

情報科学 第8回 統計計算の基本

静岡理科大学 総合情報学部
幸谷智紀

本日の内容

- 次の概念を習得し, Excelで計算が出来るようになる。
 - 平均
 - 分散
 - 標準偏差
- 度数分布表が作成できる→

元になるデータ: x_1, x_2, \dots, x_n
 ... n 個の一次元データ

平均(算術平均)

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

- 全データの和をnで割ったもの
- Excelで計算する場合
 - Sum関数とCount関数を使う方法
 - Average関数を使う方法

	A	B	C	D
1	データ	53	75	57
2	sum関数	605		
3	count関数	10 平均		60.5
4	average関数	=AVERAGE(B1 :K1)		
5		=AVERAGE(数値1, [数値2], ...)		

中央値(メディアン)

- データを大きさ順に並べた時の中間の値(偶数個の時はその平均値)

例) 3, 3, 4, 4, 5, 6, 7, 7, 10, 10

平均値: $(3+3+4+4+5+6+7+7+10+10)/10 = 5.9$

中央値: 3, 3, 4, 4, **5, 6**, 7, 7, 10, 10



→ $(5+6)/2 = 5.5$

- フィルター→昇順 or 降順で並べ換え

- Median関数を使用

6
7
7
10
10
=MEDIAN(B18:B27)

分散(1/2)

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad (1)$$

$$= \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \mu^2 \quad (2)$$

- データの「散らばり具合」を表わす値
- Excelで計算する場合
 - Sum関数とCount関数を使う方法... (1)と(2)は同じ式
 - Varp関数を使う方法

分散(2/2)

	A	B	C		A	B	
1	データ	53	75		1	データ	53
2	sum関数	605			2	sum関数	605
3	count関数	10 平均			3	count関数	10 平
4	average関数	60.5			4	average関数	60.5
5					5		
6	(1)式の場合				6	(1)式の場合	
7	$x_i - \mu$	=B1-\$E\$1	14.5		7	$x_i - \mu$	-7.5
8	$(x_i - \mu)^2$	56.25	210.25		8	$(x_i - \mu)^2$	56.25
9	$1/n \sum (x_i - \mu)^2$	483.45			9	$1/n \sum (x_i - \mu)^2$	483.45
10	varp関数	=VARP(B1 :K1)			10	varp関数	483.45
11					11		
12	(2)式の場合				12	(2)式の場合	
13	x_i^2	=B1^2			13	x_i^2	
14	$1/n \sum x_i^2$	4143.7			14	$1/n \sum x_i^2$	4143.7
15	$1/n \sum x_i^2 - \mu^2$	483.45			15	$1/n \sum x_i^2 - \mu^2$	483.45
9	$1/n \sum (x_i - \mu)^2$	483.45					
10	varp関数	=VARP(B1 :K1)					
11		=VARP(数値1, [数値2], ...)					

↑(1)の場合

(2)の場合→

VARP関数を使用↓

標準偏差

$$\sigma = \sqrt{\sigma^2}$$

- 分散の平方根・・・分散の幾何平均
- Excelで計算する場合
 - Sum関数を使う方法・・・分散を求めて平方根
 - Stdevp関数を使う方法

練習問題1

- 次の二つのデータグループ(A), (B)の中央値(メディアン), 平均, 分散, 標準偏差をそれぞれ求めよ。但し, 数式に基づく計算結果と, Excel関数による結果と両方求め, 両者に違いがないことを確認せよ。

(A) 30, 63, 58, 36, 33, 5, 87, 3, 31, 60

(B) 85, 83, 46, 88, 75, 90, 66, 78, 75, 82

ヒストグラム(度数分布表)

- 大量のデータを整理する方法の一つ
- 1. 元データ x_1, x_2, \dots, x_n が属する実数の区間 $[a, b]$ を求める。ここで

$$a \leq \min_i x_i \text{ かつ } b \geq \max_i x_i$$

2. 区間 $[a, b]$ を N 分割 (普通は等分割) し, 各小区間の中心値を x'_i とする。
3. 各小区間に属するデータの個数を f_i とする。これを度数と呼ぶ。
4. x'_i, f_i に基づいて表・グラフにする。これをヒストグラム(度数分布表)と呼ぶ

ヒストグラムの例

$$x_1, x_2, \dots, x_n \in [a, b]$$

↓

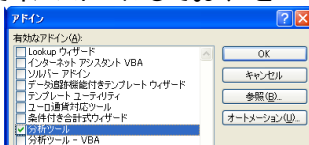
$$h = \frac{b-a}{N} \dots \text{各小区間の幅 (N等分割)}$$

小区間	級心	度数	相対度数	累積度数
$[a, a+h]$	$x'_1 = (a + (a+h))/2$	f_1	f_1/n	f_1
$[a+h, a+2h]$	$x'_2 = ((a+h) + (a+2h))/2$	f_2	f_2/n	$f_1 + f_2$
⋮	⋮	⋮	⋮	⋮
$[a + (n-1)h, b]$	$x'_N = ((a + (n-1)h) + b)/2$	f_N	f_N/n	$\sum_{i=1}^N f_i = n$

- 相対度数・・・データがその小区間に属する割合 = 確率

Excelでヒストグラムを計算するには

- メニューバーの「ツール」→「分析ツール」を用いるのが最も簡単
- 各自, 自分のExcelにこの項目があるかどうかを確認せよ!
- もしなければ, OfficeのCDを挿入し, 「ツール」→「アドイン」から「分析ツール」にチェックを入れてインストールしておくこと!



分析ツールを使わない度数分布表の作成

- Frequency関数を使う→Excelヘルプ参照
- Index関数と組み合わせて特定の区間に入るデータ数を指定して表示。

	得点分布	国語	数学	英語
	0~20	20	8	22
20	21~40	=INDEX(FREQUENCY(E5:E44,\$G\$14:\$G\$17),2)		
40	41~60	7	9	7
60	61~80	5	6	4
80	81~100	1	10	4