統計（1）

河野能知

2015.02～03

資料：
http://hpx.phys.ocha.ac.jp/wiki/Lec2015ParticleDetectors

内容

物理実験に必要な確率

- 測定と誤差
  - 物理量の測定値を無限の精度で得られることはない
  - 極端な言い方をすれば、誤差を評価していない測定値に意味は無い
  - 統計誤差
    - 同じ環境で複数回測定を行っても毎回少しずつ違う結果が得られる
  - 系統誤差
    - 実験に用いる装置の精度や解析手法における近似等
  - 誤差: 通常は、何度も測定を行うと、測定値の分布はある中心値の周りでガウス分布となる。その分布の標準偏差(σ)を誤差と呼ぶ
- 量子力学に現れる確率
  - 量子力学では、物理現象自体が確率によっている
- 実験データの解析
  - 多数の検出器を用いた場合のチャンネル間の相関の理解

確率分布

- 確率変数X: 確率的にいろいろな値を取り得る変数Xをこのように呼ぶ
- 確率分布$P(X)$:
  - 離散的な場合: 有限個の可能性X_1、...、X_nがあり、それぞれが起こる確率が$P(X)$で与えられる
  - 連続的な場合: Xは連続的な値を取り得て、Xがxから$x+dx$までの値を取る確率が$P(x)dx$で与えられる
- 平均$E(X)$、$\bar{X}$: $\sum_{i=1}^{n} X_i P(X_i)$、または$\int_{-\infty}^{\infty} xP(x)dx$
- 分散V(X): $V(X) = E\left((X-\bar{X})^2\right) = E(X^2) - E(X)^2$
- 標準偏差σ: ガウス分布の幅。その場合、$\sigma^2 = V(X)$

<u>離散的な確率分布</u>
- サイコロを振って、1～6の目が出る確率
- 放射性元素から崩壊して出てくるβ線の数の分布(一般に個数の分布)

<u>連続的な分布</u>
- ガウス分布、指数分布、χ^2分布

中心極限定理

確率変数がどんな確率分布に従っていても、複数回の独立な試行の平均
$$\overline{X_n} = (X_1 + X_2 + \cdots + X_n)/n$$
は、$n \to \infty$でガウス分布に従う。元の確率分布の平均がμ、分散がσ^2であるとき、n回の平均の分布は、平均μ、分散$n\sigma^2$のガウス分布になる。

サイコロをN回振った時の目の合計

1回目に出た目	1回目と2回目の合計										
1	2	3	4	5	6	7					
2		3	4	5	6	7	8				
3			4	5	6	7	8	9			
4				5	6	7	8	9	10		
5					6	7	8	9	10	11	
6						7	8	9	10	11	12
頻度	1	2	3	4	5	6	5	4	3	2	1

中心極限定理の照明

確率変数がどんな確率分布に従っていても、複数回の独立な試行の平均
$$\bar{X}_n = (X_1 + X_2 + \cdots + X_n)/n$$
は、$n \to \infty$でガウス分布に従う。

確率分布の**特性関数**: 確率変数をXとすると$\phi(\theta) = E\left(e^{i\theta X}\right)$

- 連続的な確率分布に対しては、これは元の分布のフーリエ変換となる
- $\phi(\theta) = E\left(e^{i\theta X}\right) = \int e^{i\theta x} P(x) dx$
- 逆変換は、$P(x) = \frac{1}{2\pi} \int \phi(\theta) e^{-i\theta x} d\theta$

X_iの平均をμ、分散をσ^2とすると、

$$\varphi_n(\theta) = E\left(e^{i\theta(X_1+\cdots+X_n)/\sqrt{n}}\right) = E\left(e^{\sqrt{n}\theta\mu} e^{i\theta(X_1+\cdots+X_n-n\mu)/\sqrt{n}}\right)$$
$$= E\left(e^{i\sqrt{n}\theta\mu}\right) E\left(e^{i\theta(X_1-\mu)/\sqrt{n}}\right) \cdots E\left(e^{i\theta(X_n-\mu)/\sqrt{n}}\right)$$

$(X_i - \mu)/\sqrt{n}$は、平均μ、分散$\sigma/\sqrt{n}$であるから、

$$\varphi_n(\theta) \cong e^{i\sqrt{n}\theta\mu} \left(1 - \frac{\sigma^2\theta^2}{2n}\right)^n \to e^{i\sqrt{n}\mu\theta - \frac{1}{2}\sigma^2\theta^2}$$

これは、ガウス分布$N(\sqrt{n}\mu, \sigma^2)$の特性関数である

測定と誤差

例として長さの測定を考える

- 測定値$x_1, x_2, \ldots\ldots, x_n$は、毎回ばらつくと思われる
- これに対して、その長さの「真の値」というものが存在する
- 測定： $x_1, x_2, \ldots\ldots, x_n$から真の値を推定すること

- 通常、誤差は平均値の周りにガウス分布するものと考えることが多い
  - 中心極限定理が根拠
- 誤差がガウス分布とは違う場合もある
  - 度数分布で数が小さい場合はポアソン分布になる

確率変数の関数

- 直接測定した量から計算して別の量を得ることがある
- $y = f(x_1, x_2, \cdots, x_n)$
- 例
  - 円の面積を求めるために、半径rを測定して$A = \pi r^2$と計算する
  - 平均身長を求めるために、1000人の身長データから平均値を求める
  - 2次元座標のデータが多数あり、それをフィットして関数のパラメータを計算する

誤差の伝播

$y = f(x_1, x_2, \cdots, x_n)$と計算するとき、各$x_i$に$\sigma_i$という誤差が存在した場合、$y$にはどれだけ誤差があるか

$\bar{y} = f(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n)$と仮定する（線形な関数の場合は正しい）
$x_i = \bar{x}_i + \Delta x_i$と平均値の周りで展開すると、

$$y = f(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i + \cdots$$

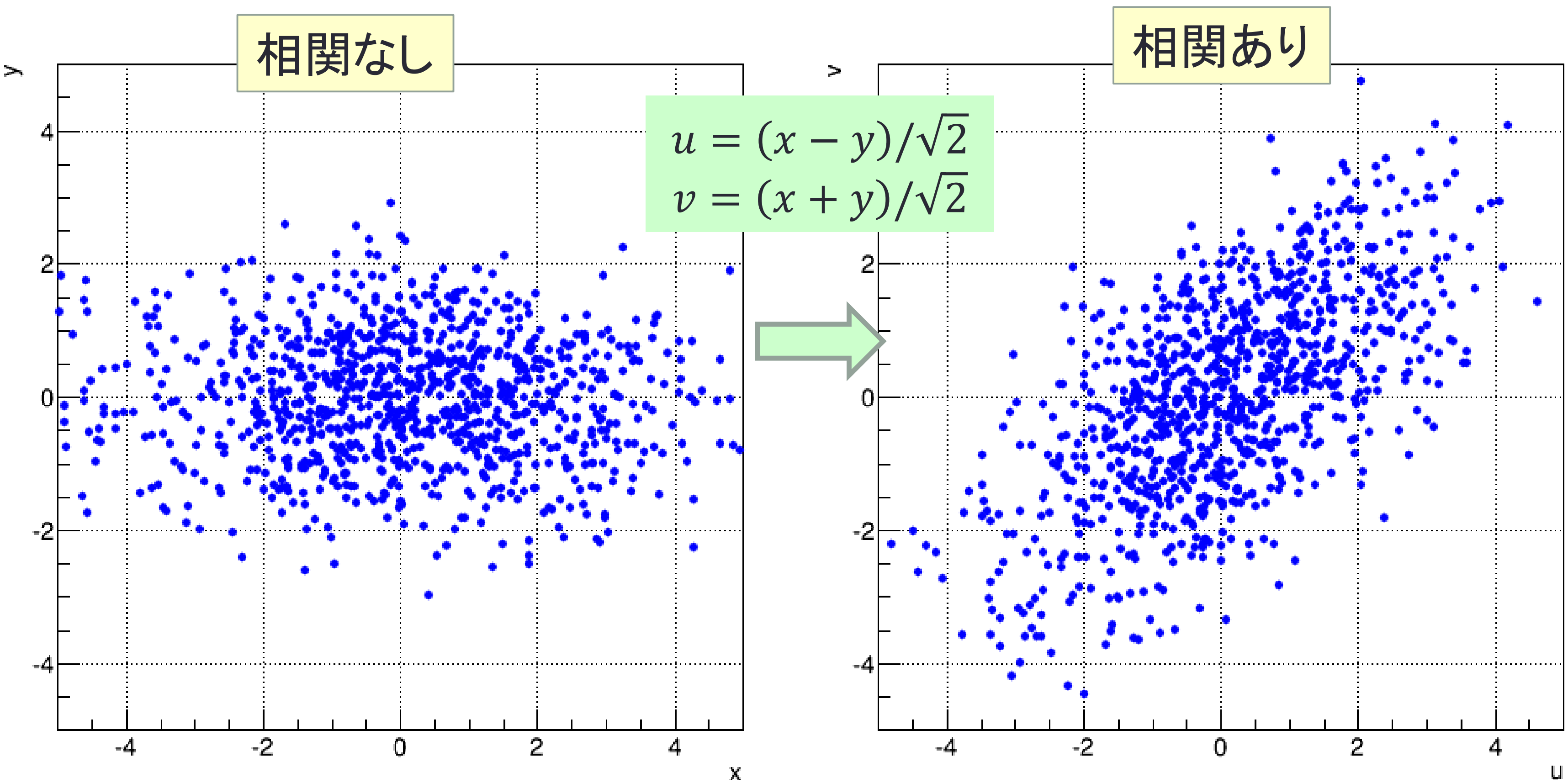
$$V(y) = E\left((y - \bar{y})^2\right) \cong E\left(\left(\sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i\right)^2\right) = E\left(\sum_{i,j=1}^{n} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \Delta x_i \Delta x_j\right)$$

$E(\Delta x_i) = 0$、$E\left(\Delta {x_i}^2\right) = V(x_i) = \sigma_i^2$であるから、変数間に相関が無ければ、

$$V(y) = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_i^2$$

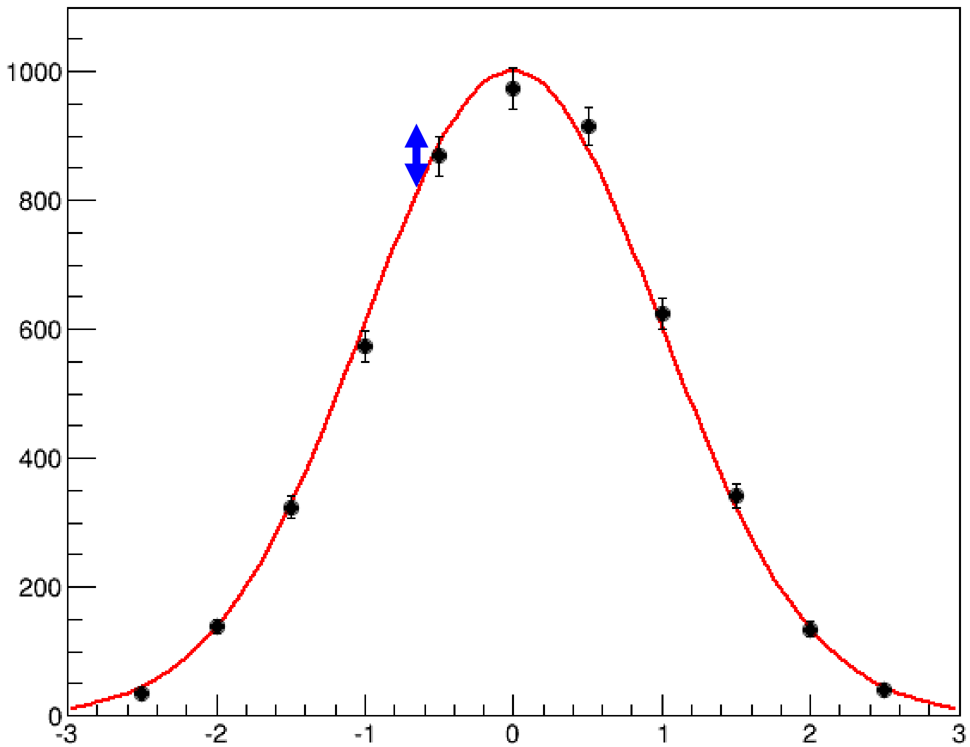
例：　$a + b$の分散は$\Delta a^2 + \Delta b^2$、誤差は$\sqrt{\Delta a^2 + \Delta b^2}$

変数の相関



- 二つの確率変数がある場合、その分布関数は$f(x, y)$となる
- 二つの変数が独立あるいは、相関が無いとは$f(x, y) = g(x)h(y)$となることである。この場合、平均や分散の計算は各変数ごとに独立に行える
  - xの分布がyの値に依らない

Likelihood法（最尤法）



$$P(y;x) \sim e^{-\frac{(y-f(x))^2}{2\sigma^2}}$$

Likelihood（尤度）

理論 ↔ データ

- 理論のパラメータを決めると測定値を予言できる
- 実験でやりたいのはこの逆である
  - 測定値が与えられた時に、理論のパラメータを決定したい

- 二つの変数（物理量）の関係が理論的に分かっていたとする
  - あるxに対して$y = f(x)$を計算できる
  - yの測定を行うと、データは$\bar{y} = f(x)$の周りに分布するはずであり、データがある値yになる確率を考える

理論のパラメータの推定

$$P(y;x) \sim e^{-\sum_{i=1}^{n} \frac{\left(y_i - f(x_i;\, p_1, p_2, p_3)\right)^2}{2\sigma_i^2}}$$

対数をとると、

$$logP \sim const. - \underbrace{\sum_{i=1}^{n} \frac{\left(y_i - f(x_i;\, p_1, p_2, p_3)\right)^2}{2\sigma_i^2}}_{\chi^2}$$

- 実験データが与えられた時、Likelihoodを最大化するパラメータを求める
  - 理論のパラメータの決定（＝物理量の測定）
- Likelihoodを最大化することと、χ^2を最小化することは誤差がガウス分布に従う場合は同等である。χ^2を最小化する方法は最小二乗法とも呼ばれる
- 最小二乗法は、パラメータについて線形であれば解析的に計算も可能である
  - $\frac{\partial \chi^2}{\partial p_i} = 0$を要請していくつかの拘束条件を導き、あとは行列の演算

検定

条件付き確率

問題

1. ガウス分布$N(\mu, \sigma^2)$の特性関数が$\phi(\theta) = e^{i\mu\theta - \frac{1}{2}\sigma^2\theta^2}$となることを示す
2. (x_i, y_i)というデータがN組与えられたとする。最小二乗法により、これを直線でフィットして直線のパラメータを求める。パラメータがどのように表せるか解析的に求める