

問題の発見・分析 ～ データマイニングとグラフ ～

情報の科学 第16回授業

04情報の蓄積・管理

対応データ 18exp16.xls

「データ」と「情報」

- データ： 単なる数字や文字の羅列
例) 8890799568 ASAP
→ 価値を見いだしていない状態
- 情報： 意味のある数字や文字の羅列
例) 英数国理社の得点(88,90,79,95,68)
できるだけ早く！(As Soon As Possible)
→ そこに何らかの価値がある(ありそう)

※この場合の「価値の大小」については、個人差がある

→ 人によって「情報」か「データ」かが異なる場合がある

データマイニングとは

膨大なデータから、何らかの役に立ちそうな情報を
発見・採掘 (mining) すること

☆ビッグデータの活用

世の中にある、膨大なさまざまなデータを、社会・経済
の問題解決や業務の効率向上に役立てよう、という考
え方。

(ビッグデータ … 数十テラバイト～数ペタバイト
＝ 単純な半角の文字数にして数十兆から数千兆)

ビッグデータの活用例

- 膨大な検索語からWebサイトの広告
- 閲覧履歴から「お勧め」を出す(リコmend)
- SNS等からトレンドを分析し、新商品を開発
- 道路のセンサーから渋滞予測、信号制御
- コンビニエンスストアの売上データから年代別の売れ筋商品を見いだす
- クレジットカードの利用履歴から、不正利用パターンを見つけ犯罪防止に役立てる

データの集計

配布されたプリントにあるA組とB組それぞれのデータについて、

- 見やすくまとめる
- わかったこと

を別紙に書く(あとで提出・発表)

※Excelでのデータを利用したい方はどうぞ

分散・標準偏差・正規分布

同じ平均値でも、集団の性質が違う

- 大きい人と小さい人との差が大きいようだ
 - データの「偏り」を客観的に表す必要性
 - 偏りを数値化する必要性

例：A組 最大177.1 最小153.2 範囲23.9

B組 最大180.3 最小149.7 範囲30.6

データの偏りを表す数字

分散:

- ・それぞれのデータの平均値との差をとり、
- ・その差を二乗し、平均をとったもの

標準偏差:

- ・分散の正の平方根

備考:

偏差値・標準偏差をもとに、平均が50になるように数値化したもの

分散と標準偏差

(数学でもやるのでこの表をよく覚えておくこと)

| | 得点 | 平均との差 | 平均との差の2乗 |
|------------------|----|-------|-------------|
| A | 67 | 13 | 169 |
| B | 55 | 1 | 1 |
| C | 42 | -12 | 144 |
| D | 57 | 3 | 9 |
| E | 49 | -5 | 25 |
| 平均 | 54 | 0 | 分散 → 69.6 |
| 標準偏差(分散の正の平方根) → | | | 8.342661446 |

基本統計量

☆他にも、いろいろな「傾向」を数値で表せる。

中央値（メジアン）

最頻値（モード）

標準偏差

分散

尖度（せんど：ヒストグラムの「とがり具合」）

歪度（わいど：ヒストグラムの「左右対称性」）

範囲（レンジ）

最小

最大

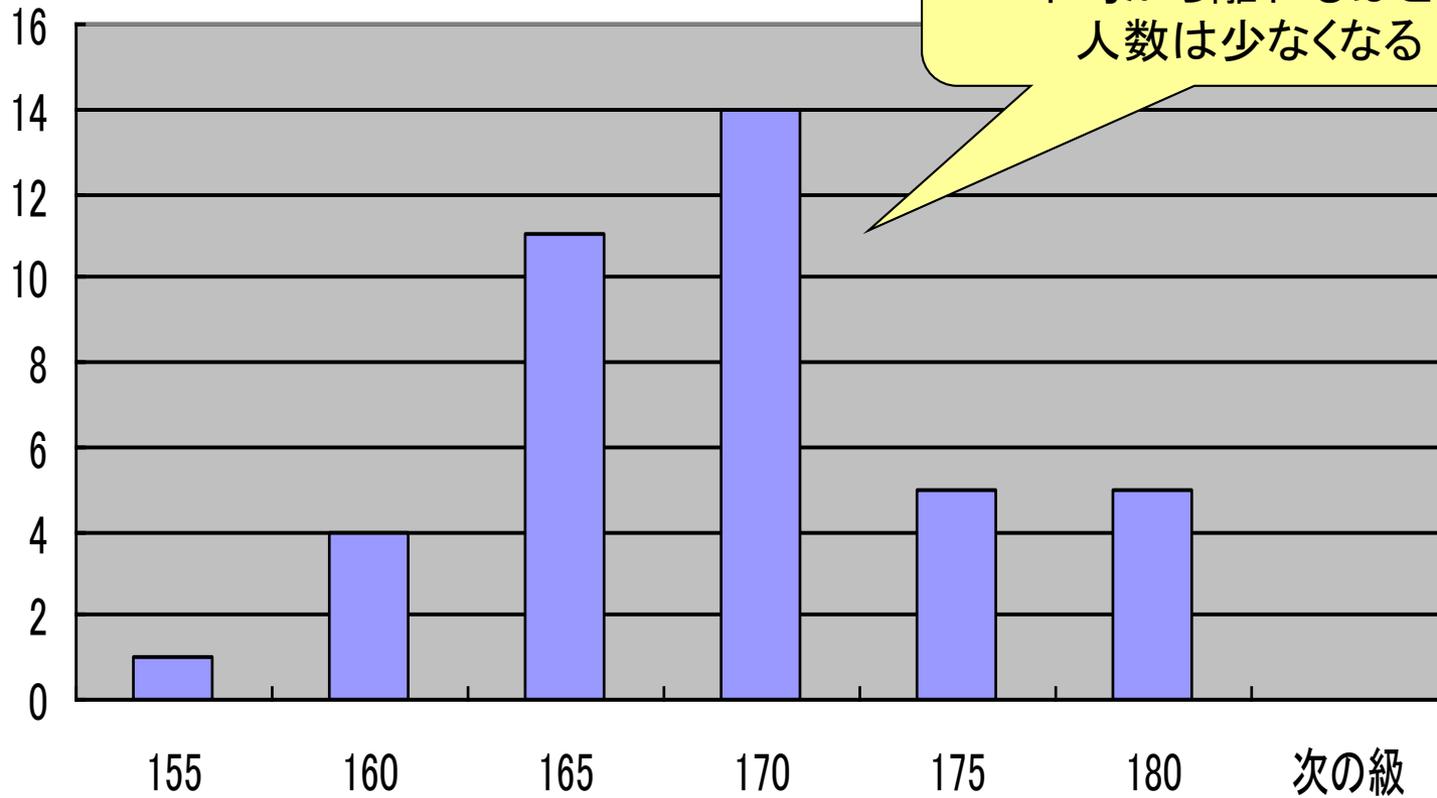
合計

標本数

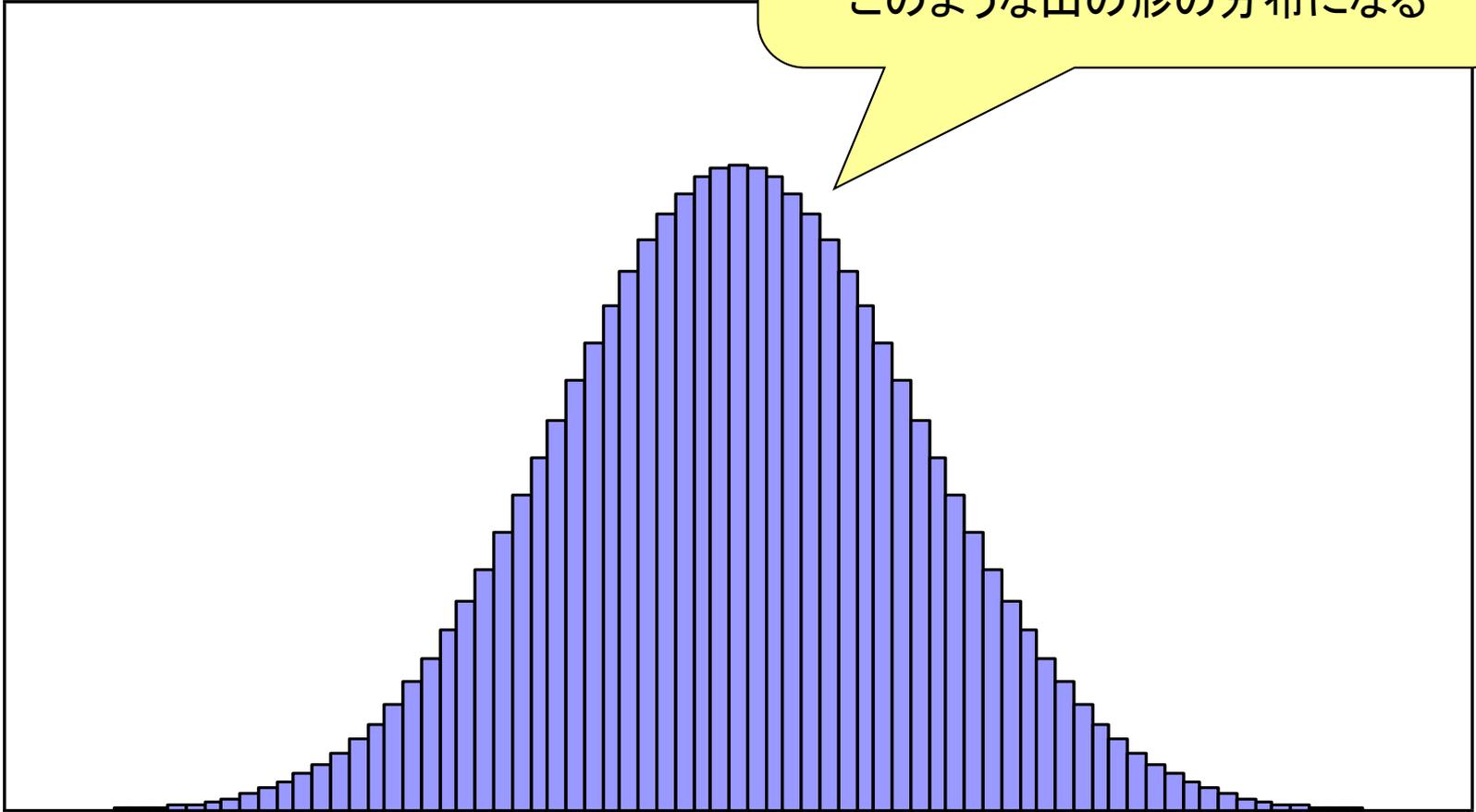
頻度

平均値付近の生徒が多く、
平均から離れるほど、
人数は少なくなる

■ 頻度



標本(サンプル)の数を増やし、
階級(それぞれの区間)の幅(差)を
限りなく小さくしていくと、
このような山の形の分布になる



標準偏差が s の正規分布

