

# 情報統計 第5回

2019年9月18日 神奈川工科大学



**櫻井 望**

国立遺伝学研究所  
生命情報・DDBJセンター

# スケジュール

|    | 17日（火）<br>データの見え<br>る化              | 18日（水）<br>検定のこれだけ<br>は  | 19日（木）<br>多変量解析の雰<br>囲気 | 23日（月）<br>データ準備<br>発表会 |
|----|-------------------------------------|-------------------------|-------------------------|------------------------|
| 1限 | 1 ガイダンス、<br>PC環境準備、<br>データの見え<br>る化 | 5 区間推定、<br>分布とその使い<br>方 | 9 相関                    | 13 自習（課題、<br>質問）       |
| 2限 | 2 統計の基本<br>と用語                      | 6 t検定                   | 10 主成分分析                | 14 自習（課題、<br>質問）       |
| 3限 | 3 プログラミ<br>ングの基礎                    | 7 検定で注意<br>すること         | 11 他の多変量<br>解析          | 15 発表会                 |
| 4限 | 4 自習（課題<br>検討、復習）                   | 8 自習（課題<br>検討、復習）       | 12 自習（課題<br>検討、復習）      |                        |

# 区間推定

## 分布とその使い方

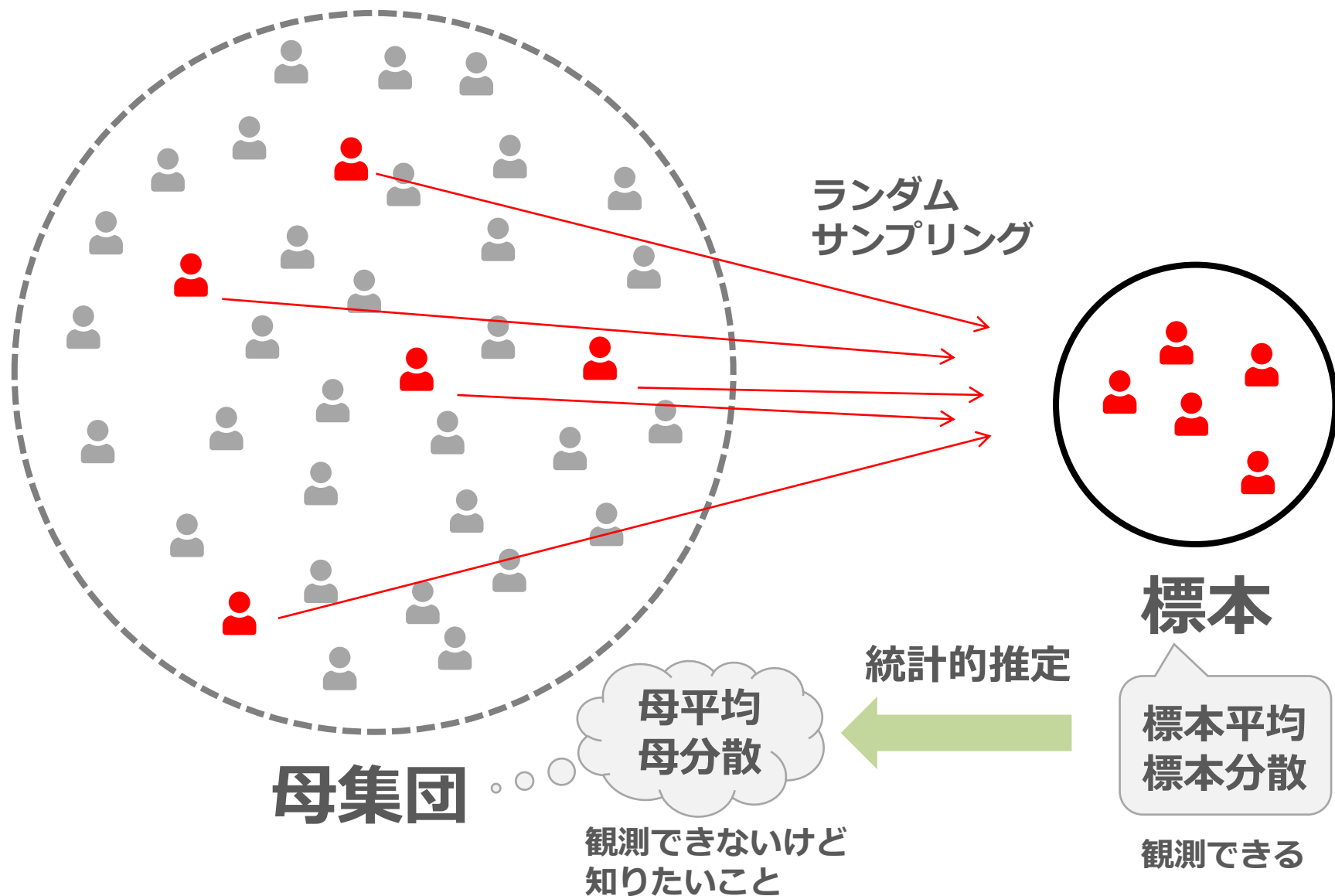
# 学習目標

区間推定を通じて、検定などの基本となる分布と、その使い方を身につけます

- ✓ 正規分布
- ✓ 標準正規分布
- ✓  $t$ 分布

# 統計的推定

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



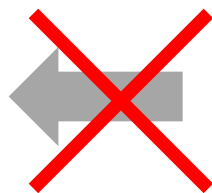
母平均 $\mu$



標本平均 $\bar{x}$

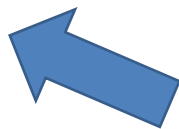
一致が期待できる

母分散 $\sigma^2$



標本分散 $s^2$

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



一致が期待できる

不偏(標本)分散 $v^2$

真の値から外れていないことを、  
**不偏性がある**と言うので。

# 点推定



「母平均 $\mu$ 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

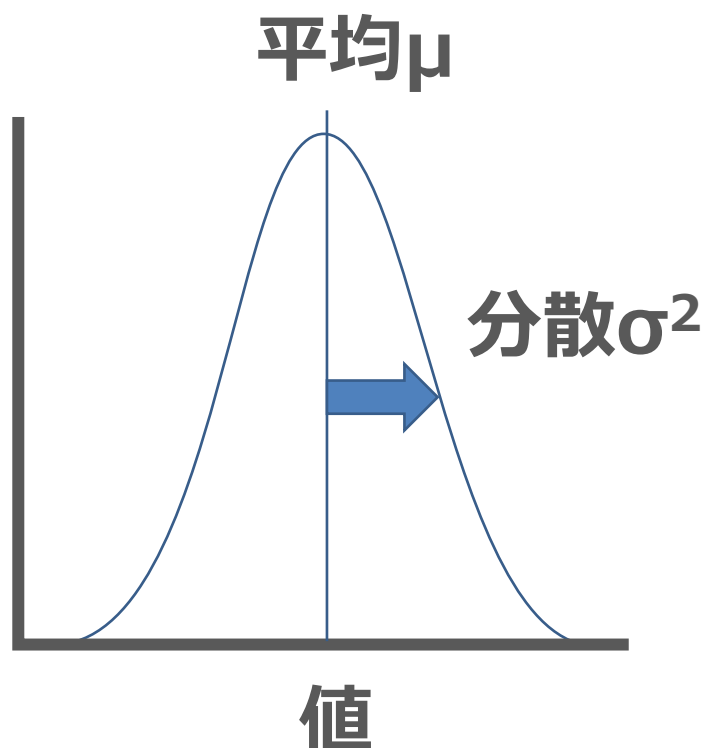


「神奈川県の子どもの平均身長は、信頼係数95%で170.2 ~ 174.6 cmである」のように、幅を持たせて表現する方法

# 標準正規分布

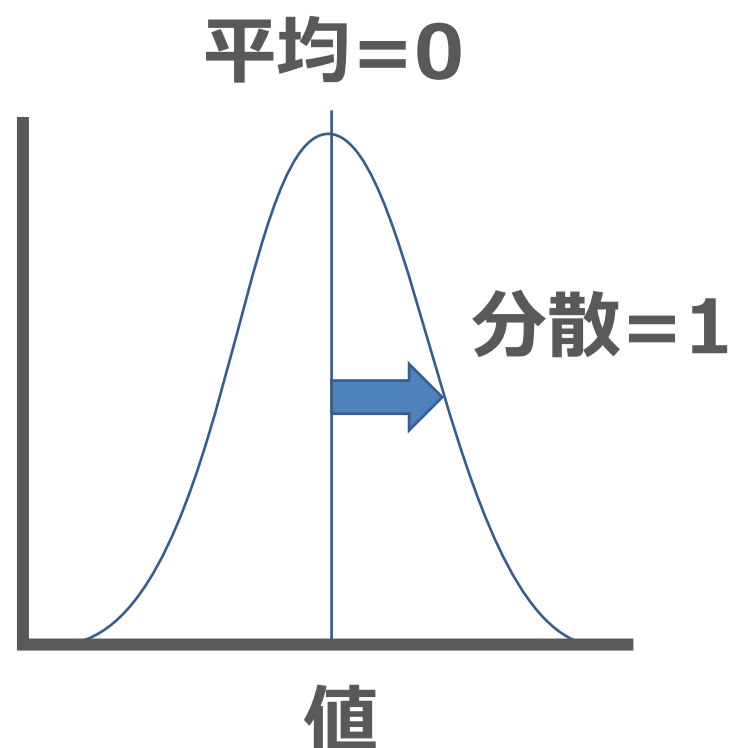


# 正規分布



平均と分散で決まる  
 $N(\mu, \sigma^2)$ と表記

# 標準正規分布

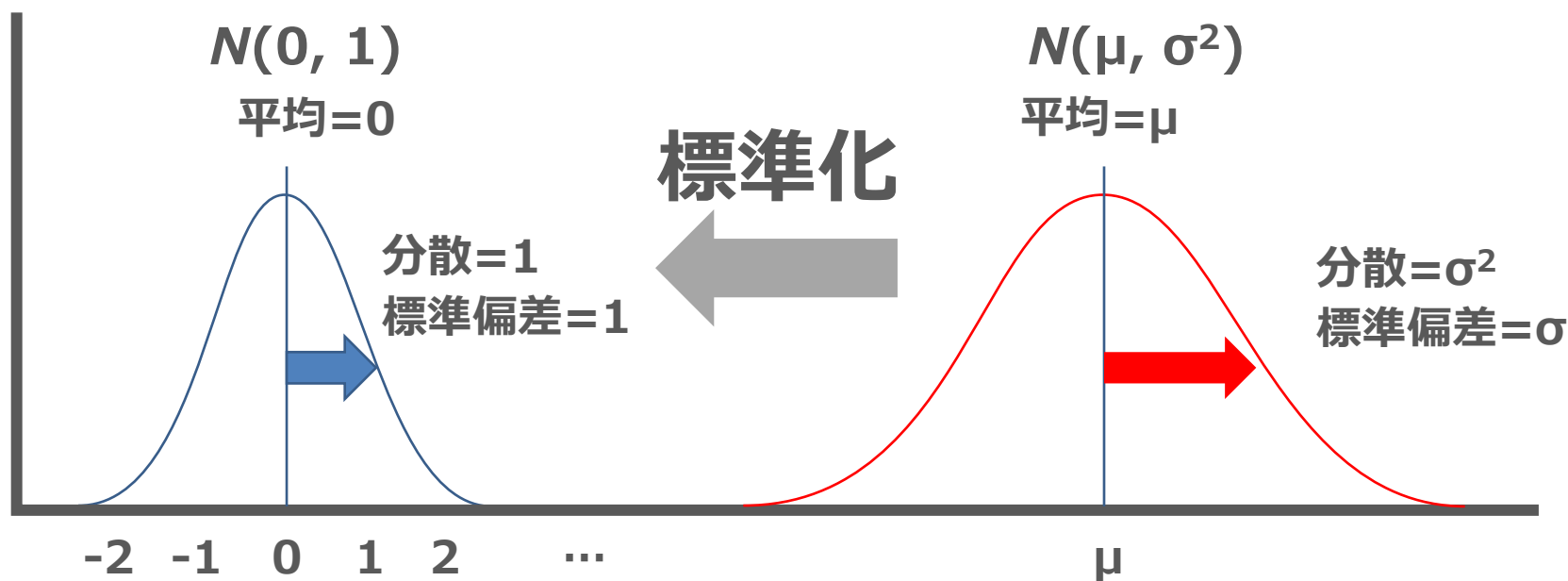


$N(0, 1)$

# 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、

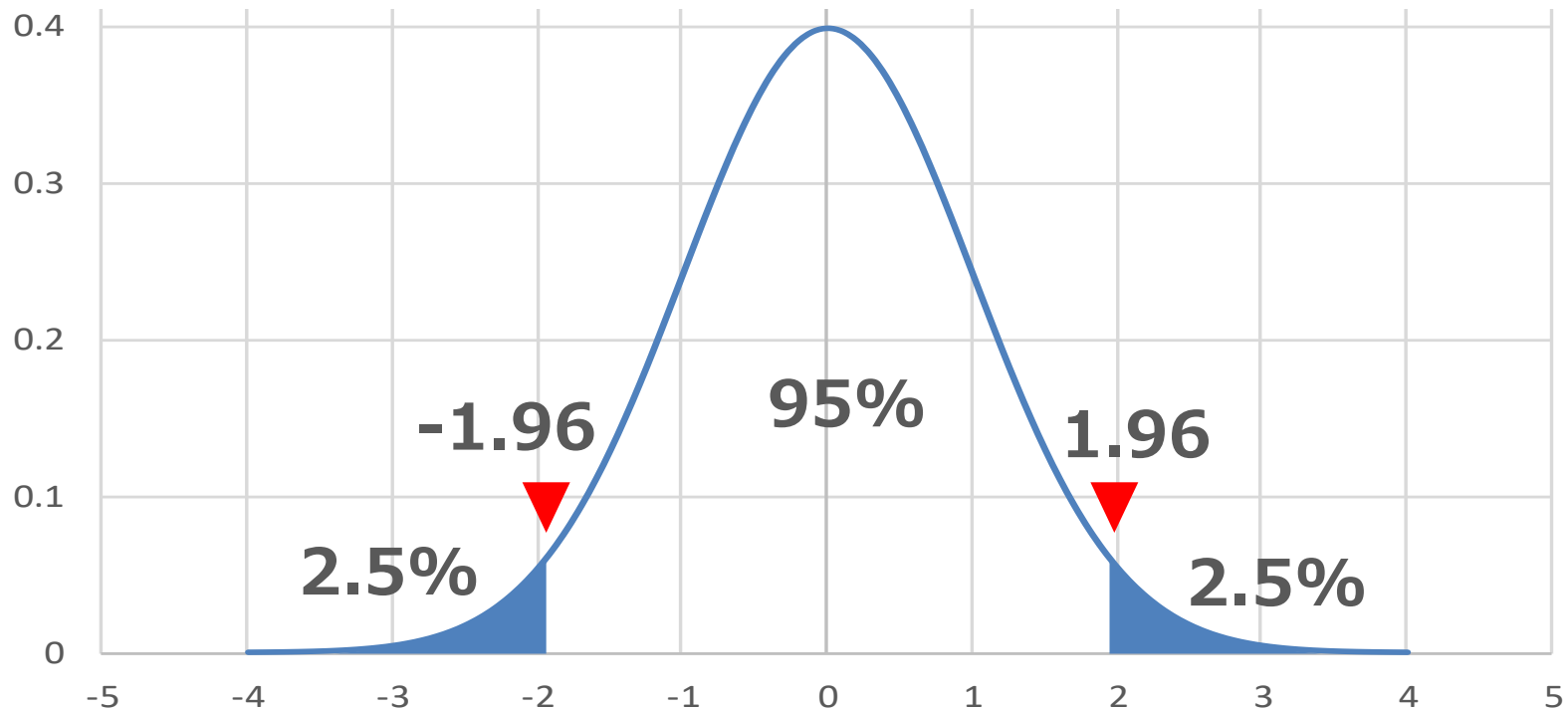
$$Z = \frac{X - \mu}{\sigma} \quad \text{と変換すると、標準正規分布になる。}$$



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

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



# 標準正規分布表

上側確率をあらかじめ  
計算したもの

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

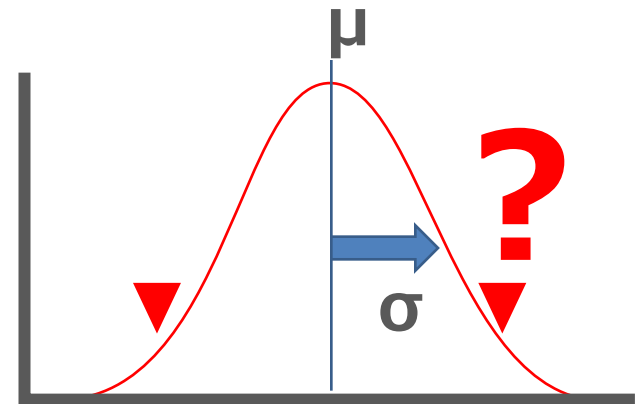
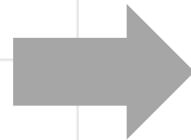
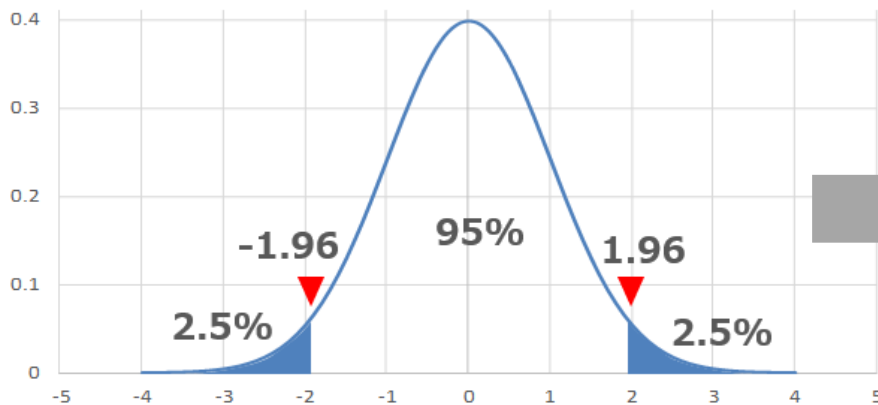
| u   | 0.00    | 0.01    | 0.02    | 0.03    | 0.04    | 0.05    | 0.06    | 0.07    | 0.08    | 0.09    |
|-----|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| 0   | 0.50000 | 0.49601 | 0.49202 | 0.48803 | 0.48405 | 0.48006 | 0.47608 | 0.47210 | 0.46812 | 0.46414 |
| 0.1 | 0.46017 | 0.45620 | 0.45224 | 0.44828 | 0.44433 | 0.44038 | 0.43644 | 0.43251 | 0.42858 | 0.42465 |
| 0.2 | 0.42074 | 0.41683 | 0.41294 | 0.40905 | 0.40517 | 0.40129 | 0.39743 | 0.39358 | 0.38974 | 0.38591 |
| 0.3 | 0.38209 | 0.37828 | 0.37448 | 0.37070 | 0.36693 | 0.36317 | 0.35942 | 0.35569 | 0.35197 | 0.34827 |
| 0.4 | 0.34458 | 0.34090 | 0.33724 | 0.33360 | 0.32997 | 0.32636 | 0.32276 | 0.31918 | 0.31561 | 0.31207 |
| 0.5 | 0.30854 | 0.30503 | 0.30153 | 0.29806 | 0.29460 | 0.29116 | 0.28774 | 0.28434 | 0.28096 | 0.27760 |
| 0.6 | 0.27425 | 0.27093 | 0.26763 | 0.26435 | 0.26109 | 0.25785 | 0.25463 | 0.25143 | 0.24825 | 0.24510 |
| 0.7 | 0.24196 | 0.23885 | 0.23576 | 0.23270 | 0.22965 | 0.22663 | 0.22363 | 0.22065 | 0.21770 | 0.21476 |
| 0.8 | 0.21186 | 0.20897 | 0.20611 | 0.20327 | 0.20045 | 0.19766 | 0.19489 | 0.19215 | 0.18943 | 0.18673 |
| 0.9 | 0.18406 | 0.18141 | 0.17879 | 0.17619 | 0.17361 | 0.17106 | 0.16853 | 0.16602 | 0.16354 | 0.16109 |
| 1.0 | 0.15866 | 0.15625 | 0.15386 | 0.15151 | 0.14917 | 0.14686 | 0.14457 | 0.14231 | 0.14007 | 0.13786 |
| 1.1 | 0.13567 | 0.13350 | 0.13136 | 0.12924 | 0.12714 | 0.12507 | 0.12302 | 0.12100 | 0.11900 | 0.11702 |
| 1.2 | 0.11507 | 0.11314 | 0.11123 | 0.10935 | 0.10749 | 0.10565 | 0.10383 | 0.10204 | 0.10027 | 0.09853 |
| 1.3 | 0.09680 | 0.09510 | 0.09342 | 0.09176 | 0.09012 | 0.08851 | 0.08691 | 0.08534 | 0.08379 | 0.08226 |
| 1.4 | 0.08076 | 0.07927 | 0.07780 | 0.07636 | 0.07493 | 0.07353 | 0.07215 | 0.07078 | 0.06944 | 0.06811 |
| 1.5 | 0.06681 | 0.06552 | 0.06426 | 0.06301 | 0.06178 | 0.06057 | 0.05938 | 0.05821 | 0.05705 | 0.05592 |
| 1.6 | 0.05480 | 0.05370 | 0.05262 | 0.05155 | 0.05050 | 0.04947 | 0.04846 | 0.04746 | 0.04648 | 0.04551 |
| 1.7 | 0.04457 | 0.04363 | 0.04272 | 0.04182 | 0.04093 | 0.04006 | 0.03920 | 0.03836 | 0.03754 | 0.03673 |
| 1.8 | 0.03593 | 0.03515 | 0.03438 | 0.03362 | 0.03288 | 0.03216 | 0.03144 | 0.03074 | 0.03005 | 0.02938 |
| 1.9 | 0.02872 | 0.02807 | 0.02743 | 0.02680 | 0.02619 | 0.02559 | 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 | 0.00452 | 0.00437 | 0.00422 | 0.00407 | 0.00392 | 0.00377 | 0.00362 | 0.00347 | 0.00332 | 0.00317 |
| 2.7 | 0.00302 | 0.00287 | 0.00272 | 0.00257 | 0.00242 | 0.00227 | 0.00212 | 0.00197 | 0.00182 | 0.00167 |
| 2.8 | 0.00152 | 0.00137 | 0.00122 | 0.00107 | 0.00092 | 0.00077 | 0.00062 | 0.00047 | 0.00032 | 0.00017 |
| 2.9 | 0.00012 | 0.00007 | 0.00002 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |

出典

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

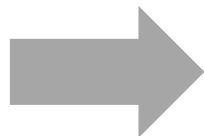
# 区間推定の考え方

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



# 標準化

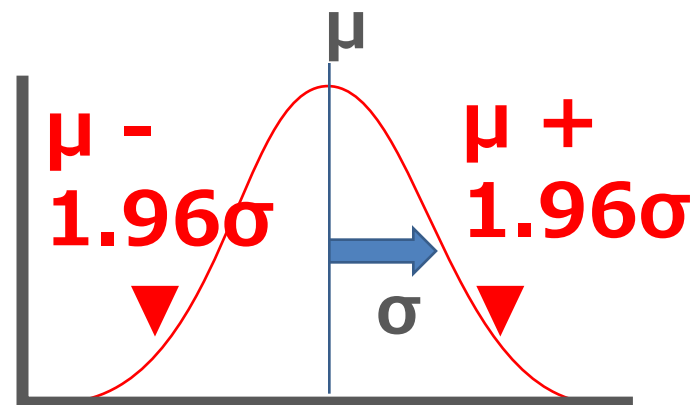
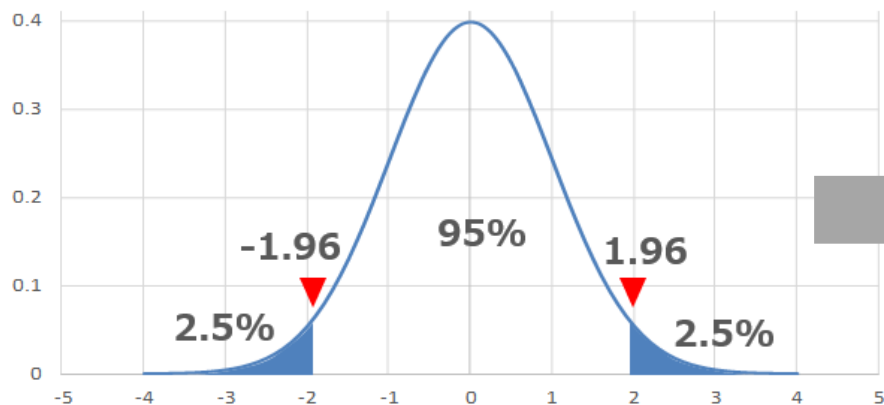
$$Z = \frac{X - \mu}{\sigma}$$



# 標準化の逆

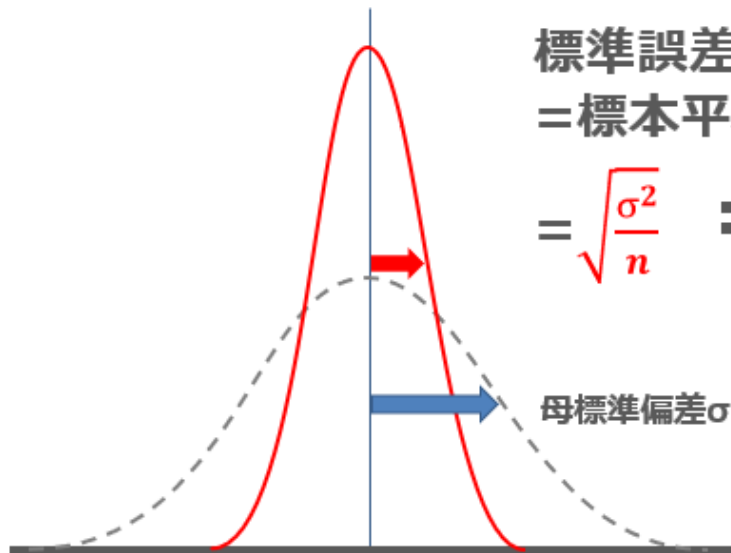
$$X = \mu + Z\sigma$$

例)  $Z = 1.96$ なら、  
 $X = \mu + 1.96 \sigma$



# 標準誤差

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



$$\begin{aligned}\text{標準誤差} &= \text{標本平均}\bar{x}\text{の標準偏差} \\ &= \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}\end{aligned}$$

$\mu$ 推定値： $\bar{x}$

標準偏差： $\frac{\sigma}{\sqrt{n}}$

を当てはめる

# 区間推定のまとめ

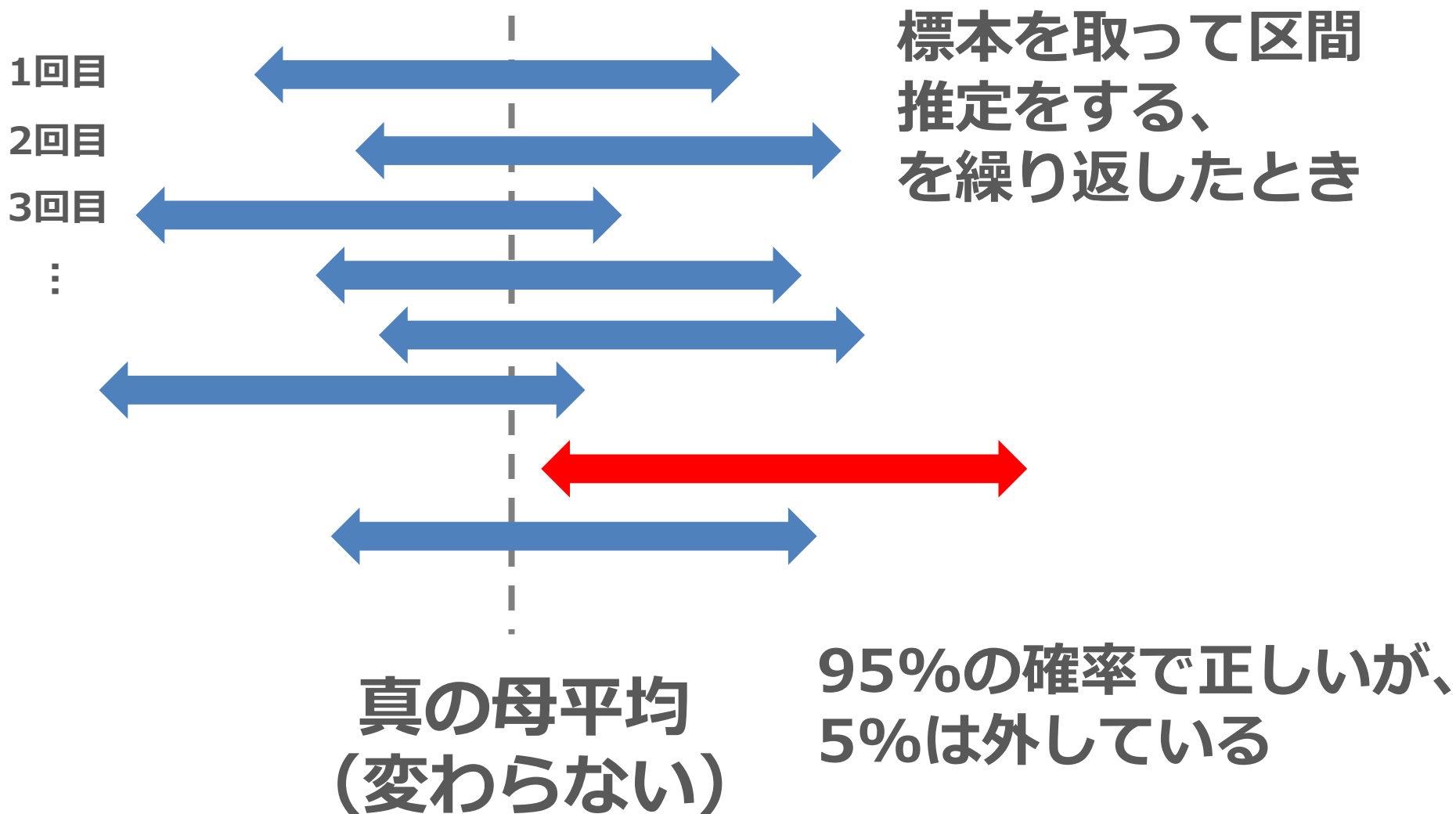
- 母平均 $\mu$ の推定値： 標本平均  $\bar{x}$
- 推定値の標準偏差： 標本平均の標準偏差  $\frac{\sigma}{\sqrt{n}}$
- の場合、95%信頼区間は、以下で求められる

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

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



# イメージ



一般化すると

## 区間推定（分散既知の場合）

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

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

ここで $A$ は、標準正規分布表から、

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

で求められる境界値

ただし...

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

## 母分散 $\sigma^2$ は不明な場合がほとんど

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



母分散が不明な場合は、正規分布ではなく、**t分布**を用いて同様に考える

# t分布

標準正規分布の、  
標本数が少ない場合の  
実用化バージョン

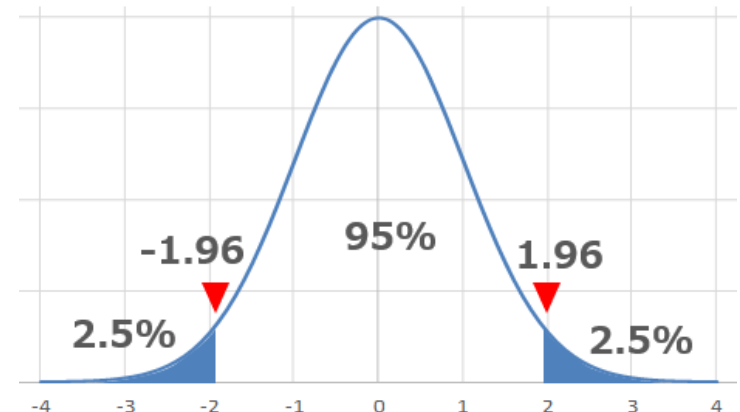
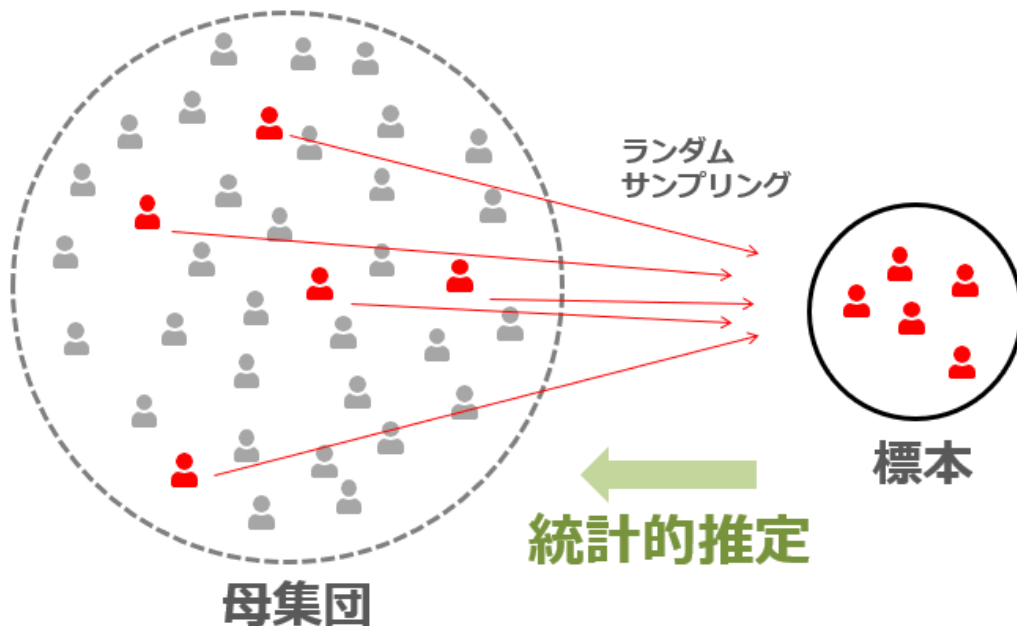
by 櫻井

# t分布

## スチューデントのt分布

正規分布する母集団から標本をとり、母平均 $\mu$ 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。**

by 櫻井



# 考えた人

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



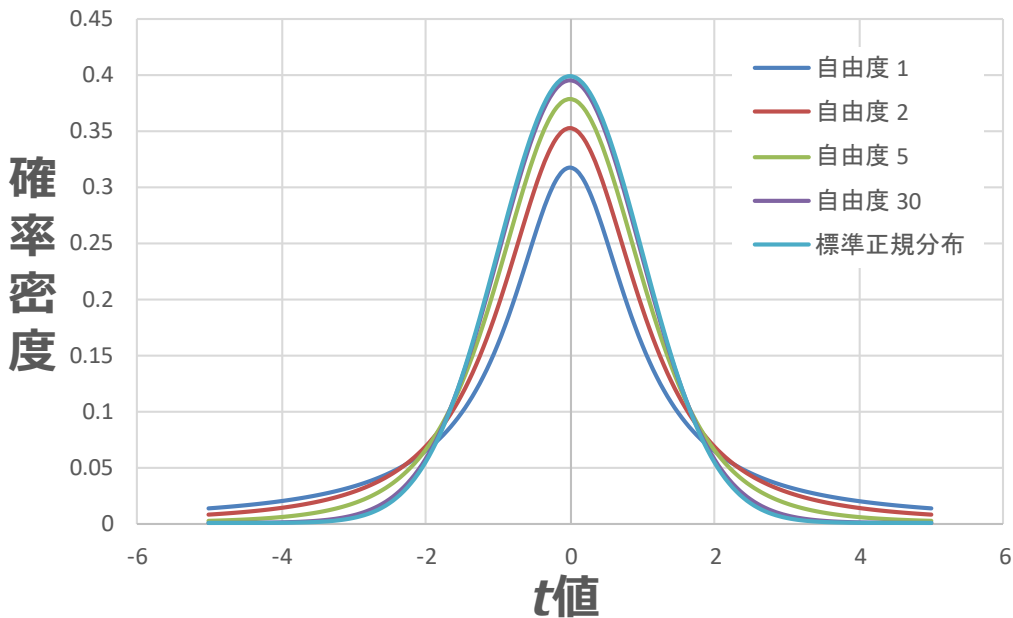
出典：Wikipedia



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

出典：ギネス社HP

# t分布



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

Excelでは、T.DIST, T.INV関数で計算できる

## t分布表

| 自由度 $\nu$ | $\alpha = 0.1$ | $\alpha = 0.05$ | $\alpha = 0.025$ | $\alpha = 0.01$ | $\alpha = 0.005$ |
|-----------|----------------|-----------------|------------------|-----------------|------------------|
| 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。

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

「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、

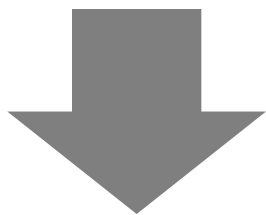
✓ 自由度 $=n-1$

✓  $\alpha$  (信頼計数)  $= (100-a)/2/100$

で求められる境界値。

# まとめ

分布（確率密度関数）



事象が起きる確率を推定できる！

# 描いてみよう

- 標準正規分布
- $t$ 分布
- 裾野の面積と境界値を計算

標準化してみよう



【参考】覚える必要はありません

正規分布の確率密度関数

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

標準正規分布の確率密度関数

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

【参考】 覚える必要はありません

## $t$ 分布の確率密度関数

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

$v$ : 自由度