統計データの扱い

樋口さぶろお

龍谷大学理工学部数理情報学科

数値計算法 L10(2010-06-25)

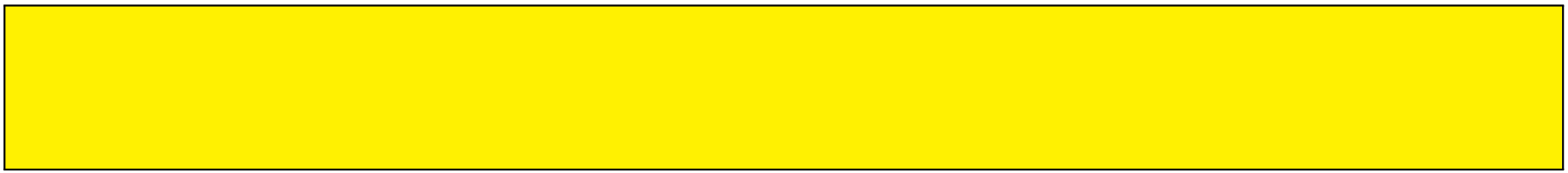
今日の目標

1. テキストファイルにフォーマットに従って保存されたデータを読み込んで使えるようになろう.
2. データから, 平均, 分散, 標準偏差を計算できるようになろう.

hig3.net

データの平均

n 個のデータがあったとしよう.
$x_1 = 10, x_2 = 6, \ldots, x_n = 89.$
データの特徴を 1 個の数値で代表させるのが平均



平均 mean (average)

$$\text{平均 } m = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

算術平均, 相加平均
別の意味で標本平均とも. ($\leftrightarrow$ 分布平均)

確率・統計

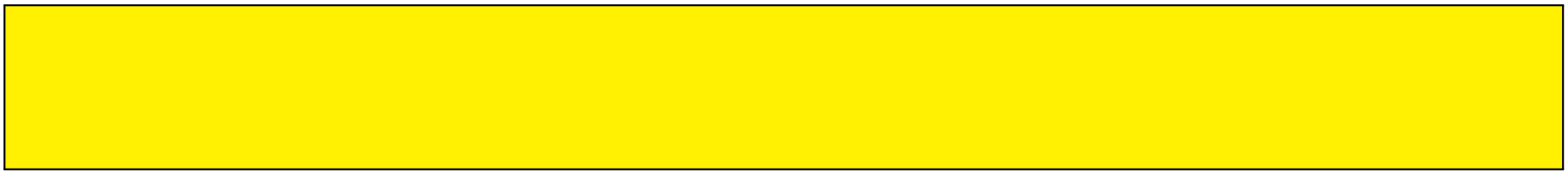
プログラミングののり
$\sum$ の計算は今までと同じのり. x_i が漸化式や $f(x)$ から計算されるのでなく, データとして与えられているところだけが異なる.

平均って本当に代表的な数値なの?

太陽系内の天体の質量の平均は?

$$\frac{1}{n}(\text{太陽質量} + \text{木星質量} + \cdots) = \frac{1}{n}(2 \times 10^{30} + 2 \times 10^{26} + \ldots + \cdots)?$$

X 帝国に住んでいる人の資産の平均は?



代用品

-  順位が真ん中のデータをそのまま答える
-  いちばん多数回現れたデータを答える

いろんな平均

4 個の星があり, それぞれ 1,2,3,6 等星. 平均は 4 等星?

明るさの平均は $\frac{1+100^{-1/5}+100^{-4/5}+100^{-5/5}}{4}$

平均は $1+5\log_{100}\frac{1+100^{-1/5}+100^{-4/5}+100^{-5/5}}{4}$?

平均は一通りじゃない. どれを使うか決める簡単な原理もない. データがどんな確率分布に従ってるかわかれば…

算術 (相加) 平均 $\frac{1}{n}\sum_{i=1}^{n} x_i$

幾何 (相乗) 平均 $\left(\prod_{i=1}^{n} x_i\right)^{1/n}$

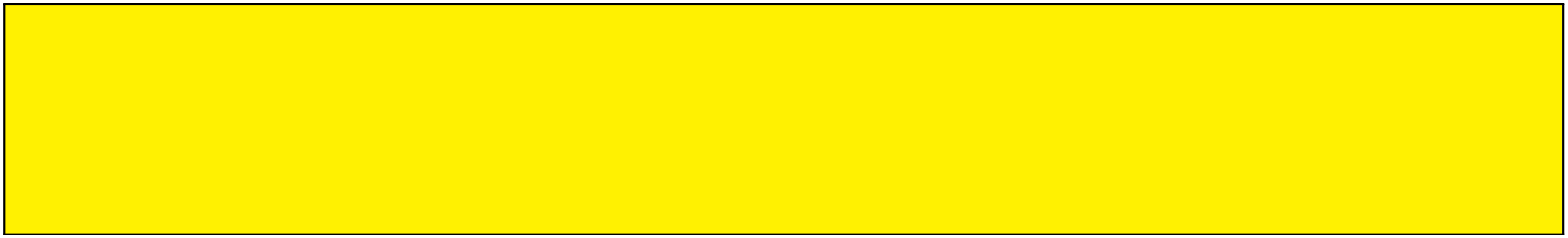
対数平均 対数をとってから平均する.

調和平均 $\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$

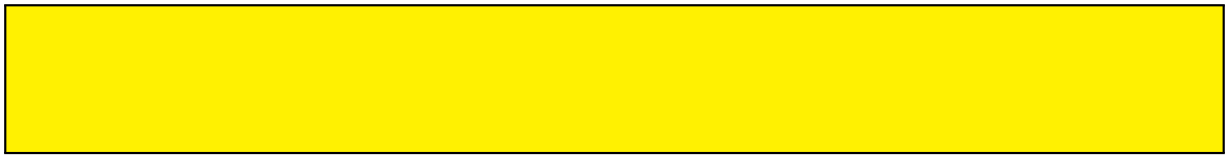
相対運動に出てくる換算質量は $n=2$ の調和平均/2.

分散

平均からのずれ具合の平均をはかろう!



案 1 $\frac{1}{n}\sum_{i=1}^{n}(x_i - m)$



案 2 $\frac{1}{n}\sum_{i=1}^{n}|x_i - m|$

案 3 $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$　単位が平均とは違う

S^2 の従う有名で便利な等式

$$
\begin{aligned}
S^2 =& \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 \\
=& \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2mx_i + m^2) \\
=& \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2m\frac{1}{n}\sum_{i=1}^{n}x_i + m^2\frac{1}{n}\sum_{i=1}^{n}1 \\
=& \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2m\cdot m + m^2 \\
=& \frac{1}{n}\sum_{i=1}^{n}x_i^2 - m^2 \quad \text{プログラムではこれを使ってみよう}
\end{aligned}
$$

この式はプログラミング的には便利. 先に平均を計算せずに, $\sum_i x_i^2, \sum_i x_i$ を求めればいい.
情報落ち, 桁落ちの可能性.

標本平均 mean, average

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

標本分散 variation

$$\sigma^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - m^2\right)$$

標本標準偏差 standard deviation

$$\sigma = \sqrt{\frac{n}{n-1}S^2}$$

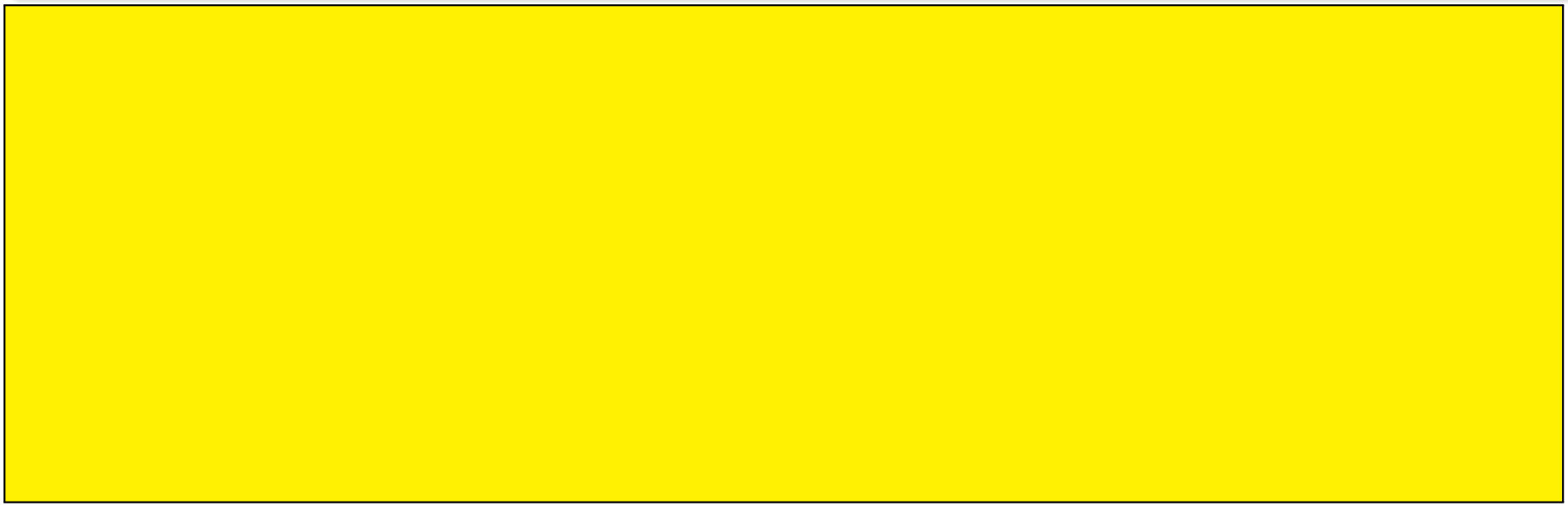
平均からの定型的なずれ. 平均と同じ単位.

σ^2 を不偏 (標本) 分散, S^2 を標本分散という流儀もある.
'分布分散'

Quiz

次のデータを考える: *9 10 12 12 12*

1. 標本平均を求めよう
2. σ^2 を両方の方法で求めよう.



Quiz

1 学年 *2* クラスの学年で期末テストが行われました. 次のようなことは可能でしょうか? '生徒を *1* 人だけクラス間で移籍させる *(*移籍していたことにする*)* ことによって, *2* クラス両方の平均点を上げる'

計算機で扱う方法

- 統計計算ソフトウェア 例 SPSS S SAS R
- 数式処理ソフトウェア 例 Mathematica Maple
- 表計算ソフトウェア 例: Excel
- 手で 例 C Java ...

平均・分散を求めるプログラムの概要

配列を使って x_i を記憶しなくてもいい.

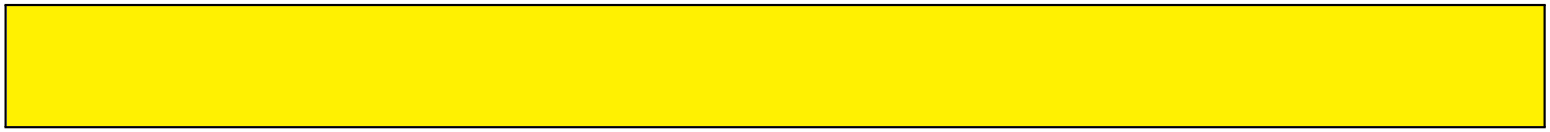
```
int main(void){
    double x,       /* データ */
    int n=0;        /* データの個数 */
    double s=0.0;       /* ∑_i x_i */
    double s2=0.0;      /* ∑_i x_i^2 */

    while( .. ){
        scanf("%lf",&x);
        s=s+x;
        s2=s2+x*x;
        n=n+1;
    }

    /* 平均・分散を s, s2から計算 */
    return 0;
}
```

プログラムにデータを与える方法

- scanf でデータを与える → 人にやさしくない　計算機基礎実習
- プログラム内に配列の初期化として書いておく →



計算機基礎実習

- 別のファイルに書いておく　今回
- 標準入力で与える → Linux(UNIX) では標準的．Windows では一般的でないかも?　そのうち

テキストファイルからのデータの読み取り

ふつうのコンパクトな書き方

```
double x;
FILE *fp; /* ファイルポインタ = ファイル型変数へのポインタ */
fp = fopen("planet.txt", "r"); /*ファイルをオープン*/

while( fscanf(fp, "%lf", &x)!=EOF ){ /*読み取ってxに代入，
                                       ファイルの最後に来たら終了*/

  /* x を使った処理をここに書く */
}
fclose(fp);                          /* ファイルをクローズ */
```

同値だけど説明の詳しいプログラム例

```
double x;
FILE *fp;
int result;

fp = fopen("planet.txt", "r"); /* ファイルをオープンする */

while( 1 ) {
  result=fscanf(fp, "%lf", &x); /*データはxに,
                                  成功失敗は result に*/
  if( result == EOF ){ /* End Of File に来た * */
    break;
  }

  /* x を使った処理  */

}
fclose(fp);                    /* ファイルをクローズする */
```

fopen

1. planet.txt は, 同じディレクトリ (フォルダ) にあるテキストファイル (メモ帳で開けるようなファイル)
2. 同じディレクトリにないときは, 相対パスまたは絶対パスで `Q:¥¥nc¥¥planet.txt` のように指定する. なぜ `¥¥`? だって `Q:¥nc` じゃ改行しちゃいそう.
3. "r" 先頭から読み込み. "w" 消去して先頭から書き込み. "w+" 追記. など. マニュアル見よう.
4. FILE 型変数へのポインタ (ファイルポインタ) を返す.

fscanf

1. scanf はコンソール (キーボード) から読み込むけど, fscanf は ファイルポインタ fp で指定されるファイルから読み込む.
2. 返り値は成功/失敗などを示す int.

fclose

1. これしないと USB フラッシュメモリ外そうとしたときにエラーになるかも.

欠席についての大事な連絡

- 就職活動, 介護体験, 病欠, 公務欠席などは, 欠席届 (教務課でもらってます) に証明する文書 (説明会案内, 病院のレシート) を添付して樋口に提出してくれれば考慮します. 教職センターの発行する欠席届はそれ 1 枚でいいです. 演習では, 欠席の次の回に, 今週, 先週, 先々週の分を 7 点で, 先々先週の分を 5 点でチェックします. TA にその旨伝えてチェックしてもらってください. ただし, 2 週連続して適用することはしません.
- 教育実習で 2 または 3 週間ぬけた人についても, 欠席届を出してくれれば考慮します. 具体的には, 復帰した最初の回の演習で, 不在の間に締切を迎えた分を減点なしでチェックします. Web 予約を使用せずに樋口に声をかけてください.

演習の課題チェックについての大事な連絡

- 演習での課題のチェックは, 時間が余っていても 12:15 までの予約の分で終了します. チェックの結果不完全であっても再度のチェックはしません.
- 演習での 12:15 以降に予約された質問は, 12:15 以前の予約すべてに対応した後, 余裕がある場合のみ対応します. ただし, 質問のみでチェックは行いません.
- 2010-07-17 土 に任意参加の演習の補講を行う予定. これ以降はチェックを行いません.