東京大学
THE UNIVERSITY OF TOKYO

科学研究のための統計入門

GSC　2019年9月28日

東京大学 生産技術研究所

Pavel Hejcik

UTokyo GSC

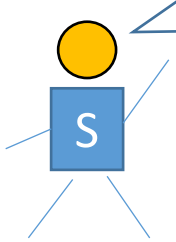
今日のメッセージ

Our knowledge is never perfect

我々の知識は決して完璧ではありません

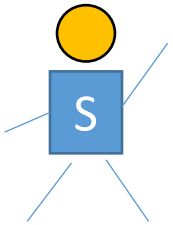
はじめに

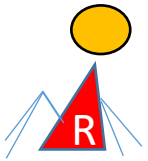
知っているということより、
なぜそれを知っているかを
知ることが大事



知っているということより、なぜそれを知っているかを知ることが大事

科学では、知識につながる主な手段は二通りある。それは、「観察」と「実験」

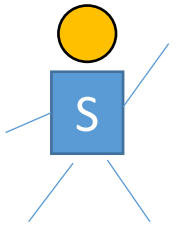


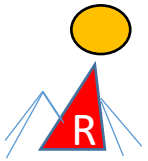


知っているということより、なぜそれを知っているかを知ることが大事

科学では、知識につながる主な手段は二通りある。それは、「観察」と「実験」

科学は客観的な証拠のもとに成り立っている。



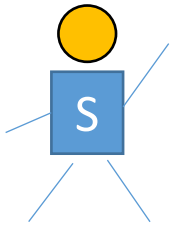


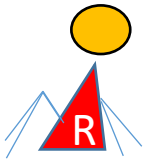
知っているということより、なぜそれを知っているかを知ることが大事

科学では、知識につながる主な手段は二通りある。それは、「観察」と「実験」

科学は客観的な証拠のもとに成り立っている

実験データは科学的証拠の一種である





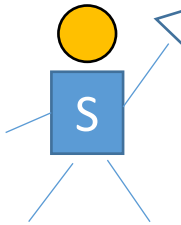
知っているということより、なぜそれを知っているかを知ることが大事

科学では、知識につながる主な手段は二通りある。それは、「観察」と「実験」

科学は客観的な証拠のもとに成り立っている

実験データは科学的証拠の一種である

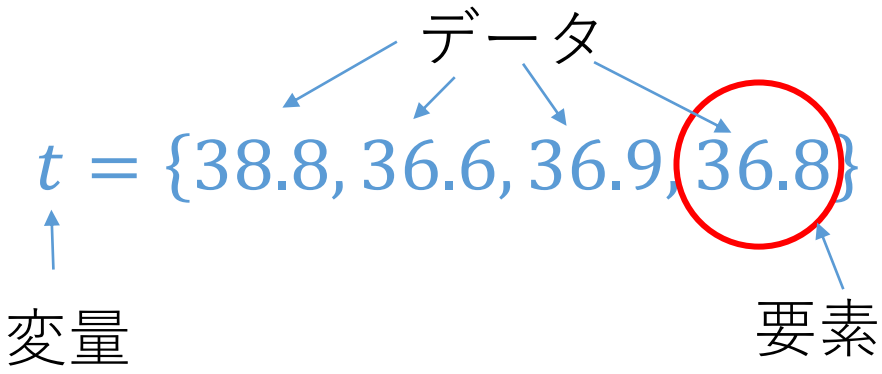
統計学とは、実験データを証拠に変えることができる学問である



だから今日統計を学ぶのね。

本日よく使われることば

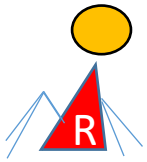
- 実験で測る物理量を「変量」と呼ぶ
- 測定の量的結果を「データ」と呼ぶ
- データの項目を「要素」と呼ぶ



測定　Measurement

- 100%正確な測定などない
- すべての測定には誤差がある
- 誤差の推定は測定の重要な部分である

知らないことを知るのが
大事、大事。



変量の種類

- 離散型

最小単位がある

取りうる値が飛び飛びになっている

例：観察中のイベントの個数

- 連続型変量

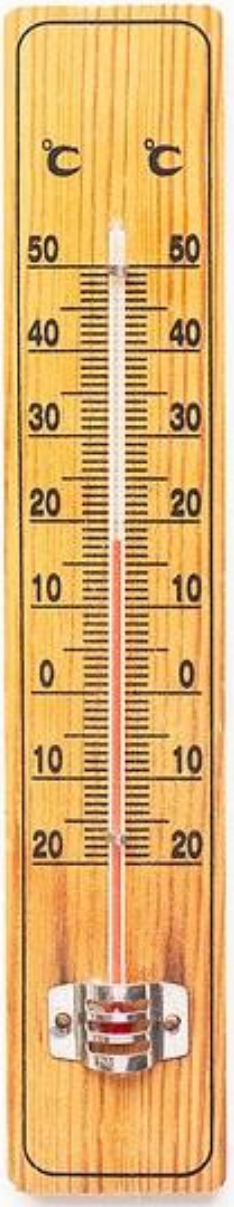
とぎれることなく、続いている

例：身長、温度など

本日の講義では連続型変量を考える

今日の気温は？

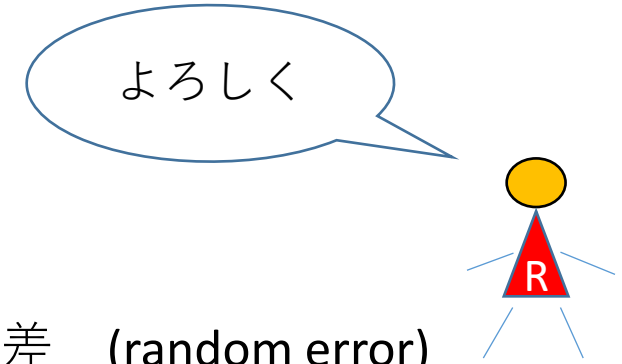
測定値　t = 17.8℃



この値はどれほど正しいだろう。

本当の値に迫る方法を探ってみよう！

誤差の種類

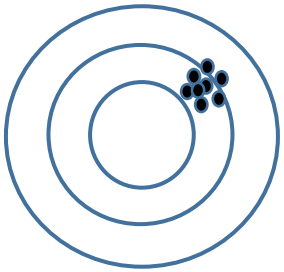


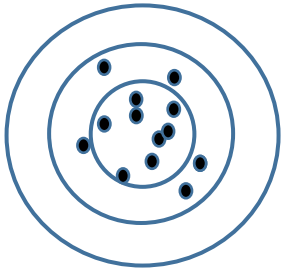
系統的誤差（systematic error)
- 気づきにくい
- 統計学で処理不可

例:
- 測定値の較正 の問題
- 測定環境の問題または個人の差

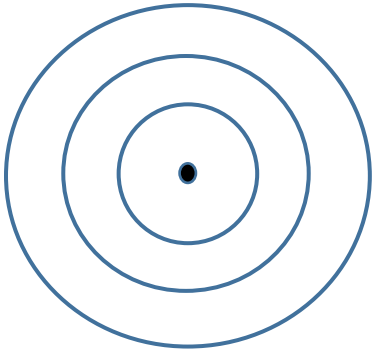
ランダム誤差　(random error)
-たくさんの小さな影響の結果
- 統計学を利用して処理可

例:
-環境の揺らぎ
-測定装置による誤差

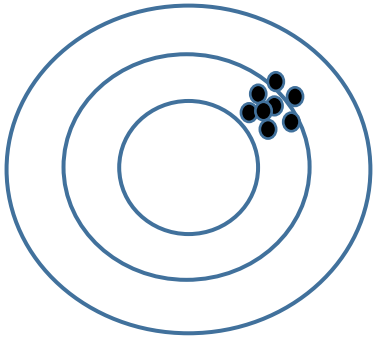




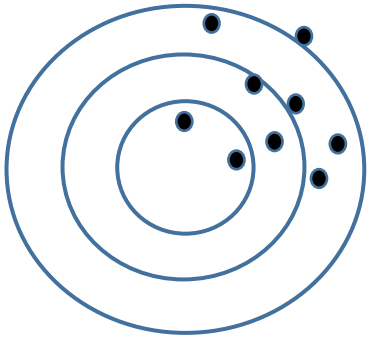
系統的とランダム誤差



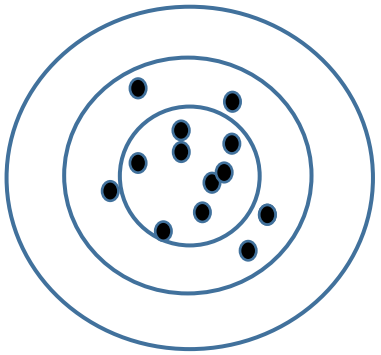
本当の値



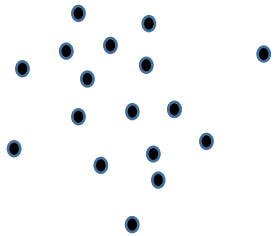
系統的



系統的＋ランダム



ランダム

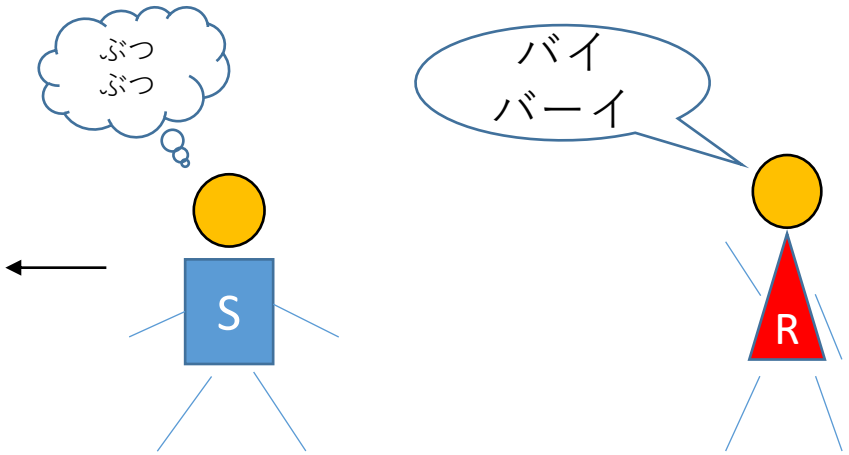


実験の現実

演習

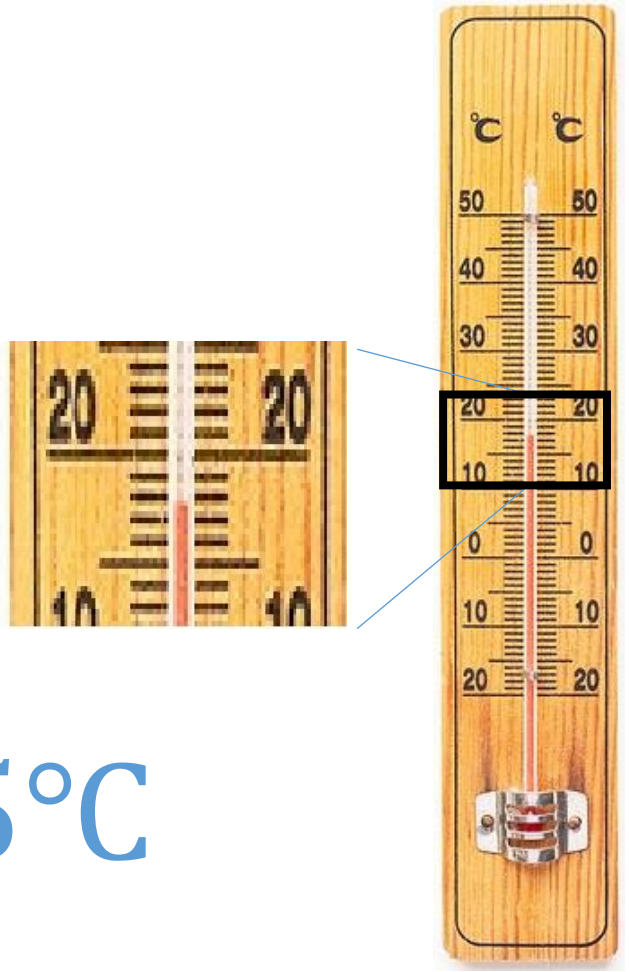
気温を測定する場合の系統的とランダム誤差の例をあげてみよう

これからは
ランダム誤差だけに注目する



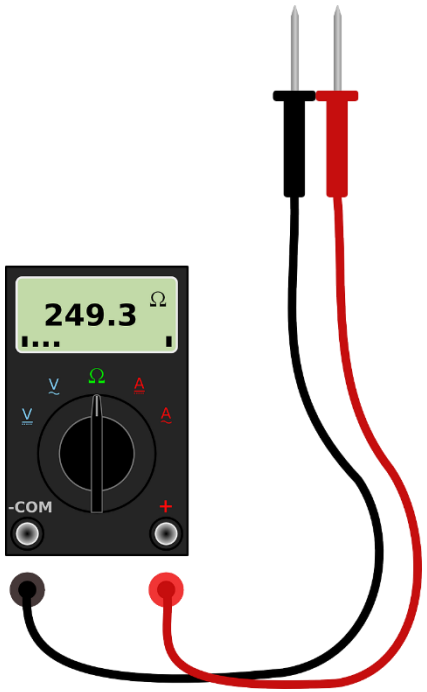
一つの測定値の誤差推定

測定の誤差は一つの測定値しかなければ、スケールの最小区分の半分にしてもよい。



$$t = 17.5℃ \mp 0.5℃$$

一つの測定値の誤差推定



電気測定装置の場合はマニュアルに測定誤差が記載されている。

$$R = 249.3\Omega \mp 0.3\Omega$$

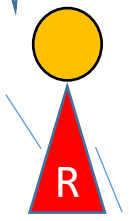
測定値の書き方

絶対誤差ともいう

$$X = X_{best} \mp \delta x$$

変量　　測定値　　誤差

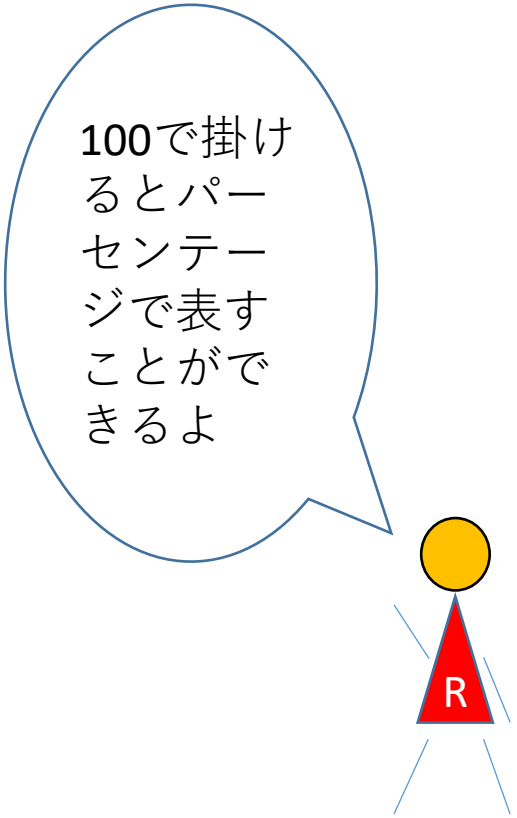
誤差は推定値の不確実性を示しているんだよ



$$t = 17.5℃ \mp 0.5℃$$

相対誤差　Relative error

$$\delta_r x = \frac{\delta x}{X_{best}}$$



相対誤差:例

R君の身長： $L = 175.0cm \mp 0.5\text{cm}$

$$\delta_r L = \frac{0.5}{175} \cong 0.003 \rightarrow 0.3\%$$

測定値の書き方

$$X = X_{best} \mp \delta_r x.100\%$$

$$T = 175cm \mp 0.3\%$$

Q:

本当の値をより正確に知ることは
可能だろうか

Q:

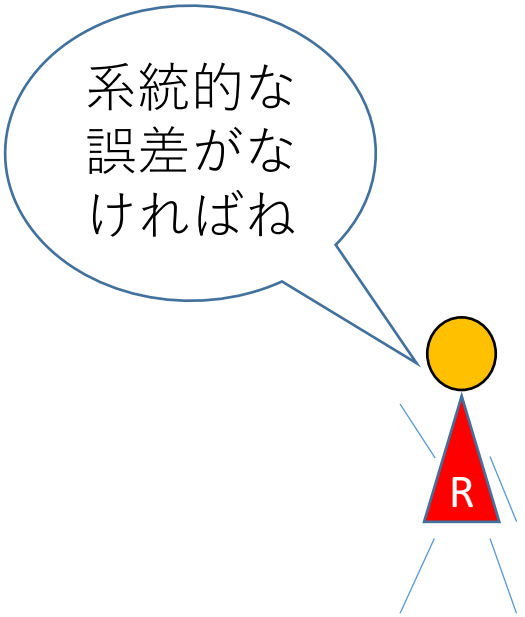
より正確な推定が可能だろうか

A:

測定を繰り返せば測定値をより正確に知ることができる

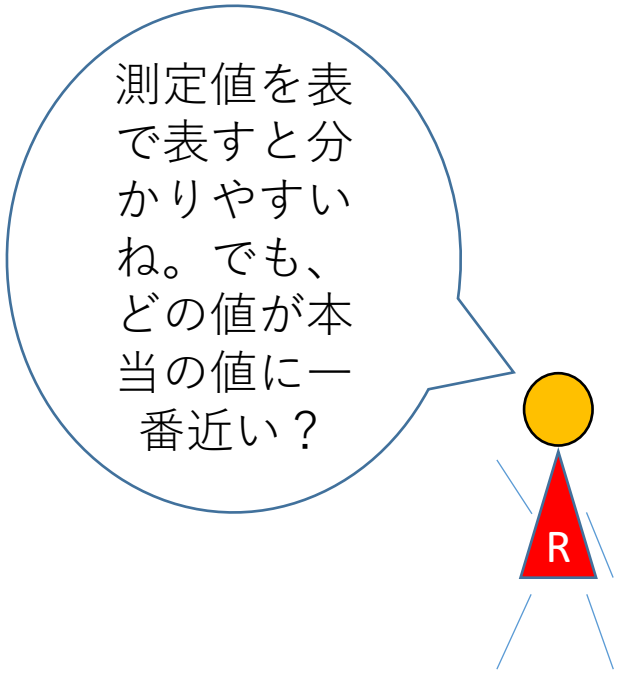
本当の値に迫る－複数の測定値

- 測定を繰り返すことでより正確にデータの性質を把握でき、また本当の値をより正確に推定できる。



複数の測定：本当の値に迫る

i	x_i
1	17.2
2	17.5
3	17.1
4	17.8
5	17.5
6	17.6
7	17.3
8	18.1
9	17.6
10	17.8
11	17.8
12	17.6



最良推定値

$$X = X_{best} \mp \delta x$$

変量　　最良推定値　　誤差

データーの3つの代表地

平均値　Mean

中央値　Median

最頻値　Mode

平均値　Mean

- 平均値の定義：

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$\bar{x}$：平均値を表す記号

N – データの数

メリット: すべてのデータを考慮しているのでデータ解析に便利

デメリット：外れ値（異常値）に強く影響を受ける

測定値:　6,12,5,3,8,4,5,11,3,3,(25)

i	x_i
1	6
2	12
3	5
4	8
5	3
6	4
7	5
8	11
9	3
10	3
11	25

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

N=10　異常値が含まれないとき

$$\sum_{i=1}^{10} x_i = 6+12+5+3+8+4+5+11+3+3=60$$

$$\bar{x} = \frac{60}{10} = 6$$

N=11　異常値が含まれるとき

$$\sum_{i=1}^{11} x_i = 6+12+5+3+8+4+5+11+3+3+25=85$$

$$\bar{x} = \frac{85}{11} \cong 7.7$$

中央値　Median

- データを半分に分ける値

奇数の場合：　$x_1, x_2, \ldots, x_{\frac{N-1}{2}}, x_{median}, x_{\frac{N+1}{2}} \ldots, x_{N-1}, x_N$

偶数の場合：　$x_1, x_2, \ldots, x_{\frac{N}{2}}, x_{\frac{N}{2}+1} \ldots, x_{N-1}, x_N$

$$x_{median} = \frac{\left(x_{\frac{N}{2}} + x_{\frac{N}{2}+1}\right)}{2}$$

メリット：　外れ値の影響を受けない

デメリット：　中央のデータの値しか考慮しない

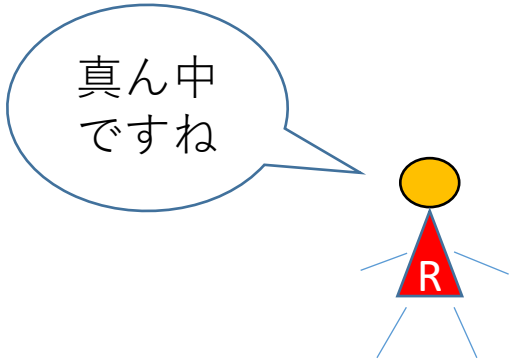
測定値の数が奇数の場合

測定値：6,12,5,3,8,25,4,11,3,3,5

並び変えると：3,3,3,4,5,5,6,8,11,12,25

3,3,3,4,5,5,6,8,11,12,25

中央値 = 5



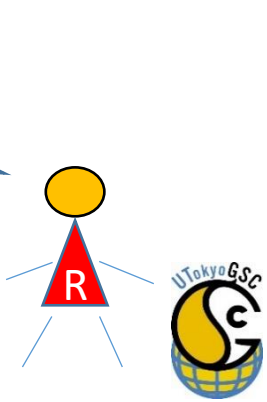
測定値の数が偶数の場合

測定値：6,12,5,3,8,4,5,11,3,3

大きさ順に並び変えると：3,3,3,4,5,5,6,8,11,12

3,3,3,4,5,5,6,8,11,12

$$\text{中央値} = \frac{5+5}{2} = 5$$



最頻値　Mode

- 頻度の最も高い要素
- 連続型データにふさわしくない

測定値：6,12,5,3,8,25,4,5,11,3,3

順番に並び変えると：3,3,3,4,5,5,6,8,11,12,25

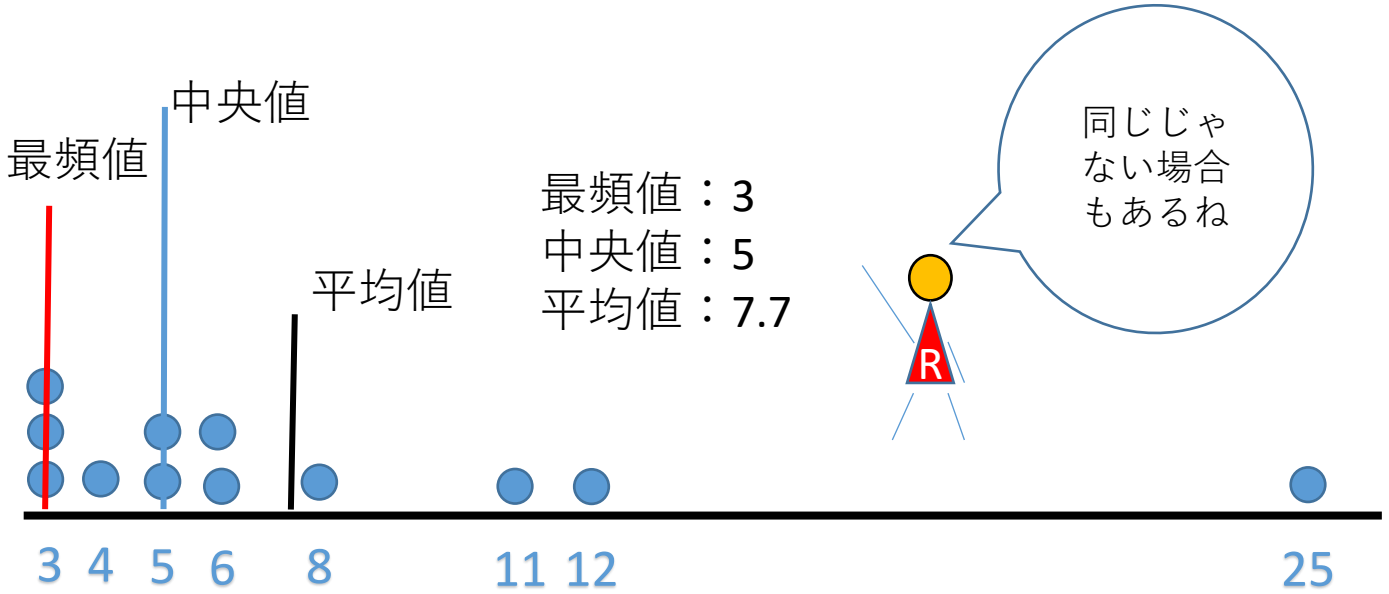
最頻値 = 3

メリット：　計算しやすい

デメリット：　データの数が少ないと意味がない、数え方によります

ミニまとめ

測定値：3,3,3,4,5,5,6,8,11,12,25



コメント

- 最頻値、中央値または平均値のなかでどれが一番適切かは研究の目的やデータの性質による。
- 科学研究では平均値が最も使われている代表値。
- 測定の本当の値の推定にも平均値が一番適切。（その理由はあとで説明します）

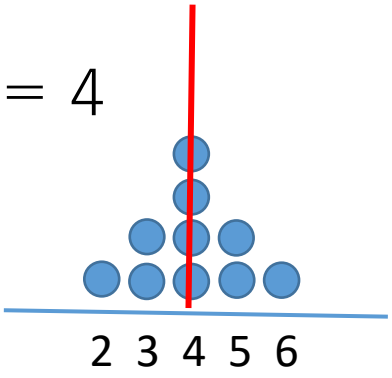
$$X_{best} = \bar{X}$$

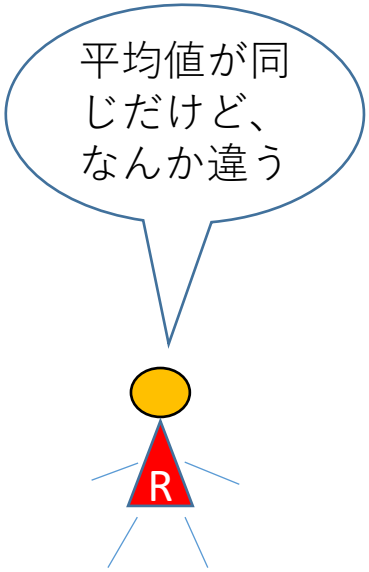
ばらつき

- 本当の値の最良推定値を平均値にしても誤差がなくならない。
- 次は平均値がどれほどいい推定かを考えよう

データA: 5,2,6,4,3,4,4,3,4,5

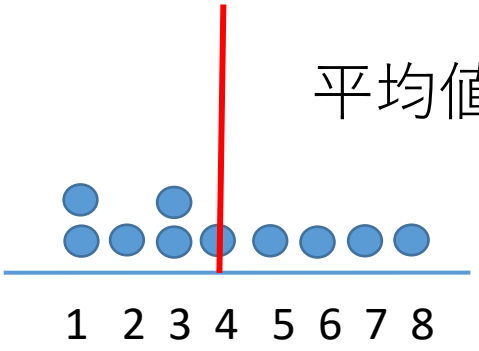
平均値A＝ 4





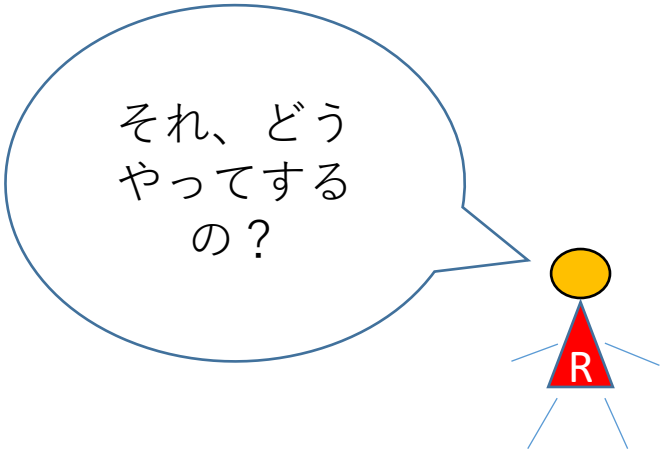
データB: 1,7,2,6,3,1,3,4,5,8

平均値B＝ 4



ばらつき

- データを理解するには平均値だけは足りない。
- 平均がどれほどデータの代表値なのかを知るために、データがどのように平均値の周りに配置されているかを把握しないといけない。

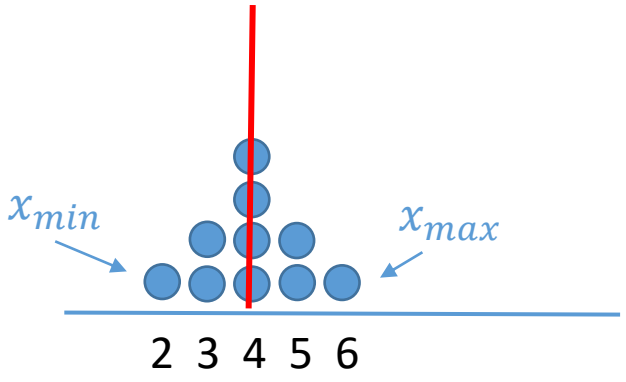


レンジ　Range

- データの範囲（最大値と最小値の差）からデータのばらつきがだいたい把握できる

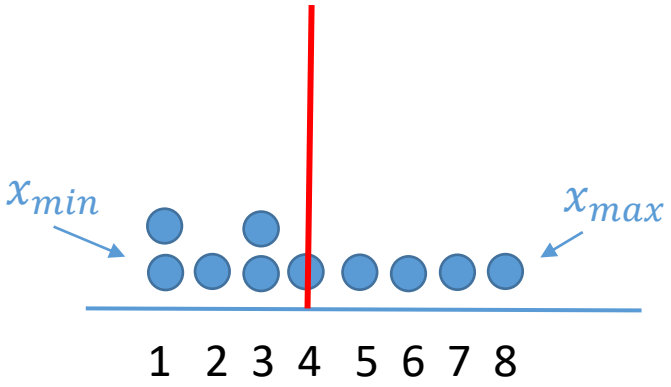
$$R = x_{max} - x_{min}$$

A: 5,2,6,4,3,4,4,3,4,5



R＝6-2＝4

B: 1,7,2,6,3,1,3,4,5,8



R＝8-1＝7

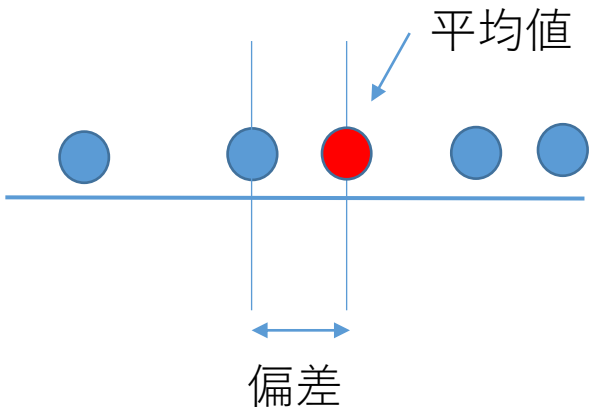
コメント

- レンジは最大値と最小値の間にデータがどのように配置されているかは考慮しない。
- すべてのデータを考慮する方法があったらいい
- 各データと平均値との平均距離はどうだろう。

平均偏差

偏差（平均値との距離）：$\Delta x_i = x_i - \bar{x}$

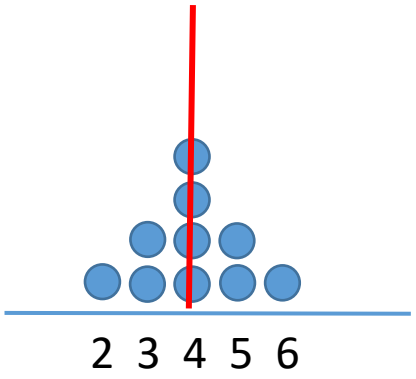
$$\text{偏差の平均} = \frac{\text{偏差の総和}}{\text{データの総数}} \qquad \overline{\Delta x} = \frac{\sum_{i=1}^{N} \Delta x_i}{N}$$



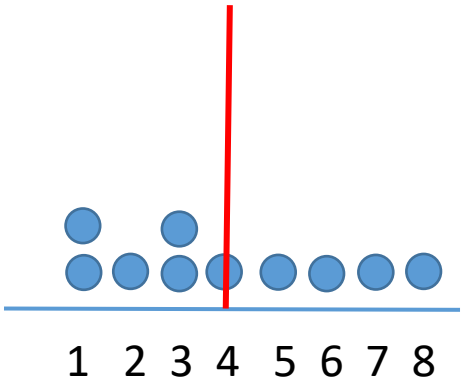
- 平均偏差を計算してみよう

i	x_i	偏差
平均値＝4		
1	5	1
2	2	-2
3	6	2
4	4	0
5	3	-1
6	4	0
7	4	0
8	3	-1
9	4	0
10	5	1
偏差の総和＝0		

A: 5,2,6,4,3,4,4,3,4,5

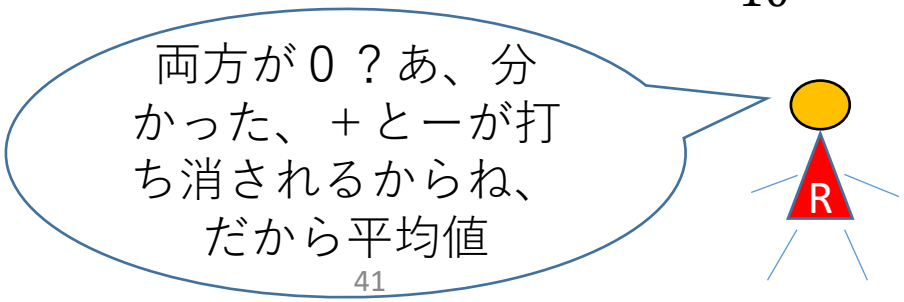


B: 1,7,2,6,3,1,3,4,5,8



$$\text{平均偏差} = \frac{\text{偏差の総和}}{\text{データの数}}$$

$$\text{平均偏差}A = \frac{0}{10} = 0 \qquad \text{平均偏差}B = \frac{0}{10} = 0$$



i	x_i	偏差
平均値＝4		
1	1	-3
2	7	3
3	2	-2
4	6	2
5	3	-1
6	1	-3
7	3	-1
8	4	0
9	5	1
10	8	4
偏差の総和＝0		

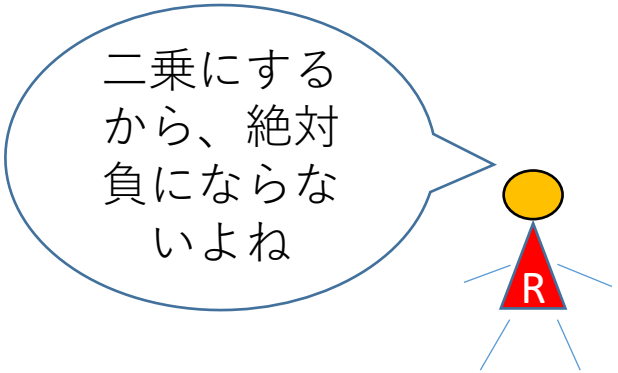
偏差平方和

- 平均偏差はうまくいかなかった。ほかの方法はないだろうか。

データの平均偏差平方和を計算してみよう

$$\text{偏差平方和} = (x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_N-\bar{x})^2$$

$$\text{偏差平方和} = \sum_{i=1}^{N} (x_i-\bar{x})^2$$



分散 variance

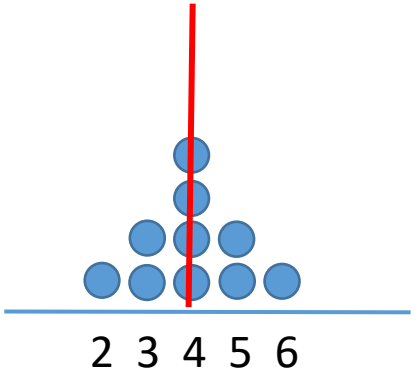
平均偏差平方和を分散と呼び、s^2と書く。

$$\text{分散} s^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_N-\bar{x})^2}{N}$$

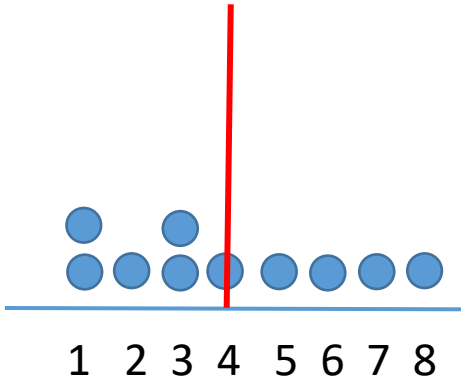
$$s^2 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N}$$

i	x_i	偏差	偏差の二乗
平均値＝4			
1	5	1	1
2	2	-2	4
3	6	2	4
4	4	0	0
5	3	-1	1
6	4	0	0
7	4	0	0
8	3	-1	1
9	4	0	0
10	5	1	1
総和		0	12

A: 5,2,6,4,3,4,4,3,4,5

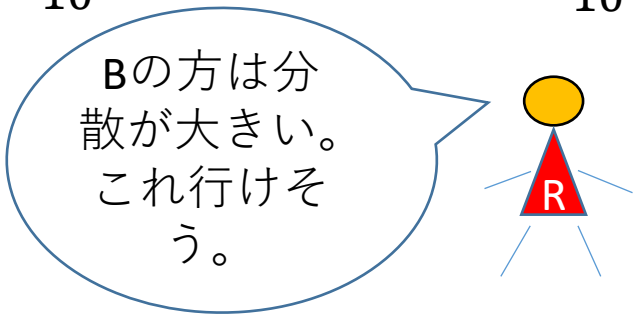


B: 1,7,2,6,3,1,3,4,5,8



$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$$

$A: s^2 = \frac{12}{10} = 1.2$　　$B: s^2 = \frac{54}{10} = 5.4$



i	x_i	偏差	偏差の二乗
平均値＝4			
1	1	-3	9
2	7	3	9
3	2	-2	4
4	6	2	4
5	3	-1	1
6	1	-3	9
7	3	-1	1
8	4	0	0
9	5	1	1
10	8	4	16
総和		0	54

標準偏差値　Standard Deviation

- 分散のルートを標準偏差と呼ぶ

$$\text{標準偏差値} s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

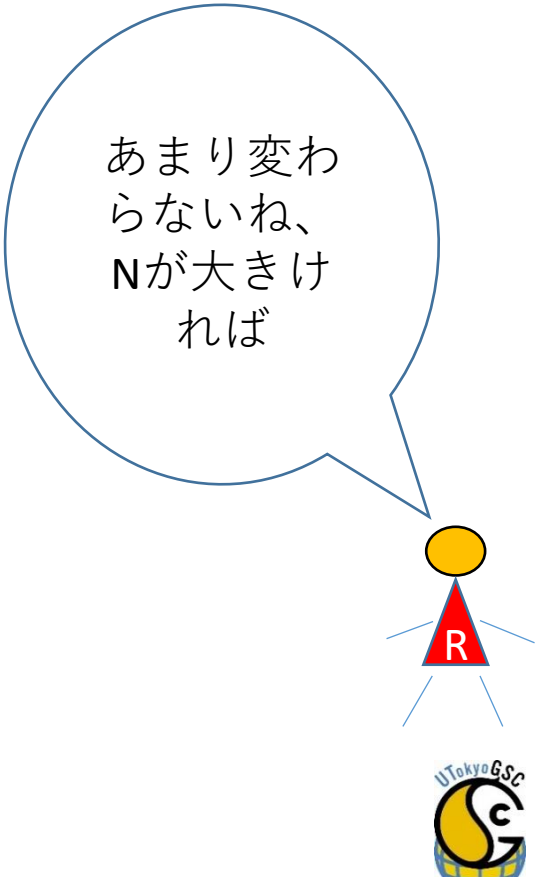
SDまたはSDTと書く場合もある

- 標準偏差値は平均値と同じ単位を持つ。

標準偏差の改善

- より複雑な数学を使えば、データの標準偏差をより正確に推定できる。

$$s(x) = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{\mathrm{N} - 1}}$$

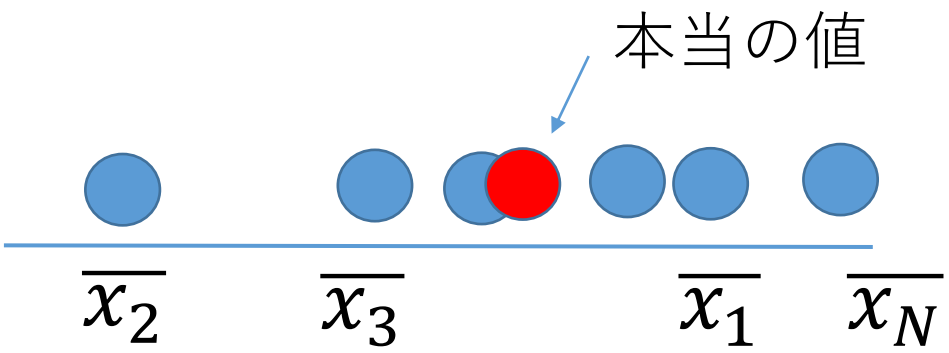


コメント

- データの分散と標準偏差を知るのはデータの性質を理解するためにとても重要。
- ただ、平均値の誤差を推定するためには、分散はそのまま使えない。
- 平均値の誤差を計算するためには、平均値が本当の値の周りにどのように配置されている（平均値の分散）かを知るのことが必要。

平均の分散

$$\mathrm{X}_1 = \{x_{11}, x_{12}, x_{13}, \ldots, x_{1N}\} \rightarrow \overline{x_1}$$
$$\mathrm{X}_2 = \{x_{21}, x_{22}, x_{23}, \ldots, x_{2N}\} \rightarrow \overline{x_2}$$
$$\mathrm{X}_3 = \{x_{31}, x_{32}, x_{33}, \ldots, x_{3N}\} \rightarrow \overline{x_3}$$
$$\vdots$$
$$\mathrm{X}_N = \{x_{N1}, x_{N2}, x_{N3}, \ldots, x_{NN}\} \rightarrow \overline{x_N}$$



平均値の標準偏差

- 平均値の分散と標準偏差を計算するには高校のレベルを超えた数学が必要。

$$\text{平均の偏差値} s(\bar{x}) = \frac{\text{データの標準偏差値}}{\sqrt{\mathrm{N}}}$$

$$s(\bar{x}) = \frac{s(x)}{\sqrt{N}} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N.(\mathrm{N}-1)}}$$

平均値の標準偏差はデータの標準偏差より小さい。

本当の値に迫る

- 測定データの平均と標準偏差値を利用して、本当の値が推定できる。

$$X = X_{best} \mp \delta x$$

↓

$$x = \bar{x} \mp s(\bar{x})$$

誤差の伝搬

- 直接測定できない、数式で定義された変量の誤差は誤差の伝搬法で求める。

誤差伝搬法

1. 変量の平均値は数式から計算する。
2. 変量の誤差は次の誤差伝搬ルールに従って計算する。

誤差伝搬法則

- 足し算、引き算

$$Z = X \mp Y$$

$$\delta Z = \delta X + \delta Y$$

絶対誤を足す

- 掛け算、割り算

$$Z = XY \quad Z = \frac{X}{Y}$$

$$\delta_r Z = \delta_r X + \delta_r Y$$

相対誤差を足す

- 関数

$$Z = g(X)$$

$$\delta Z = g(X + \delta X) - g(X)$$

δX が小さい場合

$$\delta Z = \left| \frac{d}{dX} g(X) \right| \delta X$$

例

- 高さh_0から落下する物体の位置は下記の数式で表す

$$h(h_0, t) = h_0 - 5t^2$$
$$h_0 = 10.0m \pm 0.1m$$

時間　$t = 1.3s \pm 0.03s$　が立った時の物体の位置 h を求めよう

$$\delta(h_0) = 0.1$$

$$\delta(t) = 0.03$$

$$\delta(h_0) = 0.1 \qquad \delta(t) = 0.03 \qquad \delta_r t = \frac{\delta t}{t} = \frac{0.03}{1.3} = 0.02$$

$$\delta_r t^2 = \delta_r(t \times t) = \delta_r t + \delta_r t = 2\delta_r t$$

$$\frac{\delta(t^2)}{t^2} = 2\frac{\delta t}{t} \qquad \delta(t^2) = 2t\delta t = 2 \times 1.3 \times 0.03 = 0.078$$

$$\delta h = \delta h_0 + 5 \times \delta(t^2)$$
$$\delta h = 0.1m + 5 \times 0.078m$$
$$\delta h \cong 0.50m$$

$$h(t)_{best} = 10 - 5 \times 1.3^2 = 1.55m$$

$$h(t) = 1.6m \pm 0.5m$$

測定データを違う角度から見てみよう

測定値：6,5,5,4,4,5,7,6,3,5

大きさの順に並び変えると：3,4,4,5,5,5,5,6,6,7

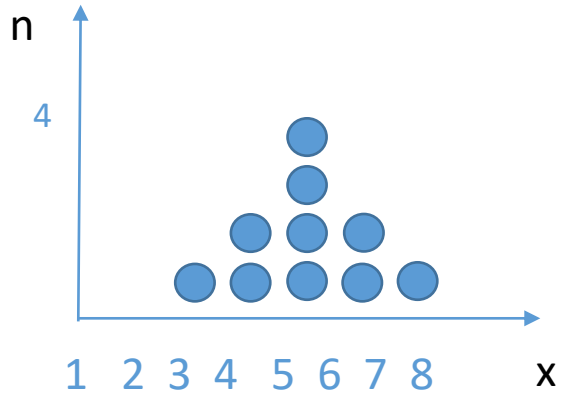
並び変えたデータの平均は下記のように計算できる

$$\bar{x} = \frac{1 \times 3 + 2 \times 4 + 4 \times 5 + 2 \times 6 + 1 \times 7}{10}$$

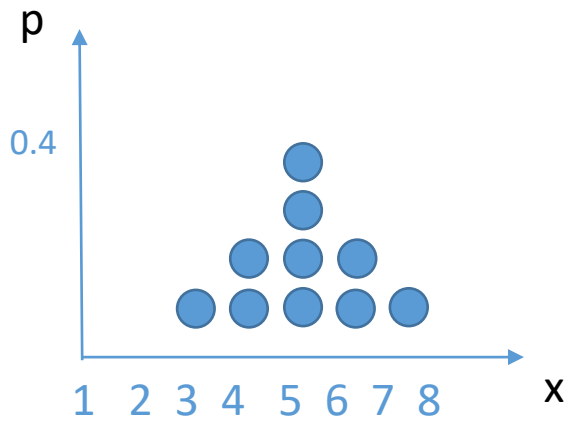
$$\bar{x} = \frac{1}{10} \times 3 + \frac{2}{10} \times 4 + \frac{4}{10} \times 5 + \frac{2}{10} \times 6 + \frac{1}{10} \times 7$$

ヒストグラムと確率分布

x_i	3	4	5	6	7
度数	1	2	4	2	1
p_i	0.1	0.2	0.4	0.2	0.1



ヒストグラム



確率分布

度数をデータの数で割ると確率分布が得られる。

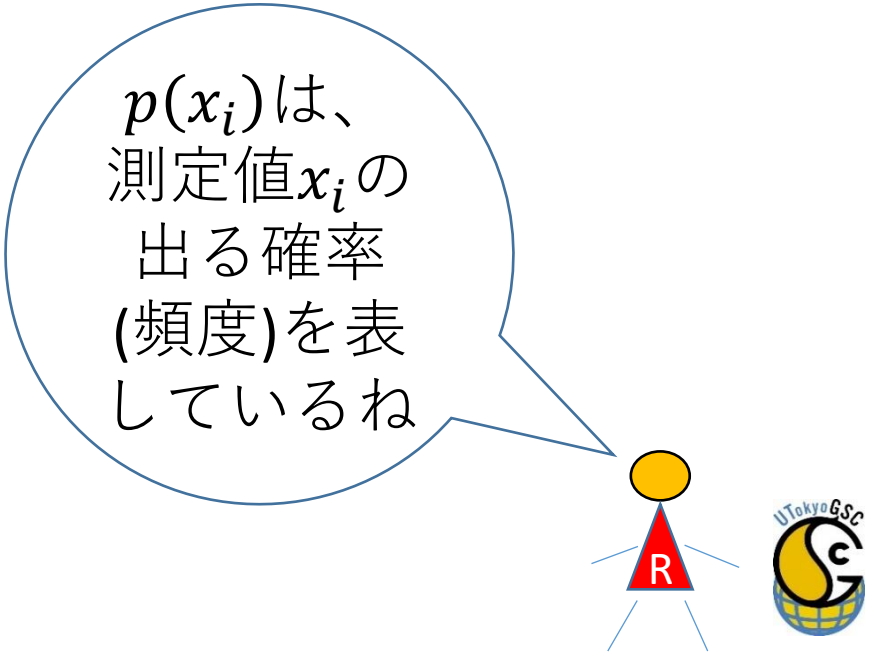
一般的に

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_M x_M}{N}$$

$$\bar{x} = \frac{n_1}{N} x_1 + \frac{n_2}{N} x_2 + \cdots + \frac{n_M}{N} x_M = \sum_{i=1}^{M} \frac{n_i}{N} x_i = \sum_{i=1}^{M} p_i x_i$$

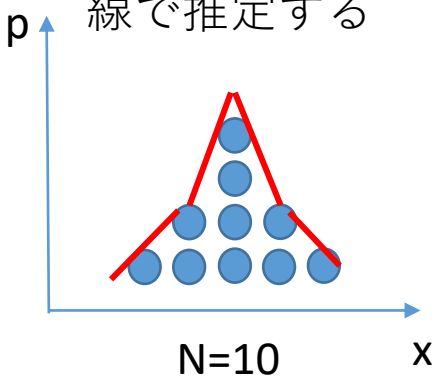
$$p_i = p(x_i) = \frac{n_i}{N}$$

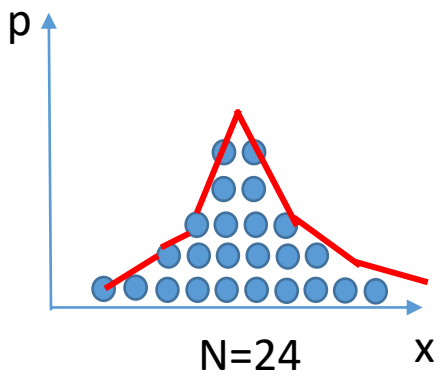
$$\sum_{i=1}^{M} n_i = N$$

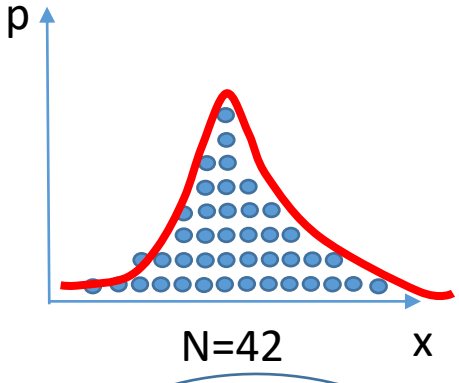


- 測定値の数が多ければ多いほど変量の確率分布がより正確に分かる。

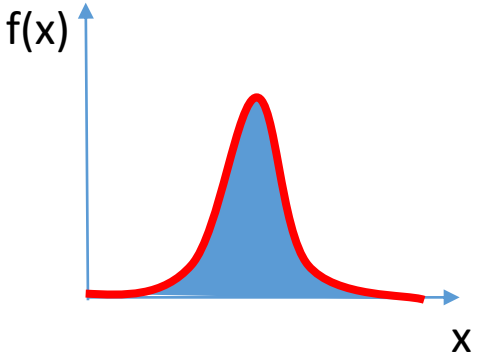
確率分布の形を折れ線で推定する





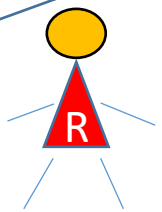


Nが無限になると確率分布が曲線になる。この曲線は確率密度関数と呼び、f（x）と書く。



$N \to \infty$

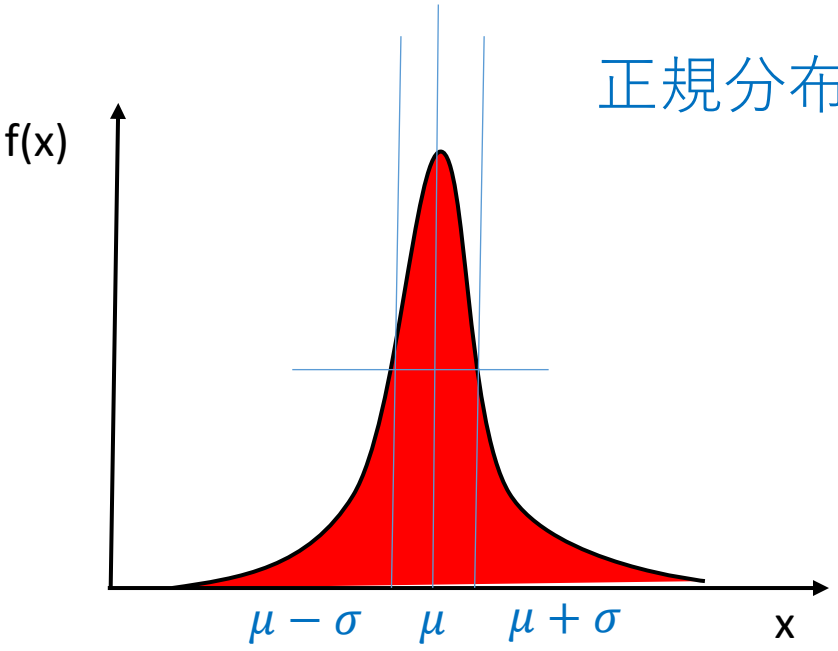
実験データをたくさん集めると確率分布の本当の形が見えてくるね



コメント

- 変量の確率分布（確率密度関数）の形は分からない場合が多いが、多くの測定値を集めれば、分布の形を推定できる。
- 確率分布の平均値は期待値と呼び、μと書く。
- 確率分布の標準偏差をσと書く。
- 統計学で一番よく使われている分布は正規分布。

正規分布 Normal Distribution



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

$f(x)=N(\mu,\sigma)$

μ：期待値

σ：分散

$x-\mu$：偏差

正規分布を知るには期待値μと分散σを知る必要がある。

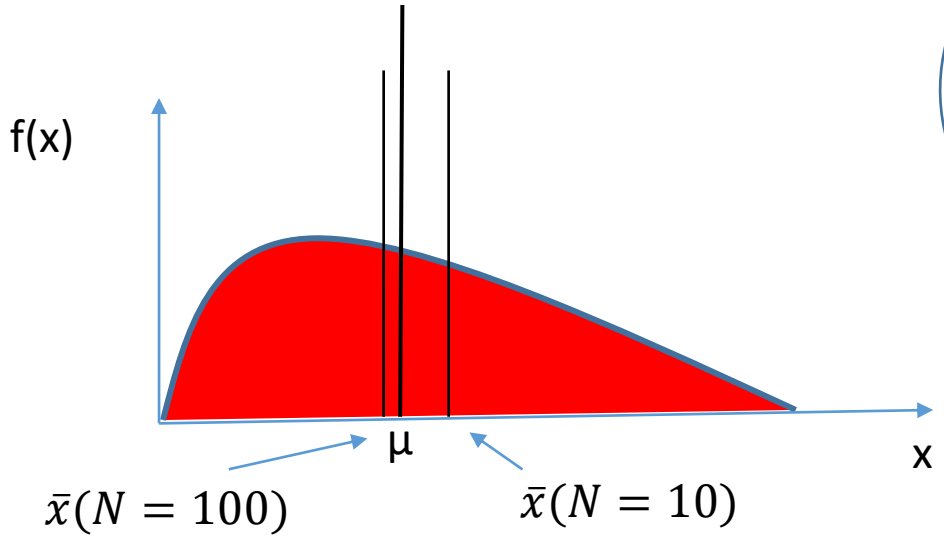
測定値の誤差は規定分布に従うと考えられる。

２つの重要な法則

第１：大数の法則　（任意の分布）

- 測定値の平均値はデータの数が多ければ多いほど期待値に近づく

変量の分布の期待値がμとする

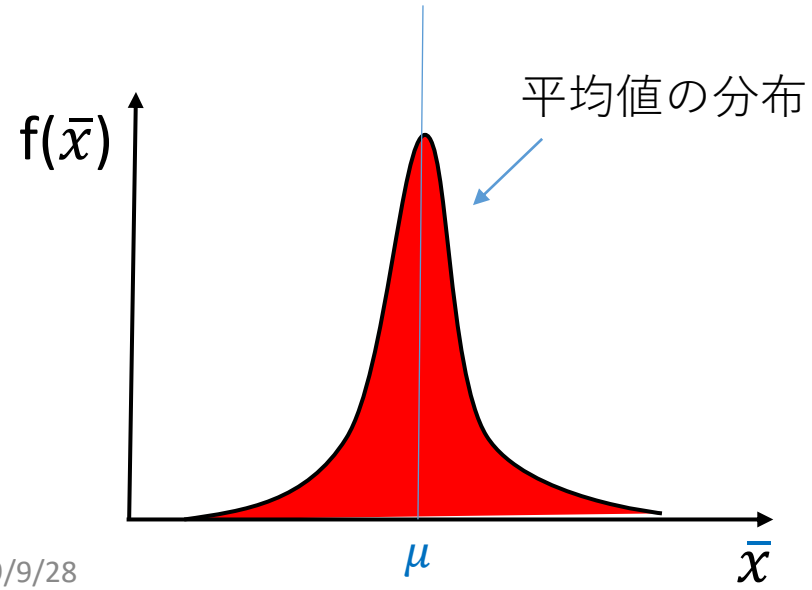


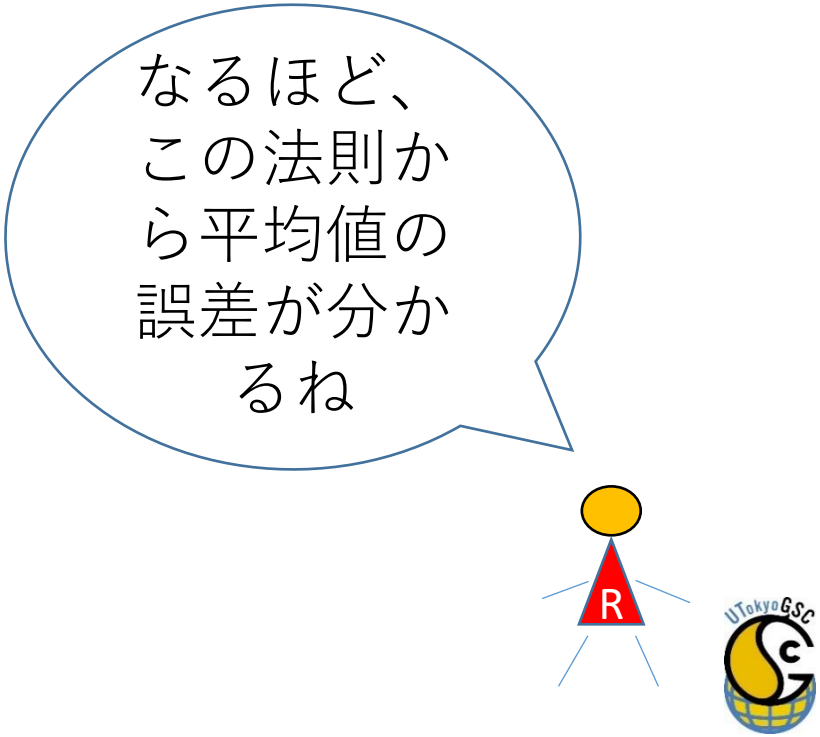
２つの重要な法則

- 第２：中心極限法則　（任意の分布）

データの平均値の分布はデータの数につれて正規分布$N(\mu, \sigma^2/n)$に近づく（μとσは変量の分布の期待値と分散）

変量の分布の期待値がμとする





まとめ

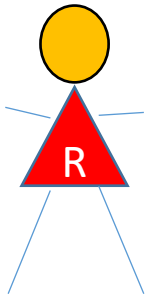
測定値の真値を推定するために複数の測定が必要

データの平均値は測定の真値の最良推定

真値の最良推定の誤差はデータの標準偏差とデータの数から推定できる

$$x = \bar{x} \mp s(\bar{x}) \qquad s(\bar{x}) = \frac{s(x)}{\sqrt{N}}$$

THE END



参考文献

- 統計学の基礎のキ 分散と相関係数編；石村卓夫、石村光資朗；東京図書
- 統計学の基礎のソ 正規分布とt分布編；石村卓夫、石村光資朗；東京図書
- 統計学の図鑑；涌井 良幸；技術評論社
- Research Methods and Statistics: A Critical Thinking Approach, Sherri L. Jackson, Wadsworth Pub Co

宿題

1. 気温を測定する場合の系統的とランダム誤差の例をあげよ。
2. 時間 t で物体が移動した距離は次の数式で表せる。

$$s(t) = s_0 + 10t^3$$

時間t_1の時の位置の推定をもとめよ。s_0とt_1の内、物体の位置への影響が大きいほうを選べ。

$s_0 = 0.60m \pm 0.01m$
$t_1 = 3.00s \pm 0.05s$

$[s(t_1) = 271\text{m} \pm 14\text{m}]$

3. 次のデータの平均値$\bar{x}$、中央値、最頻値を求めよ。

i	1	2	3	4	5
x_i	1.1	1.2	1.2	1.4	1.2

$\left[\bar{x} = 1.22, \textit{中央値} = 1.2, \textit{最頻値} = 1.2\right]$

4. 3 番のデータを使って、データの分散s^2、データの標準偏差 s 、そして平均の標準偏差$s_{\bar{x}}$を求めよ。

$[s^2 = 0.012, s \cong 0.1095, s_{\bar{x}} \cong 0.049]$

5. 平均値が測定値の真値を推定するのに適切である理由を述べよ。