統計基礎

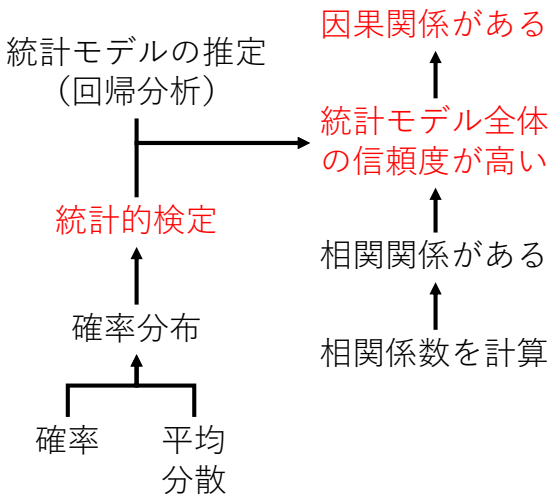
古谷知之

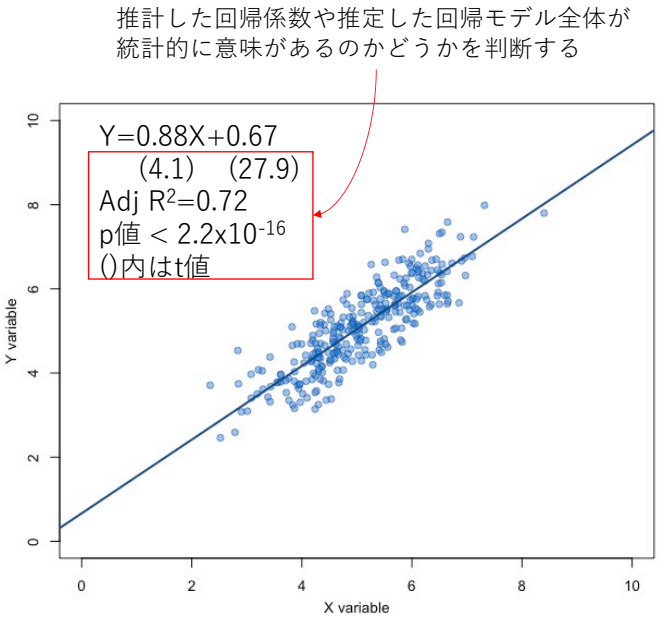
授業概要

＊履修者の状況に応じて変更される場合がありますが、概ね以下のような流れで授業を進めます。

第1回　ガイダンス・確率
第2回　確率変数と確率分布(1)
第3回　確率変数と確率分布(2)
第4回　母集団と平均
第5回　単回帰分析(1)
第6回　単回帰分析(2)
第7回　仮説検定(1)
第8回　仮説検定(2)
第9回　重回帰分析
第10回　R演習(1)
第11回　R演習(2)
第12回　R演習(3)
第13回　R演習(4)
第14回　最終試験

授業の全体像





授業内容

- 単回帰分析
- 回帰係数の統計量
- 残差の性質
- 回帰係数に関する検定
- 信頼区間
- t分布
- 仮説検定
- 帰無仮説と対立仮説
- 有意水準
- 統計的過誤
- 両側検定と片側検定

単回帰分析で検討すべきこと

- 回帰係数の値：偏回帰係数
- 回帰係数の統計的有意性：t検定
- 回帰係数の信頼度：信頼区間
- 予測への適用可能性：決定係数
- 外れ値の検出：残差解析

最小二乗法による回帰係数の推計

- 単回帰モデルの回帰係数$\hat{\beta}_1$と$\hat{\beta}_0$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\{(x_i - \bar{x})(y_i - \bar{y})\}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

- 回帰係数は不偏推定量であり、一致推定量である
- 回帰係数は正規分布する確率分布の平均値でもある

回帰係数の統計量

- 線形回帰分析では誤差ε_iが正規分布に従うという強い仮定をおいている。誤差項の分散は未知であるが、標本毎に同じであるとも仮定している。すなわち、

$$\varepsilon_i \sim N(0, \sigma^2)$$

- この仮定のもとでは、回帰係数の不偏推定量が以下のように正規分布するという性質を持つ

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right)$$

- 不偏推定量は回帰係数の分布の平均値と考えることもできる。このとき、一致推定量の性質とも合致する

回帰係数の統計量

- 回帰係数$\hat{\beta}_1$の分散の平方根を標準誤差 (Standard Error)という

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

- 標準偏差が標本のばらつきを意味するのに対し、標準誤差は統計量のばらつきを意味する

残差の性質

- 推定された回帰モデルの残差$e_i = y_i - \hat{y}_i$（実績値と予測値との差分）は、回帰分析で説明がつかない部分を意味する
- 残差e_iは互いに独立に（各i毎に）平均0、分散σ^2の正規分布に従う。$e_i \sim N(0, \sigma^2)$
- 残差平方和S_eは次のように求められる

$$S_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

- これの自由度は$n-2$なので、回帰の残差分散s_e^2は

$$s_e^2 = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n-2}$$

となる

残差の性質

- 更に残差$e_i = y_i - \hat{y}_i$は、以下の 2 つの性質を持つ

- 残差の和は0になる

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

- 残差と従属変数とは直交する

$$\sum_{i=1}^{n} e_i x_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) x_i = 0$$

残差の性質

- 誤差εが正規分布に従う仮定$\varepsilon \sim N(0, \sigma^2 I)$の下で、残差$e$も正規分布に従う

$$e \sim N\left(0, \sigma^2 (I - X(X^T X)^{-1} X^T)\right)$$

$$X^T = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}$$

- ここで、 $H = X(X^T X)^{-1} X^T$はハット行列と呼ばれる

残差の性質

- 残差を標準偏差で基準化した残差e_i/σの平方和は、自由度$n - k - 1$（単回帰分析の場合は$n - 2$）のχ^2分布に従う

$$\sum_{i=1}^{n} \left(\frac{e_i}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2(n - k - 1)$$

- （復習）残差$\varepsilon_i \sim N(0, \sigma^2)$より、$e_i/\sigma \sim N(0,1)$となることから、その二乗値$\left(\frac{e_i}{\sigma}\right)^2$は$\chi^2$分布に従う

回帰係数に関する検定

- 回帰係数が正規分布するという性質を標準化すると、以下のような性質が得られる

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim N(0,1)$$

- ここで、$\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$は$\beta_1$の標準誤差である

回帰係数に関する検定

- 残差の性質から以下の分布は自由度$n-2$のt分布に従う

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/\sum_{i=1}^{n}(x_i - \bar{x})^2}} \Big/ \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2(n-2)}}$$
$$= \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sum_{i=1}^{n} e_i^2/(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim t(n-2)$$

- ここで残差の制約条件が$\sum_{i=1}^{n} e_i = 0$と$\sum_{i=1}^{n} e_i x_i = 0$の２つあることから、自由度は2下がっている

回帰係数に関する検定

- この統計量を用いて、回帰係数に意味があるのかを検証できる
- 「回帰係数は従属変数の説明に寄与していない」という帰無仮説を棄却できれば、「回帰係数に意味がないとは言えない」ということになる→「回帰係数に意味がある」ことを直接照明するのではなく、「回帰係数に意味がない」ことを否定する

- このとき帰無仮説は以下のようになる

$$H_0: \beta_1 = 0$$

回帰係数に関する検定

- 誤差の分散σ^2をその普遍推定量$\hat{\sigma}^2 = \left(\sum_{i=1}^{n} e_i^2/(n-2)\right)^2$で置き換えると、以下のような関係性が導ける

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sum_{i=1}^{n} e_i^2/(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

回帰係数に関する検定

- 同様に、$H_0: \beta_0 = 0$に対する仮説検定は、

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} e_i^2 / n(n-2) \sum_{i=1}^{n} (x_i - \bar{x})^2}} \sim t(n-2)$$

についてt検定を行えば良い

回帰係数は統計的に信頼できるのか？

- 回帰係数$\hat{\beta}$が平均μ、分散σ^2の正規分布$N(\mu, \sigma^2)$に従う確率変数であるとき、 $\mu - 1.96\sigma^2 < \hat{\beta} < \mu + 1.96\sigma^2$が成立する確率は95%である。これを95%信頼区間という（復習）。

- ただしこのことが成立するのは、標本数nが非常に大きいとき

- 十分に大きな標本数nが得られないとき、この信頼区間はどのようにして得ることができるのか？
- また実際には、回帰係数の平均も分散も事前にはわからない

回帰係数の信頼区間

- （平均値としての）回帰係数の統計的信頼度を知りたい
- 正規分布に基づく統計的信頼区間を用いるには、回帰係数の平均と分散を知っていなくてはならない（正規分布の限界）
- しかし平均が事前には分からないので分散も分からない
  →正規分布による信頼区間をそのまま用いることができない
- シミュレーション（ベイズ推定）を用いて正規分布の平均と分散を推計することができるが、ここでは古典的な統計学のマナーに従って、正規分布に近い確率分布（t分布）を用いる

仮説検定

- 回帰係数が統計的に意味がある（有意である）ことを示すために、仮説検定と呼ばれる手法が用いられることがある

手順

- 回帰係数の分布に何らかの仮定を置く
- 回帰係数と回帰係数の分布との間に検定したい仮説を設定する
- 帰無仮説：回帰係数=0、対立仮説：回帰係数=0ではない
- 回帰係数を推計する
- 推計した回帰係数のp値を計算する
- p値が統計的有意水準α以下であれば帰無仮説を棄却する（有意水準αより大きければ回帰係数に統計的な意味はない）

仮説検定

- 一般に仮説検定は、母集団から抽出された標本について、標本の統計的性質が母集団の統計的性質と「異なる」ことを示すために用いられる

仮説検定の手順

- 検定したい統計量に何らかの確率分布を仮定する
- 「現実には起こりえない」とみなす統計的有意水準αを設定する帰無仮説と対立仮説を設定する
  - 帰無仮説H_0：棄却したい仮説、対立仮説H_1：帰無仮説と対立する仮説
- 仮定した確率分布に基づき統計量を推計する
- 帰無仮説H_0が正しい場合に標本が得られる確率p値を計算する
- p値が有意水準αより小さければ帰無仮説を棄却する

t分布

- 「帰無仮説H_0が正しい場合に標本が得られる確率p値」はどのようにして得られるか？
- 母集団から抽出された標本数nが大きく母分散σ^2が既知のときには正規分布を用いれば良い
- 標本数nが少なく母平均μが既知だが母分散σ^2が未知のときには、不偏分散s^2を用いることで、以下の統計量t値は自由度$\nu = n - 1$のt分布$t(\nu)$に従うと仮定する

$$t = \frac{\bar{X} - \mu}{\sqrt{s^2/n}}$$

- 不偏分散$s^2 = \frac{n-1}{n}\sigma^2$は母分散$\sigma^2$より少し小さい
- 不偏分散s^2は自由度$n-1$のχ^2分布に従う

t分布

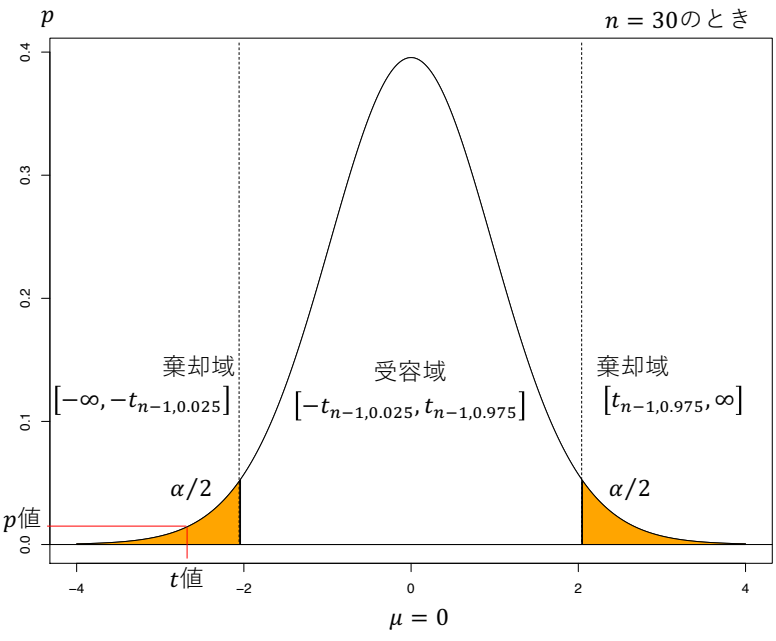
- 確率変数Xが自由度$\nu = n - 1$のt分布$t(\nu)$に従う$X \sim t(\nu)$とき、その確率関数$f(x|\nu)$は以下のように表される

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

- 一般に$\Gamma(z)$はガンマ関数と呼ばれ、 $\Gamma(z) = (z-1)!$
- 期待値$\mathrm{E}[t(\nu)] = 0$
- 分散$\mathrm{V}[t(\nu)] = \frac{\nu}{\nu-2}, \nu > 2$

t分布とp値

- t分布に従う確率変数は、受容域$[-t_{n-1,0.025}, t_{n-1,0.975}]$の中にデータの95%が収まる（$\alpha = 0.05$のとき）
- t値による検定ではαを事前に定めた上で$|t| > t_{n-1,\alpha/2}$かを判断する
- 得られたt値についてt分布上で対応する確率をp値という
- t値の絶対値が十分に大きいときp値は小さくなり、帰無仮説は棄却される
- 帰無仮説が正しいとき、 p値は大きい値をとる
- p値による検定では異常性をp値で確率的に示す



t分布とp値

- 実用的には、以下のように理解しておけば良い
- 標本数が非常に多い場合、t分布は正規分布に近づくので、95%信頼区間（$\alpha = 0.05$のとき）の絶対値が1.96より大きければ、t値が統計的に有意であると判断できる→「2 より大きければ良い」と理解しておく
- t分布は（所詮）ビッグデータやシミュレーションが無かった時代の産物なので、t分布の数式など覚えるには及ばない
- 最近の学術論文では、t値を用いた研究結果は学術論文として採用しないというものもある
- p値は様々な場面で出てくるので、理解しておくと良い

帰無仮説と対立仮説

- 帰無仮説H_0：否定したい仮説（実際に生じ得ない仮説）
- 対立仮説H_1：H_0と対立する仮説

- 有意水準αで統計的に有意 (statistically significant) であれば、帰無仮説H_0が棄却される
- 帰無仮説H_0が棄却されれば、対立仮説H_1の正当性を主張できる
- 帰無仮説H_0が棄却されなかったからと言って、帰無仮説H_0が正しいと主張することはできない

帰無仮説と対立仮説（犯罪発生）

- 帰無仮説H_0：容疑者Xは犯人ではない（無罪である）
- 対立仮説H_1：容疑者Xは犯人である（無罪ではない）

- 有意水準αで統計的に有意（statistically significant）であれば、帰無仮説H_0が棄却される
- 帰無仮説H_0が棄却されれば、対立仮説H_1の正当性（容疑者Xは犯人である）を主張できる
- 帰無仮説H_0が棄却されても、容疑者Xが犯人である可能性（冤罪となる）は否定できない（有意水準αより小さい確率で帰無仮説H_0が棄却できない可能性は否定できない）

帰無仮説と対立仮説（医療）

- 帰無仮説H_0：患者はある感染症で陽性ではない
- 対立仮説H_1：患者はある感染症で陽性である

- 有意水準αで統計的に有意（statistically significant）であれば、帰無仮説H_0が棄却される
- 帰無仮説H_0が棄却されれば、対立仮説H_1の正当性（患者は陽性である）を主張できる
- 帰無仮説H_0が棄却されても、患者が陽性である可能性（偽陽性となる）は否定できない（有意水準αより小さい確率で帰無仮説H_0が棄却できない可能性は否定できない）

帰無仮説と対立仮説（回帰係数）

- 帰無仮説H_0：不偏推定量$\hat{\beta} = 0$（不偏推定量は存在しない）
- 対立仮説H_1：不偏推定量$\hat{\beta} \neq 0$

- 有意水準$\alpha = 0.05$とする
- 標本数nのデータから回帰係数βを推計する
- 帰無仮説H_0が正しいときにp値を計算する

$$t = \frac{\hat{\beta} - 0}{\sqrt{s^2/n}}$$

- p値が有意水準αより小さければ、「帰無仮説H_0は生じ得ないこと」と判断して帰無仮説H_0を棄却し、対立仮説H_1を受容する

統計的過誤

- 第一種の過誤（偽陽性）は、有意水準αにより検出可能
- 対立仮説H_1が正しいときに第二種の過誤（偽陰性）を犯さない確率を検出力$1 - \beta$という
- 第二種の過誤を犯す確率βは直接計算できないが、有意水準αと帰無仮説H_0が主張する仮定と真実との解離度に依存する

	真実	
仮説検定の結果	H_0が正しい	H_1が正しい
H_0を棄却する	第一種の過誤 確率：有意水準α	正しい検出結果 確率：検出力$1 - \beta$
H_0を棄却しない	正しい検出結果 確率：$1 - \alpha$	第二種の過誤 確率：β

帰無仮説と対立仮説（回帰係数）

- 帰無仮説H_0：不偏推定量$\hat{\beta} = 0$（不偏推定量は存在しない）
- 対立仮説H_1：不偏推定量$\hat{\beta} \neq 0$

- 有意水準$\alpha = 0.05$のとき、単回帰モデルの不偏推定量の95%信頼区間は以下のようになる
- ただし$t_{\alpha/2}(n-k-1)$は独立変数の数kと定数項から自由度$(n-k-1)$のt値を意味する。

$$\hat{\beta} - t_{\alpha/2}(n-k-1)\sqrt{s^2/n} \leq \beta \leq \hat{\beta} + t_{\alpha/2}(n-k-1)\sqrt{s^2/n}$$

有意水準αの設定と統計的判断

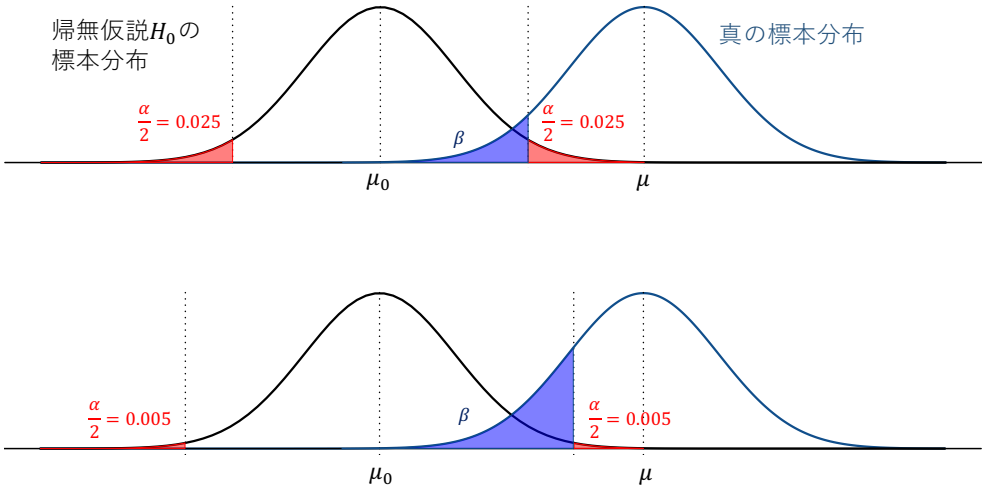
- 有意水準αは、分析者が分析の精度に応じて任意に設定できる
- この授業では$\alpha = 0.05$を多く使うが、それは信頼区間を95%に設定しておけば、そこから外れるような事象はほとんど起きないだろうと「勝手に」想定しているから
- 社会科学分野では便宜上$\alpha = 0.05$を多用するが、 $\alpha = 0.1$や$\alpha = 0.001$などであってもよい。
- 医療・薬学分野などでは、非常に小さい有意水準（$\alpha = 0.001$など）が採用されることがある
- 有意水準を小さくすれば、統計的判断が厳しくなり、例えば冤罪や偽陽性などの誤検出の可能性を少なくできる。他方、完全犯罪や偽陰性（検出失敗）を増やしてしまう。

統計的過誤

- 第一種の過誤（偽陽性）：帰無仮説H_0が正しいにもかかわらず帰無仮説H_0を棄却してしまう誤り
- 第二種の過誤（偽陰性）：帰無仮説H_0が正しくないにもかかわらず帰無仮説H_0を棄却できない誤り

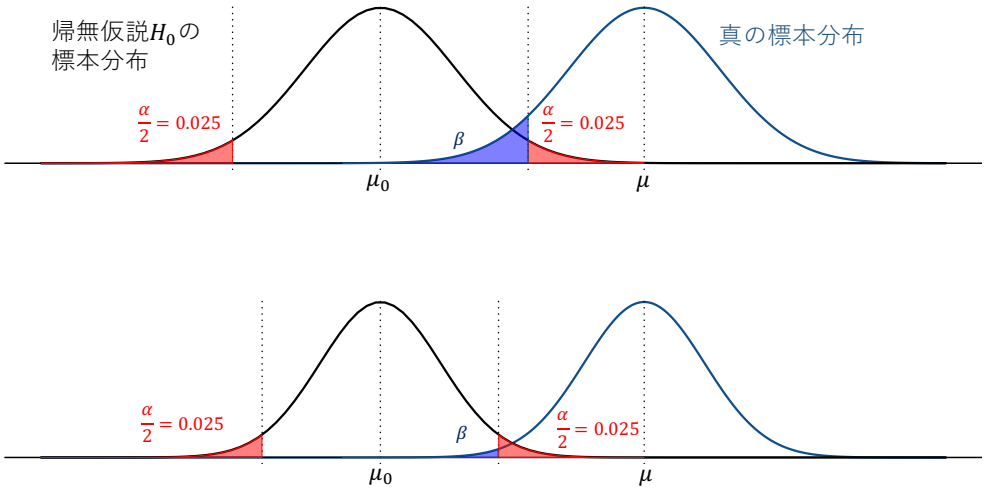
	真実	
仮説検定の結果	H_0が正しい	H_1が正しい
H_0を棄却する	第一種の過誤 確率：有意水準α	正しい検出結果 確率：検出力$1-\beta$
H_0を棄却しない	正しい検出結果 確率：$1-\alpha$	第二種の過誤 確率：β

統計的過誤と有意水準α



- 有意水準が厳しくなると、第一種の過誤の確率αは小さくなるが、第二種の過誤の確率βは大きくなる（トレードオフの関係）

統計的過誤と標本サイズ



- 帰無仮説の標本分布を知ることはできないのでを動かせないが、標本数を増やし分散を小さくすることで、αは一定でもβを小さくできる

両側検定と片側検定

- 帰無仮説H_0：$\mu = \mu_0$のとき対立仮説には以下の３つがあり得る

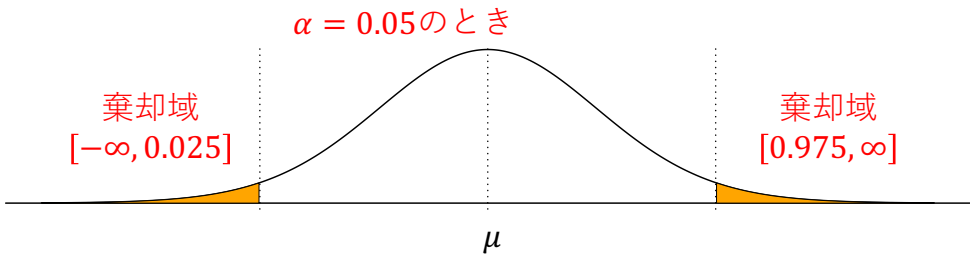
- 対立仮説H_{1a}：$\mu \neq \mu_0$
- 対立仮説H_{1b}：$\mu < \mu_0$
- 対立仮説H_{1c}：$\mu > \mu_0$

- 対立仮説の性質に応じて、両側検定（H_{1a}）か片側検定（H_{1b}、H_{1c}）かが決まる
- 片側検定を行う特段の理由がなければ両側検定を行う

両側検定の棄却域

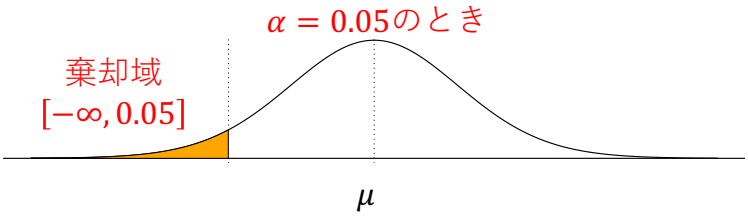
- 帰無仮説H_0：$\mu = \mu_0$
- 対立仮説H_{1a}：$\mu \neq \mu_0$
- 有意水準α

のとき、棄却域は$[-\infty, \alpha/2]$及び$[1-(\alpha/2), \infty]$となる



片側検定の棄却域

- 対立仮説H_{1b}：$\mu < \mu_0$のとき、棄却域は$[-\infty, \alpha]$となる



- 対立仮説H_{1c}：$\mu > \mu_0$のとき、棄却域は$[1-\alpha, \infty]$となる