

# 定量的なデータの分析

情報 I 第19回・第20回授業

04情報通信ネットワークとデータベース

対応データ 22exp19.xls

## 今日のテーマ

- データの種類を知る
- 数値で示す
- 数値の「意味」を知る
- わかりやすく示す

# 【復習】「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(ありそうだ)

※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(リコメンド)
- 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

# データの収集と整理 (p.192)

- オープンデータ
  - 国の行政機関や自治体、研究・教育機関が持つデータ
  - インターネットで広く公開され、容易に入手できる
- 一般的には分析する前に、必要なデータに整理する
  - 必要のある部分のみに成形したり、

# 情報分析

☆データに対し、適切な分析方法を理解する  
「定量」と「定性」

- 数値化されたもの（定量的なデータ）
  - 集計してグラフ化
  - 統計処理
- 数値化されていないもの（定性的なデータ）【次回以降】
  - 関係性や因果関係、順序などを図解
  - 同じような内容ごとや程度にまとめて数値化
  - テキストマイニングなどで数値化、分析

# 【発展】データの種類

ここに  
注目！

データの種類	尺度	意味	単位	例
質的データ (定性的)	名義尺度	区別しかできない	ない	職業区分、電話番号
	順序尺度	大小比較ができる	ない	優良可の区分、震度
量的データ (定量的)	間隔尺度	差が意味を持つ	ある	気温、偏差値
	比率尺度	比が意味を持つ	ある	長さ、重さ

# データの集計方針(3分)

ワークシートにあるA組とB組それぞれのデータを分析したい。

- 見やすくまとめ、
- 何がわかるかを確かめたい。

どのような方針で行うかを具体的に記入しよう。

# データのグラフ化：見やすくまとめる

- 時系列で示す
  - 折れ線グラフ、有向グラフ など
- 度数分布で表す
  - ヒストグラム、箱ひげ図 など
- 割合で表す
  - 円グラフ、帯グラフ など
- 2つのデータの関連性を表す
  - 散布図 など

# データの統計的分析：数値で示す

- 定量的なデータを、数値（統計量とも言う）を用いて分かりやすく示す
  - 代表的や特徴的な数値を用いる
    - 平均値、中央値、最頻値、最大値、最小値 など
  - 散らばり具合を示す
    - 分散、標準偏差、範囲、四分位偏差 など
  - 偏り具合を示す
    - 尖度（せんど）、歪度（わいど） など
  - 2変数の関係を分かりやすく示す
    - 相関、相関係数、回帰直線
  - 違いを見極める
    - 統計的仮説検定の考え方

# 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

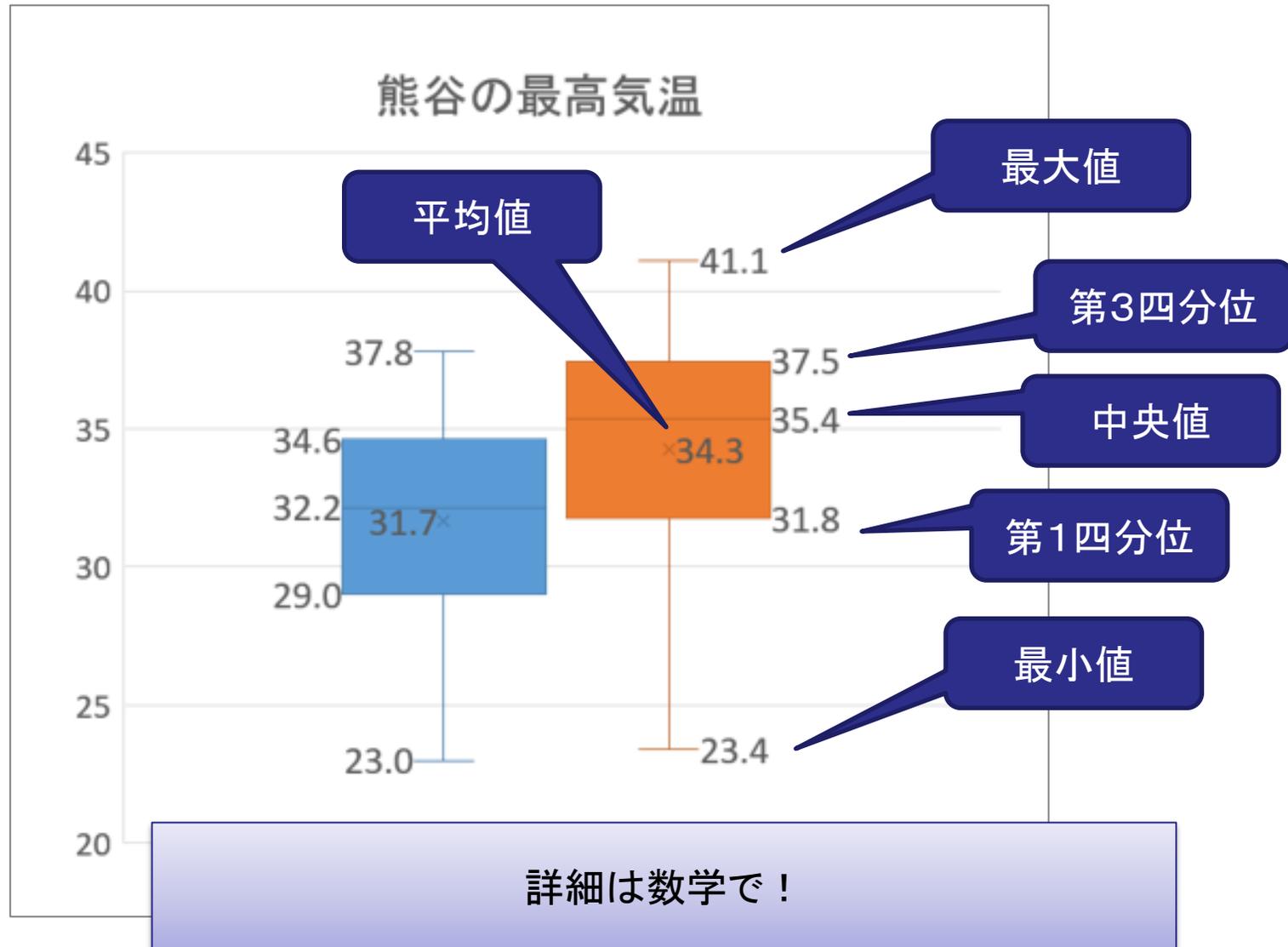
合計

標本数

# データのありかを意識 (p.193)

- 「気象庁 最高気温 データ」で検索
- DATA GO JP
- e-Stat

# 箱ひげ図(データを4分割する)



# 演習1

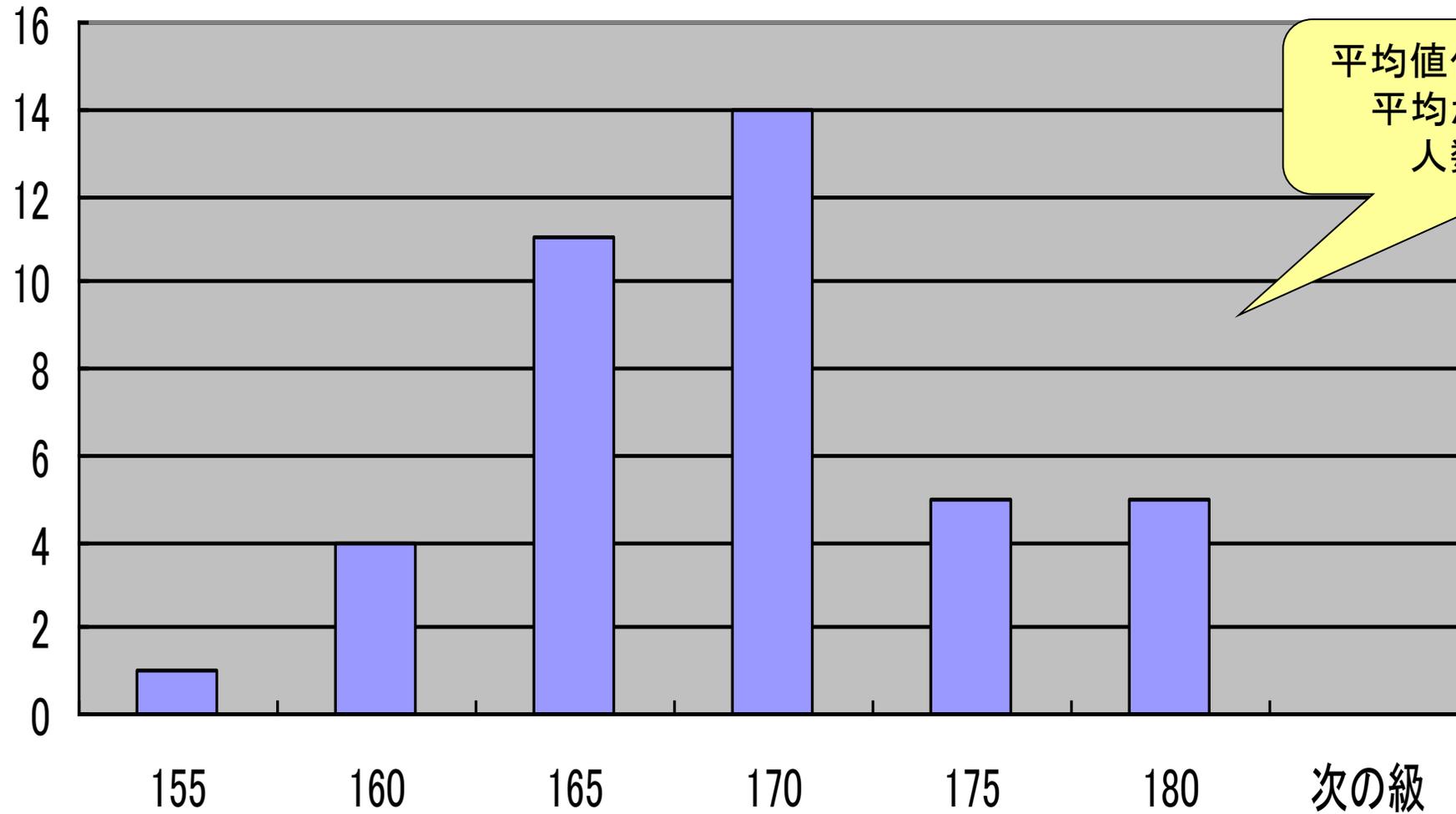
- 2017年と2018年の熊谷の最高気温について、箱ひげ図を描いてみよう。

# データの統計的な分析 ～ 仮説検定と相関 ～

情報 I 第20回授業 つづき  
04情報通信ネットワークとデータベース  
対応データ 22exp19.xls

# 1 「仮説検定」について

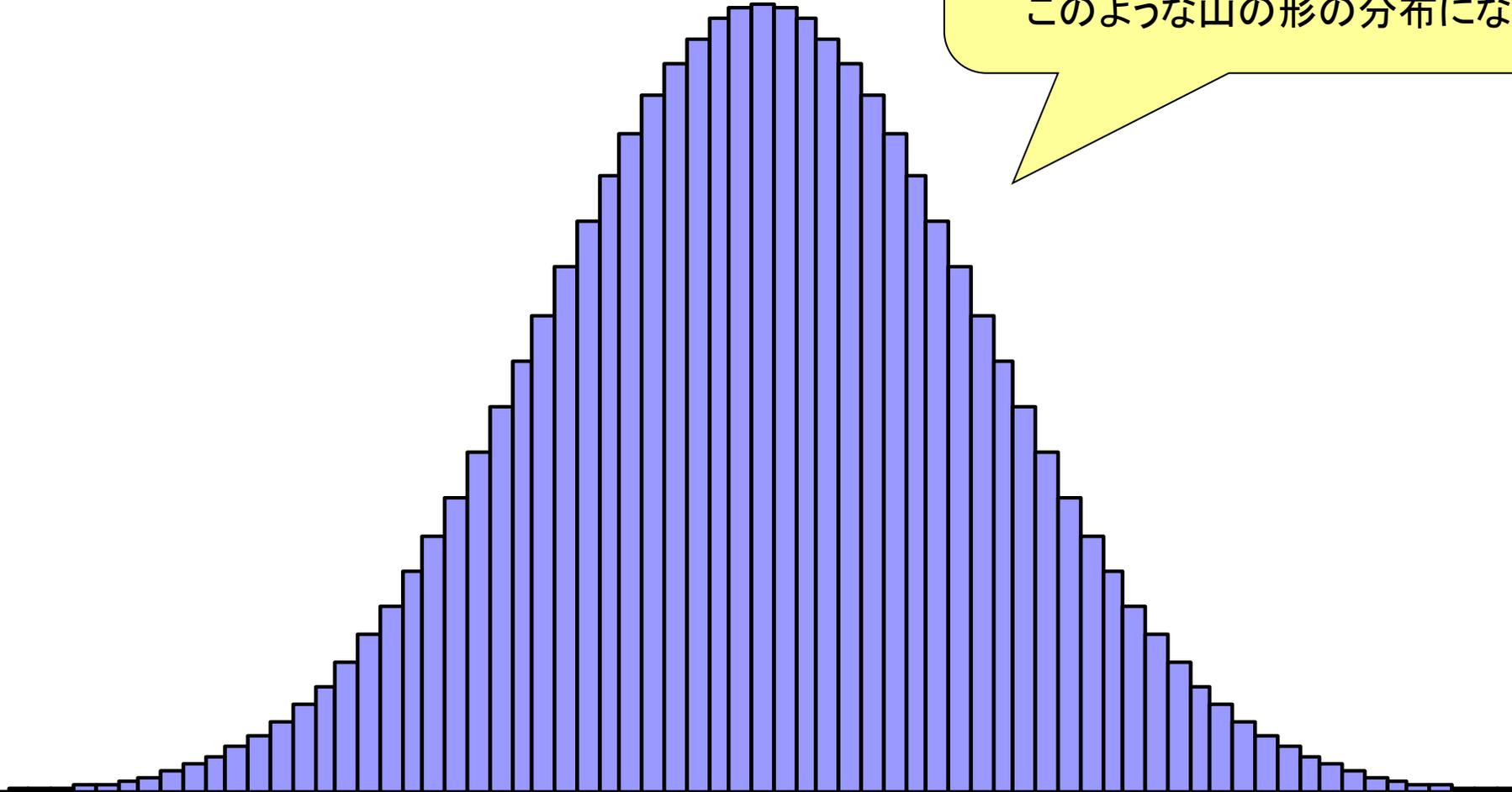
# 頻度



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

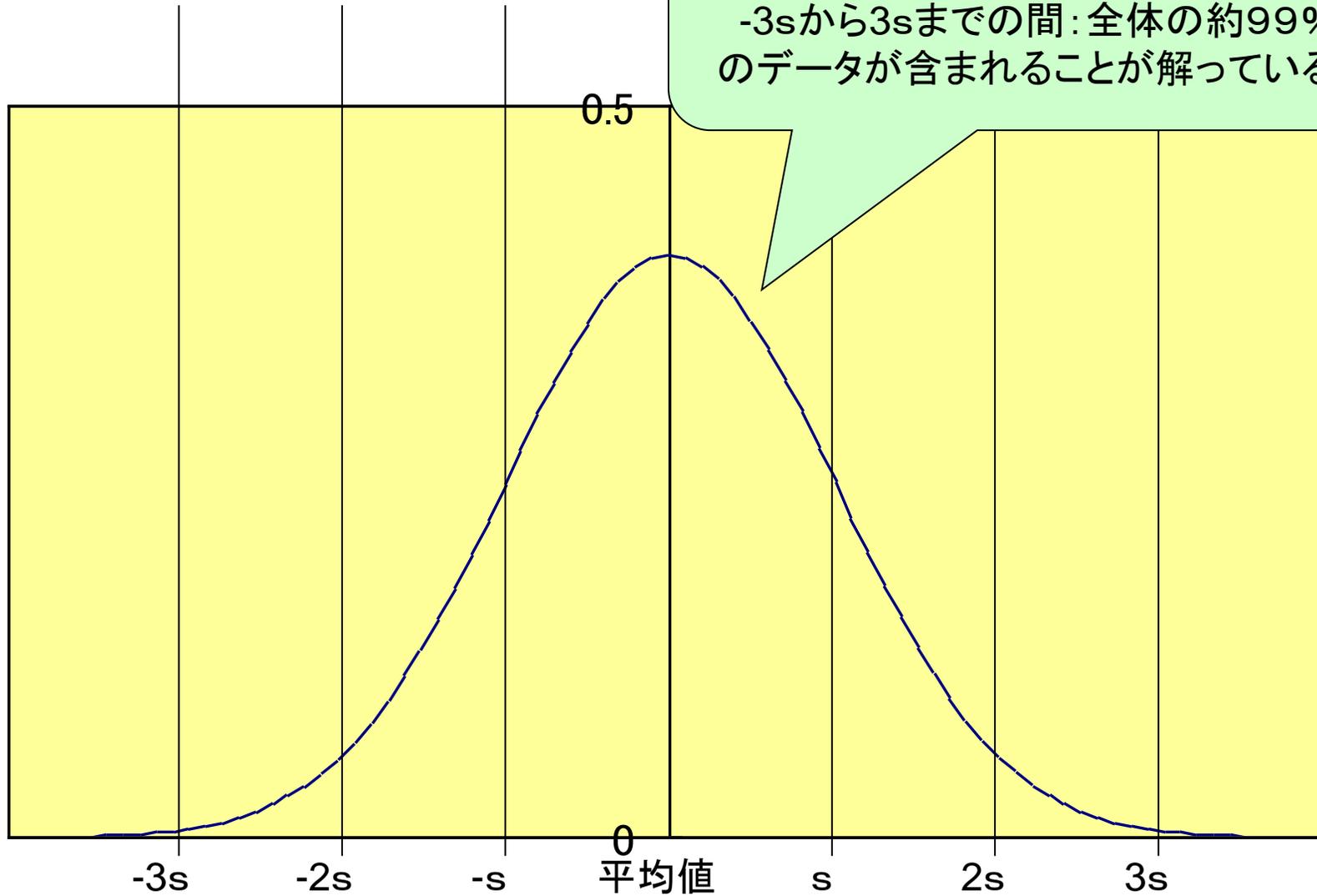
標本(サンプル)の数を増やし、  
階級(それぞれの区間)の幅(差)を  
限りなく小さくしていくと、  
このような山の形の分布になる



# 正規分布

- 平均値の周辺が最も度数が多く、
- 平均値から離れるに従って、度数が少なくなっていくような分布（山のような形）
- 世の中の多くの分布が、ほぼ正規分布のような形になると見なすことができる。
  - 模試の結果、体重、身長、・・・
- 正規分布で「分かっていること」を活用できる

# 標準偏差が $s$ の正規分布



# 統計的仮説検定

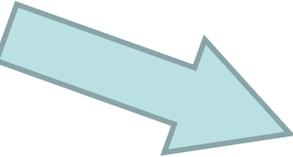
ある出来事に着目した場合、  
その起こった出来事の確率が、  
一定[有意水準といいます]  
(0.05 あるいは 0.01)以下の場合に、  
「違いがあるのでは」とする考え方

5回連続で表が出たコイン！

インチキだ！

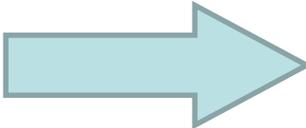
H<sub>1</sub>

偶然だ！



本物なら、そんなに連続しない！

そんなに言うなら「証明」してよ



じゃあ、そのことが起こる確率で決着つけよう！

こんなにめったにおきないことが起こるなんて！

以下！

インチキ！

H<sub>1</sub>

起きても別に不思議じゃない確率でしょ？

大きい！

偶然！

H<sub>0</sub>

例) 5回連続で表が出たコイン。  
「表が出やすい」といえるか？

$H_0$ : コイン表裏の出やすさは同様

→ 「守り」(=帰無仮説)

$H_1$ : コインは表が出やすい

→ 「示したいこと」(=対立仮説)

確率を計算し、「基準値」以下ならば、 $H_0$ は棄却( $H_1$ を採用)

→ 「示したいこと」が示される(「有意差がある」という)

# 「検定」と「過誤」

$$(0.5)^5 = 0.03125$$

有意水準 5%

$H_0$ を棄却 → このコインは表が出やすい

※もちろん「本当は表裏同じ」かもしれない！ ← 第1種の過誤

有意水準 1%

$H_1$ を棄却 → このコインは表が出やすい

とはい切り切れない

※もちろん「本当は表が出やすい」かもしれない！ ← 第2種の過誤

<注意！>この場合、決して「表裏が同じ」と言い切れるわけではない！

## <練習2>

1の目が3回連続して出たサイコロがある。  
このサイコロは1の目が出やすいと言えるか。  
有意水準1%で検定せよ。

$H_0$ :このサイコロの目の出やすさは同様

$H_1$ :このサイコロは1の目が出やすい

$$1の目が3回連続 \cdots (1/6) \times 3 \doteq 0.00463 \\ < 0.01$$

よって、 $H_0$ を棄却 → このサイコロは、有意水準1%で  
1の目が出やすい！

(発展・参考)

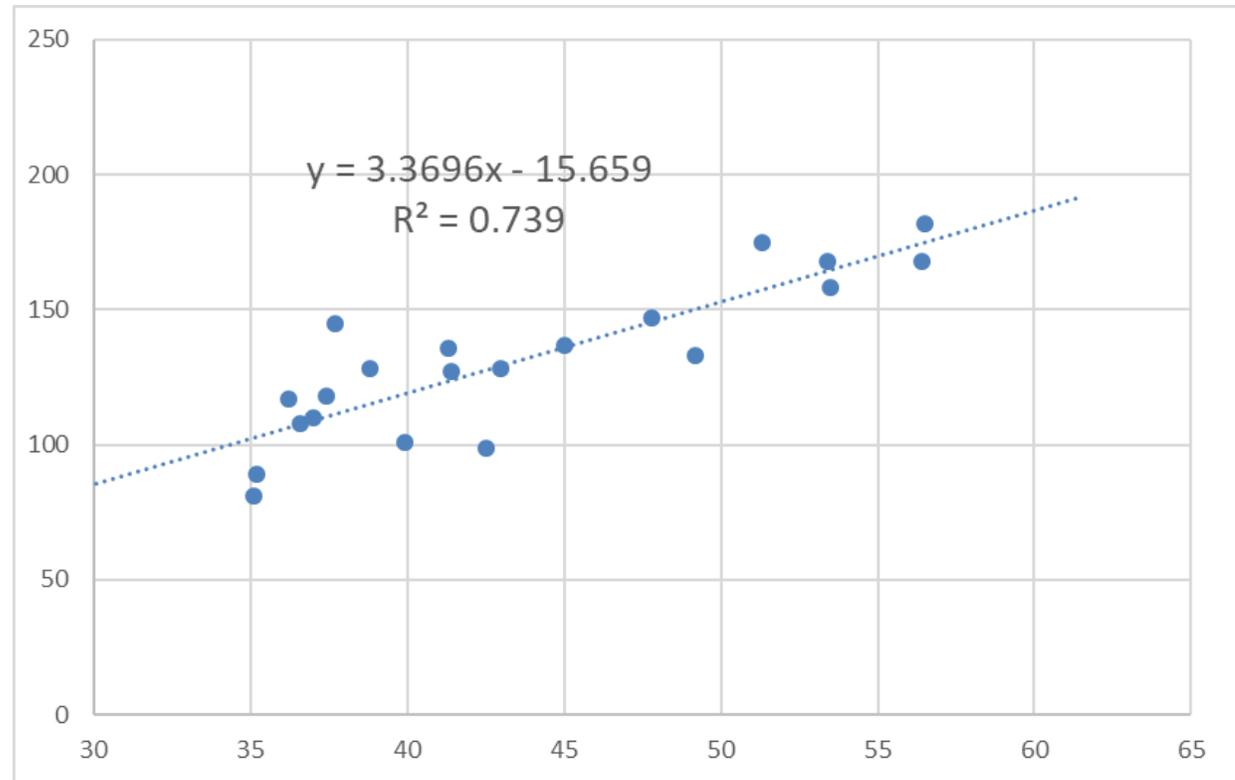
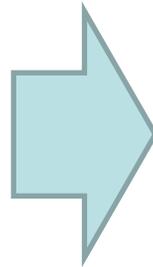
## 既に知られている分布を活用する

- 分散に「違い」があるかを確認める
  - 「F分布」の活用 → F検定
- 平均に「違い」があるかを確認める
  - 「正規分布」の活用 → Z検定 ( $n \geq 30$ )
  - 「t分布」の活用 → t検定 ( $n < 30$ )
- クロス集計表に「違い」があるかを確認める
  - 「 $\chi^2$ (カイニ乗)分布の活用」 →  $\chi^2$ 検定
  - これらの「表」から得られた確率に基づき、有意差があるかを判断している。

## 2 相関について

# 相関を調べる

名前	握力	背筋力
A	38.8	128
B	35.2	89
C	36.2	117
D	56.5	182
E	37.7	145
F	41.3	136
G	37.4	118
H	53.4	168
I	41.4	127
J	35.1	81
K	47.8	147
L	53.5	158
M	36.6	108
N	43	128
O	49.2	133
P	45	137
Q	39.9	101
R	51.3	175
S	56.4	168
T	37	110
U	42.5	99



※この直線を「回帰直線」という

## < 練習3 >

- ワークシートにある「握力と背筋力」のデータから、
  - 散布図を作成する
  - 回帰直線を表示させる
  - 回帰直線の方程式を表示させる

# 相関行列

	身長	体重	座高	握力	上体起こし	長座体前屈	反復横跳び	シャトルラン	50m走	立ち幅跳び	ハンドボール投げ
身長	1.000										
体重	0.382	1.000									
座高	0.756	0.497	1.000								
握力	0.250	0.559	0.315	1.000							
上体起こし	0.066	0.092	-0.029	0.360	1.000						
長座体前屈	0.257	0.235	0.235	0.317	0.309	1.000					
反復横跳び	0.149	0.110	0.093	0.386	0.457	0.477	1.000				
シャトルラン	0.142	-0.090	0.029	0.175	0.341	0.277	0.372	1.000			
50m走	-0.211	-0.098	-0.215	-0.454	-0.329	-0.294	-0.544	-0.553	1.000		
立ち幅跳び	0.359	0.063	0.273	0.412	0.256	0.361	0.483	0.341	-0.674	1.000	
ハンドボール投げ	0.292	0.315	0.278	0.470	0.457	0.408	0.519	0.400	-0.490	0.419	1.000

## <練習4>

- 「科学の道具箱」を開き、「高等学校体力測定データ」を確認する
- すでに整形されたデータを元に、相関行列を作ってみる
- 相関行列から相関の高い2系列を選び、散布図と回帰直線を作成する