統計基礎

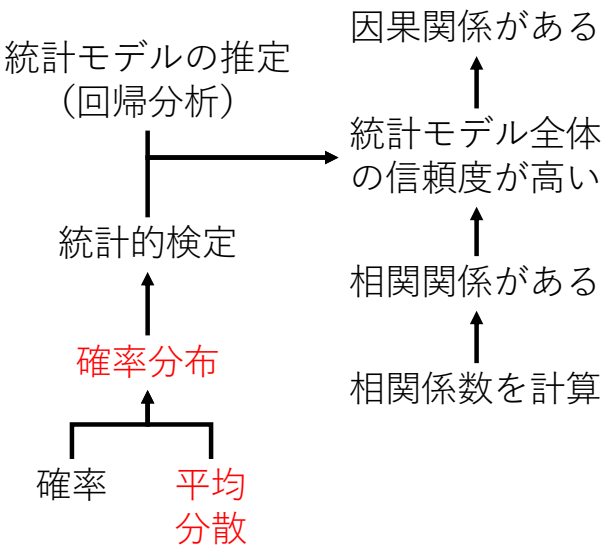
古谷知之

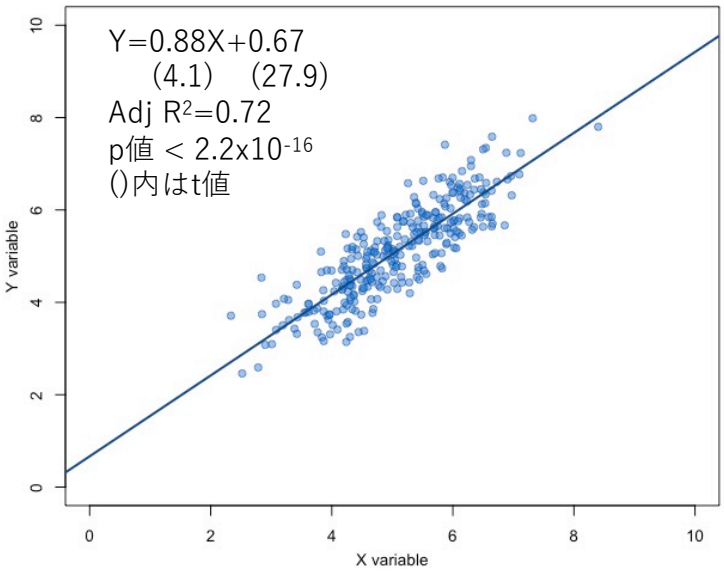
授業概要

＊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。

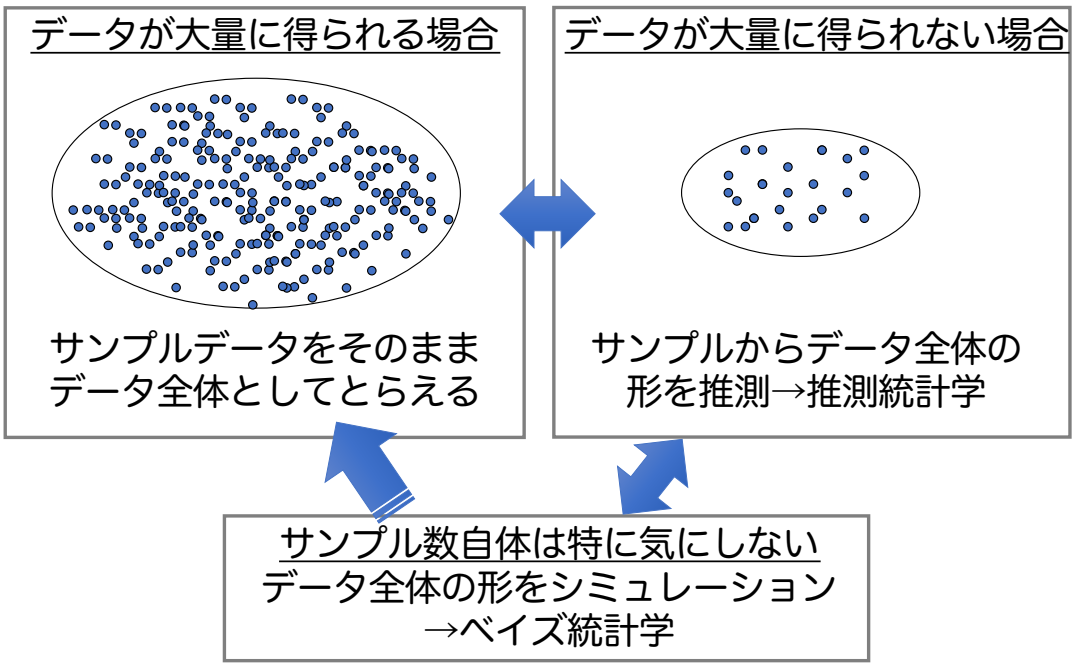
第1回　ガイダンス・確率
第2回　確率変数と確率分布(1)
第3回　確率変数と確率分布(2)
第4回　母集団と平均
第5回　単回帰分析(1)
第6回　単回帰分析(2)
第7回　仮説検定(1)
第8回　仮説検定(2)
第9回　重回帰分析
第10回　R演習(1)
第11回　R演習(2)
第12回　R演習(3)
第13回　R演習(4)
第14回　最終試験

授業の全体像





データサイエンスの基本的な考え方



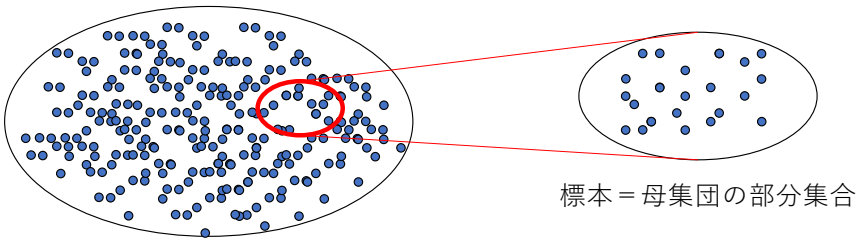
推測統計学における重要なキーワード

- 母集団と標本
- 標本平均と標本分散
- 不偏分散
- 自由度
- 大数の法則
- 中心極限定理
- 正規分布
- 信頼区間
- 標本の信頼度

今回はこれらすべてをまとめて
理解できるようにします

母集団と標本

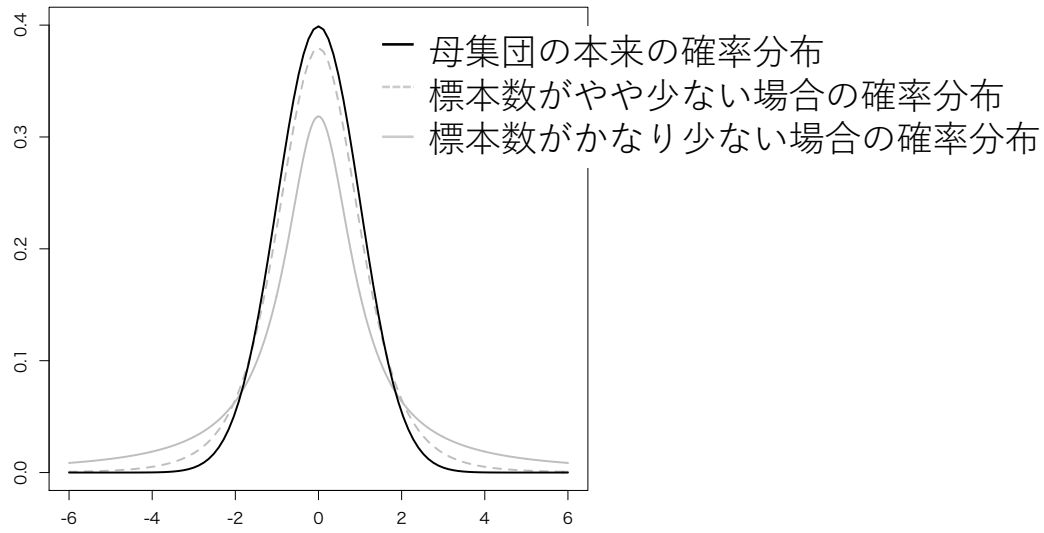
- 関心のある集合全体のことを母集団という
- 母集団の部分集合を標本（サンプル）という
- 母集団から標本を選ぶことを抽出といい、母集団のどの要素も同様に確からしく抽出されることを無作為抽出という



標本＝母集団の部分集合

母集団＝関心のある集合全体

標本数が少ないと…



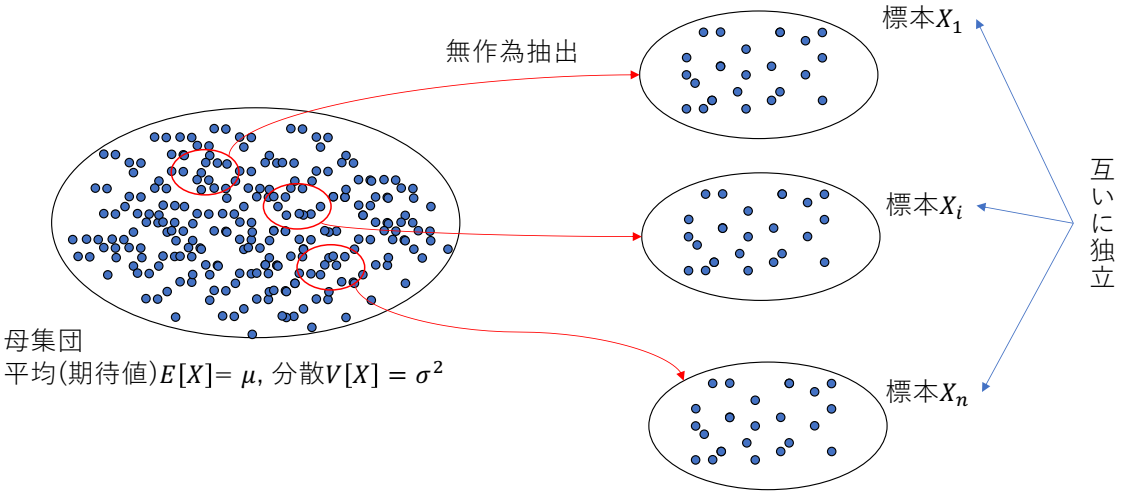
何をやりたいのか？

- 大量のデータが得られない場合、標本データから母集団の統計的性質を把握したい
- 標本データと母集団データとの統計的性質の関係を把握することで、標本データを用いた統計分析を有効なものとみなす

- そこでまず、標本と母集団の平均と分散（標準偏差）との関係について理解する
- どの標本を取り出しても母集団と同じ統計的性質をもつことが大事だが、その確証がないので、標本が母集団とどの程度異なる統計的性質を持つのかを把握する

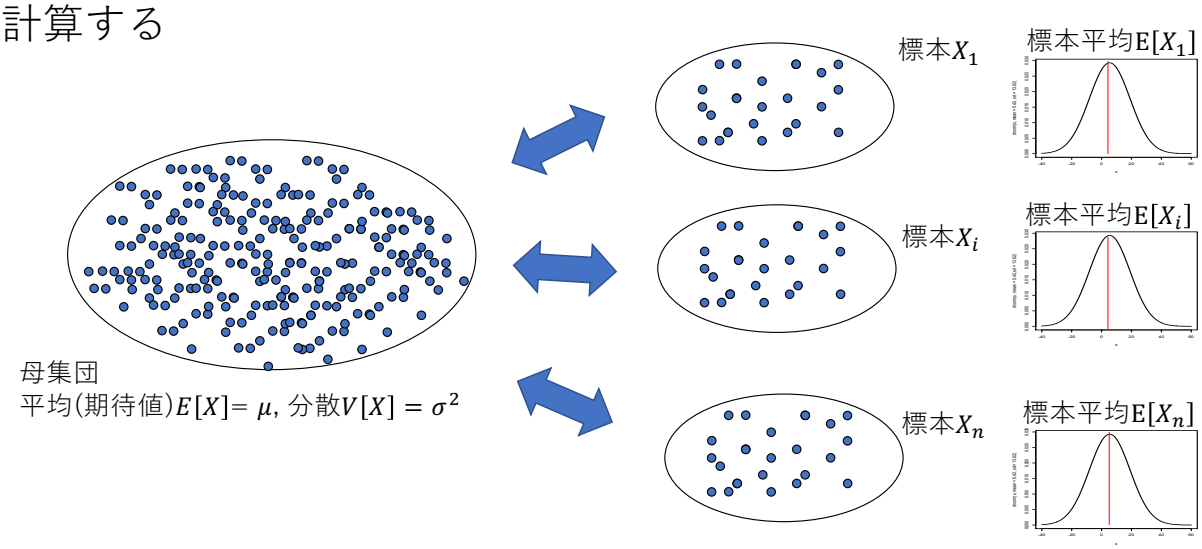
母集団の平均と標本の平均と分散

- 母集団から（互いに独立な）どの標本を（何度も）抽出しても、同じ様に母集団を再現できるのか？



母集団の平均と標本の平均と分散

- 母集団と標本との統計的性質の関係を把握する上で、各標本の平均と分散をそのまま扱うのではなく、標本平均の平均と分散を計算する



標本の平均

- 根元事象ωであるn個の標本$\{X_1, \dots, X_n\}$について、実現値$\{x_1, \dots, x_n\}$が得られたとする
- 標本$\{X_1, \dots, X_n\}$の平均$\overline{\mathrm{X}}$を以下の確率変数で捉えることにする

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + \cdots + X_n)$$

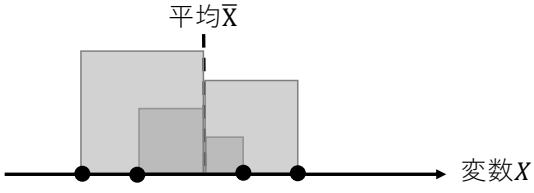
- ここで、確率変数X_iは同じ母集団から無作為抽出されたものとする。さらに、任意の確率変数X_iとX_jは互いに独立とする

標本の分散

- 標本$\{X_1, \dots, X_n\}$の分散s^2は平均$\overline{\mathrm{X}}$と標本X_iとの解離度$(\overline{\mathrm{X}} - X_i)$の平方和を標本数で割った値となる

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left((\overline{\mathrm{X}} - X_i)\right)^2$$

- イメージ的には、平均と各標本との差分を一辺とする正方形の面積の総和を標本数で割った値



- 母分散σ^2と標本分散s^2との関係性については後ほど考察する

標本平均の平均

母集団の期待値$E[X]=\mu$、分散$Var[X]=\sigma^2$とすると、
標本X_iの期待値は標本平均$E[X_i]=\mu$となる
標本平均$\bar{X}$の期待値は、

$$E[\bar{X}]=\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]=\frac{1}{n}E\left[\sum_{i=1}^{n}X_i\right]=\frac{1}{n}\sum_{i=1}^{n}E[X_i]$$

$$=\frac{1}{n}\sum_{i=1}^{n}\mu=\frac{1}{n}\times n\times\mu=\mu$$

となる。標本数nが大きくなるほど、標本平均は期待値に近づく。

標本平均の分散

他方、標本平均$\bar{X}$の分散は、

$$Var[\bar{X}]=\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]=\frac{1}{n^2}Var\left[\sum_{i=1}^{n}X_i\right]$$

$$=\frac{1}{n^2}\left[\sum_{i=1}^{n}Var[X_i]+2\sum_{i\neq j}Cov[X_i,X_j]\right]$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}s^2=\frac{s^2}{n}$$

となる。
標本数nが大きくなるほど、分散は小さくなる。

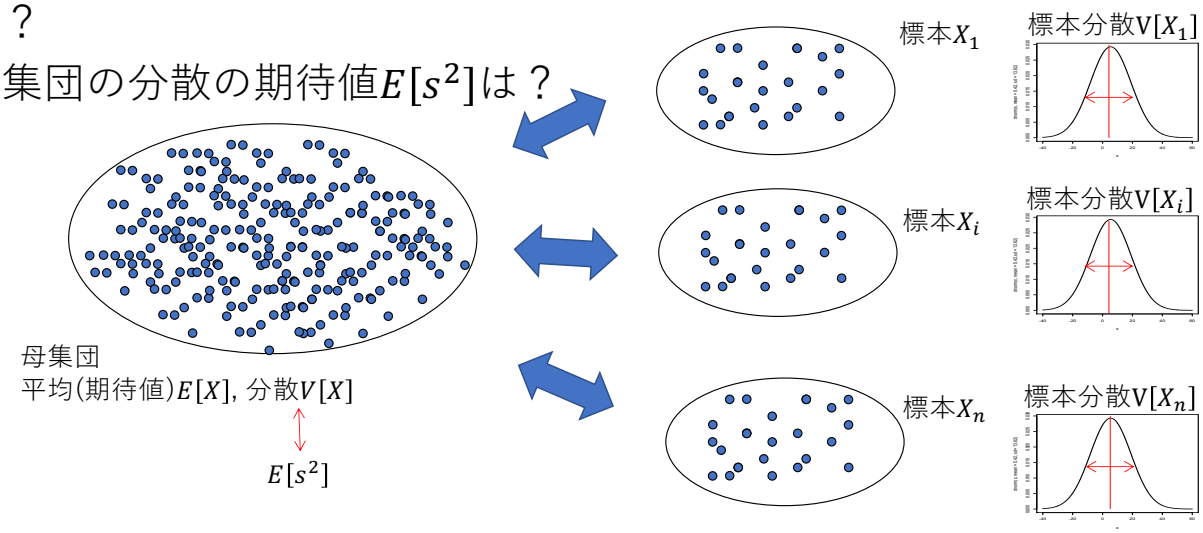
これまでに分かったことを整理すると…

- 母集団の期待値$E[X]=\mu$, 分散$Var[X]=s^2$とすると、標本X_iの期待値$E[X_i]=\mu$, 分散$Var[X_i]=s^2$となる
- 標本平均$\bar{X}$の平均（期待値） $E[\bar{X}]=\mu$である。標本数nが大きくなるほど母集団の平均（母平均に近づく）
- 標本平均$\bar{X}$の分散$Var[\bar{X}]=\frac{s^2}{n}$である。標本数$n$が大きくなるほど小さくなる

- では、母分散σ^2と標本分散s^2との関係性はどうなのか。
- 標本分散はどのような値にあることが期待されるか？＝標本分散の期待値$E[s^2]$はどのような値になるか？

母集団の分散と標本の分散

- 母集団の分散は、標本の分散s^2を用いて再現できるのか？
- 標本分散s^2と母集団の分散σ^2との間にどの程度の誤差があるのか？
- 母集団の分散の期待値$E[s^2]$は？



標本分散の期待値

- 標本分散の定義式に$\overline{\mathrm{X}} - X_i = \overline{\mathrm{X}} - \mu - (X_i - \mu)$を代入する

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left((\overline{\mathrm{X}} - X_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left((\overline{\mathrm{X}} - \mu)^2 - 2(\overline{\mathrm{X}} - \mu)(X_i - \mu) + (X_i - \mu)^2\right)$$

$$= (\overline{\mathrm{X}} - \mu)^2 - 2(\overline{\mathrm{X}} - \mu)\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) + \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$= (\overline{\mathrm{X}} - \mu)^2 - 2(\overline{\mathrm{X}} - \mu)\frac{1}{n}(X_1 - \mu + X_2 - \mu + \cdots X_n - \mu) + \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$= (\overline{\mathrm{X}} - \mu)^2 - 2(\overline{\mathrm{X}} - \mu)\frac{1}{n}(n\overline{\mathrm{X}} - n\mu) + \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$= (\overline{\mathrm{X}} - \mu)^2 - 2(\overline{\mathrm{X}} - \mu)^2 + \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

標本分散の期待値

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left((\overline{\mathrm{X}} - X_i)\right)^2 = \cdots$$

$$= (\overline{\mathrm{X}} - \mu)^2 - 2(\overline{\mathrm{X}} - \mu)^2 + \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{\mathrm{X}} - \mu)^2$$

- 標本分散の期待値$E[s^2]$を求めると、次のようになる

$$E[s^2] = \frac{1}{n}E[(X_i - \mu)^2] - E[(\overline{\mathrm{X}} - \mu)^2]$$

標本分散の期待値

- 母集団の分散$\sigma^2 = \frac{1}{n}E[(X_i - \mu)^2]$であることから、
$$E[s^2] = \sigma^2 - E[(\overline{\mathrm{X}} - \mu)^2]$$
- つまり、標本分散の期待値$E[s^2]$は母分散σ^2より$E[(\overline{\mathrm{X}} - \mu)^2]$だけ小さい
- 標本分散にこの誤差分だけ修正を加えれば、標本分散を利用して母分散を推定できるようになる
- 平均μ、分散σ^2の母集団について、さらに次の関係が成立する
$$E[(\overline{\mathrm{X}} - \mu)^2] = \frac{1}{n}E[(X_i - \mu)^2] = \frac{1}{n}\sigma^2$$
- このことから、
$$E[s^2] = \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2$$

となる

不偏分散

- 従って母分散は以下のようにして推定される
$$\sigma^2 = \frac{n}{n-1}E[s^2] = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{\mathrm{X}})^2\right)$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{\mathrm{X}})^2$$
- つまり、標本から分散を計算するときには、nで割るのではなく$n-1$で割ると母分散と等しくなる
- これを不偏分散$\hat{\sigma}^2$という
- 標本分散s^2は母分散σ^2に対して$E[(\overline{\mathrm{X}} - \mu)^2]$だけ偏りがある（小さい）が、不偏部分散$\hat{\sigma}^2$はそのような偏り（誤差）がない

自由度

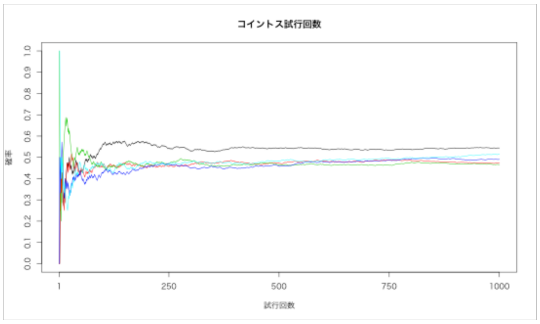
- 不偏分散は標本と標本平均の差の平方和を$n-1$で割った
- この$n-1$を自由度と呼ぶ
- 統計学では、独立に抽出された標本データがn個あるとき、このデータの自由度がnであるという
- n個のデータのうち$n-1$個のデータが決まれば残りのデータが決まるとき、自由度は１つ減って$n-1$となる

- 標本と標本平均の差の平方和を計算する際に、観測値の合計値を用いたため、自由度が１つ減ったと理解すれば良い

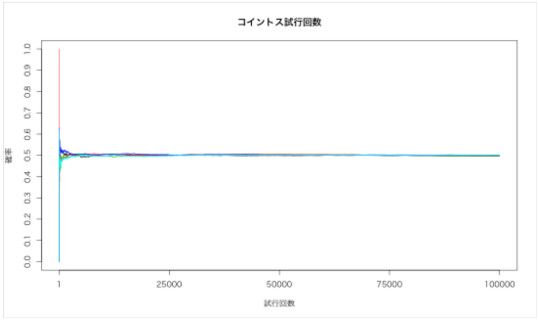
標本数を多くすると…

- コイントスをすると表が出る確率は1/2？
- 標本数が少ないと、必ずしも1/2にならない？
- 標本数を多くすると、標本の統計的性質（平均、分散）は本来の母集団の性質に近づく

5人が1000回コイントス



5人が100000回コイントス



大数の弱法則

独立で同等の確率分布$X_1, X_2, \ldots$に従う確率変数の平均をμ、分散をσ^2とする

その確率分布から標本$\{X_1, \ldots, X_n\}$を抽出した時、標本平均

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + \cdots + X_n)$$

の値が平均μから外れる確率は、分散σ^2が小さい場合$(\sigma^2 < \infty)$、標本数nを十分に多くとれば、非常に小さくなる

$$\lim_{n\to\infty} P(|\bar{X} - \mu| > \epsilon) = 0 \ (\forall \epsilon > 0)$$

標本数を十分に大きく取れば、標本平均は母集団の平均に近づく

大数の弱法則（数学的な証明）

$\{X_1, \ldots, X_n\}$が連続型確率変数の時、チェビシェフの不等式というものを用いて次式が成り立つ。

$$0 \le P(|\bar{X} - \mu| > \epsilon) = P\left((\bar{X} - \mu)^2 > \epsilon^2\right) \le \frac{1}{\epsilon^2}\frac{\sigma^2}{n}$$

このとき、 $n \to \infty$ならば$\frac{1}{\epsilon^2}\frac{\sigma^2}{n} \to 0$となるため、

$$\lim_{n\to\infty} P(|\bar{X} - \mu| > \epsilon) = 0$$

が成立する。

大数の弱法則

- コイントスによるシミュレーションで示す

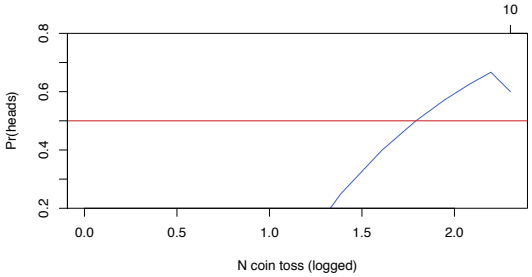
シミュレーションの手順

1. 施行回数をn回、コインの表=1、裏=0とする
2. 最初に$i = 1$回目のコイントスを試行し標本平均$\overline{X_1}$を与える
3. 試行回数$i > 1$回目のコイントスを試行し標本平均$\overline{X_i}$を計算
4. 最大の試行回数n回目に達したら、 $\overline{X_1}, \ldots, \overline{X_i}$をプロットする

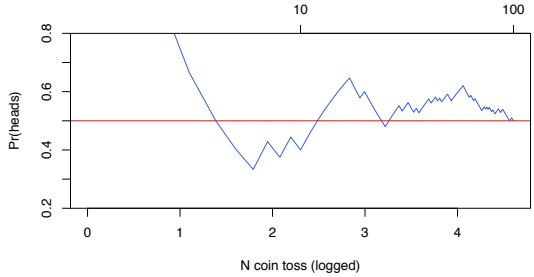
大数の弱法則

- コイントスによるシミュレーションを用いて示す

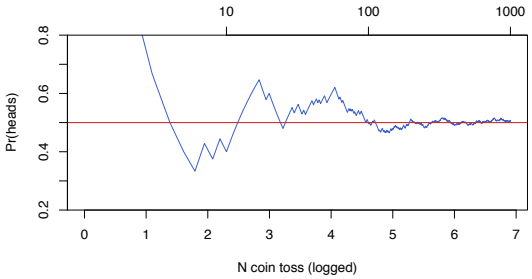
10回



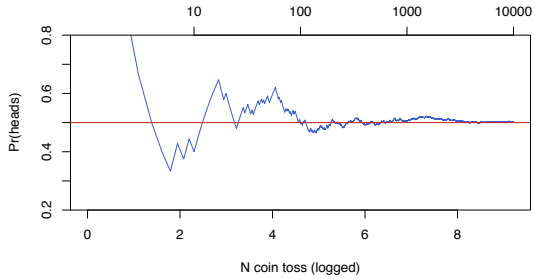
100回



1000回



10000回

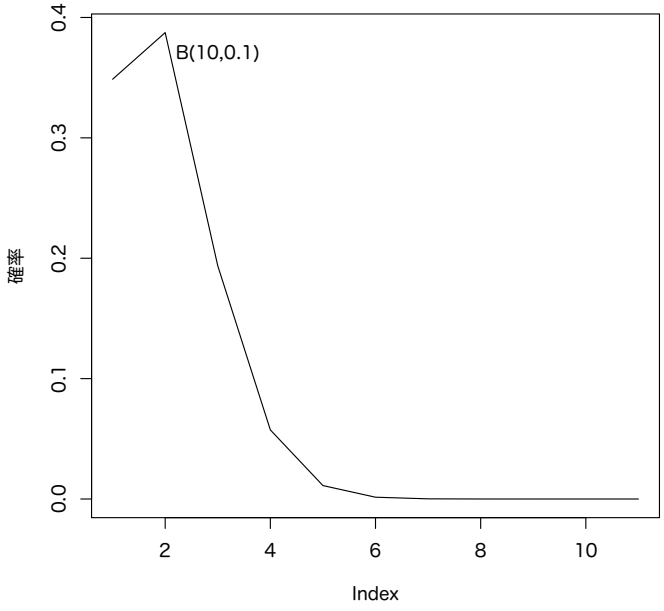


ベルヌーイ試行と二項分布

- 0か1かしかない試行において、n回の試行でs回成功し、その期待値pがわかっているとき、実験が成功する確率はベルヌーイ試行に従う
- ベルヌーイ試行の確率分布を二項分布といい、その分布は次式の確率密度関数に従う

$$Binom(n,p) = {}_nC_s \cdot p^s \cdot (1-p)^{n-s}$$
$$\approx p^s \cdot (1-p)^{n-s}$$

二項分布



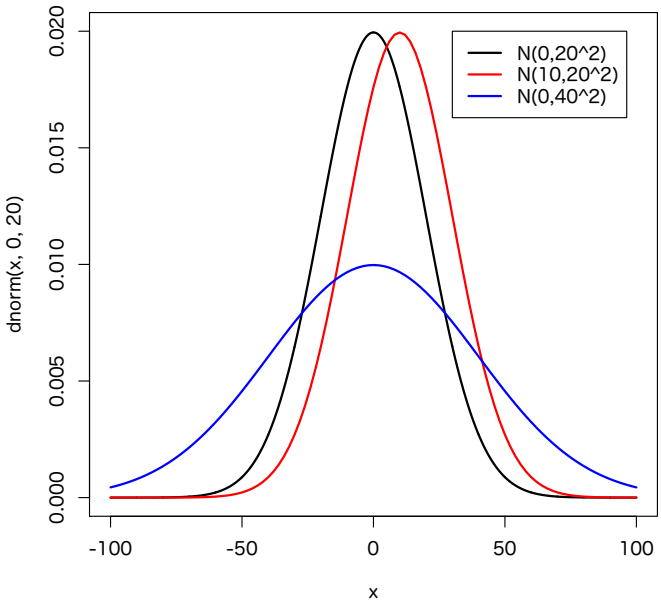
中心極限定理

- 互いに独立で同等の分布に従う確率変数$\{X_1, \ldots, X_n\}$の和$S_n = X_1 + \cdots + X_n$は、標本数が多くなるにしたがい正規分布に近づく
- 母集団の平均（母平均）と標本平均との誤差の確率分布は（母集団がどのような分布であっても多くの場合）正規分布に近づく
- 自然現象や社会現象で見られる事象やその事象を統計的に説明するための誤差の分布が正規分布に従うことを正当化するために用いられる

正規分布

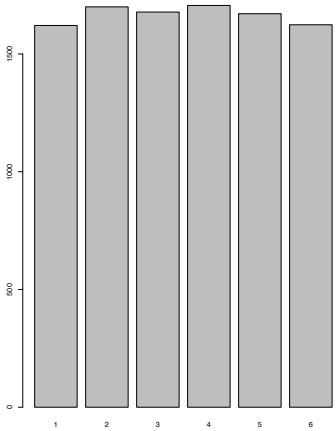
- 平均μ、分散σ^2(標準偏差σ)の正規分布$N(\mu, \sigma^2)$は、以下のように表される

$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

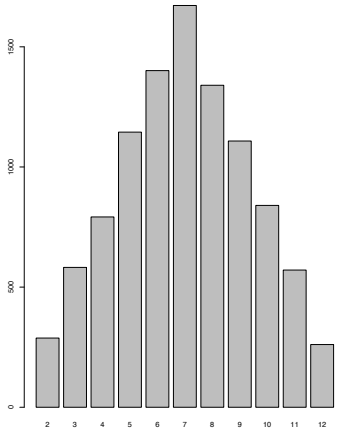


中心極限定理

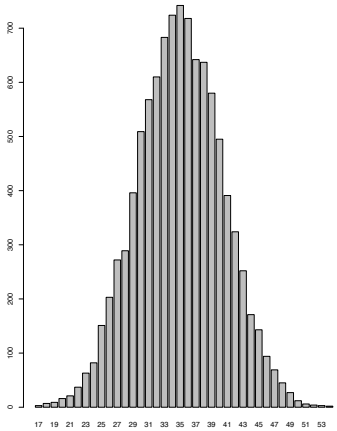
- サイコロ投げのシミュレーションで説明
- サイコロ10,000個の目の合計の分布



サイコロを１個投げた場合



サイコロを２個投げた場合



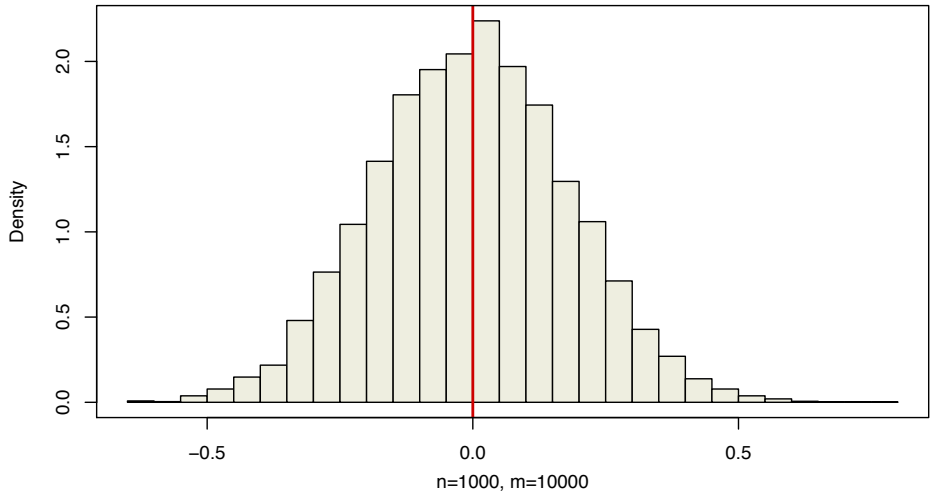
サイコロを１０個投げた場合

シミュレーションの手順

1. 母集団の確率分布を作成する（平均μ、分散σ^2）
2. 母集団から無作為に標本数nの標本を抽出し標本平均$\overline{X_n}$を計算
3. 2の作業を標本抽出回数m回繰り返す
4. 標本抽出回数m回に達したら、 m回分の標本平均$\overline{X_n}$をプロット
5. 標本数nを変化させ、2～4の作業を繰り返す

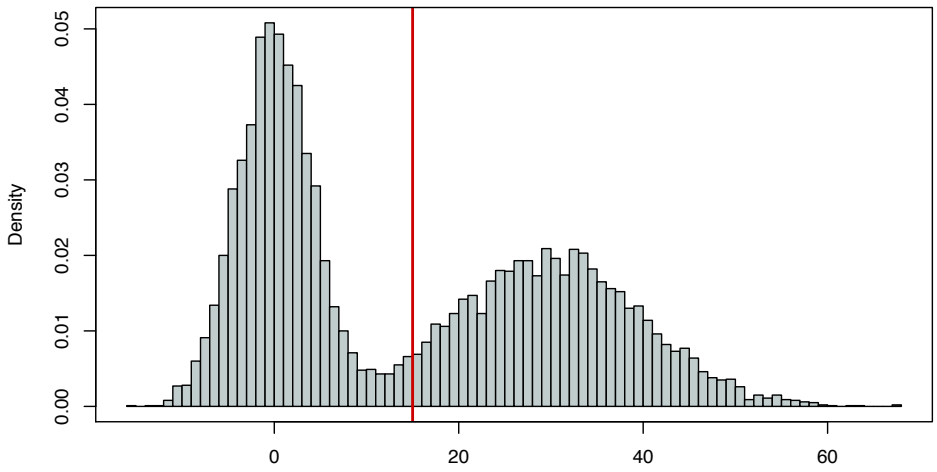
中心極限定理：シミュレーション結果

- 平均$\mu = 0$, 区間[-10, 10]の一様分布(σ^2)を母集団として、標本数$n = 1000$, 試行回数$m = 10{,}000$のシミュレーションを実施
- 標本平均$\overline{X_n}$は母集団の平均$\mu = 0$、分散σ^2/nの正規分布に近づく



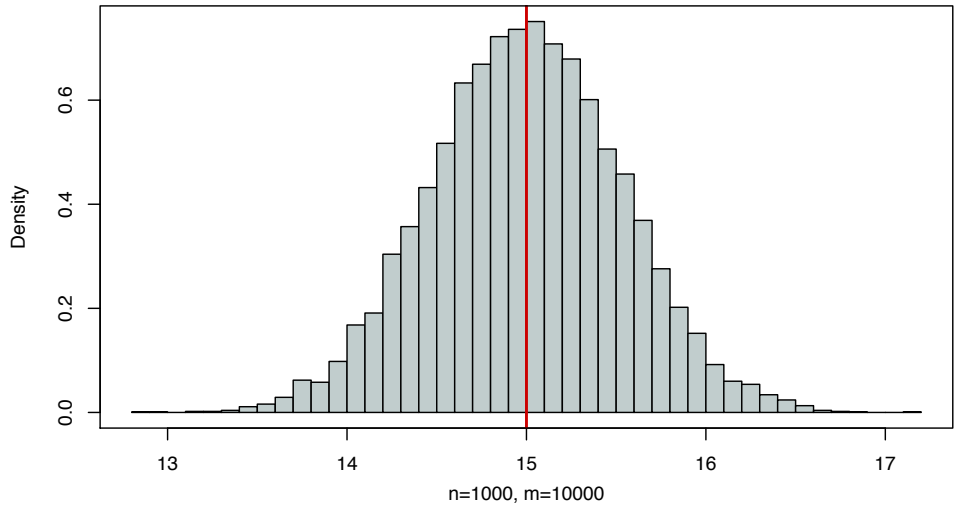
中心極限定理

- 中心極限定理は母集団がどのような確率分布でも成立する
- 例えば以下のような双峰型分布（平均μ, 分散σ^2）



中心極限定理のシミュレーション結果

- 正規分布（平均μ , 分散σ^2/n）に近づく($n = 1000, m = 10{,}000$)



大数の法則と中心極限定理

- 大数の法則：標本平均に関する法則

標本数nを非常に大きくすると($n \to \infty$)、標本平均$\bar{X}$は母集団の平均μに近づく$(\bar{X} \to \mu)$

- 中心極限定理：母集団の平均と標本平均の誤差に関する定理

標本数nを非常に大きくすると($n \to \infty$)、標本平均$\bar{X}$と母集団の平均μの誤差$(\bar{X} - \mu)$は正規分布に近づく

中心極限定理から得られる知見

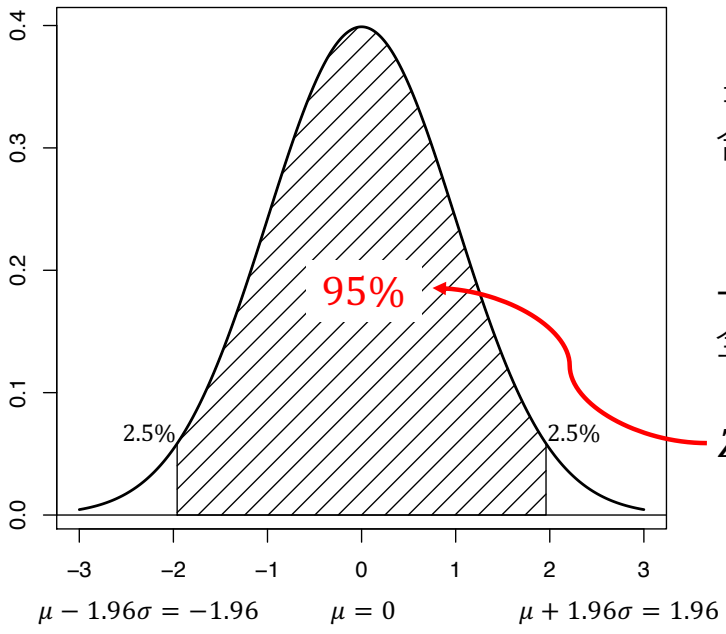
標本数nが非常に多いとき、

- 標本平均$\bar{X}$は、平均μ, 分散σ^2/nの正規分布$\mathrm{N}(\mu, \sigma^2/n)$に従う
- 標本平均と母集団の平均との差$\bar{X}-\mu$は、平均0, 分散σ^2/nの正規分布$\mathrm{N}(0, \sigma^2/n)$に従う

さらに$Z = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$とすると、

- 中心極限定理からZは平均0, 分散1の標準正規分布$\mathrm{N}(0,1)$に従う

標準正規分布$N(0,1)$



± 1.96の範囲に全体の95%が含まれる

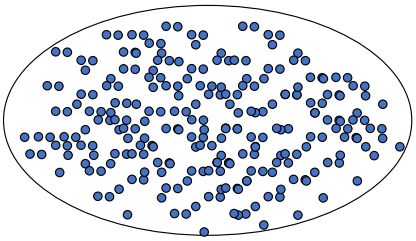
$-1.96 \leq Z \leq 1.96$のとき Z値は全体の95%の範囲内に含まれる

Z値がここに含まれる

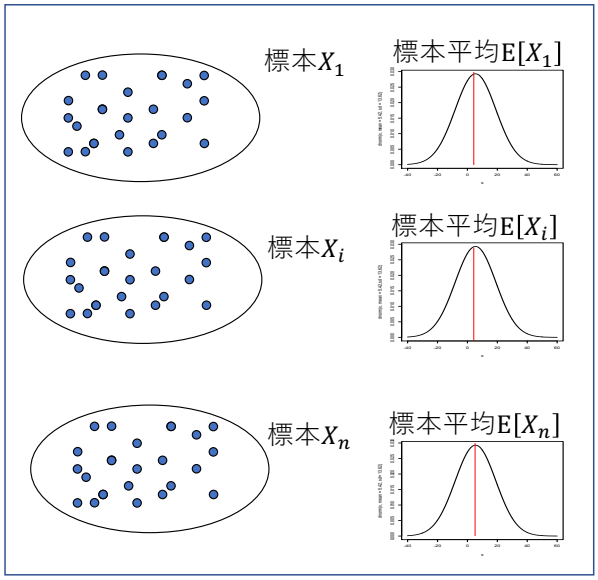
中心極限定理があると何が便利か？

- 標本と母集団の誤差が何%を正規分布を用いて検証できる

母集団の
平均(期待値)$E[X] = \mu$,
分散$V[X] = \sigma^2$



標本平均は
平均μ, 分散σ^2/n
の正規分布に従う



中心極限定理があると何が便利か？

標準正規分布の性質から$-1.96 \leq Z = \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \leq 1.96$であれば$Z$値は
標準正規分布全体の95%の範囲内に含まれる

母集団の平均が

$$\bar{X} - 1.96\sqrt{\sigma^2/n} \leq \mu \leq \bar{X} + 1.96\sqrt{\sigma^2/n}$$

の範囲に含まれるとき、母集団を95%の精度で含む信頼区間を計算することができる。

この性質は、のちに統計的検定や推定を行う際に非常に重要

信頼区間(confidential interval; CI)

$$\bar{X} - 1.96\sqrt{\sigma^2/n} \leq \mu \leq \bar{X} + 1.96\sqrt{\sigma^2/n}$$

を母集団の平均μの95%信頼区間という。

別に95%でなくてもよいが、95%信頼できれば十分な信頼度が得られているだろうという考え方（特に社会科学等の場合）。

人名が関わる医薬データなどの場合は、もっと信頼度の高い、例えば99%とか99.9%といった信頼区間を用いることがある。

一般に信頼度をαで表し、 α%信頼区間という言い方をする。

信頼区間

$$\bar{X} - 1.96\sqrt{\sigma^2/n} \leq \mu \leq \bar{X} + 1.96\sqrt{\sigma^2/n}$$

の$\sqrt{\sigma^2/n}$を標準誤差(Standard Error; SE)という

このあとに登場する統計的推定では、標本平均と母平均（母集団の平均）との誤差$(\bar{X} - \mu)$

誤差が正規分布に従うのであれば、正規分布の性質を利用して統計的推定を行うことができる

信頼区間

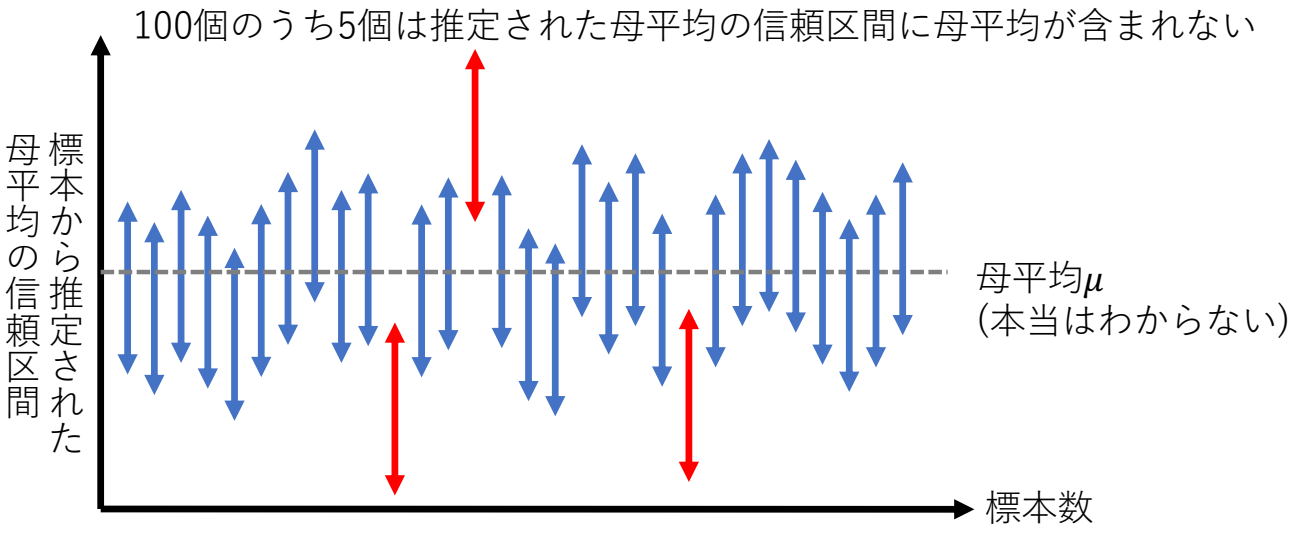
「95%信頼区間」のおおよその考え方

- 母集団から標本を100回抽出し、各標本平均から母集団の平均（母平均）の95%信頼区間を次々に計算したとき、100回中95回は95%信頼区間の中に母平均が含まれる

誤った考え方

- 標本から母集団を抽出した時、95%信頼区間の中に母集団が含まれる確率は95%となる

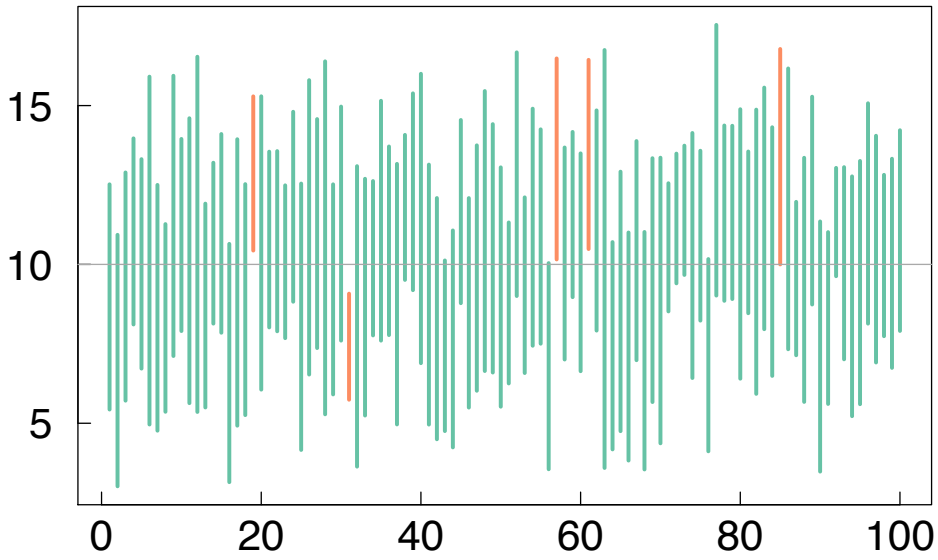
信頼区間



信頼区間のシミュレーション

1. 母集団の確率分布を作成する（平均μ、分散σ^2）
2. 母集団から無作為に標本数nの標本を抽出し標本平均$\overline{X_n}$と95%信頼区間を計算
3. 2の作業を標本抽出回数m回繰り返す
4. 標本抽出回数m回に達したらm回分の95%信頼区間をプロット

信頼区間のシミュレーション結果



標本数に対する信頼度

- 標本数を増やすと、標本への信頼度がどの程度高まるのか？
- 逆に、標本数を限定しても標本数への信頼度はある程度確保できるのか？

- 例： 100,000,000人の有権者数に対して1,600人の標本を抽出し世論調査を実施したところ、50%が政権を支持すると回答した。このとき、この世論調査（の標本）は95%信頼区間の範囲で、どの程度信頼できるのか？

サンプリング誤差と標本信頼度

- 一般に、標本数nに対して比率pが得られたとき、サンプリング誤差Eは信頼区間に対する信頼度係数kを用いて以下のように求めることができる

$$E = k\sqrt{\frac{p(1-p)}{n}}$$

- 95%信頼区間に対して信頼係数$k = 1.96$が用いられる

サンプリング誤差と標本信頼度

- 前述の例を計算すると、この世論調査のサンプリング誤差は $E = 0.0245$となる

$$E = k\sqrt{\frac{p(1-p)}{n}} = 1.96\times\sqrt{\frac{0.5\times0.5}{1600}} = 1.96\times\frac{0.5}{40} = 0.0245$$

- このとき、世論調査での政権支持率50%に対する95%信頼区間は47.5%～52.5% (50%±2.5%) であることがわかる。

サンプリング誤差とサンプル数

- 逆に、信頼区間・サンプリング誤差（許容誤差）・回答比率（想定される回答の比率）が決まっていれば、必要なサンプル数を計算できる

$$n = \left(\frac{k}{E}\right)^2 p(1-p)$$

- 信頼度係数はサンプリング誤差の早見表などを用いて得ることができる

サンプリング誤差早見表の例

https://www.biwako.shiga-u.ac.jp/sensei/mnaka/ut/samplingerrtab.html

confidence level	95%						
population proportion	1% 99%	5% 95%	10% 90%	20% 80%	30% 70%	40% 60%	50%
sample size: n = 100	±1.95	±4.27	±5.88	±7.84	±8.98	±9.60	±9.80
200	±1.38	±3.02	±4.16	±5.54	±6.35	±6.79	±6.93
300	±1.13	±2.47	±3.39	±4.53	±5.19	±5.54	±5.66
400	±0.98	±2.14	±2.94	±3.92	±4.49	±4.80	±4.90
500	±0.87	±1.91	±2.63	±3.51	±4.02	±4.29	±4.38
600	±0.80	±1.74	±2.40	±3.20	±3.67	±3.92	±4.00
700	±0.74	±1.61	±2.22	±2.96	±3.39	±3.63	±3.70
800	±0.69	±1.51	±2.08	±2.77	±3.18	±3.39	±3.46
900	±0.65	±1.42	±1.96	±2.61	±2.99	±3.20	±3.27
1,000	±0.62	±1.35	±1.86	±2.48	±2.84	±3.04	±3.10
2,000	±0.44	±0.96	±1.31	±1.75	±2.01	±2.15	±2.19
3,000	±0.36	±0.78	±1.07	±1.43	±1.64	±1.75	±1.79
4,000	±0.31	±0.68	±0.93	±1.24	±1.42	±1.52	±1.55
5,000	±0.28	±0.60	±0.83	±1.11	±1.27	±1.36	±1.39
approximation	$\pm\frac{20}{\sqrt{n}}$	$\pm\frac{43}{\sqrt{n}}$	$\pm\frac{59}{\sqrt{n}}$	$\pm\frac{78}{\sqrt{n}}$	$\pm\frac{90}{\sqrt{n}}$	$\pm\frac{96}{\sqrt{n}}$	$\pm\frac{98}{\sqrt{n}}$