入門 統計学　第3章

推定と誤差　―推測統計学―

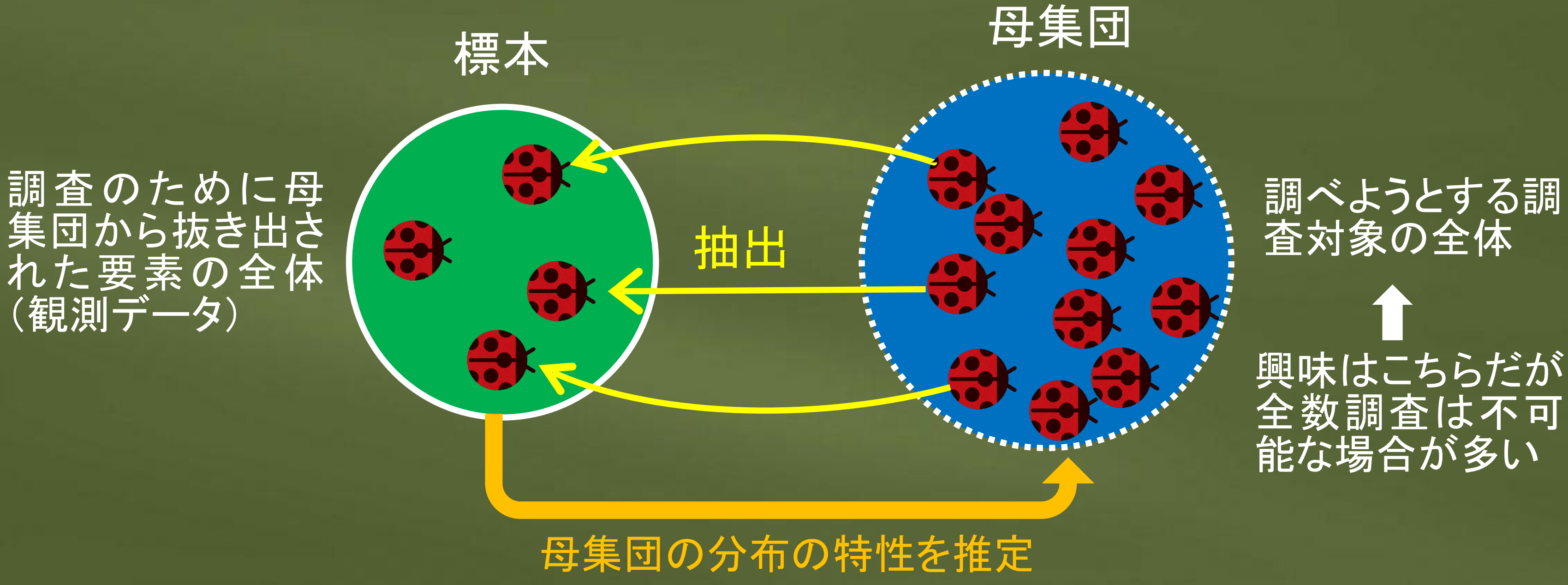
『入門 統計学 第2版　―検定から多変量解析・実験計画法・ベイズ統計学まで―』（オーム社）

※注：本書を購入された方へのサービスですので，教科書指定（参考図書は不可）していない授業での使用はお控えください。

3.1　推測統計学

- 母集団の特性（母数※）の値を，標本から推定する学問

※母数：母集団の分布形を決める平均や分散などの未知の定数



推測統計学には推定と検定がある

推定※（2種類ある）

- **点推定**：1つの値で母数を推定　　→本章
- **区間推定**：幅をとって母数を推定 →第4章

※推測する方法を推定と呼ぶ

検定

- 複群間の母数に差があることを検証
  →第5章～第9章

3.2　統計記号の使い分け方

- 母数（未知の定数）は母○○と呼び，ギリシャ小文字を使う (例：母集団の分散は母分散でσ^2)
- 標本に関する統計量※1は標本○○と呼び，アルファベット小文字を使う（標本の分散は標本分散でs^2）
- 母数の推定量※2は不偏○○と呼び，＾を付けたギリシャ小文字（点推定した分散は不偏分散で$\hat{s}^2$）

※1 統計量：標本データを計算することで，標本の特性を表した値

※2 推定量：標本から母数（母集団の特性）を点推定した値

統計記号の整理

（注：ほかにも比率や相関係数などがある）

標本の特性
（統計量）

標本平均　：　$\bar{x}$
標本分散　：　s^2
標本標準偏差：s

抽出

推定

母集団の分布の特性
（母数）

母平均　：　μ
母分散　：　σ^2
母標準偏差：σ

標本から点推定した
母数の不偏推定量

不偏平均　：　$\hat{\mu}$
不偏分散　：　$\hat{\sigma}^2$
不偏標準偏差：$\hat{\sigma}$

標本統計量≠推定量？

- 平均：標本平均$\bar{x}$が母平均の推定量$\hat{\mu}$と考える
- 分散：標本分散s^2は母分散σ^2よりも小さく偏る→標本分散を少し大きく修正する必要

$$s^2 = \frac{\Sigma(x_i - \overline{x})^2}{n} < \sigma^2 = \frac{\Sigma(x_i - \mu)^2}{n}$$

標本分散　　　母分散

標本分散が母分散より小さくなる理由

- 分散は平均に$\bar{x}$を使ったとき（つまり標本分散）が最小だから

事例：3つの標本データ｛1，2，3｝で，分散の大きさを比較する（標本平均$\bar{x}$＝2）

標本分散の偏差平方和：Σ（x_i-$\bar{x}$）^{2}=(1-2)^2+(2-2)^2+(1-2)^2=2

ほかのどんな値の平均を使ったときより小さい

母分散の偏差平方和：Σ（x_i-μ)^{2}のμは2とは限らない

μ=2.1として計算すると(1-2.1)^2+(2-2.1)^2+(1-2.1)^2=2.03

母集団のバラツキの点推定

- 計算時，標本サイズnの代わりに自由度n-1を使って少し大きくして偏りを修正する

標本分散 $s^2 = \dfrac{\Sigma (x_i - \bar{x})^2}{n}$ $\xrightarrow{\text{少し大きく}}$ 不偏分散 $\hat{\sigma}^2 = \dfrac{\Sigma (x_i - \bar{x})^2}{\boxed{n-1}}$ 自由度

不偏標準偏差は平方根を取るだけ※

不偏標準偏差 $\hat{\sigma} = \sqrt{\dfrac{\Sigma (x_i - \bar{x})^2}{n-1}}$

不偏分散のExcel関数=VAR.S

不偏標準偏差のExcel関数=STDEV.S

※厳密には，平方根を取るだけだと若干小さくなってしまうので修正係数をかけます（本書補足参照）。

（自由度が気になるので節3.6を前倒しで解説）

自由度とは？

- 推定量の計算に使う変数の数（標本サイズ）のうち，自由に値をとれる変数の数のこと
- 標本サイズnから制約条件の数を引いた値なのでn-1とは限らない（そのためv記号を当てる）
- 制約条件とは，標本を使った計算式のこと　→もとは自由に値を取れた変数同士でも，相互関係を式で決められると自由は拘束される

（不偏分散で使う標本平均は1つなのでn-1）

自由度の事例
（標本平均が制約条件の場合）

①a, b, cという，3つの変数がある（n=3）

②いずれもどのような値でも取れる（自由度=3）

③標本平均を（a+b+c)/3で計算する事を考える

④aとbはどのような値も取れるが，cは取れる値が自動的に決まる（自由度=2）

→各変数が1/3ずつの情報を出して平均を計算したため，合計で変数1個分の情報量を失った（不偏分散を計算するために残った情報量は3-1となった）

分散の分母がn-1となることを式で確認

- 母平均μの代わりに標本平均$\bar{x}$で分散を計算

→標本平均もバラつくので，その分散※も加える

→整理すると分母がn-1となる

母平均は未知

標本平均の分散

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{n} \rightarrow \sigma^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} + \boxed{\frac{\sigma^2}{n}} \rightarrow \sigma^2 = \frac{\Sigma(x_i - \bar{x})^2}{\underbrace{\boxed{n-1}}_{\text{自由度}}}$$

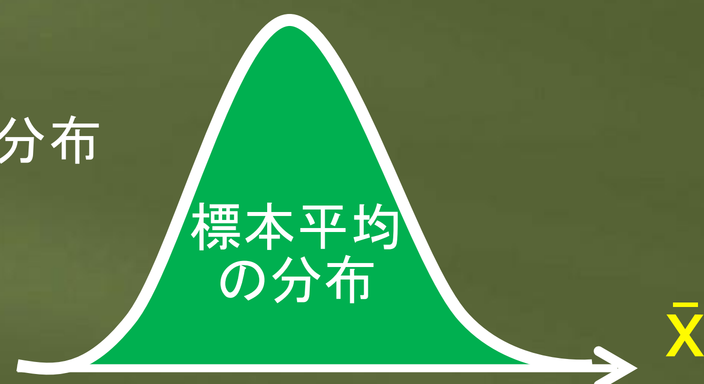
※標本平均の分散がσ^2/nとなることは，この後解説

3.3　標本分布と誤差

- 推測統計学では，標本データx個別の分布ではなく，標本統計量（標本平均$\bar{x}$や標本分散s^2など）の分布を考える

本章はこちらを扱います
（標本分散の分布は次章）

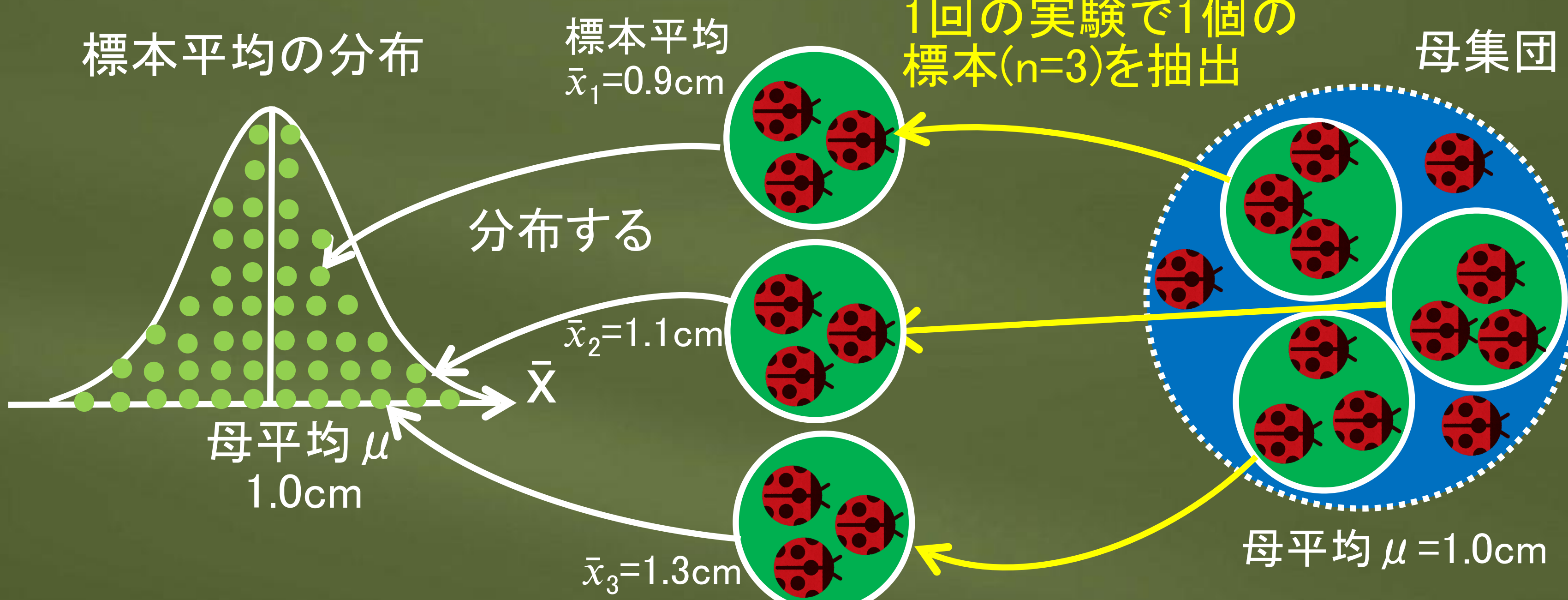
標本データの分布
といっても…



（n=1の実験の繰り返し）

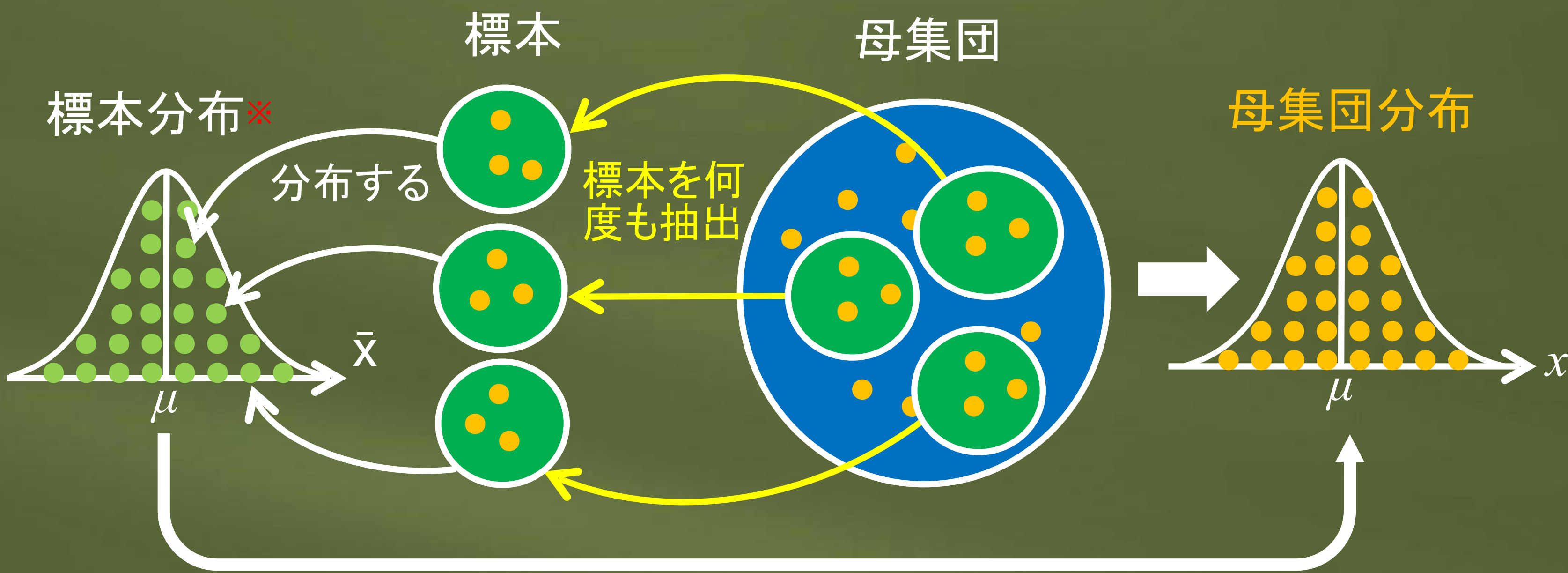
（n≧2の実験の繰り返し）

標本統計量の分布とは？



実際には実験は1回だが，理論上は実験を繰り返せば標本を沢山抽出できる

標本分布を考える理由：
推定誤差を評価するため



標本分布を考える事で推定誤差を評価できる

※標本統計量によって分布の形は異なる
（図は標本平均の例：正規分布）

誤差とは？

- いくら頑張って標本データから母数（真値）を推測しても，両者の間には必ず誤差（値のズレ）が生じる
- 誤差が小さい推定とは精度の高い推定ということ

- 誤差は標本サイズによって変化するため，データを1つしか観測していない（n=1）実験結果からは評価できない

誤差にも２種類ある

取り除くことができない誤差なので評価は重要

標本サイズと関連（標本分布で評価できる）

偶然誤差（本章）

ズレに方向性がなく，推定の精度にの指標（発生原因：標本抽出時に偶然生じる測定誤差など）

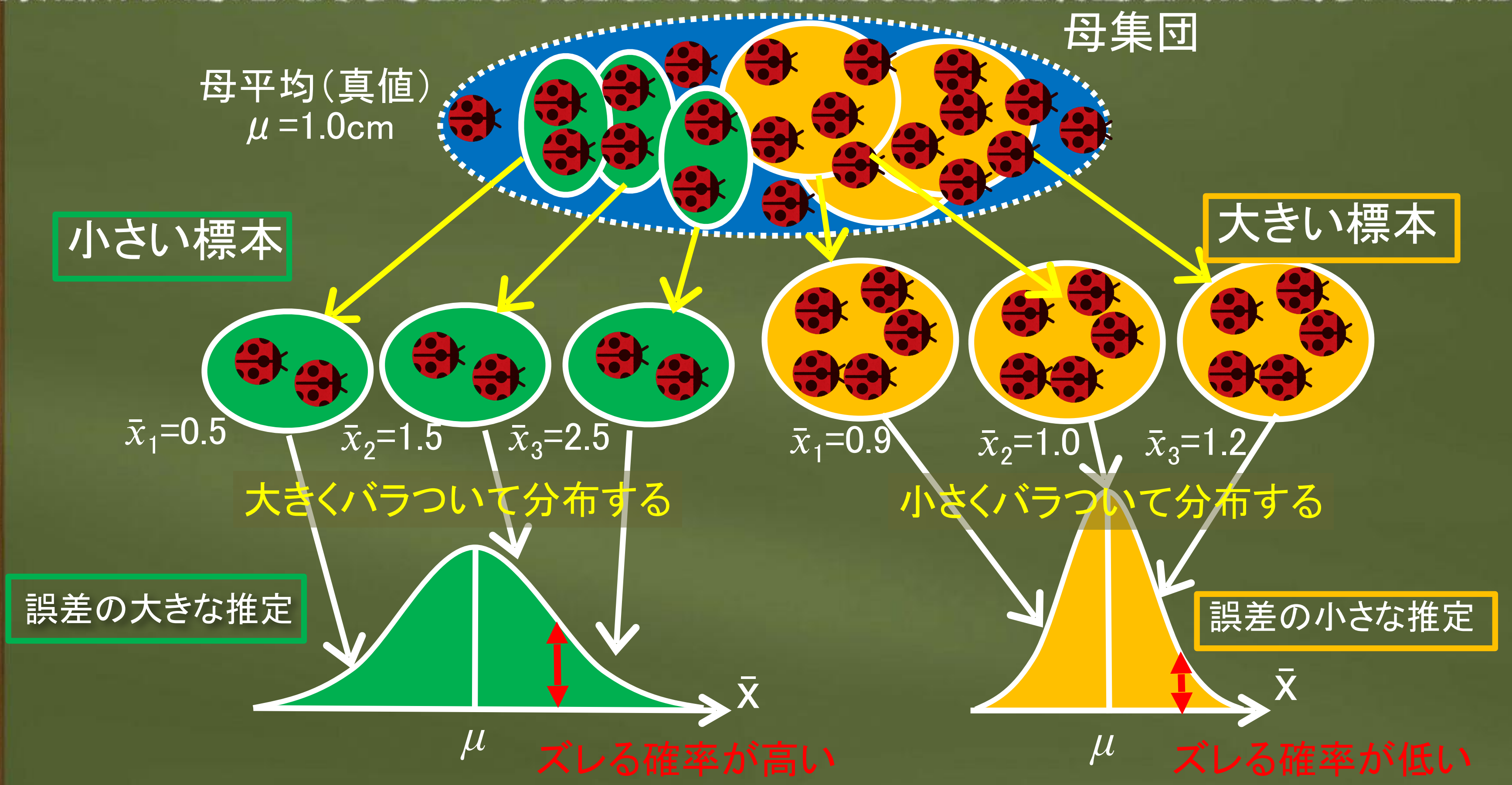
誤差（推定値と真値とのズレ）

実験計画で除去，あるいは偶然誤差へ転化

系統誤差（第10章）

ズレに方向性があり，推定の正確さの指標（発生原因：測定者や測定器の癖，実験配置の順番，欠損値など）

3.4　標本のバラツキ（サイズとの関係）



誤差の指標①　誤差分散

- 誤差は標本のバラツキ（分散・標準偏差）で評価
- 標本の分散：誤差分散→分散を標本サイズnで割るだけ

理由は次

母誤差分散：$\sigma_{\bar{x}}^2 = \dfrac{\text{母分散}}{\text{標本サイズ}}$　$\dfrac{\sigma^2}{n}$

σ²がわかる状況ならば誤差を考える必要は無いので出番はない（概念として示しただけ）

$\dfrac{\text{標本分散}}{\text{標本サイズ}}$

バラツキを過小評価するので推測統計学では使わない

標本誤差分散：$s_{\bar{x}}^2 = \dfrac{\text{不偏分散}}{\text{標本サイズ}}$　$\dfrac{s^2}{n}$

$\dfrac{\text{~~標本分散~~}}{\text{自由度}}$

不偏分散を自由度で割らない

不 偏 誤 差 分 散

$\hat{\sigma}^2$　s^2

補足　　2度もnで除す理由

（平均の分散である誤差分散は，なぜ分散を再びnで割るのか？）

- 分散に関する２つの性質を利用する

性質①：確率変数xを定数倍したものの分散は，元の確率変数xの分散を定数の2乗倍したものになる

性質②：“和の分散”は“分散の和”と等しい

同じ母集団から抽出した
標本の分散は全て等しい

標本平均式

$$V(\overline{x}) = V\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(x_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

性質①　　性質②

誤差の指標②　標準誤差

- 標本の標準偏差：標準誤差（SE)→誤差分散の平方根

母標準誤差　：$\sigma_{\bar{x}} = \dfrac{\text{母標準偏差}}{\sqrt{\text{標本サイズ}}} = \dfrac{\sigma}{\sqrt{n}}$

標本標準誤差：$s_{\bar{x}} = \dfrac{\text{標本標準偏差}}{\sqrt{\text{標本サイズ}}} = \dfrac{s}{\sqrt{n}}$

推定精度の指標として主に用いられる

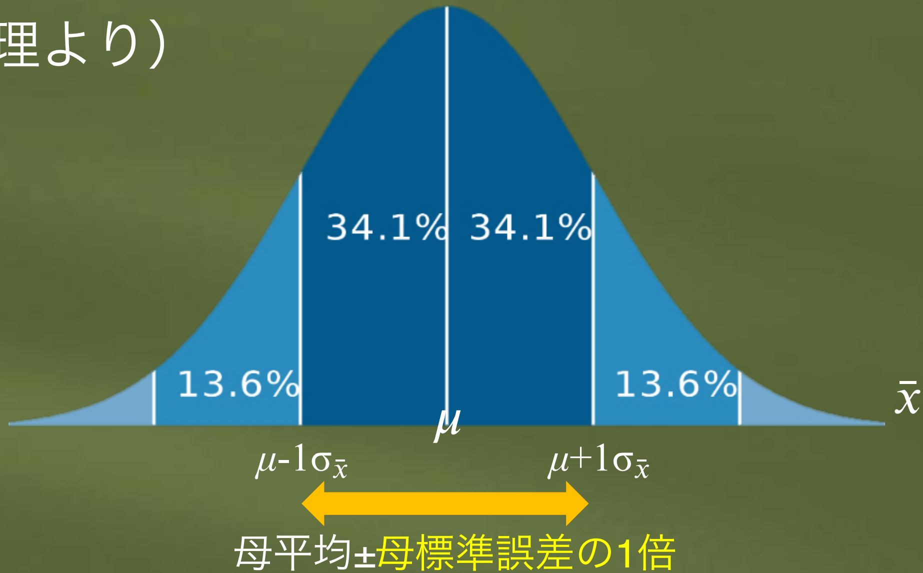
※アドバイス：点推定を行ったら，その推定の精度を表す不偏標準誤差も記しておくと親切

不偏標準誤差：$\hat{\sigma}_{\bar{x}} = \dfrac{\text{不偏標準偏差}}{\sqrt{\text{標本サイズ}}} = \dfrac{\hat{\sigma}}{}$ か $\dfrac{\text{標本標準偏差}}{\sqrt{\text{自由度}}} = \dfrac{s}{}$

再確認

標本分布における母標準誤差

母集団から繰り返し標本を抽出し，その平均を分布させると正規分布に従う（次章の中央極限定理より）



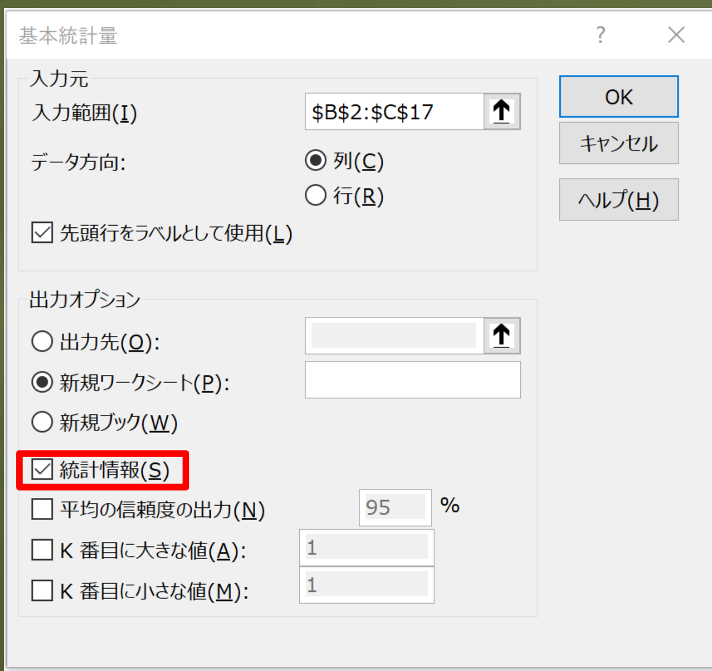
沢山ある標本平均のうちの68.2%が

母平均の±母標準誤差の範囲に含まれる

例題　キュウリ収量の不偏標準誤差を Excel分析ツールで計算

メニュー[データ]→[データ分析]で
[統計情報]を選択

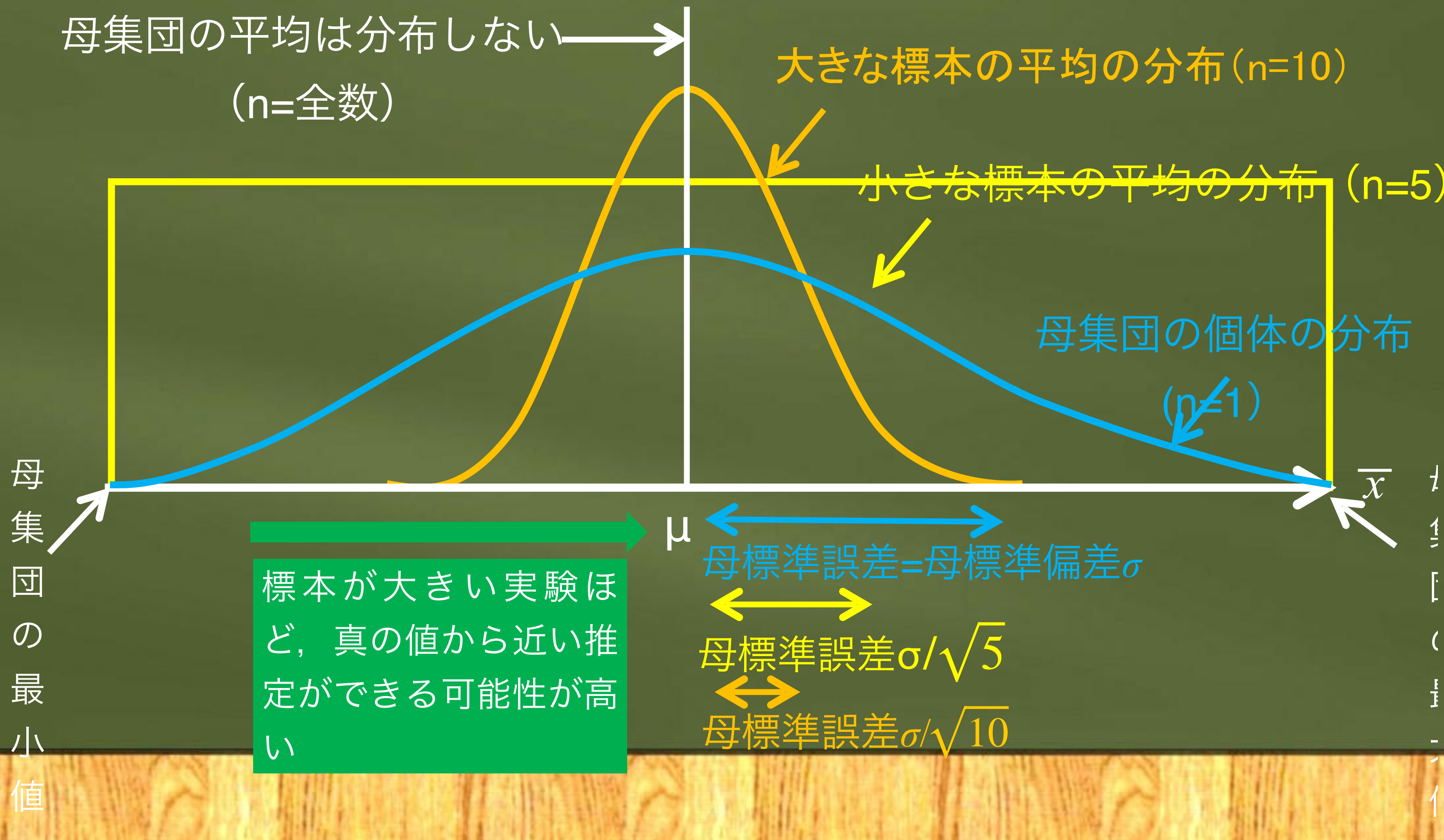
いろいろな統計量を自動で計算してくれる



	A	B	C	D	E
1	栽培法A			栽培法B	
2	平均	2443.67		平均	3124.40
3	標準誤差	110.52		標準誤差	94.35
4	中央値（メジアン）	2275		中央値（メジアン）	3203
5	最頻値（モード）	#N/A		最頻値（モード）	#N/A
6	標準偏差	428.05		標準偏差	365.40
7	分散	183226.95		分散	133515.40
8	尖度	-0.92		尖度	1.33
9	歪度	0.37		歪度	-1.10
10	範囲	1462		範囲	1414
11	最小	1757		最小	2219
12	最大	3219		最大	3633
13	合計	36655		合計	46866
14	データの個数	15		データの個数	15

注：分析ツールでは不偏標準誤差しか計算できない

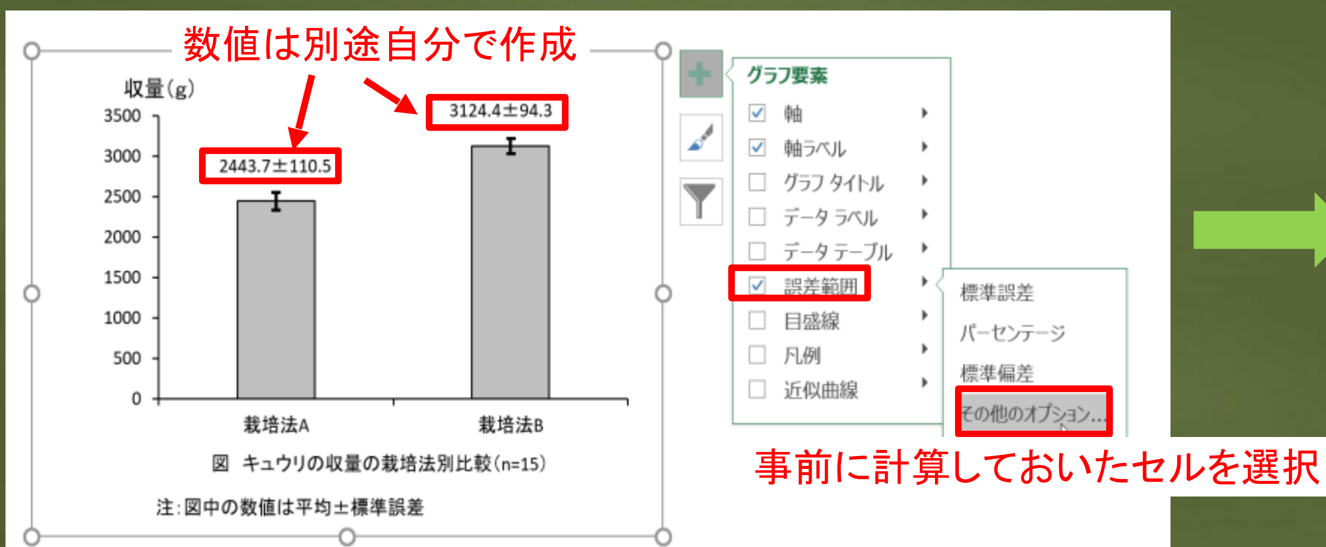
3.5 標本分布のまとめ
（標本サイズと推定誤差の大きさの関係）

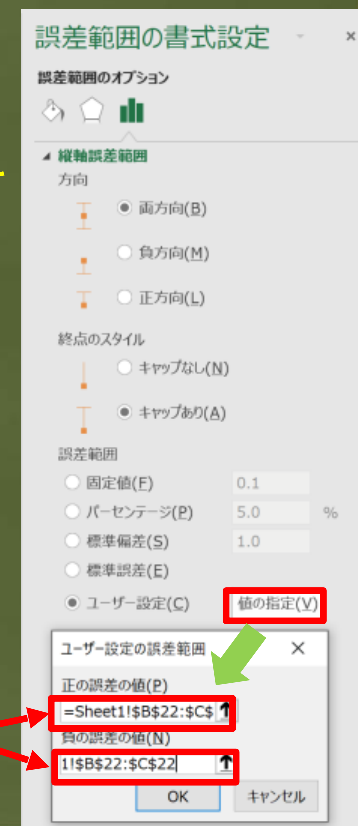


例題　Excelで棒グラフにエラーバー（誤差範囲）を付けてみよう！

栽培方法ごとに平均と不偏標準誤差を計算して棒グラフを作成しておく→［グラフエリア］を左クリック→グラフ要素を追加するための緑色の十字ボタンを左クリック→［誤差範囲］を☑→［その他のオプション］を選択

→［ユーザー設定］の「値の指定］で標準誤差が入っているセルを選択（正も負も同じセル）





（次章の区間推定のための準備）
標本平均の標準化

✿ 標準化：平均を0，母標準誤差を1に変換すること
→メリット：標準化すれば，どのような対象の標本平均に対しても同じ分布表を使うことができる

$$z_{\bar{x}_i} = \frac{\bar{x}_i - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x}_i - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}}$$

標本平均

標本平均の平均
（つまり母平均 μ）

標本平均の
標準化変量

標本平均の母標準偏差
（つまり母標準誤差）

以上で第３章は終了です。