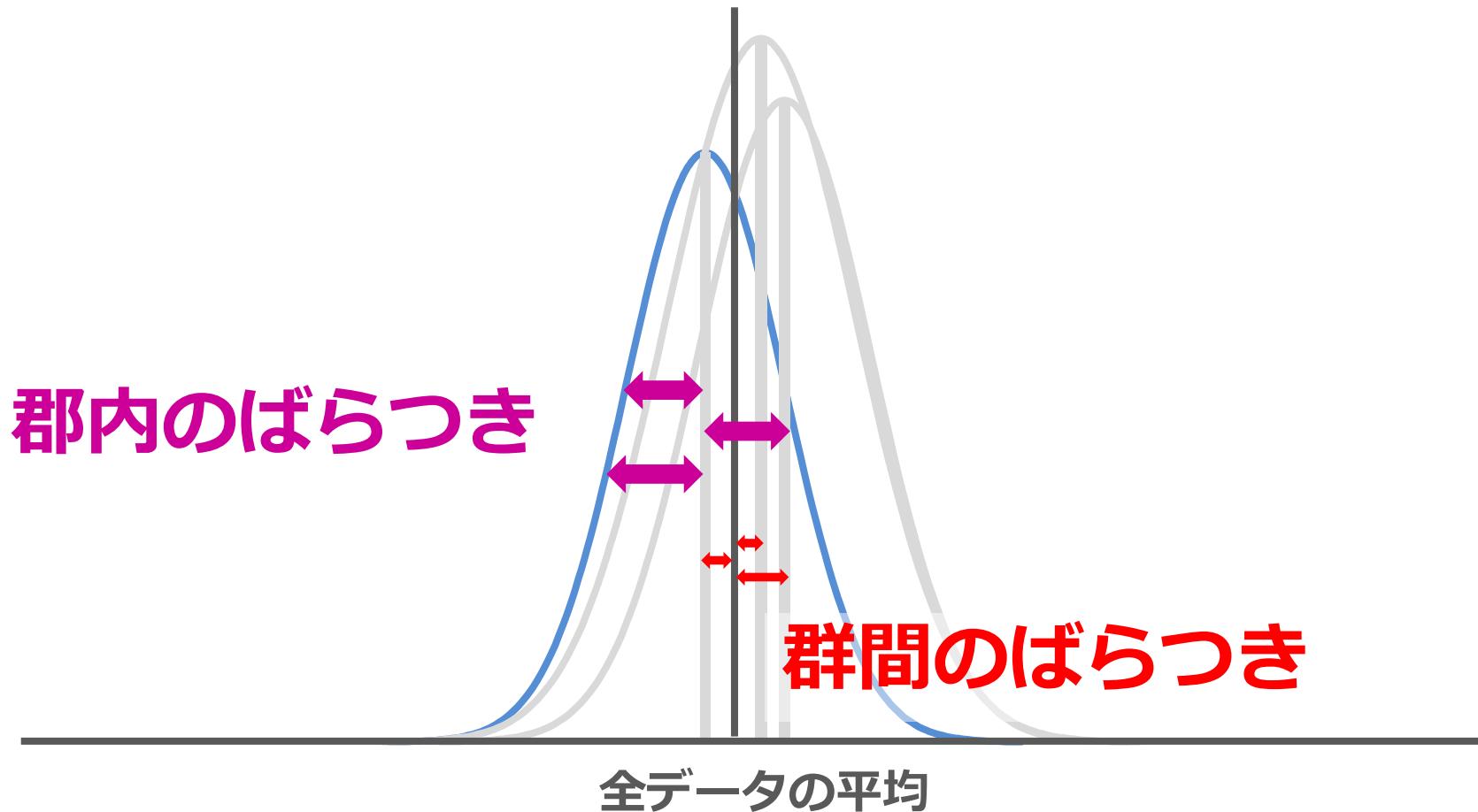
分散分析のイメージ

データのばらつきを、群間のばらつきと、偶然により起こる群内のはらつきに分けて考える



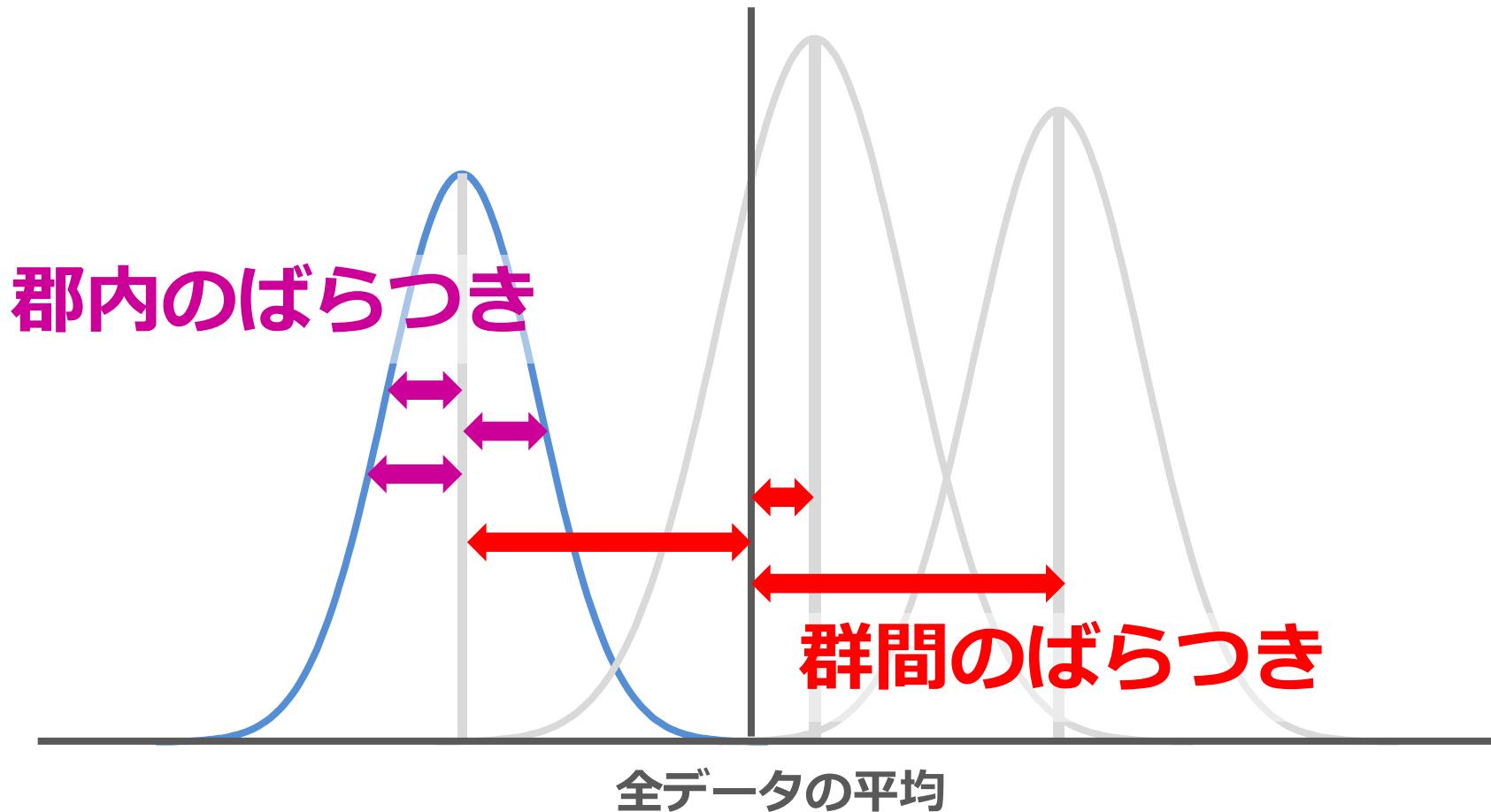
分散分析のイメージ

群の平均に差がなければ、
郡内のはらつき > 群間のはらつき



分散分析のイメージ

群の平均に差があるほど、
郡内のはらつき < 群間のはらつき



分散分析の手順

分散分析表を穴埋めしてゆく

要因	平方和 S	自由度 df	不偏標本分散 V^2	F値
群間 (因子)	$S(\text{群})$	$df(\text{群})$ =群の数-1	$V^2(\text{群})$ $=S(\text{群})/df(\text{群})$	$V^2(\text{群})/V^2(\text{残差})$
群内 (残差)	$S(\text{残差})$	$df(\text{残差})$ =全データ数-群の数	$V^2(\text{残差})$ $=S(\text{残差})/df(\text{残差})$	
全体	$S(\text{全体})$	$df(\text{全体})$		

分散分析の手順

例) A～Dの異なる生育環境で育てた植物の、ある成分の含量

A群	341	347	328	329	352
B群	305	317	342	322	319
C群	342	313	350	323	
D群	331	327	303	314	

エクセルファイル:200918_anova.xlsx

以下の基本情報を計算する

- ①群ごとのデータ数
- ②全データの個数
- ③群の平均値
- ④全データの平均値

以下の差（ずれ）を計算する

- ⑤全データについて、全体の平均からの差
- ⑥各群の平均について、全体の平均からの差
- ⑦郡内の各データについて、群平均からの差

差（ずれ）の二乗を計算する

- ⑧全データについて、全体の平均からの差の二乗
- ⑨各群の平均について、全体の平均からの差の二乗
群のデータ数を乗じる
- ⑩郡内の各データについて、群平均からの差の二乗

二乗和を計算する

- ⑪全データについての全体の平均からの差の二乗和
- ⑫各群の平均についての全体の平均からの差の二乗和
- ⑬郡内の各データについての群平均からの差の二乗和

分散分析表を埋める

⑭二乗和

$$\text{⑪} = \text{⑫} + \text{⑬} \text{となっているはず}$$

⑮自由度

全体 : ②全データ数 - 1

群間 : 群の個数 - 1

群内 : 全体の自由度 - 群間の自由度

⑯不偏標本分散（群間、群内について）

二乗和 / 自由度

⑰F値

不偏標本分散の比（群間/群内）

用語

要因 :
データに影響を与えるもの

因子 :
要因の中で特に母平均の差に
影響すると思われたため、解析
の対象とするもの

残差 :
偶然によって生じたばらつき

p値、 α のF境界値を計算する

⑯⑰で求めたF値と自由度から、F.DIST.RT関数を使って、p値を計算する

⑯有意水準 α に対応するF境界値を、F.INV.RT関数を使って計算する

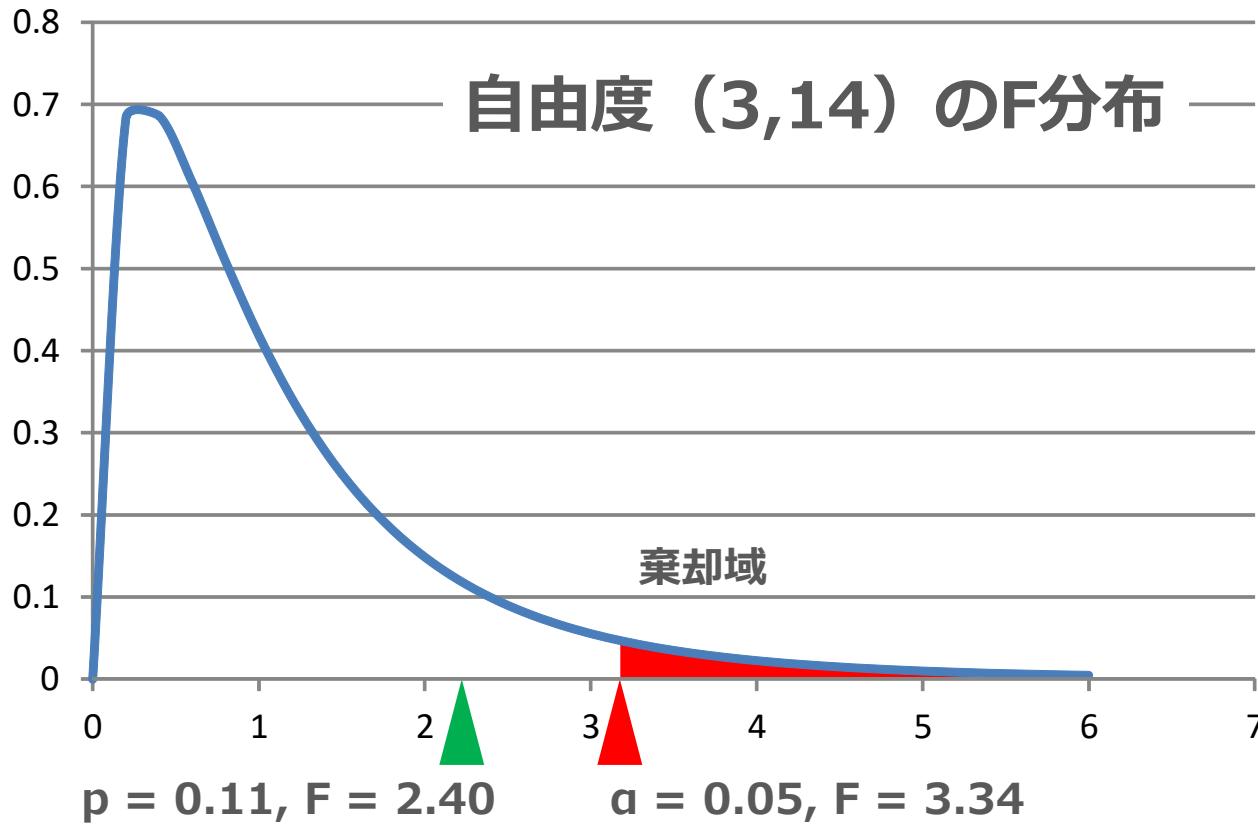
⑯F.DIST関数を用いて当該自由度のF分布を描く

p値の大きさ、 α に対応する境界値の大きさなどから、検定統計量が棄却域に入ったかどうかを判断する

結論づけをする

結論

p値は0.11となり、有意水準0.05で帰無仮説は棄却されなかった。したがって、「A～Dの生育方法によって成分の平均値に差があるとは言えない」と結論付けられた。



分散分析の種類

一元配置の分散分析 one-way ANOVA

一つの因子からなるデータを分析する方法

今回やつた
もの

二元配置の分散分析 two-way ANOVA

二つの因子からなるデータを分析する方法。例）薬剤の種類と投与量など。二つの要因が組み合わさる交互作用(相乗効果)を確認することもできる

多元配置の分散分析

情報統計 第10回

2022年8月4日 神奈川工科大学



櫻井 望
国立遺伝学研究所
生命情報・DDBJセンター

相關

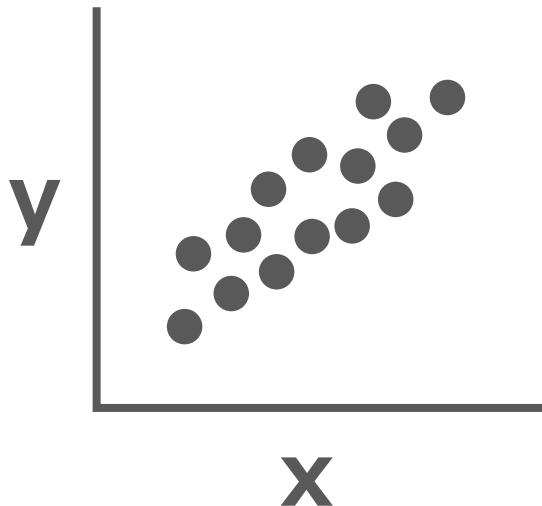
学習目標

相関のあるなしを評価できるようになる

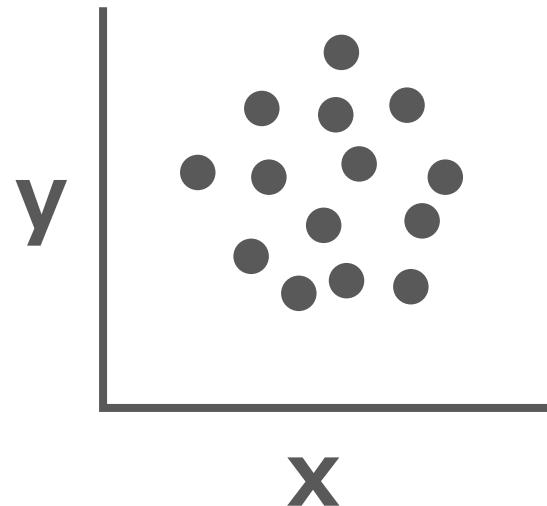
相関関係と因果関係の違いが分かる

散布図

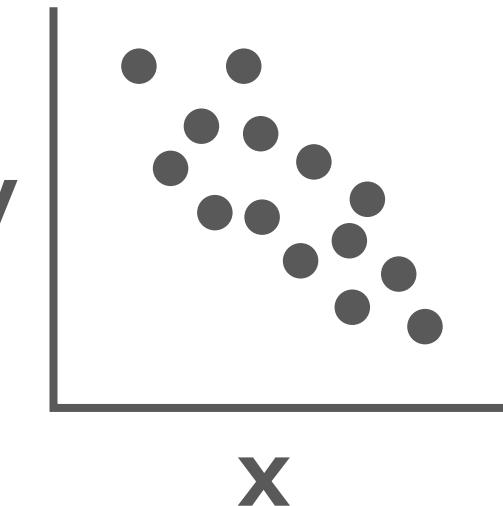
二つの変数の間の関係性を見る化する手法



正の相関がある

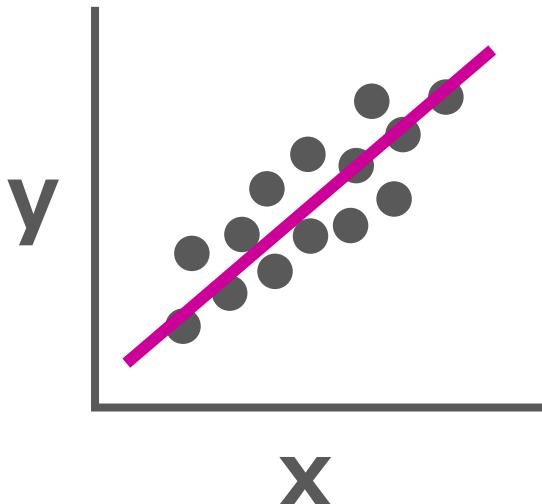


相関がない



負の相関がある

散布図の回帰曲線



エクセルのグラフ上でプロットを右クリックし、挿入できる

相関係数

- 二つの変数の間の関係性の強さを数値化したもの
- -1～1の間の値をとる

0.7～1.0：強い正の相関

0.4～0.7：中程度の正の相関

0.2～0.4：弱い正の相関

-1.0～-0.7：強い負の相関

-0.7～-0.4：中程度の負の相関

-0.4～-0.2：弱い負の相関

-0.2～0.2：相関がない

- Excelでは**PEARSON**関数で計算できる

注意点

回帰曲線のR²値は、相関係数ではありません。

R²値は、回帰曲線への当てはまり度を示すもので、「決定係数」と呼ばれます。

Excelで、原点を通らない直線近似をした場合は、ピアソン相関係数の二乗に当たります。このため、相関係数が-1～1の値を取るのに対し、R²値は0～1の値を取ります。負の相関であっても、R²が正の値を取っているのはこのためです。

生や負の相関のあるなしや、強弱を考える場合は、必ず相関係数をもとに考えましょう。

相関関係を
見てみる

都道府県別の統計

<https://todo-ran.com/>

携帯版 | スマホ版 | English

都道府県別統計とランキングで見る県民性 [とどラン]

都道府県別統計とランキングで見る県民性

<https://todo-ran.com/>

トップ	国土・インフラ	社会・政治	産業・経済	文化・くらし・健康	娯楽・スポーツ	店舗分布	その他
リクエスト	サイトについて	作者について		引用・転載について		統計八百屋	



栄養士、管理栄養士募集中

《完全無料》栄養士複数在籍、未経験歓迎など栄養士の非公開求人をご紹介



都道府県別統計を比較した都道府県ランキング。1339 ランキング掲載中

odomon@gmail.com

当サイト一番人気

都道府県
ベスト&ワースト

各都道府県の1位と47位だけを一覧表にまとめました。県民性が一目で分かります。

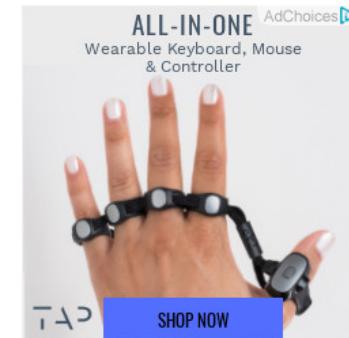
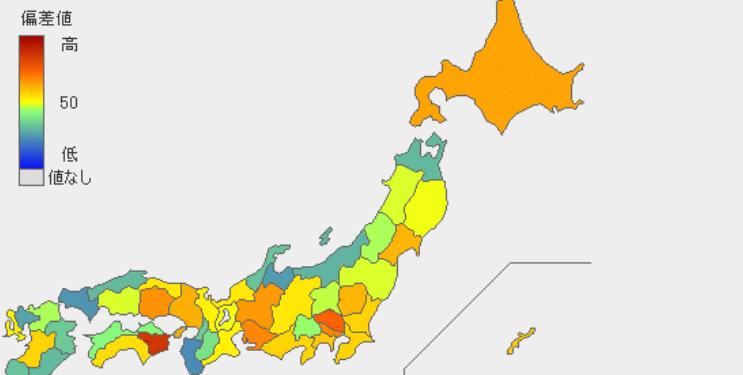
都道府県比較

東京vs大阪、埼玉vs千葉vs神奈川など任意の都道府県の似たところ、似ていないところを

トップ

最新ランキング

2019年参議院比例代表：NHKから国民を守る党得票率 [2019年 第一位 徳島県]



記事を探す

▶ 検索から探す (googleサイト内検索)

▶ カテゴリから探す

政治・経済などカテゴリ別全記事表示

▶ 新着から探す

新しい順に全記事表示

データを集めてみる

例)

神奈川県の高いランクのうち、
「しゅうまい消費量」と
「最低賃金」や「農業就業人口」との相関

- サイトでデータをコピー
- エクセルに貼り付け
- エクセルで加工（県の列で並び替え）
- 散布図を描く
- PEARSON関数で相関係数を計算する

相関係数を手で計算する

ピアソンの積率相関係数

$$r = \frac{s_{xy}}{s_x s_y}$$

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

s_{xy} : x と y の共分散

s_x : x の標準偏差

s_y : y の標準偏差

n : x と y のペアの数

無相関の検定

帰無仮説：
母集団の相関係数は0（無相関）である

分布： t 分布

検定統計量：

$$t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$$

自由度： $n-2$

※ $|r|$ は r の絶対値
エクセルでは ABS 関数
で計算できる

そのほかの相関係数

- スピアマンの順位相関係数
- コサイン相関係数

相関と因果

相関関係：

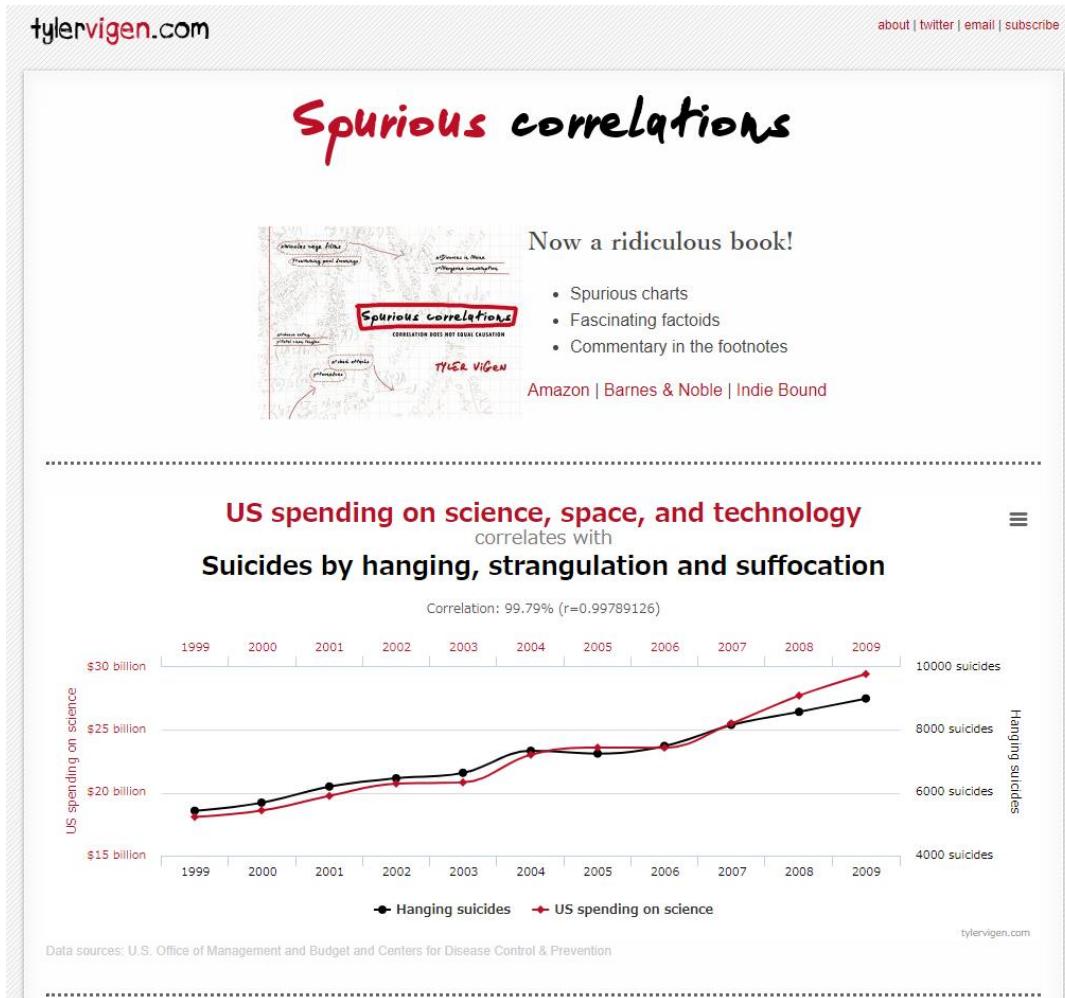
二つの事柄に関連性がある

因果関係：

二つの事柄が、原因と結果の関係である

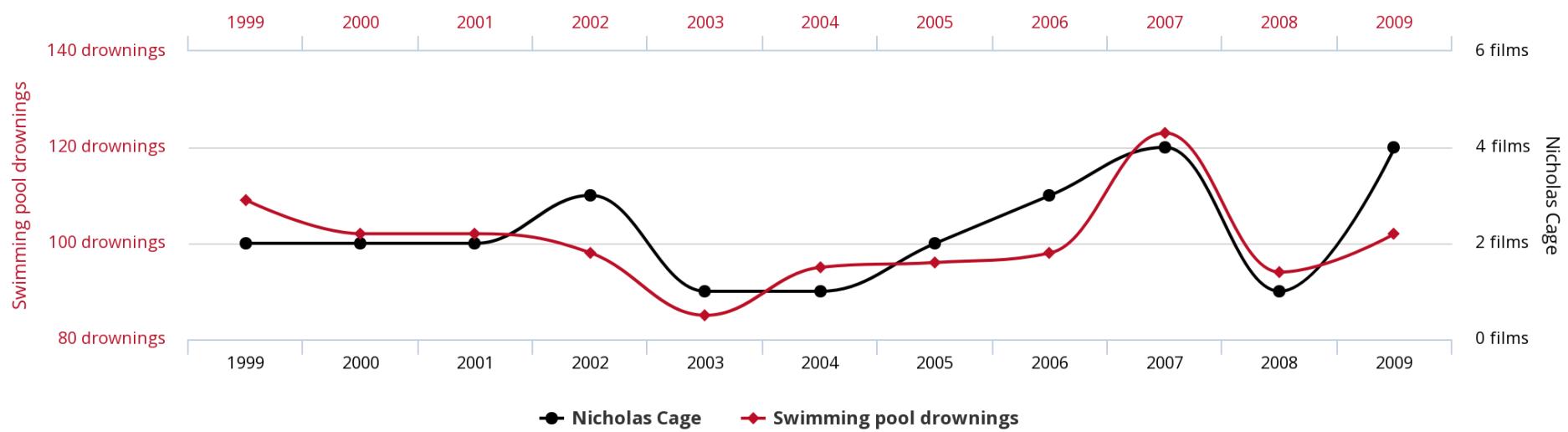
疑似相關

<https://www.tylervigen.com/spurious-correlations>



ニコラス・ケイジの映画出演本数と、 プールでおぼれた人の数に、 高い相関がある？

Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in



中室牧子
Makiko Nakamuro
津川友介
Yusuke Tsugawa

Causal
Inference
in Economics
How to measure the "causal" in everyday life

データから
真実を見抜く
思考法

「テレビを見せると子どもの学力が下がる」は
なぜ間違いなのか？世の中にあふれる
「根拠のない通説」
世界中の経済学者がこぞって用いる
最新手法をわかりやすく解説。

西内 啓



『統計学が最強の学問である』著者
統計学と経済学の最新の知見を凝縮！

原大と結果の 経済学

中室牧子、津川友介著、
ダイヤモンド社2017年

情報統計 第11回

2022年8月4日 神奈川工科大学



櫻井 望
国立遺伝学研究所
生命情報・DDBJセンター

多变量解析

学習目標

主成分分析について

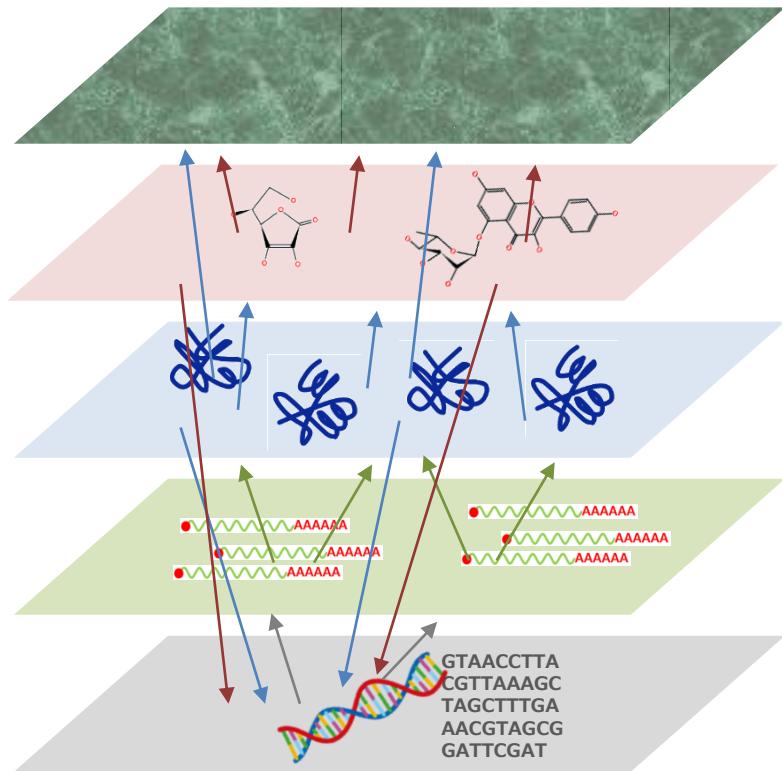
- 概念を理解する
- 結果の解釈の仕方を理解する

多変量データの例

- 大規模な疫学研究データ
- 生物等のオミクスデータ

など

生物の遺伝子情報の流れとオミクス



オミクス

表現型

代謝成分

タンパク質

転写産物

ゲノム

?

数万?

数万

数万

数万

それぞれの要素を一斉に検出
しようとする技術・学問

多変量解析の目的

- データを要約して解釈しやすくする
- データに含まれる潜在的な因子を見つける
- 状況を判別したり、分類したりする
- 状況を予測する

さまざまな多変量解析

- 似ているものをグルーピングする
クラスター解析
- データを要約する
主成分分析
- 判別、分類、予測
判別分析、PLS、PLS-DA、
重回帰分析
など

主成分分析

主成分分析で扱うデータ

組織ごとの生体試料など

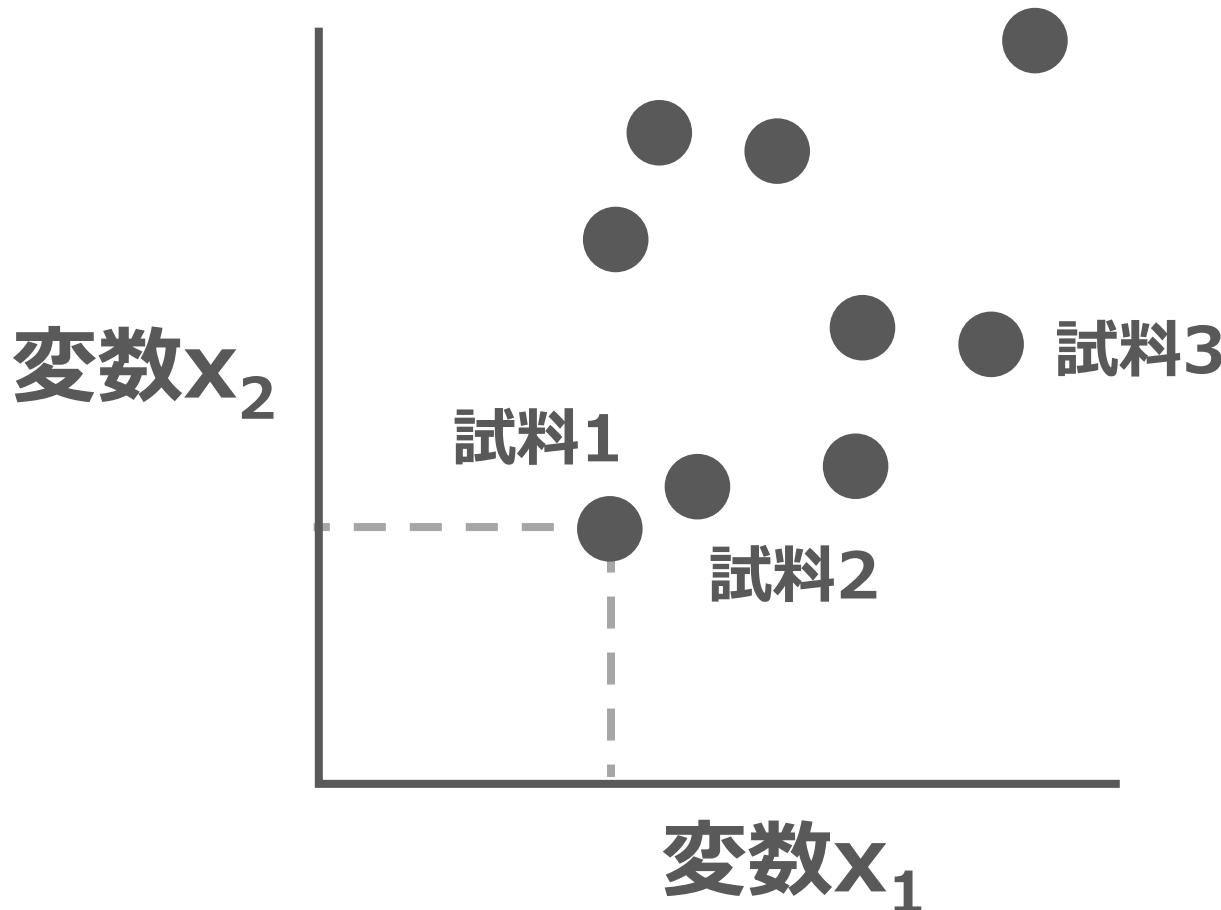
		対象				
		1	2	3	…	n
変数	X_1	X_{11}	X_{21}	X_{31}		X_{n1}
	X_2	X_{12}	X_{22}	X_{32}		X_{n2}
	X_3	X_{13}	X_{23}	X_{33}		X_{n3}
	…					
	X_m	X_{1m}	X_{2m}	X_{3m}		X_{nm}

遺伝子など
説明変数, 観測変数

遺伝子発現量など

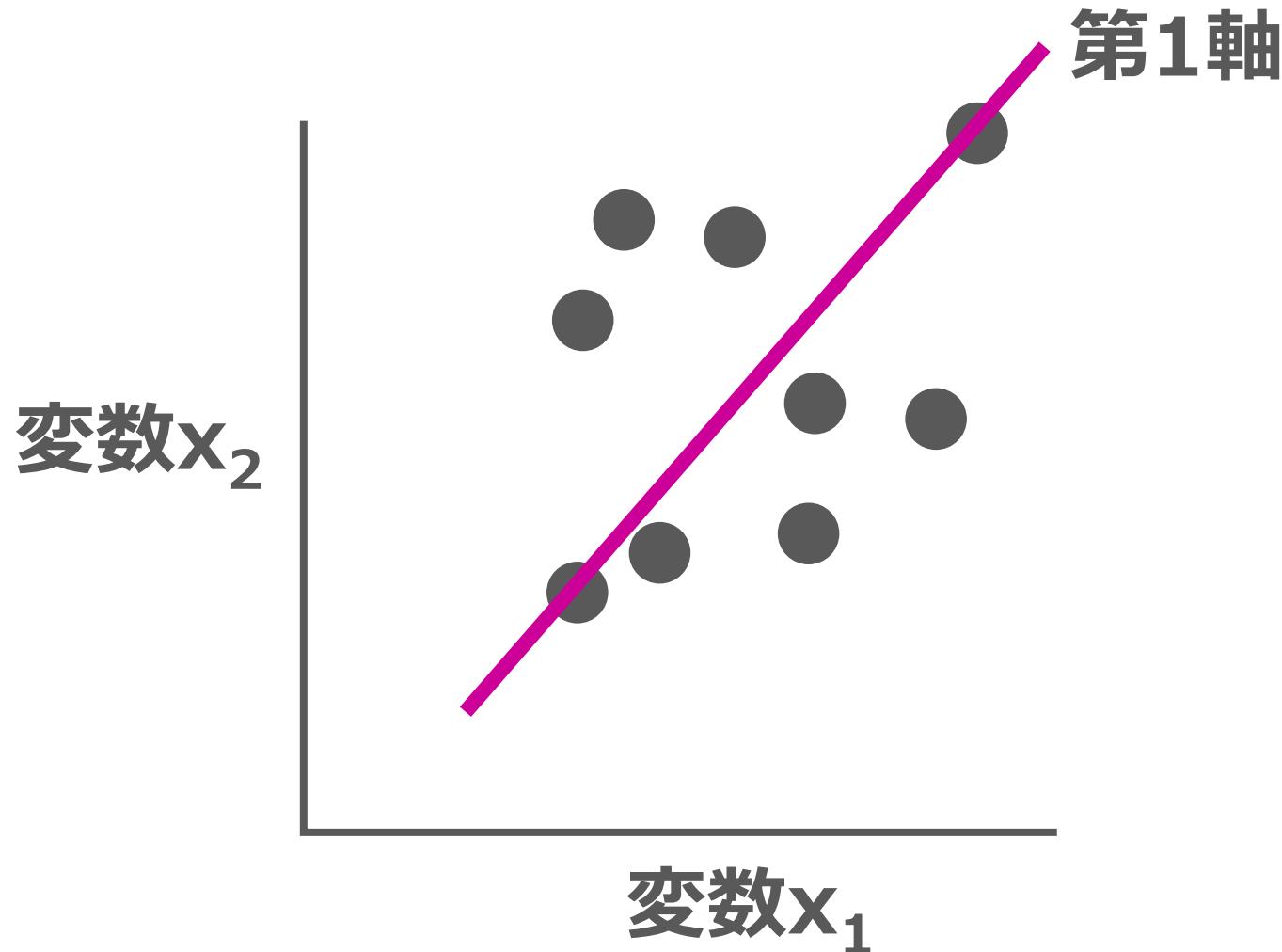
主成分分析のイメージ

- ① 例えば変数が2個しかないとき、2次元の散布図に、試料ごとに変数をプロットできる



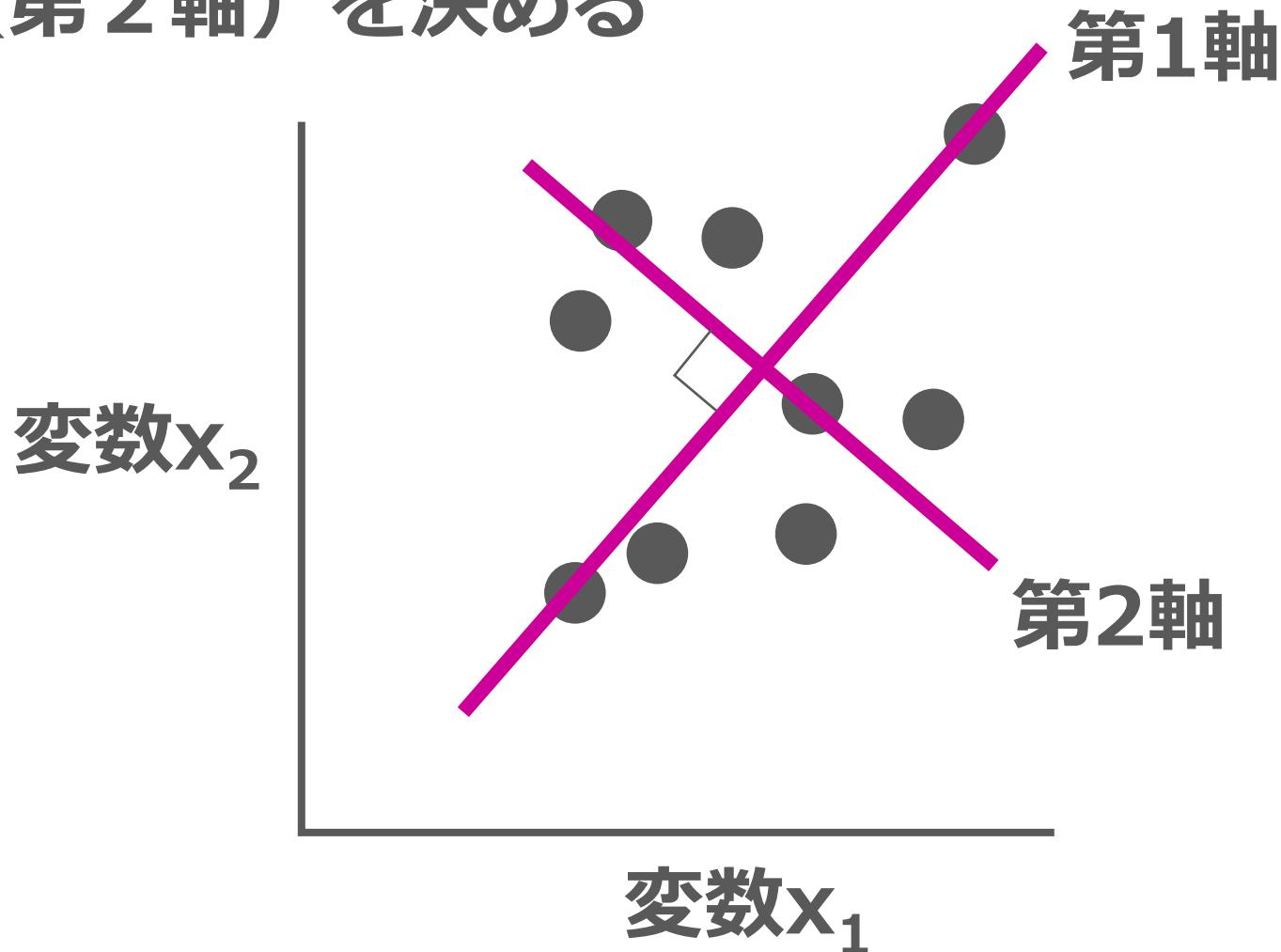
主成分分析のイメージ

②一番分散の大きい軸（第1軸）決める



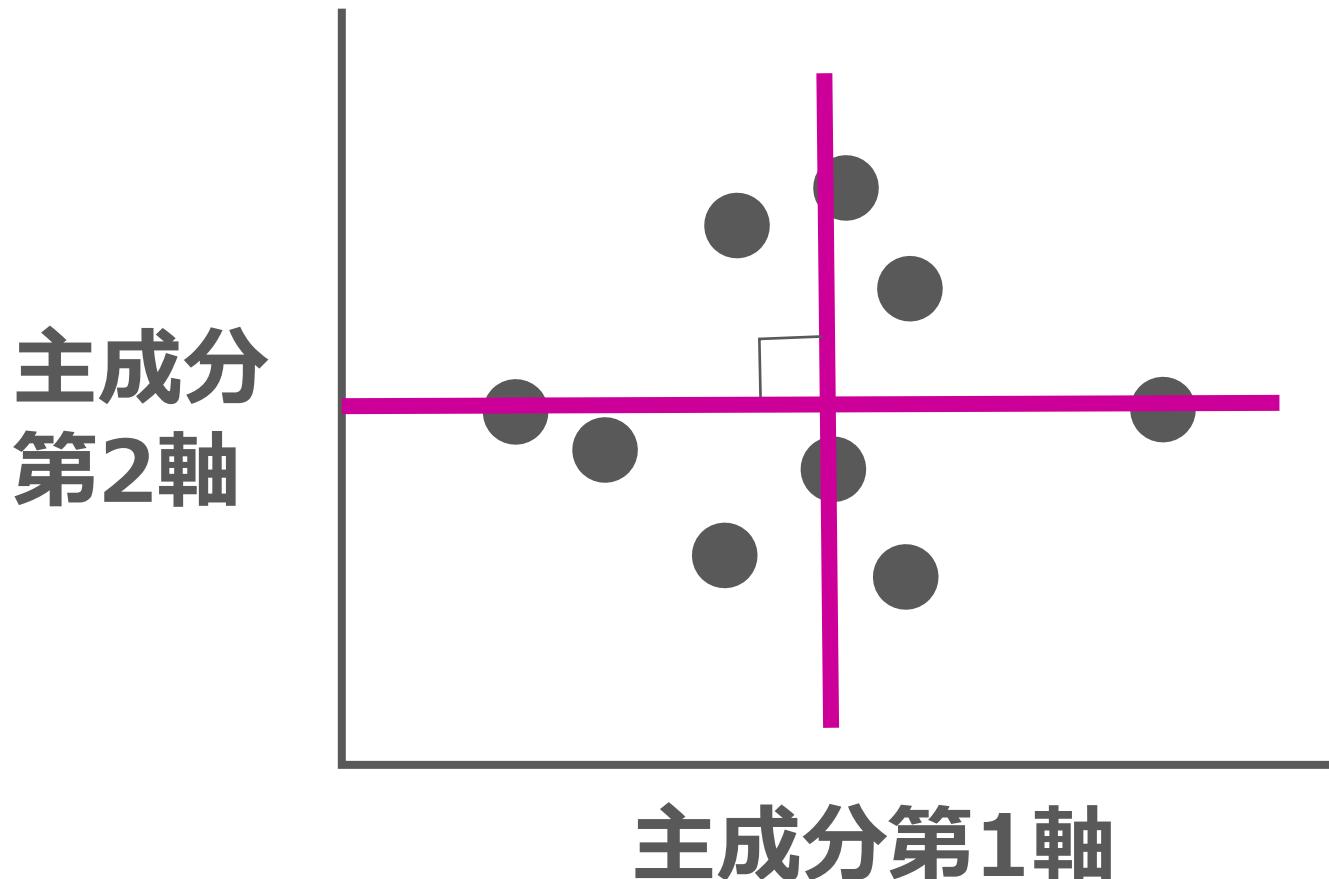
主成分分析のイメージ

③第1軸に直角に交わり、次に分散が大きい軸
(第2軸) を決める



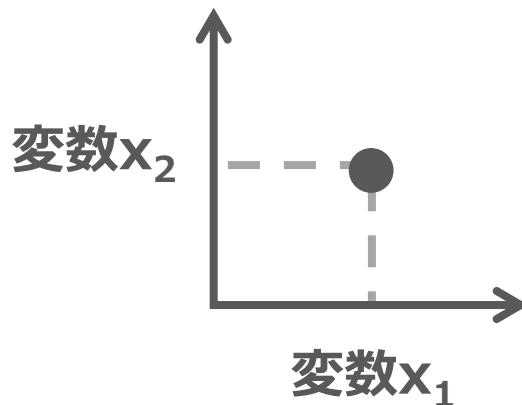
主成分分析のイメージ

④第1軸がx軸、第2軸がy軸になるように、図を回転させた新たな図を作る

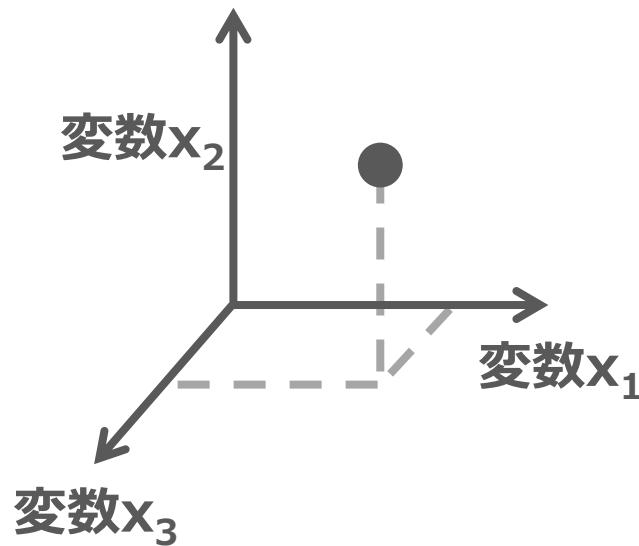


主成分分析のイメージ

m 個の変数の値を m 次元の図にプロットし、同様の計算を行うことが可能



変数2個
2次元



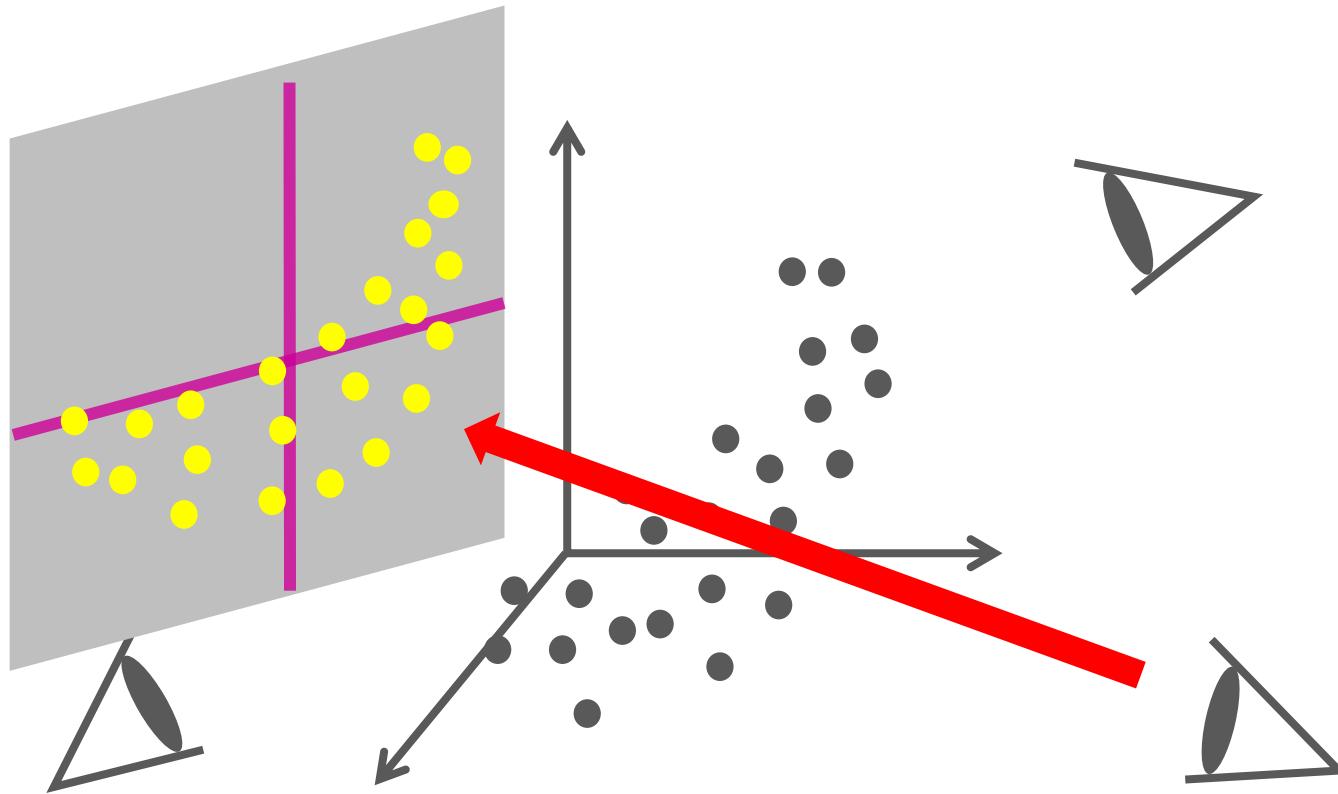
変数3個
3次元



変数 m 個
 m 次元

主成分分析のイメージ

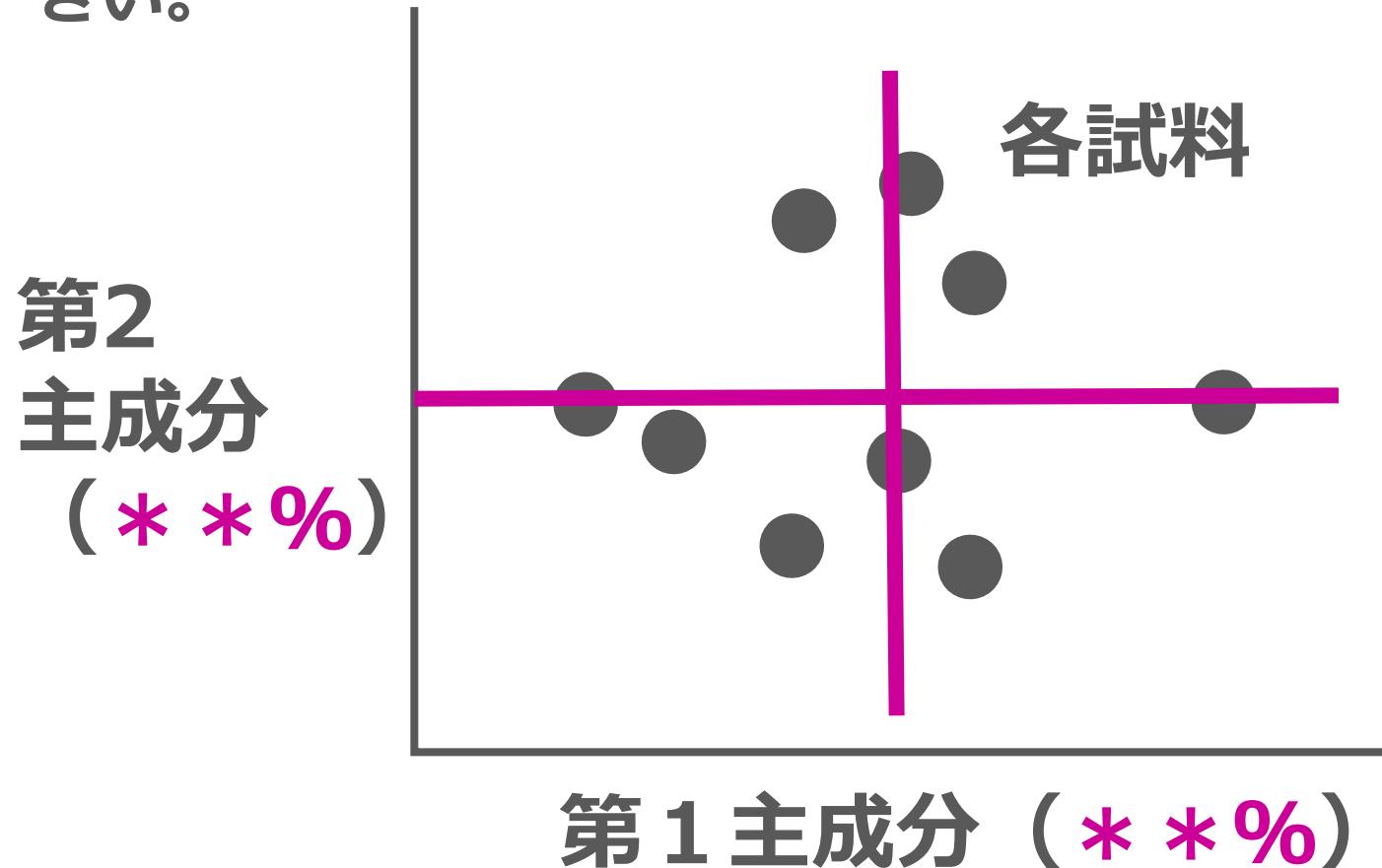
試料間の違い（特徴）が一番はっきりと見える方向から見た図が描ける



スコアプロット

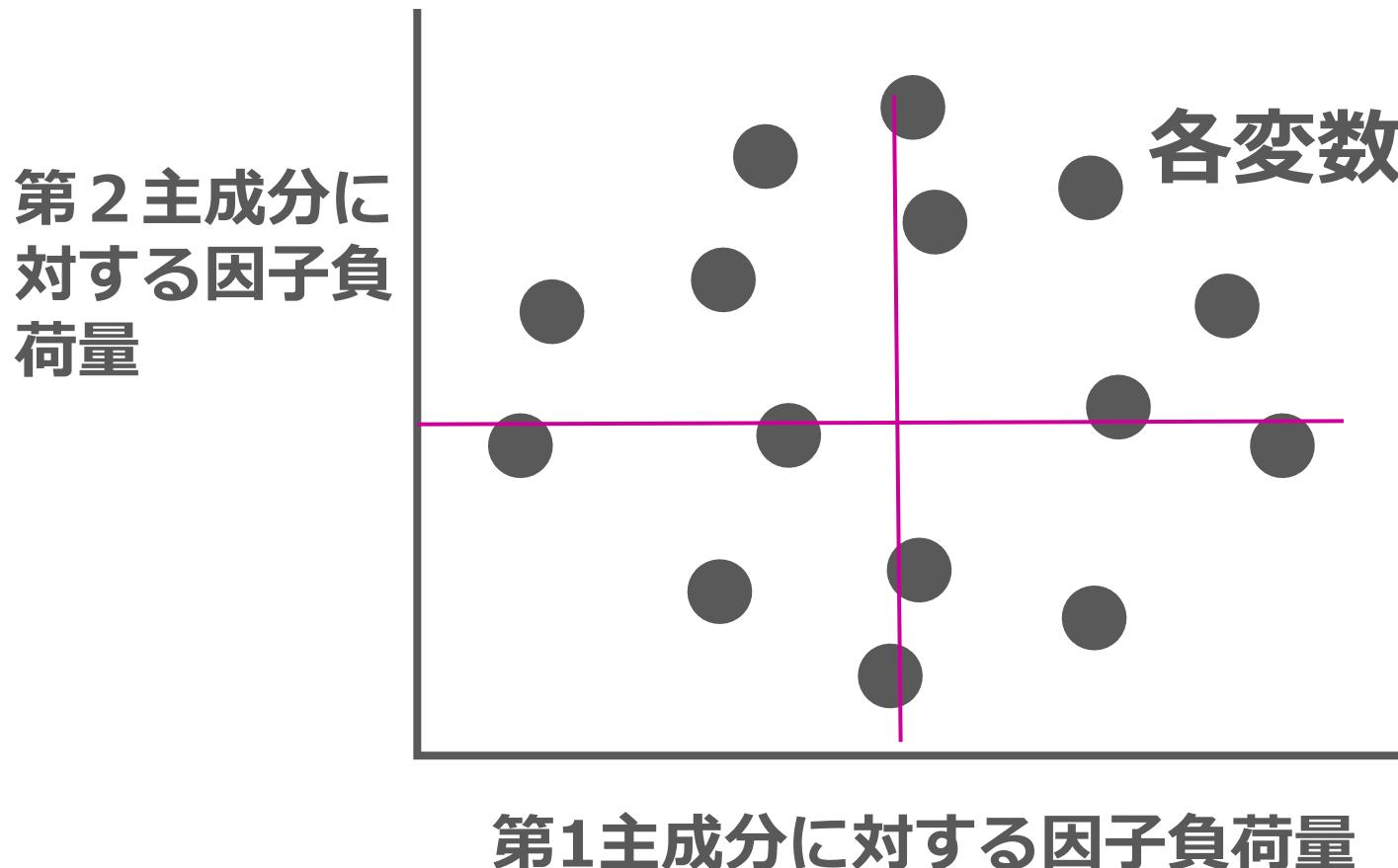
主成分軸に各試料を投影しなおした図

軸に示した%は寄与率と呼び、全体の分散のうち各主成分軸が説明する分散の比率を表す。第1主成分の寄与率が最も大きい。



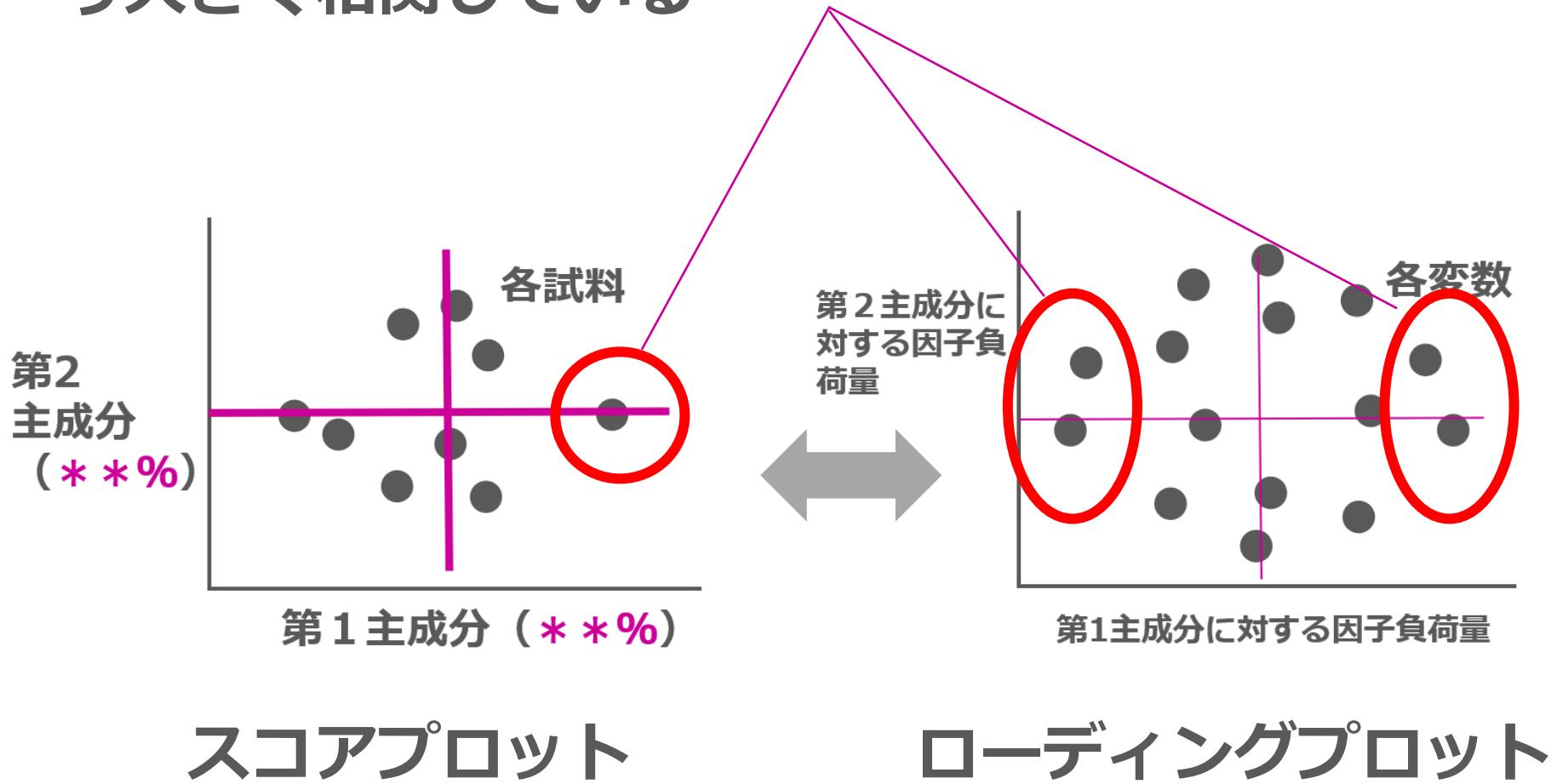
ローディングプロット

ローディングは、因子負荷量とも呼ばれ、各試料の主成分スコアと、変数の間の相関係数に相当する。
(厳密には、数値の前処理の条件などいくつか制約がある)



二つの図をセットで見る

この試料と他の試料との違いは、これらの変数がより大きく相関している



そのほかの 多変量解析

さまざまな多変量解析

- 似ているものをグルーピングする
クラスター解析
- データを要約する
主成分分析
- 判別、分類、予測
判別分析、PLS、PLS-DA、
重回帰分析
など

PLS

Partial Least Squares

部分最小二乘

PLS-DA

Partial Least Squares-Discriminant Analysis

部分最小二乘-判別分析

PLS、PLS-DAで扱うデータ

目的変数が存在する

説明変数との関連を調べたい試料の分類や、試料の特徴量など

例) 別途測定した、生理活性データなど

目的変数

組織ごとの生体試料など

		対象				
		1	2	3	…	n
変数	y_1	y_{11}	y_{21}	y_{31}		y_{n1}
	y_2	y_{12}	y_{22}	y_{32}		y_{n2}
	...					
変数	y_p	y_{1p}	y_{2p}	y_{3p}		y_{np}
	x_1	x_{11}	x_{21}	x_{31}		x_{n1}
	x_2	x_{12}	x_{22}	x_{32}		x_{n2}
	x_3	x_{13}	x_{23}	x_{33}		x_{n3}
	...					
	x_m	x_{1m}	x_{2m}	x_{3m}		x_{nm}

遺伝子など
説明変数、観測変数

遺伝子発現量など

PLS、PLS-DAで得られる結果

- PCAと類似したスコアプロットとローディングプロットが得られる
- 目的変数 (y) を説明変数 (x) で説明するためのモデルが構築される
- 目的変数を説明する変数重要度 (VIP) が計算される

情報統計 第12回

2022年8月4日 神奈川工科大学



櫻井 望
国立遺伝学研究所
生命情報・DDBJセンター

おさらい

やったこと

- 統計的手法
- 記述統計
 - ✓ 平均値等の計算
 - ✓ 相関係数、回帰式
- 推測統計
 - ✓ 推定、仮説検定
- 多変量解析
 - エクセル関数
 - プログラミング
 - Python

統計って？

集団の状況を
数値で表したもの



目的：集団の〇〇を知りたい

統計学

- データを集める
- 解析する
- 解釈する

ための方法論



結果：集団の〇〇がわかった！

重要！

結論を言う

統計的結論から、設定した目的に
対する結論を導くことが最も重要。

発表会の テンプレート

表紙 1枚

- タイトル
- 名前
- 報告日など

背景と目的 1~枚

- 何に疑問を持ち、どんな目的のためにこの課題を行ったか？
- その疑問に至った背景

方法のページ 1~枚

- どんなデータ、どんな統計的手法を使って実施したか。

だれもが追試、検証できるよう

結果のページ 1~枚

- どんな結果が得られたか
- そこから言えることは何か

結果に基づいて得られた情報について述べる

考察のページ 1~枚

- 結果を総合して、目的に対してどんな結論が得られたか

最初に掲げた疑問に対する答えや、得られた結果の価値について述べる

(将来展望のページ 1~枚)

もしあれば

- 今後こんなデータを集めれば…
- 今後こんな統計的手法を適用すれば…



もっとこんなことがわかるだろう、など

未来に対する夢を述べる

よいスライドの作り方



田中佐代子著、
講談社2013年

自習

課題準備

情報統計

第13-15回

2022年8月5日 神奈川工科大学



櫻井 望
国立遺伝学研究所
生命情報・DDBJセンター

補足

- 数学記号
- ログ変換
- 主成分分析の例

2群のt検定（独立2群）

等分散が仮定できない場合 ウエルチの方法

1群目：標本数 n_1 , 不変標本分散 s_1^2 , 標本平均 \bar{x}_1

2群目：標本数 n_2 , 不変標本分散 s_2^2 , 標本平均 \bar{x}_2

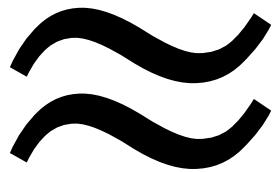
検定統計量 $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

(近似)自由度 $v \approx \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}}$

帰無仮説：2群の母集団の平均値は等しい

で、同様に検定できます

参考まで



ほぼ等しい

数学記号

○	合成写像	「 $f \circ g$ 」は写像 g と写像 f の合成を表す。すなわち $(f \circ g)(x) = f(g(x))$ である。
Im, Image, • [•]	像	写像 φ に対して、Image φ はその写像の像全体の集合（値域）を表す。写像 $\varphi: X \rightarrow Y$ に対して $\varphi[X]$ とも書く。

二項関係演算

記号	意味	解説
=	相等	$x = y$ は x と y が等しいことを表す。
≠	不一致	$x \neq y$ は x と y が等しくないことを表す。
≒, ≈	ほぼ等しい	「 $x \doteq y$ 」または「 $x \approx y$ 」は x と y がほぼ等しいことを表す。記号 \doteq は日本など少数の地域でのみ通用し、 \approx の方が標準的である。その他にも \sim , \simeq , \cong などを同様の意味で用いることもある。近似においてどのくらい違いを容認するかは文脈による。多くの場合、誤差解析的な意味で用いられ、ある誤差の見積もりの下で両者が等しいことを示すが、そのほかにも漸近解析においては漸近的に等しいという意味で用いられる。

順序構造

記号	意味	解説
<, >	大小関係、順序	「 $x < y$ 」は x と y の間に左側の方が「先」であることを示す

Wikipedia

Excelで数式表示

作図.xlsx - Excel

Nozomu Sakurai NS

共有 コメント

ファイル ホーム 挿入 描画 ページレイアウト 数式 データ 校閲 表示 ヘルプ 検索

fx オートSUM 論理 検索/行範囲 名前の定義 参照元のト雷斯

関数の 最近使った関数 文字列操作 数学/三角 フォーマットで使用 参照先のトレス

挿入 財務 日付/時刻 その他の関数 ワークシート分析 エラー チャック

関数ライブラリ 定義された名前 選択範囲から作成 ワンチ ウィンドウ

計算方法の設定 計算方法

J2

C D E F G H I J

	C	D	E	F	G	H	I	J
1								
2		平均値からの差	←の二乗					
3	148	=C3-\$H\$6	=D3^2					
4	148	=C4-\$H\$6	=D4^2	合計	=SUM(C3:C28)			
5	149	=C5-\$H\$6	=D5^2	個数	=COUNTA(C3:C28)			
6	150	=C6-\$H\$6	=D6^2	平均値	=H4/H5	手計算		
7	150	=C7-\$H\$6	=D7^2	平均値	=AVERAGE(C3:C28)	AVERAGE関数		
8	150.4	=C8-\$H\$6	=D8^2	二乗和	=SUM(E3:E28)			
9	151	=C9-\$H\$6	=D9^2	分散	=H8/H5			
10	153	=C10-\$H\$6	=D10^2	分散	=VAR.P(C3:C28)	VAR.P		
11	153	=C11-\$H\$6	=D11^2	不偏標本分散	=H8/(H5-1)			
12	153.4	=C12-\$H\$6	=D12^2	不偏標本分散	=VAR.S(C3:C28)	VAR.S		
13	155	=C13-\$H\$6	=D13^2	標準偏差	=SQRT(H9)			
14	155	=C14-\$H\$6	=D14^2	標準偏差	=STDEV.P(C3:C28)	STDEV.P		
15	155.5	=C15-\$H\$6	=D15^2	不偏標本標準偏差	=SQRT(H11)			
16	156.6	=C16-\$H\$6	=D16^2	不偏標本標準偏差	=STDEV.S(C3:C28)	STDEV.S		
17	157	=C17-\$H\$6	=D17^2					
18	157	=C18-\$H\$6	=D18^2					

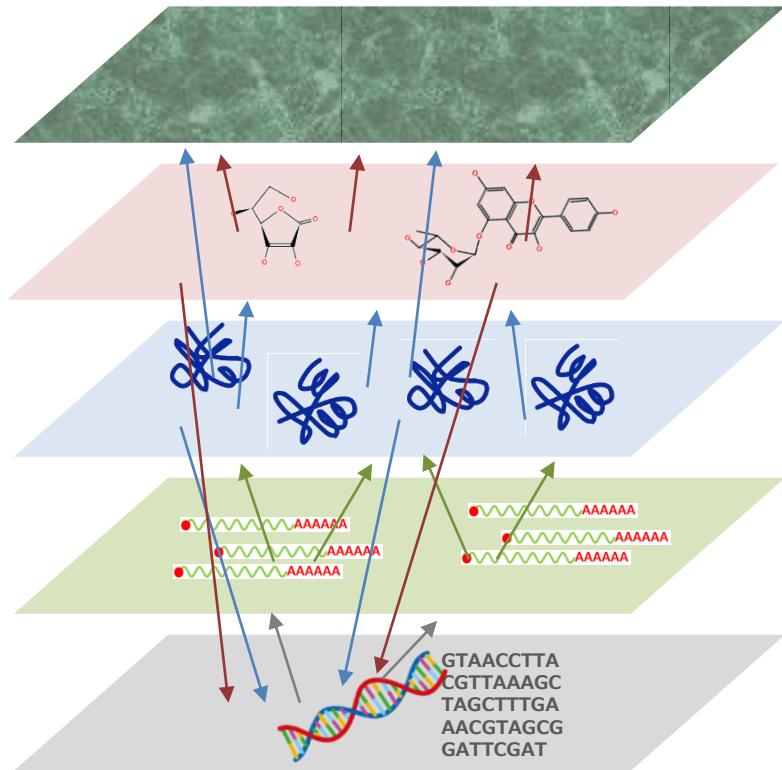
Sheet6 Sheet5 Sheet7 Sheet8 Sheet9 Sheet10 Sheet10 (2) +

100%

補足

- 数学記号
- ログ変換
- 主成分分析の例

生物の遺伝子情報の流れとオミクス



オミクス

表現型

代謝成分

タンパク質

転写産物

ゲノム

?

数万?

数万

数万

数万

それぞれの要素を一齊に検出
しようとする技術・学問

一見、正規分布のように見えないデータでも、ログスケール（対数）にすることで、正規分布に近い分布になることがある

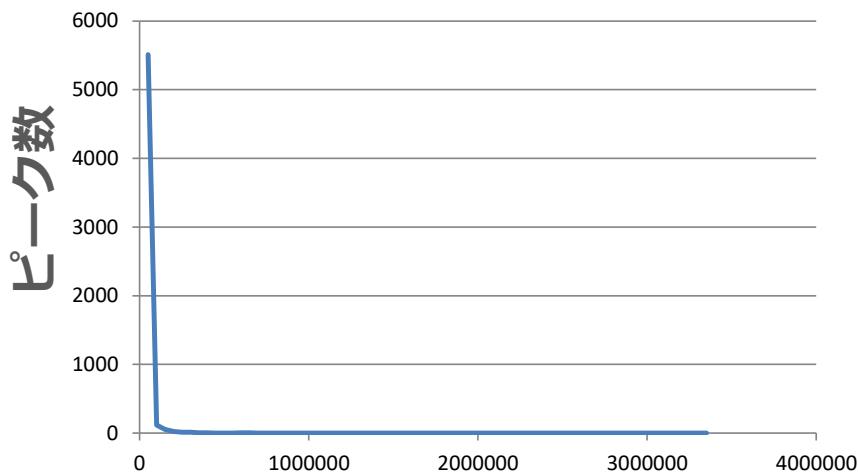
- ✓ 遺伝子発現量データ
- ✓ 質量分析での化合物検出データ

など

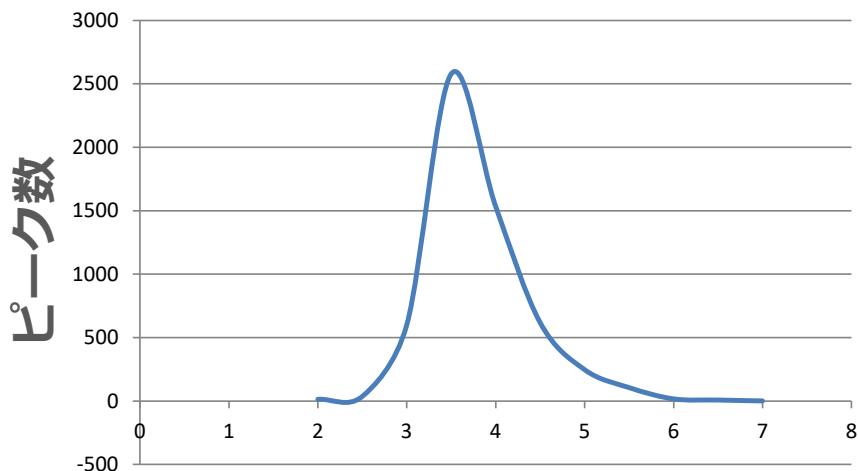
大葉（しそ）で検出された代謝物質

- 液体クロマトグラフィー-質量分析
- ESIポジティブモード

計5760ピーク



検出値
(リニアスケール)



log10変換後
(ログスケール)

Excel関数: LOGなど

ログスケールにするメリット

シグナル強度によるばらつき（分散）の変化を打ち消すことができる

例) 強度10のピークの10%のばらつきは1の差なのに対し、強度1000のピークでは、同じ10%のばらつきで100の差になる。

logに変換すると、どんな強度でも同じ数値幅のばらつきにすることができる（等分散）



データの分布をExcelで描いて判断

補足

- 数学記号
- ログ変換
- 主成分分析の例

自習

課題檢討

発表会