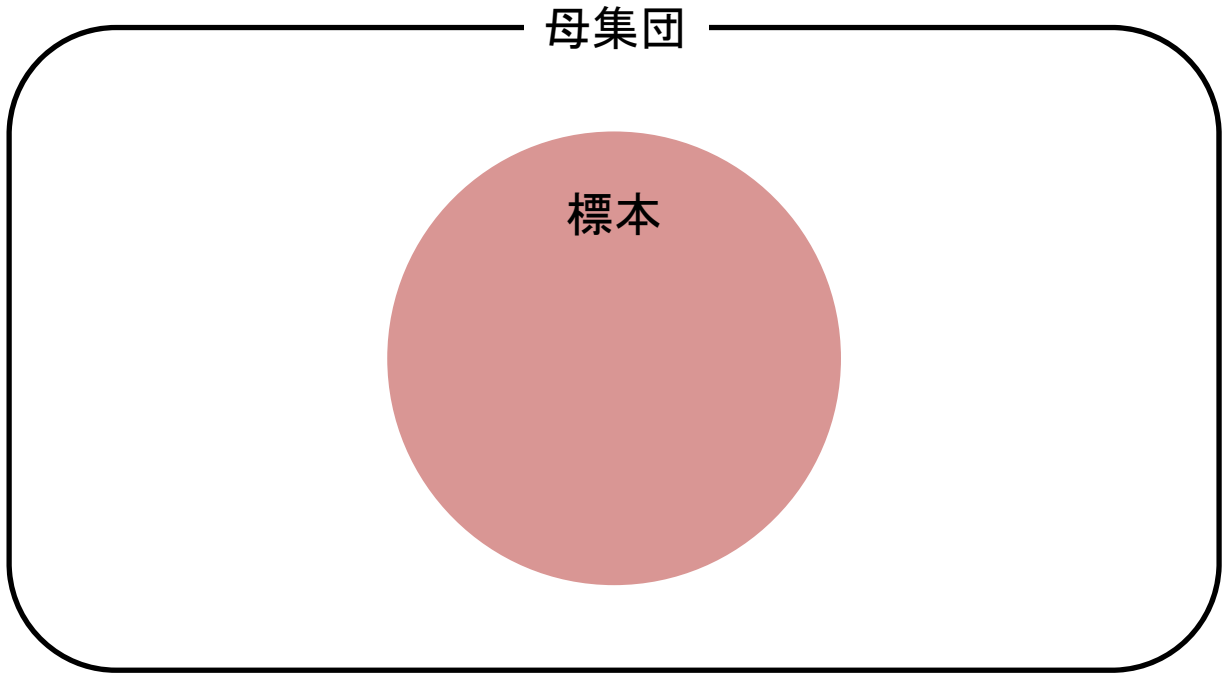
検定

データ・マイニングI

母集団と標本集団

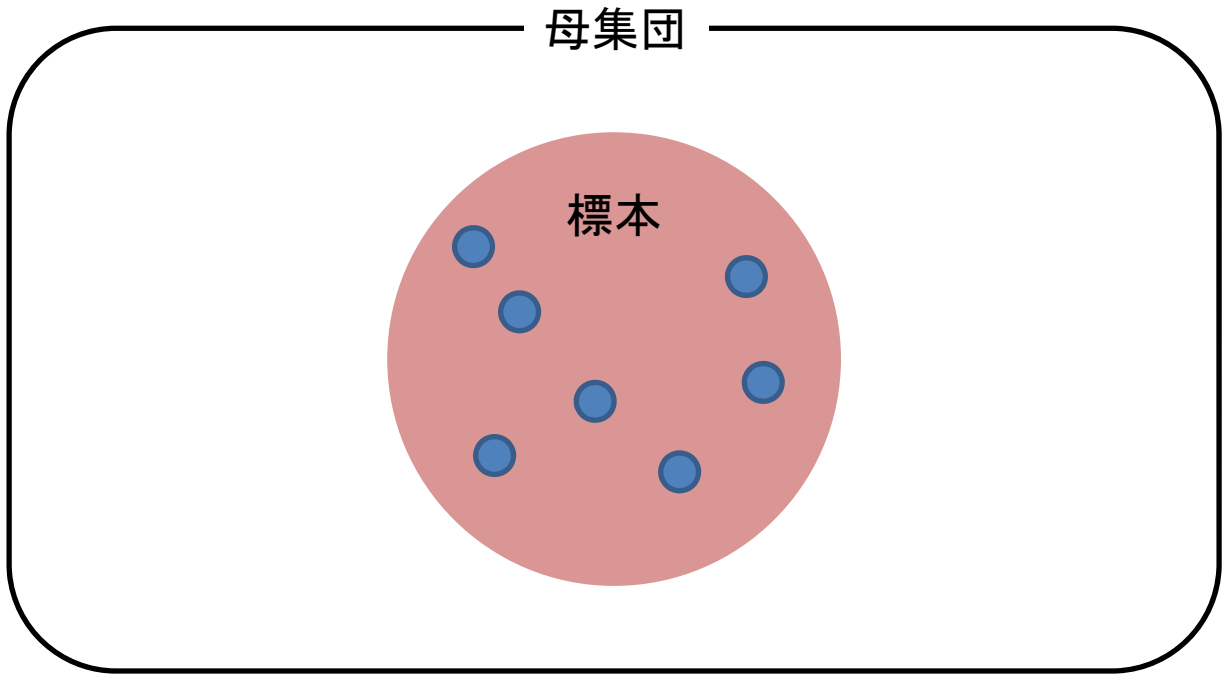
- **母集団とは、対象となるすべてのデータ**
  - 母集団が日本に住んでいる人であれば、文字通り日本に住んでいる人すべて
- **標本とは、母集団から選択されたデータ**
  - 母集団が日本に住んでいる人として、標本はそれから選択された人
  - 統計的にすべてのデータを集めて計算し、検証することはできないので、標本を使って母集団について分析する



11:29:30

母集団と標本集団

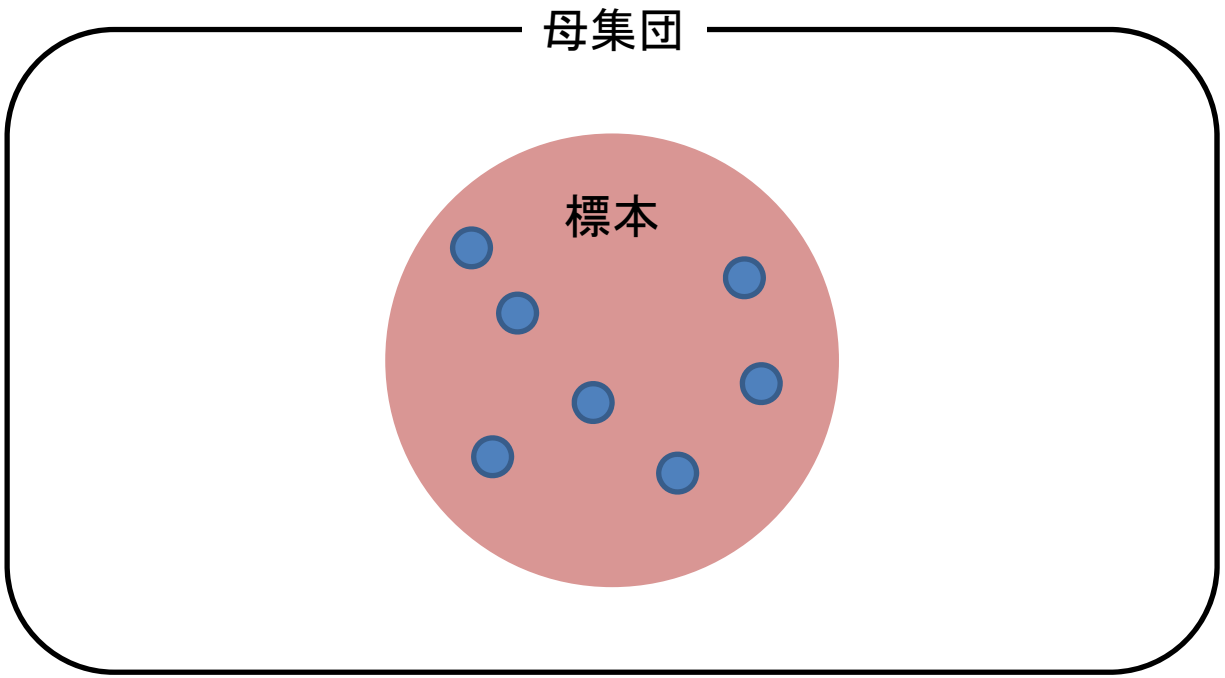
- 母集団から標本を選択する場合、ランダムに選択することが望ましい
  - 偏りをなくし、より客観的な分析を行うため
  - **標本のデータ個数をサイズという**



11:29:30

母集団と標本集団

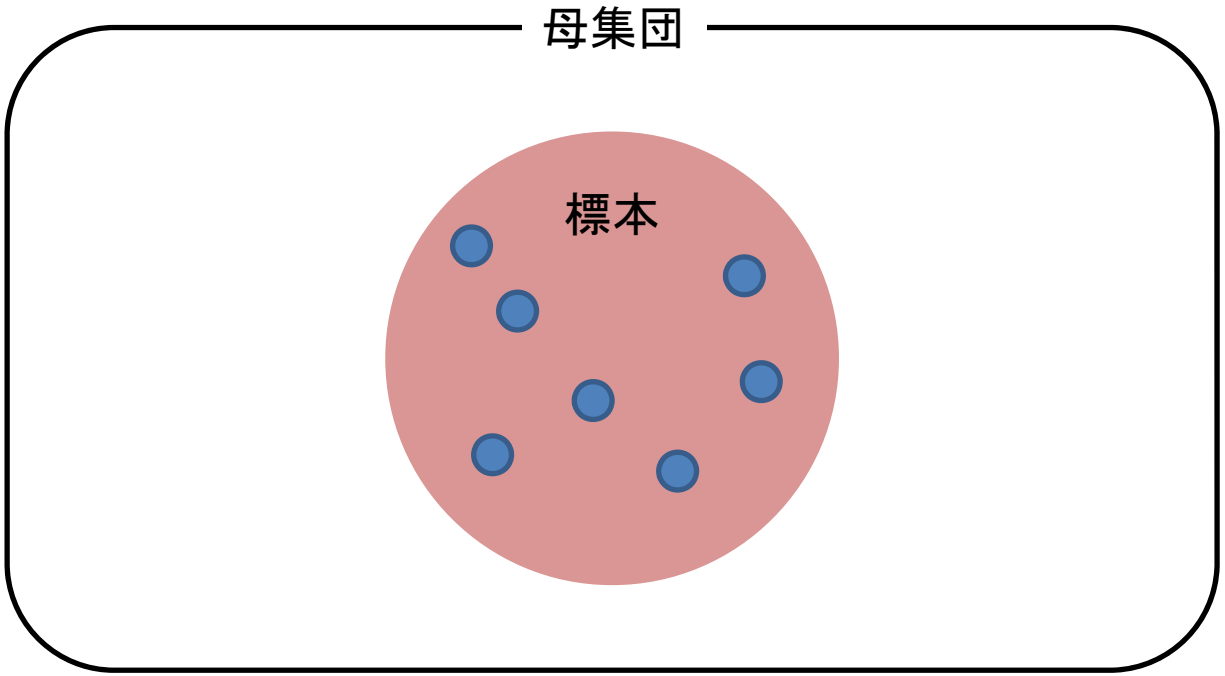
- 標本のデータを使って、母集団の特徴を分析する
  - 母集団が多すぎると使いづらいので、ちょうどよいデータ数の標本を使う
  - 例えば、母集団の平均はどうだろうか？
    - 標本平均を使えばよい！



11:29:30

標本の抽出

- 母集団から標本を抽出する
  - 日本人を分析したいが、APUの日本人学生のみを抽出すると、日本人全体の特徴を捉えることはできない！
  - 標本に偏りをなくすため、ランダムに抽出する
    - 母集団に番号つけて、乱数を発生させて、乱数と同じ番号のデータを抽出すればよい



11:29:30

YOUR TURN

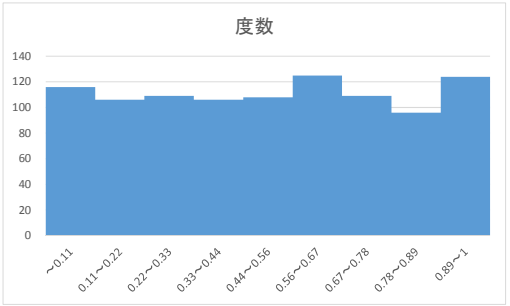
標本平均と標本のサイズ

- X_iをある分布に従う乱数として、その標本平均

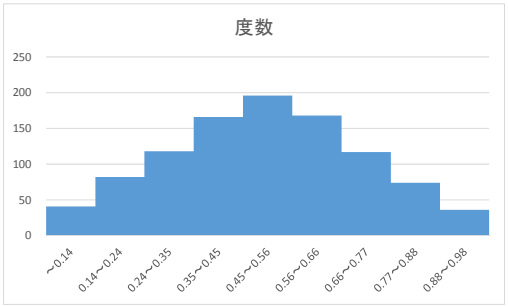
$$\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

標本サイズnを大きくすると、その分布は正規分布に近づく

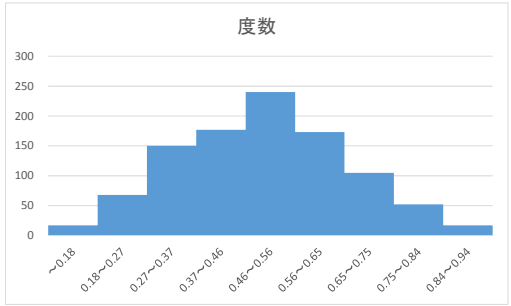
$n = 1$



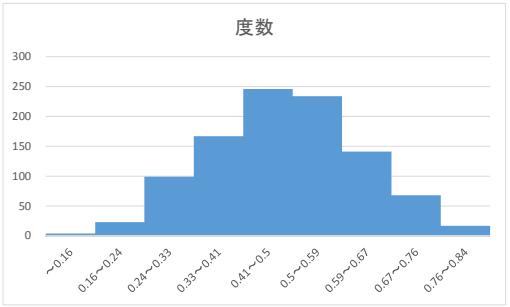
$n = 2$



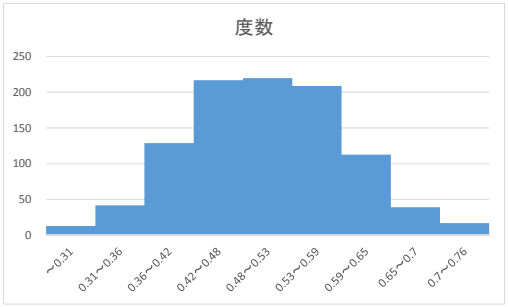
$n = 3$



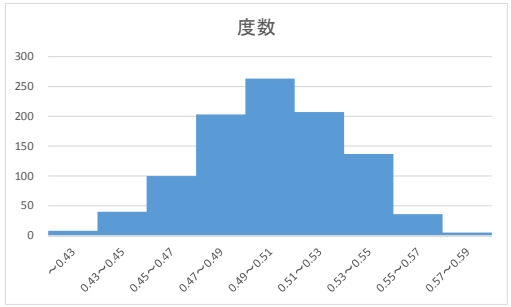
$n = 5$



$n = 10$



$n = 100$



11:29:30

標本平均の特徴:中心極限定理

- $\{X_1, X_2, \cdots, X_n\}$の独立同分布の確率変数
- 各$\{X_i\}$の期待値はμ、標準偏差はσ
- 標本サイズnが大きくなるにつれて、$\sqrt{n}(\bar{X}_n - \mu)$は正規分布$N(0, \sigma^2)$に分布収束する

$$\sqrt{n}(\bar{X}_n - \mu) \quad \xrightarrow{d} \quad N(0, \sigma^2)$$

中心極限定理

- $\{X_1, X_2, \cdots, X_n\}$の独立同分布の確率変数
- 各$\{X_i\}$の期待値はμ、標準偏差はσ
- 標本サイズnが大きくなるにつれて、$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$は正規分布$N(0,1)$に分布収束する

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \quad \xrightarrow{d} \quad N(0,1)$$

- 大量の乱数を発生させて、その平均を計算し、正規化すると、正規分布に限りなく近づく

標本平均の分布の平均

- 標本平均$\bar{X}$の分布の平均$\bar{\bar{X}}$
  - データから計算されたものを$\bar{\bar{X}}$
  - 理論的に計算されたものを$\mu_{\bar{X}}$
    - $\mu_{\bar{X}}$は母集団の平均μに等しい$\mu_{\bar{X}} = \mu$

- 標本平均$\bar{X}$の分布の標準偏差$s_{\bar{X}}$
  - データから計算されたものを$s_{\bar{X}}$
  - 理論的に計算されたものを$\sigma_{\bar{X}}$
    - 母集団のデータ数が無限ならば$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
    - 母集団のデータ数が有限ならば$\sigma_{\bar{X}} = \frac{\sqrt{N-n}}{\sqrt{N-1}}\frac{\sigma}{\sqrt{n}}$
    - nは標本サイズ、Nは有限母集団のサイズ

YOUR TURN

推定ー区間推定ー

“What! You have solved it already?”
“Well, that would be too much to say. I have discovered a suggestive fact, that is all.”
「え！もう分かったのか？」
「まあ、そこまでとはいわない。示唆に富む事実を発見しただけだよ。」

Dr. Watson and Sherlock Homes
The Sign of Four

パラメータとは

- **パラメータ（母数）とは**、以下の2つの意味を持つ

  1. **母集団の特徴を表す指標**：以下に例
     - 母平均：母集団の平均、
     - 母分散：母集団の分散、
     - 母標準偏差：母集団の標準偏差

  2. **確率分布を特徴付ける指標**：以下に例
     - 一様分布におけるaとb：$1/(b-a)$の確率密度でaとbの間の数の値を取る
     - 正規分布の平均μと分散σ^2：$N(\mu, \sigma^2)$

- ここでは、前者の意味で、特に母平均の推定を行う

パラメータの推定

- 母集団の分布の未知パラメータに対して、標本からその値を推定する

- 点推定量とは、標本に関する関数のこと
  - 推定量とは標本に関する関数であり、
    - $\frac{X_1+X_2+X_3}{3}$は推定量
  - 推定値とは推定量が取る実際の値である
    - $\frac{4+2+6}{3}=4$は推定値。ここで、$X_1=4, X_2=2, X_3=6$の値を取ったとしている

- 母平均μの推定を行う
  - 東証第1部上場している株式の収益率の平均
  - 全世界の学生のTOEFLの平均、全世界のTOEFLの平均
  - ある国に住んでいるの人全員の身長、体重の平均

11:29:31

パラメータの推定

- 母平均μの推定
  - 母標準偏差σが既知の場合
  - 母標準偏差σが未知の場合
    - 小標本の場合
    - 大標本の場合

- 母標準偏差σの推定

母平均μの推定　考え方

- 標本データから計算した標本平均$\bar{X}$は、中心極限定理によると$N(0,1)$の正規分布に近づく
  - 母集団がどのような分布でも
- この性質を利用して、母平均μを区間推定する
  - $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$を用いる

母平均μの推定：分散が既知の場合

- 標本平均$\bar{X}$の標準化

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

  - $z = \frac{X-\mu}{\sigma}$と$z = \frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}}$は異なる
    - 左辺は元のデータXを標準化したもの（Xの分布の平均はμ、標準偏差はσ）
    - 右辺は標本平均$\bar{X}$を標準化したもの（$\bar{X}$の分布の平均は$\mu_{\bar{X}}$、標準偏差は$\sigma_{\bar{X}}$）

$$\boldsymbol{z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}$$

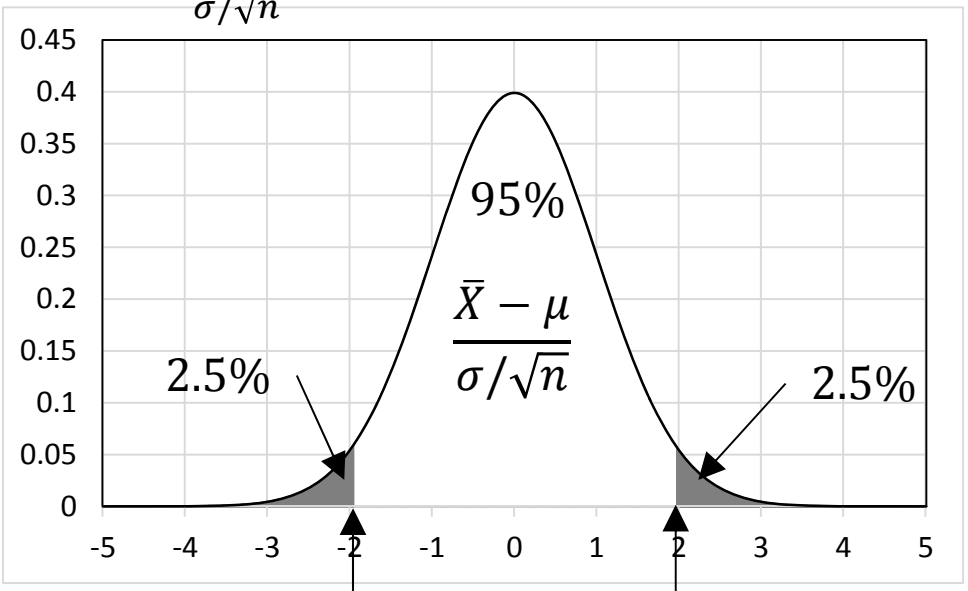
  - $\mu_{\bar{X}} = \mu$、$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$を最初の式に代入
    - $E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n}E(\sum_{i=1}^{n} X_i) = \frac{1}{n}nE(X_i) = \mu$
    - $Var(\bar{X}) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}nVar(X_i) = \frac{\sigma^2}{n}$
    - $\sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$

母平均μの区間推定のイメージ

$$P\left\{\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \leq z\right\} = P\left\{\bar{X}-\frac{z\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+\frac{z\sigma}{\sqrt{n}}\right\}$$

$$\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \leq z \quad \Leftrightarrow \quad -z \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z \quad \Leftrightarrow \quad \bar{X}-\frac{z\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+\frac{z\sigma}{\sqrt{n}}$$

$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$の分布（標準正規分布）



$-z = -1.96$　　$z = 1.96$

11:29:32

信頼区間

$$P\left\{\bar{X} - \frac{z_\alpha \sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_\alpha \sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

- **$\bar{X} - \frac{z_\alpha \sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_\alpha \sigma}{\sqrt{n}}$を信頼係数$1-\alpha$の信頼区間という**
  - **$\bar{X} + \frac{z_\alpha \sigma}{\sqrt{n}}$を上方信頼限界**
  - **$\bar{X} - \frac{z_\alpha \sigma}{\sqrt{n}}$を下方信頼限界という**
- 信頼係数$1-\alpha$というのは、ある確率変数がある範囲内に収まる確率を意味する
  - $1-\alpha$として99%、95%や90%などの確率を取る
- 例:
  - $\bar{X} - \frac{1.96\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{1.96\sigma}{\sqrt{n}}$は信頼係数95%の信頼区間
  - ここで、標準正規分布において信頼係数95%の$z_{0.05}$は1.96である

母平均μの推定：分散が既知の場合

$$\bar{X} - \frac{z\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z\sigma}{\sqrt{n}}$$

- 未知数はμのみでその他は分かっている
  - **母標準偏差σは既知**

- μの上方信頼限界（これより上は確率として2.5%の領域に入る）
  - $\mu_U = \bar{X} + \frac{z\sigma}{\sqrt{n}}$:

- μの下方信頼限界（これより下は確率として2.5%の領域に入る）
  - $\mu_L = \bar{X} - \frac{z\sigma}{\sqrt{n}}$

- 母平均μの区間推定値

$$\bar{X} - \frac{z\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z\sigma}{\sqrt{n}}$$

11:29:33

母平均μの推定：分散が既知の場合

$$\bar{X} - \frac{z\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z\sigma}{\sqrt{n}}$$

- 未知数はμのみでその他は分かっている
  - $\bar{X}$：TOEFLの平均スコアは50
  - σ：TOEFLのスコアのばらつきは10（とする）
  - n：TOEFLを受けた人の数は5
  - z：正規分布に対するしきい値
    - 上側2.5%ならば1.96、下側2.5%ならば-1.96

- μの上方信頼限界
  - $\mu_U = \bar{X} + \frac{z\sigma}{\sqrt{n}} = 50 + (1.96) \times 10/\sqrt{5}$

- μの下方信頼限界
  - $\mu_L = \bar{X} - \frac{z\sigma}{\sqrt{n}} = 50 - (1.96) \times 10/\sqrt{5}$

YOUR TURN

母平均μの推定：分散が未知の場合

- 分散（母標準偏差）が既知の場合、母平均μの推定には正規分布を使って推定する

- しかし、**分散が未知**の場合、***t*分布**を使って推定する

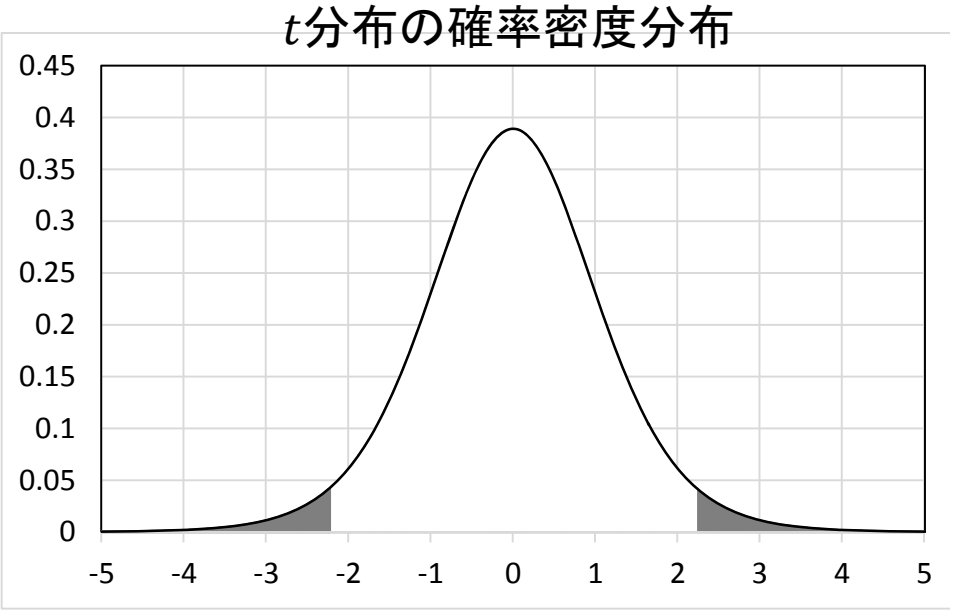
t分布

- t分布とは、

$$f_t(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{(\nu\pi)}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

が確率密度分布である分布

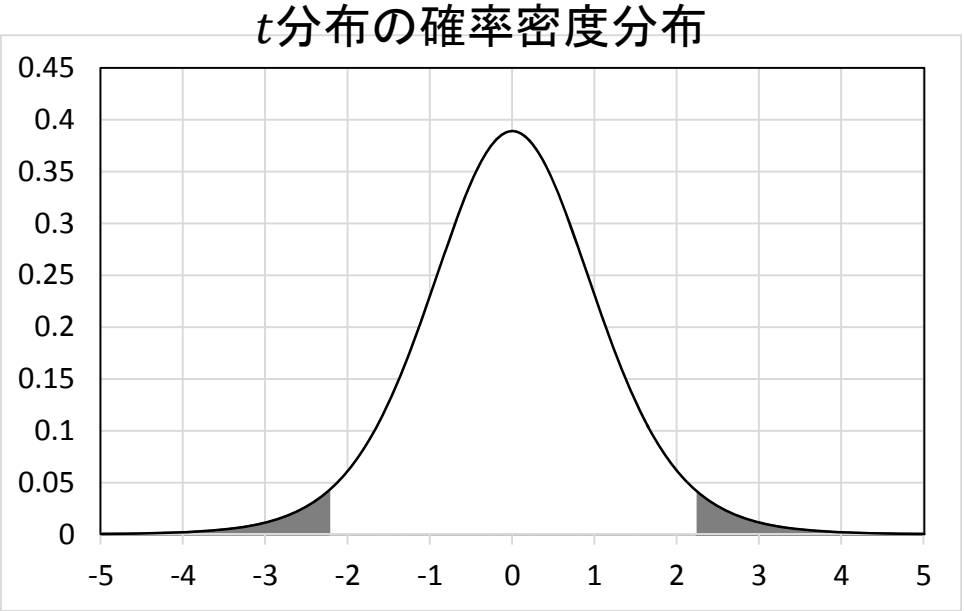
- Γはガンマ関数
- νは自由度を表すパラメータ
- 正規分布と同じbell shape
- $\nu \to \infty$となると、正規分布に収束する



11:29:33

t分布の性質

- t分布の分散は$\sigma^2 = \nu/(\nu - 2)$
- t分布の標準偏差は$\sigma = \sqrt{\nu/(\nu - 2)}$
  - 自由度νが変わると、分散と標準偏差も変わる
- t分布の自由度νは標本サイズnから1を引いたもの：$\nu = n - 1$



11:29:34

t分布に従う変数

- 標準化した標本平均は、t分布に従う変数

$$\mathbf{t} = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

  - ここで$\bar{X}$は標本平均
  - μは母平均
  - sは標本標準偏差
  - nは標本サイズ

  - 標準化した$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$との違いは母標準偏差$\sigma$と標本標準偏差$s$だけ！！
    - 母標準偏差σは未知のため、既知の標本標準偏差sを使う

$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$の応用方法

1. **母標準偏差σが未知のとき、母平均μの推定ができる**
   - 母標準偏差σが既知のとき、$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$を利用した
   - 母標準偏差σがの未知のときは、$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$の代わりに

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

     σをsに置き換えて利用する

2. **統計量の有意性の検定に使うことができる**

t分布の有意水準

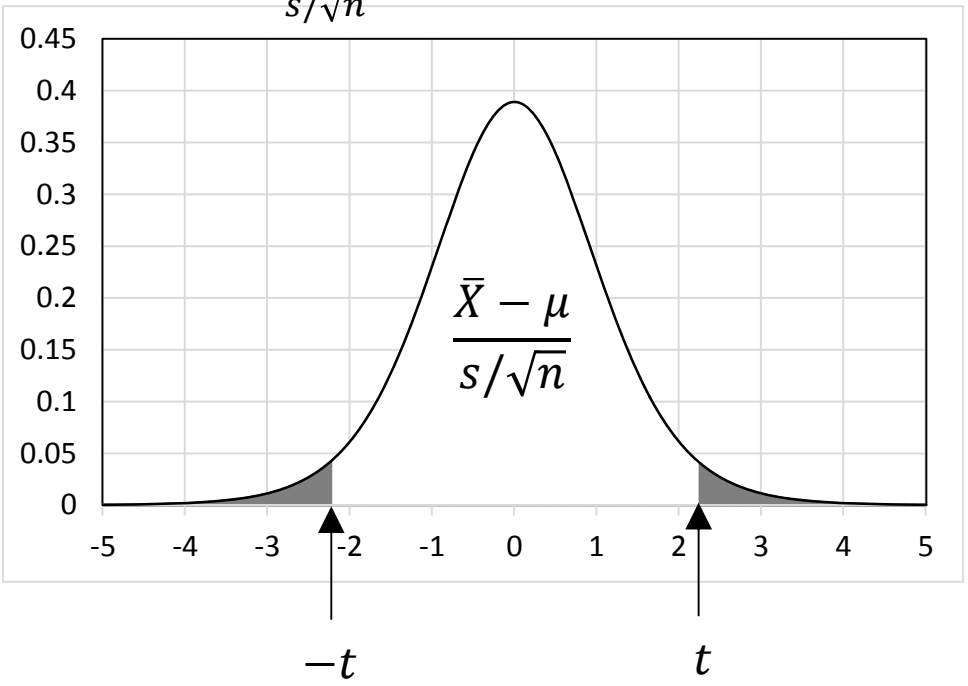
- t分布の有意水準はExcelのT.DIST関数を用いて計算できる

母平均μの推定のイメージ

$$P\left\{\left|\frac{\bar{X}-\mu}{s/\sqrt{n}}\right| \leq t\right\} = P\left\{\bar{X}-\frac{ts}{\sqrt{n}} \leq \mu \leq \bar{X}+\frac{ts}{\sqrt{n}}\right\}$$

$$\left|\frac{\bar{X}-\mu}{s/\sqrt{n}}\right| \leq t \quad \Leftrightarrow \quad -t \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq t \quad \Leftrightarrow \quad \bar{X}-\frac{ts}{\sqrt{n}} \leq \mu \leq \bar{X}+\frac{ts}{\sqrt{n}}$$

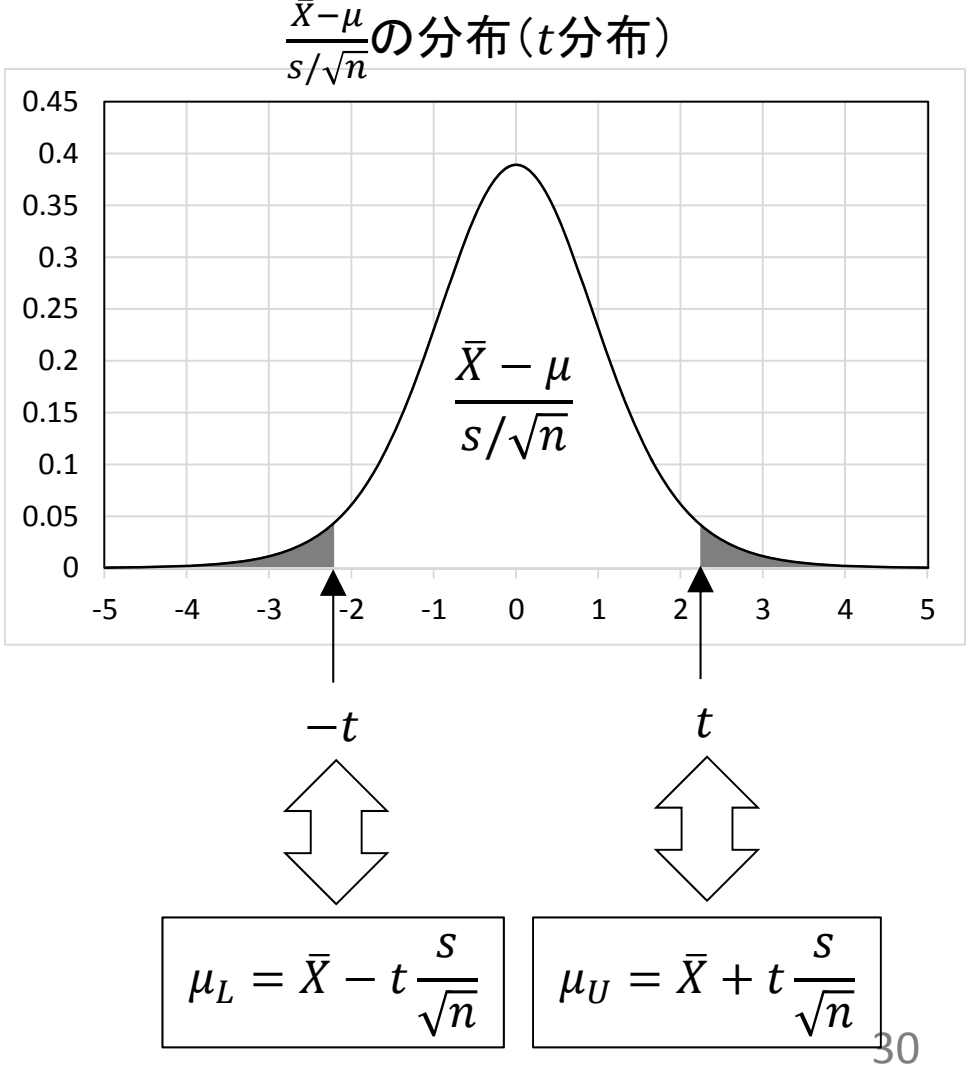
$\frac{\bar{X}-\mu}{s/\sqrt{n}}$の分布（$t$分布）



11:29:36

母平均μの推定
母標準偏差σが未知、小標本（データが少ない）

- $\mathrm{t} = \frac{\bar{X}-\mu}{s/\sqrt{n}}$を用いる
- μの上方信頼限界
  - $\mu_U = \bar{X} + t\frac{s}{\sqrt{n}}$
- μの下方信頼限界
  - $\mu_L = \bar{X} - t\frac{s}{\sqrt{n}}$



11:29:37

例：母平均μの推定
母標準偏差σが未知、小標本（データが少ない）

- 10人の学生がTOEFLを受けた
- TOEFLの点数の平均（標本平均）は$\bar{X} = 53.45$、標本標準偏差は14.52
- 母平均μを95%信頼係数のもとで推定

- 信頼係数95%として、両側有意水準0.05のtの値を計算する

$$t_{0.05} = 2.26$$

- μの上方信頼限界
  - $\mu_U = \bar{X} + t_{0.05}\frac{s}{\sqrt{n}} = 53.45 + 2.26 \times \frac{14.52}{\sqrt{10}} = 63.83$
- μの下方信頼限界
  - $\mu_L = \bar{X} - t_{0.05}\frac{s}{\sqrt{n}} = 53.45 - 2.26 \times \frac{14.52}{\sqrt{10}} = 43.07$
- 信頼係数95%の母平均μの区間推定は

$$43.07 \leq \mu \leq 63.83$$

$$P(43.07 \leq \mu \leq 63.83) = 95\%$$

母平均μの推定
母標準偏差σが未知、大標本（データが多い）

- 標本サイズが大きい場合は、t分布の代わりに、正規分布を用いる
  - 標本サイズが大きいと、t分布は標準正規分布に収束する
    - どのくらいの有意水準かにもよるが、具体的には、概ね標本サイズが600以上

- $z = \frac{\bar{X}-\mu}{s/\sqrt{n}}$を用いる

- μの上方信頼限界
  - $\mu_U = \bar{X} + z\frac{s}{\sqrt{n}}$

- μの下方信頼限界
  - $\mu_L = \bar{X} - z\frac{s}{\sqrt{n}}$

例：母平均μの推定
母標準偏差σが未知、大標本（データが多い）

- 600人の学生がTOEFLを受けた
- TOEFLの点数の平均（標本平均）は$\bar{X} = 53.45$、標本標準偏差は14.52
- 母平均μを95%信頼係数のもとで推定

- 信頼係数95%として、両側有意水準0.05のzの値を計算する

$$z_{0.05} = 1.96$$

- μの上方信頼限界
  - $\mu_U = \bar{X} + t_{0.05}\frac{s}{\sqrt{n}} = 53.45 + 1.96 \times \frac{14.52}{\sqrt{10}} = 62.45$
- μの下方信頼限界
  - $\mu_L = \bar{X} - t_{0.05}\frac{s}{\sqrt{n}} = 53.45 - 1.96 \times \frac{14.52}{\sqrt{10}} = 44.45$
- 信頼係数95%の母平均μの区間推定は

$$44.45 \leq \mu \leq 62.45$$

$$P(44.45 \leq \mu \leq 62.45) = 95\%$$

まとめ：母平均μの区間推定

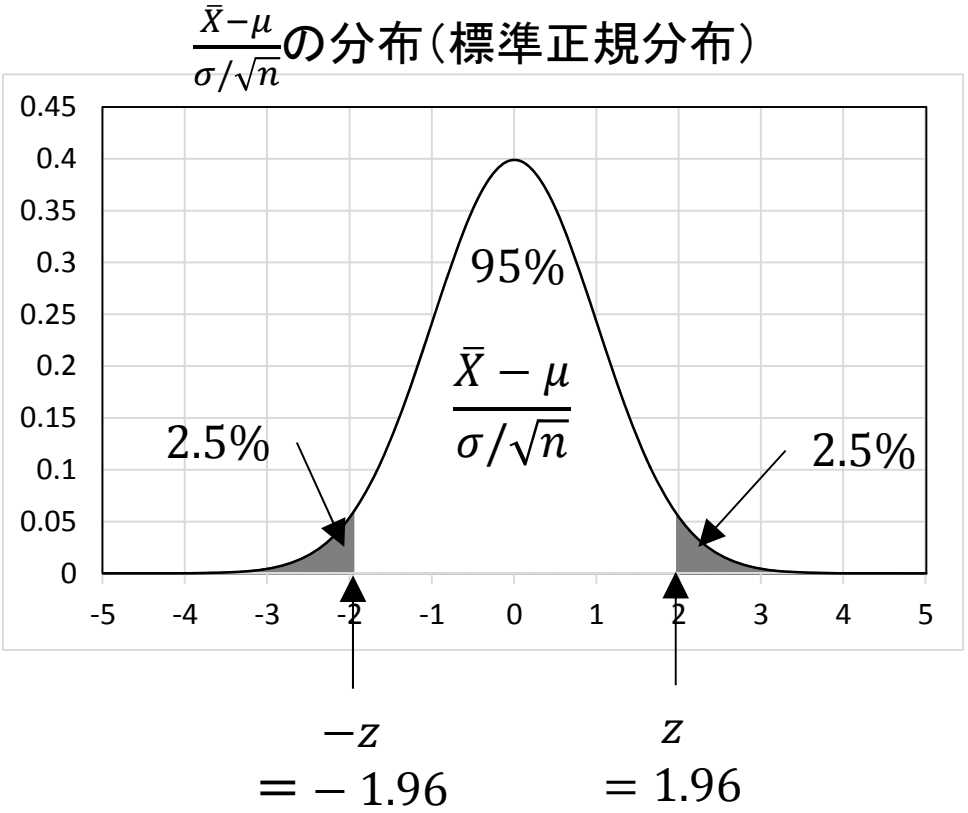
1. 母標準偏差が既知の場合は、正規分布を使って
   - $\bar{X} + \frac{z\sigma}{\sqrt{n}}$：上方信頼限界
   - $\bar{X} - \frac{z\sigma}{\sqrt{n}}$：下方信頼限界
   - $z_{0.05} = 1.96$つまり、zとして信頼係数95%であれば1.96を使う

2. 母標準偏差が未知で小標本（データが少ない）の場合は、t分布を使って
   - $\bar{X} + \frac{ts}{\sqrt{n}}$：上方信頼限界
   - $\bar{X} - \frac{ts}{\sqrt{n}}$：下方信頼限界
   - $t_{0.05} = 2.26$つまり、tとして信頼係数95%であれば2.26を使う

3. 母標準偏差が未知で大標本（データが多い）の場合は、t分布を使って
   - $\bar{X} + \frac{z\sigma}{\sqrt{n}}$：上方信頼限界
   - $\bar{X} - \frac{z\sigma}{\sqrt{n}}$：下方信頼限界

正規分布における信頼係数に対するzの値

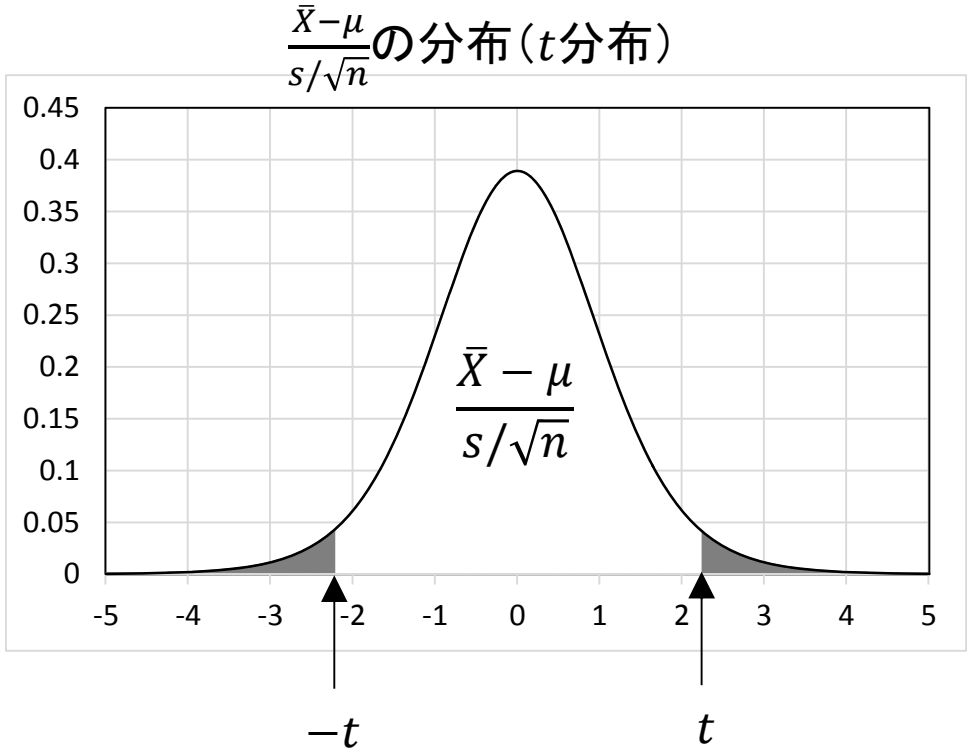
	信頼係数	z値
10%	90%	1.64
5%	95%	1.96
1%	99%	2.58



11:29:38

t分布における信頼係数に対するtの値

- 自由度によって、t分布の信頼係数に対するtの値が変わる



11:29:39

YOUR TURN

検定

“It is a mistake to confound strangeness with mystery.”
「奇異と謎を混同するのは誤りである」

Sherlock Homes
A Study in Scarlet

統計的有意性

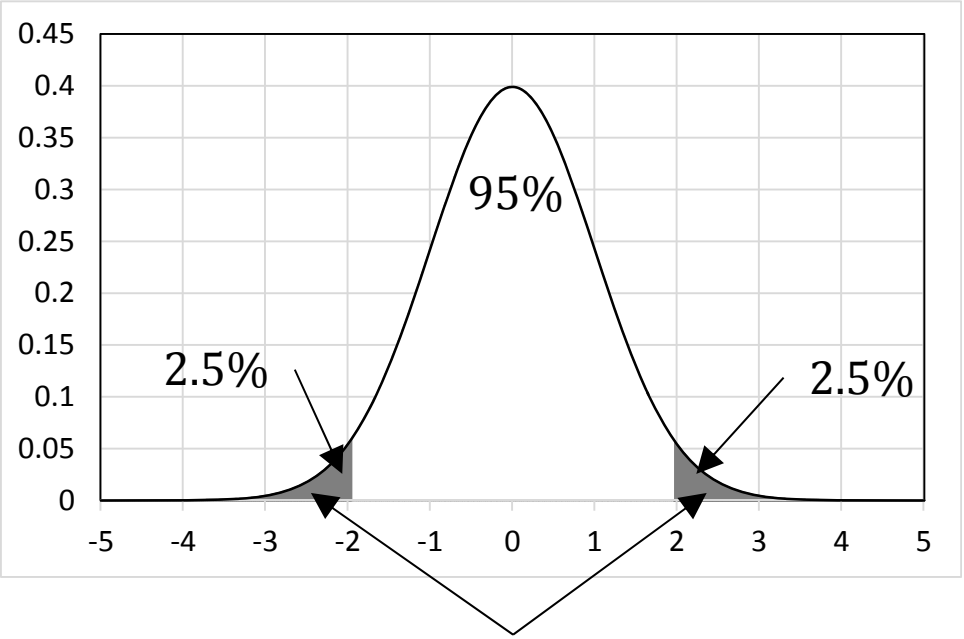
この節では、統計的有意性とは何であるかについて説明する

有意であること・有意でないこと

- **ある仮説に対応する確率がある有意水準（α）以下であるときに、その仮説は統計的に有意であるという**
  - 仮説が起こる確率を計算した結果、稀にしか起こらない水準であったようなときに用いる
  - 有意水準としては、10%、5%、1%を用いることが多い
  - 例：$P(\{\text{日本人のTOEFLのスコアが110以上}\}) \leq 5\%$だったとすると、「日本人のTOEFLのスコアが110以上」という仮説は、5%水準で統計的に有意といえる

有意であること・有意でないことのイメージ

- **ある仮説に対応する確率がある有意水準（α）以下であるときに、その仮説は統計的に有意であるという**

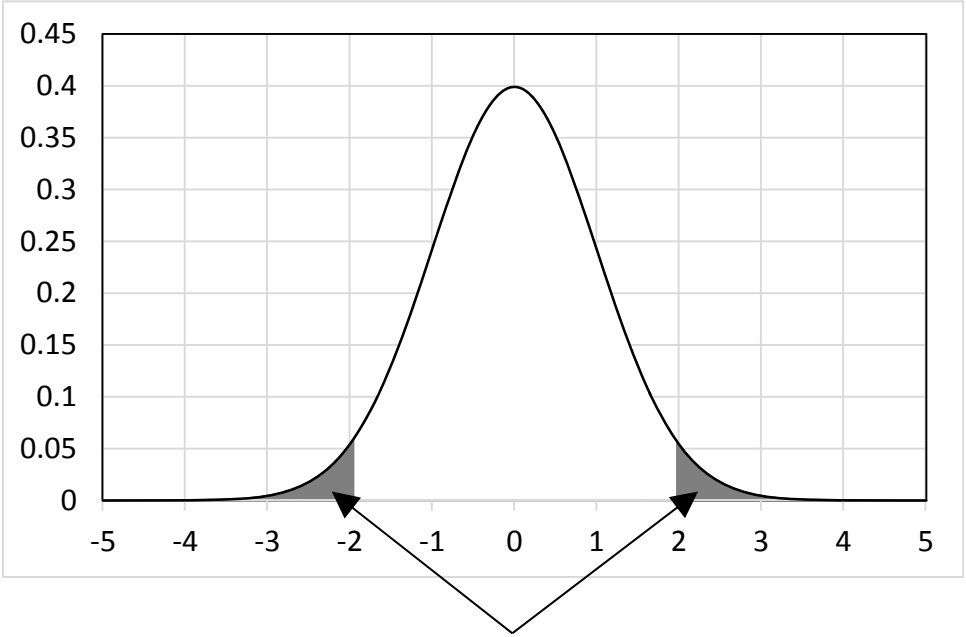


ありえないことが起こった！

11:29:40

有意であること・有意でないことのイメージ

- **ある仮説に対応する確率がある有意水準（α）以下であるときに、その仮説は統計的に有意であるという**



棄却域と呼ぶ
面積を足すとα（%）

11:29:41

統計的仮説検定

- **仮説とは、母集団のパラメータに関する主張**
  - 例えば、TOEFLを受けたすべての人の平均は50点
  - 全世界の人の身長の平均は165cm

11:29:42

統計的仮説検定の考え方

- 2つの仮説
  - 帰無仮説：H_0と表す
  - 対立仮説：H_1と表す

- **統計的仮説検定とは**
  - **標本から計算された結果により、帰無仮説H_0が正しいとして採択する**
  - **標本から計算された結果により、帰無仮説H_0が棄却され、対立仮説H_1が正しいと採択する**

  のいずれかを決めるルール

  - H_0が棄却されるような標本空間の部分集合を棄却域、
  - その補集合（棄却域以外の区間）を採択域という

統計的仮説検定の具体的なステップ

1. 仮説を設定する
   - 帰無仮説と対立仮説

2. 仮説を検定するための検定統計量を決める
   - データがどのような分布に従い、それによって検定統計量がどのような分布に従うかについても気を付ける

3. 検定統計量の棄却域を決める
   - 有意水準を設定する

4. 検定統計量の値を計算して、棄却域に入るか否かを調べる
   - 棄却域に入れば、帰無仮説が棄却される
     - 対立仮説は採択される
   - 棄却域に入らなければ、帰無仮説は棄却されない
     - 対立仮説は採択されない

仮説の設定：帰無仮説と対立仮説

- 帰無仮説H_0と対立仮説H_1
  - 帰無仮説があり、その仮説を補完する仮説を対立仮説
- 「工場で作ったLED電球の寿命の平均は1万時間より長い」という仮説を検証する
  - 帰無仮説$H_0 : \mu = 10{,}000$（平均は1万時間である）
  - 対立仮説$H_1 : \mu > 10{,}000$（平均は1万時間より長い）
- 帰無仮説を棄却（否定）されれば、対立仮説は採択される

11:29:42

なぜ帰無仮説なのか

なぜ帰無仮説を棄却するような検定をするのか？

1. 仮説が正しくないことを検証するのは比較的簡単
   - 仮説が正しいことを検証するの難しい

なぜ帰無仮説なのか

2. 統計的仮説検定の2種類の過誤が存在する
   - 第1種の過誤：帰無仮説は正しいのに棄却する誤り
     - 正しくない対立仮説を採択する誤り
   - 第2種の過誤：対立仮説が正しいのに、帰無仮説を棄却しない誤り
     - 正しいはずの対立仮説を採択しないという誤り

		本当の状態	
		帰無仮説は正しい	帰無仮説は正しくない
帰無仮説の棄却の結果	帰無仮説を棄却	第1種の過誤	○
	帰無仮説を棄却せず	○	第2種の過誤

- 第1種の過誤を起こる確率が有意水準α
  - よって有意水準をコントロールすれば、第1種の過誤をコントロールできる
  - 第2種の過誤が起こる確率はβと呼び、$1-\beta$を検出力と呼ぶ
    - ここでは第2種の過誤と検出力などについて省略する

11:29:42

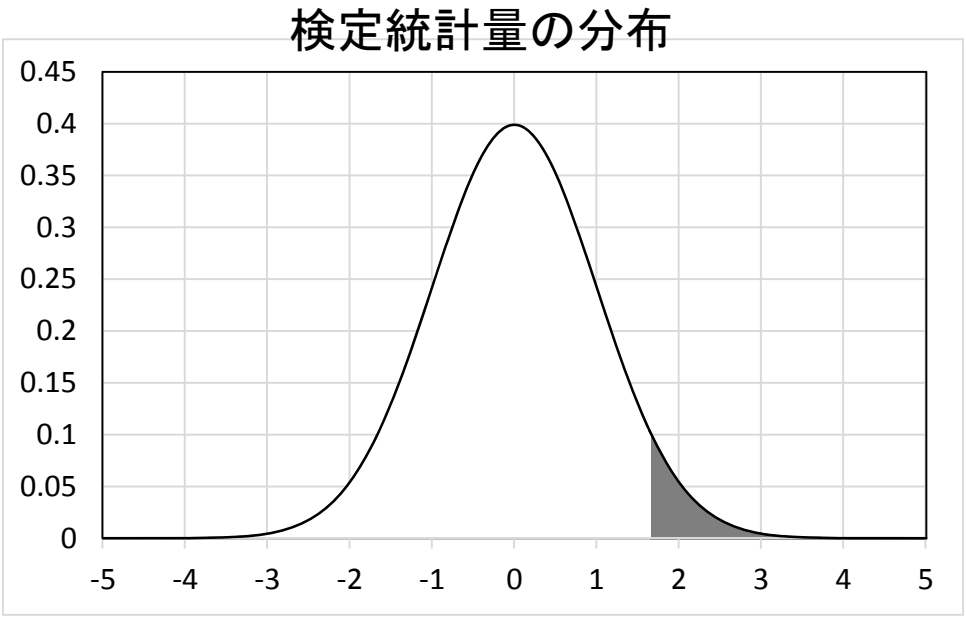
検定統計量

- **標本の関数を検定統計量という**
  - 標本から計算された値
  - 例えば、APUの学生をランダムにピックアップして、TOEFLの点数の平均
  - 検定の手がかりとなる統計量

- 「工場で作ったLED電球の寿命の平均は1万時間より長い」という仮説を検証する場合、
  - 母平均μに関する仮説検定は標本平均$\bar{X}$を使う

検定統計量の分布の計算

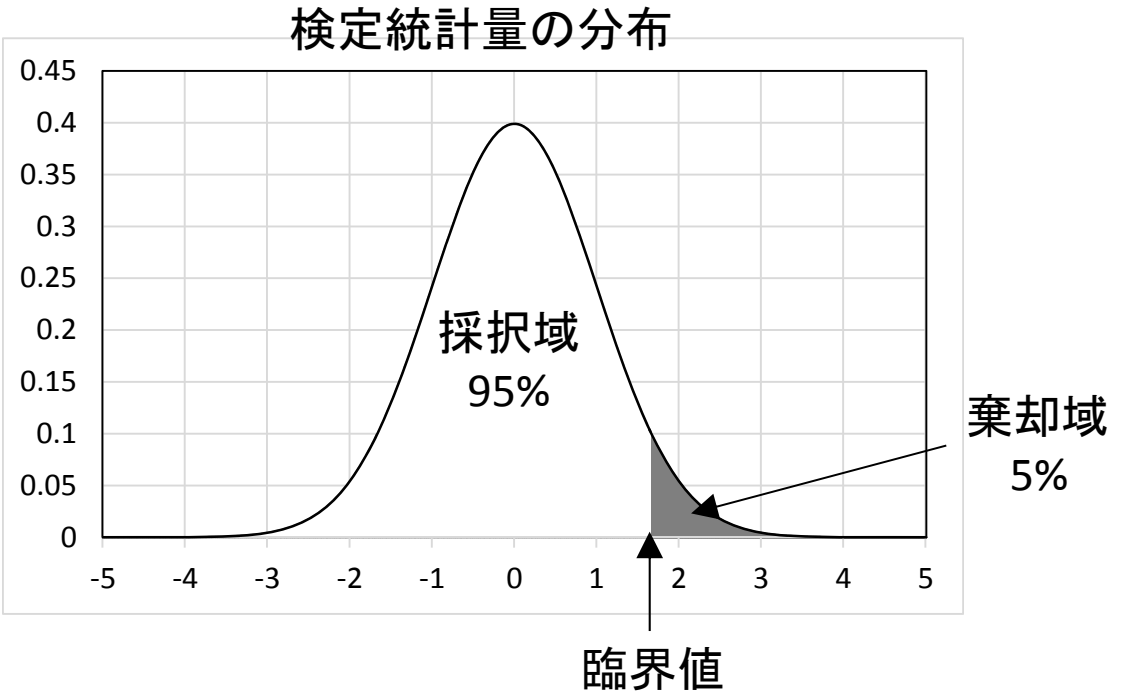
- 帰無仮説が正しいという前提で、検定統計量の分布を計算する



11:29:43

検定統計量の分布の計算棄却域

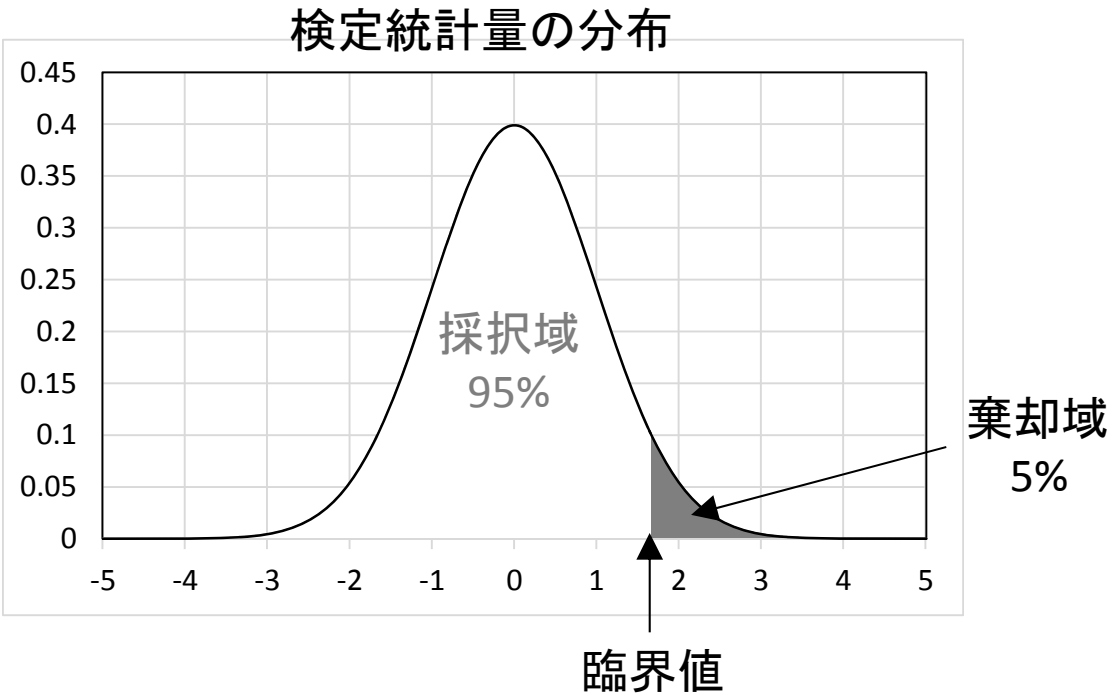
- 検定統計量を計算して、それが棄却域に入れば、帰無仮説は棄却される
  - 棄却域とは、帰無仮説が棄却される検定統計量が取りうる区間
  - 採択域とは、帰無仮説が棄却されない検定統計量が取りうる区間
  - 臨界値とは、棄却域と採択域を区切る値



臨界値の求め方

- あるデータが正規分布$N(\mu, \sigma^2)$に従うとする

1. 有意水準5%を設定する
   - 1%、5%、10%などがよく用いられている
2. 確率5%に対応する分布から求められる臨界値を計算する
3. データの水準とばらつきを戻す
   - $Z = \frac{X-\mu}{\sigma} \Rightarrow X = \sigma Z + \mu$



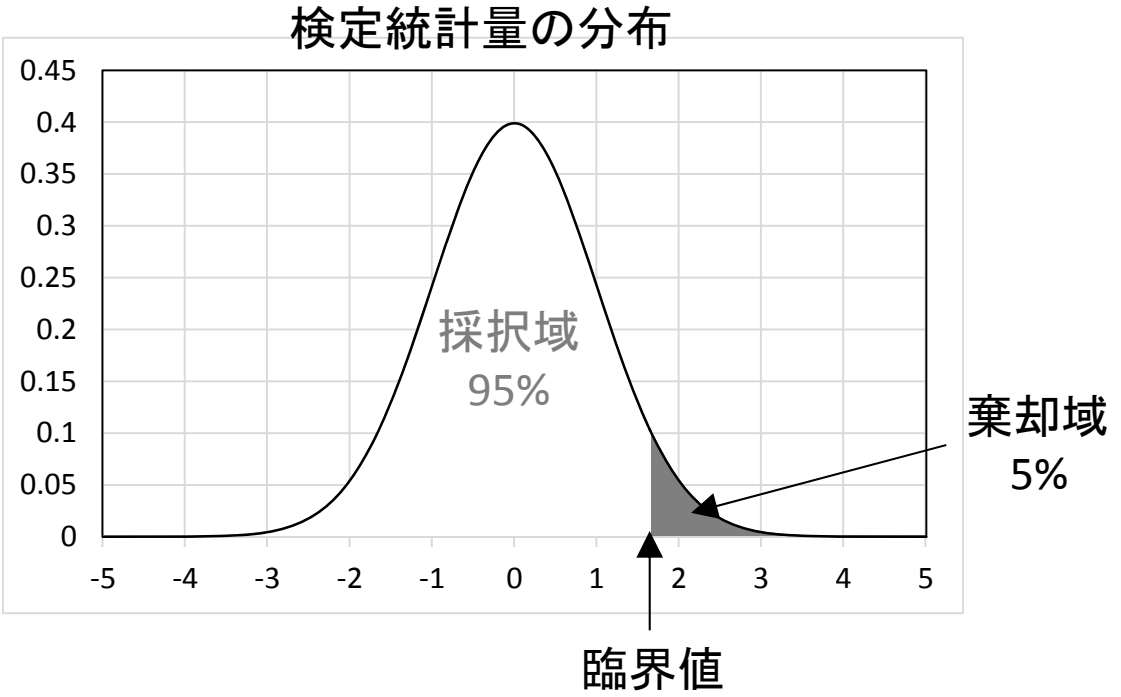
11:29:44

臨界値の求め方：例

- TOEFLのテストの結果の分布が正規分布$N(50,100)$に従うとする
  - 標準偏差は10

1. 有意水準5%を設定する
2. 確率5%に対応する標準正規分布の臨界値は1.64
   - ExcelでNORM.INV(0.95,0,1)と入力（この場合、上側5%なので、95%と入力）

3. $X = 10 \times 1.64 + 50 = 66.4$
   - 66.4点が臨界値



11:29:45

母平均に関する仮説検定

1. 母標準偏差が既知：標本平均をz変換したものは標準正規分布に従う

2. 母標準偏差が未知
   i. 小標本（データが600より少ない場合）：標本平均をt変換したものはt分布に従う
      - 正確には自由度が600より少ない場合
   ii. 大標本（データが600以上の場合）：標本平均をz変換したものは標準正規分布に従う
      - 正確には自由度が600以上の場合

母平均に関する仮説検定：母標準偏差が既知の場合

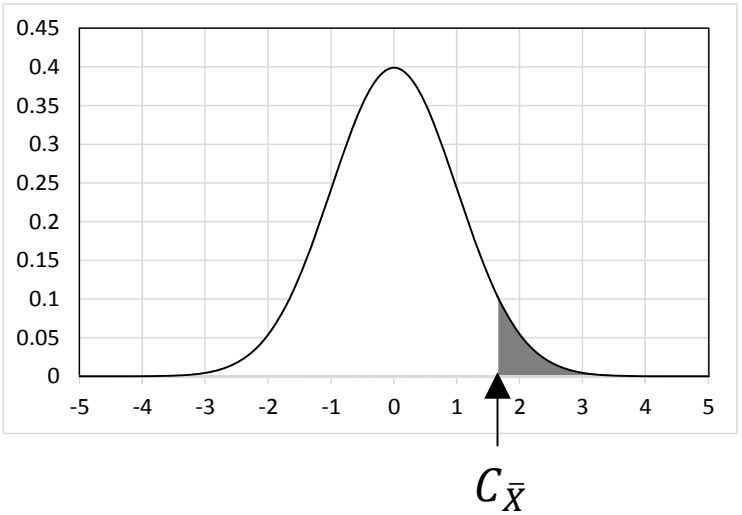
- 母平均μの推定は下方限界を利用

$$\mu = \bar{X} - \frac{z\sigma}{\sqrt{n}} \quad \Leftrightarrow \quad \bar{X} = \mu + \frac{z\sigma}{\sqrt{n}} \quad \Leftrightarrow \quad z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

  - σは母標準偏差
- 臨界値を代入すると、

$$C_{\bar{X}} = \mu + \frac{C_z\sigma}{\sqrt{n}}$$

  - $\bar{X}$の臨界値$C_{\bar{X}}$
  - zの臨界値C_z：例えば、5%有意水準のzの値
    - zは標準正規分布に従う



- $\bar{X} > C_{\bar{X}}$ならば、帰無仮説は棄却される
  - 対立仮説は採択される
- $\bar{X} \leq C_{\bar{X}}$ならば、帰無仮説は棄却されない
  - 対立仮説は採択されない

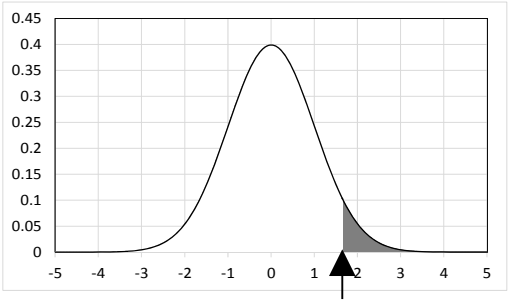
11:29:46

母平均に関する仮説検定：母標準偏差が既知の場合

- 100人の学生がTOEFLを受けた結果、点数の平均（標本平均）は53.45点
- 仮説として、
  - $H_0: \mu = 50$　帰無仮説
  - $H_1: \mu > 50$　対立仮説
- 母標準偏差は12.03
- 有意水準5%と設定
- zの5%有意水準の臨界値は$C_z = 1.645$

$$C_{\bar{X}} = \mu + \frac{C_z \sigma}{\sqrt{n}} = 50 + \frac{1.645 \times 12.03}{\sqrt{100}} = 51.978935$$

- $53.45 > 51.978935 (\bar{X} > C_{\bar{X}})$であるため、帰無仮説（$\mu = 50$）は棄却



$C_{\bar{X}}$

母平均に関する仮説検定：
母標準偏差が未知かつ小標本（データ少ない）の場合

- 母平均μの推定は下方限界を利用

$$\mu = \bar{X} - \frac{ts}{\sqrt{n}} \quad \Leftrightarrow \quad \bar{X} = \mu + \frac{ts}{\sqrt{n}} \quad \Leftrightarrow \quad t = \frac{\sqrt{n}(\bar{X} - \mu)}{s}$$

  - sは標本標準偏差
- 臨界値を代入すると、

$$C_{\bar{X}} = \mu + \frac{C_t \sigma}{\sqrt{n}}$$

  - tの臨界値C_t：例えば、5%有意水準のtの値
    - tはt分布に従う

- $\bar{X} > C_{\bar{X}}$ならば、帰無仮説は棄却される
  - 対立仮説は採択される
- $\bar{X} \leq C_{\bar{X}}$ならば、帰無仮説は棄却されない
  - 対立仮説は採択されない

11:29:47

母平均に関する仮説検定：
母標準偏差が未知かつ小標本（データ少ない）の場合

- 100人の学生がTOEFLを受けた結果、点数の平均（標本平均）は53.45点
- 仮説として、
  - $H_0: \mu = 50$　帰無仮説
  - $H_1: \mu > 50$　対立仮説
- 標本標準偏差は14.52
- 有意水準5%と設定
- tの5%有意水準の臨界値は$C_t = 1.66$

$$C_{\bar{X}} = \mu + \frac{C_t s}{\sqrt{n}} = 50 + \frac{1.66 \times 14.52}{\sqrt{100}} = 52.41032$$

- $53.45 > 52.41032 (\bar{X} > C_{\bar{X}})$であるため、帰無仮説（$\mu = 50$）は棄却

母平均に関する仮説検定：
母標準偏差が未知かつ大標本（データ多い）の場合

- 母平均μの推定は下方限界を利用

$$\mu = \bar{X} - \frac{zs}{\sqrt{n}} \quad \Leftrightarrow \quad \bar{X} = \mu + \frac{zs}{\sqrt{n}} \quad \Leftrightarrow \quad z = \frac{\sqrt{n}(\bar{X} - \mu)}{s}$$

  - sは標本標準偏差
- 臨界値を代入すると、

$$C_{\bar{X}} = \mu + \frac{C_z s}{\sqrt{n}}$$

  - zの臨界値C_z：例えば、5%有意水準のtの値
    - zは標準正規分布に従う

- $\bar{X} > C_{\bar{X}}$ならば、帰無仮説は棄却される
  - 対立仮説は採択される
- $\bar{X} > C_{\bar{X}}$ならば、帰無仮説は棄却されない
  - 対立仮説は採択されない

母平均に関する仮説検定：
母標準偏差が未知かつ大標本（データ多い）の場合

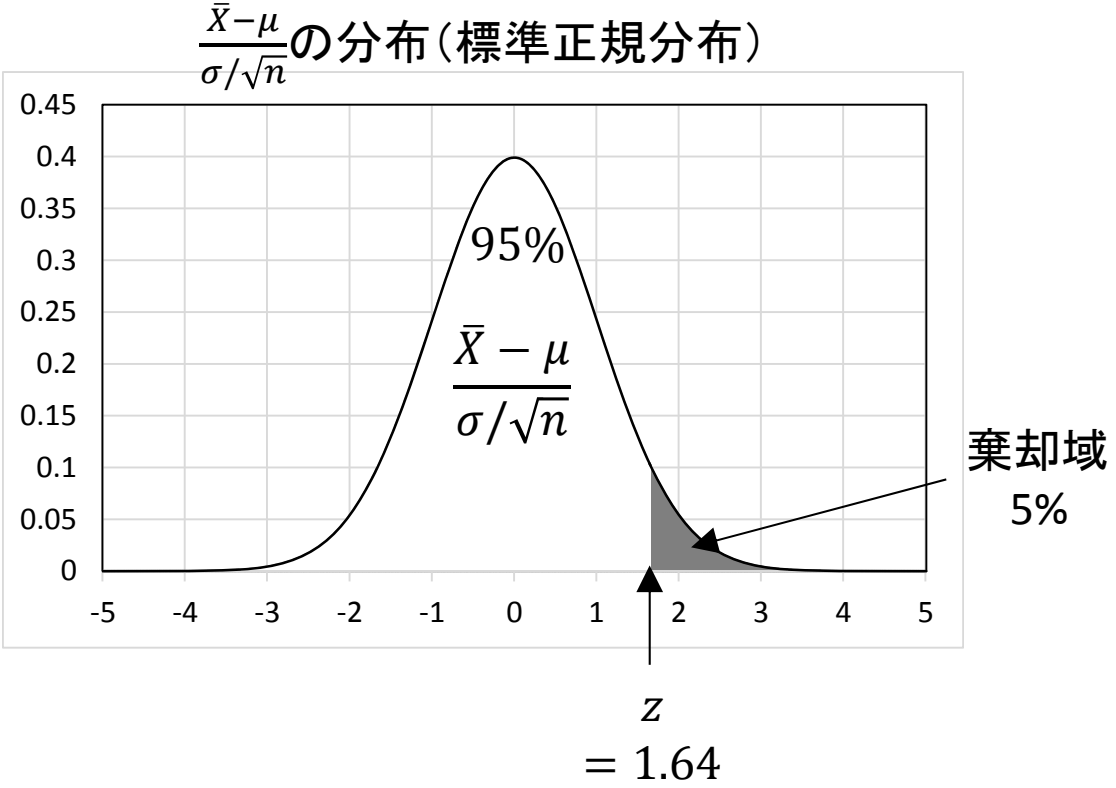
- 600人の学生がTOEFLを受けた結果、点数の平均（標本平均）は53.45点
- 仮説として、
  - $H_0: \mu = 50$　帰無仮説
  - $H_1: \mu > 50$　対立仮説
- 標本標準偏差は14.52
- 有意水準5%と設定
- zの5%有意水準の臨界値は$C_z = 1.645$

$$C_{\bar{X}} = \mu + \frac{C_z \sigma}{\sqrt{n}} = 50 + \frac{1.645 \times 14.52}{\sqrt{600}} = 50.9752$$

- $53.45 > 50.9752$ $(\bar{X} > C_{\bar{X}})$であるため、帰無仮説（$\mu = 50$）は棄却

正規分布における信頼係数に対するzの値：右側検定

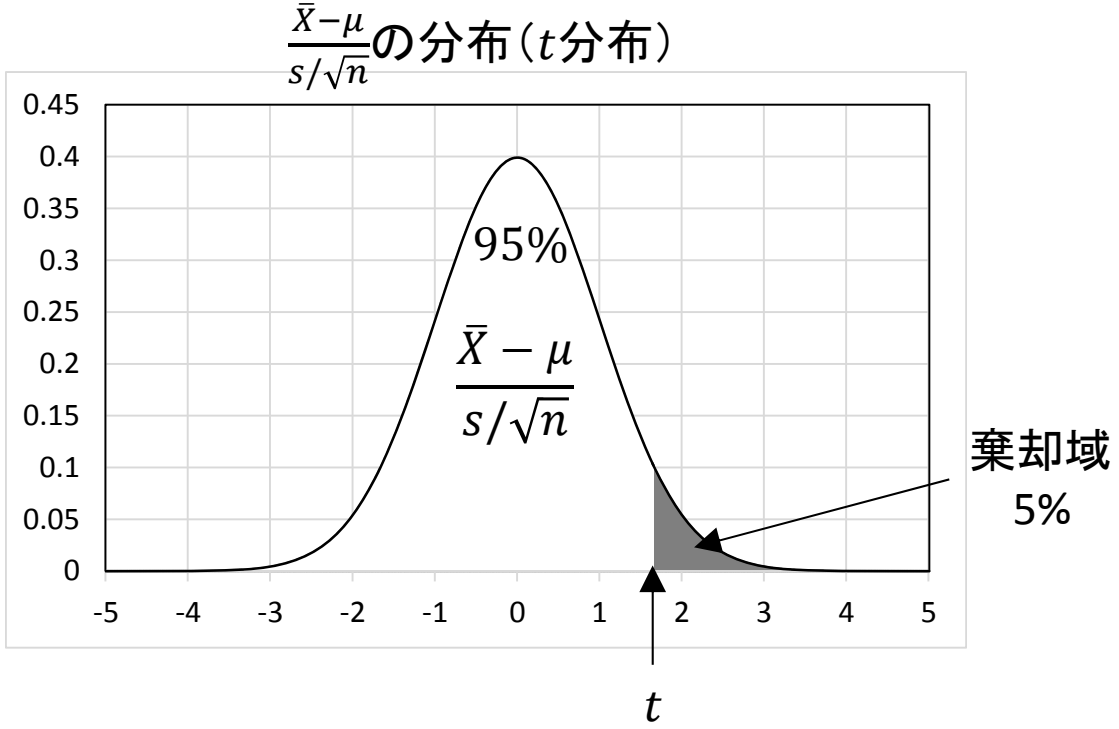
	信頼係数	z値
10%	90%	1.28
5%	95%	1.64
1%	99%	2.33



11:29:48

t分布における信頼係数に対するtの値：右側検定

- 自由度によって、t分布の信頼係数に対するtの値が変わる

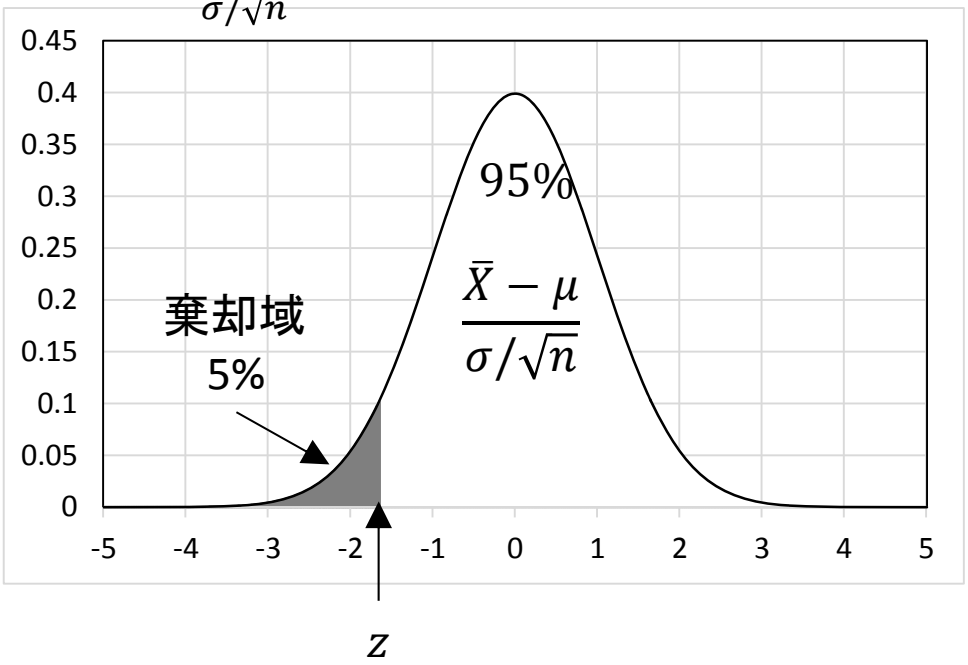


11:29:49

正規分布における信頼係数に対するzの値：左側検定

	信頼係数	z値
10%	90%	-1.28
5%	95%	-1.64
1%	99%	-2.33

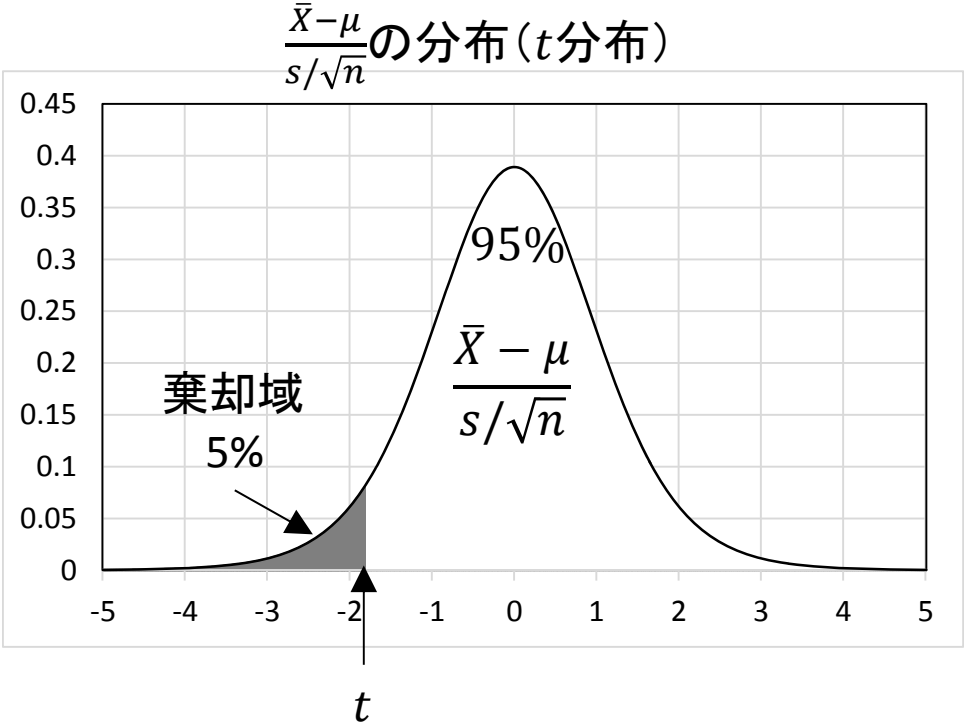
$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$の分布（標準正規分布）



11:29:49

t分布における信頼係数に対するtの値：左側検定

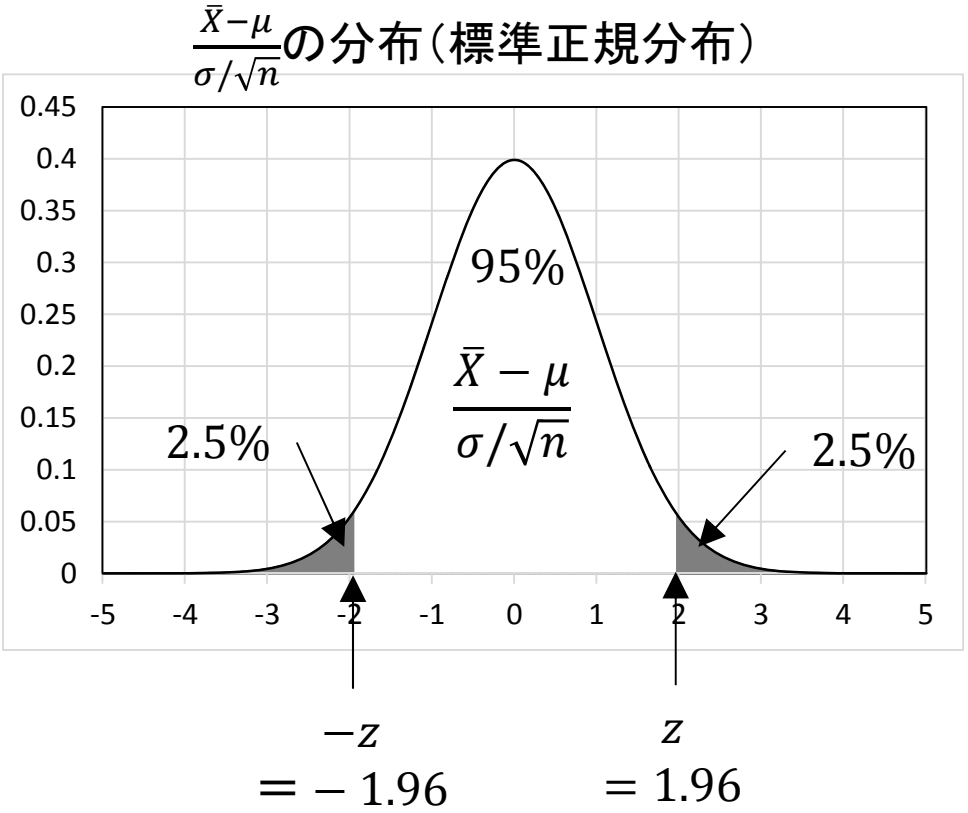
- 自由度によって、t分布の信頼係数に対するtの値が変わる



11:29:50

正規分布における信頼係数に対するzの値：両側検定

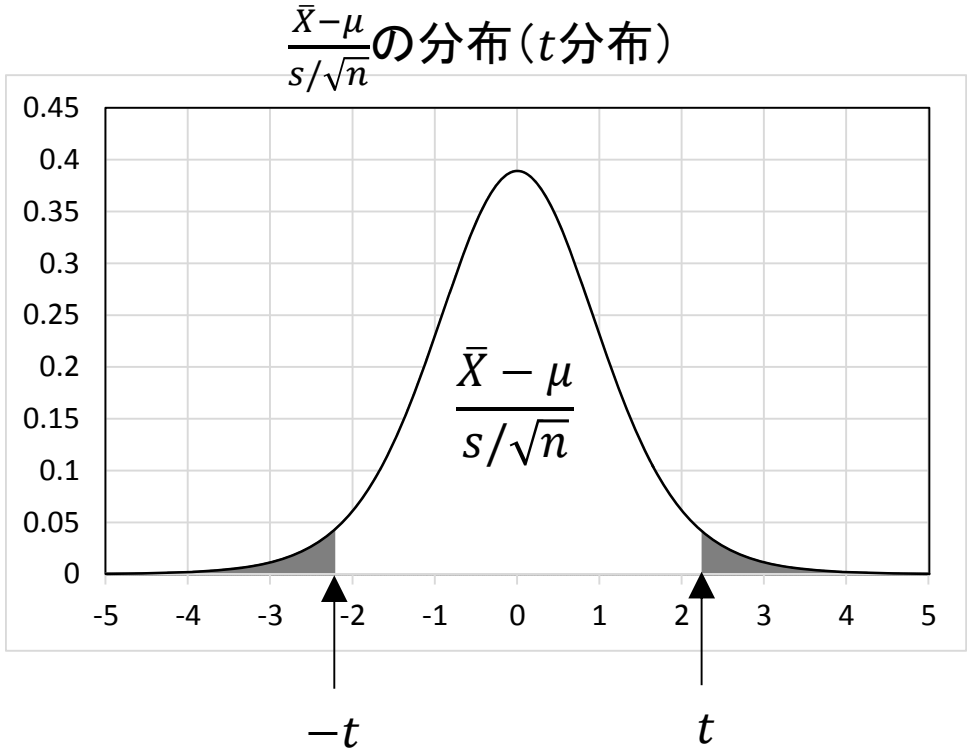
	信頼係数	z値
10%	90%	1.64
5%	95%	1.96
1%	99%	2.58



11:29:51

t分布における信頼係数に対するtの値：両側検定

- 自由度によって、t分布の信頼係数に対するtの値が変わる



11:29:52

YOUR TURN