Data Analysis Tutorials with Python

(Python で学ぶデータ解析)

2021 年 5 月 14 日

目次

第 1 章

基本事項

1.1 はじめに

通常の変数と、確率変数を明確に区別して理解を進めていこう。ここでいう「通常の」変数とは確率とは無関係に値を持つものであり、例えばその日の体重だったり身長だったり気温だったりするものである。今日は 60kg で、明日は X% の確率で 70kg になる、ということは無いため確率変数ではないと言える。ただし、解釈によっては体重も確率変数になり得る。日本人の平均体重を考えると、だいたい 60 70kg が標準体重で、数 % の確率で 100kg の人がいる、とか。

確率変数は X で表され、通常の変数は x で表される。

1.2 基礎知識

平均 $\bar{x}$

実験の n 個の測定値（$x_1, x_2, ..., x_n$）の代表値として、平均は（mean; 算術平均または相加平均）がよく用いられる：

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.1}$$

分散 V

各測定値が、平均からどれくらい離れて分布しているか（どれだけばらついているか）を表す指標として分散が用いられる。分散は平均からのズレ（偏差）の二乗和の平均であり、次のように定義される：

$$V = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \int (x - \bar{x})^2 f(x) dx \tag{1.2}$$

平均値の周りにピッタリ分布していれば、 分散は小さくなり、平均値よりも大きかったり小さかったりすると、分散 V は大きくなる。また分散の平方根を取ったものを標準偏差 S という。

期待値

ある確率変数 x に対する期待値は $E[x]$ と表され、

$$E[x] = \sum_{i=1}^{n} x_i p_i \tag{1.3}$$

と表される。ここで p_i は x_i を観測する確率である[*1]。x が全て等確率で出現するのであれば（ex. 理想的なサイコロの例)、

$$E[x] = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.4}$$

と表される。連続な値を取る確率密度関数に従う場合であれば x を観測する確率は $f(x)$ と表されるので、

$$E[x] := \int x f(x) dx \tag{1.5}$$

と積分形式で表すことができる。

Gaussian の期待値

ガウス分布に従う変数の期待値を計算すると：

$$E[x] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{1.6}$$

確率によって重み付けして平均を計算したもので、

$$E[x] = \frac{p_1 \times x_1 + p_2 \times x_2 + ... + p_n \times x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} p_i x_i \tag{1.7}$$

重要な性質として、実験結果 $x = (x_1, x_2, ..., x_n)$ の期待値は、

$$E[x] = \lim_{n\to\infty} = \frac{1}{n}\sum_{i=1}^{n} p_i x_i = \frac{1}{n}\sum_{i=1}^{n} p_i x_i \tag{1.8}$$

母集団と標本

ある母集団から確率変数 X を予想する場合。無限個のサンプル取得が可能であれば、 X の平均値（=期待値) は

$$\mu = \bar{X} = \lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \tag{1.9}$$

母平均の期待値 $E[\mu]$ は、

$$E[\mu] = \lim_{n\to\infty}\sum_{i=1}^{n} \tag{1.10}$$

で表される。母集団から無限個のサンプルを取得すれば、真の値（母平均）を計算することができることを意味する。また、母集団の分散（母分散）は

$$V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{1.11}$$

で表される。

実際には n の値は有限であり、母平均 μ や母分散 V を計算することはできない。ここで注意が必要である。

標本分散

$$E((X - \bar{X})^2) = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2 = \tag{1.12}$$

[*1] サイコロを投げる例であれば、1 回投げて x =1 の目が出る確率は p =1/6。

1.3　確率分布

ポワソン分布

平均 λ 回発生する確率事象が、x 回起こる確率

$$
\begin{aligned}
&\text{確率密度関数} \quad f(x) = \frac{e^{\lambda}\ \lambda^{x}}{x!} \\
&\text{期待値} \quad E(x) = \lambda \\
&\text{分散} \quad V(x) = \lambda
\end{aligned}
$$

事象が起こる平均値 λ が大きくなるほど、形がガウス分布に近づいていく。十分に大きくなると $\mu = \lambda$、$V = \lambda$ のガウス分布として近似できる。

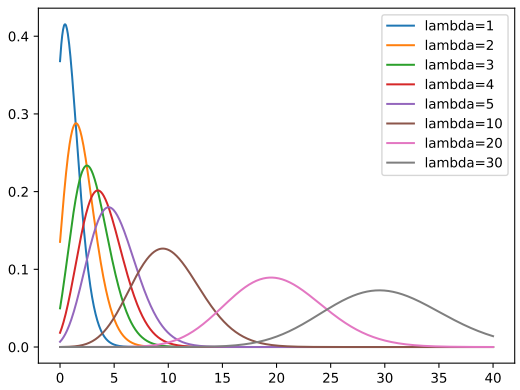

図 1.1　異なる平均値を取ったときのポワソン分布。平均 λ が大きくなるほど、ガウス分布のような形状になっていくのが分かる。

正規分布（ガウス分布）

$$
\begin{aligned}
&\text{確率密度関数} \quad f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)} \\
&\text{期待値} \quad E(x) = \mu \\
&\text{分散} \quad V(x) = \sigma
\end{aligned}
$$

ガウス分布の性質その 1

確率変数 X が $N(\mu, \sigma^2)$ の正規分布に従うなら、$aX + b$ は $N(a\mu + b, a^2\sigma^2)$ の正規分布に従う

1.4　共分散

1.4.1　モーメント

確率密度関数 $f(x)$ に対して、x^n の期待値を定義することができる（n は整数）。それらはモーメントと呼ばれ：

$$
a_n \coloneqq E[x^n] \tag{1.13}
$$

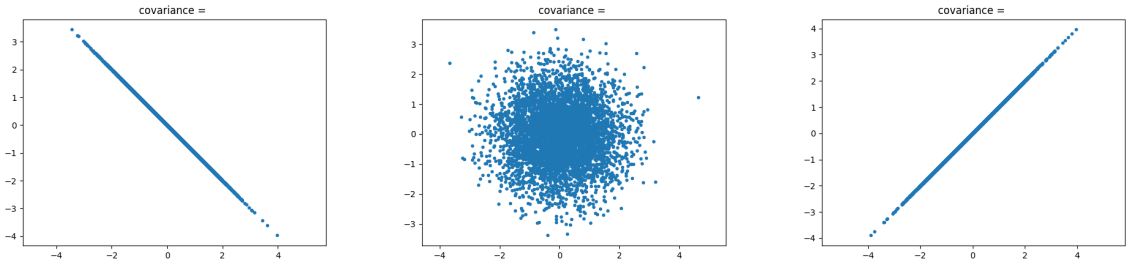

図 1.2　左から順に、cov=-1, 0, +1 の共分散の関係性を持たせてプロットした図。

と定義される。1.5 より

$$a_n = \int x^n f(x) dx \tag{1.14}$$

と表される。$n = 1$ の場合は平均値を表す。

また、x が平均値からどれだけ離れているか（散らばっているか）を表す量として中央モーメント

$$m_n \coloneqq E[(x - \mu)^n] \tag{1.15}$$

が定義されている。$n = 2$ は分散として知られている値であり、

$$m_2 = E[(x - \mu)^2] = \sigma^2 = V[x] \tag{1.16}$$

1.4.2　共分散の定義

確率密度関数が 2 変数に依存する場合（$f(x, y)$）、共分散（covariance）として変数同士の関係性を計算することができる：

$$\begin{aligned}
\mathrm{cov}[x, y] &\coloneqq \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\
&= \frac{1}{n} \sum (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y})) \\
&= \frac{1}{n} \sum x_i y_i - \frac{1}{n} \bar{y} \sum x_i - \frac{1}{n} \bar{x} \sum y_i + \bar{x}\bar{y} \\
&= \frac{1}{n} \sum x_i y_i - \frac{1}{n} \bar{y} \sum x_i \\
&= E[xy] - E[x]E[y]
\end{aligned} \tag{1.17}$$

また連続関数を用いると

$$\mathrm{cov}[x, y] = \int xy f(x, y) dx dy - \int x f(x, y) dx \int y f(x, y) dy \tag{1.18}$$

$f(x, y) = f(x) \cdot f(y)$ と表すことができるなら（互いに独立な事象の場合）、

$$\mathrm{cov}[x, y] = E[xy] - E[x]E[y] = \iint xy f(x) f(y) dx dy - \int x f(x) dx - \int y f(y) dy = 0 \tag{1.19}$$

となる。1.2 に共分散によるデータの分布の違いを示している。

また、相関（correlation）として次の量も定義される：

$$\rho \coloneqq \frac{\mathrm{cov}[x, y]}{\sigma_x \sigma_y} \tag{1.20}$$

これらは多次元への拡張も可能である。

$$V_{i,j} = \mathrm{cov}[x_i, x_j] = E[x_i x_j] - E[x_i]E[x_j], \tag{1.21}$$

$$\rho_{i,j} = \frac{V_{ij}}{\sigma_i \sigma_j} \tag{1.22}$$

1.4.3　共分散行列

1.4.4　多変量正規分布

Multivariate nominal distribution（多変量正規分布）は：

$$f(\boldsymbol{x}|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu)^T \Sigma^{-1}(\boldsymbol{x}-\mu)\right) \coloneqq \mathcal{N}(\mu, \Sigma) \tag{1.23}$$

と定義される。$\boldsymbol{x}$ は多次元の変数、μ は平均値ベクトル、Σ は $N \times N$ 共分散行列、$|\Sigma|$ は行列式を表す。

第 2 章

検定

2.1 概要

帰無仮説 (H_0）と対立仮説（H_1）を用意する。帰無仮説の元での検定量（test statistics）を、実験結果を元に計算する。その検定量が H_0 の棄却領域にいれば、帰無仮説を棄却。棄却領域に入らなければ帰無仮説を採択する。その判定のために、検定量から p-value を計算して、棄却領域を定義する p-value（$\alpha = 0.05$）と比較して棄却領域にいるかどうかを検討すればよい。

H_0 が棄却されるということは、p-value がしきい値以下であることを示しており、これは「帰無仮説が現実世界で正しい仮説であると、めったに起こり得ない結果が本実験結果から得られている」とということであり、それは帰無仮説が正しい、という仮定が間違っていると考えるのが自然であろう。そのため棄却されるということである。

ここで重要なことは、どんな量を検定量とするかということと、p-value を計算するために、検定量の従う分布（sampling distribution）の形が既知であることである。特にその分布を表式できなければならず、方程式は分かるけれどパラメーターが推測できないということはあまり意味がない。そこでスチューデントの t-検定やカイ二乗検定が出てくるのである。

2.2 カイ自乗検定

$$\chi^2 = \frac{(x_1 - \mu_1)}{\sigma_1^2} + \frac{(x_2 - \mu_2)}{\sigma_2^2} + ... + \frac{(x_n - \mu_n)}{\sigma_n^2} = \sum_{i=1}^{n} \frac{(x_i - \mu_i)}{\sigma_i^2} \tag{2.1}$$

カイ自乗の確率関数

2.2.1 どうやってフィットの妥当性を評価するか？

異なる実験で Z ボソンの質量が測定された。これらの実験結果を尤もらしく一つの数値で表現することはできるだろうか？

L3	91.161 ± 0.013
OPAL	91.174 ± 0.011
Aleph	91.186 ± 0.013
Delphi	91.188 ± 0.013

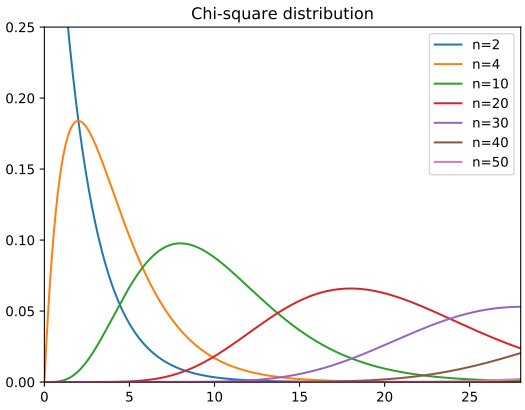

図 2.1　χ^2 の確率分布。chi^2 が大きくなるに従って、関数の形が右へシフトしていく。

第 3 章

誤差論

3.0.1 誤差の伝播

N 点の計測（$x_1, x_2, ..., x_n$）を行う。測定値それぞれに誤差がついている場合（$x_1 \pm \Delta x_1, x_2 \pm \Delta x_2, ..., x_n \pm \Delta x_n$）、測定結果の合計、平均にはどのような誤差が付くか。

加減乗除

$$(x \pm \sigma_x) \pm (y \pm \sigma_y) = (x \pm y) \pm \sqrt{\sigma_x^2 + \sigma_y^2} \tag{3.1}$$

$$(x \pm \sigma_x)/(y \pm \sigma_y) = \frac{x}{y} \pm \sqrt{\left(\frac{\sigma_x}{y}\right)^2 + \left(\frac{x \cdot \sigma_y}{y^2}\right)^2} \tag{3.2}$$

合計の誤差

$$y = \Sigma_i \ x_i = x_1 + x_2 + ... + x_n = n \times \frac{1}{n}(x_1 + x_2 + ... + x_n) = n\bar{x} \tag{3.3}$$

誤差の和はyに対して、

$$\Delta y = \sqrt{\left(\frac{\partial}{\partial x_1} y\right)^2 \Delta x_1^2 + ... +} = \sqrt{\Sigma_{i=1} \left(\frac{dy}{dx_i}\right)^2 \Delta x_i^2} \tag{3.4}$$

として伝搬する (誤差伝搬の公式). 今の場合、

$$\frac{\partial}{\partial x_1} n \times \frac{1}{n}(x_1 + x_2 + ... + x_n) = 1 \tag{3.5}$$

となるので、

$$\Delta y = \sqrt{(\Sigma(\Delta x_i)^2)} \tag{3.6}$$

測定値 x_i の分散は、測定した n 点から事前に計算される。なので誤差の伝播の式を進めると、

$$\Delta y = \sqrt{ns^2} = \sqrt{n}s \tag{3.7}$$

となる。

平均の誤差

測定値の平均値には、

$$\bar{x} = \frac{1}{n}\Sigma x_i \tag{3.8}$$

誤差伝搬の式を適用して、

$$\Delta\bar{x} = \sqrt{n \times \left(\frac{1}{n}\right)^2 s^2} = \frac{1}{\sqrt{n}} \tag{3.9}$$

この式は、平均値の誤差は測定点の数のルート N 倍で小さくなっていくことを示している。測定を繰り返し行い、平均を得た場合にどのように誤差が小さくなっていくかを示している。

- 計数の誤差- 計数 N の統計誤差は $\sqrt{N}$ で表される

ヒストグラムにおける誤差

ここまでの議論は、例えば粒子の不変質量を測定した場合に、あるイベントでは m_1[GeV]、あるイベントでは m_2[GeV]、... と測定した時に、その測定量に対してどれくらい誤差が付いているかを論じてきた。議論をもう一度繰り返すと、n 回測定した質量の平均値は、

$$\bar{m} = \frac{1}{n}\Sigma m_i \tag{3.10}$$

測定結果の分散は、

$$V = \Sigma(m_i - \bar{\mu})・・・(☆) \tag{3.11}$$

で表される。しかし、これらの議論はヒストグラムの場合には少し特殊な議論となる。よく不変質量の分布をヒストグラムにしたりして議論を進めていくが、この時によく「ヒストグラムにエラーバーを付けて」と言われるが、このエラーバーはここまでの誤差とは若干違っている。(☆) の誤差は不変質量の測定値に付くものであって、ヒストグラムのそれぞれのビンにつくものではない。

3.0.2　ビンにつく誤差

「ヒストグラムのエラー」は、そのビンにいる統計数に対して考える。よくみるヒッグス粒子の不変質量を例に取ってみよう。（☆）の議論はヒッグス質量 125GeV に対して $125GeV \pm \sigma_H$ という形で付くものである。これは強いて言うなら下図の横方向 (x 軸方向) に対しての誤差である、

では、図に示されている縦方向の誤差は一体なんでしょうか？これこそが高エネ実験領域でよく耳にする「統計誤差」というものである。

考え方

例えば 1 番目のビンに注目する。このビンでカウントされている事象数を仮に 460 事象と読み取ることにする。これを「とある確率分布に従う事象を観測し、平均値 460 事象観測した」と捉える。よく用いられる確率分布はポワソン分布である (1.3 節)。簡単に分散 $V = 460$、$\sigma = \sqrt{460}$ と求めることができる。これがヒストグラムの計量の際の誤差である。感覚的には、縦方向に（各ビンに）ある確率分布があって、それらの平均値をつなぐようにしてヒッグス質量分布が計算されているとすればいい。

誤差はどうなっていくか

統計量が溜まってくると、1 ビン〜10000 事象とかがあるかもしれない。こういうときにポワソン分布でヒストグラムのエラーバーを考えていいのか？という疑問が湧いてくる。答えは**ガウス分布**を使う、が、「考え方」は変わらないということである。

- 二項分布→ポワソン分布- 二項分布において λ を一定にして、n を大きく、p を小さくするとポワソン分布になる- ポワソン分布→ガウス分布- ポワソン分布で λ を大きくすると、期待値 λ、分散 λ のガウス分布になる- 二項分布→ガウス分布- 期待値 np、分散 $np(1-p)$ の両方が大きい場合、期待値 np、分散 $np(1-p)$ のガウス分布になる

上記の関係性を思い出せば、1 ビンあたりの統計量（＝1 ビンで観測したイベントの期待数）が大きくなっても、結局は分散 λ のガウス分布に従っているので、ヒストグラムのエラーバーはやはり $\sqrt{N}$ になる。

つまりヒストグラムのエラーバーは、

$N \pm \sqrt{N}$

の関係で成り立っている（ポワソン分布 or ガウス分布に由来するものであることに留意）。

3.0.3　「誤差は $\frac{1}{\sqrt{N}}$ で小さくなっていく」の真実

誤差の大きさそのものが $\frac{1}{\sqrt{N}}$ なのではなく、ここで言っている誤差は「XX%」の誤差に相当するものである。ヒストグラムの誤差は、

$$N \pm \sqrt{N} \tag{3.12}$$

で付くものだったから、1 ビンに対する誤差は

$$\frac{\sqrt{N}}{N} \times 100 = \frac{1}{\sqrt{N}}\% \tag{3.13}$$

である。つまり、1 ビンに入る統計数が多くなればなるほど、統計量が増えれば増えるほど、誤差は、$\frac{1}{\sqrt{N}}$ で小さくなっていきます。なので、できる限り統計量を貯める、という至極まっとうな感想が生まれてきます。

ちなみに、例えばあるビンの統計数が 1 事象だった場合の誤差は、

$$1 \pm 1 \tag{3.14}$$

で、「なんだ、誤差の大きさは 1 じゃん、小さいじゃん」と思うかもしれないが、

$$\frac{\sqrt{1}}{1} \times 100 = 100\% \tag{3.15}$$

の誤差が付いているという事実を忘れてはならない。ちなみに 100 事象貯めれば、

$$\frac{\sqrt{100}}{100} \times 100 = 10\% \tag{3.16}$$

10% の誤差にまで削減することができるのである。

第 4 章

RooFit で学ぶ統計処理

https://twiki.cern.ch/twiki/bin/view/Main/LearningRoostats

4.1 ジャーゴン

- フロートにする：定数として扱わないこと。MLE の時に、何を変数として扱って、何を定数として固定するかに関する話題でよく出る。setConstant(0);

4.2 関数

第 5 章

HistFactory

RooFit や RooStat で処理を行っていくためには、RooWorkspace を作成する必要がある。簡単なチュートリアルレベルであれば RooWorkspace の作成は簡単な作業であるが、実際の解析ではかなり骨の折れる作業である。そこで HistFactory と呼ばれるパッケージが ROOT では提供されていて、これを用いることでユーザーは自分のヒストグラムから簡単に RooWorkspace(＝確率密度関数) を計算して保存することができる。

5.1 使い方

ROOT には hist2workspace というコマンドが用意されていて[*1]、これに「どのヒストグラムをどういう設定で使用するか」を記した XML ファイルを食わせることで RooWorkspace がアウトプットされる。

5.2 marked Poisson model

marked poisson model と呼ばれる確率密度関数をヒストグラムの情報から計算する。ヒストグラムのビンの情報はシグナル数 S と背景事象数 B との間に次の関係を定義する。ν_b^{sig} は着目する b 番目のビンに含まれるイベント数で、次の等式が成り立つ（単にビンを端から端まで足し合わせたら全イベント数になるということを言っている)。

$$S = \sum_b \nu_b^{sig} \tag{5.1}$$

$$B = \sum_b \nu_b^{bkg} \tag{5.2}$$

シグナルと背景事象の”shape”は $f_S(x_e)$、$f_B(x_e)$ で表現する。ヒストグラムを hist-¿Scale(1/hist-¿Integral()) することと等しく、そのイベントがどのような確率で出現するかを表すことができる。

$$f_S(x_e) = \frac{\nu_{be}^{sig}}{S\Delta_{be}} \tag{5.3}$$

$$f_B(x_e) = \frac{\nu_{be}^{bkg}}{S\Delta_{b_e}} \tag{5.4}$$

以上の定義した式を用いて次の marked poisson model を計算する。

$$P(x_1, .., x_n|\mu) = \mathrm{Pois}(n|\mu S + B) \left[\prod_{e=1}^{n} \frac{\mu S f_S(x_e) + B f_B(x_e)}{\mu S + B} \right] \tag{5.5}$$

[*1] ローカルで ROOT を触っている人はインストールしていないかもしれない。ROOT を自分の環境でビルドしたときのことを思い出しましょう。

この式は、イベント数やその他のパラメータが固定（観測からわかっていれば）であれば、μ にのみ依存し、これはまさしく Likelihood を表している。

5.3　HistFactory のレンプレート

はじめに使用する添字について簡単な説明を行う。

- e：イベント
- b：ビン
- c：チャンネル
- s：サンプル
- p：パラメーター

さらに、

- ϕ_p：パラメーター毎の規格化定数（NormFactor）
- α_p：系統誤差に関連するパラメーター、それ自身は事前に CR 等で見積もっておくもの（OverallSys, HistoSys）
- γ_{csb}：ビンごとの系統誤差（ShapeSys、）

を定義する。 実際に HistFactory が計算する数式は次のものである。

$$P(n_c, x_e, a_p|\phi_p, \alpha_p, \gamma_b) = \prod_c \left[\mathrm{Pois}(n_c|\nu_c) \prod_{e=1}^{n_c} f_c(x_e|\alpha) \right] \times G(L_0|\lambda, \Delta_L) \times \prod_p f_p(a_p|\alpha_p) \tag{5.6}$$

第 6 章

Python と学ぶ統計

6.1 最小二乗法 (least square)

python/chi_square.py N 組のデータ (x_i, y_i) を測定した $(i = 1, ..., N)$。これらを尤もらしく表すことができる関数 $y = f(x)$ を求めたい。想定する状況は、実験者が x_i を固定して、y_i を測定していく、というもの。各 y_i はとある平均値の周りにふらつきを持って測定される。以下では簡単に線形近似の場合を考える。

$$y = ax + b \tag{6.1}$$

の係数を実験データから求めたい。そこで次のような値を考える。

$$\Phi(a, b) = \sum_{i=1}^{N} w_i r_i^2 = \sum_{i=1}^{n} w_i [f(x_i) - y_i]^2 \tag{6.2}$$

r_i は残差（residual）と呼ばれる値であり、データ点とモデル曲線との差を表している。w_i は各データ点に対する重みであり、分散の逆数に比例した値を持つ。仮に w_i が $1/\sigma_i^2$ に全く等しい場合、Φ は χ^2 に等しくなるため、最小二乗フィットはカイ自乗フィットとも呼ばれる。

誤差を含むような実験データ値から、もっともらしい関数の形を決定する時に用いられる手法。n 個のデータ点 $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$ をフィットする際に（関数の形は経験的に決定する）指標として用いる。

第 7 章

統計処理の応用；高エネルギー実験

高エネルギー素粒子実験では、未知の素粒子の兆候を掴むために様々な測定を行っている。新しい物理模型が存在すればこの様な兆候として発見することができる、という事前の調査に基づいて研究を進めていく。信号事象として新粒子の兆候、背景事象として既存の理論に起因する兆候を割り当てて考えていくことが多い。新粒子の発見が勿論究極の目標であるが、そこに至るまでの過程で様々な理論模型を棄却して可能性のある領域を絞っていく作業も非常に重要である。高エネルギー実験ではこれらを統計処理で発見・棄却の評価を行っていくが、通常の検定とは少し毛色が異なるため独立した章として、本章でその概要と詳細について議論する。

7.1 はじめに

高エネルギー実験では仮定する新物理事象から予測される兆候を掴むために、粒子のエネルギーや角度情報を測定する。例えばヒッグス粒子であれば、粒子の不変質量を計算することで（$H \to \gamma\gamma$ などの反応）125GeV 付近にピークを持つ不変質量分布を観測することができる。また、ヒッグス粒子の反応とは全く関係のない 2 光子で不変質量を組んでしまうことも予想される。そのため、得られる不変質量分布としては背景事象（ここでは標準理論で予想される通常の反応過程）の分布の上に、ピークが乗った様なヒストグラムを描くことができる。実験屋としては、そのピークの幅（分解能）を絞るための手法の開発・改良を行ったり、背景事象をいかに削減するか・信号事象をいかに無駄なく取得するかに工夫を凝らしていく。そのうえでピークが見つかれば「発見」という結論となる。

問題は、背景事象から計算した分布が偶然 125GeV にピークを持っただけではないのか、等どのような確率の上に成り立っていることであるかを評価していく必要がある。

素粒子実験では目当ての事象をカウントして、そのカウント数 s で理論モデルの検証を行っていく。その際には背景事象もあるカウント数 b だけ含まれてしまうことから、たとえ信号事象を観測できたとしてもノイズに埋もれてしまったら発見にはつながらない。そこで、最終的には b に対して s が優位に大きいかどうかを議論する必要があるのだが、この手法にはいくつかの種類があり

- カウンティング
- シェイプフィット

が主な手法である。大層な名前がついているが特に難しいわけではないのでじっくり考えれば理解できるはず。信号の発見を行うには、$s = 0$ の仮説の $p - value$ を計算することが一般的である。

これは、目的の新物理は世の中に存在しない場合（$s = 0$）に、実験で得られたイベント数がどれくらい普通に起きるかの目安となる。

7.2　検定について

??で述べた検定について、再度確認を行う。ここでは帰無仮説 H_0 として「信号事象は存在する」を選択し、対立仮説として「信号事象は存在しない、背景事象で世の中は記述される」を選択する。しかし帰無仮説が棄却されたからといって、直ちに仮定する信号事象が存在する、という結論にはならない。

7.3　p-value

$$p = \int_{\alpha}^{\infty} f(x)dx \tag{7.1}$$

例えばガウス分布であれば、p 値が小さければ小さいほど起こり得ない事象であると理解できる。高エネ業界では専ら、p 値を標準化された正規分布における significance Z に焼き直して議論する。

$$Z = \Phi^{-1}(1-p) \tag{7.2}$$

$p = 0.5$ のときに $Z = 0$、$p = 0.05$ のときに $Z = 1.64$、$p = 2.9 \times 10^{-7}$ の時 $Z = 5$ となる。つまり、標準化された正規分布の平均値から何 σ 離れた場所に位置しているかの指標となる。

7.4　$S/\sqrt{B}$ の導出

背景事象 B に対して信号事象 S がどれだけ優位に多いかを表す指標であり、発見感度の議論の仕方としては最も基本的なものの一つである。ポワソン分布の平均値 λ が大きいときにガウス分布で近似できる性質を使う。素粒子実験においてポワソン分布の平均値とは、「実験で観測できるとされる事象数 $s+b$」を意味する。ゆえに十分な統計を溜めることのできる実験では以下の近似式を用いることができる。

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi(s+b)}} e^{\frac{x-(s+b)^2}{2(s+b)^2}} \tag{7.3}$$

ここでポワソン分布は、 $\mu = s+b$、$\sigma = \sqrt{s+b}$ のパラメータを持つガウス分布で記述されている。

この近似式を用いた場合に、信号事象がゼロの仮説（background only hypothesis）に対する p 値は次のように計算できる（$s = 0$)。

$$p = 1 - \Phi\left(\frac{x-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{x-b}{\sqrt{b}}\right) \tag{7.4}$$

ここから、予想される significance を求めることができて、

$$\Phi\left(\frac{x-b}{\sqrt{b}}\right) = 1 - p \tag{7.5}$$

$$\frac{x-b}{\sqrt{b}} = \Phi^{-1}(1-p) = Z \tag{7.6}$$

（8.6）式より、$x = s+b$ の場合に予想される siginificance は、

$$\mathrm{med}[Z|s] = \frac{s}{\sqrt{b}} \tag{7.7}$$

この式は信号事象の優位性を議論する時に広く用いられているものである。

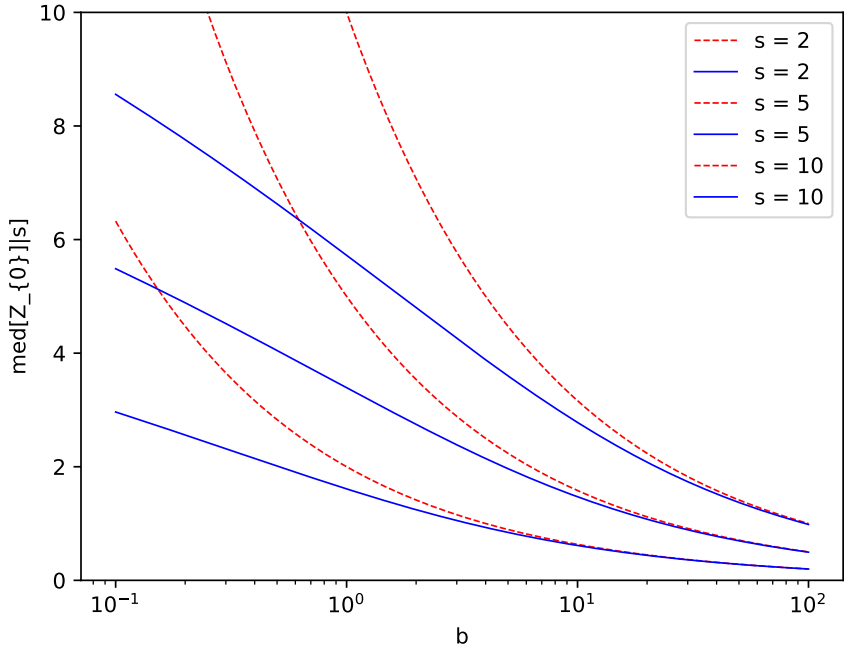

図 7.1　予想される siginificance の分布。赤線は $Z = s/\sqrt{b}$、青線は $Z = \sqrt{2\left((s+b)\log\left(1+\frac{s}{b}\right)-s\right)}$ から見積もった曲線。Asimov siginificance の方が背景事象が少なくなってくる領域でも正しく expected limit を見積もることが可能である。基本的には $s/\sqrt{b}$ はそのような領域では overestimation になっているだけで、$s/\sqrt{b}$ 同士で比較する分には多少の意味はある。

7.5　Statistical test

$$\lambda(s) = \frac{L(s, \hat{\hat{\theta}}(s))}{L(\hat{s}, \hat{\theta})} \tag{7.8}$$

7.6　Asymptotic formulae

7.6.1　導入

ある確率変数 x を測定して N ビンのヒストグラムを作成した（$n_1, n_2, ..., n_N$）。i 番目のビン内のイベント数の期待値は次のように表される。

$$E[n_i] = \mu s_i + b_i \tag{7.9}$$

ここで s_i と b_i は i 番目のビンにおける信号数と背景事象数の平均値である。

$$s_i = s_{tot} \int_{\text{bin } i} f_s(x;\theta_s)dx \tag{7.10}$$

$$b_i = b_{tot} \int_{\text{bin } i} f_b(x;\theta_b)dx \tag{7.11}$$

$f(x;\theta)$ は各モデルの確率密度関数を表しており、積分することで注目するビンに事象が得られる確率を計算することができる。また、μ は信号強度であり、$\mu = 0$ の時には bacground only hypothesis、$\mu = 1$ のときには信号を含ん

だ仮説を表現することができる。θ は確率密度関数の形状を決めるパラメーターで、$(\theta_s, \theta_b, b_{tot})$ をまとめて nuisance parameter として扱う。

7.6.2 likelihood

作成したヒストグラムに対する尤度関数は、とあるビンに事象が観測される確率（ポワソン分布）の積で表すことができる。

$$L(\mu, \theta) = \Pi_{\mathrm{j}=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \tag{7.12}$$

さらに、control region を定義した解析では、CR における尤度関数も式に含めることができる。

7.6.3 Profile likelihood ratio

信号強度 μ を試験するために、profile likelihood ratio $\lambda(\mu)$ を用いる。

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})} \tag{7.13}$$

分母は Likelihood の最大値、分子は μ をとある値に固定した時に Likelihood が最大となるように求めた θ を用いたときの値。尤度関数を最大にするパラメータの組み合わせは $(\hat{\mu}, \hat{\theta})$ であるため、常に分子は分母より小さいか分母に等しくなるので、$0 < \lambda < 1$ が成り立つ。$\lambda = 1$ はデータと μ がよく一致することを表していて、0 に近づくほど実験データと乖離している（μ を仮定した理論では実験データを説明できない、棄却される対象となる）ことが表される。

7.6.4 test statistic $t_\mu = -2 \ln \lambda$

statistical test を行う際には、λ をさらに次の様に変換して用いることが多い。

$$t_\mu = -2 \ln \lambda(\mu) = -2 \left(\ln L(\mu, \hat{\hat{\theta}}) - L(\hat{\mu}, \hat{\theta}) \right) \tag{7.14}$$

先程の λ の範囲の議論より、t_μ は $-2 < t_\mu < 0$ の範囲を動く。よって 0 に近いほど（値が大きくなるほど）、データと μ の整合性が取れないことを表している。

$$p_\mu = \int_{t_{\mu,obs}}^{\infty} f(t_\mu | \mu) dt_\mu \tag{7.15}$$

$t_{\mu,obs}$ は実験データから見積もった値、$f(t_\mu|\mu)$ は t_μ が得られる確率密度関数を表している。

7.6.5 Test statistics t_μ for $\mu > 0$

一般的に信号モデルは、既存のモデルに対してさらに数イベント信号が観測できると予測する。そのため信号強度は正の値をとる。その場合を試験量も記述しておく。$\mu > 0$ の領域では今まで通りの likelihood ratio を用いて、0 以下の領域では μ の値を 0 に固定した likelihood ratio を定義しておく。

7.6.6 Test statistics q_μ, q_0

高エネ実験では新物理 $\mu = 1$ を探している。実際の統計処理では $\mu = 0$ の background only hypothesis を棄却する方向で計算を進めていく[*1]。よって方針は、$\mu = 0$ の test statistics を正しく評価していくこととなる。（で、問題は）

[*1] ここが少し混乱の元。示したいのは信号を含んだモデルが存在することであるが、統計処理では「信号を含まないモデル H_0 では実験データを説明できない → 新しいモデルが必要となる」というロジックで進んでいく。

ここから定義した t_μ をさらにいろいろ条件を変えたときの式として、q_μ に置き換えて議論していくので、本書もそれに倣う[*2]。

$$q_0 = -2\ln\lambda(0) : \hat{\mu} > 0 \tag{7.16}$$

$$q_0 = 0 : \hat{\mu} < 0 \tag{7.17}$$

$$q_\mu = -2\ln\lambda(\mu) : \hat{\mu} < \mu \tag{7.18}$$

$$q_\mu = 0 : \hat{\mu} > \mu \tag{7.19}$$

7.7　Profile likelihood ratio の近似式

信号強度 μ を試験する場合を考え、データ点は μ' に従って分布しているとする。

$$-2\ln\lambda(\mu) = \frac{(\mu-\hat{\mu})^2}{\sigma^2} + O(1/\sqrt{N}) \tag{7.20}$$

ここで、$\hat{\mu}$ は mean μ'、標準偏差 σ のガウス分布に従う。N はサンプル数を表す。σ は共分散から簡単に求めることができて、

$$V_{ij} = \mathrm{cov}[\hat{\theta_i}, \hat{\theta_j}] \tag{7.21}$$

7.8　カウンティング（1 ビンフィット）

ポワソン分布に従う事象を n 事象観測して、1 ビンのヒストグラムで評価する場合を考える。SR における期待値は

$$E[n] = \mu s + b \tag{7.22}$$

s は信号モデルから予想される信号事象数の平均値、b は背景事象数の期待値、μ は信号強度を表す。b は nuisance parameter [*3]であり、CR における背景事象数の分布から制限をかけることができる。CR における背景事象の期待値は、SR と CR における背景数の違いを表す scale factor を τ で表すと、

$$E[m] = \tau b \tag{7.23}$$

と記述することができる。実際の解析ではこの τ も nuisance parameter としてフィットから求める手順を踏むが、ここでは簡単のために既知の値として話を進める。

実験から得られる値は SR の事象数 n、CR における自少数 m、興味のあるパラメータ（parameter of intereset; POI）μ、そして nuisance parameter b である。ポワソン分布に従うような場合に、μ と b に対して実験結果から次のような尤度関数を定義することができる。

[*2] 式の形は t_μ の議論の時に示したとおりだが、q_μ と明記したときには「μ の値に関して条件を掛けて立式していきますよ」という意思表示だと思ってもらえればいい。

[*3] nuisance とは迷惑なこと、厄介なこと、の意味を持つ単語。高エネの人が真に興味があるのは信号数なので、まぁ背景事象数は邪魔だということか。

7.9　シェイプフィット

1 ビンではなく、分布の形を反映した統計処理を行いたい場合、こちらの手法を用いる。俗語的に

- シェイプフィット
- シェイプでフィットした

とか呼ばれる。

7.10　CLs 法

以下で定義する CL_s と呼ばれる量を test statistic として用いて解析感度を評価する手法。

$$CL_s = \frac{CL_{s+b}}{CL_b} \tag{7.24}$$

CL_x はたどっていくと μ に依存する。よく用いられるのは 95%CLs であり、$CL_s = 0.05$ となる μ の値を探してそれを μ_{up} として解釈し、最終的に断面積の上限値設定に使用する手法である。

第 8 章

高エネルギー実験における統計処理

8.1 Likelihood

私達が興味のあるパラメータは（究極的には）信号事象の数に相当する信号強度 μ である。自然界では既に μ は定まった値を持っており、それを人間が知らないだけである。そのため実験データが得られる確率は

$$P(\text{data}|\mu) \tag{8.1}$$

と表すことができる。実験屋が持っている情報は実験データ（とシミュレーションサンプル）であり、データから確率密度関数を求める必要がある。これが Likelihood の意味するところである。

$$L(\mu) \tag{8.2}$$

Likelihood 関数を用いると、信号強度 μ の推定も最大尤度法（Maximum likelihood; ML）を用いることで可能となる。慣習的に推定量はハット記号で表し

$$\hat{\mu} = argmax L(\mu) \tag{8.3}$$

と信号強度を推定することができる。統計量が増加すると $\hat{\mu} \to \mu^{\text{truth}}$ に近づく。

8.1.1 Likelihood を計算する

実際の実験では例えば質量の分布（であったり、p_T、E であったり）のヒストグラムを計算して、そこに背景事象とデータとの間に統計的に優位な乖離があるかどうかを判別することになる。つまり、

- H_0：（この世の中には新物理などなく）データと背景事象は一致している
- H_1：（この世の中に新物理はある！）データと背景事象には何らかの乖離が生じている

とする 2 つの仮説を検討する、仮説検定の議論に持ち込むこととなる。このときに Neyman-Pearson の補題（検定量として Likelihood ratio を用いるのが最も性能が良い）によると、2 つの仮説に基づいた Likelihood funciton を計算して、それらの比を検定量とした仮説検定を行うことになる。ヒストグラムに基づいた Likelihood の計算方法には二種類ある。

8.2 Likelihood の使い方

ヒストグラムさえ作ることができれば、あとは何らかのツール（もしくは手計算？）で likelihood を計算することができる。計算した likelihood は主に 3 つの使い方が想定されている。

8.2.1 Maximum likelihood parameter estimation

8.2.2 Frequentist confidence interval

8.2.3 Beysian credible interval

8.3 Profile likelihood

あるビン i に期待される事象数は、

$$E(n_i) = \mu s_i + b_i \tag{8.4}$$

で表される。s_i はそのビンに信号事象が何イベント存在するかを表しており、b_i は背景事象が何イベントあるかを表している。よって i ビンに n_i 事象観測される確率はポワソン分布を用いると

$$\mathcal{P} = \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)} \tag{8.5}$$

と表される。全ビンに対して積を取ると Likelihood function が定義でき

$$L(\mu) = \prod_{i \in bins} \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)} \tag{8.6}$$

と表される。これが最も単純な形の Likelihood function であり、μ にのみ依存している。これは信号事象も背景事象も完全に 100% その性質が分かっている実験（現実的にはありえないが）でのみ使用できる likelihood 関数である。

もちろん実際には s_i や b_i は系統誤差の影響（規格化定数や分布の形等の影響、実験家がコントロールできない影響）を受けるため、それらを表すパラメーター（nuisance parameter）として $\boldsymbol{\theta}$ を用いて、

$$s_i \to s_i(\boldsymbol{\theta}), \quad b_i \to b_i(\boldsymbol{\theta}) \tag{8.7}$$

と定義し直すと、likelihood は

$$L(\mu, \boldsymbol{\theta}) = \prod_{i \in bins} \frac{(\mu s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta}))^{n_i}}{n_i!} e^{-(\mu s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta}))} \tag{8.8}$$

と書き直せる（式の形としては何も変わっていないが)。この様に、真に興味のある信号強度 μ 以外のパラメーター（nuisance parameters;NPs）に依存させた likelihood を profile likelihood と呼ぶ。

8.3.1 likelihood ratio

検定量として profile likelihood を用いた尤度比を考える。

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} \tag{8.9}$$

分子はある μ の値に対して likelihood を最大にする NP の値 $\hat{\hat{\boldsymbol{\theta}}}$ を求めた likelihood である（confitional maximum-likelihood)。また分母は全てのパラメーターをスキャンして likelihood を最大にする値 $\hat{\mu}$, $\hat{\boldsymbol{\theta}}$ を求めた likelihood である（unconditional maximum-likelihood)。ゆえに定義から

$$0 < \lambda < 1 \tag{8.10}$$

の範囲を取る。

8.3.2 系統誤差の扱い

系統誤差は基本的には全く値のわからないものであり、別の実験ないし作業から求める必要がある。つまり何らかの分布を使って「constrain」する必要がある。Profile likelihood は、系統誤差（systematics uncertainties）を likelihood 関数の中に含めることができる。系統誤差は "constrained" nuisance parameter として式に含める。

$$L(n, \theta^0|\mu, \theta) = \prod_{i \in bins} \mathcal{P}(n_i|\mu \times S_i(\theta) + B(\theta)) \times \prod_{j \in syst} \mathcal{G}(\theta_j^0|\theta_j, \Delta\theta) \tag{8.11}$$

ここで系統誤差は（慣習的に）平均値 $\theta^0 = 0$、分散 $\Delta\theta = 1$ の正規ガウス分布に従うものとされる。系統誤差の影響とは、あるビン i に対して $+1$〜-1 の値を取らせたときの影響を言う。

系統誤差が Nuisance parameter として尤度関数に含まれたように、その他のパラメータも尤度関数に「free parameter」として含めることができる[*1]。この nuisance parameter を「Normalizatoin factors（NF）」と呼び、

$$B(\theta, k) = kB(\theta) \tag{8.12}$$

の様に表現することができる（$\mu S(\theta)$ と同じ発想）。

8.3.3 Fit

もう一度系統誤差を含めた profile likelihood を眺めてみる。

$$L(n, \theta^0|\mu, \theta) = \prod_{i \in bins} \mathcal{P}(n_i|\mu \times S_i(\theta) + B(\theta)) \times \prod_{j \in syst} \mathcal{G}(\theta_j^0|\theta_j, \Delta\theta) \tag{8.13}$$

この likelihood は $L(\boldsymbol{n}, \boldsymbol{\theta}^0|\mu, \boldsymbol{\theta})$ から分かるように、ある信号強度である系統誤差の値をとっている時に、データセット $\boldsymbol{n}$、系統誤差 $\boldsymbol{\theta}^0$ の値を取っている関数である。事前に分かっているのはデータセット $\boldsymbol{n}$、系統誤差 $\boldsymbol{\theta}^0$ である[*2]。

maximum-likelihood estimation でパラメータ推定をした際には、$(\mu, \theta_1, ..., \theta_N)$ の複数セットの推定量を計算することになる。

通常の likelihood は μ に対する関数だったので、μ をスキャンしていき尤度関数が最大値を取るところを見つければよかった。言い換えると一次元の関数の最大値を見つける問題に相当する。対して Profile likilihood は、nuisance parameter として複数の free parameter に依存しているので、N 次元の関数となっている。

8.3.4 Likelihood ratio

profile likelihood はこれまで見てきたように NP（系統誤差、規格化定数 etc）を含んでいる。例えばある仮説（つまり μ の値を何かに固定する、ex. H_0）を選んだ時に、どの NP の値を使えばよいのだろうか？データに選ばせよう、というのが Profile の哲学である。

[*1] free parameter とは、と思うかもしれないが値が実験から決めない変量のこと。例えば μ は free parameter で、尤度関数を最大にする値を推定値として取るという点で、μ も free parameter である。

[*2] 繰り返しになるが、慣習的に constrain term のガウス分布は正規化されているので、$\theta^0 = 0$、$\Delta\theta = 1$ を意味している。そのため $\mathcal{G}(0|\theta, 1)$ として見ても良い。

$$t_{\mu_0} = -2 \ln \frac{L(\mu = \mu_0, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})} \tag{8.14}$$

を Profile likelihood ratio（PLR）と呼ぶ。

- $\hat{\hat{\theta}}$ は $\mu = \mu_0$ に対する best-fit（conditional MLE）
- $\hat{\theta}$ は overall （global）な best-fit (unconditional MLE)

Wilk's theorem によると、PLR は χ^2 分布に従うとされる。

第 9 章

Asimptotic Formulae

9.1 物理探索における検定

通常の検定と同様に帰無仮説と対立仮説を用意するが、どのような検定を行うかによって真逆の定義となる。

discovering

- H_0：背景事象のみとする仮説（background-only）
- H_1：背景事象 + 信号事象の仮説（signal + background）

limit settting

- H_0：背景事象 + 信号事象の仮説（signal + background）
- H_1：背景事象のみとする仮説（background-only）

また通常の検定と同様に、実験データと仮説の合い具合は p-value で評価される。

物理学において p-value は、それと同値の significance Z に変換して議論を進めることが一般的である。これは

$$Z = \Phi^{-1}(1-p) \tag{9.1}$$

と定義され、ガウス分布を用いた片側検定[*1]における累積分布関数に等しい。例えばヒッグス粒子の「発見」の場合[*2]、背景事象仮説 H_0 を $Z = 5$ の閾値以上で棄却している。これは $p = 2.87 \times 10^{7}$ に相当し、H_0 を仮定した場合に実験結果が得られる可能性は非常に（非常に非常に）小さいことが分かる。これがよく言う 5σ で発見、という意味合いである。

ヒッグス粒子の例は「発見」についての考え方であったが、実験を稼働させる段階（もしくは前段階）において、その実験が持つと期待される感度を正しく評価しておくことも重要である。

[*1] 物理学では信号事象があるかないか、つまり観測された実験結果が予想値よりも大きいかどうかを評価するため、右側の片側検定が通常である。

[*2] 先に導入した H_0 と H_1 の組み合わせに注意

第 10 章

Asimov data

Likelihood ratio から計算した検定量（test statistics）は次の形をしていた。

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu) dq_\mu \tag{10.1}$$

検定量の sampling distribution を知る必要がある。上限値の設定のためには $f(q_\mu|\mu)$ のがどのような分布になるかを事前に知っておく必要がある。