

統計学

データの比較

「教科書 第3章 P60～81」の内容

京都駅前校 木曜4限
担当: 酒井辰也

復 習

●平均(算術平均)(average)

$$\text{平均} = \frac{(\text{データ値の合計})}{(\text{データ値の個数})}$$

$$\text{平均} = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum_{i=1}^N x_i}{N}$$

復習



母平均

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

「ミュー」

違いに
注意！

標本平均

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

「xバー」

- 真の平均値
- 統計調査で一番知りたい平均値

- サンプル調査で得られる平均値
- 母平均とは異なる事が多い

復習

母集団

全データ数 N
母平均 μ

n

サンプリング調査

標本数 n

標本平均 \bar{x}

標本

母分散

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

「シグマの2乗」

※ なぜ、**不偏分散**の場合は $n-1$ で割るのか？は、高度な数学計算による証明が必要なため、当授業では省略します。

※ 教科書では、「標本数が $n < 30$ のときは $n-1$ で割った不偏分散にする」と書いてあります。
標本数が非常に大きければ母分散に近づきます。

標本分散

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

「シグマハットの2乗」

不偏分散

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

「エスの2乗」

復習

母集団

全データ数 N
母平均 μ

n

サンプリング調査

標本数 n

標本平均 \bar{x}

標本

母
標準偏差

σ =
「シグマ」

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

標本
標準偏差

$\hat{\sigma}$ =
「シグマハット」

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

分散のルートを取ったものを「標準偏差」と呼ぶ。

分散は「データ値の2乗」に相当する値だったので、ルートを取ると「データ値の散らばり度合い」をより適切に表現できる。

標準偏差

s =
「エス」

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

確認演習

次の問題を、 ● 電卓 ● ノート にて計算してください。

作業

確認演習①

作業

次のデータの ●平均 ●母分散 ●不偏分散 ●母標準偏差 ●標準偏差
を求めよ。

データ
56
58
60
61
63

各自の電卓で
計算せよ

計算法は次ページ



母分散

$$29.2 / 5 = \underline{5.84}$$

不偏分散

$$29.2 / (5-1) = \underline{7.30}$$

母標準偏差

$$\sqrt{5.84} \doteq \underline{2.42}$$

標準偏差

$$\sqrt{7.30} \doteq \underline{2.70}$$

平均

59.6

確認演習①

作業

次のデータの ●平均 ●母分散 ●不偏分散 ●母標準偏差 ●標準偏差
を求めよ。

5個

データ
56
58
60
61
63

偏差
$56 - 59.6 = -3.6$
$58 - 59.6 = -1.6$
$60 - 59.6 = 0.4$
$61 - 59.6 = 1.4$
$63 - 59.6 = 3.4$

(偏差) ²
12.96
2.56
0.16
1.96
11.56

合計	298
----	-----

平均	<u>59.6</u>
----	-------------

偏差平方和	29.2
-------	------

母分散

$$29.2 / 5 = \underline{5.84}$$

不偏分散

$$29.2 / (5-1) = \underline{7.30}$$

母標準偏差

$$\sqrt{5.84} \doteq \underline{2.42}$$

標準偏差

$$\sqrt{7.30} \doteq \underline{2.70}$$

確認演習②

作業

次のデータの ●平均 ●母分散 ●不偏分散 ●母標準偏差 ●標準偏差
を求めよ。

データ
70
70
70
70
70

偏差
$70 - 70 = 0$
$70 - 70 = 0$
$70 - 70 = 0$
$70 - 70 = 0$
$70 - 70 = 0$

$(\text{偏差})^2$
0
0
0
0
0

合計	350
----	-----

偏差平方和	0
-------	---

平均	<u>70</u>
----	-----------

母分散
$0 / 5 = \underline{0}$

不偏分散
$0 / (5-1) = \underline{0}$

母標準偏差
$\sqrt{0} \doteq \underline{0}$

母標準偏差
$\sqrt{0} \doteq \underline{0}$

データの標準化

- 標準化変量(Z値)
- 偏差値(T値)
- 変動係数

散布図

- 散布図
- 共分散
- 相関係数
- 正の相関・負の相関

回帰分析 (教科書未掲載)

- 回帰直線
- 最小二乗法

質問:

ある人物の期末試験点数は以下の通りだった。

科目	点数
数学	70点
英語	70点

この人は
「数学」「英語」のどちらも「**同程度の学力である**」と言えるか？

科目	点数	平均
数学	70点	? 点
英語	70点	? 点

平均点が分からないので
判断できない。

ならば……

科目	点数	平均
数学	70点	65点
英語	70点	75点

この人の「数学」「英語」の実力は？

点数の「偏差」を計算してみると……

$$\begin{array}{l} \text{数学: } 70 - 65 = +5 \text{ 点} \\ \text{英語: } 70 - 75 = -5 \text{ 点} \end{array}$$

この人は「数学」の方が得意と言える。

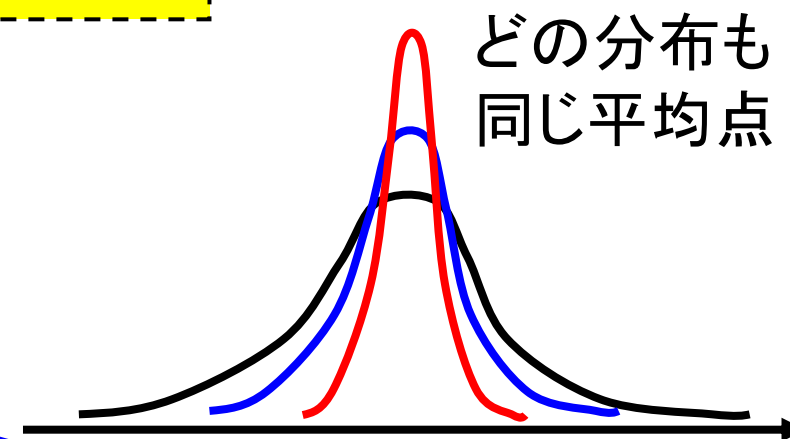
では……

科目	点数	平均
数学	70点	65点
英語	70点	65点

どちらも平均から
+5点

この人の「数学」「英語」の実力は？

科目	点数	平均	標準偏差
数学	70点	65点	? 点
英語	70点	65点	? 点

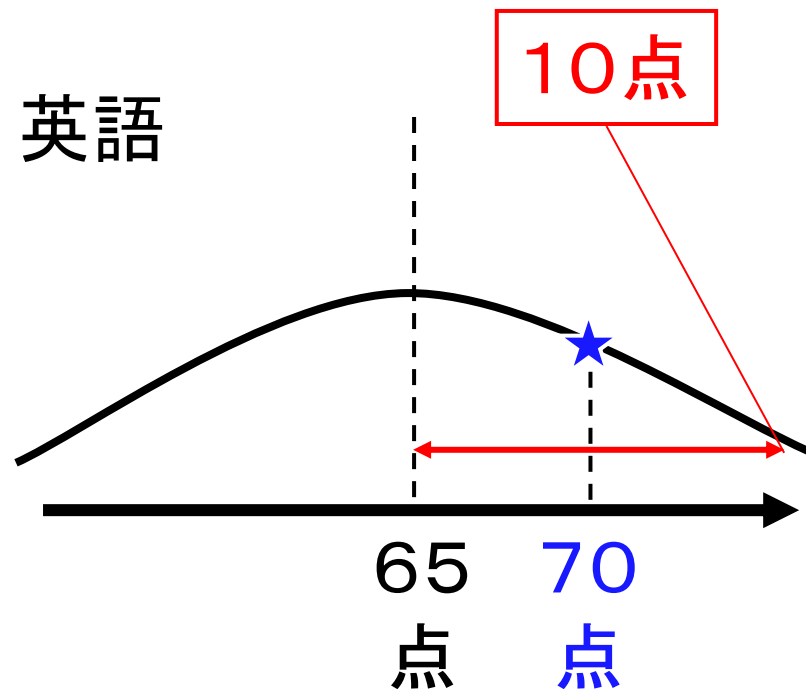
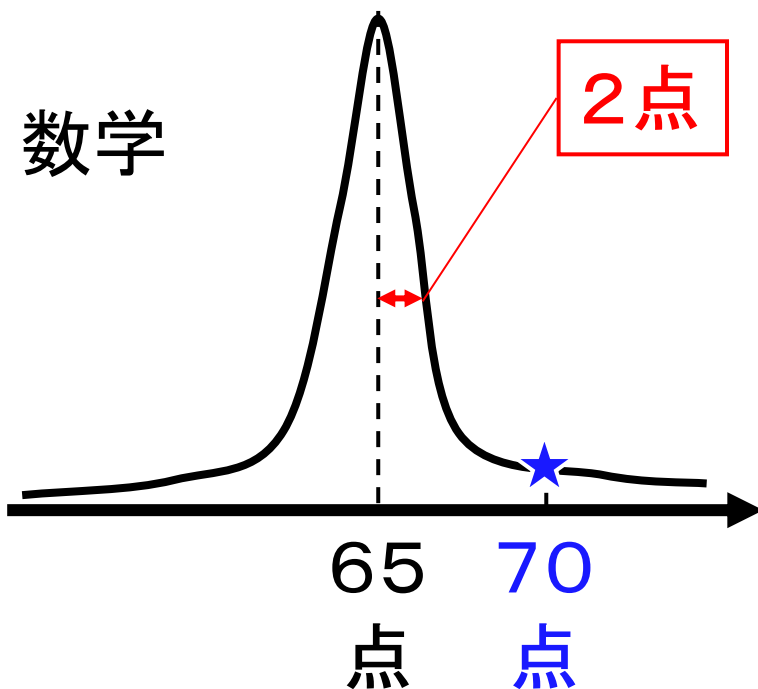


全生徒の点数の、平均からの散らばり具合（標準偏差）が
分からないので、判断できない。

ならば……

科目	点数	平均	標準偏差
数学	70点	65点	2点
英語	70点	65点	10点

この人の「数学」「英語」の実力は？



「数学」の方が順位が高い

データの標準化

- 標準化変量(Z値)
- 偏差値(T値)
- 変動係数

散布図

- 散布図
- 共分散
- 相関係数
- 正の相関・負の相関

回帰分析 (教科書未掲載)

- 回帰直線
- 最小二乗法

データの標準化

元のデータ値を代表値や散布度を用いて変換し、データ全体における相対的な位置づけを比べられるようにする事。

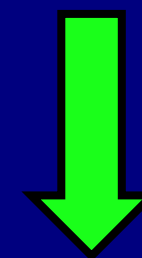
例えば、さきほどの
「英語」と「数学」の実力がどちらが高いか？を
わかりやすく表現するための数値変換です。

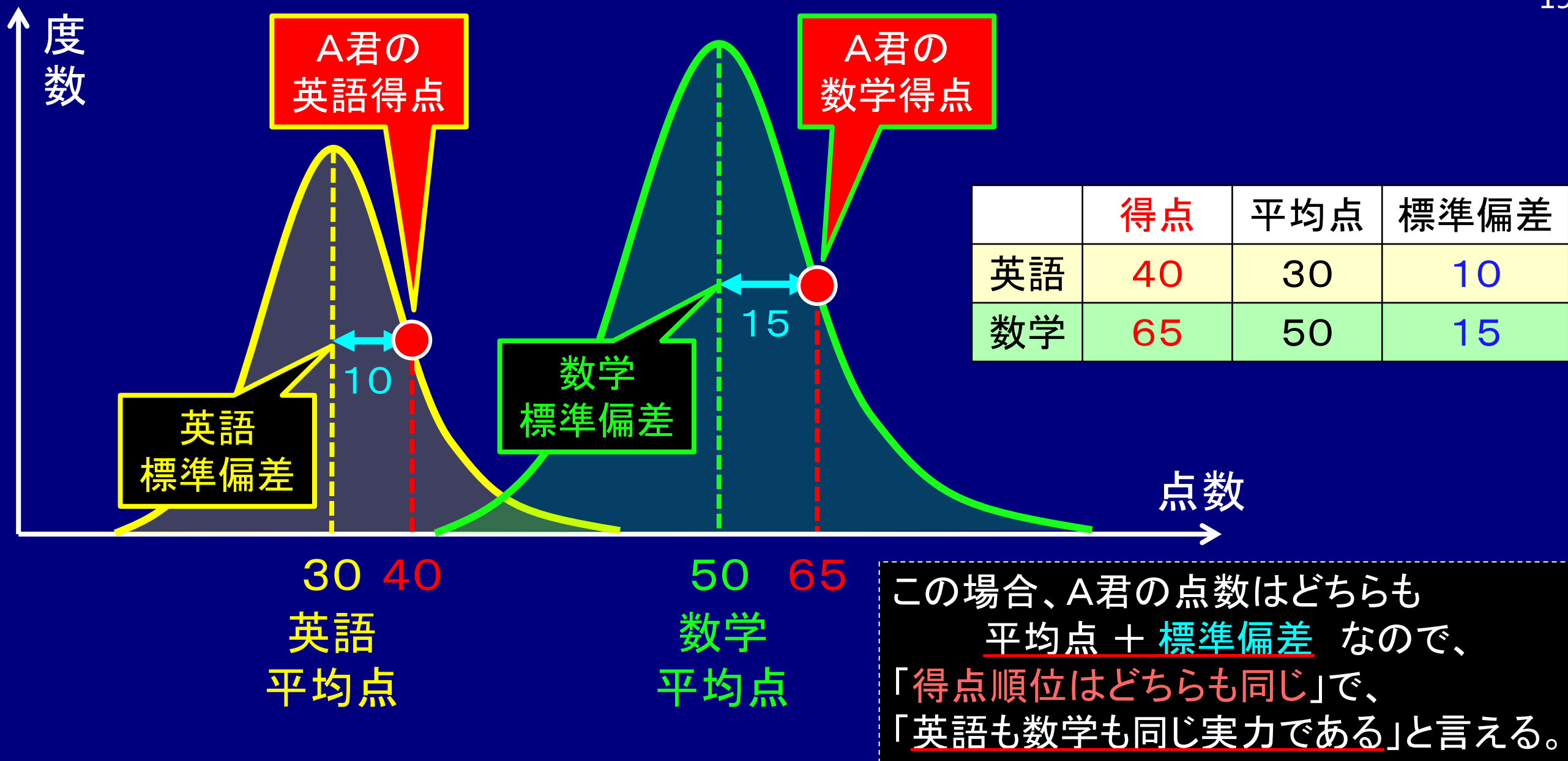
A君の英語と数学の試験点数は以下であった。

A君は英語と数学のどちらが優秀であるか？

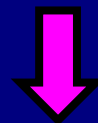
	得点	平均点	標準偏差
英語	40	30	10
数学	65	50	15

グラフ化してみると……





……ということで、次のような値を導入する。



標準化変量 (Z値)

$$Z_i = \frac{x_i - \bar{x}}{\sigma_x}$$

\bar{x} : x の平均

σ_x : x の標準偏差

- 各データ値 x_i の偏差を、標準偏差で割ったものを「標準化変量 (Z値)」と呼ぶ。
- 標準化変量 (Z値) は、「各データ値 x_i の偏差」が
標準偏差の何倍かを示す値となる。
- 標準化変量 (Z値) は、必ず「平均が“0”」「標準偏差が“1”」となる。

(注意) 教科書では(P62)、標準化変量を“ u ”や“ v ”と書いているが、
呼び名が「Z値」である事から、当スライドでは“Z”と表示する事にします。

No.	データ値	偏差	標準化変量 (Z値)
1	44	44 - 42 = 2	2 / 10.8 = 0.19
2	55	55 - 42 = 13	13 / 10.8 = 1.20
3	53	53 - 42 = 11	11 / 10.8 = 1.02
4	29	29 - 42 = -13	-13 / 10.8 = -1.20
5	37	37 - 42 = -5	-5 / 10.8 = -0.46
6	42	42 - 42 = 0	0 / 10.8 = 0.00
7	25	25 - 42 = -17	-17 / 10.8 = -1.57
8	38	38 - 42 = -4	-4 / 10.8 = -0.37
9	58	58 - 42 = 16	16 / 10.8 = 1.48
10	39	39 - 42 = -3	-3 / 10.8 = -0.28

平均	42.0
----	------

標準偏差	10.8
------	------

(数値は四捨五入しています)

標準化変量 (Z値)

$$Z_i = \frac{x_i - \bar{x}}{\sigma_x}$$

標準化変量 (Z値) は、元のデータ値が、

- 平均より大きいと“正”
- 平均と等しければ“0”
- 平均より小さいと“負”

となる。

標準化変量(Z値)を用いる事で……

① 同じ集団内での平均値に対する値の大小が判る。(偏差だけでも判るが)

	試験の 点数	受験者の 平均点	受験者の 標準偏差		Z値
A君	80	65	10	$(80 - 65) \div 10 = 15 \div 10 =$	1.5
B君	70			$(70 - 65) \div 10 = 5 \div 10 =$	0.5
C君	60			$(60 - 65) \div 10 = -5 \div 10 =$	-0.5

優秀

② 異なる集団でも、値の相対的な高さの比較が可能になる。

	A君の 点数	各科目の 平均点	各科目の 標準偏差		Z値
国語	80	65	15	$(80 - 65) \div 15 = 15 \div 15 =$	1.0
英語	50	60	20	$(50 - 60) \div 20 = -10 \div 20 =$	-0.5
数学	45	25	10	$(45 - 25) \div 10 = 20 \div 10 =$	2.0

得意

標準偏差
1

Z値

元データ

平均0

標準偏差
1

Z値

平均0

「値の大小」や
「値の高さ」の比較が
容易になる

標準偏差
1

Z値

元データ

平均0

標準偏差
1

Z値

平均0

標準化変量 (Z値)

$$Z_i = \frac{x_i - \bar{x}}{\sigma_x}$$

\bar{x} : x の平均

σ_x : x の標準偏差

標準化変量 (Z値) は

- 平均が“0”
- 標準偏差が“1”

このような値は、他にも以下のものがある。

偏差値 (T値)

$$T_i = 10 \times \frac{x_i - \bar{x}}{\sigma_x} + 50$$

偏差値 (T値) は

- 平均が“50”
- 標準偏差が“10”

標準偏差は「平均値を中心としたデータの散らばり具合」を表す数値だが、異なるデータ集団を比較する際、

(例：大学生10000人の身長データ vs 小学生10000人の身長データ)
標準偏差だけで大学生と小学生のデータの散らばり度合いを比較する事は出来ず、以下の**変動係数**を用いる。

変動係数 (the coefficient of variation)

$$CV = \frac{\sigma_x}{\bar{x}}$$

σ_x : x の標準偏差

\bar{x} : x の平均

統計学を学ぶには数学に慣れていきましょう。

- 統計学でよく登場するシグマ記号 Σ に慣れる
- 標準化変量(Z値)の平均が“0”、標準偏差が“1”になる理由を考える。

●まず、シグマ記号とは、**数の和**を計算する記号です。

$$\sum_{i=1}^5 i = 1 + 2 + 3 + 4 + 5 = \underline{15} \blacksquare$$

$$\sum_{i=1}^5 i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 1 + 4 + 9 + 16 + 25 = \underline{55} \blacksquare$$

$$\begin{aligned} \sum_{i=1}^5 8 &= 8 + 8 + 8 + 8 + 8 \\ &= 8 \times 5\text{個} = \underline{40} \blacksquare \end{aligned}$$

$$\sum_{i=1}^N (\text{定数}) = (\text{定数}) \times N$$

$$\sum_{i=1}^N x_i = x_1 + x_2 + \cdots + x_N$$

●まず、シグマ記号とは、**数の和**を計算する記号です。

$$\begin{aligned}\sum_{i=1}^N (x_i + y_i) &= (x_1 + y_1) + (x_2 + y_2) + \cdots + (x_N + y_N) \\ &= (x_1 + x_2 + \cdots + x_N) + (y_1 + y_2 + \cdots + y_N) \\ &= \sum_{i=1}^N x_i + \sum_{i=1}^N y_i\end{aligned}$$

★「足し算」の**総和(シグマ)**は、個別の項の総和に分解できる。

$$\sum_{i=1}^N (x_i + y_i) = \sum_{i=1}^N x_i + \sum_{i=1}^N y_i$$

●まず、シグマ記号とは、**数の和**を計算する記号です。

$$\begin{aligned}\sum_{i=1}^N 3x_i &= 3x_1 + 3x_2 + \cdots + 3x_N \\ &= 3(x_1 + x_2 + \cdots + x_N) \\ &= 3 \sum_{i=1}^N x_i\end{aligned}$$

★「**定数倍**」は、**シグマ記号**の外に出す事ができる。

$$\sum_{i=1}^N a x_i = a \sum_{i=1}^N x_i \quad (a: \text{定数})$$

★「**足し算**」の**総和(シグマ)**は、個別の項の総和に分解できる。

$$\sum_{i=1}^N (a x_i + b y_i + c) = a \sum_{i=1}^N x_i + b \sum_{i=1}^N y_i + N c$$

以下のような間違いをしないように！

$$\sum_{i=1}^N \frac{ax_i + b}{c} = \frac{\sum_{i=1}^N (ax_i + b)}{\sum_{i=1}^N c}$$

正しくは

$$\begin{aligned} \sum_{i=1}^N \frac{ax_i + b}{c} &= \sum_{i=1}^N \left[\frac{ax_i + b}{c} \right] = \sum_{i=1}^N \left[\frac{a}{c} x_i + \frac{b}{c} \right] \\ &= \sum_{i=1}^N \frac{a}{c} x_i + \sum_{i=1}^N \frac{b}{c} = \frac{a}{c} \sum_{i=1}^N x_i + N \frac{b}{c} \end{aligned}$$

● N 個のデータ $\{x_1, x_2, \dots, x_N\}$ の平均は

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{1}{N} \sum_{i=1}^N x_i$$

$$\boxed{\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i} \dots \textcircled{1} \quad \boxed{\sum_{i=1}^N x_i = N \bar{x}} \dots \textcircled{1}'$$

● N 個のデータ $\{x_1, x_2, \dots, x_N\}$ をすべて a 倍した
 $\{ax_1, ax_2, \dots, ax_N\}$ の平均は……

$$\begin{aligned}\overline{(ax)} &= \frac{ax_1 + ax_2 + \dots + ax_N}{N} \\ &= \frac{1}{N} \sum_{i=1}^N ax_i = a \frac{1}{N} \sum_{i=1}^N x_i = a \cdot \bar{x}\end{aligned}$$

★すべて a 倍したデータの平均は、元データの平均の a 倍になる。

$$\boxed{\overline{(ax)} = a \cdot \bar{x}} \quad \dots \textcircled{2}$$

(例) 全商品の価格を1.5倍にすると、平均価格も1.5倍になる。

● **N個**のデータ $\{x_1, x_2, \dots, x_N\}$ すべてに b を加算した、
 $\{x_1 + b, x_2 + b, \dots, x_N + b\}$ の平均は……

$$\begin{aligned} \overline{(x+b)} &= \frac{1}{N} \sum_{i=1}^N (x_i + b) = \frac{1}{N} \left\{ \sum_{i=1}^N x_i + \sum_{i=1}^N b \right\} \\ &= \frac{1}{N} \left\{ \sum_{i=1}^N x_i + Nb \right\} = \frac{1}{N} \left\{ N\bar{x} + Nb \right\} \\ &= \bar{x} + b \end{aligned}$$

★すべてに b を足した データの平均は、
 元データの平均に b を足した ものになる。

$$\begin{aligned} \overline{(x+b)} \\ &= \bar{x} + b \end{aligned}$$

(例) 全商品の価格を20円値上げすると、平均価格も20円上がる。…… ③

● **N**個のデータ $\{x_1, x_2, \dots, x_N\}$ から作った、

$\{ax_1 + b, ax_2 + b, \dots, ax_N + b\}$ の平均は……

$$\begin{aligned} \overline{(ax + b)} &= \frac{1}{N} \sum_{i=1}^N (ax_i + b) = \frac{1}{N} \left\{ \sum_{i=1}^N ax_i + \sum_{i=1}^N b \right\} \\ &= \frac{1}{N} \left\{ a \sum_{i=1}^N x_i + Nb \right\} = \frac{1}{N} \left\{ a N \bar{x} + Nb \right\} \\ &= a \bar{x} + b \end{aligned}$$

★すべて a 倍 して b を足した データの平均は、
元データの平均を a 倍 して b を足した ものになる。

$$\begin{aligned} \overline{(ax + b)} \\ &= a \bar{x} + b \end{aligned}$$

…… ③

商品の元値 x_i		税込価格 $1.08 \cdot x_i$		値上げ後の売値 $1.08 \cdot x_i + 20$
100		108		128
200		216		236
300		324		344
400		432		452
500		540		560

消費税加算
 $\times 1.08$

20円値上げ
 $+ 20$

元値の平均

「値上げ後の売値」の平均

$$\begin{aligned}\bar{x} &= \frac{100 + 200 + 300 + 400 + 500}{5} \\ &= 300 \blacksquare\end{aligned}$$

$$\begin{aligned}& \frac{128 + 236 + 344 + 452 + 560}{5} \\ &= \frac{1720}{5} = 344 \blacksquare\end{aligned}$$

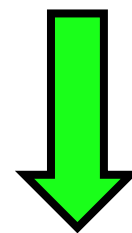
$$1.08 \times 300 + 20 = 344 \text{ と一致する。}$$

● N 個のデータ $\{x_1, x_2, \dots, x_N\}$ の標準偏差は……

$$\sigma_x = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \dots\dots \textcircled{4}$$

● N 個のデータ $\{x_1, x_2, \dots, x_N\}$ から作った、
 $\{ax_1 + b, ax_2 + b, \dots, ax_N + b\}$ の標準偏差は……

$$\sigma_{ax+b} = \sqrt{\frac{1}{N} \sum_{i=1}^N \{ (ax_i + b) - \overline{(ax + b)} \}^2}$$

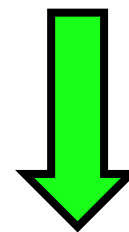


$$\sigma_{ax+b} = \sqrt{\frac{1}{N} \sum_{i=1}^N \{ (ax_i + b) - \overline{(ax + b)} \}^2}$$

$$= \sqrt{\frac{1}{N} \sum_{i=1}^N \{ (ax_i + \cancel{b}) - (a\bar{x} + \cancel{b}) \}^2}$$

$$= \sqrt{\frac{1}{N} \sum_{i=1}^N \{ ax_i - a\bar{x} \}^2}$$

$$= \sqrt{\frac{1}{N} \sum_{i=1}^N \{ a(x_i - \bar{x}) \}^2}$$



$$\begin{aligned}
\sigma_{ax+b} &= \sqrt{\frac{1}{N} \sum_{i=1}^N \{ a(x_i - \bar{x}) \}^2} \\
&= \sqrt{\frac{1}{N} \sum_{i=1}^N a^2(x_i - \bar{x})^2} \\
&= \sqrt{\frac{a^2}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \\
&= a \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \\
&= a \cdot \sigma_x
\end{aligned}$$

★すべて a 倍 して b を足した
 データの標準偏差は、
 元データの標準偏差を a 倍
 したものになる。

$$\sigma_{ax+b} = a \cdot \sigma_x$$

..... ⑤

標準化変量 (Z値) $Z_i = \frac{x_i - \bar{x}}{\sigma_x} = \frac{1}{\sigma_x} x_i - \frac{\bar{x}}{\sigma_x}$ なので..

●平均は $\overline{Z} = \overline{\left\{ \frac{1}{\sigma_x} x_i - \frac{\bar{x}}{\sigma_x} \right\}} = \frac{1}{\sigma_x} \bar{x} - \frac{\bar{x}}{\sigma_x} = \underline{0}$ ■

●標準偏差は $\sigma_Z = \left\{ \frac{1}{\sigma_x} x_i - \frac{\bar{x}}{\sigma_x} \right\}$ の標準偏差

$$= \frac{1}{\sigma_x} \cdot \sigma_x = \underline{1}$$

⑤式より

標準化変量 (Z値) は
 ●平均が“0”
 ●標準偏差が“1”

データの標準化

- 標準化変量(Z値)
- 偏差値(T値)
- 変動係数

散布図

- 散布図
- 共分散
- 相関係数
- 正の相関・負の相関

回帰分析 (教科書未掲載)

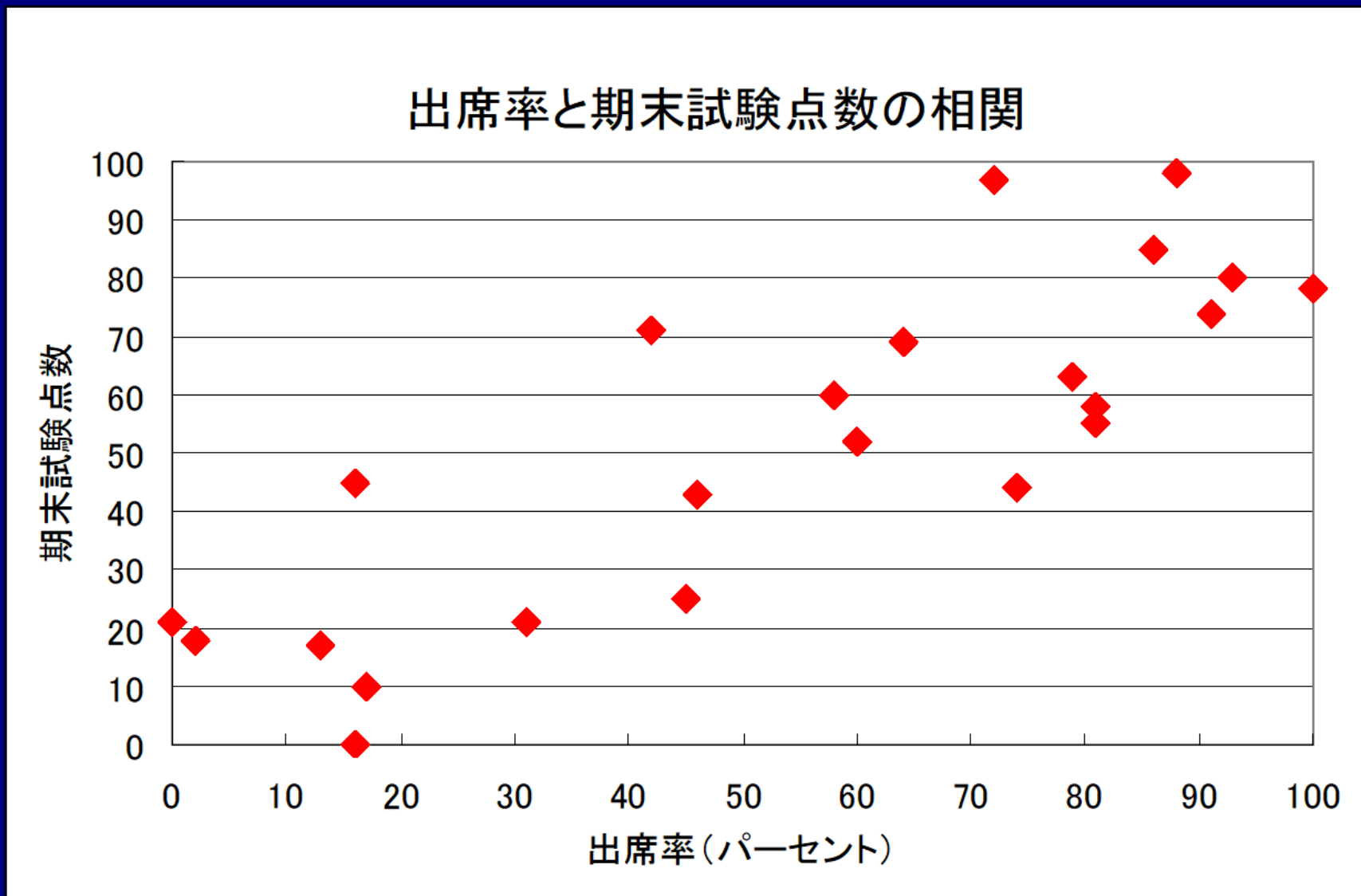
- 回帰直線
- 最小二乗法

散布図

各データが2つ変数で構成される値の場合、
2変数それぞれを X 座標、 Y 座標として、
データを XY 平面にプロットした図の事。

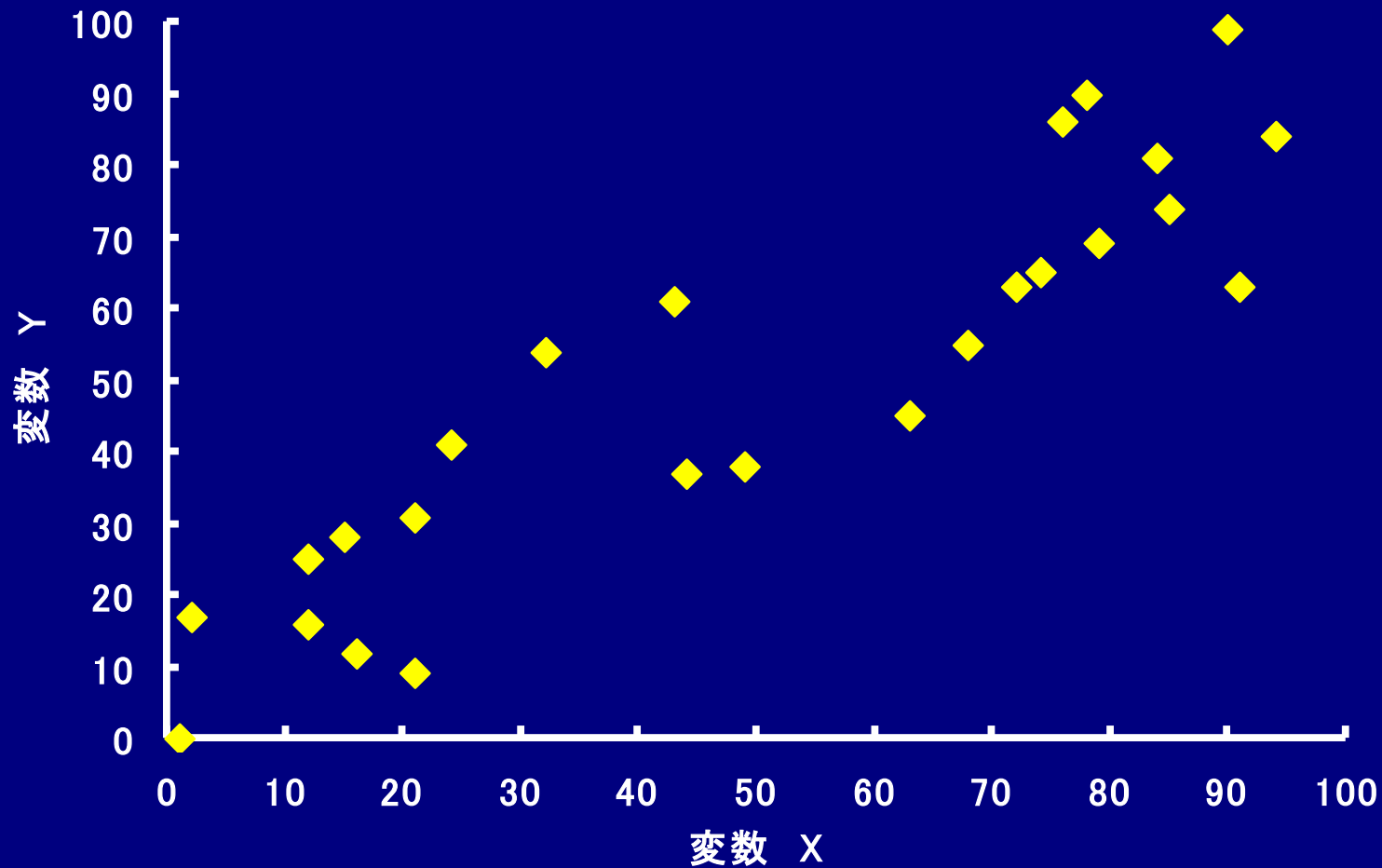
散布図

散布図の例



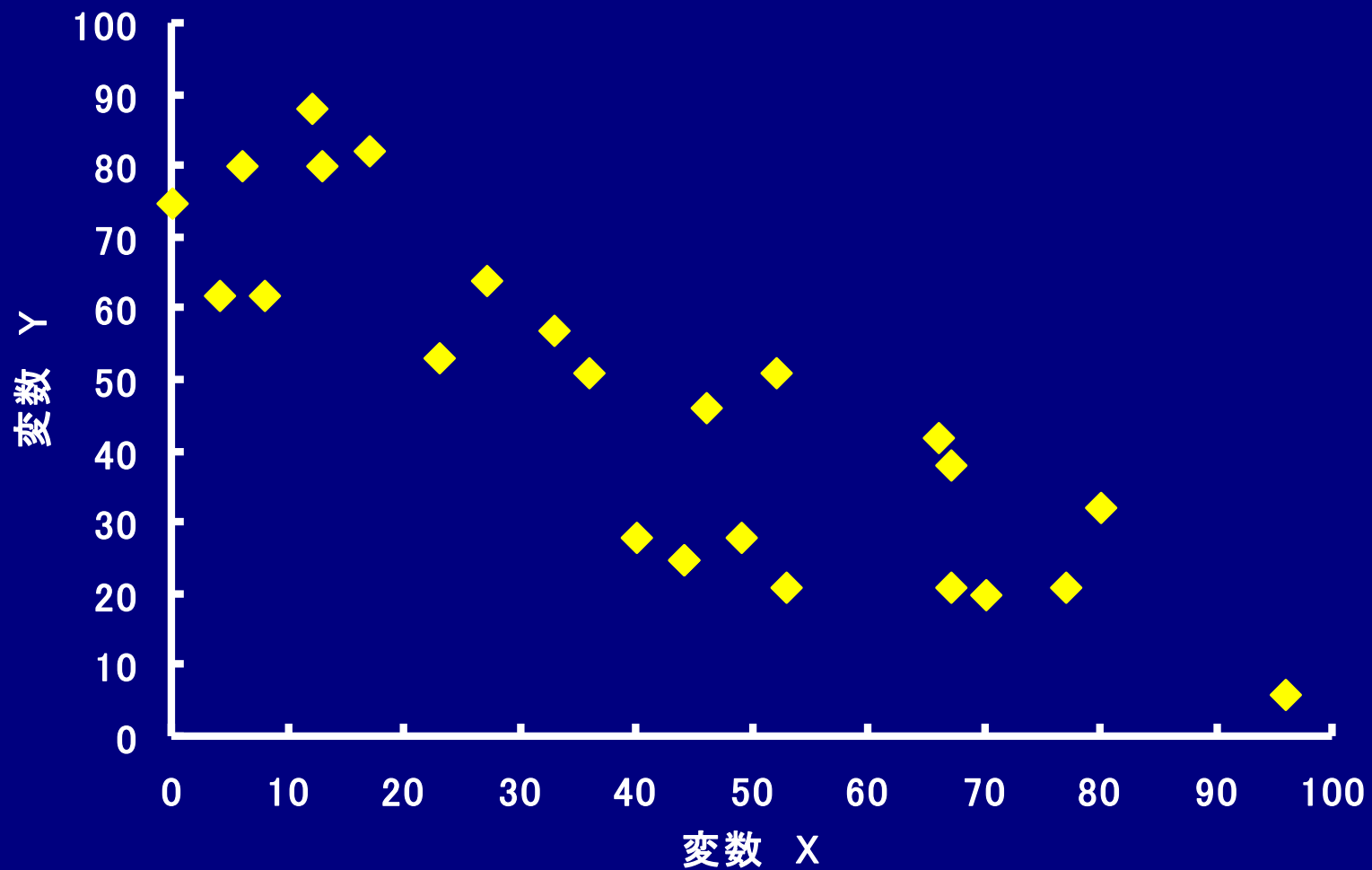
散布図

A. 正の相関がある例



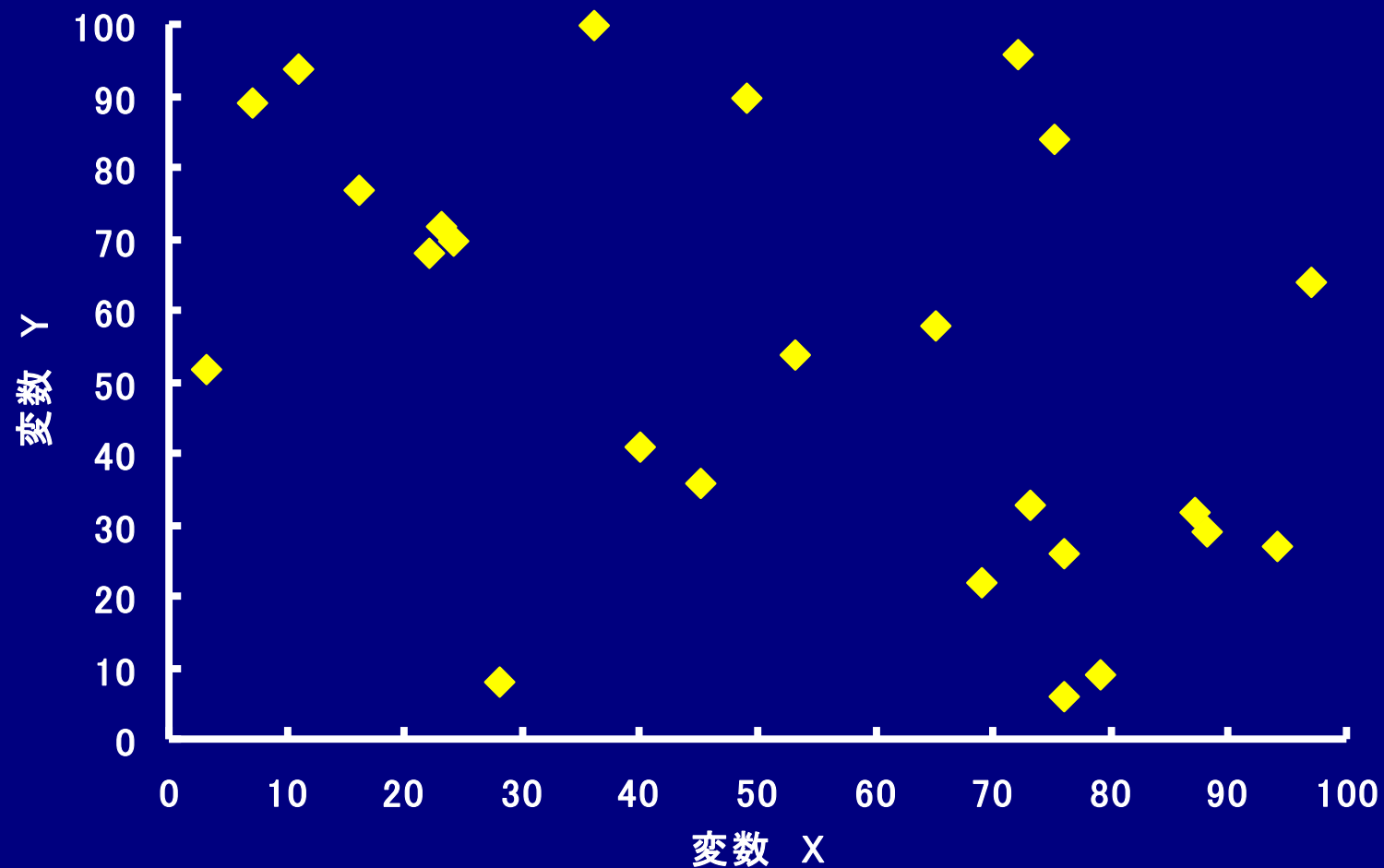
散布図

B. 負の相関がある例



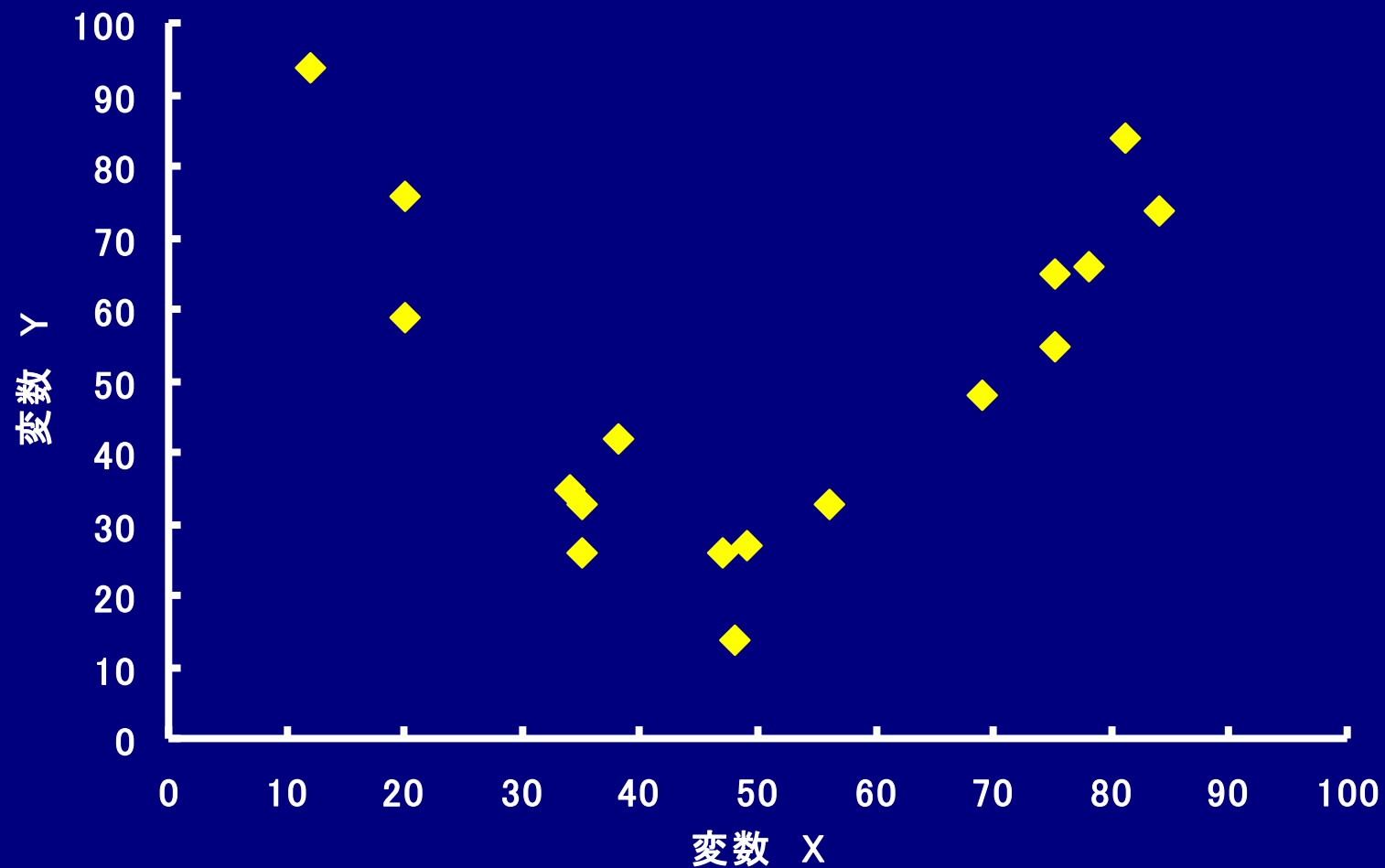
散布図

C. 無相関(相関がない)の例



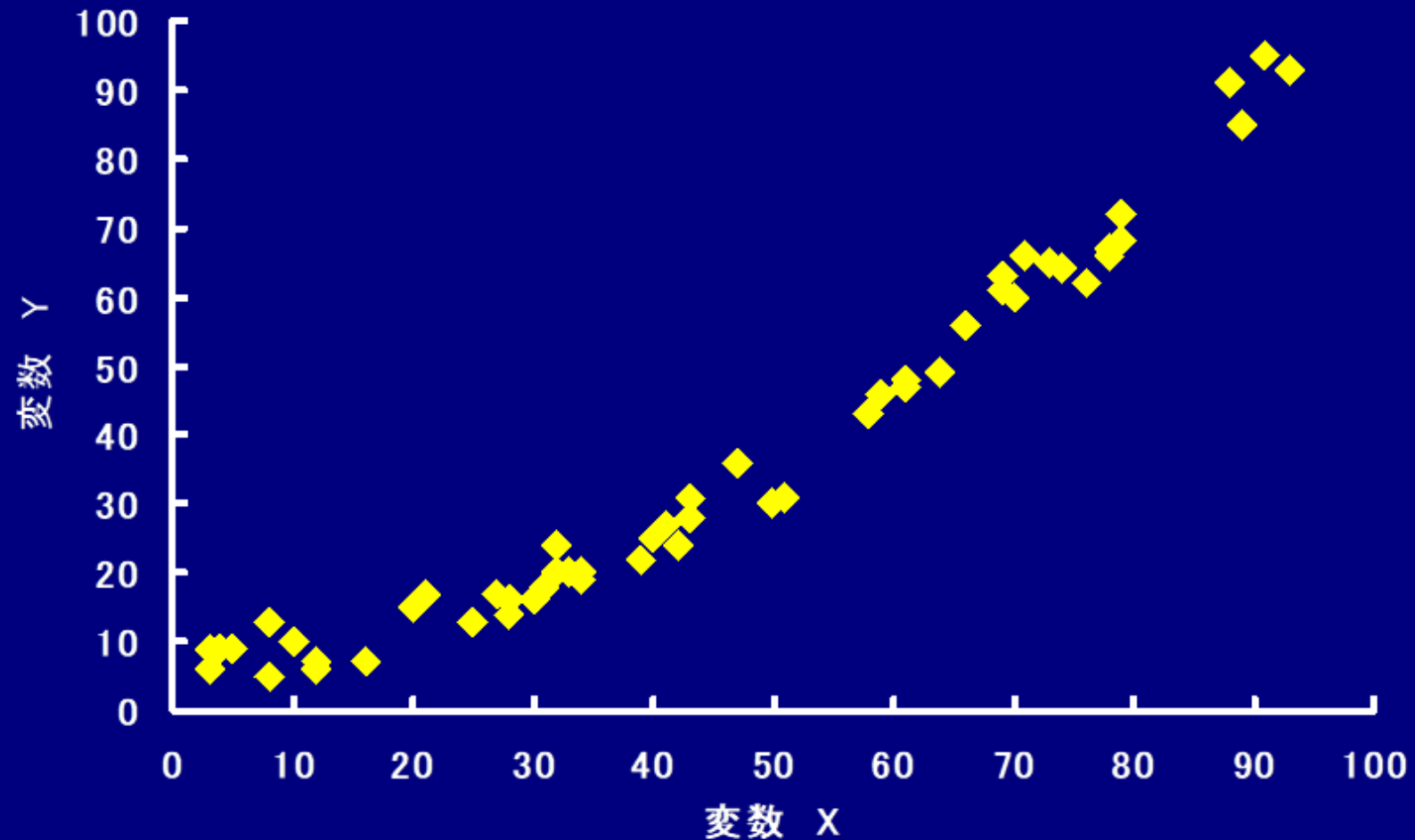
散布図

D. 2次式的関係の例①



散布図

E. 2次式的関係の例②



ならば……

被験者 番号	ビタミンA 摂取量	成長ホルモン 増加量
A001	284	451
A002	395	589
A003	827	875
A004	484	981
A005	30	593
A006	78	727
A007	1103	450
A008	853	589
A009	312	1203
A010	493	534
A011	47	539
A012	356	874
A013	45	238
A014	2	639
A015	674	347
A016	467	1237
A017	538	896
A018	777	653
A019	854	468
A020	1342	623

この表から

「**ビタミンA摂取量**」と「**成長ホルモン増加量**」との間に、

- 相関関係があるか？
- どのような相関関係か？（正？負？）
- 相関関係はどれくらいの度合いか？

……と言われればどうする？

散布図を描いて、目で見るだけの判断をしても、
正確には判りづらい

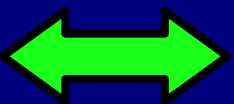
そこでこれらを、「**ある数値**」を計算する事によって
考えてみる ↓

2変数を持つデータ(全N個)があるとき、

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$$

相関係数 (ピアソンの積率相関係数)

$$r_{xy} = \frac{\sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}$$

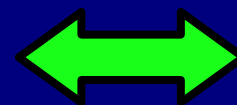


2変数を持つデータ(全N個)があるとき、

(x_1, y_1) 、 (x_2, y_2) 、 $\dots\dots\dots$ 、 (x_N, y_N)

相関係数 (ピアソンの積率相関係数)

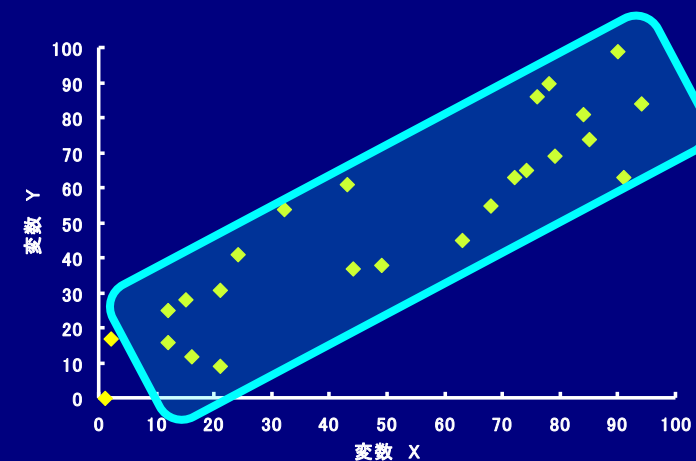
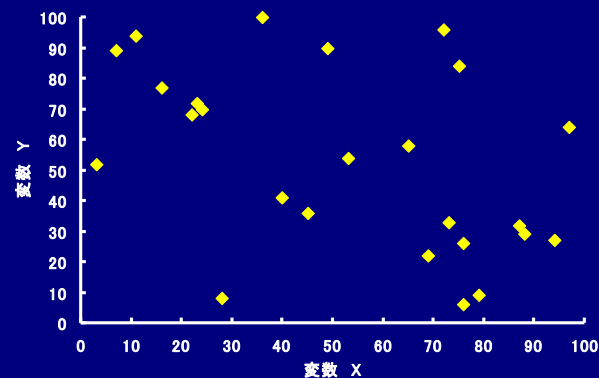
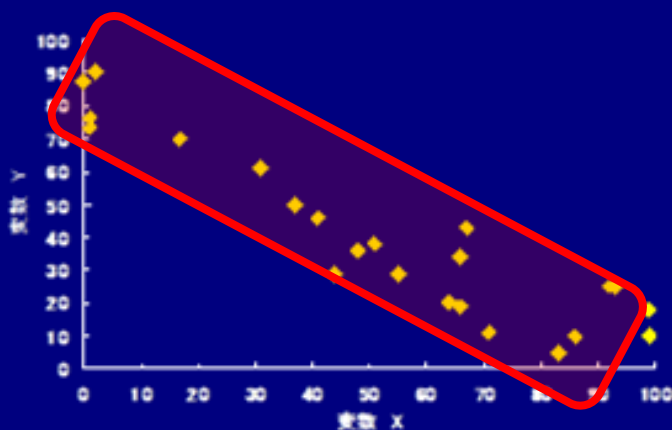
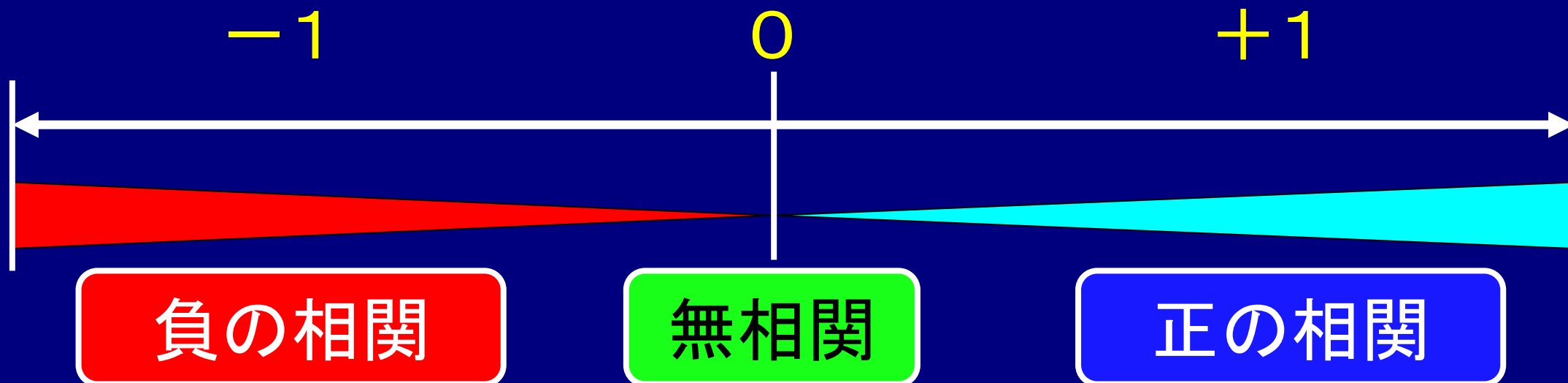
$$r_{xy} = \frac{\begin{array}{c} (x\text{の偏差}) \cdot (y\text{の偏差}) \\ \text{の合計} \end{array}}{\begin{array}{c} (x\text{の偏差})\text{の2乗} \\ \text{の合計} \\ (x\text{の偏差平方和}) \end{array} \begin{array}{c} (y\text{の偏差})\text{の2乗} \\ \text{の合計} \\ (y\text{の偏差平方和}) \end{array}}$$



相関係数 r_{xy} は、

- 取り得る範囲 $-1 \leq r_{xy} \leq +1$
- 「正の相関」なら $0 < r_{xy} \leq +1$
★ $+1$ に近いほど強い正の相関
- 「負の相関」のなら $-1 \leq r_{xy} < 0$
★ -1 に近いほど強い負の相関
- r_{xy} が 0 に近いほど 「無相関」


相関係数 r_{xy} は、



相関係数 r_{xy} の数式の意味は？

分子・分母に以下のような
同じ数を掛けてみると……

$$r_{xy} = \frac{\sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}} \times \frac{\left[\frac{1}{N} \right]}{\left[\frac{1}{N} \right]}$$

$$= \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2}}$$


相関係数 r_{xy} の数式の意味は？

$$\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

x と y の「**共分散**」

$r_{xy} =$

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

x の標準偏差

$$\sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2}$$

y の標準偏差

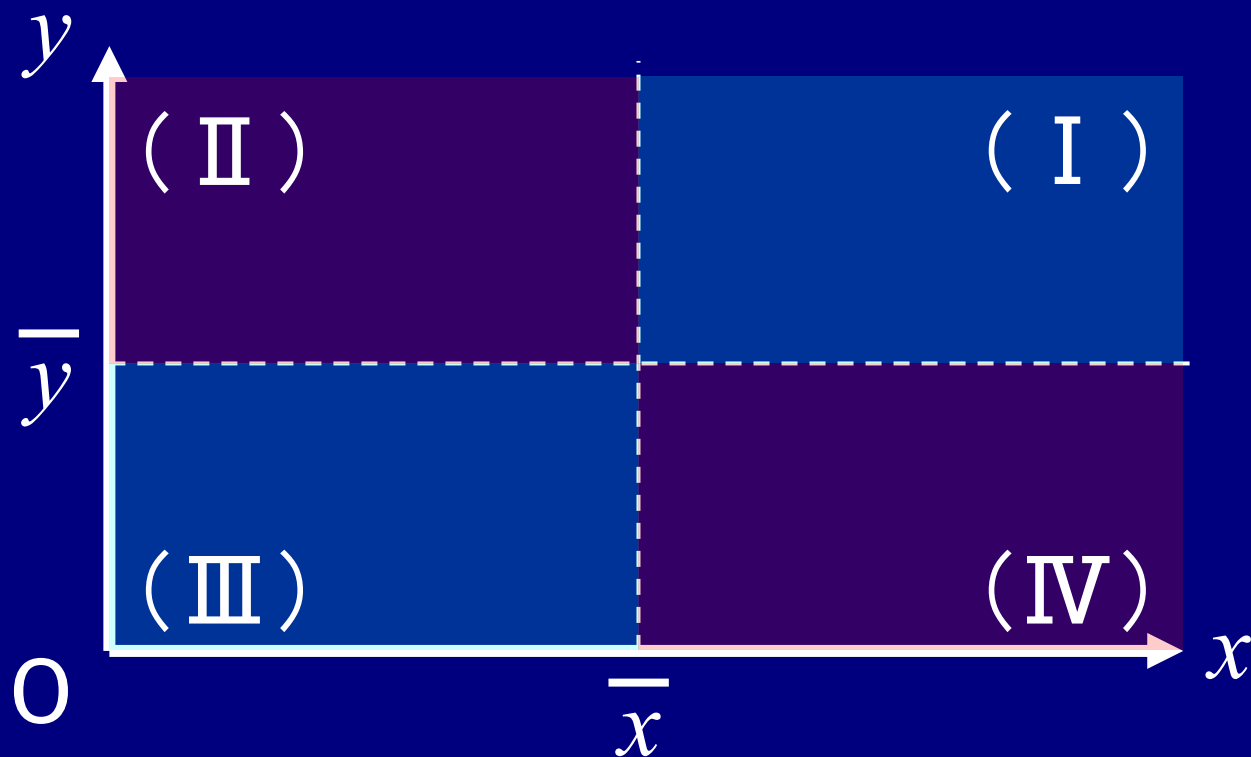
$$\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

$=$

$$\sigma_x \cdot \sigma_y$$



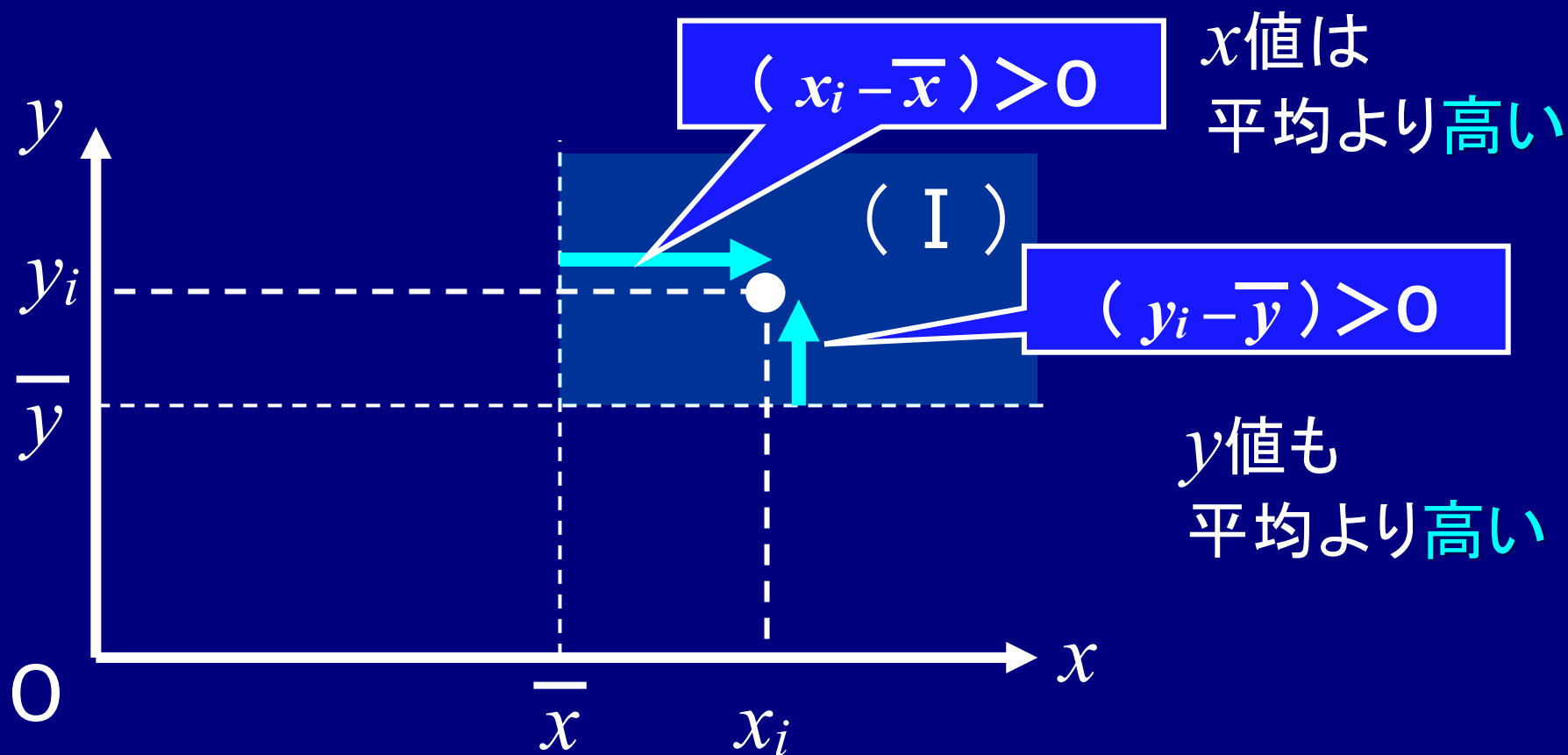
$$r_{xy} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sigma_x \cdot \sigma_y}$$



$$\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y}) > 0 \text{ (プラス) の寄与}$$

$$r_{xy} = \frac{\sigma_x \cdot \sigma_y}{\sigma_x \cdot \sigma_y}$$

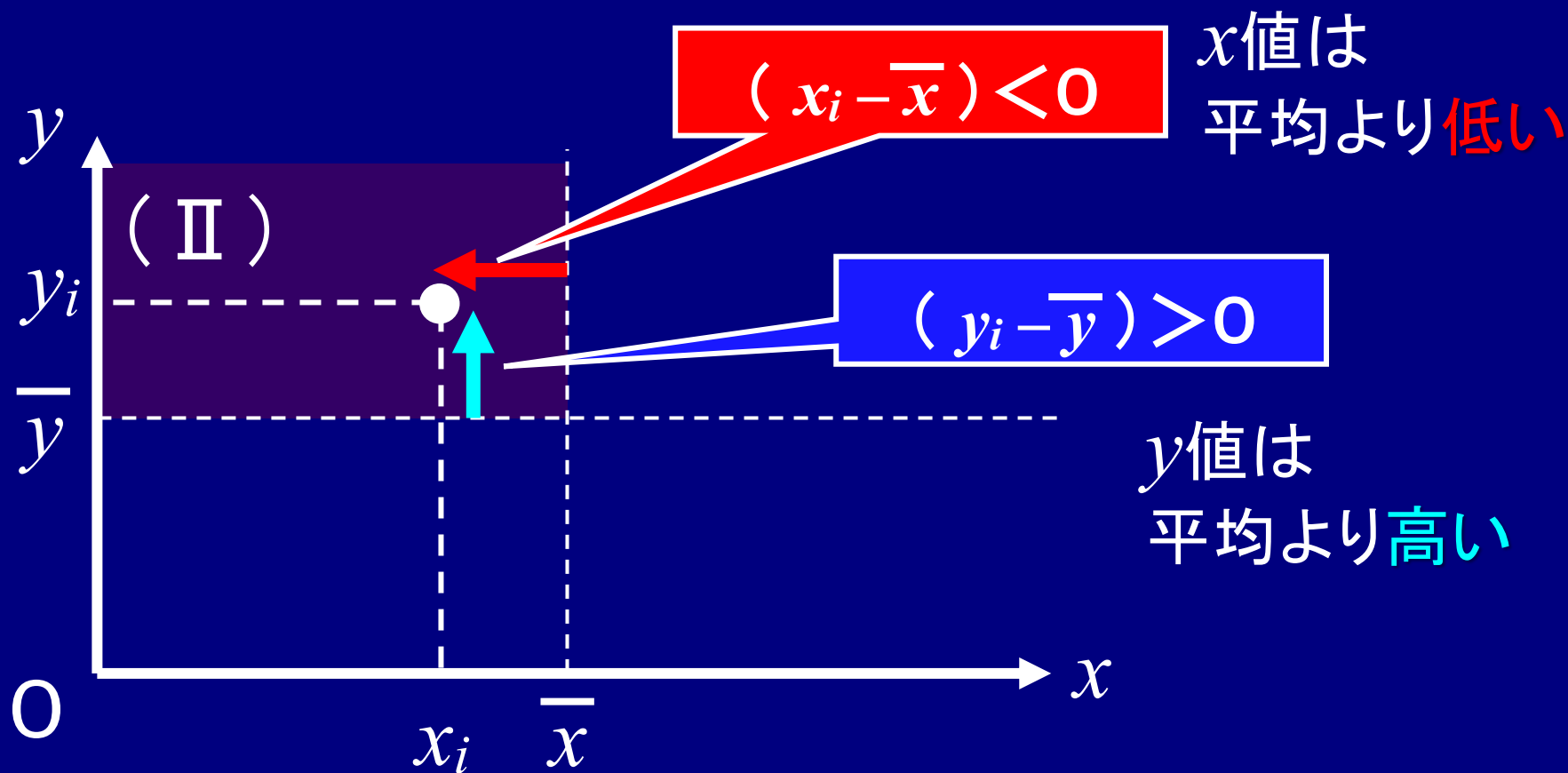
(プラス) × (プラス)
= (プラス)



$$\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y}) < 0 \text{ (マイナス)の寄与}$$

$$r_{xy} = \frac{\quad}{\sigma_x \cdot \sigma_y}$$

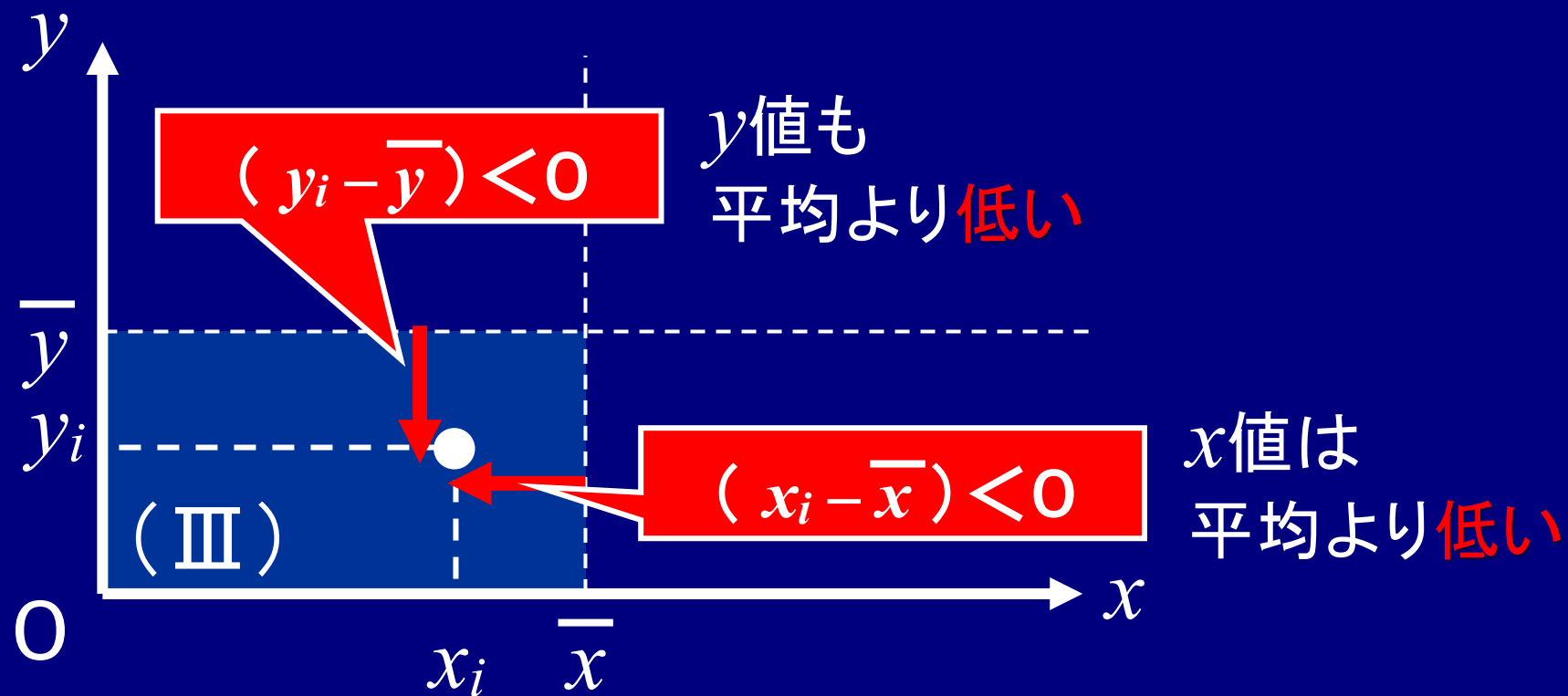
(マイナス) × (プラス)
= (マイナス)



$$\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y}) > 0 \text{ (プラス) の寄与}$$

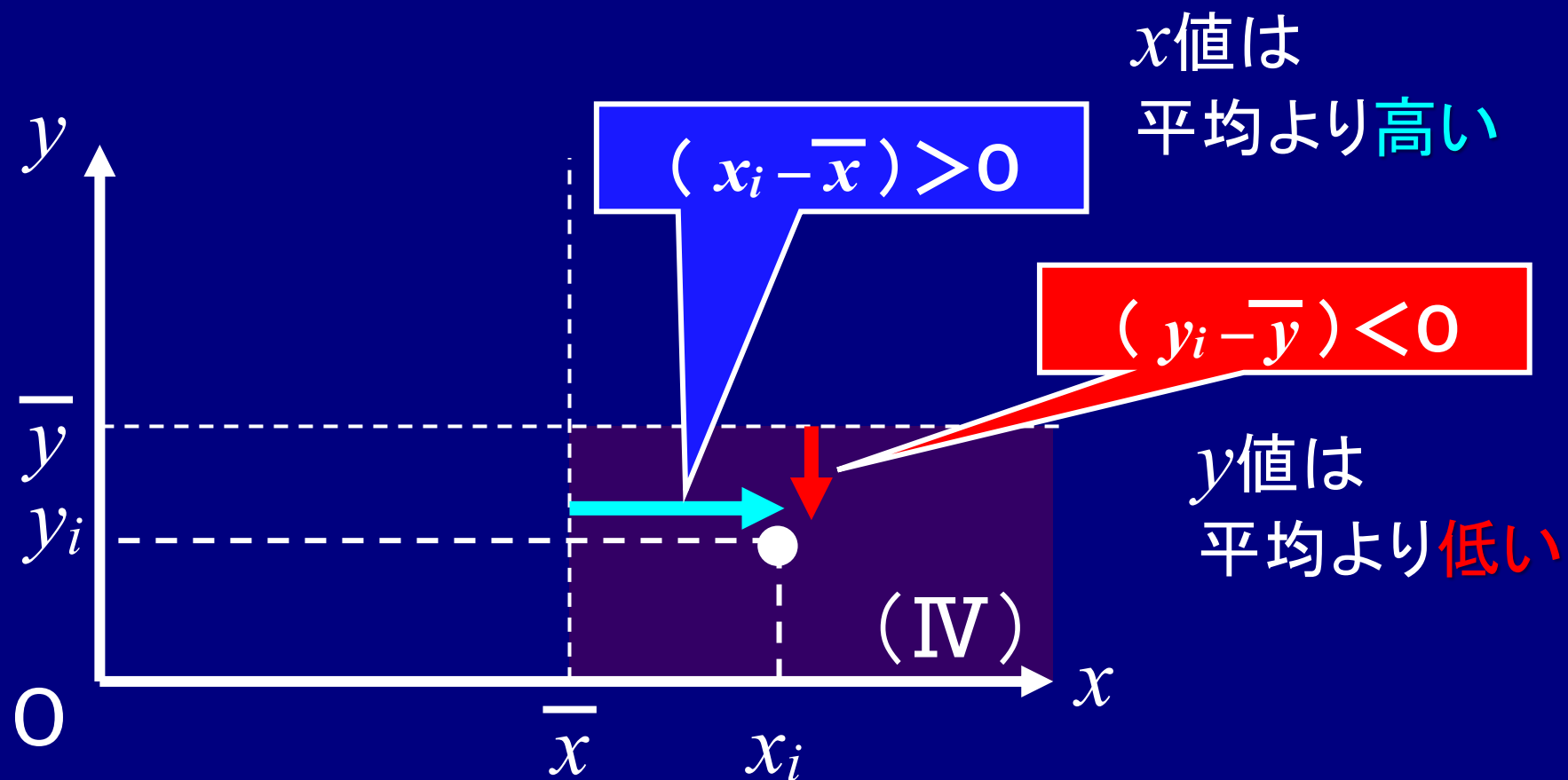
$$r_{xy} = \frac{\sigma_x \cdot \sigma_y}{\sigma_x \cdot \sigma_y}$$

(マイナス) × (マイナス)
= (プラス)



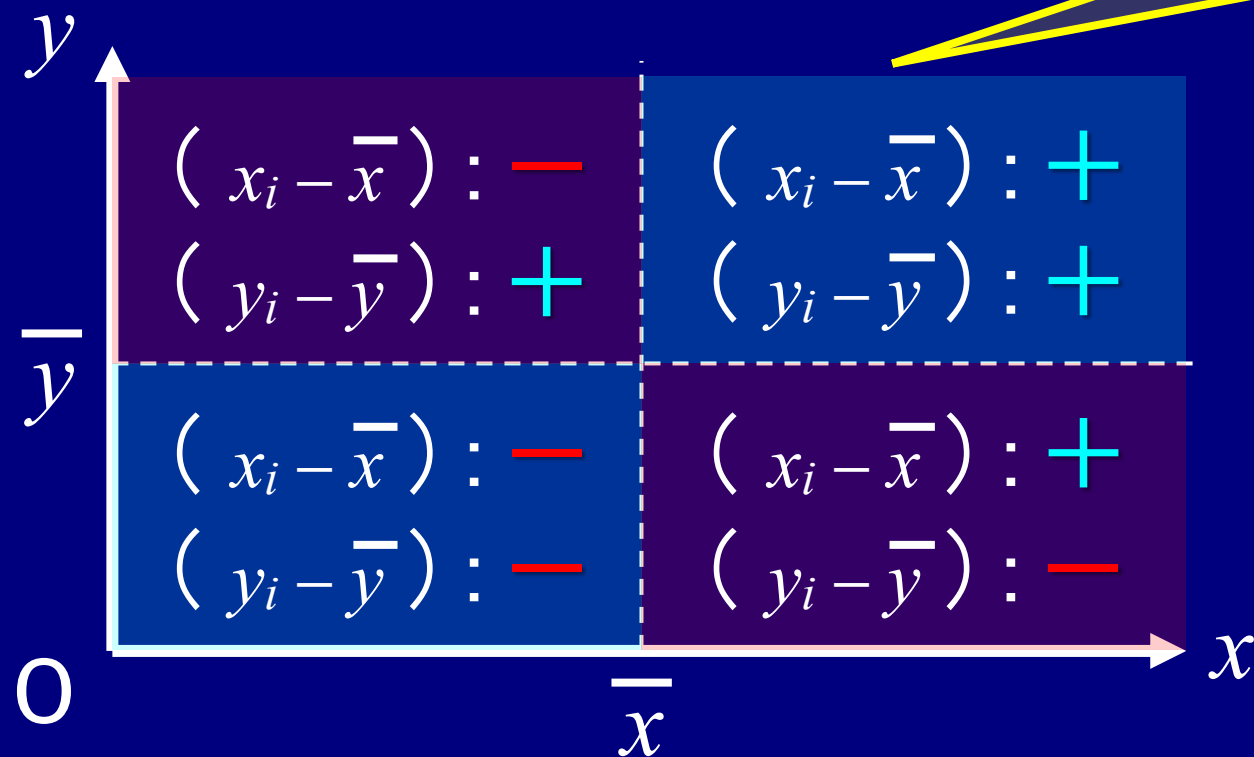
$$r_{xy} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sigma_x \cdot \sigma_y} < 0 \text{ (マイナス)の寄与}$$

(プラス) × (マイナス)
= (マイナス)



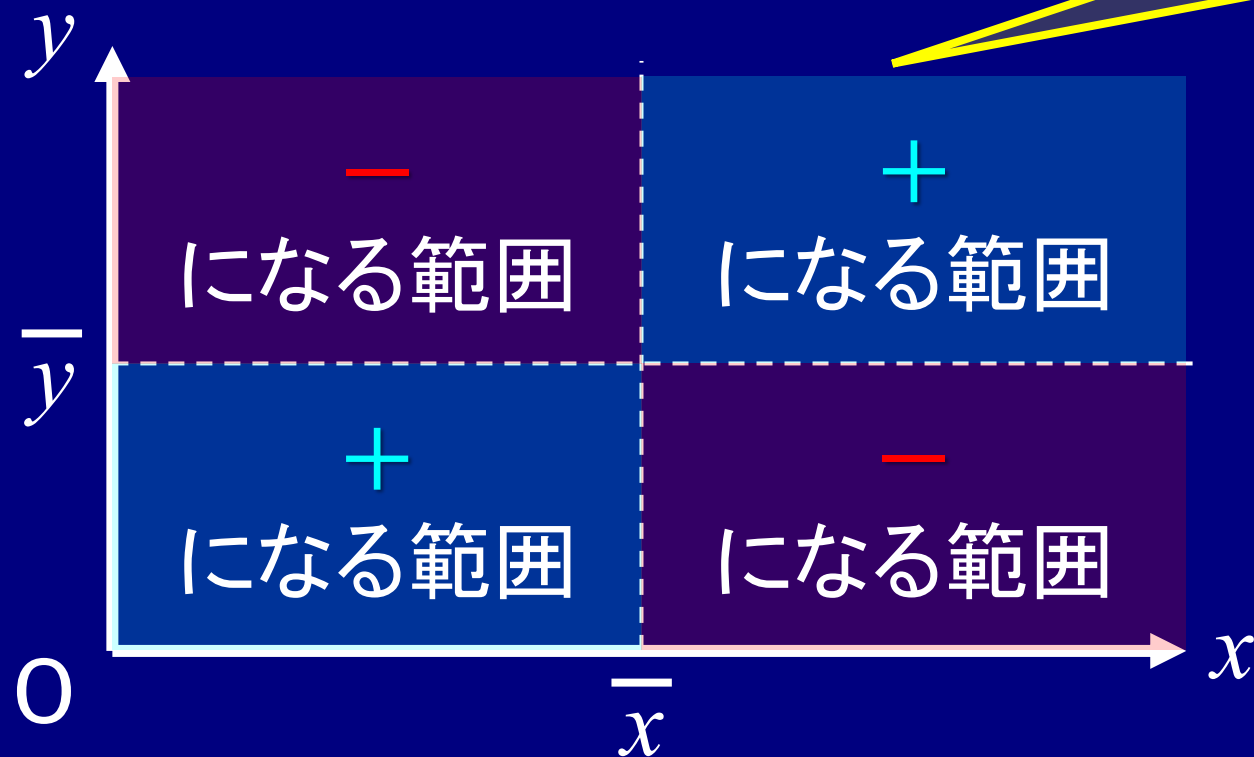
$$r_{xy} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sigma_x \cdot \sigma_y}$$

$(x_i - \bar{x})(y_i - \bar{y})$ が

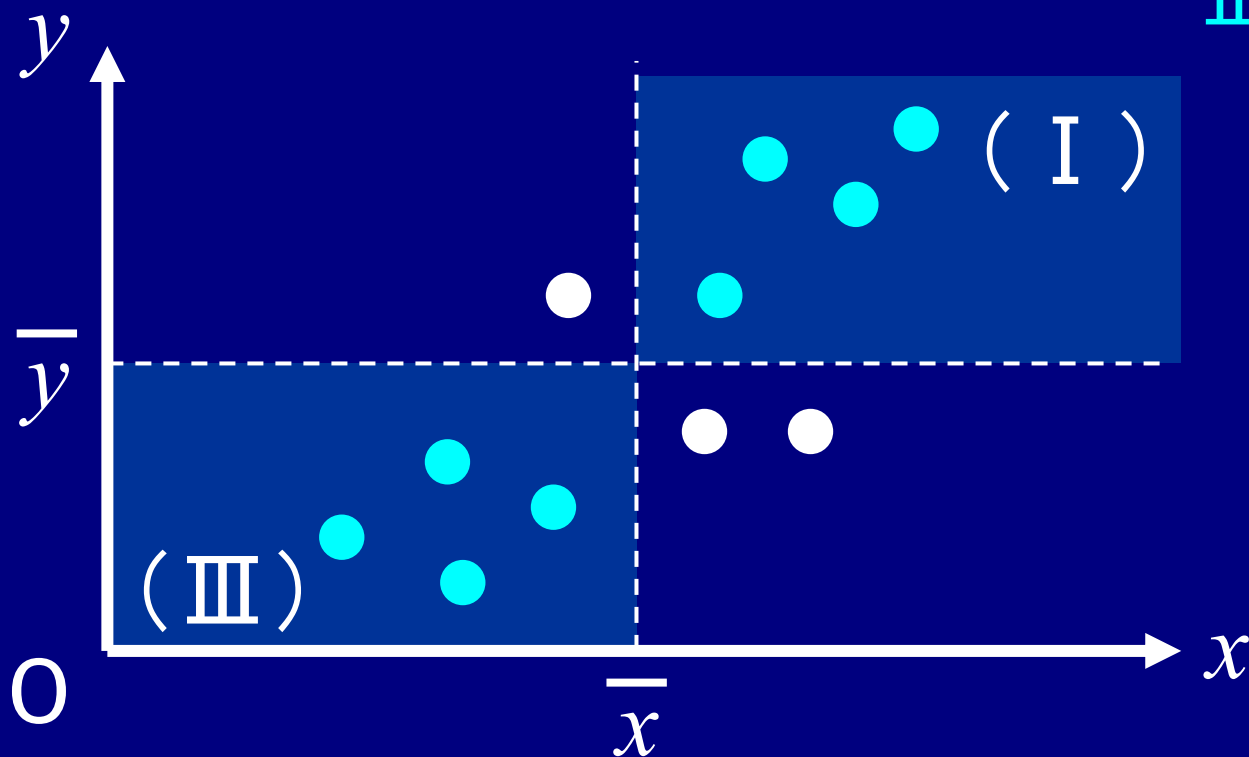


$$r_{xy} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sigma_x \cdot \sigma_y}$$

$(x_i - \bar{x})(y_i - \bar{y})$ が



$$r_{xy} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sigma_x \cdot \sigma_y} > 0$$

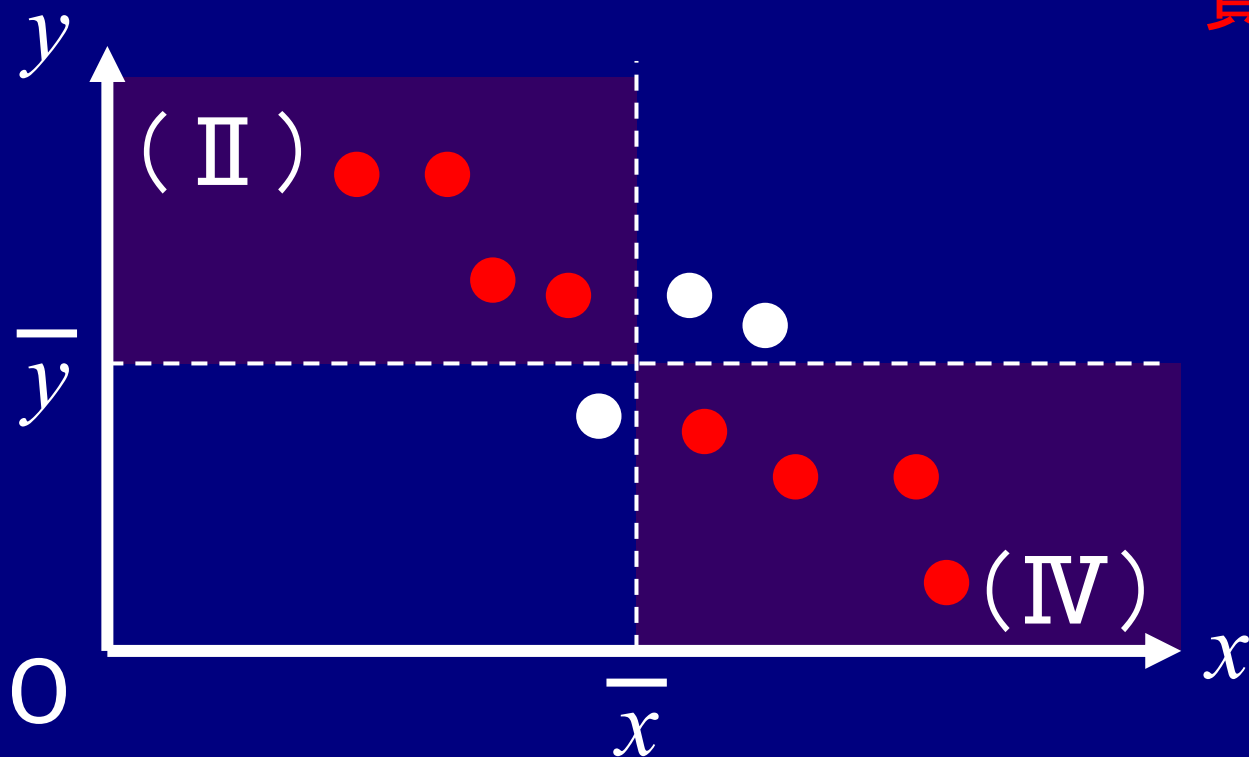


正の相関があるデータ群のときは、

(x 偏差)(y 偏差) が正になる
データが多いので、

$$r_{xy} > 0 \text{ となる}$$

$$r_{xy} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sigma_x \cdot \sigma_y} < 0$$



負の相関があるデータ群のときは、

(x 偏差)(y 偏差) が負になる
データが多いので、

$$r_{xy} < 0 \text{ となる}$$

データの標準化

- 標準化変量(Z値)
- 偏差値(T値)
- 変動係数

散布図

- 散布図
- 共分散
- 相関係数
- 正の相関・負の相関

回帰分析 (教科書未掲載)

- 回帰直線
- 最小二乗法

回帰分析

散布図を描いた際、

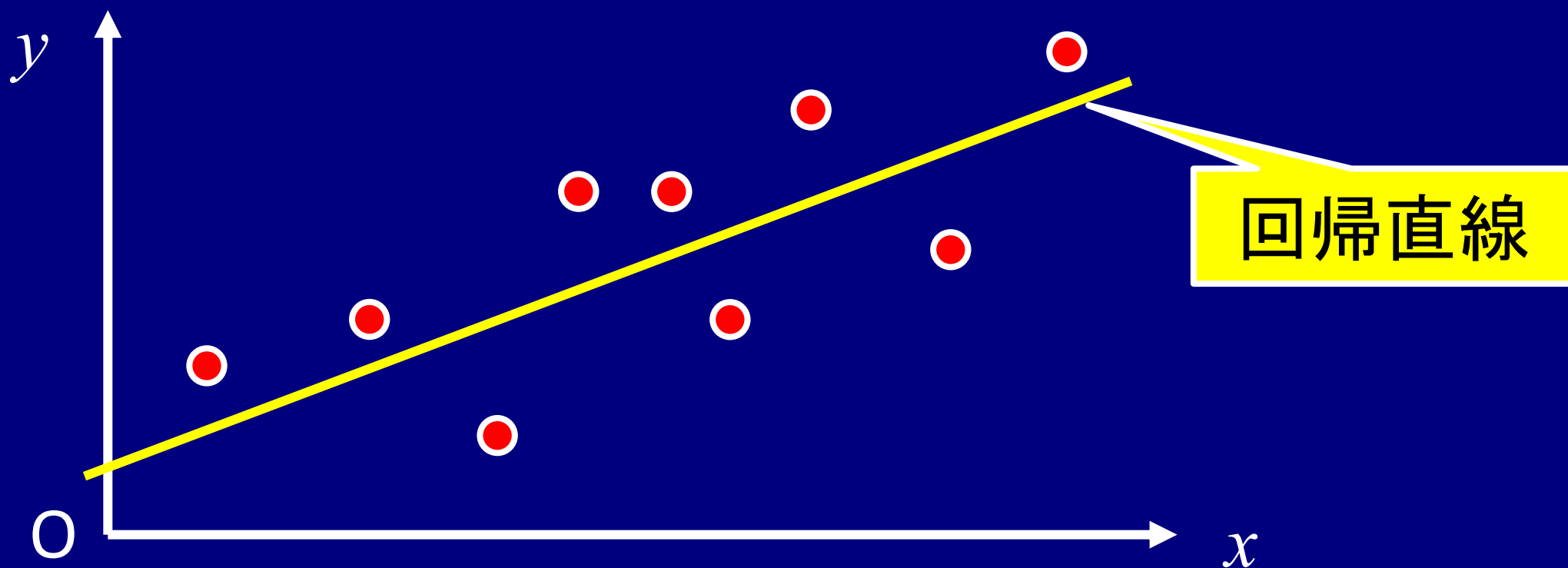
- x と y の2変数の間にどのような関係があるのかを推定し、
- x と y との真の関係を求めて、

分析する事。

ここでは回帰直線(1次式の関係)について述べます。

回帰分析

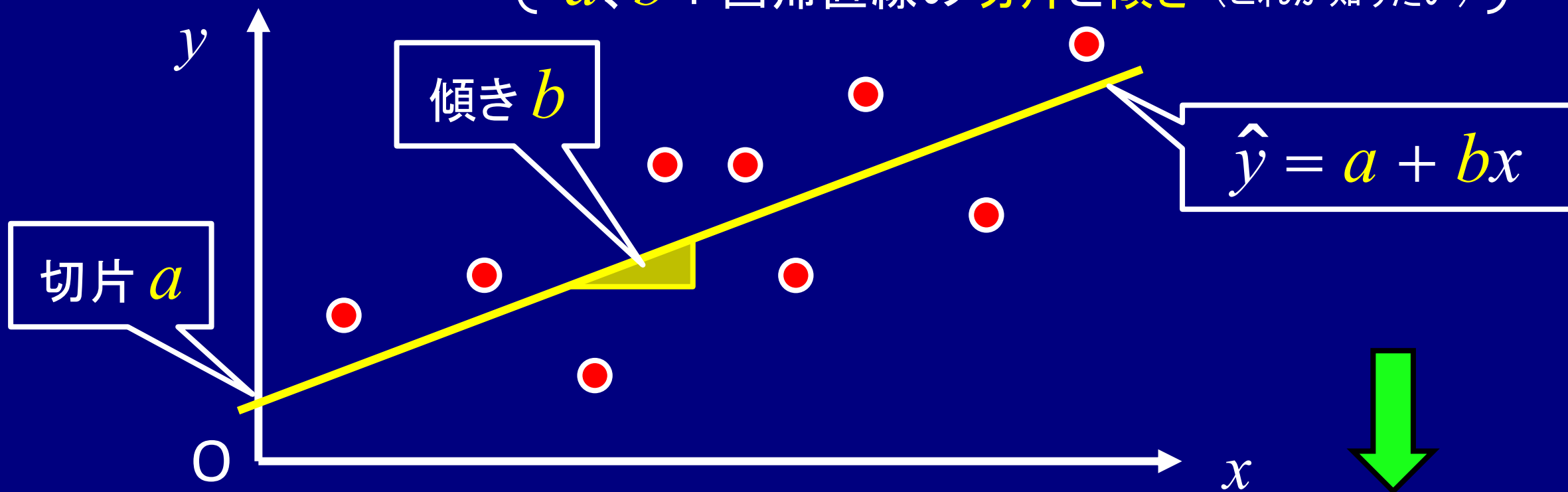
ある2変数のデータ組の散布図を描いたとき、
これらのデータが、直線的な相関にあるとしたとき、
直線的相関の元の直線が回帰直線となる



回帰直線の求め方(最小二乗法)

求めたい回帰直線を $\hat{y} = a + bx$ とすると、

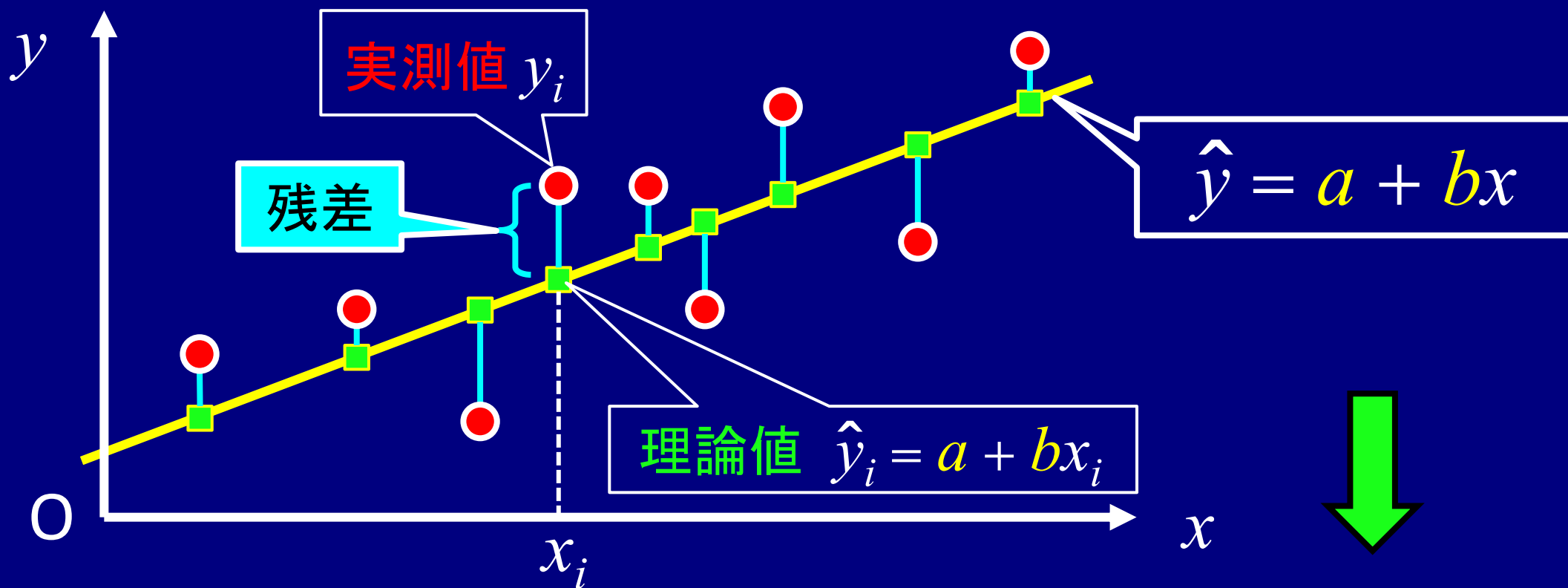
\hat{y} : 回帰直線上の理論値
 a, b : 回帰直線の切片と傾き (これが知りたい)



回帰直線の求め方(最小二乗法)

残差ができるだけ小さくなるような回帰直線の「切片 a 」「傾き b 」を求めるために、「残差の2乗和」を計算し……

$$D = \sum_{i=1}^N \{y_i - \hat{y}_i\}^2 = \sum_{i=1}^N \{y_i - (a + bx_i)\}^2$$



回帰直線の求め方(最小二乗法)

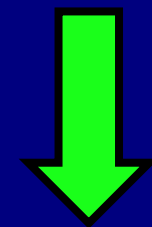
残差ができるだけ小さくなるような回帰直線の「切片 a 」「傾き b 」を求めるために、
「残差の2乗和」を計算し……

$$D = \sum_{i=1}^N \{y_i - \hat{y}_i\}^2 = \sum_{i=1}^N \{y_i - (a + bx_i)\}^2$$

「残差の2乗和 D 」が最小となる a と b を見つけるために、 D を a と b で偏微分し、

$$\frac{\partial D}{\partial a} = 0 \quad \frac{\partial D}{\partial b} = 0$$

によって、 D が極小となる a と b を求めると…… (大学理系レベルなので、計算結果だけを見ればよい。)



(計算法) ※わかる人だけ参考にすればよい

$$D = \sum \{ y_i - (a + bx_i) \}^2$$

Dを a と b で偏微分して、Dが極小となる所を調べると、

$$\frac{\partial D}{\partial a} = \sum 2\{ y_i - (a + bx_i) \} \cdot (-1) = 0 \text{ より}$$

$$\sum \{ y_i - a - bx_i \} = 0$$

$$\sum y_i - \sum a - \sum bx_i = 0$$

$$\sum y_i - Na - b \sum x_i = 0 \quad \dots\dots ①$$

$$\frac{\partial D}{\partial b} = \sum 2\{ y_i - (a + bx_i) \} \cdot (-x_i) = 0 \text{ より}$$

$$\sum \{ x_i y_i - ax_i + bx_i^2 \} = 0$$

$$\sum x_i y_i - a \sum x_i - b \sum x_i^2 = 0 \quad \dots\dots ②$$

$$① \text{より、} a = \frac{1}{N} \sum y_i - \frac{1}{N} b \sum x_i \quad \dots\dots ③$$

③を、②に代入すると、

$$\sum x_i y_i - \left\{ \frac{1}{N} \sum y_i - \frac{1}{N} b \sum x_i \right\} \sum x_i - b \sum x_i^2 = 0$$

$$N \sum x_i y_i - \left\{ \sum y_i - b \sum x_i \right\} \sum x_i - b N \sum x_i^2 = 0$$

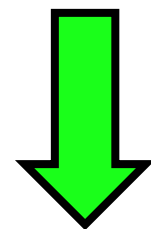
$$N \sum x_i y_i - \left(\sum x_i \right) \left(\sum y_i \right) = b \left\{ N \sum x_i^2 - \left(\sum x_i \right)^2 \right\}$$

∴

$$b = \frac{N \sum x_i y_i - \left(\sum x_i \right) \left(\sum y_i \right)}{N \sum x_i^2 - \left(\sum x_i \right)^2}$$

a は、③より、

$$a = \bar{y} - b \bar{x}$$



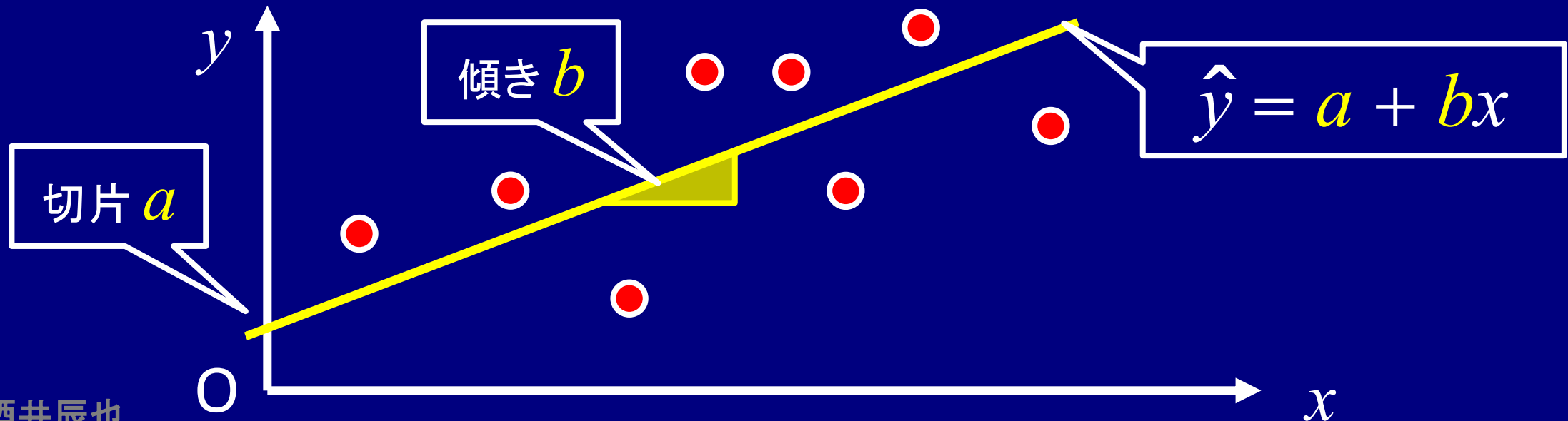
回帰直線の求め方(最小二乗法)

$$b = \frac{N \sum_{i=1}^N x_i y_i - \left[\sum_{i=1}^N x_i \right] \left[\sum_{i=1}^N y_i \right]}{N \sum_{i=1}^N x_i^2 - \left[\sum_{i=1}^N x_i \right]^2}$$

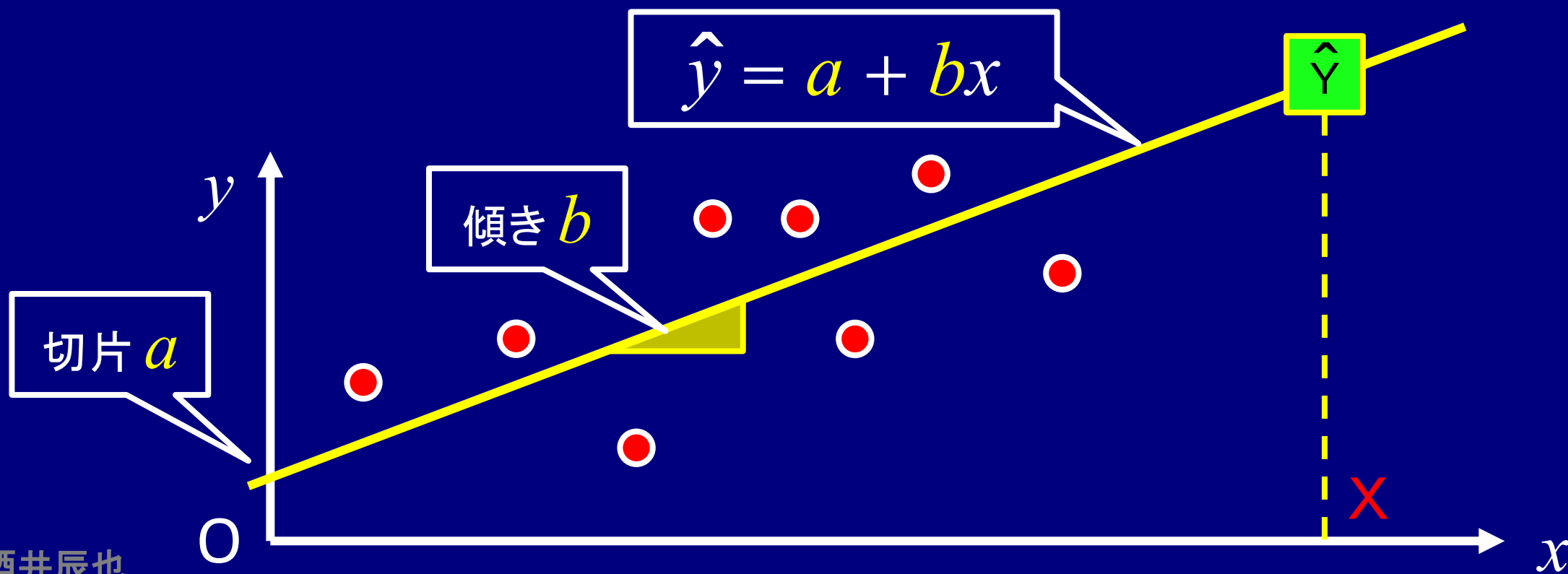
$$a = \bar{y} - b \bar{x}$$

※これらは通常、Excelや統計処理ソフトなどを用いて計算します。

(電卓計算は面倒)



回帰直線の切片 a と傾き b が判明すれば、
ある x 値に対する理論値 \hat{y} を類推する事ができる。



$$b = \frac{N \sum_{i=1}^N x_i y_i - \left[\sum_{i=1}^N x_i \right] \left[\sum_{i=1}^N y_i \right]}{N \sum_{i=1}^N x_i^2 - \left[\sum_{i=1}^N x_i \right]^2}$$

$$a = \bar{y} - b \bar{x}$$

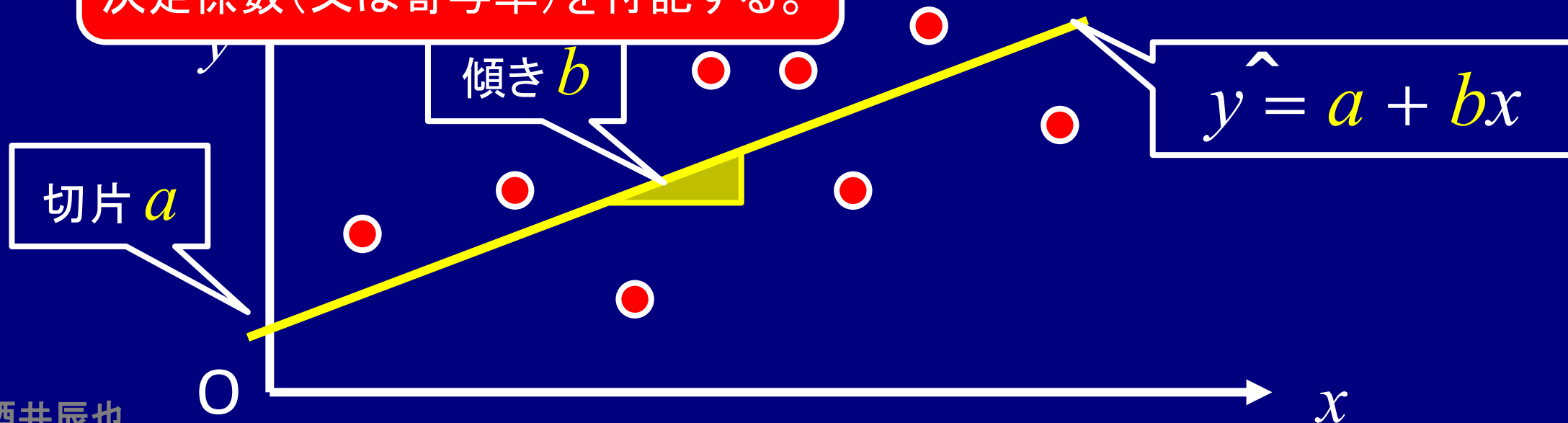
決定係数 (0~1)

$$r^2 = (\text{相関係数})^2$$

寄与率

$$r^2 \times 100 (\%)$$

回帰分析の結果には、
決定係数(又は寄与率)を付記する。



2変数を持つデータ(全N個)があるとき、

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$$

相関係数 (ピアソンの積率相関係数)

$$r_{xy} = \frac{\sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}$$

