

データサイエンス概論I & II データサイエンス総論I & II

信頼区間と統計的検定

九州大学 数理・データサイエンス教育研究センター

信頼区間

この「信頼区間」で学ぶこと

存在しうるすべてのデータ

実際に手に
入ったデータ

例えば
「平均値」

あなた、この平均値、真の平均値と同じと思います？

いや、ある程度近いとは思うけど、多少違うかも...



※なお以下では平均の信頼区間について論じますが、分散の信頼区間もあります

こういう状況と少し似ている

こないだのテスト，みんな40点ぐらいだよ～

みんなって誰よ？名前言ってみなさい！

AちゃんとB君とCさんが40点だよお

それじゃみんなじゃないでしょ！

じゃあ，平均は30点から60点ぐらいだよお

ウソ言いなさい！平均90点なんじゃないの？

90点はあり得ないよお～



信頼区間①

母集団と標本化

全部が無理なら一部だけ

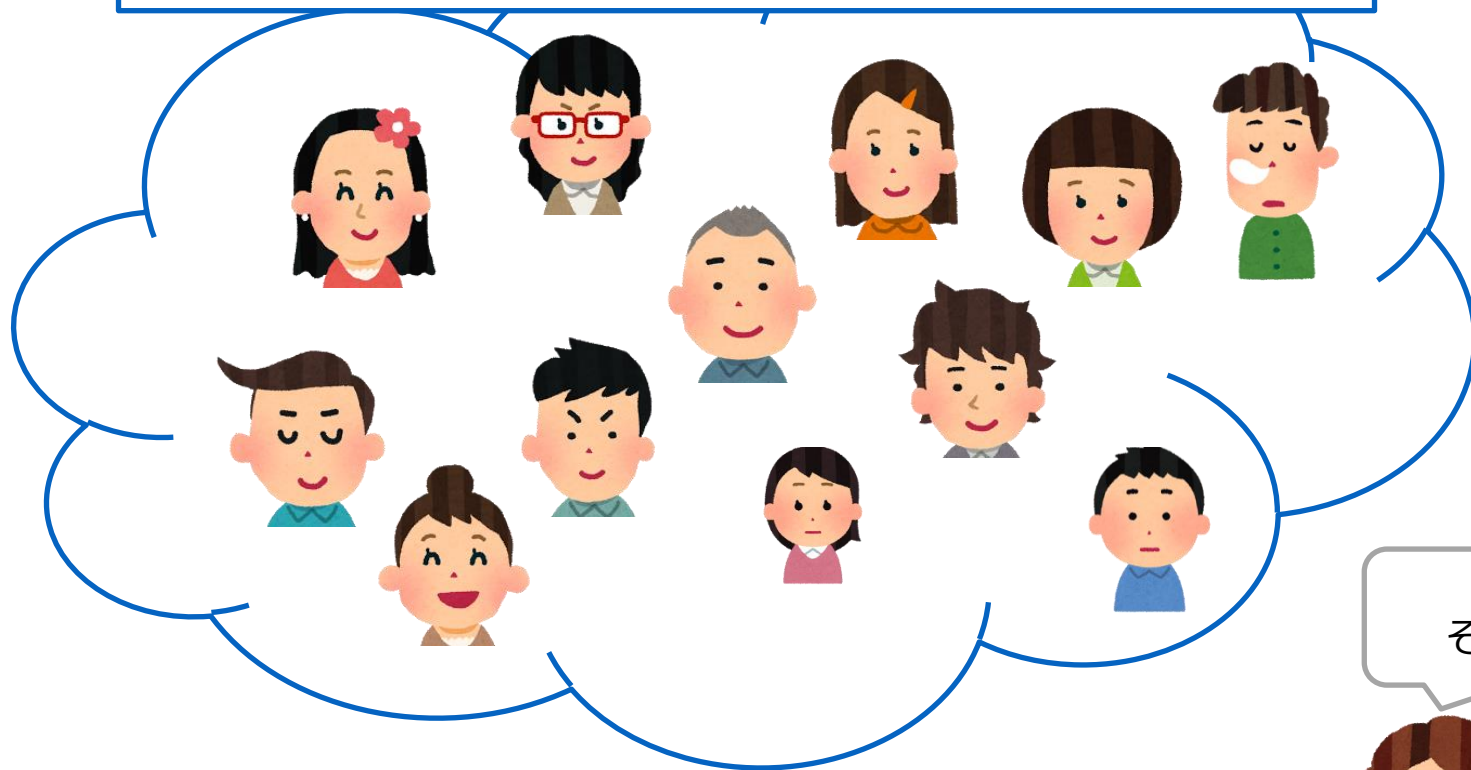
真の平均点を求めたい：例題

- お母さんは、子供が受けたテストの
真の平均点を知りたくなった
- しかし、テストを受けた生徒は相当多いので、
全員のテストの点数を聞くのは無理っぽい
 - 数人なら聞けるかも...
- 真の平均点はわかるのだろうか？



もちろん、**ありとあらゆる生徒**にテストの点数を
聞けば、真の平均はわかるだろう

ありとあらゆる生徒にテストの点数を聞いて回る



さすがに
それは無理



すべてのデータの集合 = 母集団： 母集団全体は見えないことが多い

想定される
数が多い

データ取得中に
増減が起こりうる

全員は無理

コストが
かかる



ちなみに父集団ってのではありません。
お母さんとも無関係。
母集団の英訳は「population」です。
(mothers' groupではないです)

全部が無理なら一部だけ：標本

母集団：ありとあらゆる生徒

無作為抽出
(ランダムサンプリング)

標本

81点

67点

77点

こちらは
わかる

知りようが
ない

知りようが
ない

μ : 母平均

(ありとあらゆる生徒の平均点)

σ^2 : 母分散

(ありとあらゆる生徒の平均点からのばらつき)

\bar{x} : 標本平均

$$= (81 + 67 + 77) / 3 = 75$$

s^2 : 標本分散

$$= (6^2 + 8^2 + 2^2) / 3 \approx 34.7$$

N : 標本サイズ = 3

やりたいこと： 標本平均から母平均を推定したい

母集団：ありとあらゆる生徒

無作為抽出
(ランダムサンプリング)

標本

①これを知りたいのだが、
全生徒(母集団)が
不可知なので知りようがない

②標本を用いて
計算できます

③なるべく
正確に
推定したい

μ : 母平均

(ありとあらゆる生徒の平均点)

σ^2 : 母分散

(ありとあらゆる生徒の平均点からの)

\bar{x} : 標本平均

s^2 : 標本分散

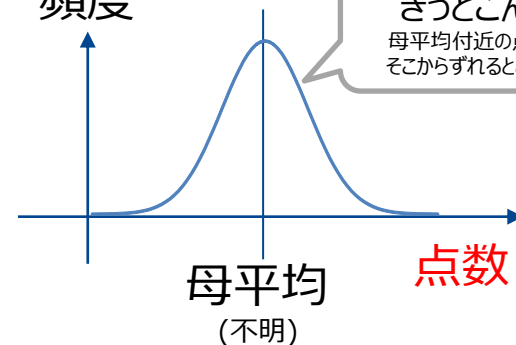
N : 標本サイズ

この「信頼区間と統計的検定」のスライドでの仮定： 母集団の分布は正規分布

母集団：ありとあらゆる生徒



頻度



- この仮定により以下の議論がずいぶん簡潔に！

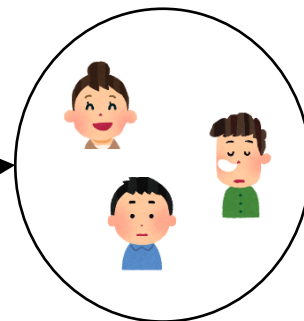
信頼区間② 標本平均の分布



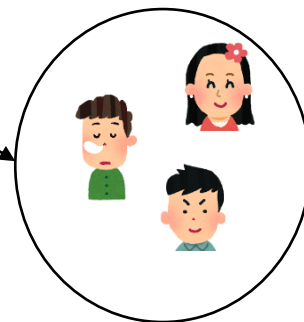
「分布の平均」じゃなくて「平均の分布」

異なる標本からは異なる標本平均が出る
→ T 回標本抽出すると T 個の標本平均が出る

母集団：ありとあらゆる生徒

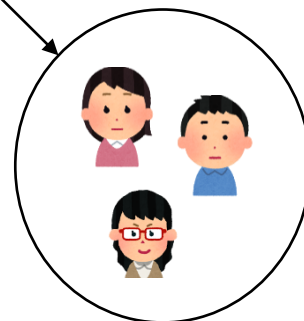


標本 1
標本平均 64点



標本 2
標本平均 71点

⋮

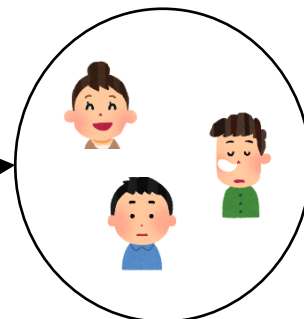


標本 T
標本平均 95点

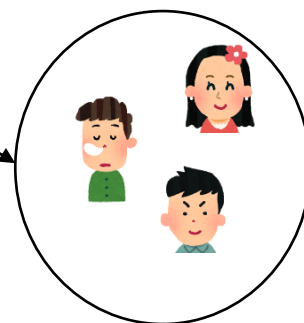
うーむ、どれが
母平均に近いのか!?
(母平均はわからないのだけど...)

とりあえず「標本平均の分布」を考えてみる： 「分布の平均」じゃなくて「平均の分布」

母集団：ありとあらゆる生徒

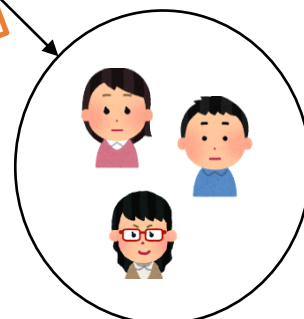


標本 1
標本平均 64点



標本 2
標本平均 71点

⋮



標本 T
標本平均 95点

頻度

どんな分布と
思います？

標本平均の
分布

標本平均 \bar{x}



ここでもし、母集団の分布が(母平均は不明だが) 正規分布とわかってるなら…

母集団：ありとあらゆる生徒

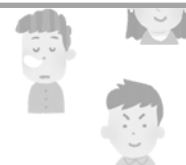


頻度



母平均
(不明)

点数



標本 2

標本平均 71点

⋮



標本 T

標本平均 95点

頻度

どんな分布と
思います?

標本平均の
分布

標本平均 \bar{x}



なんと「標本平均の分布」も、 母平均を中心とした正規分布に！

母集団：ありとあらゆる生徒



頻度



母平均
(不明)

点数

標本 2

標本平均 71点

⋮

標本 T

標本平均 95点

頻度



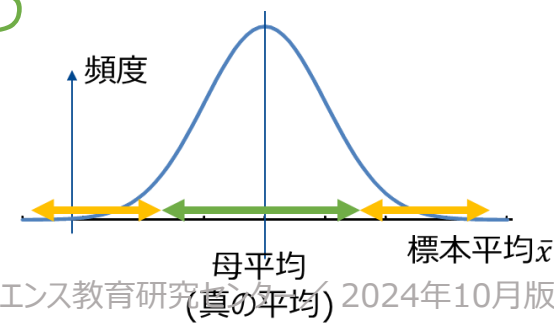
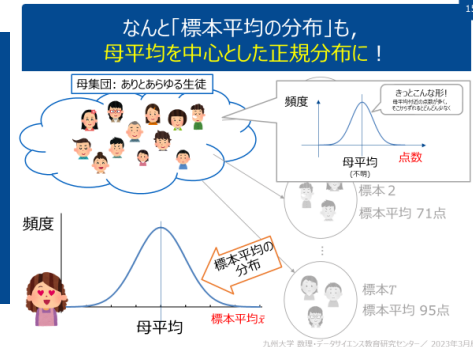
標本平均の
分布

母平均

標本平均 \bar{x}

皆さんがうっすら考えていることと一致しませんか？

- 何人か選んで平均(=標本平均)をとる
- それは真の平均(=母平均)ではないはず
- 運が悪ければ、真の平均と全く違うこともあるだろう
- でも、真の平均と全く違うことは珍しいだろう
- どちらかと言えば真の平均に近いことのほうが多そう
- ということは、標本平均の分布はこんな感じ→



疑問：どちらも母平均を中心とした正規分布だが、同じものなのか？

母集団：ありとあらゆる生徒



頻度

きっとこんな形！
母平均付近の点数が多く、
そこからずれるとどんどん少なく

母集団の分布

母平均
(不明)

点数

頻度

標本平均の分布

母平均

標本平均 \bar{x}

同じもの？

答：分散(ばらつき)が違います

母集団：ありとあらゆる生徒



頻度

きっとこんな形！
母平均付近の点数が多く、
そこからずれるとどんどん少なく

母集団の分布

母平均
(不明)

点数

分散(ばらつき)は「あるもの」によって変わります

頻度

標本平均の分布

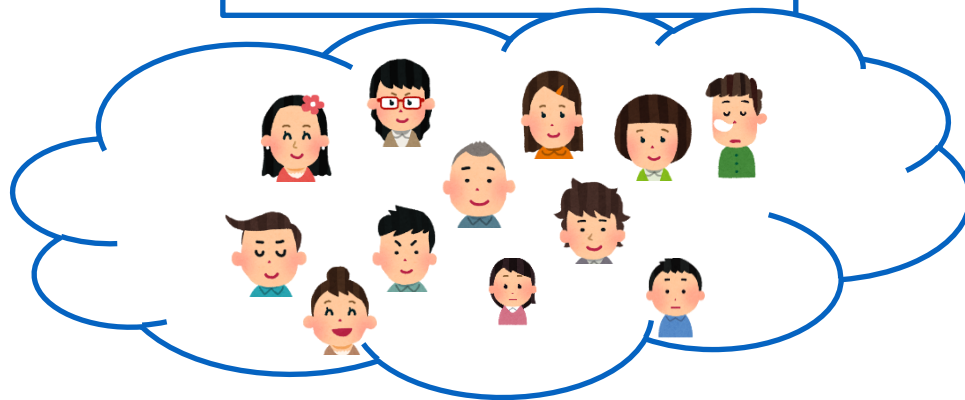
母平均

標本平均 \bar{x}

分散が
違うのか！

「標本平均の分布」のばらつきは 標本サイズ N に依存 (1/3)

母集団：ありとあらゆる生徒



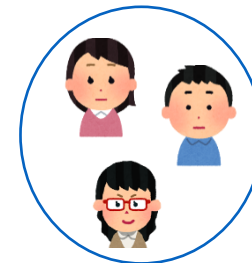
標本サイズ小



標本平均
64点



標本平均
71点



標本平均
95点

標本サイズ大



標本平均
72点



標本平均
70点

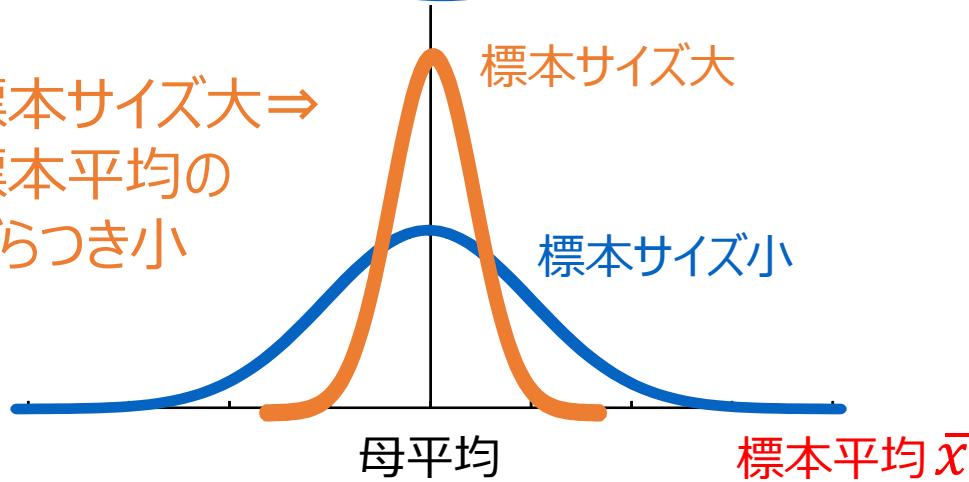


標本平均
68点

標本サイズ大⇒
標本平均の
ばらつき小

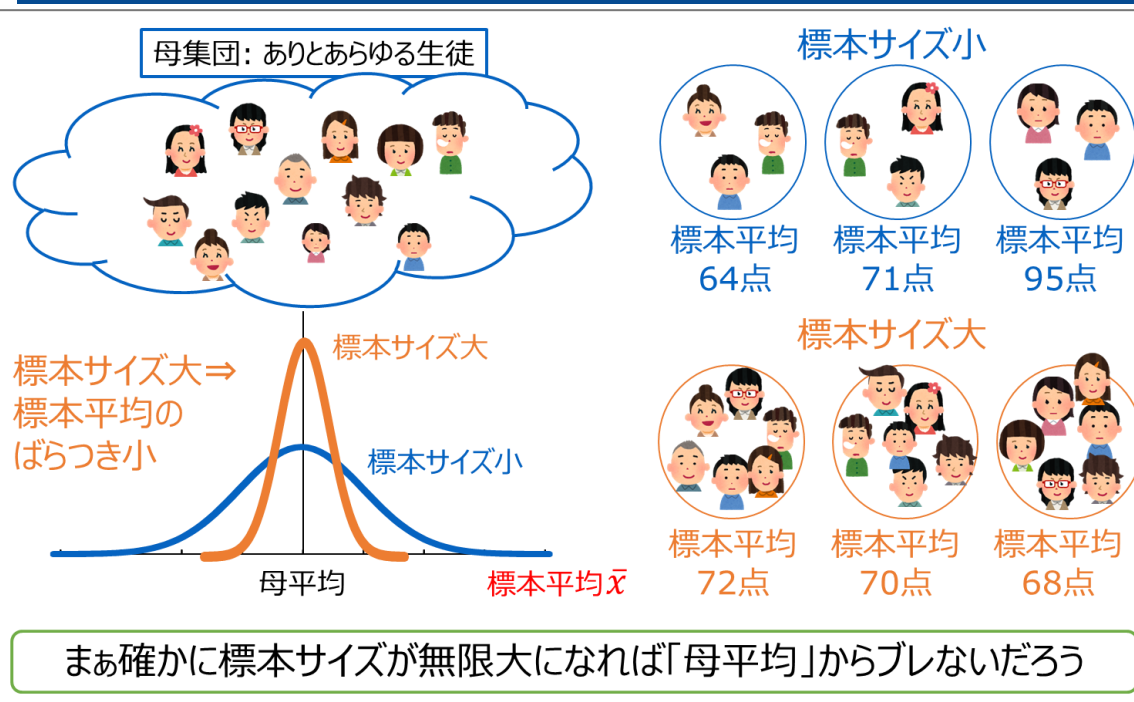
標本サイズ大

標本サイズ小



まあ確かに標本サイズが無限大になれば「母平均」からブレないだろう

「標本平均の分布」のばらつきは 標本サイズ N に依存 (2/3)



「標本サイズを大きくして平均を求めれば、母平均に近づく
(=母平均からのブレは少なくなる)」

この一見アタリマエにも見える傾向を、「大数の法則」と呼ぶ

「標本平均の分布」のばらつきは 標本サイズ N に依存 (3/3)

- 標本平均の分散（標本平均のばらつき）

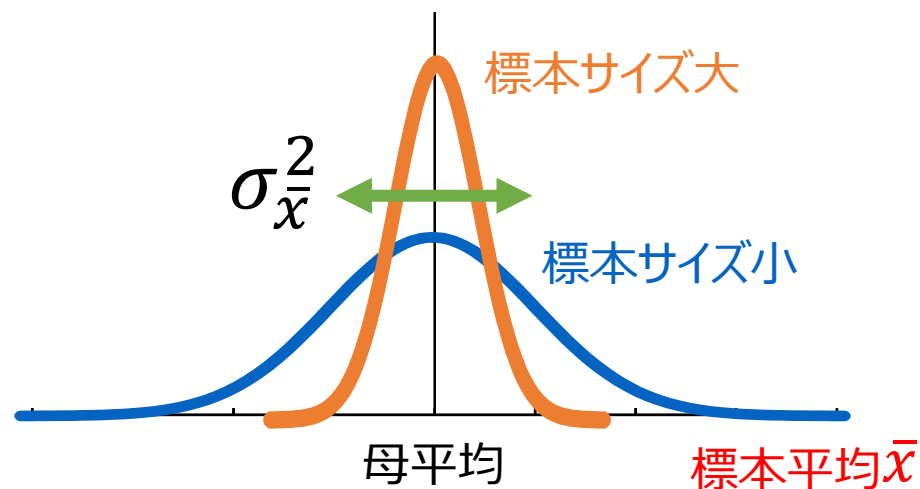
$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{N}$$

σ^2 : 母分散
 N : 標本サイズ

標本サイズ N が大きくなるほど、
ばらつき $\sigma_{\bar{x}}^2$ はどんどん小さく



というわけで、基本的には
データは多いほうがよい



ここまでわかったこと

- 全データ(母集団)から真の平均を求めたい
- 全データ調査は非現実的．普通は限られた標本しかない
- 残念ながら，標本平均は，毎回変わり，真の値ではない
- ただし，母集団が正規分布ならば，標本平均は，平均＝「母平均 μ 」，分散＝「母分散 σ^2/N 」の正規分布に従う

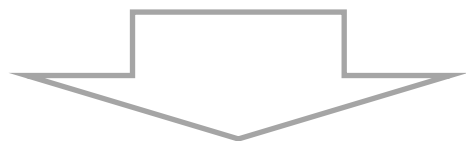
参考：母集団が正規分布でなくても， N が大きくなれば，標本平均の分布は正規分布に近づきます！→【付録1】中心極限定理

この事実をなんとか使いたい

平均90点
ぐらいでしょ！



- 全データ(母集団)から真の平均を求めたい
- 全データ調査は非現実的．普通は限られた標本しかない
- 残念ながら，標本平均は，毎回変わり，真の値ではない
- ただし，母集団が正規分布ならば，標本平均は，平均＝「母平均 μ 」，分散＝「母分散 σ^2/N 」の正規分布に従う



- この事実を使って，「この範囲に真の平均はありそう」
ぐらいは言えないか？

みんなの平均も
30～60点だよ



信頼区間③

母分散がわかっている場合の 信頼区間

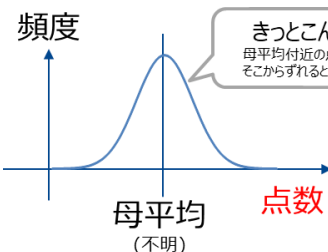
真の平均をピッタリ答えるのは無理だが、
(ある程度の確からしさで)「この範囲にありそう!」と答えることはできる

さあ、いま標本平均と標本分散は分かった！ これからどうする？

母集団：ありとあらゆる生徒

無作為抽出
(ランダムサンプリング)

標本

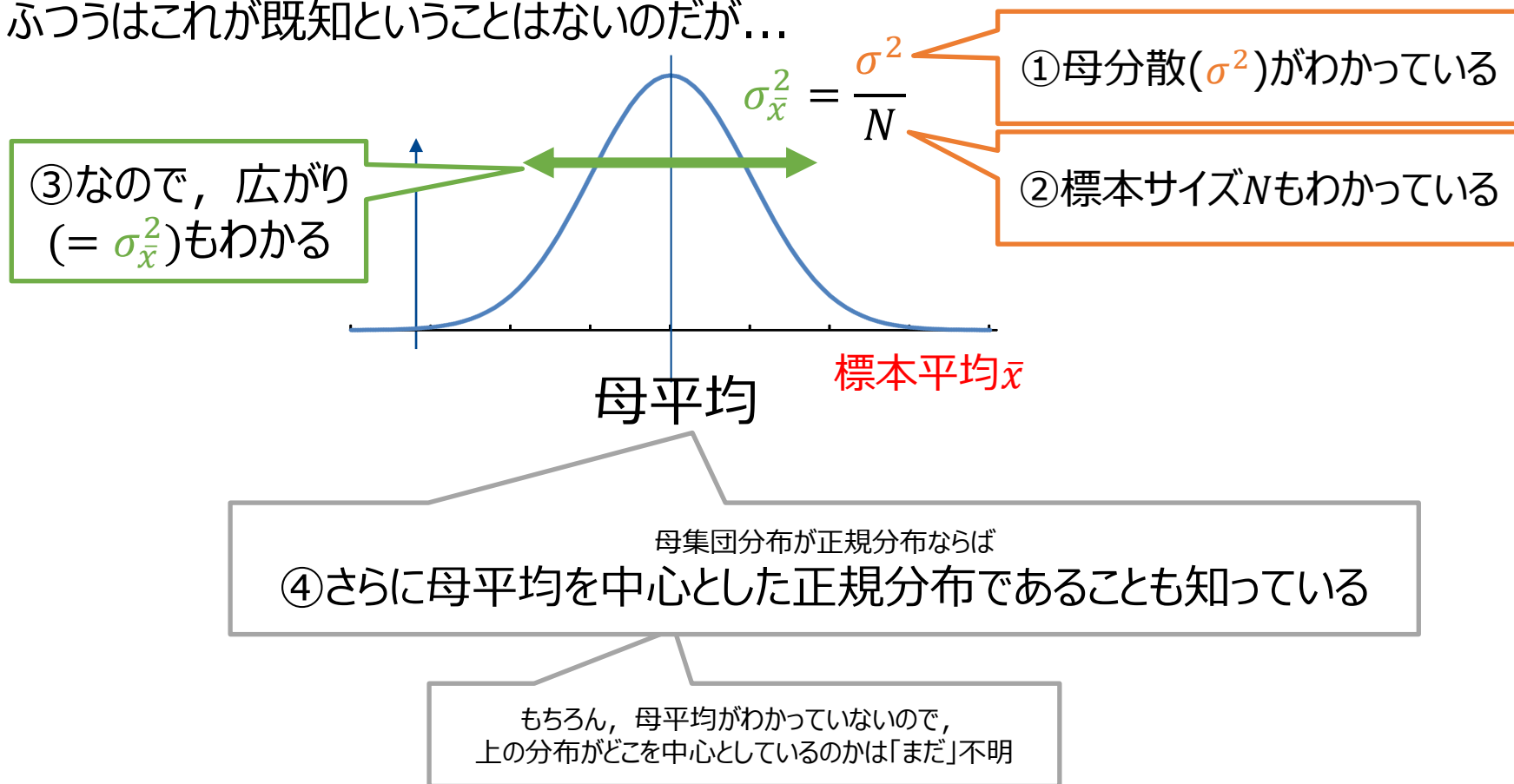


とりあえず
わかっていること

標本平均 64点
標本分散 78
標本サイズ $N = 3$

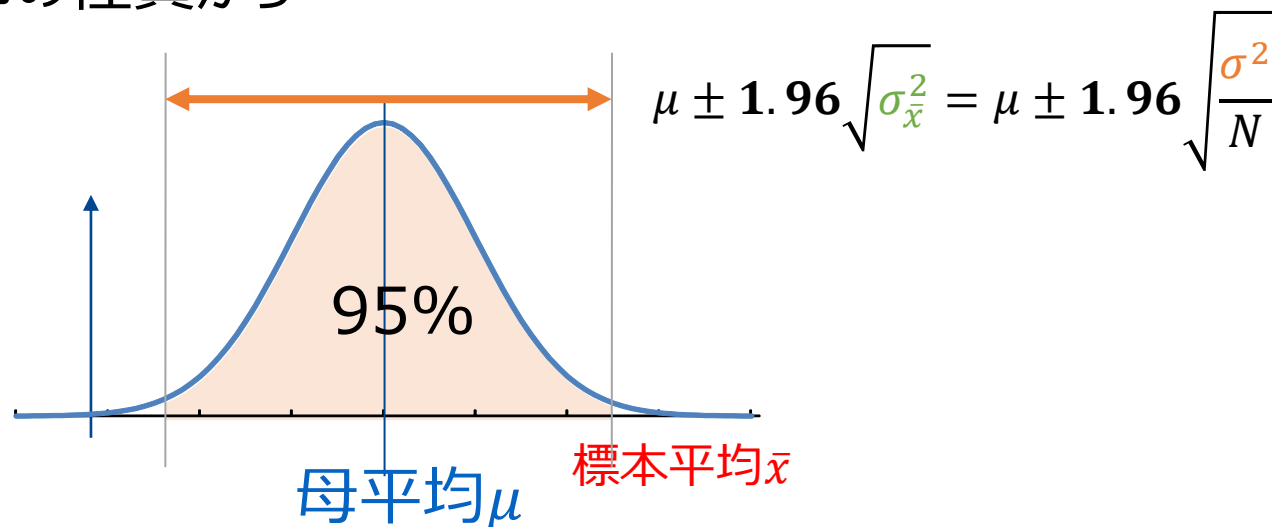
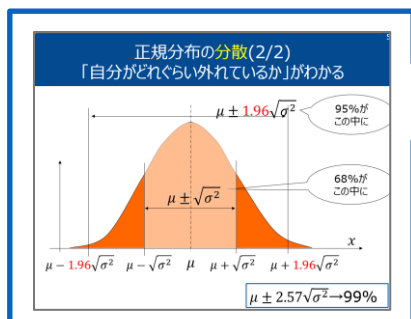
母平均(真の平均)はどの範囲にある？ : さらに母分散も(なぜか)わかっているなら (1/4)

- 母分散 σ^2 = すべての生徒のテストの点数の分散
- ふつうはこれが既知ということはないのだが...



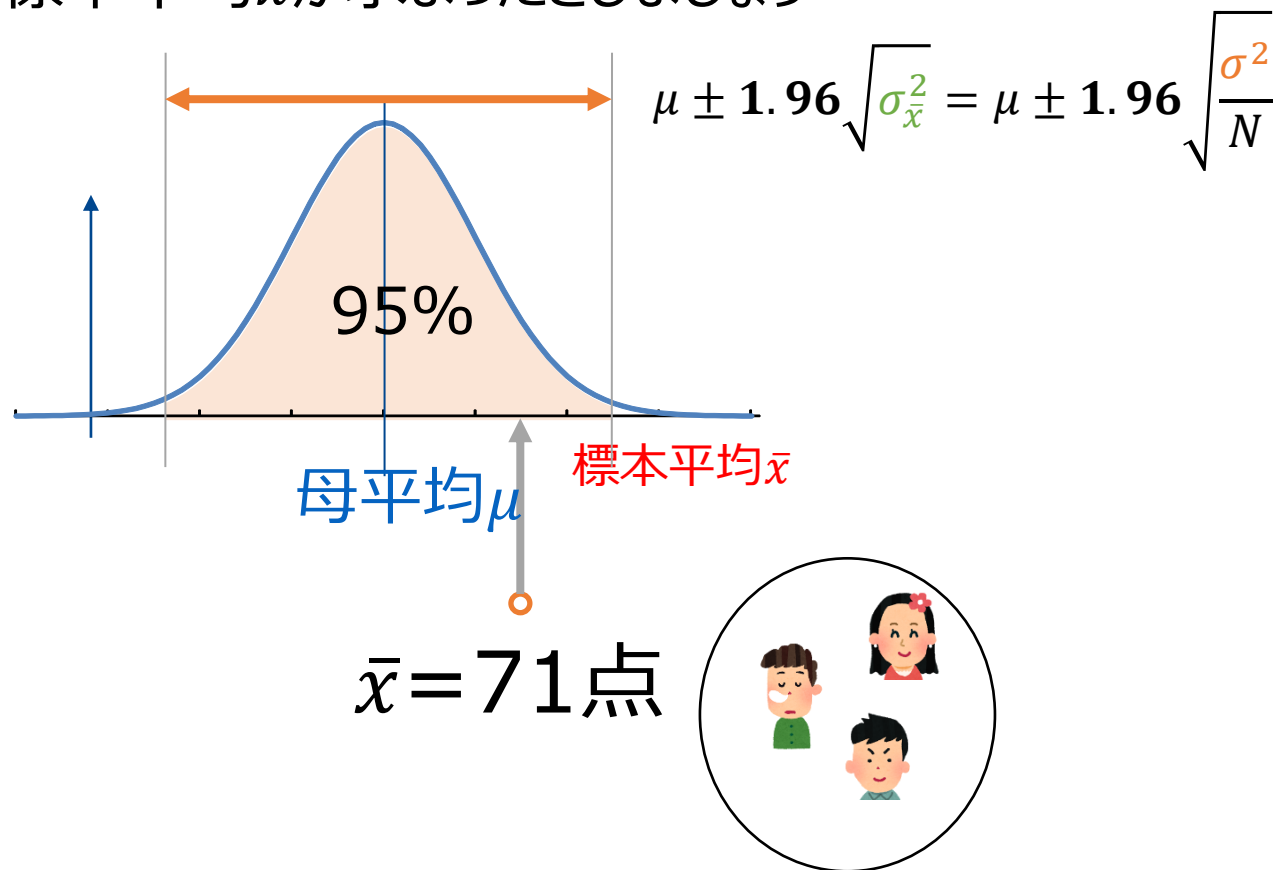
母平均(真の平均)はどの範囲にある？ : さらに母分散も(なぜか)わかっているなら (2/4)

- ところで正規分布の性質から...



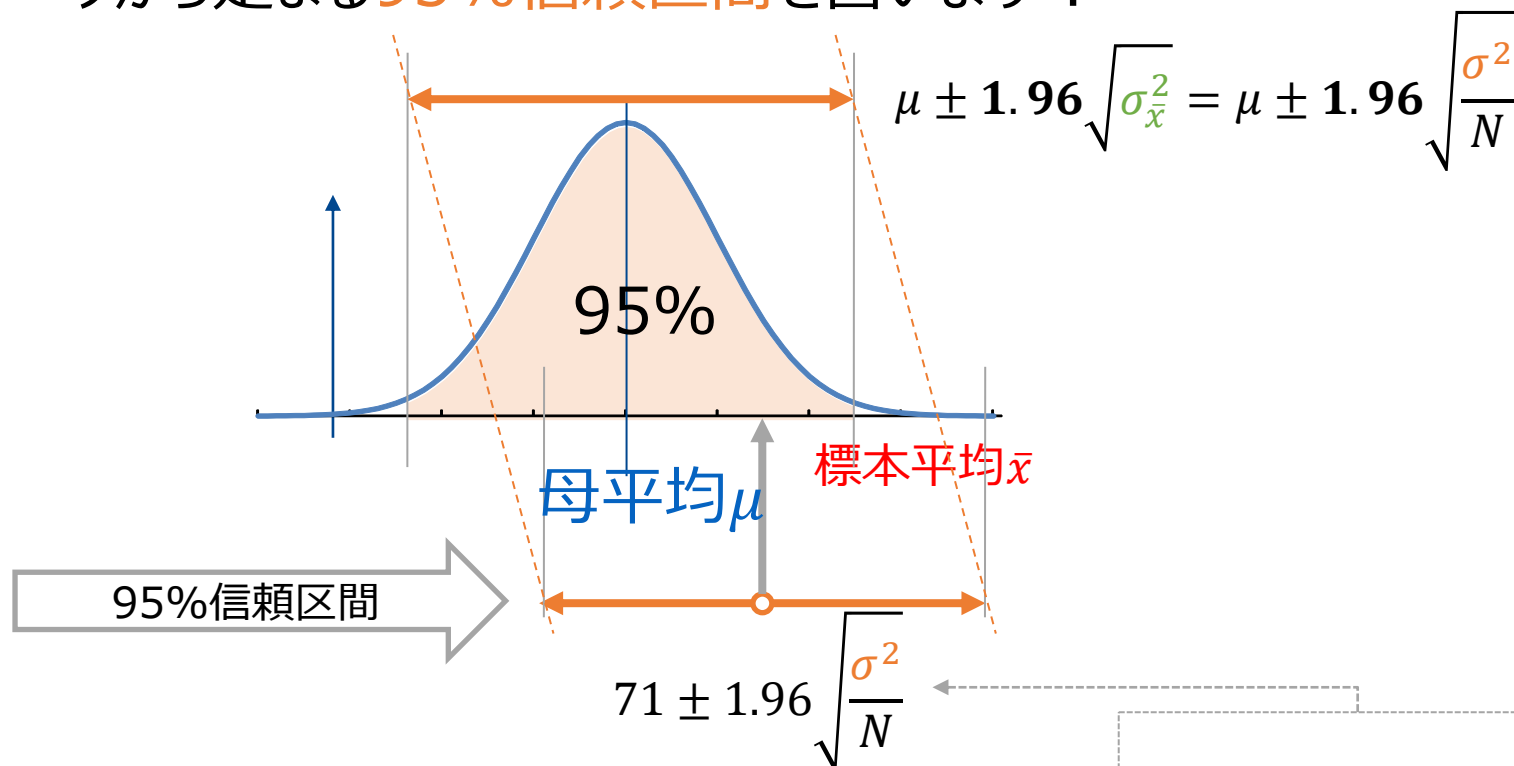
母平均(真の平均)はどの範囲にある？： さらに母分散も(なぜか)わかっているなら (3/4)

- さて, いま一つの標本平均 \bar{x} が求まったとしましょう



母平均(真の平均)はどの範囲にある？ : さらに母分散も(なぜか)わかっているなら (4/4)

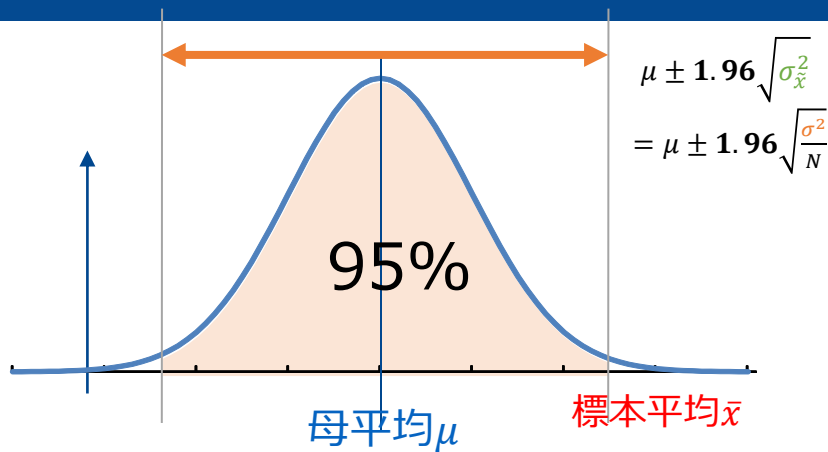
- これをデータから定まる95%信頼区間を言います！



ちゃんと区間として書くと
 $\left[71 - 1.96 \sqrt{\frac{\sigma^2}{N}}, 71 + 1.96 \sqrt{\frac{\sigma^2}{N}} \right]$

- = 「母平均の存在範囲」の推定結果！
- より正確な意味は後述

ん？ 標本平均(ex. 71点)から, 95%信頼区間をどうやって定めた??

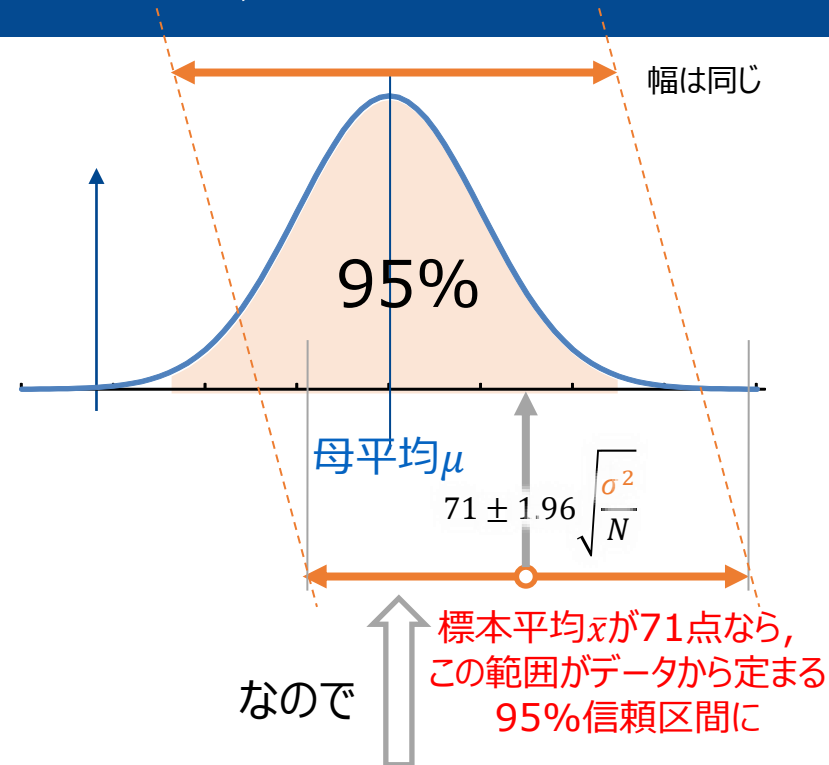


なので

確率95%で次が成り立つ

$$\text{母平均} - 1.96\sqrt{\frac{\sigma^2}{N}} \leq \text{標本平均} \leq \text{母平均} + 1.96\sqrt{\frac{\sigma^2}{N}}$$

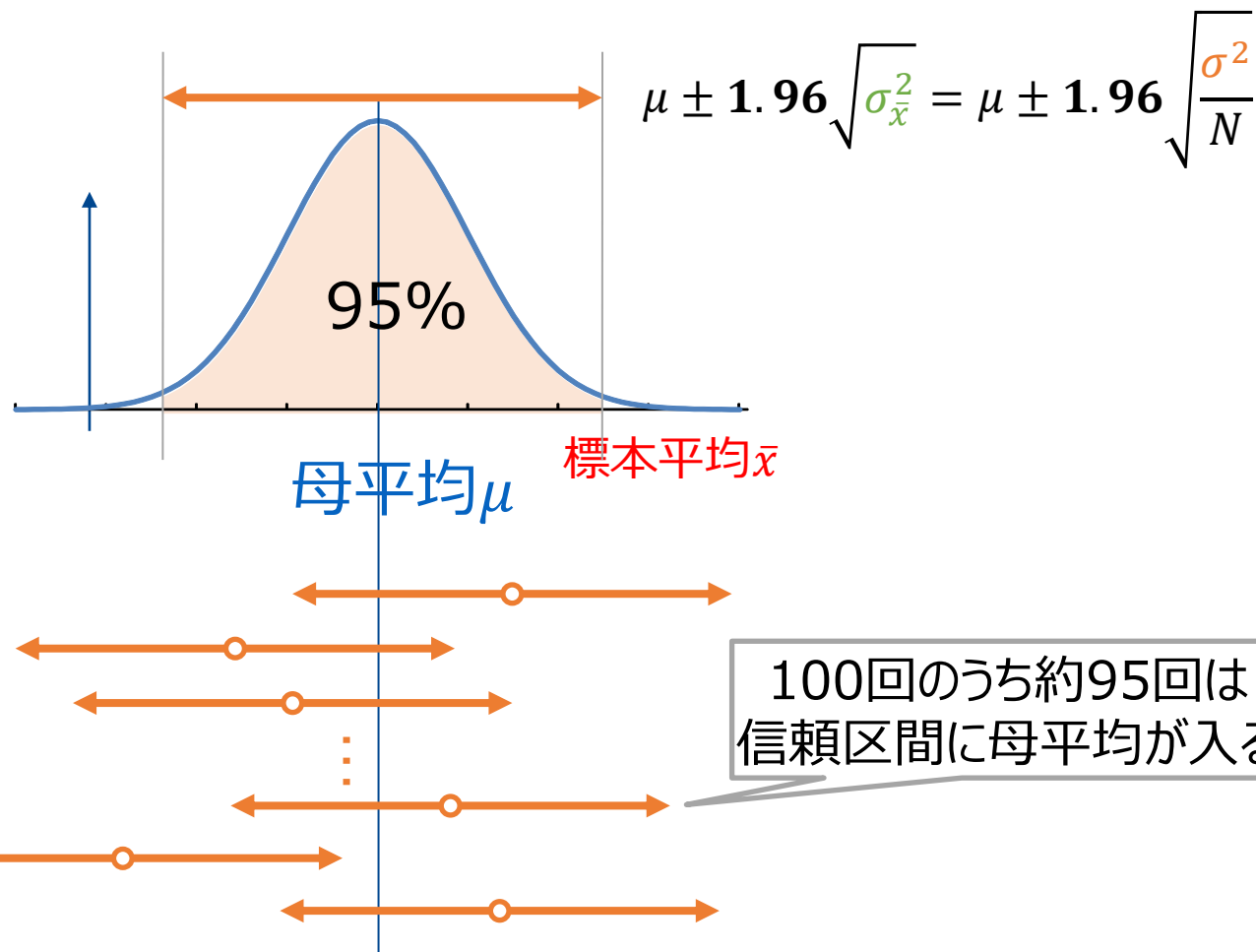
単なる
移項



$$\text{標本平均} - 1.96\sqrt{\frac{\sigma^2}{N}} \leq \text{母平均} \leq \text{標本平均} + 1.96\sqrt{\frac{\sigma^2}{N}}$$

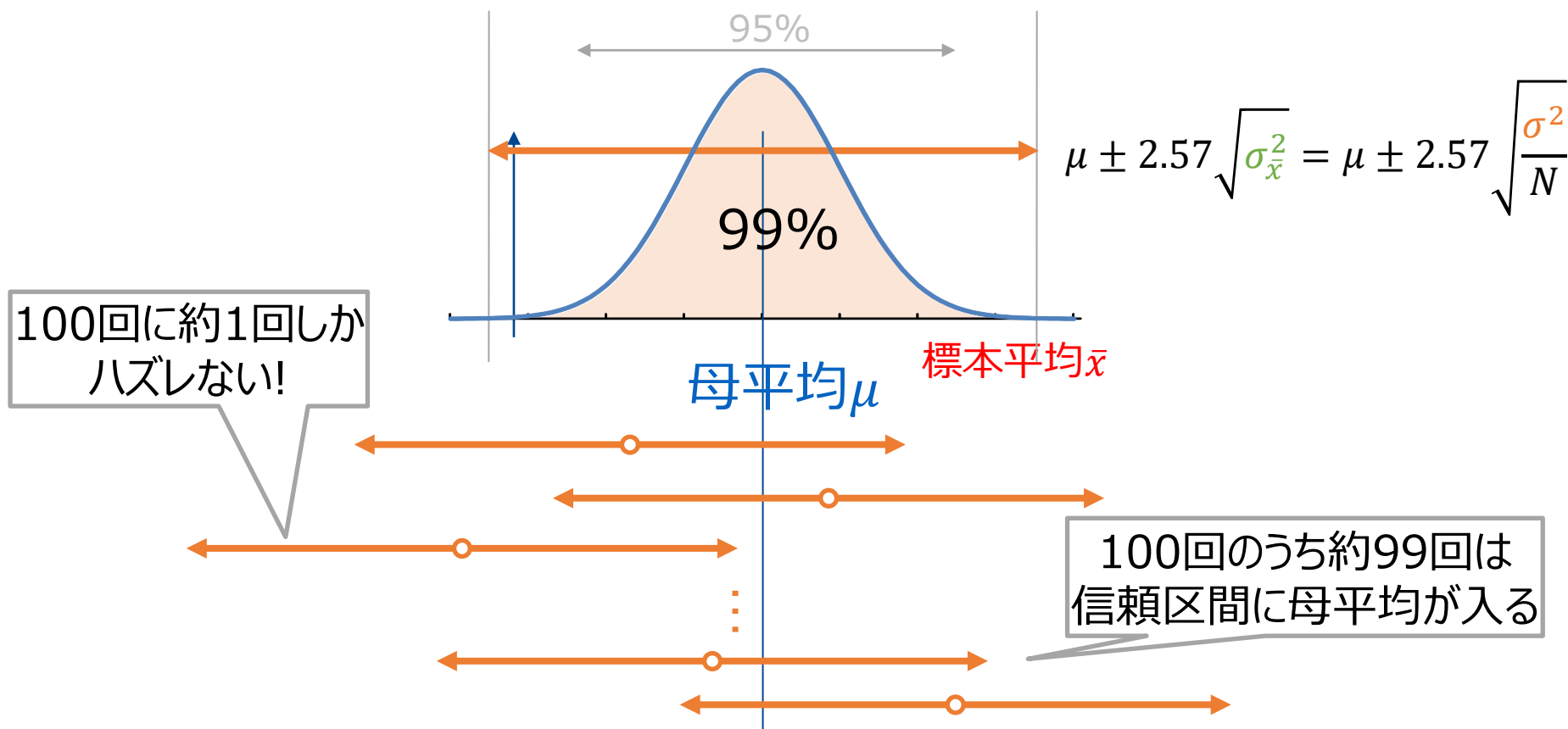
$p\%$ 信頼区間の意味： 何度も標本平均を求め、その度にその区間を求めたら、それらのうち $p\%$ は母平均を含む

- $p = 95\%$ の場合



$p\%$ 信頼区間の意味： 何度も標本平均を求め、その度にその区間を求めたら、それらのうち $p\%$ は母平均を含む

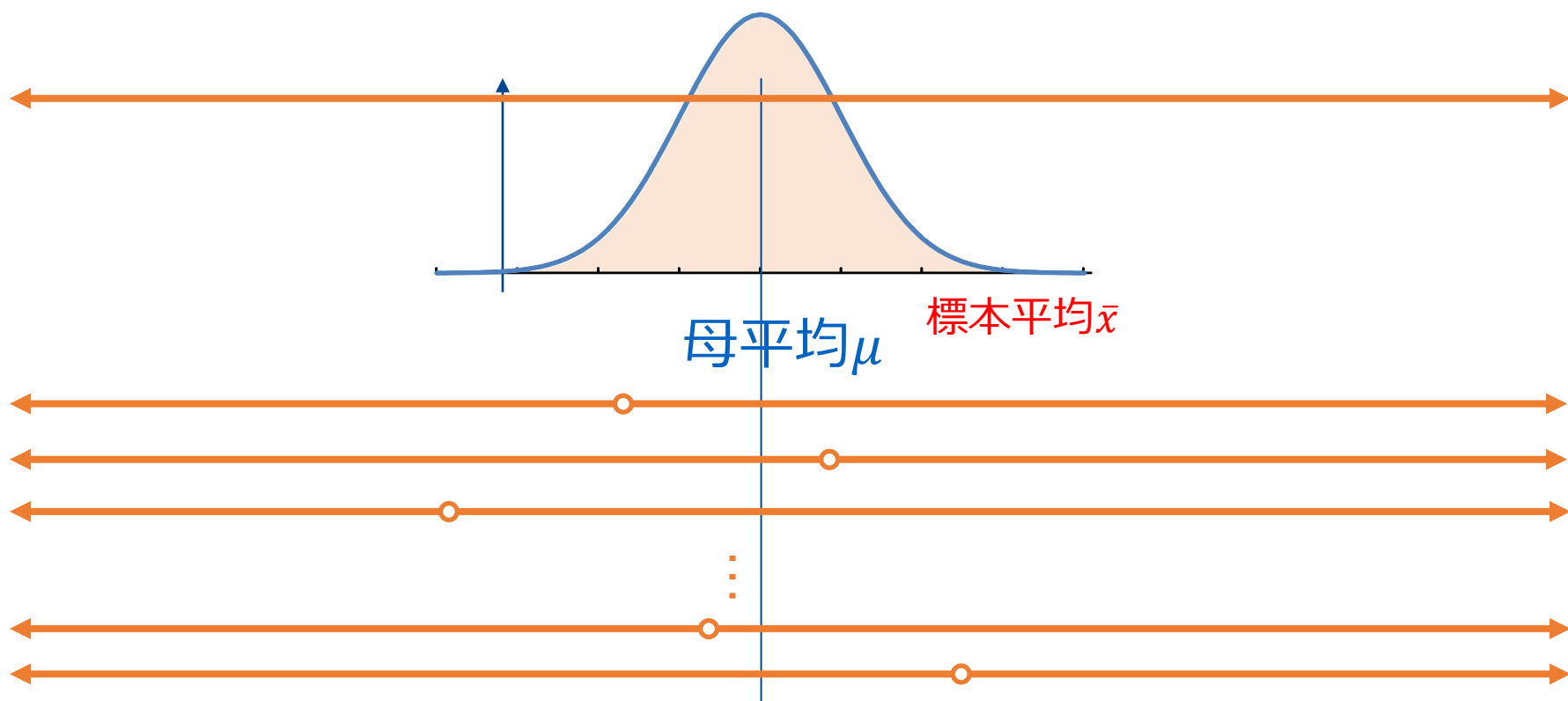
- $p = 95\% \rightarrow 99\%$ 信頼区間になると...



- ハズレないのはうれしいけど、区間幅が増加。母平均の姿はとらえにくくなる... なので...

100%信頼区間がいいなあ～

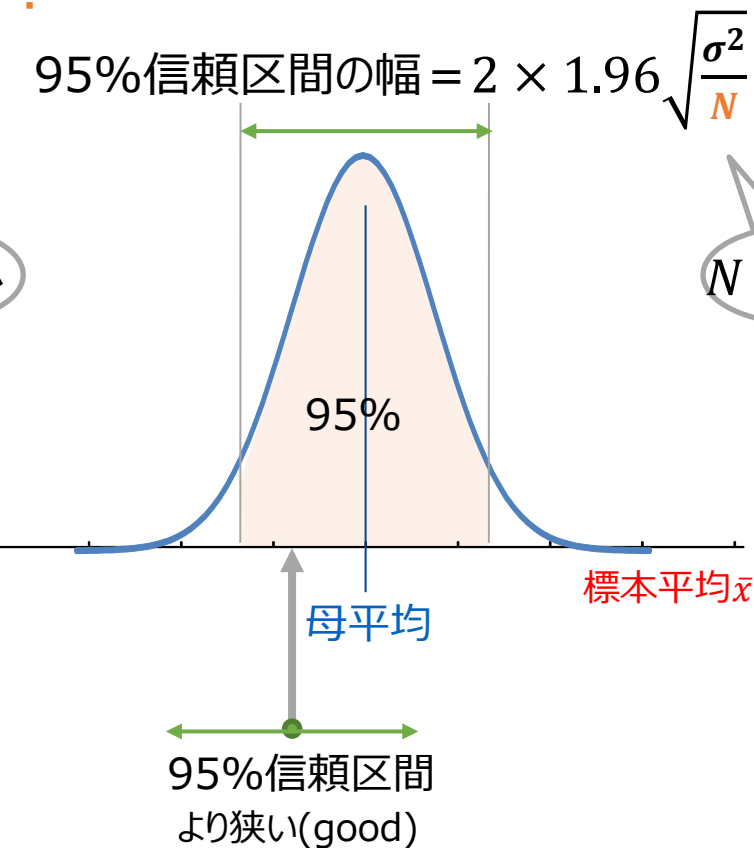
- 区間幅を無限大にすれば100%に
- しかし母平均がどの辺か全くわからないので無意味！



- これに対し「5%犠牲にする」だけで(95%信頼区間), 範囲をグッと絞り込める

では、どうすれば信頼区間を狭くできる？

- p の値を小さくするのは、本末転倒…
- ならば標本サイズ N を増やせばよい！



信頼区間④

母分散がわからない場合の 信頼区間

手元にある「標本分散」だけでなんとかしないと！

Q: 普通は母分散がわからない. どうするか?

A: 標本サイズ N が大きいなら標本分散で代用

標本サイズ N
= (例えば)100



標本平均 72点

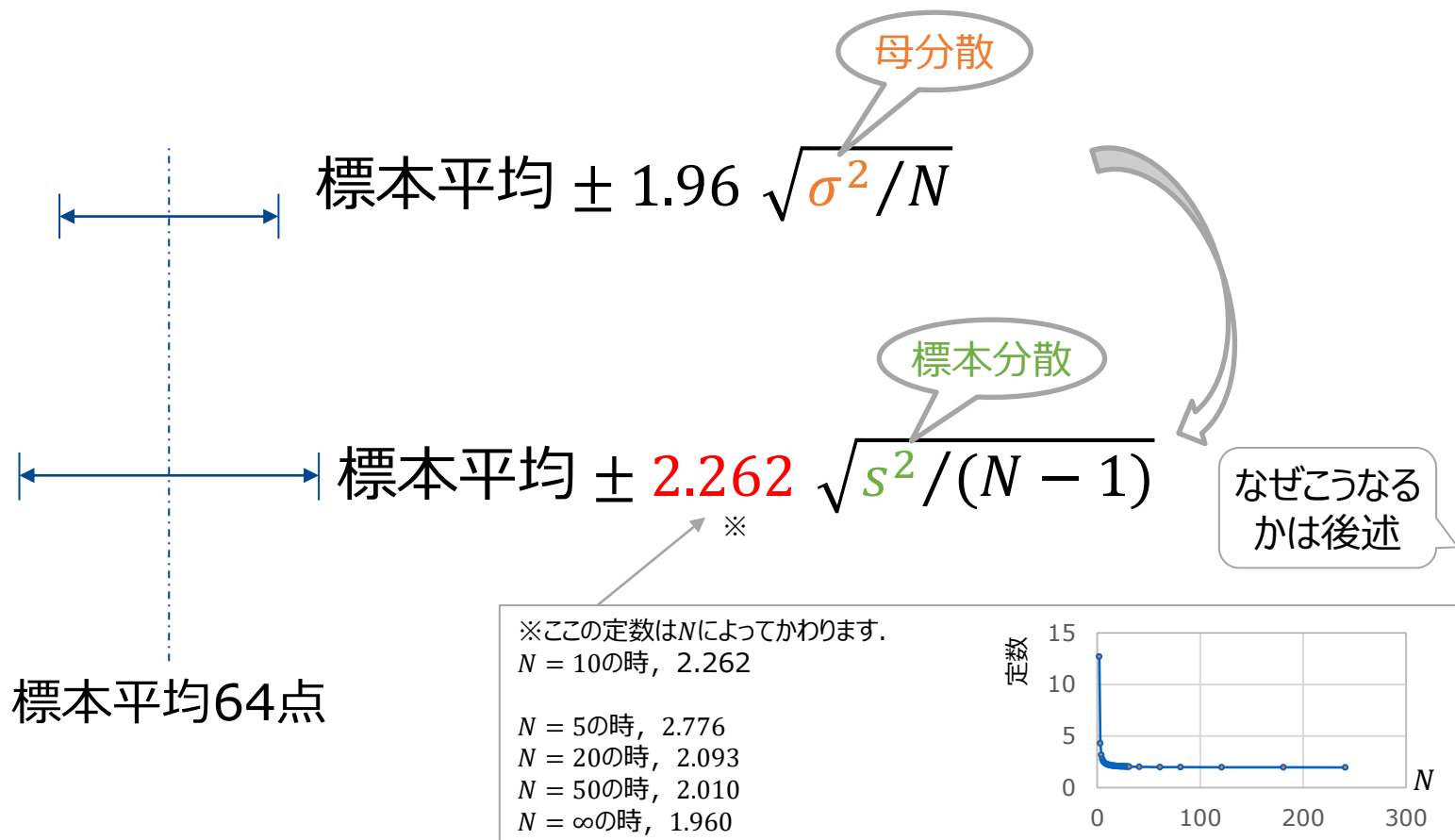
標本分散 s^2 78 \doteq 母分散 σ^2 と「みなす」



あとは「母分散がわかっているケース」と同じ

Q:母分散は不明, N も小さい. さて, どうする?

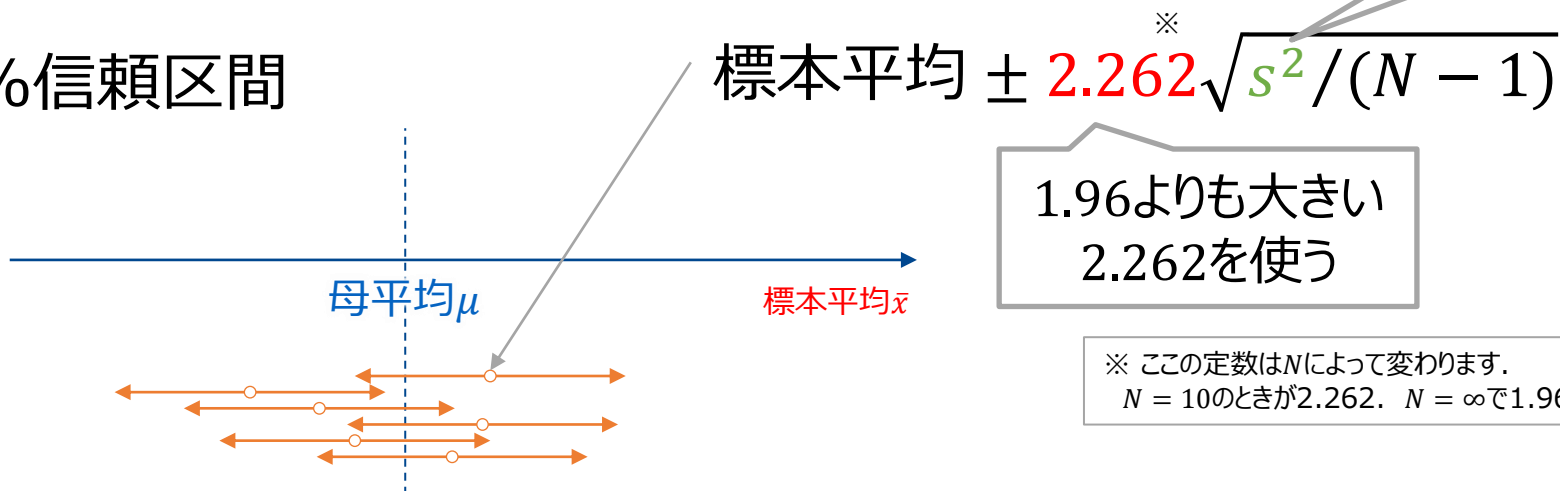
- 95%信頼区間なら, 標本分散を使う代償に, 1.96をもう少し大きくして, 勘弁してもらう!



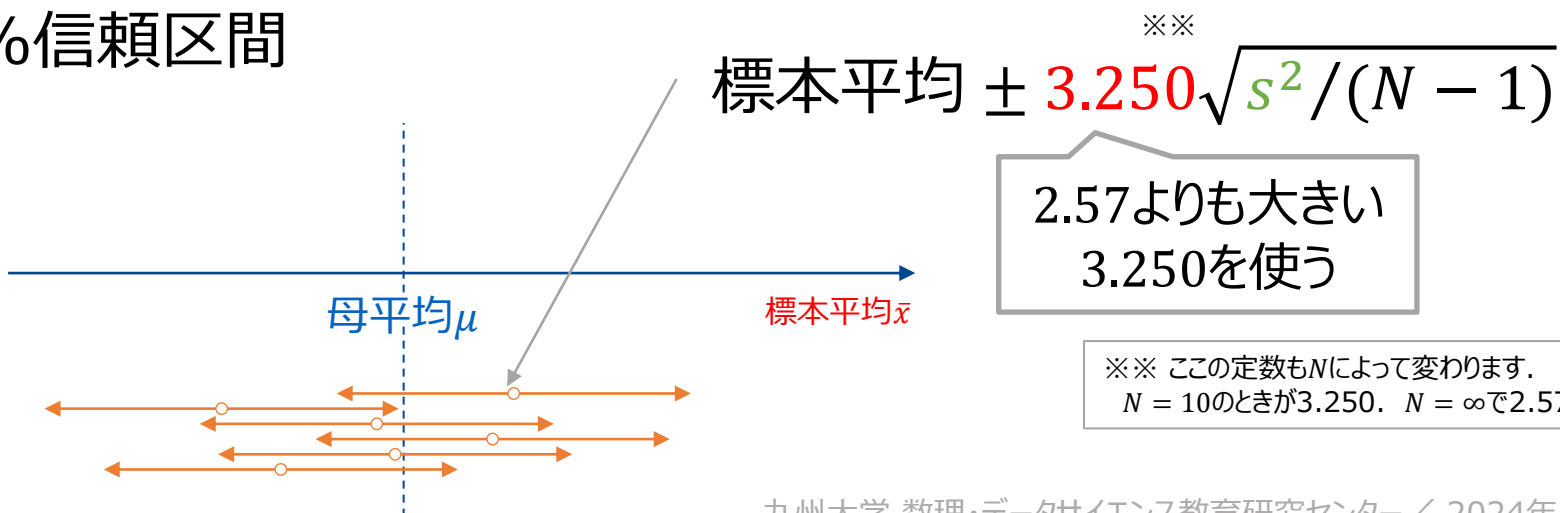
「母分散も不明， N も小さい」場合の 95%信頼区間と99%信頼区間

標本分散

● 95%信頼区間



● 99%信頼区間



平均90点
ぐらいでしょ！



ここまでのまとめ

90点は
信頼区間の
外側だよお



- 標本平均から「 $p\%$ 信頼区間」がわかる
 - 100回標本抽出すると、約 p 回はその区間内に母平均がある
- 母分散 σ^2 がわかっているのなら信頼区間は 標本平均 \pm 定数 $\sqrt{\frac{\sigma^2}{N}}$
 - $p = 95(\%)$ なら、定数=1.96
- 母分散 σ^2 がわからないときは、標本分散 s^2 を代わりにつかう
 - 標本平均 \pm 定数 $\sqrt{\frac{s^2}{N-1}}$
 - $p = 95(\%)$ なら、定数=2.262 ($N = 10$ のとき)
 - N が小さいほど定数は大きく→範囲は広くなる。逆に $N \rightarrow \infty$ で定数→ 1.96

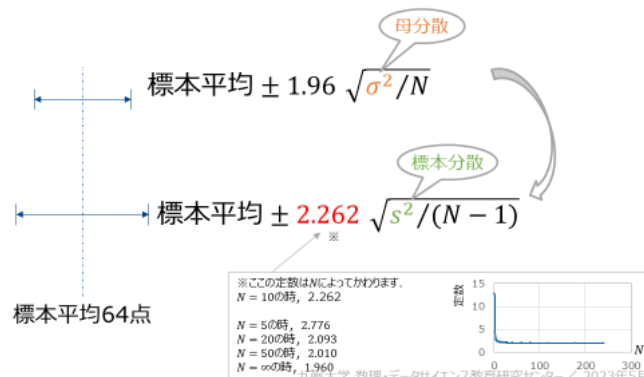
信頼区間⑤

母分散がわからない場合の詳細と t 分布

なぜこんな
対応になったのか？

Q: 母分散は不明, N も小さい. さて, どうする？

- 95% 信頼区間なら, 標本分散を使う代償に, 1.96 をもう少し大きくして, 勘弁してもらう！



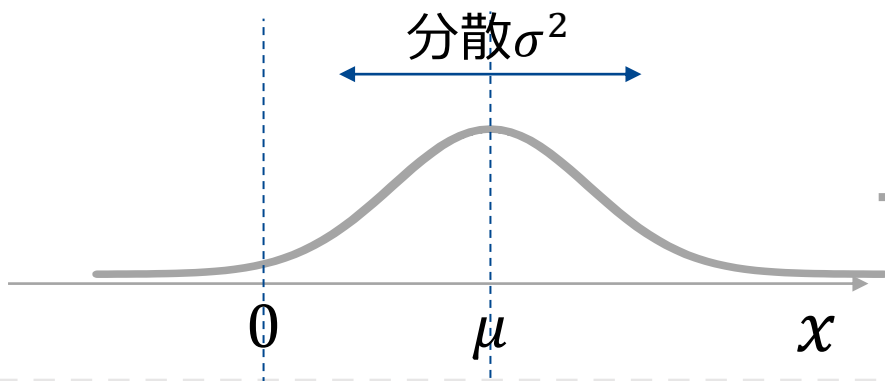
よくわからなければ, ここは読み飛ばしても何とかなる (?)

標準正規分布と標準化

- 平均0で分散1の正規分布を「標準正規分布」と呼ぶ
- x が平均 μ で分散 σ^2 の正規分布
 $\rightarrow z = \frac{x - \mu}{\sqrt{\sigma^2}}$ とすれば標準正規分布に！

任意の正規分布

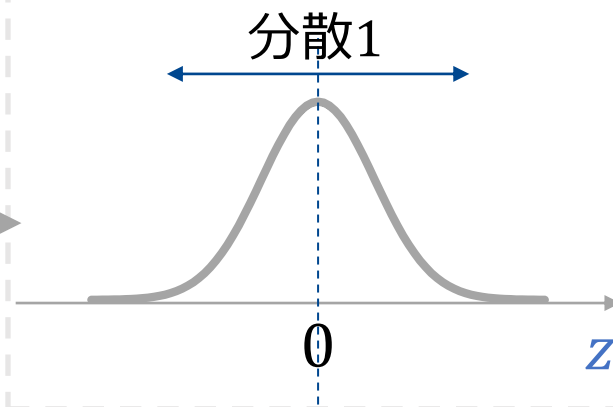
平均 μ で分散 σ^2 の正規分布



$$z = \frac{x - \mu}{\sqrt{\sigma^2}}$$

標準正規分布

平均0で分散1の正規分布



- この変換を「 $(xの)$ 標準化」と呼ぶ
- x のスケールや単位に依存しない分布となる

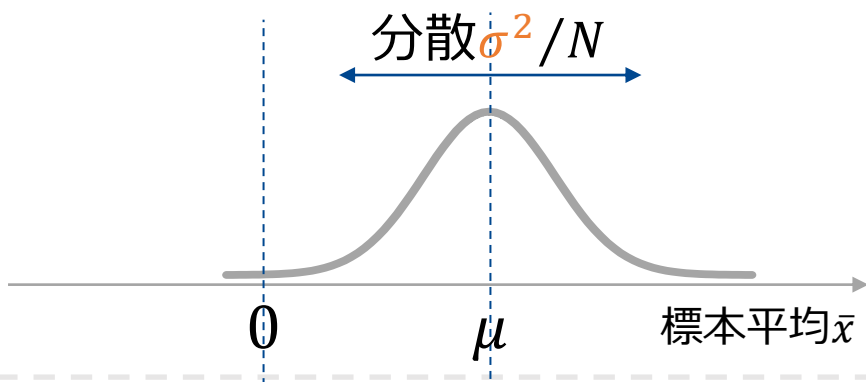
標本平均 \bar{x} の分布の標準化

- 「標本平均 \bar{x} の分布」=「平均 μ で分散 σ^2/N の正規分布」
- なので, $z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/N}}$ とすれば標準正規分布に

母分散 σ^2 を N で割ったもの

標本平均の分布

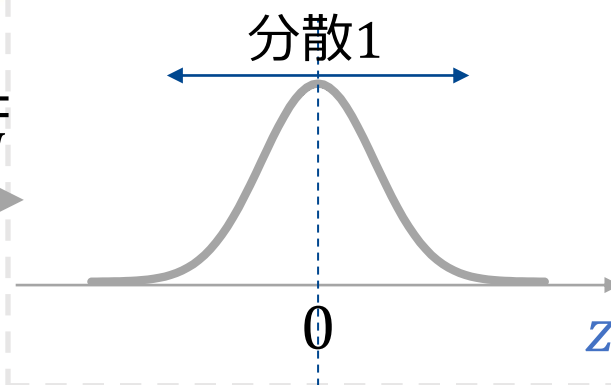
平均 μ で分散 σ^2/N の正規分布



σ^2 を用いて
 \bar{x} を z に変換
(標準化)

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/N}}$$

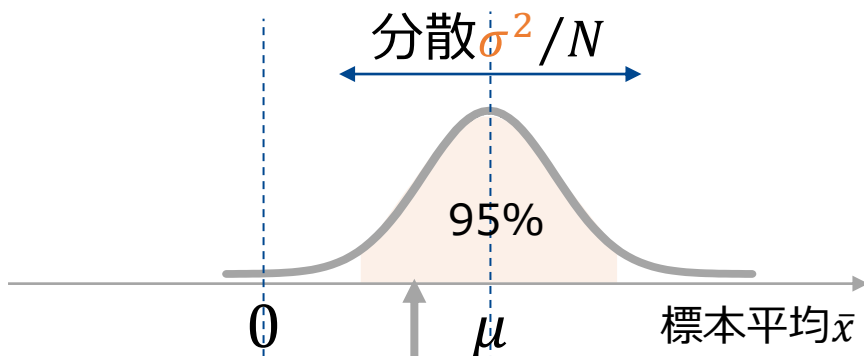
標準正規分布



標準正規分布から 母平均 μ の信頼区間を求める

標本平均の分布

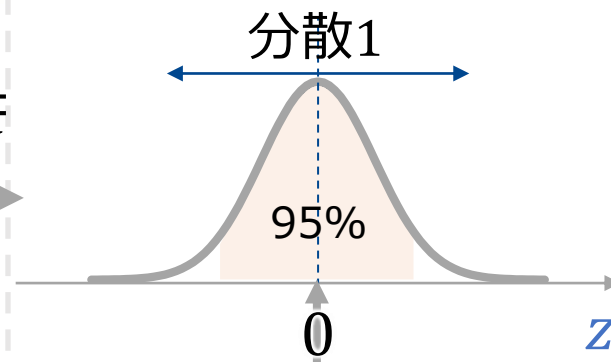
平均 μ で分散 σ^2/N の正規分布



σ^2 を用いて
 \bar{x} を z に変換
(標準化)

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/N}}$$

標準正規分布



μ の95%信頼区間は $\bar{x} \pm 1.96\sqrt{\sigma^2/N}$

ある標本から得た
標本平均 \bar{x}

z が ± 1.96 の範囲内にある
確率が95%

というわけで、信頼区間の求め方をまとめると (母分散 σ^2 がわかっている場合)

- $z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/N}}$ で標準化したと考える
- z が $p\%$ 入る範囲を,
「標準正規分布表」で調べる
 - $p = 95\%$ なら ± 1.96 であることがわかる

正規分布表 [編集]

引用元: 成美 清松、坂井 忠次『数理統計学要説』培風館、1952年。doi:10.11501/1371195。

標準正規分布 $X \sim N(0, 1)$ における確率 $P(0 \leq X \leq Z)$ の値をまとめた。

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2421	.2453	.2484	.2515	.2546
0.7	.2580	.2611	.2642	.2673	.2703	.2733	.2762	.2791	.2819	.2849
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3079	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3341	.3367	.3391
1.0	.3413	.3438	.3461	.3485	.3508	.3529	.3551	.3572	.3593	.3613
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3769	.3788	.3808	.3827
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3979	.3996	.4013
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177

標準化されているので
 σ^2 や N によらず
いつも同じ表が使える
(標準化のメリット)

Wikipedia “正規分布”より

- $z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/N}}$ が $[-1.96, 1.96]$ に95%の確率で入ることがわかる
- $\bar{x} - \mu$ が $[-1.96\sqrt{\sigma^2/N}, 1.96\sqrt{\sigma^2/N}]$ に95%の確率で入ることがわかる
- 母平均 μ の95%信頼区間が $[\bar{x} - 1.96\sqrt{\sigma^2/N}, \bar{x} + 1.96\sqrt{\sigma^2/N}]$ とわかる

さてさて、母分散 σ^2 が不明な場合、
(z への標準化ではなく) t へのスチューデント化を行う

- スチューデント化： $t = \frac{\bar{x} - \mu}{\sqrt{s^2 / (N-1)}}$

何となく標準化 $z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2 / N}}$ と似てる

標本平均の分布

平均 μ で分散 σ^2 / N の正規分布

分散 σ^2 / N

母分散は不明

0

μ

標本平均 \bar{x}

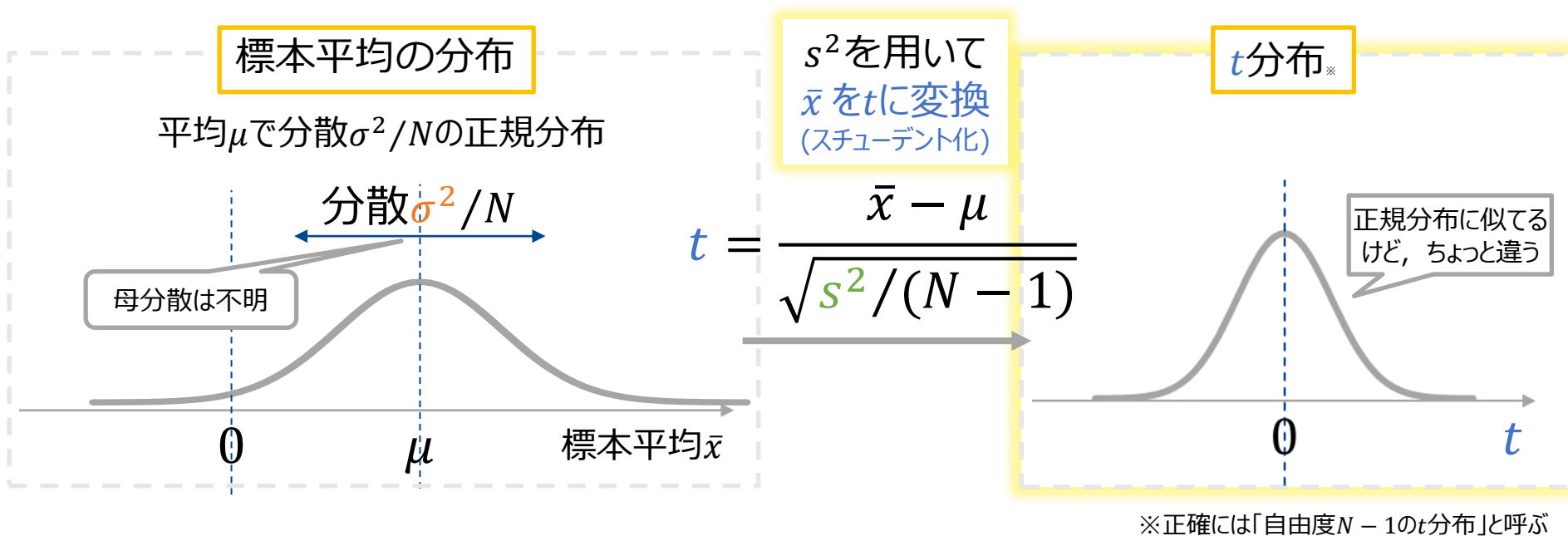
s^2 を用いて
 \bar{x} を t に変換
(スチューデント化)

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2 / (N - 1)}}$$

- すると…

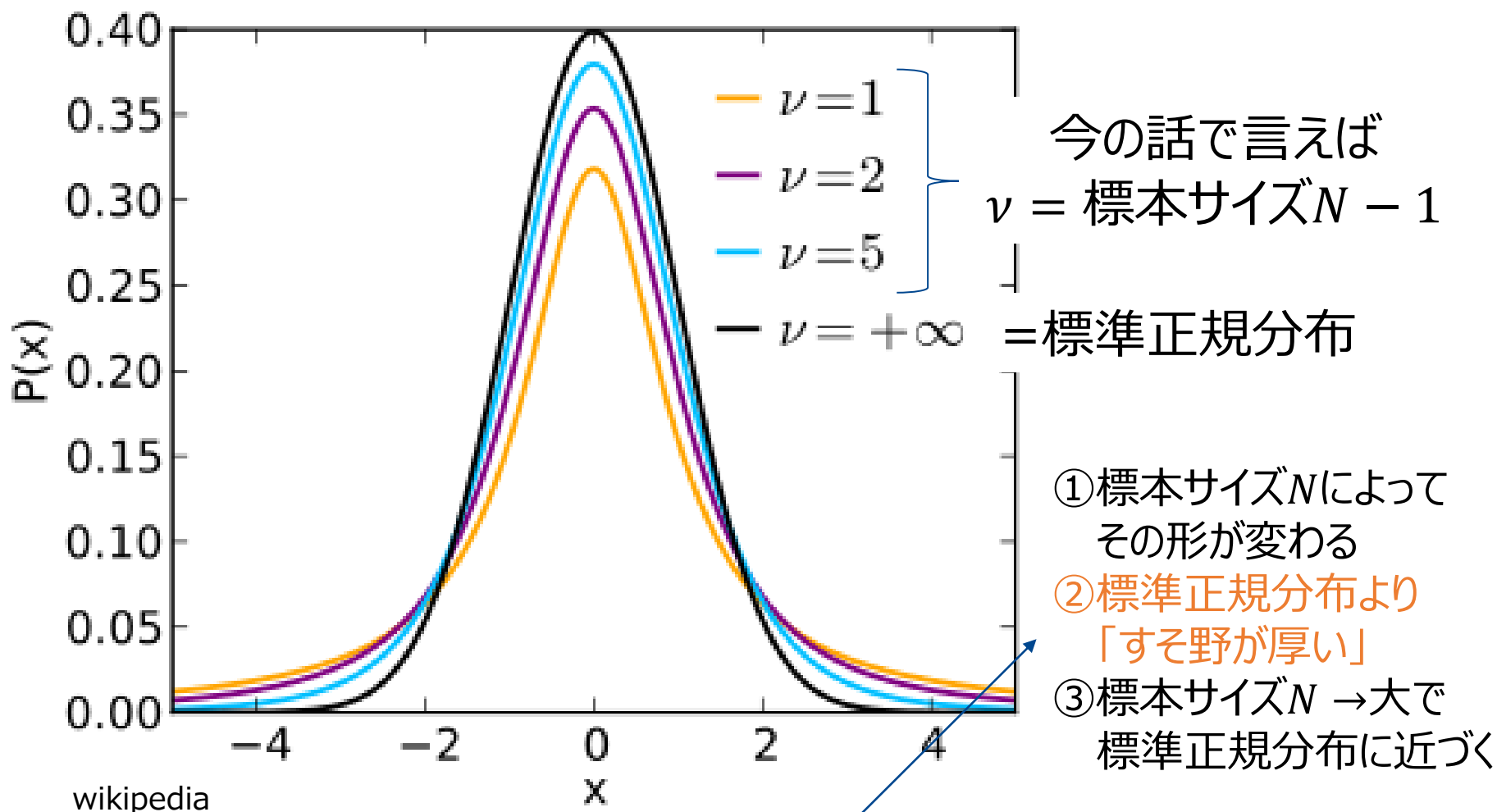
さてさて、母分散 σ^2 が不明な場合、 (z への標準化ではなく) t へのスチューデント化を行う

- t は（正規分布によく似た） **t 分布**に従うことが知られている



- t 分布の式は、正規分布以上に「複雑」なのでここでは割愛…
- 複雑具合を見てみたい人は、検索してみましょう！

t 分布は標準正規分布と似た形。
ただし、標本サイズ N に応じて形が変わる



だから区間幅が広がる！

t 分布で求めた母平均 μ の信頼区間

標本平均の分布

平均 μ で分散 σ^2/N の正規分布

分散 σ^2/N

母分散は不明

0

μ

標本平均 \bar{x}

95%信頼区間は $\bar{x} \pm 2.262\sqrt{s^2/(N-1)}$

ある標本から得た
標本平均 \bar{x}

s^2 を用いて
 \bar{x} を t に変換
(スチューデント化)

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/(N-1)}}$$

t 分布*

正規分布に似てる
けど、ちょっと違う

0

t

$N = 10$ のとき、 t が 0 ± 2.262 の範囲に入る確率は95%

というわけで、信頼区間の求め方をまとめると (母分散 σ^2 が不明の場合)

- $t = \frac{\bar{x} - \mu}{\sqrt{s^2/(N-1)}}$ でスチューデント化したと考える

- 標本数 N のとき(= 自由度 $N - 1$ のとき)
 t が $p\%$ 入る範囲を「 t 分布表」で調べる
 - $N = 10, p = 95\%$ なら ± 2.262 と書かれている

Table of selected values [\[edit \]](#)

The following table lists values for t -distributions with v degrees of freedom for a range of one-sided or two-sided percentages along the top are confidence levels, and the numbers in the body of the table are the $t_{\alpha, n-1}$ fact intervals.

The last row with infinite v gives critical points for a normal distribution since a t -distribution with infinitely many (See [Related distributions](#) above).

One-sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two-sided	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.713	0.896	1.118	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.709	0.891	1.106	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.705	0.888	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073

Wikipedia "Student's t-distribution"より

- $t = \frac{\bar{x} - \mu}{\sqrt{s^2/(N-1)}}$ が $[-2.262, 2.262]$ に95%の確率で入ることがわかる
- $\bar{x} - \mu$ が $[-2.262\sqrt{s^2/(N-1)}, 2.262\sqrt{s^2/(N-1)}]$ に95%の確率で入ることがわかる
- 母平均 μ の95%信頼区間が $[\bar{x} - 2.262\sqrt{s^2/(N-1)}, \bar{x} + 2.262\sqrt{s^2/(N-1)}]$ とわかる

統計的検定

「平均の差の検定」を例に

母平均は違うか？

母集団A

違う？



母集団B

例えば...



全非喫煙者の
寿命データA



違う
平均寿命？

全喫煙者の
寿命データB



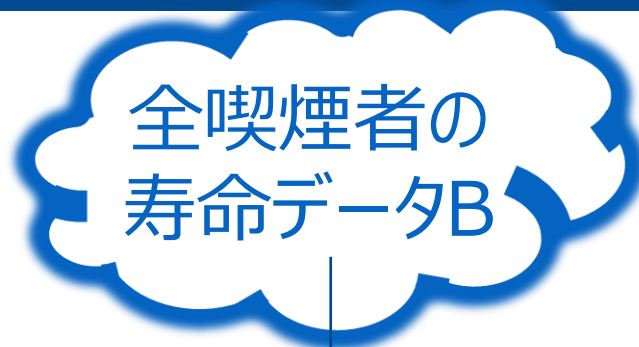
とはいえ、標本平均しか手に入らないわけで…



全非喫煙者の
寿命データA



違う
平均寿命？



全喫煙者の
寿命データB

手に入った
標本平均 \bar{x}_A

手に入った
標本平均 \bar{x}_B

じゃあ、

$$\bar{x}_A \neq \bar{x}_B$$

となれば「違う平均」かな…



でも、標本平均って揺らぎますよね…



全非喫煙者の
寿命データA



違う
平均寿命？

全喫煙者の
寿命データB



手に入った
標本平均 \bar{x}_A

手に入った
標本平均 \bar{x}_B

そうか、そうすると

$\bar{x}_A \neq \bar{x}_B$ となっただけでは
本当に違うのか判断できないよね…



うーん、だとすれば、差がある程度より大きければ「違う」とするしかないか…



全非喫煙者の
寿命データA



違う
平均寿命？

全喫煙者の
寿命データB



手に入った
標本平均 \bar{x}_A

手に入った
標本平均 \bar{x}_B

$|\bar{x}_A - \bar{x}_B| > \text{ある値}$

…でもこの「ある値」って
どうやって設定するの？



実は



全非喫煙者の
寿命データA



違う
平均寿命？

全喫煙者の
寿命データB



手に入った
標本平均 \bar{x}_A

手に入った
標本平均 \bar{x}_B

\bar{x}_A と \bar{x}_B にどれぐらい差があれば、
どれぐらい自信をもって「平均は
同じじゃない！」と言えるか、
実はいい方法があるんです！





統計的検定①

検定の考え方と 基本的手順



用語が面倒だったりしますが...

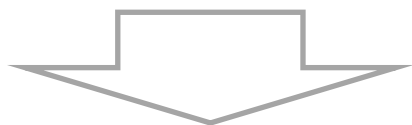
統計的検定とは何で、いつ使うのか？

- 「AとBは差がある」と統計的に言いたいときに使う
- 事例
 - 東京と福岡で平均所得に差があるか？
 - あるトレーニングの有無で能力に差が出たか？
 - あるダイエット食品に効果があるか？
 - ダイエット食品と食べた群と食べなかった群に差があるか？
 - ある遺伝子がある病気の原因になるか？
 - 「野生型」と「その遺伝子をノックアウトした型」で、病気の罹患率に差があるか？

統計的検定の基本手順 (1/3)

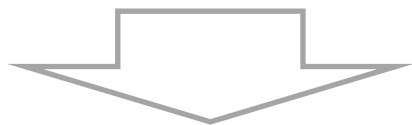
● 帰無仮説の設定

- 言いたいことと反対の仮説(「AとBに**差はない**」)を立てる



● 仮説の下で矛盾が起こることを証明

- = そもそも仮説が間違っていた



● 帰無仮説(「AとBに**差はない**」)を棄却

- 「差がない」ってのは間違い



喫煙者と非喫煙者の
寿命に差はないと考える



差がないなら、
この標本平均の差は
なんかおかしい



喫煙者と非喫煙者の
寿命に差はないと考える

なんかすごく回りくどく見える
最初から「差がある」ことを示せばいいのでは？



統計的検定の基本手順 (2/3)

回りくどく見えますが...

- 「差がある」ことを直接証明するよりも
- 「差がない」ことを前提にして矛盾を導くほうが楽
- いわゆる「背理法」
 - 「 $\sqrt{2}$ が有理数でない」ことを直接証明するよりも
 - 「 $\sqrt{2}$ が有理数である」ことを前提にして矛盾を導くほうが楽
- 1つでも矛盾が見つければ, それで否定できる点がありがたい！



正面から試行錯誤しながら追及するより、
容疑者に自分の正しさを証明させながら、
そのほころびを(1つでも)見つけたほうが早い

統計的検定の基本手順 (3/3)

もうちょっと詳細

- 帰無仮説の設定
 - 言いたいことと反対の仮説(「AとBに差はない」)を立てる
- 検定統計量の値を計算
 - 例えば標本平均の値
- 確率の計算
 - 上記の値が「どの程度起こりうるものなのかどうか」
- 仮説の判定
 - そもそも仮説が間違っていた
- 帰無仮説(「AとBに差がない」)を棄却

喫煙者と非喫煙者の
寿命に差はないと考える



喫煙者と非喫煙者
それぞれの
標本平均を得る

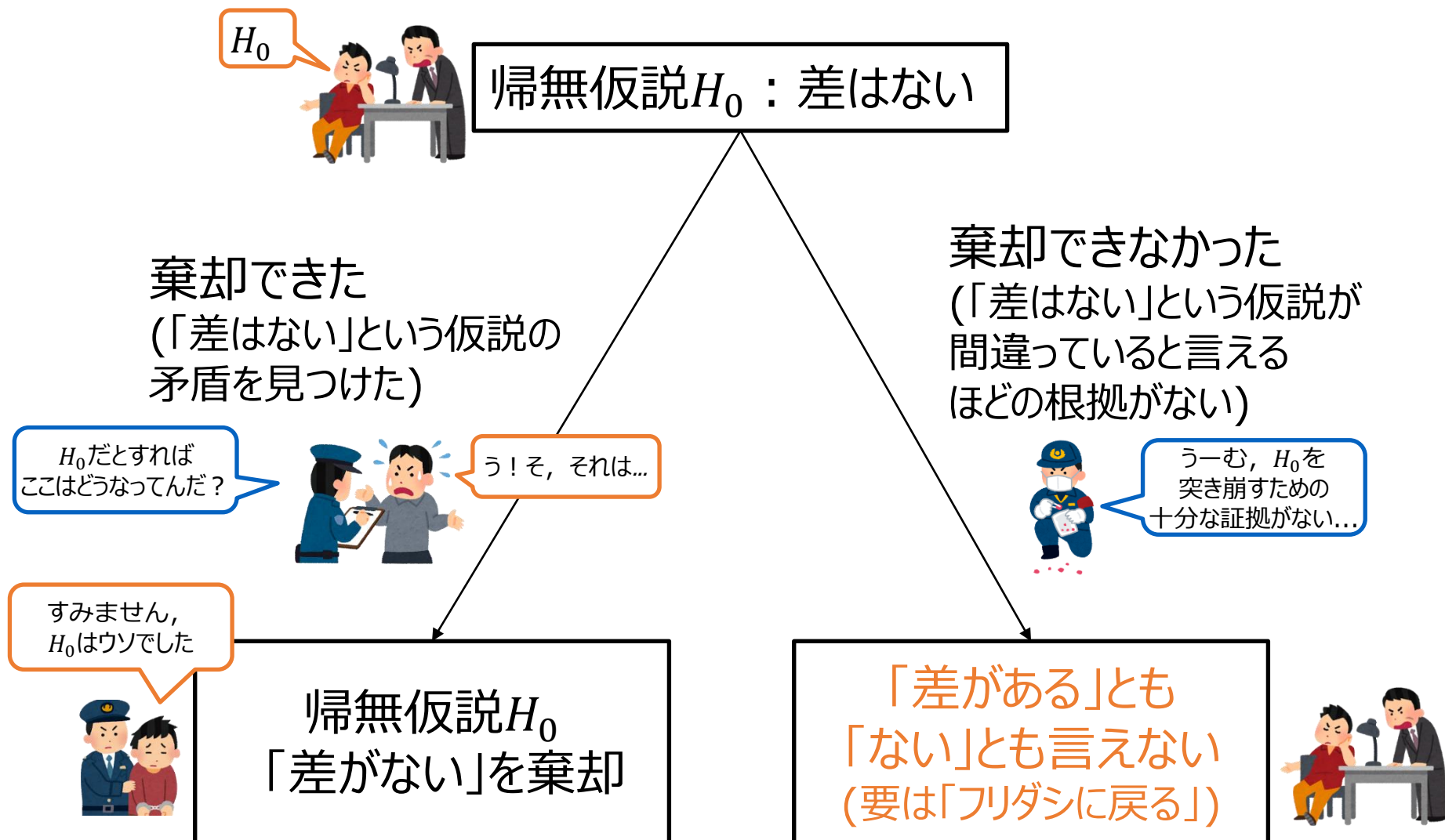


差がないのに、標本平均の差がこんなに大きくなるのは、珍しすぎ



差がないと考えたのが
間違いだった

帰無仮説の矛盾が見つからなかった (=棄却できなかった)場合の考え方

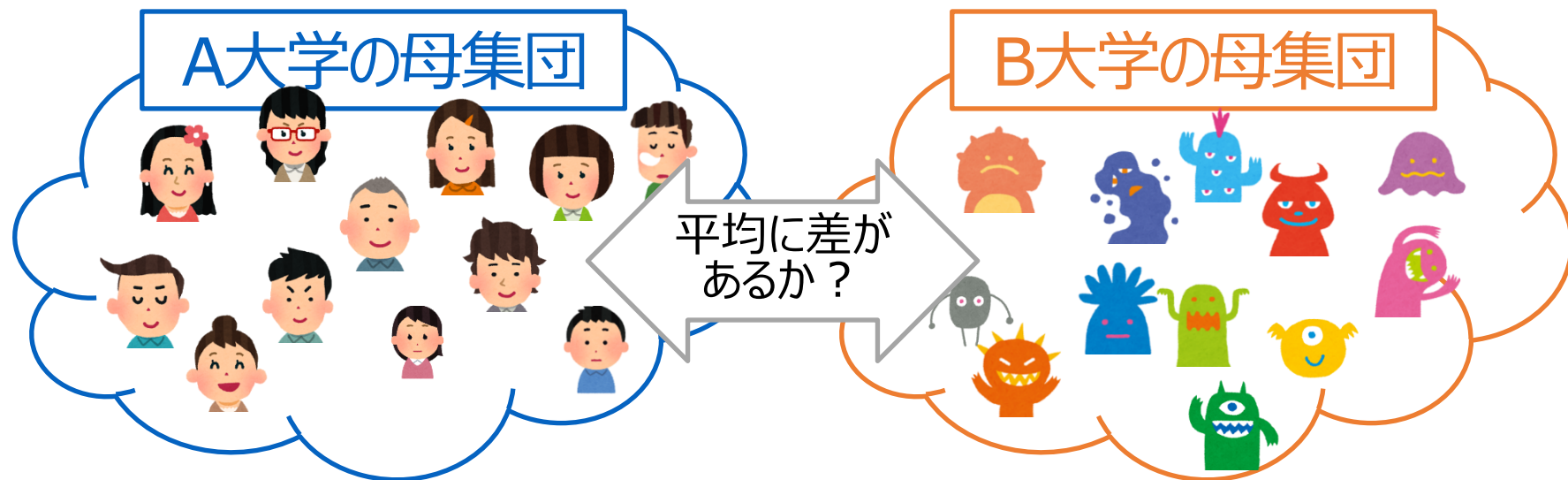


統計的検定②

2群の平均の差の検定： 目的と攻め方

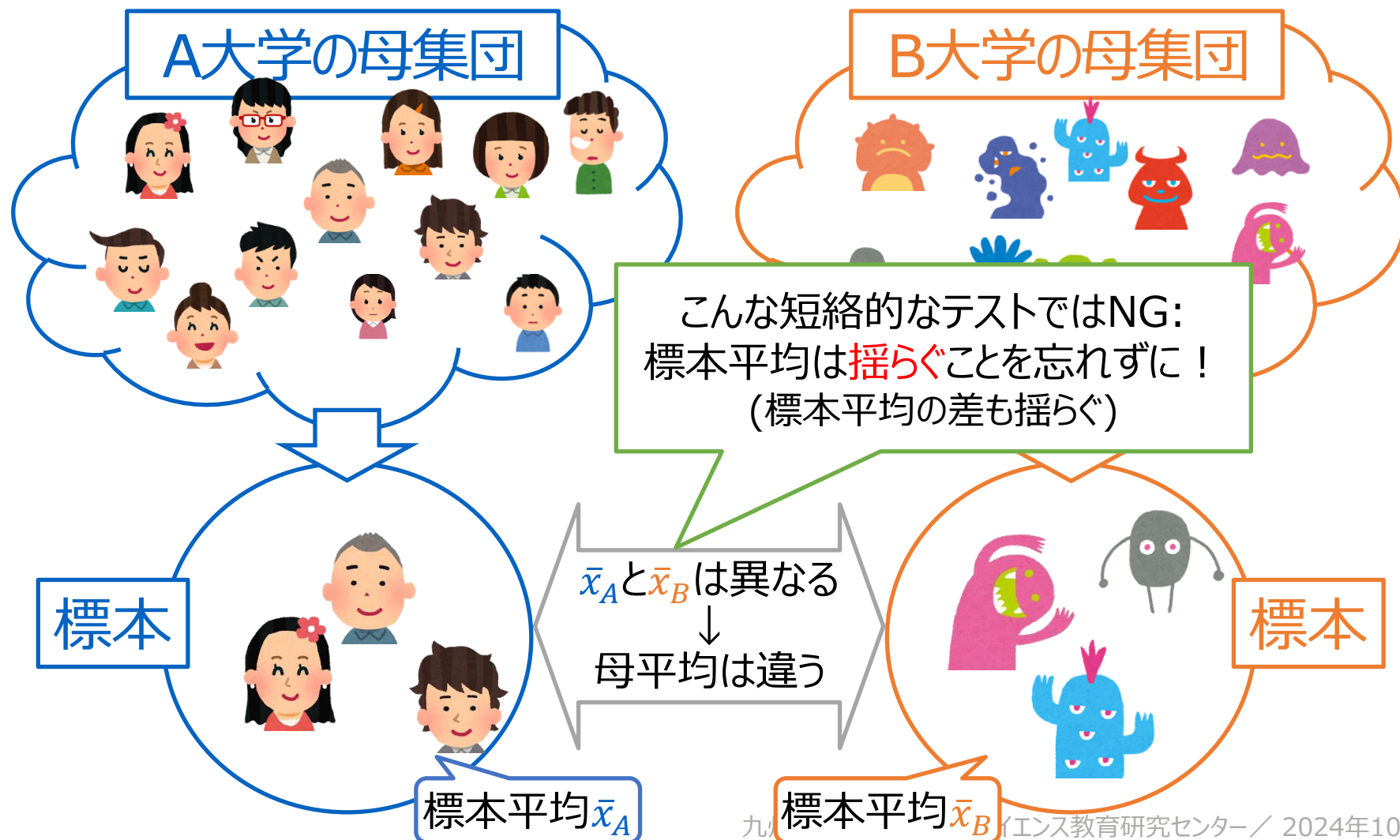
二つの母集団の母平均に差があるかを，標本の平均を利用して検定！
最もよくある検定問題

もっと一般的な検定課題：平均の差の検定 (A大学もB大学も母平均は不明)

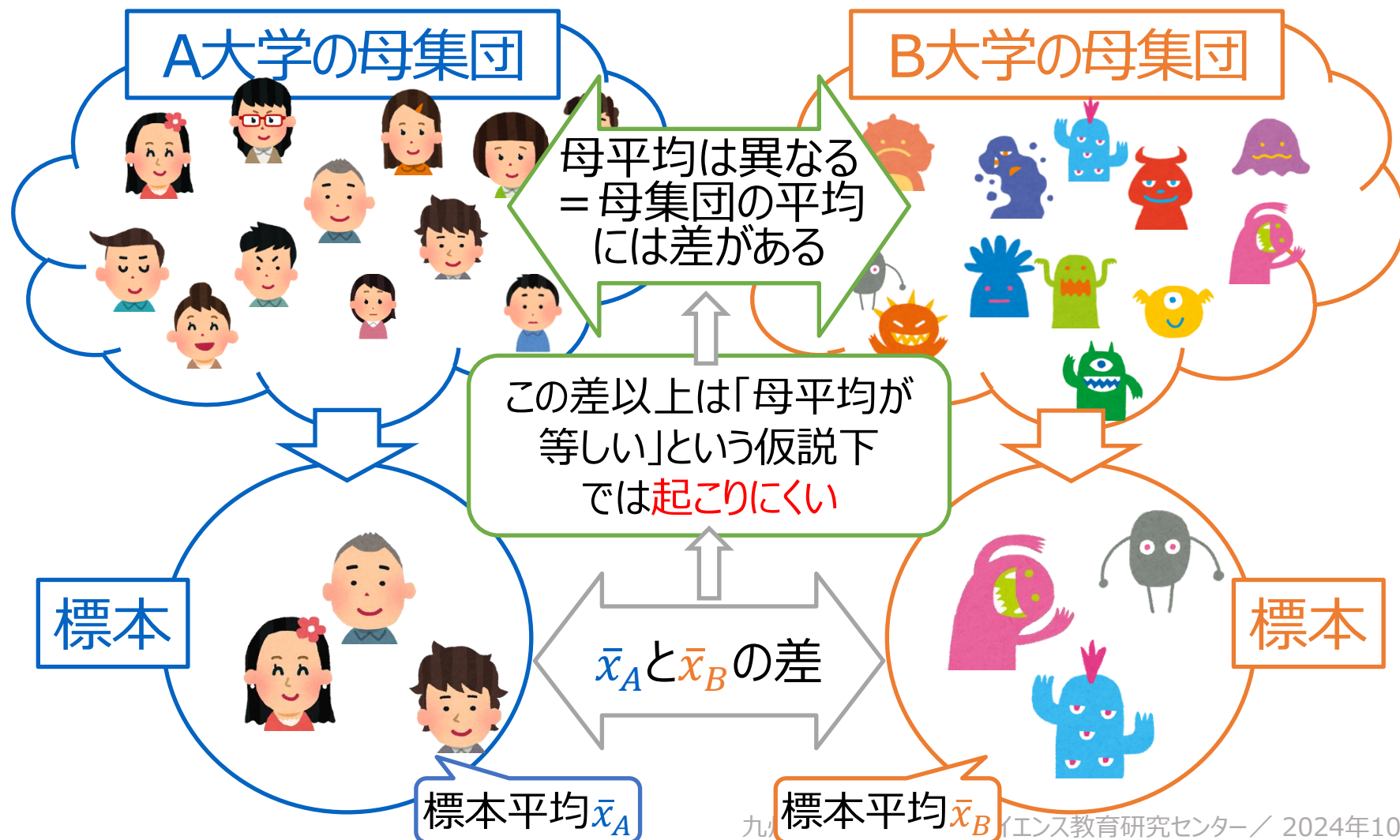


- 先に述べた以下の検定課題は，すべてこのタイプ
 - 東京と福岡で平均所得に差があるか？
 - あるトレーニングの有無で差が出たか？
 - あるダイエット食品に効果があるか？
 - ある遺伝子がある病気の原因になるか？

平均の差の検定～ダメな方法



平均の差の検定～実際の攻め方



統計的検定③
2群の平均の差の検定：
2群の母分散 σ^2 が既知の場合

説明を簡単にするため、まず色々と仮定します

A大学の母集団



B大学の母集団



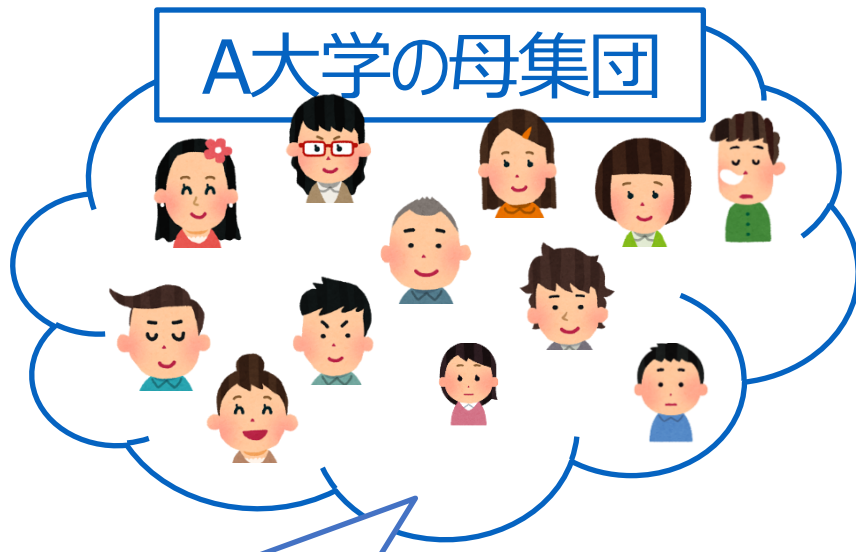
話を簡単にするための仮定

- これまでと同様, 「母集団 A, B の点数は, 共に正規分布」
 - 「正規性の仮定」と呼ばれる
- さらに, 母集団 A, B からの標本数は共に N で等しい
- さらに, 母集団 A, B の分散は共に σ^2 で等しい
 - すなわち, $\sigma_A^2 = \sigma_B^2 = \sigma^2$
 - 「等分散性の仮定」と呼ばれる
- さらに, ここ③では, 母分散 σ^2 の値がわかっていることも仮定
 - 次スライドで図示

ここでは、「母分散 σ^2 の値は既知」

単に等しいことだけでなく、具体的な値までわかっている

A大学の母集団



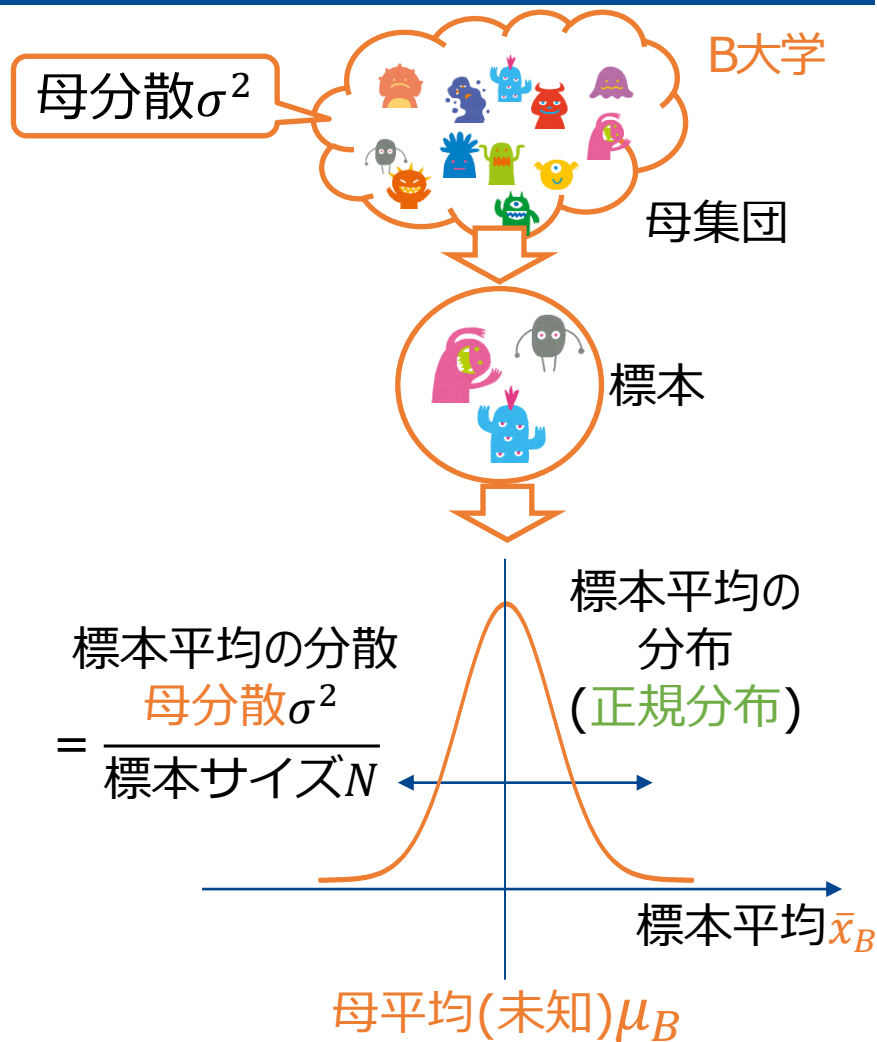
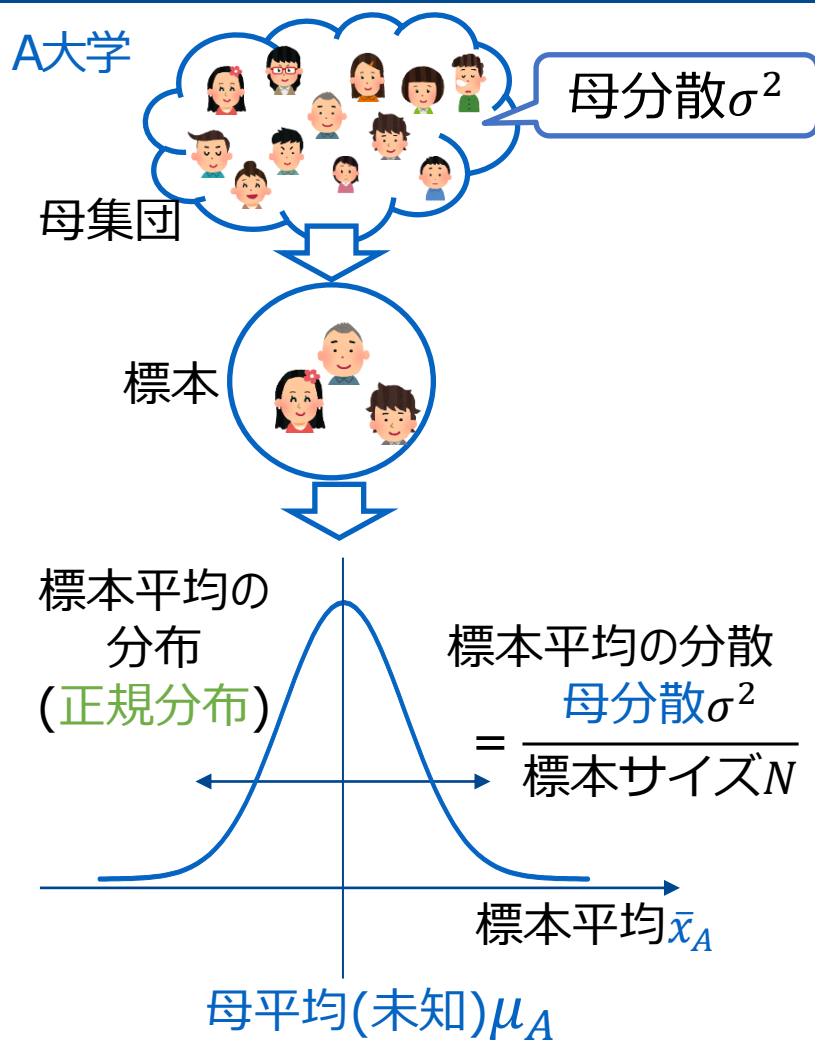
母平均 μ_A は未知だが
母分散 σ^2 は既知

B大学の母集団



母平均 μ_B は未知だが
母分散 σ^2 は既知

先述の通り，標本平均の分布は正規分布



標本平均の分布から 標本平均の「差」の分布もわかる

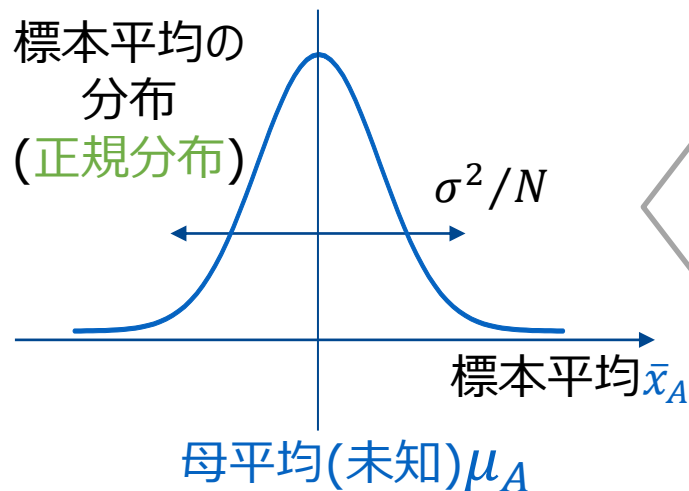
A大学

B大学

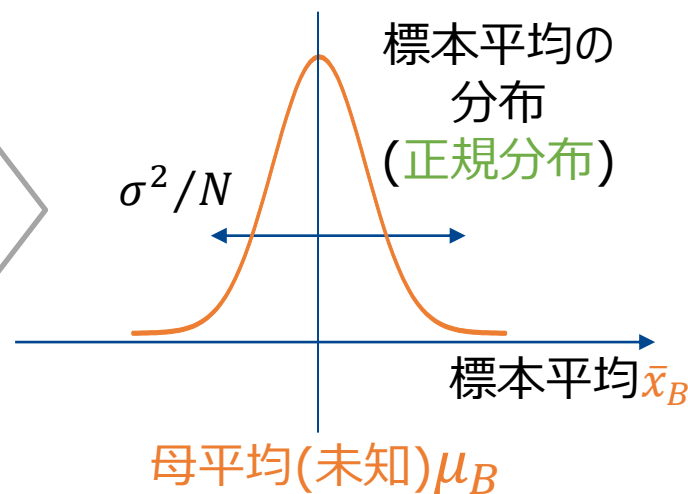


どんな分布？

標本平均の差 $\bar{x}_A - \bar{x}_B$



\bar{x}_A と \bar{x}_B の差



標本平均の分布から 標本平均の「差」の分布もわかる

A大学

B大学

扱いやすい正規
分布で助かったー



標本平均の
差の分布
(正規分布)

$$2\sigma^2/N$$

$$\mu_A - \mu_B$$

標本平均の差 $\bar{x}_A - \bar{x}_B$

標本平均の
分布
(正規分布)

$$\sigma^2/N$$

標本平均 \bar{x}_A

母平均(未知) μ_A

\bar{x}_A と \bar{x}_B の差

標本平均の
分布
(正規分布)

$$\sigma^2/N$$

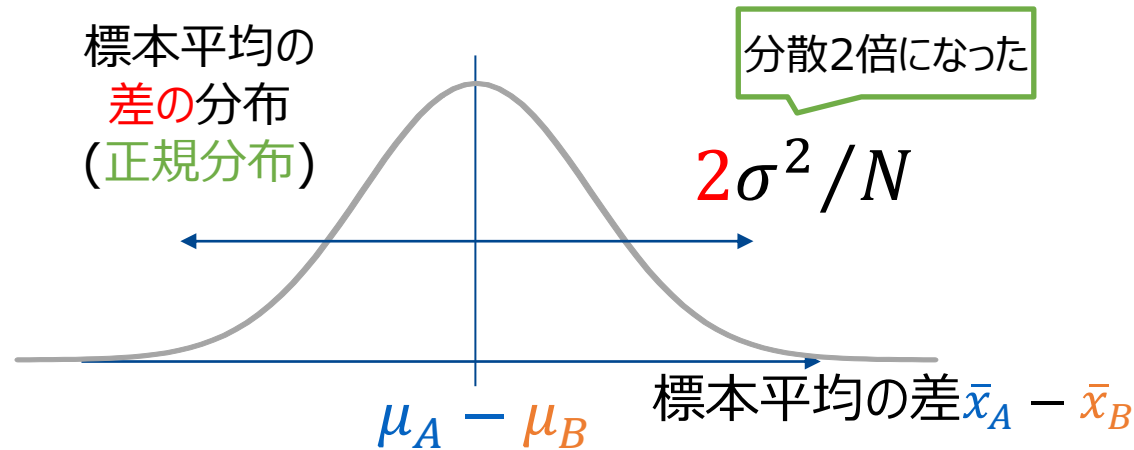
標本平均 \bar{x}_B

母平均(未知) μ_B

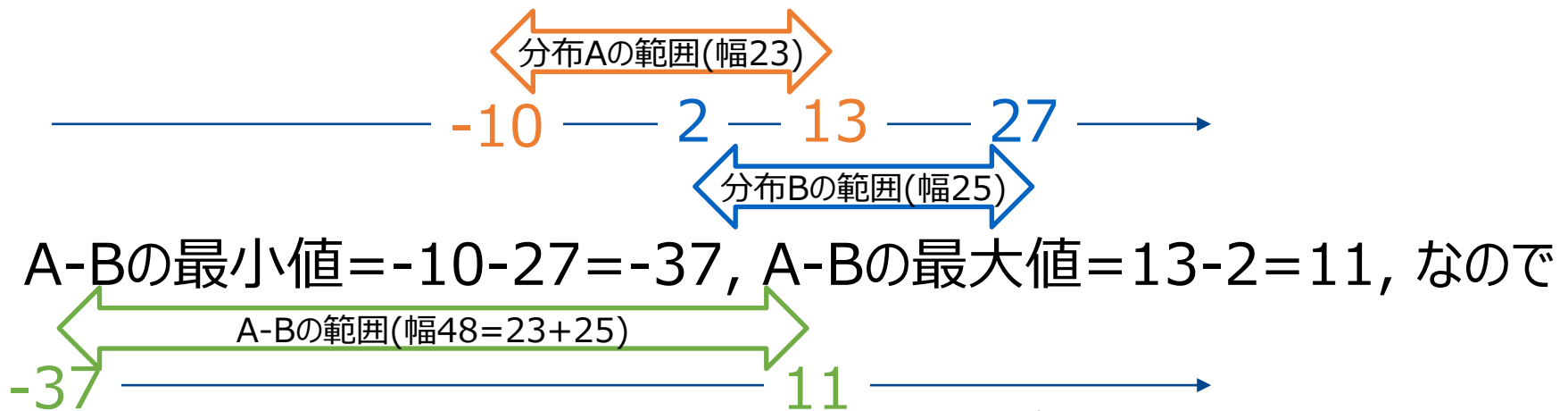
ちょっと解説： 差の分布，分散は大きくなります

A大学

B大学



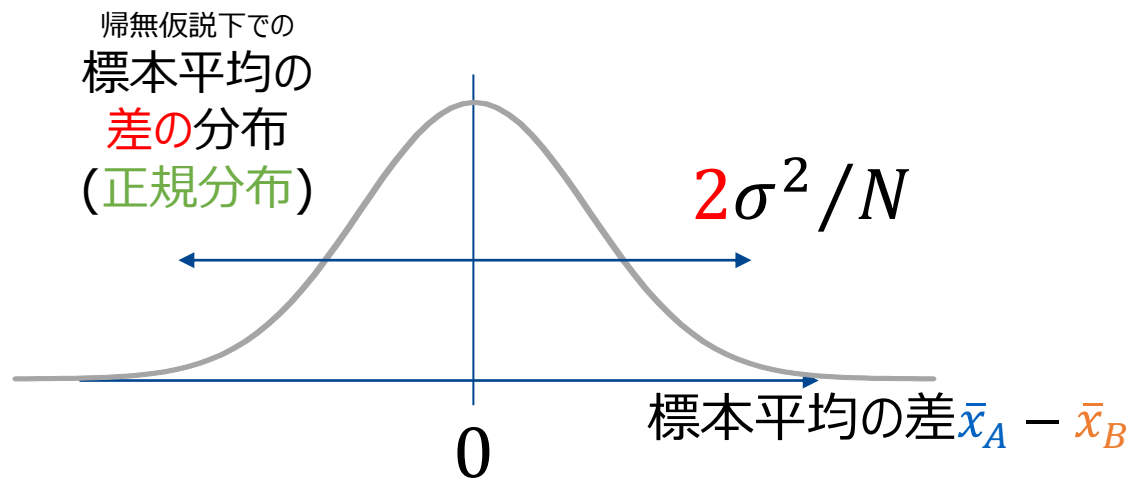
なぜ差の分散は大きくなる？ (一様分布を用いた，非常に) 直観的な説明



さあ、いよいよ帰無仮説： 「2大学の母平均には差がない」＝母平均は等しい！

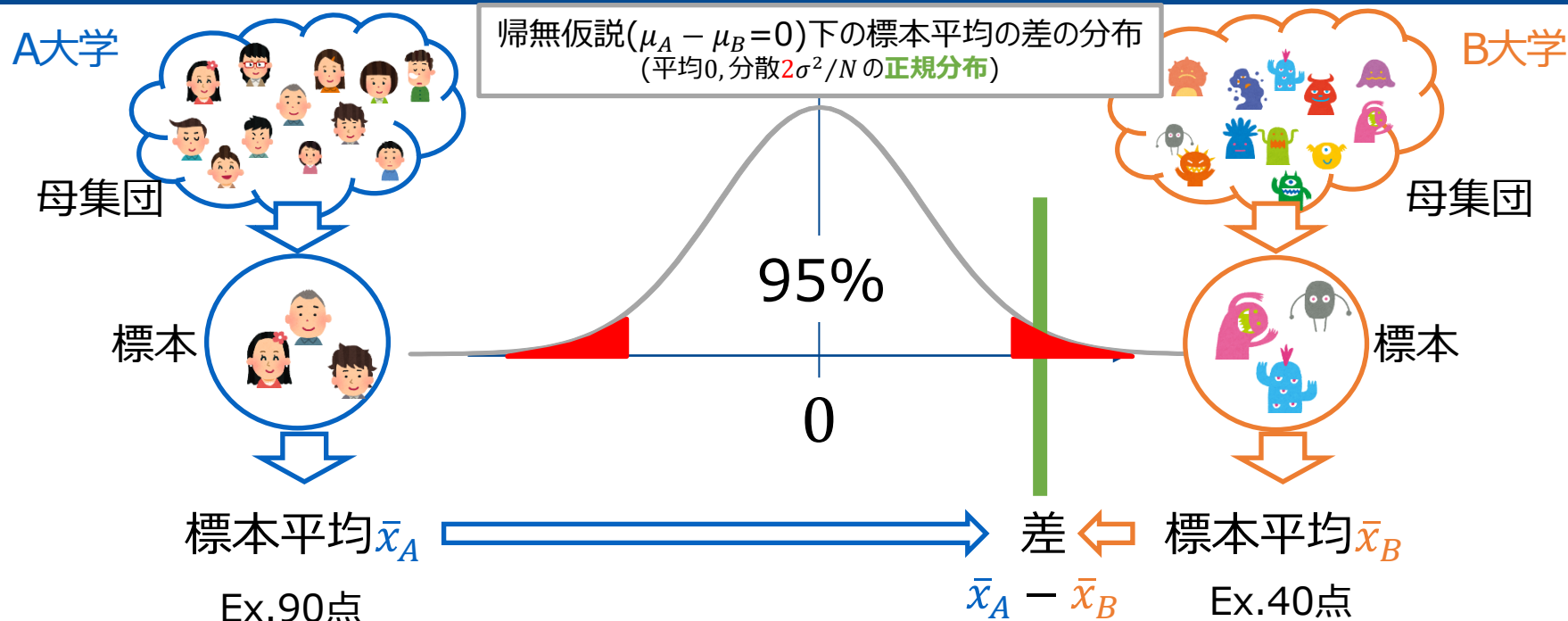
A大学

B大学



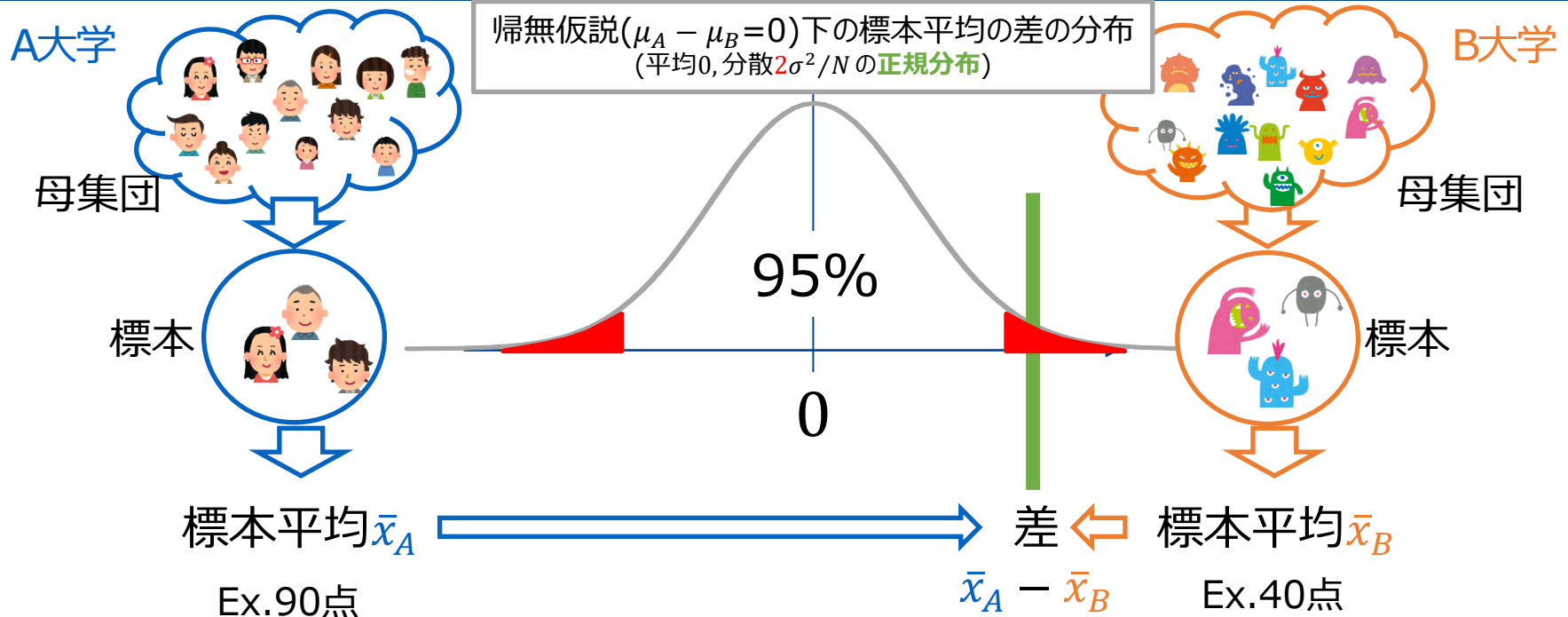
母平均は等しい，すなわち
 $\mu_A = \mu_B$ という仮説の下では
 $\mu_A - \mu_B = 0$
 (ちょうどいい具合に未知項 μ_A, μ_B がなくなった!)

実際の差が分布のどこに来るかをテスト！ 果たして帰無仮説は棄却されるか？



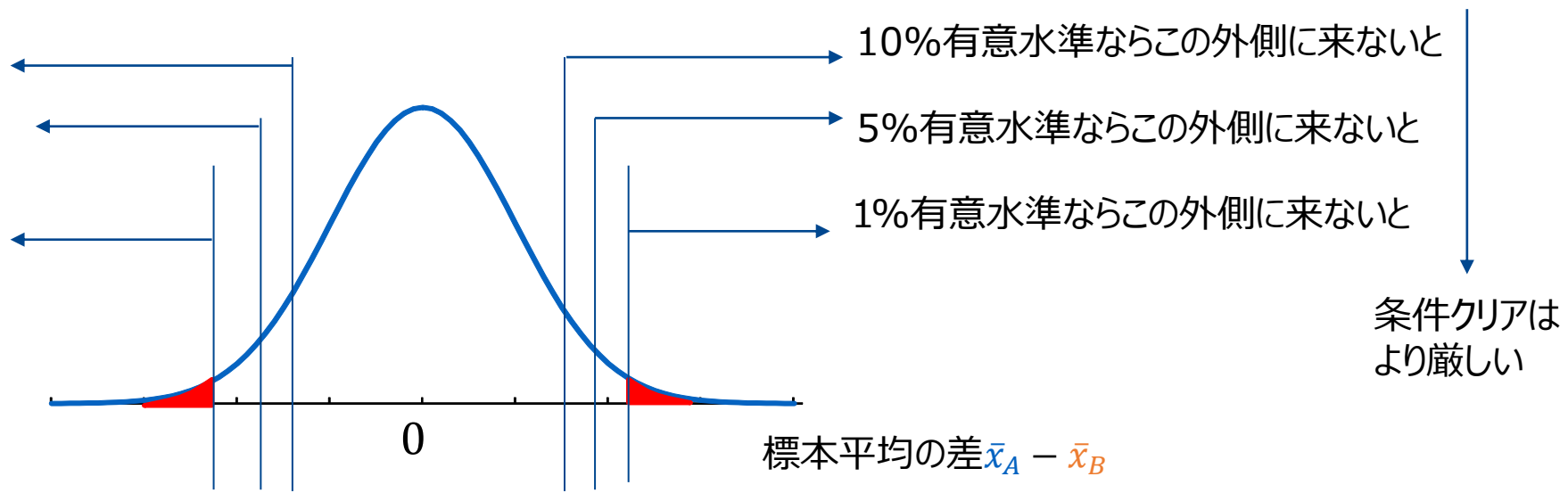
上図のようになれば(めったに起きない程度の差が生じているので)
帰無仮説「差はない」は棄却, **有意水準5%で「差がある」**

どういう場合に「『差はない』帰無仮説」は 棄却されやすいか？



- 場合1: \bar{x}_A と \bar{x}_B の差(の絶対値)が大きい
- 場合2: 母分散 σ^2 が小さい
- 場合3: 標本サイズ N が大きい

有意水準はどうやって決める？



- 例えば命に係わるような話で、絶対安心して「差がある」と言いたい場合ほど、厳しく（有意水準を小さく. 5%とかじゃなく, 1%, もしくはそれ以下）
- 厳しくなるほど、大きな差が要求される
- 分野によっては、ある程度定められている場合も

統計的検定④
2群の平均の差の検定：
母分散 σ^2 がわからない場合

A大学の母集団



B大学の母集団



話を簡単にするための仮定

- これまでと同様, 「母集団A, Bの点数は, 共に正規分布」
 - 「正規性の仮定」と呼ばれる

- さらに, 母集団A, Bからの標本数は共に N で等しい

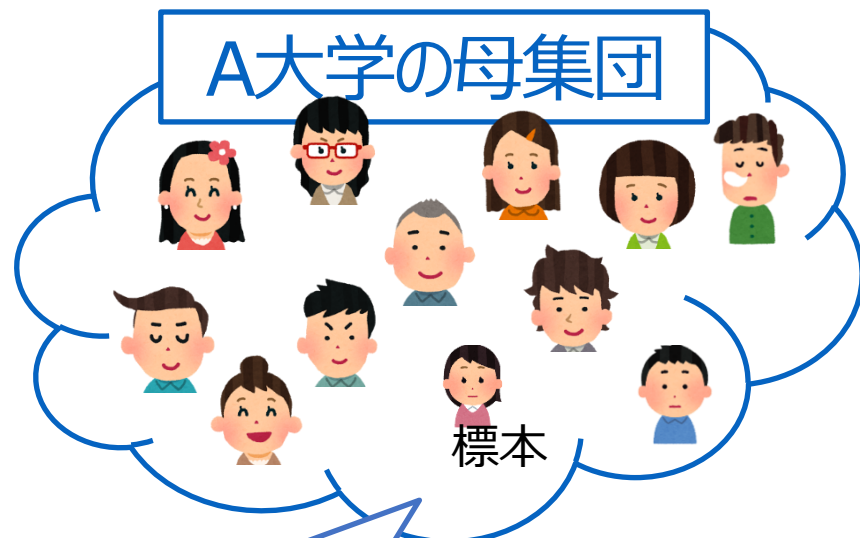
この仮定は,
簡単な変更で
なくすことができます
→【付録2】

- さらに, 母集団A, Bの分散は共に σ^2 で等しい
 - すなわち, $\sigma_A^2 = \sigma_B^2 = \sigma^2$
 - 「等分散性の仮定」と呼ばれる

● ~~さらに, ここ③では, 母分散 σ^2 の値がわかっていることも仮定~~

- ここ④は, 母分散の値まではわからないという, より一般的な状況

母分散がわからない。ピンチか？ (1/3)

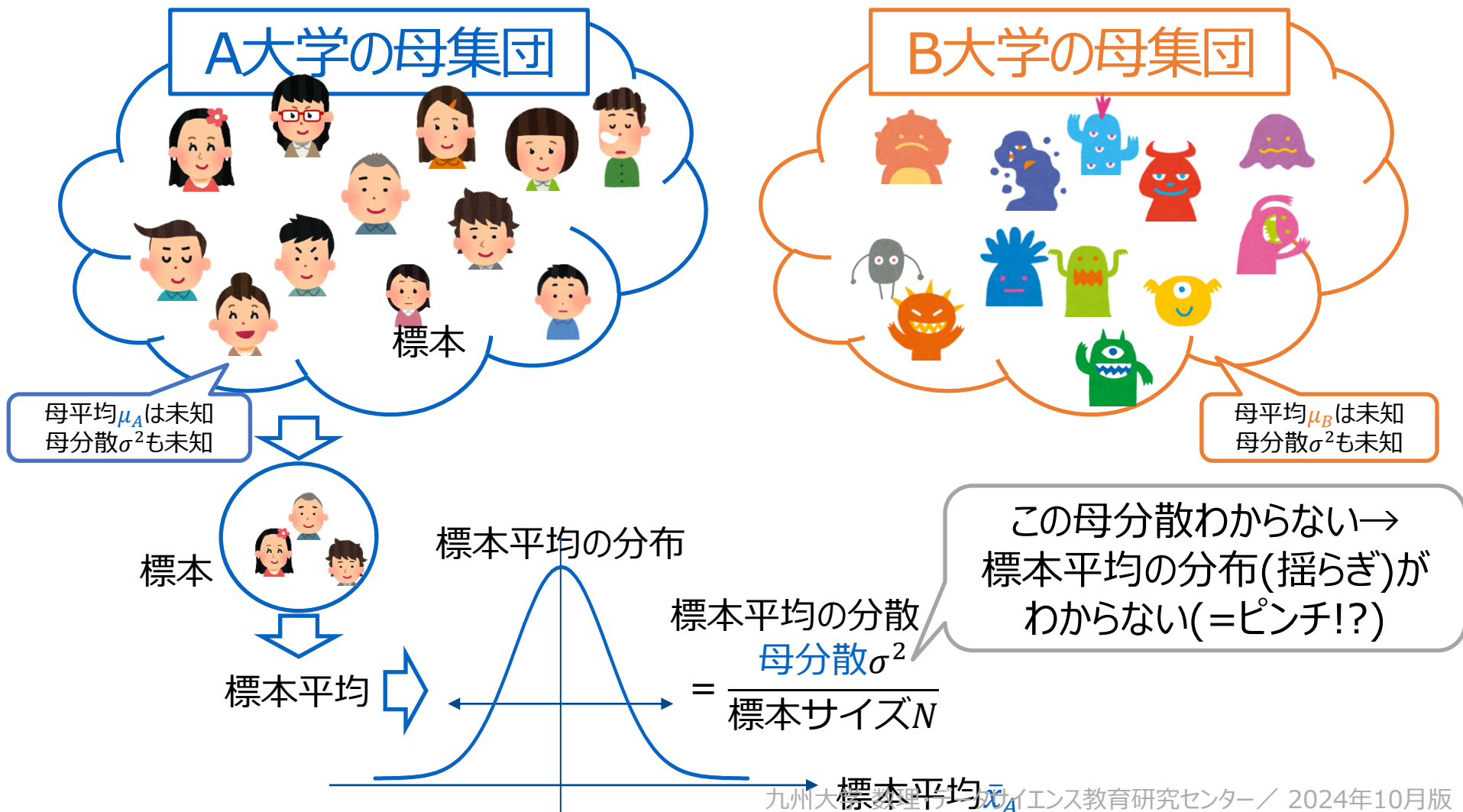


母平均 μ_A は未知
母分散 σ^2 も未知

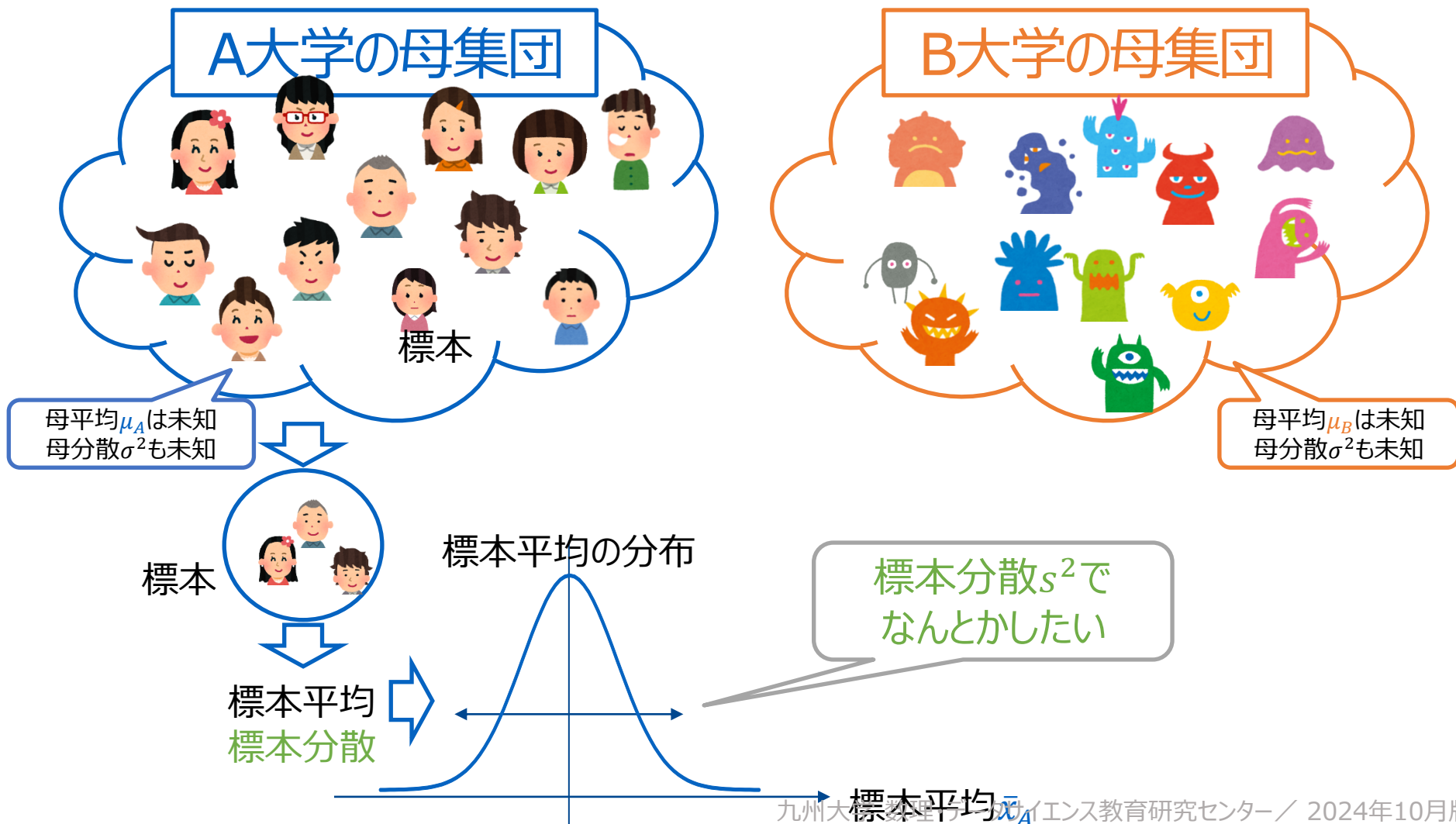


母平均 μ_B は未知
母分散 σ^2 も未知

母分散がわからない。ピンチか？ (2/3)



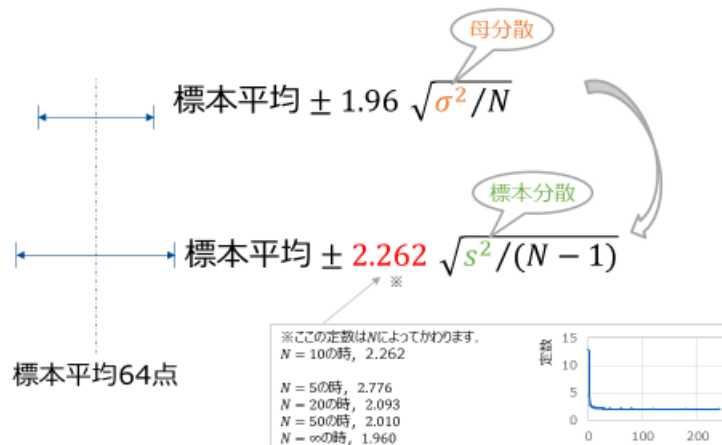
母分散がわからない。ピンチか？ (3/3)



ん？ 似た話，どっかで聞いた！？

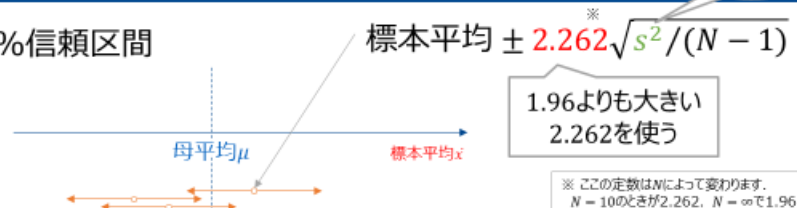
Q: 母分散は不明, N も小さい. さて, どうする？

- 95%信頼区間なら, 標本分散を使う代償に, 1.96をもう少し大きくして, 勘弁してもらう！

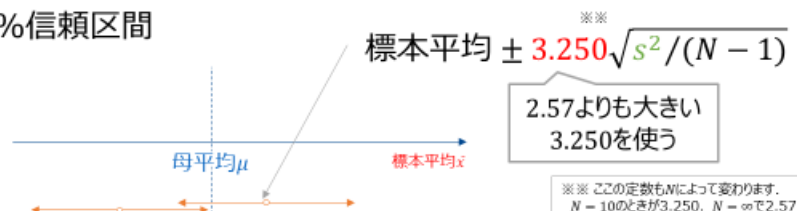


「母分散も不明, N も小さい」場合の
95%信頼区間と99%信頼区間

- 95%信頼区間



- 99%信頼区間



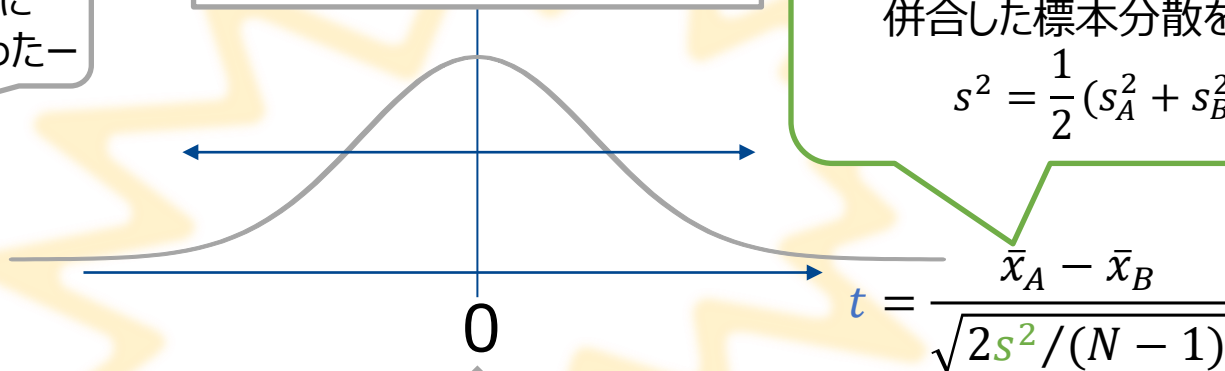
母分散不明でも，母平均に差があるかを検定したい： t 分布を使えば検定可能！（ t 検定）

母分散がわからない場合

信頼区間と同様に
 t 分布が使えてよかったー



帰無仮説($\mu_A - \mu_B = 0$)下の t の分布
(自由度 $2N - 2$ の t 分布)

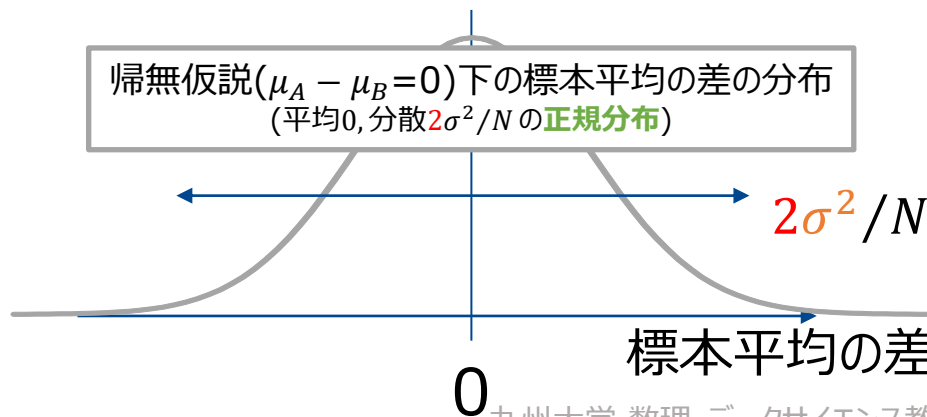


群AとBの各標本分散 s_A^2 と s_B^2 を
併合した標本分散を利用

$$s^2 = \frac{1}{2}(s_A^2 + s_B^2)$$

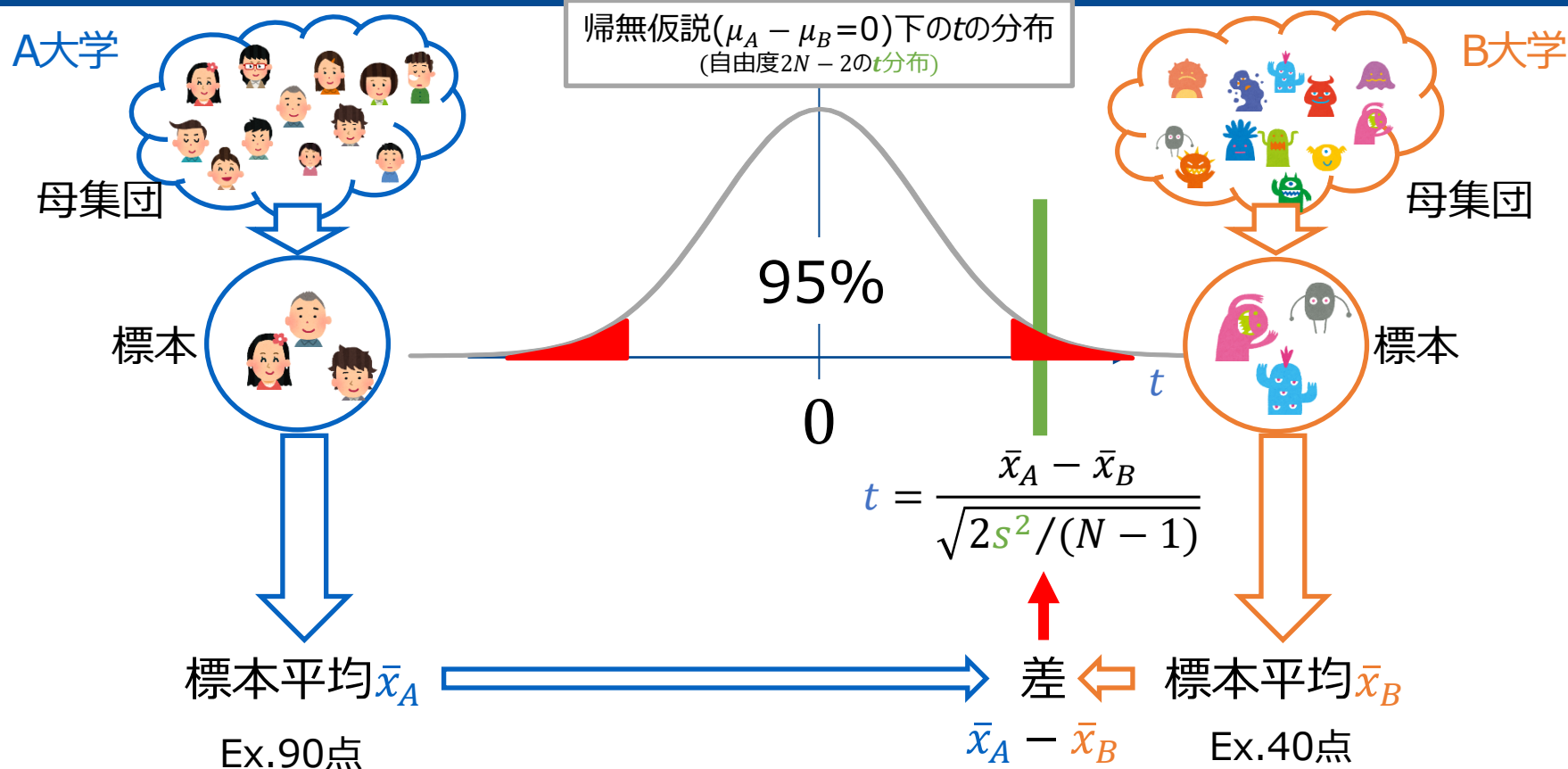
スチューデント化

帰無仮説($\mu_A - \mu_B = 0$)下の標本平均の差の分布
(平均0, 分散 $2\sigma^2 / N$ の正規分布)



母分散がわかる場合

というわけでは母分散既知の場合と同様、
帰無仮説を棄却できればOK



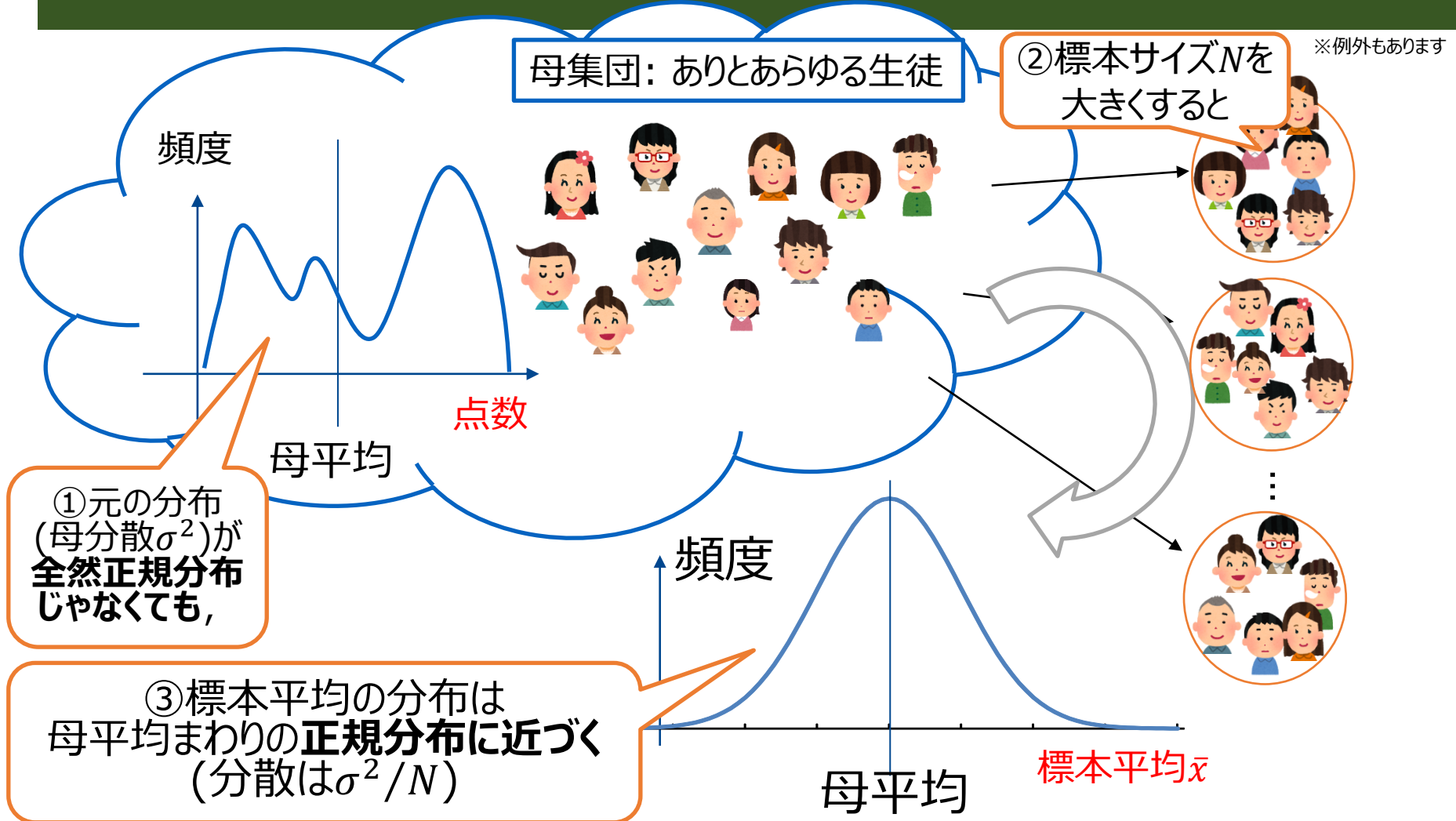
上図のようになれば(めったに起きない程度の差が生じているので)
帰無仮説「差はない」は棄却, **有意水準5%で「差がある」**

※ t 分布は標準正規分布よりつぶれているので、
同じ有意水準のためには母分散既知の場合より大きな差が必要(特に N が小さいとき)

【付録1】 中心極限定理

標本サイズ N が大きくなると,
母集団の分布が正規分布でなくても
標本平均の分布は正規分布に近づく！

面白い事実: 標本サイズ N が大きくなると, 元の分布が
どんな形※でも, 「標本平均の分布」は正規分布に近づく!



なんだかスゴイこの事実を「中心極限定理」と呼びます

【付録2】 2群の標本サイズが異なる場合の t 検定

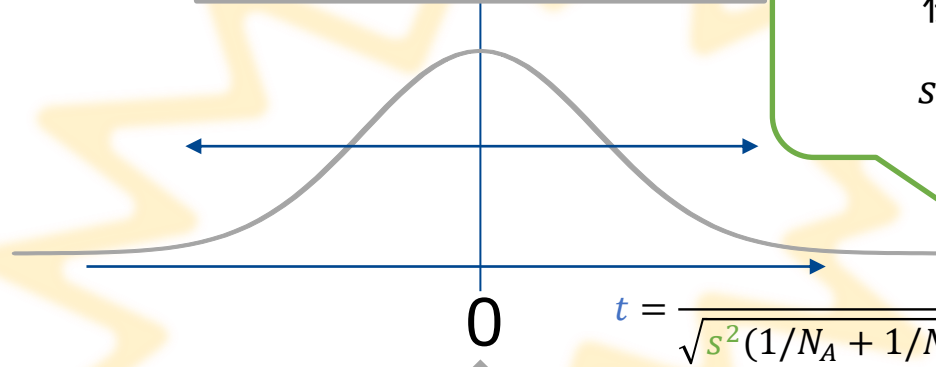
2群で標本サイズが異なる場合

A群の標本サイズ N_A , B群は N_B

母分散がわからない場合

母分散がわかる場合

帰無仮説($\mu_A - \mu_B = 0$)下の t の分布
(自由度 $N_A + N_B - 2$ の t 分布)



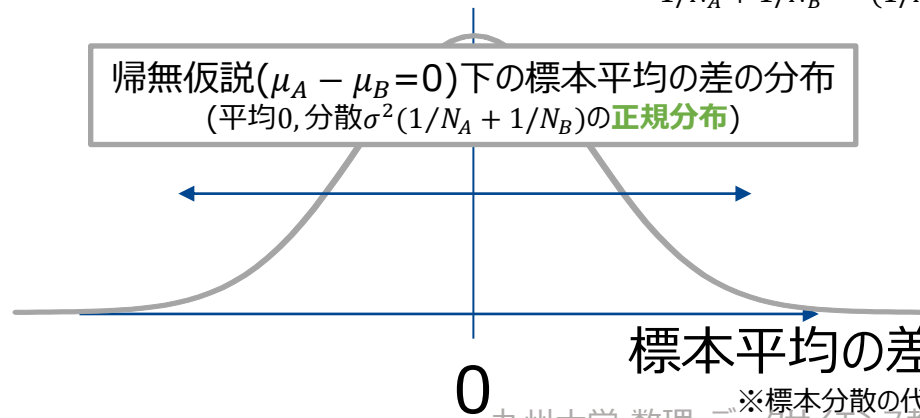
群AとBの各標本分散 s_A^2 と s_B^2 を
併合した標本分散

$$s^2 = \frac{N_A s_A^2 + N_B s_B^2}{N_A + N_B}$$

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{s^2(1/N_A + 1/N_B)(N_A + N_B)/(N_A + N_B - 2)}}$$

※

帰無仮説($\mu_A - \mu_B = 0$)下の標本平均の差の分布
(平均0, 分散 $\sigma^2(1/N_A + 1/N_B)$ の正規分布)



標準化

母分散 → 標本分散

$$1/N_A + 1/N_B \rightarrow (1/N_A + 1/N_B)(N_A + N_B)/(N_A + N_B - 2)$$

標本平均の差 $\bar{x}_A - \bar{x}_B$

※標本分散の代わりに不偏標本分散というのをいれば、
この式がもう少しシンプルになる