01. データ解析の準備

京都大学　　加納　学

Division of Process Control & Process Systems Engineering
Department of Chemical Engineering, Kyoto University

manabu@cheme.kyoto-u.ac.jp
http://www-pse.cheme.kyoto-u.ac.jp/~kano/

あなたはどっち？

担当する製造装置において，製品品質特性を数回測定したところ，いずれも規格内であった．

- この製造装置は素晴らしい．
  規格品の製造が目的であり，その目的を達していることが測定データから確かめられたのだから，もう何も心配する必要はない．

  ➡ 天下太平派　たぶん，ハッピーな人生を送れる♪

- 規格外品を製造してしまうリスクを評価すべき．
  たまたま規格外にならなかっただけで，さらに測定を続ければ規格外品が検出されるかもしれない

  ➡ 油断大敵派　たぶん，人間ドックでボロボロ orz

ちょっと皆さんの知識を確認！

- 平均を知ってますか？
- 平均を計算したことはありますか？
- 分散を知ってますか？
- 分散を計算したことはありますか？

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

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

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

なぜ平均や分散を計算するの？

- 製品品質特性を数回測定しただけで，製造装置の特徴を完全に把握できるわけではない．
- 完全に把握するためには，測定を無限回繰り返し，その結果を評価する必要がある．
- しかし，無限回の測定というのは非現実的であり，我々としては，有限回の測定データに基づいて，無限回測定したときの結果を想像するしかない．

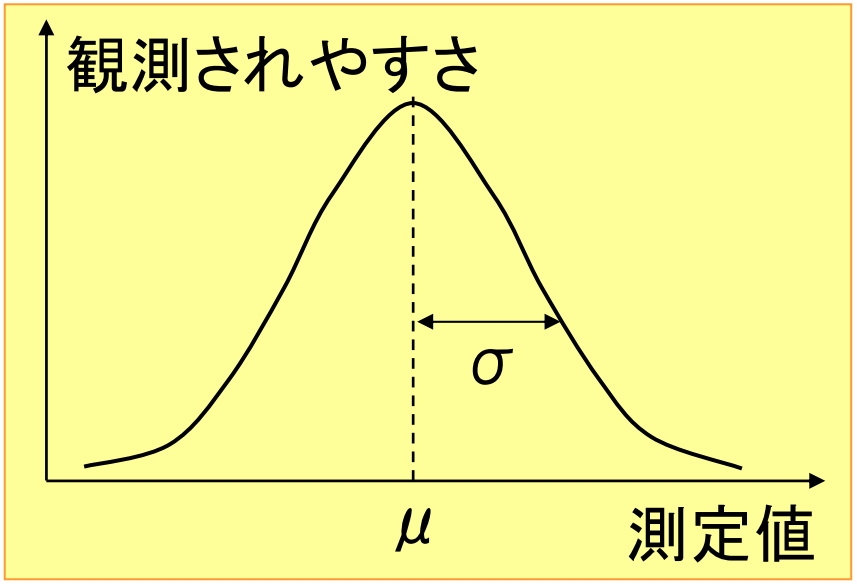
これがデータ解析

統計的推測

- **母集団**
  対象とする集団全体，つまり無限回の測定データ.

- **母集団分布**
  母集団のバラツキ具合.
  これを知ることが，データ解析の目的.

- **母数**
  母集団分布を特徴づける値
  - 母平均 μ
  - 母分散 σ^2
  - 母標準偏差 σ



- **統計的推測**
  データから母集団を特徴づける母数を推測する作業

統計的推測成功の秘訣

- 無作為抽出
  - データが母集団全体の特徴を反映していなければならない.
  - 無作為(ランダム) にデータを抽出する必要がある.

- ダメな例（日本人の服装の好みを調査しようとして・・・）
  - ヒョウ柄好きの大阪のおばちゃんばかりを調査対象に. . .
  - 銀座のブティックでショッピングしているセレブばかりを調査対象に. . .

分布の特徴を捉える

- 標本平均

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

- 標本分散

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

偏差

偏差平方和

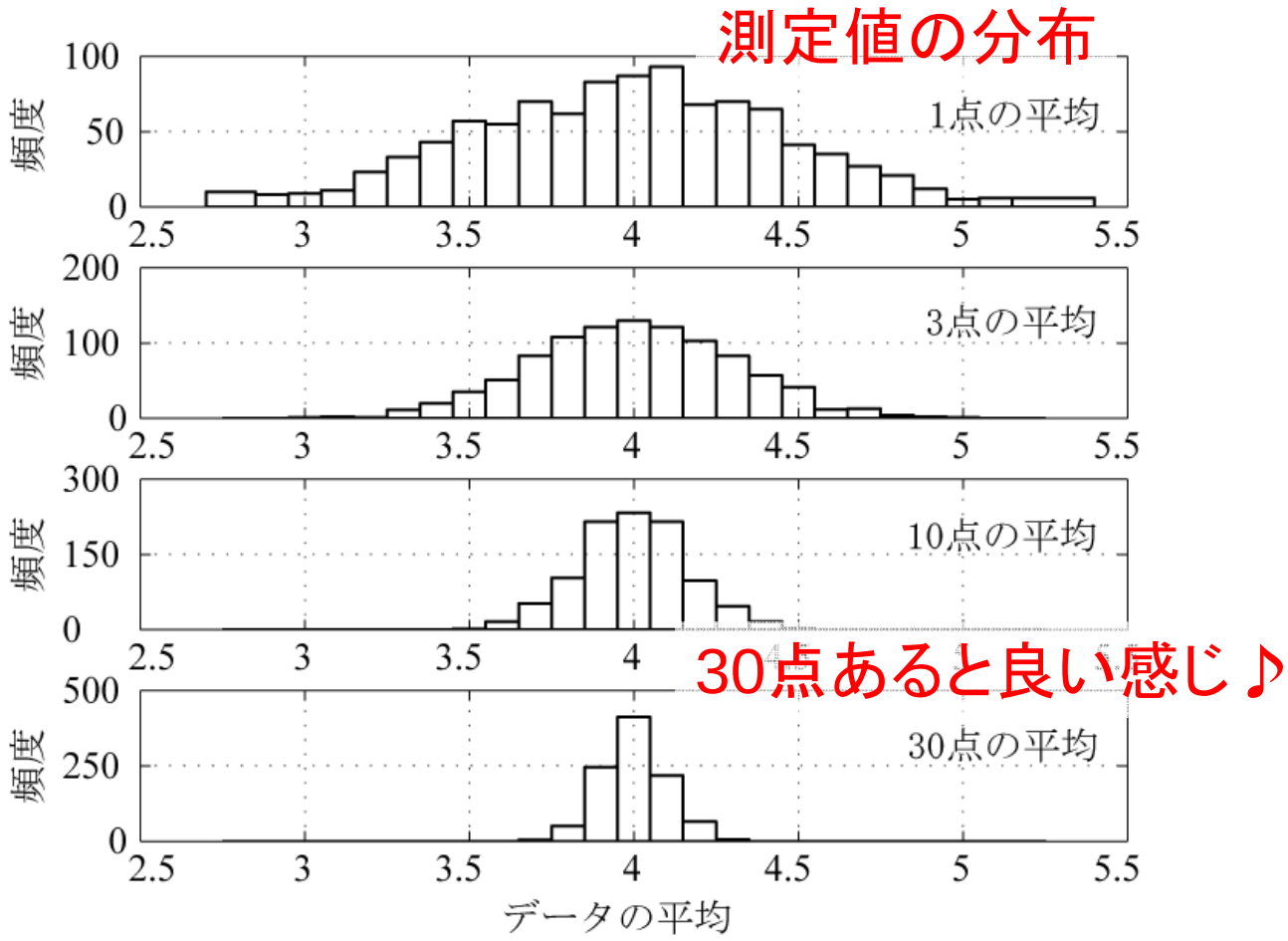
- 標本標準偏差

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

データが何個あれば，まともに計算できるでしょうか？

実際に平均を計算してみると

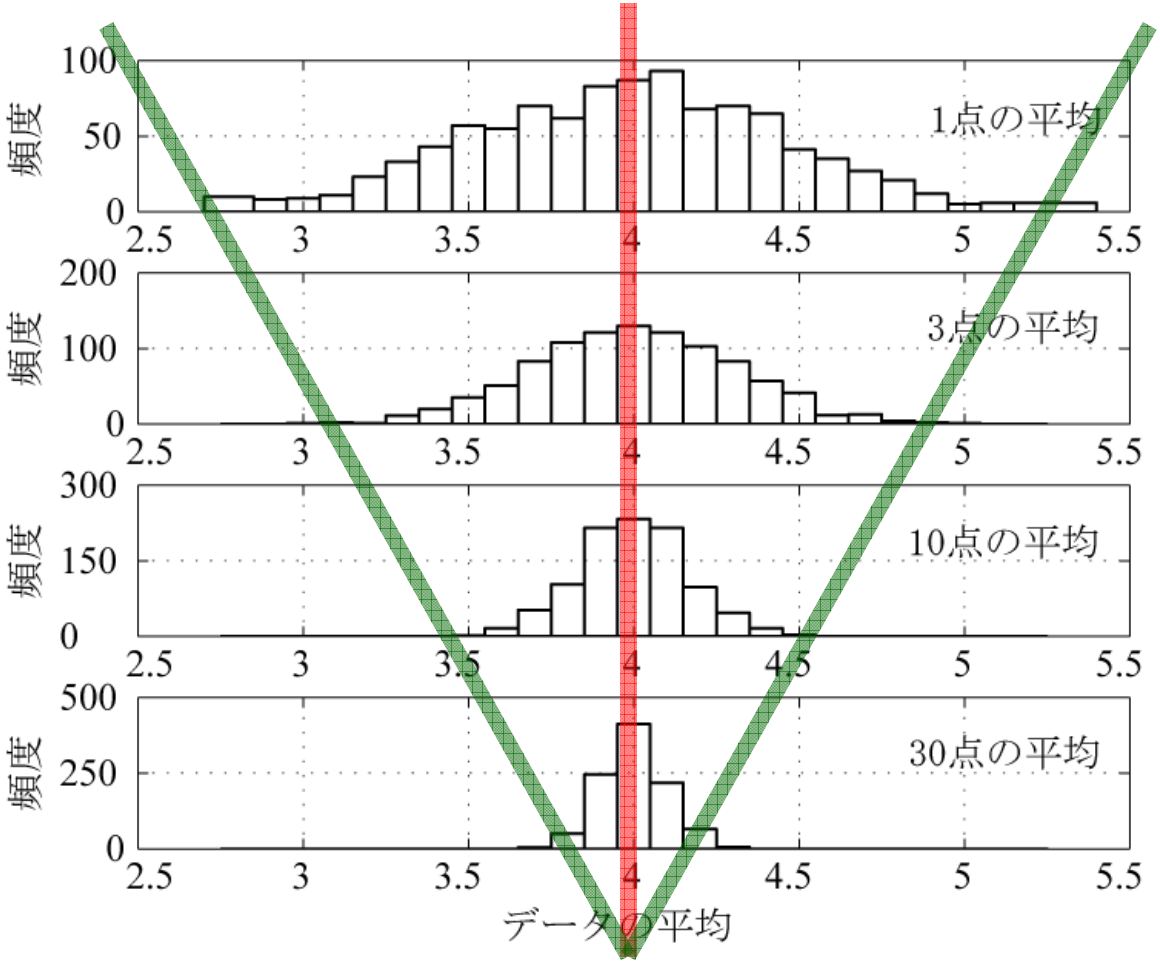
- 1000回の計算結果をヒストグラムにまとめた



不偏性と一致性

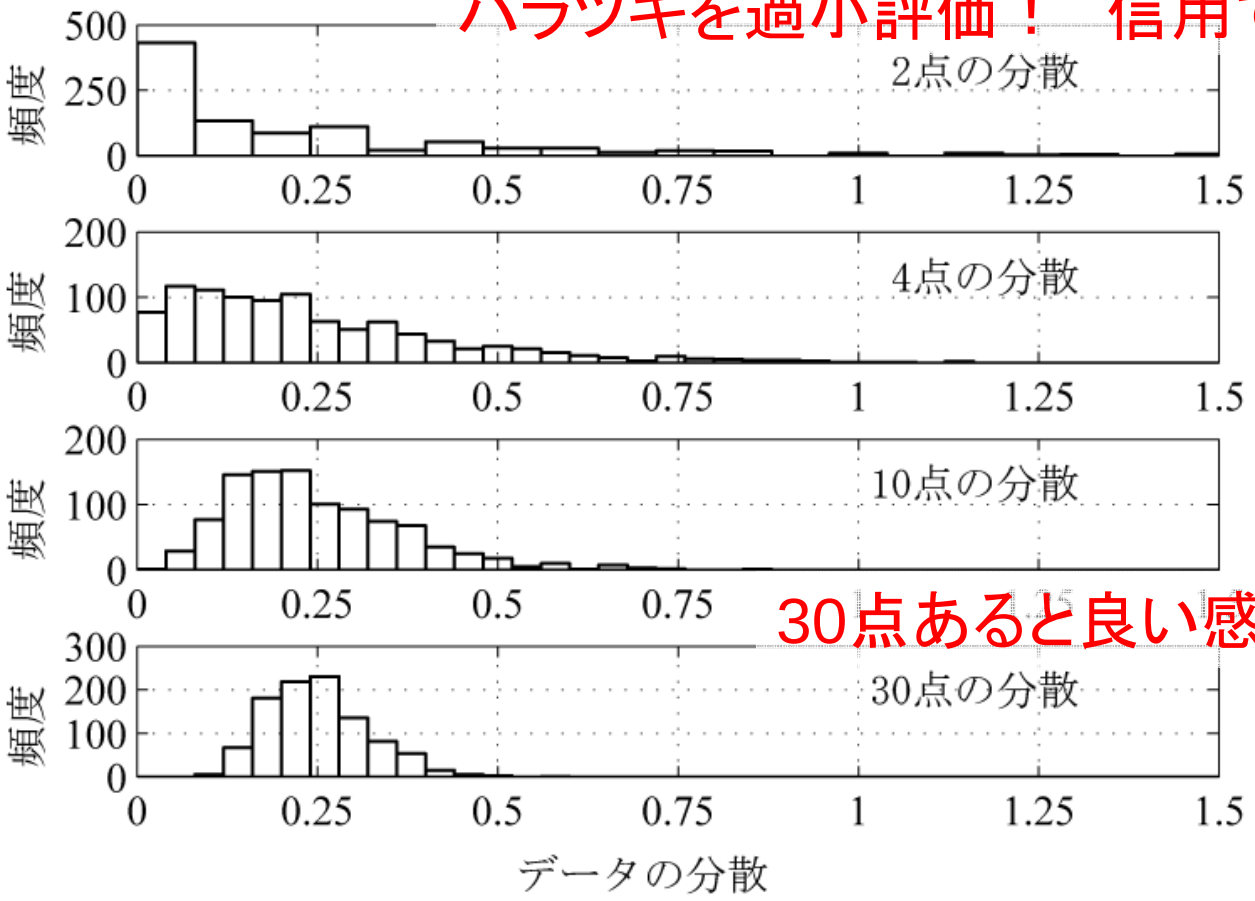
不偏性：データ数によらず，平均の平均が真値になる．

一致性：データ数が増えると平均のバラツキが小さくなる．



実際に分散を計算してみると

- 1000回の計算結果をヒストグラムにまとめた



分散をN-1で割って求める理由

- 自由度が $N-1$ だから
  - 平均を求める場合，N 個のデータはすべて自由に値を取ることができた．
  - 分散を求める場合，N 個の偏差すべてが自由に値を取ることはできない．なぜなら，偏差の総和は0でなければならないからである．つまり，$N-1$個の偏差の値が与えられると，N 個目の偏差の値は決まってしまう．そこに自由はない．
  - そこで，平均の N や分散の $N-1$ を自由度と呼び，平均化する際にはこの自由度で総和を割る．

- 不偏推定量にするため
  - $N-1$ で割ると不偏分散になる．
  - つまり，計算を無限回繰り返すと，その平均は母分散 σ^2 に一致する．

データの標準化

- どちらが重大事件でしょうか？
  - 平均温度より 3℃ 高い.
  - 平均重量より 3kg 重い.
- どちらが重大事件でしょうか？
  - 平均重量より 3kg 重い.
  - 平均重量より 3000g 重い.
- 標準化
  - 種類の異なる変数を同じ土俵にのせる.
  - 平均 0, 分散 1 に変換する.

$$u_n = \frac{x_n - \bar{x}}{s}$$

平均

標準偏差

変数間の相関を捉える

- 共分散

$$s_{xy}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})$$

- 相関係数
  - 標準化したデータの共分散に等しい.
  - $-1 \leq r_{xy} \leq 1$

$$r_{xy} = \frac{s_{xy}^2}{s_x s_y} = \frac{1}{N-1}\sum_{n=1}^{N} u_n v_n$$

標準化した x

標準化した y

$$\frac{1}{N-1}\sum_{n=1}^{N}(u_n \pm v_n)^2 = \frac{1}{N-1}\sum_{n=1}^{N}(u_n^2 \pm 2u_n v_n + v_n^2) = 2 \pm 2r_{xy} \geq 0$$

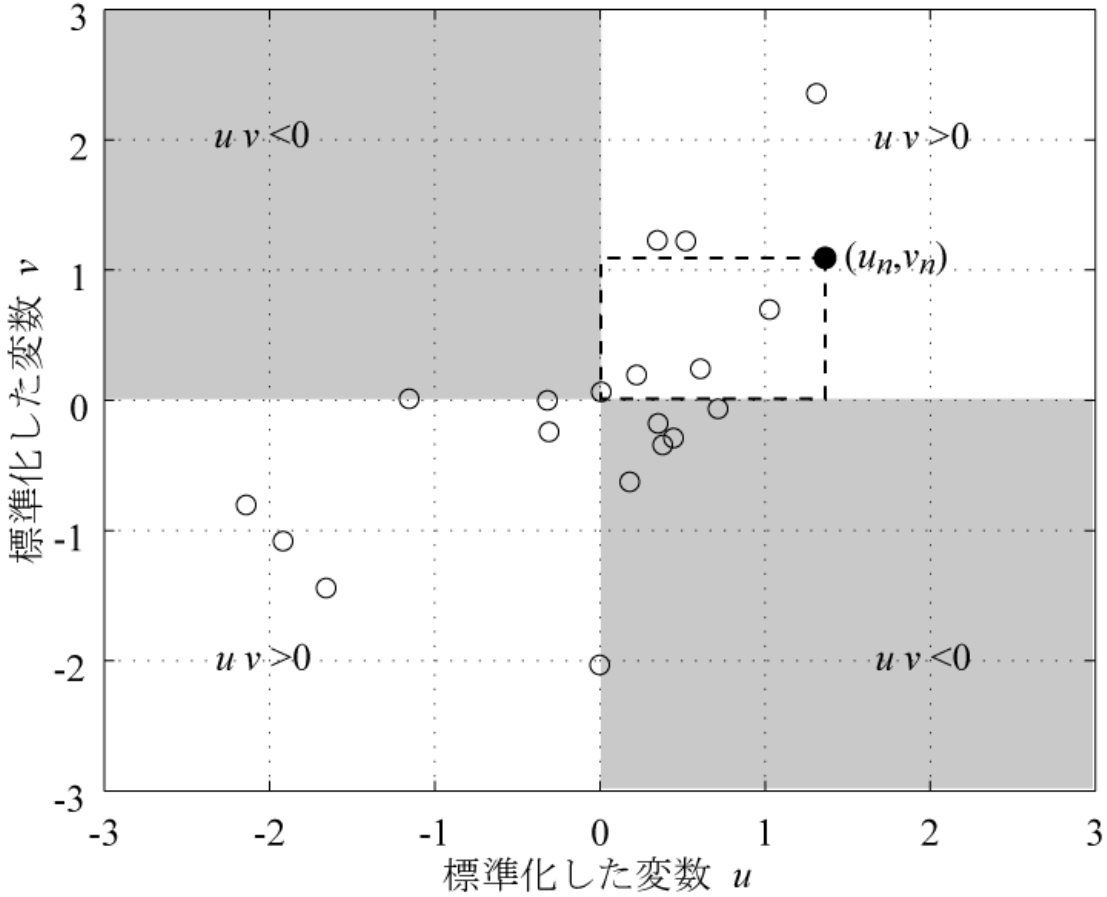
相関係数の解釈

- 相関係数
  - 1 に近ければ，正の相関がある．
  - −1 に近ければ，負の相関がある．
  - 0 に近ければ，無相関である．

r_{xy}	相関
$+0.7 \sim +1.0$	強い正の相関
$+0.4 \sim +0.7$	弱い正の相関
$-0.4 \sim +0.4$	無相関
$-0.7 \sim -0.4$	弱い負の相関
$-1.0 \sim -0.7$	強い負の相関

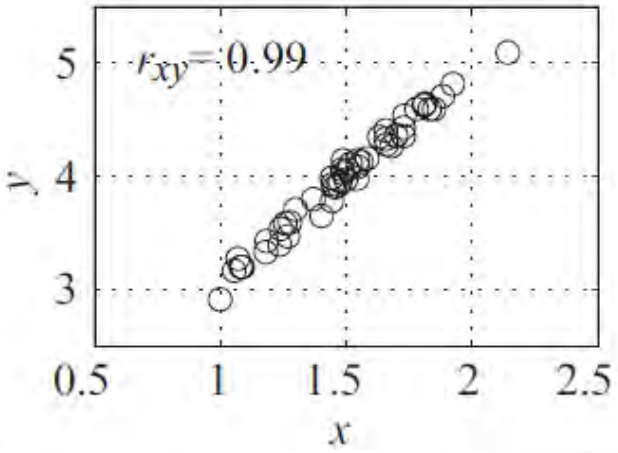
相関係数の図的解釈

- 全体としてデータが右上がりの傾向を持てば，$uv>0$ となる（白色の領域に存在する）データが多数派となり，その和は正，つまりデータは正の相関を持つ．

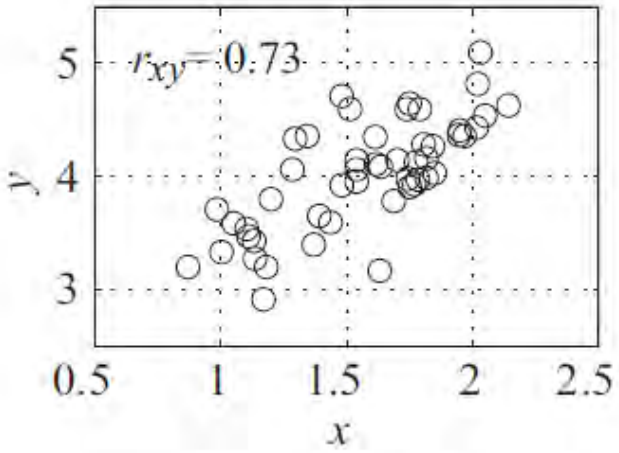


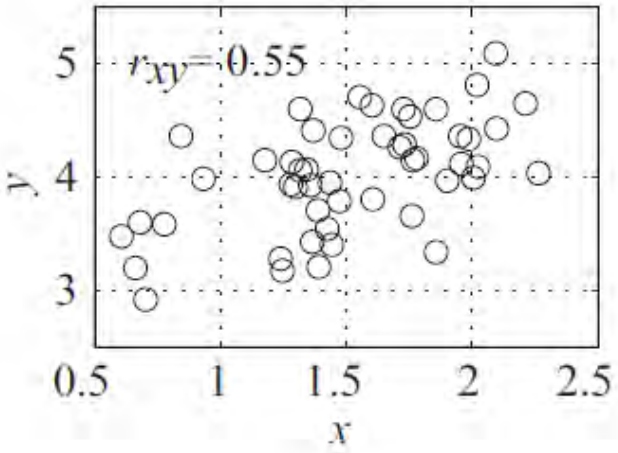
相関係数の具体例

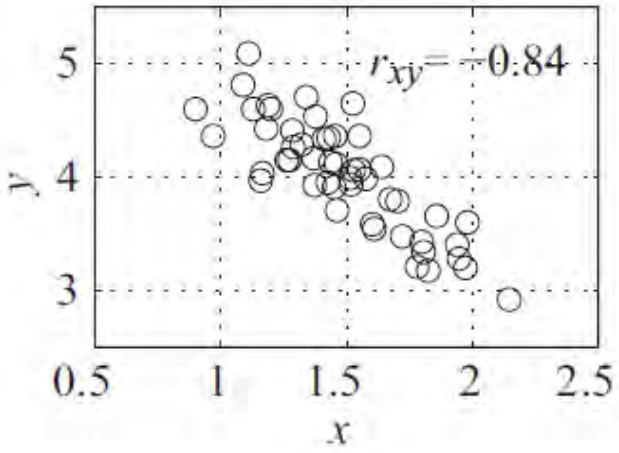
強い正の相関



やや強い正の相関





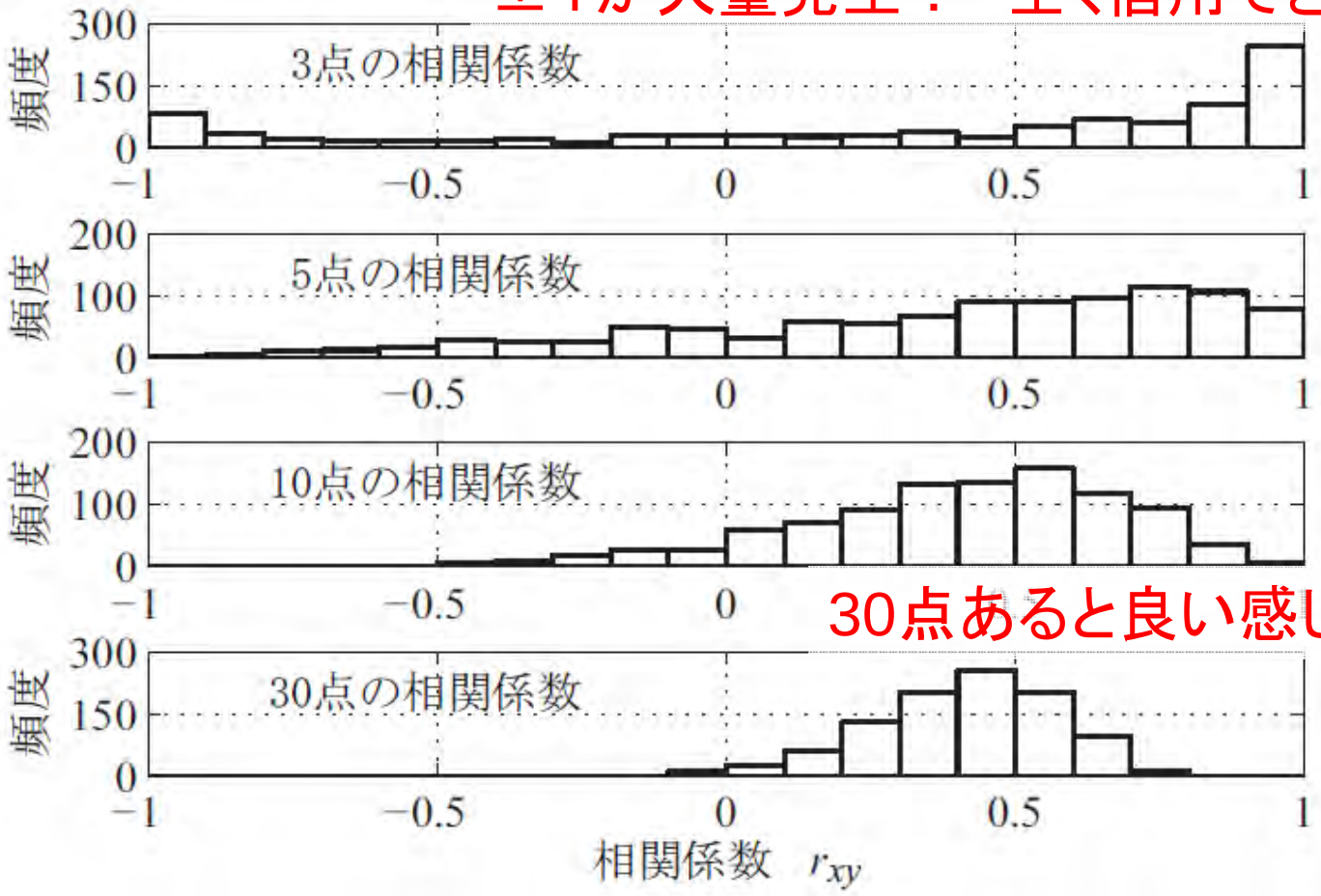


弱い正の相関

強い負の相関

実際に相関係数を計算してみると

- 1000回の計算結果をヒストグラムにまとめた



データが生まれる背景

- 確率

  事象の起こりやすさを定量的に表すものであり，
  事象 A の起こる確率を Pr(A) と書く.
  - 事象 A が絶対に起こらないなら Pr(A) = 0
  - 事象 A が必ず起こるなら Pr(A) = 1
  - 常に $0 \leq Pr(A) \leq 1$

コイン投げをした場合
Pr(表が上) = Pr(裏が上) = 0.5
Pr(裏と表が同時に上) = 0

確率変数と確率分布

- **確率変数**
  コインの裏表やサイコロの目のように，それぞれの事象が起こる確率が決まっている変数

- **確率分布**
  確率変数がそれぞれの値をとる確率を述べたもの

サイコロの目を x とすれば，x は確率変数であり，
その確率分布は

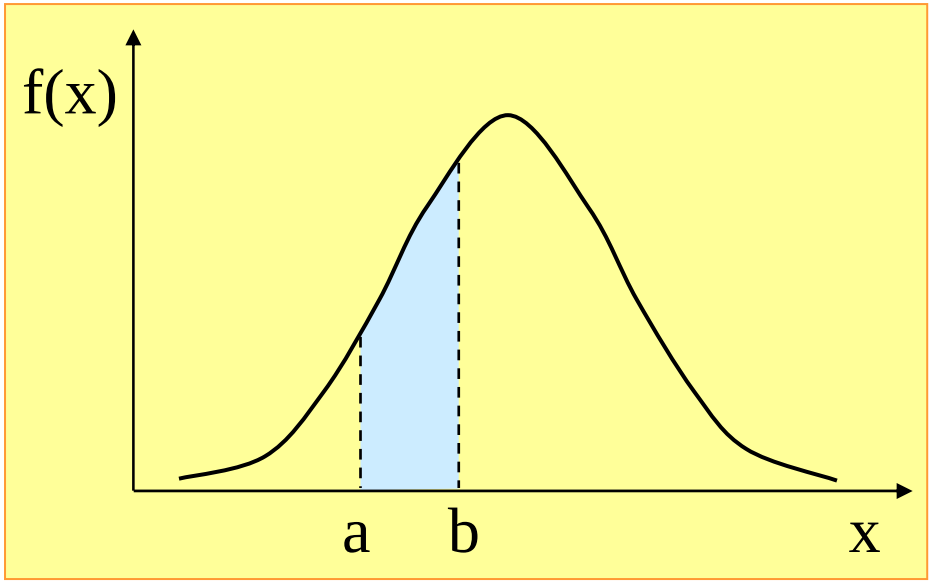
$$\Pr(x = i) = 1/6 \qquad (i = 1, 2, 3, 4, 5, 6)$$

確率密度関数

- **確率密度関数**
  確率変数 x の確率分布が

$$\Pr(a \leq x \leq b) = \int_a^b f(x)dx$$

で与えられるとき，$f(x)$ を
x の確率密度関数という.



$$f(x) \geq 0$$

ある区間内の値をとる確率は 0 以上

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

確率を全部あわせると 1

$$\Pr(x = a) = \int_a^a f(x)dx = 0$$

値 a をとる確率 0

同時確率密度関数

- **同時確率密度関数**
  確率変数 x, y の同時確率分布が

$$\Pr(a \leq x \leq b, c \leq y \leq d) = \int_c^d \int_a^b f(x,y) dx dy$$

で与えられるとき，$f(x, y)$ を2次元確率変数 (x, y) の同時確率密度関数という.

$$f(x,y) \geq 0$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) dx dy = 1$$

独立

- 独立

$$f(x,y) = f(x)f(y)$$

が成り立つとき，2次元確率変数 (x, y) は独立である．

金銀2つのコインを投げた場合

$$\Pr(\text{どちらも表}) = \Pr(\text{金貨が表})\Pr(\text{銀貨が表}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$\Pr(\text{金貨が表，銀貨が裏}) = \Pr(\text{金貨が表})\Pr(\text{銀貨が裏}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$\Pr(\text{金貨が裏，銀貨が表}) = \Pr(\text{金貨が裏})\Pr(\text{銀貨が表}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$\Pr(\text{どちらも裏}) = \Pr(\text{金貨が裏})\Pr(\text{銀貨が裏}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

独立と無相関

独立　≠　無相関

無相関は「相関がない」というだけであって，
「何の関係もない」ことを意味しない．

関係はあるが無相関という状況は起こりうる．

独立とは，正真正銘，何の関係もないということである．

無限回繰り返すと見えるもの

- 期待値
  確率分布を重みとする,
  確率変数がとる値の重み付き平均

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx$$

$$E(ax + b) = aE(x) + b$$

$$E(x + y) = E(x) + E(y)$$

期待値は確率変数の実現値を無限回サンプリングしたときの平均.

標本を無限回抽出すれば, 母集団の特徴を完全に捉えられるので, 確率変数の期待値は母平均に等しい.

期待値と母数

母平均　$\mu_x = E(x)$

母分散　$\sigma_x^2 = V(x) = E((x-\mu_x)^2)$

母共分散　$\sigma_{xy}^2 = C(x,y) = E((x-\mu_x)(y-\mu_y))$

分散や共分散の性質

$$V(x) = E(x^2) - E(x)^2$$

$$V(ax+b) = a^2 V(x)$$

$$V(x+y) = V(x) + V(y) + 2C(x,y)$$

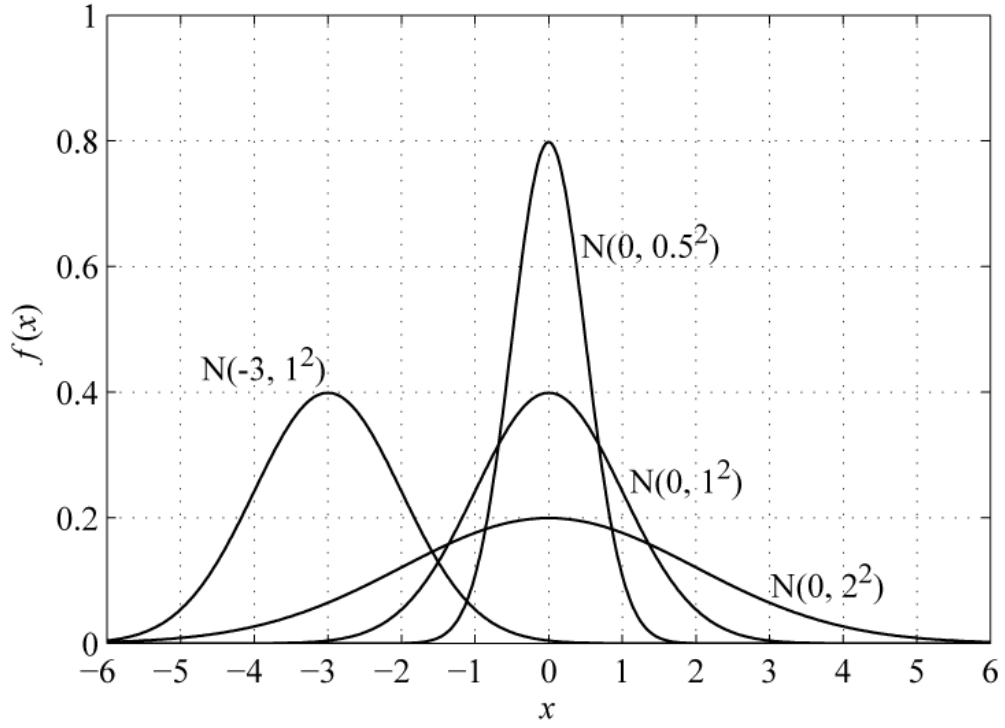
$$C(x,y) = E(xy) - E(x)E(y)$$

最も重要な確率分布

- 正規分布
  - 確率密度関数

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

  - 平均 μ, 分散 σ^2 の正規分布を $N(\mu, \sigma^2)$ と表す.



正規分布の性質

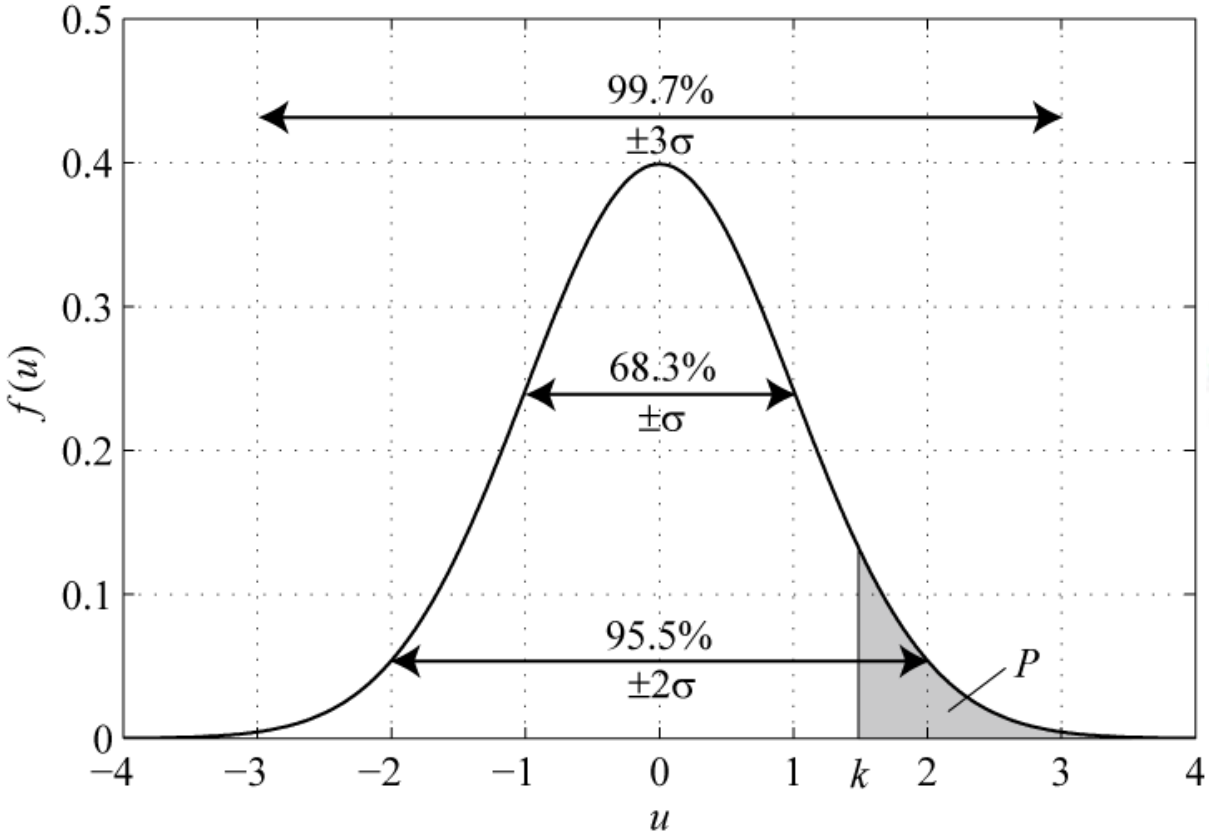
- 確率変数 x_n が互いに独立あるいは無相関にそれぞれ $\mathrm{N}(\mu_n, \sigma_n^2)$ に従うなら,

$$\sum_{n=1}^{N} a_n x_n \sim \mathrm{N}\left(\sum_{n=1}^{N} a_n \mu_n, \sum_{n=1}^{N} a_n^2 \sigma_n^2\right)$$

正規分布に従う確率変数の和は正規分布に従う.

標準正規分布

- 標準正規分布
  - 平均 0, 分散 1 の正規分布 $\mathrm{N}(0, 1^2)$
  - 確率変数 x が正規分布に従うとき,
    x を標準化した変数は標準正規分布に従う.



$$u = \frac{x - \mu}{\sigma} \sim \mathrm{N}(0, 1^2)$$

具体的な確率の計算

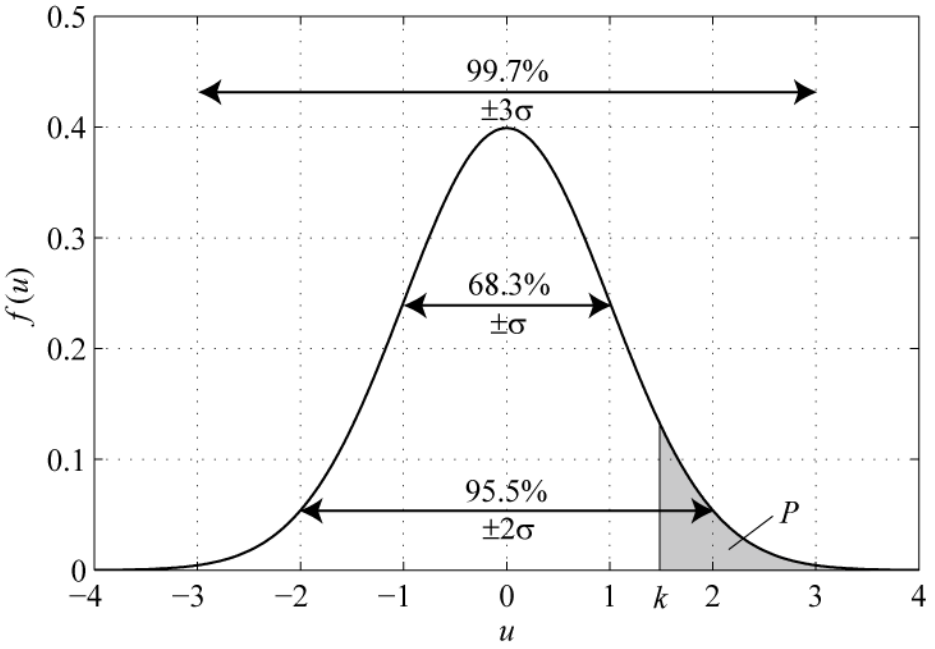
- 標準正規分布の上側確率
  標準化された確率変数 u が k 以上の値をとる確率

$$P = \Pr(u \geq k) = \int_k^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du$$

正規分布表

k	0	1.0	1.645	1.960
P	0.5000	0.1587	0.0500	0.0250

2.0	2.576	3.0	4.5
0.0228	0.0050	0.0013	0.000003



$$\Pr(-3 \leq u \leq 3) = 1 - 2 \times 0.0013 = 0.9974$$

なぜ正規分布が重要か？

- 中心極限定理
  - データ数 N が大きければ，母集団の確率分布がどのような形状をしていようとも，確率変数の和あるいは平均は正規分布に従う.
  - 確率変数 x_n が互いに独立で，$E(x_n) = \mu, V(x_n) = \sigma^2$ であるとき，N が大きければ，近似的に

$$u = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}} \sim \mathrm{N}(0, 1^2)$$

が成り立つ.