情報統計
第1回

2025年8月4日　神奈川工科大学

櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

スケジュール

	8月4日(月) データの見える化、統計の基本	8月5日(火) 確率分布の使い方、プログラミング	8月6日(水) 検定、分散分析、多変量解析の雰囲気	8月7日(木) 解析の例、まとめ、発表会
1限 9:30 ～ 11:10			7　t検定、検定で注意すること	11 主成分分析やデータ解析の例
2限 11:20 ～ 13:00		4　区間推定、確率分布とその使い方	8 分布の仲間と分散分析	12　授業のまとめ
昼休み 13:00～13:40				
3限 13:40 ～ 15:20	1　ガイダンス、PC環境準備、データの見える化、相関	5　プログラミングの基礎	9 多変量解析、主成分分析	13　自習(課題、質問)
4限 15:30 ～ 17:10	2　統計の基本、平均値の推定(中心極限定理)	6　自習(課題検討、復習)	10　自習(課題検討、復習)	14　発表会
5限 17:20 ～ 19:00	3　自習(課題検討、復習)			

授業で使うサイト

https://github.com/nsaku/kait2025/wiki

要パスワードサイト

ID:	
PW:	

注意　必ずログアウト！

教室の端末を使って、ご自分のGoogleアカウントでブラウザでにログインしたら、使用後に忘れずにログアウトしてください。

ガイダンス

統計の知識は

大学で

- 実験の計画
- データの分析・評価

仕事で

管理栄養士さん

- 調査研究等の設計
- データの分析・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

統計は

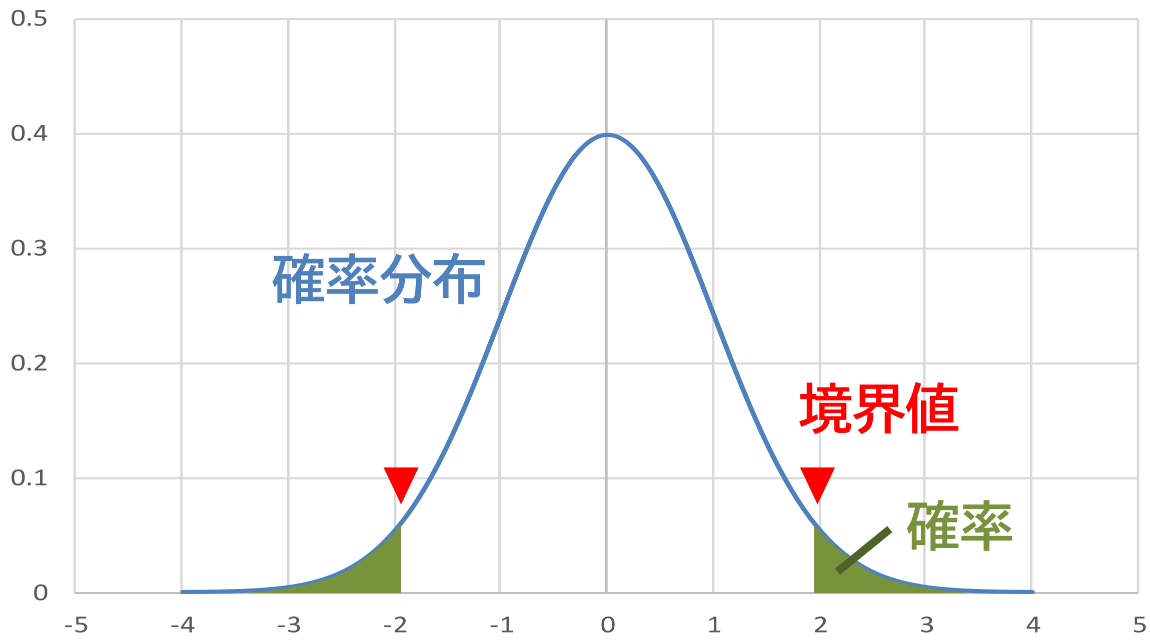
無関心ではいられるが
無関係ではいられない

学習すること

- データの見える化
- 検定の基礎
- 分散分析
- 多変量解析の雰囲気

これだけはおさえておきたい大事なポイント

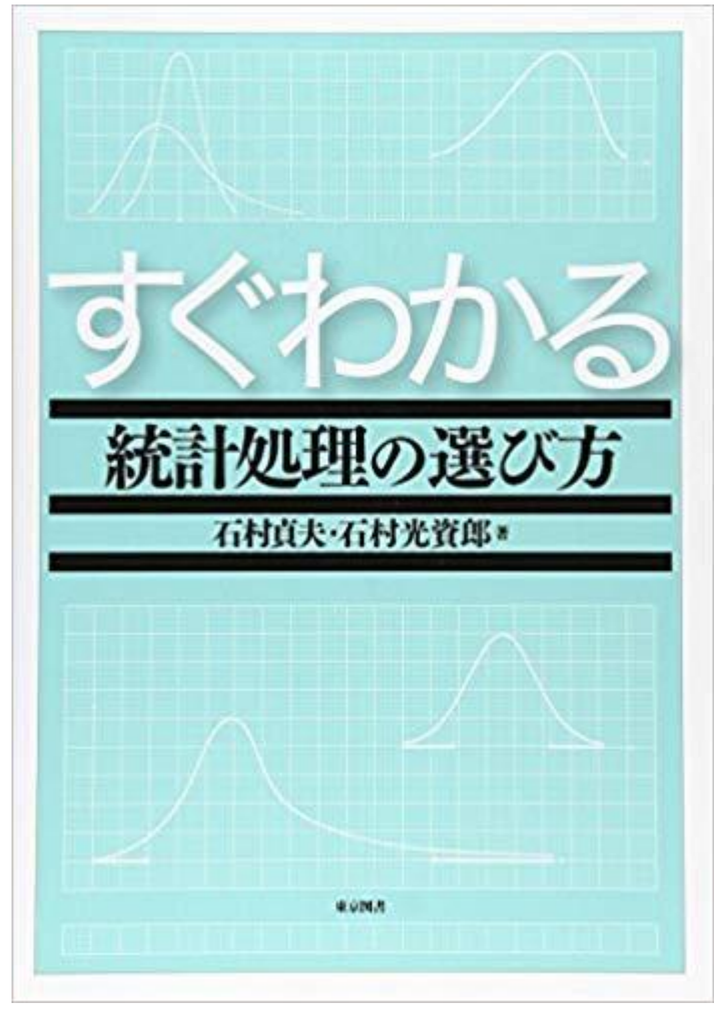
- 中心極限定理
- 確率分布とその使い方



参考図書は

参考図書は


すぐわかる
統計用語の基礎知識
石村貞夫＋デズモンド・アレン＋劉晟 著


すぐわかる
統計処理の選び方
石村貞夫・石村光資郎 著

参考動画

3brown 1blue Japan

https://www.youtube.com/@3Blue1BrownJapan

数学のいろいろなトピックを、直感的にわかるアニメーションを使って視覚化しながら詳しく解説するチャンネル。
（少し難しい）

中心極限定理

https://youtu.be/XXsTu66VB-E?si=4r17NgQ-PWGKp6dD

参考情報

AI: ChatGPT, Gemini, Copilot, Claude, Elicitなど

統計を基礎から勉強しています。中心極限定理とは何ですか？

いいですね、統計の勉強を基礎から進めるのはとても大切です。
「**中心極限定理（ちゅうしんきょくげんていり、Central Limit Theorem：CLT）**」は、統計学の中でも
非常に重要な定理の1つです。以下に、できるだけわかりやすく説明します。

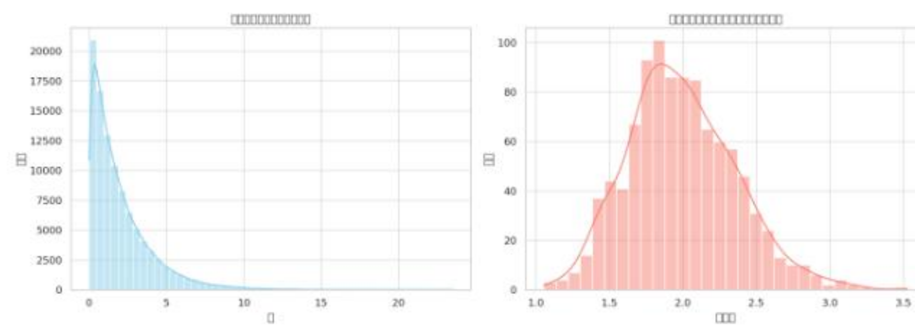
■ 中心極限定理とは？

> **たとえ元のデータ（母集団）がどんな分布であっても、平均をたくさん集めていく**
> **布は正規分布に近づく**という定理です。

■ もう少し詳しく説明すると…

1. ある母集団（分布の形は問わない、偏っていてもOK）があるとします。
2. この母集団から **n個のデータを取り出して、その平均を計算する。**
   → これを **何回も繰り返す。**
3. そうやって集めた「平均の集まり（＝標本平均の分布）」を調べると…
4. **その分布は「正規分布（ガウス分布）」に近づいていく**のです（nが大きくなると）

視覚的にわかるシミュレーションを見たいです。



こちらが視覚的なシミュレーションです：

- **左側**のグラフは「母集団の分布」を示しており、**指数分布**という偏った形をしています。これは現実のデータのように、偏りのある分布を想定したものです。
- **右側**のグラフは、その母集団から「30個ずつの標本を1000回」取り出して**平均を求めた値の分布**です。こちらはほぼ**正規分布の形**に近づいています。

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る

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

Step1 情報収集： 1～4のいずれかでデータを集める

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、データを取得する
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、さらに元データを入手する
3. 独自のアンケートを作成して、データを収集する
4. 1～3に代わる何か

Step2 解析・考察： 集めたデータで独自のグラフなどを作成しながら、統計的に解析して、結果を得て、考察する。

Step3 発表： 行ったことを発表する。

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）
- 相関

統計
学

統計って？

コトバンク　コトバンクを検索　検索

統計 （読み）トウケイ

デジタル大辞泉の解説

とう‐けい【統計】

［名］(スル)集団の個々の構成要素の分布を調べ、その集団の属性を数量的に把握すること。また、その結果を数値や図表で表現したもの。「統計をとる」「統計を出す」「就業人口を統計する」

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。集団を一時点で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　株式会社平凡社 / 百科事典マイペディアについて　情報

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を統計とよんでいる。［泉　俊衛］

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、主として狭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。自然現象や単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つ意義や特質は初めからとくに問題に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な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、特定の時点ないし時間と場所ないし範囲の規定）に基づいてとらえられる。その場合、統計集団をいかに構成するか、より具体的にい

ChatGPT

「統計（とうけい）」とは、**データを集めて整理・分析し、そこから意味のある情報や傾向を導き出す方法や学問のこと**です。

もう少し具体的に説明すると：

🌟 統計の主な目的

1. **データの要約（記述統計）**
   多くのデータを平均・中央値・分散などの指標を使って分かりやすくまとめます。
   例：クラスのテストの平均点や最高点・最低点などを出す。
2. **データからの推測（推測統計）**
   一部のデータ（標本）から全体（母集団）について推測すること。
   例：100人にアンケートして、日本全体の意見を予測する。

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！

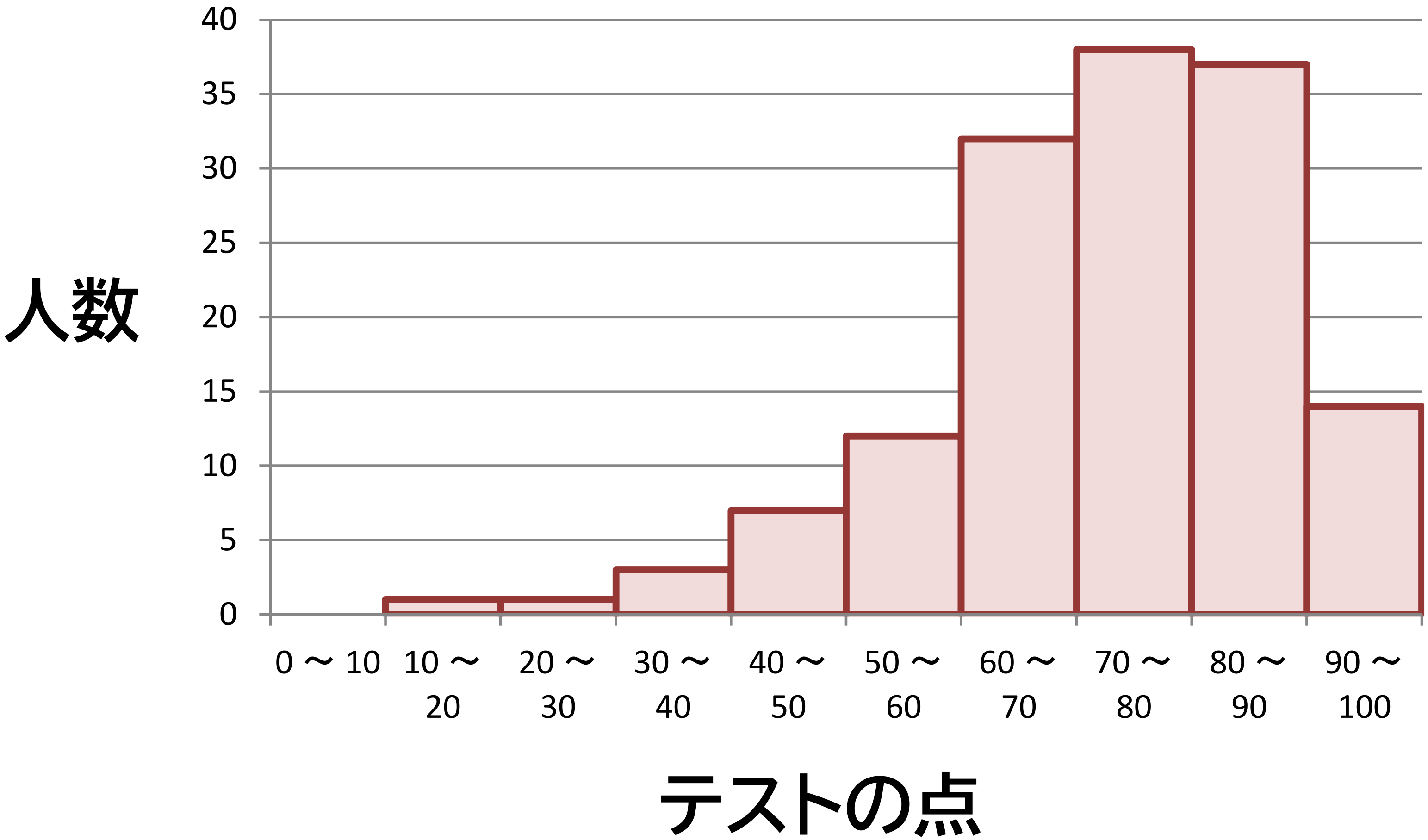
データを集める準備

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
- プログラミングに使う

【注意】授業が終わったら必ずログアウト

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. それぞれの区間に入った数を数える

目的に応じて、細かなアレンジは可

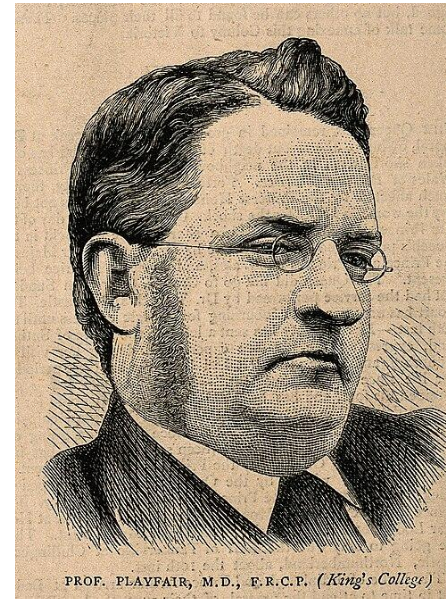
様々な見える化手法

- 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

近代的なグラフの発明

ウィリアム・スモールト・プレイフェア（1759-1823）
William Smoult Palyfair

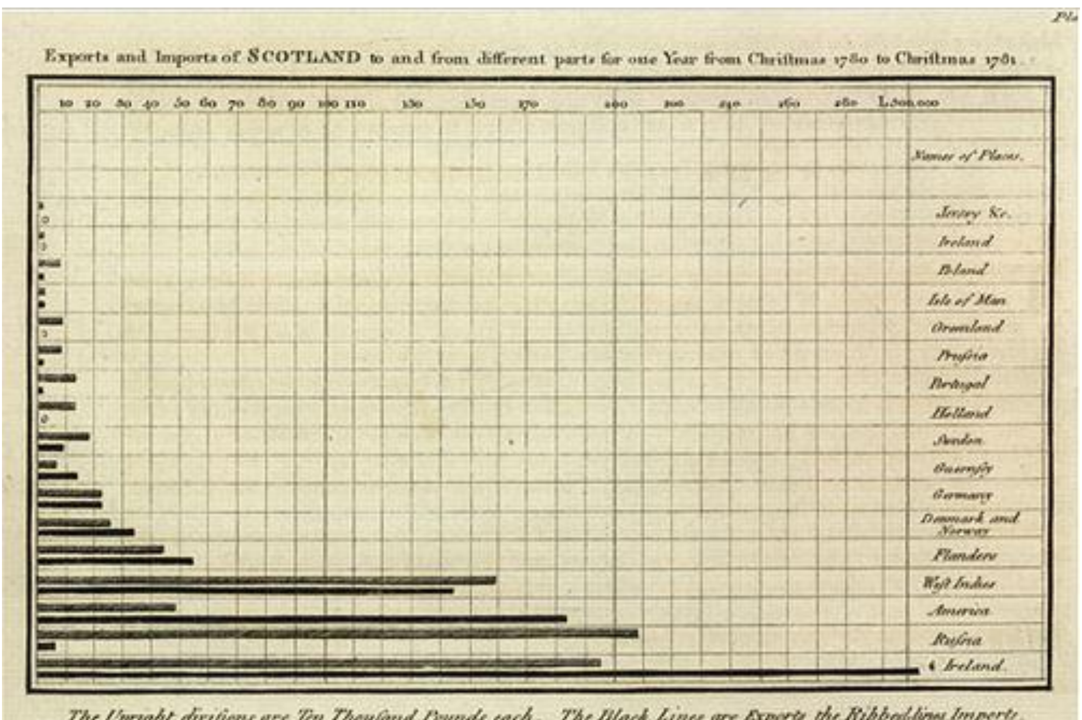
イギリスの政治経済学者



出典：wellcomecollection.org

- 棒グラフ
- 折れ線グラフ
- 円グラフ
- 面グラフ

などを発明。「統計グラフの父」と呼ばれる。



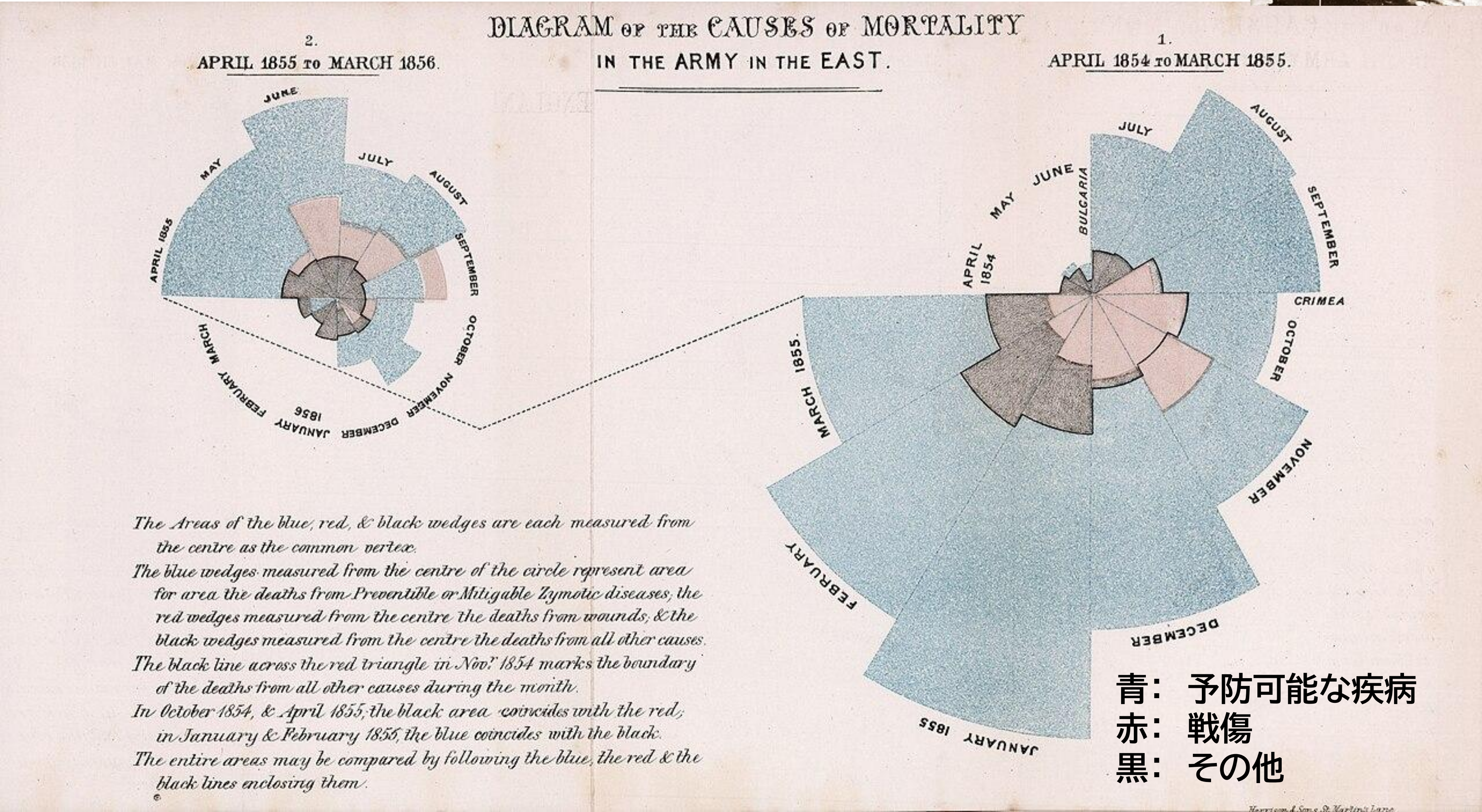
出典：wikipedia

スコットランドと17か国との間の輸出入の状況を可視化

ナイチンゲールが作成した鶏頭図

クリミア戦争での死因を視覚化し、衛生環境の改善の重要性を示した

出典: wikipedia
Florence Nightingale（1820-1910）イギリス



青: 予防可能な疾病
赤: 戦傷
黒: その他

厚生労働省　統計

で検索

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

2024年7月1日掲載	令和６年毎月勤労統計調査特別調査を実施します
2024年6月19日掲載	令和６年労使関係総合調査にご協力をお願いします
2024年6月19日掲載	令和６年賃金引上げ等の実態に関する調査にご協力をお願いします
2024年6月17日掲載	労働経済動向調査にご協力をお願いします　～令和6年８月調査～
2024年5月24日掲載	令和６年雇用動向調査(上半期)　ご協力のお願い

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
▶ 統計関連サイトリンク

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.社会保険
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：222KB］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.縦断調査（パネル調査）

1.1.人口

統計・調査名	統計・調査内容
人口動態調査 NEW 7月25日	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
人口動態職業・産業別統計	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
人口動態統計特殊報告	人口動態統計を基に、特定のテーマについてとりまとめたもの
生命表 NEW 7月25日	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（生命関数）によって表したもの

人口動態調査

お知らせ

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25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

オーダーメード集計・匿名データ提供
政府統計の統一ロゴタイプ

厚生労働省
Ministry of Health, Labour and Welfare

テーマ別に探す | 報道・広報 | 政策について

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 > 結果の概要

人口動態調査

結果の概要

◎人口動態調査の結果

- 確定数
- 速報
- 月報年計(概
- 報告書

◎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

●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

【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
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5年 令和4年 令和3年
令和2年 令和元年 平成30年 平成29年 平成28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

人口動態統計月報年計(概数)の概況

●月報(概数)の年間合計です。毎年、調査年の翌年6月上旬頃に公表しています。

令和6年 令和5年 令和4年 令和3年
令和2年 令和元年 平成30年 平成29年 平成28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

人口動態統計月報(概数)

●月毎の概数です。調査月の約5ヶ月後に公表しています。

令和7年	1月	2月 New 7月25日			
令和6年	1月	2月	3月	4月	5月
令和5年	1月	2月	3月	4月	5月
令和4年	1月	2月	3月	4月	5月
令和3年	1月	2月	3月	4月	5月
令和2年	1月	2月	3月	4月	5月
平成31年・令和元年	1月	2月	3月	4月	5月

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。調査年が国勢調査年次のときは翌々年10月に刊行し、国勢調査年次以外のときは翌々年3月に刊行しています。

令和5年 令和4年 令和3年 令和2年 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。

「我が国の人口動態」は、平成30年版をもって刊行を終了いたしました。一部のグラフは「人口動態統計(報告書)」に掲載されております。

平成30年我が国の人口動態(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。
●社会の状況が変わっていく中で、令和2年及び3年の数値に大きく増減があり、これまで用いてきた機械的な算出方法により算出した推計値は、実態と乖離する恐れがあることから、令和元年を最後に推計しないこととしました。
(年間推計の算出方法等の詳細は「令和3年」をご覧ください。)
●月間の数値については、速報を調査月の2か月後、月報(概数)を約5か月後に、年間の数値については、12月速報を調査年の翌年2月、月報年計(概数)を翌年6月、確定数を翌年9月に、従来どおり公表する予定としております(公表予定は こちら)。

令和3年 令和2年 令和元年 平成30年 平成29年
平成28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

ISSN 1345-5222

政府統計

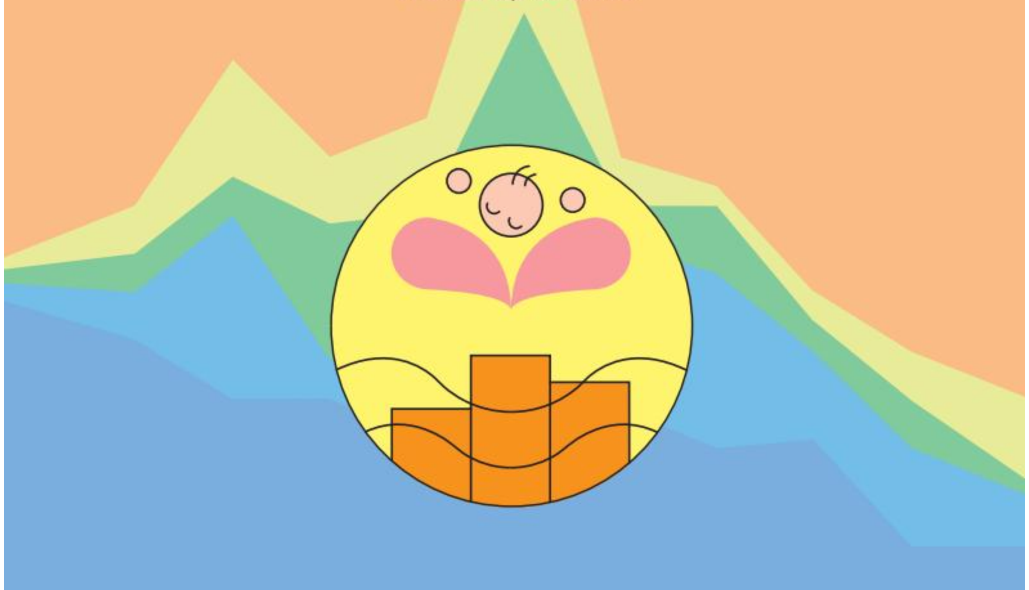
平成30年

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,

お探しのページが見つかりません（404 Not Found）。

申し訳ありません。お探しのページは見つかりませんでした。

- お探しのページが移動または削除されている場合があります。
- アドレス（URL）が間違っている可能性があります。目的のアドレス（URL）をお確かめの上、もう一度アクセスし直してみてください。

水道対策は、令和６年４月１日より、国土交通省及び環境省へ移管されました。

内容に応じて、国土交通省または環境省ホームページをご覧下さい。

- 国土交通省（水道整備・管理のうち水質・衛生以外）
- 環境省（水質基準の策定、水道整備・管理のうち水質・衛生）

お探しのページが見つからない場合は、厚生労働省ホームページ内を検索するか以下のリンクをご利用ください。

- 厚生労働省トップページはこちら
- サイトマップはこちら

厚生労働省ホームページ内の検索（お探しのページに含まれるキーワードなどで検索してください。）

Google カスタム検索 検索


amazon.co.jp
お届け先: [redacted]
場所を更新する
本
Amazon.co.jpを検索
JP
こんにちは, ログイン
アカウント＆リスト
返品もこちら
注文履歴
0
カート
すべて
Amazonポイント: 残高を確認
ヘルプ
ランキング
Amazon Basics
新着商品
プライム特典
対象の映画・TV番組が見放題
本
詳細検索
ジャンル一覧
新刊・予約
Amazonランキング
コミック・ラノベ・BL
雑誌
文庫・新書
洋書
Prime Reading
本のキャンペーン
本 › 社会・政治 › 社会学
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）
一般財団法人　厚生労働統計協会
平成30年　我が国の人口動態　平成28年までの動向 単行本 – 2018/3/29
厚生労働省政策統括官（統計・情報政策担当）（編集）
本書は、平成28年までの人口動態統計の主な内容、人口の動きや寿命について、グラフを中心に時系列観察、地域別観察、諸外国との比較を行っており、頁毎に簡単な解説と主な統計表を掲載するなど、人口動態の概要ができるだけ平易に分かるように構成されています。
この商品に関する問題を報告する
本の長さ
57ページ
言語
日本語
出版社
一般財団法人　厚生労働統計協会
すべての詳細を表示
最大12%ポイント還元:
紙書籍 真夏の涼読まとめ買いキャンペーン
ポイント還元率
2~4冊 最大2%
5~9冊 最大5%
10冊以上 最大12%
9/4(木)まで 今すぐエントリー
単行本
￥4,785 より
その他の新品 ￥4,785から
お届け先 [redacted] -お届け先の更新
すべての出品を見る
リストに追加

お探しのページが見つかりません（404 Not Found）。

申し訳ありません。お探しのページは見つかりませんでした。

- お探しのページが移動または削除されている場合があります。
- アドレス（URL）が間違っている可能性があります。目的のアドレス（URL）をお確かめの上、もう一度アクセスし直してみてください。

水道対策は、令和６年４月１日より、国土交通省及び環境省へ移管されました。

内容に応じて、国土交通省または環境省ホームページをご覧下さい。

- 国土交通省（水道整備・管理のうち水質・衛生以外）
- 環境省（水質基準の策定、水道整備・管理のうち水質・衛生）

お探しのページが見つからない場合は、厚生労働省ホームページ内を検索するか以下のリンクをご利用ください。

- 厚生労働省トップページはこちら
- サイトマップはこちら

厚生労働省ホームページ内の検索（お探しのページに含まれるキーワードなどで検索してください。）

Google カスタム検索　検索

過去に掲載されていた記事等について

過去に掲載されていた記事等は「国立国会図書館インターネット資料収集保存事業」Webサイトにて以前の内容を閲覧可能な場合があります。

参考:検索の手順

1. 国立国会図書館インターネット資料収集保存事業 のページにアクセスしてください
2. 詳細検索画面の「ページURL」欄に「**www.mhlw.go.jp/toukei/list/dl/81-1a2.pdf**」と入力してください


WARP Web Archiving Project
国立国会図書館
インターネット資料収集保存事業
言語(Language): 日本語
よくあるご質問
ヘルプ
サイトマップ
トップ > 詳細検索
詳細検索
全文
本文
平成30年我が国の人口動態
AND
ページURL
AND
ファイル種別(拡張子)
html pdf doc,docx xls,xlsx ppt,pptx その他
すべて選択/解除
インターネット閲覧
インターネットで閲覧できるものに絞る
PDFだけに
メタデータ
タイトル
編者
公開者(出版者)
起点URL
ISSN/ISBN
書誌ID
項目間を AND で検索
範囲指定
保存日
年 月 ～ 年 月
コレクション
国の機関 自治体(都道府県 政令指定都市 市町村 市町村合併 特別地方公共団体)
法人・機構 大学・学術機関 政党 イベント・スポーツ 電子雑誌 その他
すべて選択/解除
検 索
クリア
国の機関だけに

検索結果

【検索条件】
(コレクション種別 = 国の機関) AND (本文、ページタイトル = 平成30年我が国の人口動態) AND (ファイル種別(拡張子) = pdf)

絞込／再検索

メタデータの検索結果 0件 | 全文の検索結果 19件

1

適合度順 | 降順 | 20

000636869.pdf / 000636869.pdf [PDF] [保存日:2020/06/14 - 2025/03/11]
厚生労働省
www.mhlw.go.jp/content/11920000/000636869.pdf
国民健康・栄養調査 厚生労働省:平成27年国民健康・栄養調査 図3 エネルギー摂取量の年次推移(20歳以上女性) 図4 BMI実測値 図5 BMI理想値 図2 厚生労働省 平成30年我が国の人口動態より

Microsoft PowerPoint - 20230424E ×ì³ó˚©1D˜nê .å“F Rev.3 - ³Ôü (1) / %E5%88%86%E6%8B%85%E7%A0%94%E7%A9%B6%E5%A0%B1%E5%91%8A%E6%9B%B8%EF%BC%9A20DA1006-buntan4-4%E3%80%80P63-85_0.pdf [PDF] [保存日:2024/03/01 - 2025/03/02]
厚生労働省
mhlw-grants.niph.go.jp/system/files/report_pdf/%E5%88%86%E6%8B%85%E7%A0%94%E7%A9%B6%E5%A0%B1%E5%91%8A%E6%9B%B8%EF%BC%9A20DA1006-buntan4-4%E3%80%80P63-85_0.pdf
出産が増加している 厚生労働省 平成30年 我が国の人口動態 20歳代30歳代 64 2023/5/243 ← やせ → ← 普通体重 → ← 肥満 → 回答数200人 平均20.3

我が国の人口動態平成30年-本省用.indb / 81-1a2.pdf [PDF] [保存日:2018/05/02 - 2024/09/04]
厚生労働省
www.mhlw.go.jp/toukei/list/dl/81-1a2.pdf

我が国の人口動態平成30年-本省用.indb / 81-1a2en.pdf [PDF] [保存日:2018/05/01 - 2019/05/06]
厚生労働省
www.mhlw.go.jp/english/database/db-hw/dl/81-1a2en.pdf

絞込／再検索

1

適合度順 | 降順 | 20

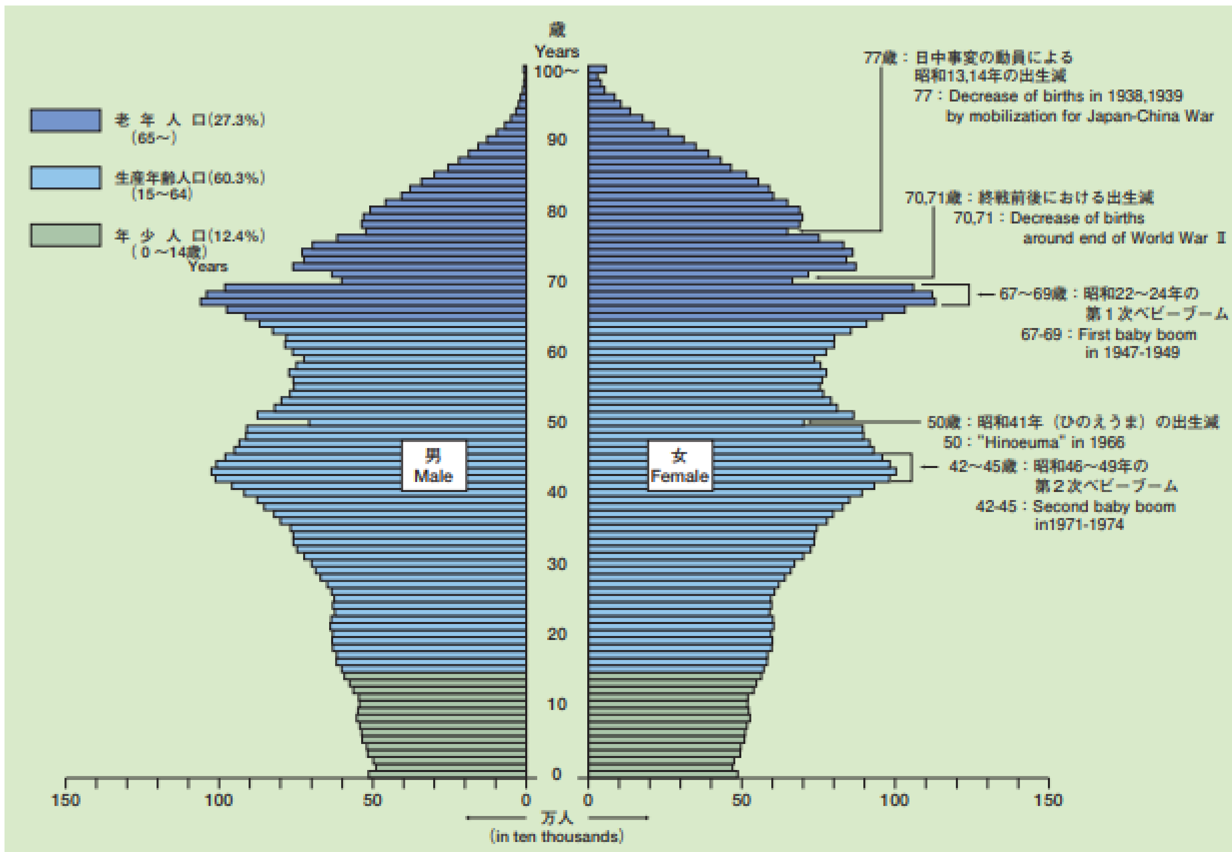
授業のページから
ダウンロード可
（要PW）

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%

我が国の人口ピラミッドー平成 28 年 10 月 1 日現在ー
Population pyramid as of Oct.1, 2016



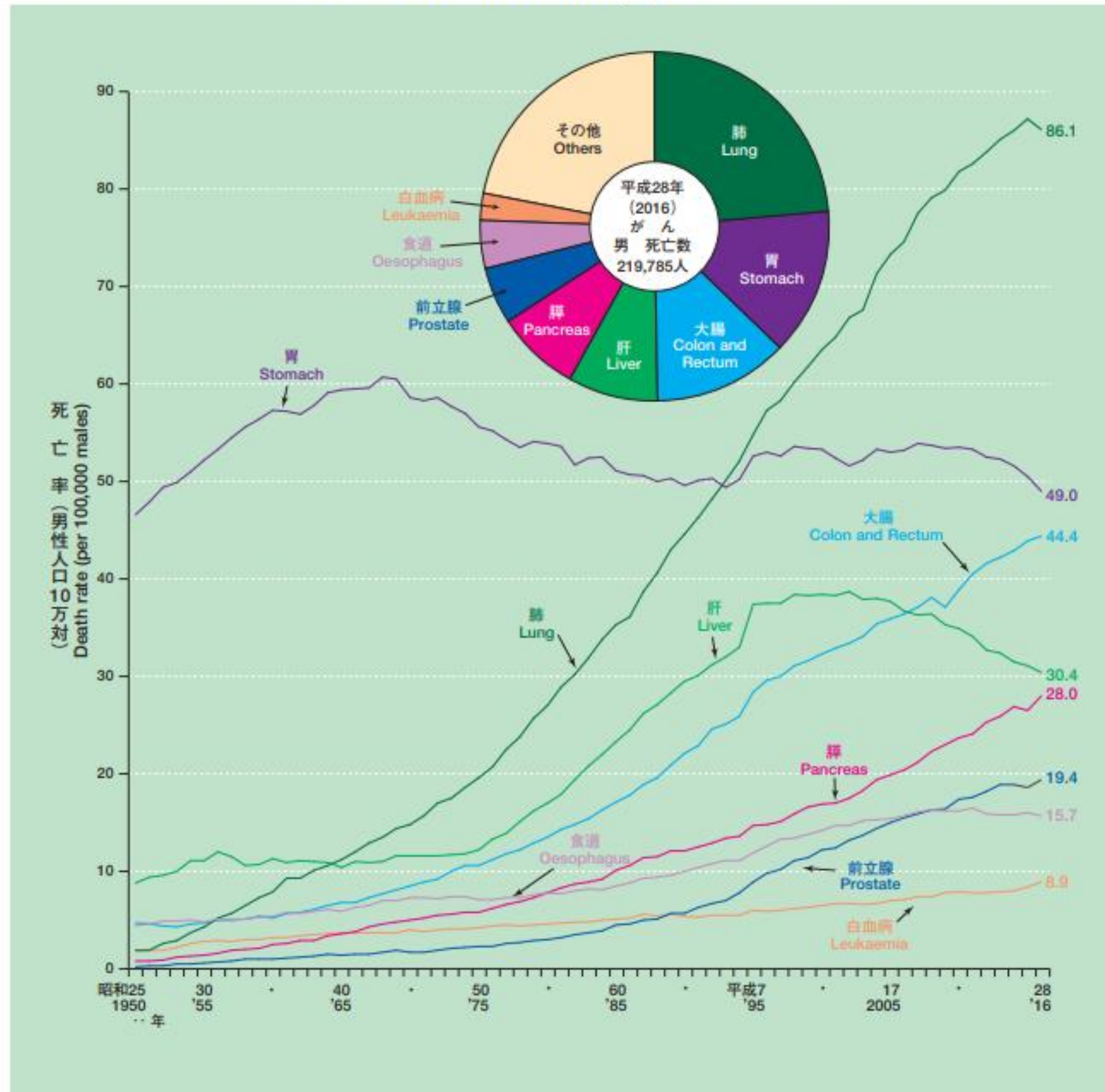
資料：総務省統計局 「人口推計（平成28年10月1日現在）」(総人口)

p.6

- 折れ線グラフ
- 円グラフ

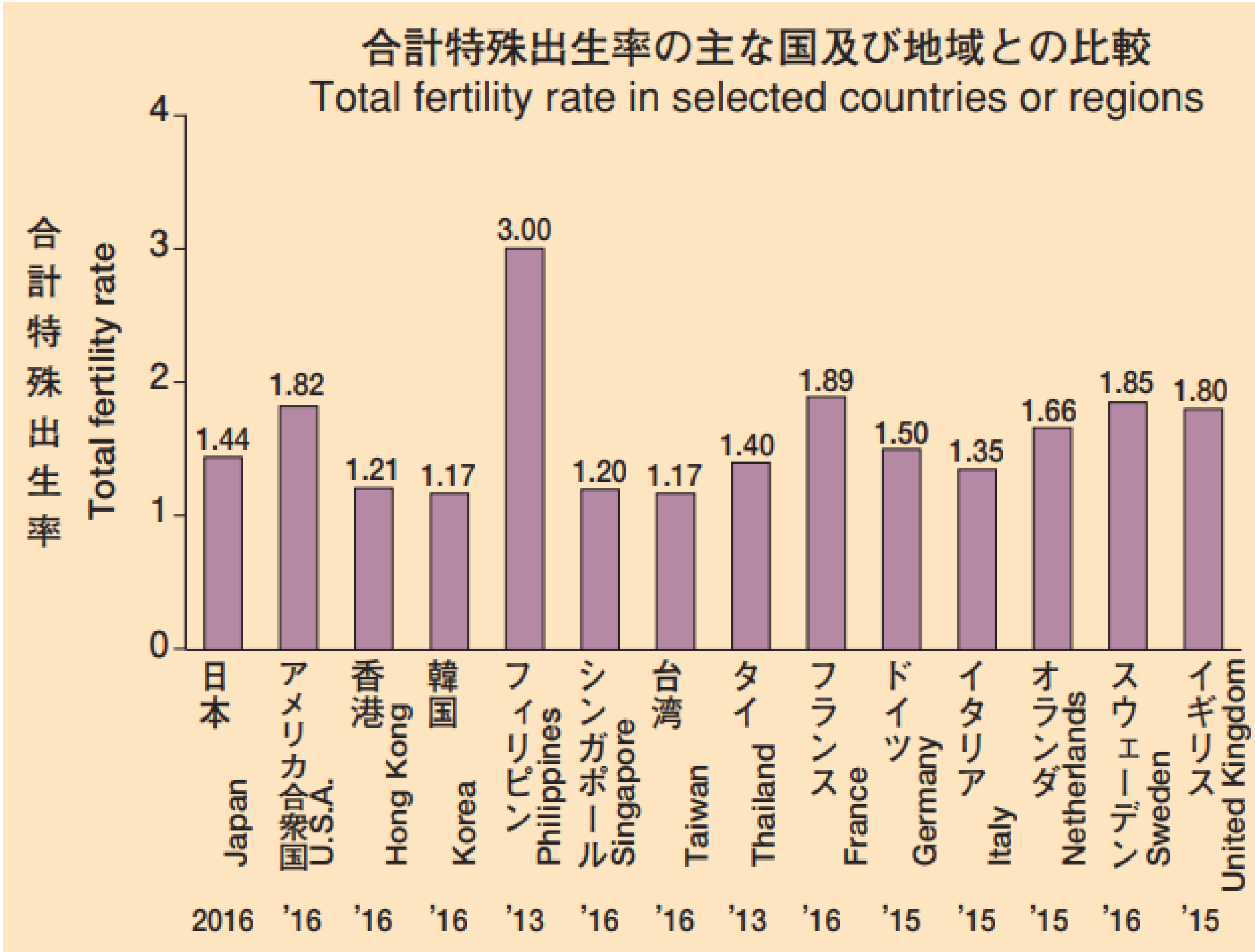
部位別にみたがんの死亡率の年次推移，男ー昭和25～平成28年ー

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

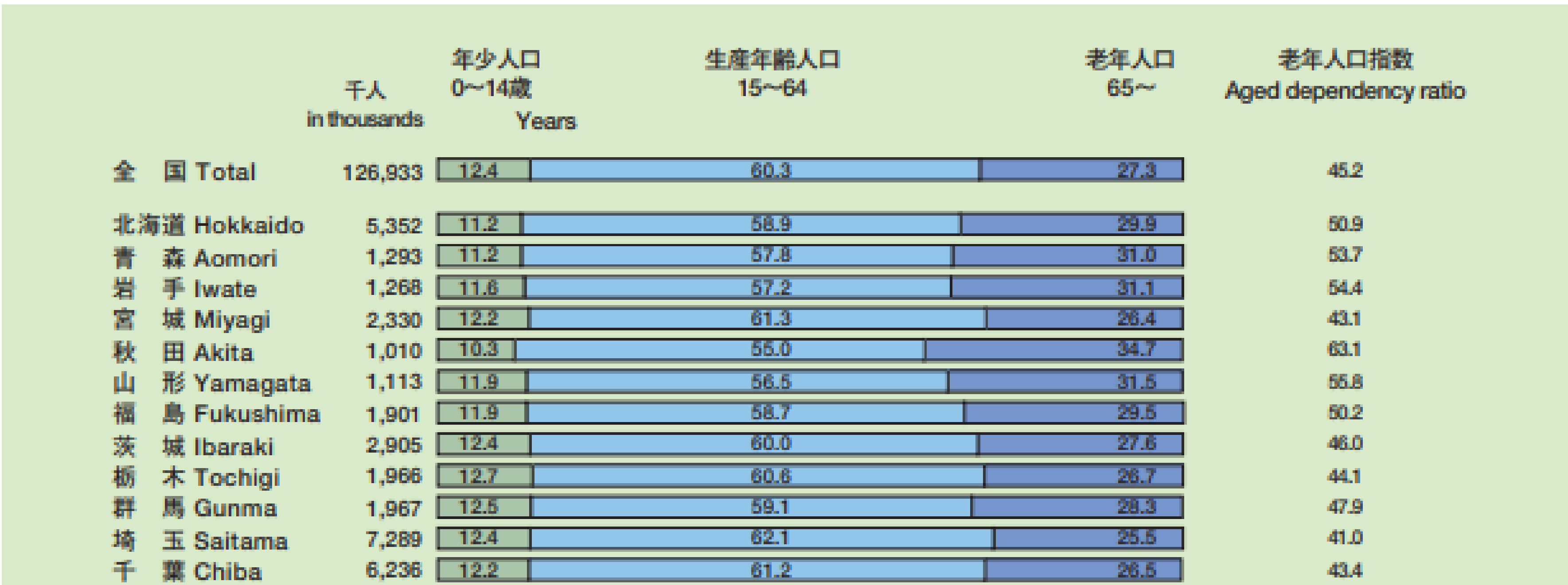

合計特殊出生率の主な国及び地域との比較
Total fertility rate in selected countries or regions
合計特殊出生率
Total fertility rate
0
1
2
3
4
1.44
1.82
1.21
1.17
3.00
1.20
1.17
1.40
1.89
1.50
1.35
1.66
1.85
1.80
日本 Japan 2016
アメリカ合衆国 U.S.A. '16
香港 Hong Kong '16
韓国 Korea '16
フィリピン Philippines '13
シンガポール Singapore '16
台湾 Taiwan '16
タイ Thailand '13
フランス France '16
ドイツ Germany '15
イタリア Italy '15
オランダ Netherlands '15
スウェーデン Sweden '16
イギリス United Kingdom '15

p.14

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年人口指数－平成28年－

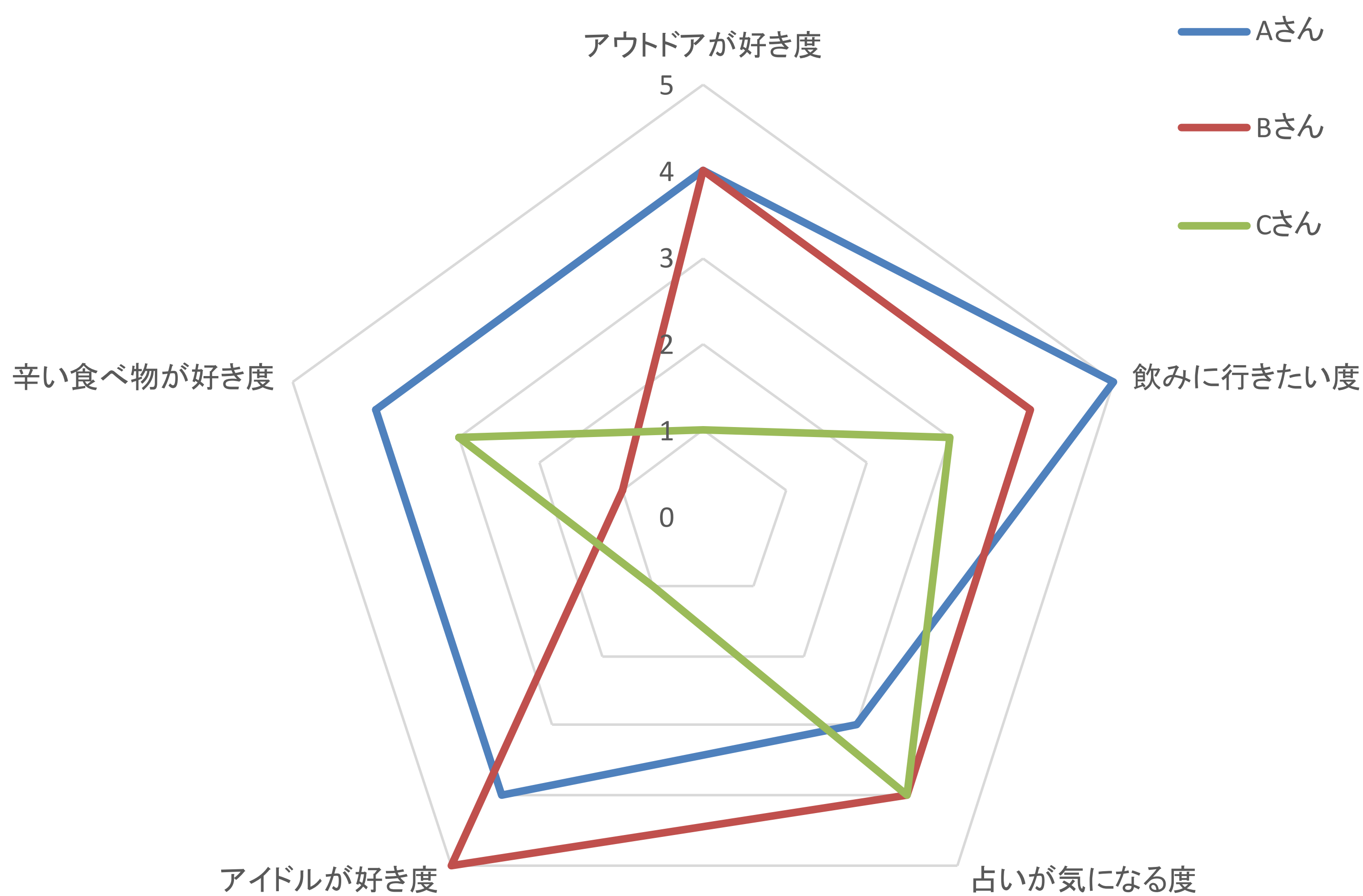
Percent distribution of population by 3 age groups and aged dependency ratio,by prefecture，2016



p.7

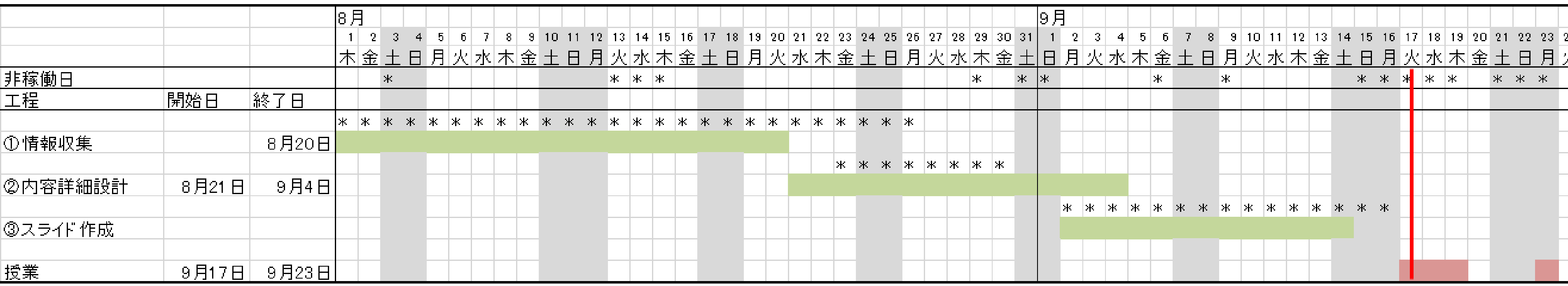
最新の情報を入手するには…

その他：レーダーチャート

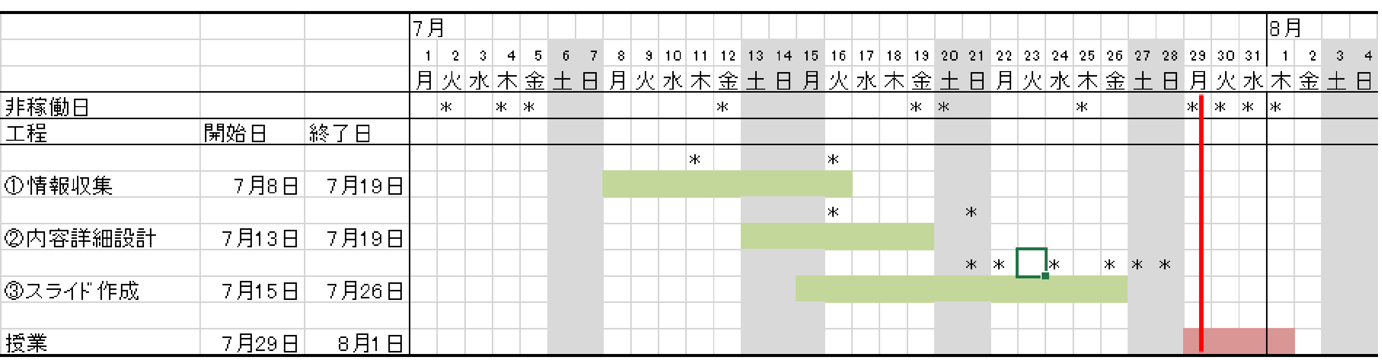


その他：ガントチャート

2019年



2024年



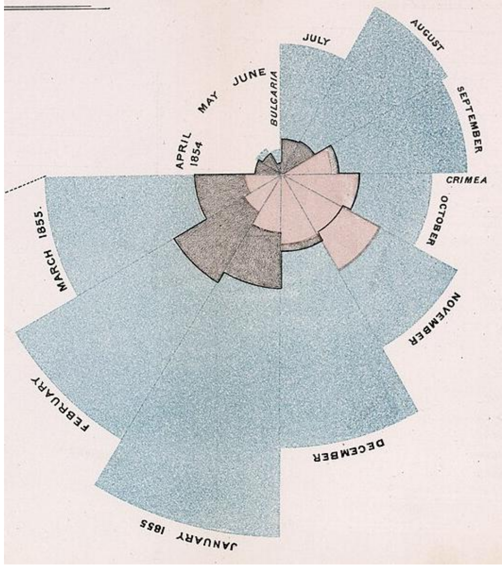
計画

＊ 実際

可視化のまとめ

- いろいろなグラフ(可視化方法)がある

重要となるスキル



- 正しく使う
- 正しく理解する

超重要!! だまされない だまさない
故意でなくても

- 必要に応じて新しいグラフを考える

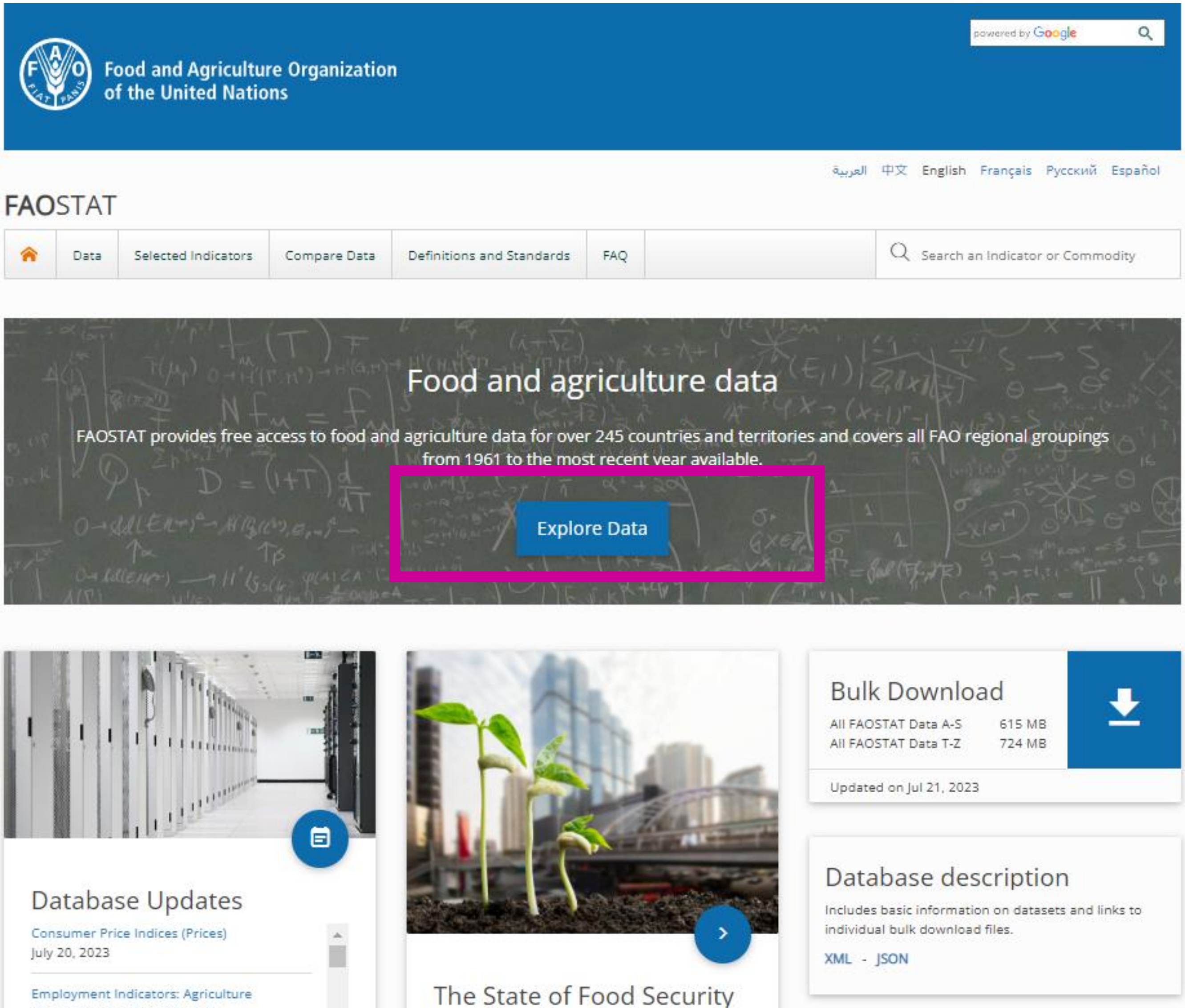
色々な
統計サイト

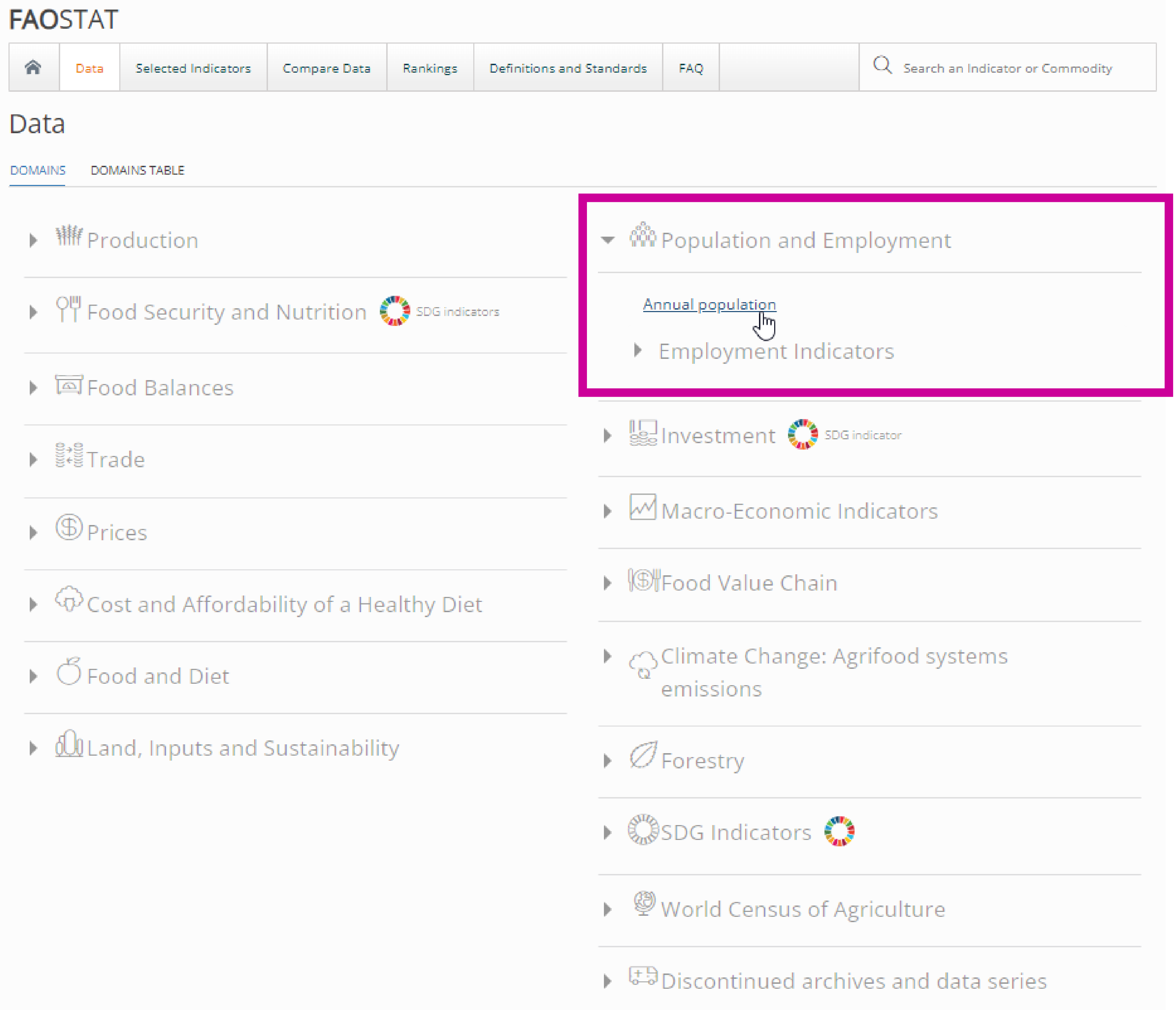
都道府県別の統計

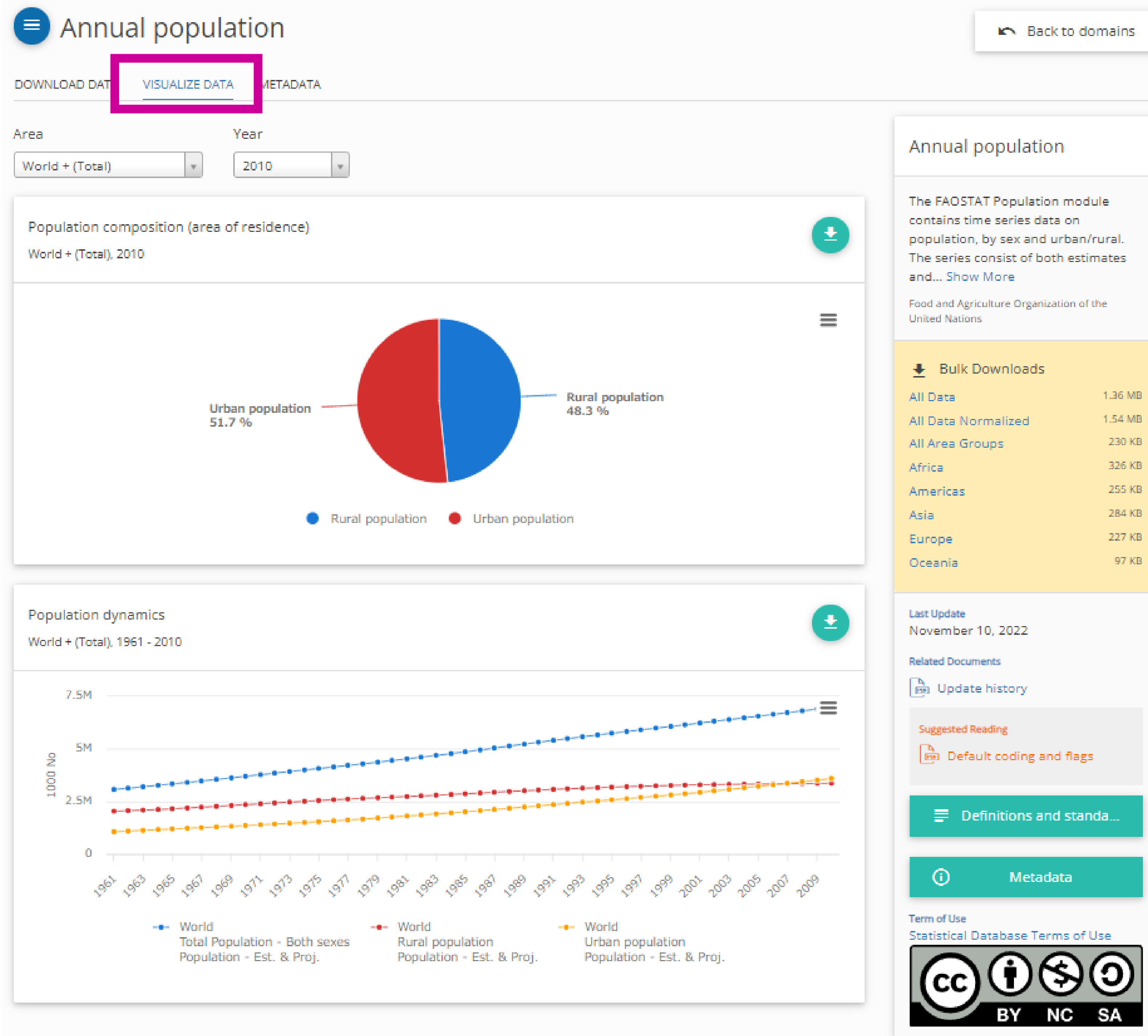
https://todo-ran.com/

アメリカ農務省の統計

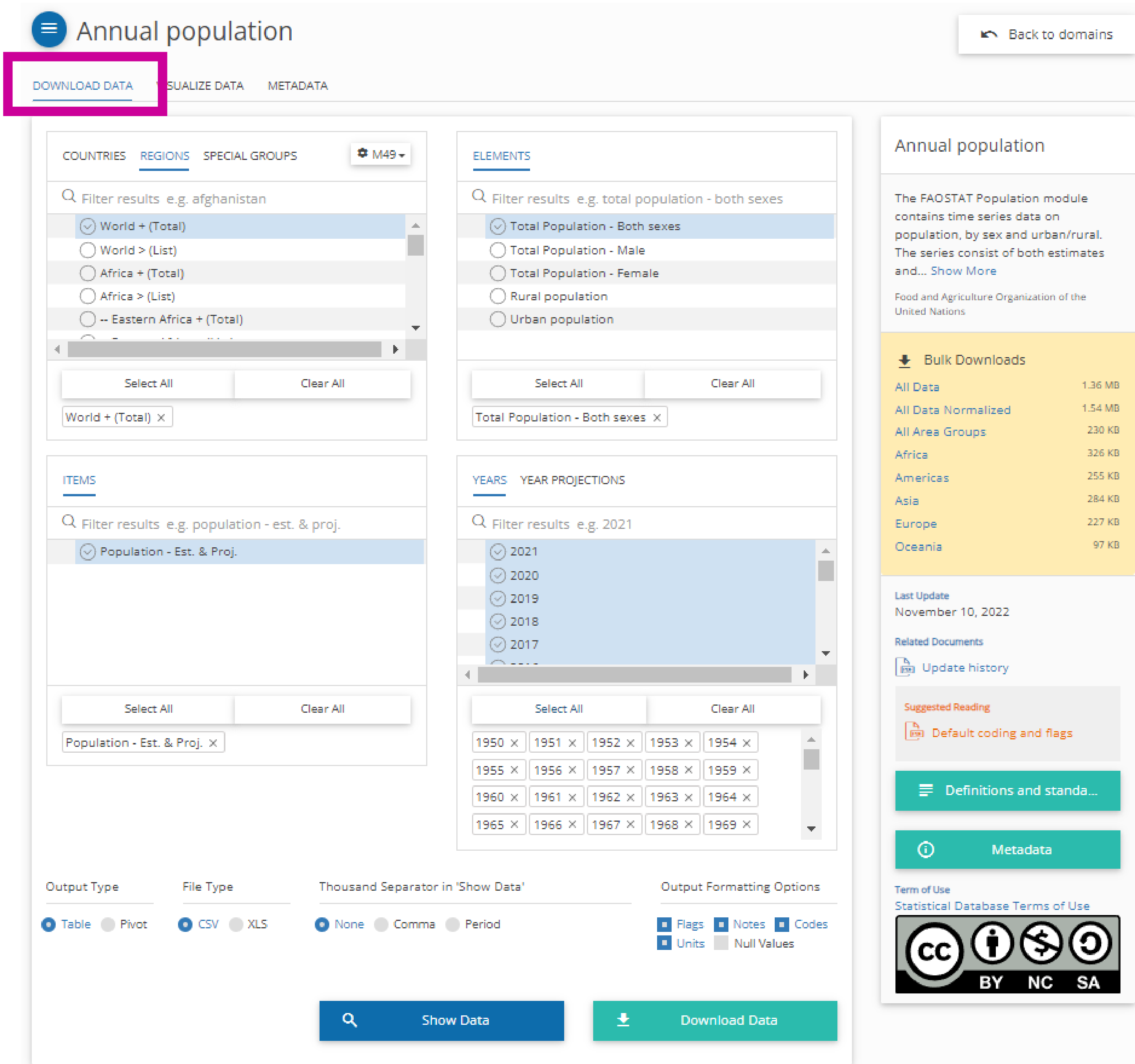
http://www.fao.org/faostat/en/




FAOSTAT
Data
Selected Indicators
Compare Data
Rankings
Definitions and Standards
FAQ
Search an Indicator or Commodity
Data
DOMAINS
DOMAINS TABLE
Production
Food Security and Nutrition
SDG indicators
Food Balances
Trade
Prices
Cost and Affordability of a Healthy Diet
Food and Diet
Land, Inputs and Sustainability
Population and Employment
Annual population
Employment Indicators
Investment
SDG indicator
Macro-Economic Indicators
Food Value Chain
Climate Change: Agrifood systems emissions
Forestry
SDG Indicators
World Census of Agriculture
Discontinued archives and data series

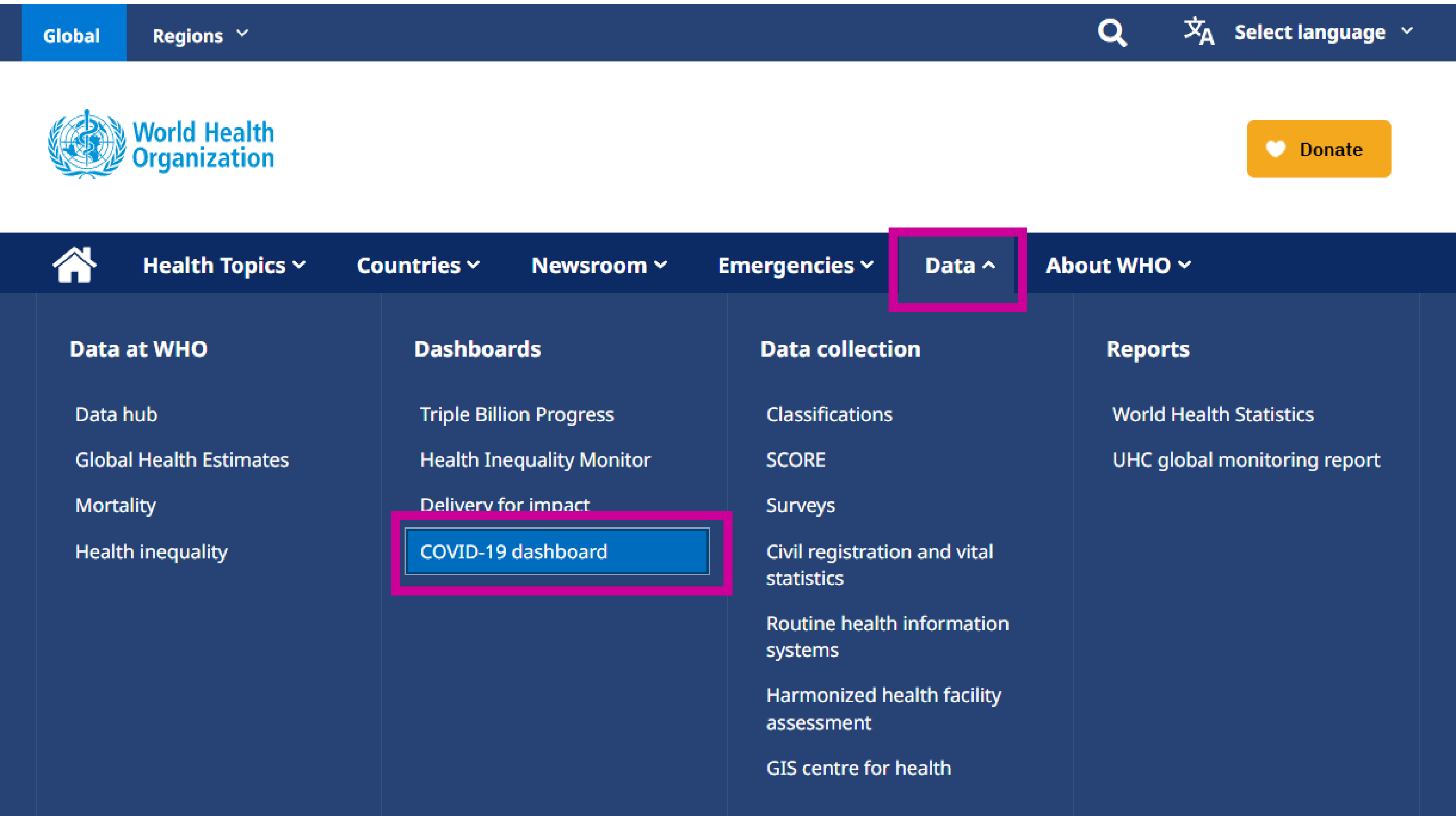


Annual population
Back to domains
DOWNLOAD DATA
VISUALIZE DATA
METADATA
Area
World + (Total)
Year
2010
Population composition (area of residence)
World + (Total), 2010
Urban population 51.7 %
Rural population 48.3 %
Rural population
Urban population
Population dynamics
World + (Total), 1961 - 2010
7.5M
5M
2.5M
0
1000 No
1961 1963 1965 1967 1969 1971 1973 1975 1977 1979 1981 1983 1985 1987 1989 1991 1993 1995 1997 1999 2001 2003 2005 2007 2009
World Total Population - Both sexes Population - Est. & Proj.
World Rural population Population - Est. & Proj.
World Urban population Population - Est. & Proj.
Annual population
The FAOSTAT Population module contains time series data on population, by sex and urban/rural. The series consist of both estimates and... Show More
Food and Agriculture Organization of the United Nations
Bulk Downloads
All Data 1.36 MB
All Data Normalized 1.54 MB
All Area Groups 230 KB
Africa 326 KB
Americas 255 KB
Asia 284 KB
Europe 227 KB
Oceania 97 KB
Last Update
November 10, 2022
Related Documents
Update history
Suggested Reading
Default coding and flags
Definitions and standa...
Metadata
Term of Use
Statistical Database Terms of Use
cc BY NC SA


Annual population
Back to domains
DOWNLOAD DATA
VISUALIZE DATA
METADATA
COUNTRIES
REGIONS
SPECIAL GROUPS
M49
Filter results e.g. afghanistan
World + (Total)
World > (List)
Africa + (Total)
Africa > (List)
-- Eastern Africa + (Total)
Select All
Clear All
World + (Total)
ELEMENTS
Filter results e.g. total population - both sexes
Total Population - Both sexes
Total Population - Male
Total Population - Female
Rural population
Urban population
Select All
Clear All
Total Population - Both sexes
ITEMS
Filter results e.g. population - est. & proj.
Population - Est. & Proj.
Select All
Clear All
Population - Est. & Proj.
YEARS
YEAR PROJECTIONS
Filter results e.g. 2021
2021
2020
2019
2018
2017
Select All
Clear All
1950 1951 1952 1953 1954
1955 1956 1957 1958 1959
1960 1961 1962 1963 1964
1965 1966 1967 1968 1969
Output Type
Table
Pivot
File Type
CSV
XLS
Thousand Separator in 'Show Data'
None
Comma
Period
Output Formatting Options
Flags
Notes
Codes
Units
Null Values
Show Data
Download Data
Annual population
The FAOSTAT Population module contains time series data on population, by sex and urban/rural. The series consist of both estimates and... Show More
Food and Agriculture Organization of the United Nations
Bulk Downloads
All Data 1.36 MB
All Data Normalized 1.54 MB
All Area Groups 230 KB
Africa 326 KB
Americas 255 KB
Asia 284 KB
Europe 227 KB
Oceania 97 KB
Last Update
November 10, 2022
Related Documents
Update history
Suggested Reading
Default coding and flags
Definitions and standa...
Metadata
Term of Use
Statistical Database Terms of Use
BY NC SA

WHOの統計

https://www.who.int/



World Health Organization | Data

WHO Global Language Show navigation

WHO COVID-19 dashboard

WHO Health Emergencies Programme

World

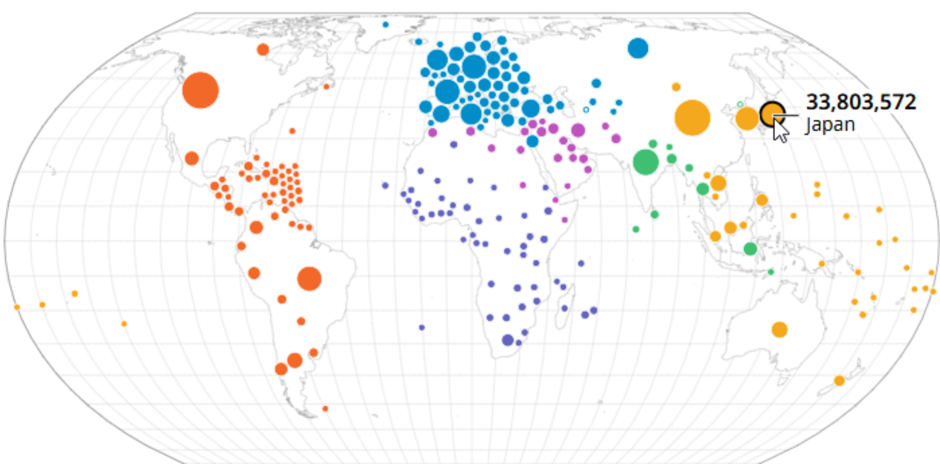
Summary Circulation **Cases** Deaths More

Last 7 days | Last 28 days | **Total cumulative**

Count | Rate per 100 000

Number of COVID-19 cases reported to WHO (cumulative total)

World



WHO Regions

Africa Americas Eastern Mediterranean Europe South-East Asia Western Pacific

778,417,964 +8,399 increase on previous 7 days

Reported COVID-19 cases

World, 7 days to 13 July 2025

Number of COVID-19 cases reported to WHO (cumulative total)

World

Country	Cases
World	**778m**
United States of America	103m
China	99.4m

Show 229 more

Pitcairn	4

こんな統計はどこかにないかな・・・？

一日の中でスマホを操作している時間に関する統計情報はありますか？

はい、**日本におけるスマートフォンの1日あたりの使用時間**について、信頼性のある統計情報をもとにご紹介します。

📱 日本国内の平均スマホ使用時間

● メディア総接触時間調査（Impress Web担当者フォーラム）

- 2025年の週平均（1日あたり）で、**携帯電話・スマートフォンへの接触時間は過去最高の約165.1分（約2時間45分）**でした。 Web担当者Forum +8

これは、テレビや雑誌などを含むメディア全体の消費時間中の数値の一部として示されています。

👥 利用者意識と実態のギャップ

● Mikkeプロジェクト調査（2021年）

- 日本人の53.6%が「1日3時間以上スマホを使っている」と回答していますが、彼ら自身の意識と実際の利用時間には

統計情報ソースまとめ

- 統計情報は色々なところにある
- 情報を探すのにAIは役立つ

重要となるスキル

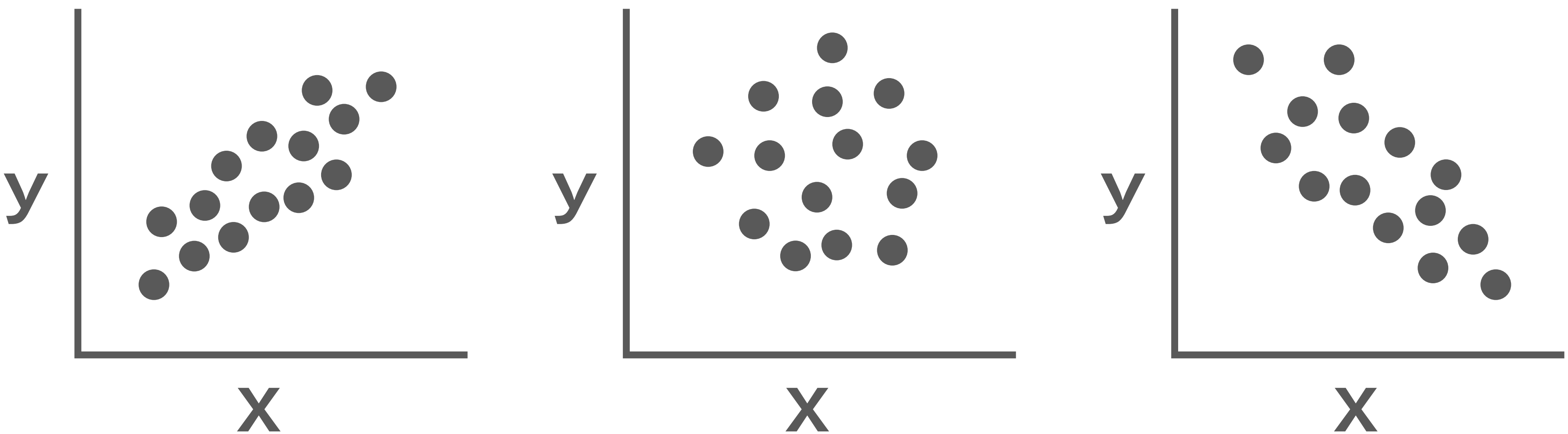
- 「これを知りたい」という発想力
- ファクトチェック（情報をうのみにしない）

相関

- ✓相関の値、計算方法を理解する
- ✓回帰曲線とは異なる
- ✓相関関係と因果関係は異なる

散布図

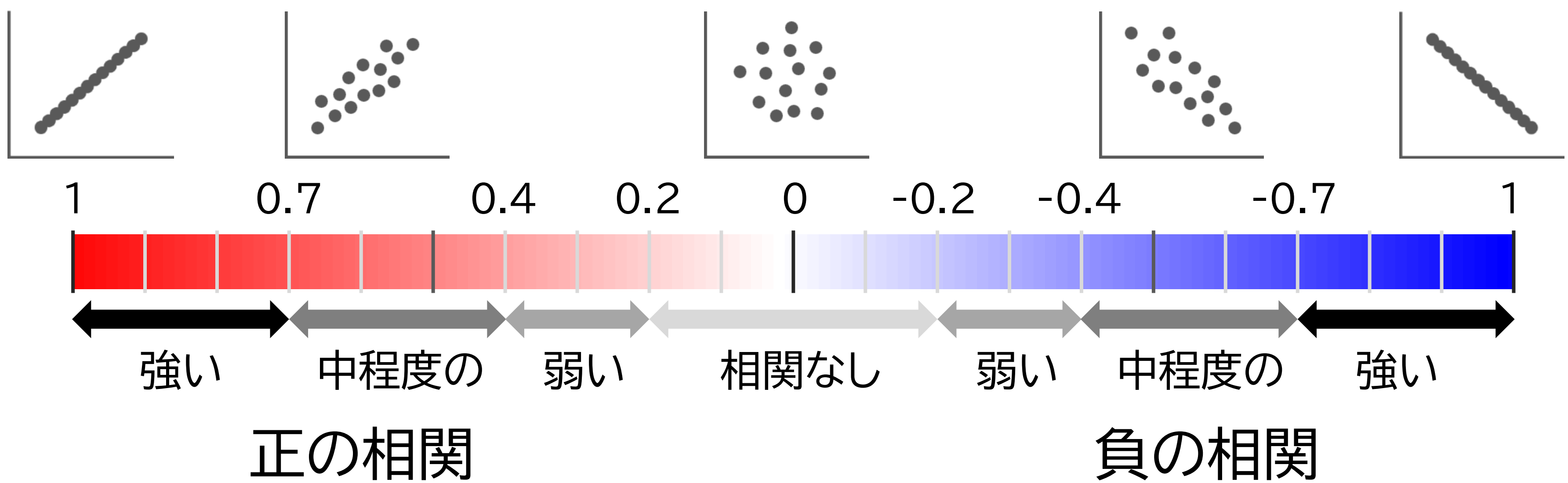
二つの変数の間の関係性を見える化する手法



正の相関がある　　相関がない　　負の相関がある

相関係数

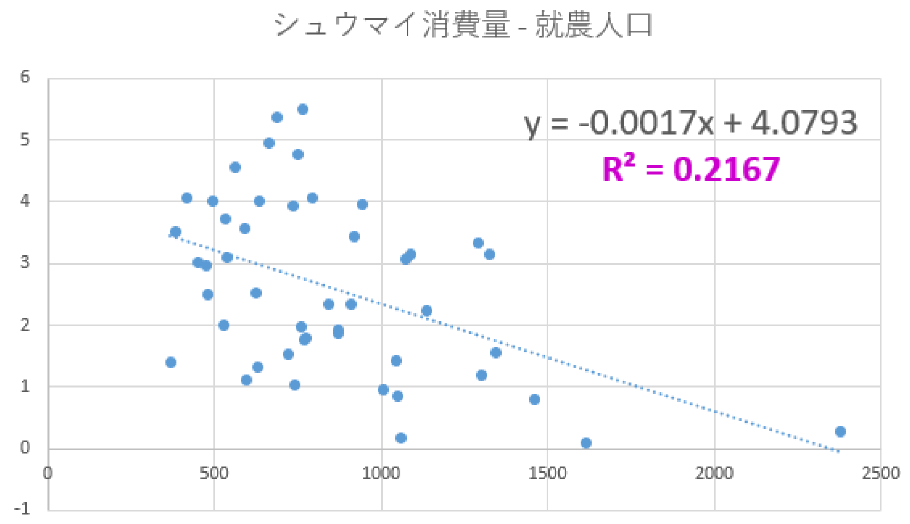
- 二つの変数の間の関係性の強さを**数値化**したもの
- 1～-1の間の値をとる



※数字の区切りはあくまで目安

- Excelでは**PEARSON関数**で計算できる

注意点



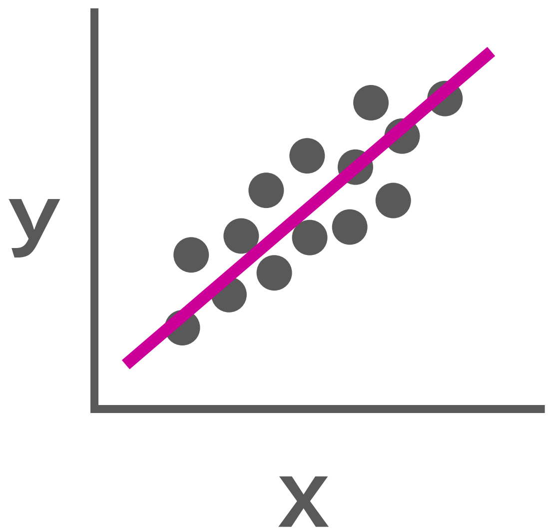
回帰曲線のR^2値は、相関係数ではない！

R^2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、R^2値はピアソン相関係数の二乗に当たります。相関係数が-1～1の値を取るのに対し、R^2値は0～1の値を取ります。負の相関であっても、R^2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。

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

相関関係を
見てみる

都道府県別の統計

https://todo-ran.com/

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

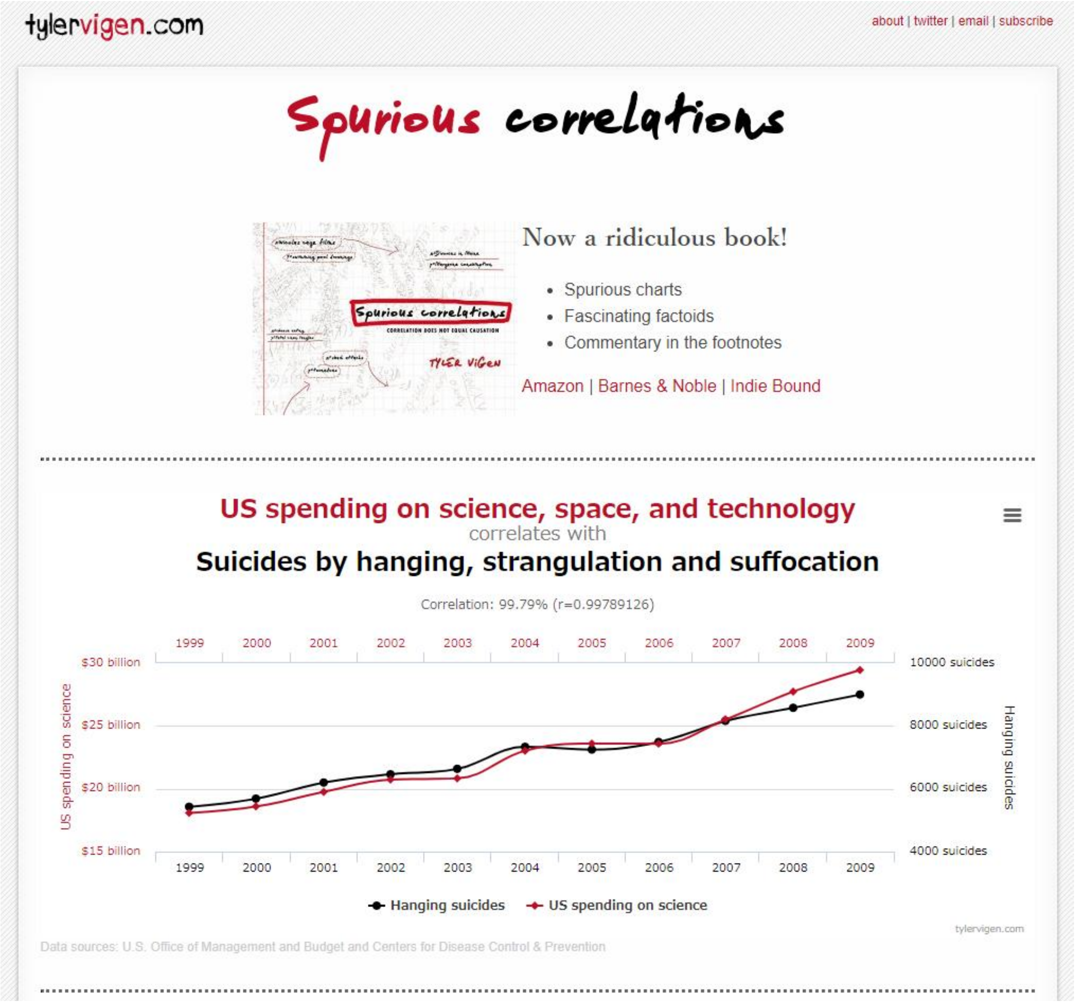
相関と因果

相関関係:
二つの事柄に関連性がある

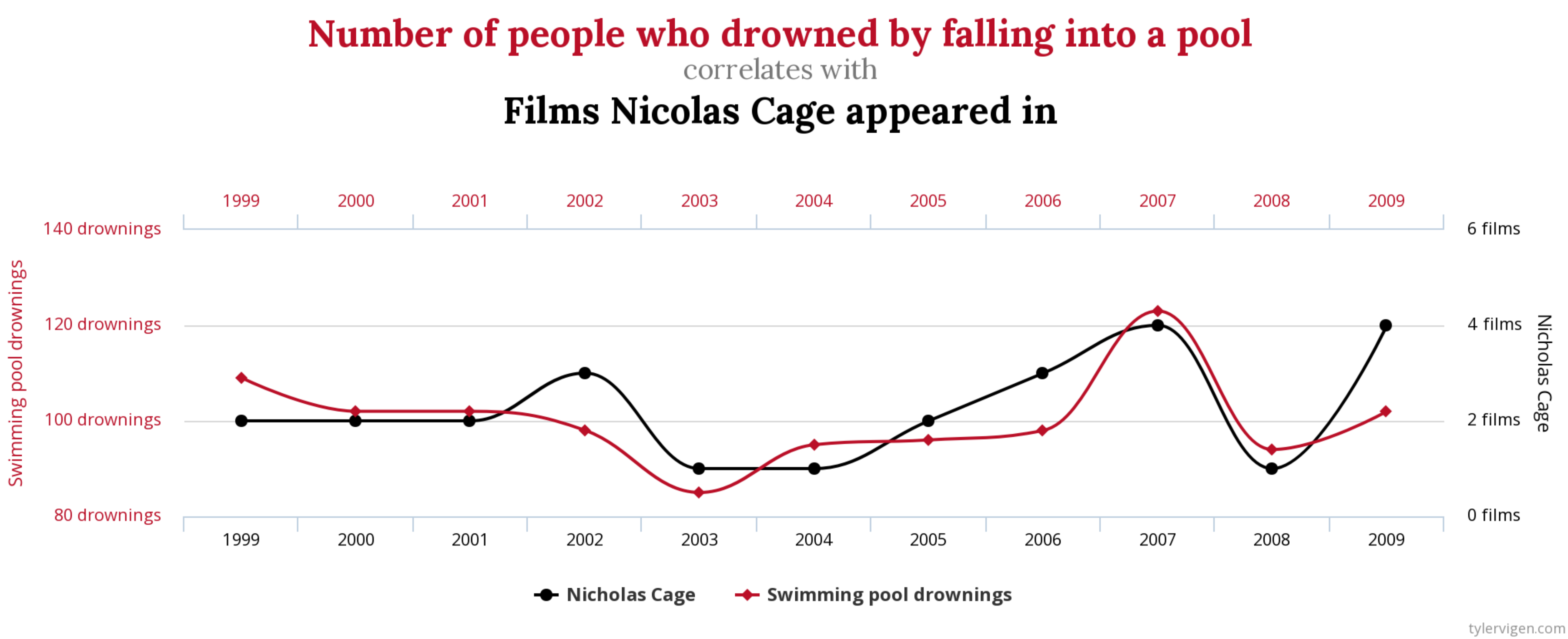
因果関係:
二つの事柄が、原因と結果の関係
である

疑似相関

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？



中室牧子, 津川友介著、ダイヤモンド社2017年

情報統計
第2回

2025年8月4日　神奈川工科大学

櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

統計の基本

学習目標

平均値の推定方法を理解する

ポイント

中心極限定理と分布

学習目標2

以下の統計用語を理解する

- 平均値、中央値
- 分散、標準偏差
- 統計量
- 分布

- 統計的推定
- 母集団
- ランダムサンプリング
- 標本

- 母平均、母分散
- 標本平均、標本分散、不偏標本分散
- 中心極限定理
- 正規分布（ガウス分布）
- 標準誤差

復習

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

復習

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！

目的:
このクラスの人の身長は
どのくらい？

集めたデータ

↓

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

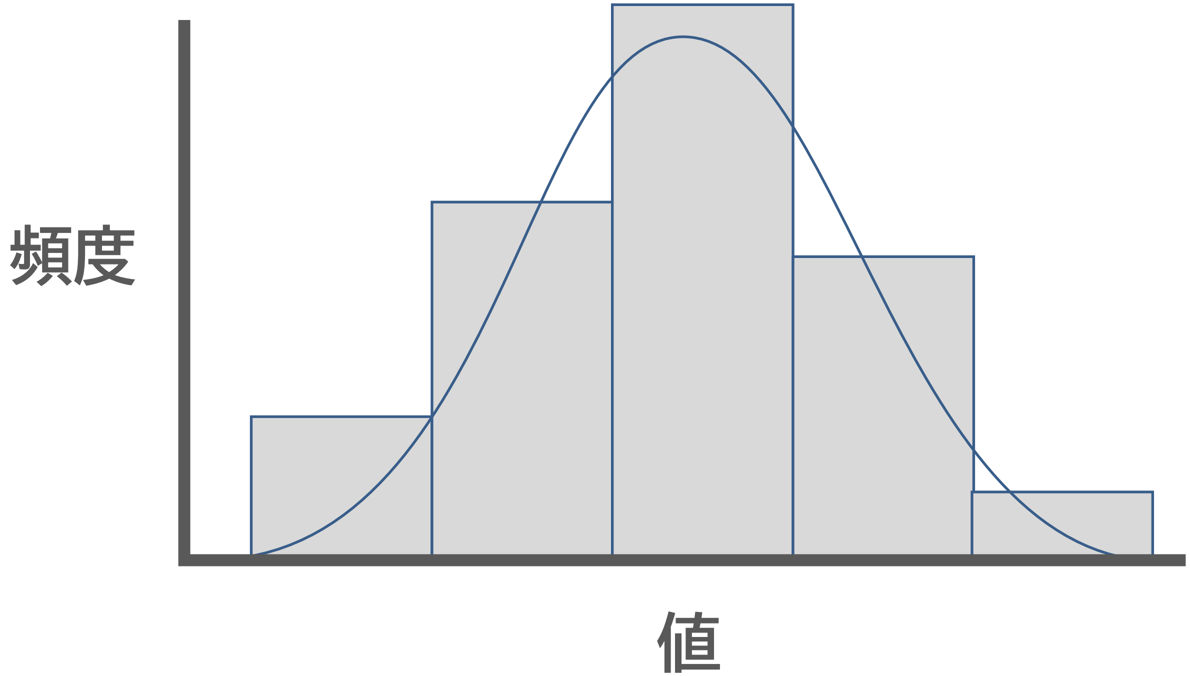
分散
標準偏差

ばらつきを表す値

統計量（基本統計量、基礎統計量とも）

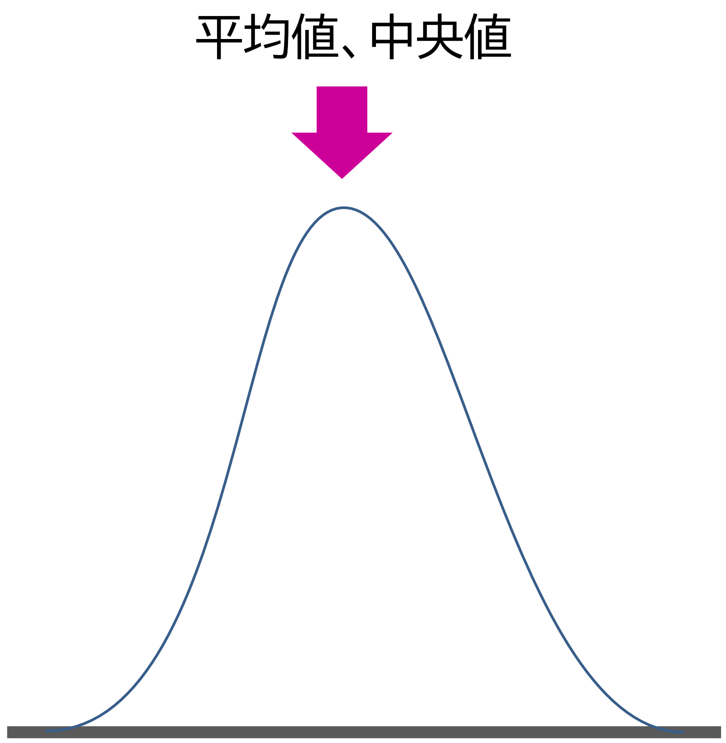
分布

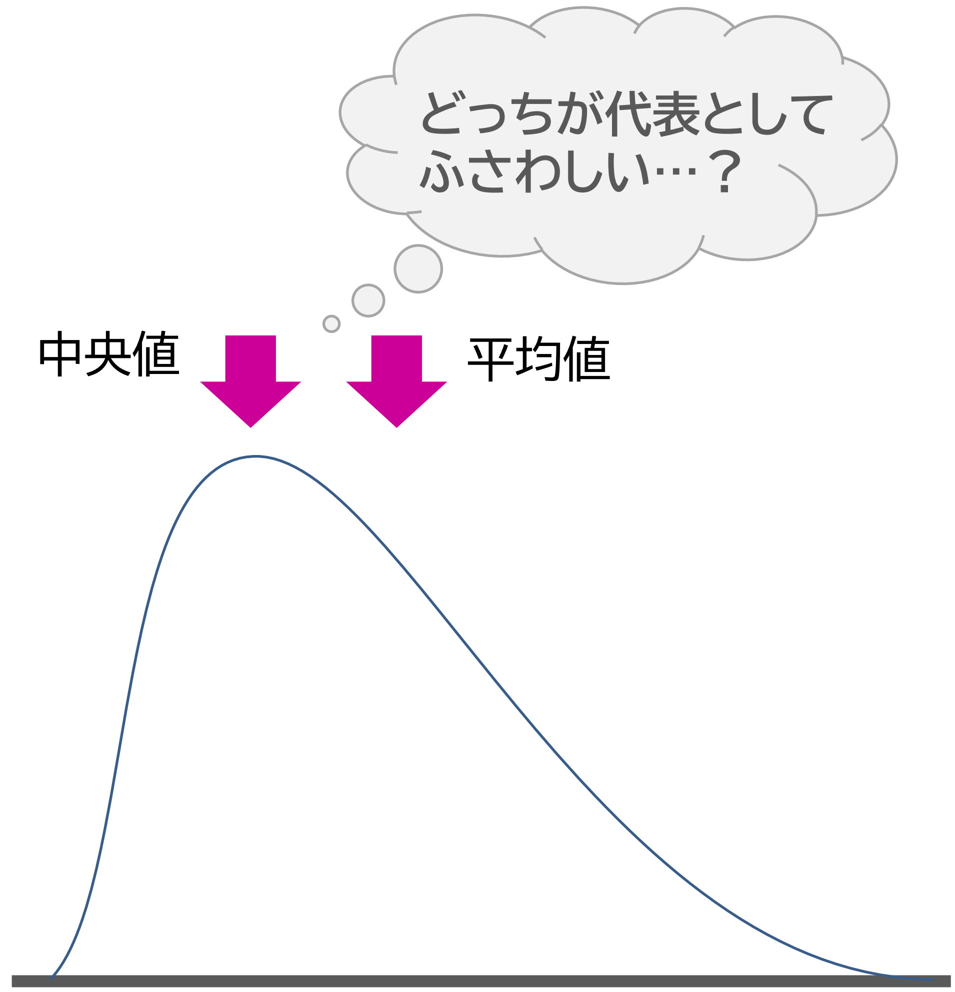
データの全体的な偏り具合を表したもの



ヒストグラム（頻度分布図）

データの中心





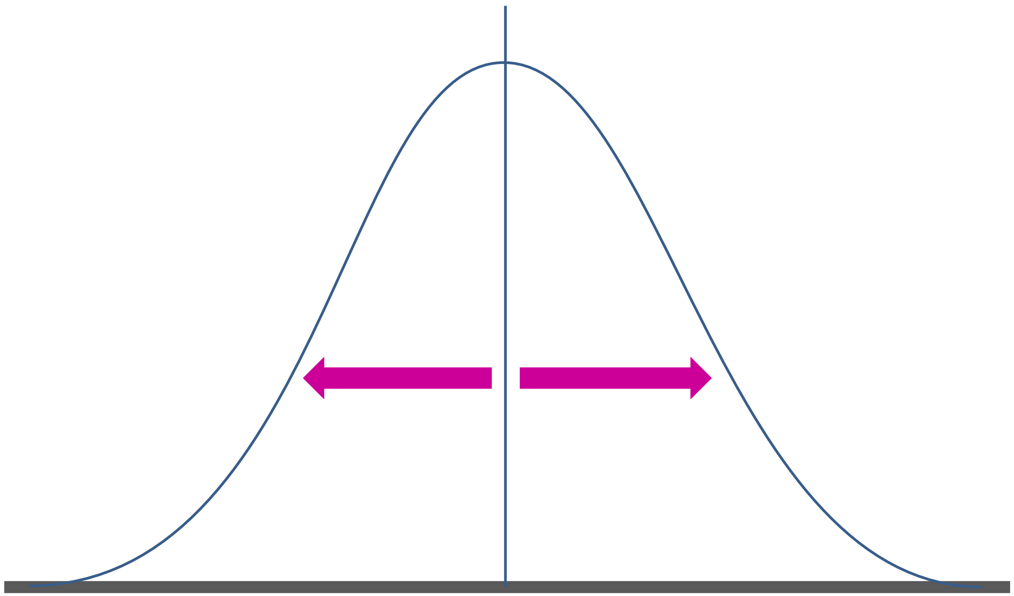
偏りのないデータ

- 身長の分布など

偏っているデータ

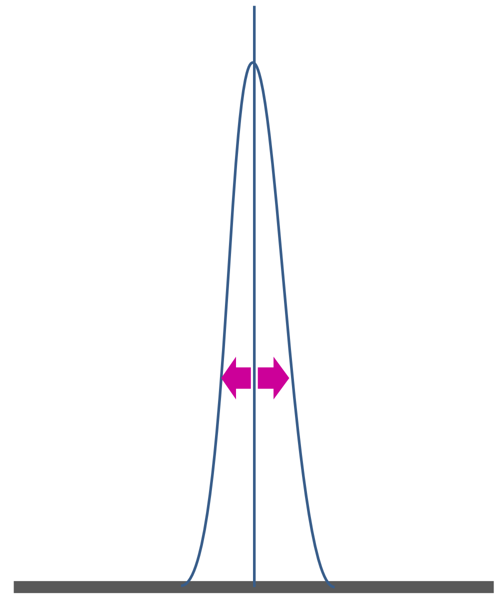
- 体重の分布
- 所得の分布など

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

計算方法

平均値

- 合計を計算
- 要素数で割る

計算方法

中央値

小さい順(大きい順)にならべて、
真ん中の値を取る

- 要素が奇数の場合、真ん中の値を採用
- 要素が偶数の場合、真ん中の2要素の平均値を計算

計算方法

分散、標準偏差

ばらつき

＝

平均値からのずれの大きさ

中央値ではなく

計算方法

分散

- 平均値を計算
- 各要素の値-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

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

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

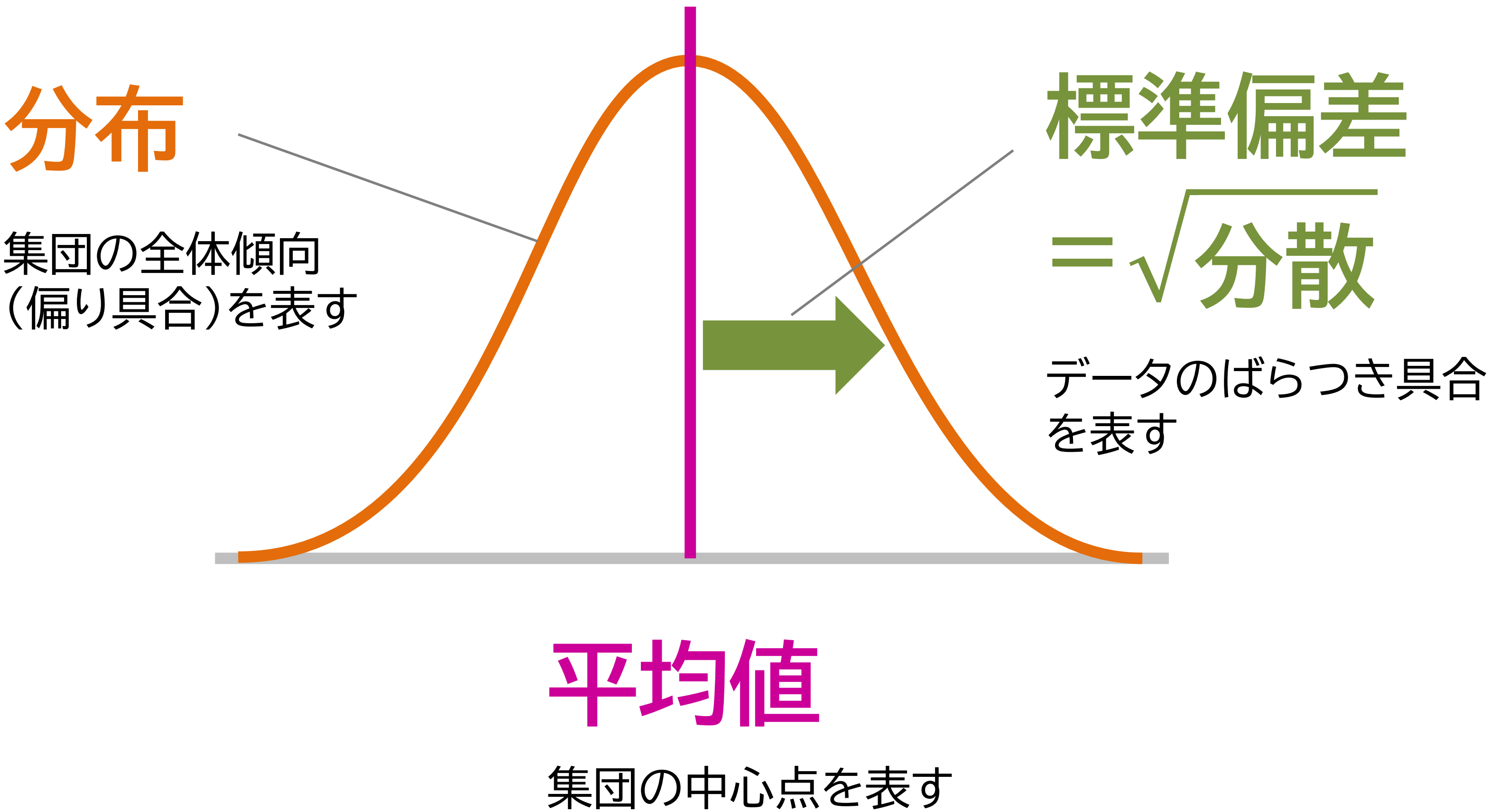
計測した値と単位を
そろえるため、
平方根を計算

標準偏差

目的:
このクラスの人の身長は
どのくらい？

	平均		標準偏差
男性		±	
女性		±	

集団を可視化したイメージ



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？

復習

統計学

- データを集める
- 解析する
- 解釈する

ための**方法論**

⇩

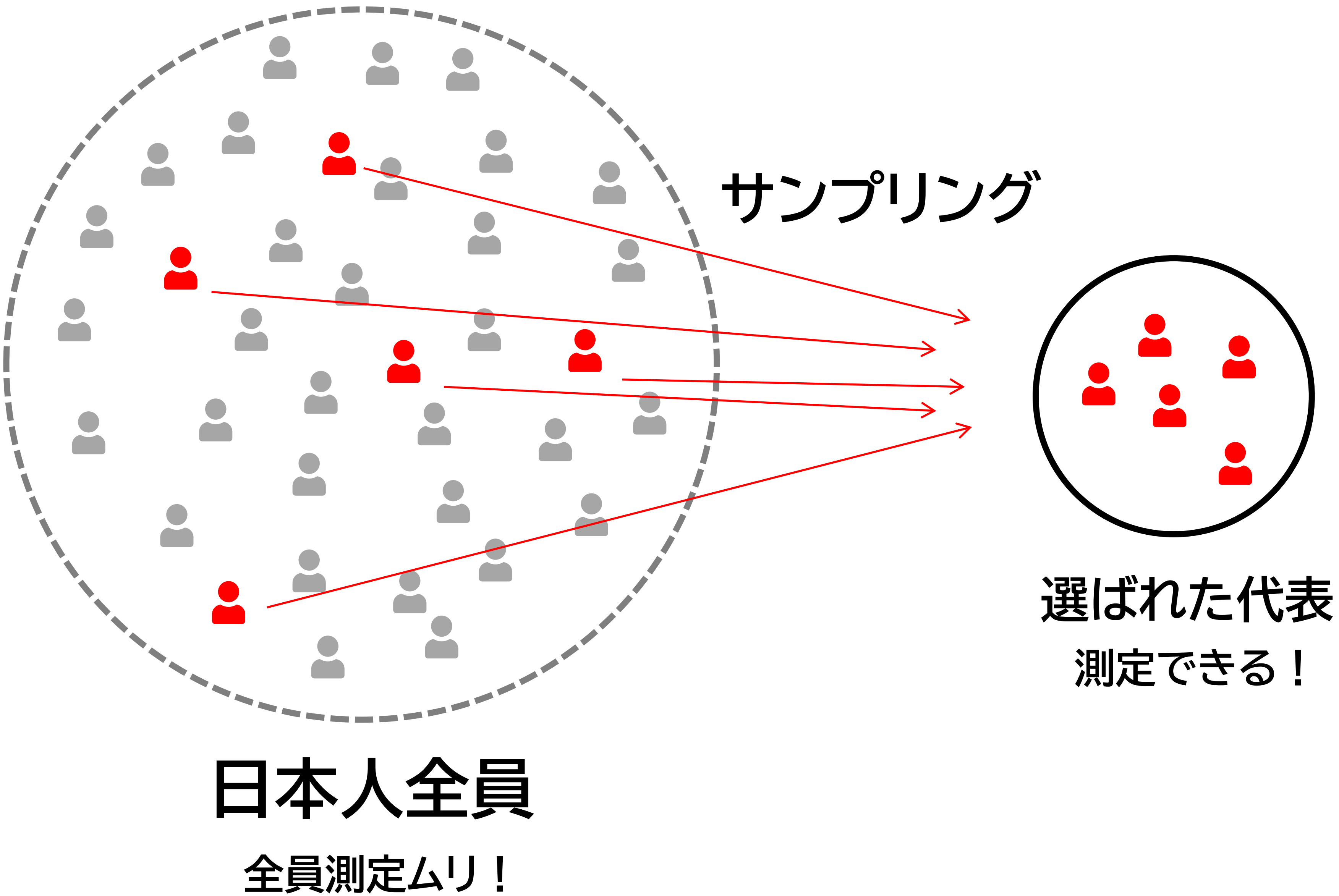
結果：集団の〇〇がわかった！

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
選ばれた代表
測定できる！
日本人全員
全員測定ムリ！

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?

↓

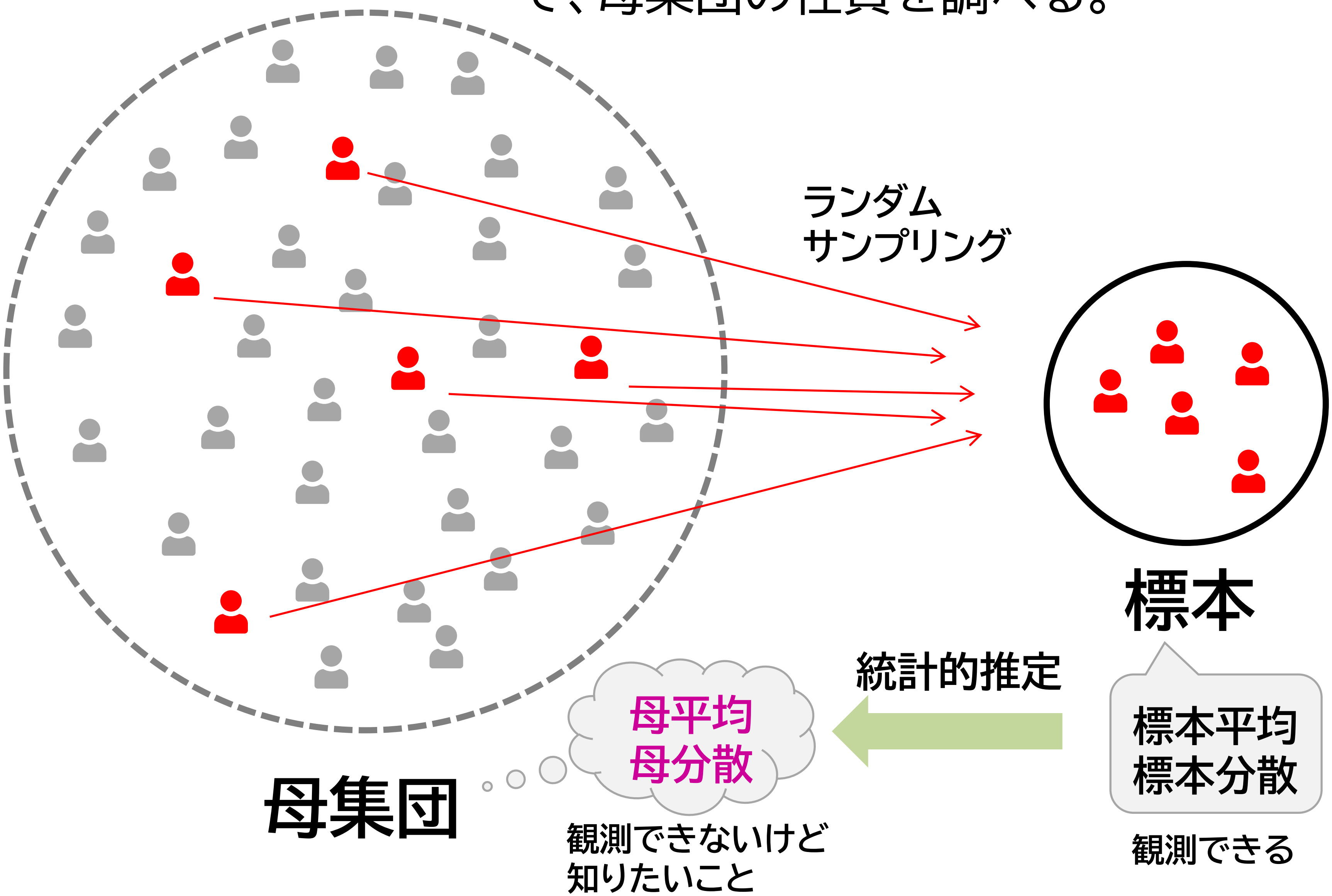
限られた標本を使って
母集団(日本人全体)の

- **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

実は一致が期待できない!!

一致が期待できるのは、母集団の全標本を観測できる場合(全数検査)だけ

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

μ: 平均(mean)のm
σ: 標準偏差(standard deviation)のs
に相当するギリシャ文字

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

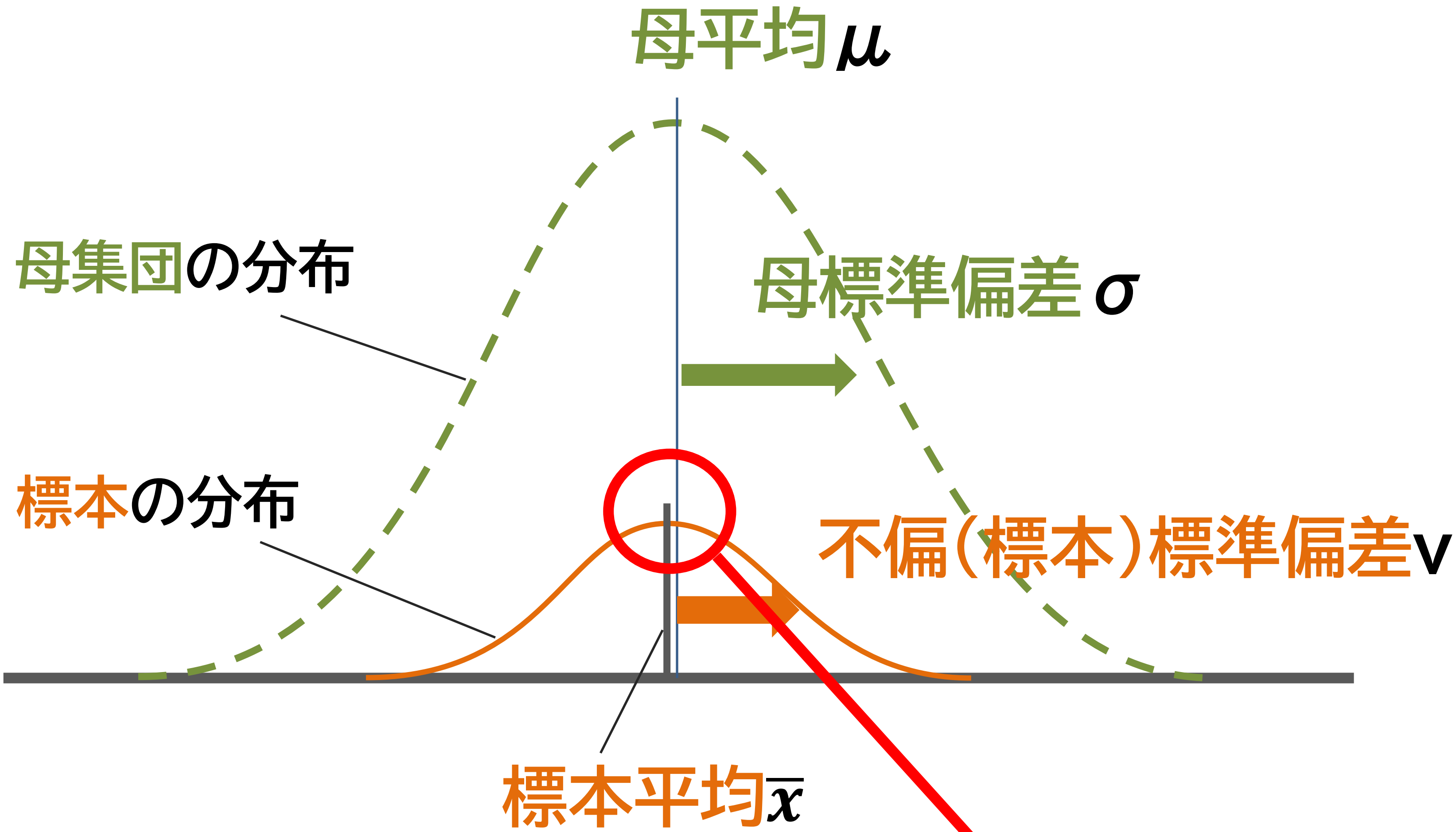
④その全要素(iが1からnまで)の合計

不偏(標本)分散

⑤n-1で割る

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

母平均μ（真実）と、標本平均$\overline{x}$（推定）のズレ

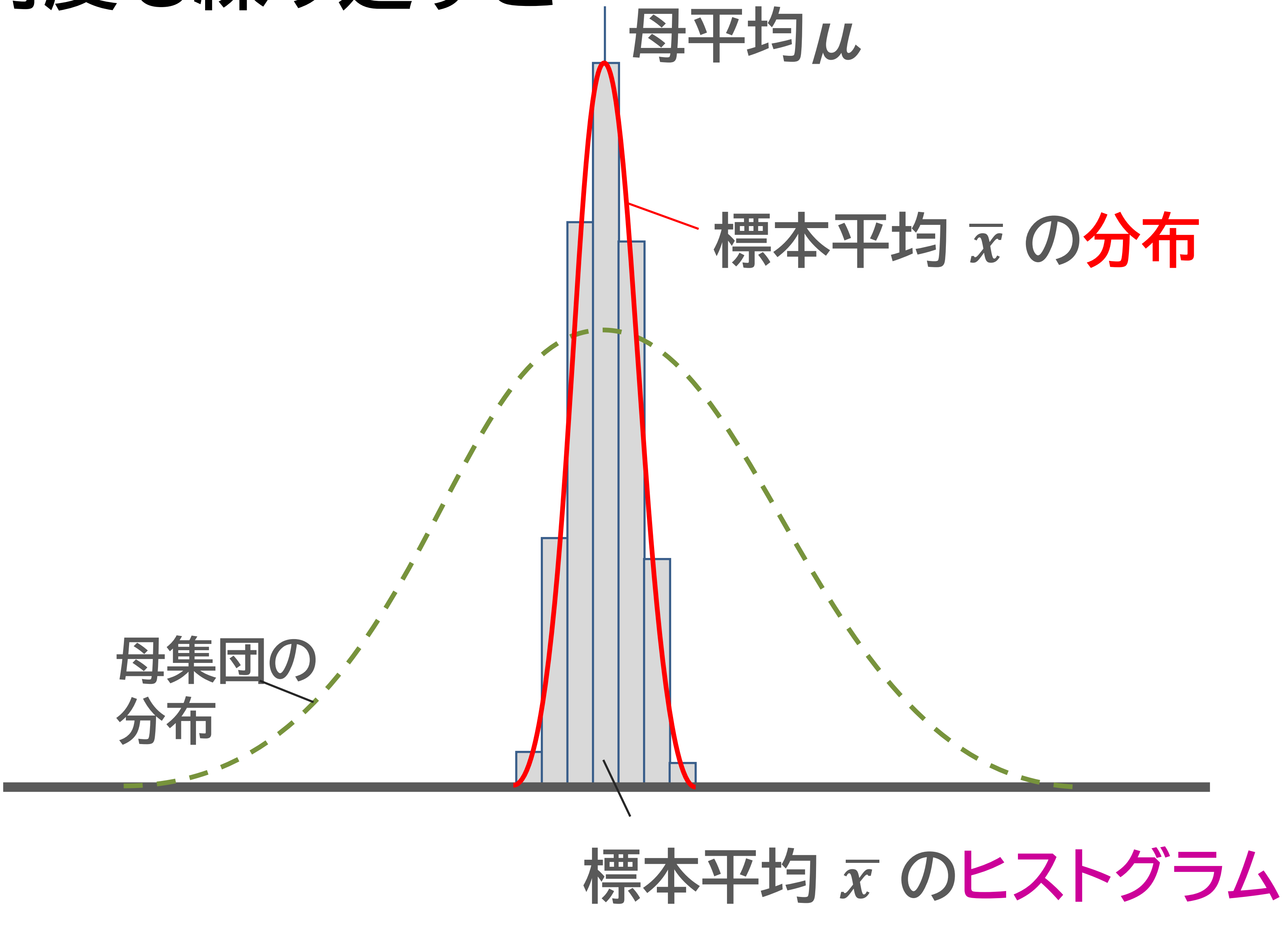


どれだけズレているか？

誤差

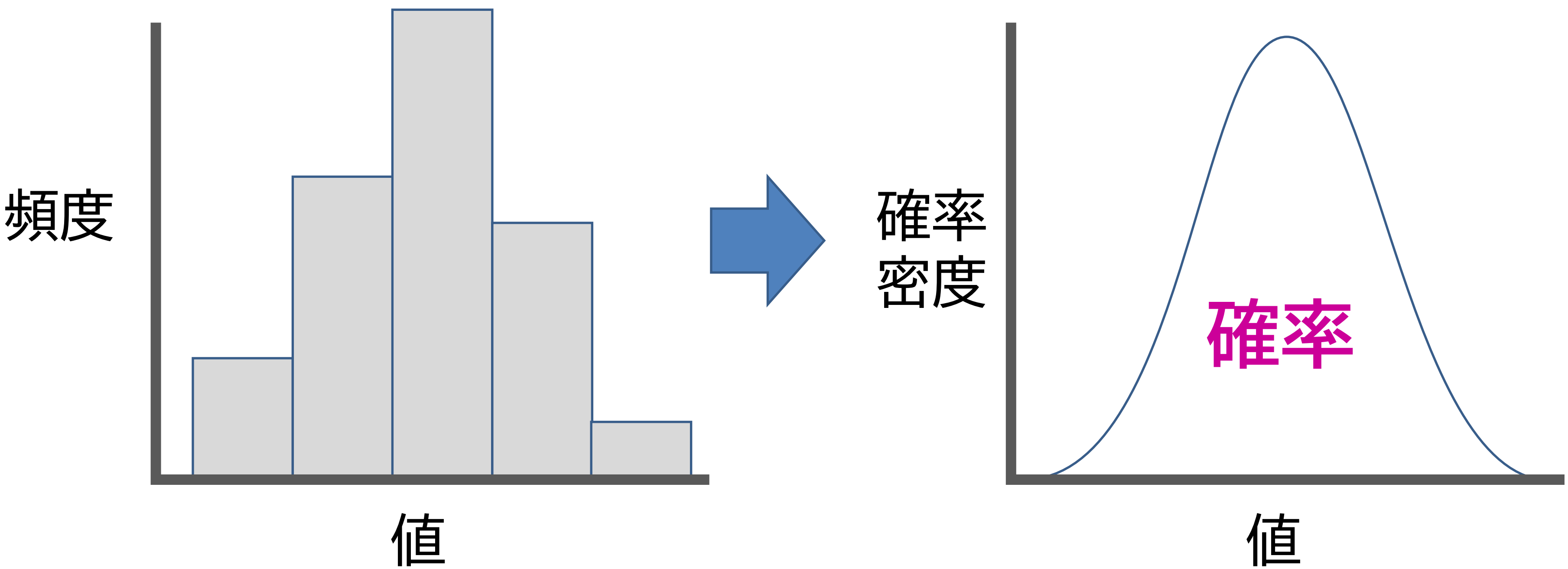
- サンプリング誤差
- 測定誤差

サンプリングして標本平均$\bar{x}$を算出して、を何度も繰り返すと…



分布

データの全体的な偏り具合を表したもの



ヒストグラム（頻度分布図）

①観測結果を表す図

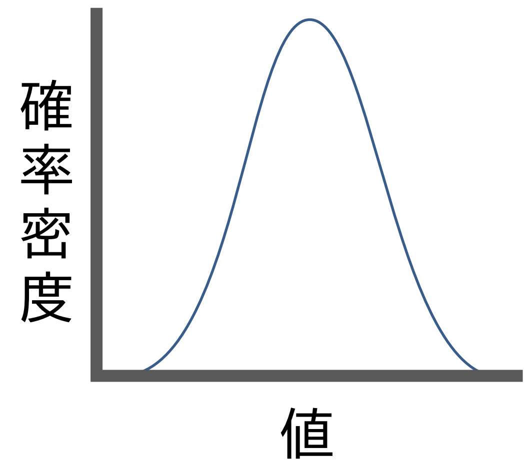
確率密度関数

②事象の起こる確率を表す関数

代表的な確率密度関数

正規分布（ガウス分布）

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



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

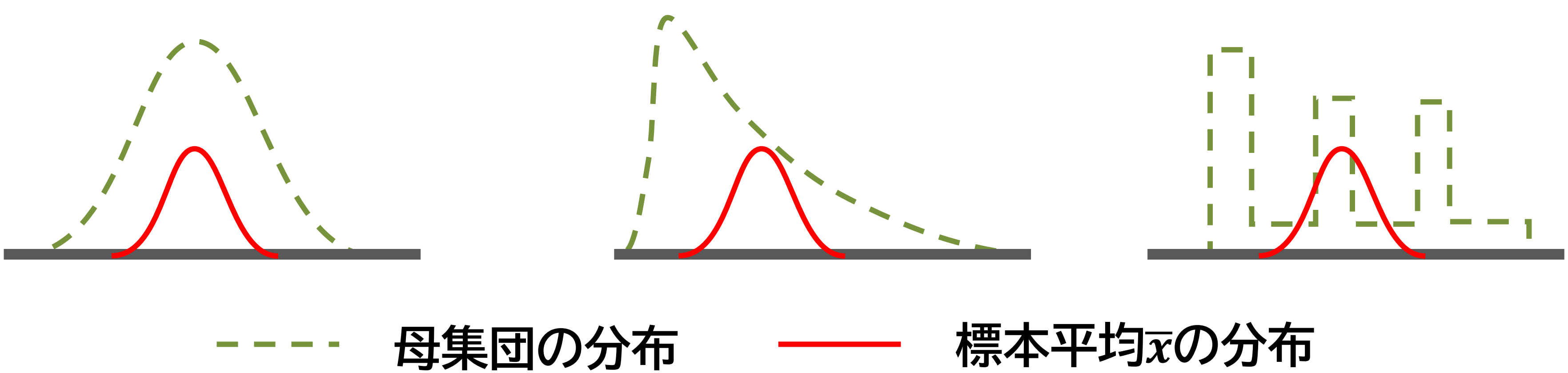
均等な確率で生じたばらつきの場合にとる分布

✓ 身長の分布
✓ 測定誤差の分布
✓ 自然界で起こるゆらぎ
✓ 標本平均$\bar{x}$ の分布

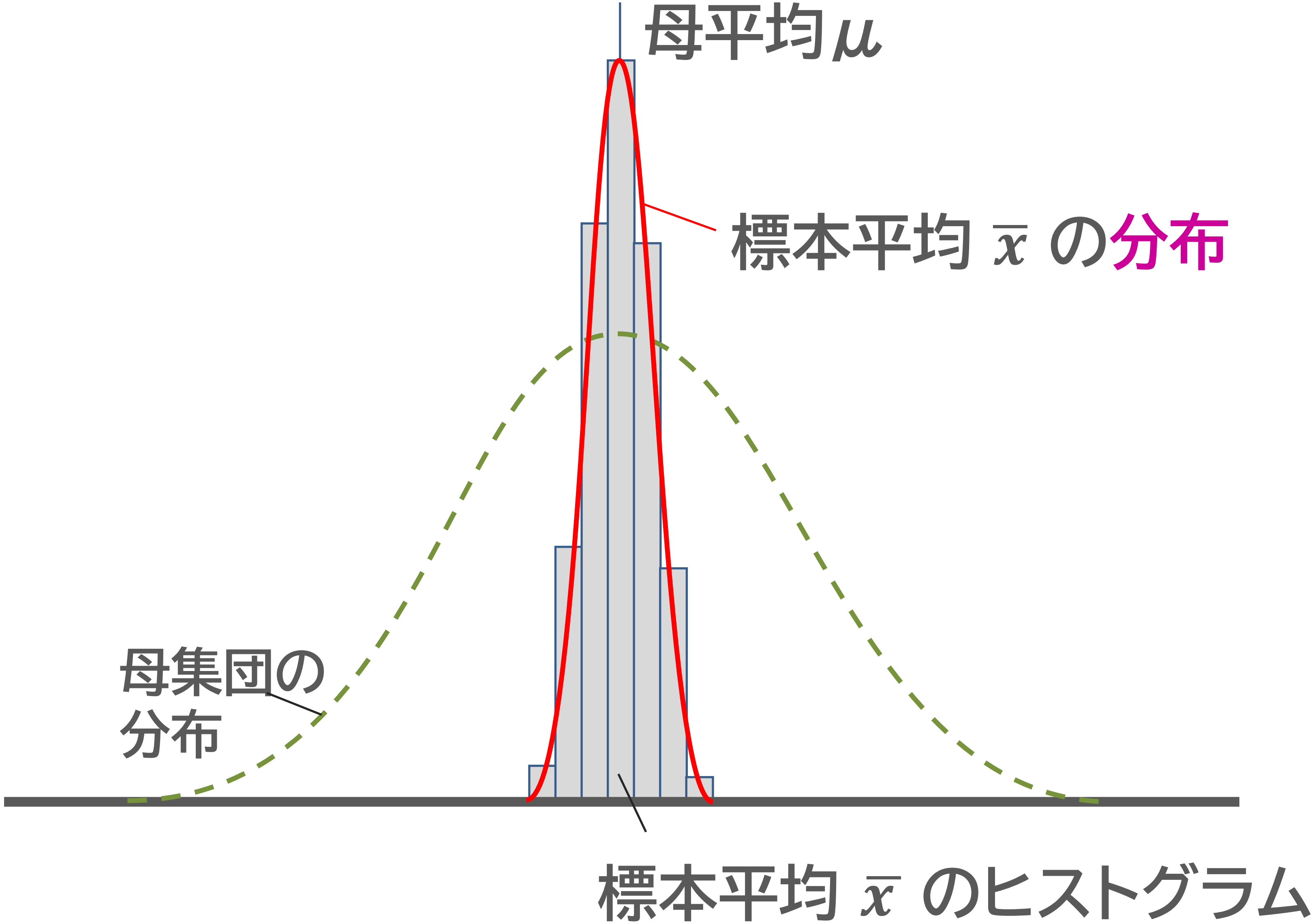
など

中心極限定理

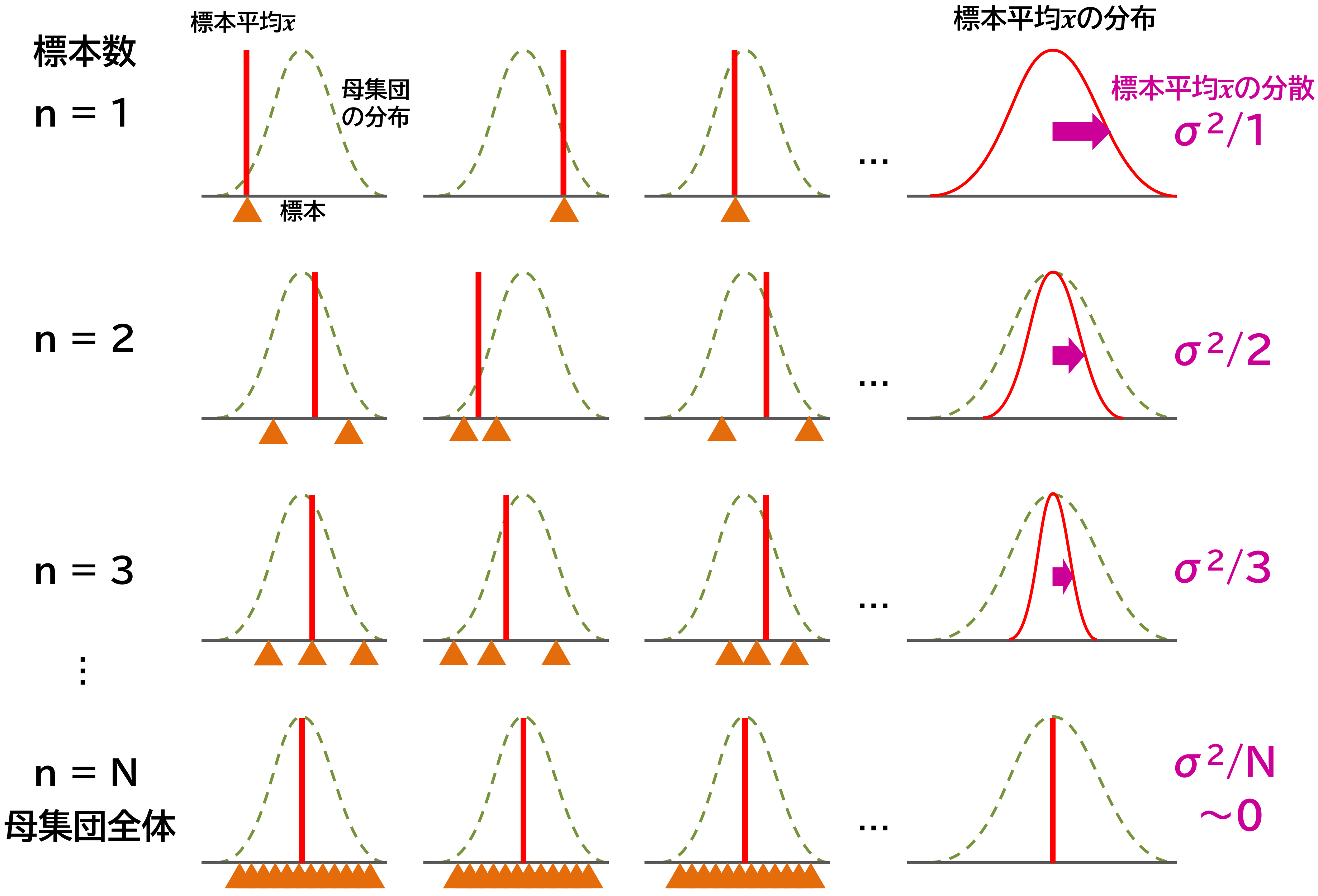
母集団から標本をサンプリングして、標本平均$\overline{x}$を計算することを繰り返すと、標本平均$\overline{x}$の分布は、正規分布に近づく。
母集団がどんな分布であっても成り立つ。
分散が無限でなければ



何個の標本をサンプリングしたらよいか？
標本数を変えたとき、分布はどうなるか？



標本数を変えると…


標本数
n = 1
n = 2
n = 3
n = N
母集団全体
標本平均$\bar{x}$
母集団
の分布
標本
…
標本平均$\bar{x}$の分布
標本平均$\bar{x}$の分散
$\sigma^2/1$
$\sigma^2/2$
$\sigma^2/3$
σ^2/N
~ 0

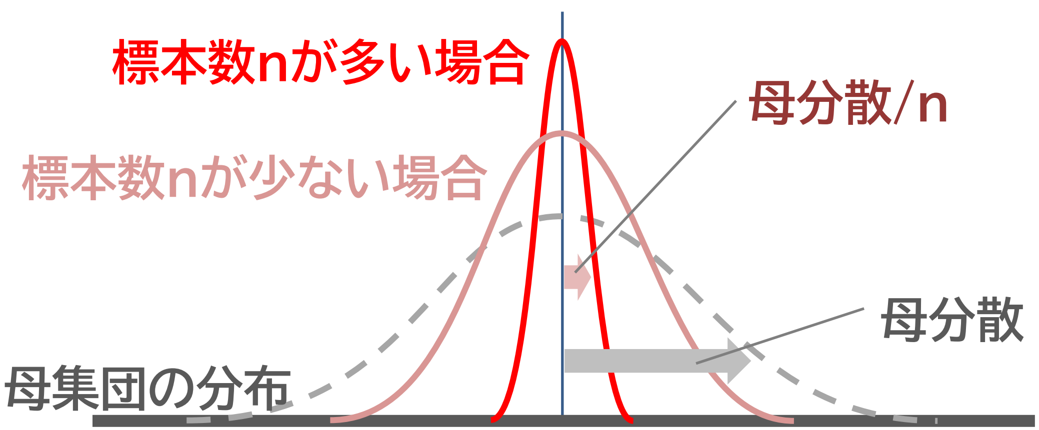
標本平均$\bar{x}$の分布の性質

- 正規分布に従う
- 分散は、標本数 n が大きいほど、小さくなる

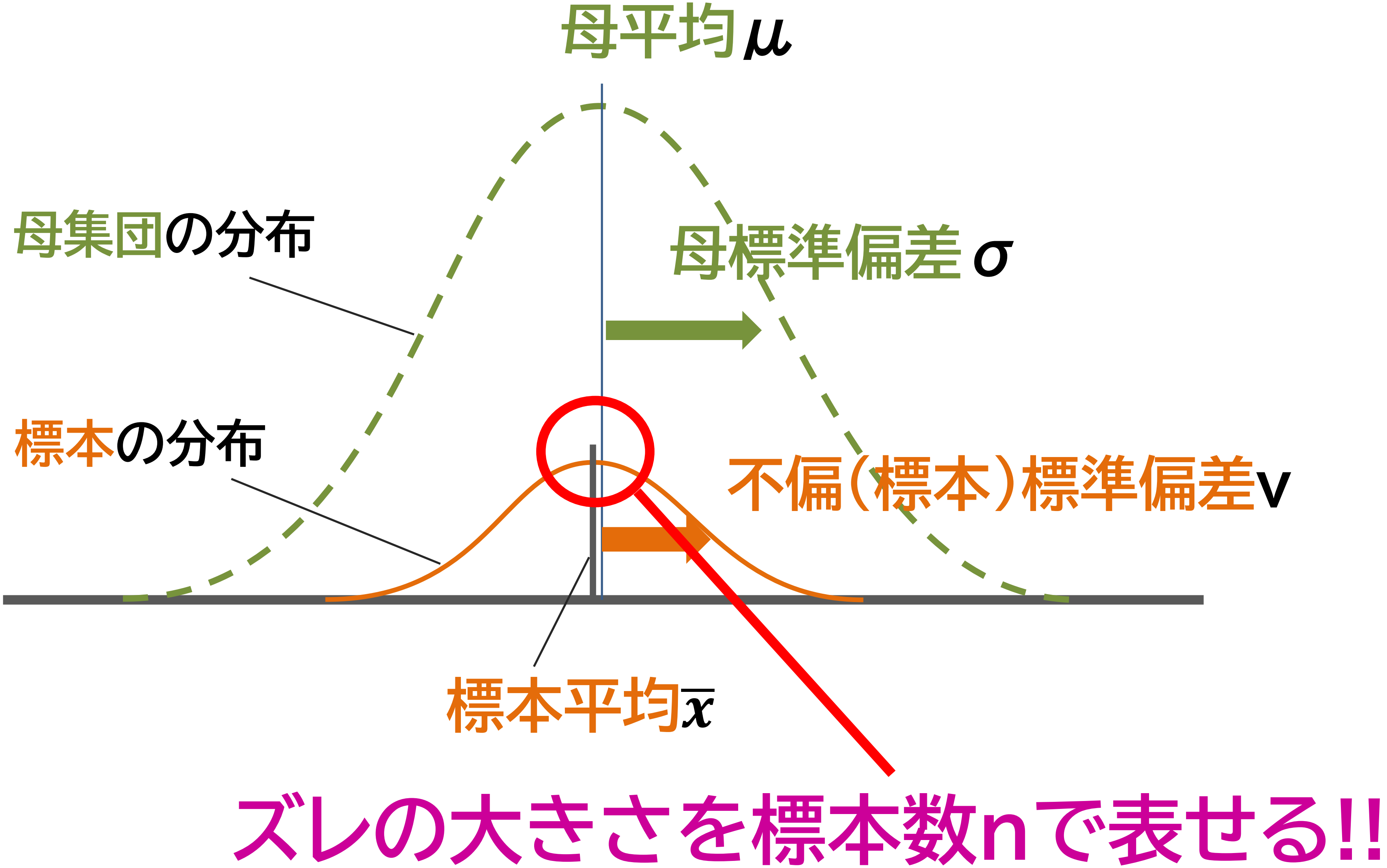
  n=1 なら、母集団のうち一つずつを測定するのと同じなので、分散も同じ。
  n=母集団数N なら、全数検査なので、母平均μとのずれはゼロになる。

- 分散は、**母分散σ^2の1/n**になる

  標本平均$\bar{x}$を計算したときの、母分散とのズレの大きさ

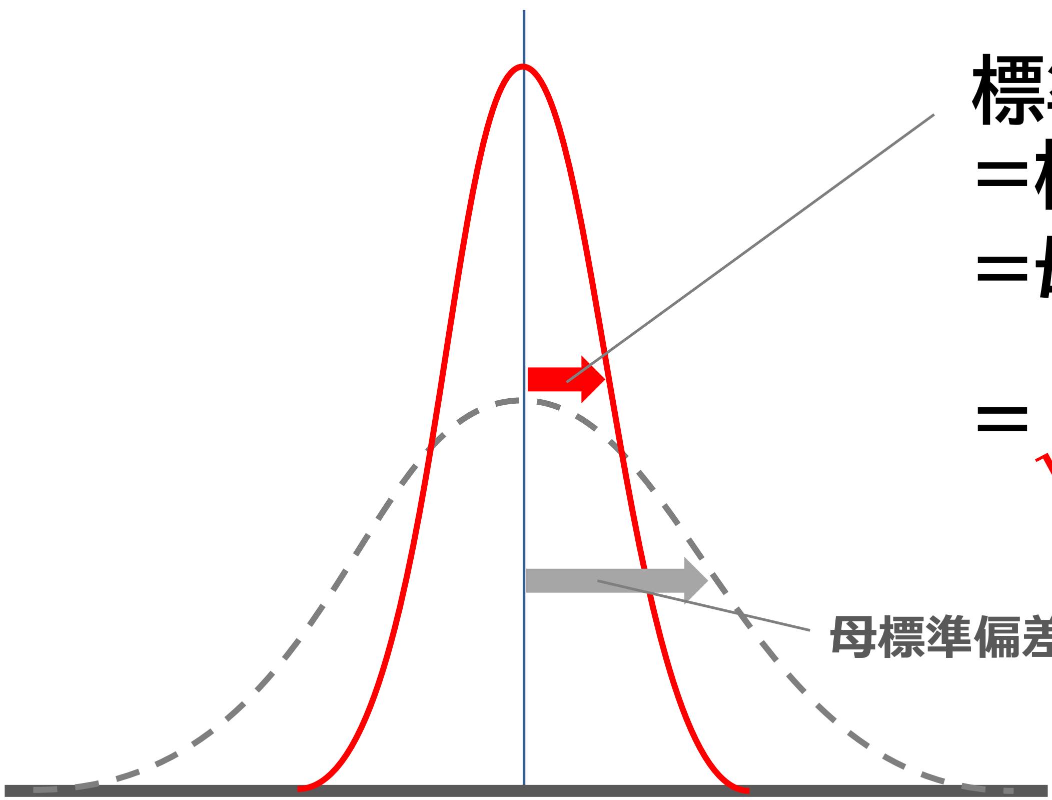


母平均μ（真実）と、標本平均$\bar{x}$（推定）のズレ



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



母標準偏差σ

標準誤差
＝標本平均$\bar{x}$の標準偏差
＝母分散σ^2/nの平方根

$$=\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

$$=\frac{v}{\sqrt{n}}$$

不偏標本
標準偏差

標準偏差 と 標準誤差

論文などでよく見る図

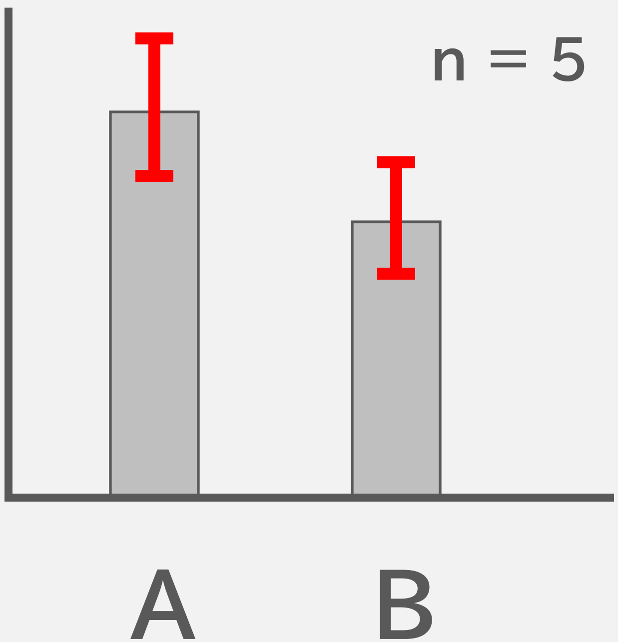


図1 A群とB群の＊＊の違い
それぞれ5個体を測定した。エラーバーは標準偏差を表す

エラーバーが**標準偏差**

↓

測定した標本自体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

【ここに注意！】

標準誤差は$1/\sqrt{n}$が掛けられているので、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当なのかどうかを、正しく判断しながらデータを解釈しましょう。

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

実は一致が期待できない!!

一致が期待できるのは、母集団の全標本を観測できる場合（全数検査）だけ

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

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

母分散 σ^2

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$\geqq$$

標本分散 s^2

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

$\frac{\sigma^2}{n}$ のばらつきを持つ

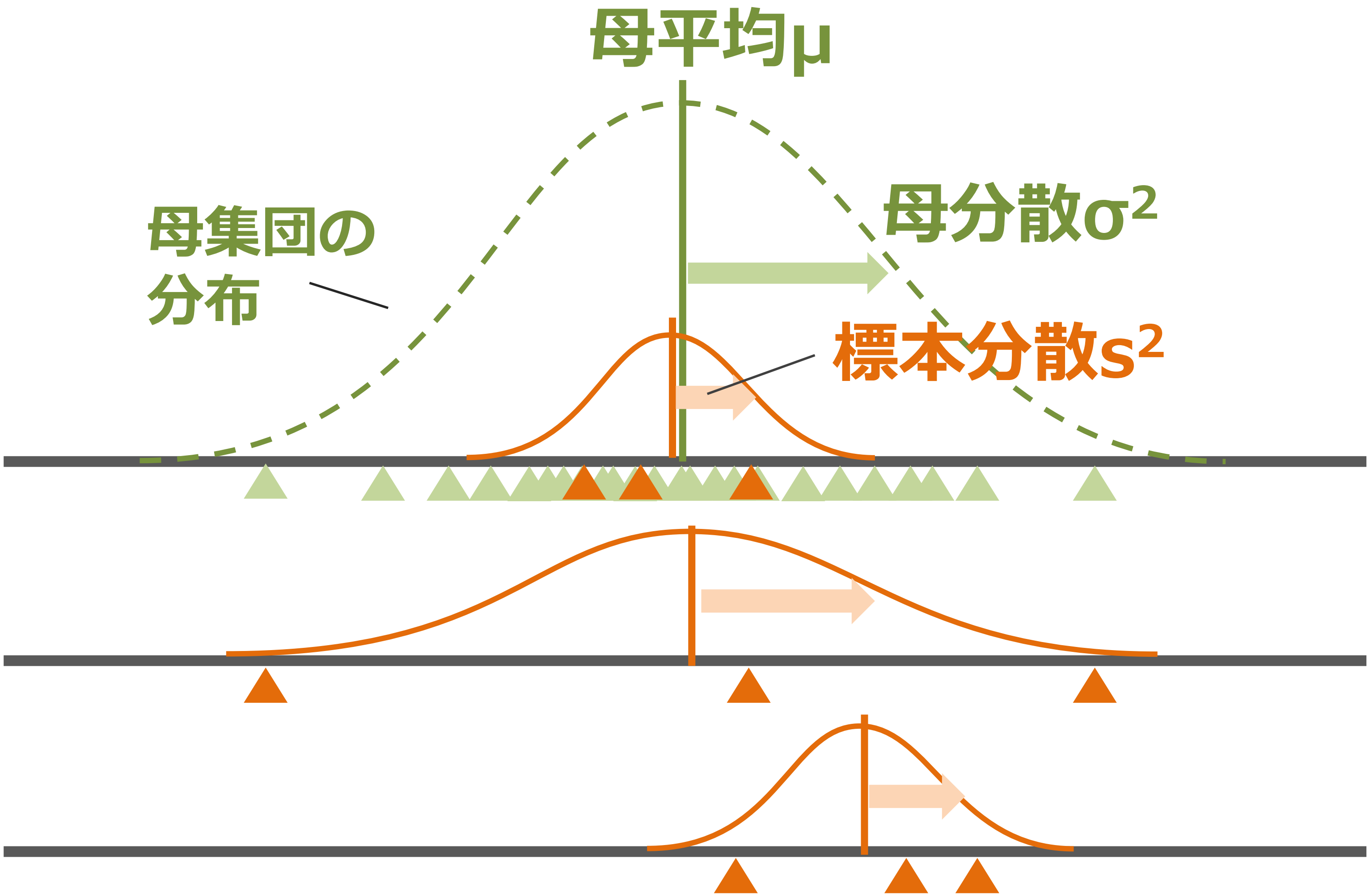
$$\approx \frac{1}{\boldsymbol{n-1}}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

不偏(標本)分散 v^2

標本分散は、計算に標本平均$\bar{x}$が使われ、$\bar{x}$は一定のばらつきを持つため、母平均μを使った本来の計算より小さく見積もられてしまう。それを補正するためにn-1で割った不変標本分散を使う。

なぜn-1で割ると補正されるのか？

標本分散s^2は、母分散σ^2よりも小さくなる



どのくらい小さくなっているか？

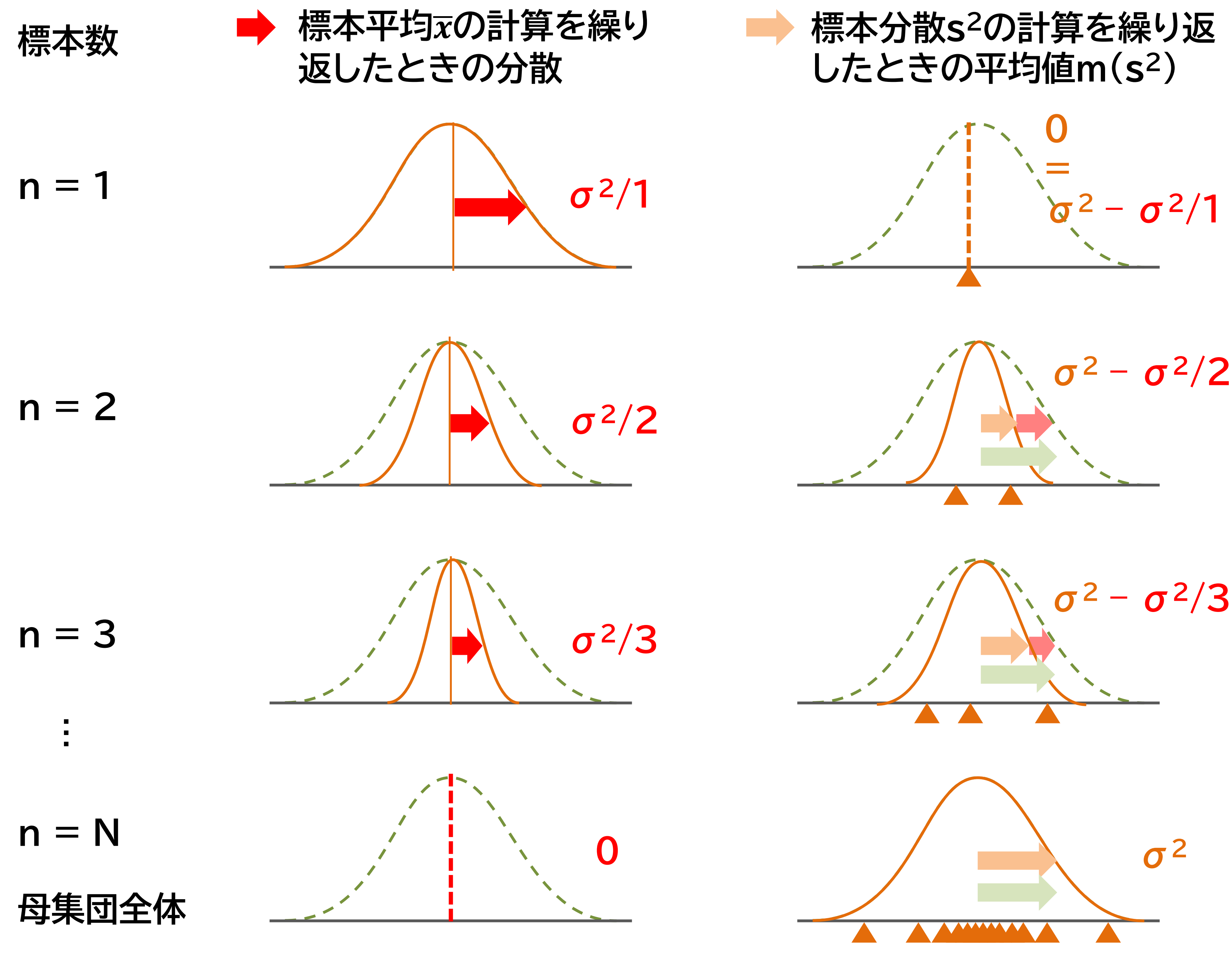
サンプリングして標本分散s^2を算出して、を何度も繰り返すと…

…

標本分散s^2

$m(s^2)$ = 母分散σ^2 - X

標本分散s^2の平均値



標本数
標本平均x̄の計算を繰り返したときの分散
標本分散s²の計算を繰り返したときの平均値m(s²)
n = 1
σ²/1
0
=
σ² − σ²/1
n = 2
σ²/2
σ² − σ²/2
n = 3
σ²/3
σ² − σ²/3
⋮
n = N
母集団全体
0
σ²

どのくらい小さくなっているか？

サンプリングして標本分散s^2を算出して、を何度も繰り返すと…

…

標本分散s^2

$$m(s^2) = \text{母分散}\sigma^2 - \frac{\sigma^2}{n}$$

標本分散s^2の平均値

式を変形して、**母分散σ^2**を出してみると…

$$\mathrm{m}(\mathrm{s}^2) = \sigma^2 - \frac{\sigma^2}{n}$$

$$\mathrm{m}(\mathrm{s}^2) = \frac{n-1}{n}\sigma^2$$

$$\sigma^2 = \frac{n}{n-1}\mathrm{m}(\mathrm{s}^2)$$

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

$$\sigma^2 = \frac{1}{\boldsymbol{n-1}}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \mathrm{v}^2$$ （不偏標本分散）

不偏標本分散

n-1の意味

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

- **推定のあいまいさをうまく表現している**

  標本数nが少ないと、母分散とのズレが大きく、標本数が母集団の数N(大きな数)に近づくと、母分散に近づく

- **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)値の個数

  上の式では、一度 $\bar{x}$ を計算しているため、一つの観測値 x_i 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ためです。
　Excel関数も、標準偏差を求めるstdev関数は、不偏標本標準偏差(n-1で割る)を計算しています。

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

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

2025年8月4日　神奈川工科大学

櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

自習

統計サイトの検索
→　課題のテーマ探し、調査

情報統計
第4回

2025年8月5日　神奈川工科大学

櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

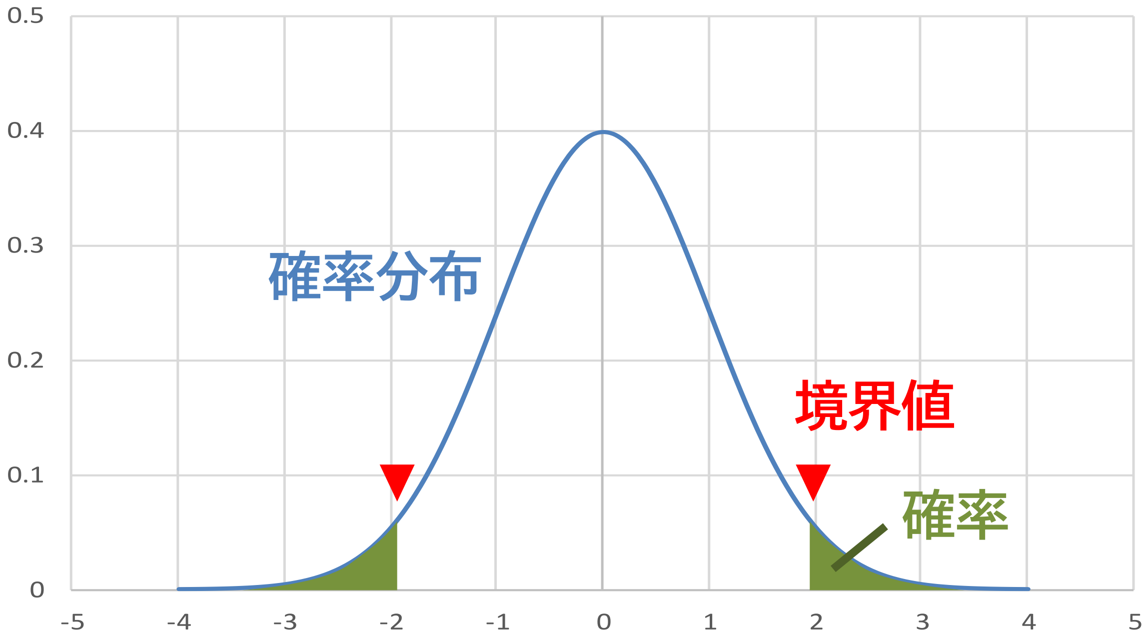
学習目標

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

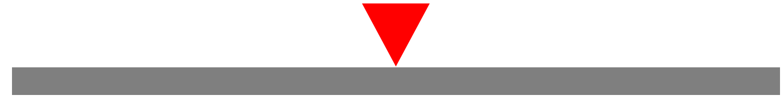
ポイント

- 確率分布とその使い方



昨日の復習

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

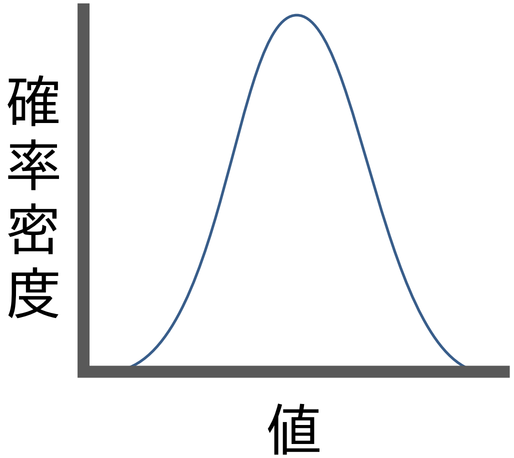
【注意】 以下の意味とは異なる
「神奈川県の95%の男子の身長は170.2～174.6 cmである」

標準正規分布

代表的な分布

復習

正規分布（ガウス分布）



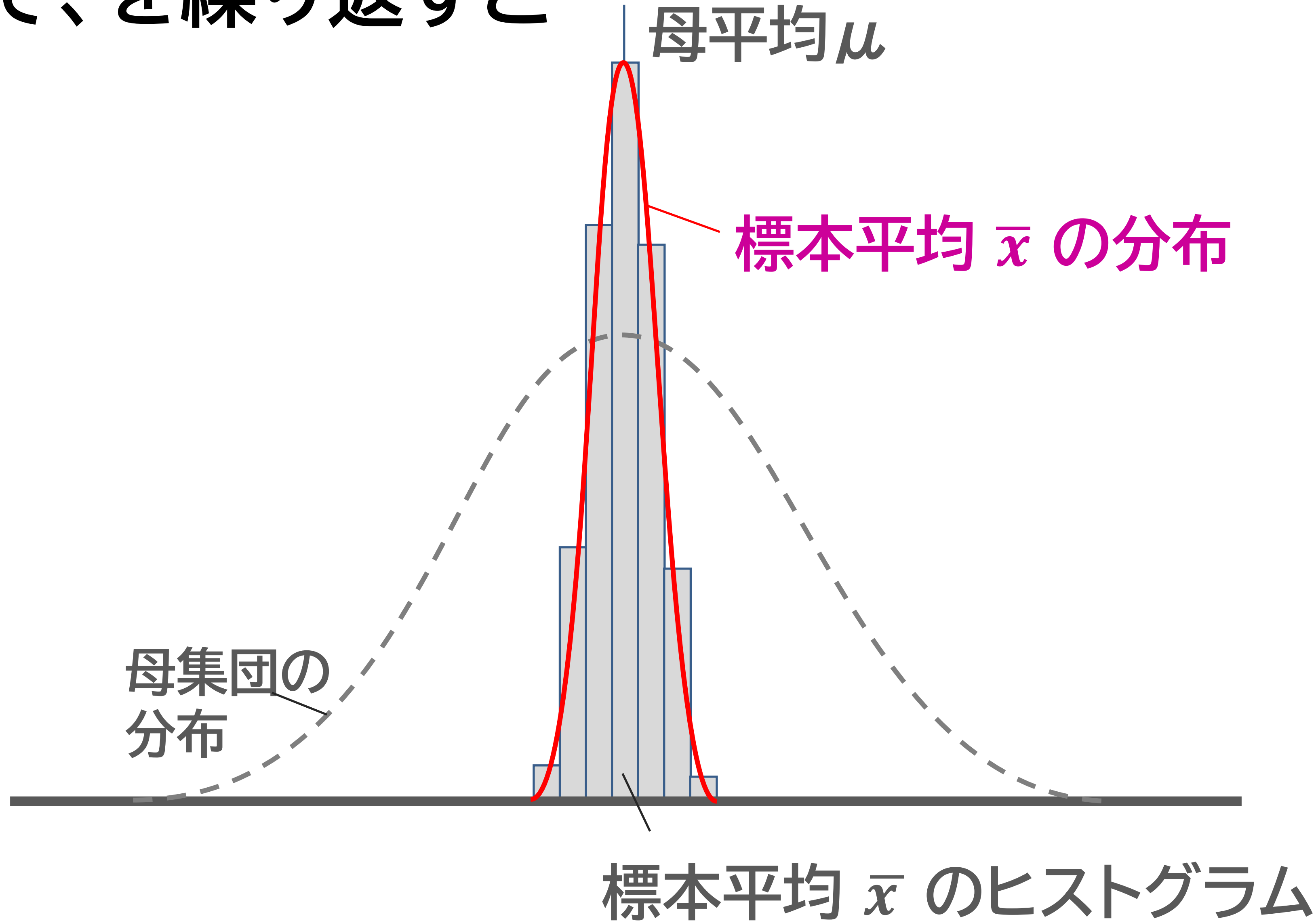
- 平均値が中心で、
- 平均値に近いものが多く、
- 左右に均等な釣り鐘状の分布

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

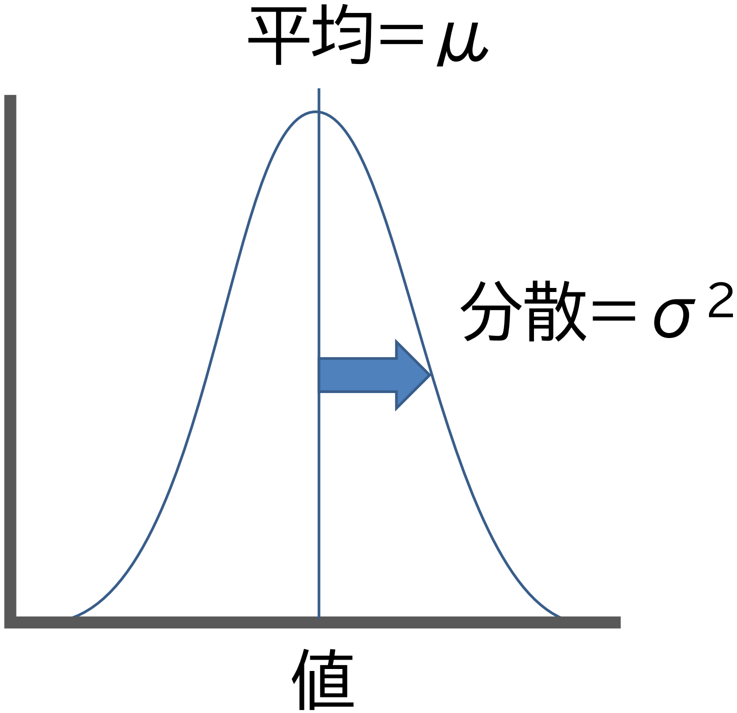
復習

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…



正規分布

無限にある

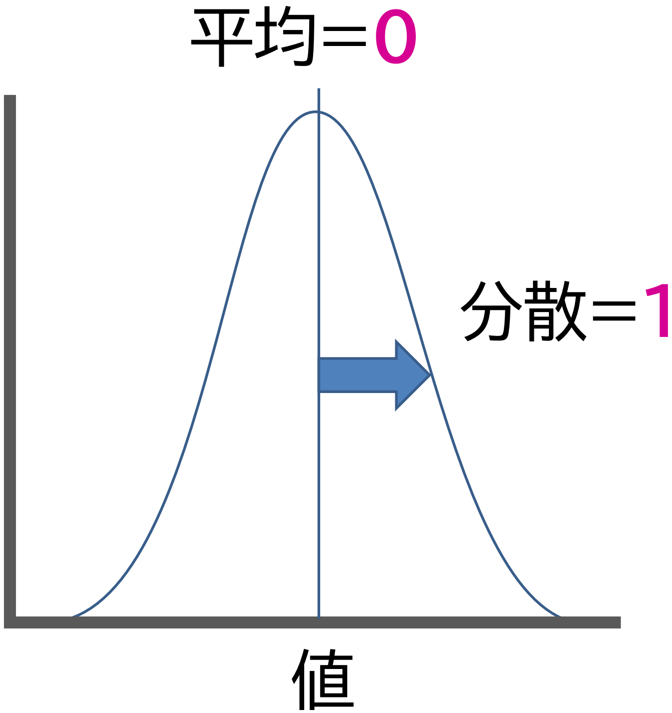


平均と分散で決まる

$N(\mu, \sigma^2)$と表記

標準正規分布

一つしかない

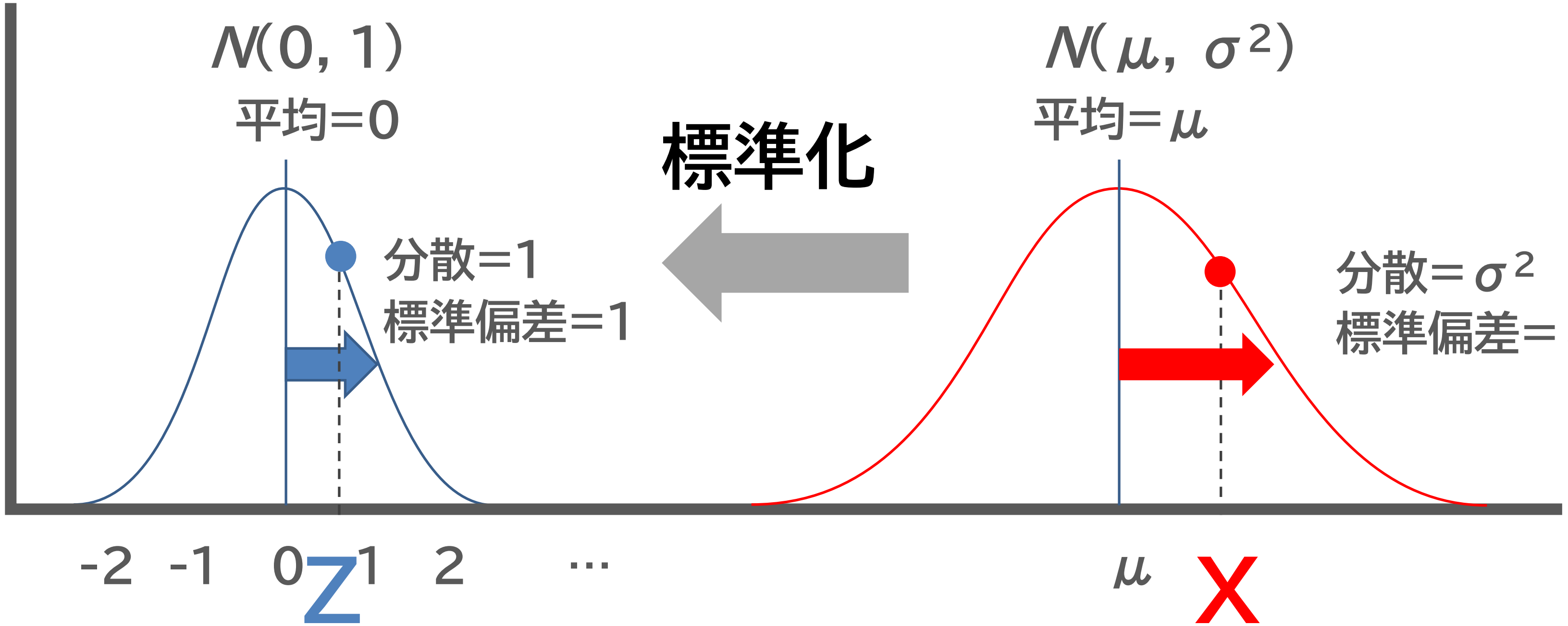


$N(0, 1)$

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、

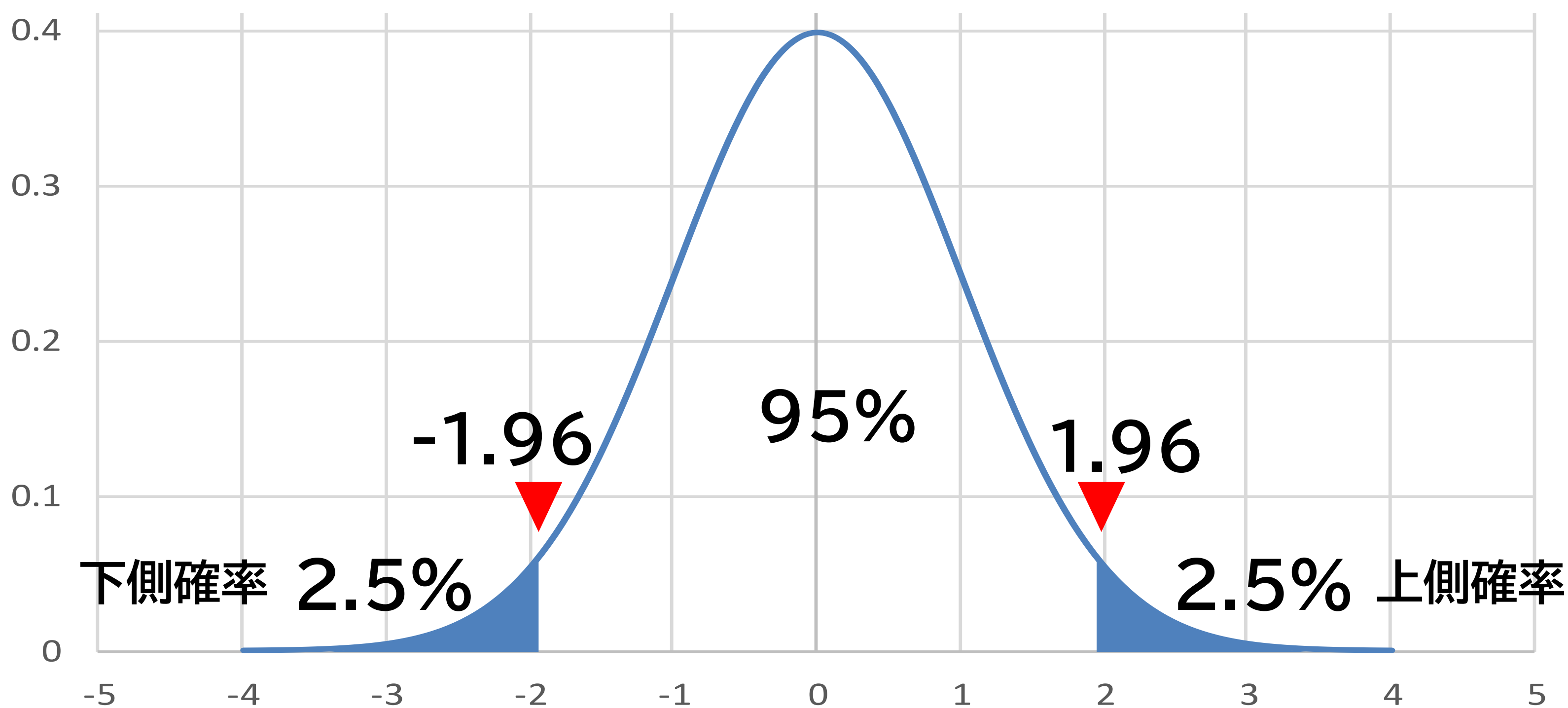
$$Z = \frac{X-\mu}{\sigma}$$ と変換すると、標準正規分布になる。



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。ある値より外側の面積が一意に計算できる
  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

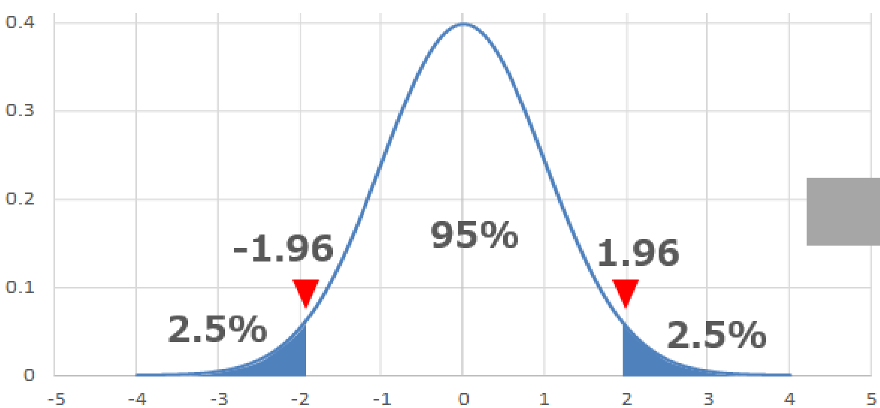
u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
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	$\boxed{0.02500}$	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
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.00539	0.00523	0.00508	0.00494	0.00480
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.00402	0.00391	0.00379	0.00368	0.00357

出典
https://to-kei.net/distribution/normal-distribution/table/

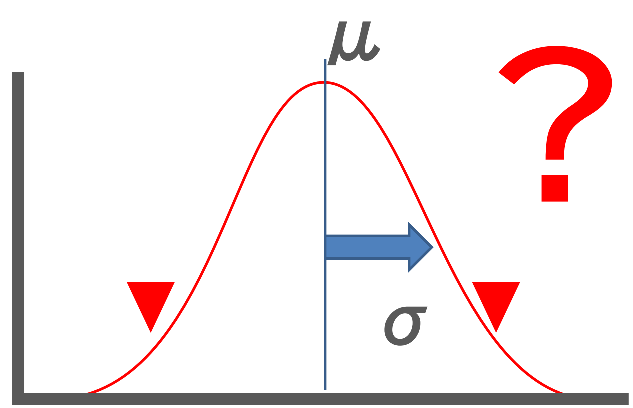
区間推定の考え方

- 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確率a%のときの境界値を用いて、もとの正規分布の境界値を計算する

（この境界値の間の区間を、a%信頼区間という）



標準正規分布



もとの正規分布

標準化

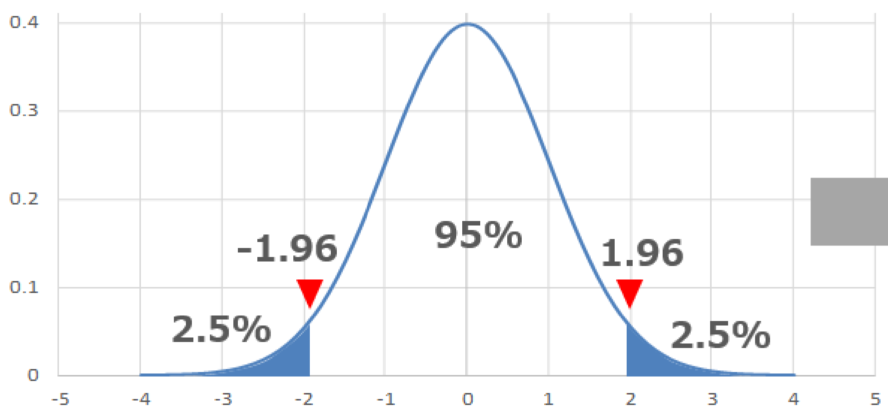
$$Z = \frac{X - \mu}{\sigma}$$

標準化の逆

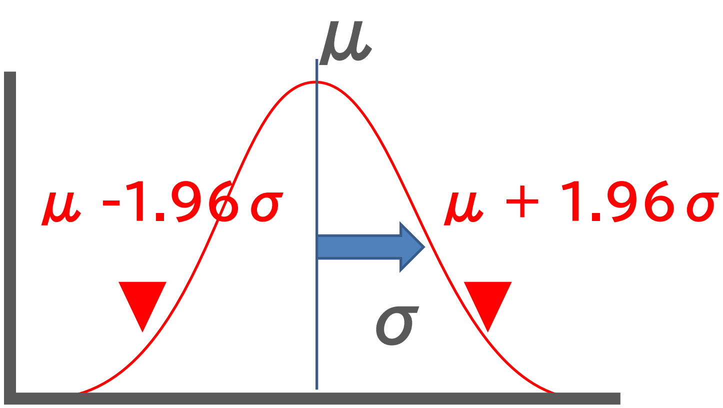
$$X = \mu + Z\sigma$$

例）Z ＝ 1.96なら、

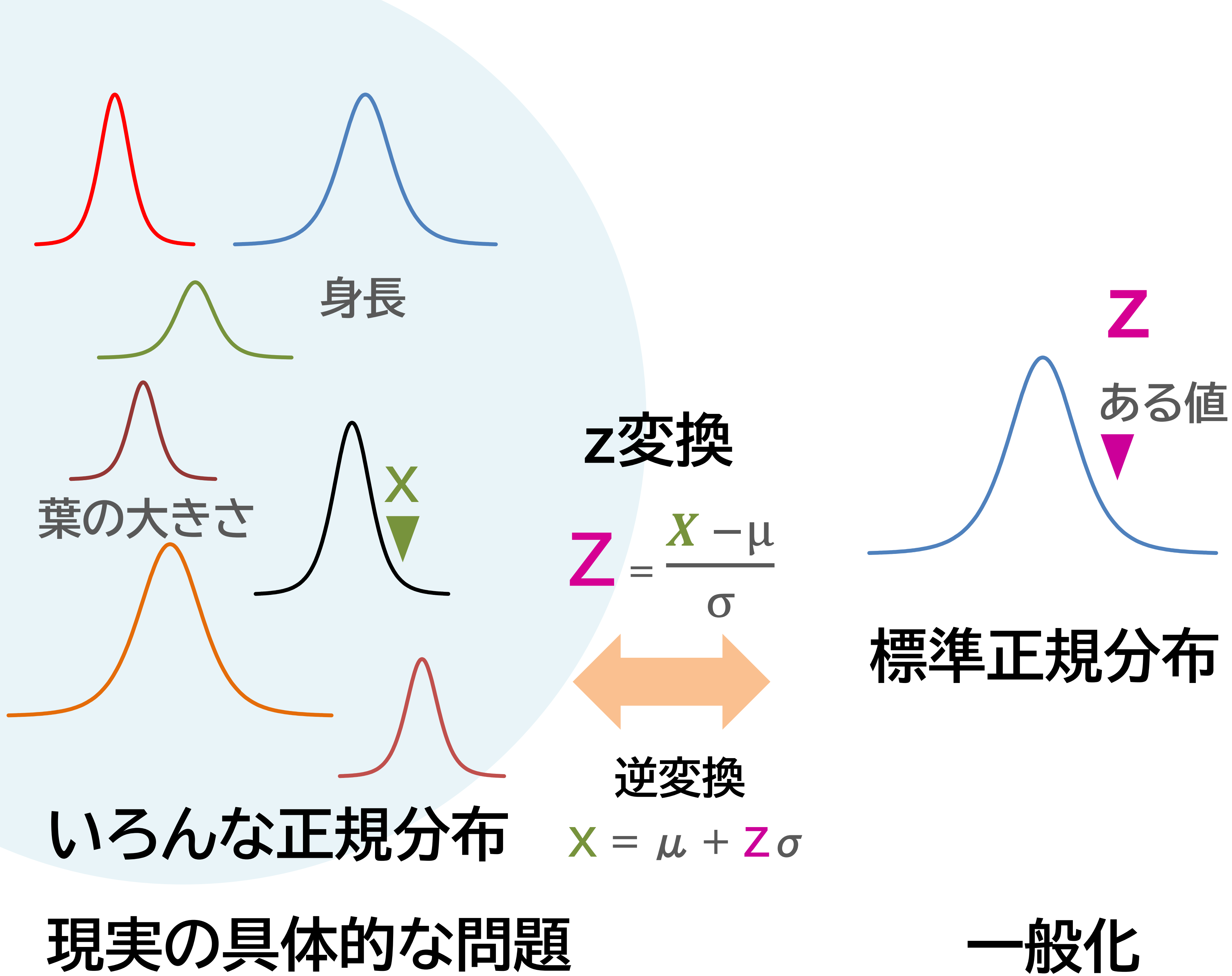
$$X = \mu + 1.96\sigma$$



標準正規分布



もとの正規分布



身長
葉の大きさ
x
z変換
$Z = \frac{X - \mu}{\sigma}$
逆変換
$x = \mu + Z\sigma$
いろんな正規分布
現実の具体的な問題
Z
ある値
標準正規分布
一般化

母集団の平均値を推定する問題の場合

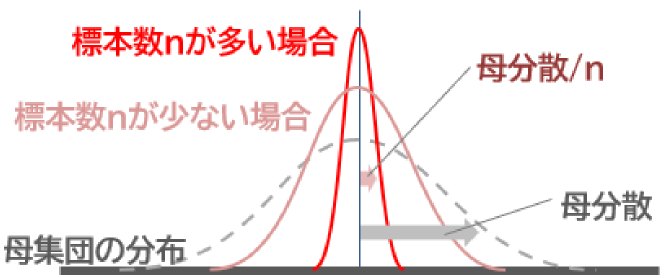
標本平均$\bar{x}$の分布の性質

- 正規分布に従う
- 分散は、標本数 n が大きいほど、小さくなる

  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。
  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。

- 分散は、**母分散σ^2の1/n**になる

  標本平均$\bar{x}$を計算したときの、母分散とのズレの大きさ



μ推定値: $\bar{x}$

標準偏差: $\dfrac{\sigma}{\sqrt{n}}$

を当てはめると…

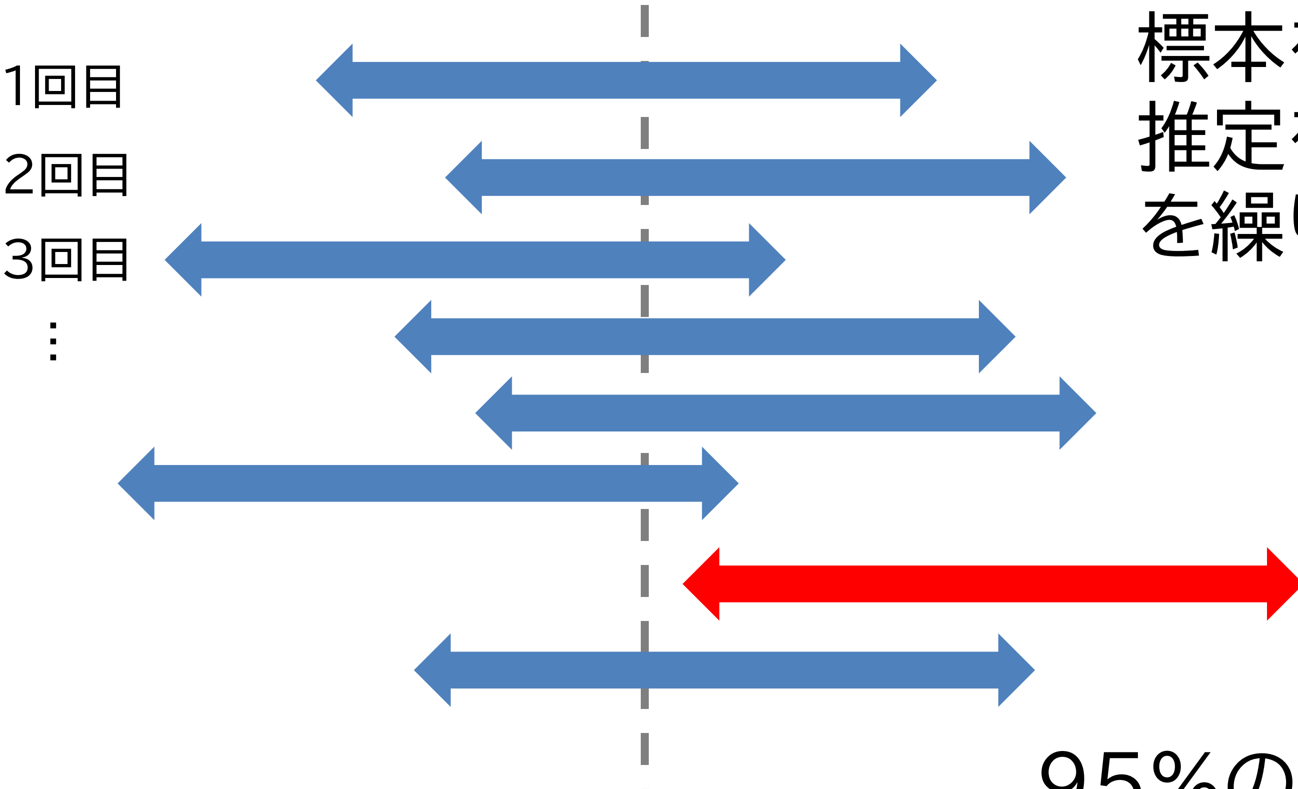
95%信頼区間は、以下で求められる

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

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 $\frac{\sigma}{\sqrt{n}}$

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、母平均がその区間内に含まれる場合が95回おこる」

イメージ



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

真の母平均
（変わらない）

一般化すると

区間推定（分散既知の場合）

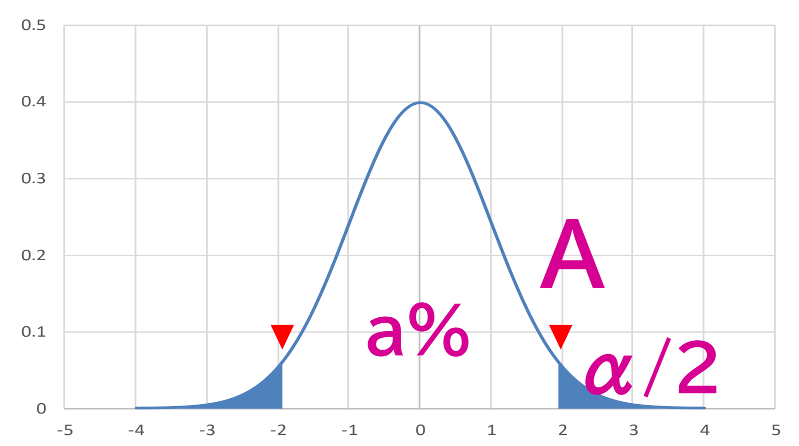
母平均μ、母分散σ^2の正規分布する母集団から抽出したn個の標本から求められる、a%信頼区間は以下となる。

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

ここでAは、標準正規分布表から、

確率α ＝ (100-a)/100

で求められる境界値



ただし…

以上のアプローチは、
母分散σ^2がわかっていると仮定した場合の話

現実の問題では、母分散σ^2は不明な場合がほとんど

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

母平均μが不明なので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。

↓

母分散の代わりに、不偏標本分散を使えばよい。
その場合、正規分布を仮定してもよいのか？

↓

t分布を使う

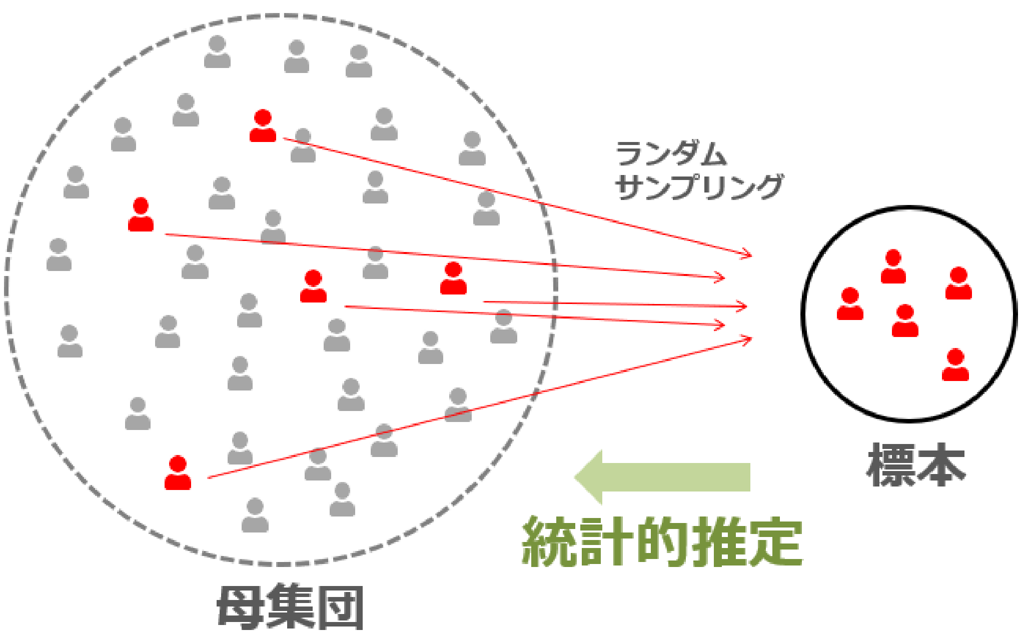
t 分布

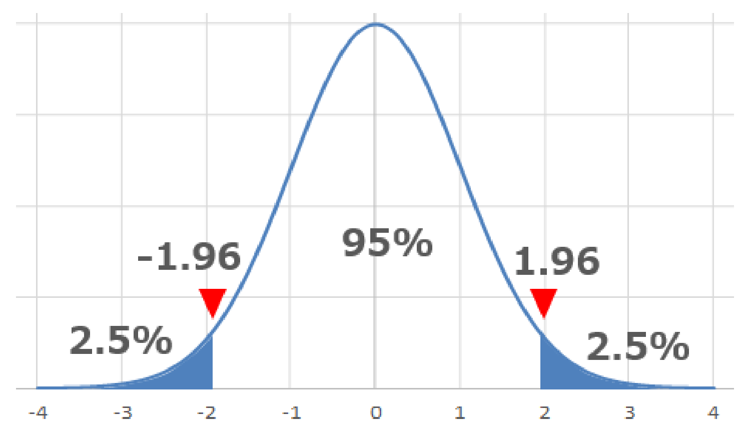
標準正規分布の、
標本数が少ない場合の
実用化バージョン

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。





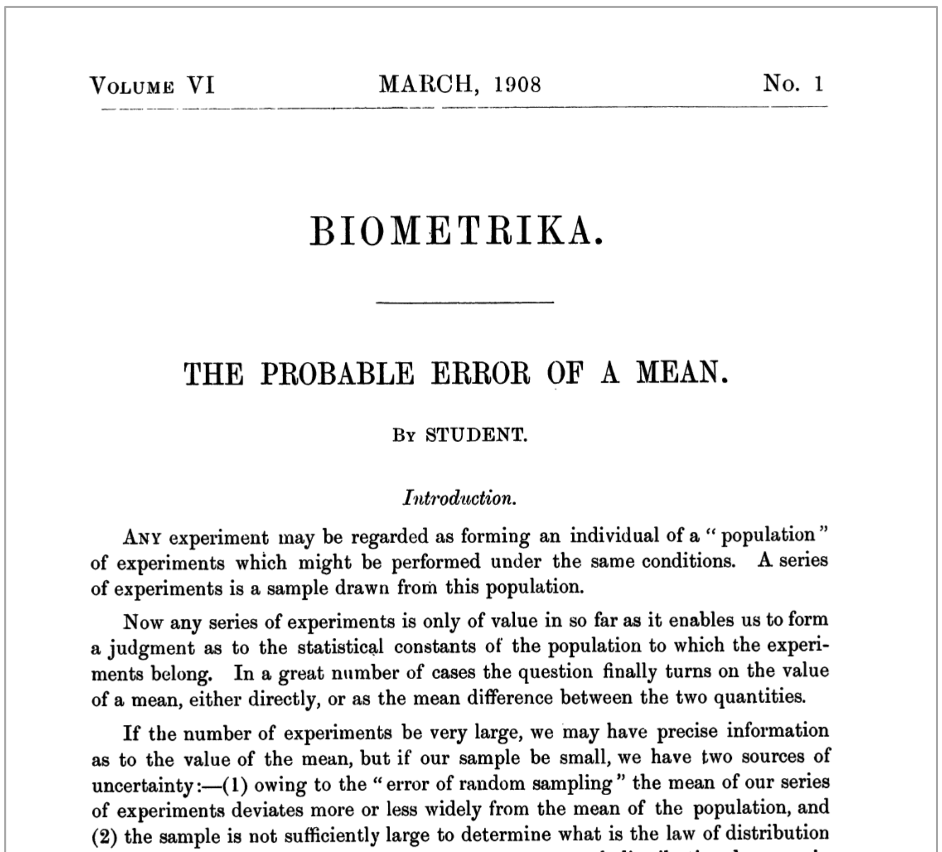
考えた人

ウィリアム・シーリー・ゴセット
William Sealy Gosset (1876-1937)

イギリスの統計学者

出典：Wikipedia

出典：ギネス社HP



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

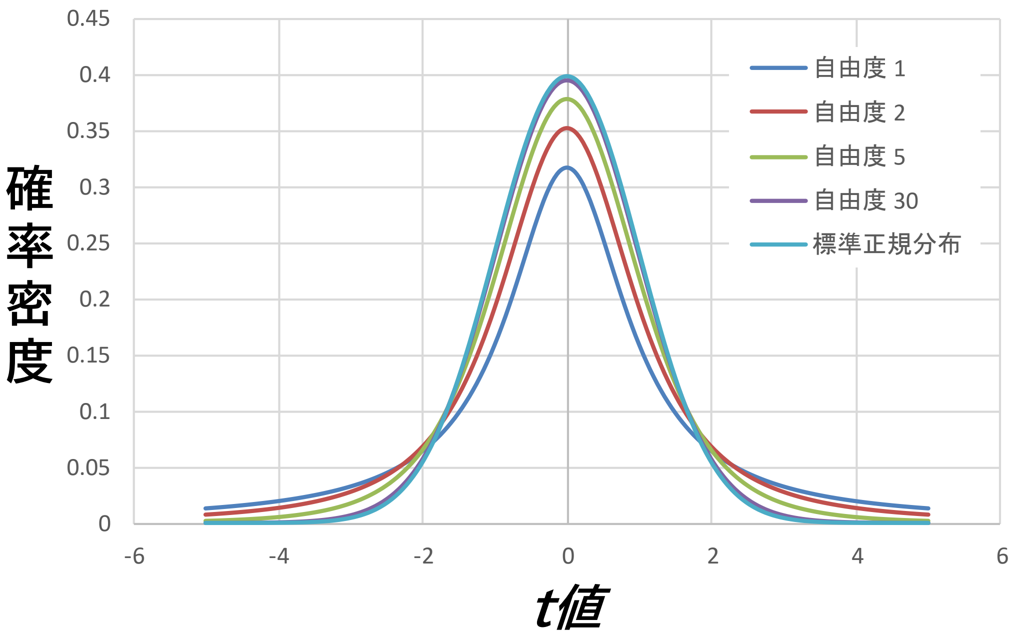
ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution

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）。

t 分布



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

t 分布表

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
22	1.321	1.717	2.074	2.508	2.819

出典
https://to-kei.net/distribution/t-distribution/t-table/

t 分布

性質：母平均μ、不偏分散v^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。

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

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

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、

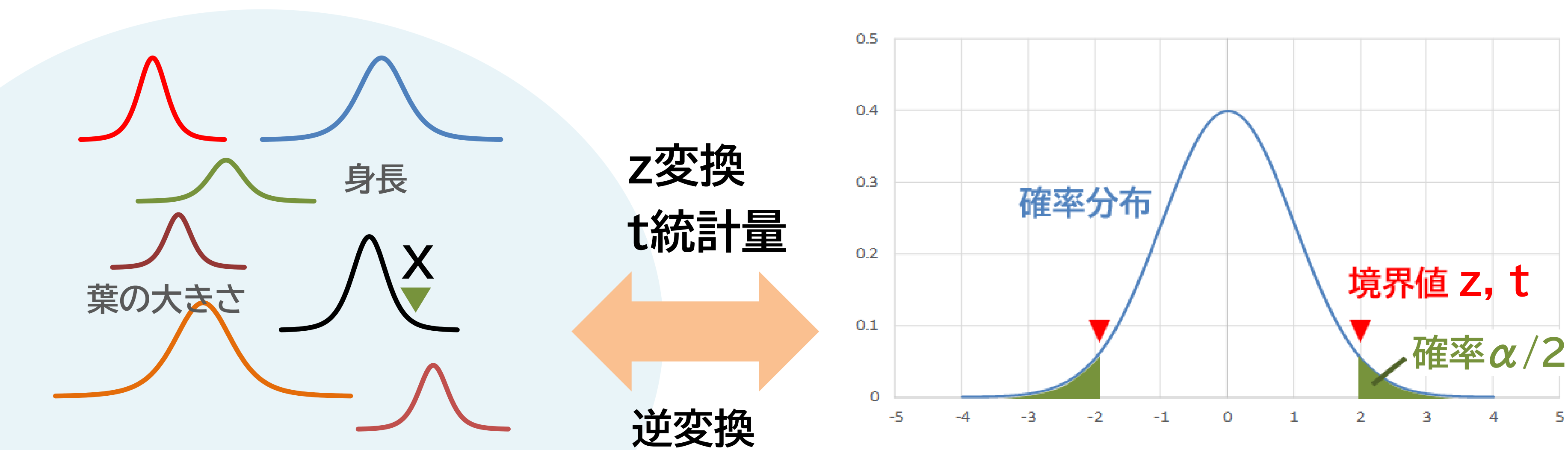
✓自由度＝n-1
✓確率α＝（100-a)/100

で求められる境界値。

まとめ

分布（確率密度関数）があれば、事象が起きる確率を推定できる！

使い方： 現実の問題を標準的な分布に当てはめ、
確率から境界値、境界値から確率を求める



現実の具体的な問題
（正規分布）

標準的な分布
（t分布、標準正規分布）

描いてみよう

- 標準正規分布
- *t*分布
- 裾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5回

2025年8月5日　神奈川工科大学

櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

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って動かし、授業に役立てます

プログラミング言語の種類

<u>汎用</u>

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

<u>専用</u>

R, Matlab（統計）,
Unity（ゲーム）

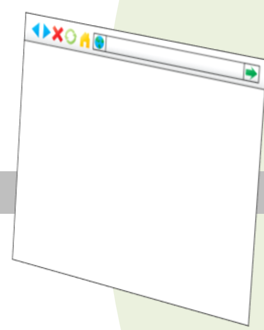
インターネット

JavaScript,
HTML,
CSS

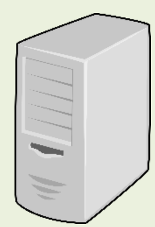
PHP,
Java,
Python,
Ruby, Perl,
Go, Rust,
JavaScript,
Shell Script
Kotlin

SQL

PC、スマホ



ウェブブラウザ



ウェブ
サーバー

データベース
サーバー

python™

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

  ✓数値計算、機械学習、ディープラーニング、生成AI

- 簡単（覚えやすい）
- はやっている（情報が多く、困ったときに解決しやすい）

若干のデメリット

- コードの書き方が、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戸惑う、、、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

具体的なやりたい目標を持つことが必須。
なんとなく「覚えなきゃ」と本読みから始めても、身につきません（時間の無駄）。

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

情報統計
第6回

2025年8月5日　神奈川工科大学

櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

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
- アンケートを作ってみる

などで、どんな課題発表にするか考えてみましょう

課題発表のしかた

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、検証できるように

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

情報統計
第7回

2025年8月6日　神奈川工科大学

櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

t 検定

この時間の内容

- *t* 検定
  - ✓ 新しい数式などは出てこない
  - ✓ 分布の使い方のお作法、用語だけ
- 検定で気を付けること、検定のいろいろ
  - ✓ 注意点や、関連するトピックを俯瞰
  - ✓ 広がりの中で、t検定の位置づけを確認

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

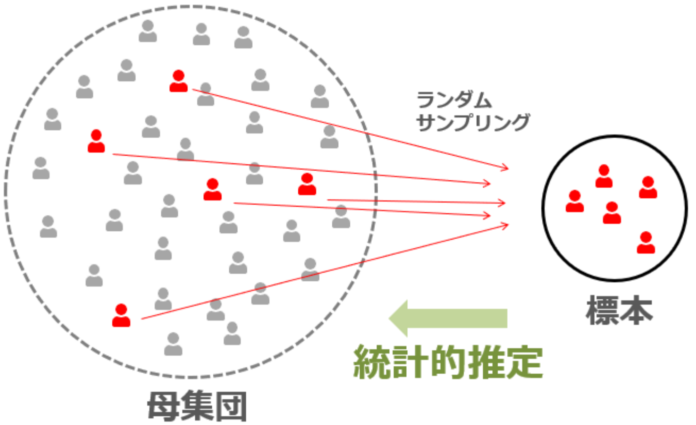
t 検定ツールの使い方を覚えるのではなく、Excelで自分で計算してみます

有意水準5％で
帰無仮説は棄却されました。
従って、＊＊＊です。

検定

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μ=95点

- 実際の点数を、複数回にわたりこっ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。

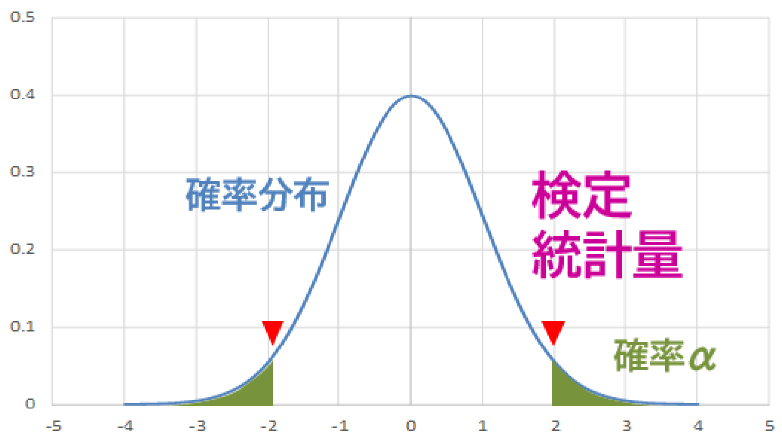
②検定統計量の計算

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。

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



②検定統計量の計算

標本平均	$\overline{x}$	91.5
不偏標本分散	v^2	10.7
母平均	μ	95

$$t = \frac{\overline{x} - \mu}{\frac{v}{\sqrt{n}}}$$

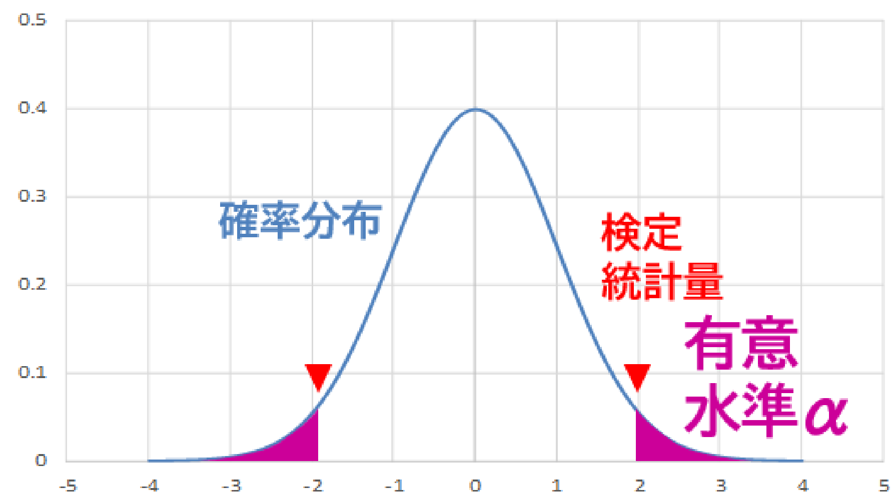
-2.62

③仮説採否の評価

有意水準 α を0.05とする

有意水準 α

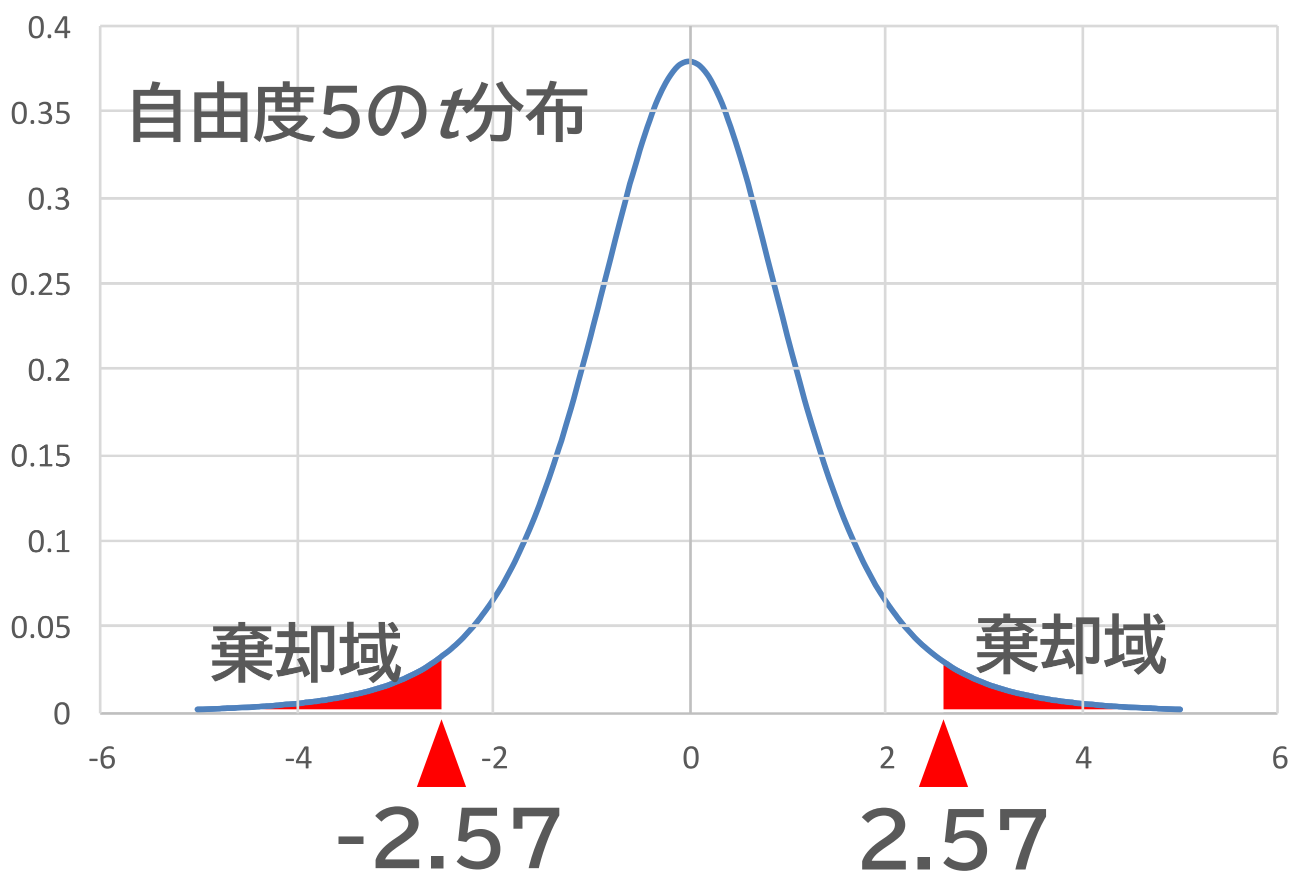
仮説を棄却するかどうかを決める基準の確率。これよりも小さい確率を持つ場合は、めったに起こらないことが起きていると考えられるため、帰無仮説(普通、変化がない)が棄却される。



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

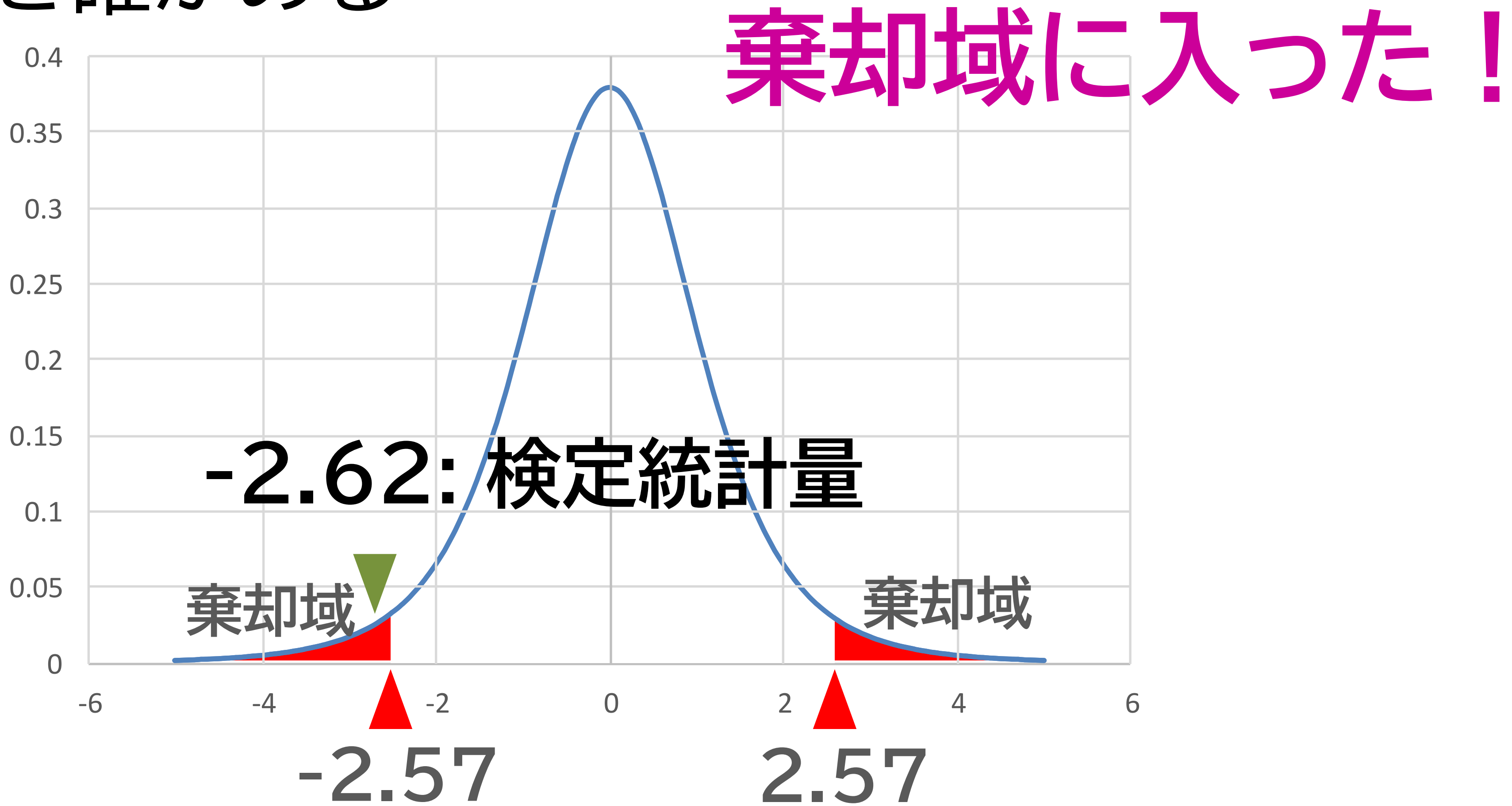
2.57



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



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…

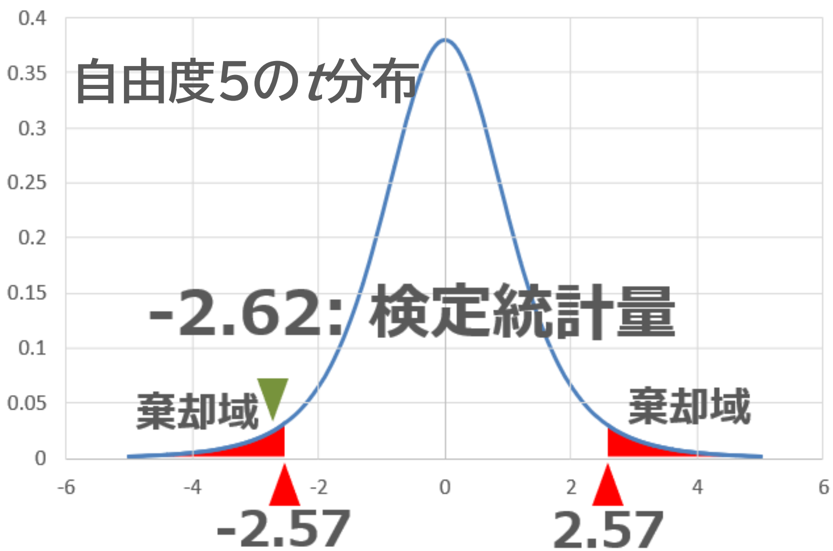
「帰無仮説が正しい」と安易に結論付けてはいけない。 ×

「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点ではないかもしれない。

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。

0.0235

※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、他より優勢である状態を「**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

検定と確率分布との関係

①仮説を立てる

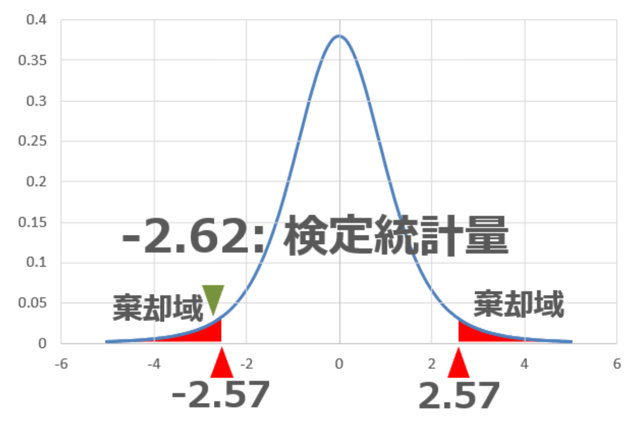
主張したいことを「対立仮説」に

②検定統計量を計算

- ✓ 適切な分布を選ぶ

  母集団の平均を推定する問題なら、t分布

- ✓ 分布に合った検定統計量を計算　　t値



③評価

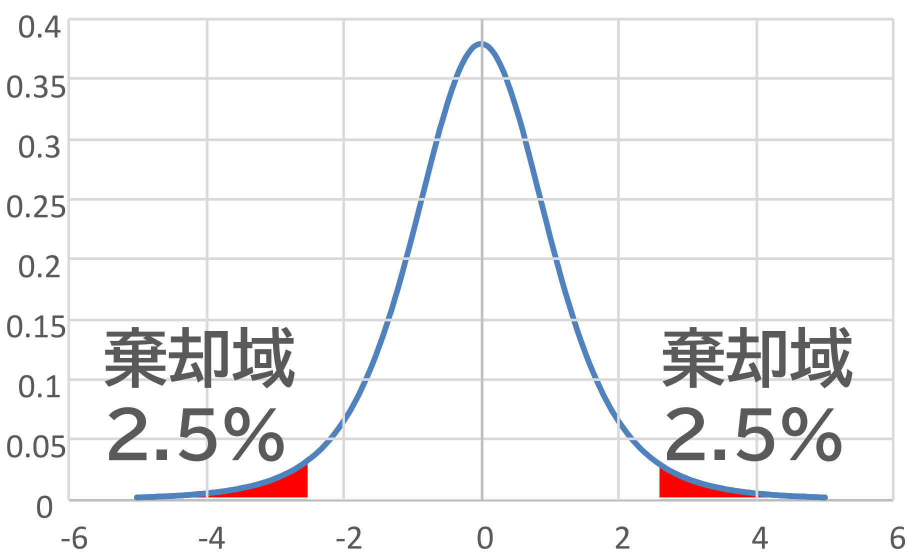
分布の境界値を超えているか？

検定の補足

✓両側検定と片側検定
✓t 検定のいろいろ

両側検定 と 片側検定

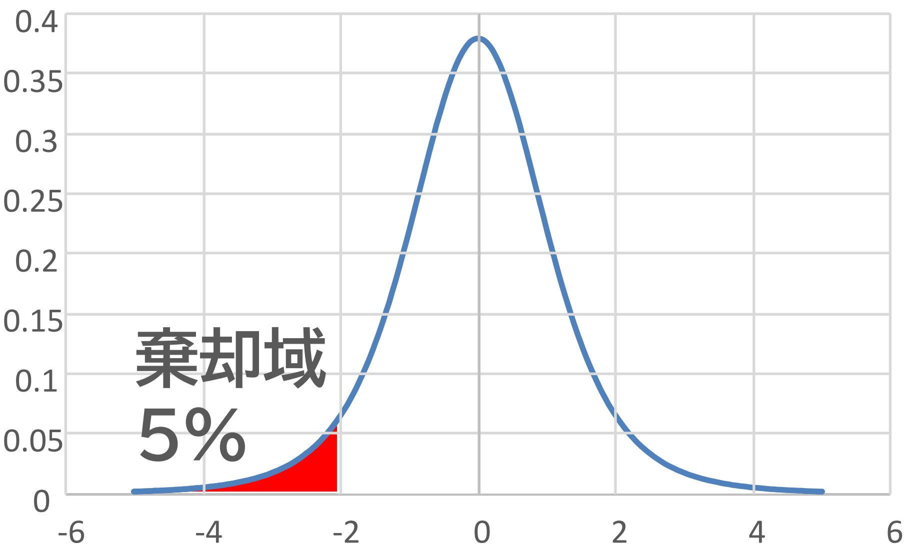
Aさんのカラオケ問題では、有意水準0.05を、その半分の0.025ずつに分け、*t* 分布の両側に割り当てて考えました

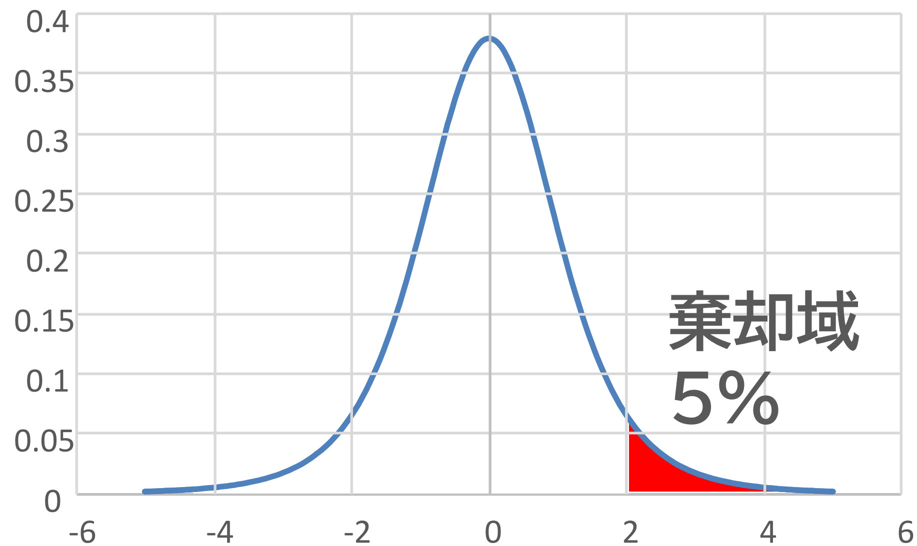


これは
両側検定
と呼ばれます

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もでき、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 n1, 不変標本分散 s1, 標本平均$\overline{x1}$

2群目：標本数 n2, 不変標本分散 s2, 標本平均$\overline{x2}$

プール分散 $$s^2 = \frac{(n1-1)\times s1^2 + (n2-1)\times s2^2}{n1+n2\ -2}$$

検定統計量 $$t = \frac{\overline{x1}-\overline{x2}}{s\sqrt{\frac{1}{n1}+\frac{1}{n2}}}$$

自由度： n1 + n2 -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　ウェルチの方法

1群目:標本数 n1, 不変標本分散 s1, 標本平均$\overline{x1}$

2群目:標本数 n2, 不変標本分散 s2, 標本平均$\overline{x2}$

検定統計量　$$t = \frac{\overline{x1} - \overline{x2}}{\sqrt{\frac{s1^2}{n1} + \frac{s2^2}{n2}}}$$

(近似)自由度 $$v \approx \frac{(\frac{s1^2}{n1} + \frac{s2^2}{n2})^2}{\frac{s1^4}{n1^2(n1-1)} + \frac{s2^4}{n2^2(n2-1)}}$$

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

- ✓ 検定の間違い
- ✓ 多重性の問題、FDR（偽発見率）
- ✓ p値 ＜ 0.05にとらわれるな！

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

有意水準 α

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。

よく似た表

スクリーニング検査

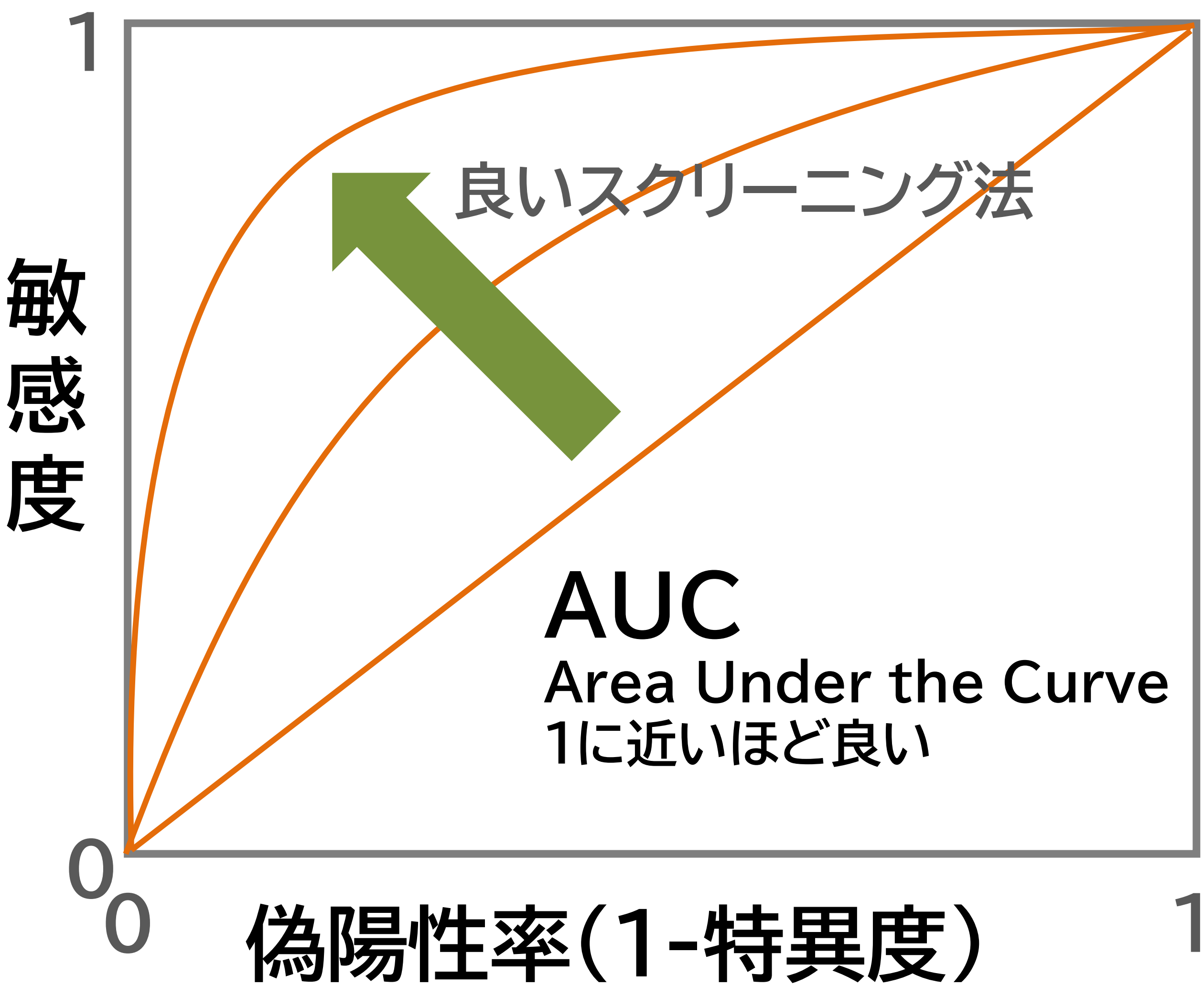
＊＊の スクリーニング		本当は	
		病気	健康
結果	陽性 （＋）	真陽性 True Positive 敏感度	偽陽性 False Positive 偽陽性度
	陰性 （-）	偽陰性 False Negative 偽陰性度	真陰性 True Negative 特異度

カットオフ値

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

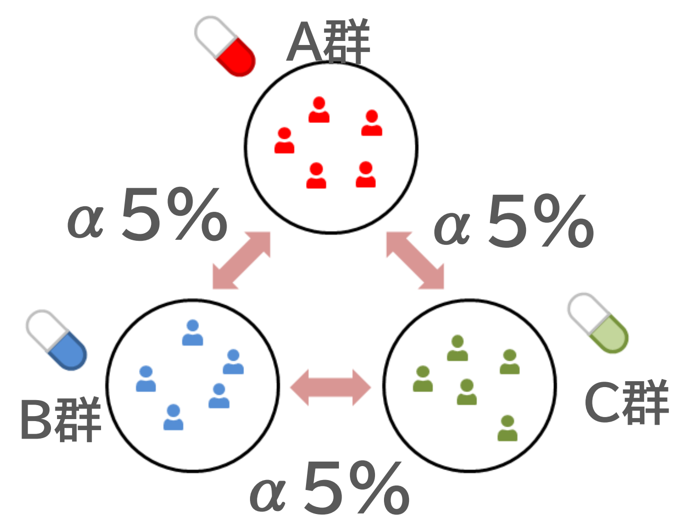
検定を繰り返すと、誤りが大きくなる

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α5%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 ＝ 0.14。



数打てば当たる状況！

対策

1. 多重比較のための検定法を使う
2. Bonferroniの補正
3. False Discovery Rateの調整

1.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

2. 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）
α ＝ 0.05で3回検定を繰り返す場合、
α’ ＝ 0.05 ÷ 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を引き下げるので、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。

3. 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(誤った姿)	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

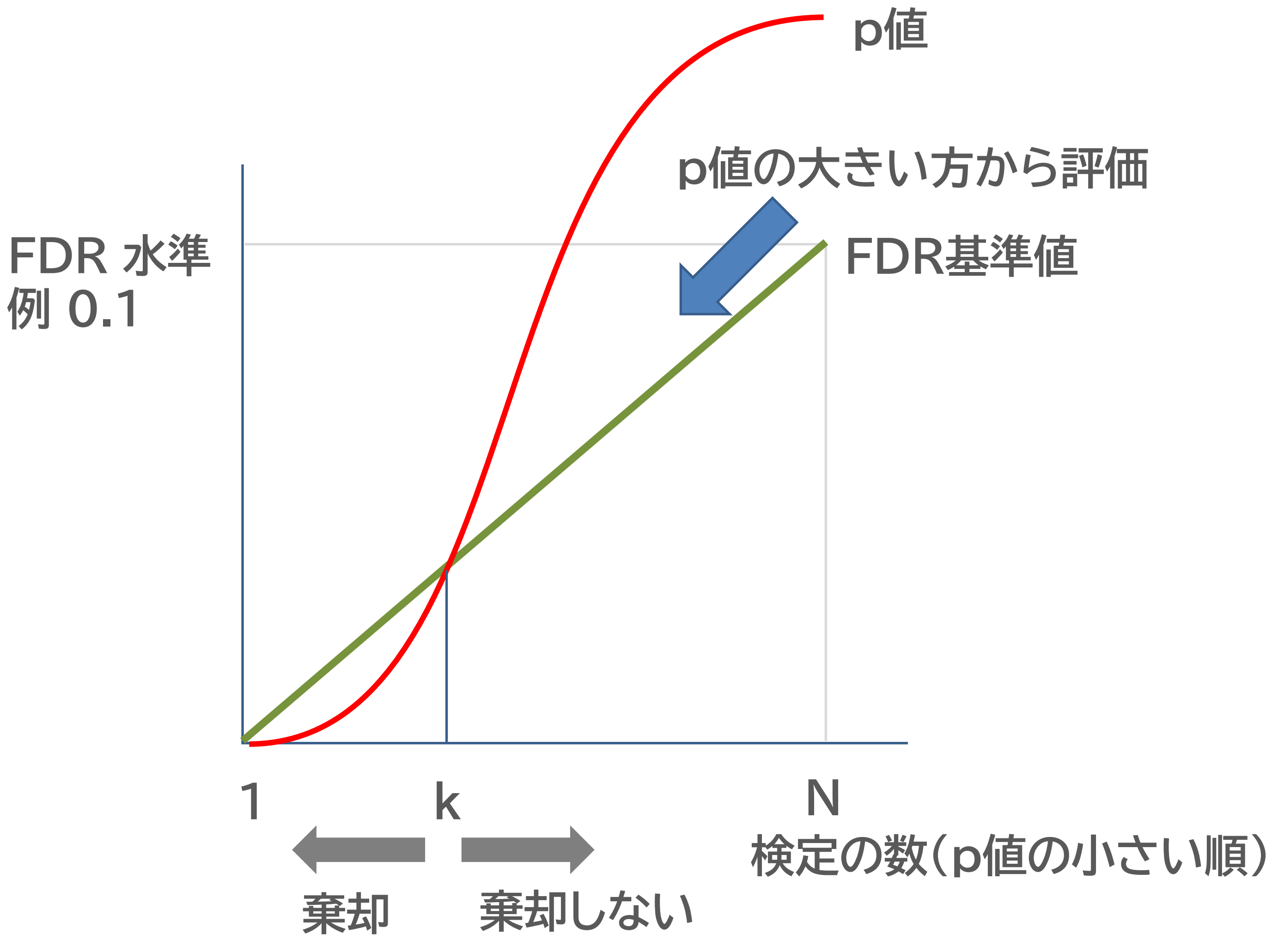
これを、一定水準例えば0.05にする方法

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 x 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

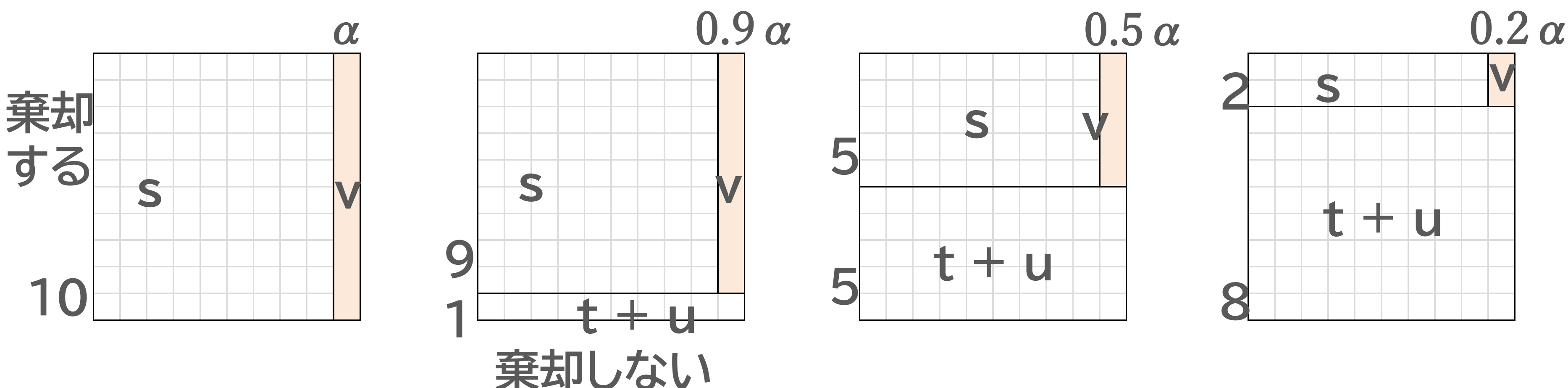


FDR調整のイメージ　その2

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 82016) 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）

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　坂巻顕太郎（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 勲（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（英国ラフバラ大学）
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　井桁正尭（兵庫医科大学）

- 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

やってはいけない不正行為

- t 検定で有意にならなかったので、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

そのほかの検定

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

- カイ二乗検定
- 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	合計
朝食を食べる			
朝食を食べない			
合計			

など

情報統計
第8回

2025年8月6日　神奈川工科大学

櫻井 望

公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

分布の仲間と
分散分析

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）

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： $F = \dfrac{v^2{}_a}{v^2{}_b}$

※$v^2{}_a$, $v^2{}_b$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

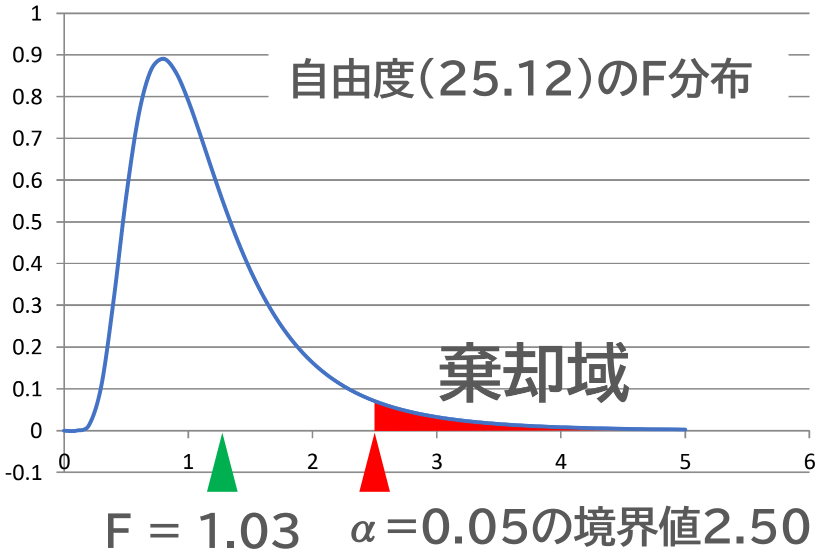
例）身長データの場合

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

$F = 23.63$(女性) / 23.02(男性) $= 1.03$

自由度(25, 12)の**F分布**から、F.DIST.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

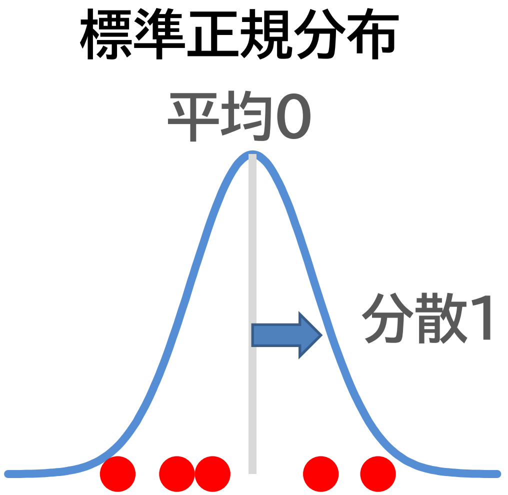
F検定で「分散に差がある」という結論を得たのち、2群の平均値に差があるかどうかをt検定すると、**「検定の多重性」**の問題にあたってしまう。

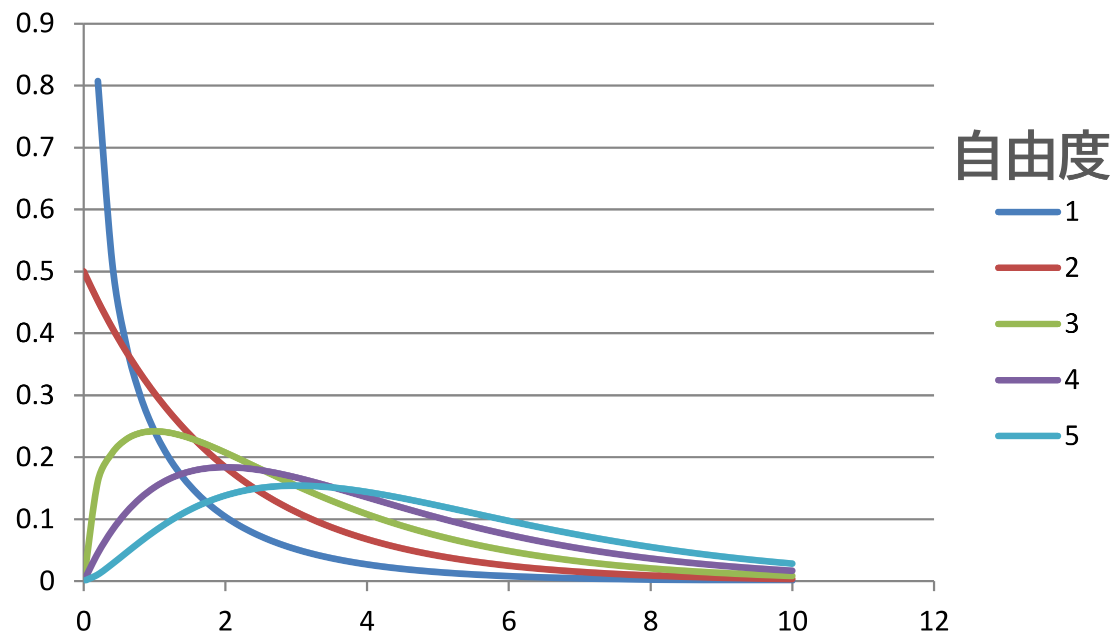
近年では、等分散かどうかに関係なく適用できるウェルチの検定を最初から行うことが望ましいという考えも出てきている。


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
複数の変数
（分散を考える）
標準正規分布
ひとつの変数

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	69	36
女性	21	24

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	ビールあんまり	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	ビールあんまり	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

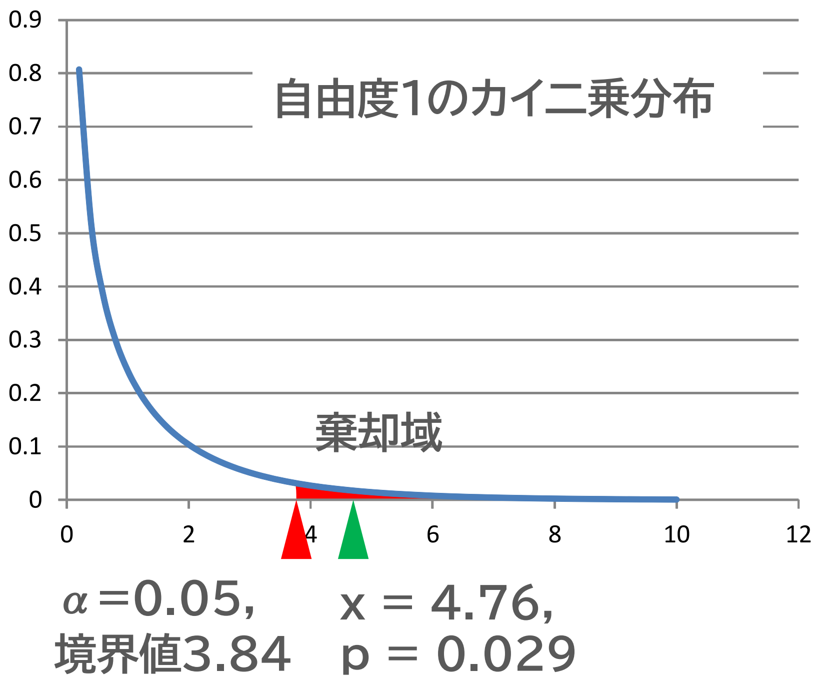
x = 0.57 + 0.86 + 1.33 + 2 = 4.76

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 < α)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

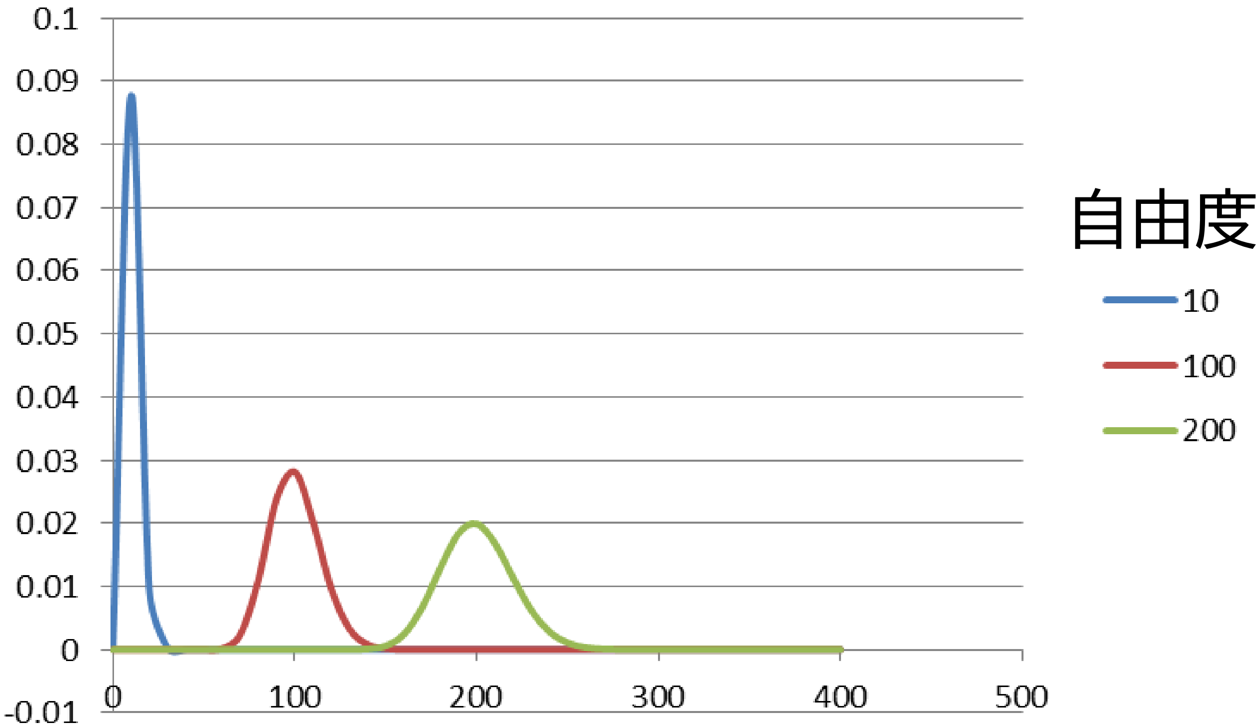
カイ二乗検定の留意点

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分布とカイ二乗分布の関係

自由度k_1のカイ二乗分布$\chi^2{}_1$

自由度k_2のカイ二乗分布$\chi^2{}_2$

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**$\sigma^2{}_1$と$\sigma^2{}_2$が等しいと仮定できる場合**は、

$$F = \frac{v^2{}_1}{v^2{}_2}$$

これをF検定で利用している！

F分布の活用（式の確認）

カイ二乗分布の性質

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

この式を変形すると、

不偏標本分散v²になっている！（k = n-1）

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

したがって、

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

分散分析

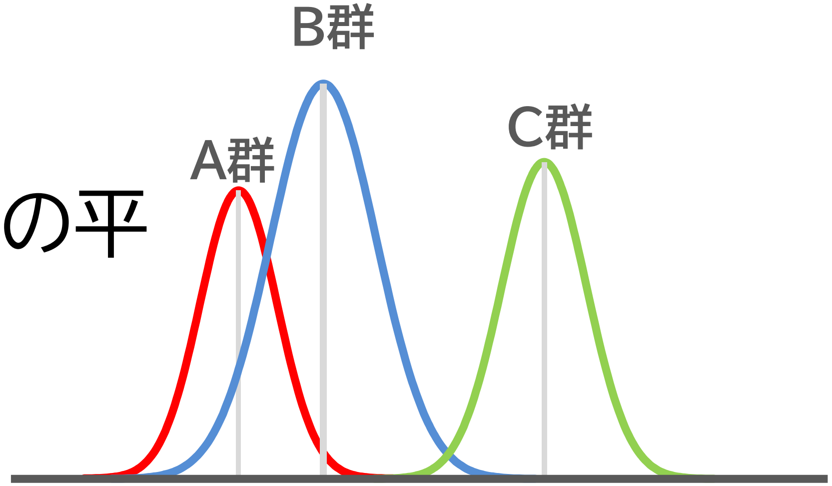
Analysis of Variance

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

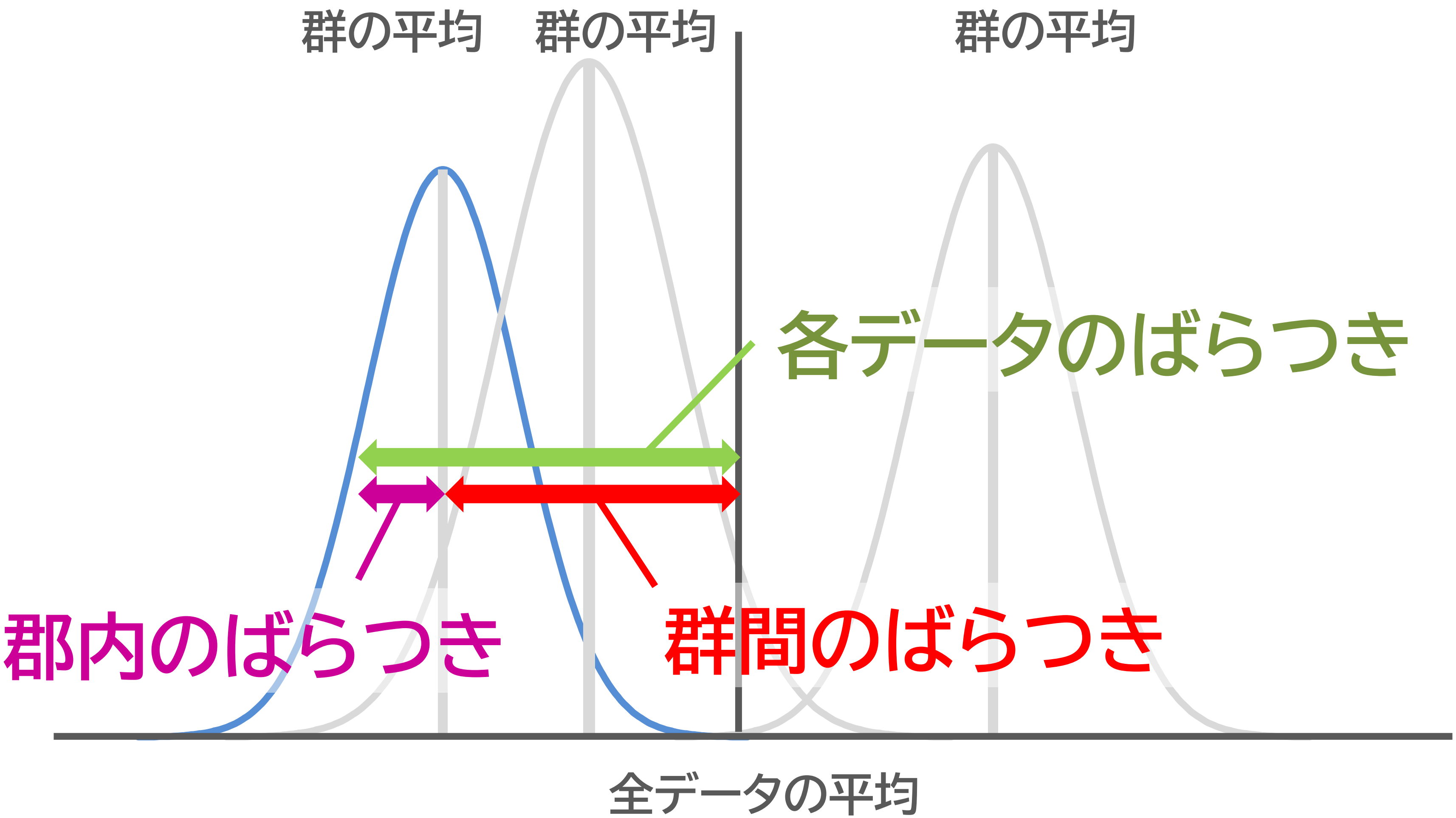
↓

どれが異なってい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

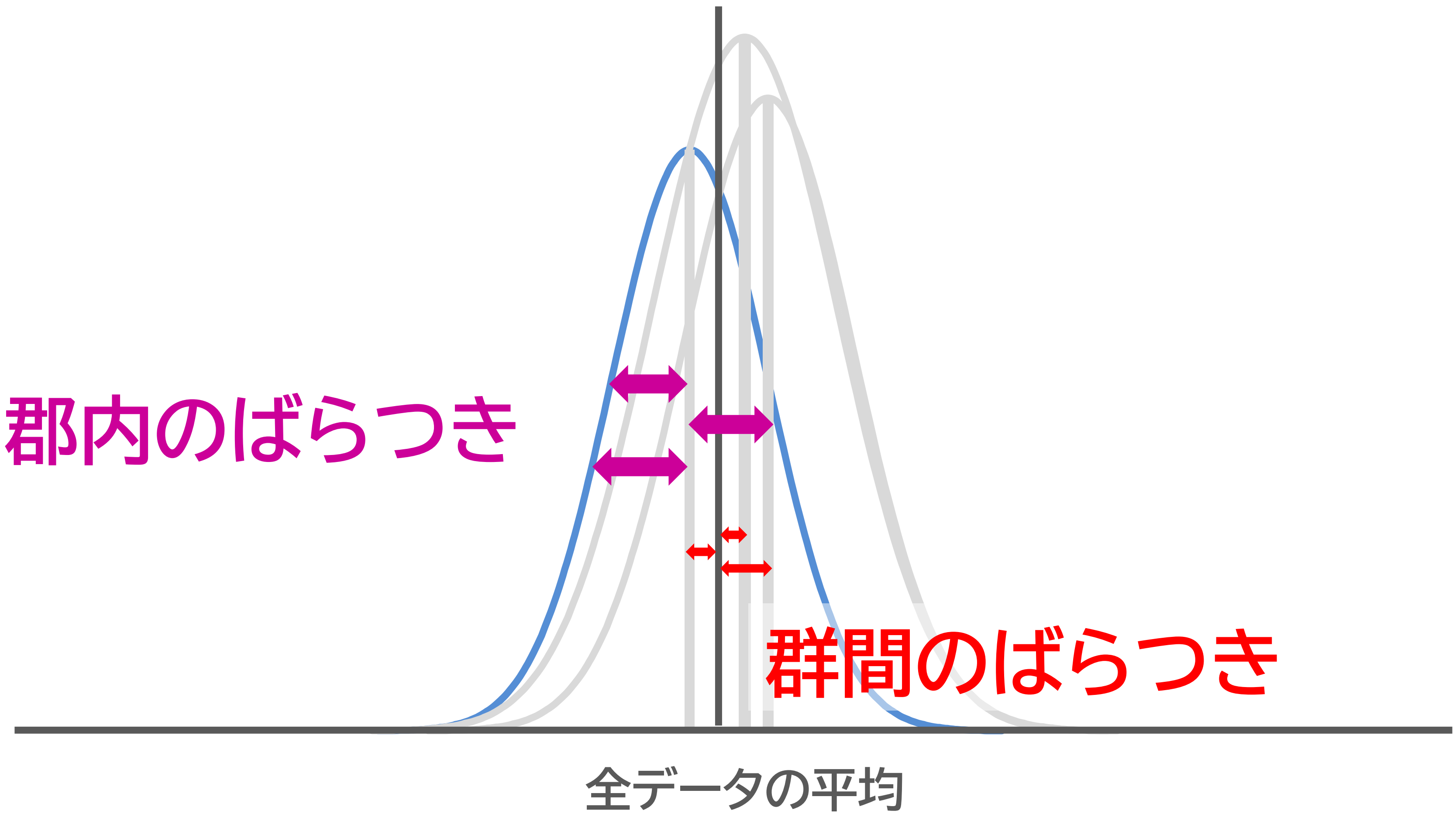
分散分析のイメージ

データのばらつきを、**群間**のばらつきと、**偶然により起こる郡内**のばらつきに分けて考える



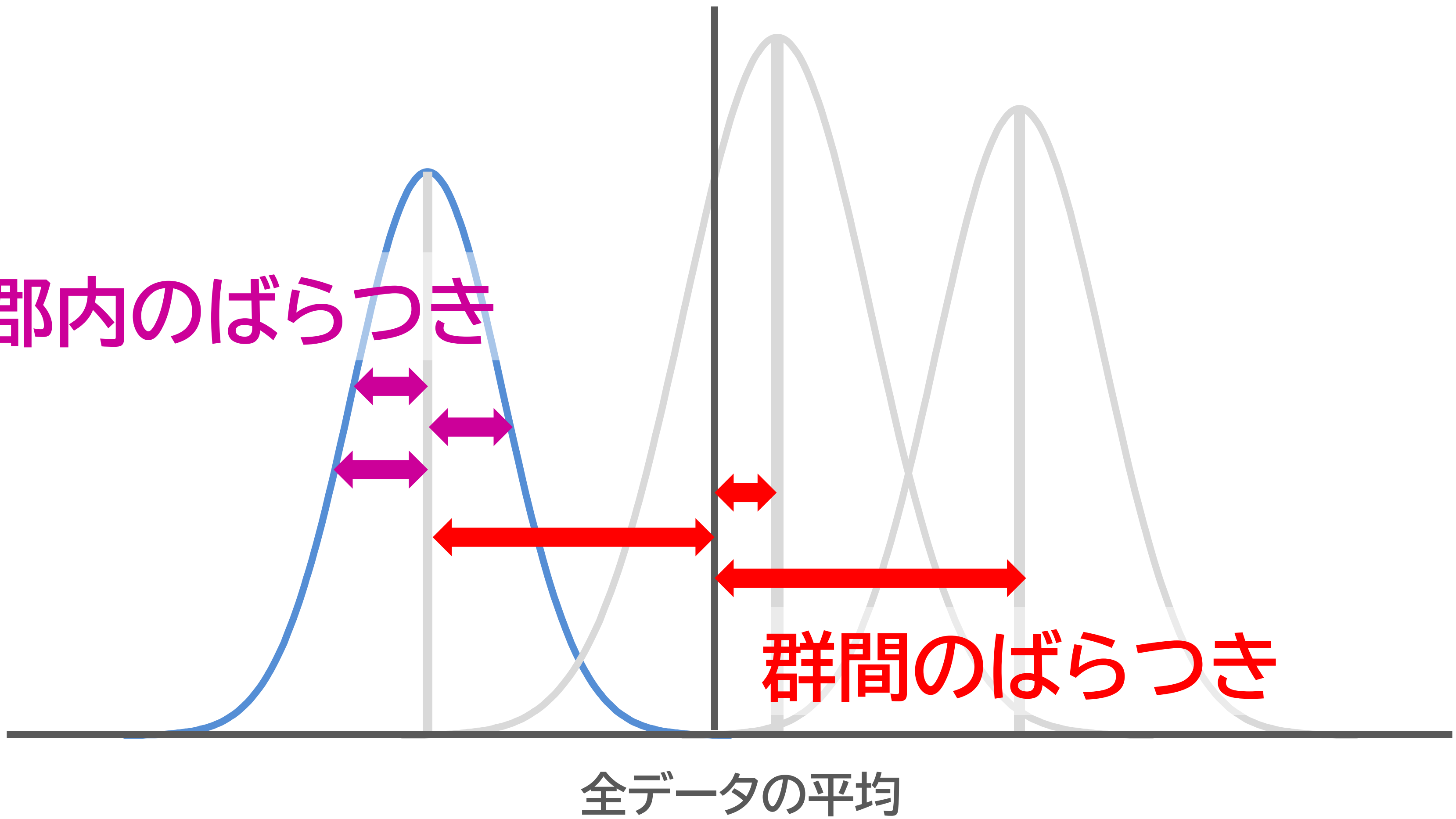
分散分析のイメージ

群の平均に差がなければ、
郡内のばらつき > 群間のばらつき



分散分析のイメージ

群の平均に差があるほど、
郡内のばらつき ＜ 群間のばらつき



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（群）	df（群） =群の数-1	V^2（群） =S（群）/df（群）	V^2（群）/V^2（残差）
群内 （残差）	S（残差）	df（残差） =全データ数-群の数	V^2 （残差） =S（残差）/df（残差）	
全体	S（全体）	df（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	

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
　⑪ ＝ ⑫ ＋ ⑬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

⑱⑰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
⑲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

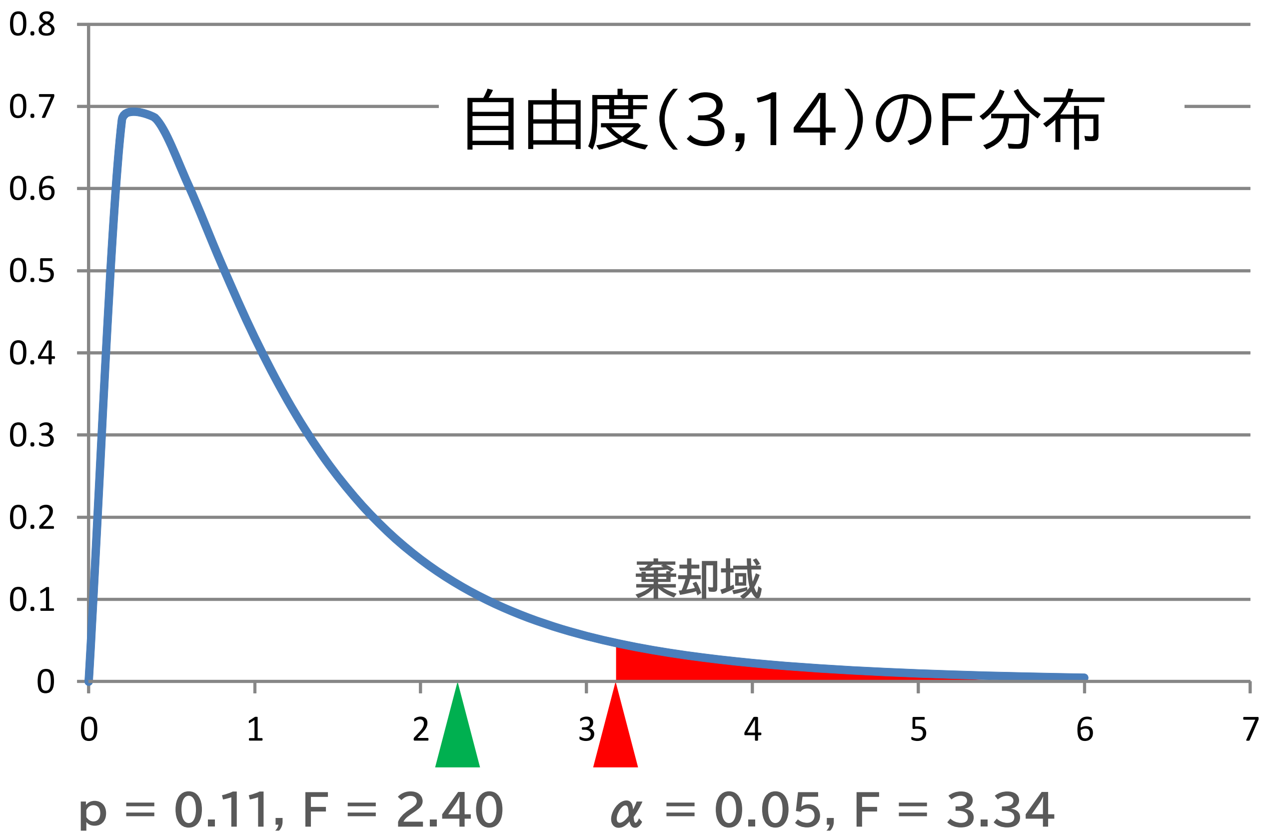
⑳F.DIST関数を用いて当該自由度のF分布を描く

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

今回やったもの

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

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9回

2025年8月6日　神奈川工科大学

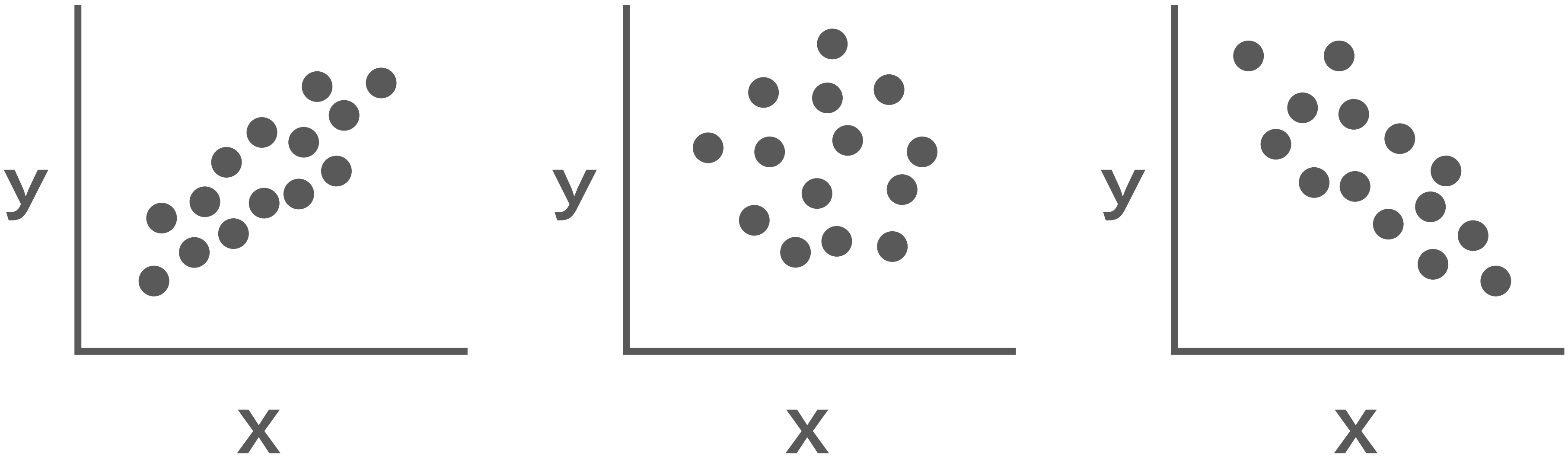
櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

相関

復習

散布図

二つの変数の間の関係性を見える化する手法

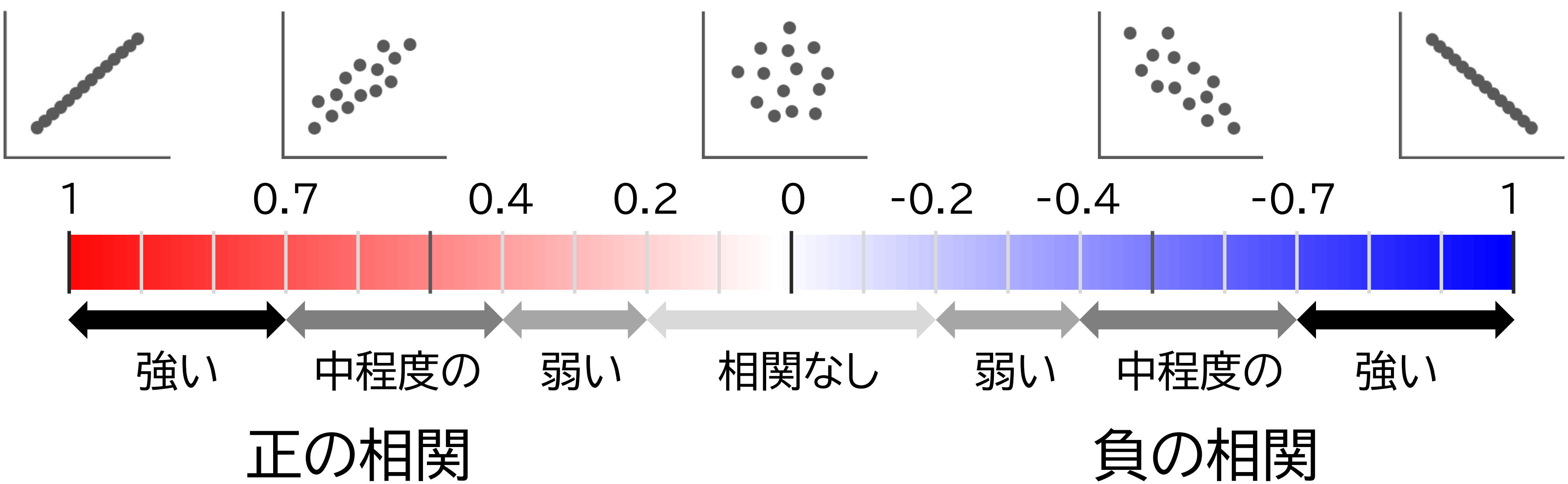


正の相関がある　　相関がない　　負の相関がある

相関係数

復習

- 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



※数字の区切りはあくまで目安

- Excelでは**PEARSON関数**で計算できる

相関係数を手で計算する

ピアソンの積率相関係数

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

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

sxy: xとyの**共分散**
sx: xの標準偏差
sy: 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

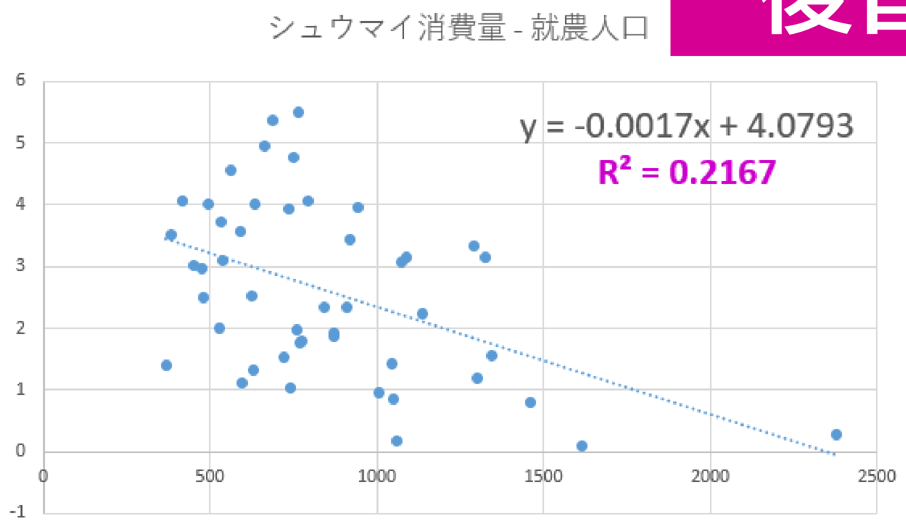
検定統計量：　$t = \dfrac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$

※|r|はrの絶対値
　エクセルではABS関数
　で計算できる

自由度：n-2

復習

注意点



回帰曲線のR^2値は、相関係数ではない！

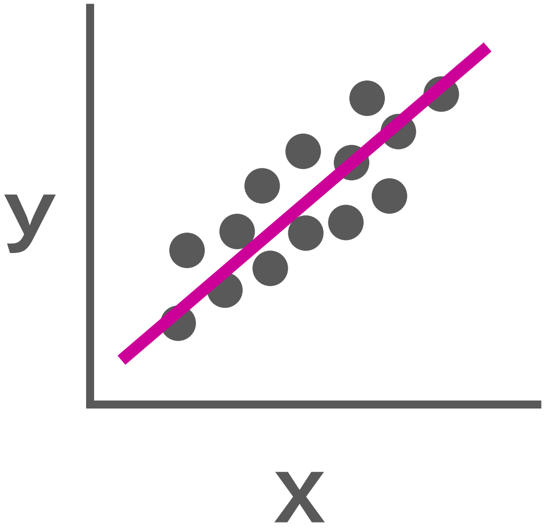
R^2値は、回帰曲線への当てはまり度を示すもので、「決定係数」と呼ばれます。

Excelで、原点を通らない直線近似をした場合は、R^2値はピアソン相関係数の二乗に当たります。相関係数が-1～1の値を取るのに対し、R^2値は0～1の値を取ります。負の相関であっても、R^2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。

復習

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

そのほかの相関係数

- スピアマンの順位相関係数
  - ✓ 少数の極端な外れ値の影響をおさえ、全体的な傾向を重視したい場合に有利
  - ✓ 正規分布を仮定していなくてもよい
- コサイン相関係数
  - ✓ 数値そのものではなく、パターンを重視したい場合に有利

スピアマンの順位相関係数

	評価点	
ワインの銘柄	Aさん	Bさん
A	100	60
B	90	58
C	85	100
D	80	55
E	75	54
F	70	53
G	65	52
H	60	0

順位に
置き換え

順位点	
Aさん	Bさん
1	2
2	3
3	1
4	4
5	5
6	6
7	7
8	8

ピアソン相関係数
0.599

スピアマン順位相関係数
0.929

Excelでは、数値などの値を順位に置き換えたのち、PEARSON関数（CORREL関数）で計算できる。（計算例は、補足資料 todoran.xlsxに）

きき酒大会

1回目

- 7種ほどの異なるお酒が出される(①～⑦など)
- 味、香り、色などの好みで、順位をつける

2回目

- 同じお酒が、1回目とは異なるラベルで出される(A～Gなど)
- 同様に、順位をつける

採点

1回目と2回目の順位の差の少なさを評価する

各お酒で順位の差を計算し、その二乗和を得点として、点数の低い人(完全一致で0点)を勝者とする。

- スピアマンの順位相関
- 分散(ばらつきの大きさ)　の考え方を利用

多変量解析

学習目標

主成分分析について

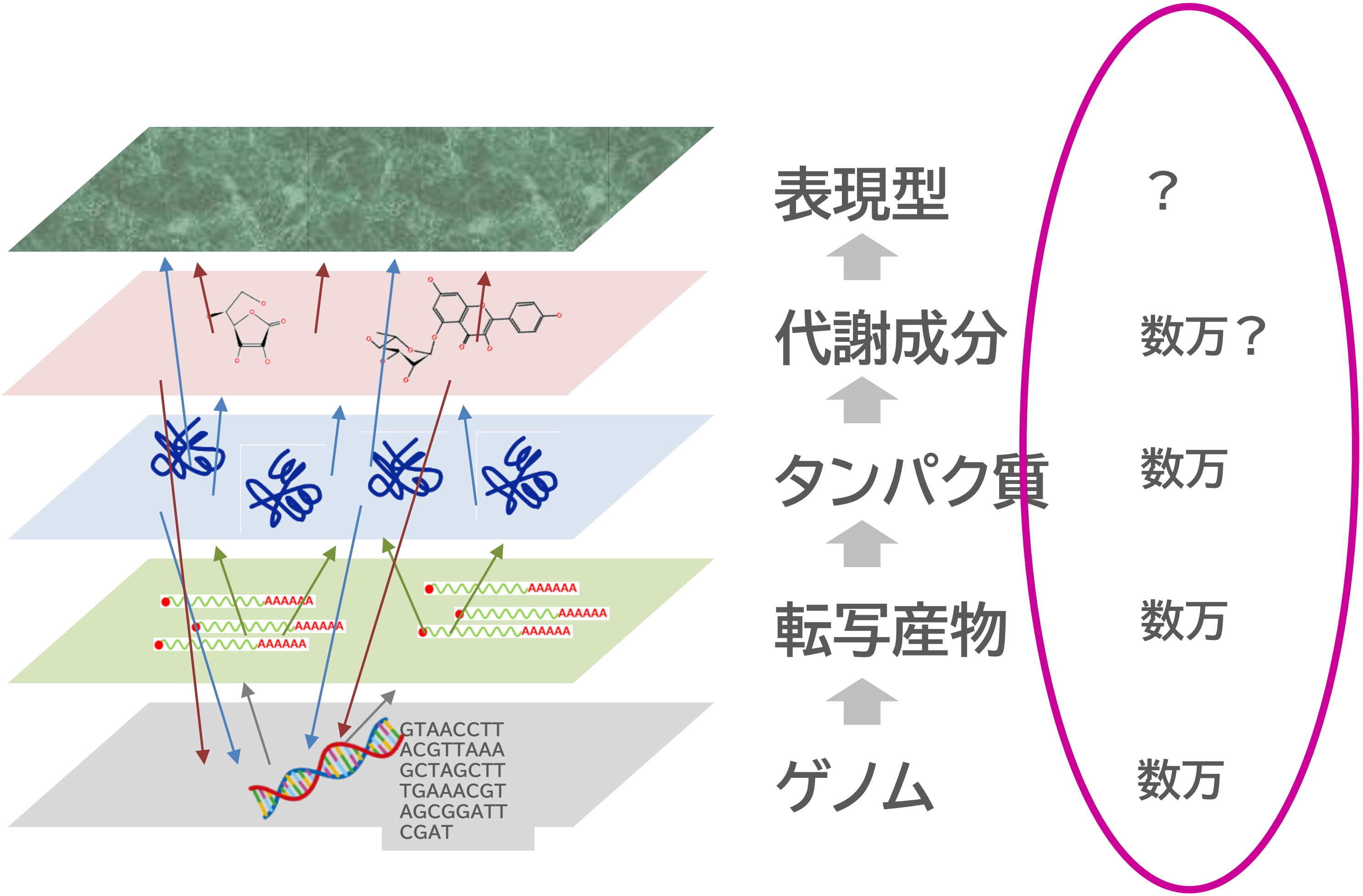
- 概念を理解する
- 結果の解釈の仕方を理解する

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　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}

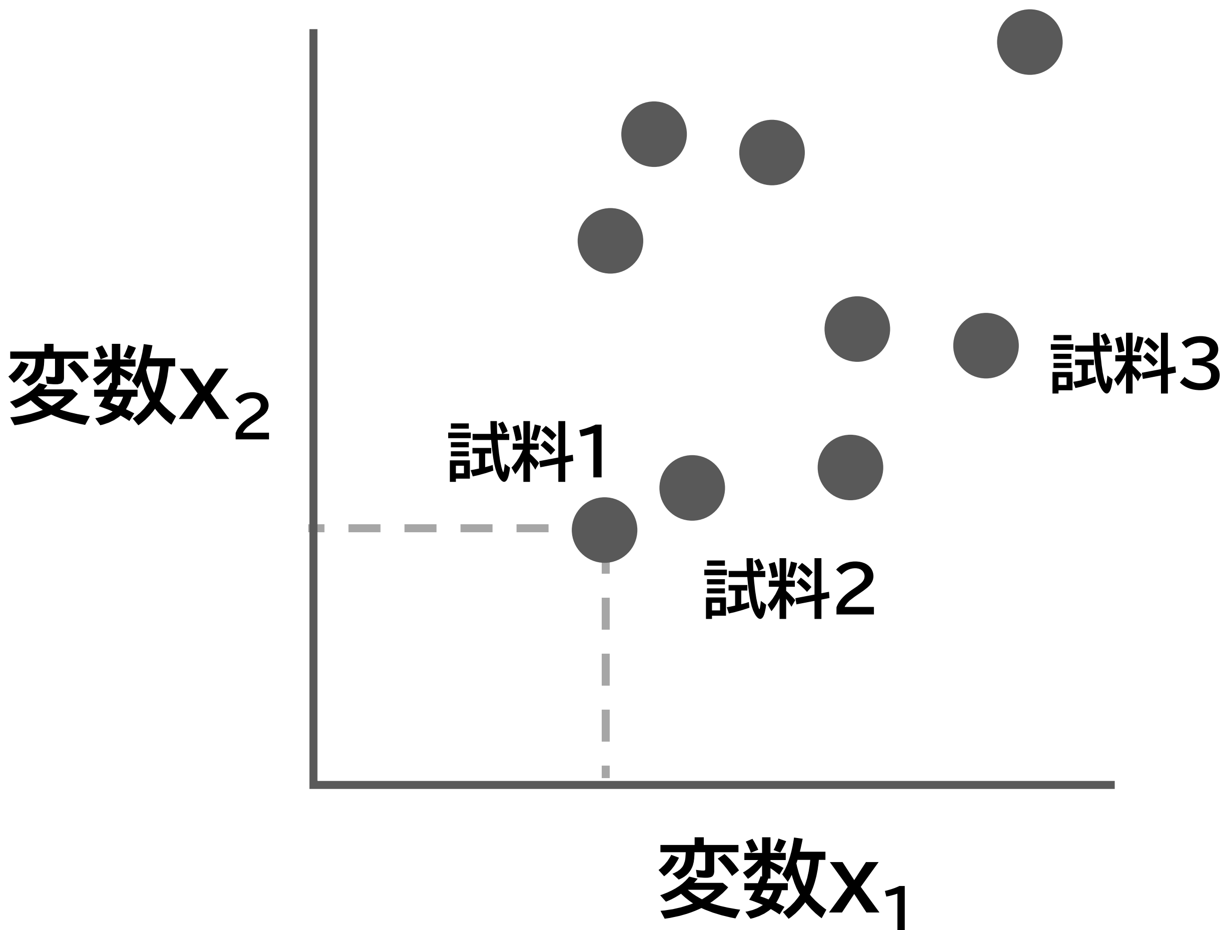
遺伝子など

説明変数, 観測変数

遺伝子発現量など

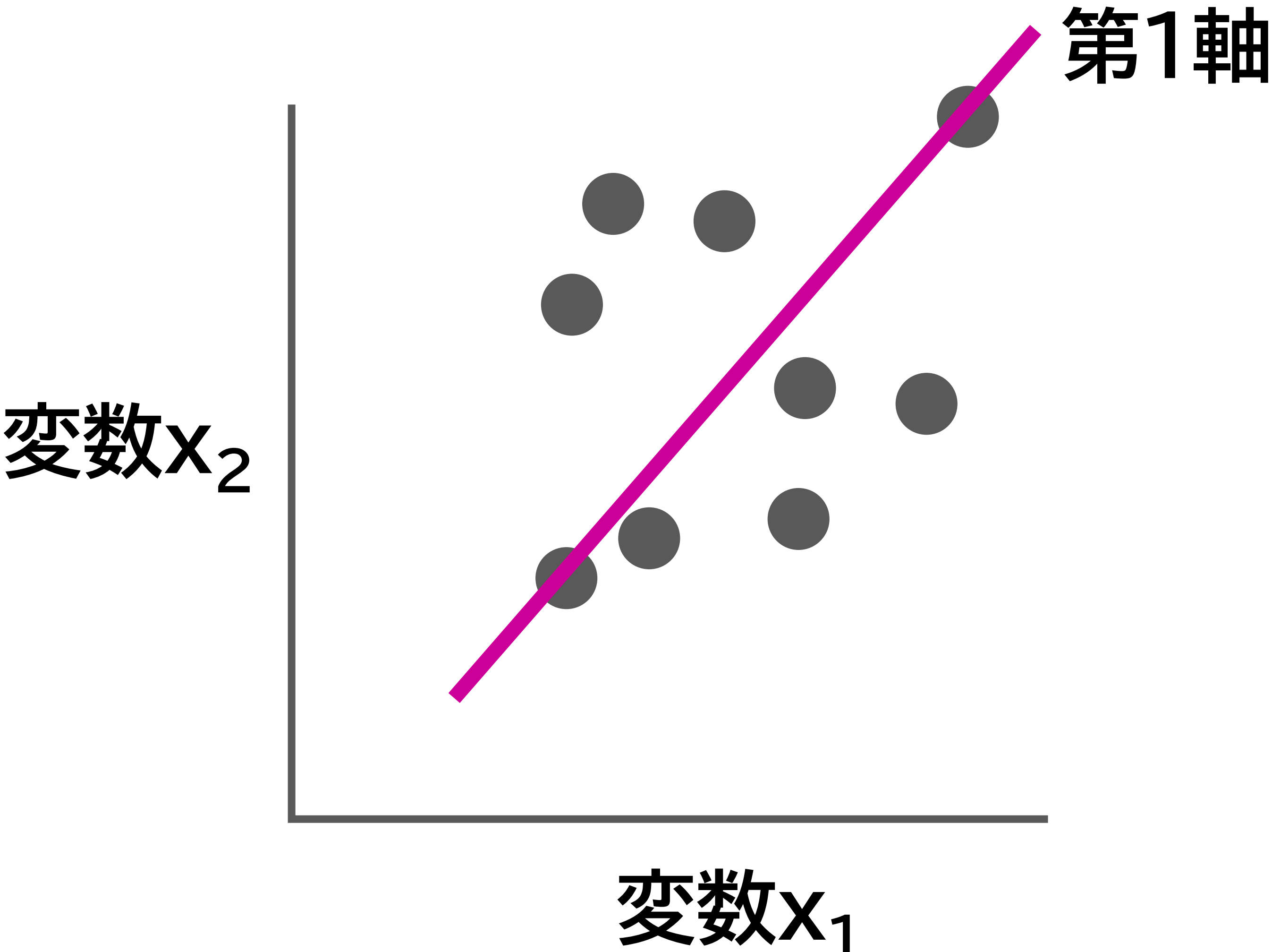
主成分分析のイメージ

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



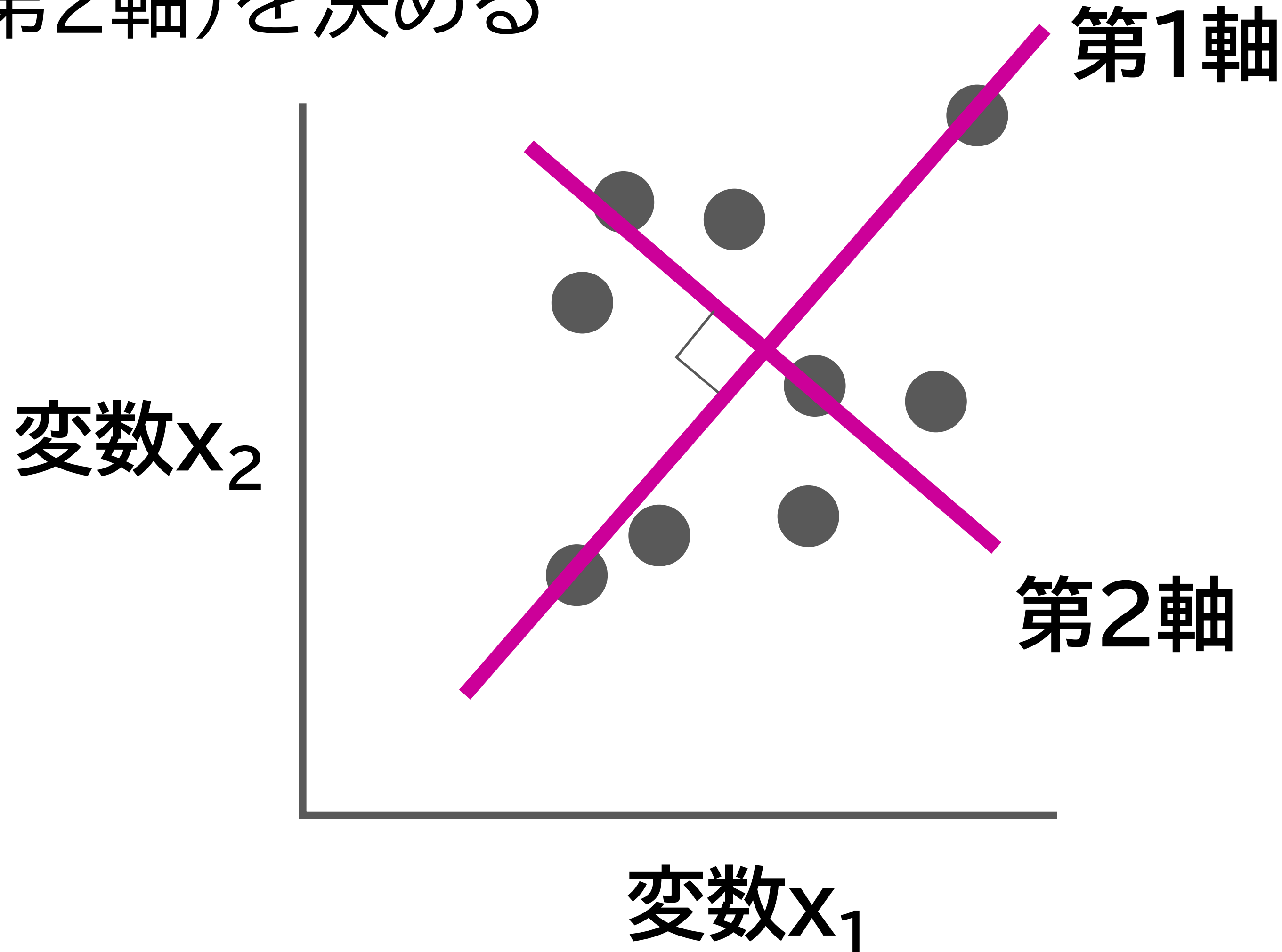
主成分分析のイメージ

②一番分散の大きい軸(第1軸)決める



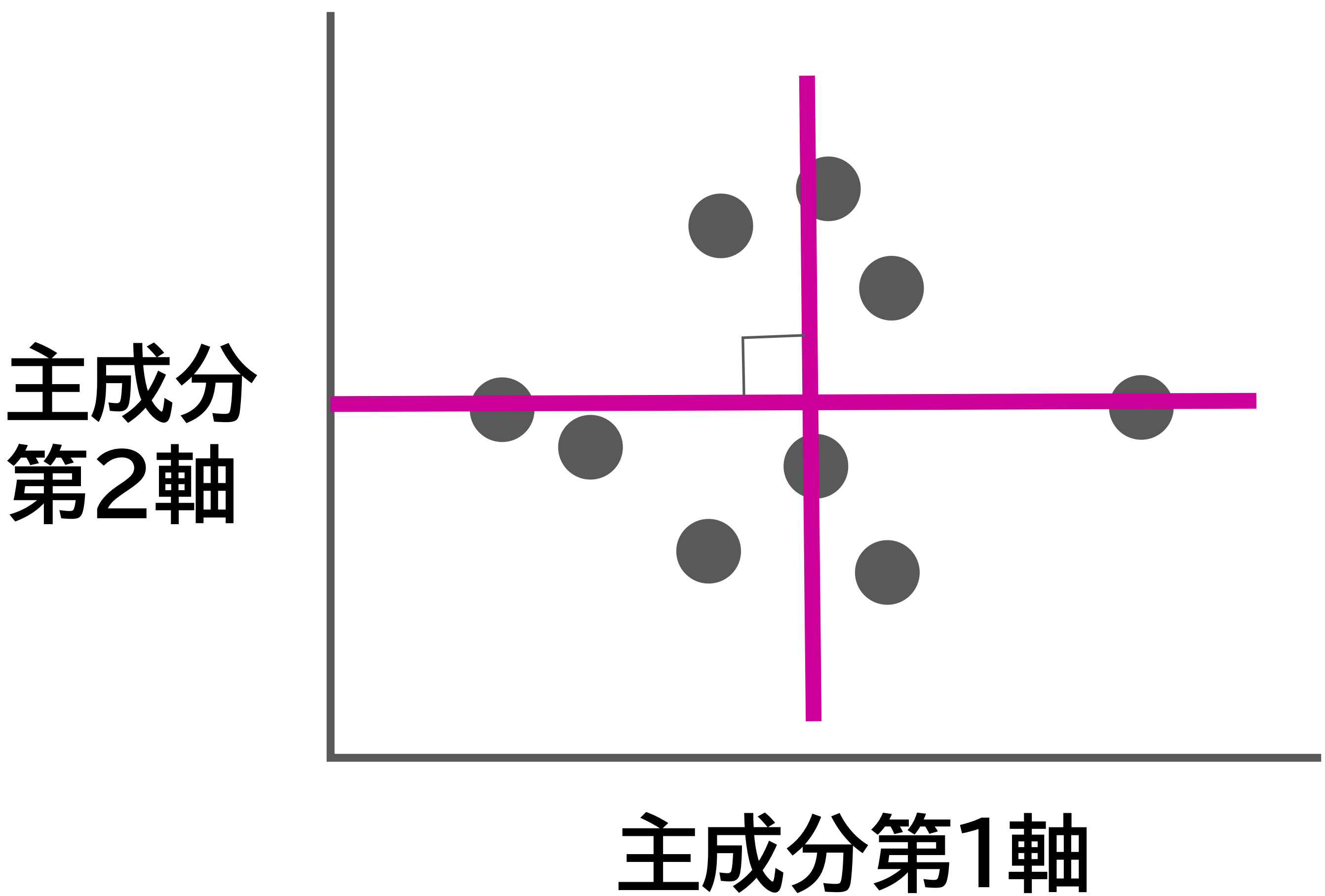
主成分分析のイメージ

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



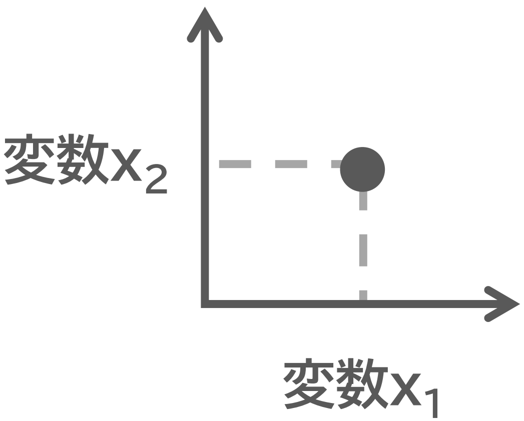
主成分分析のイメージ

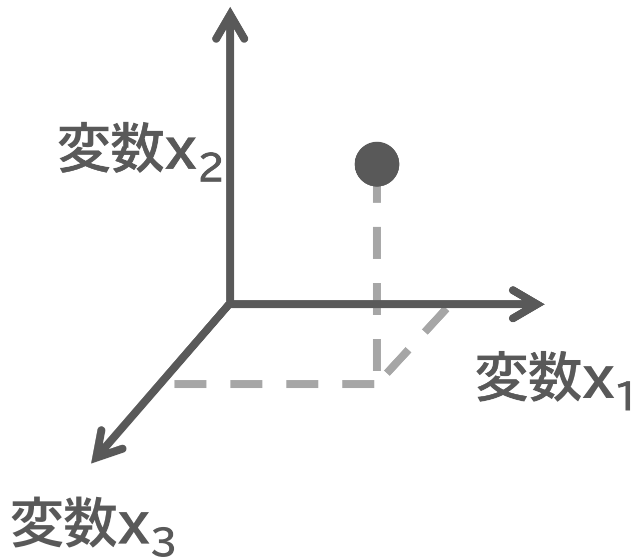
④第1軸がx軸、第2軸がy軸になるように、図を回転させた新たな図を作る



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





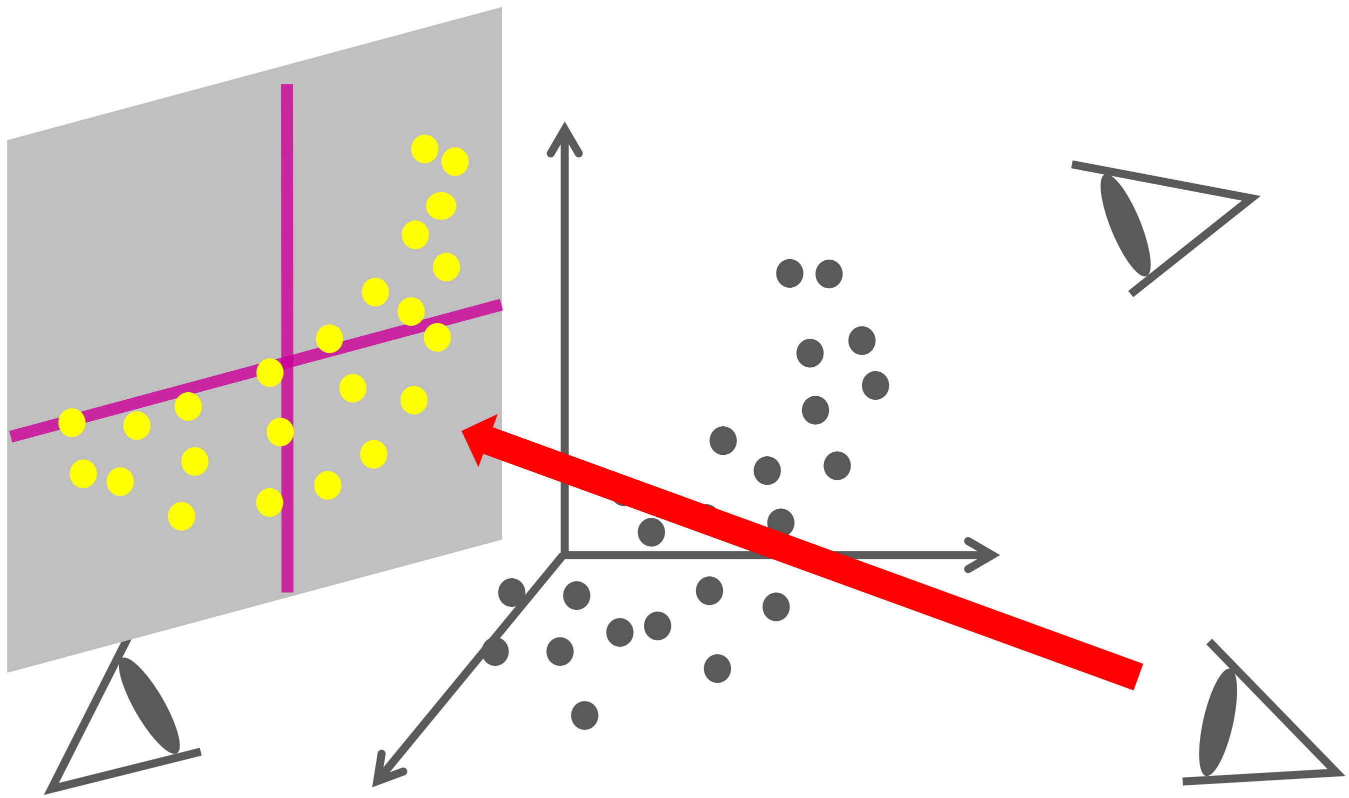
変数2個
2次元

変数3個
3次元

変数m個
m次元

主成分分析のイメージ

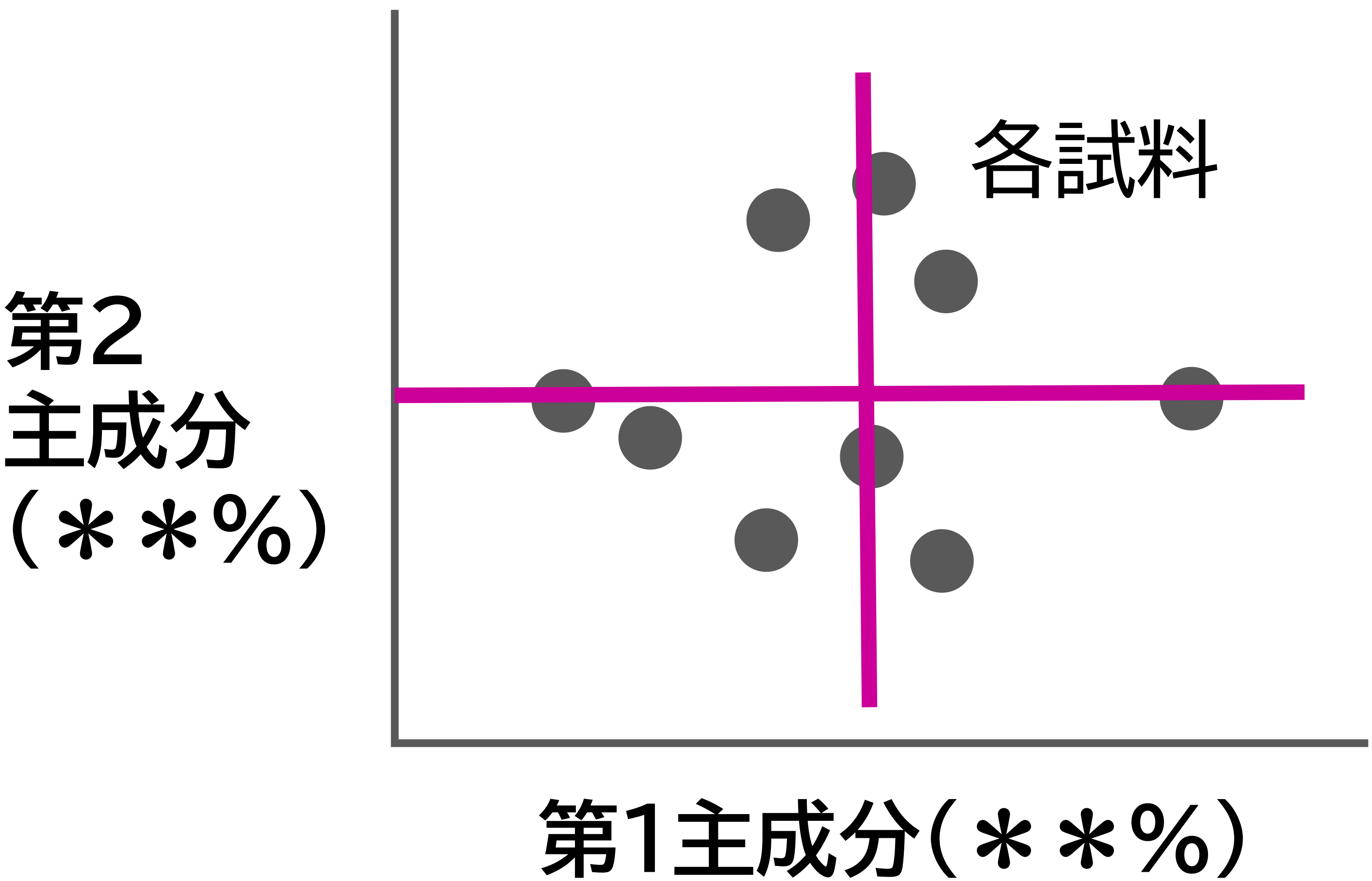
試料間の違い(特徴)が一番はっきりと見える方向から見た図が描ける



スコアプロット

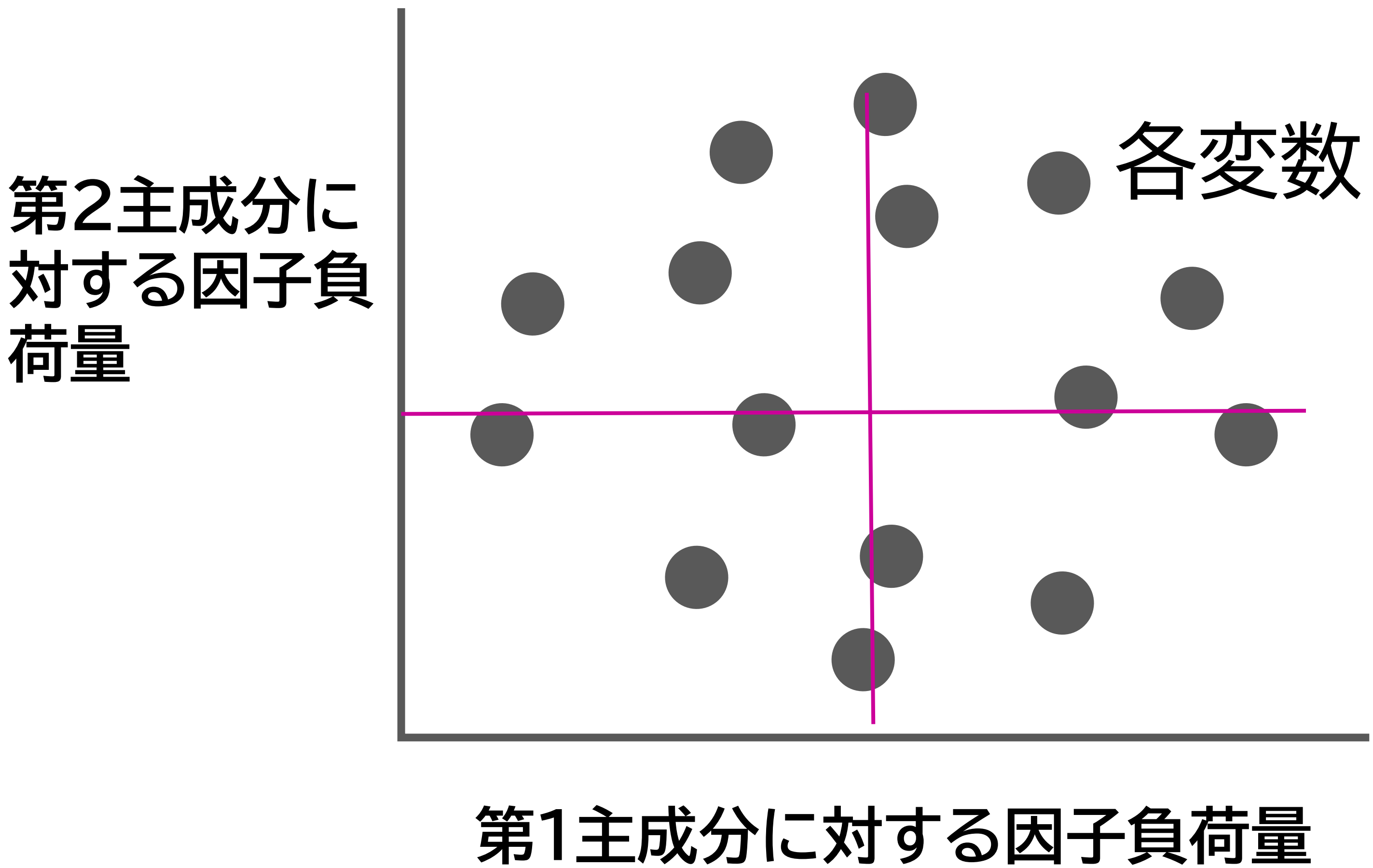
主成分軸に各試料を投影しなおした図

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。



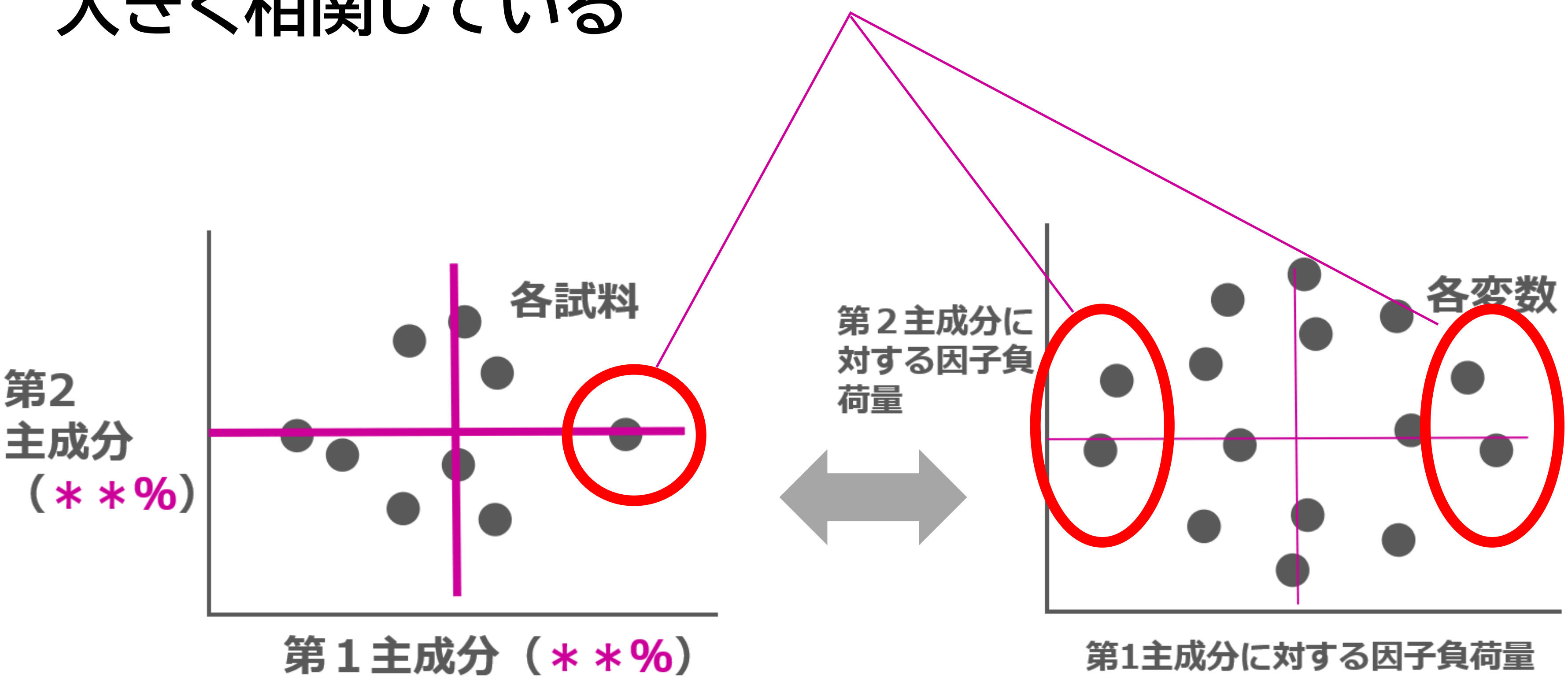
ローディングプロット

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



スコアプロット　　ローディングプロット

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

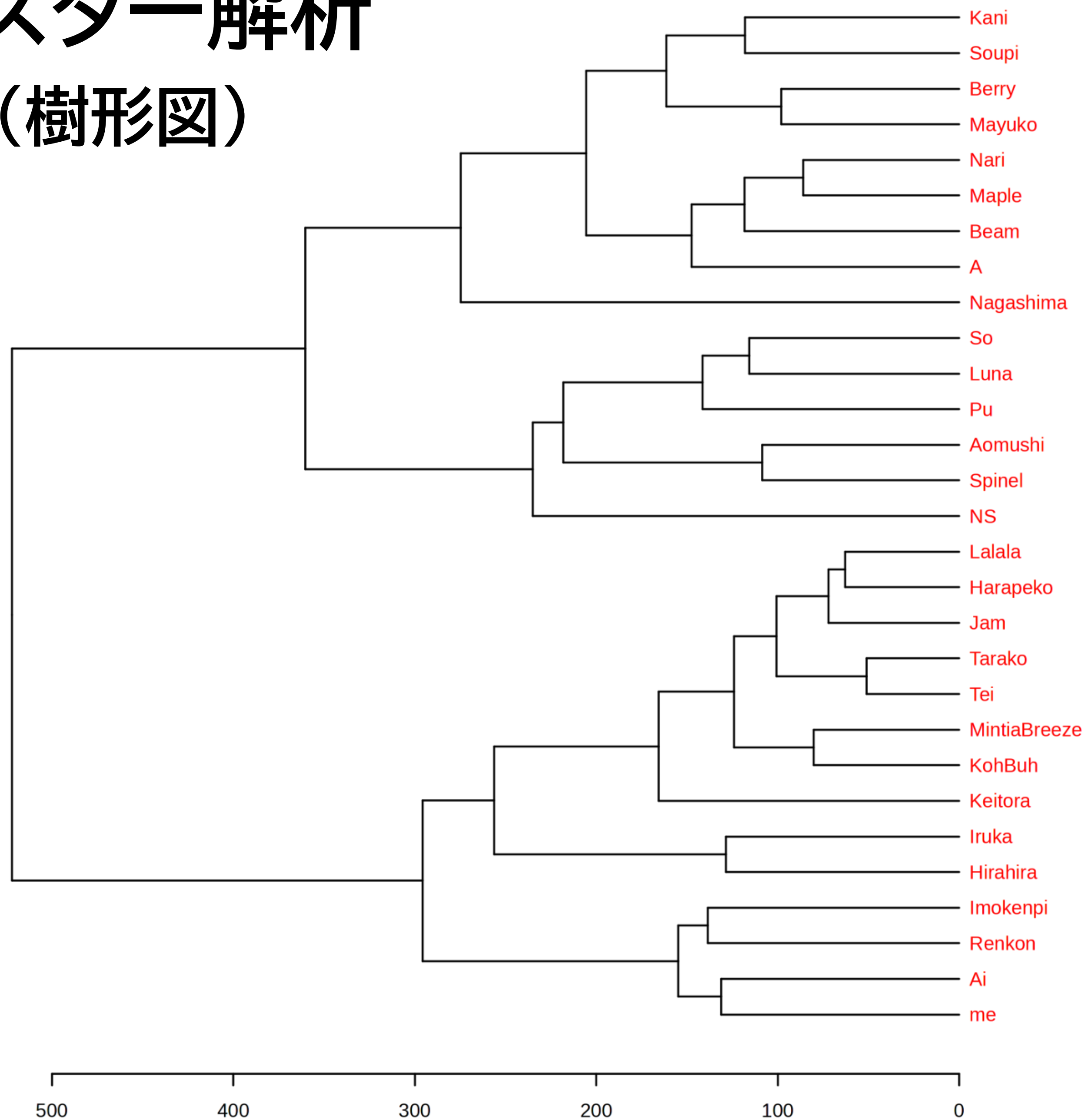
- 判別、分類、予測

  判別分析、PLS、PLS-DA、重回帰分析

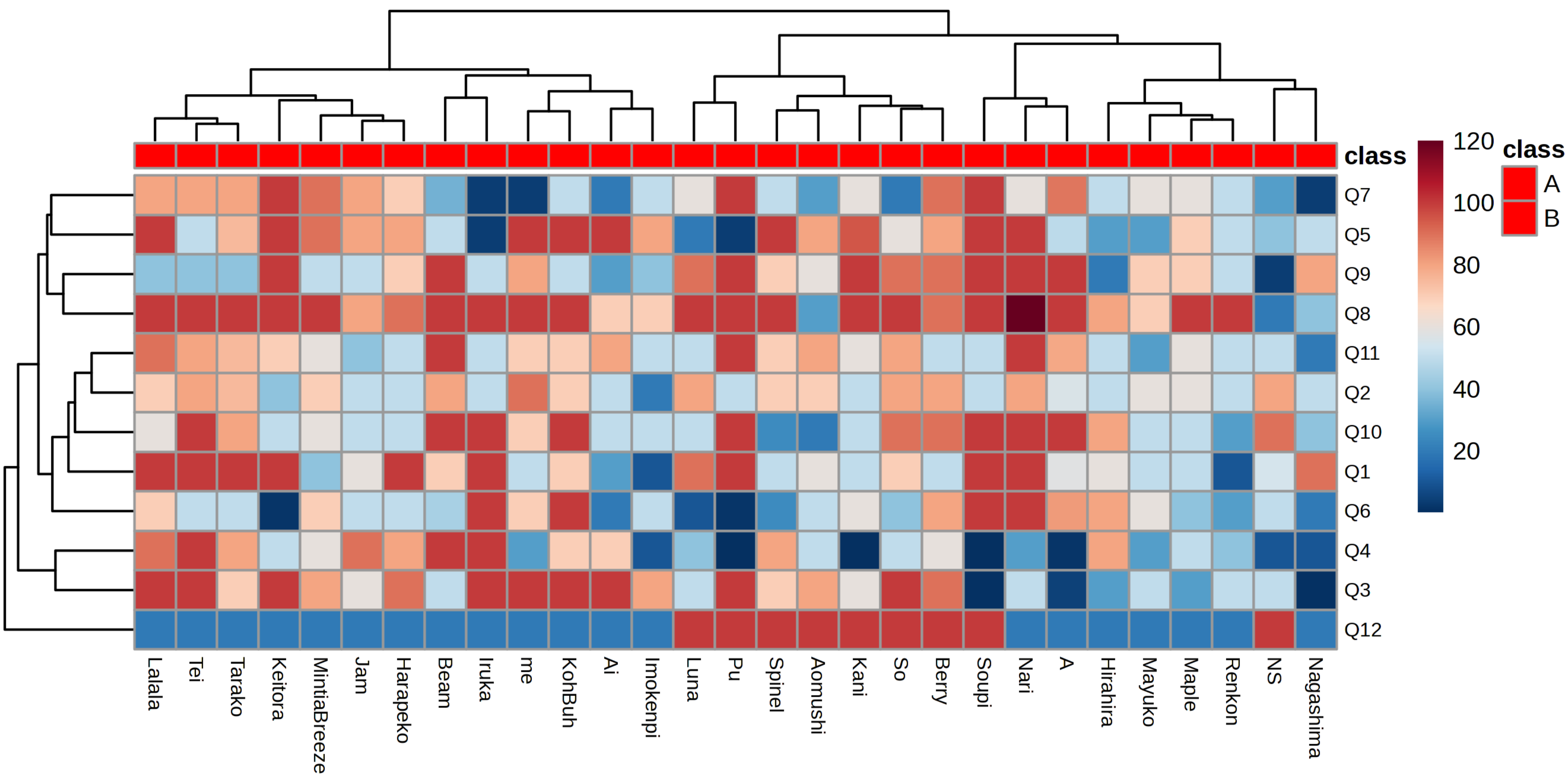
など

階層的クラスター解析

デンドログラム（樹形図）



ヒートマップ



Normalizeしない設定で解析。このためサンプルのデンドログラムが前スライドとは異なる。

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-判别分析

PLS、PLS-DAで扱うデータ

目的変数が存在する

組織ごとの生体試料など

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

		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}
	…					
	y_p	y_{1p}	y_{2p}	y_{3p}		y_{np}
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

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10回

2025年8月6日　神奈川工科大学

櫻井 望

公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

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

自習

課題準備

情報統計
第11回

2025年8月7日　神奈川工科大学

櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

補足

- 数学記号
- ログ変換

2群のt検定(独立2群)

等分散が仮定できない場合　ウェルチの方法

1群目：標本数 n1, 不変標本分散 s1, 標本平均$\overline{x1}$

2群目：標本数 n2, 不変標本分散 s2, 標本平均$\overline{x2}$

検定統計量　$$t = \frac{\overline{x1} - \overline{x2}}{\sqrt{\frac{s1^2}{n1} + \frac{s2^2}{n2}}}$$

(近似)自由度　$$v \approx \frac{\left(\frac{s1^2}{n1} + \frac{s2^2}{n2}\right)^2}{\frac{s1^4}{n1^2(n1-1)} + \frac{s2^4}{n2^2(n2-1)}}$$

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

$\approx$

ほぼ等しい

数学記号

$\circ$	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。
$\mathrm{Im},\ \mathrm{Image},\ \bullet[\bullet]$	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$\varphi\colon X \to Y$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。
$\neq$	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	「$x ≒ y$」または「$x \approx y$」は x と y がほぼ等しいことを表す。記号 ≒ は日本など少数の地域でのみ通用し、≈ 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

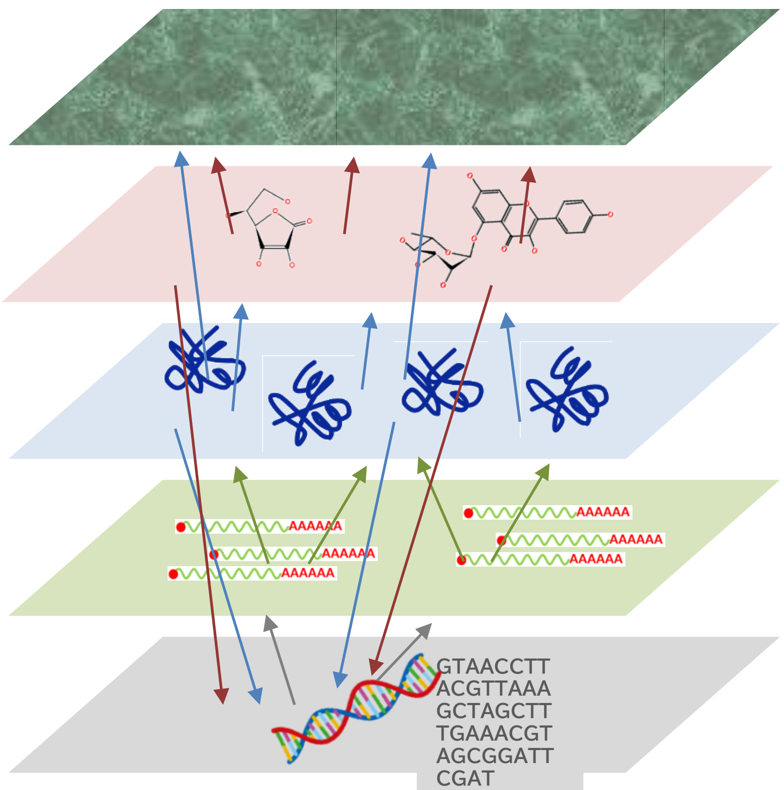
記号	意味	解説
<, >	大小関係、順序	「$x < y$」は　x と y の間に何 方が「先」であることを示す。

Wikipedia

補足

- 数学記号
- ログ変換

生物の遺伝子情報の流れとオミクス



表現型	？
代謝成分	数万？
タンパク質	数万
転写産物	数万
ゲノム	数万

オミクス　それぞれの要素を一斉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

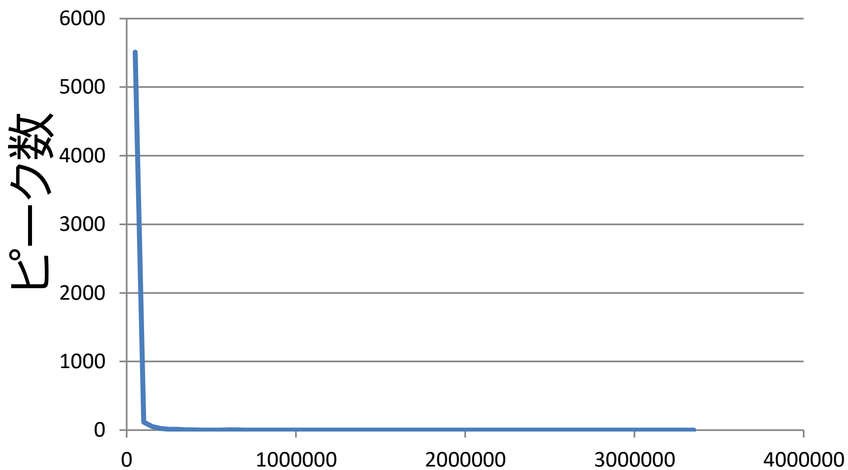
- ✓ 遺伝子発現量データ
- ✓ 質量分析での化合物検出データ

など

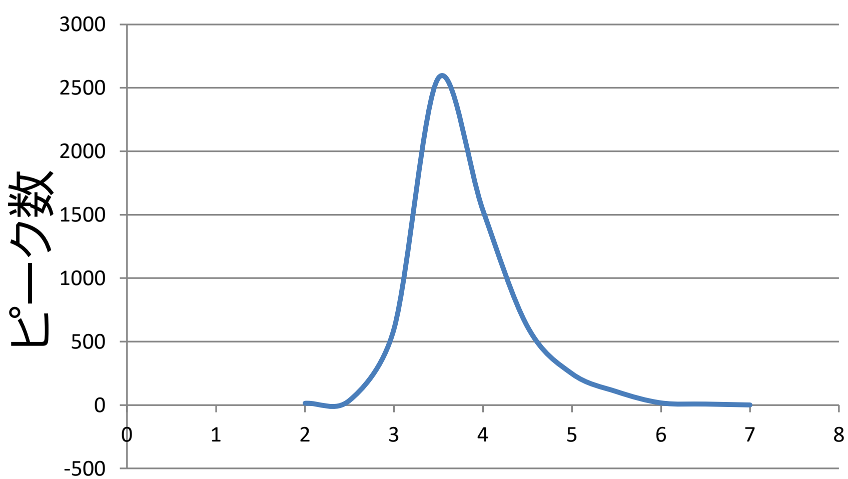
大葉(しそ)で検出された代謝物質

- 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

解析の例

主成分分析、散布図などの例

情報統計
第12-14回

2025年8月7日　神奈川工科大学

櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

- ✓ 授業のおさらい
- ✓ 発表会

お疲れさまでした