九州大学 工学部地球環境工学科
船舶海洋システム工学コース

海事統計学 （担当：木村）

1）データの整理と表現

火曜1・2限（8：40～12：00）
場所： 船2講義室

授業の資料等は

http://sysplan.nams.kyushu-u.ac.jp/gen/index.html

海事統計学

授業内容： 海事における不確実性問題の表現と扱い方について述べる
（海事を具体例に取り上げる確率・統計学）

- データの整理と表現 （統計処理）
- 確率
- 確率変数と確率分布
- 標本分布
- 推定
- 仮説検定

授業の資料等は

http://sysplan.nams.kyushu-u.ac.jp/gen/index.html

成績評価方法：

- テーマの区切り毎に**レポート**
- ほぼ毎回の**演習** （1回2点）

} 合計40点満点

期末試験が60点未満の場合、演習やレポートの点を足して60点を超えたら、60点として評価

- **期末試験** → 100点満点 → 期末試験の得点が成績

期末試験の受験資格：**講義への出席2／3以上**　5回休んだらアウト

統計学の必要性

例１）　船の位置の特定

観測によって得られる位置は、実は真の位置の**推定**

誤差を見越した安全な航路をとるには？

適切なマージンは？

例２）　クレーン等に用いるワイヤロープ

長時間使用すると疲労などにより切れるので、

切れる前に交換したい

何時間で交換すべきか？

●マージンが少なすぎれば危険・大きすぎても資源の無駄遣い

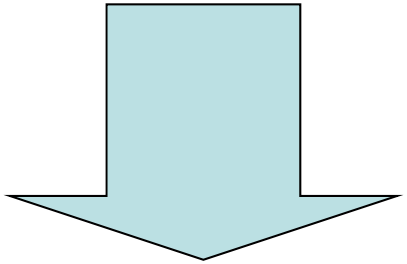
●実験によりデータを収集するが、回数が限られる

このような「不確実な現象」をどのように扱うか？

「統計学」は、これらに合理的な答えを与える

例）ワイヤの細かいキズ・材質の不均一さ
観測装置の各部が持つ誤差
各個人がとる行動など

正確な観測が不可能な要因の影響を受ける現象を予測・対処するには？



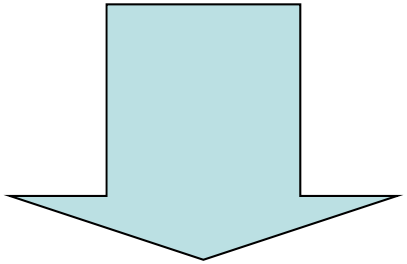
1つ1つの要因や現象は、予測や観測が不可能でも、
同じ要因が多数あるいは様々な要因が多数重なると、
非常にきれいな性質が表れる → 「**統計**」を利用して予測

計測のたびに小数点以下の値が変動する

【例題】 誤差 10 [mm] 程度の精度しかない測定装置を用いて、
誤差 1 [mm] 程度以下の精度で位置や長さを計測するには？

例）ワイヤの細かいキズ・材質の不均一さ
観測装置の各部が持つ誤差
各個人がとる行動など

正確な観測が不可能な要因の影響を受ける現象を予測・対処するには？



1つ1つの要因や現象は、予測や観測が不可能でも、
同じ要因が多数あるいは様々な要因が多数重なると、
非常にきれいな性質が表れる → 「**統計**」を利用して予測

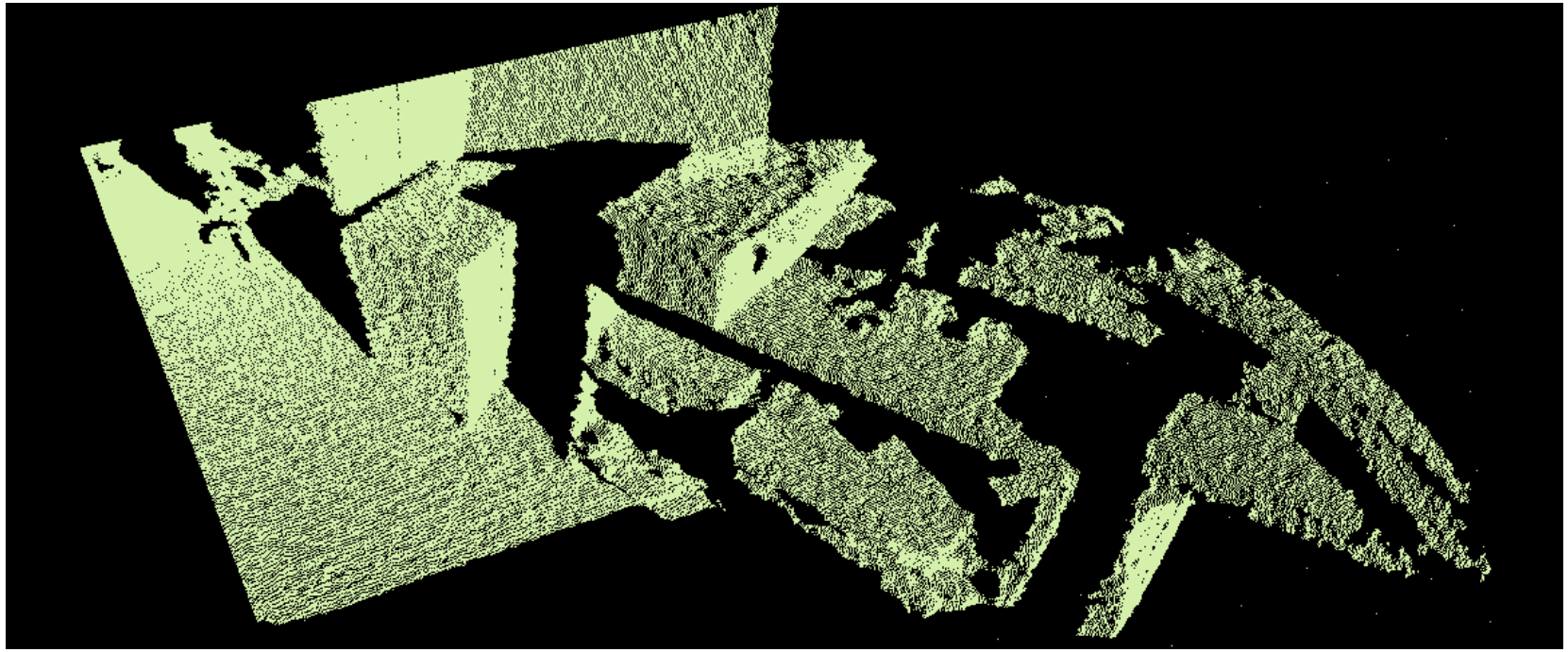
計測のたびに
小数点以下の
値が変動する

【例題】 誤差 10 [mm] 程度の精度しかない測定装置を用いて、
誤差 1 [mm] 程度以下の精度で位置や長さを計測するには？

答え：100回以上測定を繰り返して平均をとる　　なぜ10回でなく100回？　今後の講義で明らかに

XBOXのゲームコントローラ「キネクト」

＝　安価な3Dレーザスキャナの一種

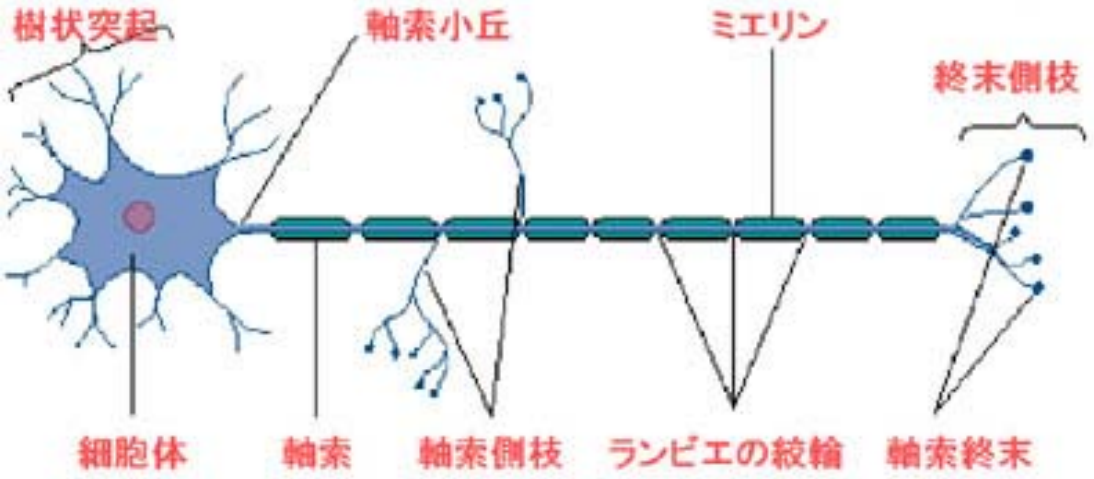


60度程度の視野角で奥行き 0.3～4[m]の空間内の物体の表面を
「深度マップ（濃淡画像）」として1秒間に20フレーム取得可能
ただし、**点1コあたり±10[mm]程度の誤差**

これこそが
「統計学」の力

人体の上半身を計測して**毎フレーム全点の平均値**をとり変動を
検出することで**鼓動・脈拍を非接触で計測可能**（0.2[mm]程度の精度）

神経細胞



ある値以上の電圧の
信号が入ると、
パルス状の信号を
発生して伝播

しかし…

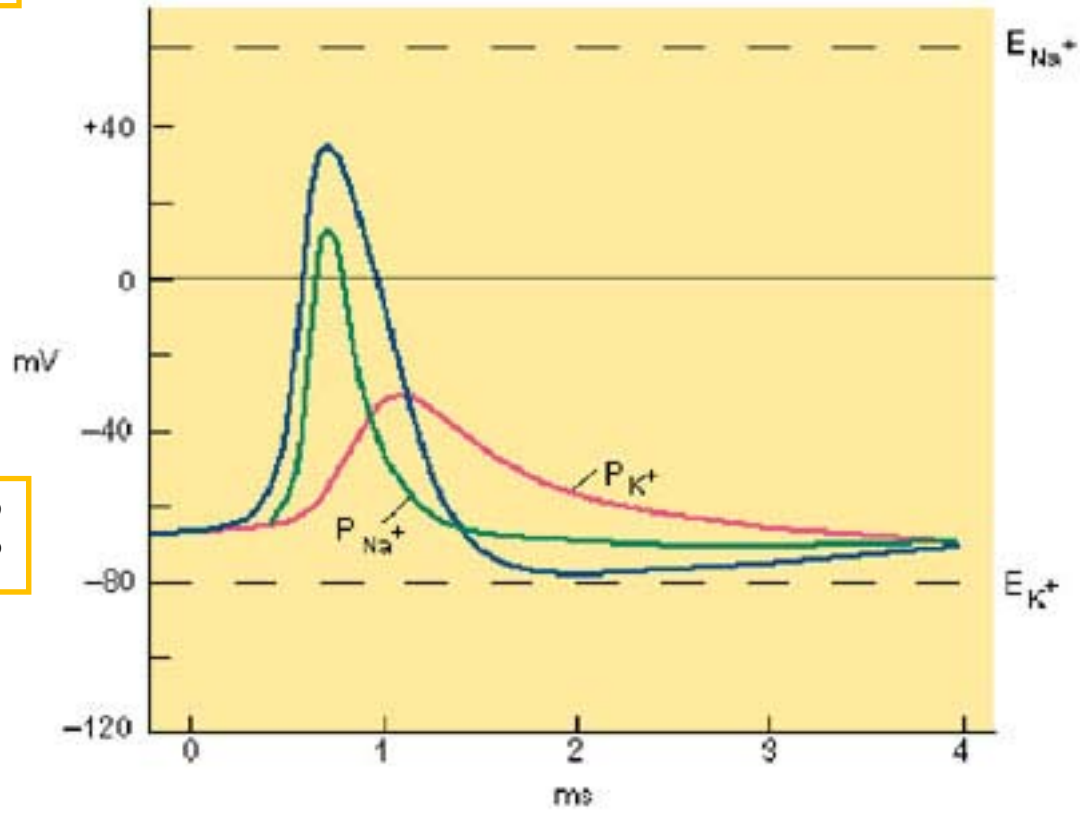
細胞の出力＝信号＋雑音

細胞自身が信号と同じ
大きさの雑音を発生

1つの神経細胞の
出力をみても、
ほとんどランダムなパルス
非常に粗悪な信号伝達素子

正規分布に従う
乱数を平均する
とゼロ

神経線維を束にして、
多数の出力を合計することで、
信号を正確に伝えている
（100本でS／N比が10倍）



英文における文字の出現確率

文字	確率		文字	確率		文字	確率
A	8.29%		J	0.21%		S	6.33%
B	1.43%		K	0.48%		T	9.27%
C	3.68%		L	3.68%		U	2.53%
D	4.29%		M	3.23%		V	1.03%
E	12.08%		N	7.16%		W	1.62%
F	2.20%		O	7.28%		X	0.20%
G	1.71%		P	2.93%		Y	1.57%
H	4.54%		Q	0.11%		Z	0.09%
I	7.16%		R	6.90%			

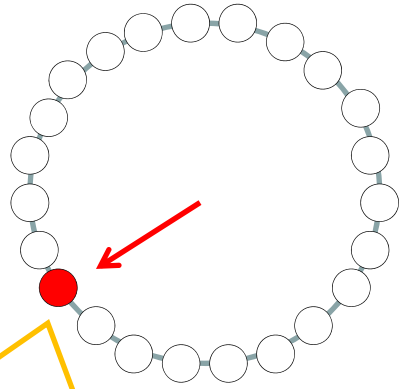
モールス符号などで文字を変換し、通信する場合、
出現確率の高い文字ほど**短い符号**で表す　→　通信時間が短くて済む

画像圧縮や音声・動画圧縮も原理的に同じ

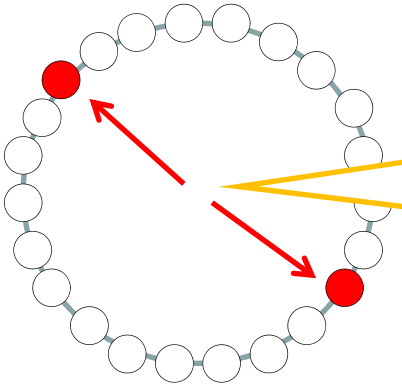
例）640x426　24bitカラー画像
無圧縮：798KB
JPEG圧縮：62.3KB

独立な事象の同時発生確率に基づく関連事象の検出

素粒子の崩壊検出



素粒子の検出はポアソン分布
（発生時間間隔が指数分布の乱数）
ほとんどの場合、同時に反応するのは1個

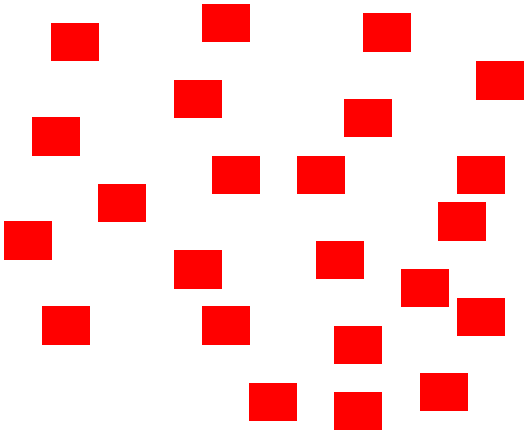


検出器2コ同時に反応
しかも互いが反対側

素粒子の崩壊

独立な事象が同時に発生したと
仮定すると、確率論的な頻度としては
ありえない＝2つの事象は関連

生体視覚情報処理



同じタイミングで点滅している部分を
結んだ全体を1つのまとまりとして知覚

一部しか見えていないのに
全体の形や動きが把握できる

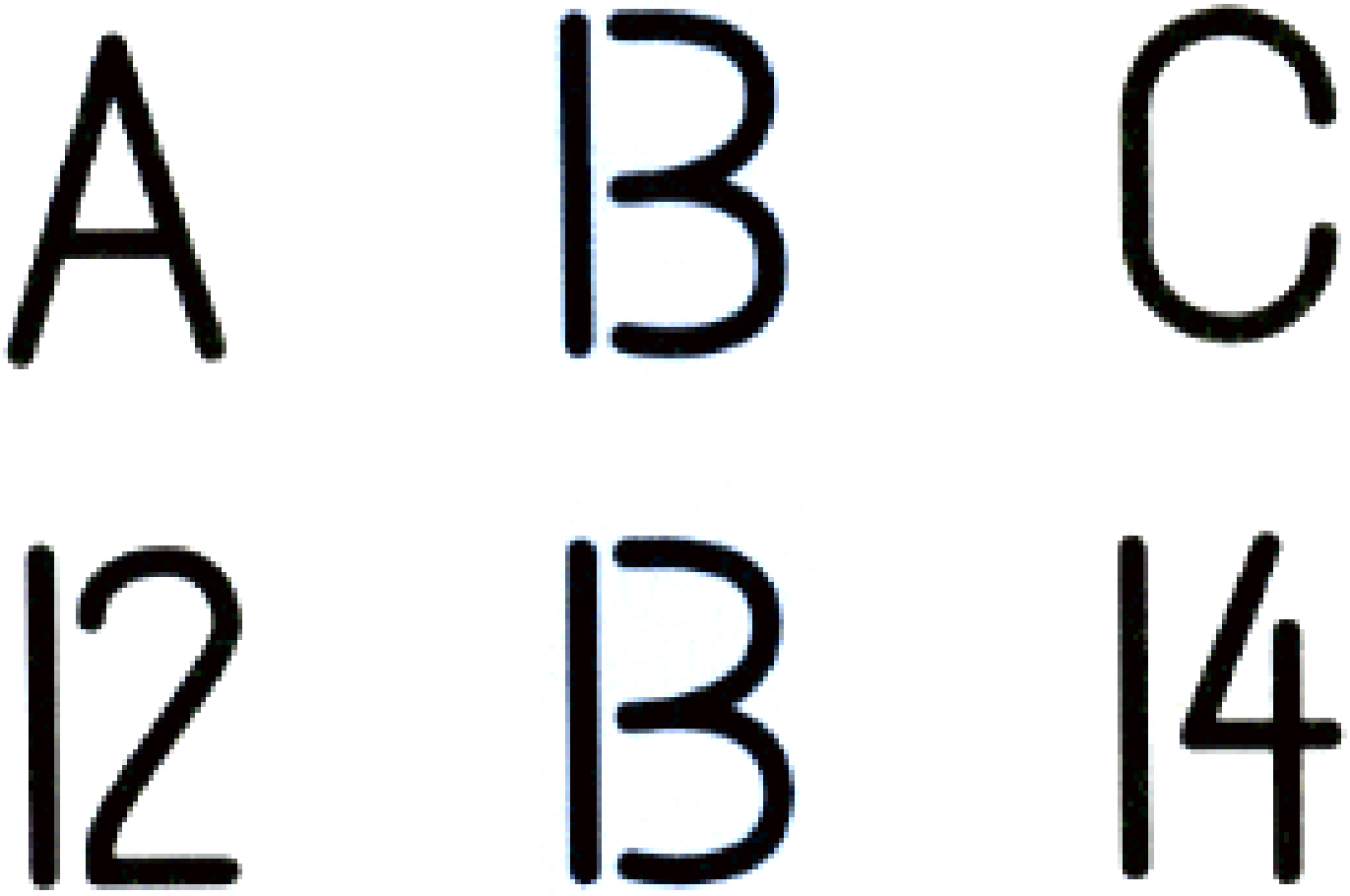
経験に基づく高度な情報処理

経験に基づく高度な情報処理

経験に基づく高度な情報処理

経験に基づく高度な情報処理

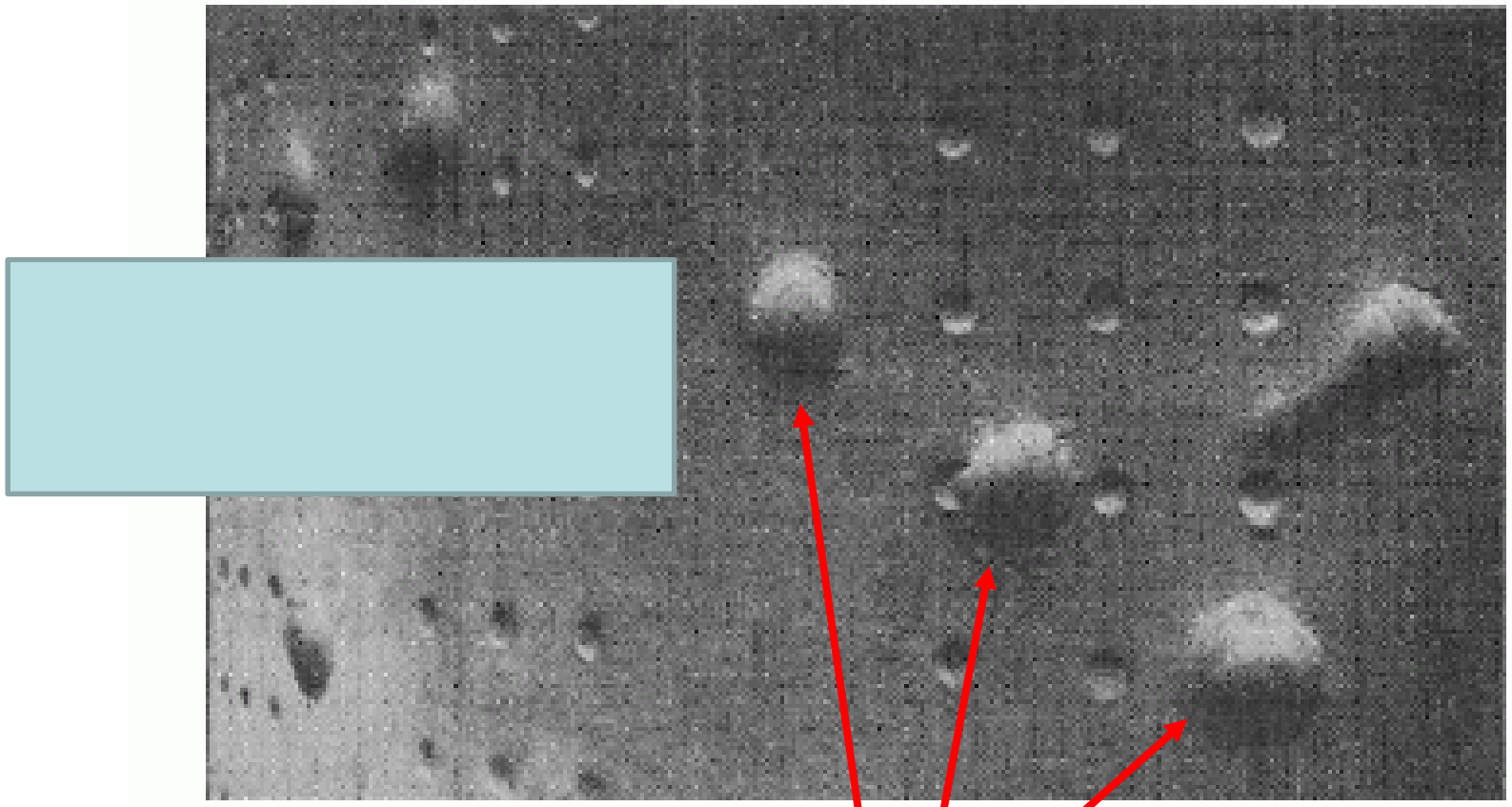
経験に基づく高度な情報処理



経験に基づく高度な情報処理

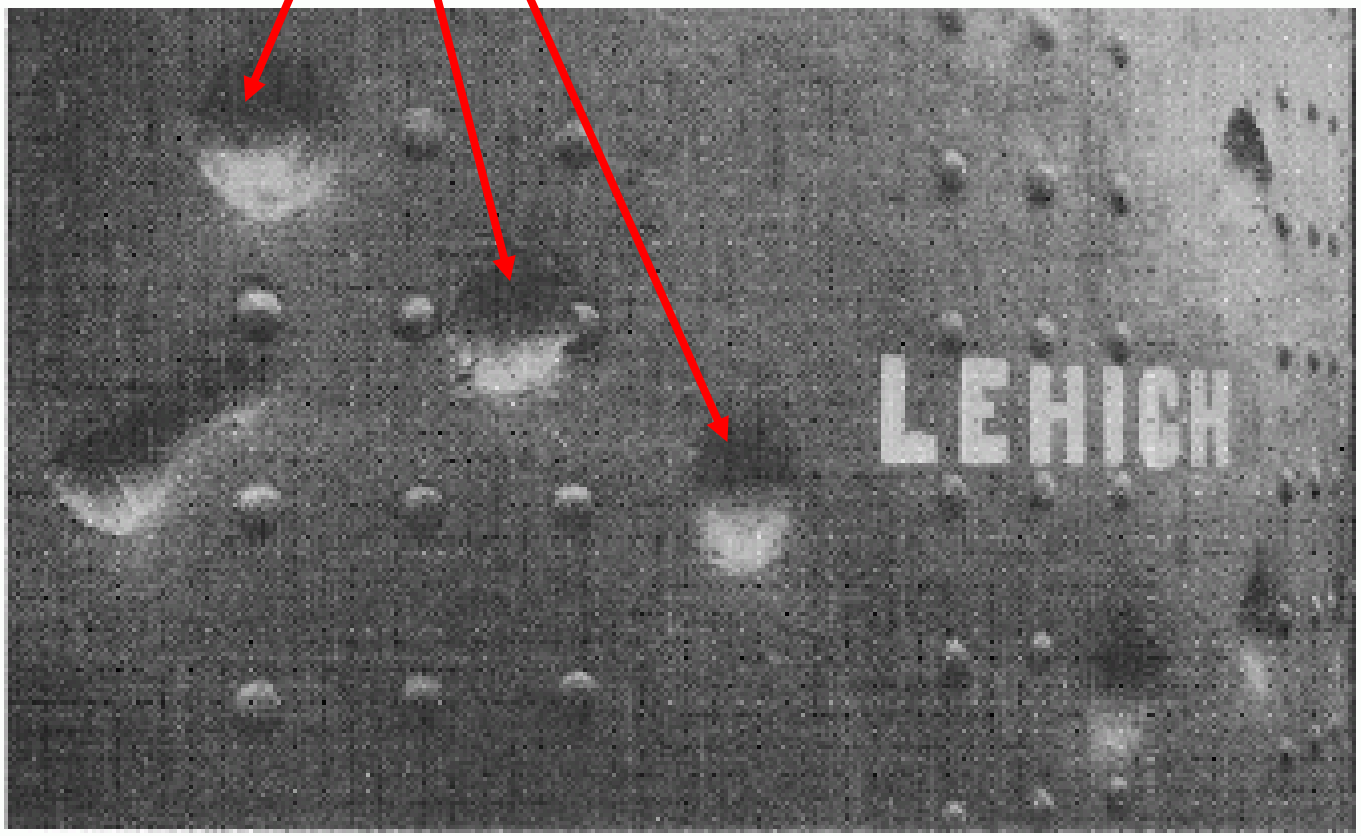
情報が足りない場合、前後のパターンから**統計的に可能性の高そう**な文字で補ってしまう

経験に基づく高度な情報処理



出っ張って見える？

凹んで見える？



単一の明暗画像から立体構造を復元する場合、
数学的には明らかに情報不足であり、いくつもの解が存在しうる

視覚情報処理は、「光は上からくる場合が多い」といった
統計的な性質を気づかないうちに利用し、足りない情報を補っている

データの整理と表現

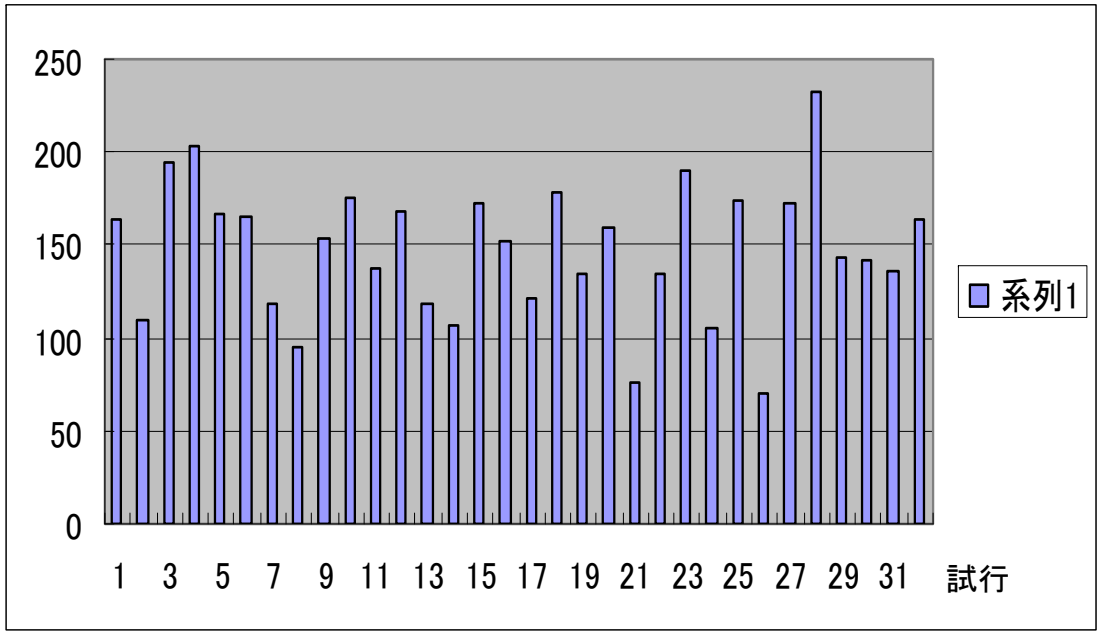
例）
負荷をかけたワイヤロープが
切れるまでの時間

単位：10時間
32試行

163.4	120.9
109.4	177.7
194.2	134.8
203.5	159.3
166.8	75.9
165.6	134.7
118.5	189.8
94.4	105.3
153.4	174.4
176.1	70.8
137.5	173.0
168.5	232.6
117.9	143.5
107.3	141.5
172.2	135.4
151.7	163.2

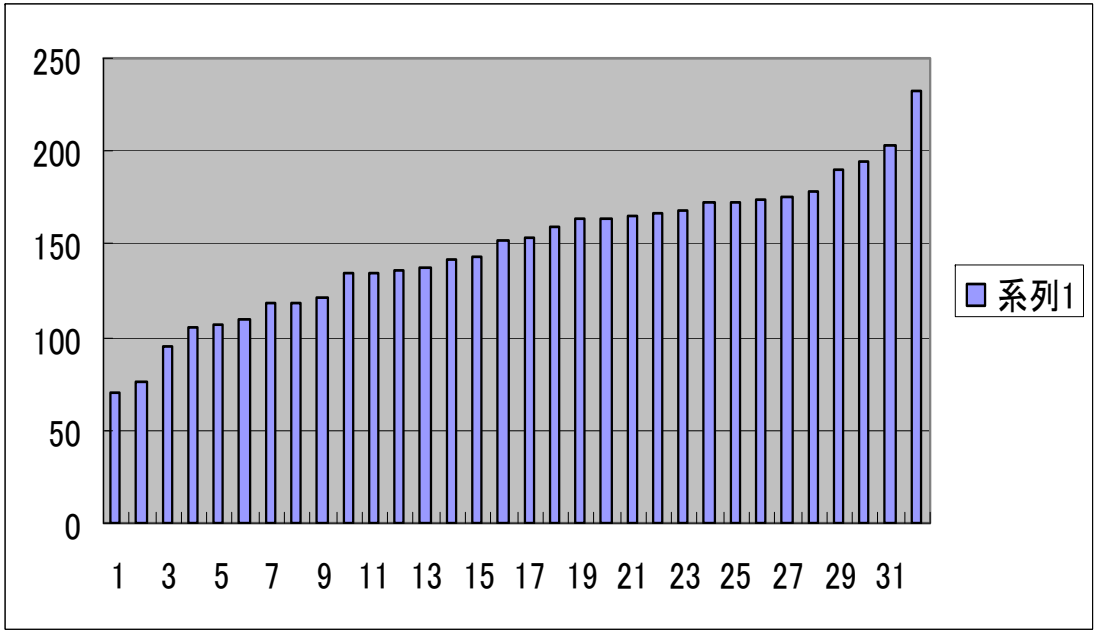
データをただ眺めるだけでなく、
見やすく整理したほうが
データ全体の情報を正確に
把握できる

切れるまでの時間

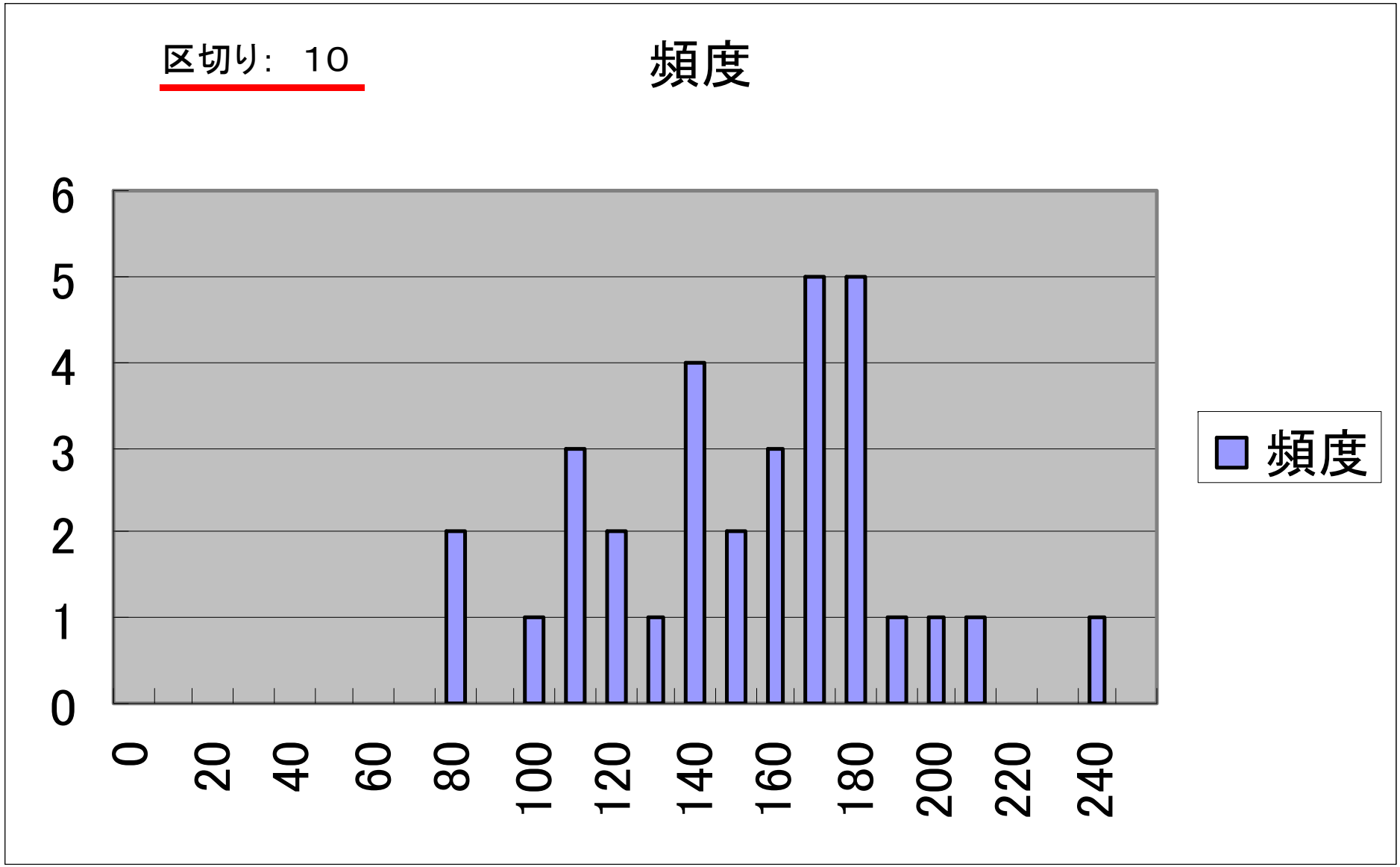


とりあえずグラフで表す

降順で並べてみる

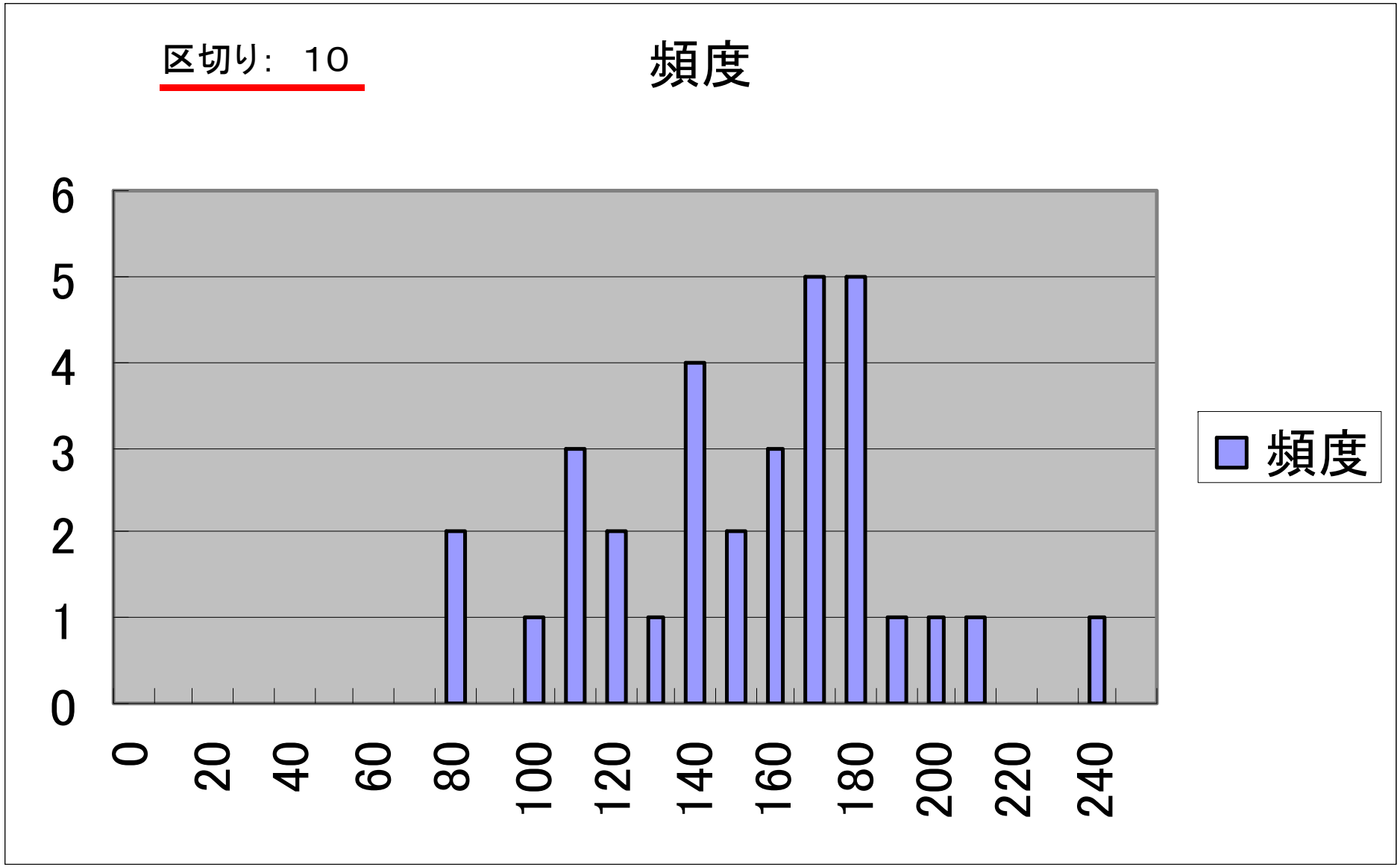


ある区切りでクラス分けするデータ整理法



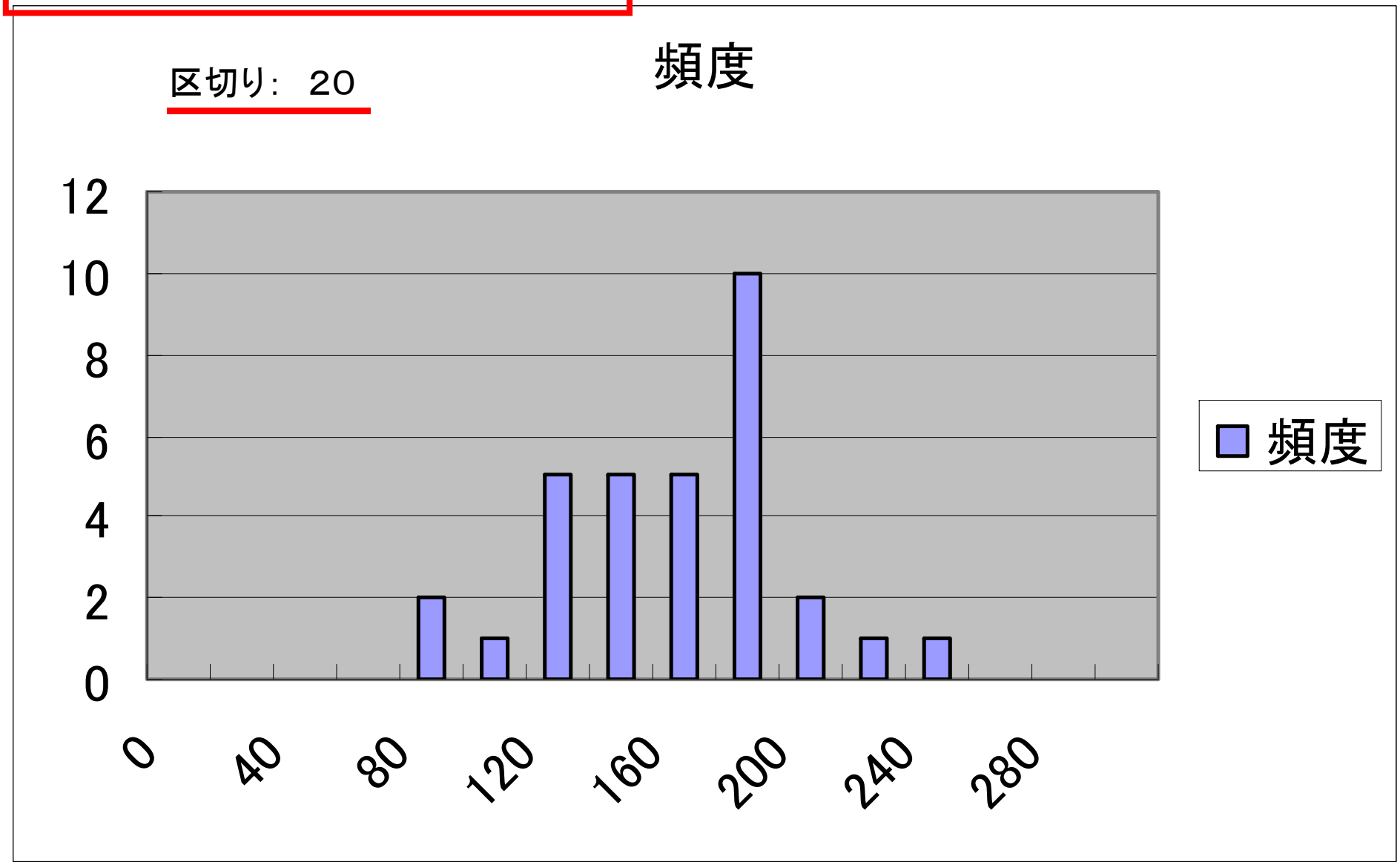
ヒストグラム（度数分布）

ある区切りでクラス分けするデータ整理法



ヒストグラム（度数分布）

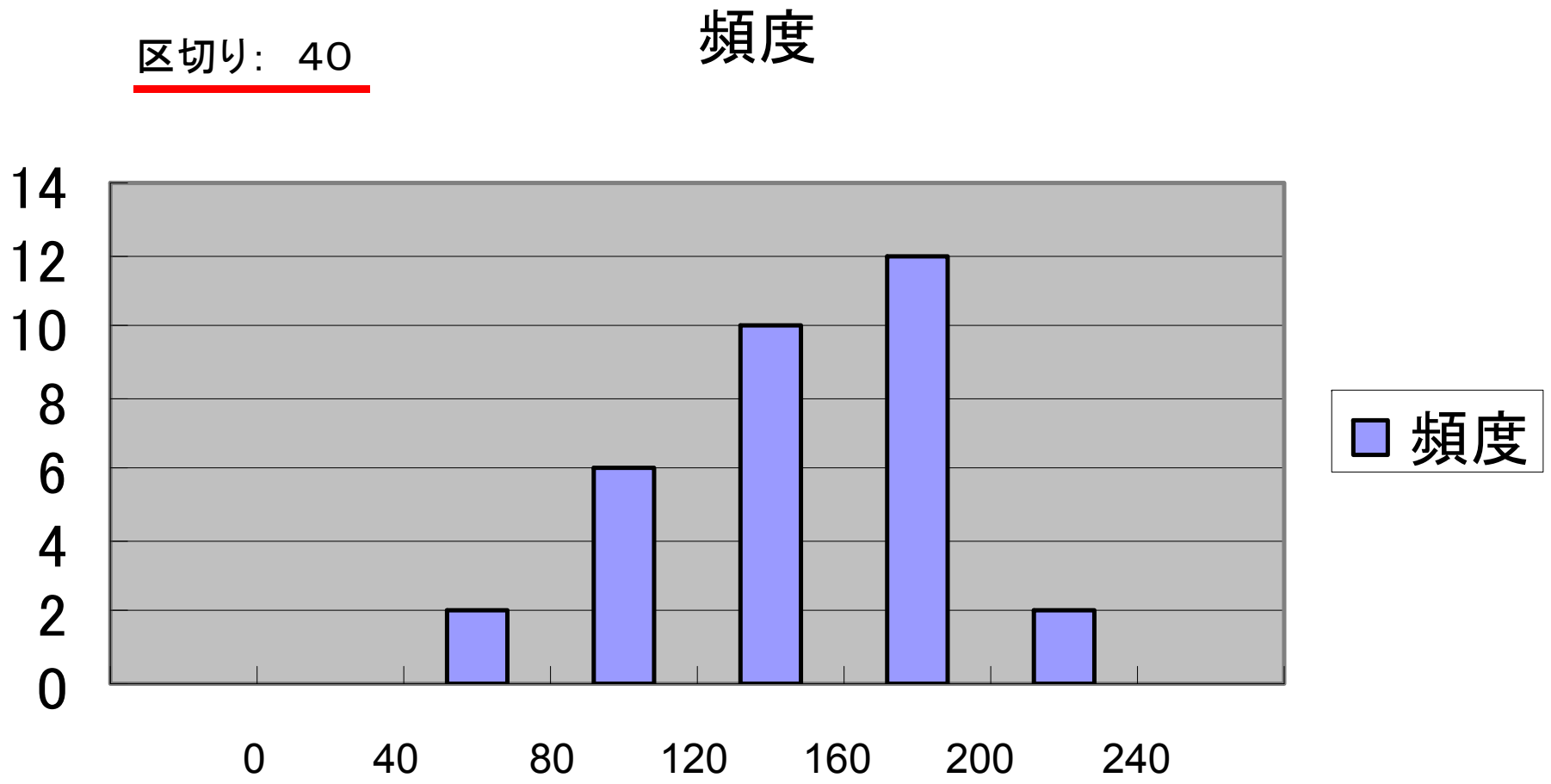
ある区切りでクラス分けするデータ整理法



ヒストグラム（度数分布）

ある区切りでクラス分けするデータ整理法

区切り： 40



処理のやり方によって、データの解釈が大きく異なってくる

切れるまでの時間

データの特性値

統計データ：観察の対象について得られた測定値の集合

統計分析：統計データに含まれる規則性を見出す

平均(mean) 分散(variance)
標準偏差(standard deviation)　} 特性値・統計量

標本（サンプル）

nコのデータ：　$x_1, x_2, x_3, \cdots, x_n$

平均：　$\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

分散：

標準偏差：　$\sigma = \sqrt{\sigma^2}$

} データの散らばり具合を表現する

データの特性値

統計データ：観察の対象について得られた測定値の集合

統計分析：統計データに含まれる規則性を見出す

平均(mean) 分散(variance)
標準偏差(standard deviation)　} 特性値・統計量

標本（サンプル）

nコのデータ：　$x_1, x_2, x_3, \cdots, x_n$

平均：　$\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

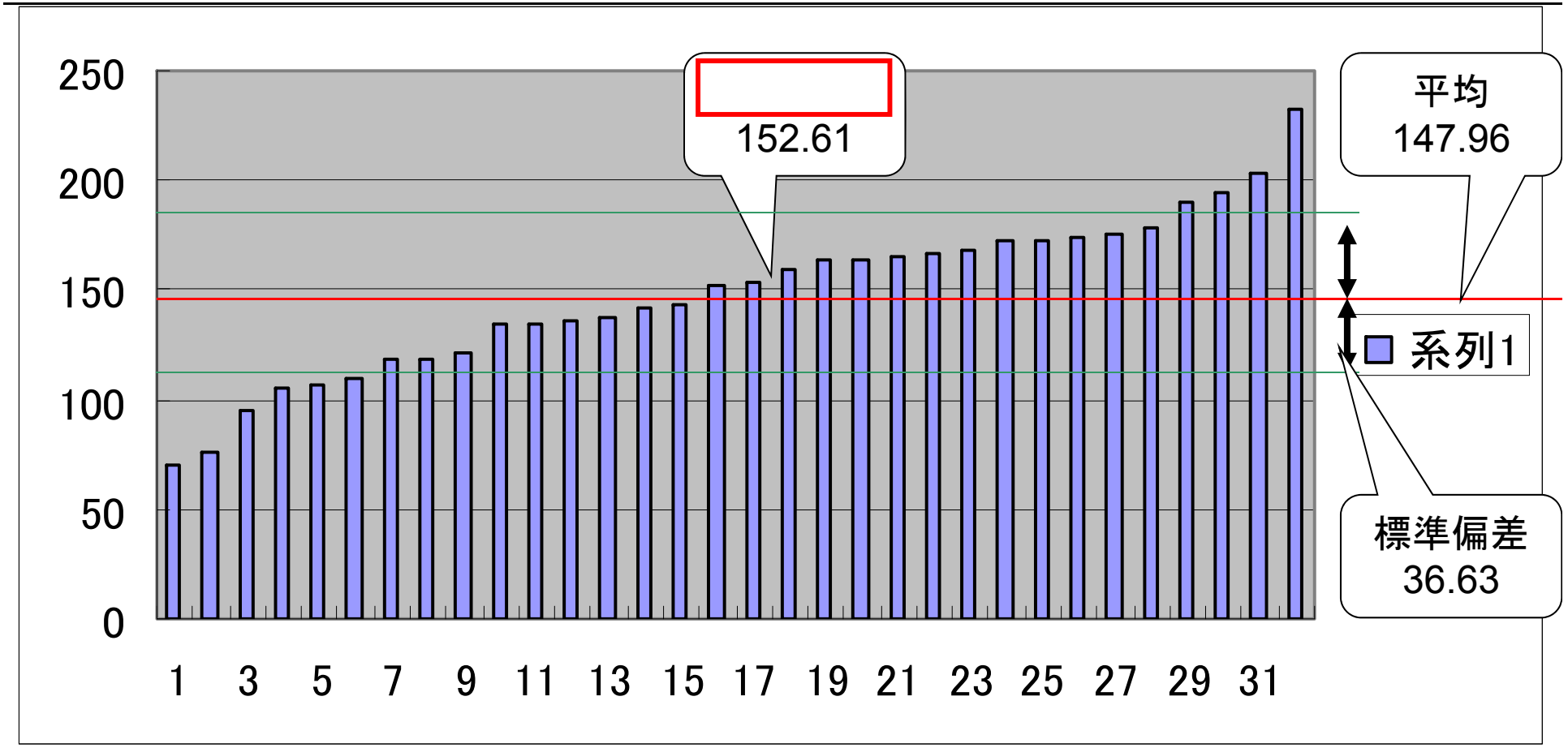
分散：　$\sigma^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

標準偏差：　$\sigma = \sqrt{\sigma^2}$

} データの散らばり具合を表現する

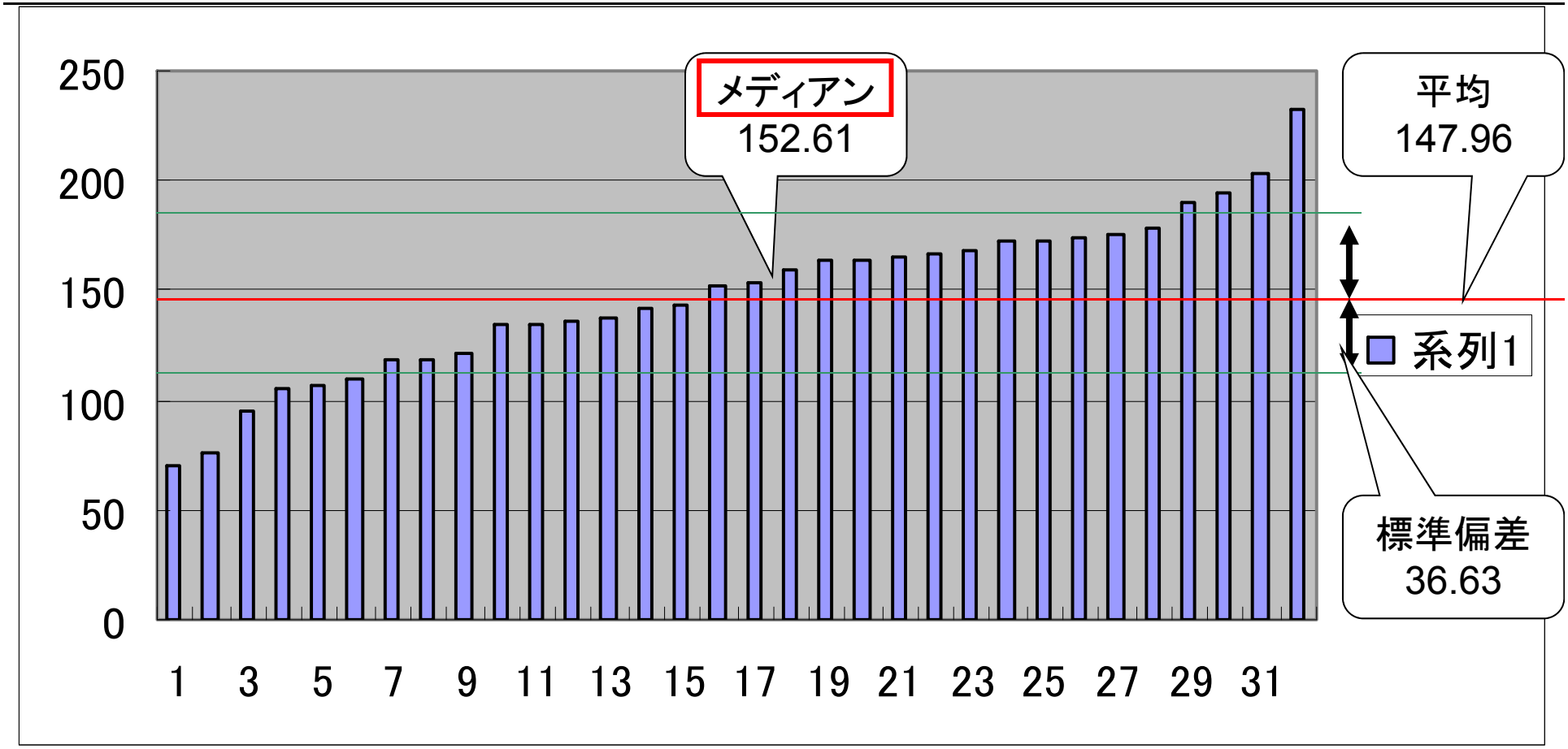
中位数

データを大きさ順に並べたとき、ちょうど中央に位置するデータの値
　データの数nが奇数のとき、(n+1)/2 番目の値、
　データの数nが偶数のとき、n/2 番目と (n/2)+1 番目の中間の値



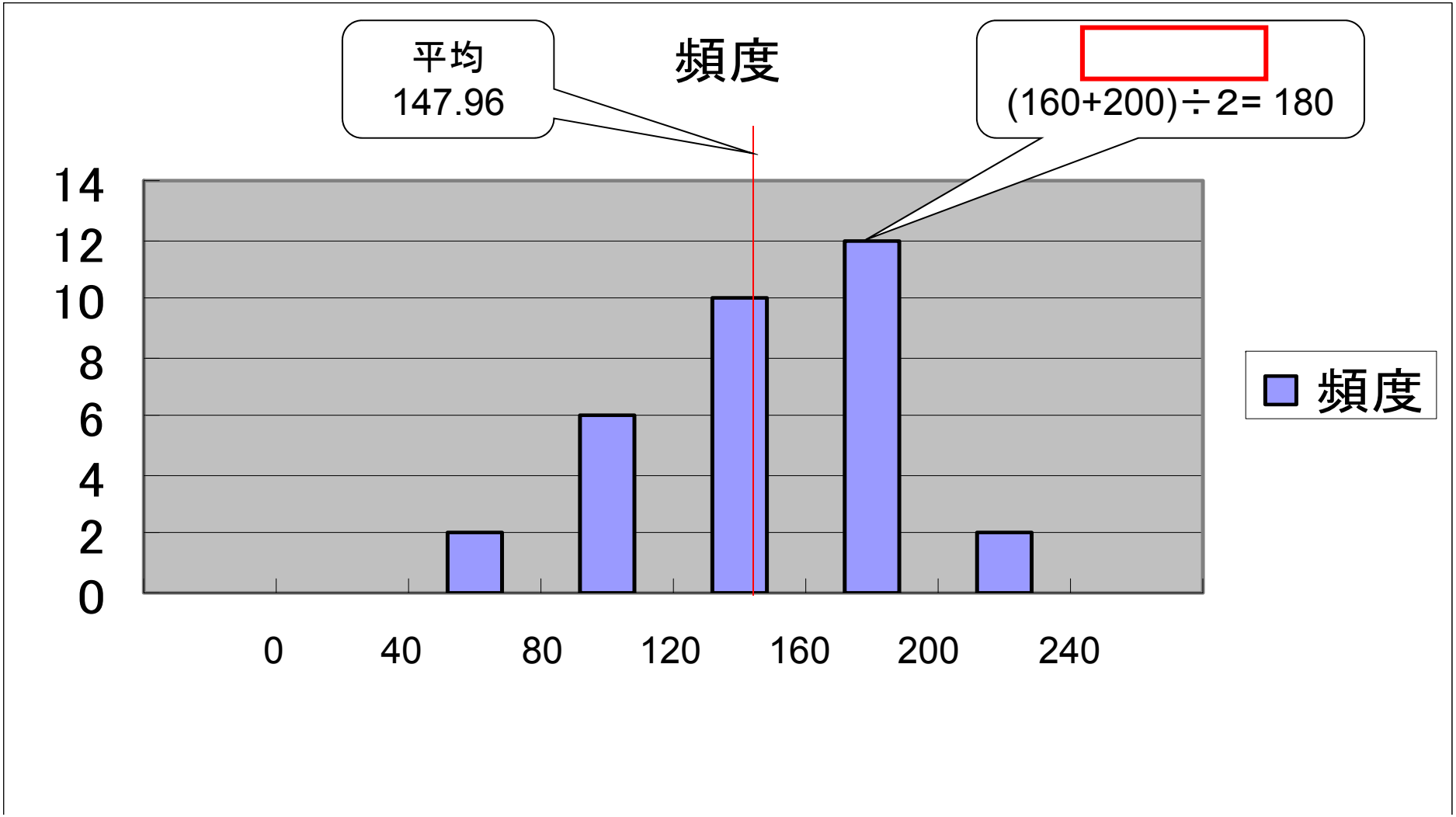
中位数（メディアン：median）

データを大きさ順に並べたとき、ちょうど中央に位置するデータの値
　データの数nが奇数のとき、(n+1)/2 番目の値、
　データの数nが偶数のとき、n/2 番目と (n/2)+1 番目の中間の値



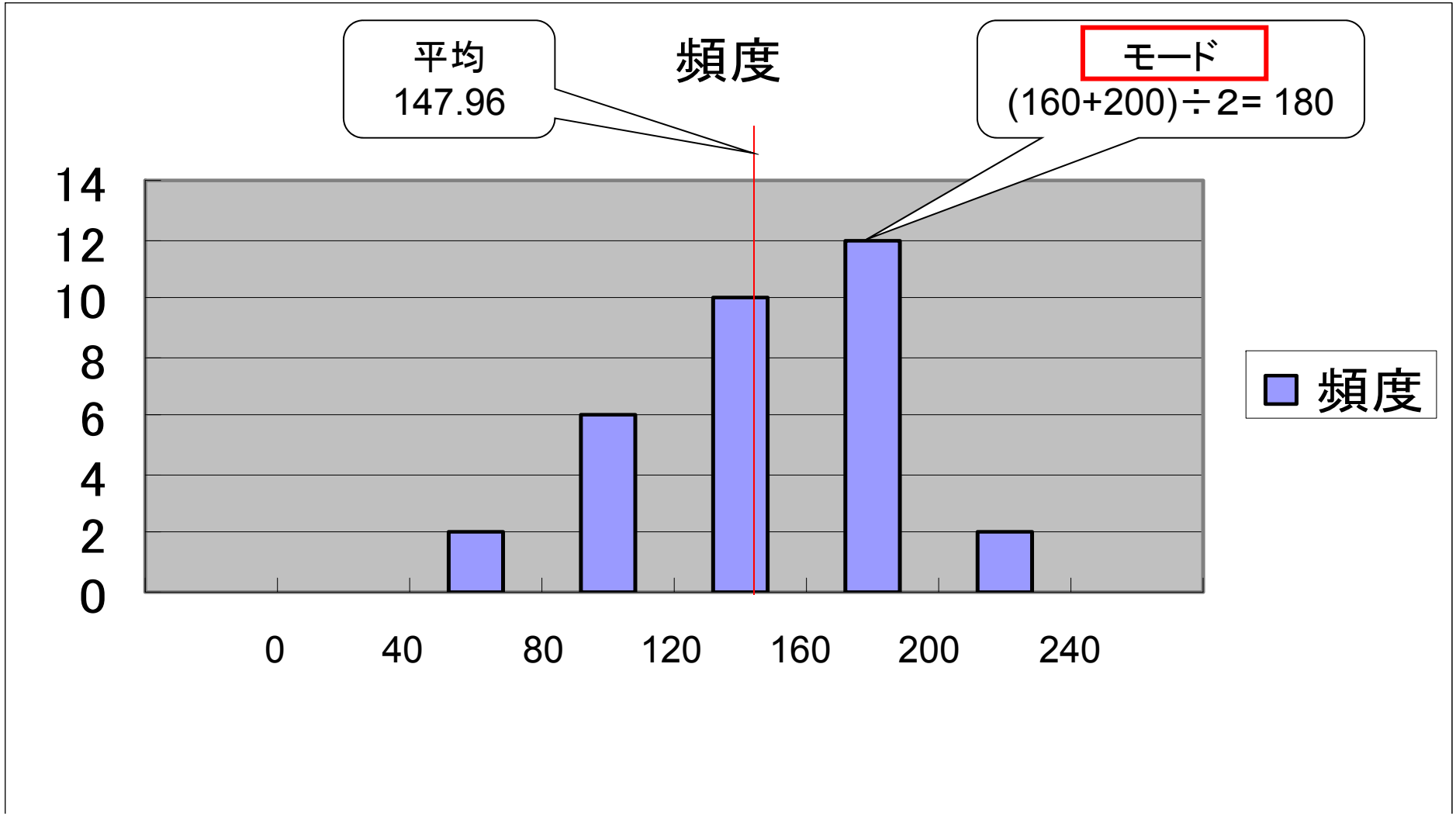
最頻値

度数分布表で度数が最も多い階級値
2つ以上存在することもある



最頻値（モード：mode）

度数分布表で度数が最も多い階級値
2つ以上存在することもある



平均 （mean）

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

数学的には、各要素からの2乗距離の差 [　　] を最小化する c を
平均 $\bar{x}$ としている。

中位数 （中央値，メディアン：median）

データを大きさ順に並べたとき、ちょうど中央に位置するデータの値
　データの数nが奇数のとき、(n+1)/2 番目の値、
　データの数nが偶数のとき、n/2 番目と (n/2)+1 番目の中間の値

数学的には、各要素からの絶対距離の和 [　　] を最小化する c を
中位数 としている。

この統計量は、解析が困難だが [　　] という特徴がある。

最頻値 [　　]

度数分布表で度数が最も多い階級値
2つ以上存在することもある

工学的に非常に有用

平均 （mean）

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

数学的には、各要素からの2乗距離の差 $\sum_{i=1}^{n}(x_i - c)^2$ を最小化する c を
平均 $\bar{x}$ としている。

中位数 （中央値，メディアン：median）

データを大きさ順に並べたとき、ちょうど中央に位置するデータの値
　データの数nが奇数のとき、(n+1)/2 番目の値、
　データの数nが偶数のとき、n/2 番目と (n/2)+1 番目の中間の値

数学的には、各要素からの絶対距離の和 $\sum_{i=1}^{n}|x_i - c|$ を最小化する c を
中位数 としている。

この統計量は、解析が困難だが データの例外値（ノイズ）に影響されない という特徴がある。

最頻値 （モード：mode）

度数分布表で度数が最も多い階級値
2つ以上存在することもある

工学的に非常に有用

このほかよく使われる統計値

範囲（レンジ：range）
データの最高値と最低の差

４分位数（quartile）
全体のデータを大きい順に並べたとき、データを4等分する位置の値であり、3つの値がある。（中央の値は中位数）
最小・最大値が他のデータからとび離れ、分散での表現が不適切な場合に用いられる

【小話】　平均時速について
100km の距離を往路では時速50km/h,
復路では時速100km/hで走った。 往復での平均時速は？

このほかよく使われる統計値

範囲（レンジ：range）
データの最高値と最低の差

４分位数（quartile）
全体のデータを大きい順に並べたとき、データを4等分する位置の値であり、3つの値がある。（中央の値は中位数）
最小・最大値が他のデータからとび離れ、分散での表現が不適切な場合に用いられる

【小話】　平均時速について
100km の距離を往路では時速50km/h,
復路では時速100km/hで走った。 往復での平均時速は？

~~$\frac{50+100}{2}=75 \quad (km/h)$~~

往復の所要時間は 3h

$$\frac{100+100}{3}=66.66\cdots \quad (km/h)$$

参考： 画像のノイズ除去

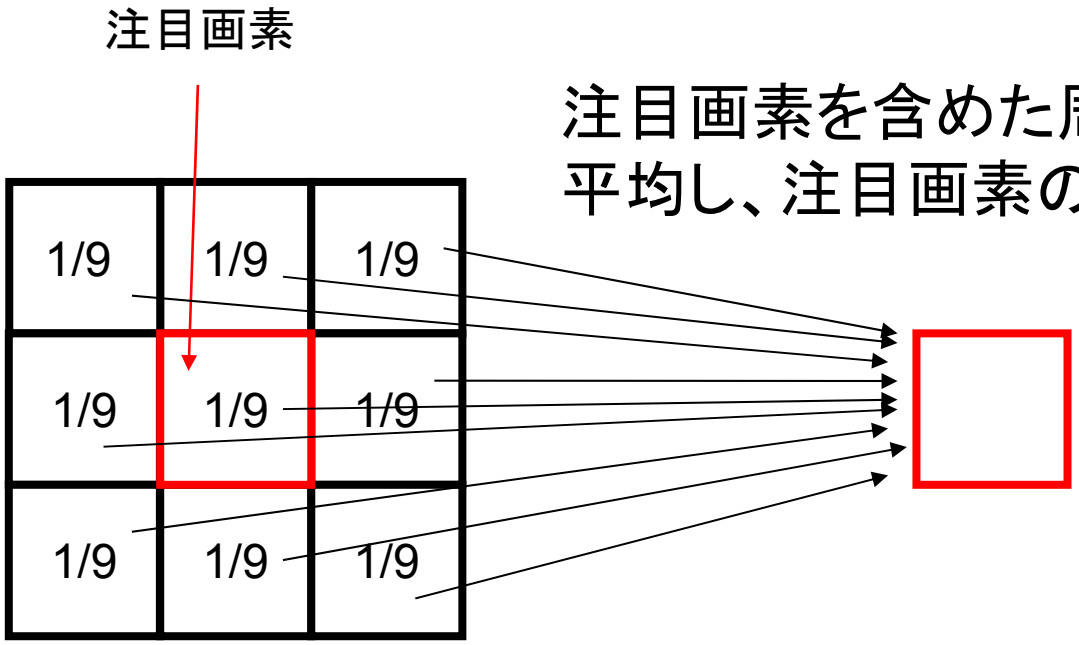
http://cutie.dip.jp/pc/image/
より引用

【元画像】

ノイズ

【ゴマ塩ノイズ画像】

画像のノイズ除去（1）
平均化（平滑化）フィルタによるぼかし処理



注目画素を含めた周囲の画素の明るさを
平均し、注目画素の明るさを置き換える

この置き換え処理を画像中の全ての画素に対して実行する

【ゴマ塩ノイズ画像】

【元画像】

【ぼかし1回】

【ゴマ塩ノイズ画像】

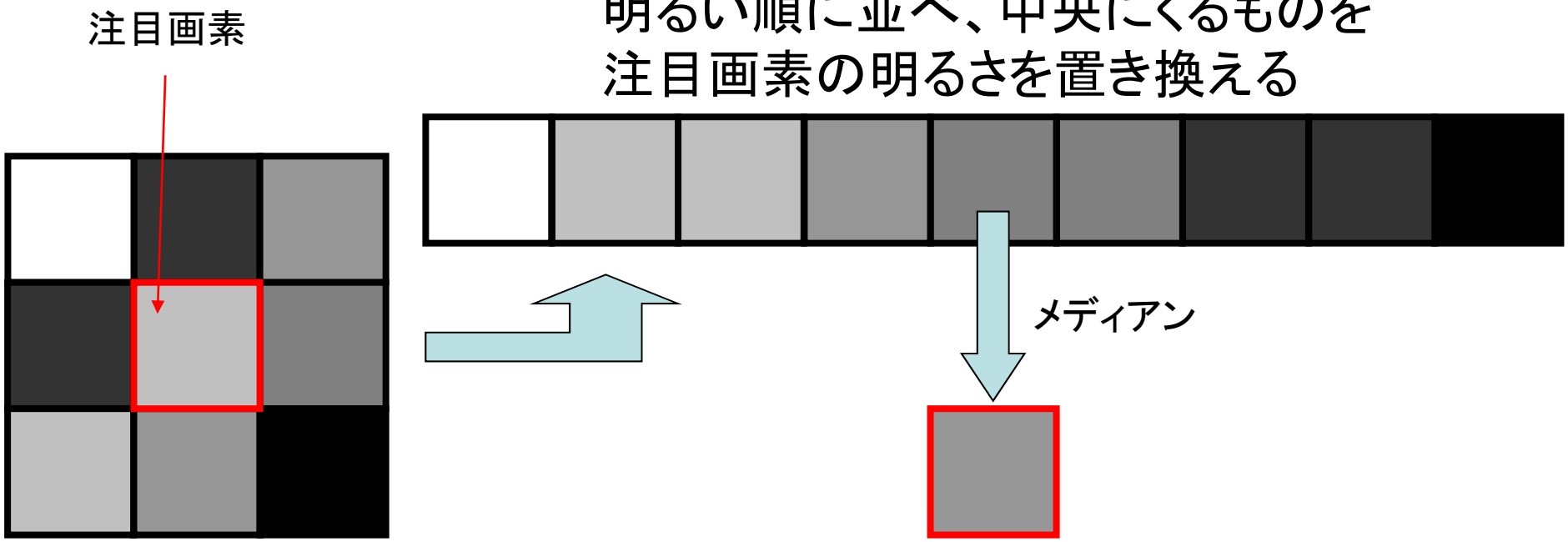
元の画像の
エッジが薄れる

【元画像】

【ぼかし10回】

画像のノイズ除去（2）
メディアンフィルタ

注目画素を含めた周囲の画素の明るさを
明るい順に並べ、中央にくるものを
注目画素の明るさを置き換える



この置き換え処理を画像中の全ての画素に対して実行する

【ゴマ塩ノイズ画像】

【元画像】

【メディアン1回】

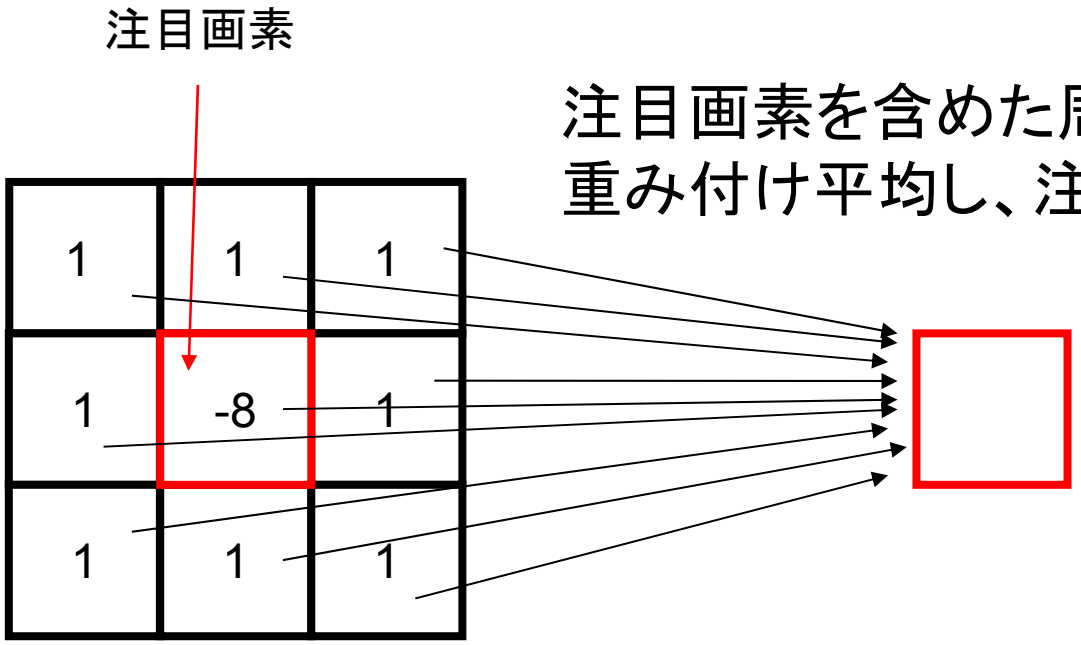
【ゴマ塩ノイズ画像】

元の画像の
エッジを保存

【元画像】

【メディアン10回】

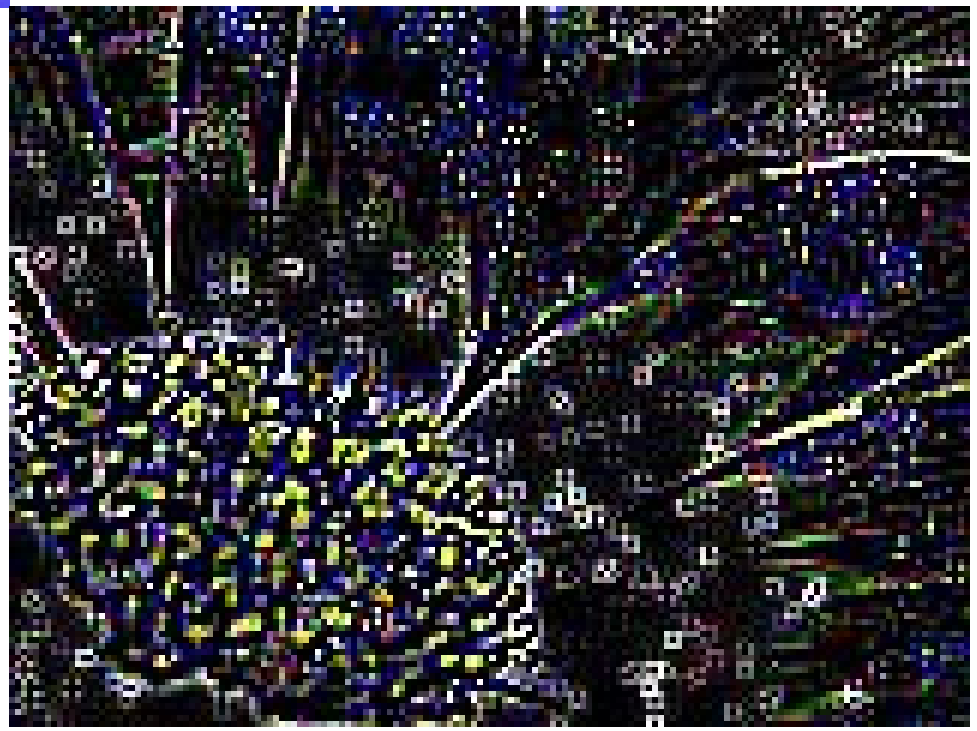
ちなみに…
エッジフィルタ（ラプラシアンフィルタ）



注目画素を含めた周囲の画素の明るさを
重み付け平均し、注目画素と置き換える

この置き換え処理を画像中の全ての画素に対して実行する

【ゴマ塩ノイズ画像】



分散の性質

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \boxed{}$$

分散の定義式

各データの2乗の平均値

（平均値）の2乗

【証明】

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

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

$$=$$

分散の性質

平均値を計算後でないと
分散を計算できない
→各データを2回読む

平均値と分散を同時に計算できる
→各データを1回読込めばよい

工学的
に有用

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

分散の定義式

2乗の平均値

（平均値）の2乗

【証明】

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

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

$$= \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - 2\bar{x}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \left(\frac{1}{n} n (\bar{x})^2\right)$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - 2(\bar{x})^2 + (\bar{x})^2$$

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

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

まとめ

データの整理と表現

1） ヒストグラム（度数分布）

2） **平均・分散・標準偏差**

3） 中位数＝中央値（メジアン）

4） 最頻値（モード）

5） 範囲（レンジ）・4分位数

【演習問題】　2018.04.10　　学籍番号　　氏名

（1）「データの散らばり具合い」を表す統計量は何か？
　またn個のデータを　$x_1, x_2, x_3, \cdots, x_n$　と表すとき、
　上記統計量の定義式を示せ.

（2）次のデータの 平均、分散、メディアン、モードをそれぞれ求めよ。

５　８　３　７　３　９　１０　３　１　９

平均　＝

分散　＝

メディアン＝

モード＝

【演習問題】　　学籍番号　　氏名

（1）「データの散らばり具合い」を表す統計量は何か？
　またn個のデータを　$x_1, x_2, x_3, \cdots, x_n$　と表すとき、
　上記統計量の定義式を示せ.

標本（サンプル）

nコのデータ：　$x_1, x_2, x_3, \cdots, x_n$

平均：　$\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

分散：　$\sigma^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

標準偏差：　$\sigma = \sqrt{\sigma^2}$

データの散らばり具合を表現する

【演習問題】　　学籍番号　　氏名

（2）次のデータの平均、分散、メディアン、モードを求めよ。

5　8　3　7　3　9　10　3　1　9

平均　5．8
分散　42.8　- 5.8^2 = 9.16

2乗の平均値から
平均値の2乗を引く

小さい順に並べる

メディアン（5＋7）／2＝6

1　3　3　3　5　7　8　9　9　10

モード（最頻値）＝3

ヒストグラム（度数分布）を作る