13 データ処理

この章では、実験データを処理して物理量を決定する際にしばしば必要となる、データへの関数の当てはめ (フィッティング (fitting)) について述べる。

13.1 実験データと誤差

13.1.1 統計誤差と系統誤差

実験データには**統計誤差**(statistical error) と**系統誤差**(systematic error) がほぼ必ず伴う。データの処理に際しては、これらの誤差の性質と大きさを正しく理解し、適切な処理を行う必要がある[1]。

統計誤差は、放射線源の崩壊の測定 (例えば α 線のトンネル効果) に伴う単位時間当たりの計数率の揺らぎなど、頻度が統計的にばらついている現象を測定したときに伴うものである。多くの場合、ばらつきの従う統計分布は分かっており[2]、測定データから物理量を計算したり、複数の測定データを組み合わせたりする操作は、通常の統計学の手法で可能である。

他方、系統誤差は、測定装置の較正 (calibration)、検出効率 (efficiency) などの決定に伴う誤りや不定性に起因する誤差で、一般には、誤差がどのような分布関数に従うかは分からない。便宜的には、Gauss 型の分布をすると仮定し、複数の系統誤差の要因が起因している場合の全系統誤差 σ_{syst} を、個々の系統誤差 $\sigma_1, \sigma_2, \ldots$ から、

$$\sigma_{syst} = \sqrt{\sigma_1^2 + \sigma_2^2 + \ldots} \tag{13.1}$$

と推測することが多い[3]。

以下の説明では、系統誤差を無視し、統計誤差についてのみ考える。

13.1.2 確率密度関数

宇宙線が単位時間毎に検出器で検出される個数や、ヒストグラムに区分されたデータのチャンネル毎のカウント数は**Poisson 分布**で与えられ、その確率密度関数は

$$P(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{13.2}$$

で与えられる。ここで、平均値は λ で、測定値が x になる確率が P である。Poisson 分布では、平均値と分散が共に λ であるので、例えばヒストグラムのあるチャンネルのカウント数が n の場合、その統計誤差は $\sqrt{n}$ である[4]。

実験データを扱う際に重要な**Gauss 分布**の確率密度関数は

$$G(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{13.3}$$

[1] 例えば、誤差 A が誤差 B より桁で小さければ、誤差 A の性質を詳しく知る必要は一般にあまりない (誤差 B を考えれば十分)。

[2] 例えば、Poisson 分布や Gauss 分布

[3] 各誤差の要因の間に相関が無いとしたことに相当する。

[4] 厳密には $\sqrt{\lambda}$ であるが、一般に分布の真の平均値 λ は既知では無いので、測定値 n をもって λ に代える。

で与えられる。ここで、平均値は μ、分散は σ^2 で、測定値が x になる確率が G である。なお、平均値 λ が大きい Poisson 分布は、$\mu = \lambda$、$\sigma^2 = \lambda$ の Gauss 分布に漸近する (各人確かめよ)。

通常、誤差を伴う測定値を $A \pm \sigma$ と表示する。これは、平均値 μ の推定値が A、標準偏差の推定値が σ であることを示している (1σ 誤差表記)。Gauss 分布の場合、確率変数 x が $\mu \pm \sigma$ の間に含まれる確率は 68.3%、$\mu \pm 2\sigma$ の間に含まれる確率は 95.3%、$\mu \pm 3\sigma$ の間に含まれる確率は 99.7% である。

13.2 Maximum Likelihood (最尤) 法

実験により、ある分布に従ってばらついている独立な n 個のデータ

$$x_1, x_2, \ldots, x_n$$

が得られたとし、このデータをもとに物理量 a を推定する事を考える。ばらつきの原因としては、物理的要因 (例えば、寿命を持った粒子の崩壊時間の測定) や測定装置の性質 (測定の分解能) などが考えられる。ここでは、ばらつきは前述のような原因により**ランダム**に起き、その確率密度関数 $p(x; a)$(例えば、$P(x; \lambda)$ や $G(x; \mu, \sigma)$) は決定したい物理量 a を与えれば計算できるものとする。

1 回の測定で測定値が x_i になる確率が $p(x_i; a)$ であるので、独立な n 回の測定でデータが

$$x_1, x_2, \ldots, x_n$$

となる確率は、

$$L(a) = \prod_{i=1}^{n} p(x_i; a) \tag{13.4}$$

で与えられる。この $L(a)$ を**Likelihood(確からしさ) 関数**という。

> パラメータ a の推定値として最も確からしい値 (**最尤値**)$\hat{a}$ は、Likelihood 関数 $L(a)$ が最大になる時の値である

として、実験データから最尤値 $\hat{a}$ を求める方法を**Maximum Likelihood (最尤) 法**という。実際には、$L(a)$ ではなく $\ln L(a)$ を計算する方が容易である場合が多い。

13.2.1 μ^+ 粒子の寿命測定

Maximum Likelihood による物理量の推定を、μ^+ 粒子の寿命測定を例に説明する。

μ^+ 粒子は弱い相互作用により、ある寿命 τ で、

$$\mu^+ \to e^+ + \nu_e + \bar{\nu_e} \tag{13.5}$$

と崩解する。実際に μ^+ が崩壊する時間を測定したところ、10 崩解事象に対して、

```
3.70, 4.51, 2.04, 5.94, 0.85, 3.12, 3.70, 0.26, 0.71, 0.14
```

というデータを得たとする (単位は μs)。μ^+ 粒子の崩解が指数分布

$$\begin{aligned} f(t; \tau) &= \frac{1}{\tau} e^{-t/\tau} \\ &= \lambda e^{-\lambda t} \quad (\lambda \equiv 1/\tau) \end{aligned} \tag{13.6}$$

に従うとして、データから寿命 τ を推定し、その誤差を検討する。

指数分布であるので、Likelihood 関数 $L(\lambda)$ は

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} \lambda e^{-\lambda t_i} \\ \ln L(\lambda) &= \sum_{i=1}^{n} \ln\left(\lambda e^{-\lambda t_i}\right) \\ &= n \ln \lambda - \lambda \sum_{i=1}^{n} t_i \end{aligned} \tag{13.7}$$

となる。Likelihood 関数 $\ln L(\lambda)$ が極大となる λ の値を求めるために、λ で微分して、

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i = 0 \tag{13.8}$$

この式を解くと、

$$\hat{\tau} = \frac{1}{\tau\lambda} = \frac{\sum_{i=1}^{n} t_i}{n} \tag{13.9}$$

を得る。すなわち、寿命の最尤値 $\hat{\tau}$ は測定データの加算平均であることが分かり、実際に求めると、

$$\hat{\tau} = 2.5\,\mu\mathrm{s} \tag{13.10}$$

という結果を得る (各人文献値と比較してみよ)。

$\hat{\tau}$ の誤差については、Likelihood 関数 $L(\tau)$ から計算できる確率分布関数

$$\frac{L(\tau)}{\int L(\tau')d\tau'} \tag{13.11}$$

を用いて、**信頼区間 (Confidence Interval)** あるいは**信頼度 (Confidence Level)** として表す。具体的には、

信頼度 90%で寿命 $\tau = \hat{\tau}^{+\sigma'}_{-\sigma}$

と書かれている場合、これは

寿命 (の真の値)τ が $\hat{\tau} - \sigma < \tau < \hat{\tau} + \sigma'$ の間に入る確率が 90%

を意味する。σ、σ' は

$$\frac{\int_0^{\hat{\tau}-\sigma} L(\tau')\tau'}{\int_0^{\infty} L(\tau')\tau'} = 0.05, \quad \frac{\int_{\hat{\tau}+\sigma'}^{\infty} L(\tau')\tau'}{\int_0^{\infty} L(\tau')\tau'} = 0.05 \tag{13.12}$$

から求める事ができる。

課題

与えられた μ^+ 粒子の崩解時間分布から、図 13.1 にあるような確率分布関数 $L(\tau)/\int L(\tau')d\tau'$ を数値積分を用いて計算し図示せよ ($L(\tau)$ は Likelihood 関数である)。また、信頼度 (Confidence Level) 90%で寿命 τ を推定せよ。

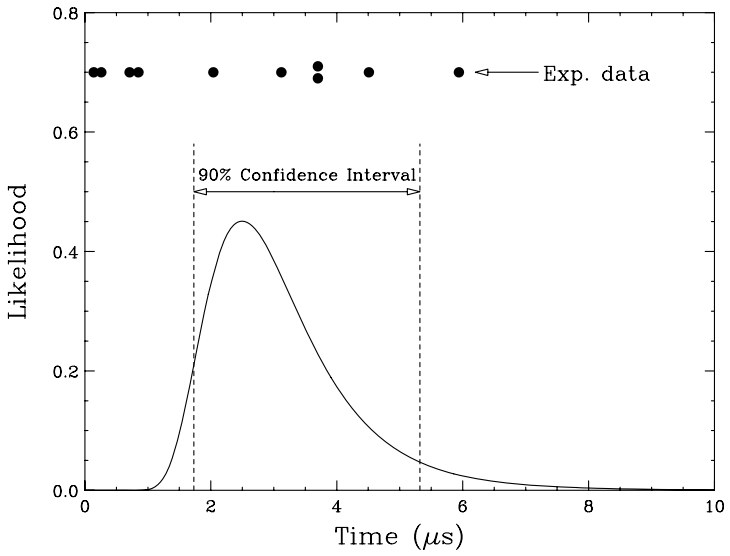

図 13.1: μ^+ 粒子の寿命測定データが指数分布に従うと仮定して計算した規格化された確率分布関数 $L(\tau)/\int L(\tau')d\tau'$。

13.3 線形関数の最小 2 乗法

最小 2 乗法は、独立変数 x_i に対して測定データ y_i が誤差 σ_i をともなって n 個得られたが、**分布関数がよく分かっていないために、前述の Maximum Likelihood 法が適用できない**ような場合に、データに理論曲線を当てはめる手法として有効である。

簡単のため、k 次の多項式

$$f(x) = \sum_{j=0}^{k} a_j x^j \tag{13.13}$$

をデータに当てはめる (データを最も良く再現する a_j を求める) ことを考える。最小 2 乗法では、関数 $f(x)$ とデータ y_i とのズレの度合いを、

$$\chi^2 \equiv \sum_{i=1}^{n} \frac{(y_i - f(x_i))^2}{\sigma_i^2} \tag{13.14}$$

で表し、**この χ^2 を係数 a_j を調整して最小にする**。したがって、各 a_j に対する偏微分が 0 になるような a_j の組を求めれば良いので、解くべき問題は

$$\frac{\partial \chi^2}{\partial a_j} = \sum_{i=1}^{n} \frac{2(y_i - f(x_i))}{\sigma_i^2} x_i^j = 0 \quad (j = 0, \ldots, k) \tag{13.15}$$

で与えられる $(k+1)$ 元連立 1 次方程式を解くことに帰着する。

1 次式 $f(x) = a + b(x)$ の場合について具体的な形を書き下すと、

$$\begin{pmatrix} \sum_{i=1}^n \frac{1}{\sigma_i^2} & \sum_{i=1}^n \frac{x_i}{\sigma_i^2} \\ \sum_{i=1}^n \frac{x_i}{\sigma_i^2} & \sum_{i=1}^n \frac{x_i^2}{\sigma_i^2} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = A \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \frac{y_i}{\sigma_i^2} \\ \sum_{i=1}^n \frac{x_i y_i}{\sigma_i^2} \end{pmatrix} \tag{13.16}$$

となる。a、b の誤差は、行列 A の逆行列の対角項により与えられることが

“Data Reduction and Error Analysis for the Physical Sciences”, P.R. Bevinton, (McGraw-Hill, 1969)

に詳しく説明されている。興味のあるものは読んでみると良い。

行列 A の逆行列 A^{-1} は、Gauss-Jordan の消去法を用いることにより数値計算で求めることももちろん可能だが、このような次元の低い場合は解析解を書き下して、それを計算させる方がよい。具体的には、

$$\begin{aligned}\sigma_a^2 &\approx \frac{1}{\Delta}\sum_{i=1}^{n}\frac{x_i^2}{\sigma_i^2}\\ \sigma_b^2 &\approx \frac{1}{\Delta}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\\ \Delta &= \sum_{i=1}^{n}\frac{1}{\sigma_i^2}\sum_{i=1}^{n}\frac{x_i^2}{\sigma_i^2}-\left(\sum_{i=1}^{n}\frac{x_i}{\sigma_i^2}\right)^2\end{aligned} \tag{13.17}$$

を計算すればよい。

13.4 最小 2 乗法と Maximum Likelihood 法

最小 2 乗法は、分布関数の形を知らなくても適用できる利点がある反面、fit するデータ y_i には誤差 σ_i を与えてやる必要がある。この誤差は Maximum Likelihood 法の場合は不要である。データが Gauss 分布で表される場合、最小 2 乗法と Maximum Likelihood が等価であることは以下のようにして示すことができる。

x_i における測定値の平均値が $f(x_i;a)$ で与えられる場合、$\ln L$ は

$$\begin{aligned}\ln L &= \ln\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(y_i-f(x_i;a))^2}{2\sigma_i^2}\right)\\ &= \sum_{i=1}^{n}\ln\left(\frac{1}{\sqrt{2\pi}\sigma_i}\right)-\sum_{i=1}^{n}\frac{(y_i-f(x_i;a))^2}{2\sigma_i^2}\\ &= \sum_{i=1}^{n}\ln\left(\frac{1}{\sqrt{2\pi}\sigma_i}\right)-\frac{1}{2}\chi^2\end{aligned} \tag{13.18}$$

となる。第 1 項はパラメータ a に依存しない量であり、第 2 項は χ^2 の $-1/2$ である。従って、χ^2 を最小にすることにより L が最大になり、Gauss 分布に従うデータに Maximum Likelihood 法で関数を当てはめるのは、最小 2 乗法と等価であることが確かめられた。

課題

表 13.1 に μ^+ 粒子の 0.5 μs 毎の崩解数の測定例を載せる。

このデータのバックグラウンドが無視できるほど小さいならば、崩解数 $f(t)$ は時間 t の関数として

$$f(t) = a\exp(-\lambda t)\quad(\lambda\equiv 1/\tau) \tag{13.19}$$

で与えられる。この関数をデータに当てはめることにより、μ^+ の寿命 τ を決める事ができる。

表 13.1: μ^+ 粒子の寿命測定のデータ。

時間 (μs)	崩解数	時間 (μs)	崩解数
0.25	999	5.25	98
0.75	818	5.75	86
1.25	662	6.25	76
1.75	501	6.75	53
2.25	419	7.25	43
2.75	307	7.75	24
3.25	262	8.25	24
3.75	208	8.75	21
4.25	171	9.25	16
4.75	131	9.75	22

この fit は、データの ln と取ると、1 次式の最小 2 乗法に帰着する (誤差も適切に変換する必要がある)。最小 2 乗法のプログラムを書き、寿命 τ とその誤差を求めよ。また、`gnuplot` に組み込まれた最小 2 乗法フィットを行うコマンドでもフィットを行い、両者を比較せよ。