統計のイメージ

2023年10月4日　工学院大学　飯島研究室セミナー（八王子）

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

櫻井　望

2018
2023

国立遺伝学研究所

- t-検定
- 分散分析（ANOVA）
- 階層的クラスター解析（HCA）
- 主成分分析（PCA）
- PLS-DA

推定

- t-検定
- 分散分析（ANOVA）
- 階層的クラスター解析（HCA）
- 主成分分析（PCA）
- PLS-DA

分類・要約

予測

多変量解析

今年

今日の内容

- 検定の基礎
- 多変量解析のイメージ
- 多変量解析の実習

✓ 用語
✓ 使うときのポイント

スケジュール

今年

- 検定の基礎
  - ✓ 統計の大事な考え方
平均値〜t分布まで
  - ✓ t検定
検定のやりかた
気を付けること
  - ✓ いろんな検定とANOVA
- 多変量解析のイメージ
- 多変量解析の実習

午前中

〜90分

〜60分

〜夕方まで

統計

学

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！

目的：

研究室メンバーの身長について知りたい。

代表的な数値：

平均値、最大値、

そのほか：

（基本・基礎）統計量

平均値
中央値

中心を表す値

最大値
最小値

一番を表す値

目的：

研究室メンバーの身長について知りたい。

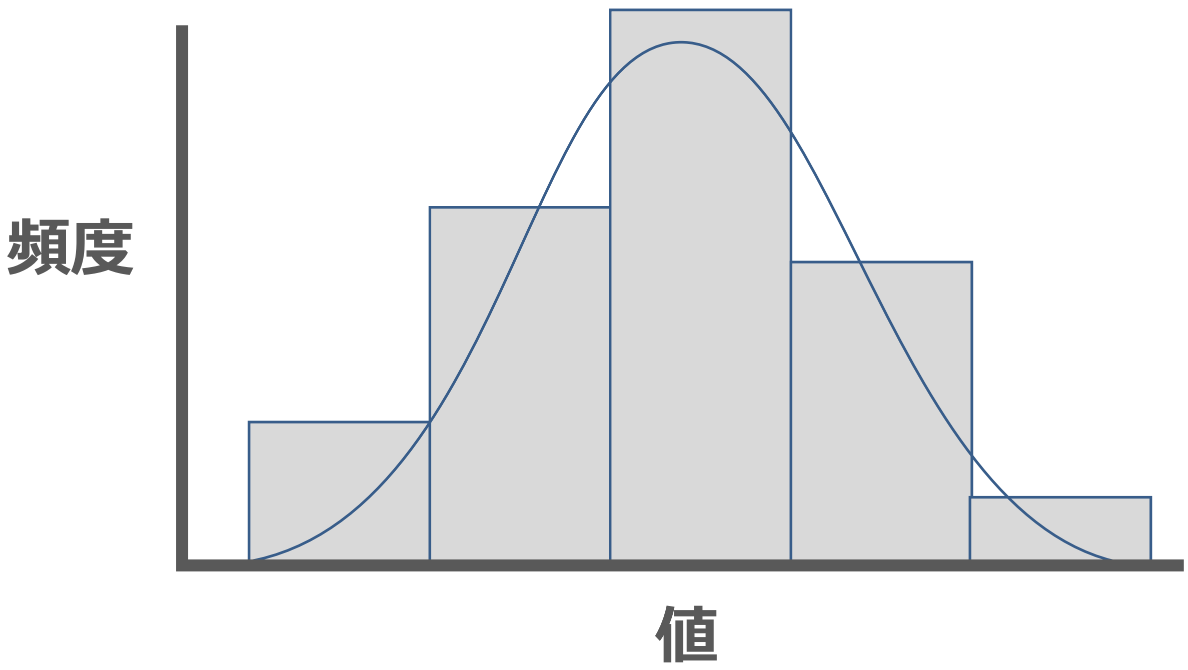
代表的な数値：

平均値、最大値、

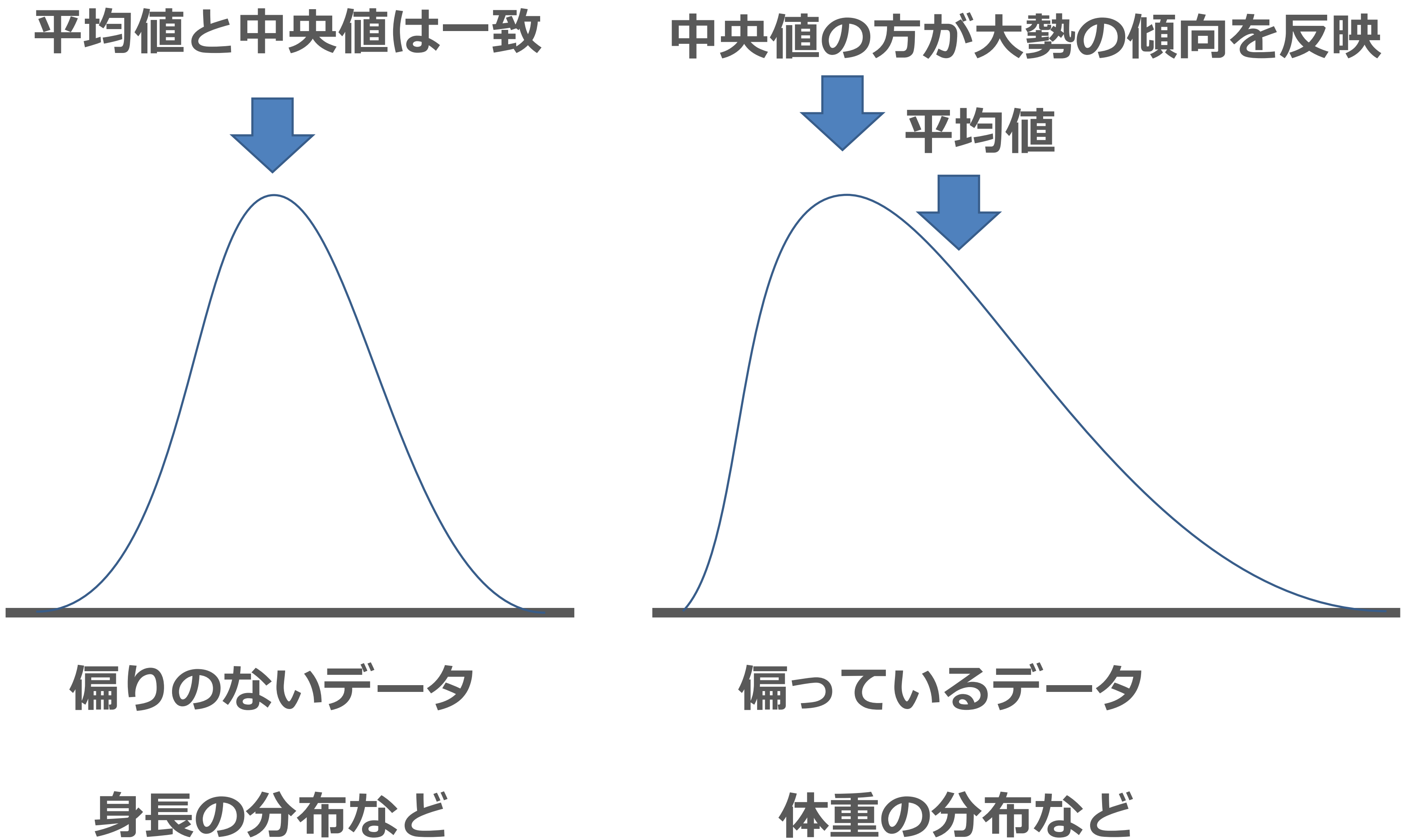
そのほか：

グラフ（図）

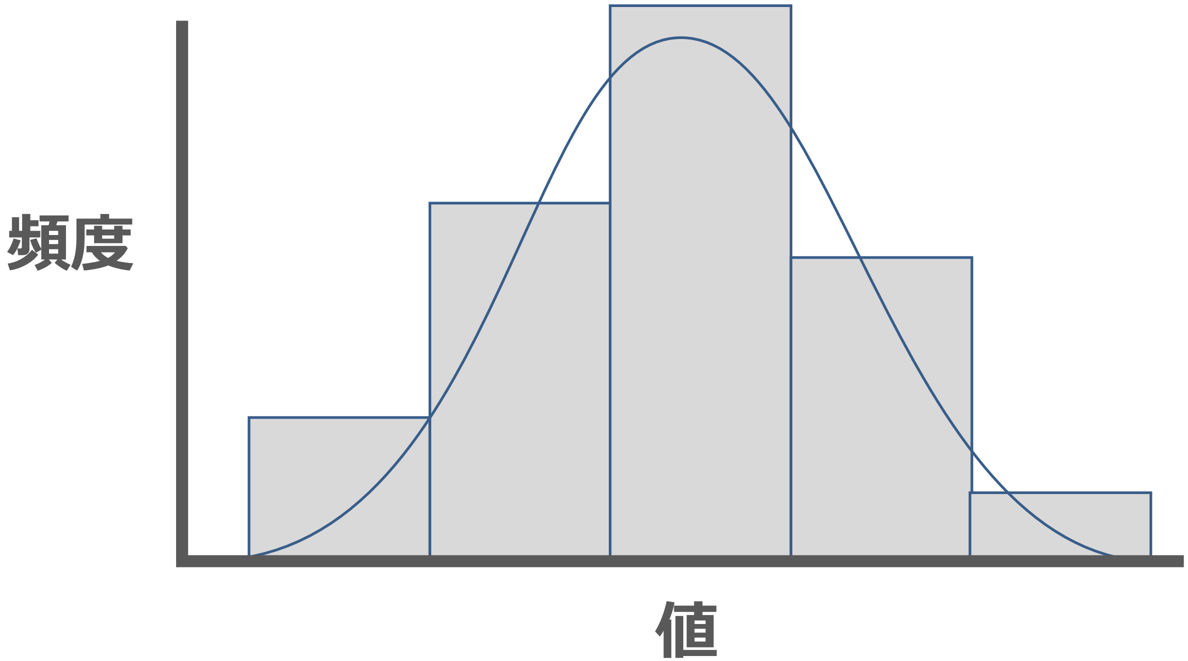
グラフの例



ヒストグラム（頻度分布図）


平均値と中央値は一致
中央値の方が大勢の傾向を反映
平均値
偏りのないデータ
偏っているデータ
身長の分布など
体重の分布など

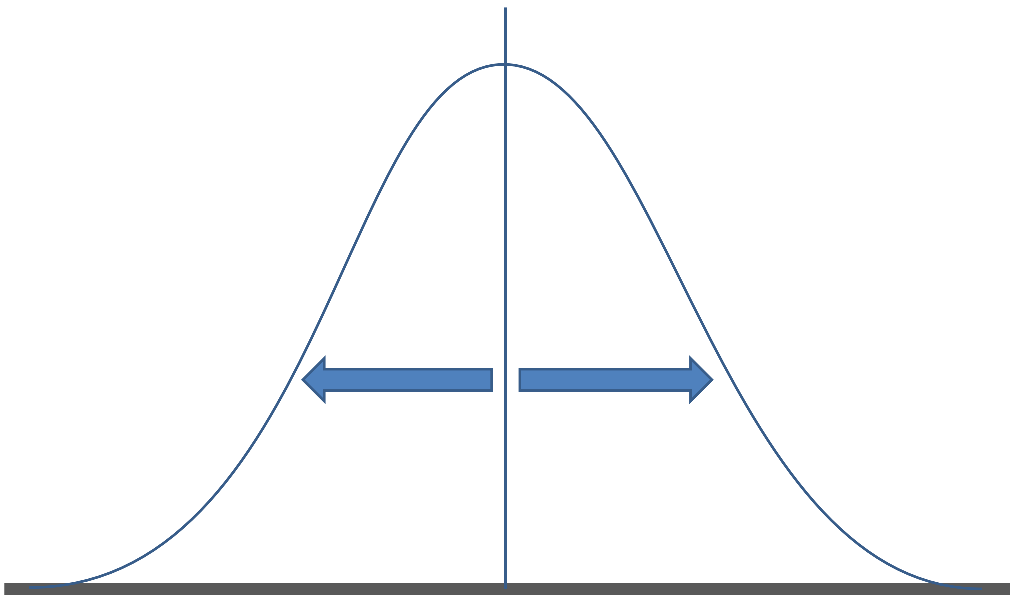
グラフの例



ヒストグラム（頻度分布図）

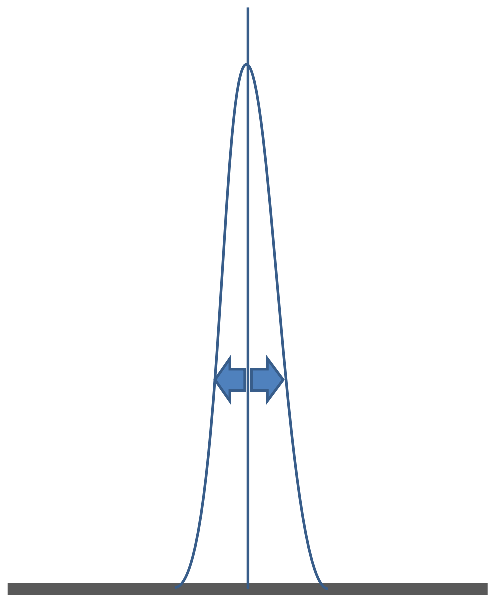
データの散らばり具合がわかる　分布

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

（基本・基礎）統計量

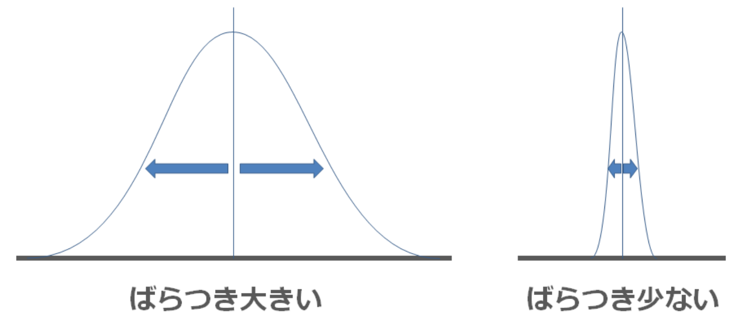
分散

標準偏差
＝分散の平方根

ばらつきを
表す値

分散

その集団の、平均値からのずれの大きさ



- 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

⑤要素数nで
割って平均
にする

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

①平均値

③その2乗

②要素iと平均値の差

④その全要素(iが1からnまで)の合計

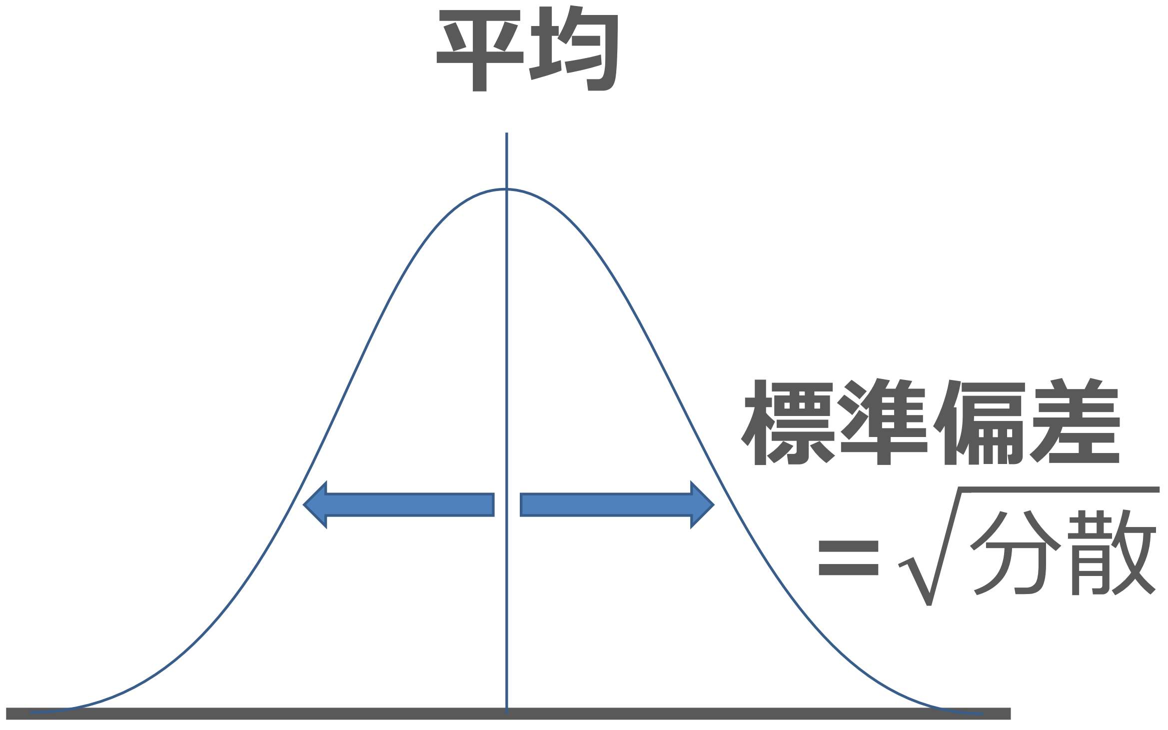
分散…2乗された値

計測値と単位を
そろえるため

平方根を計算

標準偏差

イメージ



目的：この研究室の人の身長はどのくらい？

もっと広い
世界が知りたい

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！

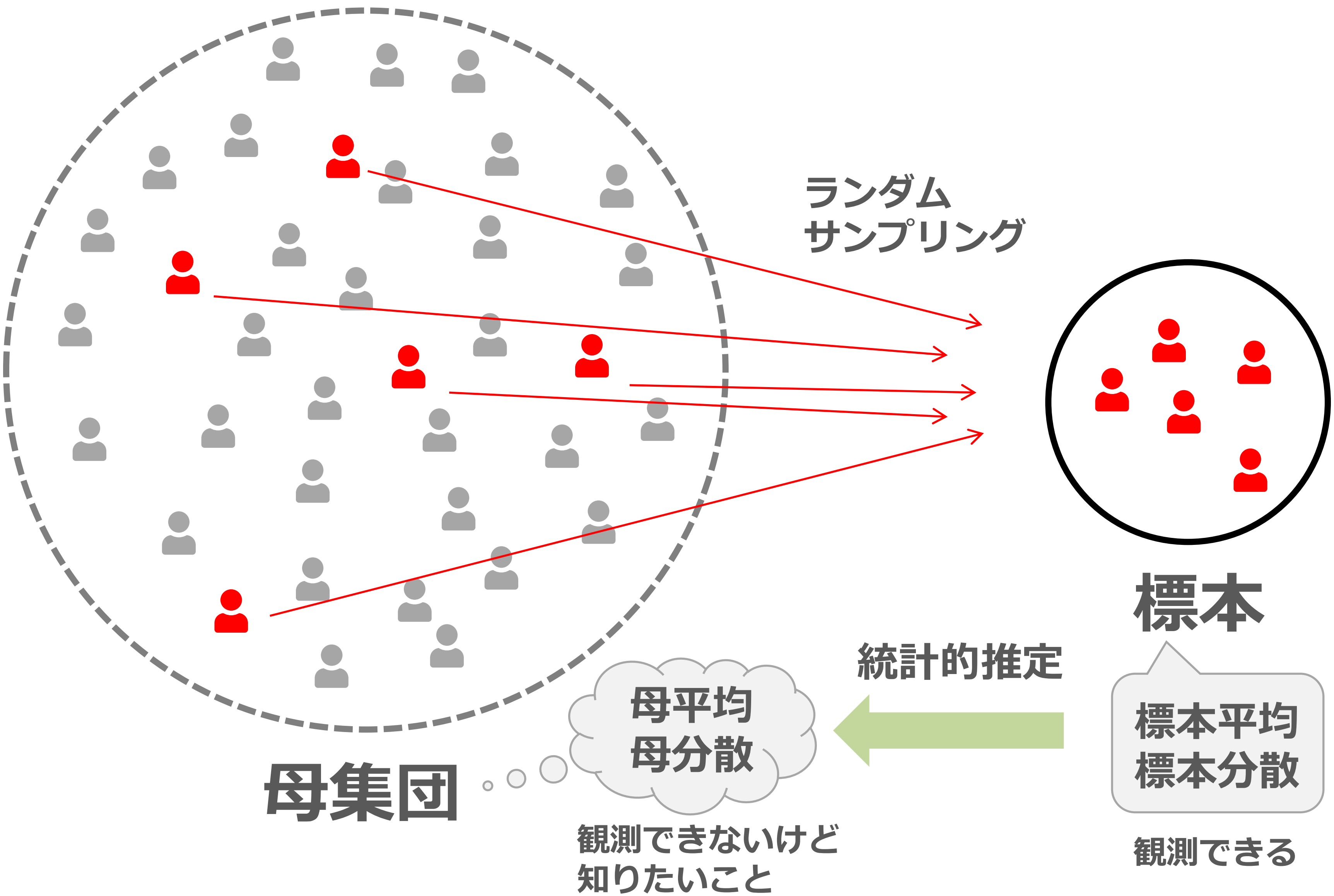
目的：日本人の身長はどのくらい？

↓

限られた標本から
母集団（日本人全体）の
平均値やばらつきを
推定する、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$\overline{x}$

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

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

②要素iと平均値の差

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

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

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ためです。

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

母平均μ $\leftarrow$ 標本平均$\bar{x}$

一致が期待できる

母分散σ^2 $\cancel{\leftarrow}$ 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**

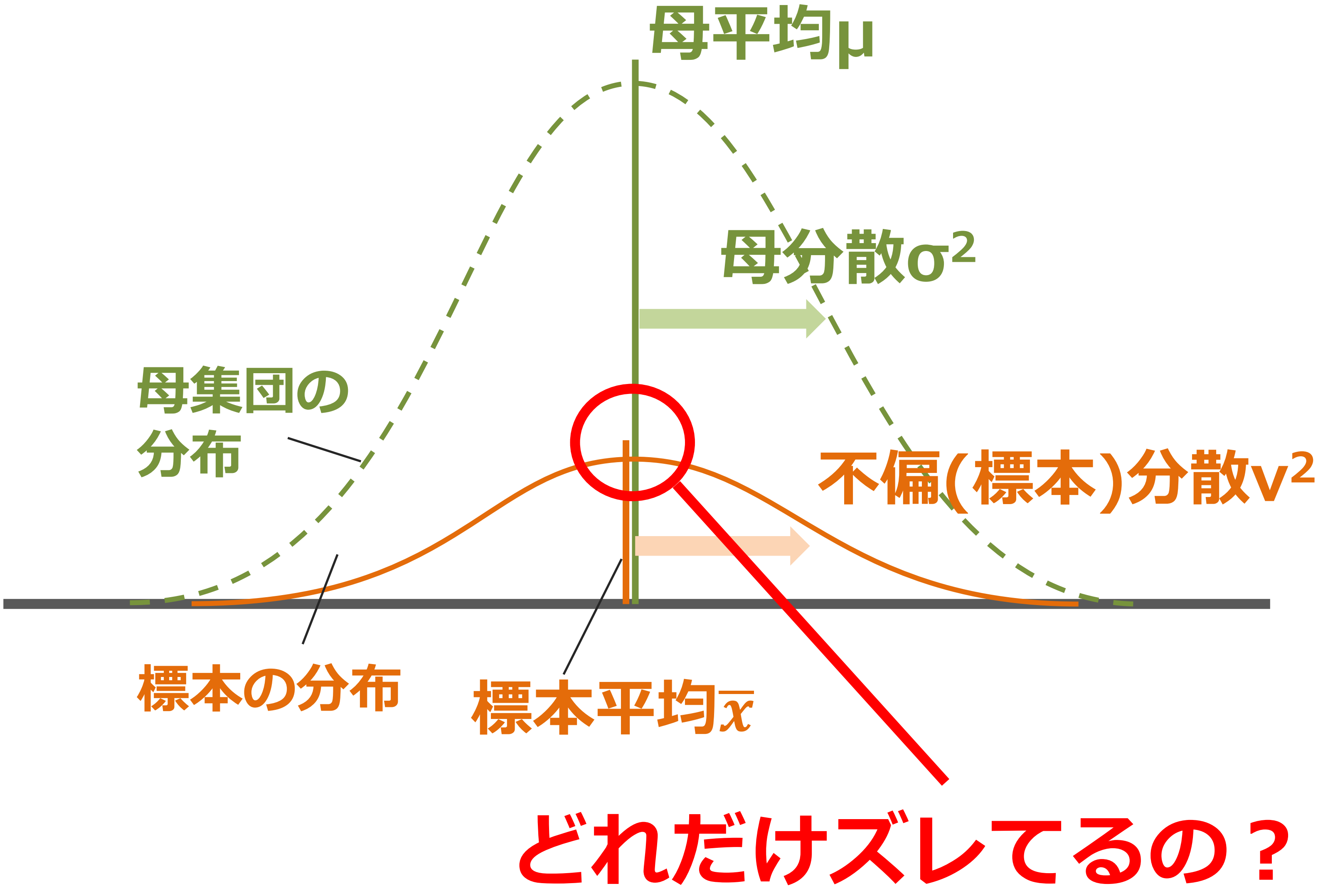
$\nwarrow$

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。

イメージ

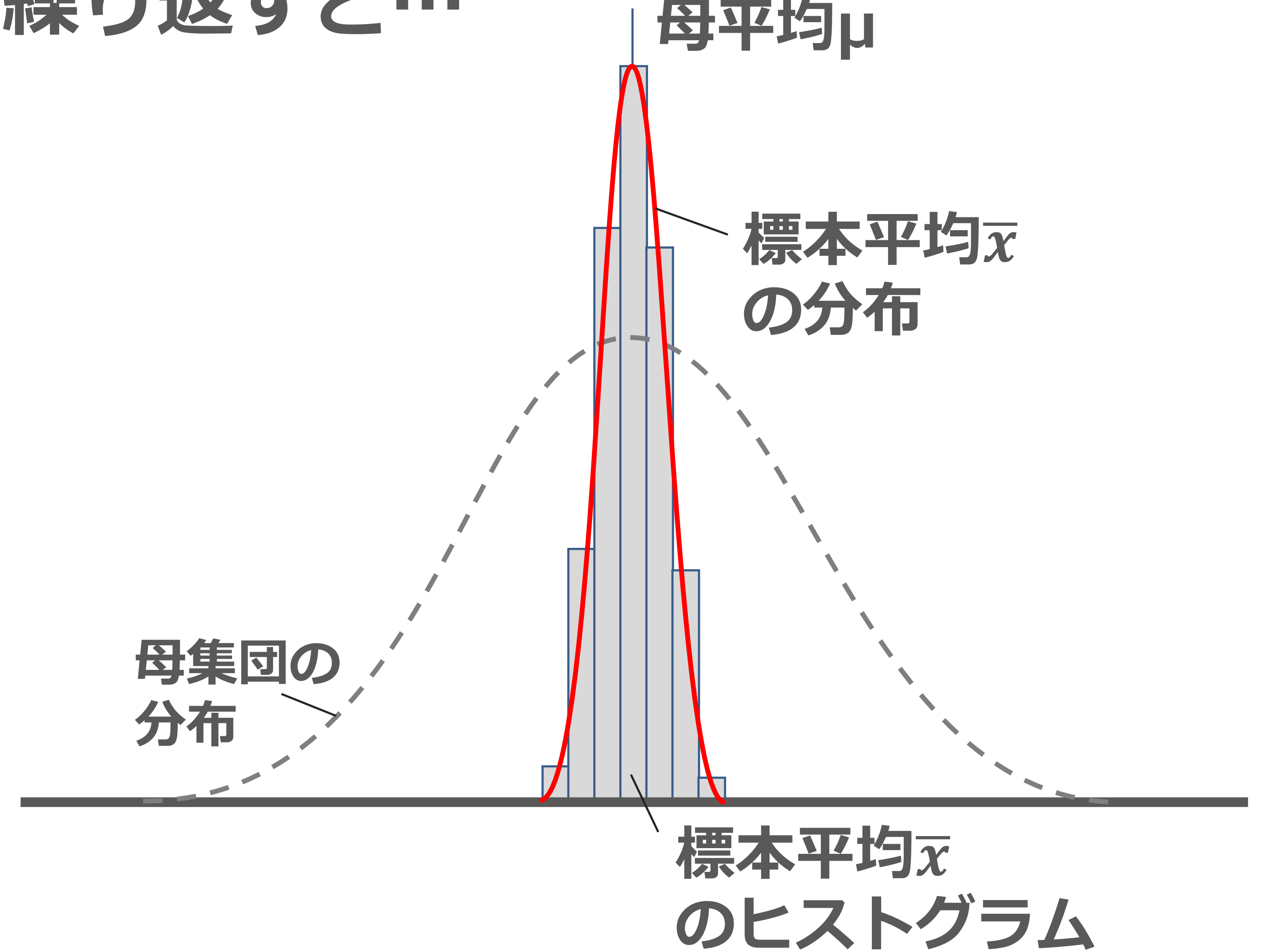


母平均μ
母分散σ²
母集団の
分布
不偏(標本)分散v²
標本の分布
標本平均$\overline{x}$
どれだけズレてるの？

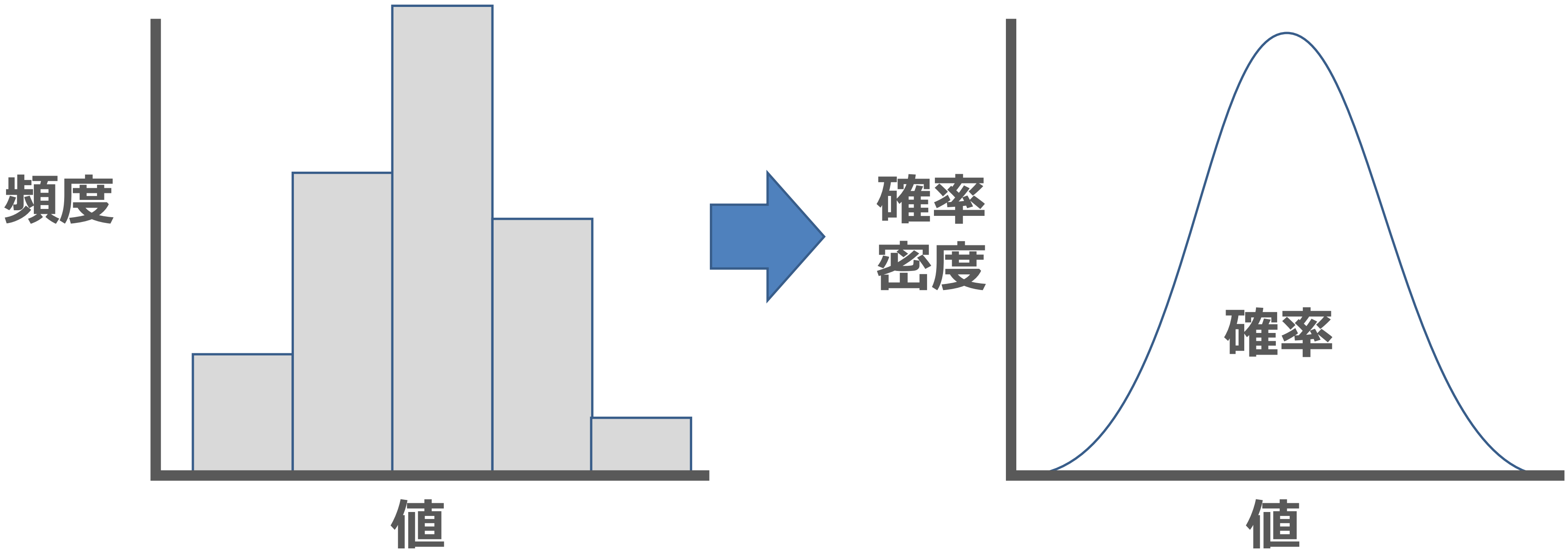
誤差

- サンプリング誤差
- 測定誤差

サンプリングして標本平均$\overline{x}$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

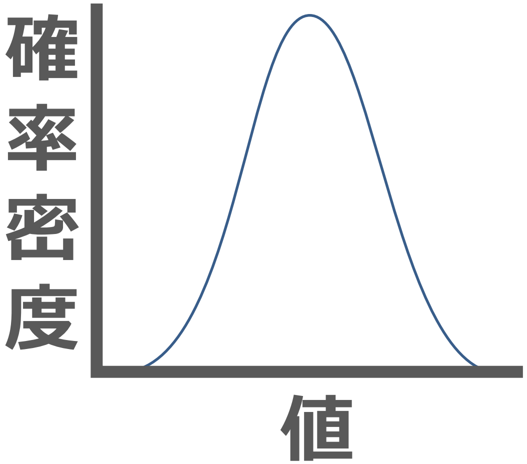
観測結果

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



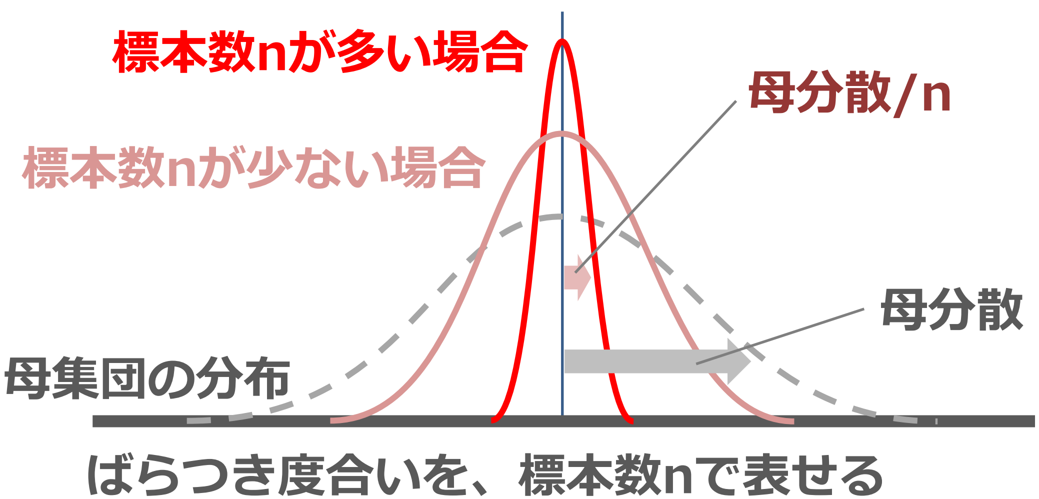
均等な確率で生じたばらつき
の場合にとる分布

✓ 身長の分布
✓ 測定誤差の分布
✓ 自然界で起こるゆらぎ　　など

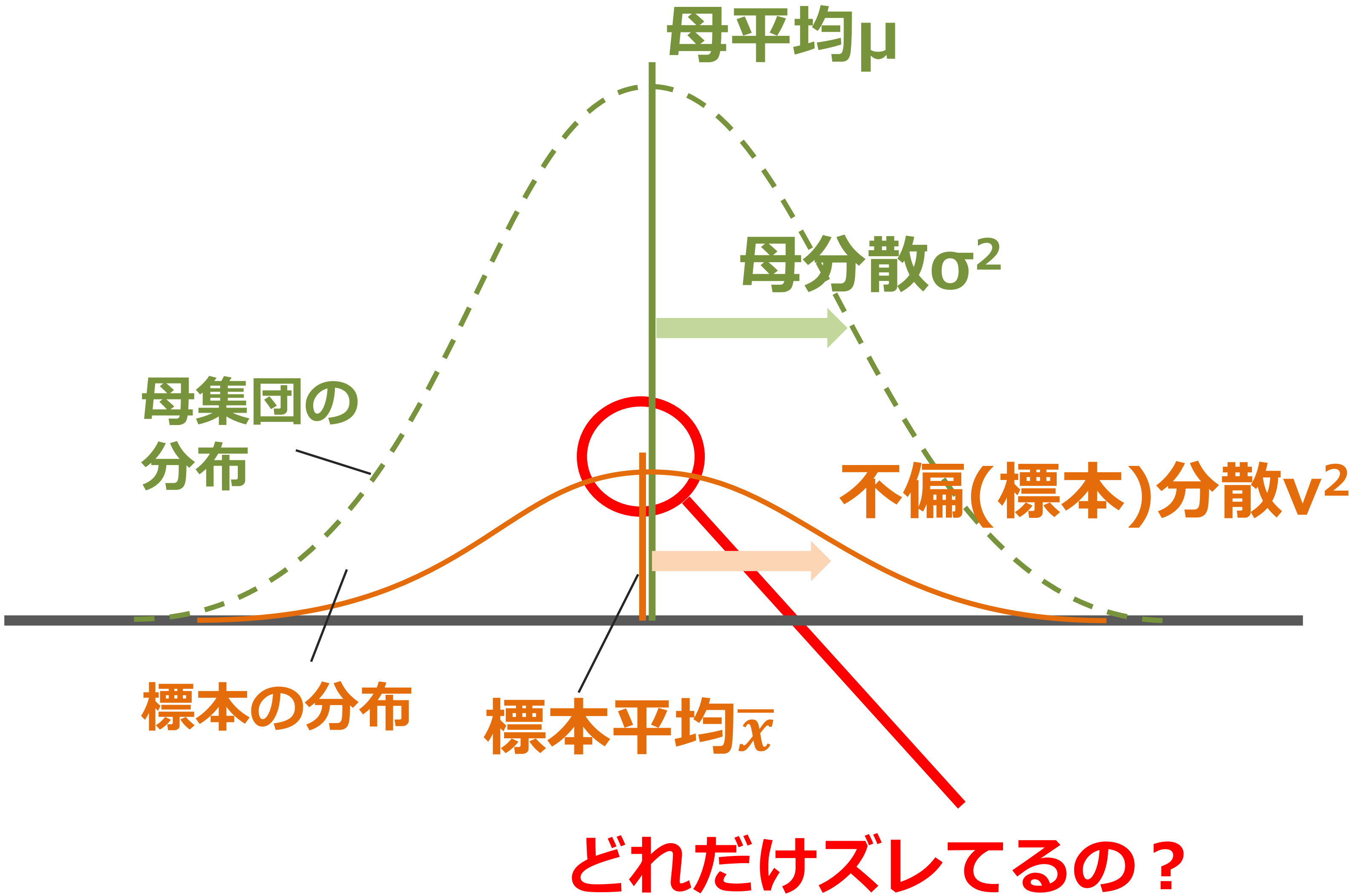
標本平均$\overline{x}$の分布

- 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。



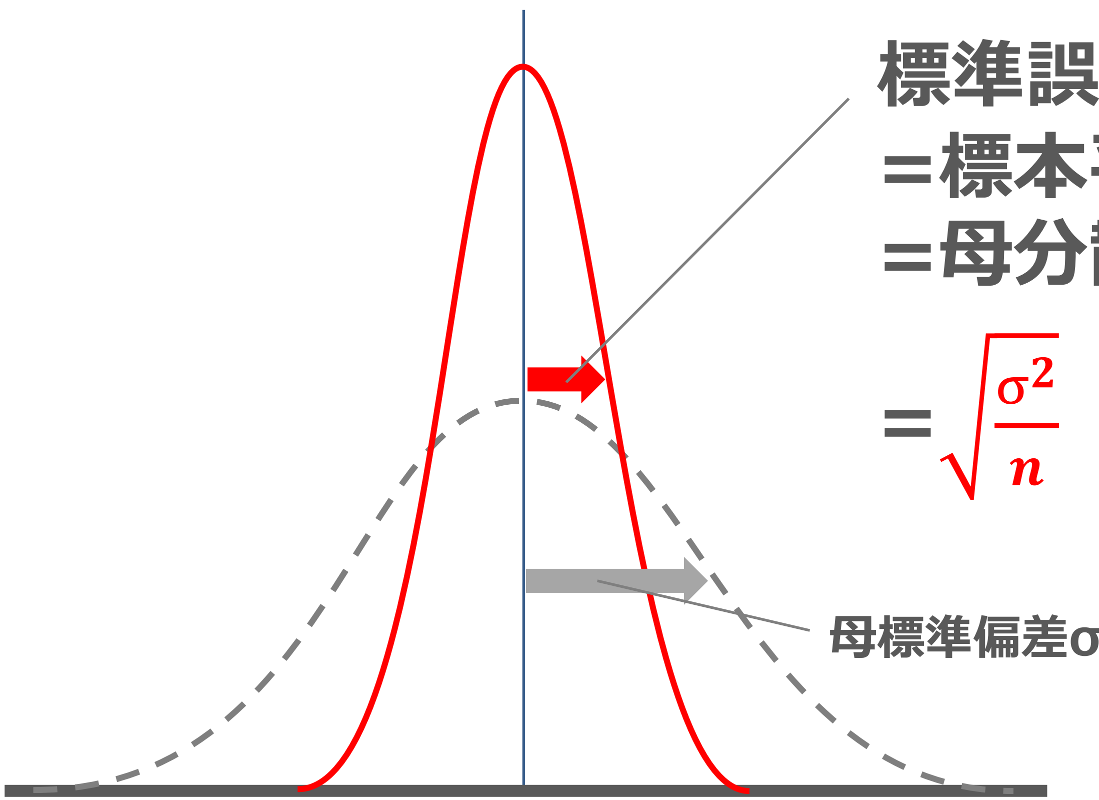
中心極限定理


母平均μ
母分散σ²
母集団の
分布
不偏(標本)分散v²
標本の分布
標本平均$\overline{x}$
どれだけズレてるの？

➡ **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

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

標準偏差と標準誤差

論文などでよく見る図

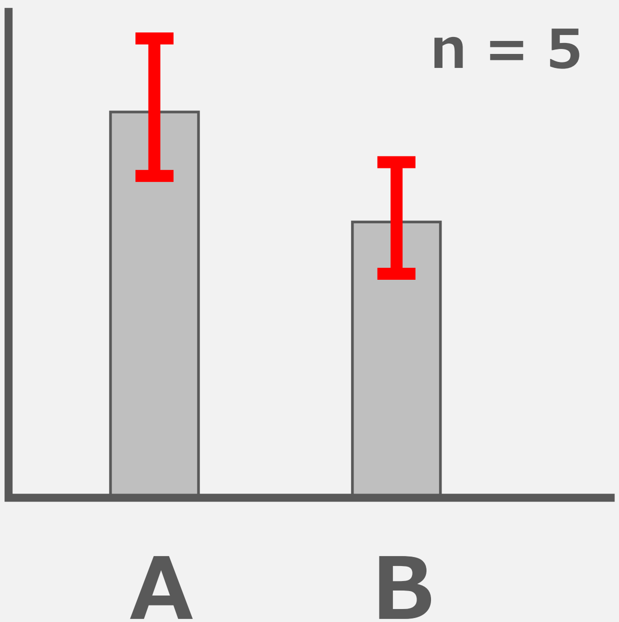


図１ A群とB群の＊＊の違い
それぞれ5個体を測定した。
エラーバーは標準偏差を表す

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

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$\overline{x}$

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**

母分散σ^2 ← **不偏(標本)分散v^2**

一致が期待できる

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 s^2

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 v^2

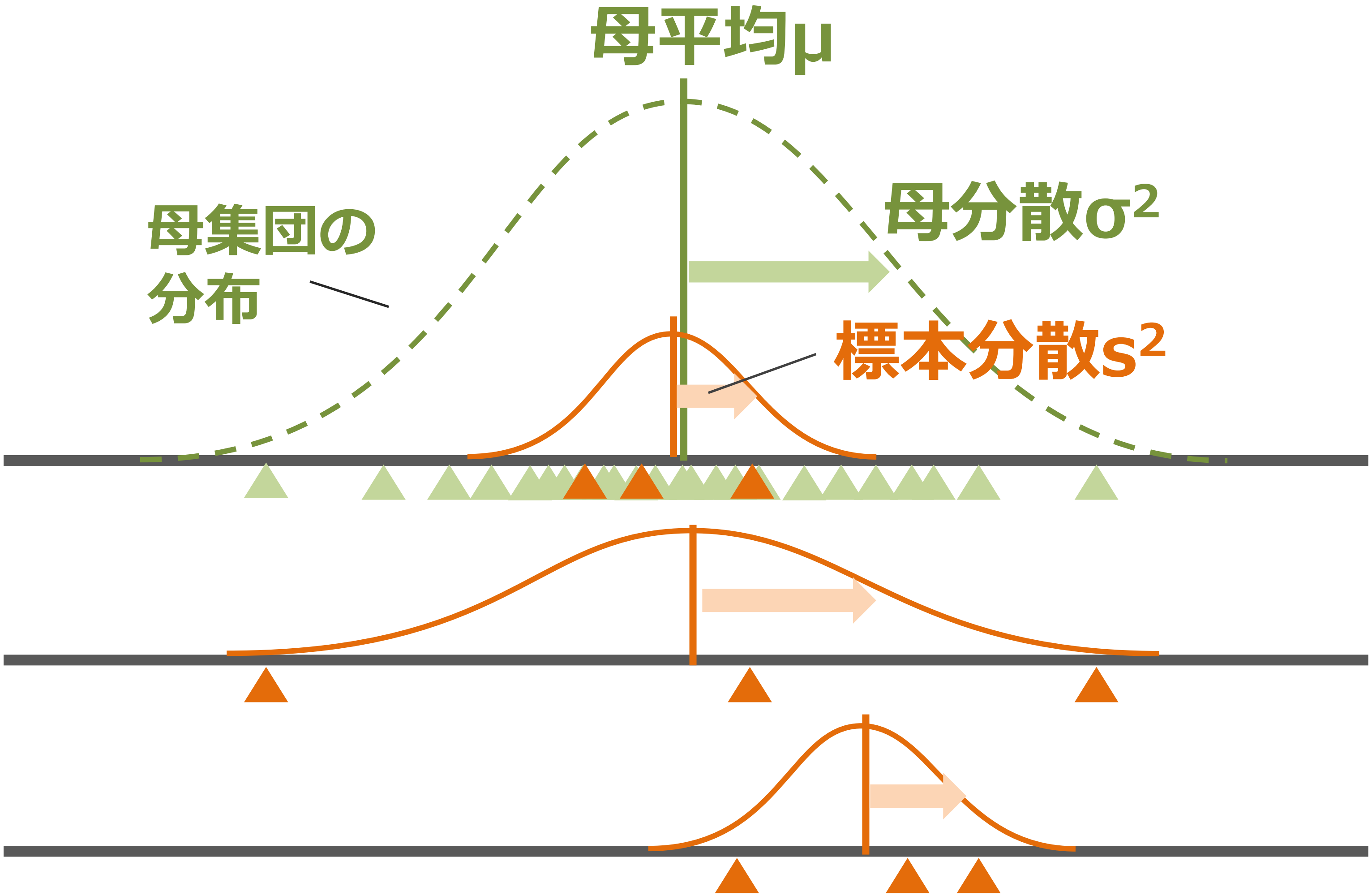
⑤n-1で割る

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

→ 少し大きな値に補正されている

標本分散s^2は、母分散σ^2よりも小さくなる

- 標本数n=N(母集団の要素全数)なら、標本分散s^2=母分散σ^2
- n < Nなら、必ず標本分散は母分散より小さくなる



どのくらい小さくなっているか？

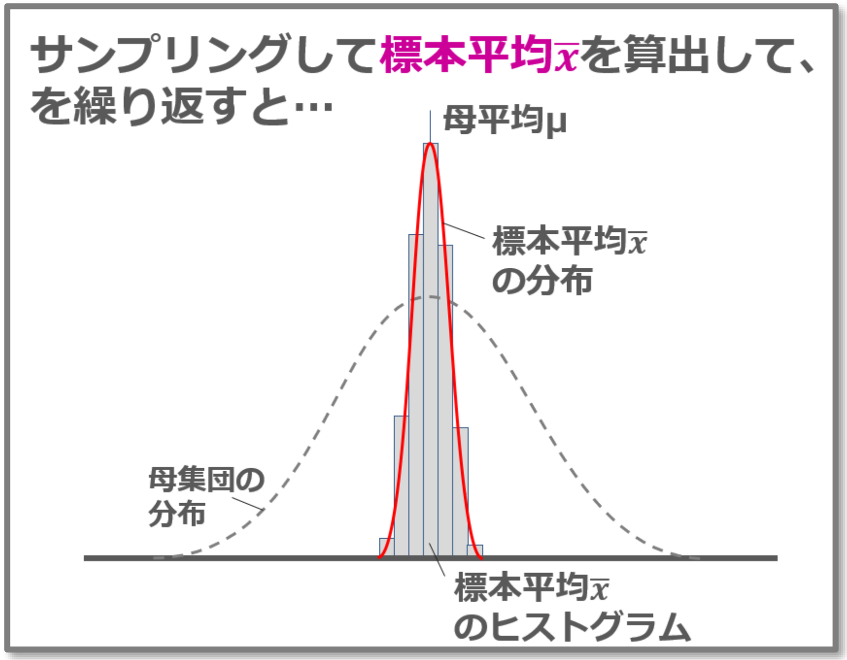
サンプリングして標本分散s^2を算出して、を繰り返すと…

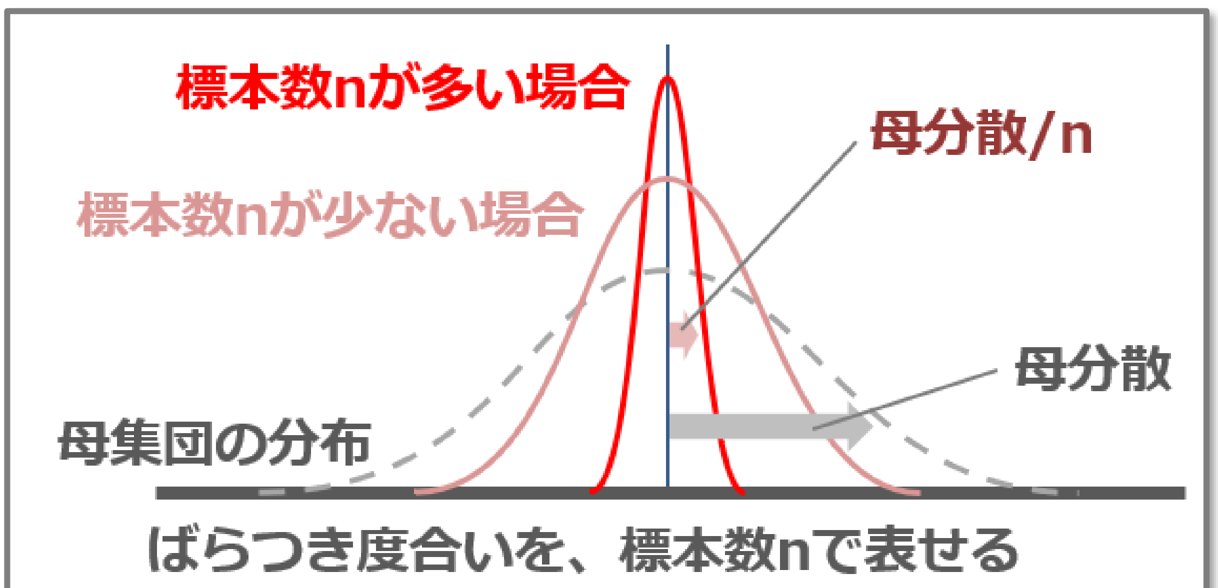
…

標本分散s^2

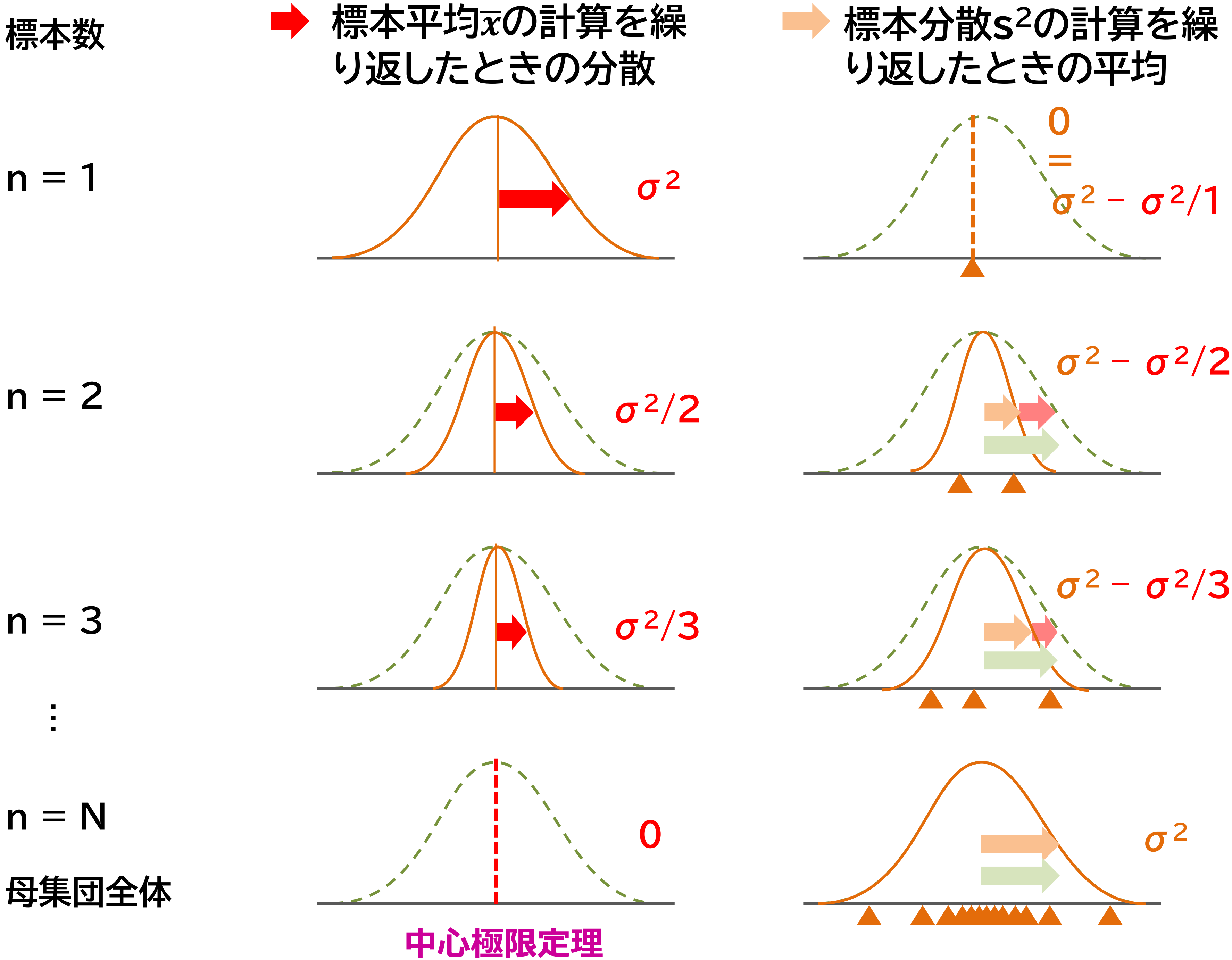
$m(s^2)$ = 母分散σ^2 - X

標本分散s^2の平均値


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…
母平均μ
標本平均$\bar{x}$
の分布
母集団の
分布
標本平均$\bar{x}$
のヒストグラム


標本数nが多い場合
母分散/n
標本数nが少ない場合
母分散
母集団の分布
ばらつき度合いを、標本数nで表せる

中心極限定理


標本数
標本平均$\bar{x}$の計算を繰り返したときの分散
標本分散s^2の計算を繰り返したときの平均
n = 1
σ^2
0
=
$\sigma^2 - \sigma^2/1$
n = 2
$\sigma^2/2$
$\sigma^2 - \sigma^2/2$
n = 3
$\sigma^2/3$
$\sigma^2 - \sigma^2/3$
⋮
n = N
母集団全体
0
σ^2
中心極限定理

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、を繰り返すと…

…

標本分散s^2

$m(s^2)$　＝　母分散σ^2　$\boxed{-\ \dfrac{\sigma^2}{n}}$

標本分散s^2の平均値

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

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

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

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

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

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**

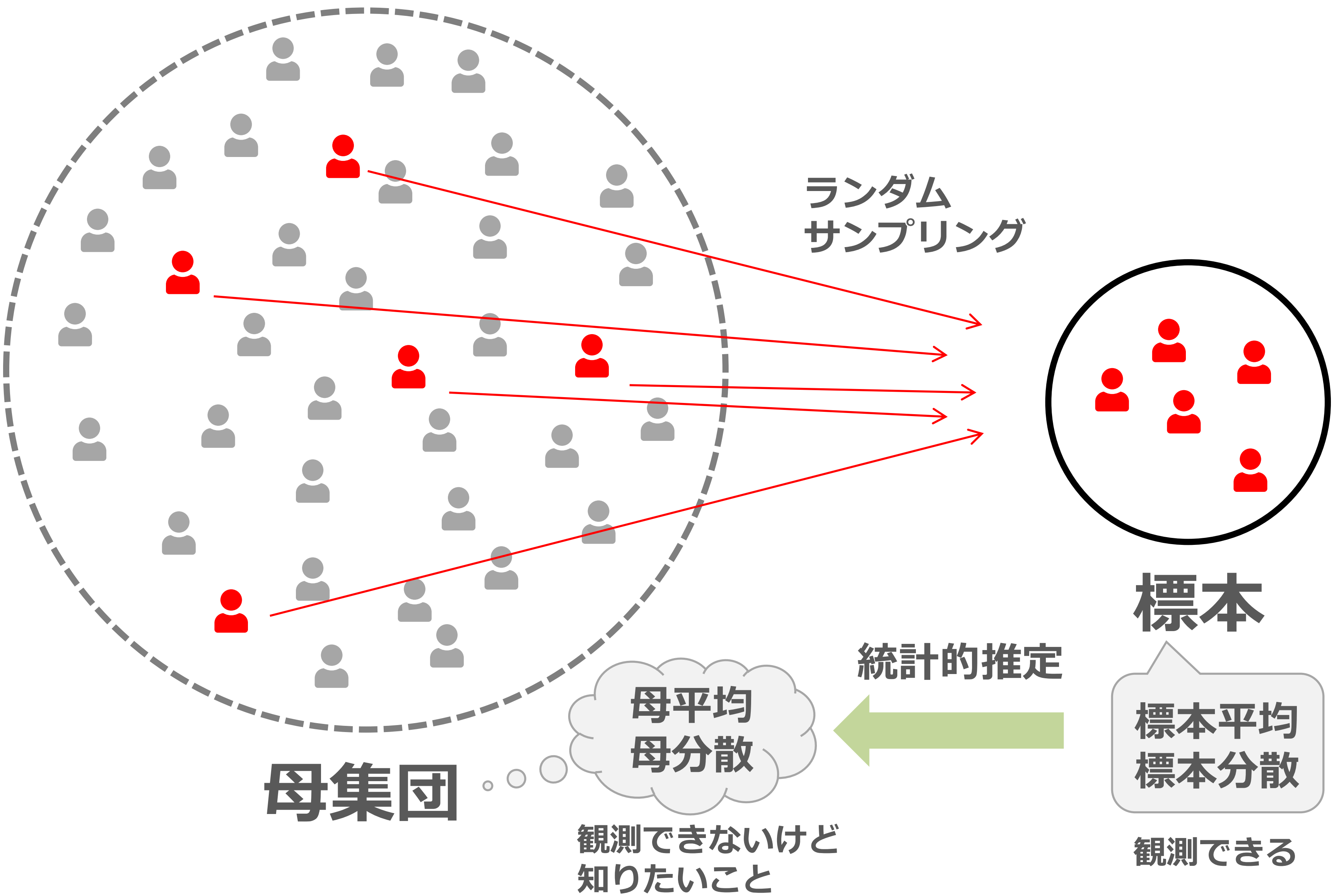
一致が期待できる

不偏(標本)分散v^2

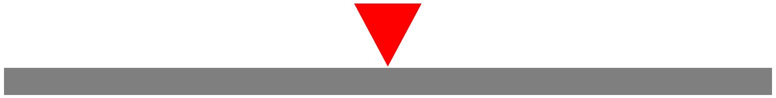
真の値から外れていないことを、
不偏性があると言うので。

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。



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

「日本の男子の平均身長は、95%の信頼区間で170.2 ~174.6 cmである」のように、幅を持たせて表現する方法

今年

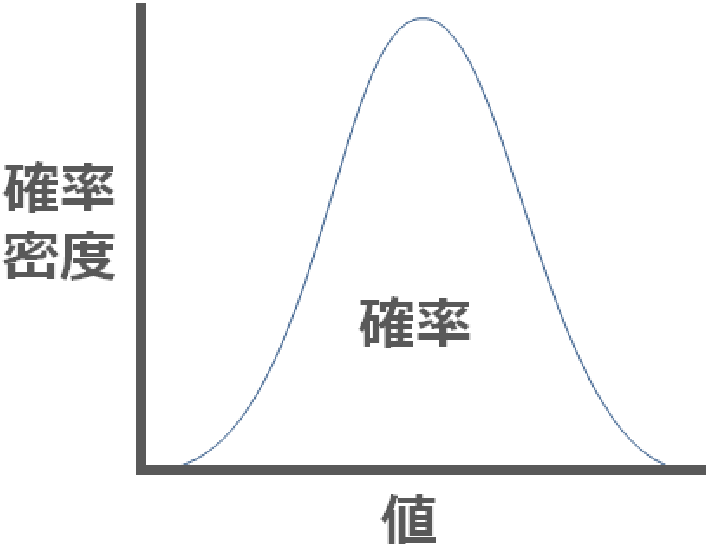
分布

を使うと、ある事象が起こる

確率

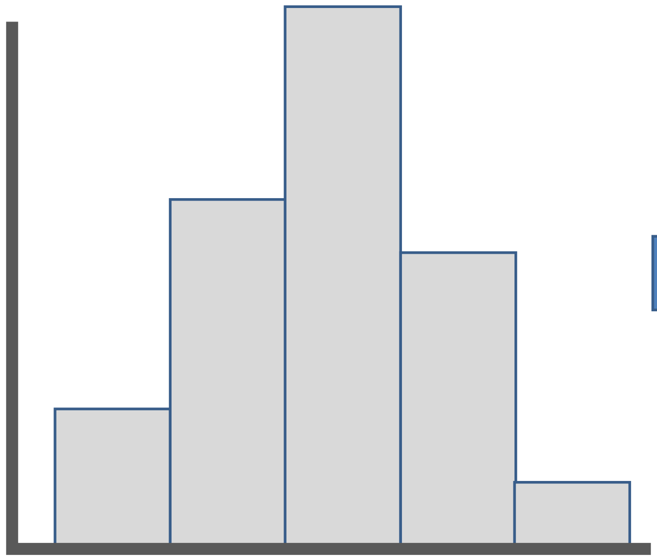
が分かる

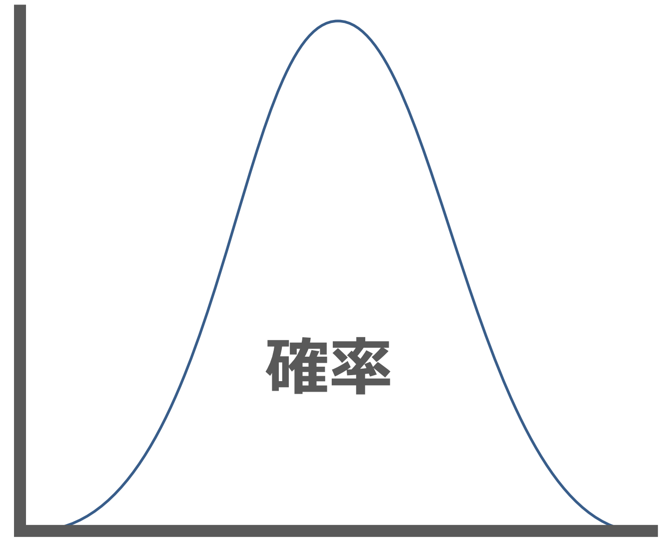
検定の基本的な考え方



復習

分布　データの散らばり具合





ヒストグラム

観測結果

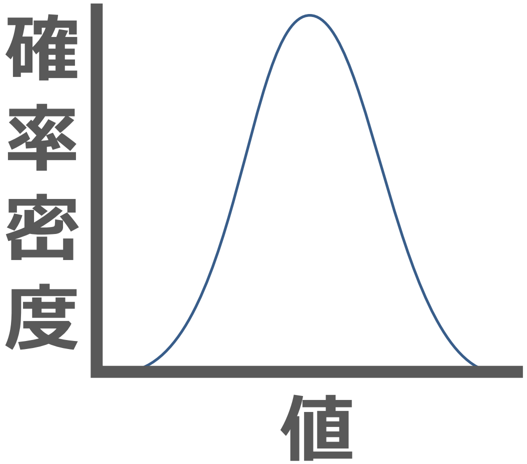
確率密度関数

事象の起こる確率を表すモデル

正規分布（ガウス分布）

復習

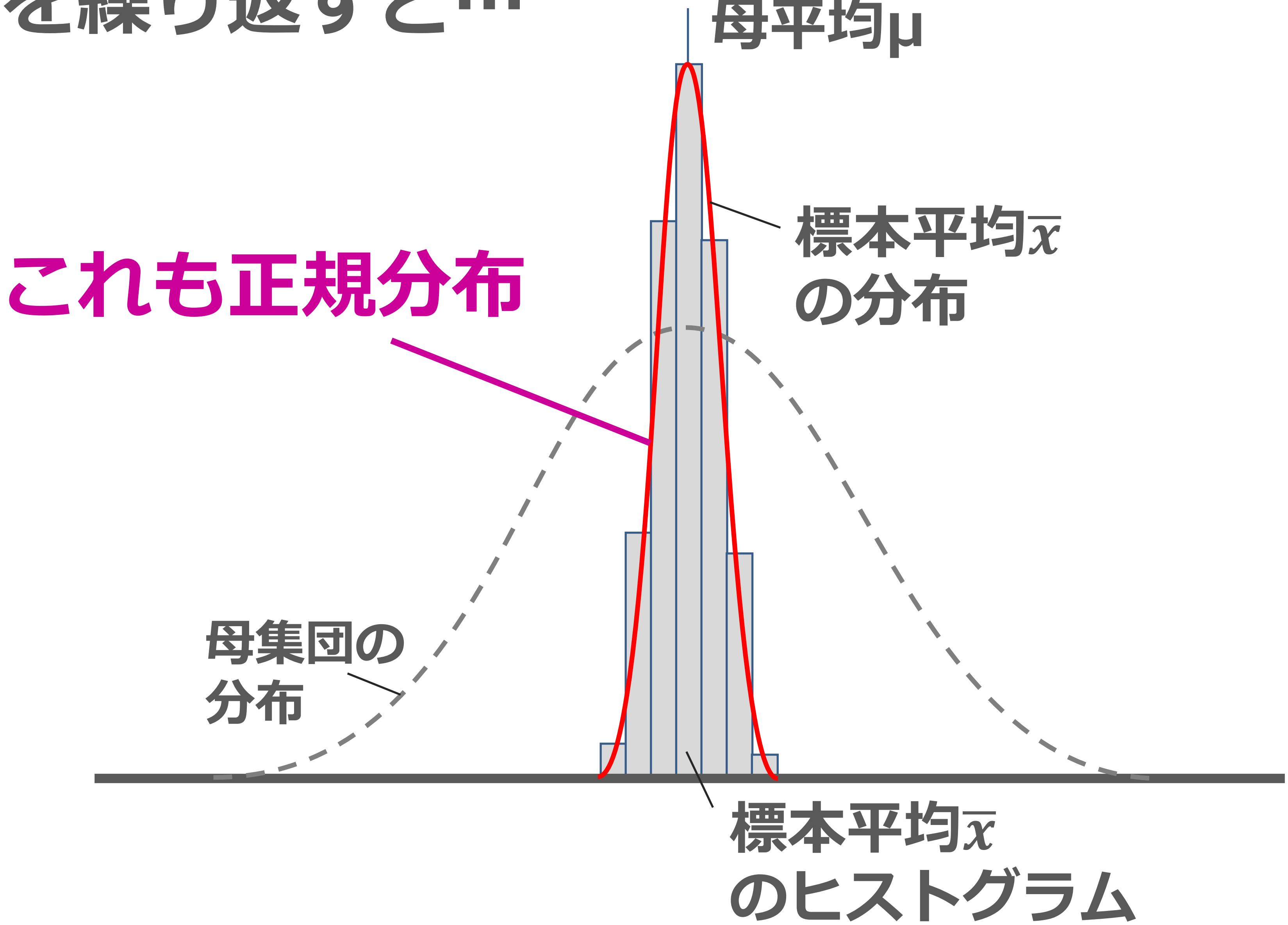
- 平均値が中心で、
- 平均値に近いものが多く、
- 左右に均等な釣り鐘状の分布



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

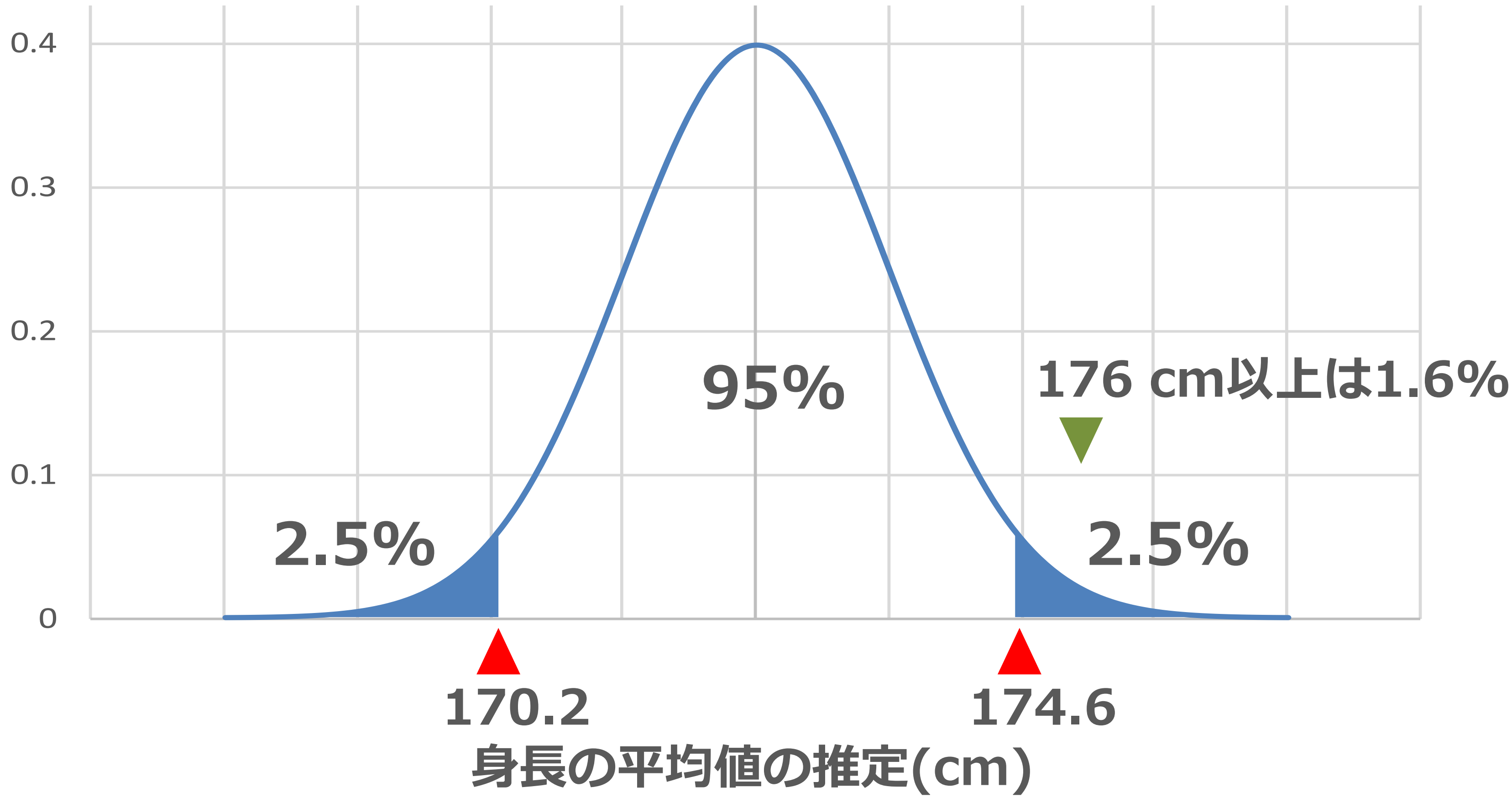
サンプリングして標本平均$\bar{x}$を算出して、を繰り返すと…



分布を使うと

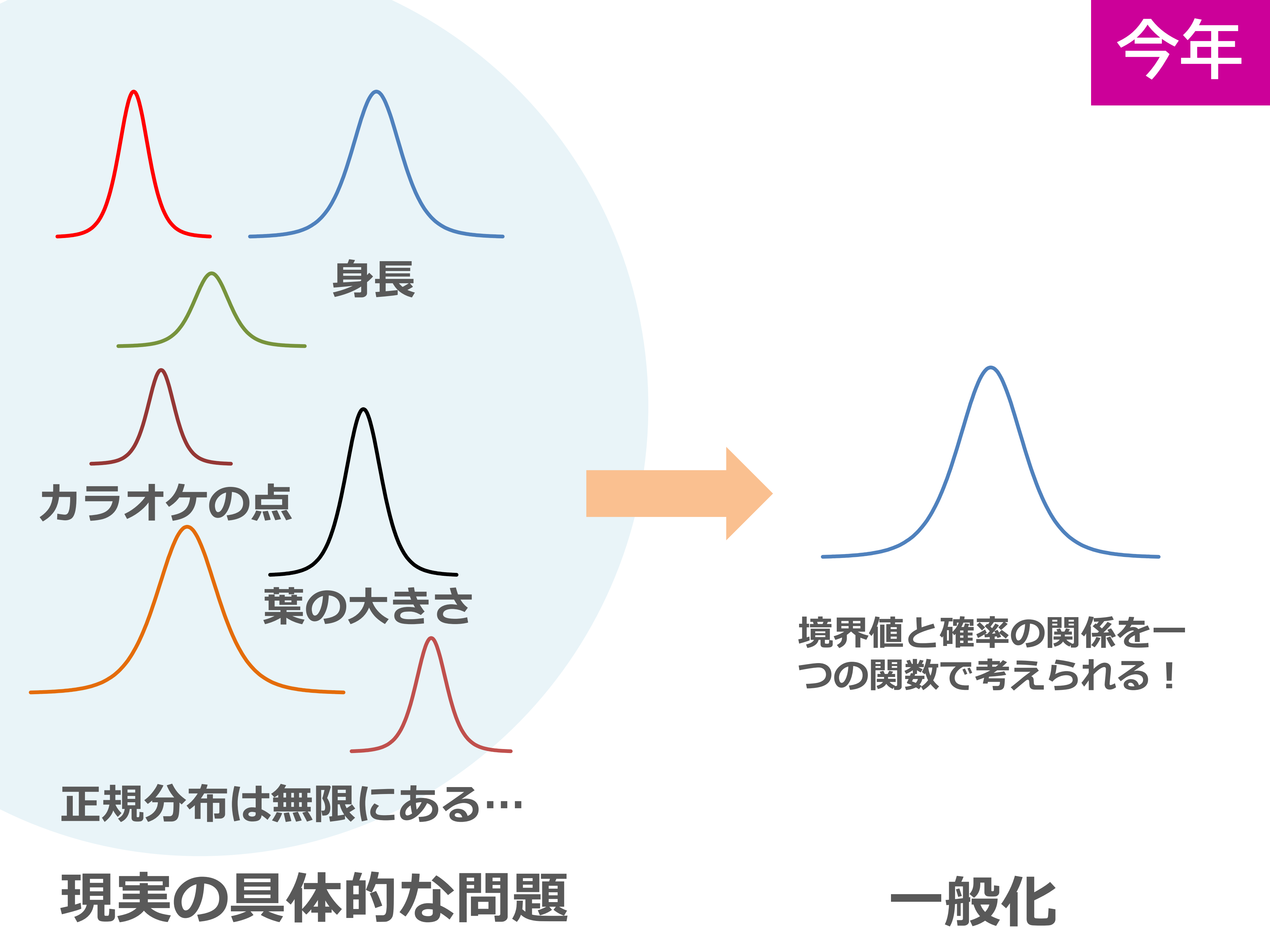
・ある事象がある確率で起きるときの境界値▲
・ある境界値▼以上（以下）が起こる確率

が分かる



「日本の男子の平均身長は、95%の信頼区間で170.2 ~174.6 cmである」

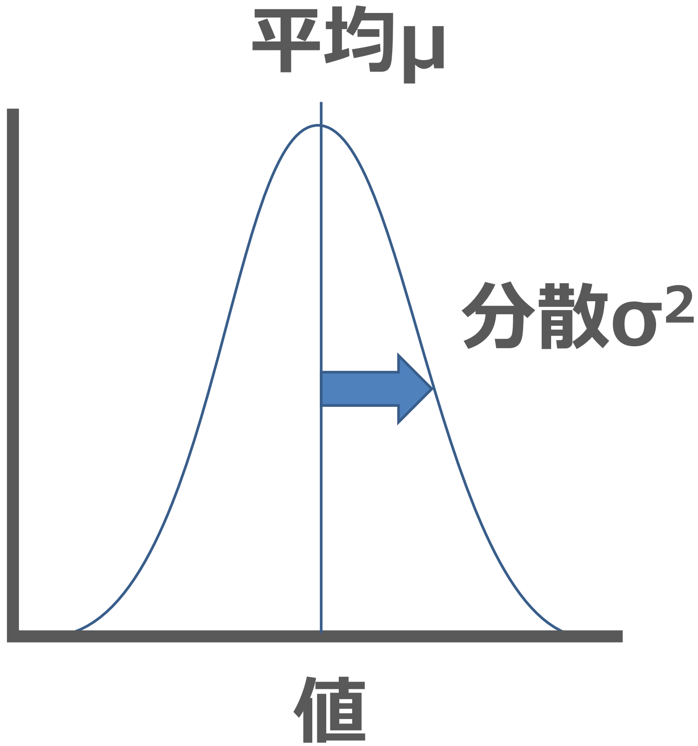
のような**区間推定**ができる



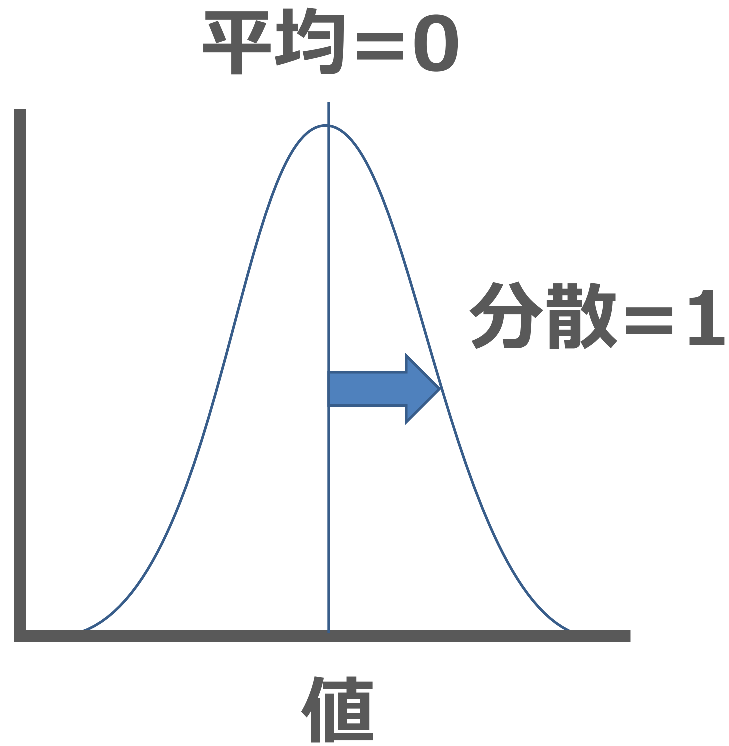
今年
身長
カラオケの点
葉の大きさ
正規分布は無限にある…
現実の具体的な問題
境界値と確率の関係を一
つの関数で考えられる！
一般化

標準正規分布

正規分布　　　標準正規分布



平均と分散で決まる
$N(\mu, \sigma^2)$と表記

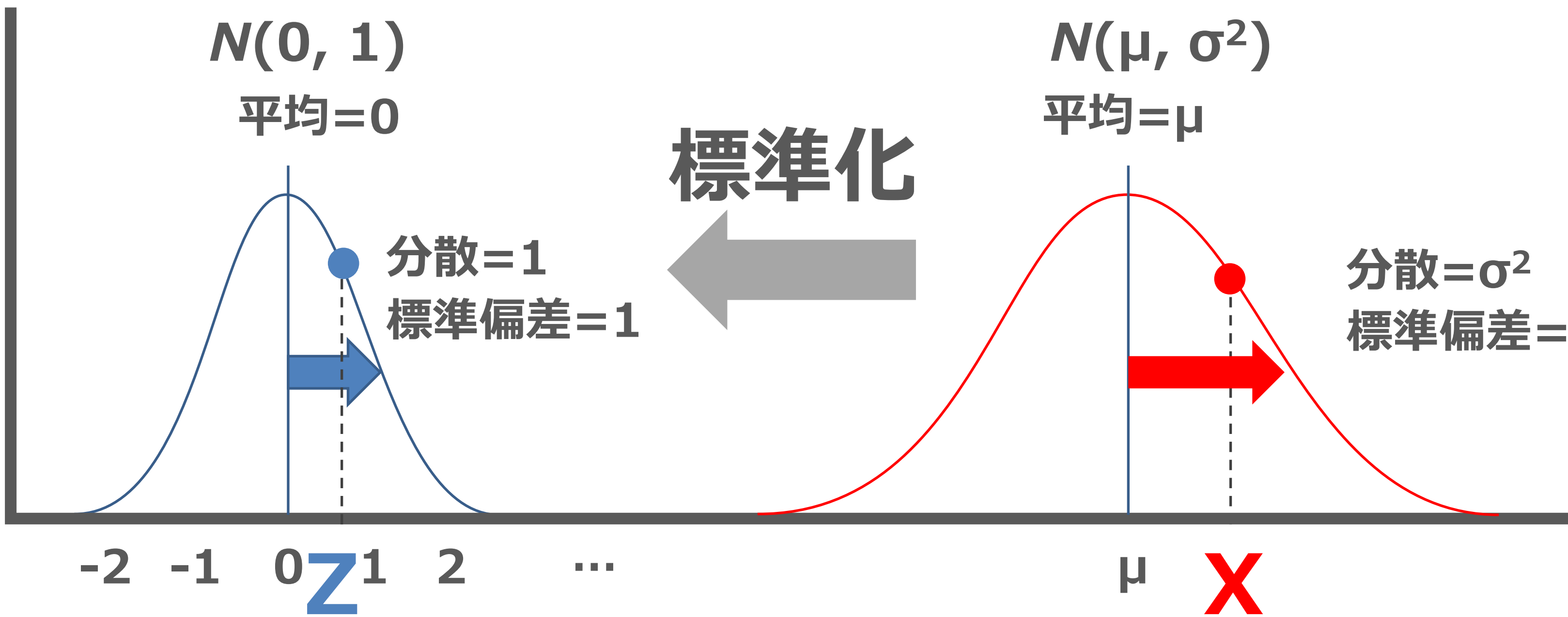


$N(0, 1)$

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、

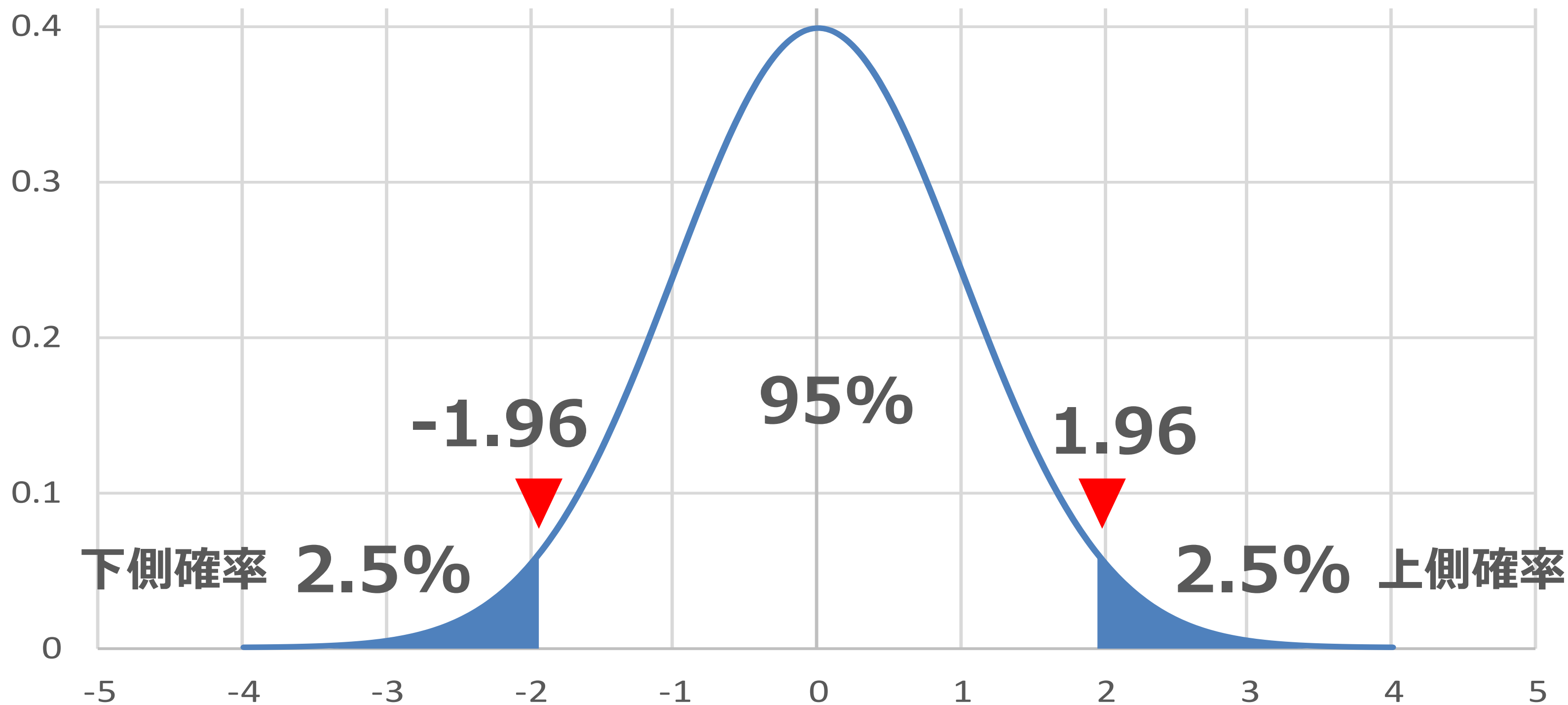
$$Z = \frac{X - \mu}{\sigma}$$ と変換すると、標準正規分布になる。



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

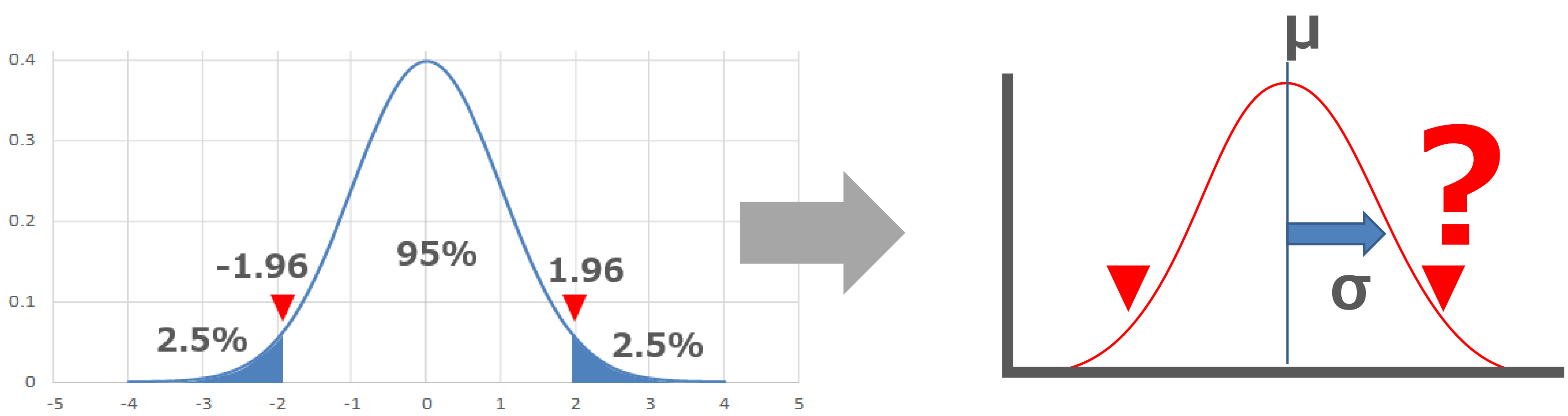
u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
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
						.00539	0.00523	0.00508	0.00494	0.00480
						.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

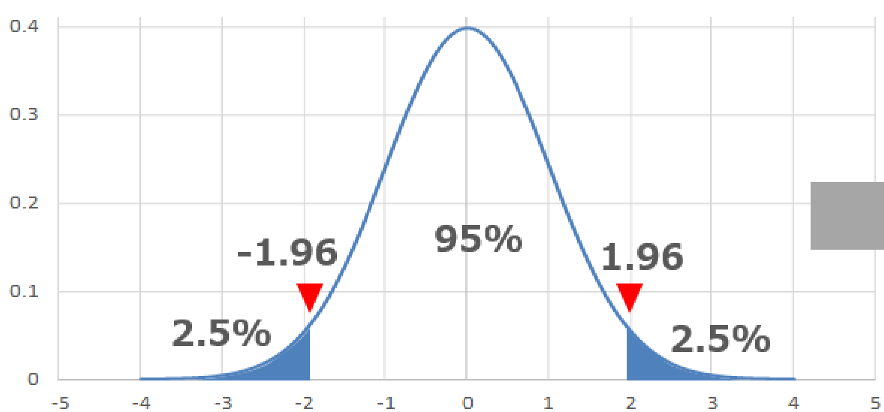
$$Z = \frac{X - \mu}{\sigma}$$

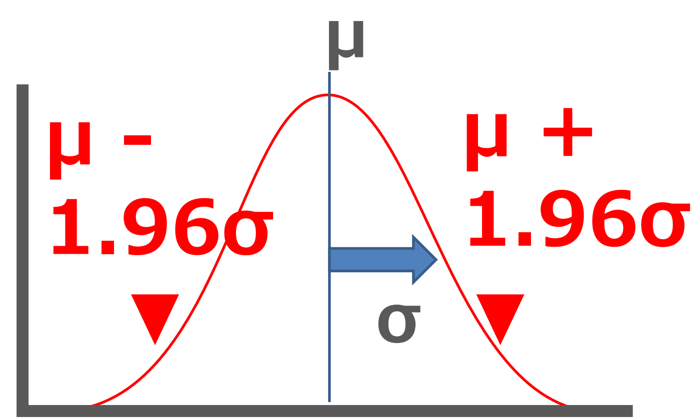
標準化の逆

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

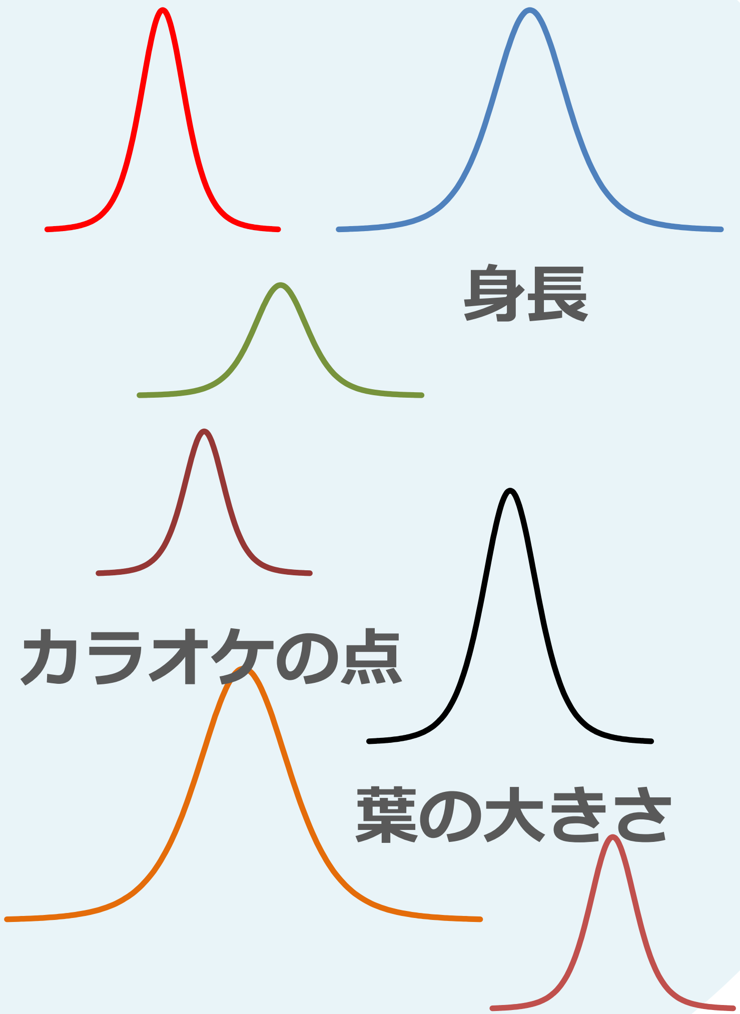
例）$Z = 1.96$なら、

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





今年



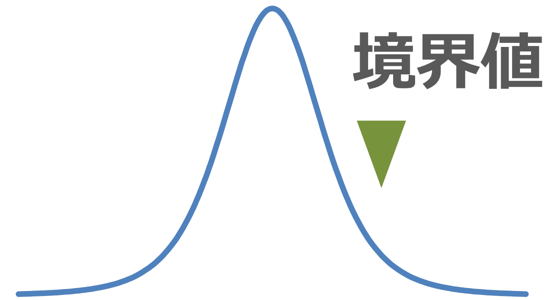
いろんな正規分布

z変換

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

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

逆変換して
適用



✓ **標準**正規分布

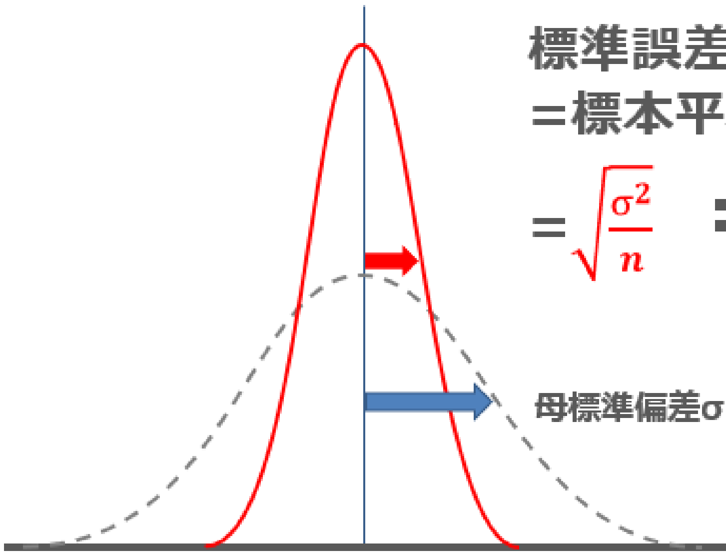
境界値と確率
を考える

現実の具体的な問題

一般化

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



標準誤差
＝標本平均$\bar{x}$の標準偏差

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

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 $\bar{x}$

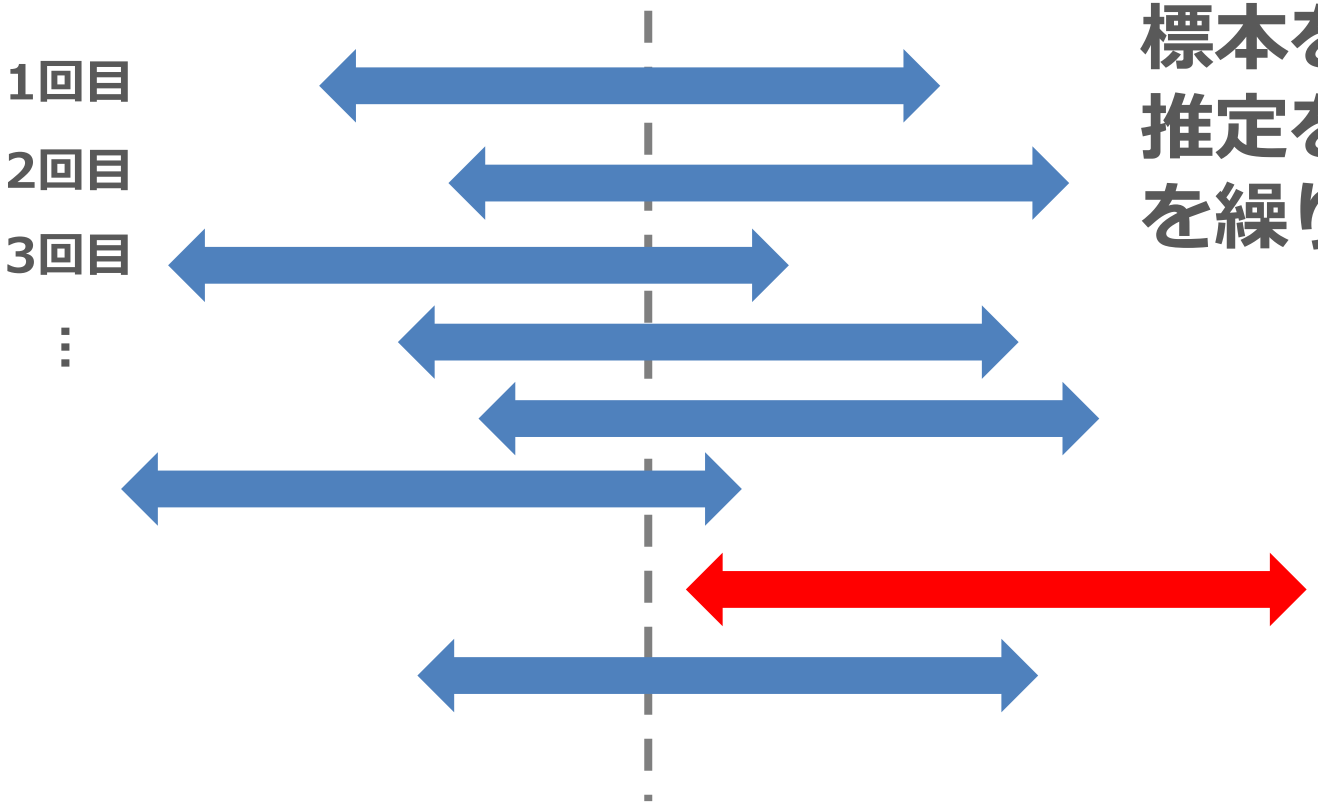
推定値の標準偏差：　標本平均の標準偏差 $\frac{\sigma}{\sqrt{n}}$

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、母平均がその区間内に含まれるのが95回になる」

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

真の母平均
（変わらない）

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σ²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。

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

ここでAは、標準正規分布表から、

α(信頼係数) = (100-a)/2/100

で求められる境界値

ただし…

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

母分散σ²は不明な場合がほとんど

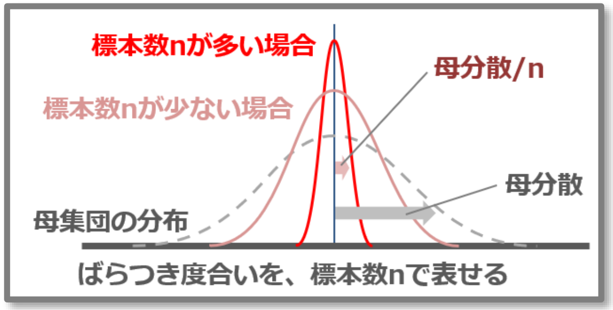
母平均μが不明（推定したい）のに母分散σ²だけ分かっているって、
どういうこと？　そんな状況はほとんどない！

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

不偏標本分散に置き換える

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

では、その時の分布は？　正規分布でいいの？
特にnが小さいとき、$\bar{x}$ とμのずれは大きいが…



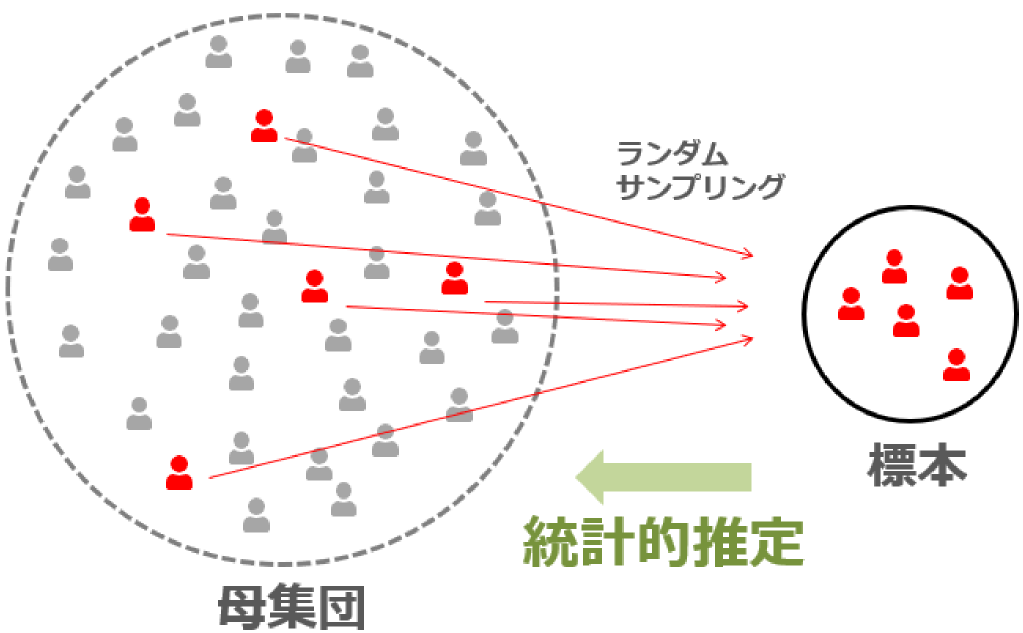
中心極限定理

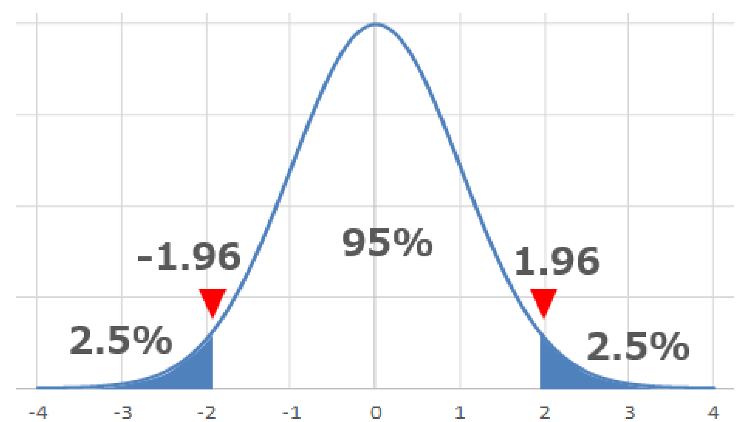
t分布

*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

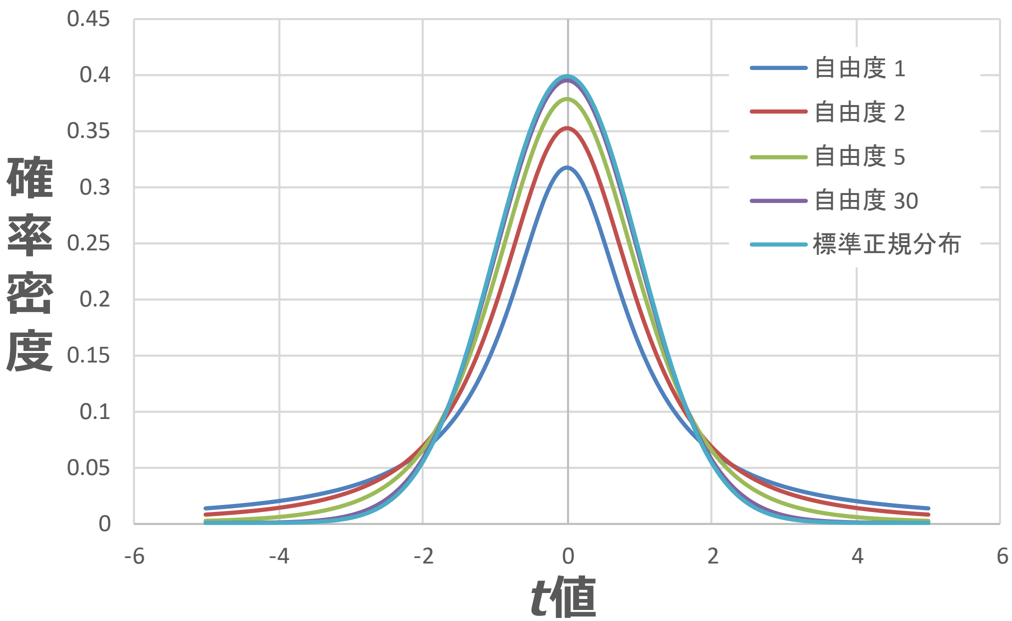
出典：Wikipedia

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：ギネス社HP

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
[illegible]	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v²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。

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

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

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。

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

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。

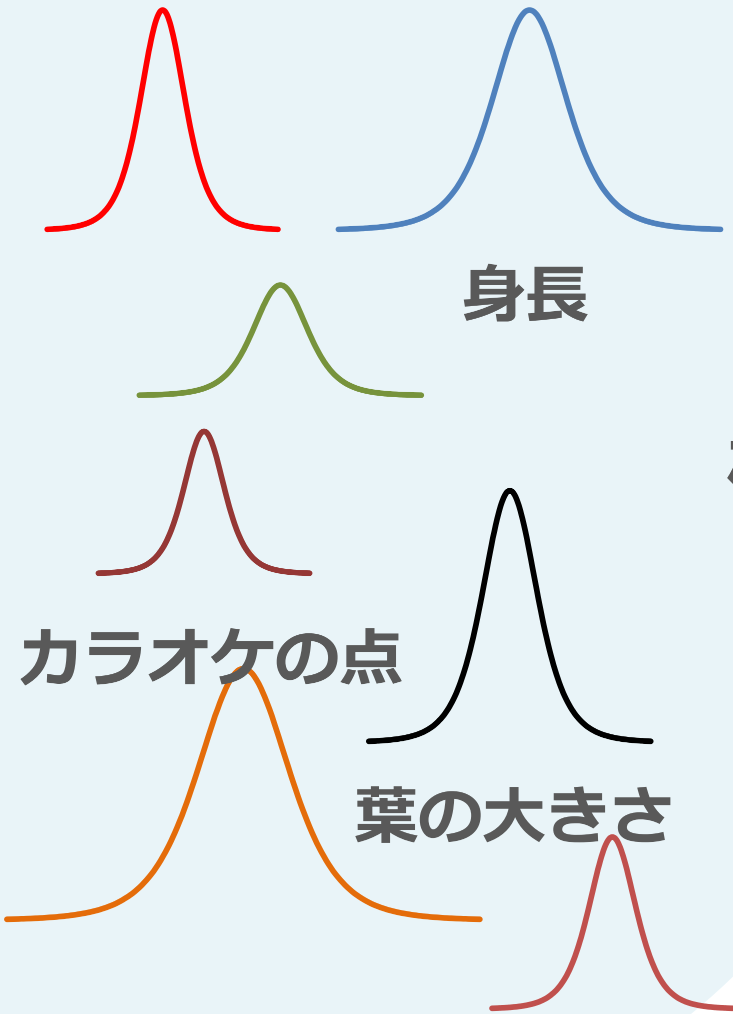
今年

z変換

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

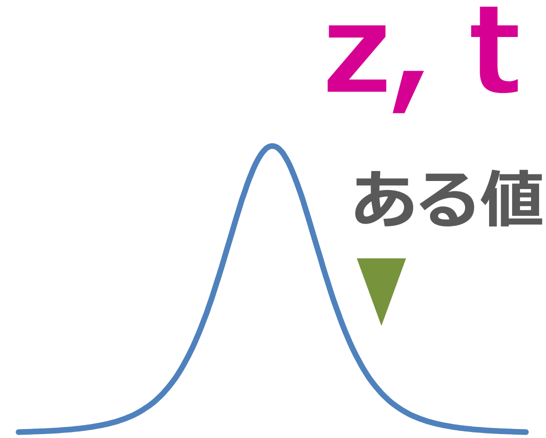
検定統計量

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



いろんな正規分布

現実の具体的な問題



- ✓ 標準正規分布
- ✓ *t*分布

一般化

逆変換して
適用

【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 (ニュー) ：自由度

スケジュール

今年

- 検定の基礎
  - ✓ 統計の大事な考え方
    平均値〜t分布まで
  - ✓ t検定
    検定のやりかた
    気を付けること
  - ✓ いろんな検定とANOVA
- 多変量解析のイメージ
- 多変量解析の実習

午前中

〜90分

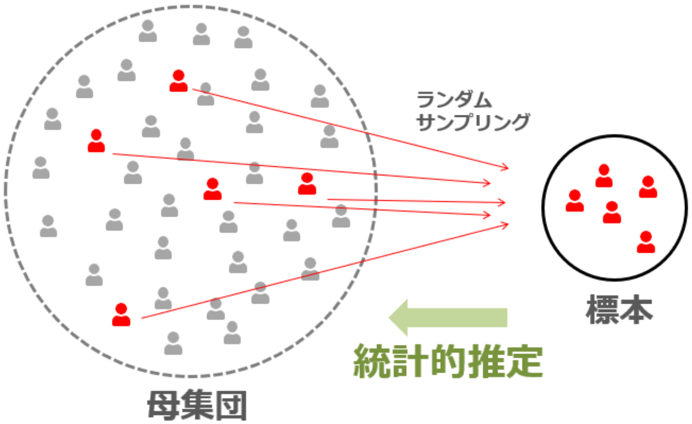
〜60分

〜夕方まで

t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

検定

帰無仮説○○は、有意水準△△で棄却されました。したがって、□□という結論を得ました。

みたいなやつ

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。

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

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

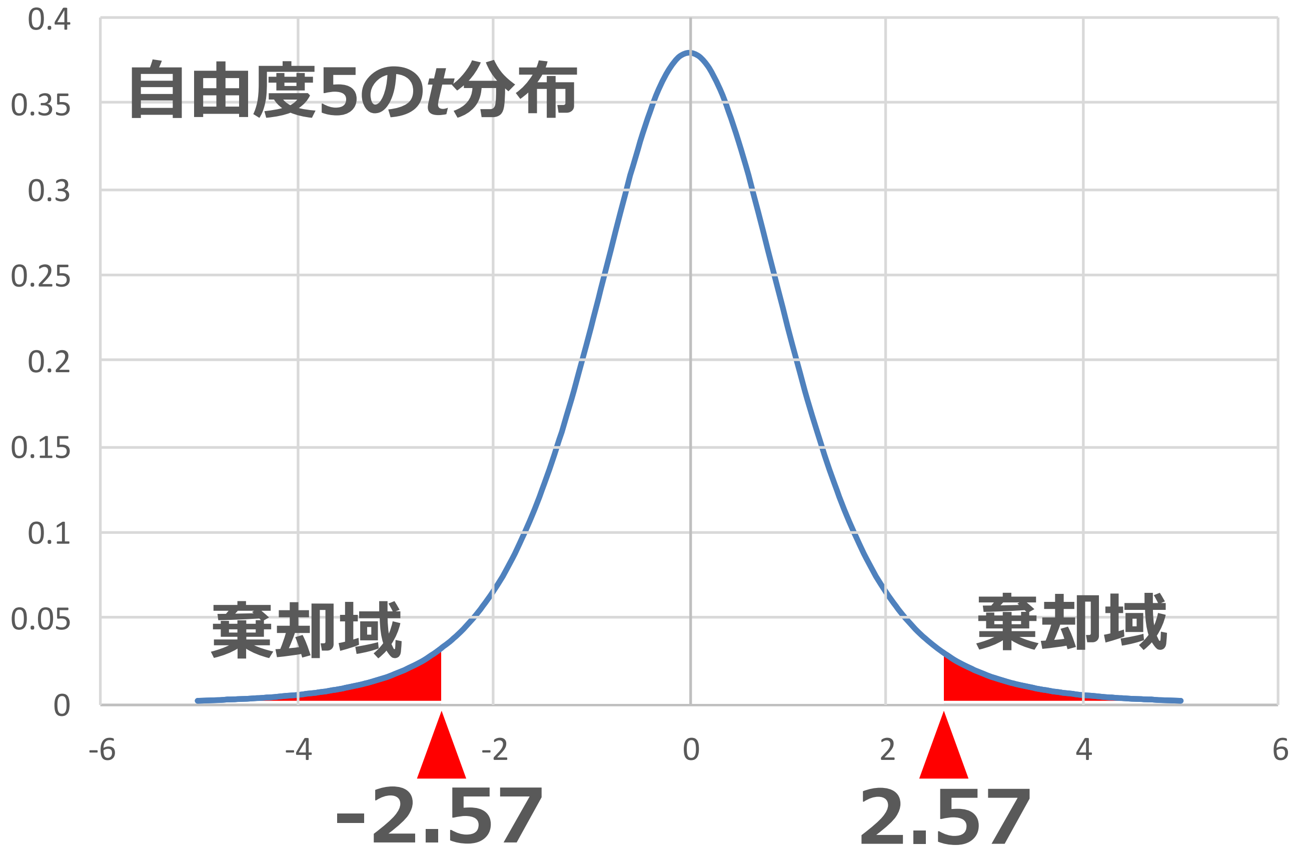
有意水準α

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

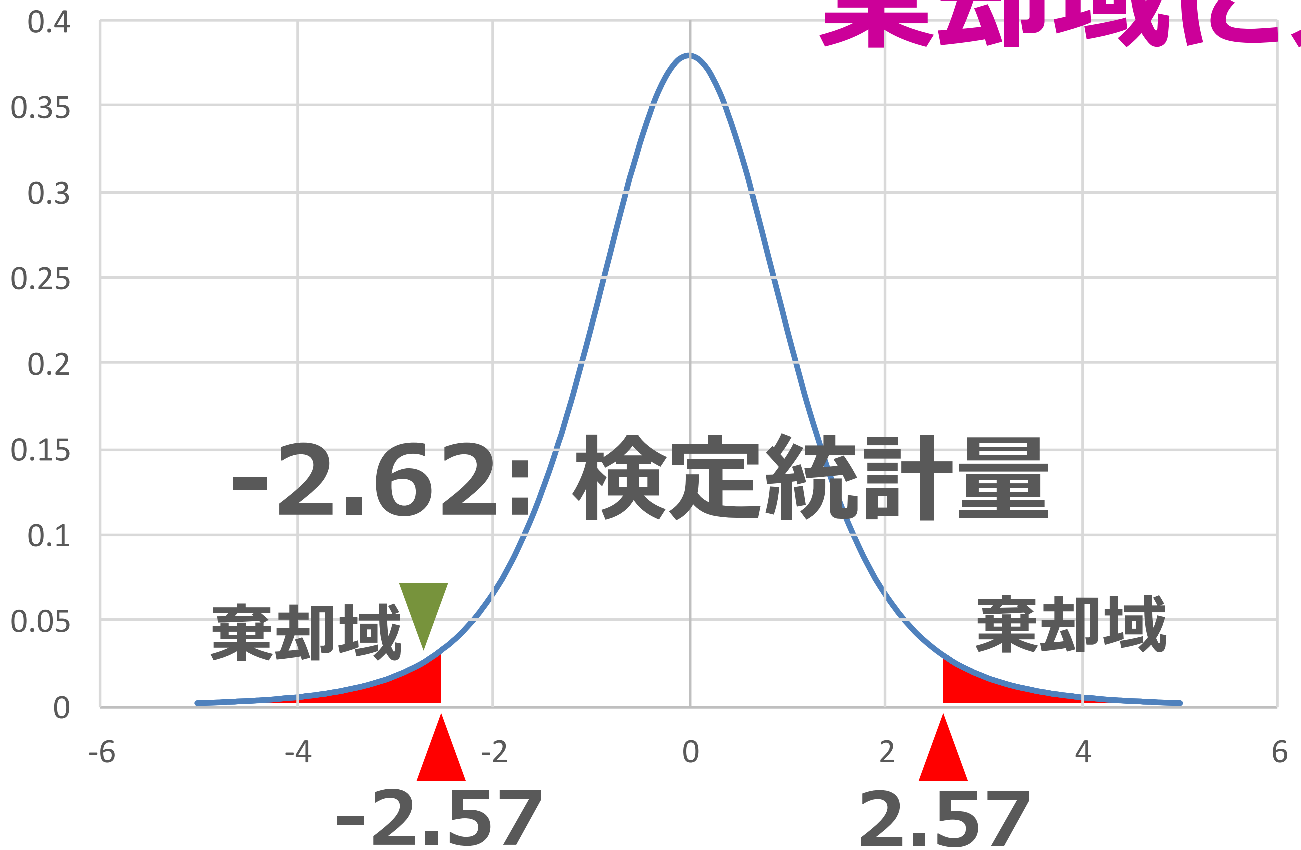


Excelで計算してもよい

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

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。

検定と確率分布との関係

①仮説を立てる

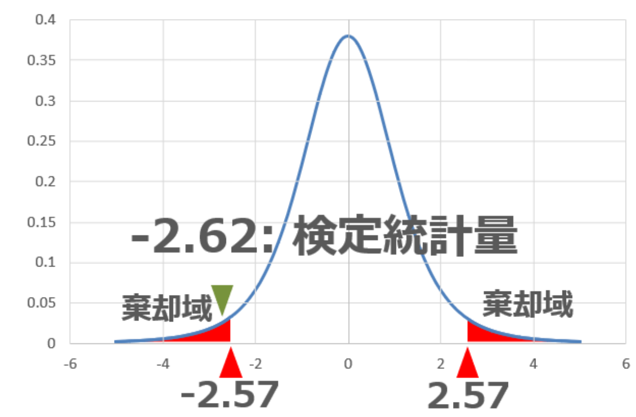
主張したいことを「対立仮説」に

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？

注意点

帰無仮説が棄却されないとき…

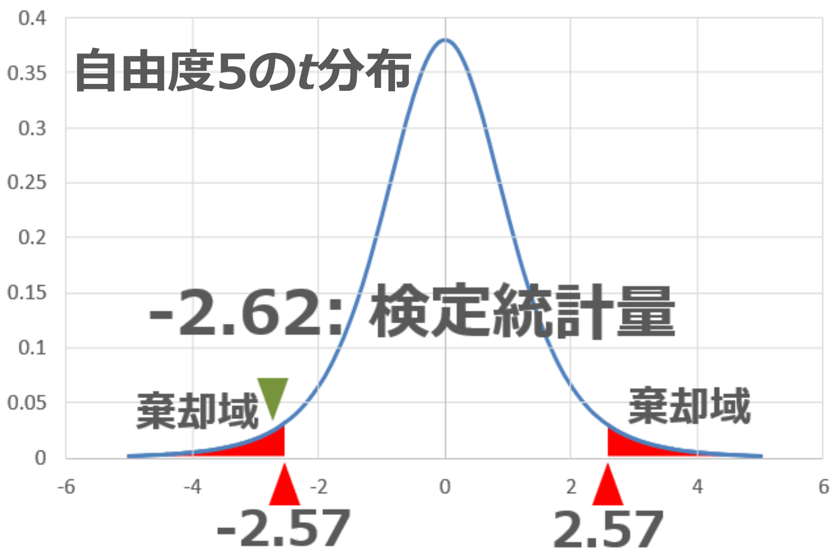
「帰無仮説が正しい」と安易に結論付けてはいけない。 ×

「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、α= 0.025の時に計算された値。*t*値2.62より外側の面積（**p値=0.024**）も、この分布から求めることができる。

※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。

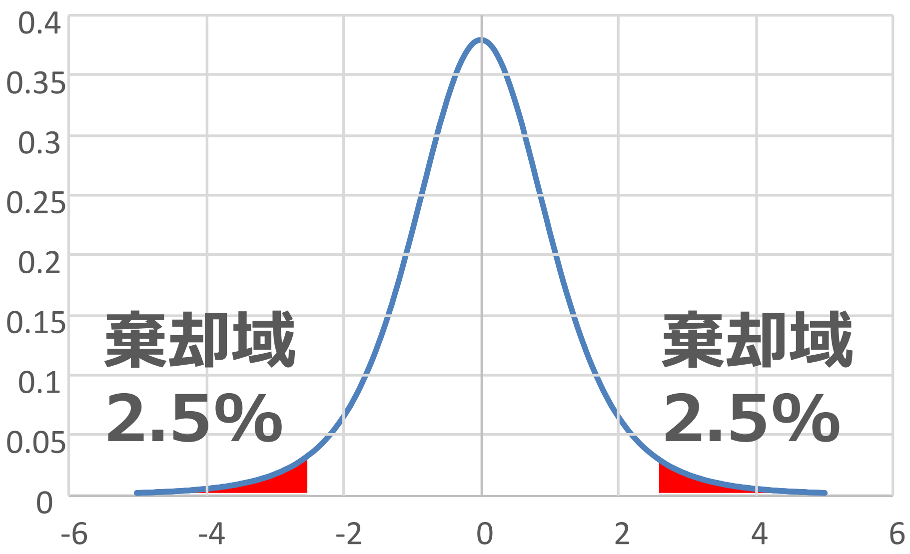
検定では、確率的にまれに起こる事象かどうか、つまり「意味ありげ（有意）」かどうかを調べるからです。

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。

両側検定と片側検定

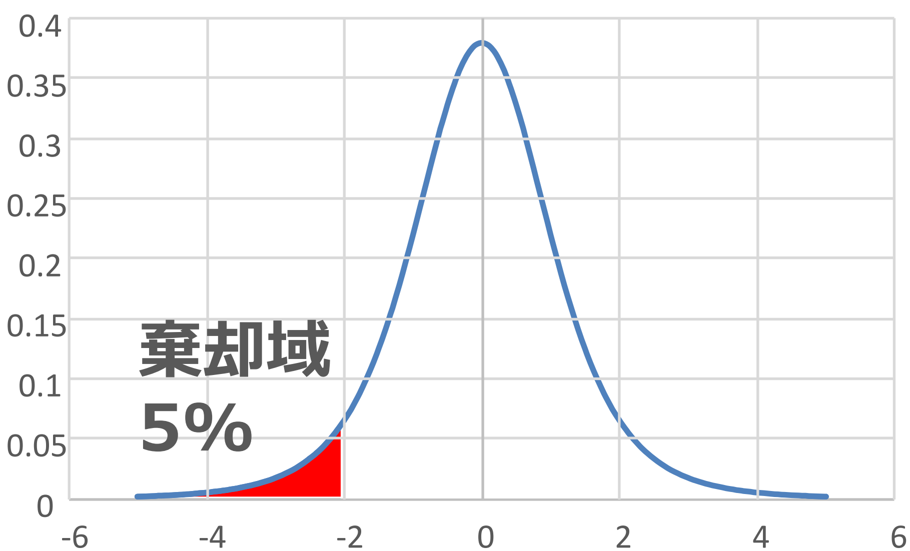
Aさんのカラオケ問題では、有意水準0.05を、その半分の0.025ずつに分け、*t*分布の両側に割り当てて考えました

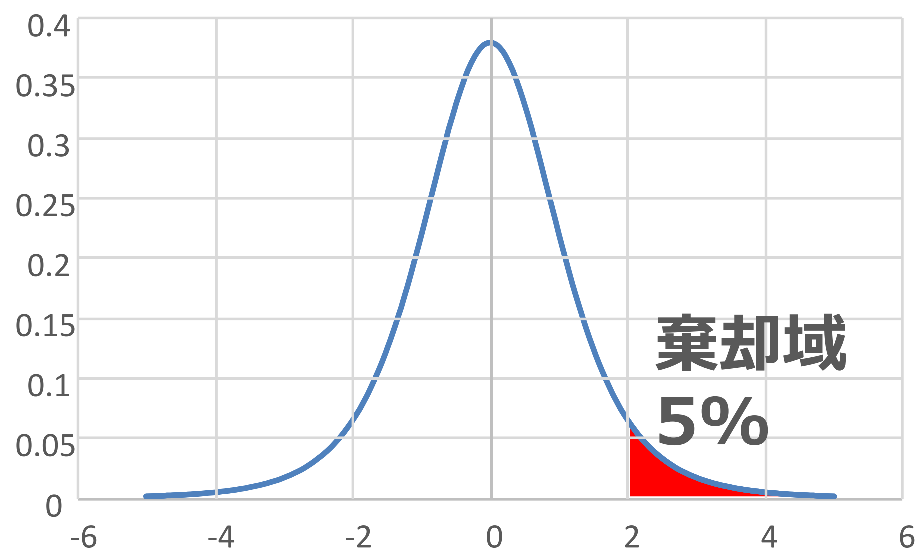


これは
両側検定
と呼ばれます

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, 不偏標本分散 s1, 標本平均$\overline{x1}$

2群目：標本数 n2, 不偏標本分散 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, 不偏標本分散 s1, 標本平均$\overline{x1}$

2群目：標本数 n2, 不偏標本分散 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　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 (誤った姿)
検定結果	棄却する (陽性)	1-β (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-β）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。

スクリーニング検査

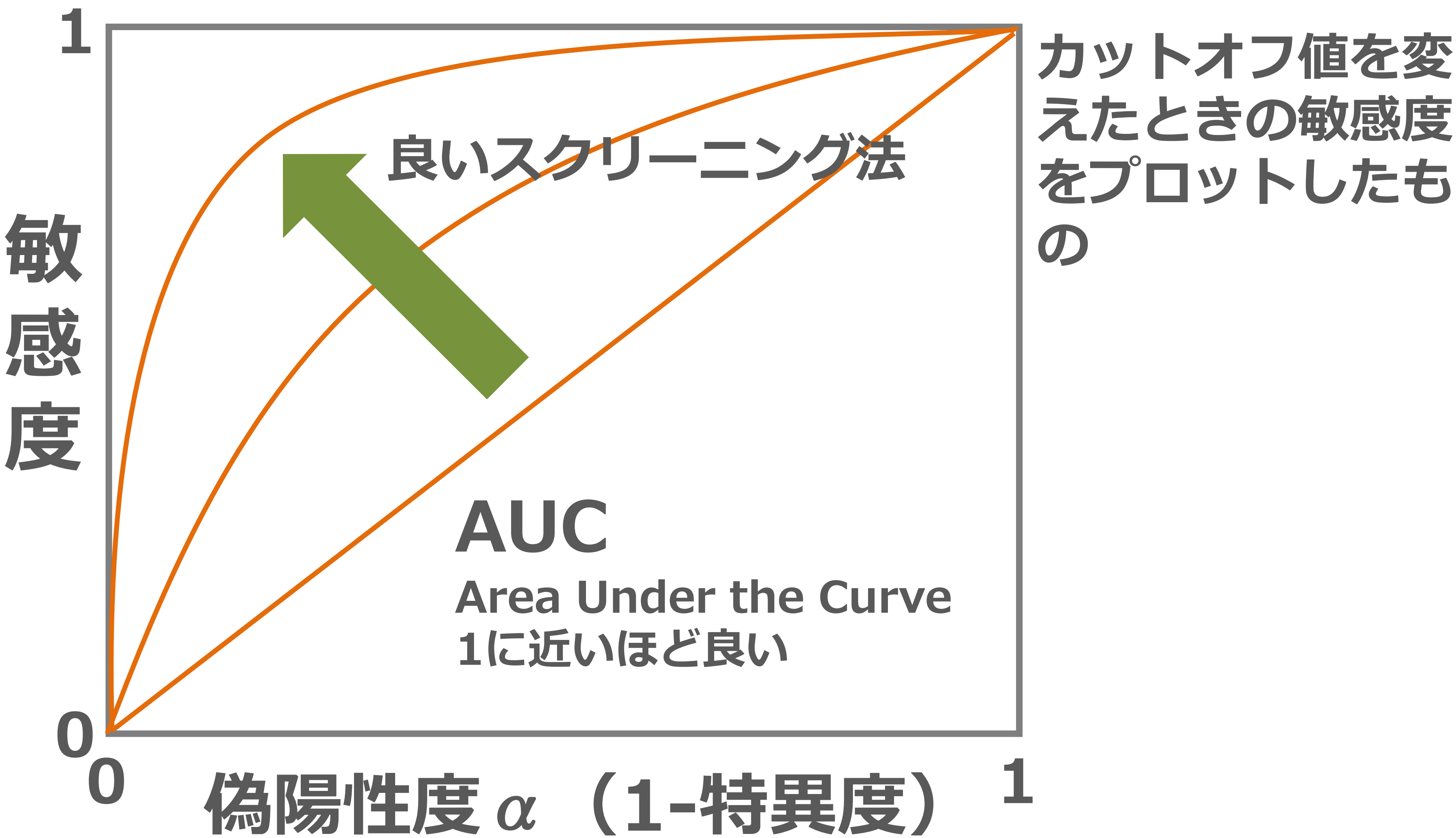
＊＊のスクリーニング		本当は	
		病気	健康
結果	陽性（+）	真陽性 True Positive 敏感度	偽陽性 False Positive 偽陽性度
	陰性（-）	偽陰性 False Negative 偽陰性度	真陰性 True Negative 特異度

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。

ROC曲線

Receiver Operating Characteristic curve



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

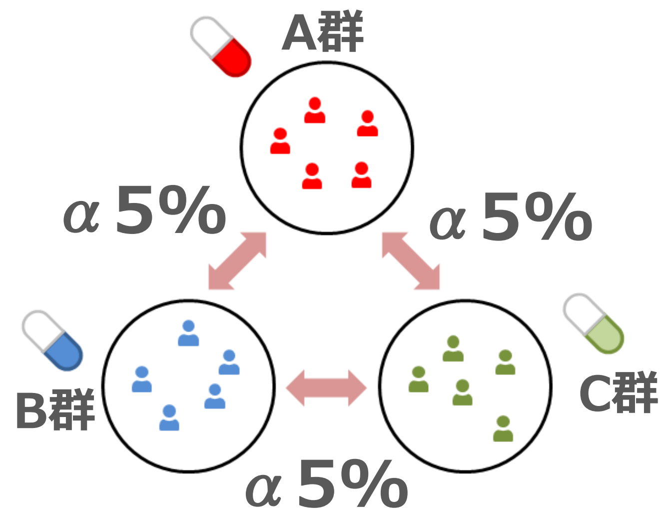
検定を繰り返すと、誤りが大きくなる

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！

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。

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

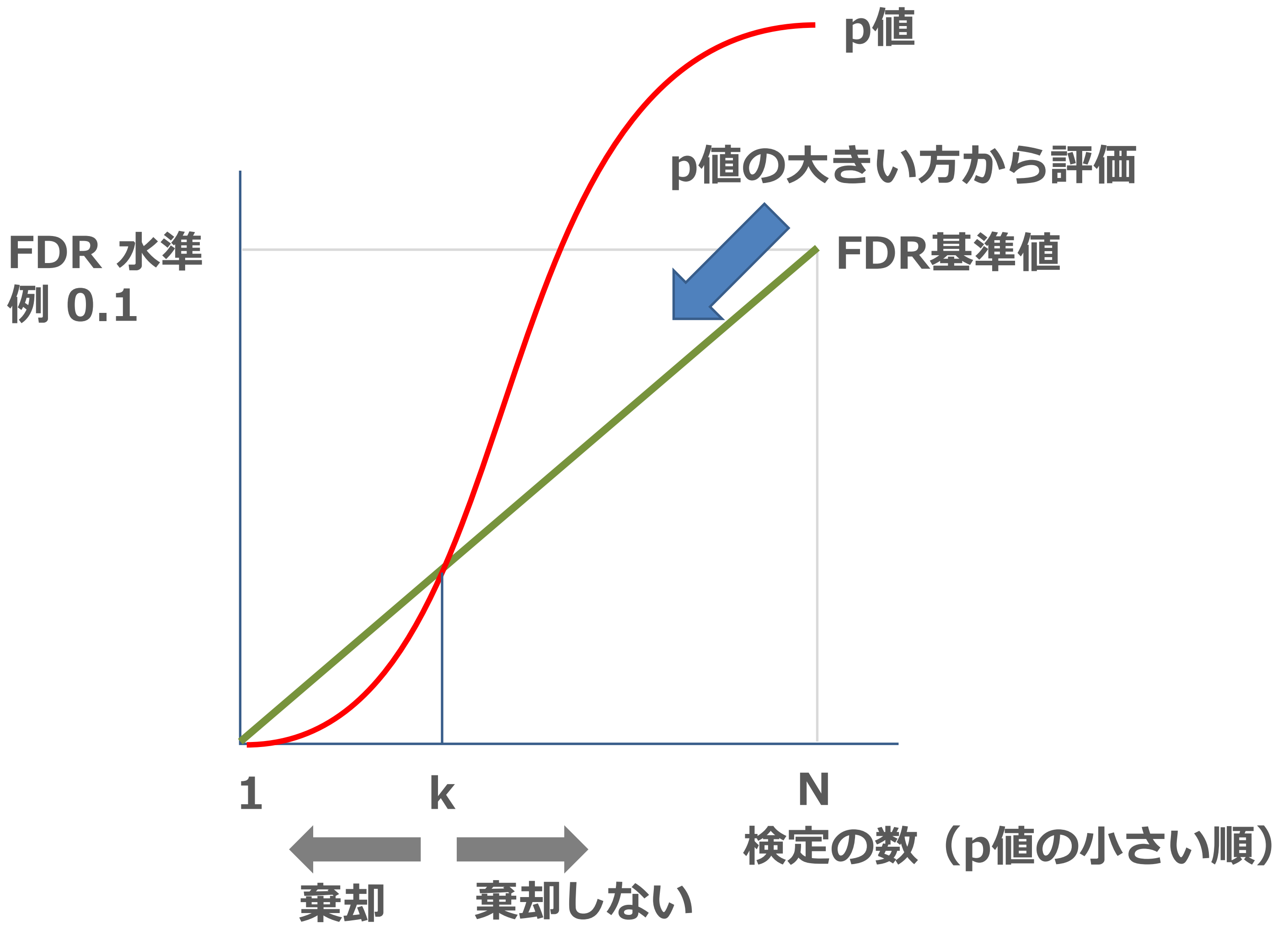
FDR q = v/R: 棄却したもののうち、偽陽性の率
これを、一定水準（例えば0.05）にする方法

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

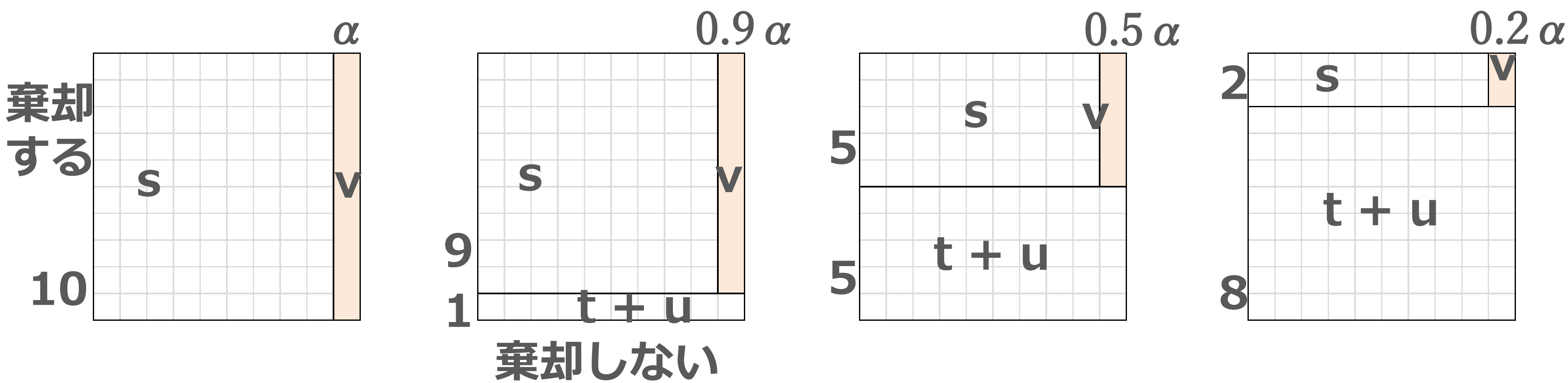


FDR調整のイメージ　by 櫻井

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 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

スケジュール

今年

- 検定の基礎
  - ✓ 統計の大事な考え方
    平均値～t分布まで
  - ✓ t検定
    検定のやりかた
    気を付けること
  - ✓ いろんな検定とANOVA
- 多変量解析のイメージ
- 多変量解析の実習

午前中

～90分

～60分

～夕方まで

いろいろな検定

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

F検定　等分散性の検定

1群目：標本数 n_1, 不偏標本分散 $v^2{}_1$

2群目：標本数 n_2,不偏標本分散 $v^2{}_2$

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

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。

分散分析

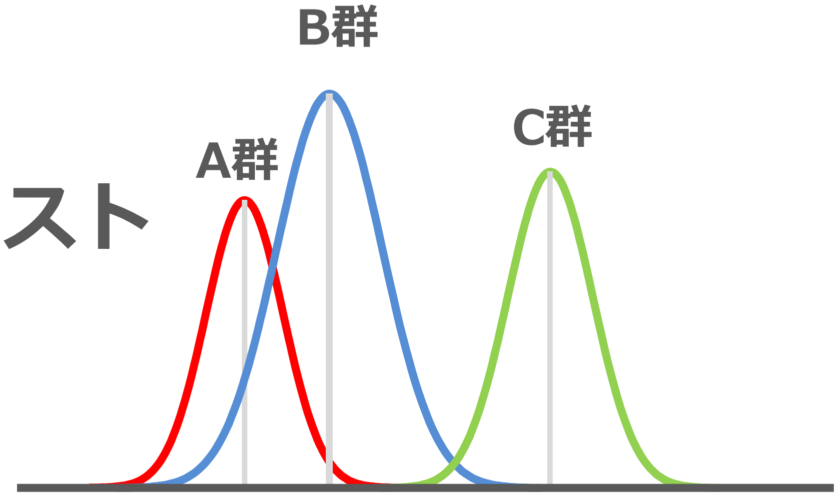
Analysis of Variance

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

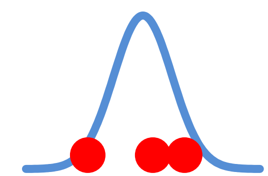
例）1組、2組、3組で、テストの平均点に差があるか？



F分布

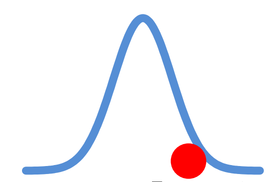
複数の変数を含む2群
（分散の比を考える）

カイ二乗分布



複数の変数
（分散を考える）

標準正規分布



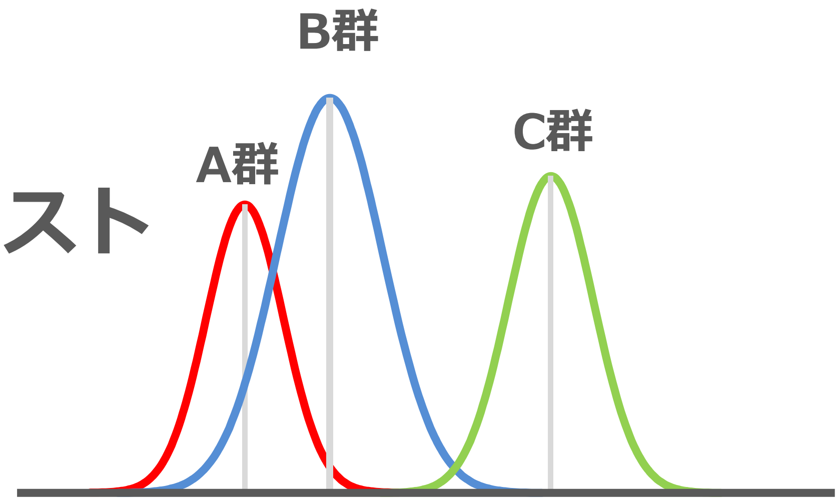
ひとつの変数

by 櫻井

- ✓ 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群の母平均は等しくない（異なる値が含まれる）

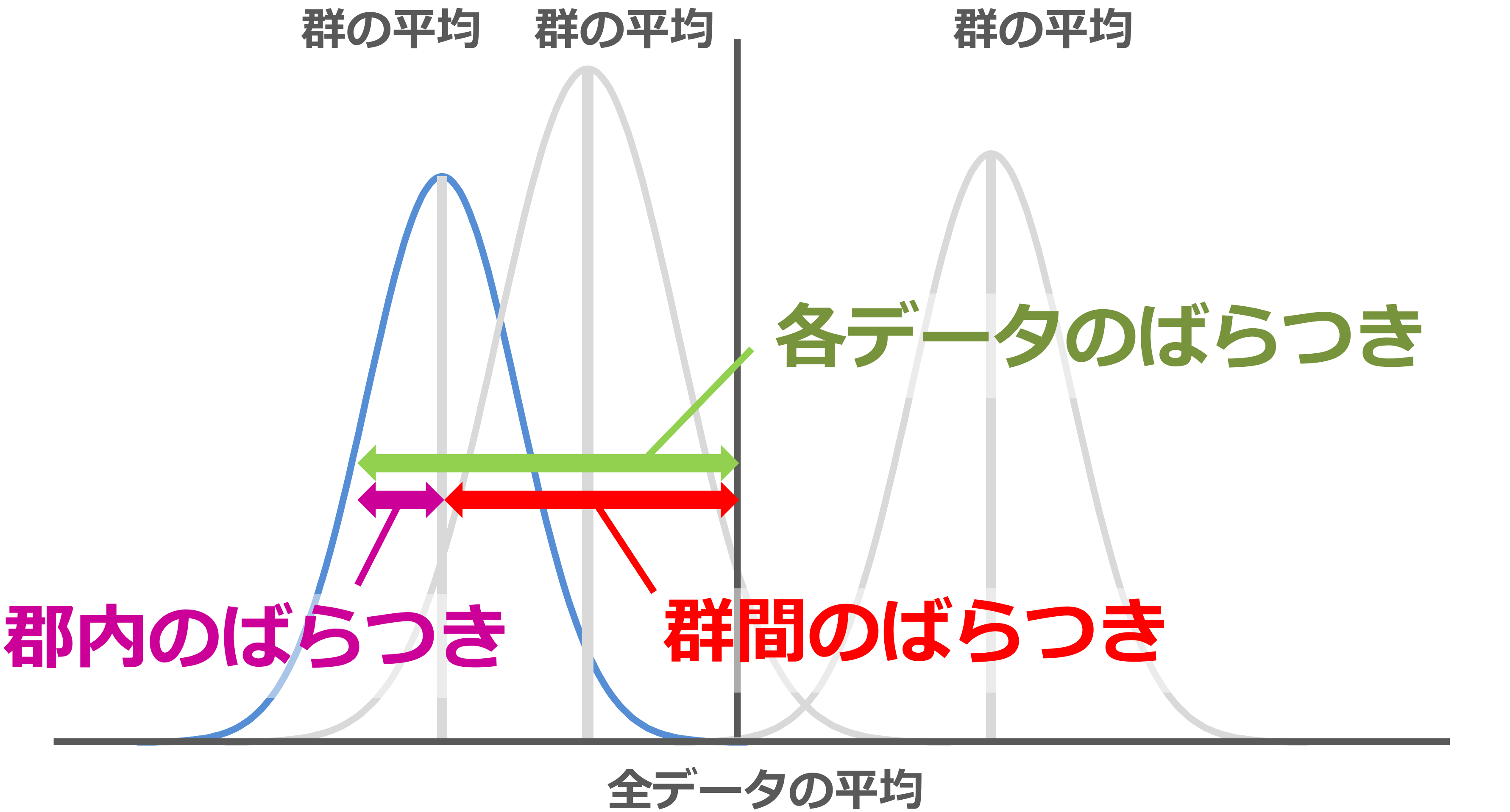
↓

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

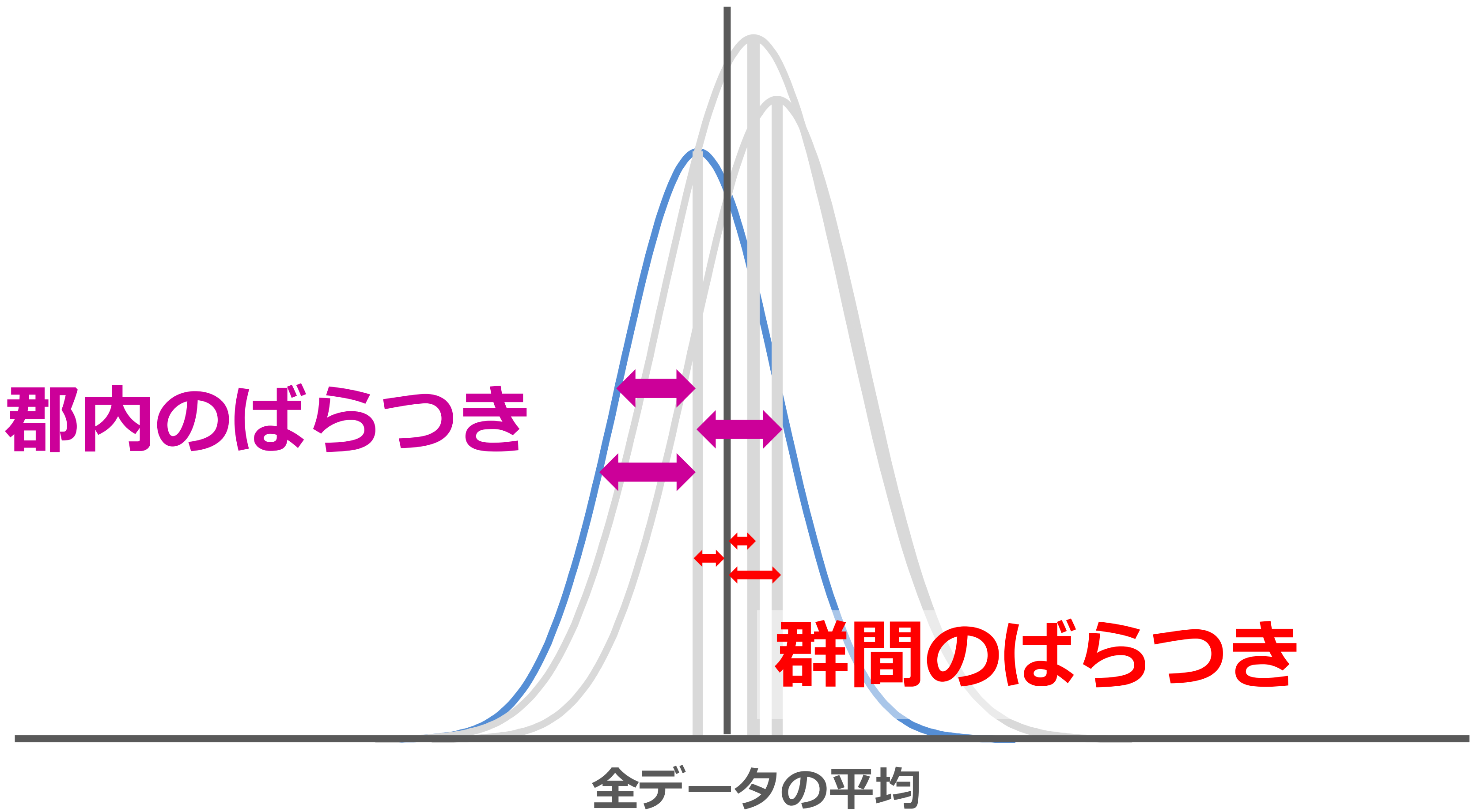
分散分析のイメージ

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



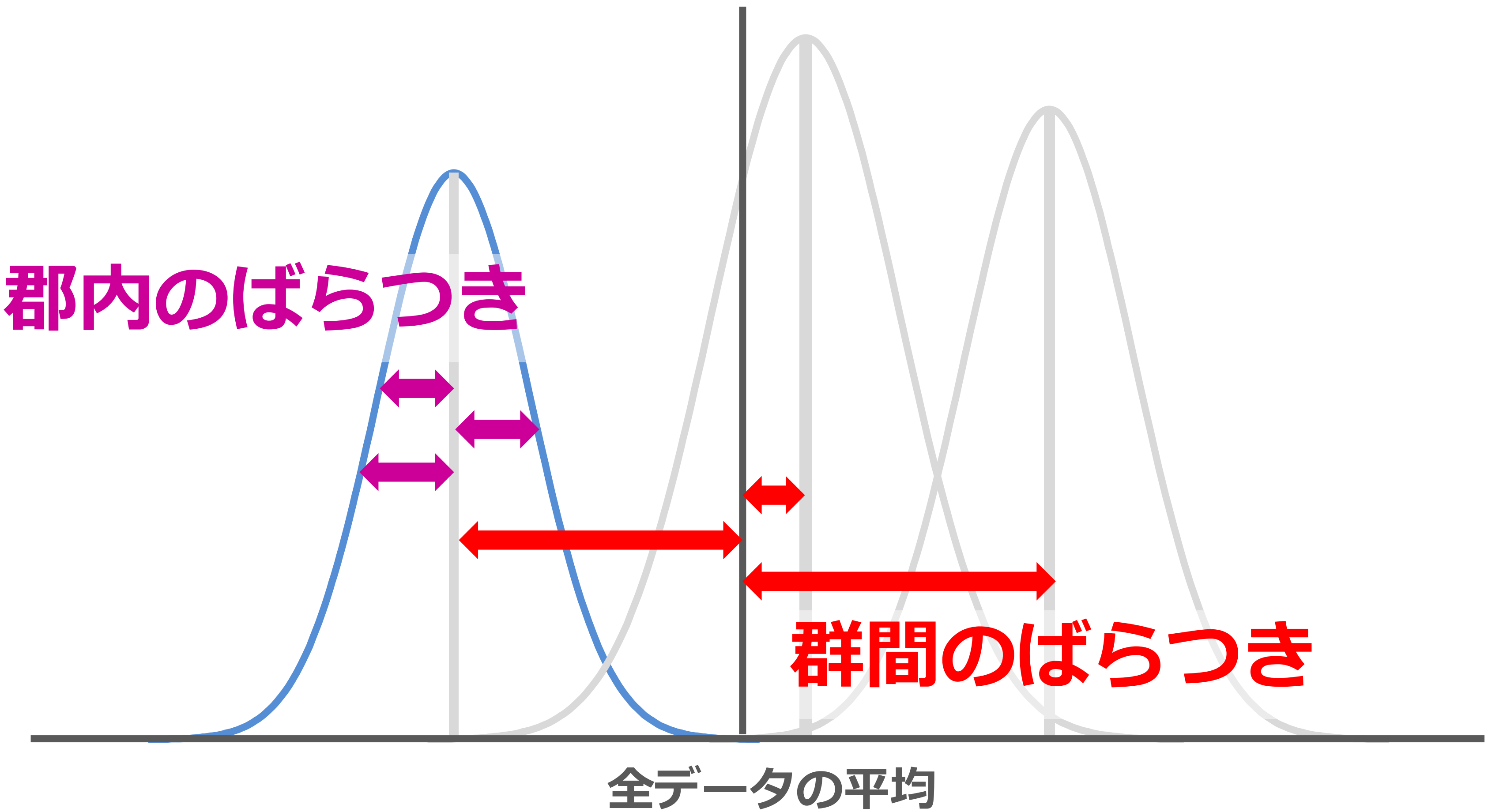
分散分析のイメージ

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	

210915_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
　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

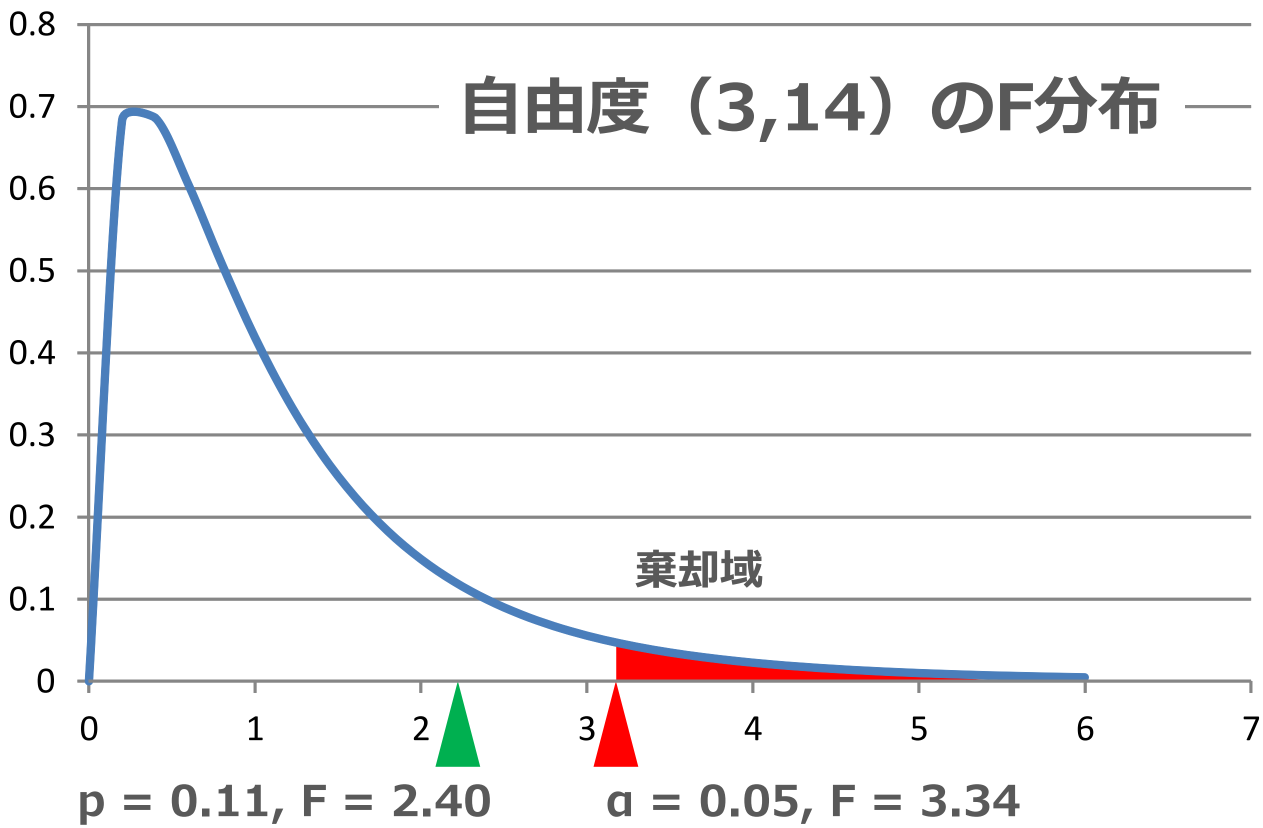
⑳F.DIST関数を用いて当該自由度のF分布を描く

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

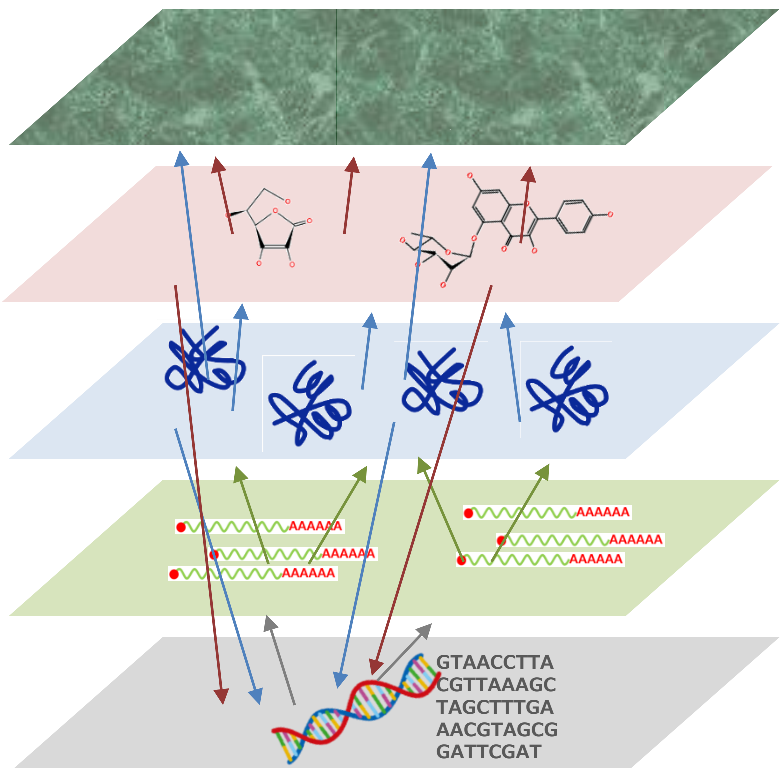
多変量解析

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



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

扱うデータ

生体試料など

		対象				
		1	2	3	$\cdots$	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}
	$\cdots$					
	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

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

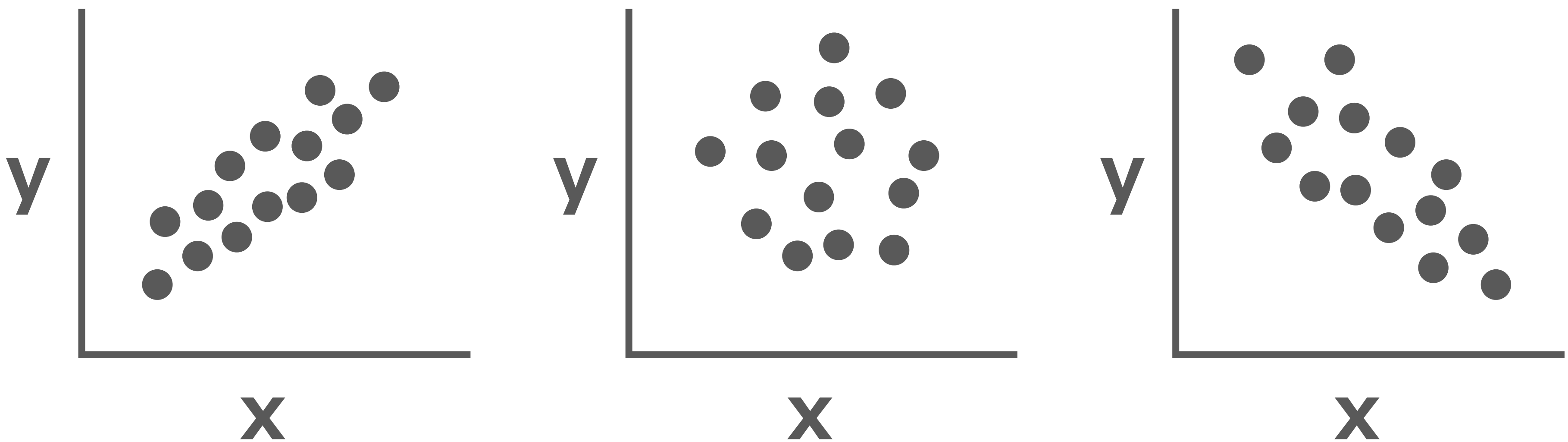
- データの類似性を考える
  （相関解析）
- 似ているものをグルーピングする
  クラスター解析
- データを要約する
  主成分分析
- 判別、分類、予測
  判別分析、PLS、PLS-DA、
  重回帰分析

など

相関

散布図

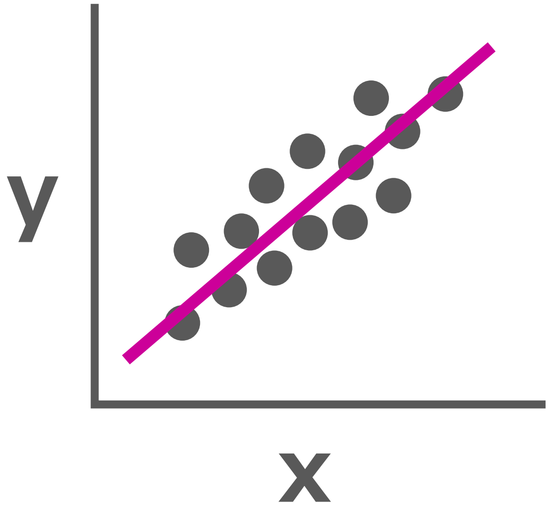
二つのデータ間の関係性を見える化する手法



正の相関がある　相関がない　負の相関がある

散布図の回帰曲線

データに最もフィットする関数



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

- 二つのデータ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

目安：

0.7～1.0：強い正の相関	-1.0～-0.7：強い負の相関
0.4～0.7：中程度の正の相関	-0.7～-0.4：中程度の負の相関
0.2～0.4：弱い正の相関	-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、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、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。

正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。

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：

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

※|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

そのほかの相関係数

- スピアマンの順位相関係数
- コサイン類似度

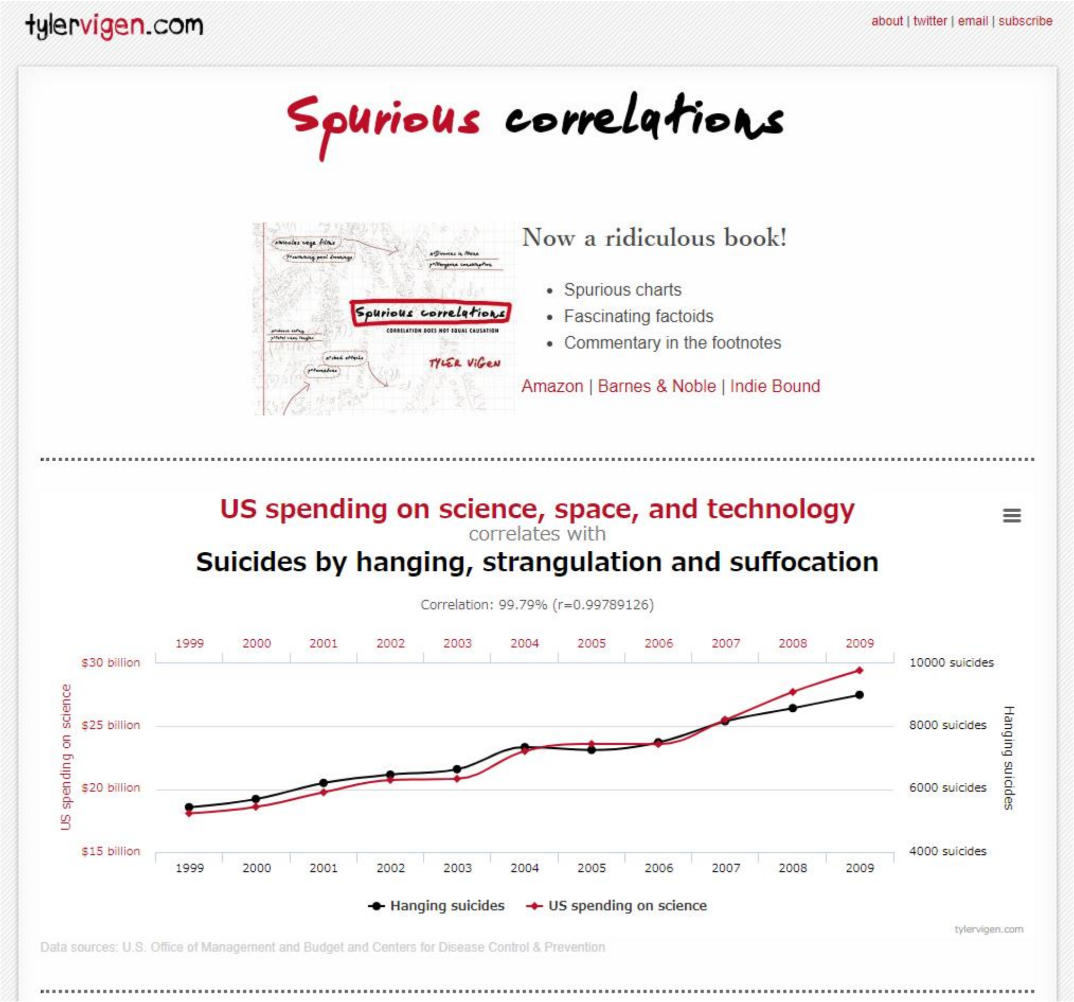
相関と因果

相関関係：
二つの事柄に関連性がある

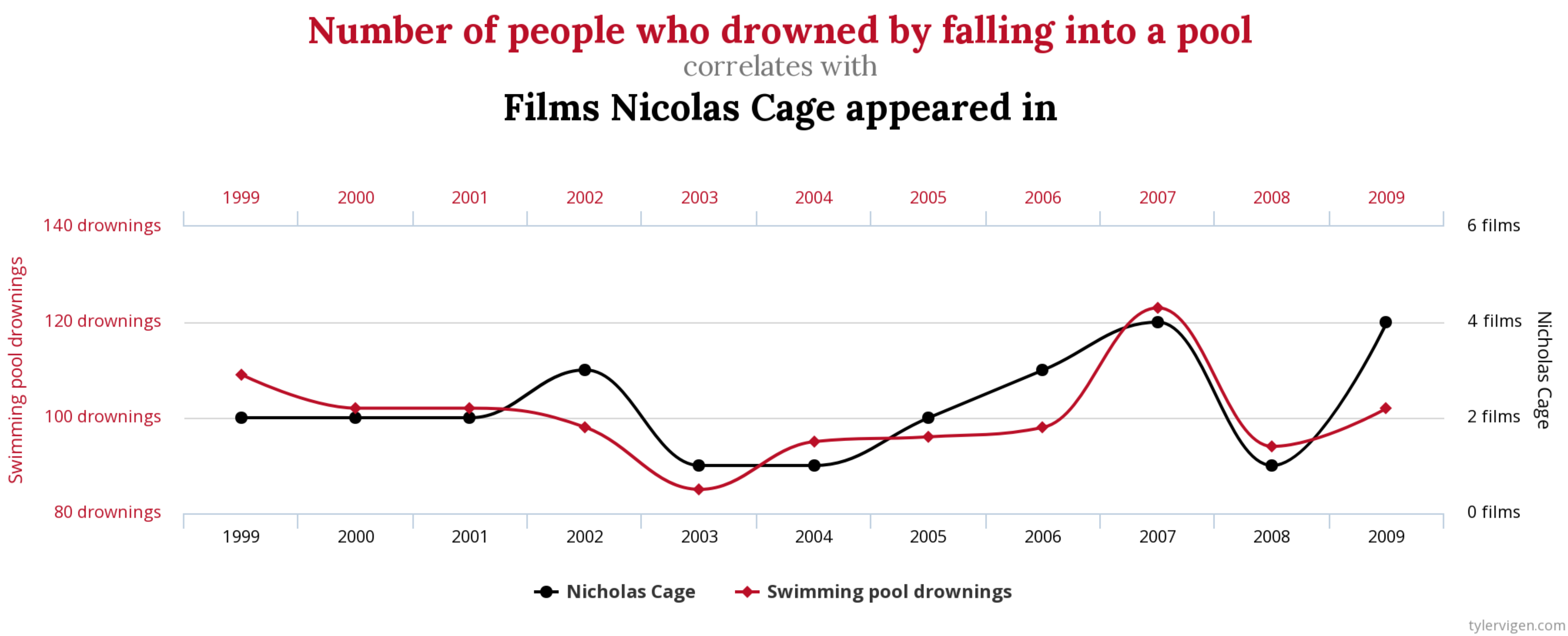
因果関係：
二つの事柄が、原因と結果の関係である

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

さまざまな多変量解析

- データの類似性を考える
  （相関解析）
- 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

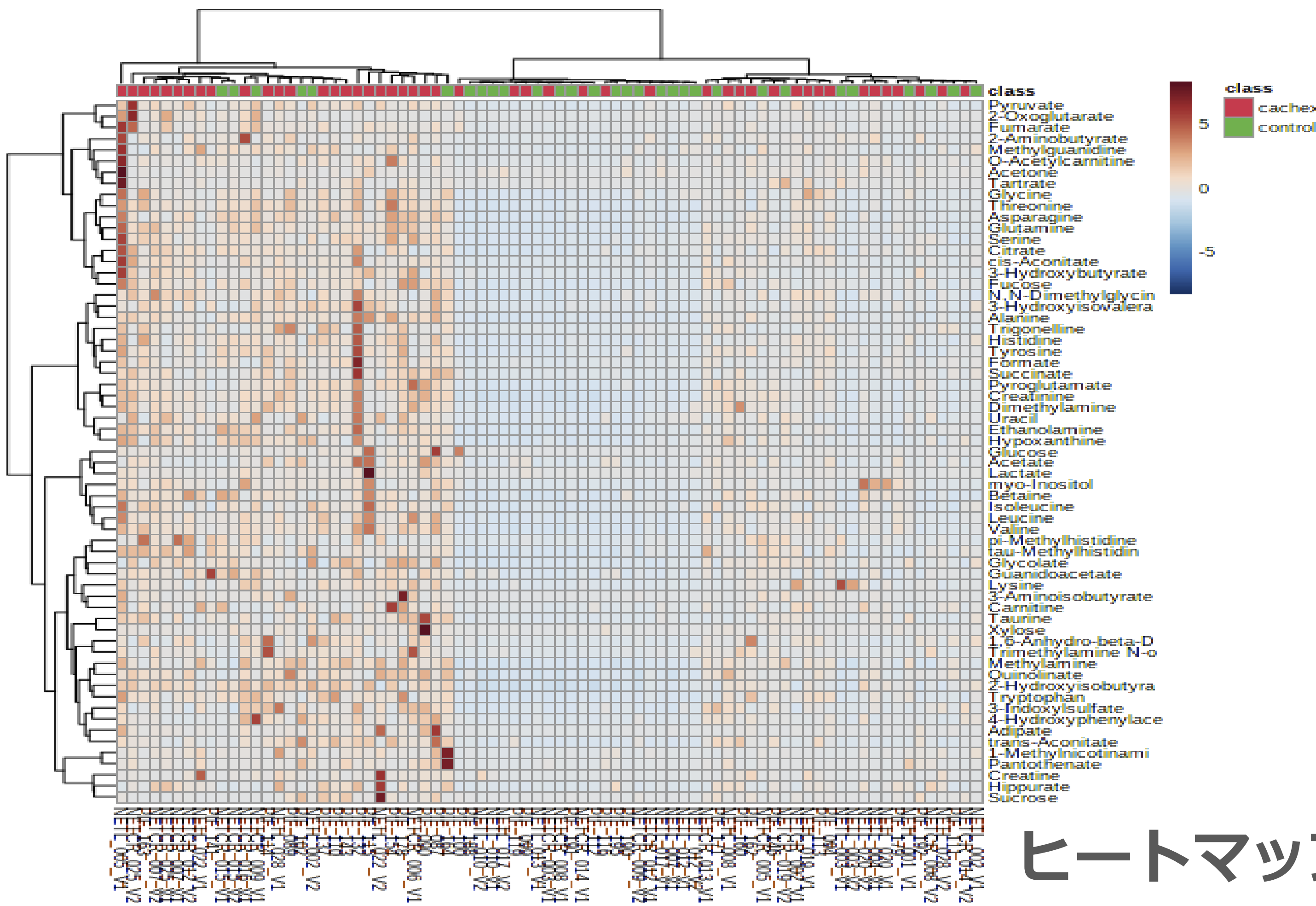
階層的クラスター解析

Hierarchical Cluster Analysis

似ているものを近くに配置する

サンプル

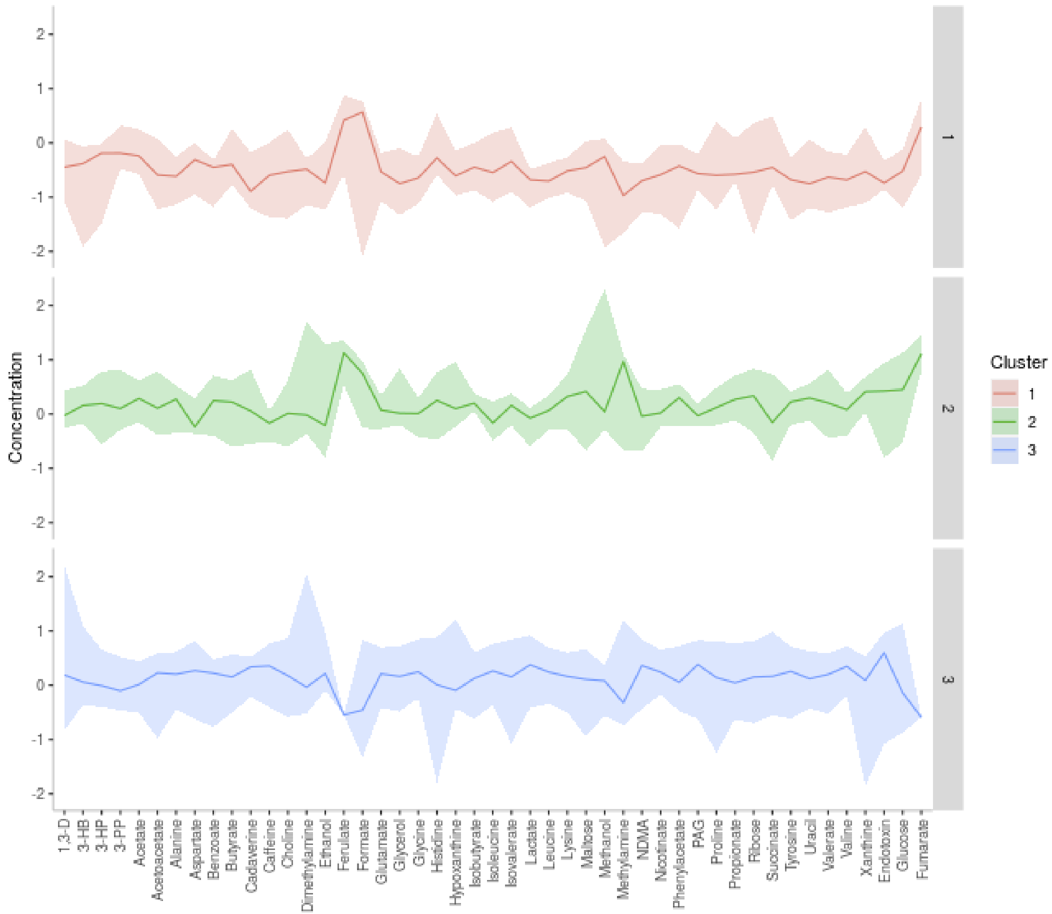
検出
化合物



ヒートマップ

k-平均法 (k-means)

似ているものをk個のパターンに分類する

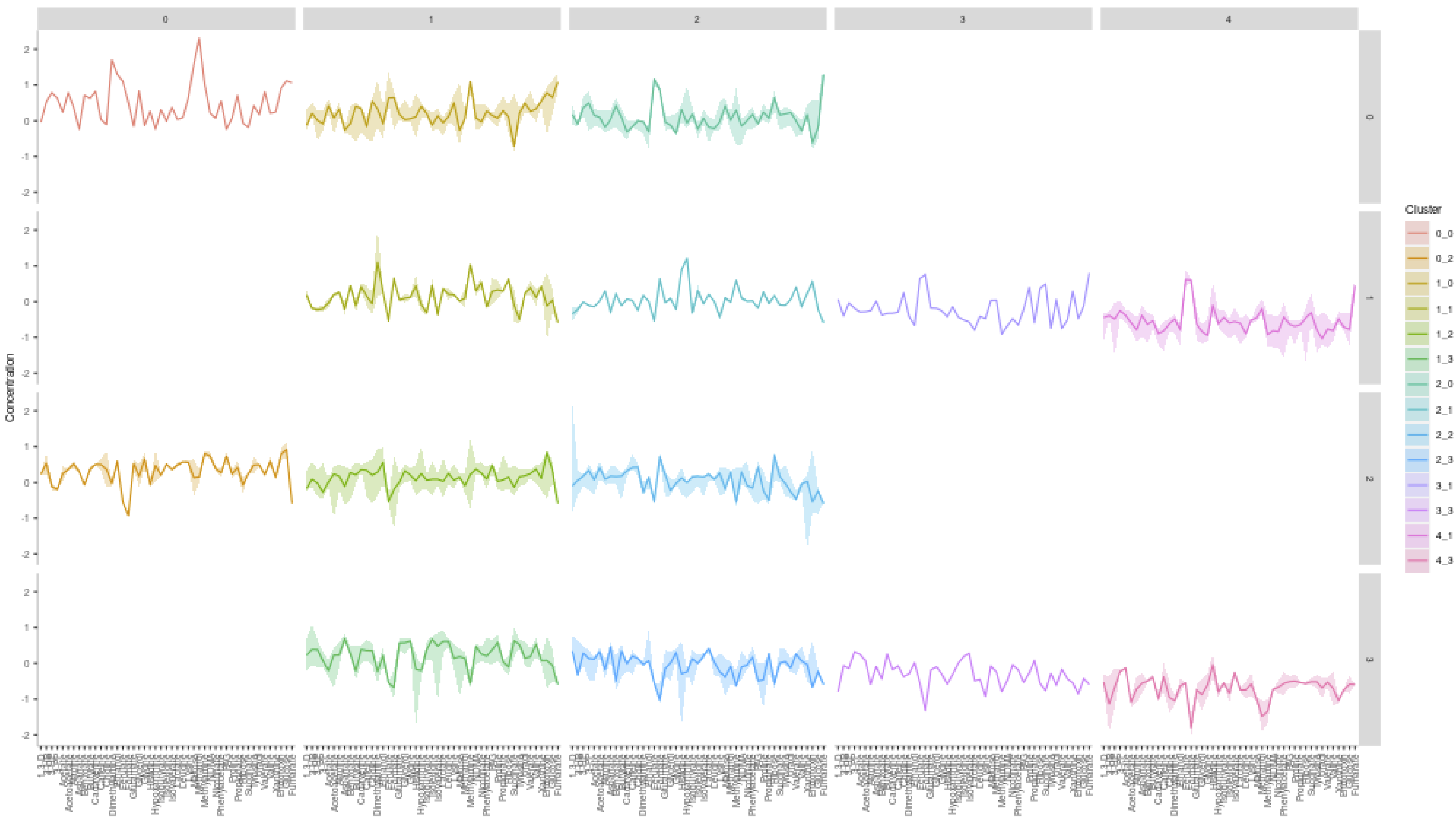


39サンプルを3個のパターンに分けた例

自己組織化マップ

Self-Organizing Map

似ているものを2次元マップの近くのグリッドに配置させる



39サンプルを5×4個のパターンに分けた例

数値の前処理

- transformation（変形）

  ログ化、平方根化など

  変数が持つ値の分布に偏りがある場合などに、偏った値の影響が出すぎたりしないよう、適当な重みづけに直す。

- normalization（正規化）

  平均値補正、中央値補正、内部標準補正など

  サンプル間で値の分布が異なっている場合に、適切な比較ができるように直す。

- scaling（スケーリング）

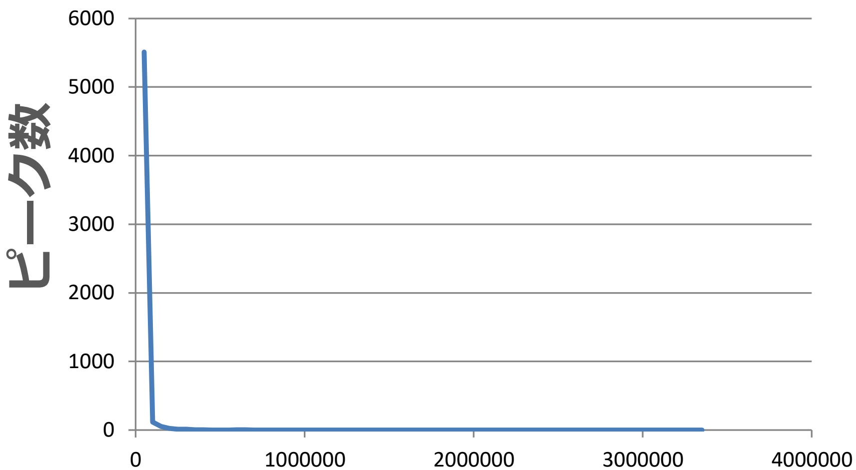
  平均値補正、分散を1にする、それらの組み合わせなど。

  それぞれの変数で、サンプル間での変動に大きな差がある場合などに、変動の幅を一定にするなどして、結果に対する変数の影響を調整する。

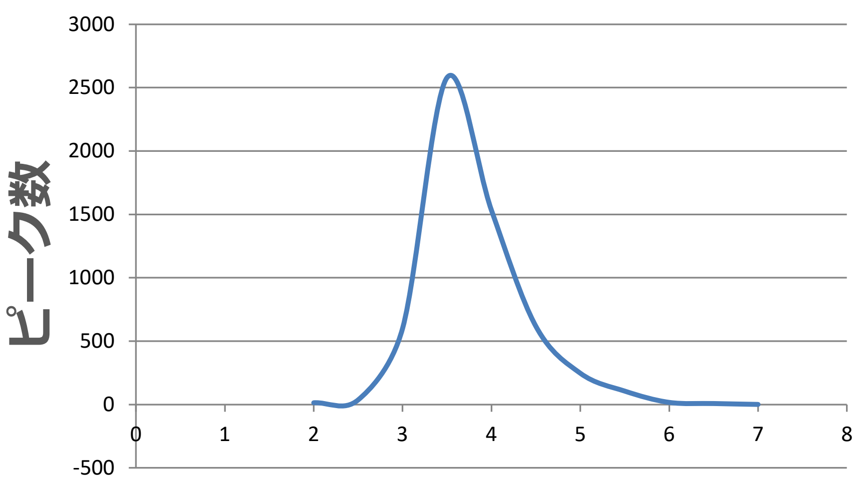
大葉（しそ）で検出された代謝物質

- 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

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

数値の前処理

- **transformation（変形）**

  ログ化、平方根化など

  変数が持つ値の分布に偏りがある場合などに、偏った値の影響が出すぎたりしないよう、適当な重みづけに直す。

- **normalization（正規化）**

  平均値補正、中央値補正、内部標準補正など

  サンプル間で値の分布が異なっている場合に、適切な比較ができるように直す。

- **scaling（スケーリング）**

  平均値補正、分散を1にする、それらの組み合わせなど。

  それぞれの変数で、サンプル間での変動に大きな差がある場合などに、変動の幅を一定にするなどして、結果に対する変数の影響を調整する。

さまざまな多変量解析

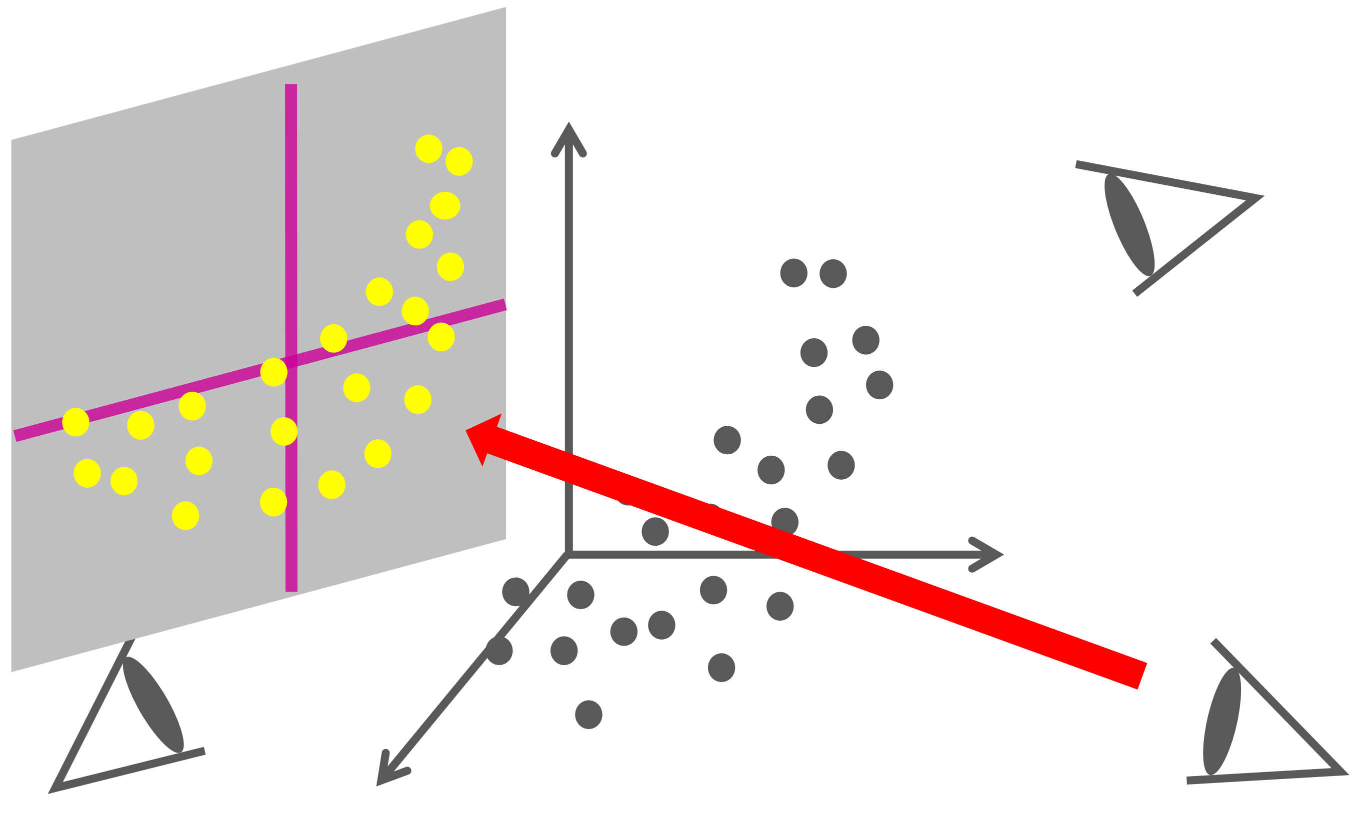
- データの類似性を考える
  （相関解析）
- 似ているものをグルーピングする
  クラスター解析
- データを要約する
  主成分分析
- 判別、分類、予測
  判別分析、PLS、PLS-DA、
  重回帰分析

など

主成分分析

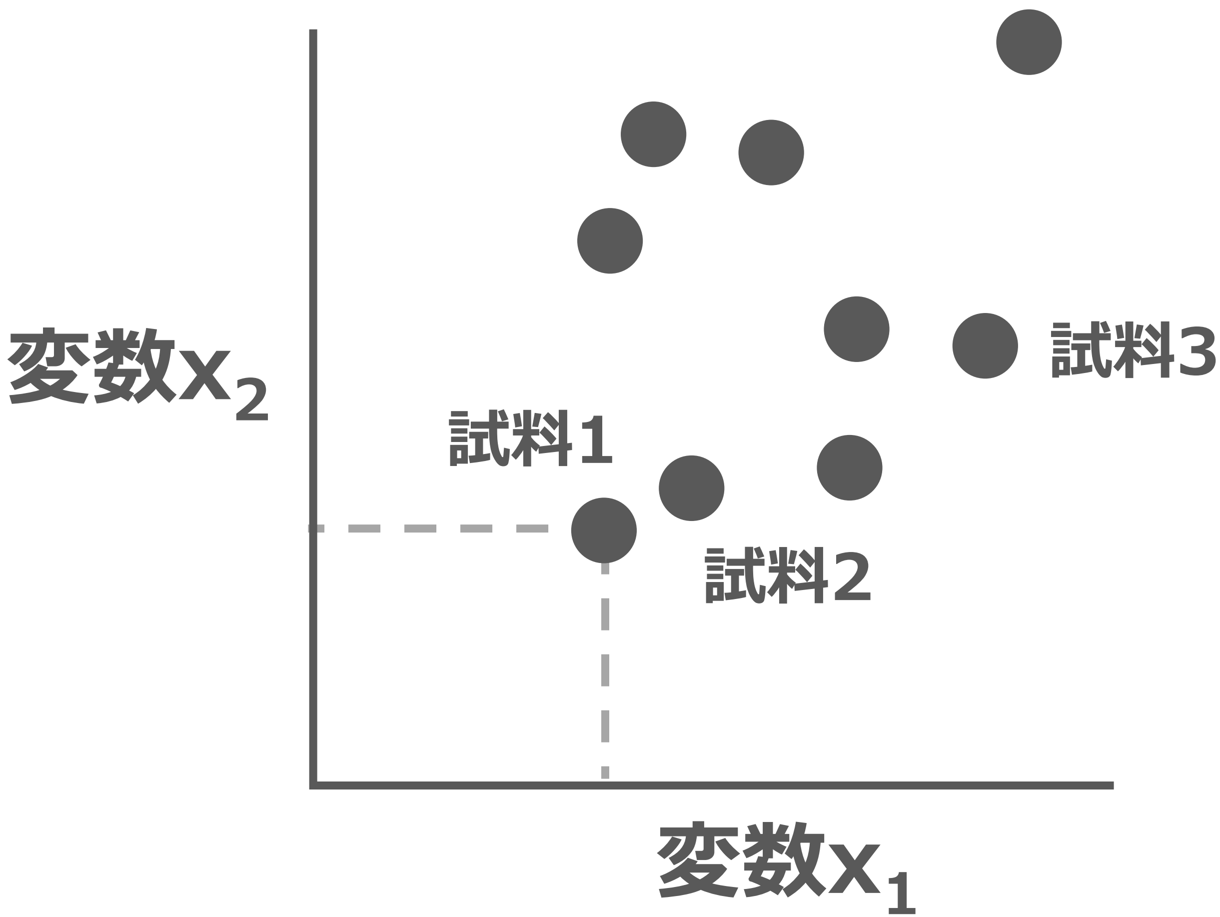
主成分分析のイメージ

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を描く



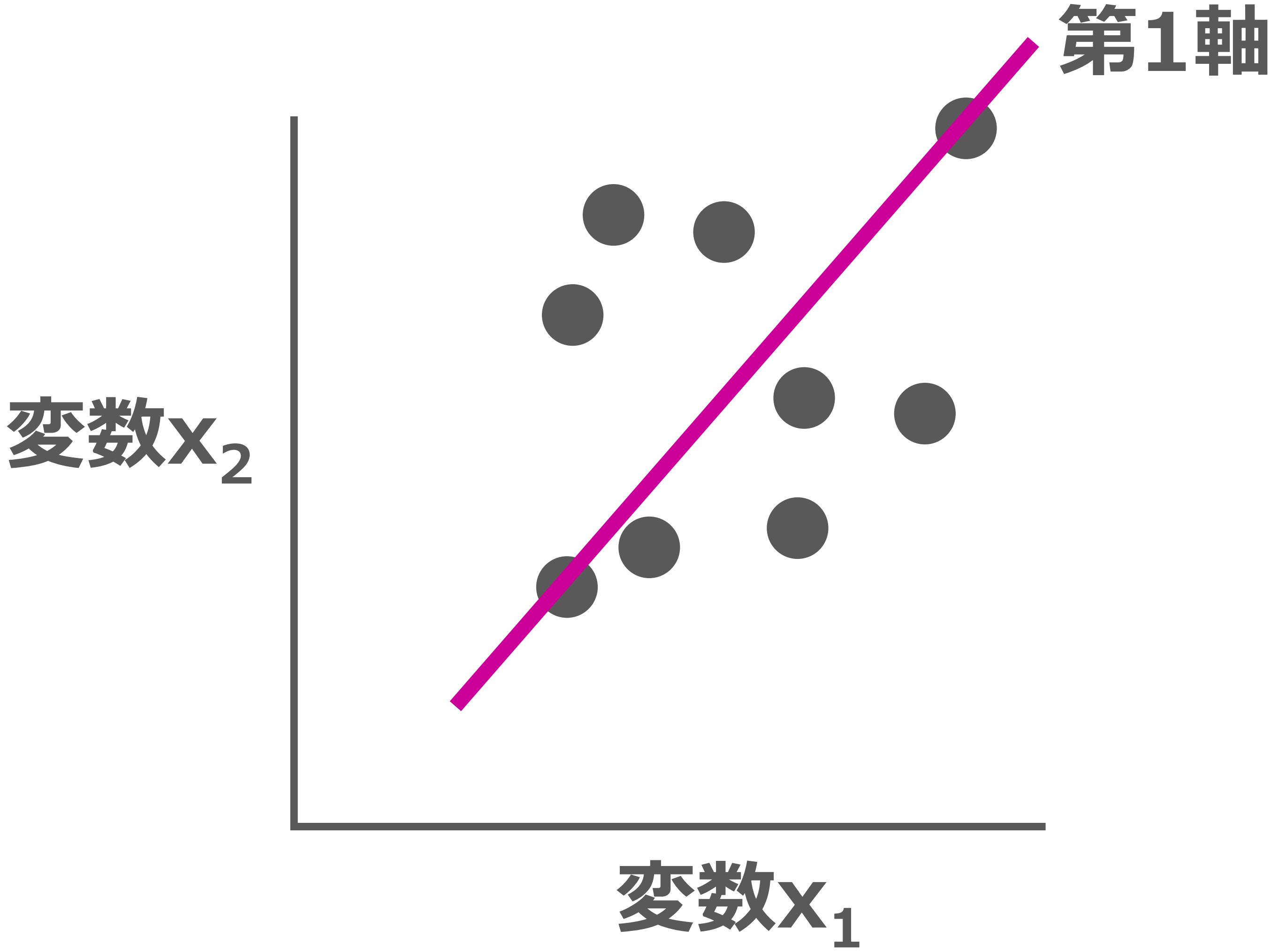
主成分分析のイメージ

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



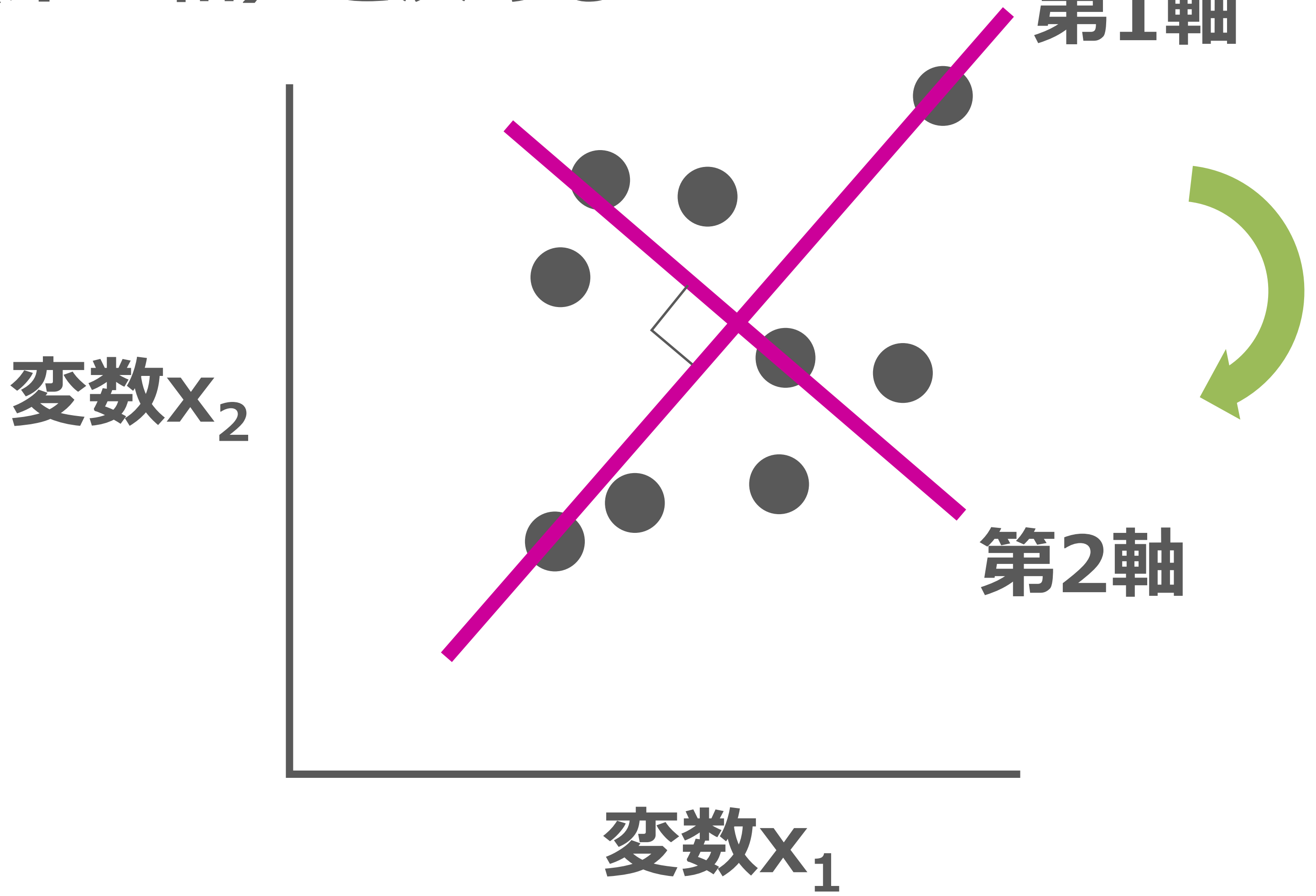
主成分分析のイメージ

②一番分散の大きい軸（第１軸）決める



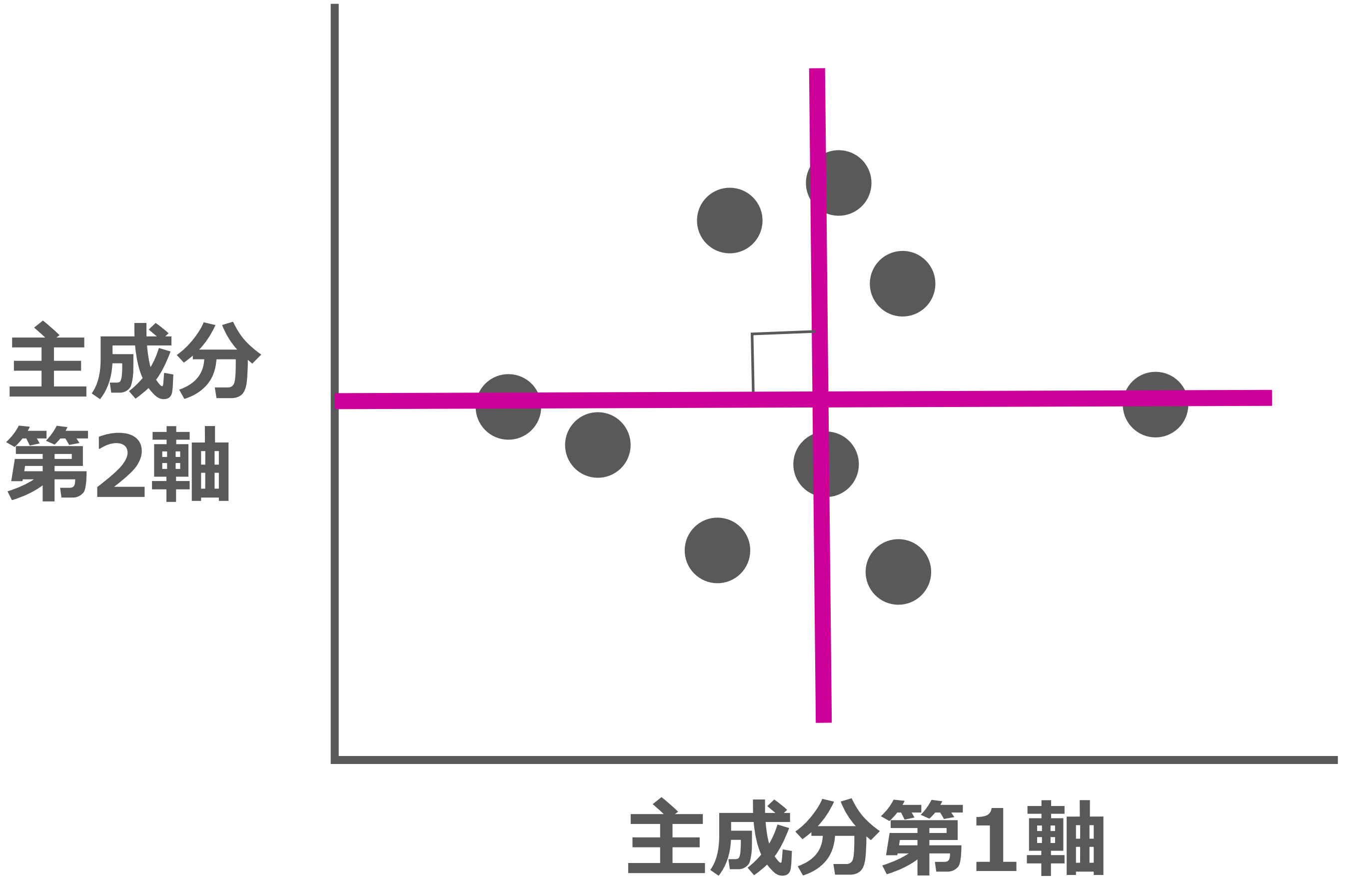
主成分分析のイメージ

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



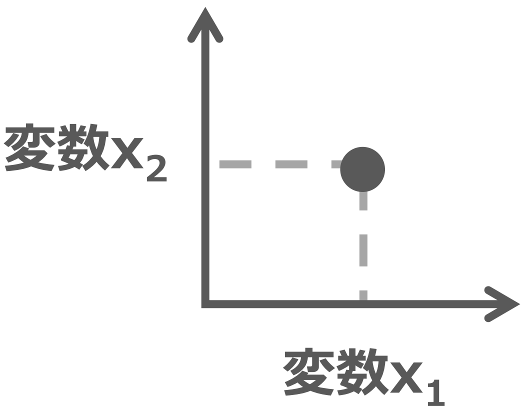
主成分分析のイメージ

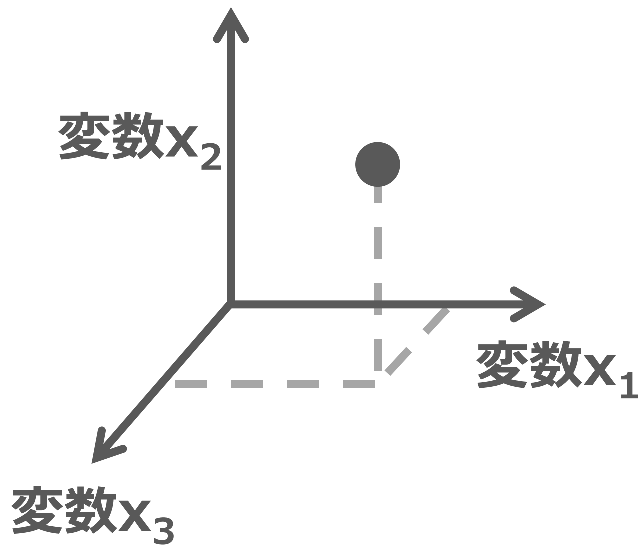
④第1軸がx軸、第2軸がy軸になるように、図を回転させた新たな図を作る



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





描けないが計算上は可能

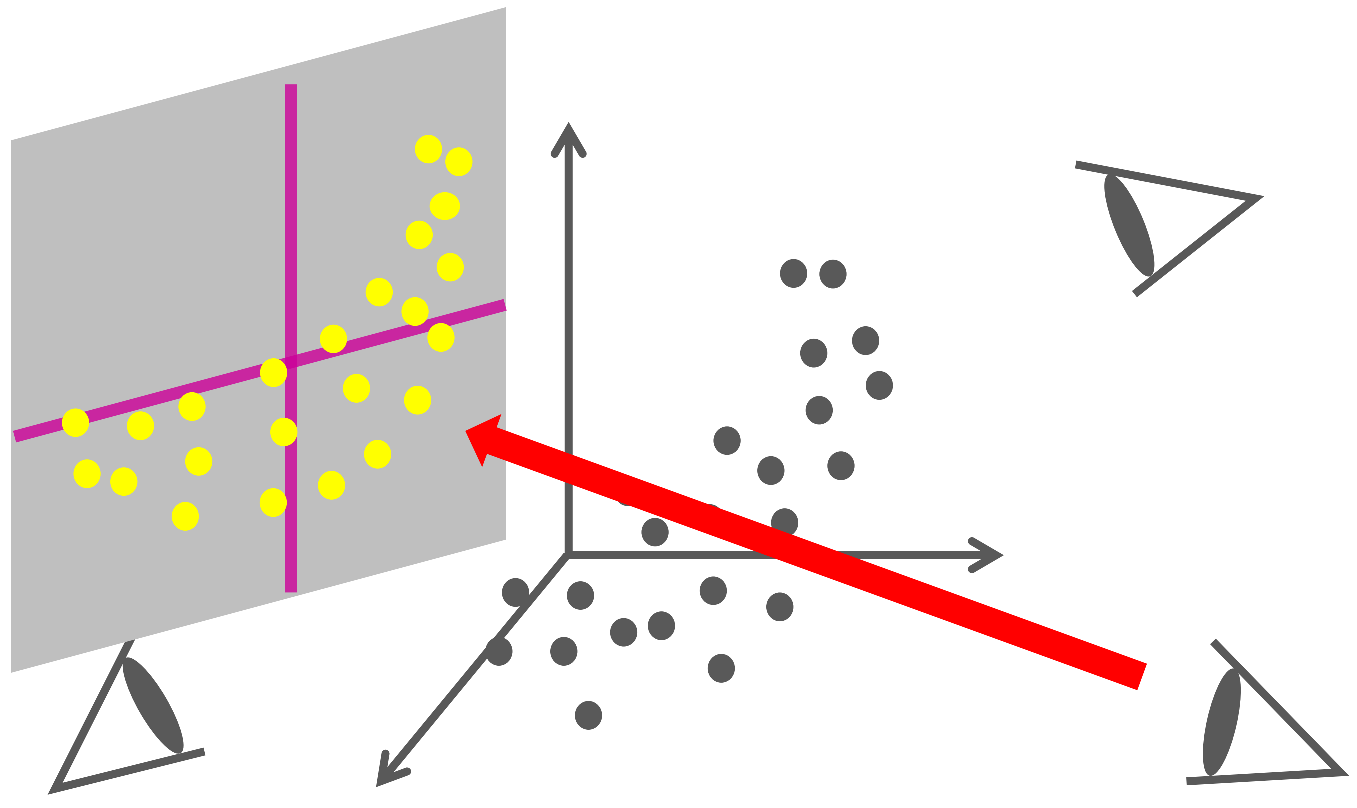
変数2個
2次元

変数3個
3次元

変数m個
m次元

主成分分析のイメージ

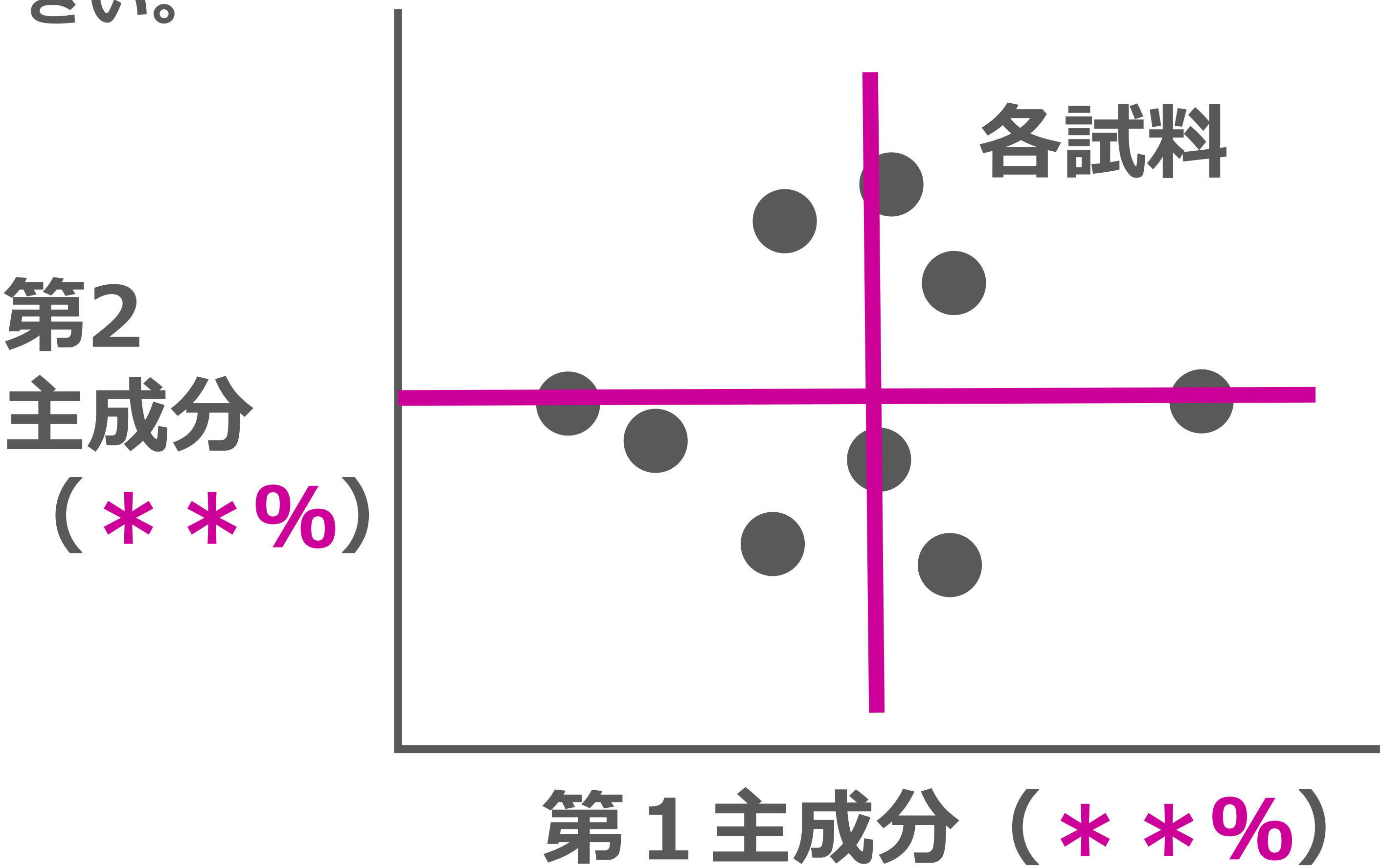
試料間の違い（特徴）が一番はっきりと見える方向から見た図が描ける



スコアプロット

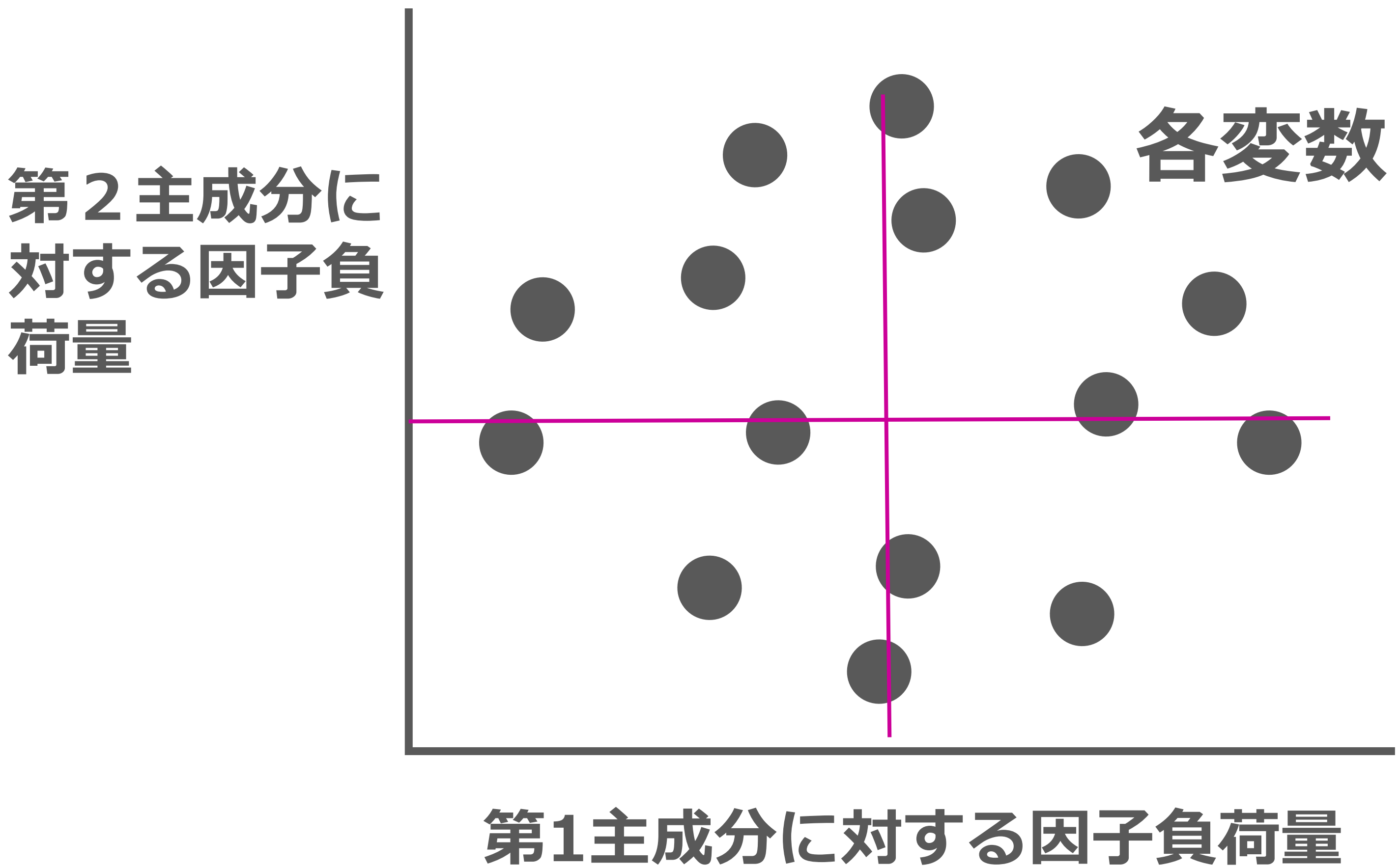
主成分軸に各試料を投影しなおした図

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。



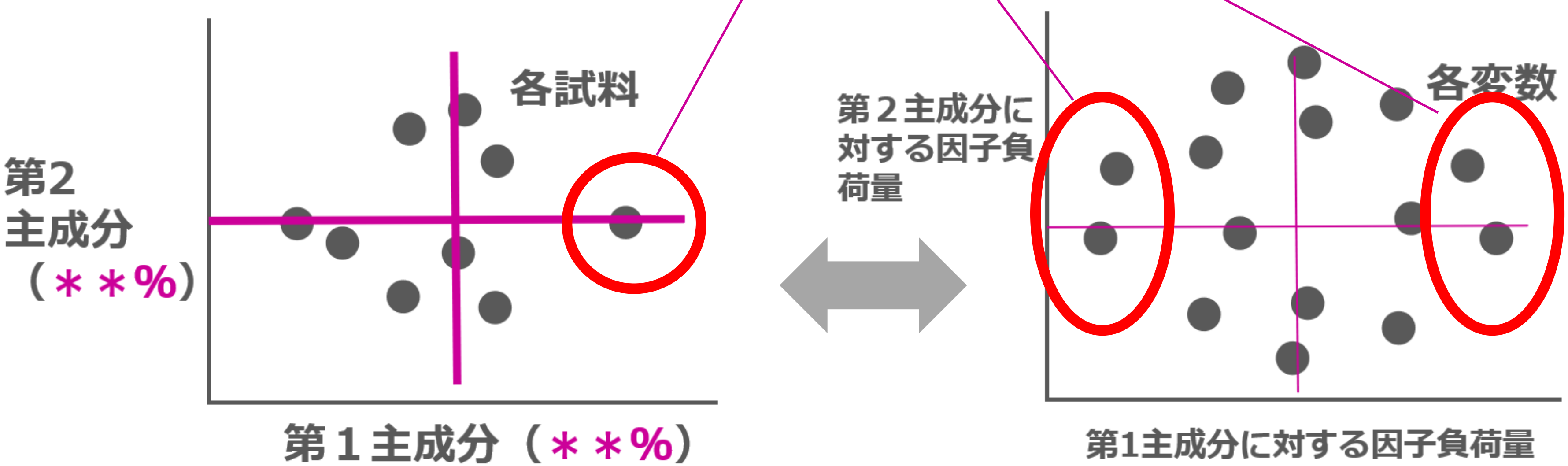
ローディングプロット

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

PCAの例

17名に、お好みのアンケートを実施

Q1. お酒が好き度　嫌い：0%、普通：50%、すごい好き：100%
Q2. 自炊をする頻度(%)→全くしない0%、毎日する100%
Q3. 映画を見るのが好き度
Q4. ところてんの好き度
Q5. 就職活動の進行度　めっちゃ進んでる：100%
Q6. 金欠度　やっべぇ：100%
Q7. きのこの山よりたけのこの里派　きのこの山：0%、たけのこの里：100%
Q8. 自分の将来の夢に近づいているか？　　今夢の中：100%
Q9. 今の満腹度　もう食えない：100%
Q10. 今の生活に満足しているか　大満足：100%
Q11. 食べることが好き　好きではない：０％、すごく好き：100％

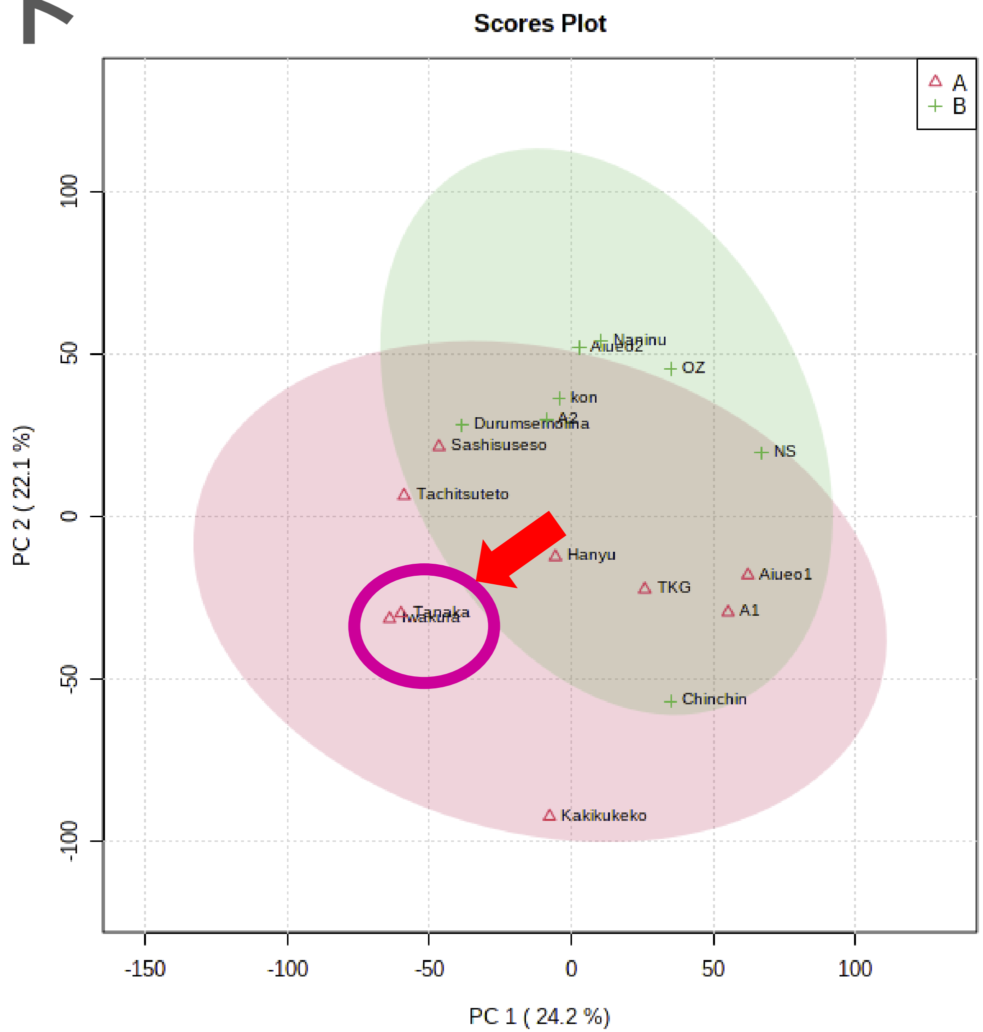
PCAの例

アンケート結果

id	class	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
Tanaka	A	60	75	100	30	0	99	100	0	60	20	100
A1	A	80	2	90	0	0	100	0	20	40	99	100
TKG	A	50	70	60	80	30	75	50	65	10	68	90
Hanyu	A	5	2	90	3	0	80	50	20	30	20	100
Kakikukeko	A	50	60	100	75	0	100	0	35	10	5	100
Iwakura	A	100	50	100	50	20	100	100	0	50	10	100
Sashisuses	A	50	0	50	0	0	90	100	30	40	0	100
Aiueo1	A	50	0	80	5	20	20	0	0	50	50	100
Tachitsute	A	50	0	70	30	0	100	100	20	40	0	100
Chinchin	B	50	70	50	30	30	50	0	0	70	30	100
NS	B	90	10	50	30	50	50	40	80	40	80	75
Naninu	B	50	40	10	10	10	40	100	40	50	60	80
Durumsem	B	50	0	50	50	0	80	100	0	50	30	50
A2	B	50	0	100	80	0	40	100	30	40	60	100
OZ	B	10	20	100	20	30	20	90	40	50	85	100
Aiueo2	B	50	10	50	0	0	60	100	50	30	60	80
kon	B	50	0	80	80	10	80	100	60	40	80	100

PCAの例

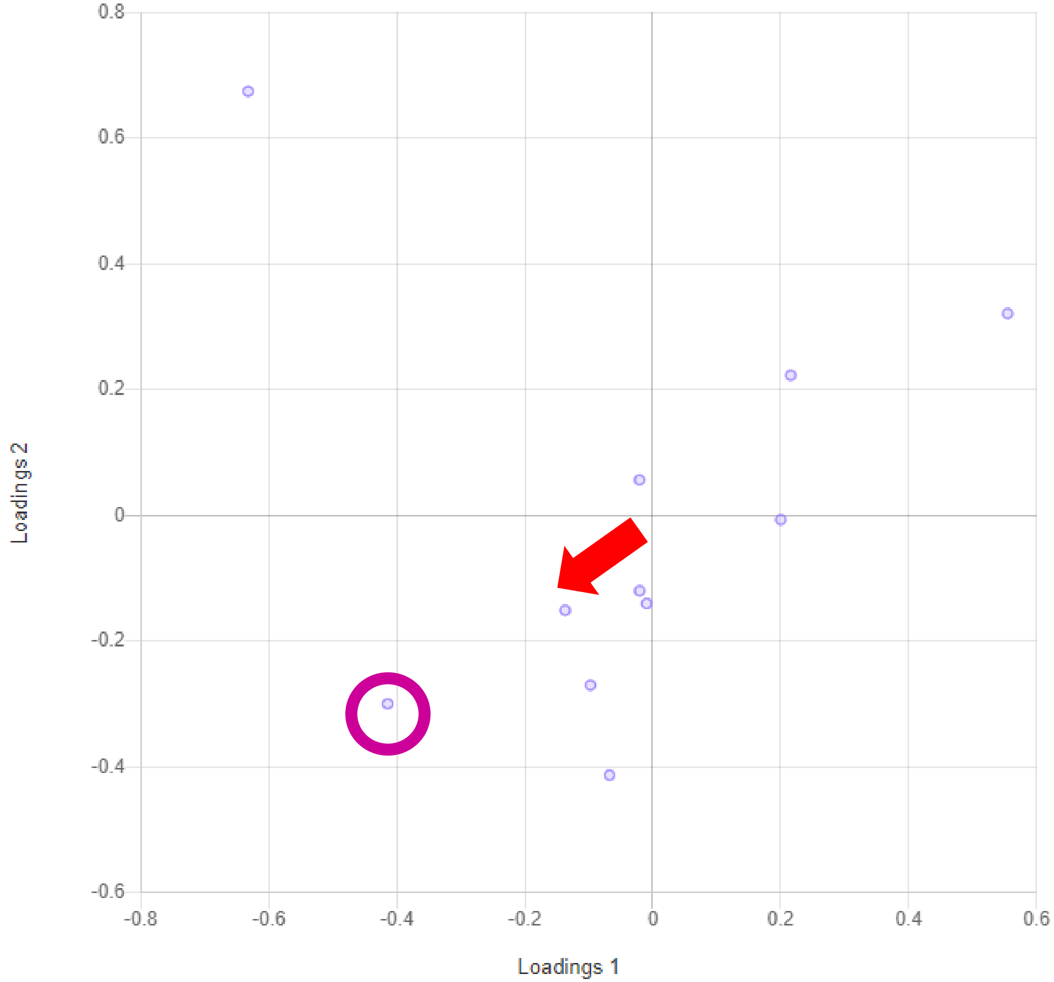
スコアプロット

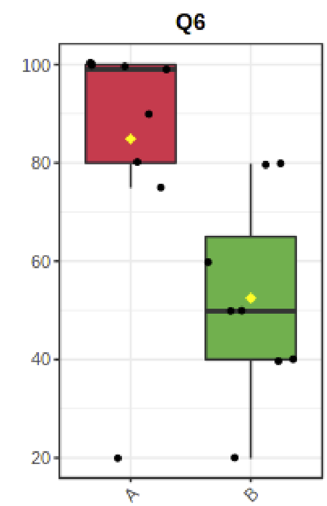


A, Bの区分、色は無視してください。

PCAの例

ローディングプロット





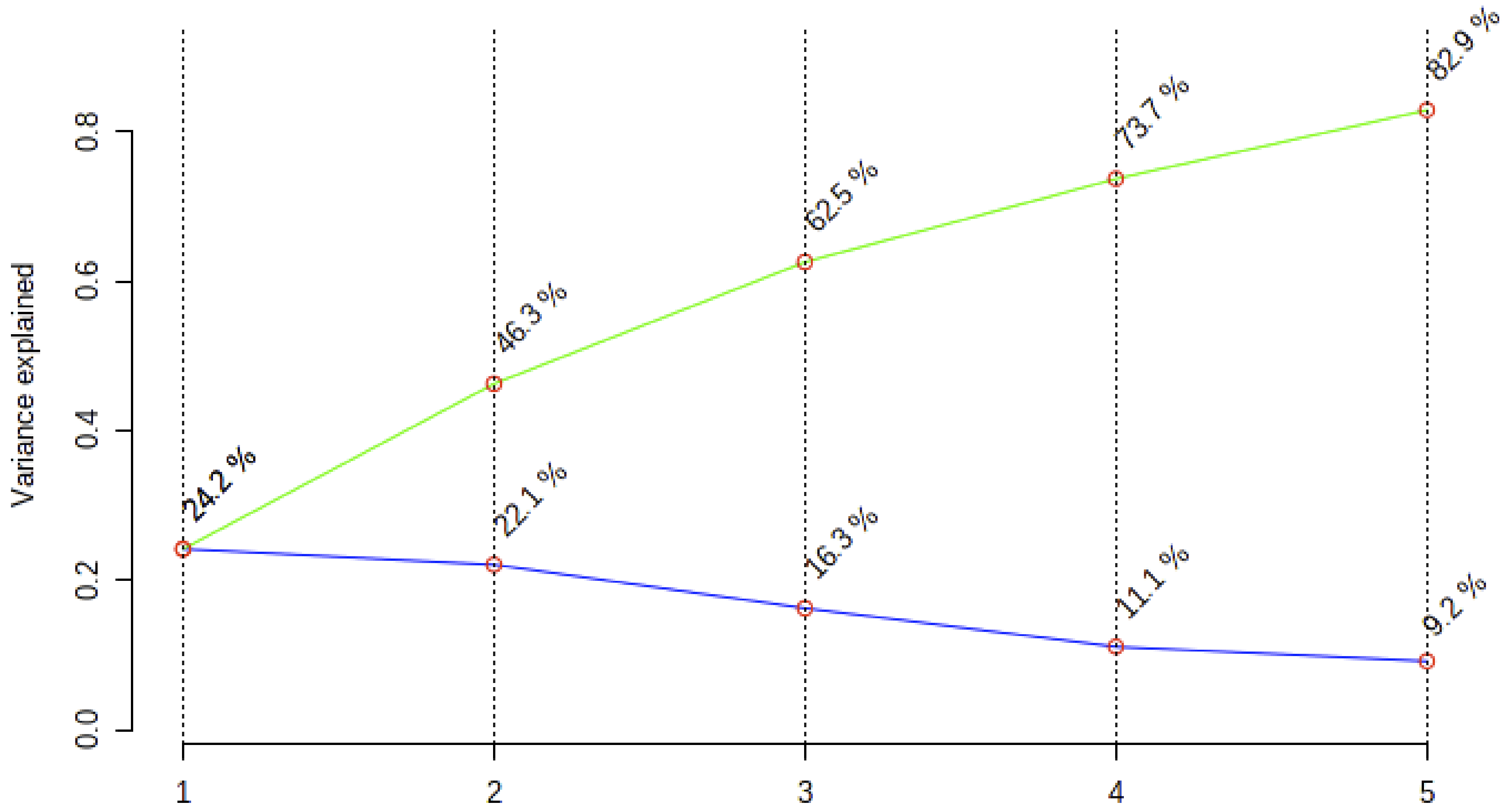
PCAの例

17名に、お好みのアンケートを実施

Q1. お酒が好き度　嫌い：0%、普通：50%、すごい好き：100%
Q2. 自炊をする頻度(%)→全くしない0%、毎日する100%
Q3. 映画を見るのが好き度
Q4. ところてんの好き度
Q5. 就職活動の進行度　めっちゃ進んでる：100%
Q6. 金欠度　やっべぇ：100%
Q7. きのこの山よりたけのこの里派　きのこの山：0%、たけのこの里：100%
Q8. 自分の将来の夢に近づいているか？　　今夢の中：100%
Q9. 今の満腹度　もう食えない：100%
Q10. 今の生活に満足しているか　大満足：100%
Q11. 食べることが好き　好きではない：０%、すごく好き：100%

PCAの例

寄与率と累積寄与率



主成分軸

PLS

Partial Least Squares

部分最小二乗

PLS-DA

Partial Least Squares-Discriminant Analysis

部分最小二乗-判別分析

PLS、PLS-DAで扱うデータ

目的変数が存在する

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

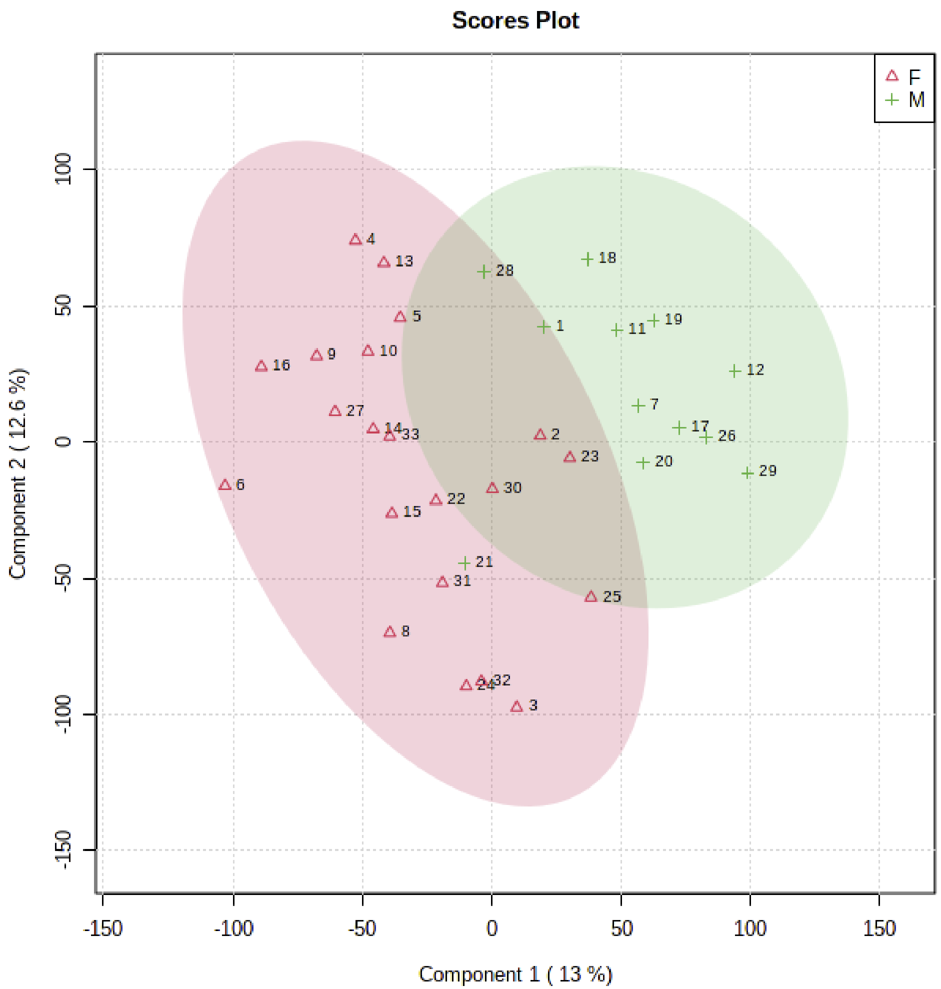
遺伝子発現量など

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

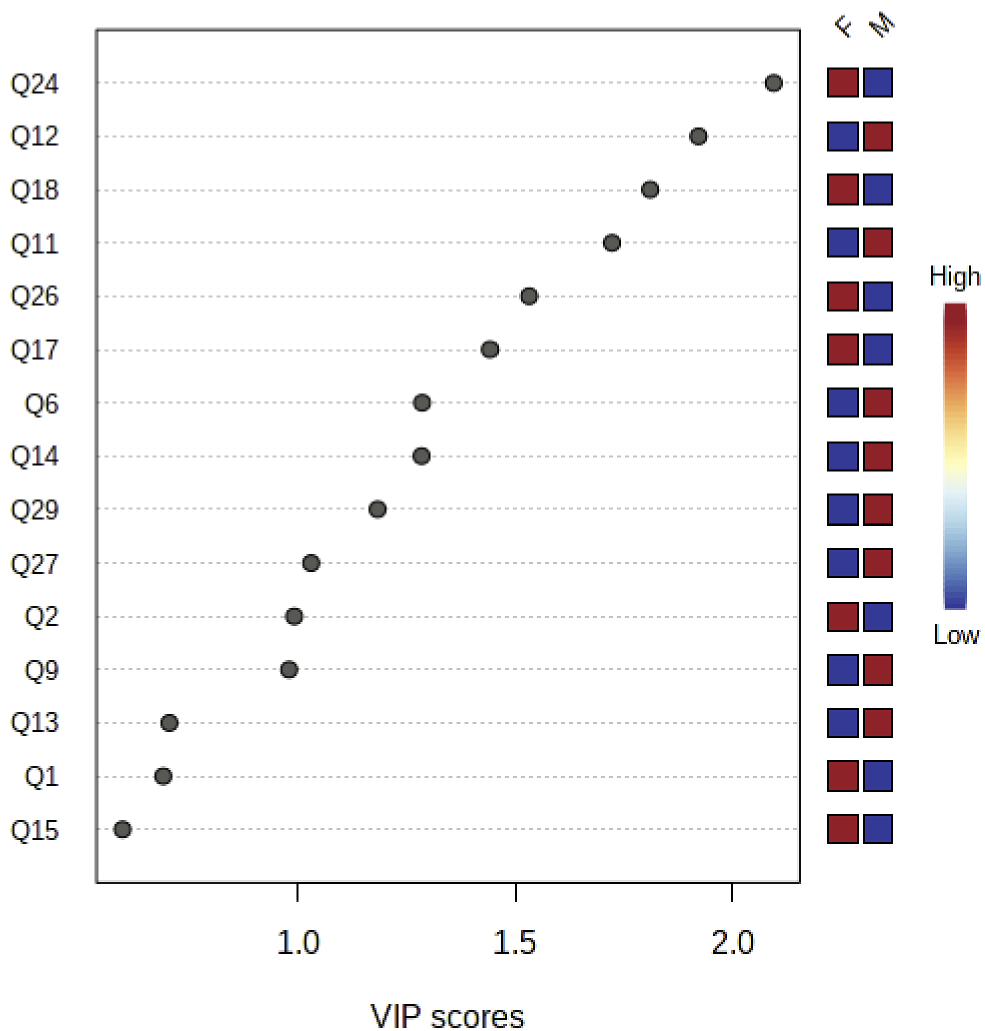
PLSの例

別のお好みアンケート。33名回答。
目的変数に性別を入れ、男女の好みの差を調べた。



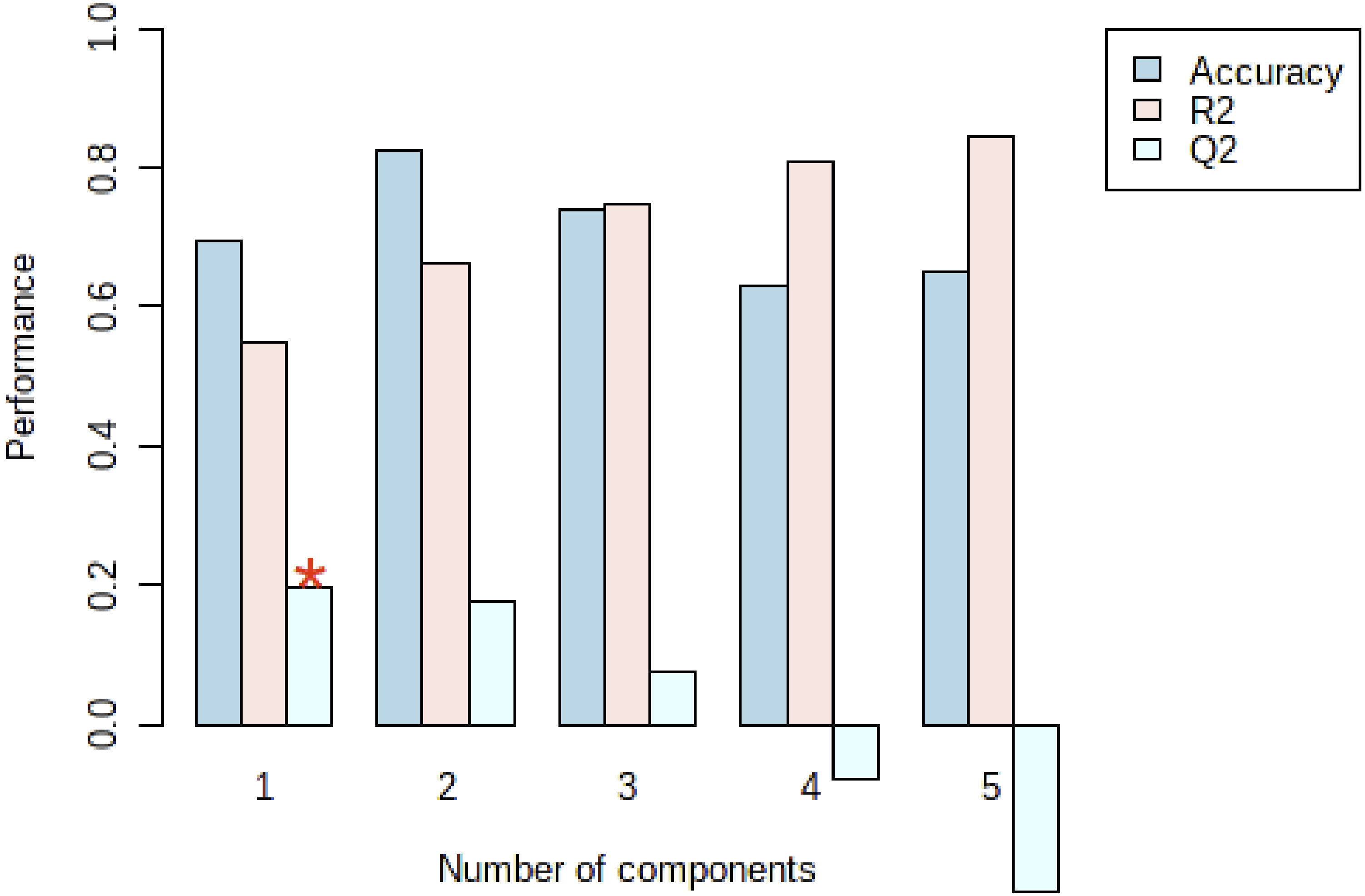
PLSの例

VIP


F
M
Q24
Q12
Q18
Q11
Q26
Q17
Q6
Q14
Q29
Q27
Q2
Q9
Q13
Q1
Q15
High
Low
1.0
1.5
2.0
VIP scores

PLSの例

Q^2: PLS-DAのための、作られたモデルの予測性能



スケジュール

今年

- 検定の基礎
  - ✓ 統計の大事な考え方
    平均値～t分布まで
  - ✓ t検定
    検定のやりかた
    気を付けること
  - ✓ いろんな検定とANOVA
- 多変量解析のイメージ
- 多変量解析の実習

午前中

～90分

～60分

～夕方まで