

# 統計の基礎

2025年9月8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

# 資料のサイト

<https://github.com/nsaku/ku2025/wiki>

# パスワード



推定

- 検定とは？ t-検定
- 分散分析（ANOVA）
- 相関
- 主成分分析（PCA）
- 回帰（PLS回帰）

要約・分類

予測

多変量解析

# 今日の内容

- 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

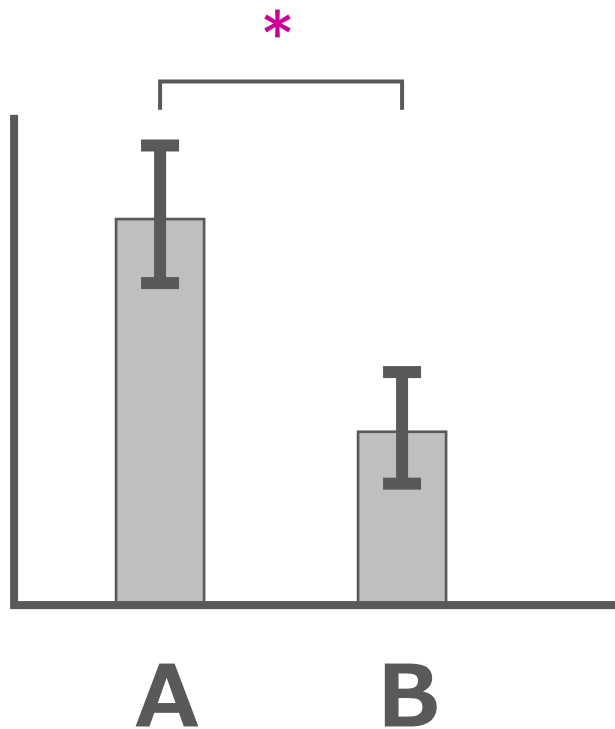
# 全体の目標

- 用語
- 分布とその使い方
- 多変量解析の概念

→ 研究に活かせるように

# 検定の基礎

目標： この意味が分かるようになる



n = 5

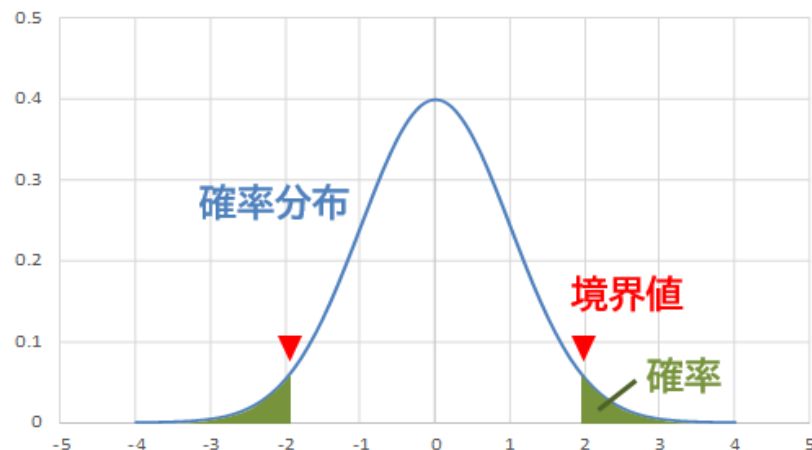
\* *t*-test  $p < 0.05$

# ポイント

検定の多くは、**平均値に差があるかどうか**を確率的に評価したもの、またはその派生



- 平均値を推定したとき、それがどれだけばらつくか(真の値から外れるか)には法則性があり、ある**分布**に従う
- 分布は、事象が起こる**確率**を表したものの





統計



# 統計って？

**集団**の状況を  
数値で表したものの



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

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



**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：**

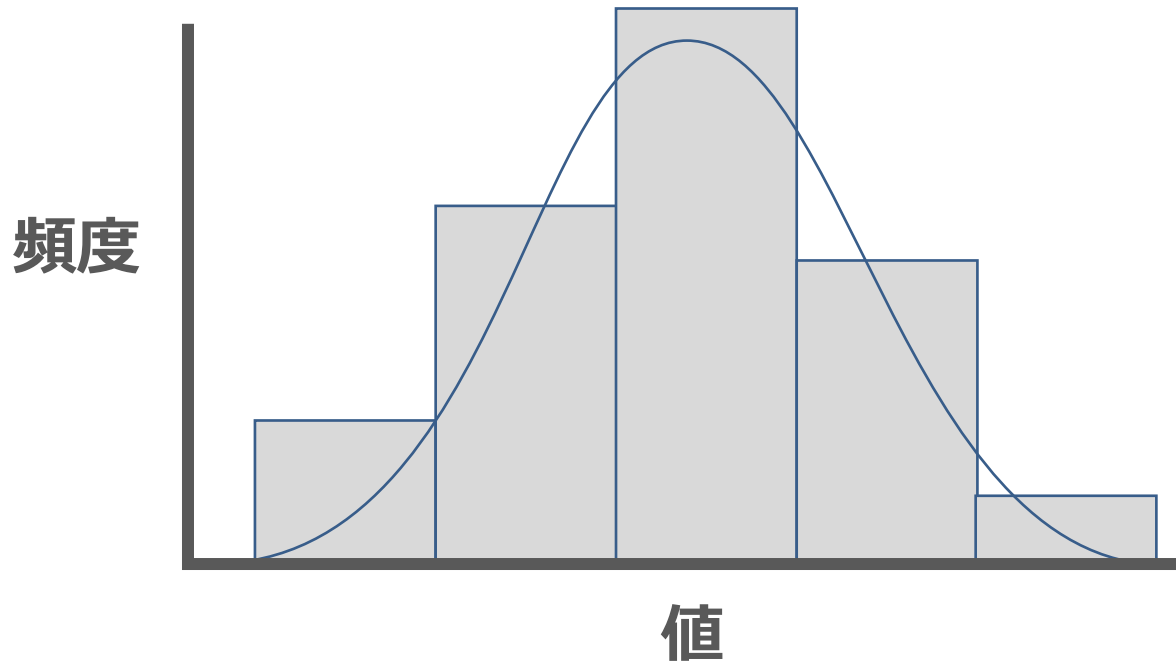
平均値、最大値、

**そのほか：**

グラフ（図）

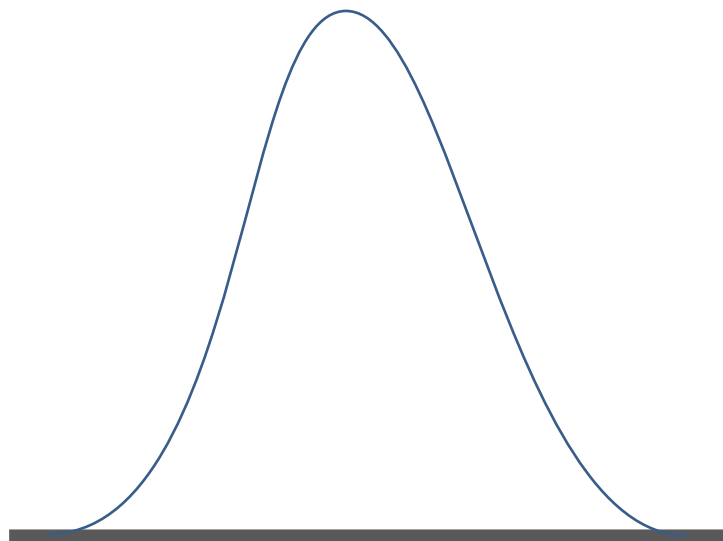


# グラフの例



ヒストグラム（頻度分布図）

平均値と中央値は一致



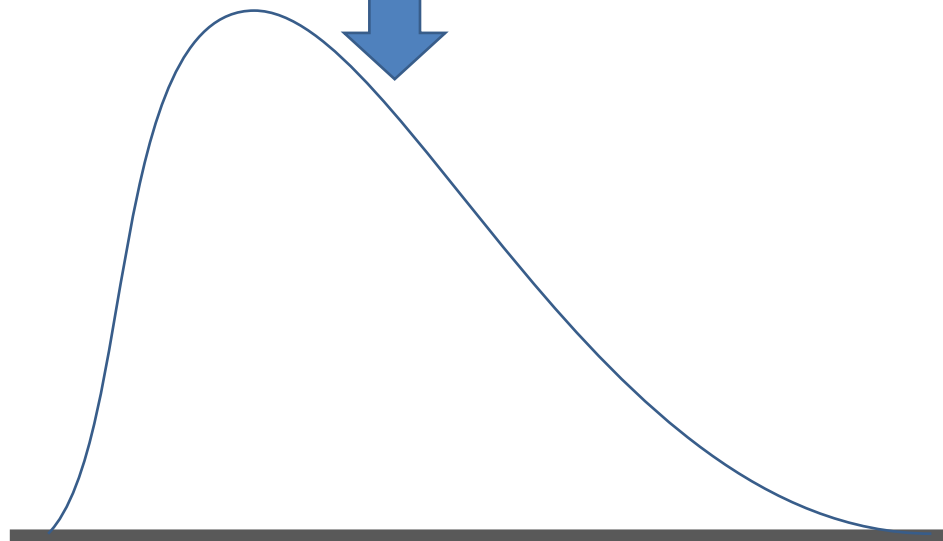
偏りのないデータ

身長分布など

中央値の方が大勢の傾向を反映



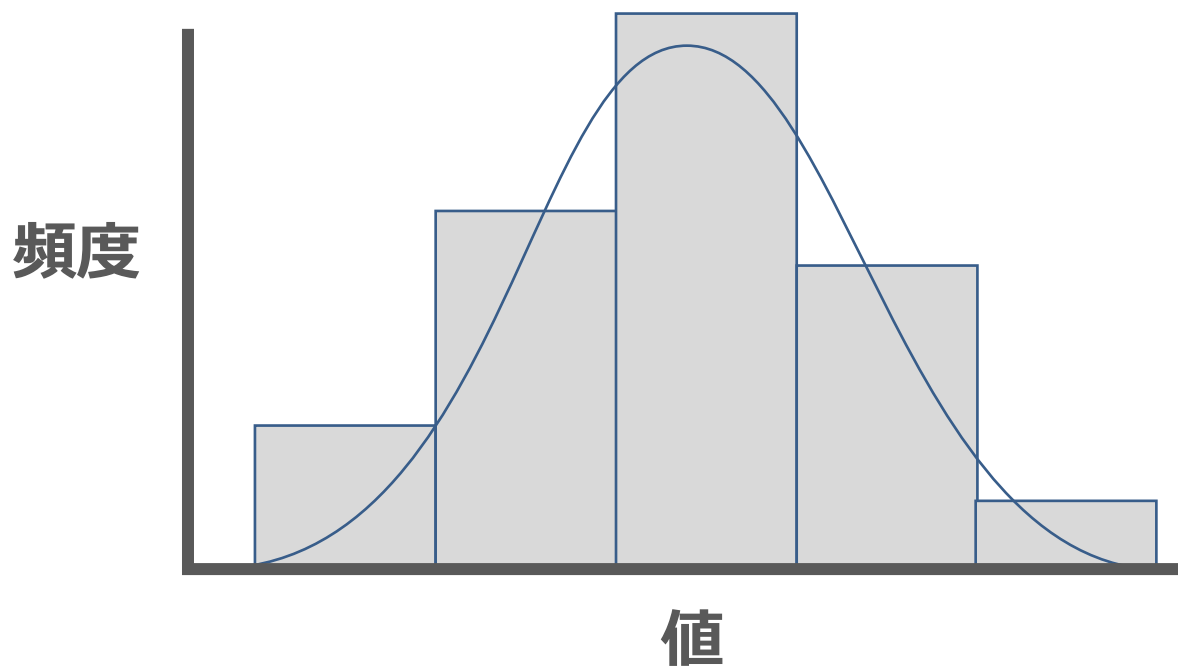
平均値



偏っているデータ

体重分布など

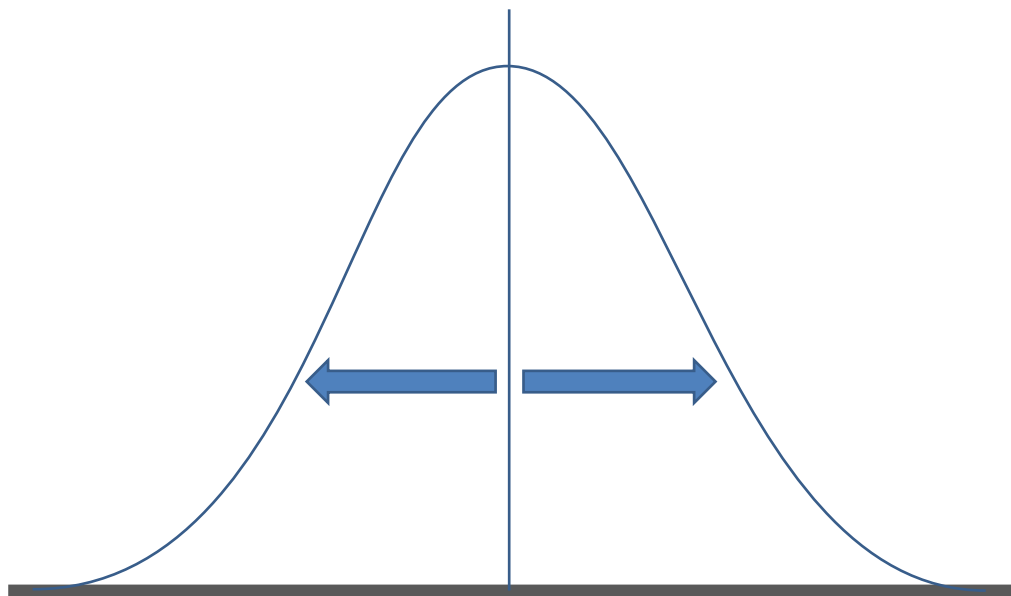
# グラフの例



ヒストグラム（頻度分布図）

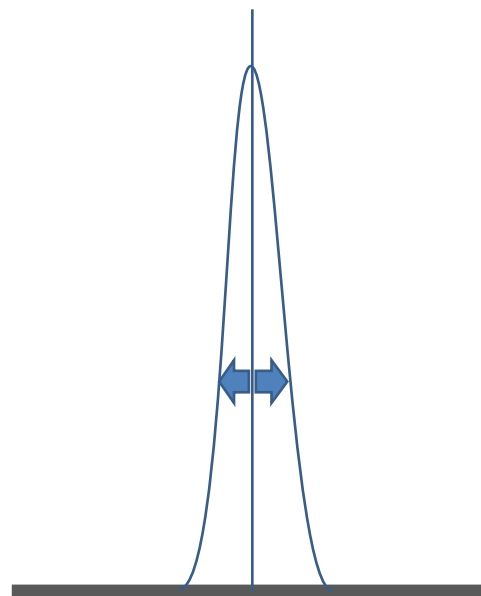
データのばらつき具合がわかる **分布**

# 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

# 計算方法

# 分散

- 平均値を計算

※中央値ではなく、必ず平均値

- (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





# 集団を可視化したイメージ

分布

集団の全体傾向  
(偏り具合)を表す

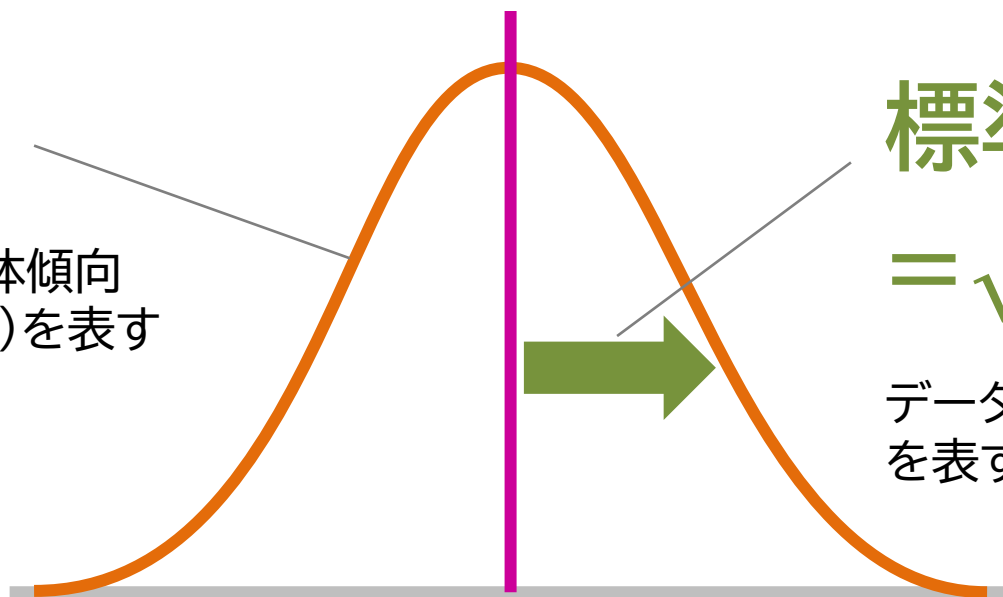
標準偏差

$$= \sqrt{\text{分散}}$$

データのばらつき具合  
を表す

平均値

集団の中心点を表す

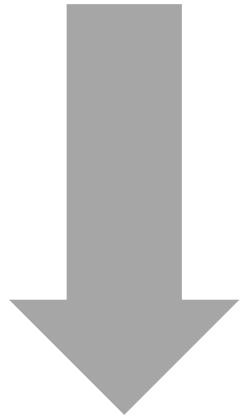


目的：この研究室の人の  
身長はどのくらい？

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

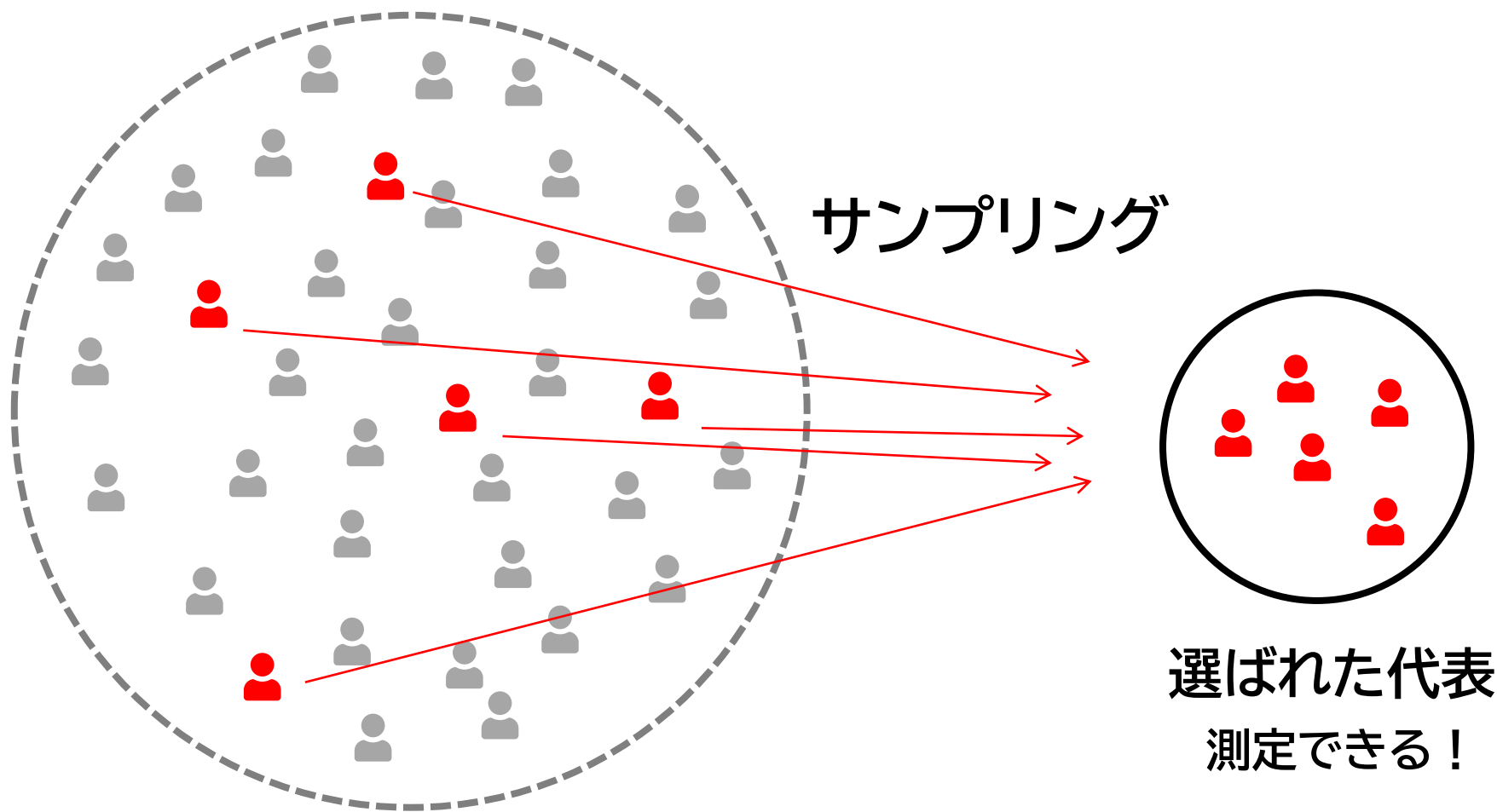


- ✓ 現実的ではない
- ✓ コストもかかる

## 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

日本人全員

全員測定ムリ！

選ばれた代表  
測定できる！

# 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



今回の目的の場合、  
サンプリングされた人のこと

標本  
(サンプル)

# 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？



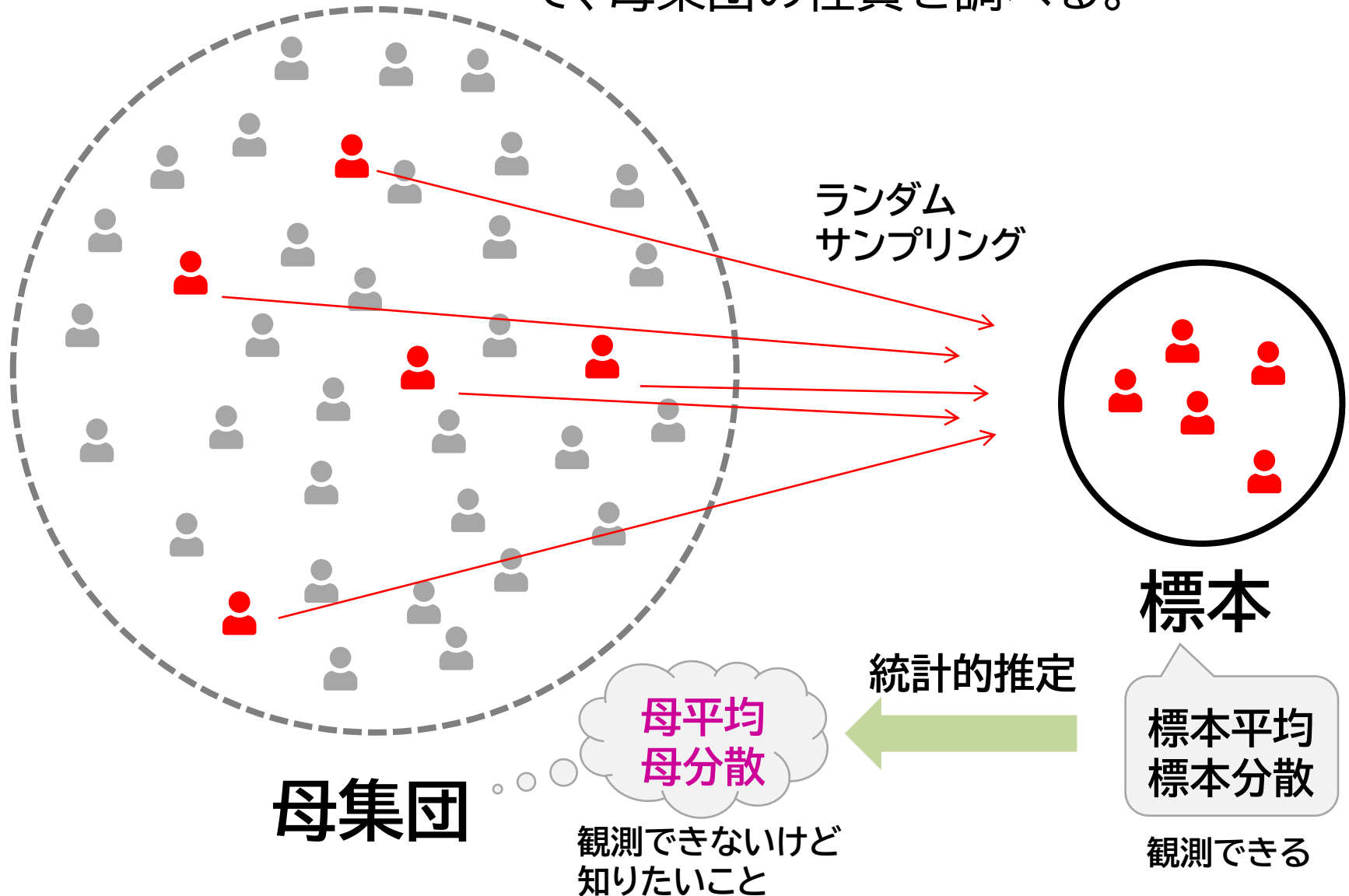
限られた**標本**を使って  
**母集団**(日本人全体)の

- **推定の平均値**や
- **推定のばらつき**を

見積もる、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  $\mu$  ← 標本平均  $\bar{x}$   
一致が期待できる

母分散  $\sigma^2$  ~~←~~ 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

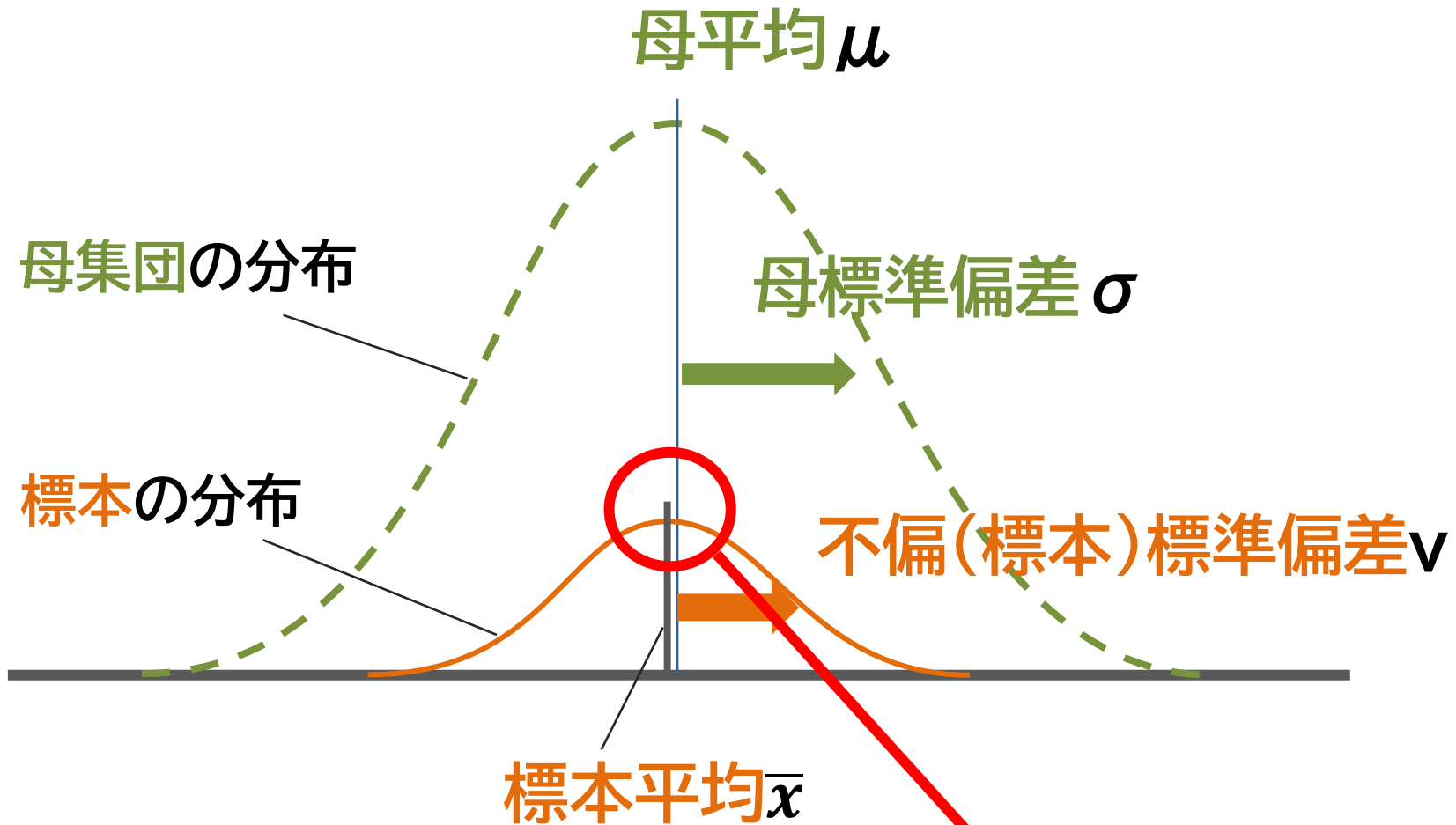
不偏(標本)分散  $v^2$

$\mu$ : 平均(mean)のm  
 $\sigma$ : 標準偏差(standard deviation)のs  
に相当するギリシャ文字

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

母平均  $\mu$  の推定

# 母平均 $\mu$ (真実)と、標本平均 $\bar{x}$ (推定)のズレ

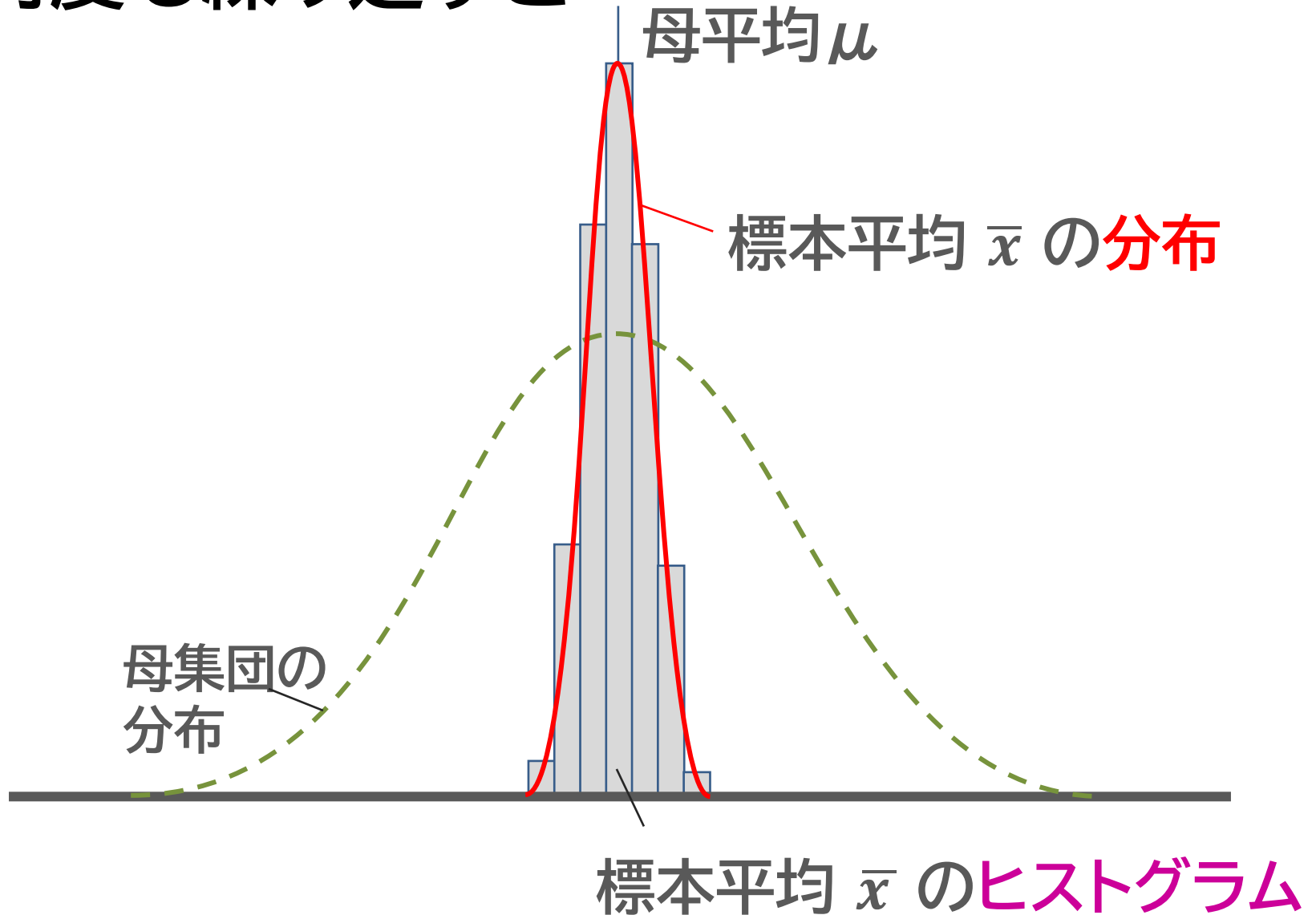


どれだけズレているか？

# 誤差

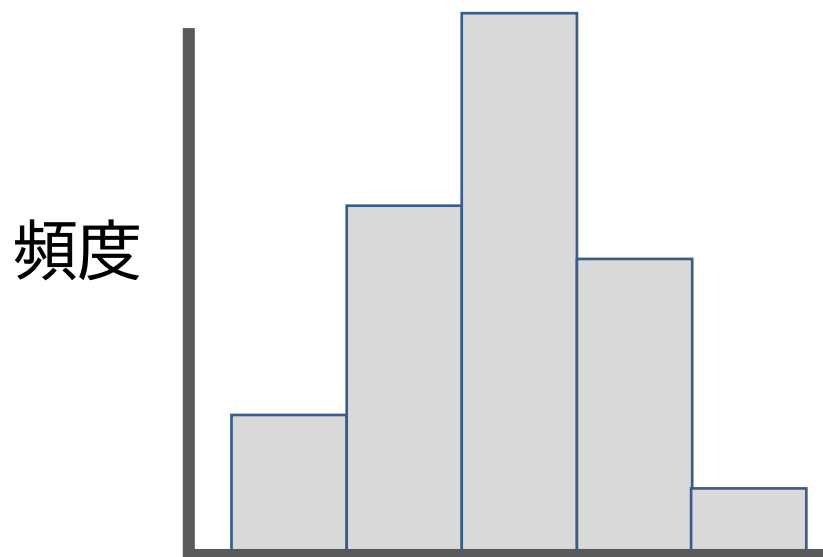
- サンプルリング誤差
- 測定誤差
- 推定誤差  
など...

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...



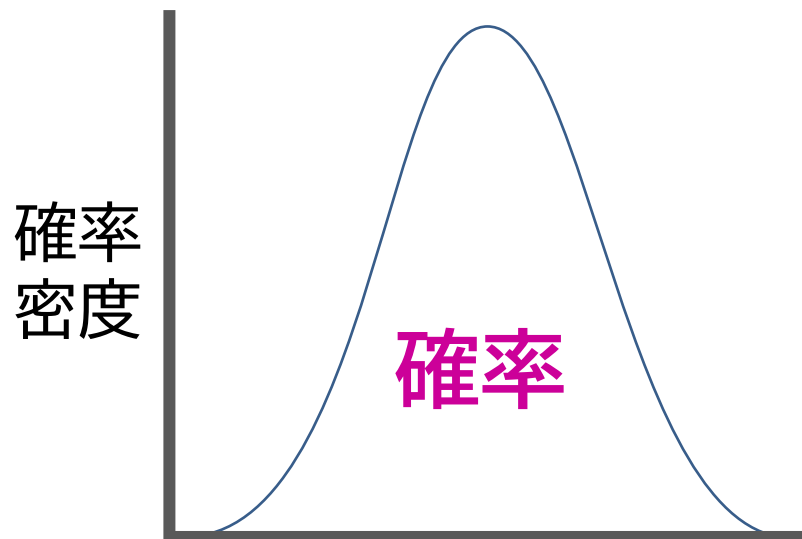
# 分布

データの全体的な偏り具合を表したものの



値  
ヒストグラム  
(頻度分布図)

①観測結果を表す図



値  
確率密度関数

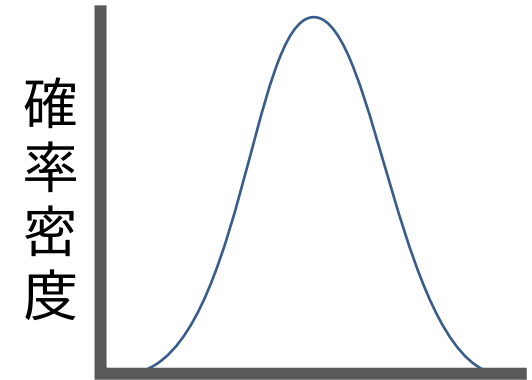
②事象の起こる確率を表す関数



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



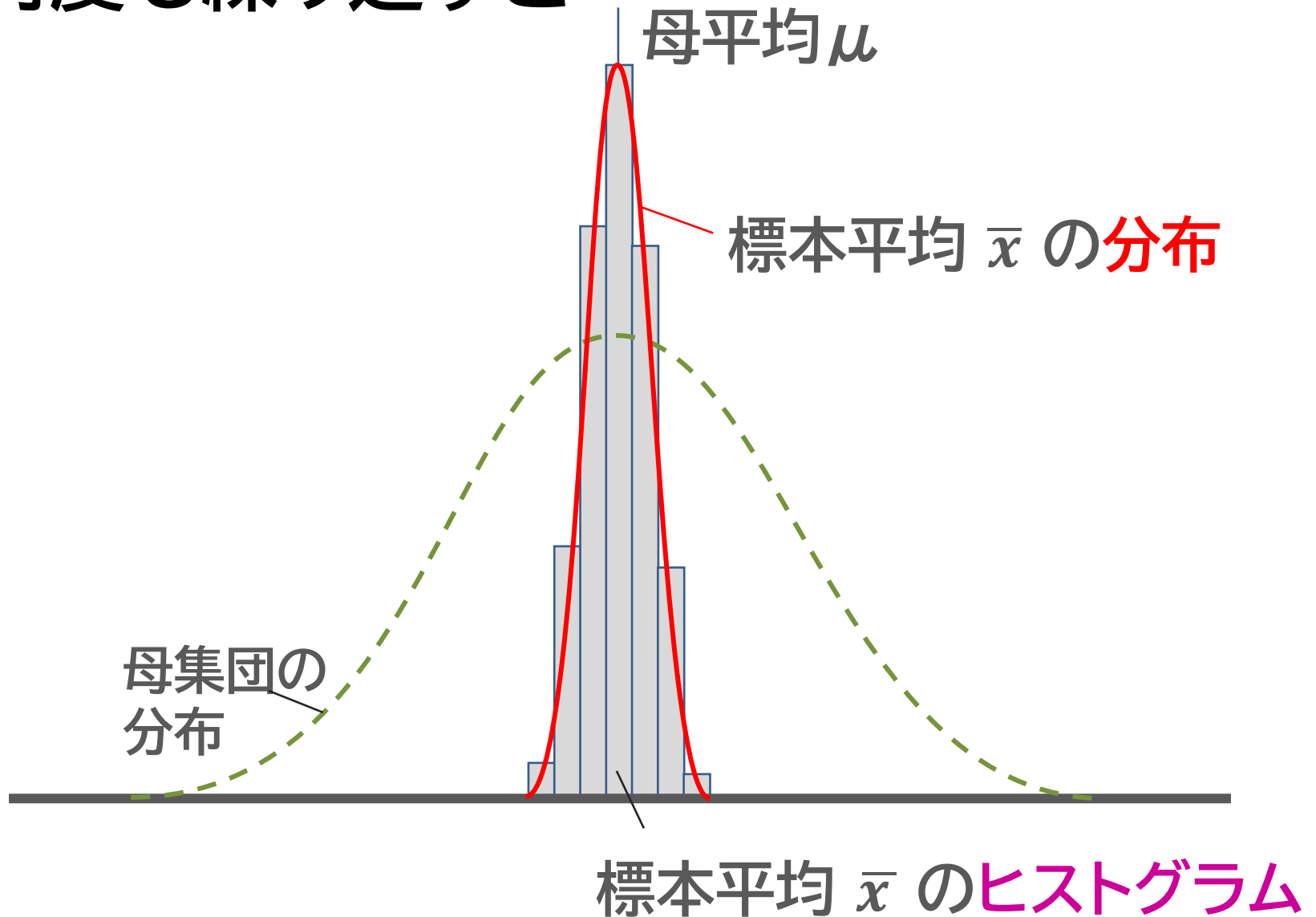
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

均等な確率で生じたばらつきの場合にとる分布

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

など

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...

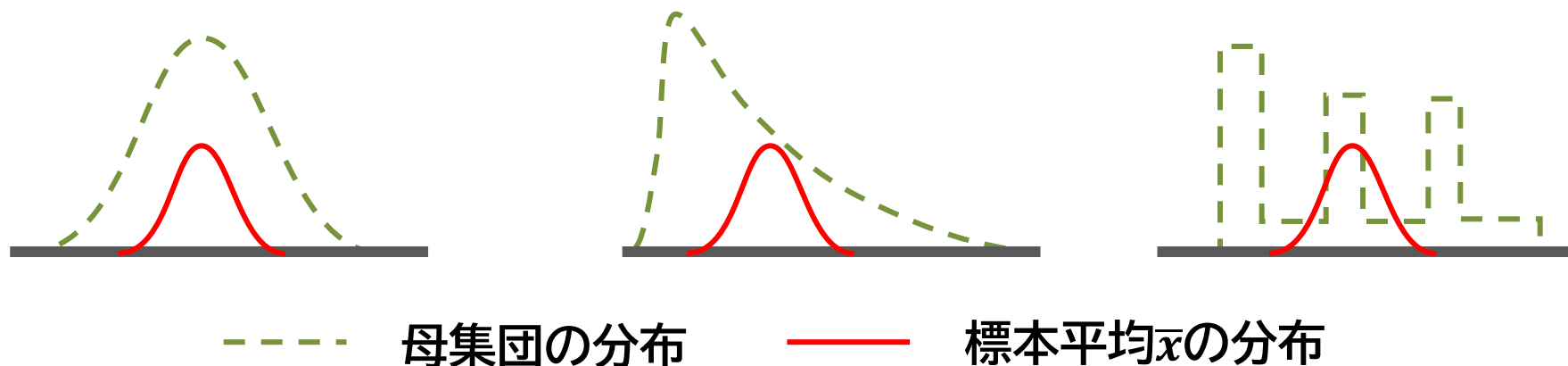


# 中心極限定理

母集団から標本をサンプリングして、標本平均 $\bar{x}$ を計算することを繰り返すと、標本平均 $\bar{x}$ 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

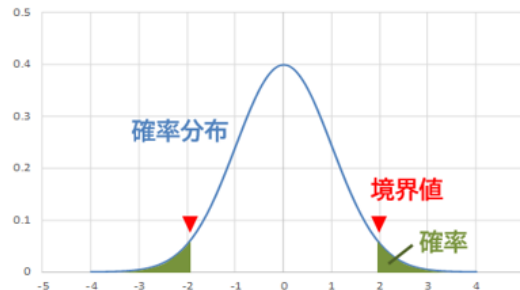


## ポイント

検定の多くは、**平均値に差があるかどうか**を確率的に評価したもの、またはその派生

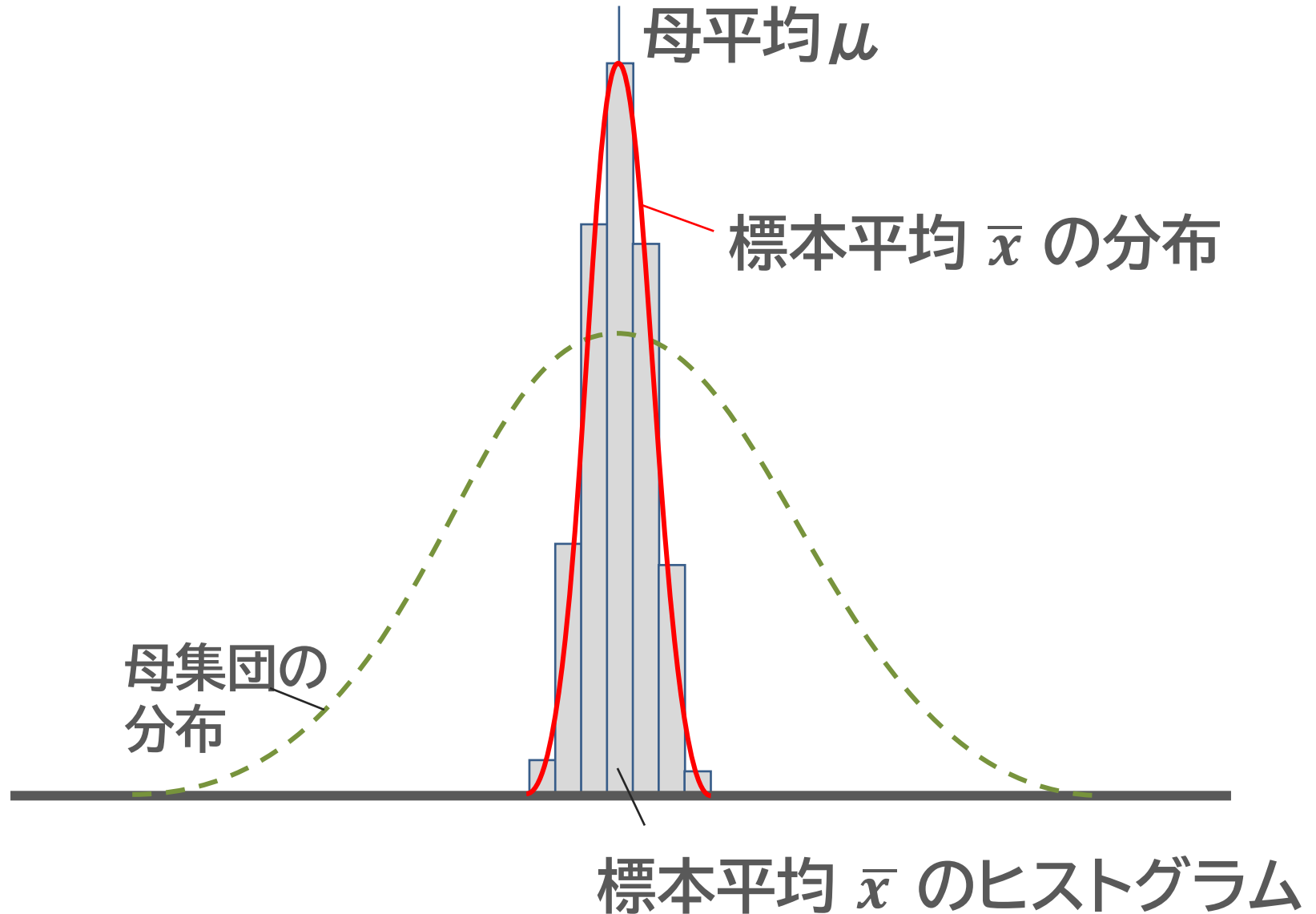


- 平均値を推定したとき、それがどれだけばらつくか(真の値から外れるか)には法則性があり、ある**分布**に従う
- 分布は、事象が起こる**確率**を表したもの



- 中心極限定理
- 正規分布

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？

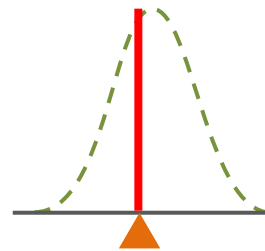
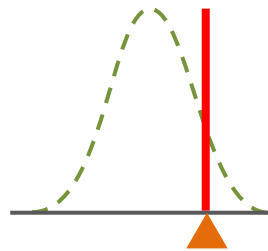
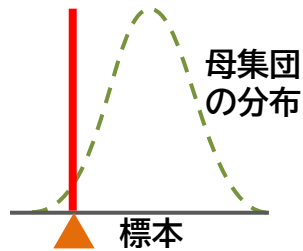


# 標本数を変えると...

標本数

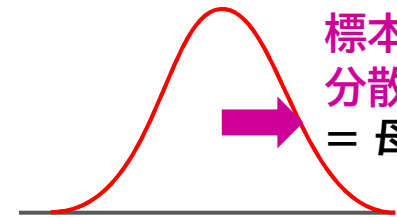
$n = 1$

標本平均 $\bar{x}$

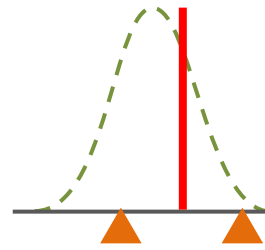
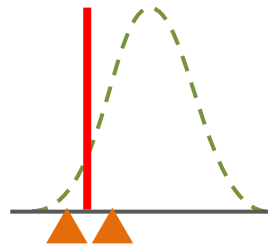
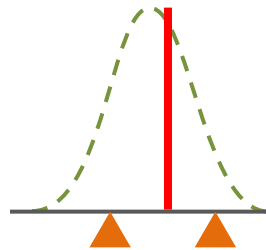


...

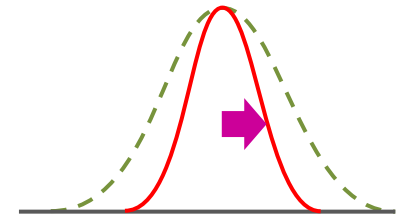
標本平均 $\bar{x}$ の分布



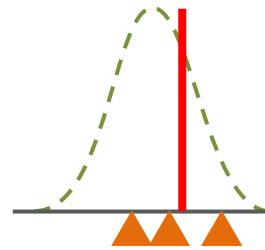
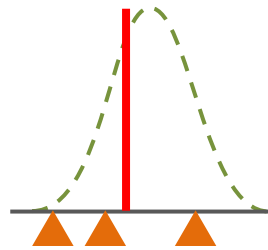
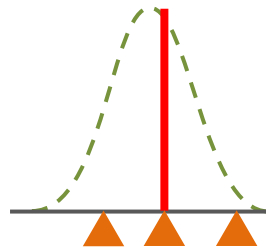
$n = 2$



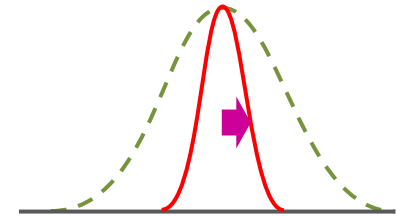
...



$n = 3$



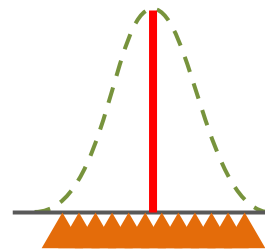
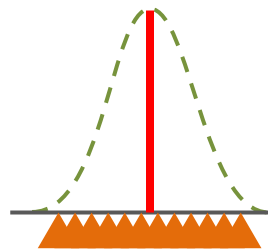
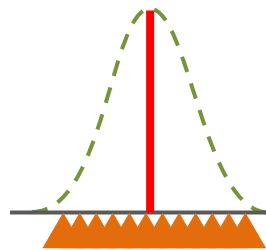
...



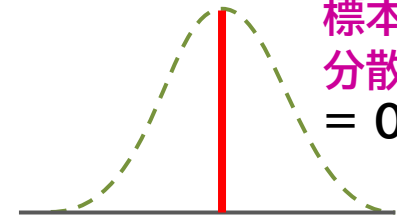
⋮

$n = N$

母集団全体



...

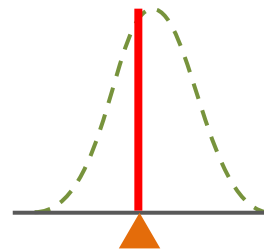
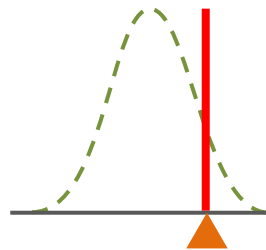
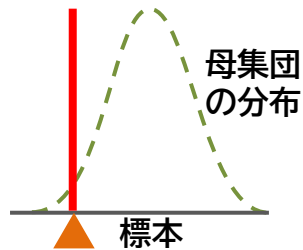


# 標本数を変えると...

標本数

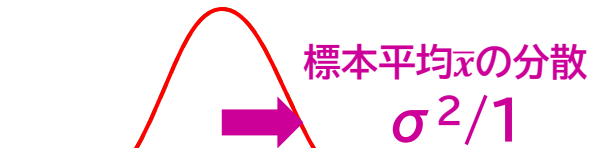
$n = 1$

標本平均 $\bar{x}$

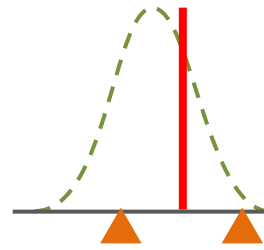
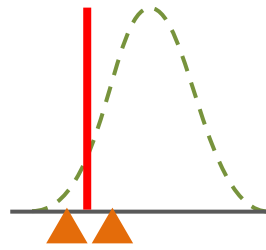
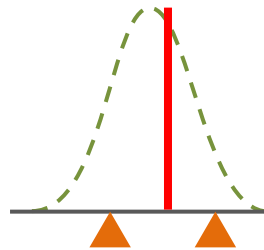


...

標本平均 $\bar{x}$ の分布



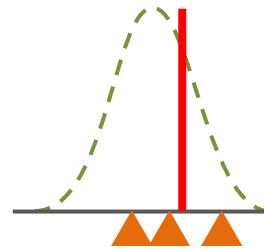
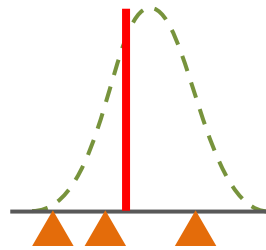
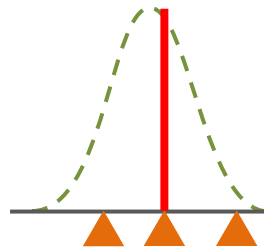
$n = 2$



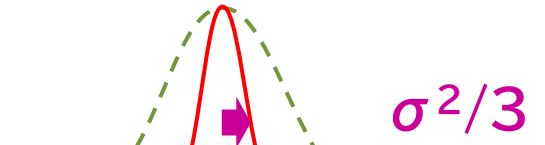
...



$n = 3$

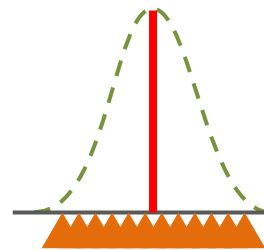
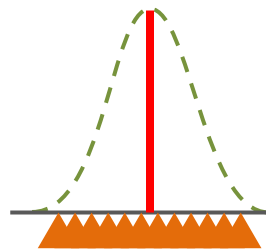
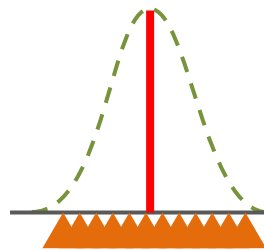


...

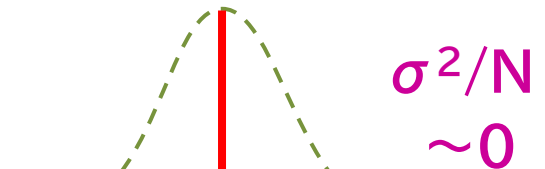


⋮

$n = N$   
母集団全体



...



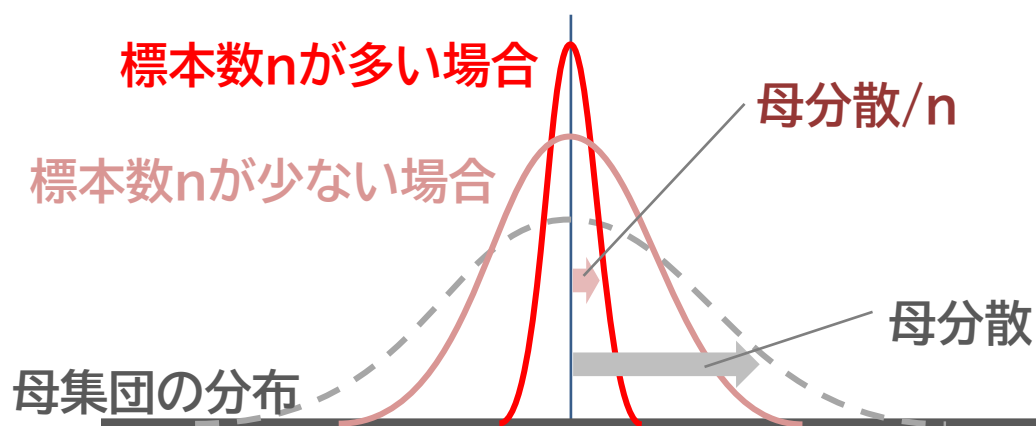
# 標本平均 $\bar{x}$ の分布の性質

- 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。  
 $n=\text{母集団数}N$  なら、全数検査なので、母平均 $\mu$ とのずれはゼロになる。

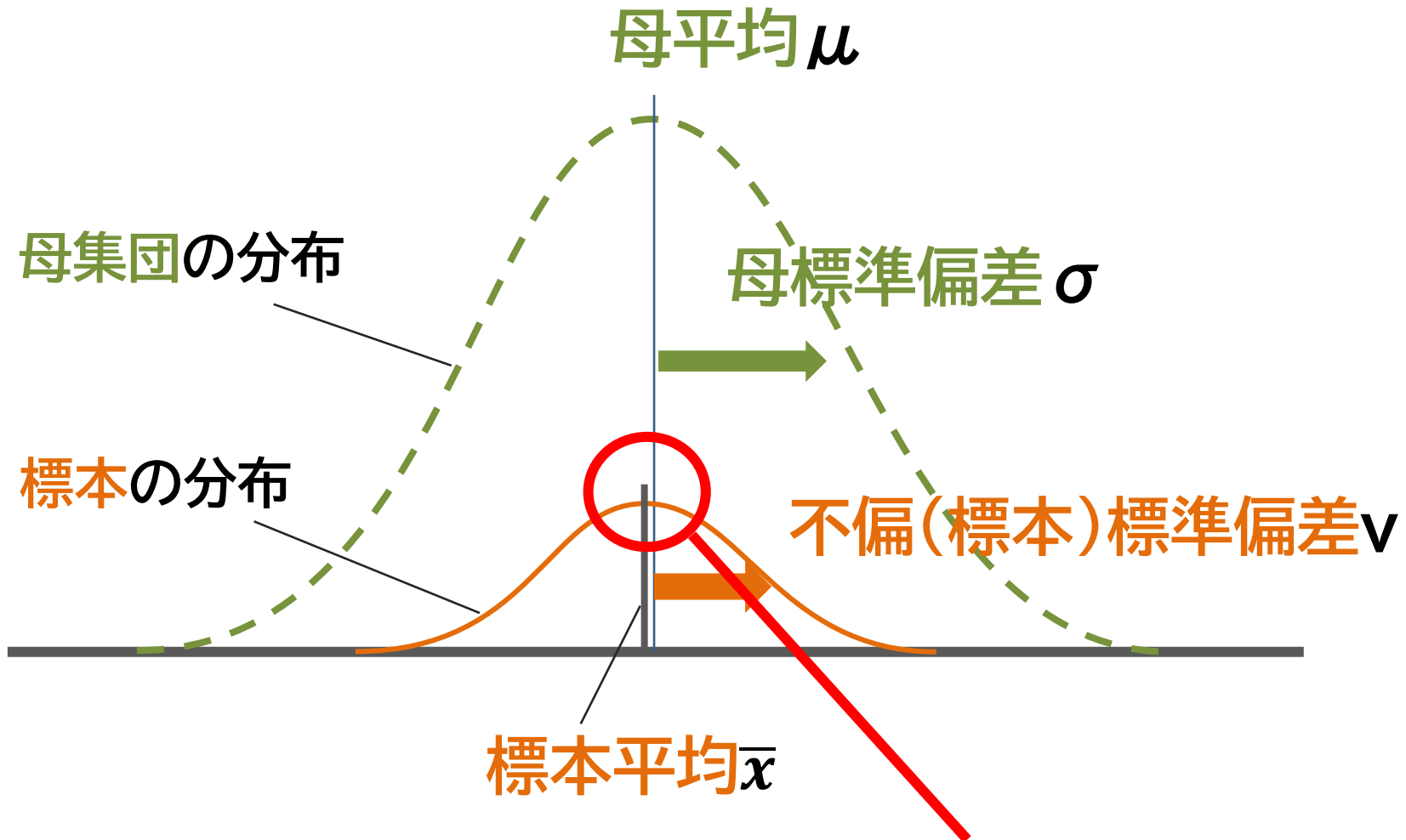
- 分散は、**母分散 $\sigma^2$ の $1/n$** になる

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 $\mu$ (真実)と、標本平均 $\bar{x}$ (推定)のズレ

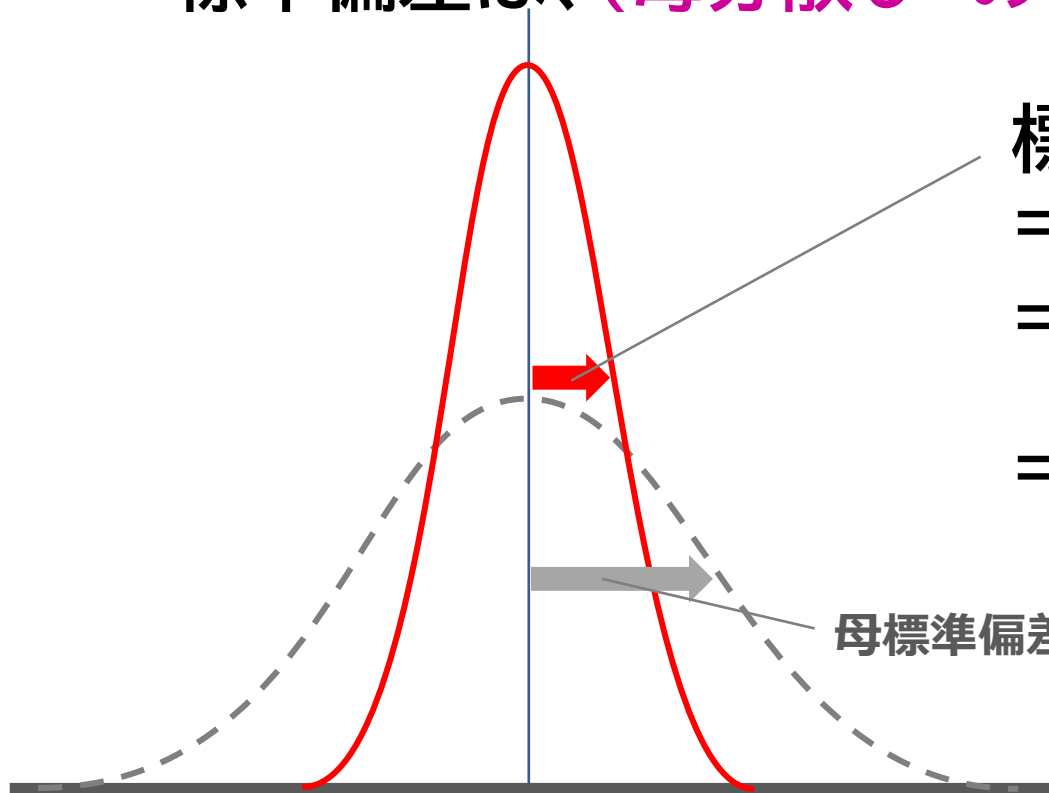


ズレの大きさを標本数 $n$ で表せる!!

# 標準誤差

- 標本平均 $\bar{x}$ のばらつきを標準偏差で表したものの分散は、母分散 $\sigma^2$ の  $1/n$

標準偏差は、(母分散 $\sigma^2$ 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 $\sigma^2/n$ の平方根  
$$= \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

母標準偏差 $\sigma$

# 母分散 $\sigma^2$ の推定

母平均  $\mu$  ← 標本平均  $\bar{x}$   
一致が期待できる

母分散  $\sigma^2$  ~~←~~ 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  $v^2$

$\mu$ : 平均(mean)のm  
 $\sigma$ : 標準偏差(standard deviation)のs  
に相当するギリシャ文字

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

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標本平均  
③その2乗

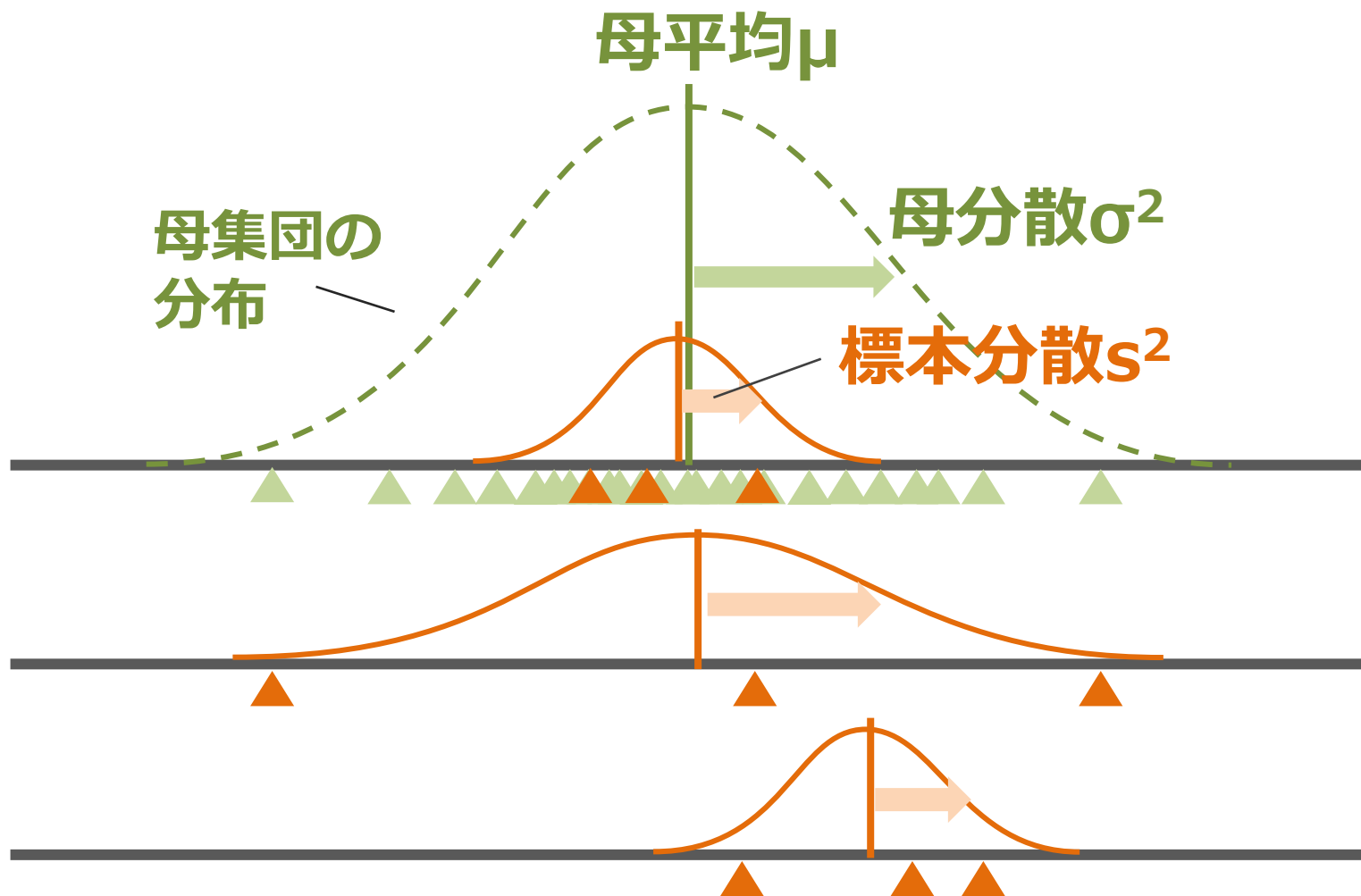
④その全要素(iが1からnまで)の合計

# 不偏(標本)分散

⑤n-1で割る

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

標本分散 $s^2$ は、母分散 $\sigma^2$ よりも小さくなる



母分散  $\sigma^2$

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

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

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

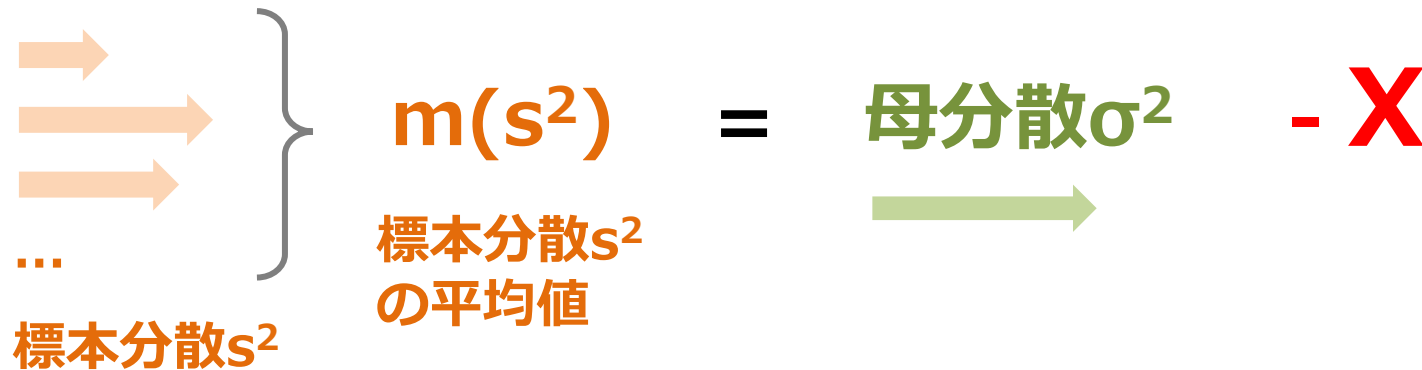
不偏(標本)分散  $v^2$

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 $\mu$ 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？

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…



The diagram illustrates the relationship between sample variance and population variance. On the left, three orange arrows of increasing length point to the right, with an ellipsis (...) below them. A large curly bracket groups these arrows. Below the arrows is the text '標本分散 $s^2$ '. To the right of the bracket is the expression  $m(s^2)$  in orange, followed by an equals sign. Below  $m(s^2)$  is the text '標本分散 $s^2$ の平均値'. To the right of the equals sign is the expression '母分散 $\sigma^2$ ' in green, with a green arrow pointing to it from below. To the right of this is a minus sign and a large red 'X'.

$$\left. \begin{array}{c} \text{→} \\ \text{→} \\ \text{→} \\ \dots \end{array} \right\} m(s^2) = \text{母分散}\sigma^2 - X$$

標本分散 $s^2$

標本分散 $s^2$ の平均値

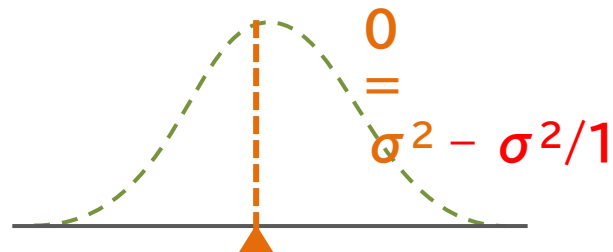
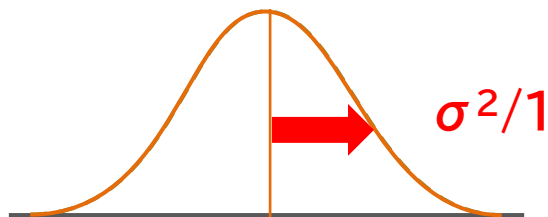


標本数

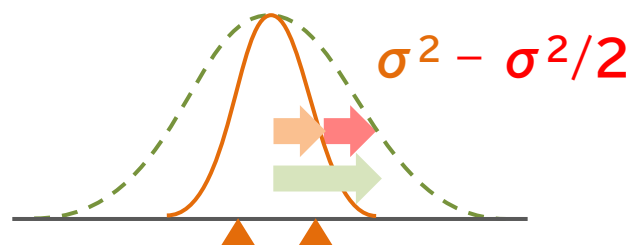
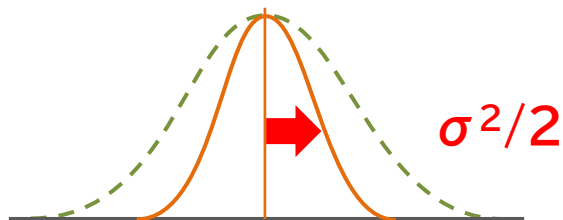
➡ 標本平均 $\bar{x}$ の計算を繰り返したときの分散

➡ 標本分散 $s^2$ の計算を繰り返したときの平均値 $m(s^2)$

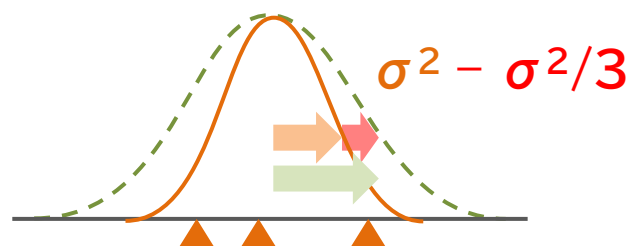
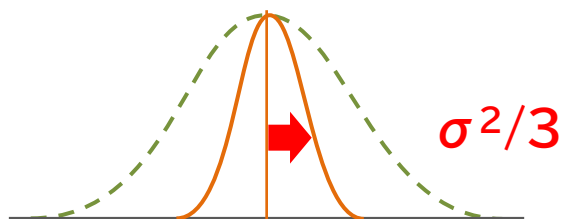
$n = 1$



$n = 2$

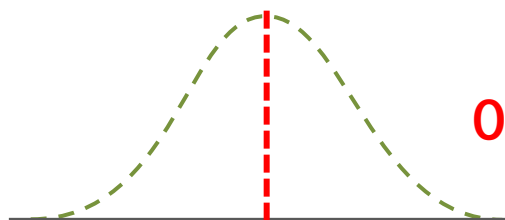


$n = 3$

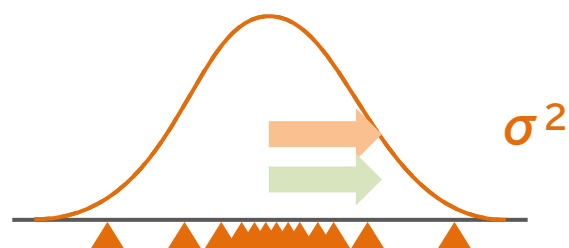


⋮

$n = N$

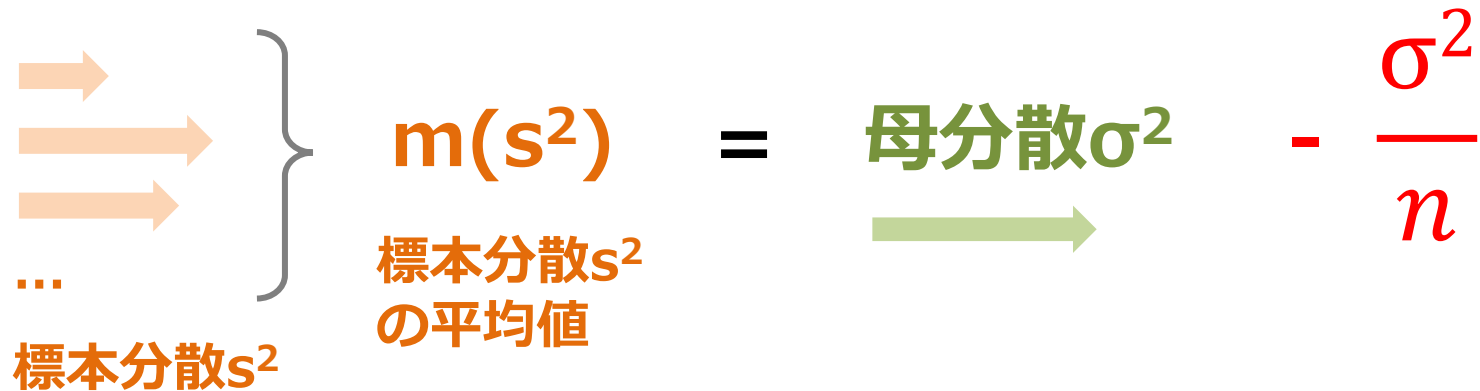


母集団全体



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…


$$\left. \begin{array}{l} \text{→} \\ \text{→} \\ \text{→} \\ \dots \\ \text{標本分散}s^2 \end{array} \right\} m(s^2) = \text{母分散}\sigma^2 - \frac{\sigma^2}{n}$$

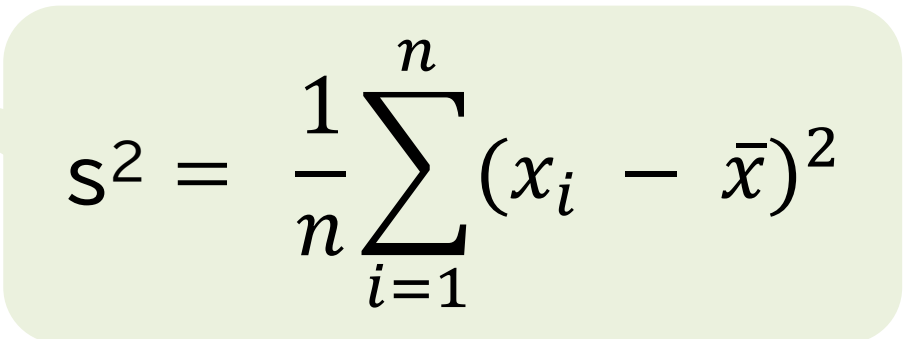
標本分散 $s^2$ の平均値

式を変形して、**母分散  $\sigma^2$** を出してみると…

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

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

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


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

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

母平均  $\mu$  ← 標本平均  $\bar{x}$   
一致が期待できる

母分散  $\sigma^2$  ~~←~~ 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  $v^2$

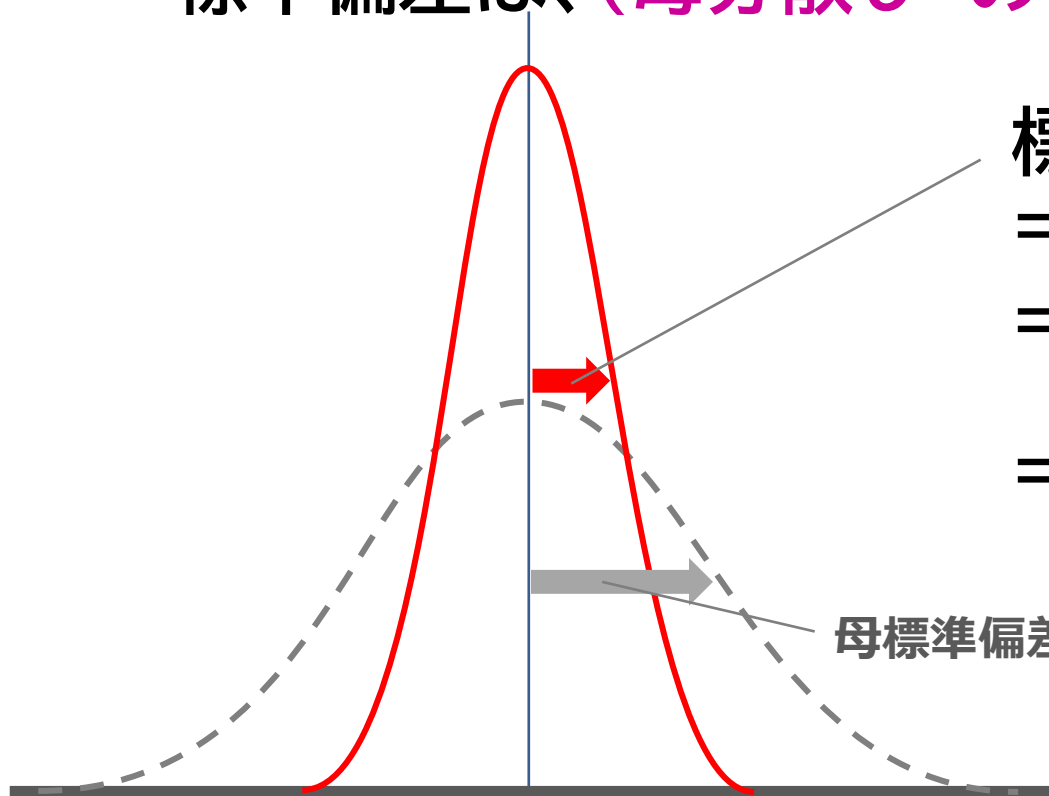
$\mu$ : 平均(mean)のm  
 $\sigma$ : 標準偏差(standard deviation)のs  
に相当するギリシャ文字

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

# 標準誤差

- 標本平均 $\bar{x}$ のばらつきを標準偏差で表したものの分散は、母分散 $\sigma^2$ の  $1/n$

標準偏差は、(母分散 $\sigma^2$ 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 $\sigma^2/n$ 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

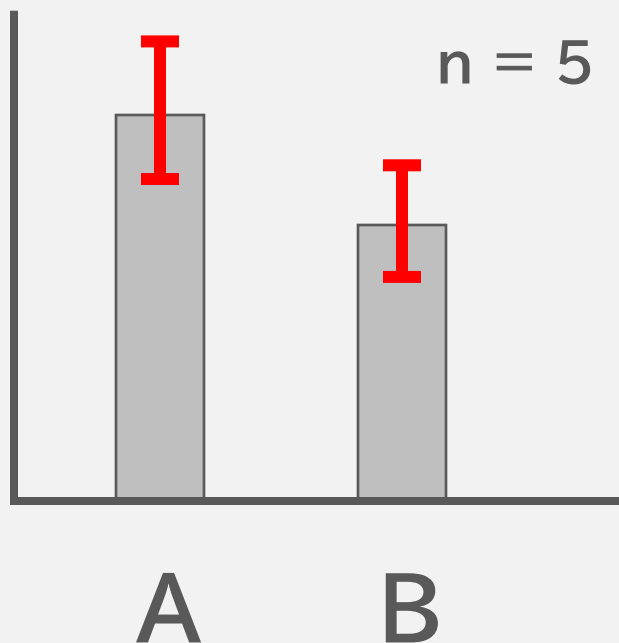


図1 A群とB群の\*\*の違い  
それぞれ5個体を測定した。**エ**  
**ラ**バーは**標準偏差**を表す

エラーバーが**標準偏差**  $s$



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**  $v / \sqrt{n}$



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！】

$n \geq 3$ のとき、標準偏差  $>$  標準誤差となり、エラーバーが短くなるので、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。

# n-1の意味

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

- 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からです。

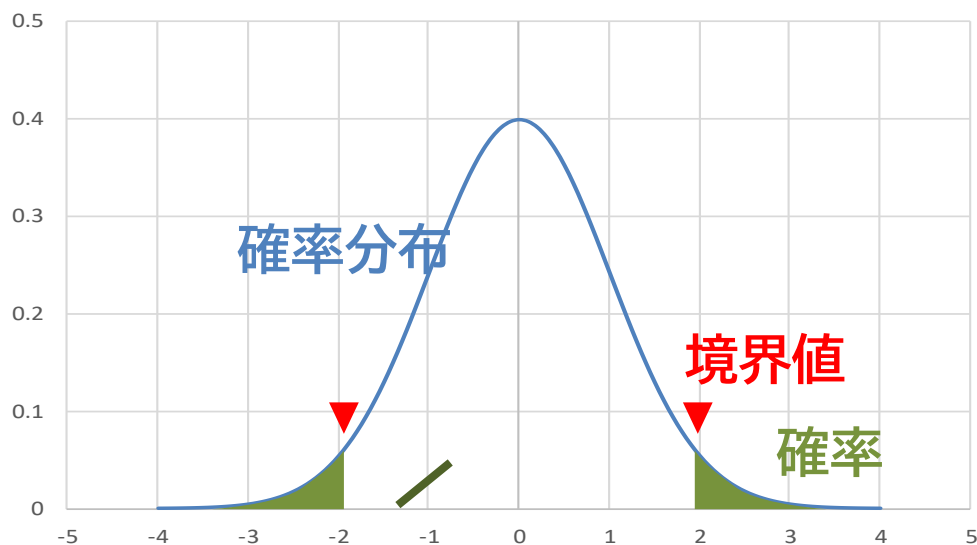
Excel関数も、標準偏差を求めるstdev関数は、不偏標本標準偏差( $n-1$ で割る)を計算しています。

$n$ 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

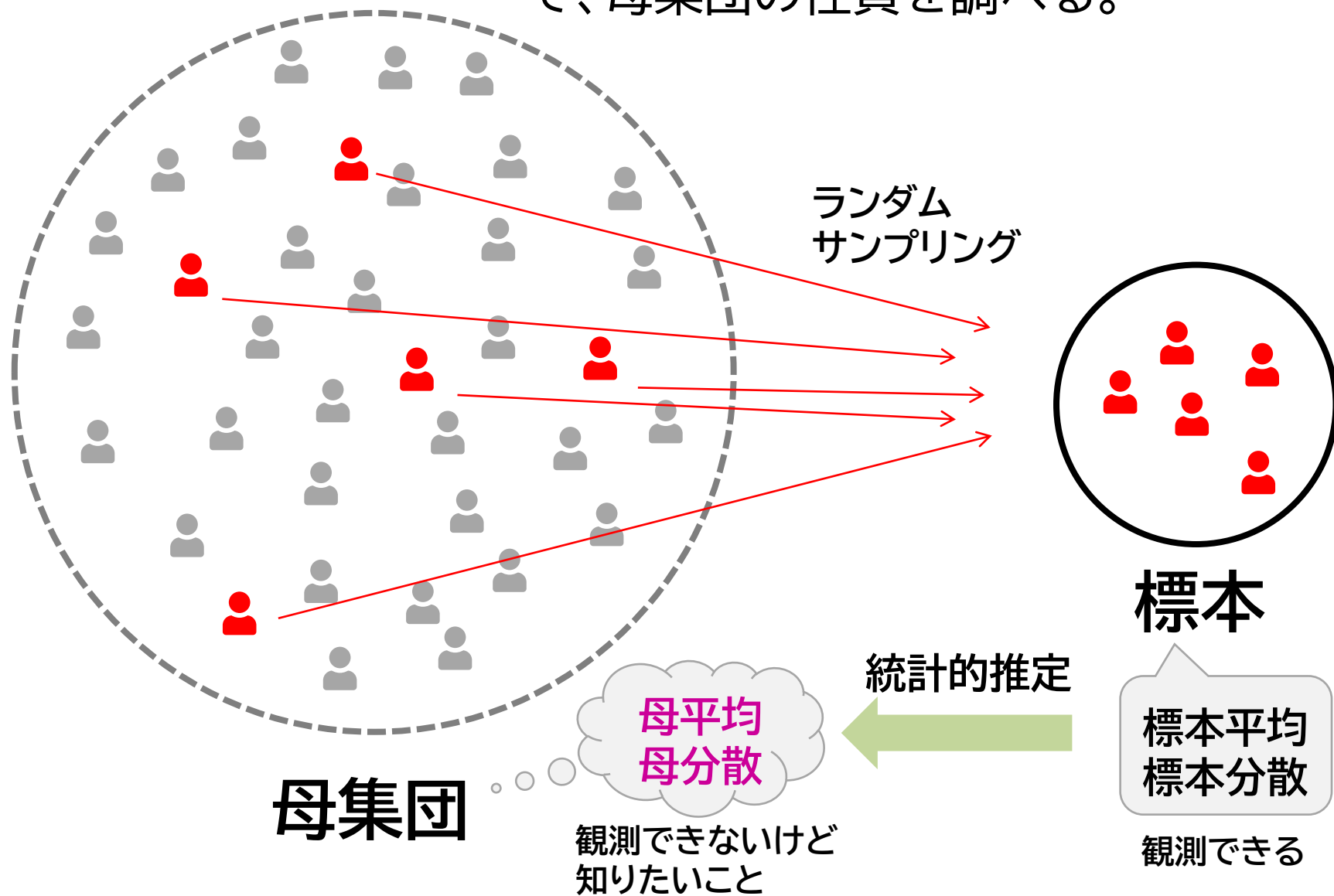


# 確率分布と その使い方



# 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 $\mu$ はこの値」、「母分散 $\sigma^2$ はこの値」のように、一つの代表値を決める方法

# 区間推定



「東京都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

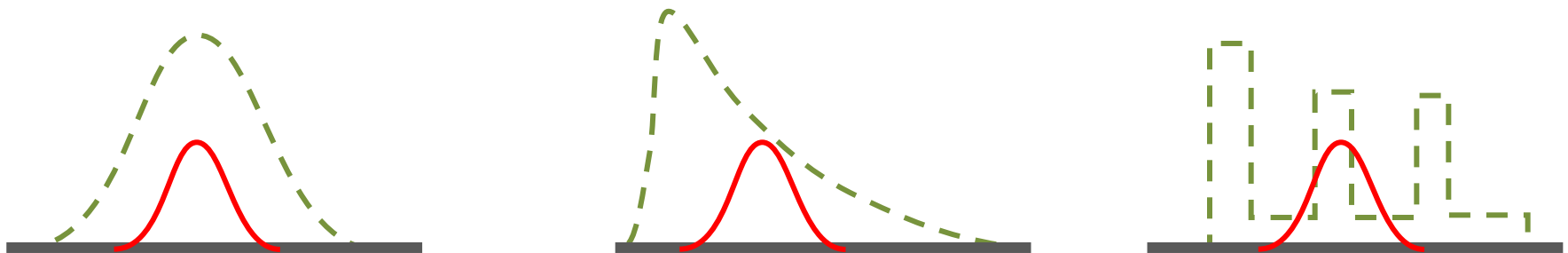
「東京都の95%の男子の身長は170.2～174.6 cmである」

# 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 $\bar{x}$ を計算することを繰り返すと、標本平均 $\bar{x}$ 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

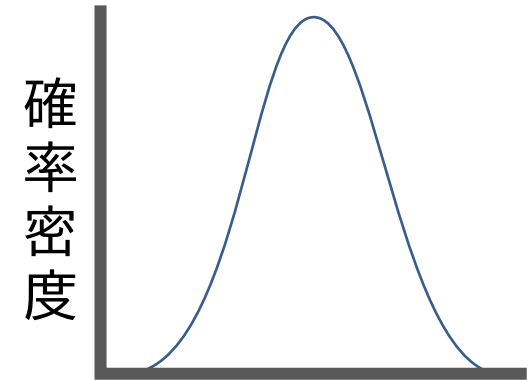


--- 母集団の分布

— 標本平均 $\bar{x}$ の分布

# 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

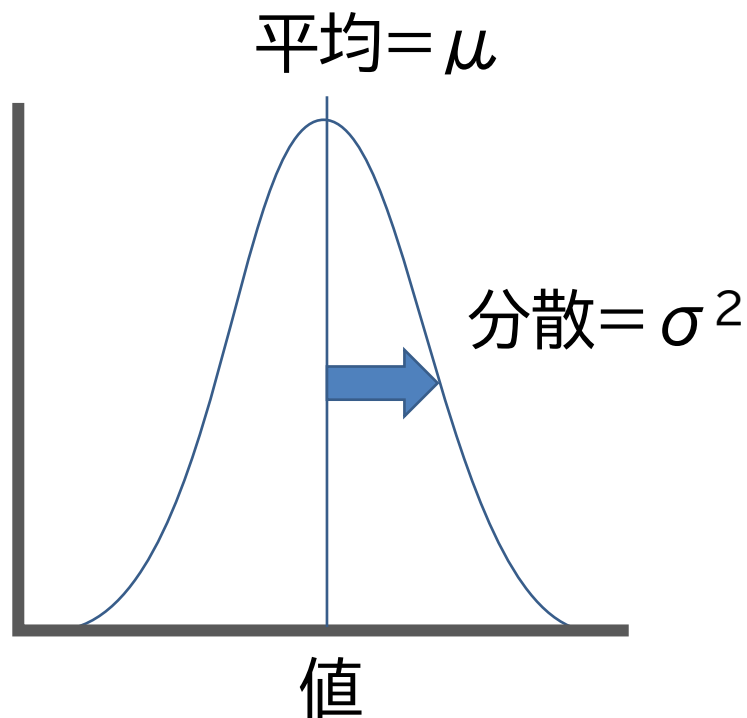
- ✓ 身長分布
- ✓ 測定誤差分布
- ✓ 自然界で起こるゆらぎ
- ✓ 標本平均 $\bar{x}$ の分布

など

# 標準正規分布

# 正規分布

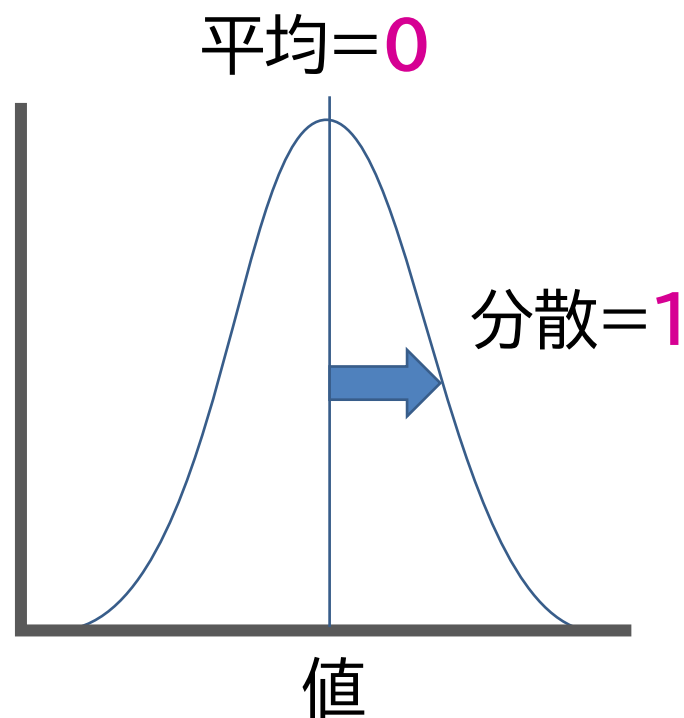
無限にある



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

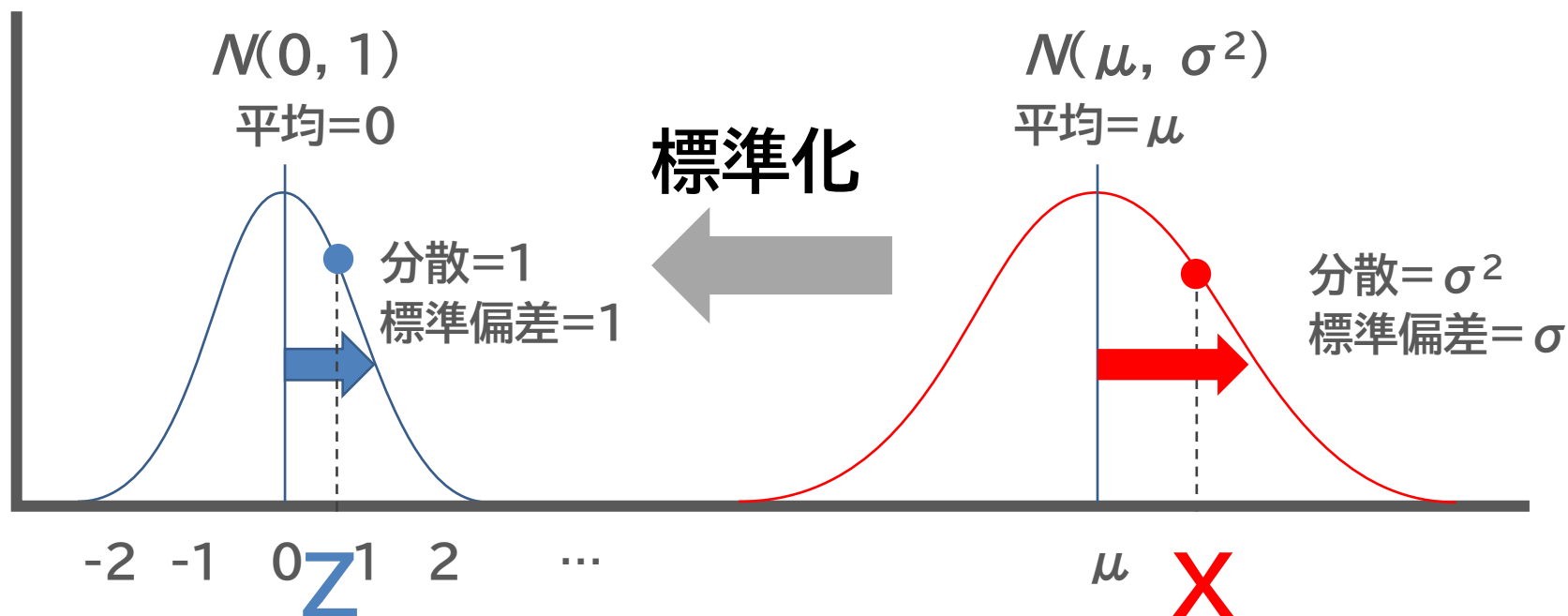


$N(0, 1)$

# 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。

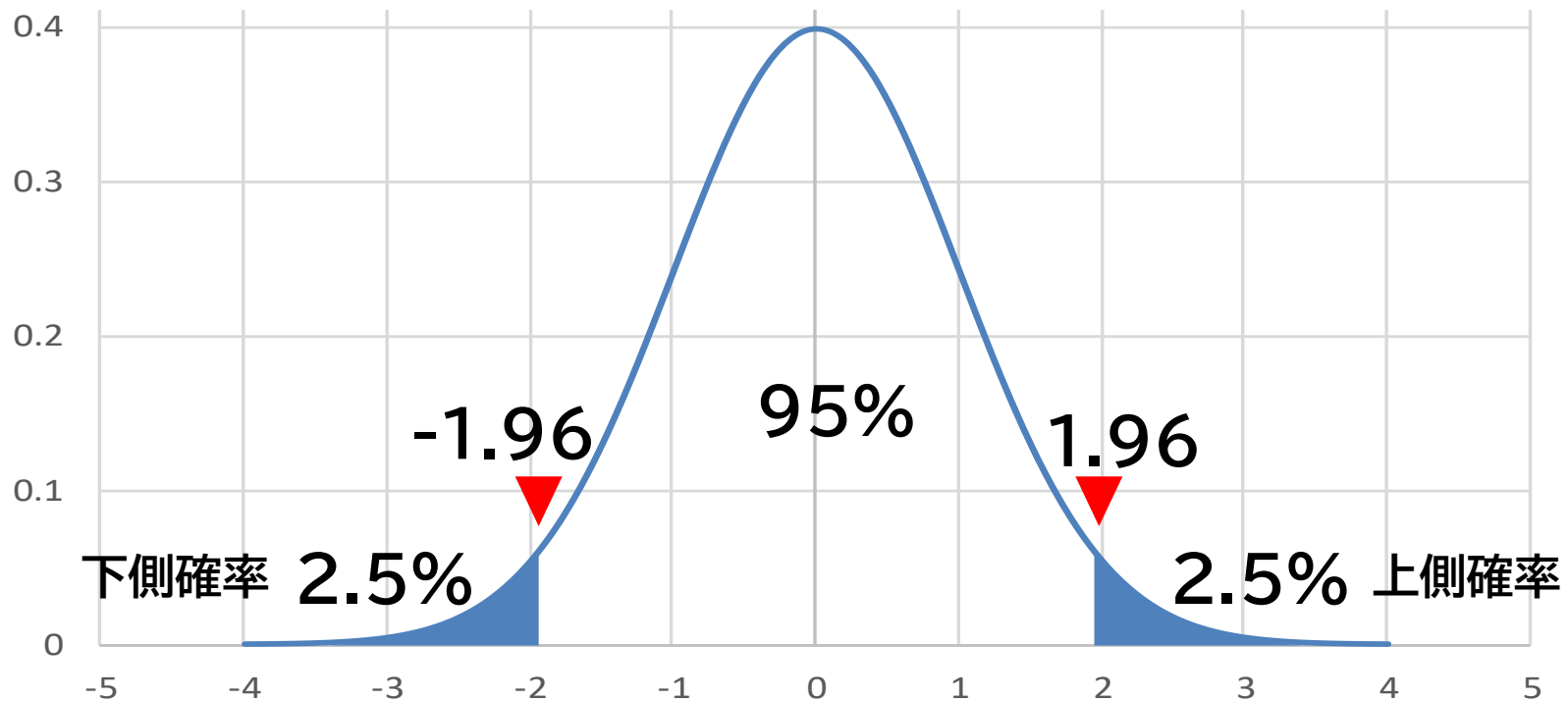


中央を $\mu$ ずらして、幅を1に合わせているだけ！



# 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

出典

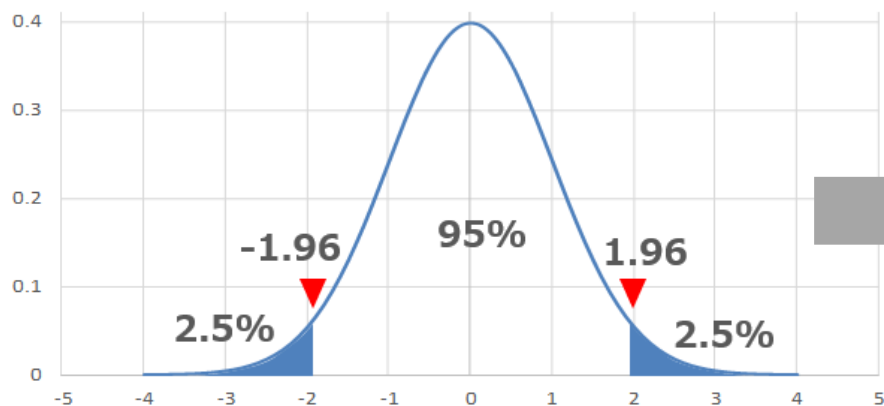
<https://to-kei.net/distribution/normal-distribution/table/>

u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
2.5	0.00620	0.00602	0.00584	0.00566	0.00549	0.00533	0.00517	0.00500	0.00484	0.00468
2.6	0.00453	0.00437	0.00421	0.00405	0.00390	0.00375	0.00359	0.00344	0.00329	0.00314
2.7	0.00299	0.00284	0.00269	0.00254	0.00239	0.00225	0.00210	0.00196	0.00182	0.00168
2.8	0.00154	0.00140	0.00126	0.00112	0.00100	0.00087	0.00075	0.00063	0.00052	0.00041
2.9	0.00030	0.00020	0.00010	0.00005	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

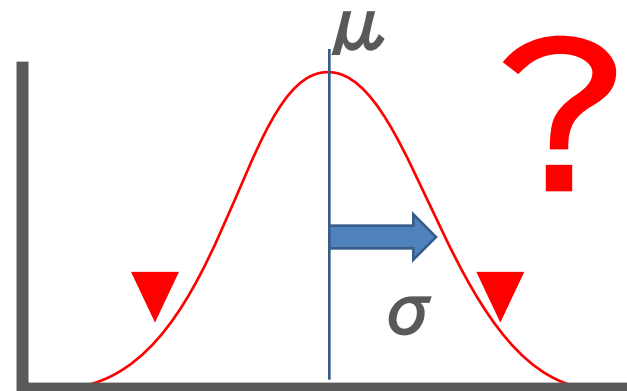
# 区間推定の考え方

- 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 $\mu$ 、分散 $\sigma^2$ 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)



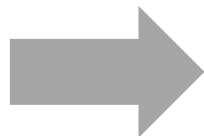
標準正規分布



もとの正規分布

# 標準化

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

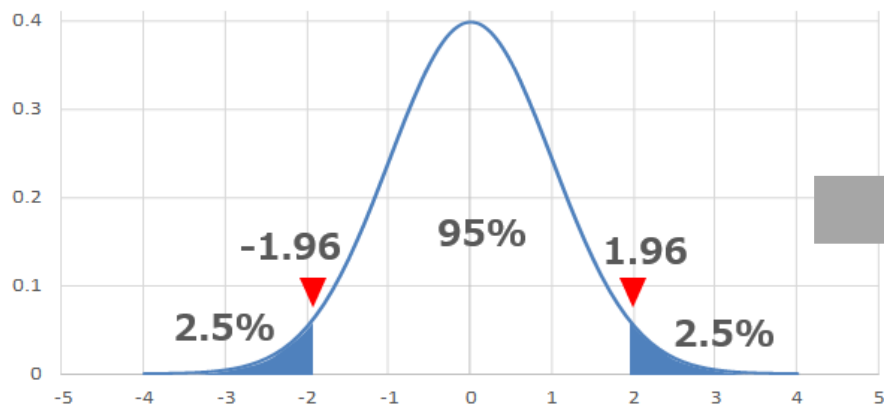


# 標準化の逆

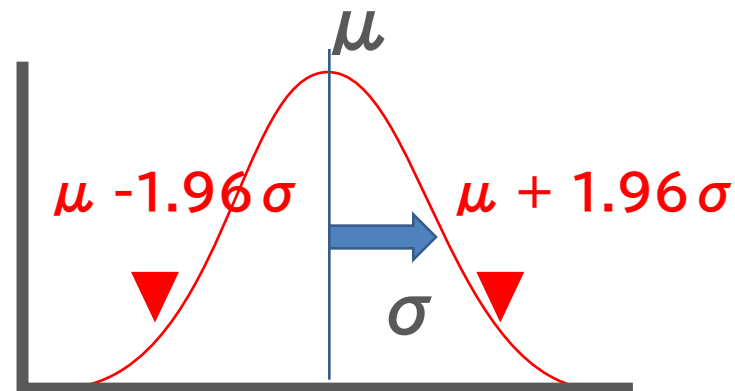
$$X = \mu + Z\sigma$$

例)  $Z = 1.96$ なら、

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



標準正規分布



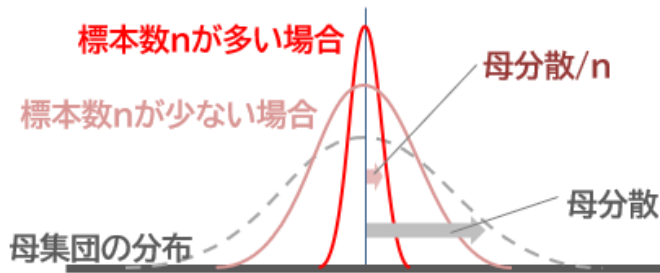
もとの正規分布

# 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 $\mu$ とのずれはゼロになる。
- 分散は、**母分散 $\sigma^2$ 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



$\mu$ 推定値:  $\bar{x}$

不偏標本標準偏差:  $\frac{s}{\sqrt{n}}$

を当てはめると...

95%信頼区間は、以下で求められる

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

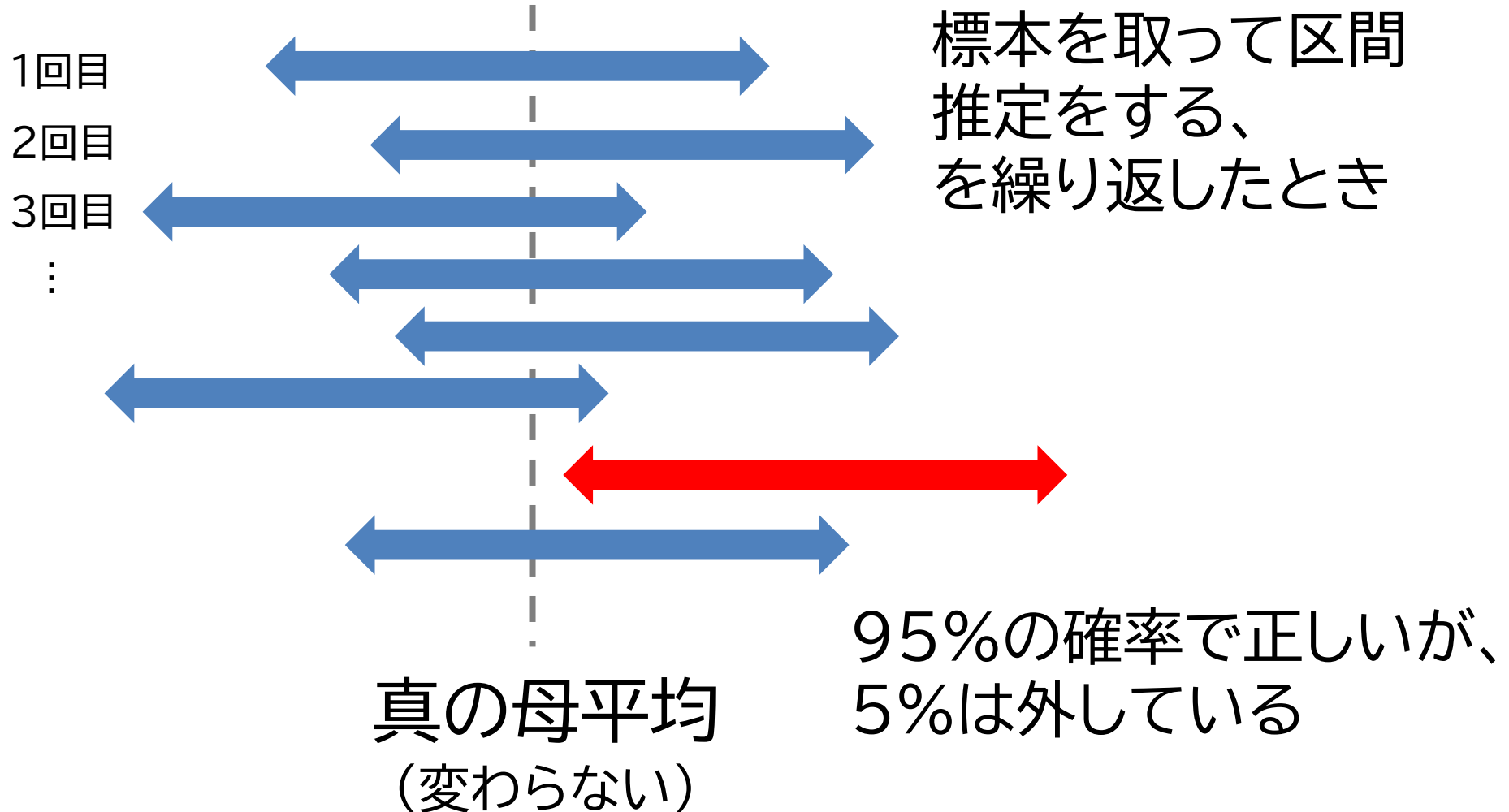
母平均 $\mu$ の推定値：標本平均  $\bar{x}$

推定値の標準偏差：不偏標本標準偏差  $\frac{v}{\sqrt{n}}$

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



# 点推定



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

# 区間推定

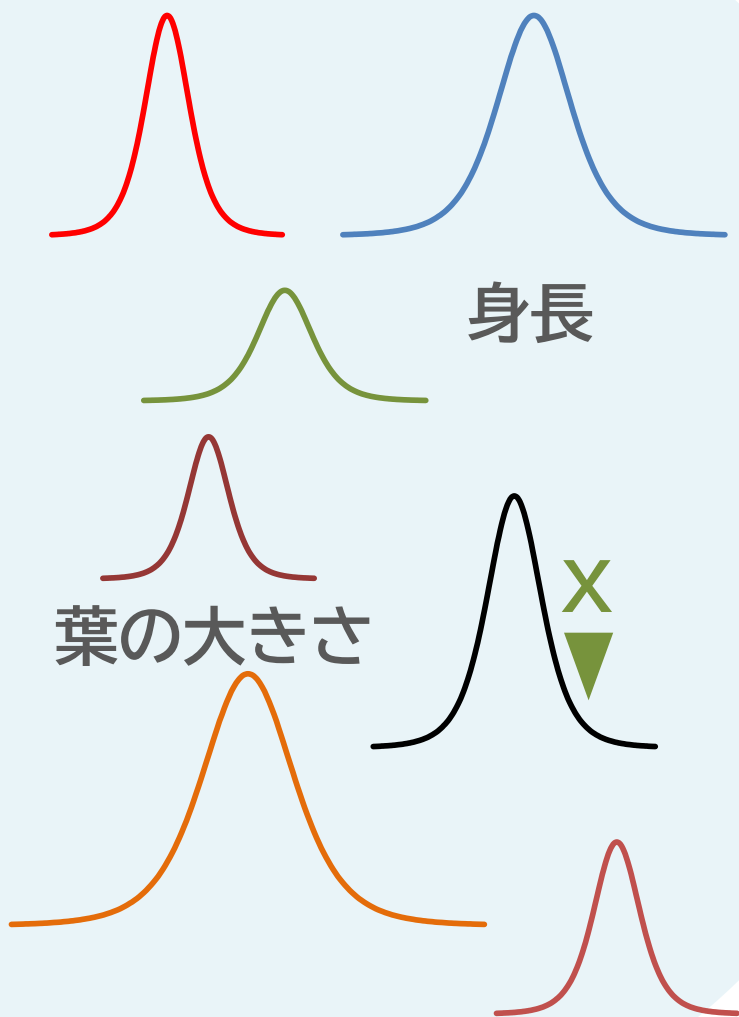


「東京都の男子の平均身長は、信頼係数95%で170.2～174.6 cmである」のように、幅を持たせて表現する方法

【注意】 以下の意味とは異なる

「東京都の95%の男子の身長は170.2～174.6 cmである」





身長

葉の大きさ

$X$

いろんな正規分布

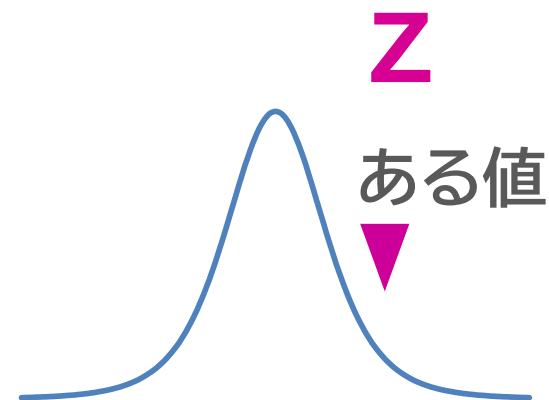
z変換

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



逆変換

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



標準正規分布

現実の具体的な問題

一般化

**正規分布  
万歳！**

# ただし現実は…

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

母平均 $\mu$ の推定値：標本平均  $\bar{x}$

推定値の標準偏差：不偏標本標準偏差  $\frac{v}{\sqrt{n}}$

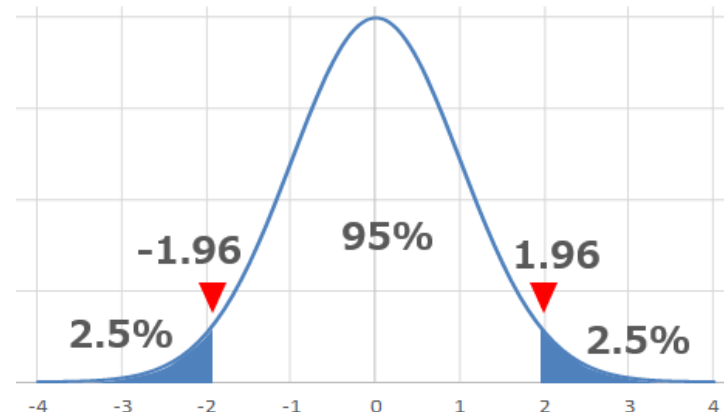
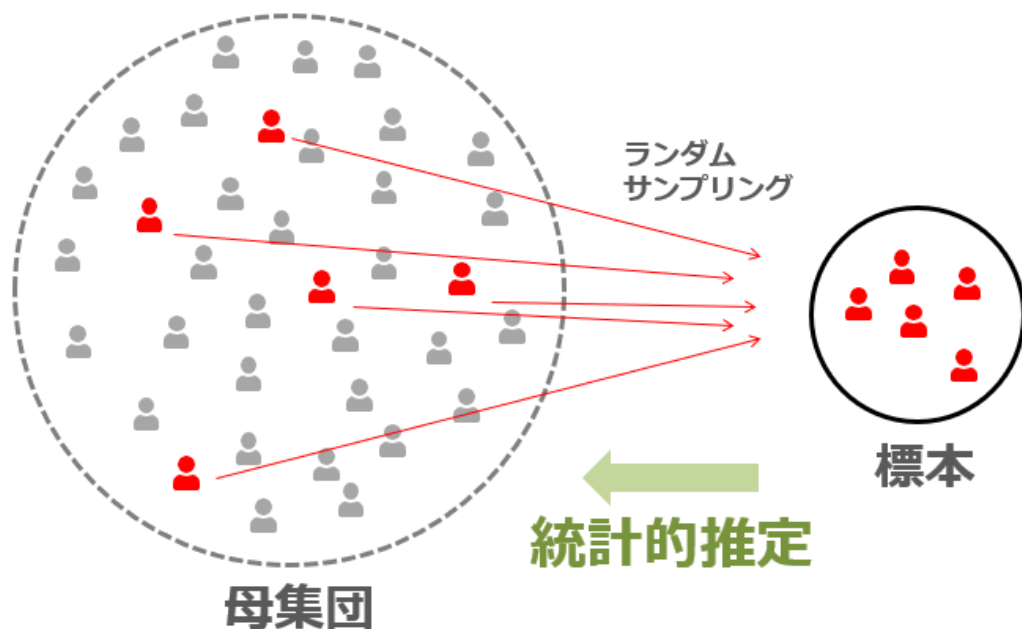
標本が少ない場合に、この推定精度が低く、  
正規分布と乖離してしまう。

# $t$ 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 $\mu$ 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.

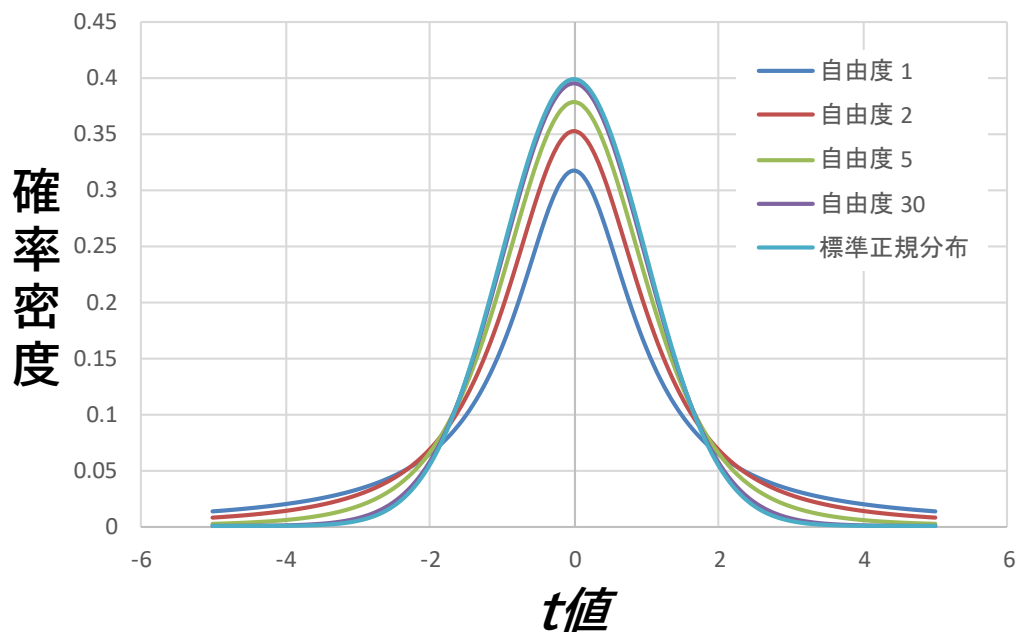
Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between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

## $t$ 分布表



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

Excelでは、T.DIST、T.INV関数で計算できる

自由度 $\nu$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
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 $\mu$ 、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。

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

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

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 $\mu$ 、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、

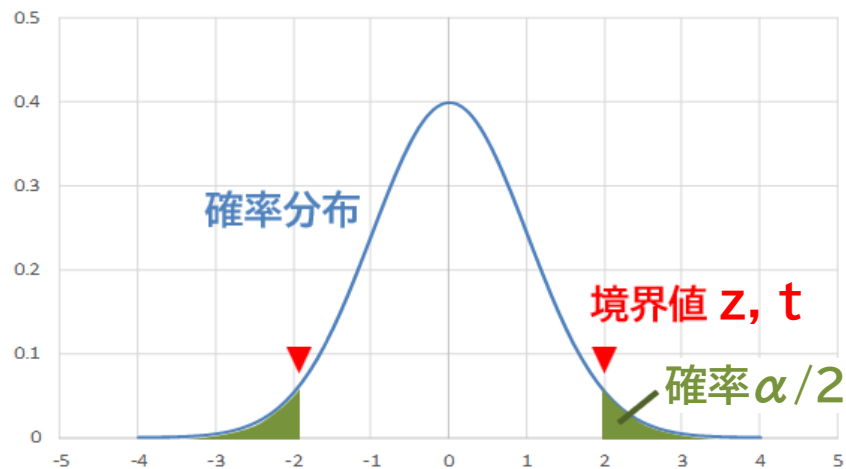
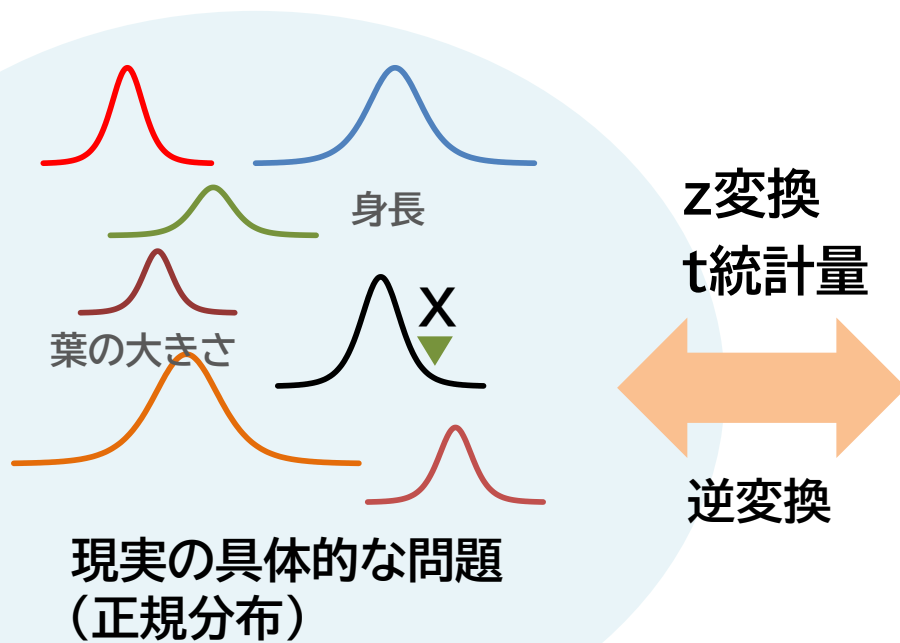
- ✓ 自由度 $=n-1$
- ✓ 確率 $\alpha = (100-a)/100$

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



標準的な分布  
( $t$ 分布、標準正規分布)

【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

# $t$ 検定

- ✓ 新しい数式などは出てこない
- ✓ 使い方のお作法だけ

# 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

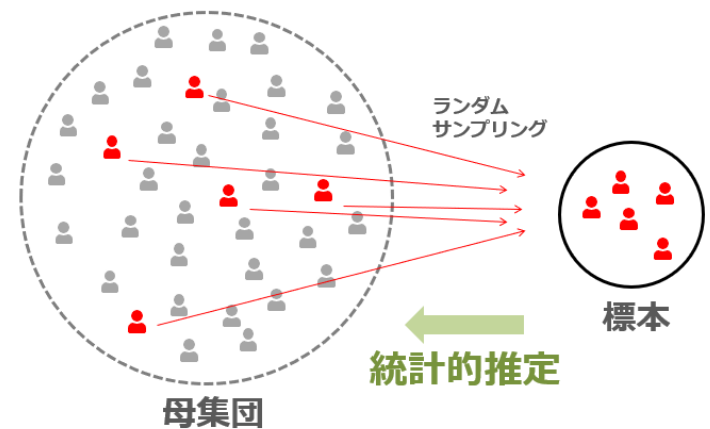
有意水準5%で  
帰無仮説は棄却されました。  
従って、\*\*\*です。

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 $\mu=95$ 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

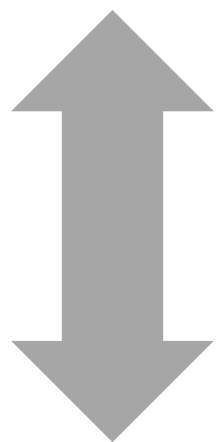
91, 90, 95, 88, 96, 89

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



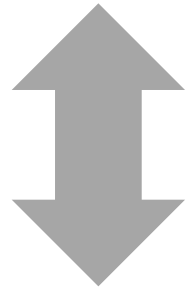
どちらでもよさそう  
だが…

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。

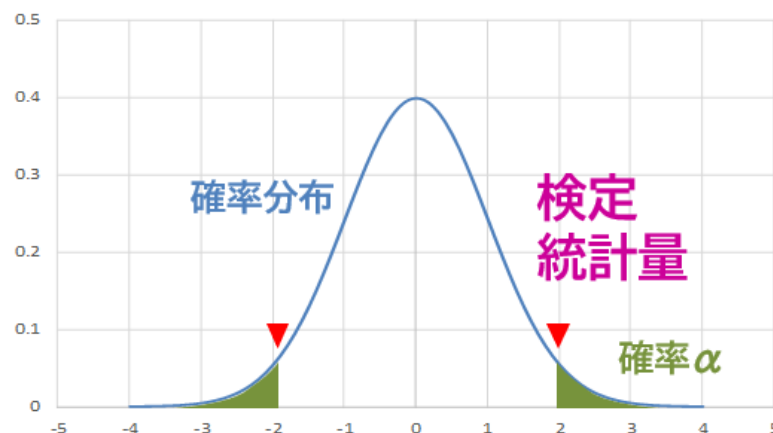
# ②検定統計量の計算

## 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。

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



## ②検定統計量の計算

標本平均

$\bar{x}$

91.5

不偏標本分散

$v^2$

10.7

母平均

$\mu$

95

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

-2.62

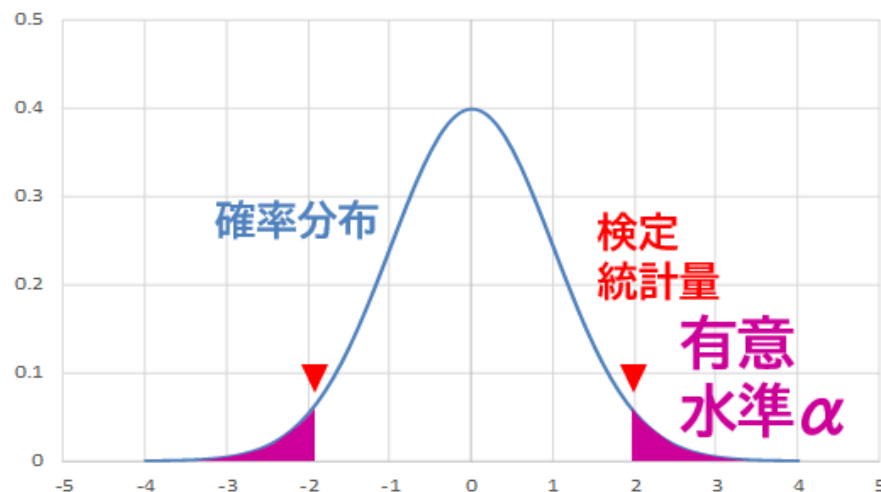


# ③仮説採否の評価

有意水準  $\alpha$  を0.05とする

## 有意水準 $\alpha$

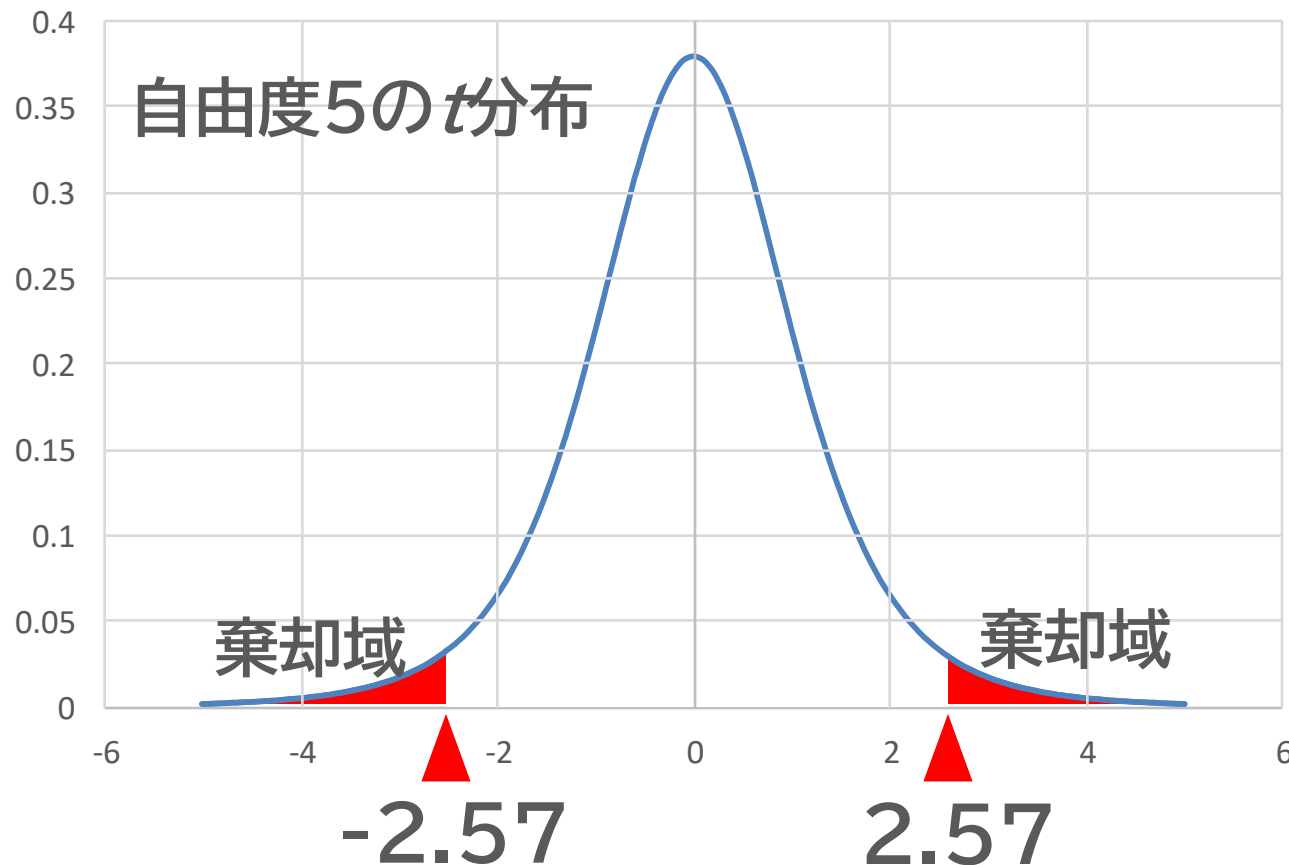
仮説を棄却するかどうかを決める基準の確率。これよりも小さい確率を持つ場合は、めったに起こらないことが起きていると考えられるため、帰無仮説(普通、変化がない)が棄却される。



# 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



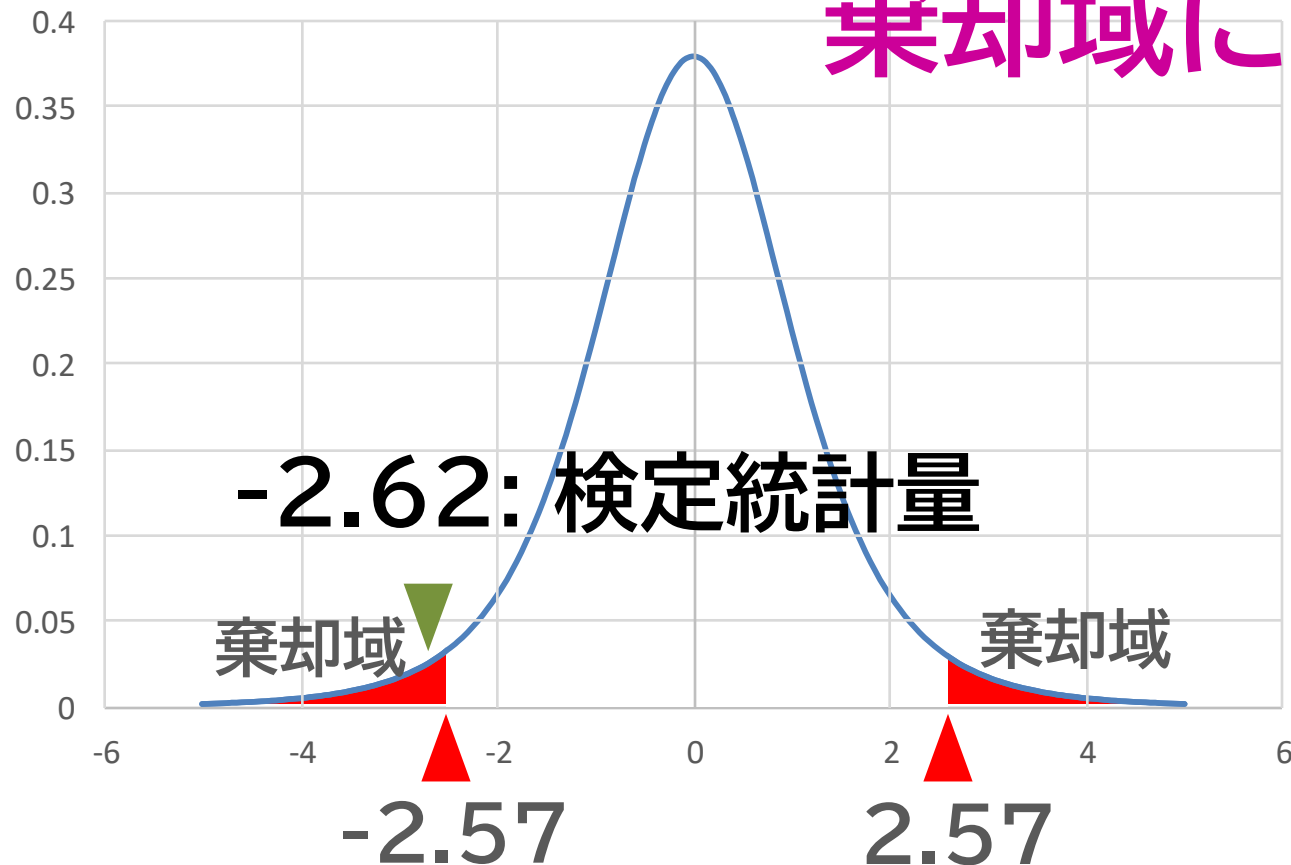
Excelで計算  
してもよい



# 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

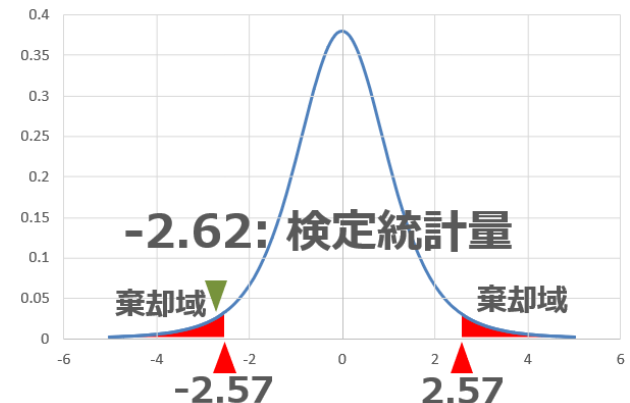
✓ 適切な分布を選ぶ

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？



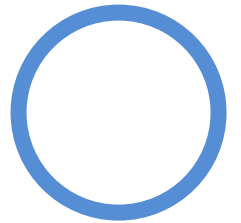
# 注意点

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。



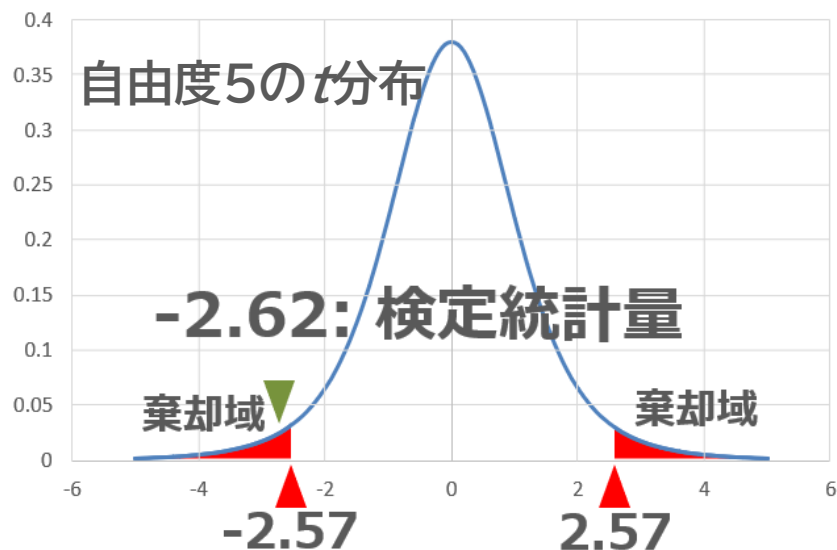
「帰無仮説が誤っているとは言えない」とは言える。



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、 $\alpha = 0.025$ 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。

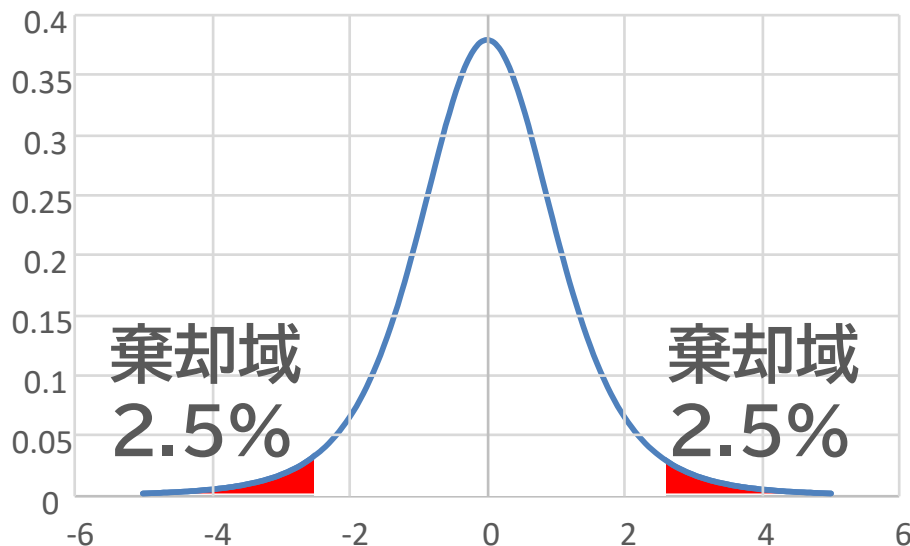
この違いに気を付けて正しく使い分けましょう。

# 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

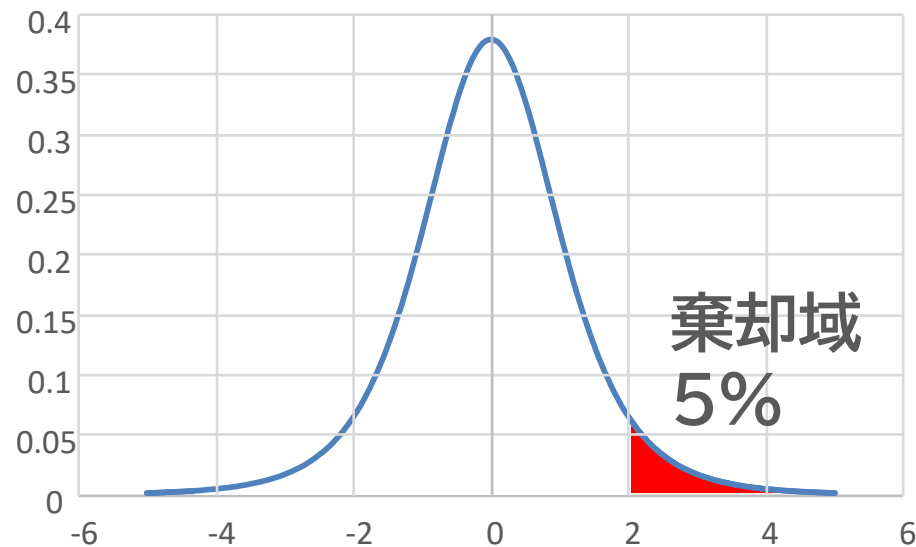
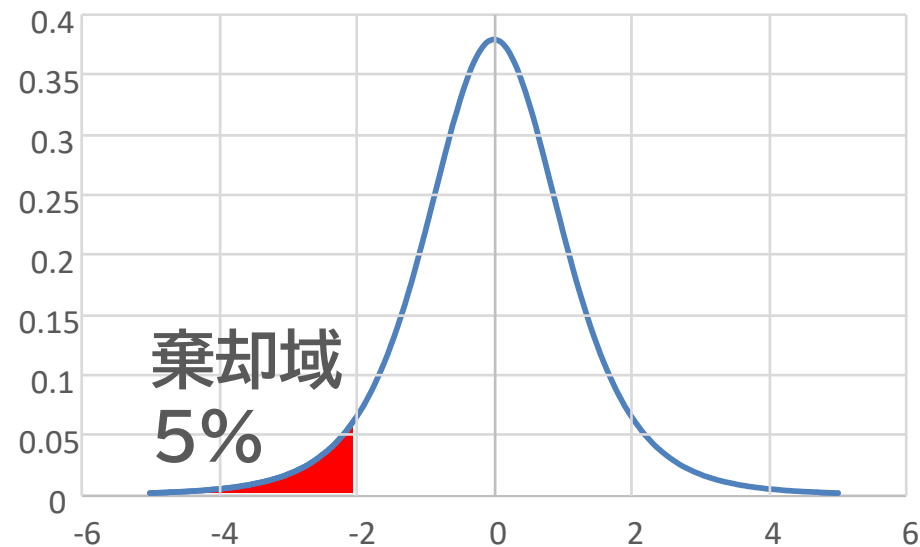
Aさんのカラオケ問題では、有意水準0.05を、その半分の0.025ずつに分け、 $t$ 分布の両側に割り当てて考えました



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？
  - 今年の給料は去年の＊＊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ります。

- 1群の $t$  検定

母集団の平均値が特定の値であるかどうかの検定

- 2群の $t$  検定

2つの群の平均値に差があるかどうかの検定

- ✓ 対応のある2群の場合
- ✓ 独立した2群の場合

# 1群の $t$ 検定

母集団の平均値が、特定の値かどうかを検定します

Aさんのカラオケ平均点が95点かどうかで行ったのは、  
実は、1群の $t$ 検定です

他の例)

工場のラインで規格どおりに製品が製造されているかどうか？

# 2群の $t$ 検定(対応あり)

「対応がある」とは、例えば以下のような場合です。

介入試験をおこない、試験食の摂取前後で数値を測定した

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

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

# 2群の $t$ 検定(対応あり)

実はこの問題は、次の手順で、1群の $t$ 検定として処理できます

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

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

# 2群の $t$ 検定(独立2群)

実験科学の分野などでよく使われます

例)

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

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

# 2群の $t$ 検定 (独立2群)

## 等分散の場合

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

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

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

検定統計量  $t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

自由度:  $n_1 + n_2 - 2$

帰無仮説: 2群の母集団の平均値は等しい

で、同様に検定できます

参考まで

# 2群の $t$ 検定(独立2群)

等分散が仮定できない場合    **ウェルチの方法**

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

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

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

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

帰無仮説: 2群の母集団の平均値は等しい

で、同様に検定できます

**参考まで**



# メッセージ

どんな検定でも

- 検定統計量
- 自由度
- 分布の計算方法

などさえ分かれば、身につけたステップで、**自分でできる！**

# 検定で 注意すること

- ✓ 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。

# 検定の二つの間違い

## 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、間違いだとして棄却してしまうこと]

この過誤を犯す確率は  $\alpha$  で表され、実は、その値のことを有意水準と呼んでいる。

$\alpha$ : 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いなのに、正しいとして棄却しないこと]

この過誤を犯す確率は  $\beta$  で表され、 $(1-\beta)$ 、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( $\beta$  が小さい)のが、よい検定とされる。

$\beta$ : ぼーんやりものの見逃し率

		帰無仮説が本当は	
		間違い (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 $\alpha$
	棄却 しない (陰性)	第二種の過誤 偽陰性 $\beta$	OK

  
 有意水準  $\alpha$   


第一種の過誤を起こさないように  $\alpha$  を下げて厳しく判定すると、 $\beta$  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  $\alpha$  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

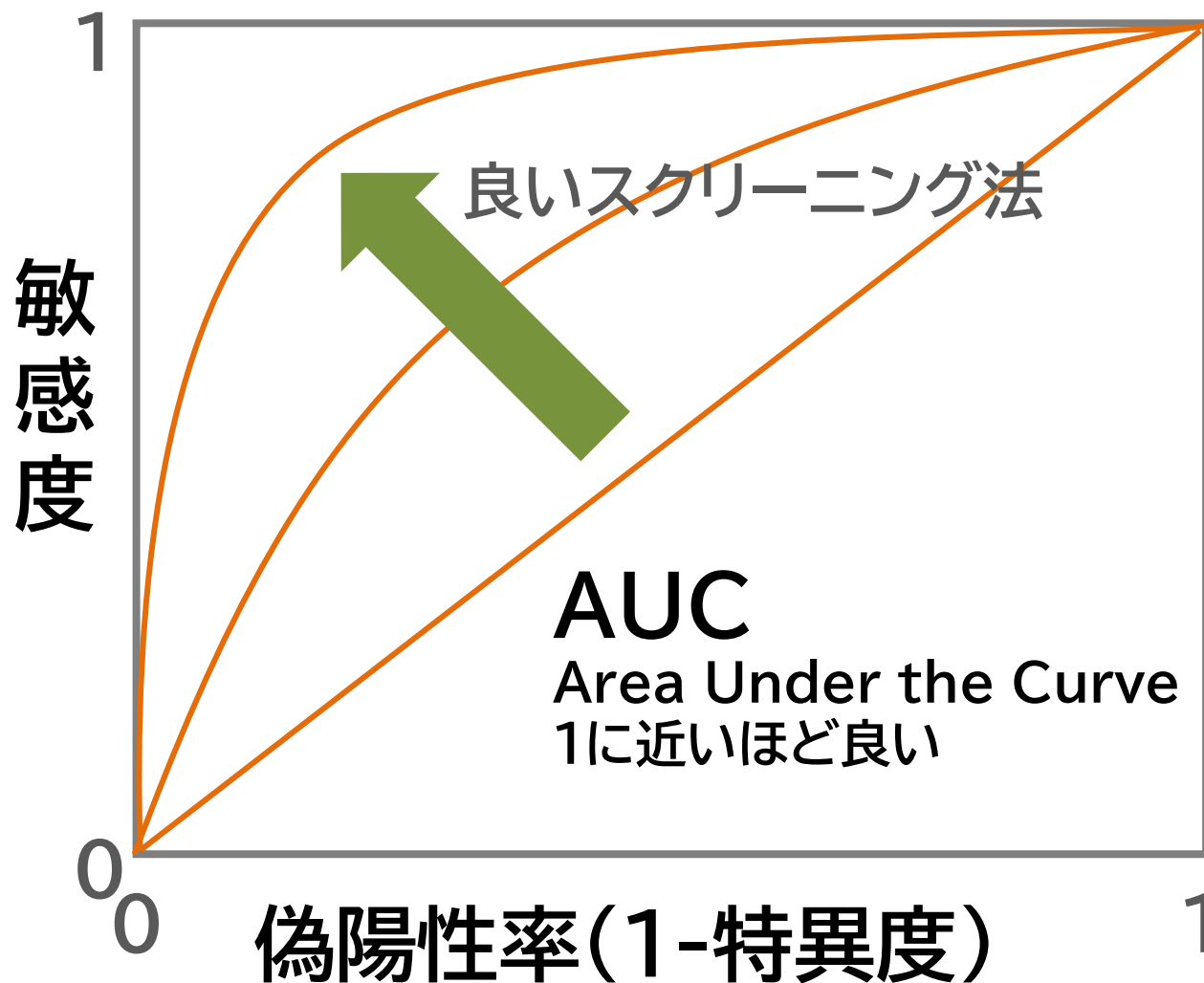
↑  
カットオフ値  
↓

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。

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

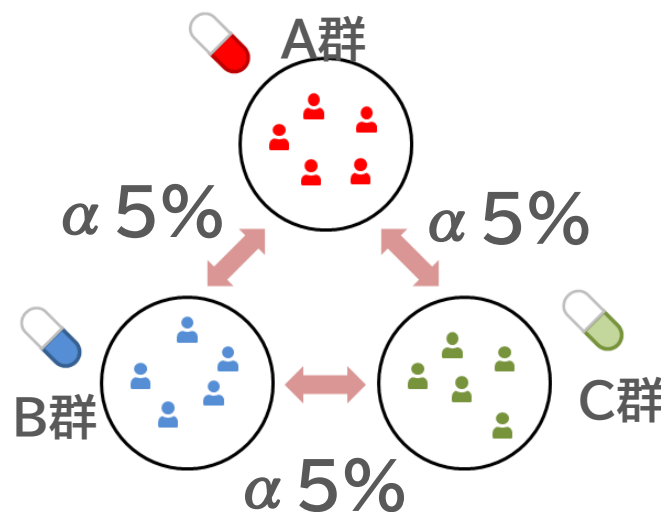
# 検定を繰り返すと、誤りが大きくなる

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 $\alpha$  5%で検定する。  
3つの薬に差がないことを主張したい。

1回の検定で差がないという結果になる確率は0.95。

3回の検定でどれもが差がない結果となる確率は、0.95の3乗で、0.86。

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



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  $\alpha$  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  $\alpha$  (お手つき率) が決して水準を超えないように、むりやり  $\alpha$  を引き下げるので、第二種の過誤の率(見逃し率)  $\beta$  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  $\alpha$  が上がるのを許容しながら、 $\beta$  を小さく抑える方法。

		帰無仮説が本当は		
		間違い(正しい姿)	正しい(誤った姿)	計
検定結果	棄却する(陽)	s	v $\alpha$ 偽陽性	R
	棄却しない(陰)	t $\beta$ 偽陰性	u	N-R
計		N-n	n	N

FDR  $q = v/R$  棄却したもののうち、偽陽性の率

これを、一定水準例えば0.05にする方法

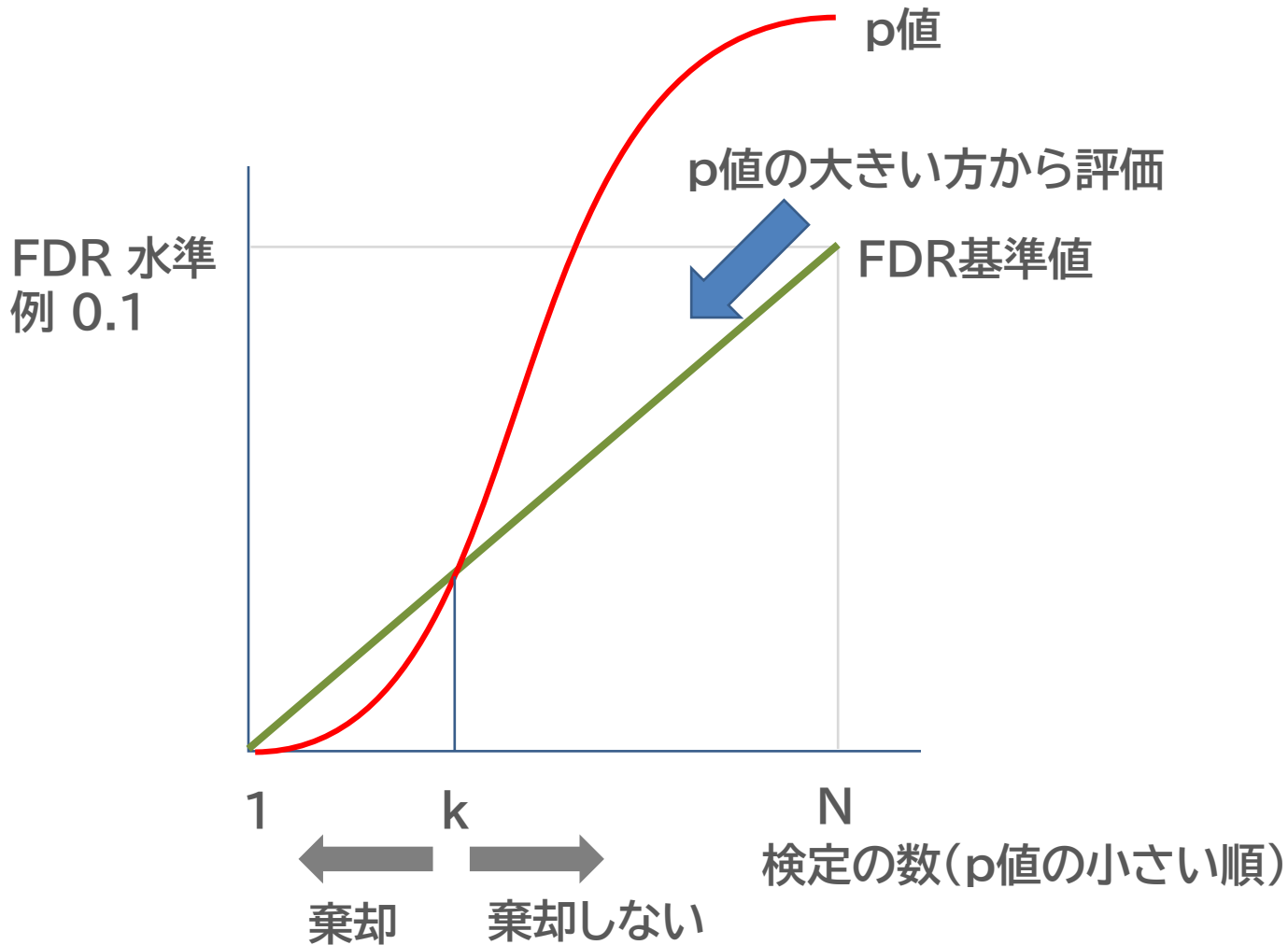


# FDR調整の手順

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

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

# FDRのイメージ



# FDR調整のイメージ

## その2

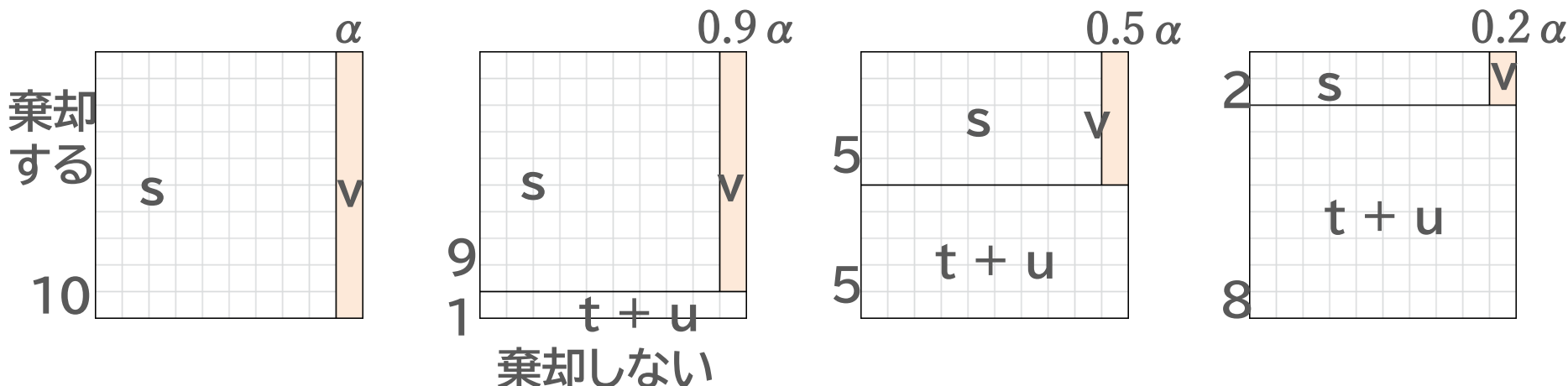
p値は、検定を繰り返したときに誤る確率でもあるので、複数回検定を繰り返したときに、最大のp値が有意水準 $\alpha$ を下回っているなら、すべての検定が十分有意であると判断してもよいものとする(甘いが)。

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、 $\alpha$ 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、 $\alpha$ 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、 $\alpha$ 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  $\alpha$  としてよく使われる0.05  
という数字に、特に深い意味はない

起こりにくい確率のひとつの基準として  
使われているだけ

# アメリカ統計学会の声明

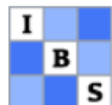
Wasserstein and Lazar (2016) The American statistician 70: 129-133 Editorial  
Wasserstein et al (2019) The American Statistician 73 (S1): 1-19 Editorial

- p値が特定の値以下だったことで「統計的に有意であった」と言うてはいけない
- 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](#) 坂巻頼太郎(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](#) 井桁正亮(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に、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



その他のほかの検定

# 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  $n_1$ , 不変標本分散  $v^2_1$

2群目: 標本数  $n_2$ , 不変標本分散  $v^2_2$

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

自由度:  $n_1 - 1, n_2 - 1$  ※分子と分母に対応させて、二つ与える

帰無仮説: 2群の分散は等しい

F分布を扱うExcel関数: F.DIST, F.DIST.RTなど

# 例)身長データの場合

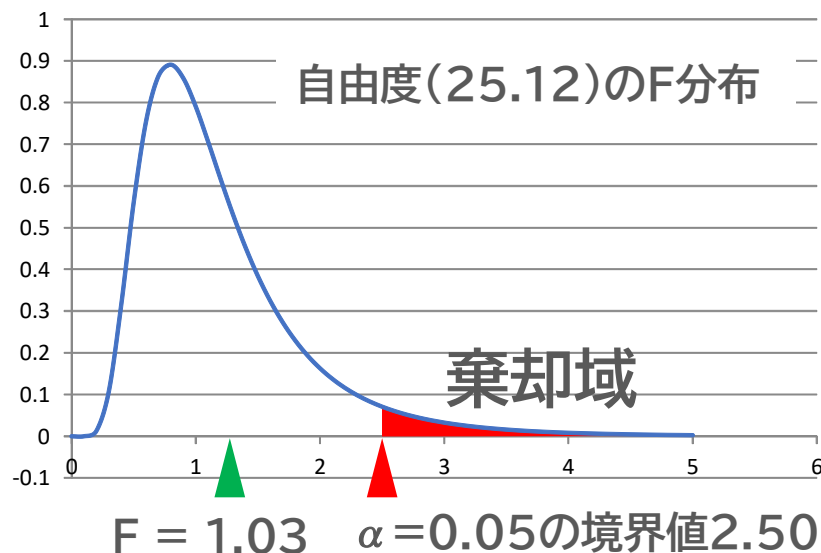
女性:  $n_1 = 26$ ,  $v^2_1 = 23.63$

男性:  $n_2 = 13$ ,  $v^2_2 = 23.02$

有意水準: 0.05とする

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

自由度(25, 12)のF分布から、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。

# 分散分析

Analysis of Variance

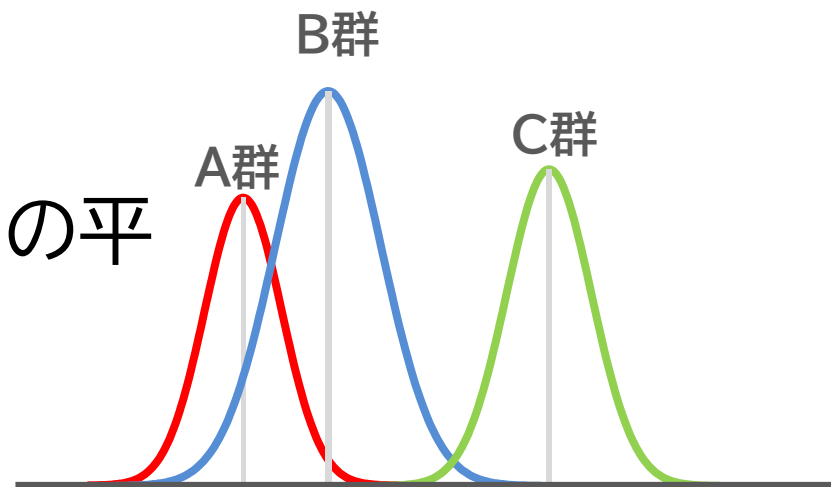
# ANOVA



- ✓ 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

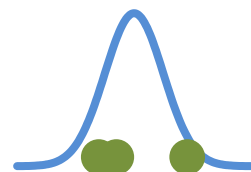
例) 1組、2組、3組で、テストの平均点に差があるか？



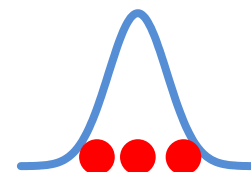
# F分布

## カイ二乗分布

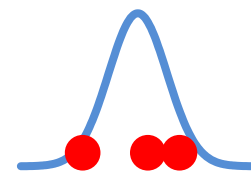
## 標準正規分布



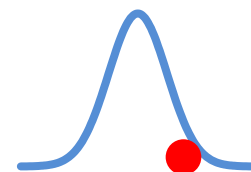
複数の変数を含む**2群**  
(分散の比を考える)



**複数の変数**  
(分散を考える)



**ひとつの変数**



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

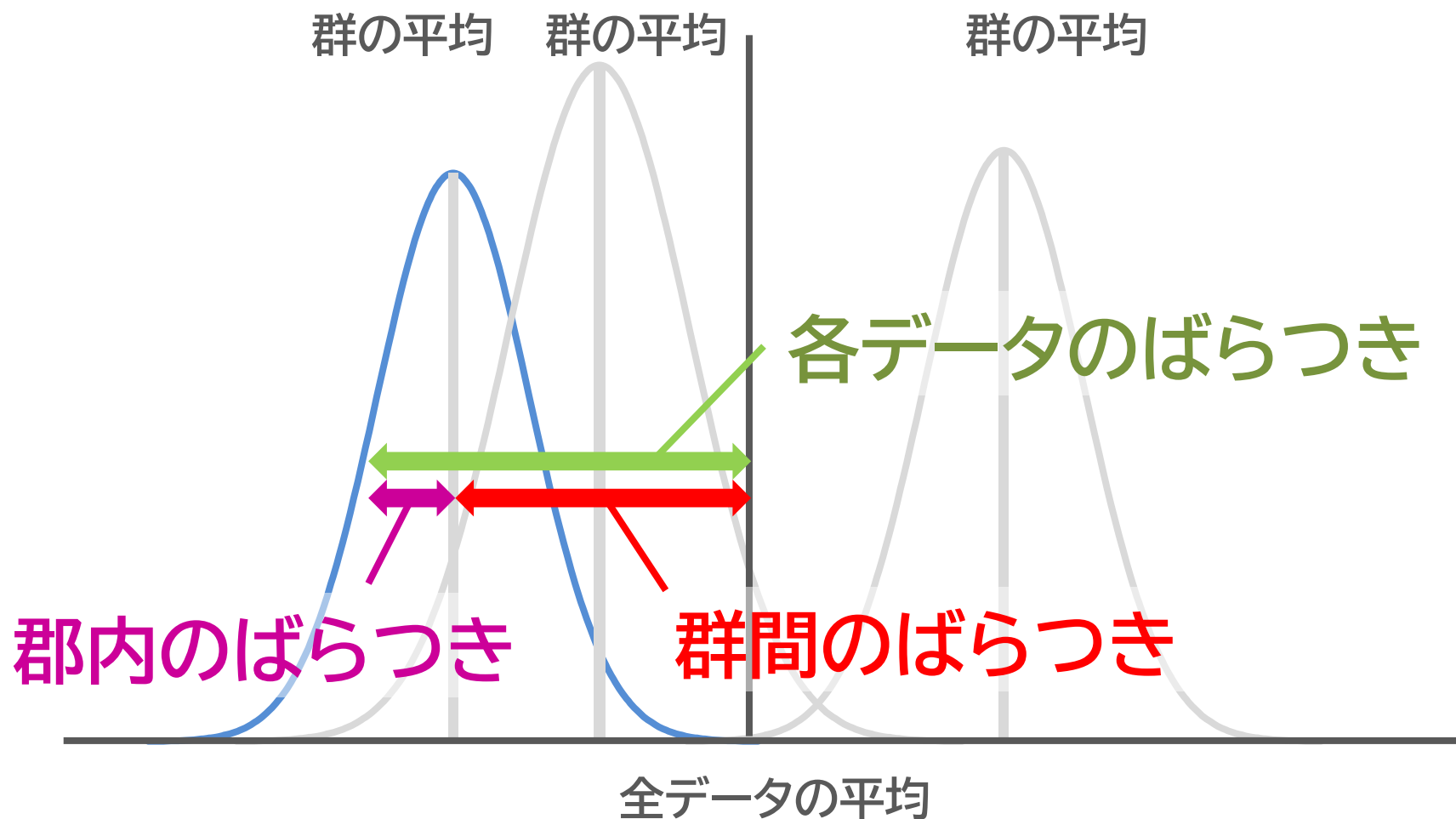


どれが異なっているかまではわからない！

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

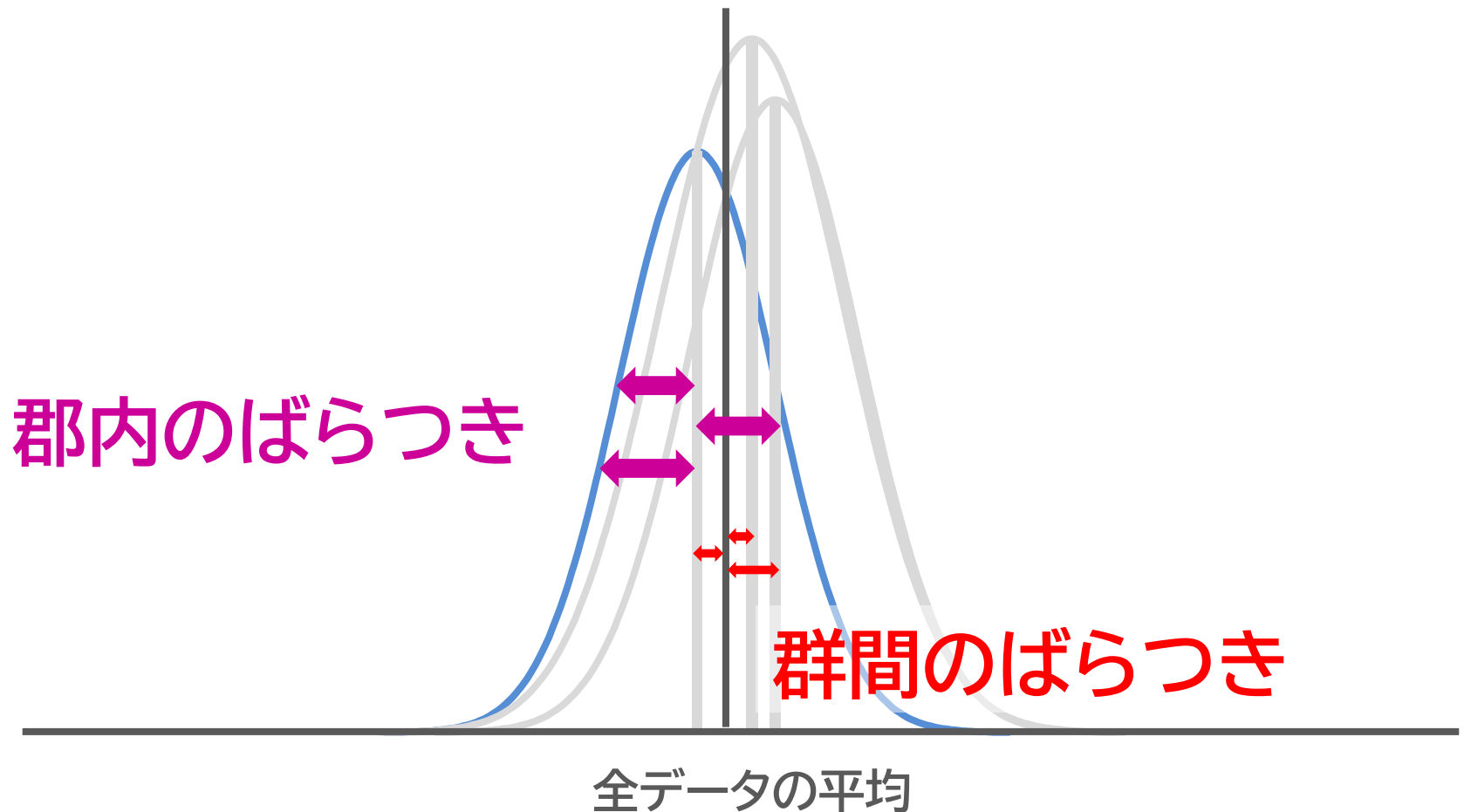
データのばらつきを、**群間**のばらつきと、**偶然により起こる群内**のばらつきに分けて考える



# 分散分析のイメージ

群の平均に差がなければ、

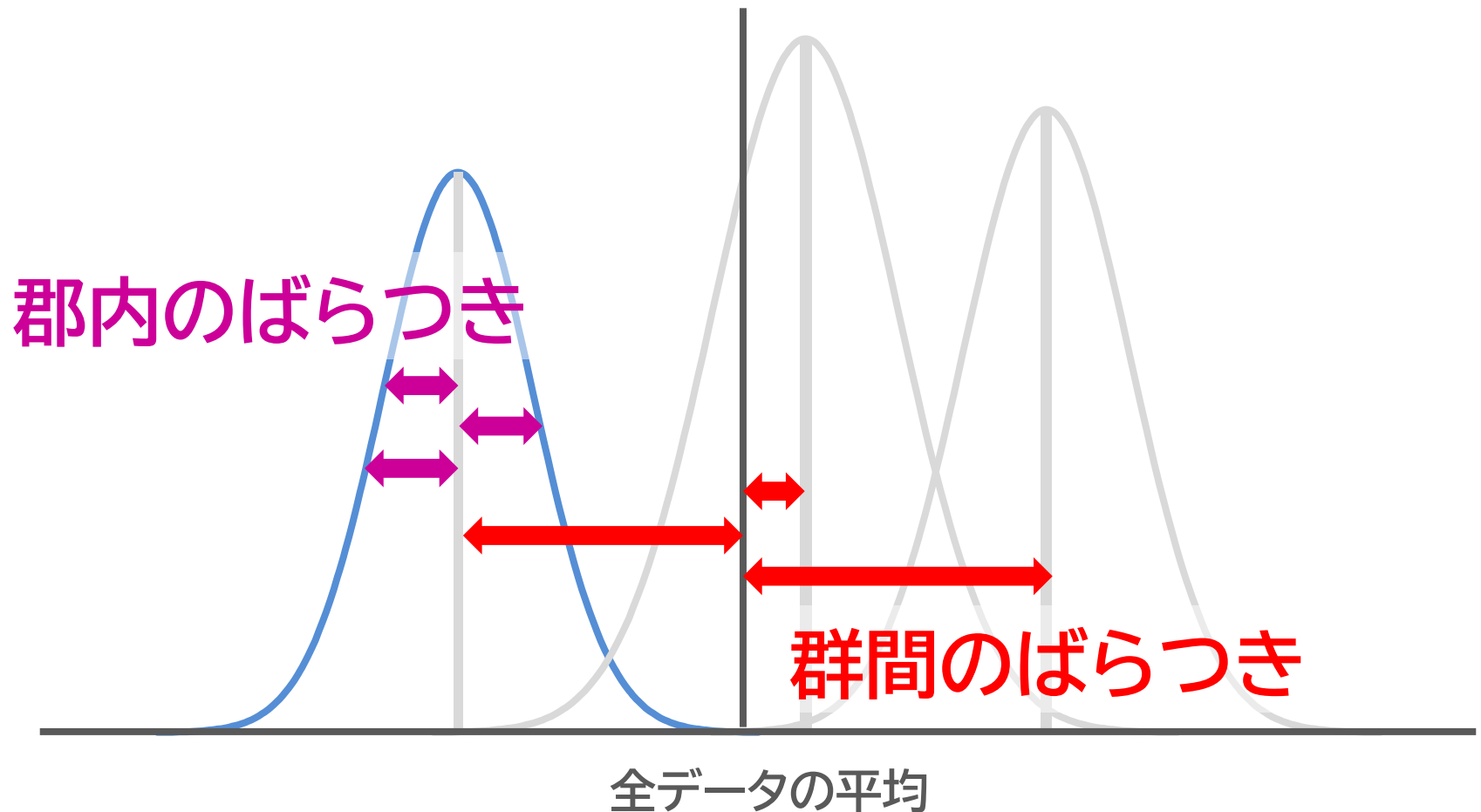
**群内**のばらつき > **群間**のばらつき



# 分散分析のイメージ

群の平均に差があるほど、

**群内**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 <sup>2</sup>	F値
群間 (因子)	S(群)	df(群) =群の数-1	V <sup>2</sup> (群) =S(群)/df(群)	V <sup>2</sup> (群)/V <sup>2</sup> (残差)
群内 (残差)	S(残差)	df(残差) =全データ数-群 の数	V <sup>2</sup> (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群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群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群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、 $\alpha$ のF境界値を計算する

⑮⑯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

⑰有意水準  $\alpha$  に対応するF境界値を、F.INV.RT関数を使って計算する

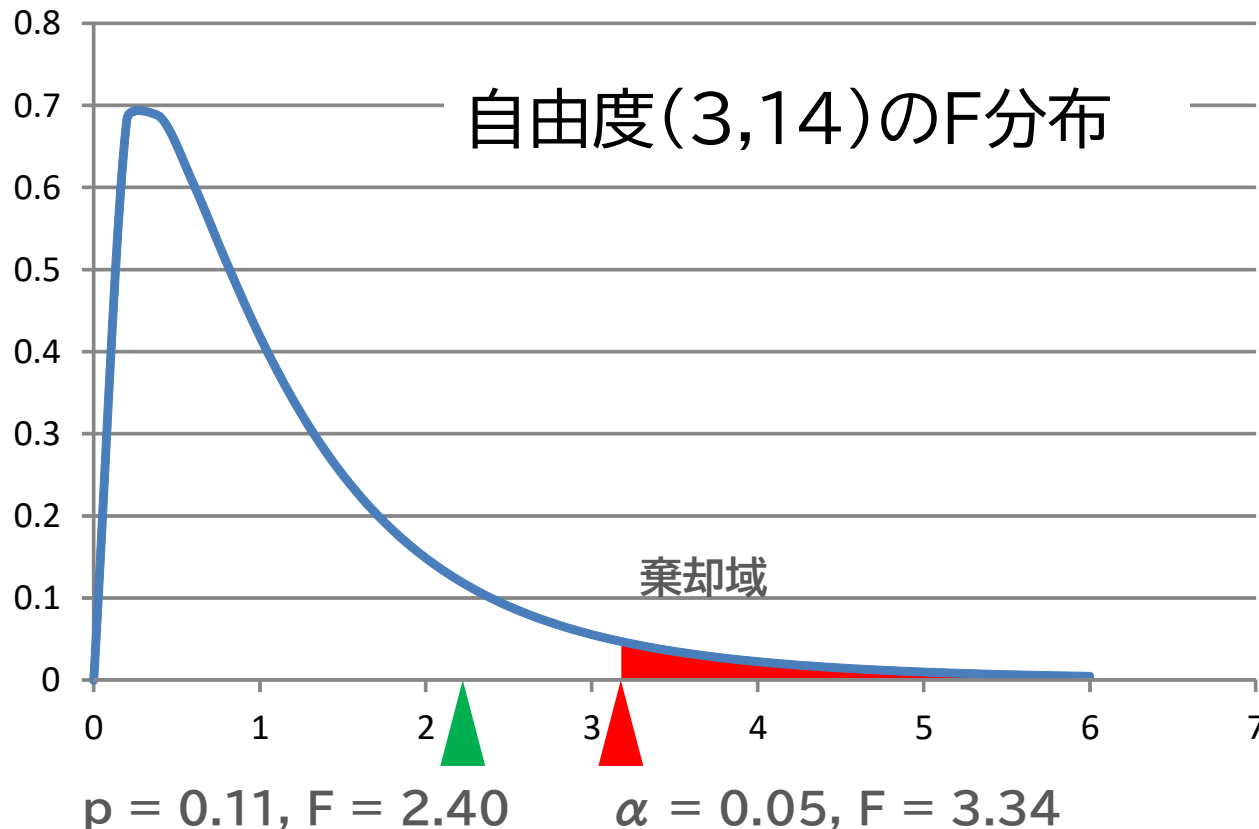
⑱F.DIST関数を用いて当該自由度のF分布を描く

p値の大きさ、 $\alpha$  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

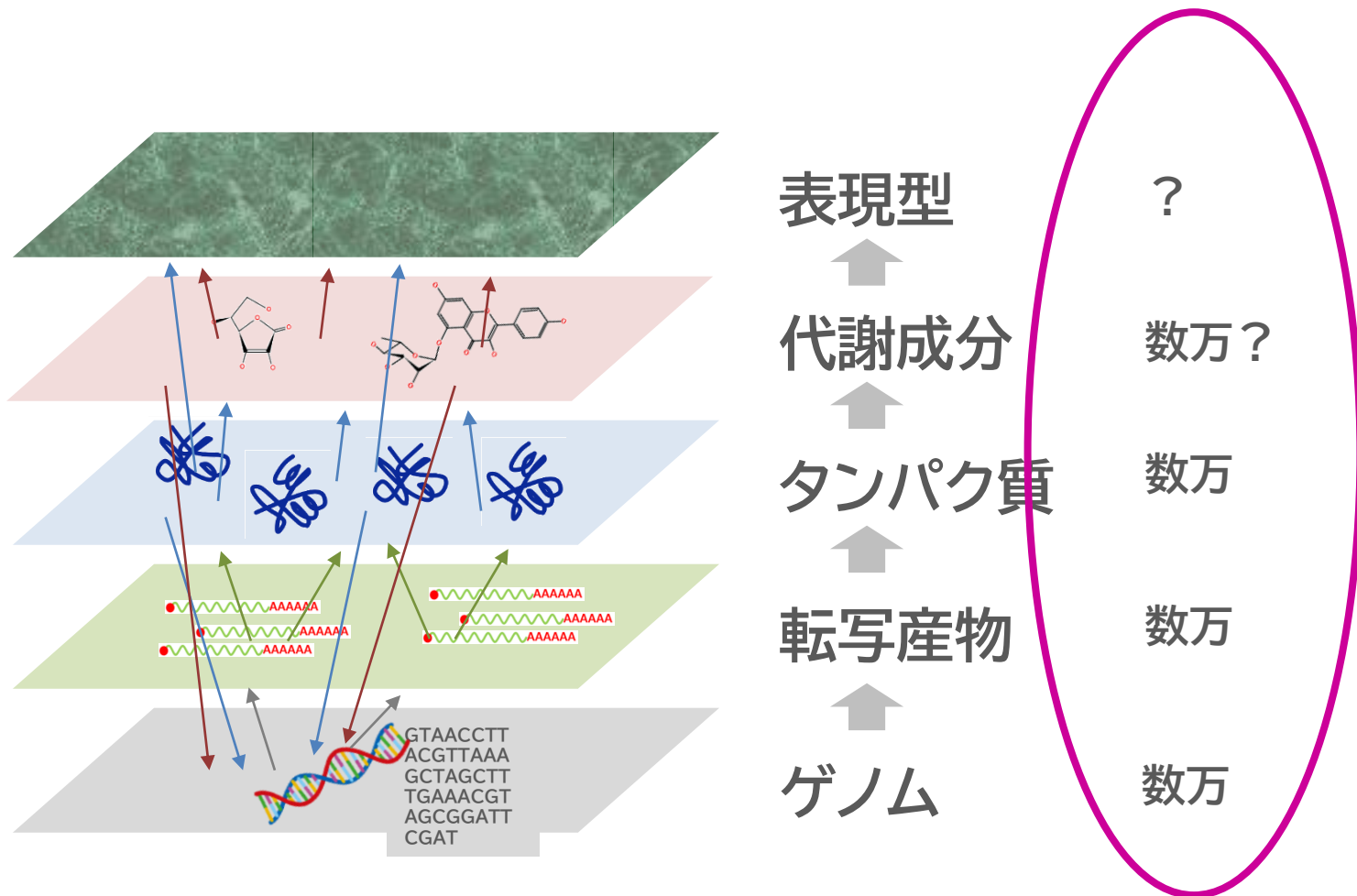
# 多变量解析

# 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



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

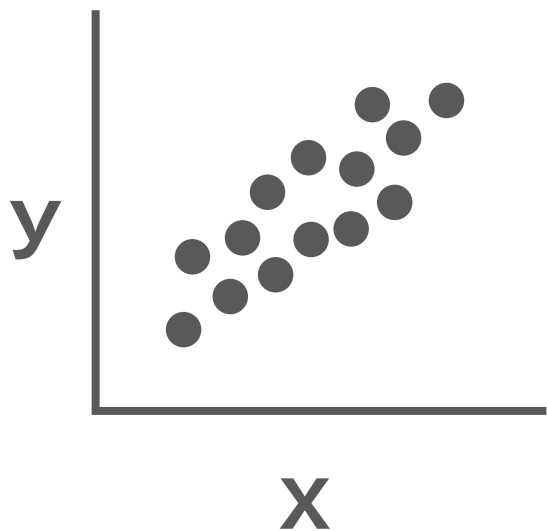
- 似ているものをグルーピングする  
クラスター解析
- データを要約する  
主成分分析
- 判別、分類、予測  
判別分析、PLS、PLS-DA、重回帰分析

など

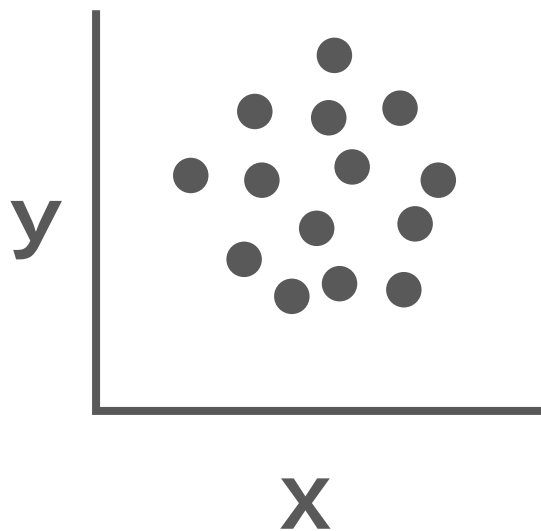
相関

# 散布図

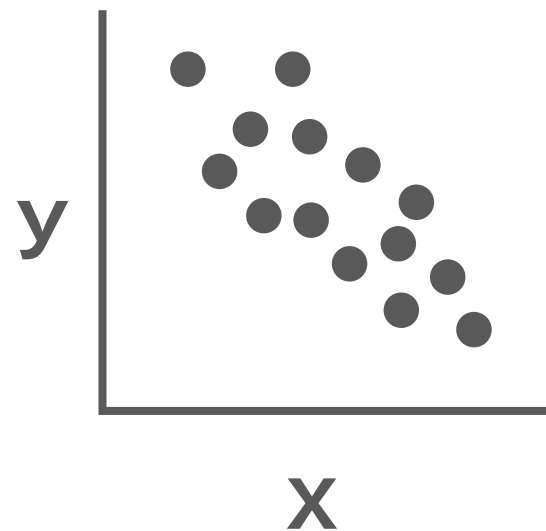
二つの変数の間の関係性を見える化する手法



正の相関がある



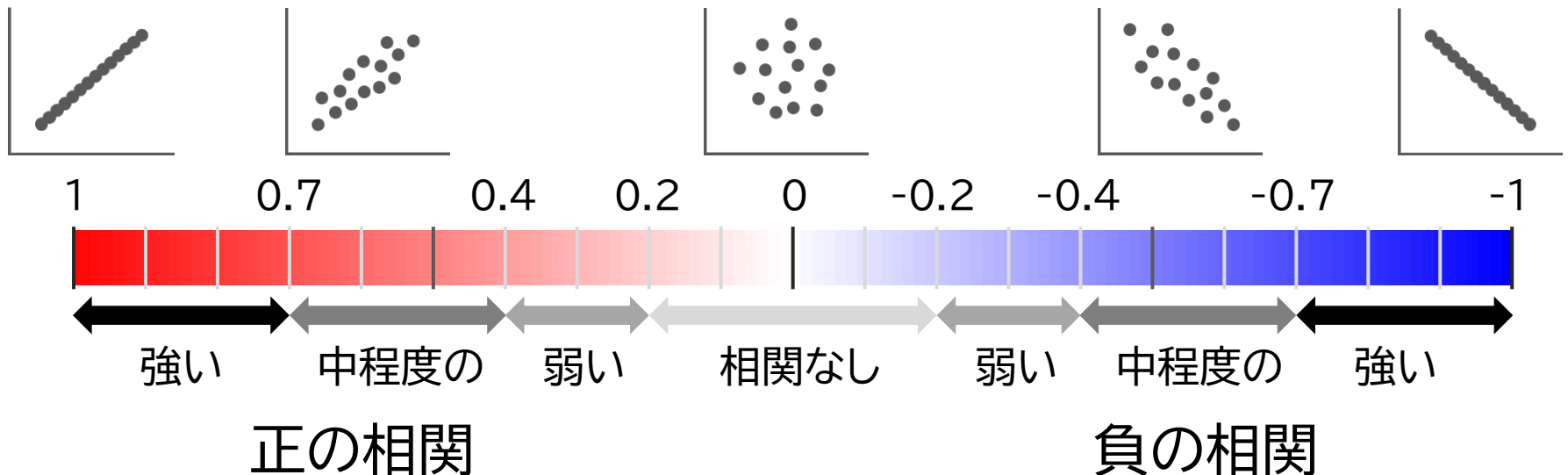
相関がない



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

# 相関係数を手で計算する

ピアソンの積率相関係数

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

$s_{xy}$ : xとyの共分散

$s_x$ : xの標準偏差

$s_y$ : yの標準偏差

$n$ : xとyのペアの数

# 無相関の検定

帰無仮説:

母集団の相関係数は0(無相関)である

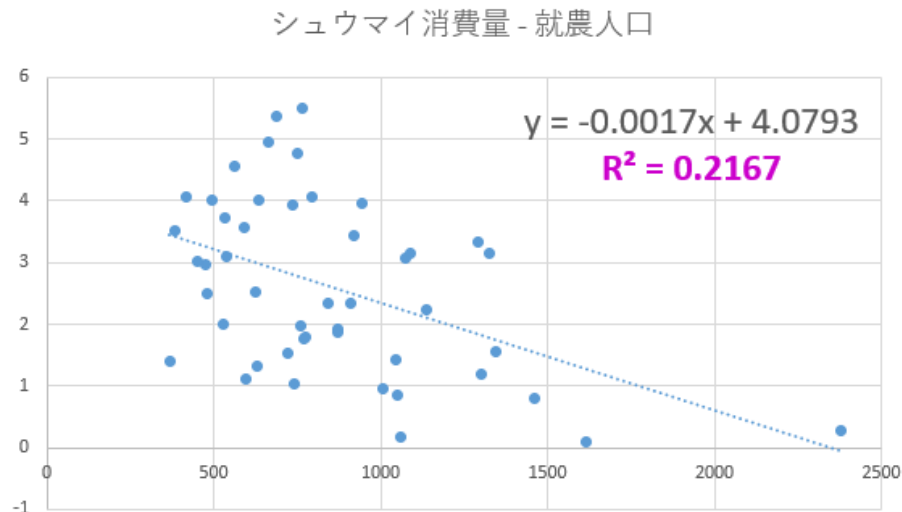
分布:  $t$ 分布

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

自由度:  $n-2$

※ $|r|$ は $r$ の絶対値  
エクセルではABS関数  
で計算できる

# 注意点



## 回帰曲線の $R^2$ 値は、相関係数ではない！

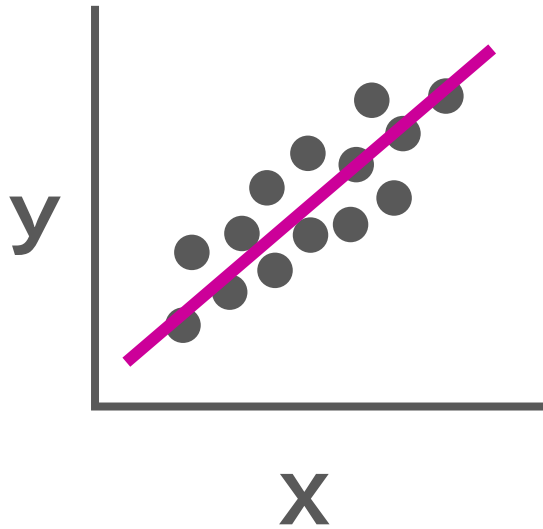
$R^2$ 値は、回帰曲線への当てはまり度を示すもので、「決定係数」と呼ばれます。

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 $-1 \sim 1$ の値を取るのに対し、 $R^2$ 値は $0 \sim 1$ 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

# その他の相関係数

## ● 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
A	1	2
B	2	3
C	3	1
D	4	4
E	5	5
F	6	6
G	7	7
H	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

# 相関と因果

## 相関関係：

二つの事柄に関連性がある

## 因果関係：

二つの事柄が、原因と結果の関係である

# 疑似相關

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

tylervigen.com

[about](#) | [twitter](#) | [email](#) | [subscribe](#)

## Spurious correlations



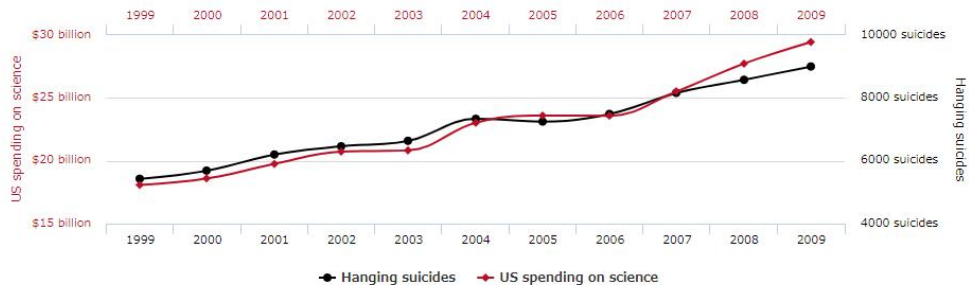
Now a ridiculous book!

- Spurious charts
- Fascinating factoids
- Commentary in the footnotes

[Amazon](#) | [Barnes & Noble](#) | [Indie Bound](#)

### US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation

Correlation: 99.79% ( $r=0.99789126$ )

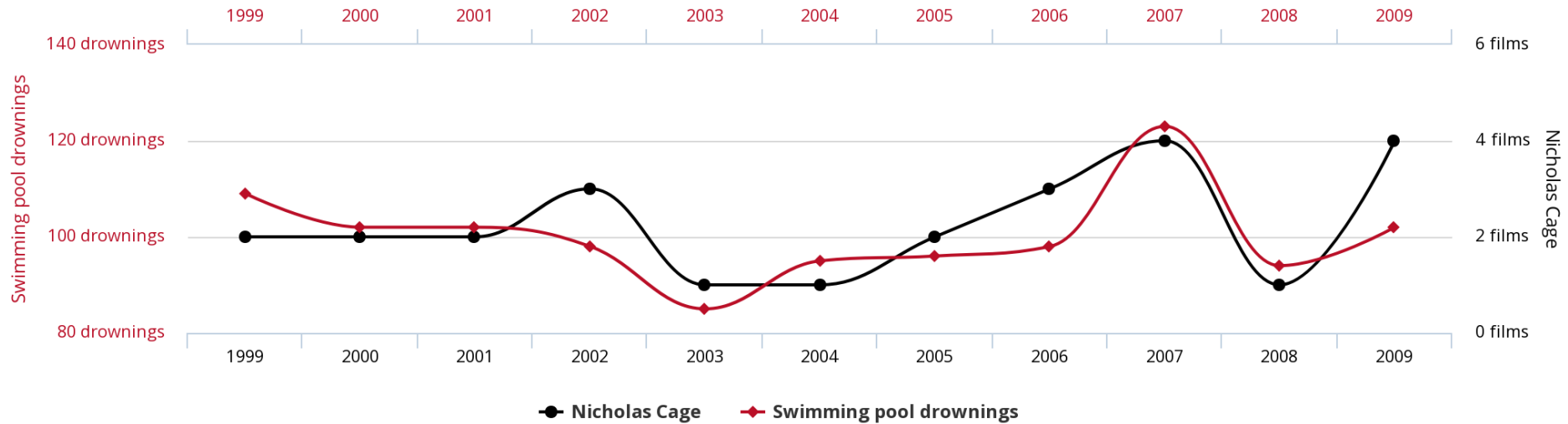


Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

tylervigen.com

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

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



中室牧子  
Makiko Nakamura  
津川友介  
Yusuke Tsugawa

Causal  
Inference  
in Economics  
*How to uncover the "causal" in everyday life*

データから  
真実を見抜く  
思考法

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

西内 啓

推薦  
します

『統計学が最強の学問である』著者

統計学と経済学の最新の知見を凝縮！

# 原因と結果の 経済学

中室牧子, 津川友介著、ダイアモンド社2017年



# 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	$X_1$	$X_{11}$	$X_{21}$	$X_{31}$		$X_{n1}$
	$X_2$	$X_{12}$	$X_{22}$	$X_{32}$		$X_{n2}$
	$X_3$	$X_{13}$	$X_{23}$	$X_{33}$		$X_{n3}$
	...					
	$X_m$	$X_{1m}$	$X_{2m}$	$X_{3m}$		$X_{nm}$

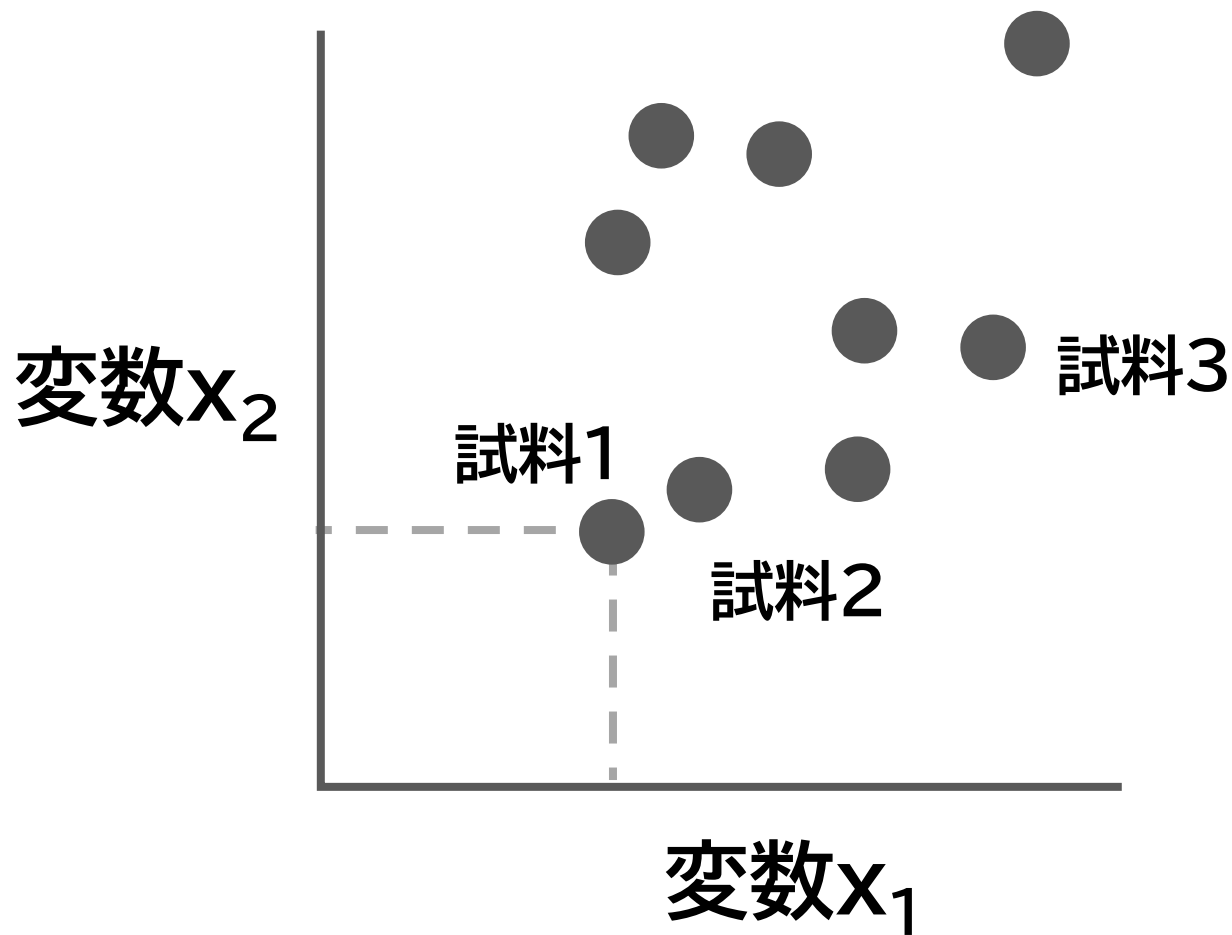
遺伝子など

説明変数, 観測変数

遺伝子発現量など

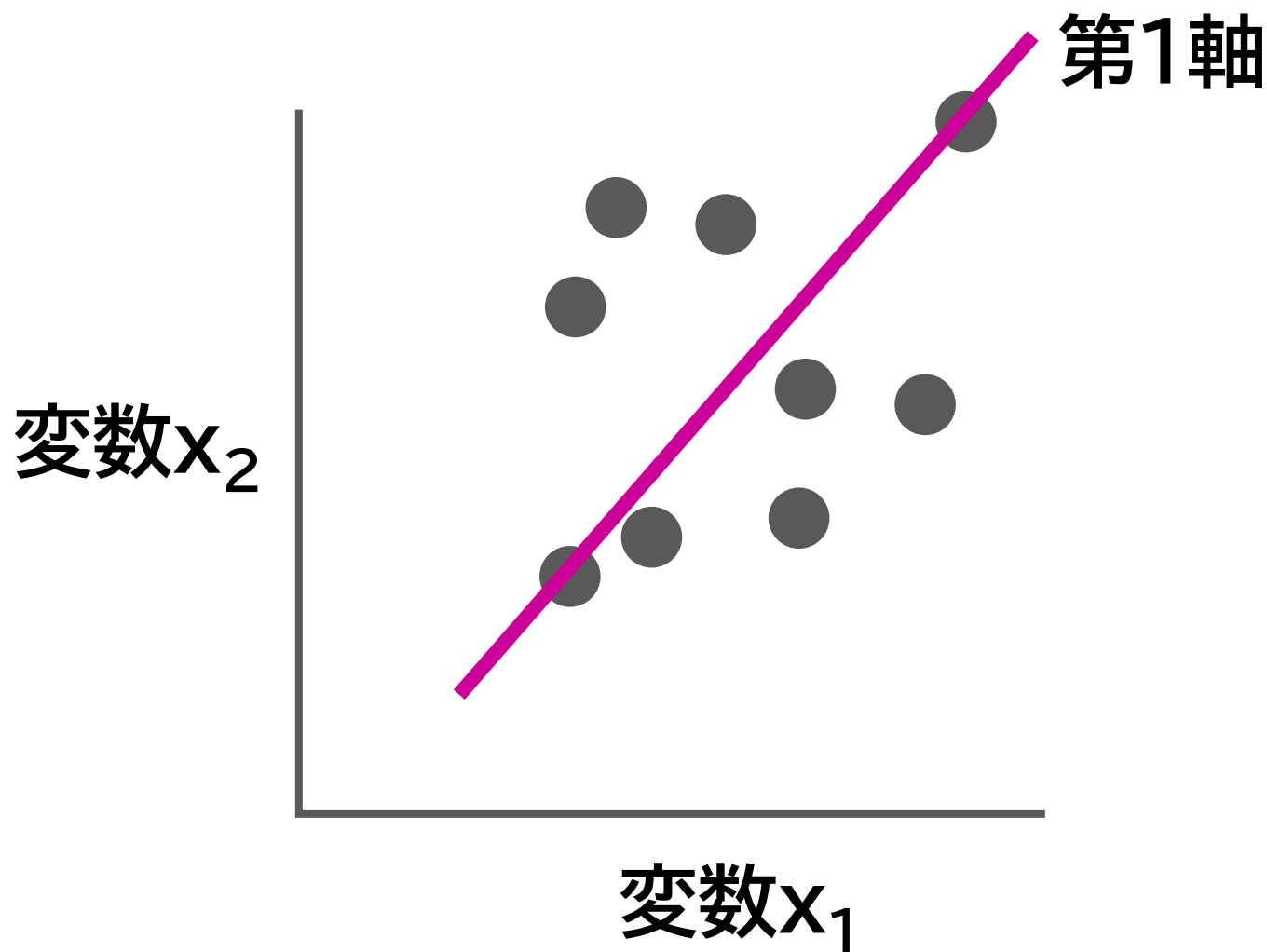
# 主成分分析のイメージ

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



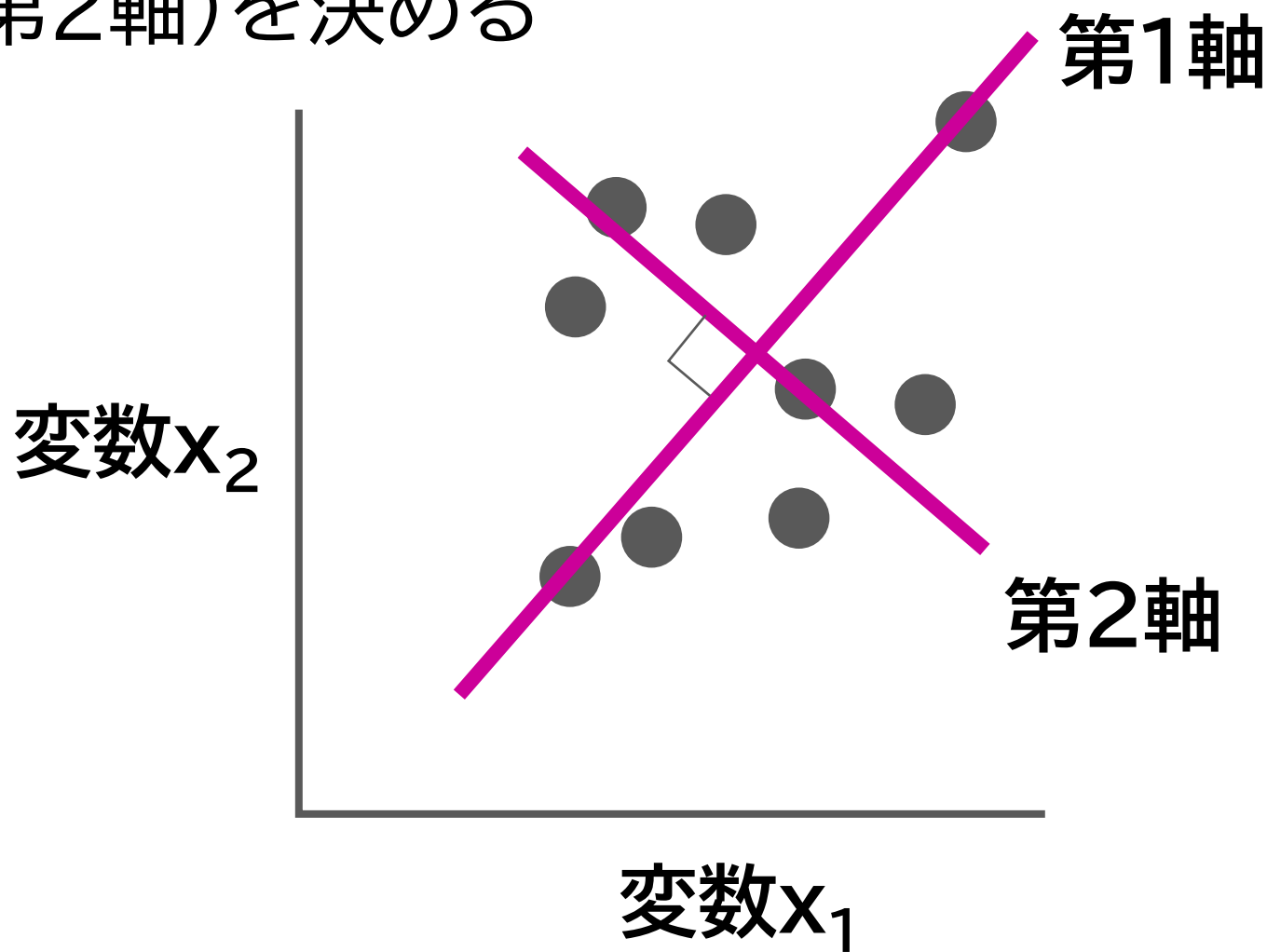
# 主成分分析のイメージ

②一番分散の大きい軸(第1軸)決める



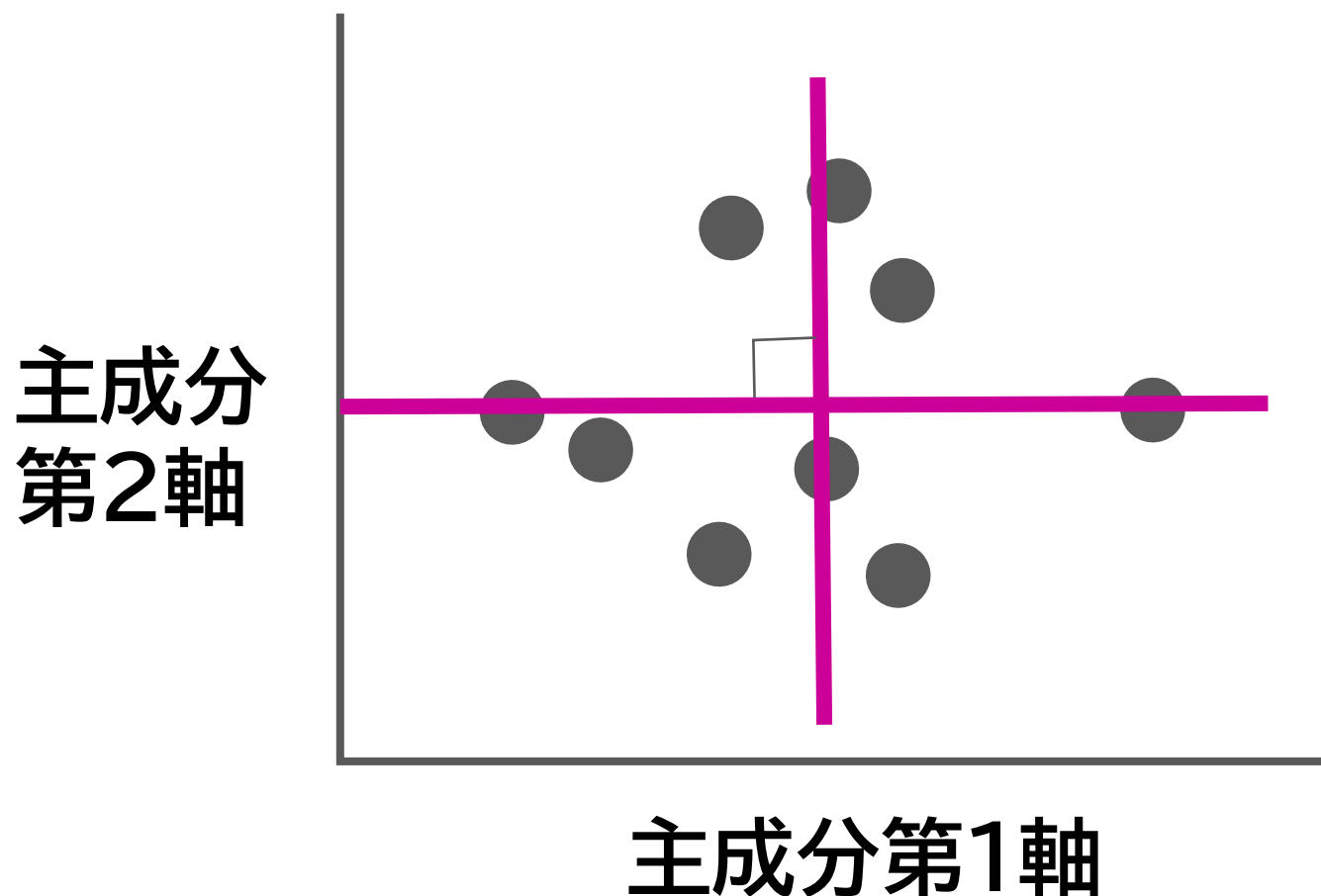
# 主成分分析のイメージ

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



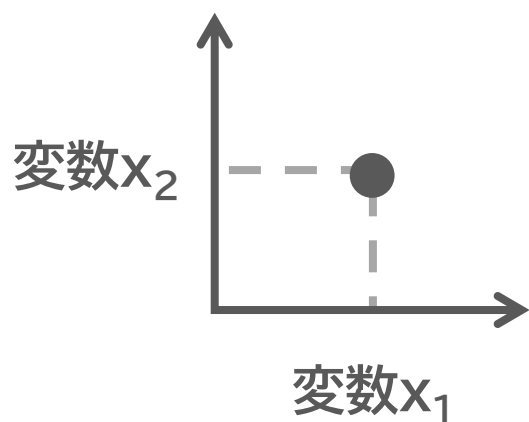
# 主成分分析のイメージ

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

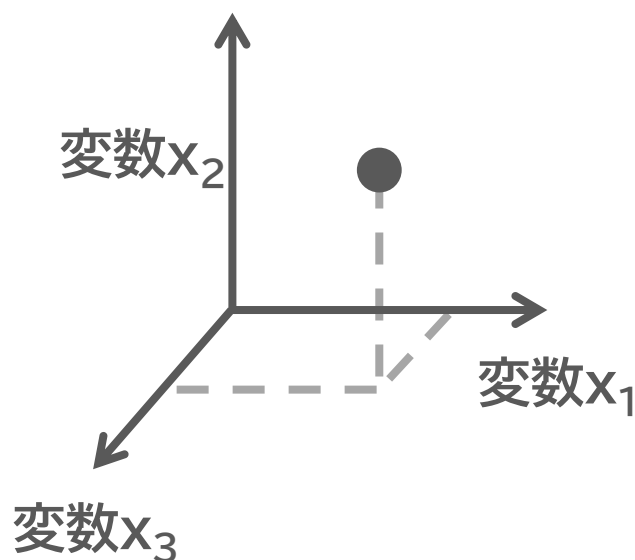


# 主成分分析のイメージ

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



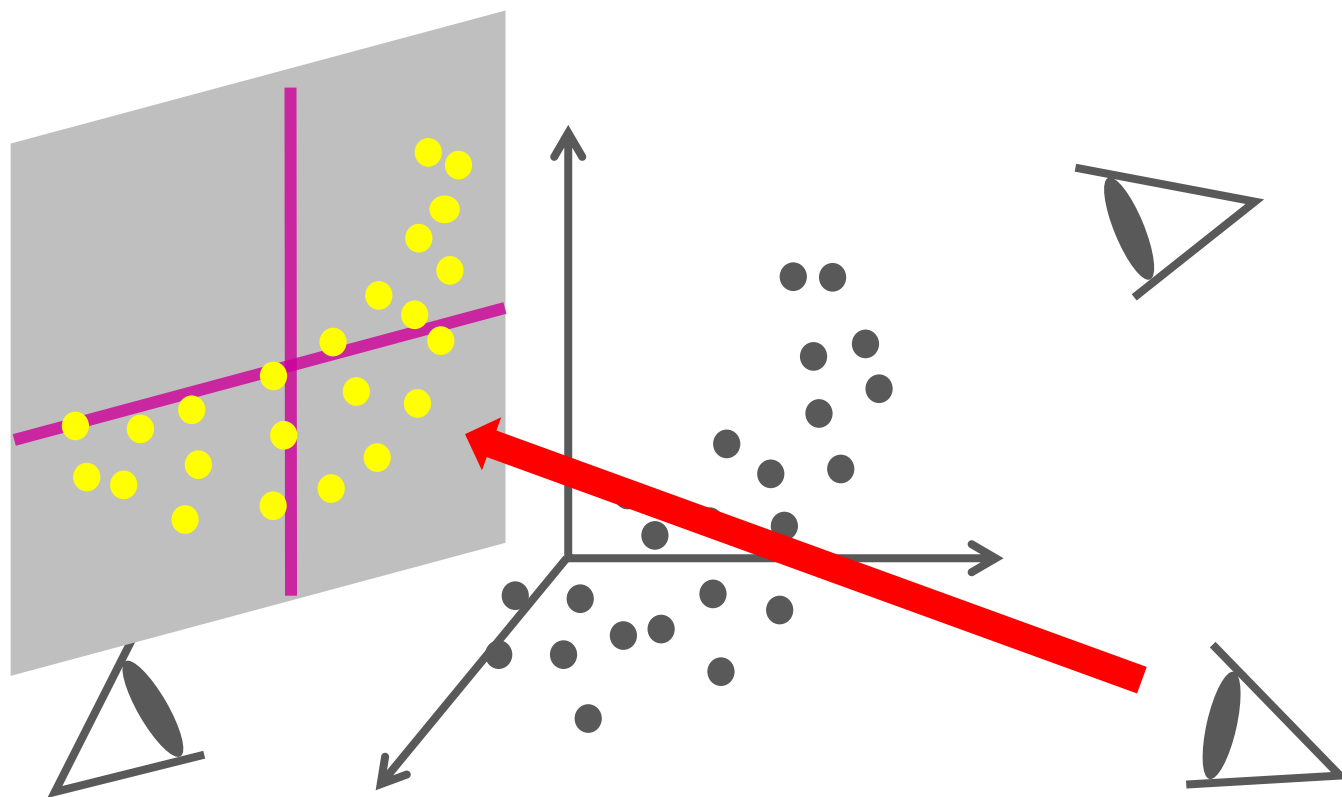
変数3個  
3次元



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

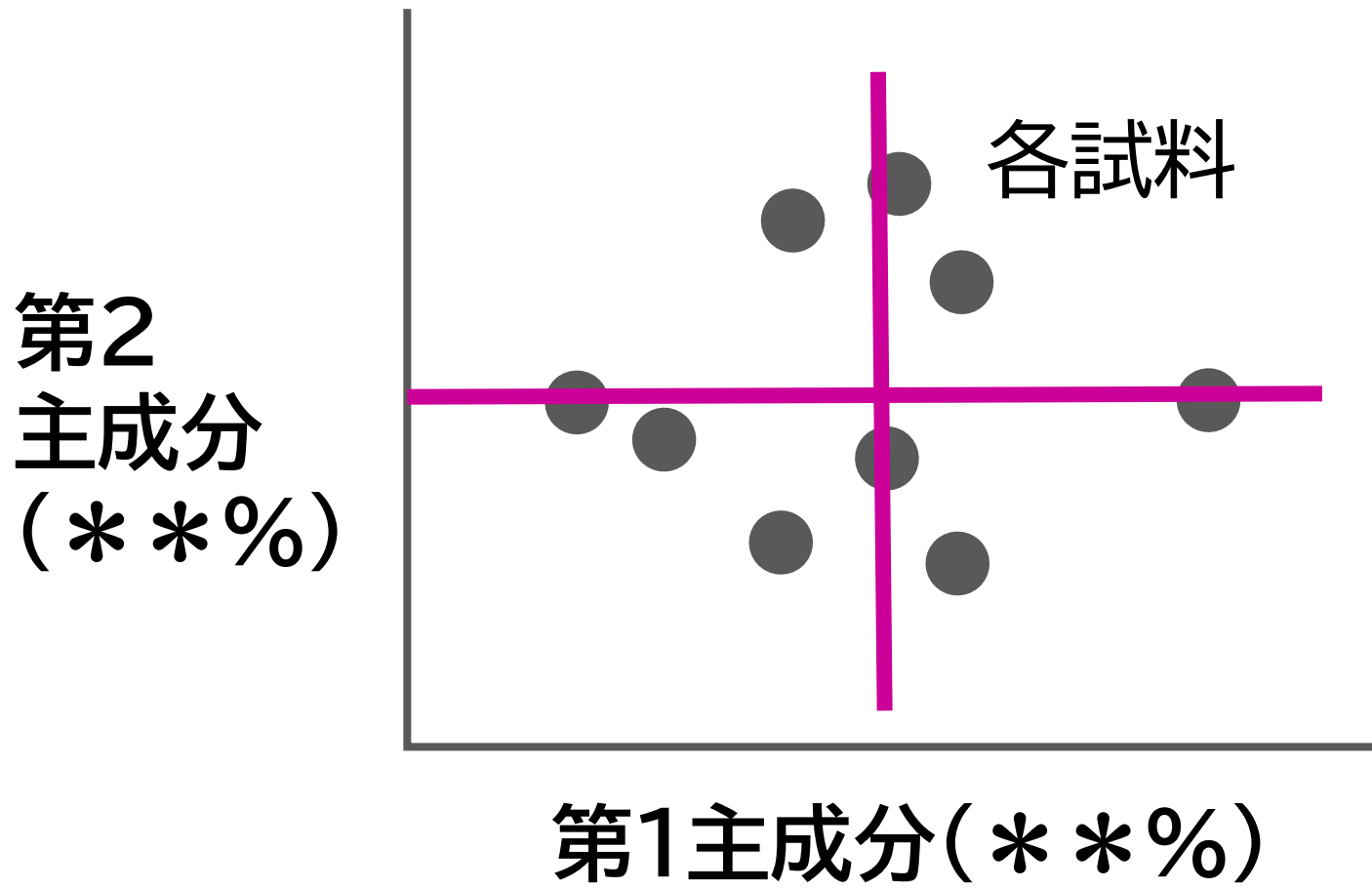




# スコアプロット

## 主成分軸に各試料を投影しなおした図

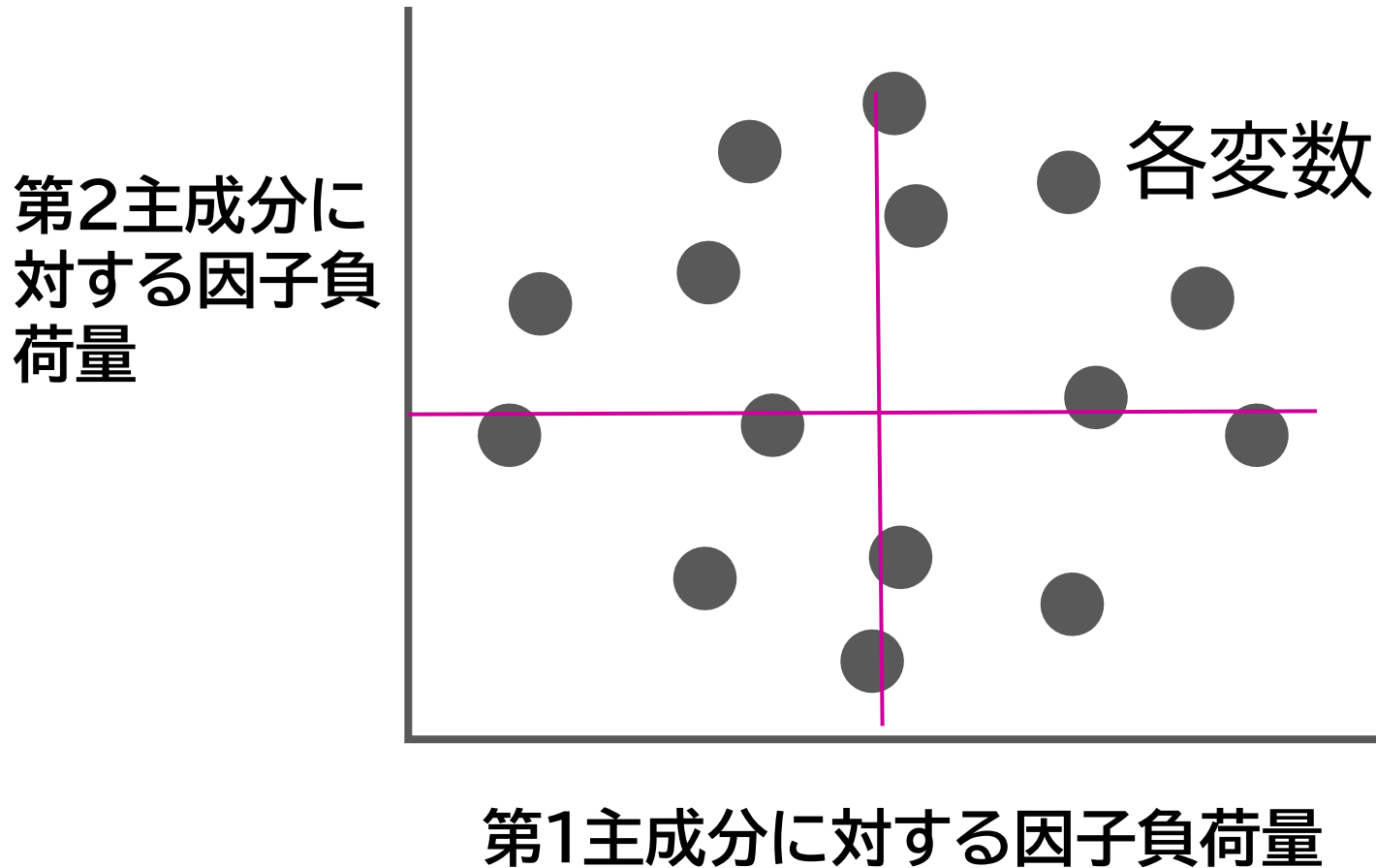
軸に示した%は**寄与率**と呼び、全体の分散のうち各主成分軸が説明する分散の比率を表す。第1主成分の寄与率が最も大きい。



# ローディングプロット

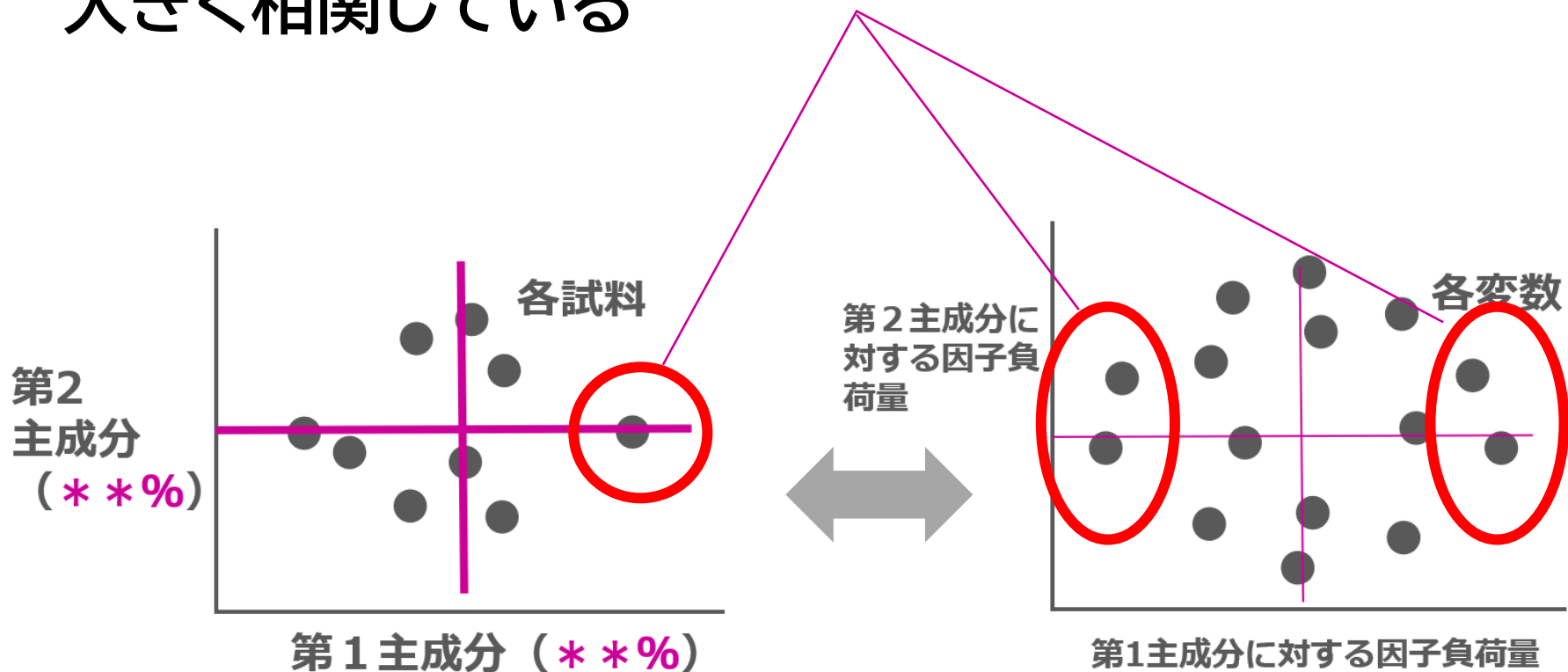
ローディングは、因子負荷量とも呼ばれ、各試料の主成分スコアと、変数の間の相関係数に相当する。

(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

# MetaboAnalystで解析



**MetaboAnalyst 6.0** - from raw spectra to biomarkers, patterns, functions and systems biology

- [Home](#)
- [Data Formats](#)
- [Tutorials](#)
- [User Forum](#)
- [MetaboAnalystR](#)
- [Publications](#)
- [Update History](#)
- [Databases](#)
- [APIs](#)
- [User Stats](#)
- [Data Policy](#)
- [About](#)
- [Contact](#)



[Manage Cookies](#)


### News & Updates

- Registration is now open for our **Omics Data Science Course** 1 week bootcamp (Aug. 5-9) or 12 weeks (Sep - Nov). Early bird discount ends by June 1, 2024; [NEW](#);
- Our paper [MetaboAnalystR 4.0: a unified LC-MS workflow for global metabolomics](#) is now available on **Nature Communications**; [NEW](#);
- Our paper [MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation](#) is now available on **Nucleic Acid Research**; [NEW](#);
- Check out our latest **Nature Protocols**: [Web-based multi-omics integration using the Analyst software suite](#); [NEW](#);
- Added support for testing the significance of different groups shown in PCA score plot based on PERMANOVA (06/25/2024); [NEW](#);
- Enhanced support for multivariate time series analysis based on user feedback (05/30/2024); [NEW](#);
- Enhanced support for interactive visualization of t-tests and volcano plots based on user feedback (05/23/2024); [NEW](#);
- Enhanced support for customizing colors and shapes in 2D score plots in Statistical Analysis modules (04/16/2024); [NEW](#);
- Fix the issue on name mapping for fumarate based on user feedback (04/10/2024);
- Updated the tutorials for all new and updated modules in MetaboAnalyst version 6.0 (03/12/2024);

[Read more .....](#)

[Click here to start](#)

### Overview



MetaboAnalyst is a web-based platform dedicated for comprehensive metabolomics data analysis, interpretation and integration with other omics data. Over the past decade, MetaboAnalyst has evolved from statistical and functional analysis for targeted metabolomics data, towards more streamlined analysis for both quantitative and untargeted metabolomics data. In addition to many feature enhancements, the version 6.0 contains three new modules - tandem MS spectral processing and **compound annotation**, dose response analysis for **chemical risk assessment**, and leveraging

サンプルデータ: 231004\_sampledata.xlsx

入力用加工データ: 231004\_sampledata\_mod.csv

解析方法: 解説資料\*

# 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)

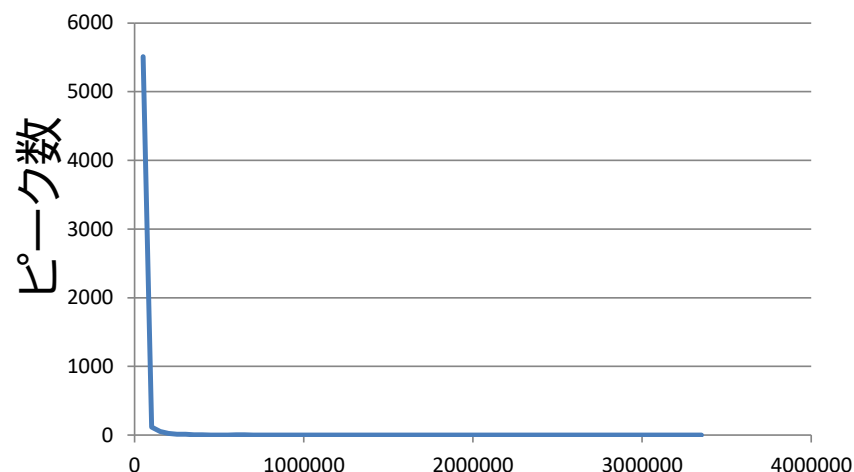
平均値補正、分散を1にする、それらの組み合わせなど。

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。

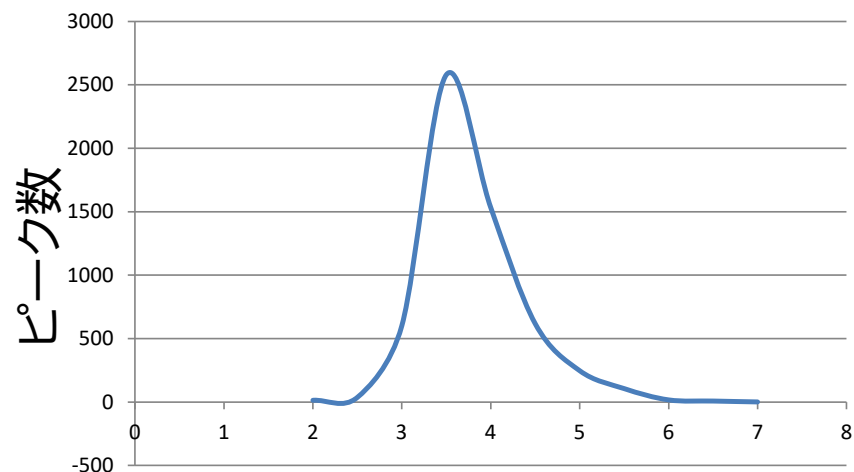
# 大葉(しそ)で検出された代謝物質

- 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

# MetaboAnalystデータのノーマライズ画面

## Normalization Overview:

The normalization procedures are grouped into three categories. You can use one or combine them to achieve better results.

- Sample normalization is for general-purpose adjustment for systematic differences among samples;
- Data transformation applies a mathematical transformation on individual values themselves. A simple mathematical approach is used to deal with negative values in log and square root. Please search OmicsForum using "normalization #metaboanalyst" to find more information.
- Data scaling adjusts each variable/feature by a scaling factor computed based on the dispersion of the variable.

**Sample normalization**

- ☒ None
- ☐ Sample-specific normalization (i.e. weight, volume) [Specify](#)
- ☐ Normalization by sum
- ☐ Normalization by median
- ☐ Normalization by a reference sample (PQN) [Specify](#)
- ☐ Normalization by a pooled sample from group (group PQN) [Specify](#)
- ☐ Normalization by reference feature [Specify](#)
- ☐ Quantile normalization (suggested only for > 1000 features)

**Data transformation**

- ☐ None
- ☒ Log transformation (base 10)
- ☐ Log transformation (base 2)
- ☐ Square root transformation (square root of data values)
- ☐ Cube root transformation (cube root of data values)
- ☐ Variance stabilizing normalization (data-adaptive transformation)

**Data scaling**

- ☐ None
- ☐ Mean centering (mean-centered only)
- ☒ Auto scaling (mean-centered and divided by the standard deviation of each variable)
- ☐ Pareto scaling (mean-centered and divided by the square root of the standard deviation of each variable)
- ☐ Range scaling (mean-centered and divided by the range of each variable)

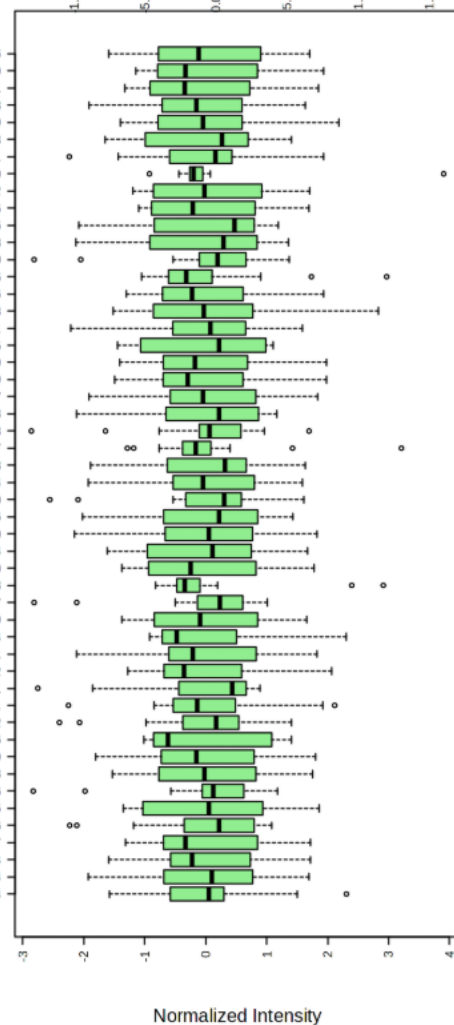
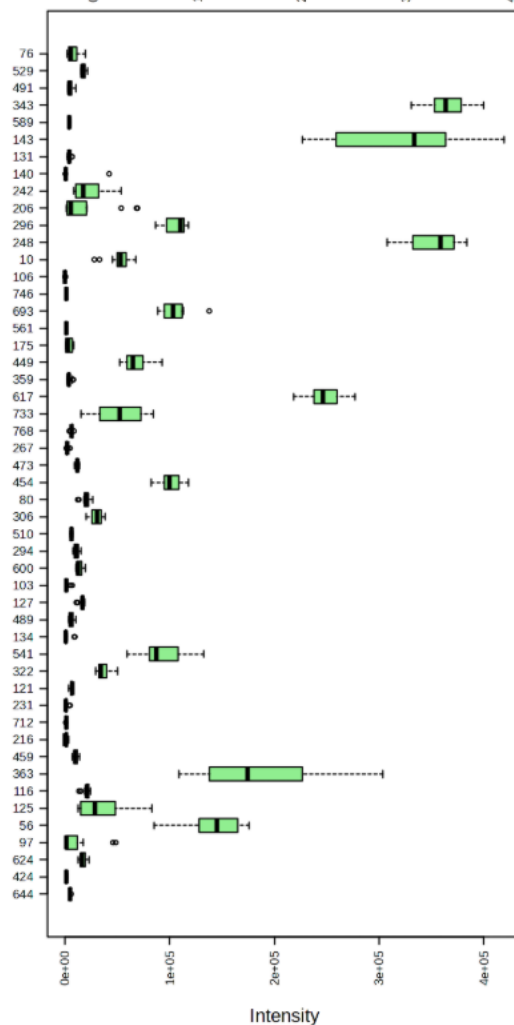
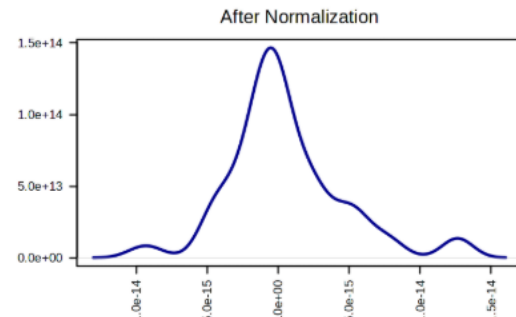
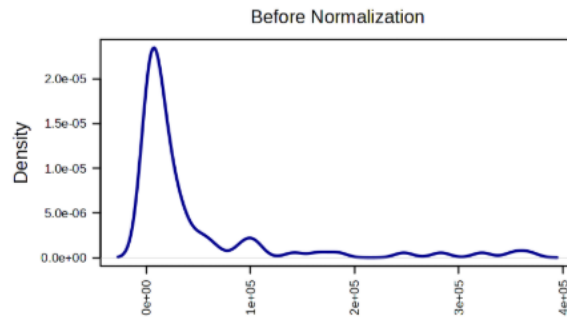
Normalize

View Result

Proceed



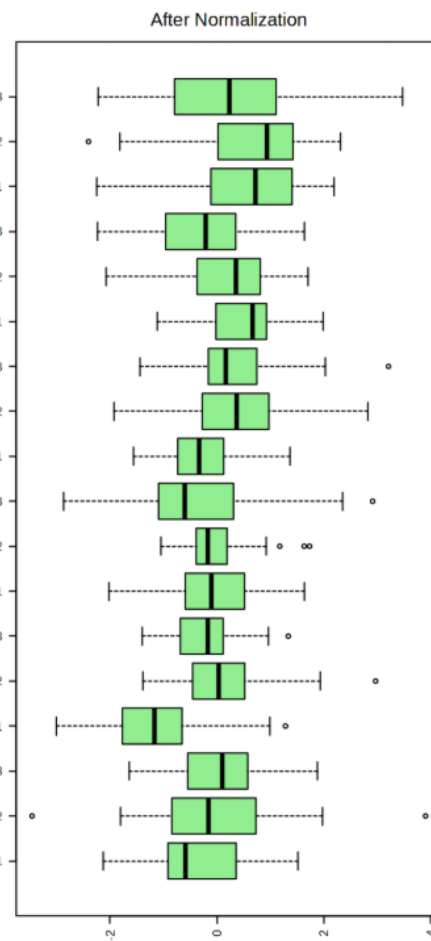
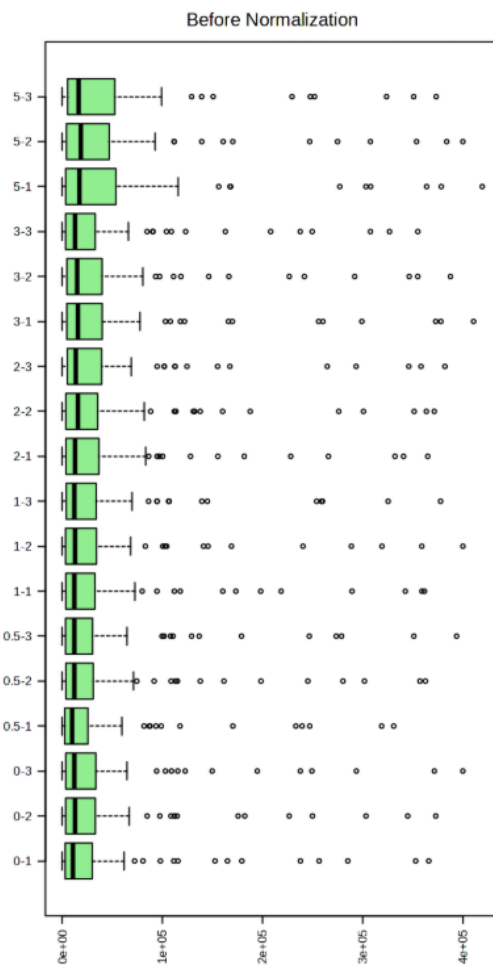
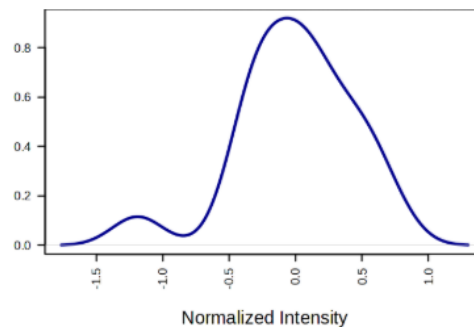
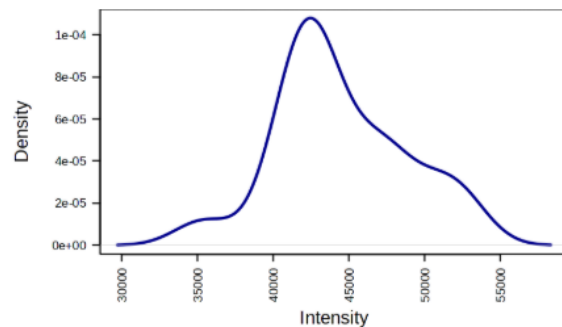
# 説明変数の ノーマライズ



ノーマライズ前

ノーマライズ後

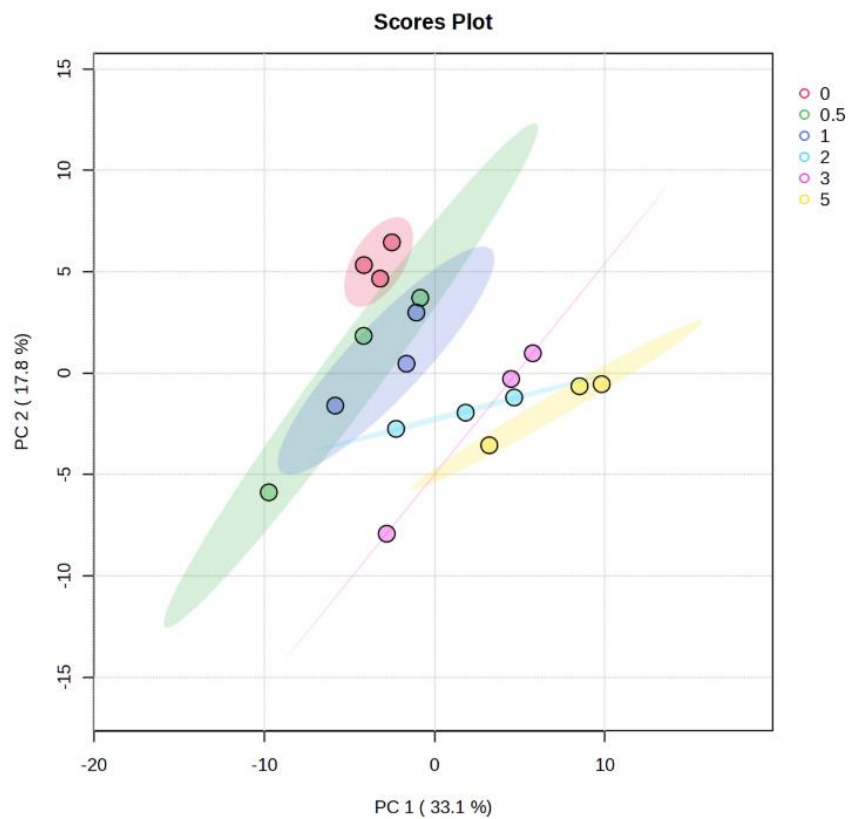
# サンプルの ノーマライズ



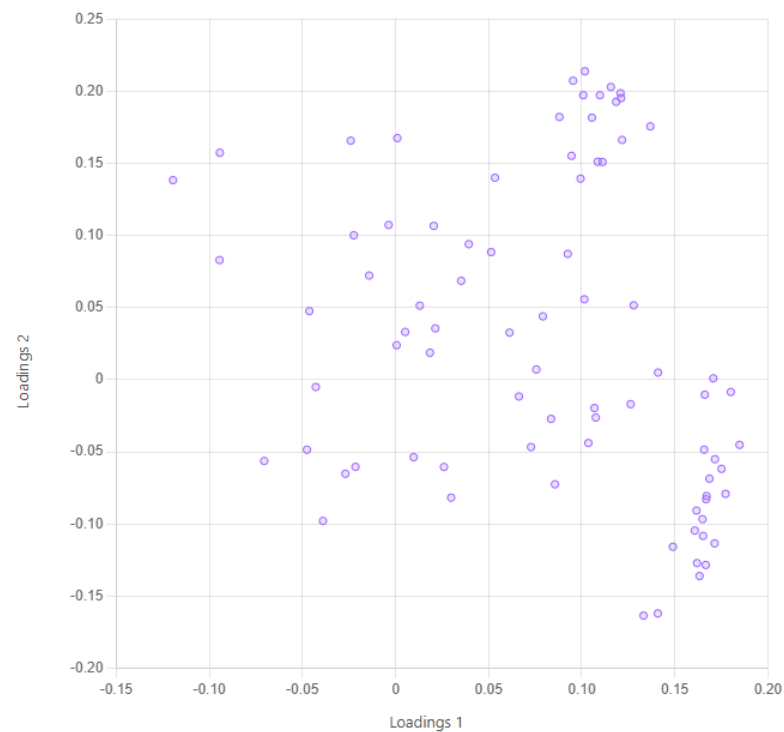
ノーマライズ前

ノーマライズ後

# PCA結果

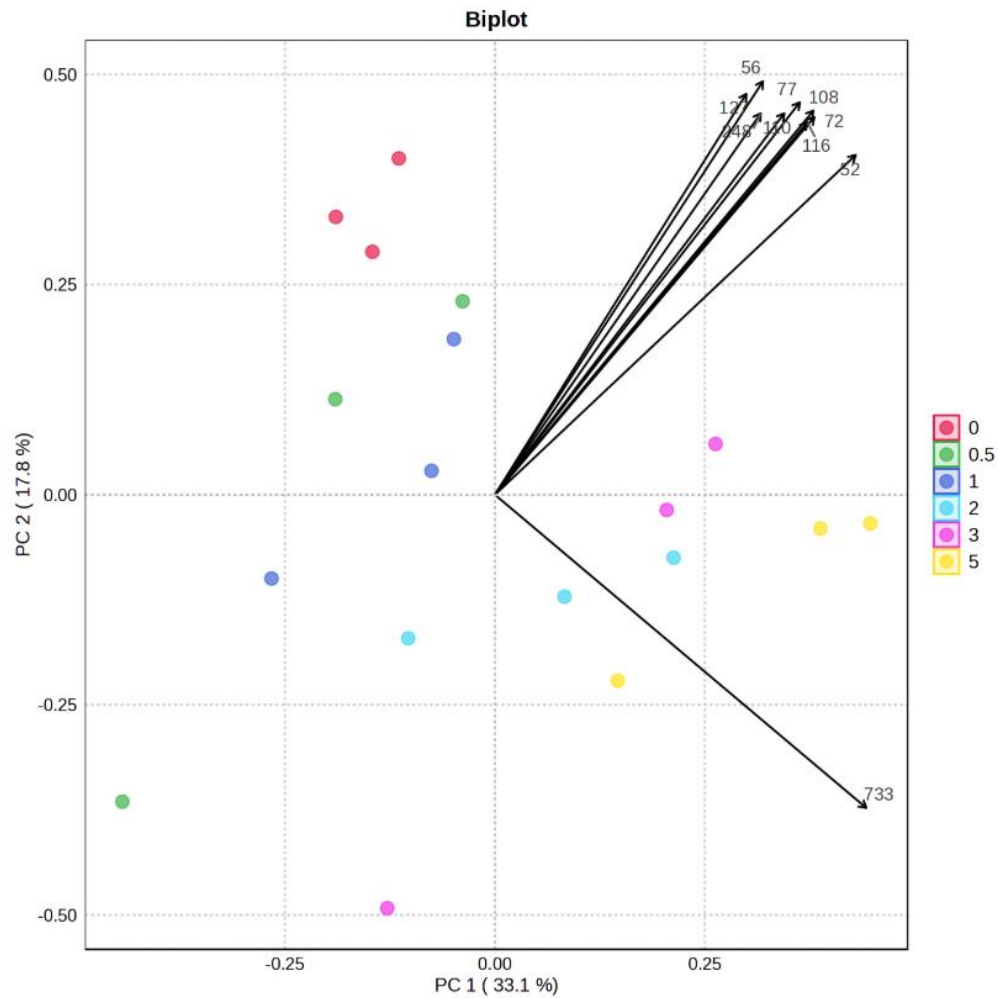


スコアプロット



ローディングプロット

# PCA結果



バイプロット

# 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

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

# 回帰分析の一種

回帰分析とは

あるデータから、別のデータの結果を予測したり説明したりする手法

**PLS回帰:**

PLSで作られたモデルで、他のデータの値を予測する方法  
(量的な予測)

**PLS判別分析(PLS-DA):**

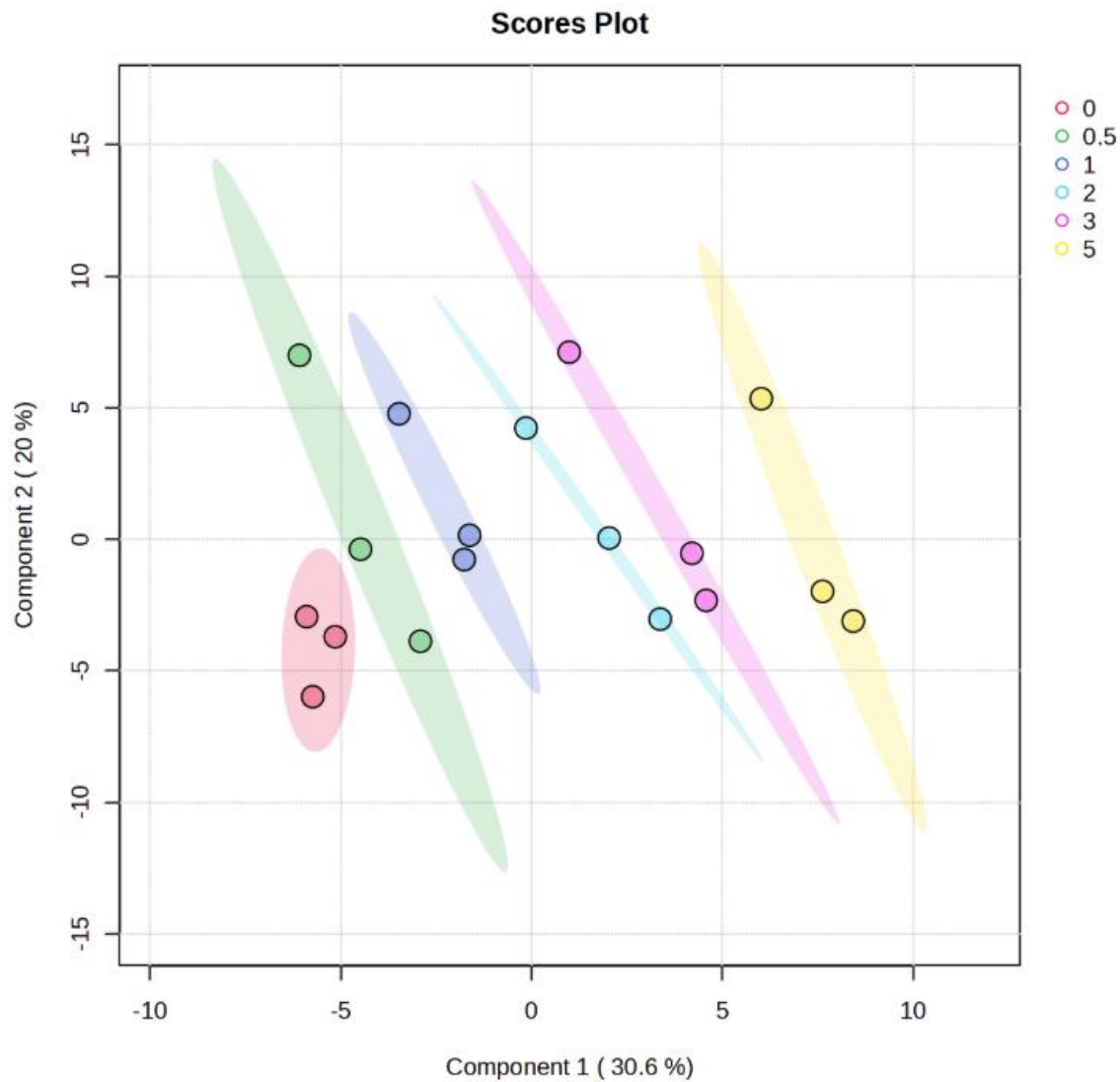
PLSで作られたモデルで、他のデータの分類を予測する方法  
(質的な予測)

# 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

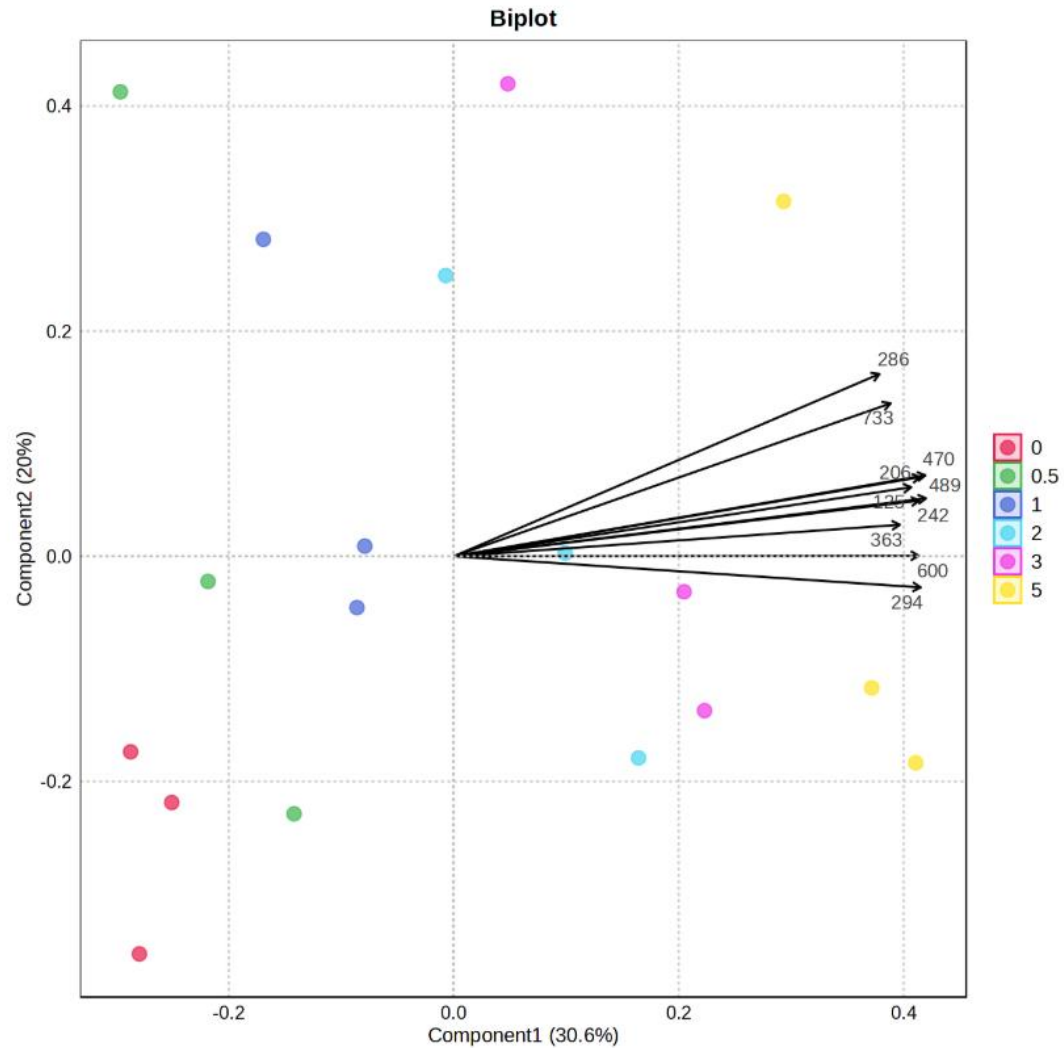


# PLS結果



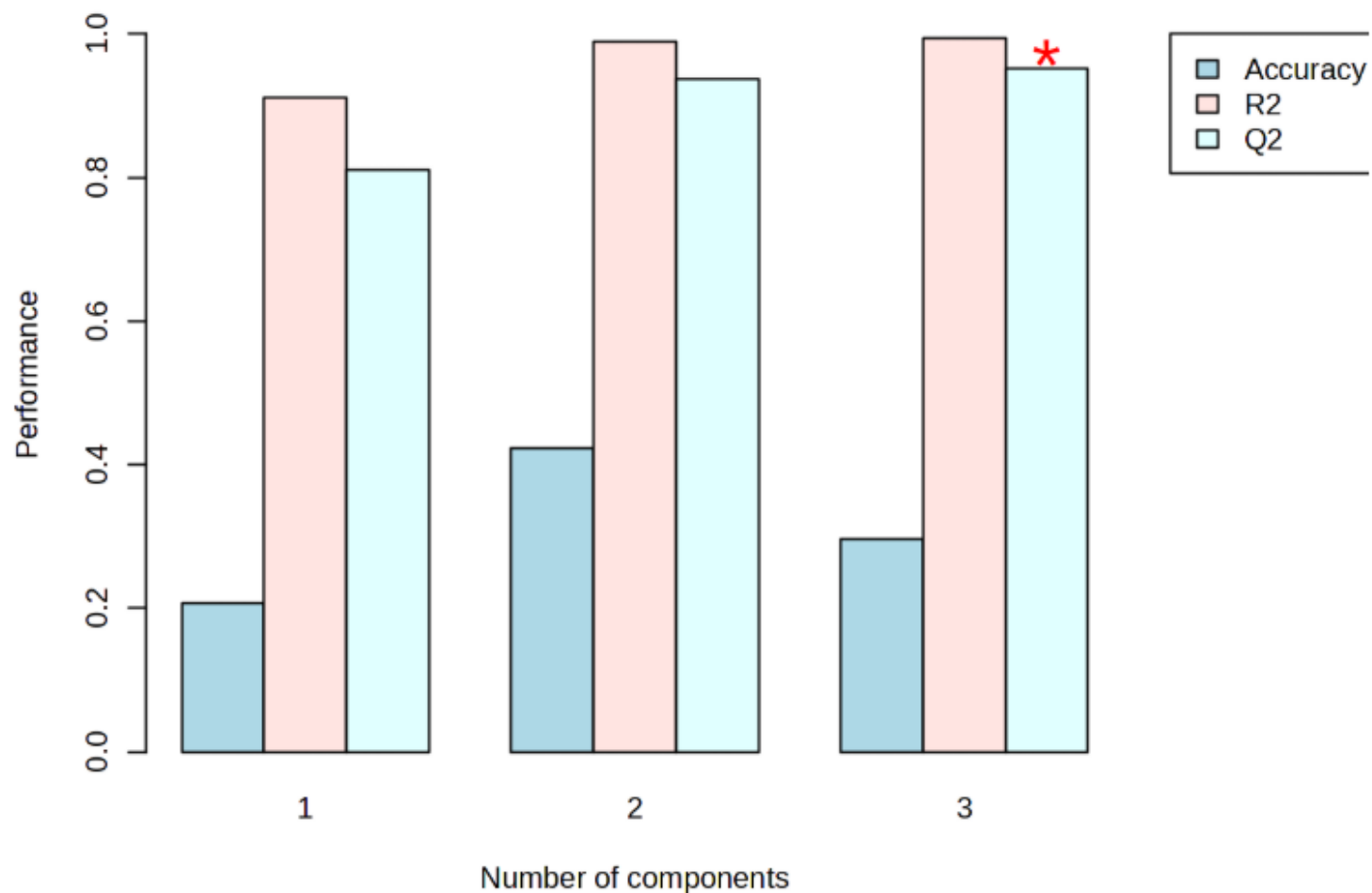
スコアプロット

# PLS結果



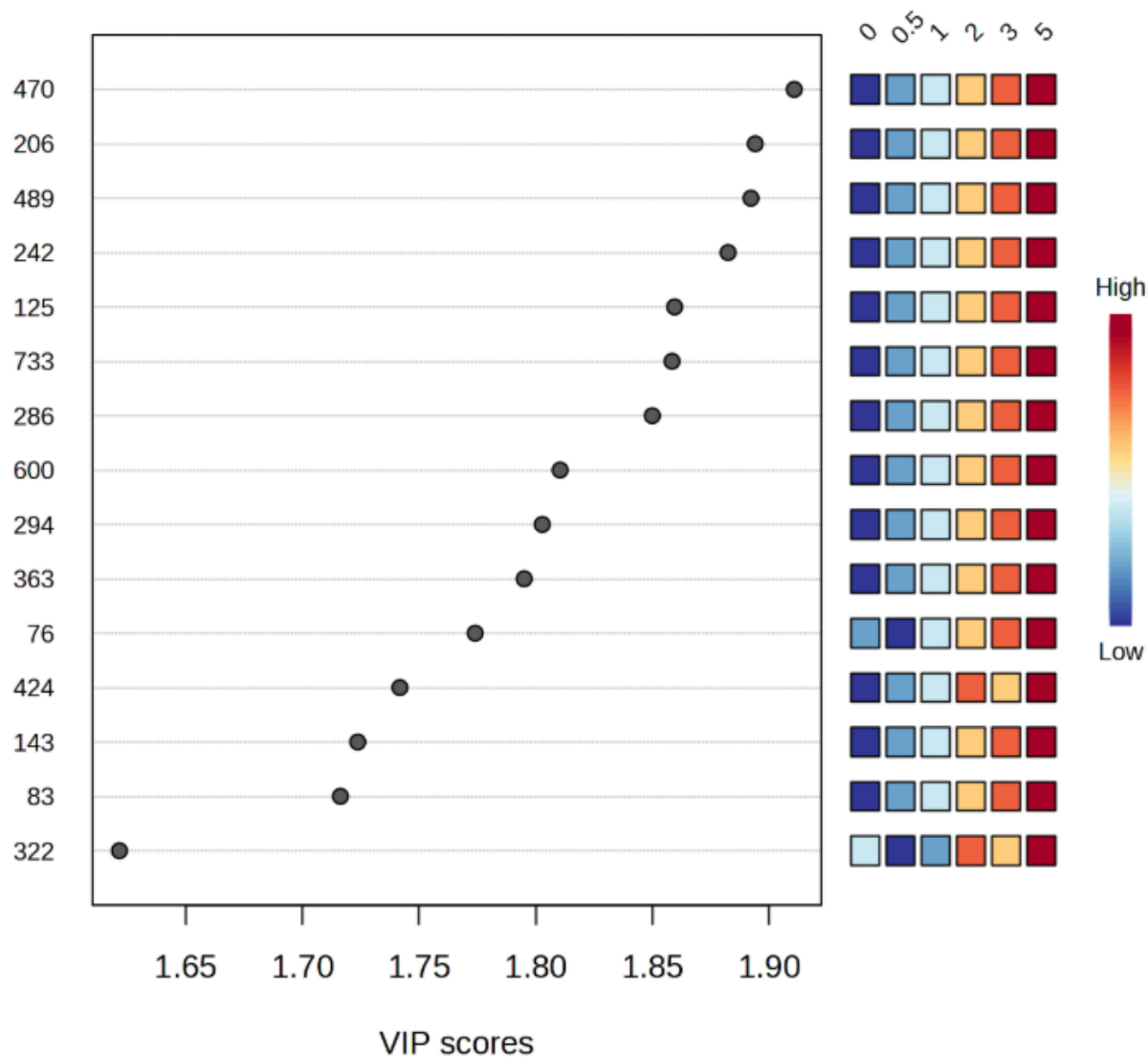
バイプロット

# PLS結果



判別モデルの予測性能評価

# PLS結果



寄与の大きい説明変数