

# 高エネルギー宇宙物理学 のための ROOT 入門

– 第 2 回 –

奥村 暁

名古屋大学 宇宙地球環境研究所

2016 年 4 月 27 日

# 前回の補足

# 前回の補足 (1)

---

- ❖ 前回の講義資料の表紙で所属が間違っていました…
  - ▶ 誤：太陽地球環境研究所
  - ▶ 正：宇宙地球環境研究所 (※ 2015 年 10 月に改組)
  
- ❖ 言い忘れたこと
  - ▶ 学会とかで見かけたら気軽に声をおかけください
  - ▶ 「先生」じゃなくて「さん」が良いです

## 前回の補足 (2)

---

### ■ 講義中の質問について

- ▶ 言い忘れましたが、講義の最中は自由に質問してください
- ▶ ググレカスな質問も OK (その場で解決するのが重要)

### ■ 質問が出ない場合

- ▶ やる気がない、理解が追いついてない、全部知っているのどれかとみなします
- ▶ こちらもやる気が出ないので、質問歓迎です

### ■ 前回参加者数およそ 70 人

- ▶ 対して、git clone の行われた数が 25 くらい
- ▶ 今日は git clone しない人はいついていけないので、今すぐ

## 前回の補足 (3)

---

- RHEA.pdf の「付録 A Mac での研究環境の構築」という章は記述がかなり古いので (多分 2009 年頃) 注意
  - ▶ 「A.4 KeyRemap4MacBook」は Karabiner という名前に変更されました
  - ▶ 「A.6 MacPorts」は、最近だと homebrew を使うほうが好まれています
  - ▶ 「A.6.3 LaTeX」は、最近だと pTeX よりも TeX Live が好まれています

## 前回の補足 (4)

---

- 「すぐに図が出せるから」という理由で Excel や gnuplot を使い続けるのはやめる
- 大事なのはすぐに図を出すことではなく、得られたデータから現象を読み解くこと
- Excel や gnuplot で宇宙線・素粒子物理のデータ解析はできません
- 最初は時間がかかっても ROOT で作図する習慣をつける

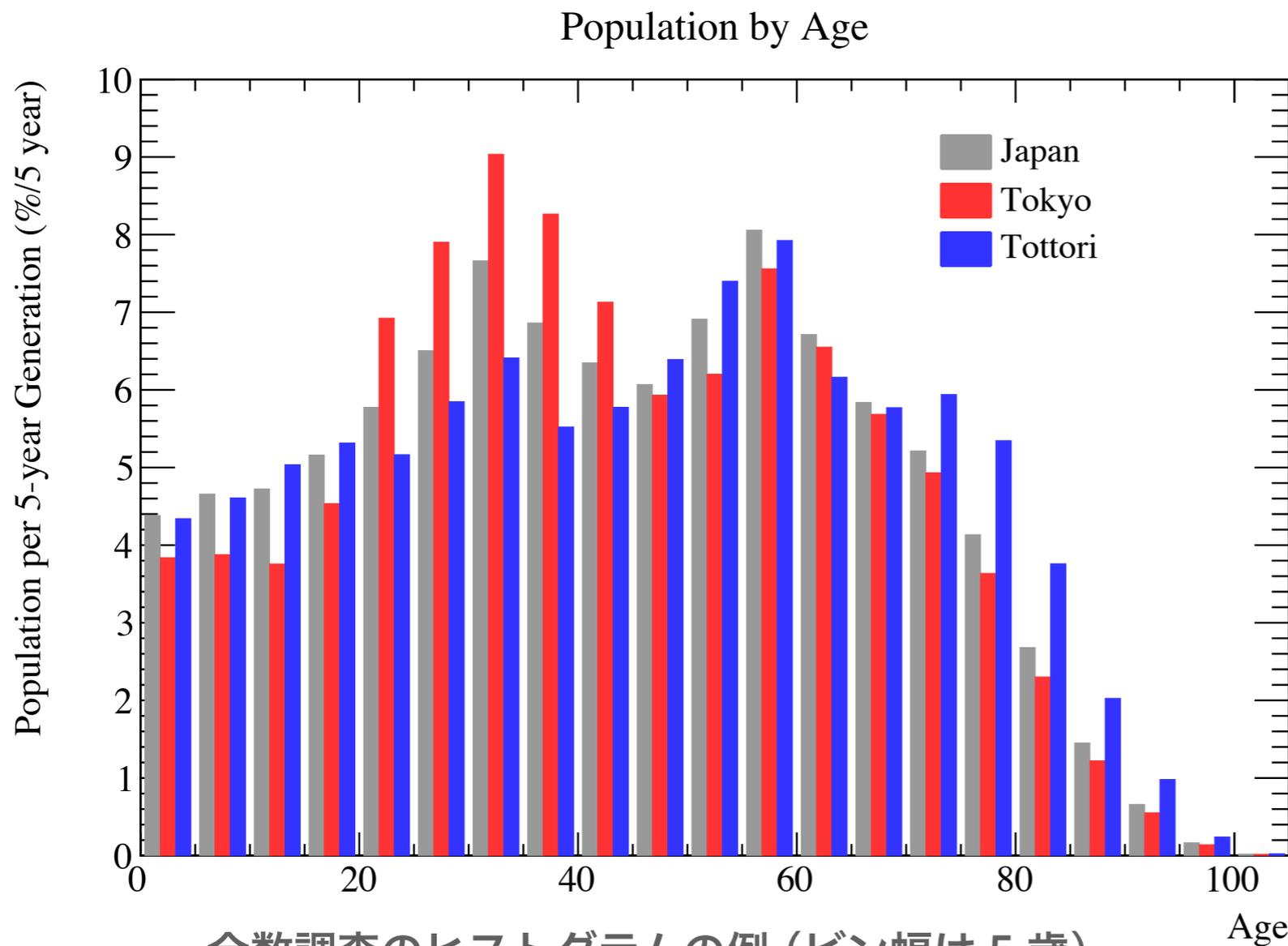
## 前回の補足 (5)

---

- C++ の用語が分からない人は、RHEA.pdf の C++ の章を参照すること
- 書きかけですが、クラス (class) とは何かくらいは理解できるはず

# ヒストグラム

# ヒストグラム (histogram) とはなにか？



全数調査のヒストグラムの例 (ビン幅は 5 歳)  
(データ出典：国勢調査 2005)

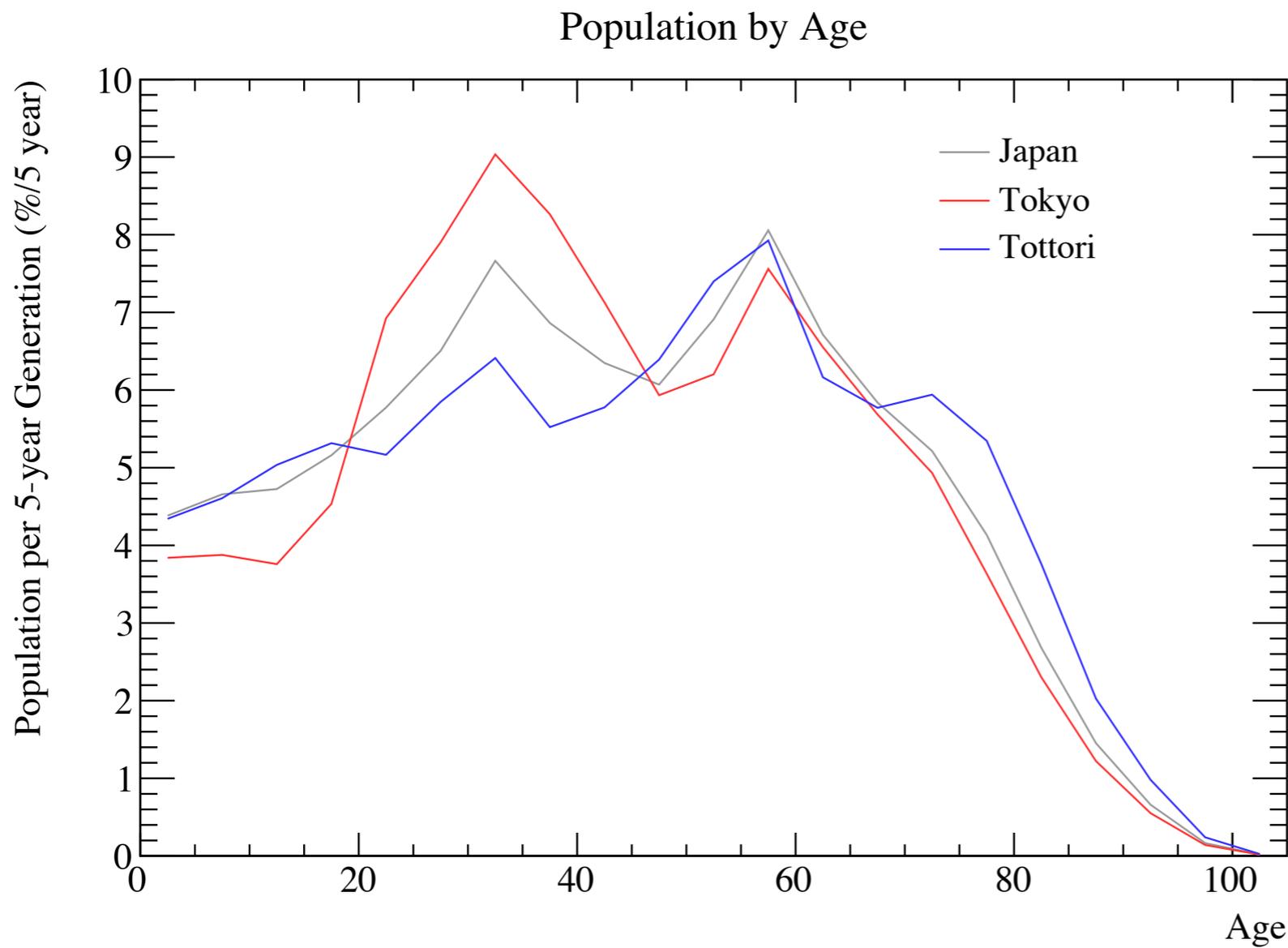
- 度数分布図
- ある物理量がどのように分布しているかを、値の範囲を**ビン (bin)**に区切って表示したもの
- 実験での使用例
  - ▶ 光検出器の波高分布 (ポアソン分布と正規分布)
  - ▶ 崩壊時間や飛程の分布 (指数分布)
- 分布同士の比較、理論曲線との比較によく使われる

# 大事なこと

---

- 積分すると総数になる
  - ▶ 標本の大きさ (sample size)
    - ▶ 総測定回数や総発生事象 (トリガーした宇宙線粒子のエネルギー分布など)
    - ▶ 全数測定の総数 (国勢調査、実験装置の全数調査など)
    - ▶ 標本数 (number of samples) とは言わないので注意
  - ▶ 確率密度関数の場合は 100% や 1
    - ▶ 十分に標本が大きい (=統計誤差の小さい) MC シミュレーションで得られた物理量の分布や理論曲線など
- 面積に意味があるので原則として縦軸のゼロを表示する (対数表示の場合はもちろん不可能)
- 全数調査と標本調査は分布が異なる

# 間違った表示の例



- ヒストグラムを折れ線グラフにしない
  - ▶ ビンの中心値はそのビンの代表値ではない
  - ▶ 面積が保存しない
  - ▶ (多くの場合) 折れ線の傾きに物理的な意味がない
  - ▶ 誤差棒が大きい場合、傾きを見せるのは読者の誤解を誘発する

# 1 次元ヒストグラム

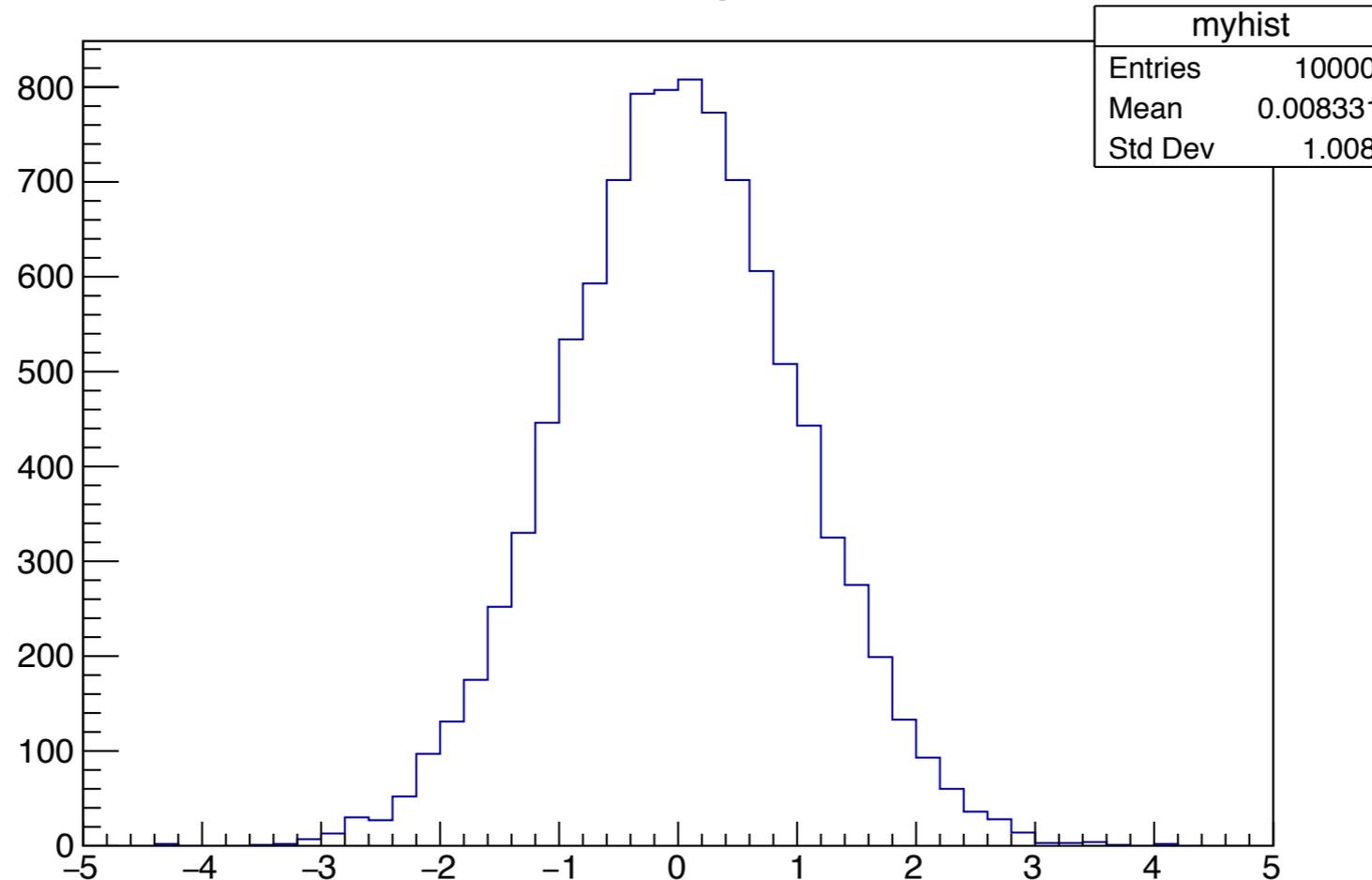
# TH1 クラス

---

- ROOT の 1 次元ヒストグラムは TH1 というクラス
- ヒストグラムの縦軸のデータ型に応じて複数の派生クラスがある
  - ▶ **TH1D** – double (14 桁まで扱える、多分一番よく使う)
  - ▶ TH1F – float (7 桁)
  - ▶ TH1C – char (-128 – 127)
  - ▶ TH1S - 16 bit int (short) (-32768 – 32767)
  - ▶ TH1I – 32 bit int (-2147483648 – 2147483647)
- TH1D 以外はひとまず忘れて良い

# 前回の復習

Gaussian Histogram ( $\sigma = 1$ )

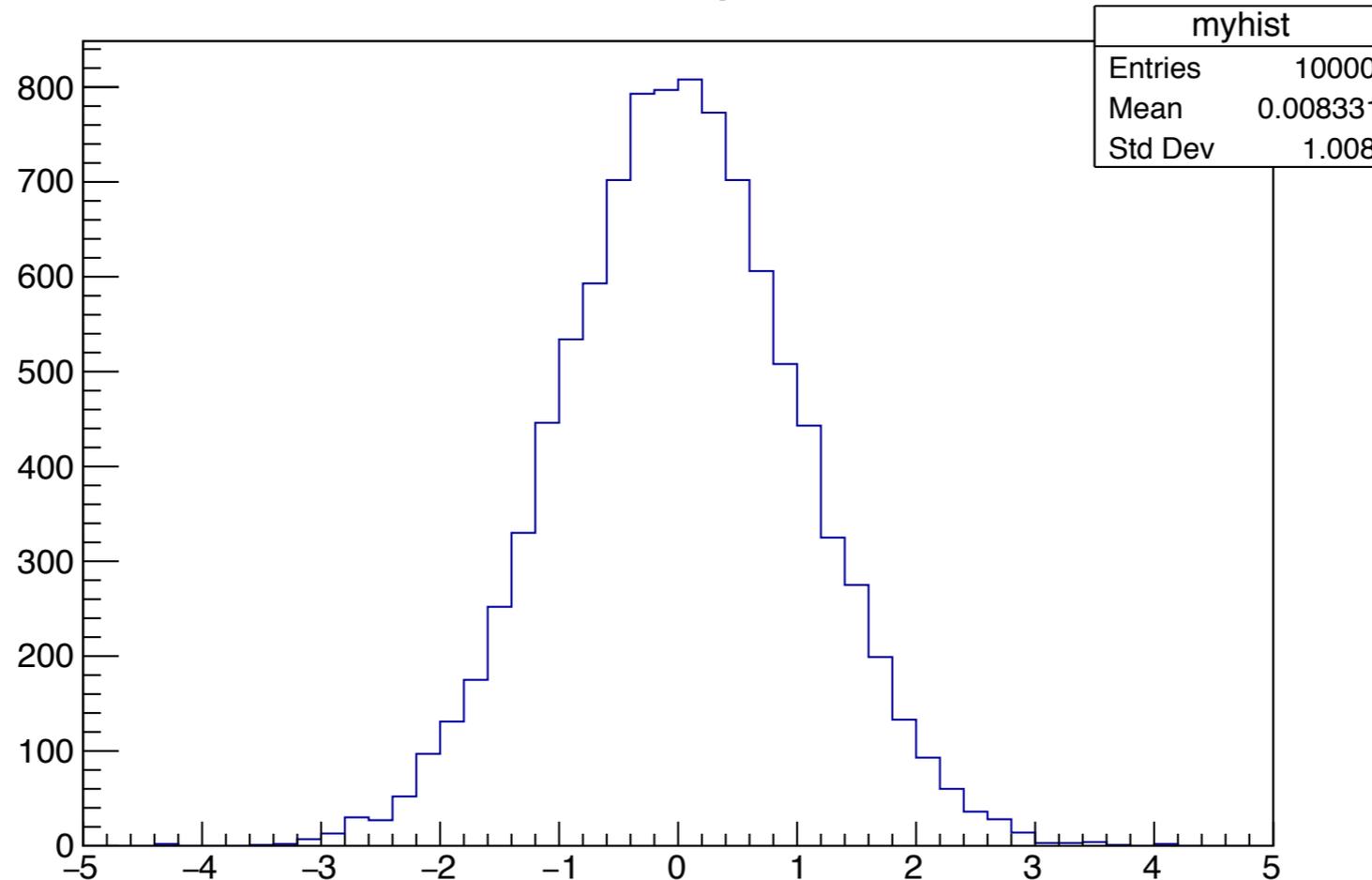


```
$ root
root [0] TH1D* hist
         = new TH1D("myhist",
                   "Gaussian Histogram (#sigma = 1)",
                   50,
                   -5,
                   5)
root [1] hist->FillRandom("gaus", 10000)
root [2] hist->Draw()
```

- ① 名前 (なくてもよい)
- ② タイトル (なくてもよい)
- ③ ビン数
- ④ 下限値
- ⑤ 上限値
- ⑥ 標準偏差 1 の正規分布を乱数で  $10^4$  回詰める

# 自分で $10^4$ 回詰める

Gaussian Histogram ( $\sigma = 1$ )



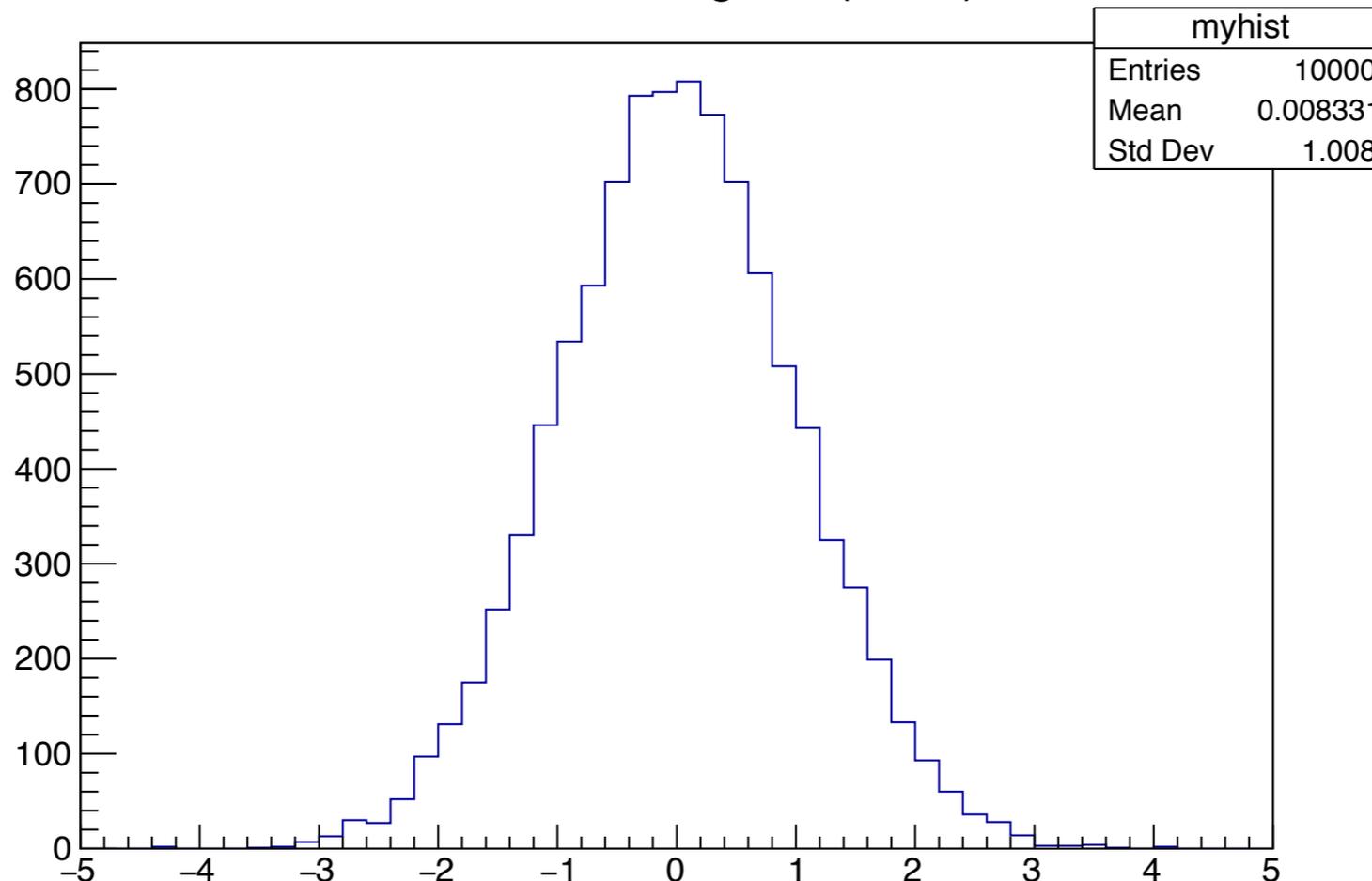
```
$ root
root [0] TH1D* hist = new TH1D("myhist", "Gaussian Histogram (#sigma = 1)", 50,
-5, 5)
root [1] for(Int_t i = 0; i < 10000; i++){
root (cont'ed, cancel with .@) [2] Double_t x = gRandom->Gaus();
root (cont'ed, cancel with .@) [3] hist->Fill(x);
root (cont'ed, cancel with .@) [4]}
root [5] hist->Draw()
```

- ① 乱数を生成し
- ② 詰める

※ 実際には、測定値などを詰める

# ヒストグラムの基本的な量

Gaussian Histogram ( $\sigma = 1$ )



```
root [6] hist->GetEntries()
(Double_t) 10000.0
root [7] hist->GetMean()
(Double_t) 0.008331
root [8] hist->GetStdDev()
(Double_t) 1.008
root [9] hist->GetMeanError()
(Double_t) 0.00997350
root [10] hist->GetStdDevError()
(Double_t) 0.00705233
```

① 総数

② 標本の平均値 (母集団の真の平均値ではない)

③ 標本の標準偏差 (standard deviation)

④ 平均値の統計誤差

⑤ 標準偏差の統計誤差

# 平均値、分散、標準偏差

- 平均値：通常、ある物理量の相加平均（母平均は  $\mu$ ）

$$\text{標本平均} \quad \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

$$\text{母平均} \quad \mu = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N x_i$$

- 分散：その分布の散らばり具合を示す

$$\text{(不偏)標本分散} \quad s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \quad \text{※ ROOT は } N \text{ で割っている}$$

$$\text{母分散} \quad \sigma^2 = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

- 標準偏差：散らばり具合を物理量と同じ次元で示す

標本の標準偏差  $S$       母集団の標準偏差  $\sigma$

# RMS と混同しないこと

---

- RMS（二乗平均平方根）と標準偏差は定義が異なります

$$\text{RMS} = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2} \quad ※ \text{平均を引かない}$$

- PAW や ROOT ユーザの多くが混同しているので注意
  - ▶ PAW が最初に間違い、ROOT は意図的に (?) 間違いを継承した
  - ▶ 最新の ROOT では、RMS という言葉はもう使われな

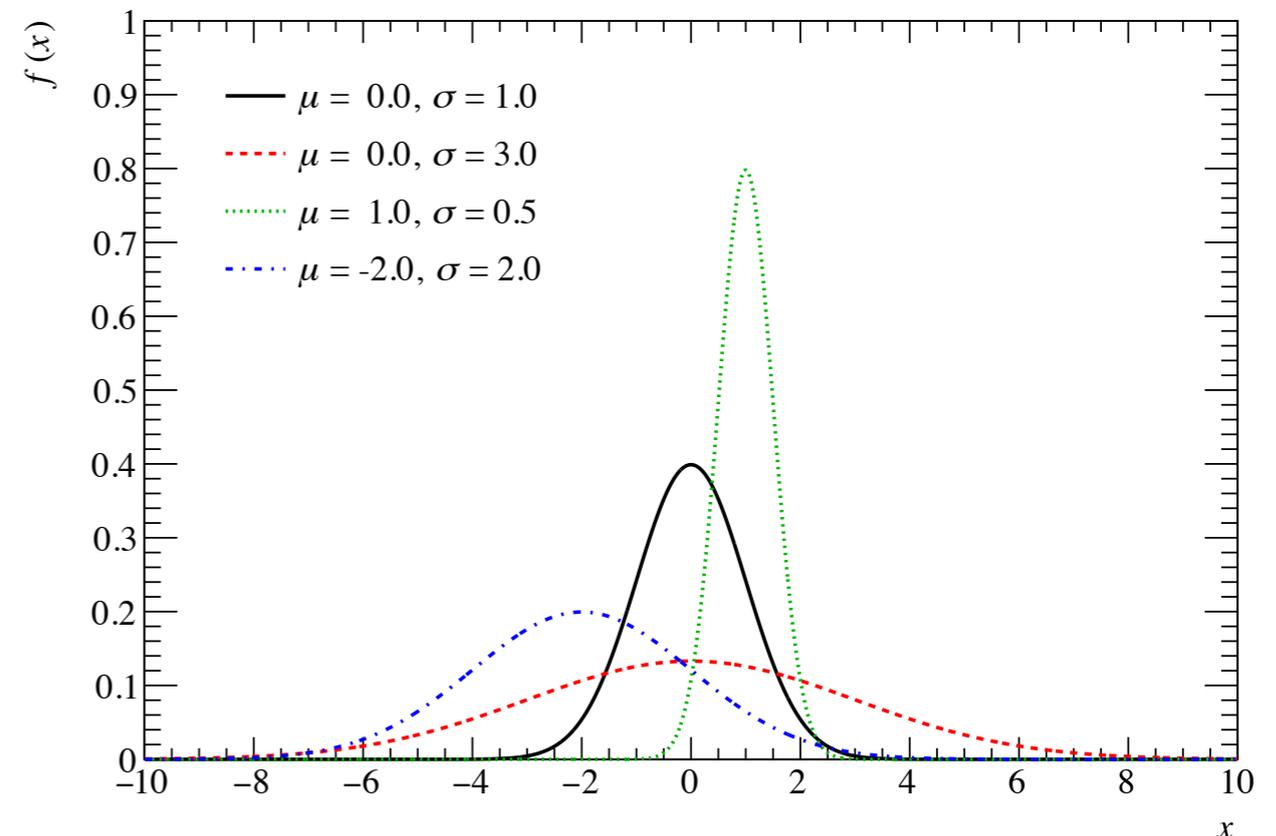
# 正規分布

# 正規分布 (Normal Distribution) とは

- ガウス分布 (Gaussian distribution) とも
- 平均値  $\mu$  と分散  $\sigma^2$  (もしくは標準偏差  $\sigma$ ) で表される

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

- 我々が最も頻繁に使う分布
- 多数の確率過程が組み合わさった場合、結果として出てくる物理量が正規分布に従う (中心極限定理)
- 面積一定の場合、高さとは幅は  $1/\sigma$  と  $\sigma$  に比例する



# GetMeanError と GetStdDevError

---

- 母集団の分布や理論的な分布が正規分布であったとしても、限られた実験データ（標本）は母集団を完全に再現しない
- 標本から得られる平均値や標準偏差は、真の値とはずれる
- TH1::GetMeanError と GetStdDevError は、それぞれの推定量を返す
- 正規分布の場合、物理量  $x$  に対し次の推定量の誤差がある

$$\delta \bar{x} = \frac{s}{\sqrt{N}}$$

$$\delta s = \frac{s}{\sqrt{2N}}$$

# 確かめてみる

```
$ cat StandardError.C
void StandardError() {
  const Int_t kSampleSize = 10000;
  const Int_t kRepeat = 10000;
  const Double_t kMean = 0.;
  const Double_t kSigma = 1.;

  TH1D* hMeanError = new TH1D("hMeanError", ";<math>x</math>", 100, -0.05, 0.05);
  TH1D* hStdDevError = new TH1D("hStdDevError", ";#math>#sigma_{x}</math>", 100, -0.05,
0.05);

  for(Int_t i = 0; i < kRepeat; i++){
    TH1D h("", "", 100, -5, 5);
    for(Int_t j = 0; j < kSampleSize; j++){
      Double_t x = gRandom->Gaus(kMean, kSigma);
      h.Fill(x);
    }
    hMeanError->Fill(h.GetMean() - kMean);
    hStdDevError->Fill(h.GetStdDev() - kSigma);
  }

  TCanvas* can = new TCanvas("can", "can", 1200, 600);
  can->Divide(2, 1, 1e-10, 1e-10);
  can->cd(1);
  hMeanError->Draw();
  can->cd(2);
  hStdDevError->Draw();
}
```

① 平均  $\mu = 0$ 、標準偏差  $\sigma = 1$

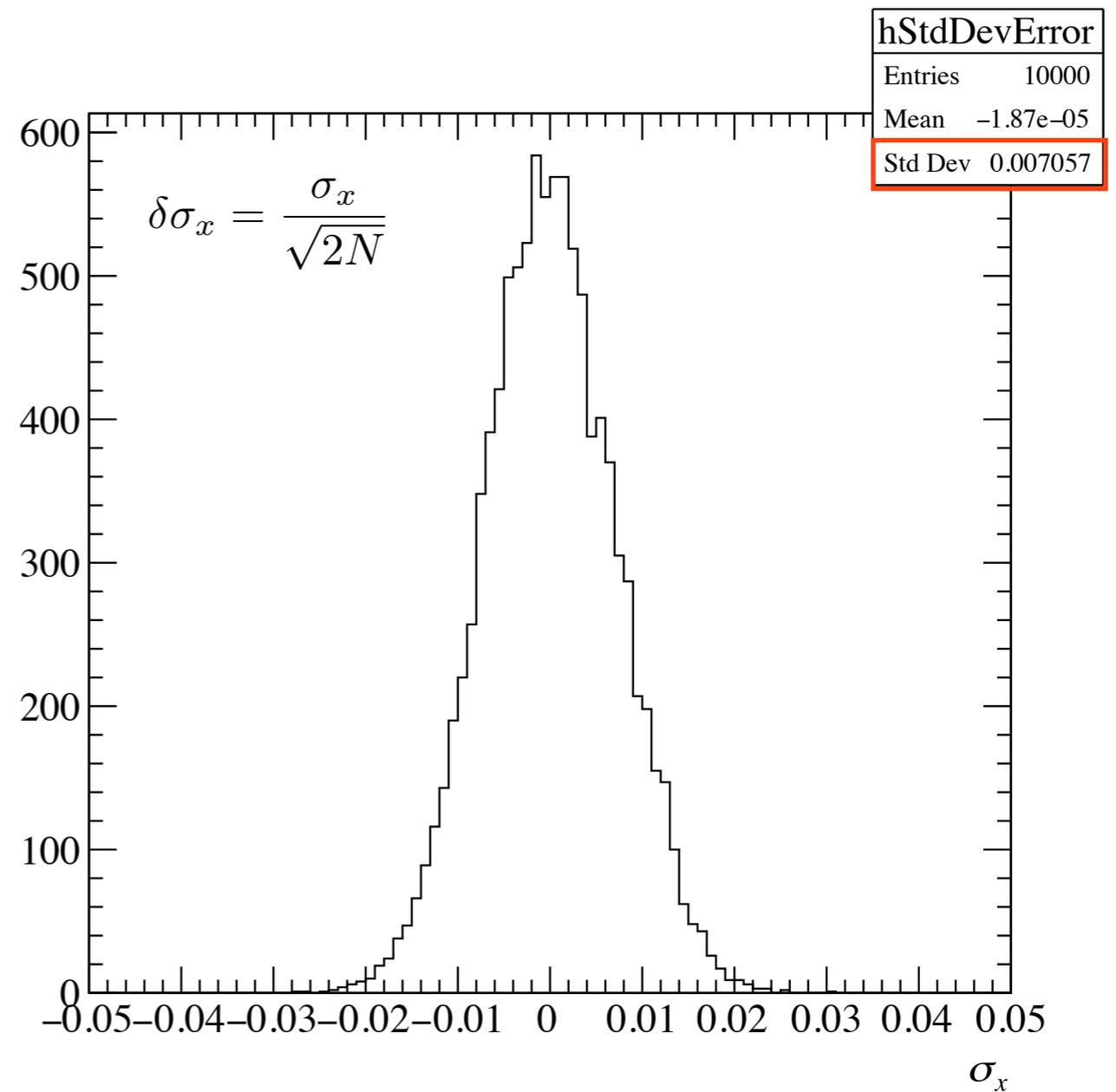
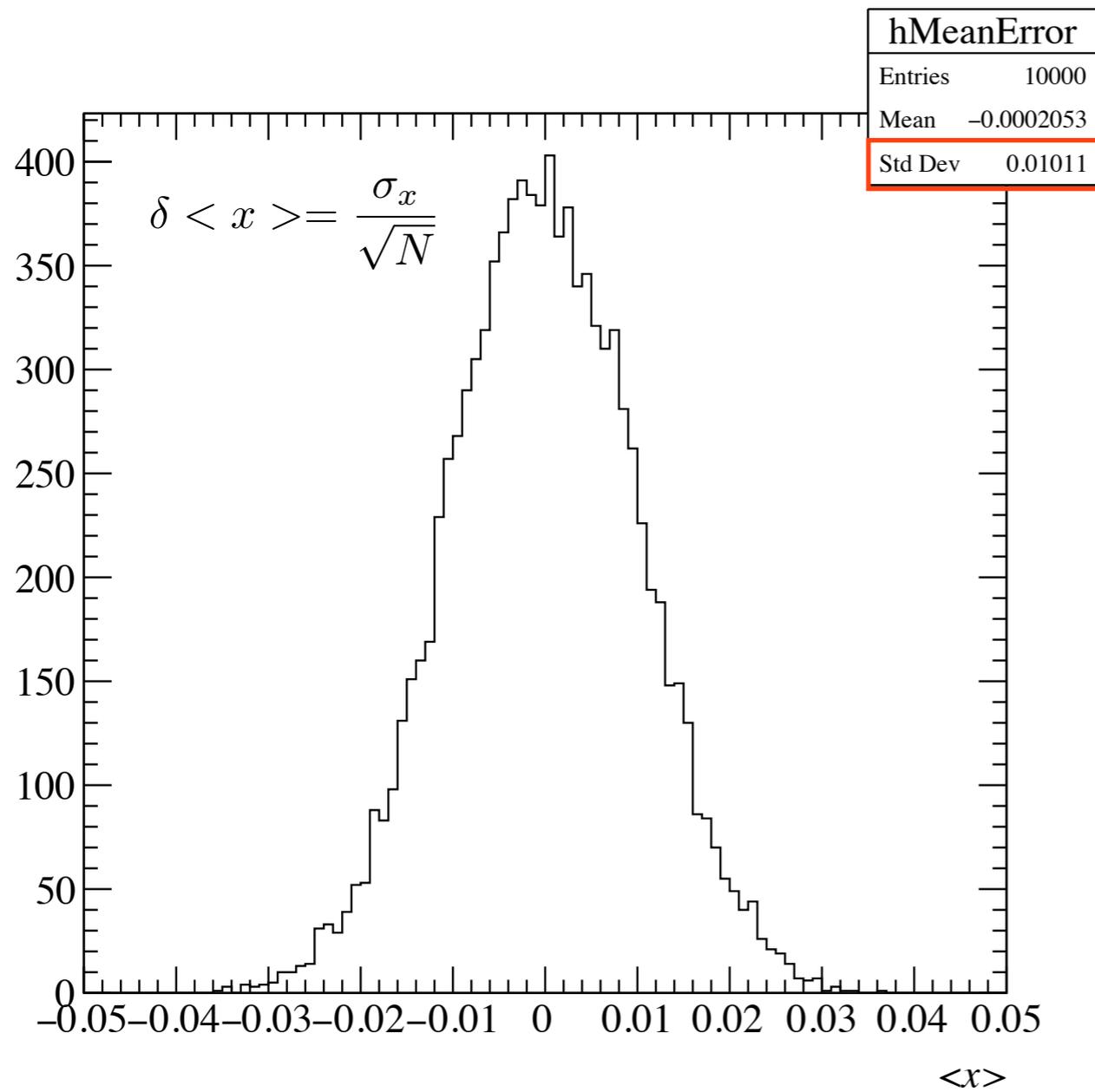
② 真の値からどれだけずれたかを詰めるヒストグラム

③  $\mu = 0$ 、 $\sigma = 1$  で乱数を 10,000 回生成

④ 標本で得られた  $\bar{x}$  と  $\sigma_x$  の、真値との差を詰める  
これも 10,000 回繰り返し

⑤ Draw する

# 確かめてみる



- $\sigma_x / \sqrt{N} = 1/100 = 0.01$
- $\sigma_x / \sqrt{2N} = 1/(1.414 \times 100) = 0.00707$
- 誤差の範囲で一致している

# 大事なこと

---

- 通常の測定は母集団から標本を抜き出しているだけ
- 真の分布は知りえないので標本から推定する
- 平均値や標準偏差は、標本から計算されたもの
  - ▶ 真の平均値からの誤差は  $\sigma_x/\sqrt{N}$
  - ▶ 真の標準偏差からの誤差は  $\sigma_x/\sqrt{2N}$
- ある確率分布に従う測定があった場合、統計誤差はその分布の標準偏差
- 多数の測定から平均値を求める場合は、統計誤差は  $\sigma_x/\sqrt{N}$

# 注意事項

---

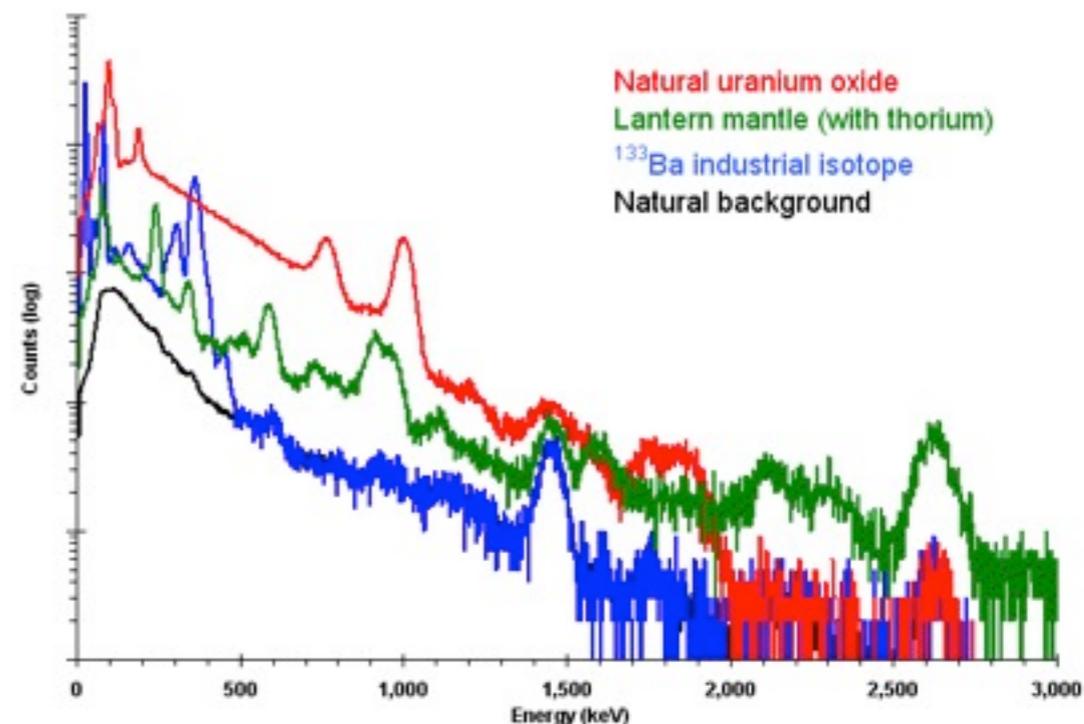
- 実際の実験データを解析する場合、真に正規分布であることはほとんどない
  - ▶ 正規分布は正負の無限大の値を取りうるが、実際の測定でそのような値は取りえない
  - ▶ 光電子増倍管の出力波高を正規分布と仮定することがあるが、負のゲインはありえない
- ROOT で横軸の表示範囲を変更すると、平均値や標準偏差が表示範囲のみで再計算される

# ポアソン分布 (準備間に合わず)

# フィッティング

# ヒストグラムのフィッティング

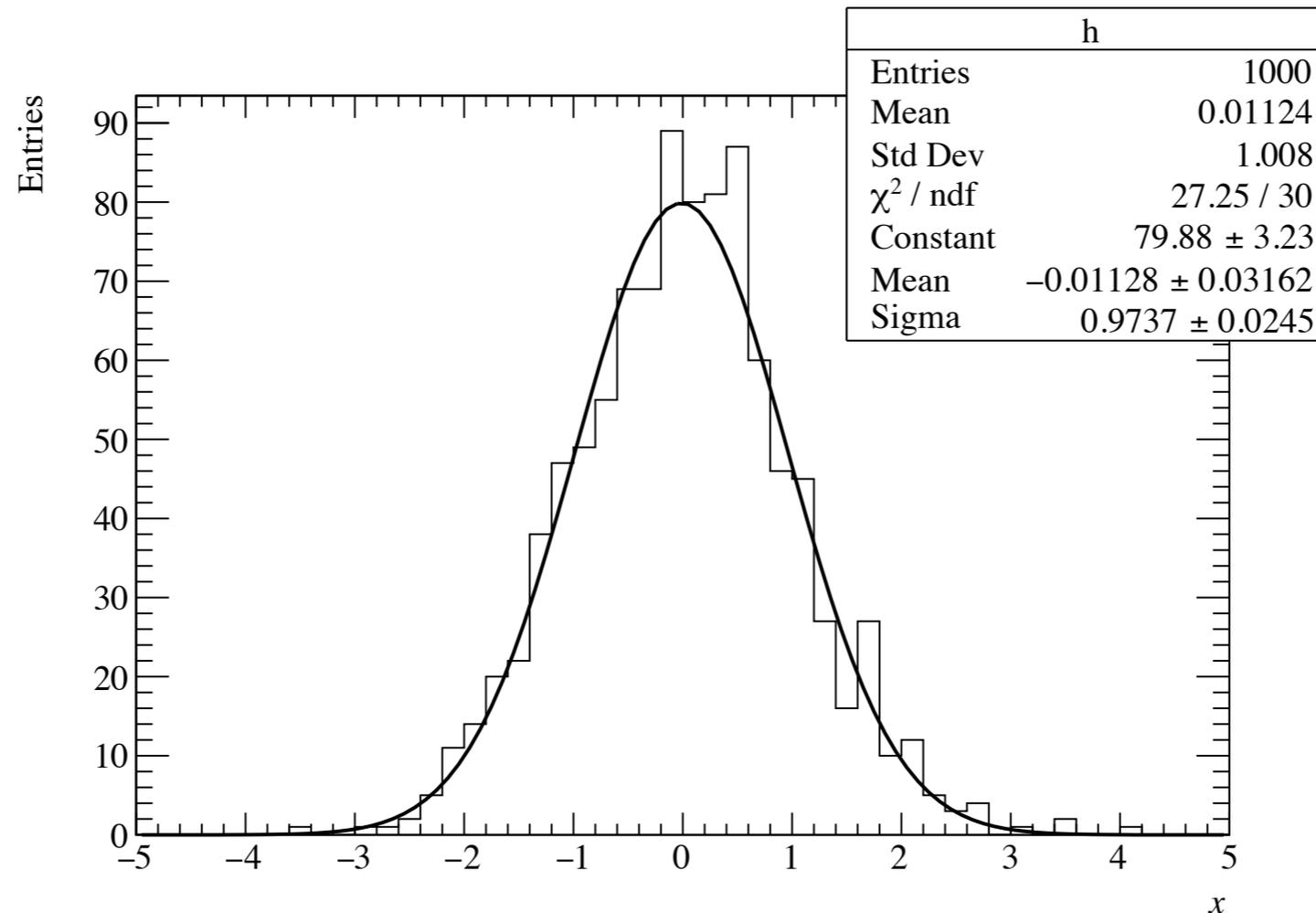
- 実験で得られたヒストグラムから物理量を抜き出すとき、単純な1つの正規分布であることは少ない
  - ▶ 複数のピークの存在するデータ
  - ▶ バックグラウンドを含むデータ
- ヒストグラムをよく再現するモデル関数を作り、フィッティング (fitting、曲線のアてはめ) を行うことで変数 (parameter) を得る



出典 Amptek

<http://amptek.com/products/gamma-rad5-gamma-ray-detection-system/>

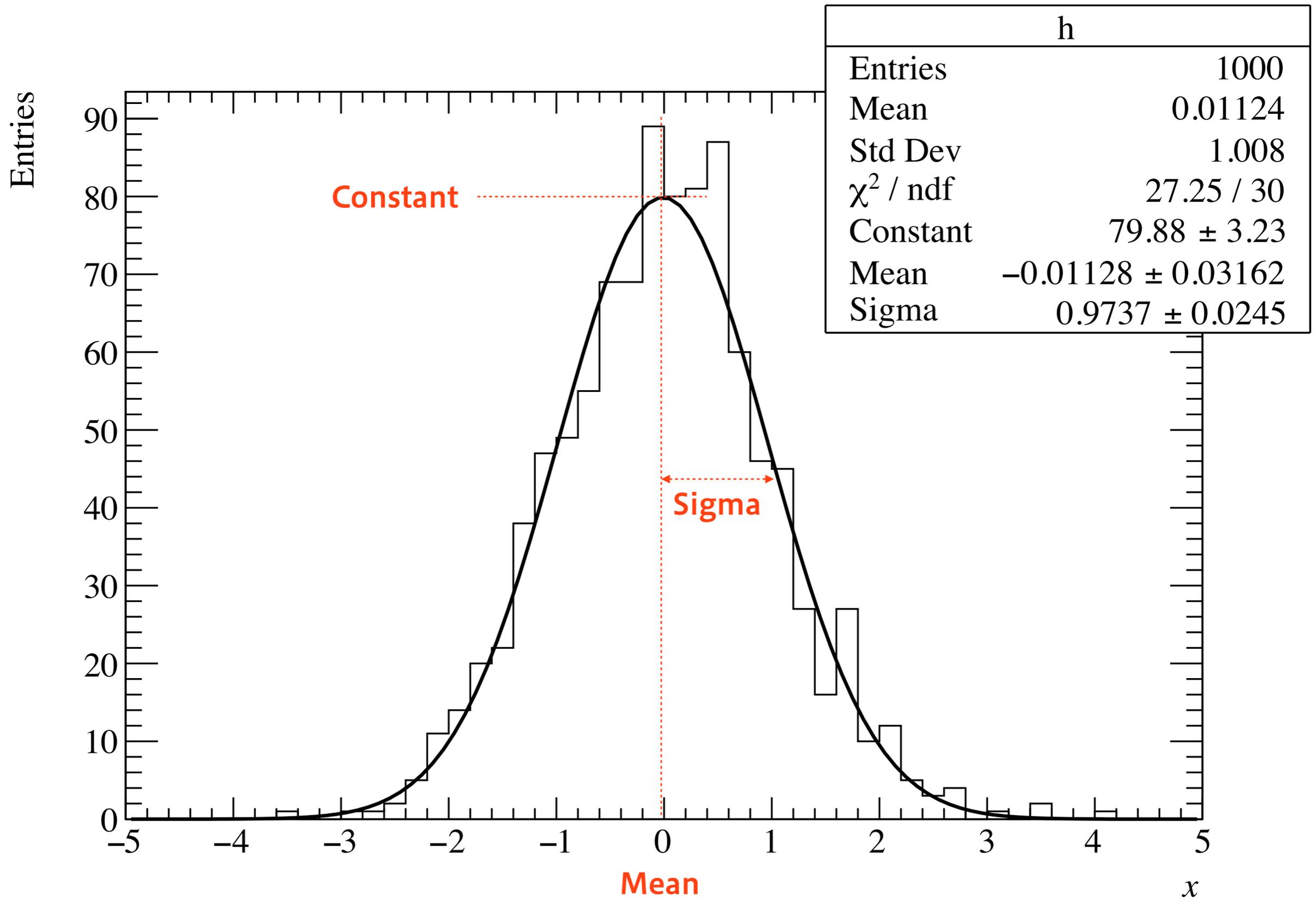
# 単純な例



```
root [0] TH1D* hist = new TH1D("h", ";#it{x};Entries", 50, -5, 5)
root [1] hist->FillRandom("gaus", 1000)
root [2] hist->Fit("gaus")
```

```
FCN=27.2533 FROM MIGRAD      STATUS=CONVERGED      60 CALLS      61 TOTAL
                        EDM=1.22437e-07      STRATEGY= 1      ERROR MATRIX ACCURATE
```

EXT NO.	PARAMETER NAME	VALUE	ERROR	STEP SIZE	FIRST DERIVATIVE
1	Constant	7.98846e+01	3.22837e+00	6.64782e-03	-1.29981e-05
2	Mean	-1.12836e-02	3.16206e-02	8.19052e-05	-1.55071e-02
3	Sigma	9.73719e-01	2.44588e-02	1.69219e-05	-7.15963e-03



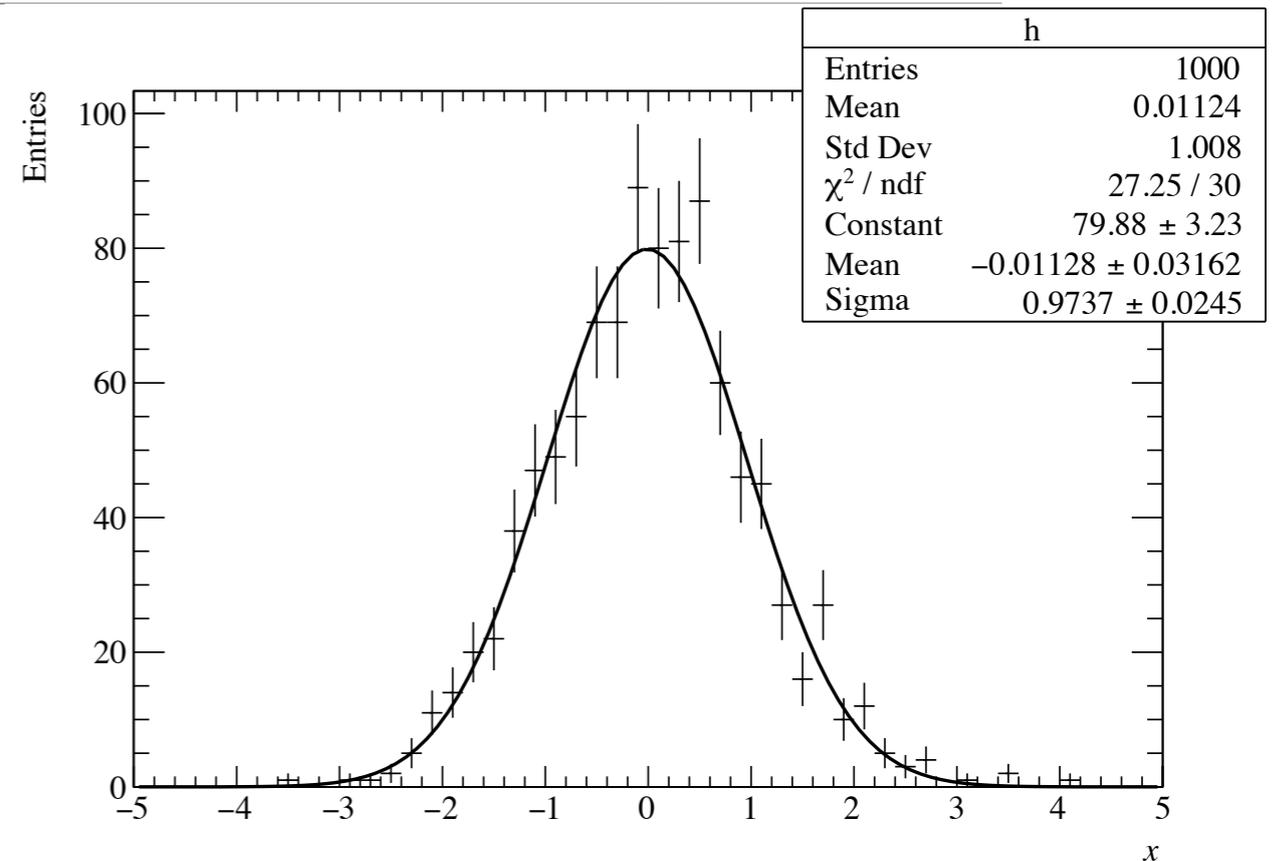
# 変数の比較

	平均	標準偏差
真値	0	1
ヒストグラム	0.011±0.032	1.008±0.023
フィット	-0.011±0.032	0.974±0.025

- 両者とも誤差の範囲で真値を推定できている
- 誤差の大きさは両者で同程度

# ROOT は内部で何をしているか

- 各ビンには統計誤差が存在
  - ▶ そのビンに入る標本の大きさはポアソン分布に従う
  - ▶  $N > 20$  で正規分布と見なせる
  - ▶  $\delta N = \sqrt{N}$  と近似できる



- 最小二乗法を用いて、カイ二乗 ( $\chi^2$ ) を最小にするように、モデル関数の変数空間を探索する

$$\chi^2 = \sum_{i=1}^N \frac{(y_i - f(x_i))^2}{\delta y_i^2}$$

$x_i$  : ビンの中心値

$y_i$  : 各ビンの計数

$f(x_i)$  :  $x_i$  におけるモデル関数の値

$\delta y_i$  :  $y_i$  の誤差

N - 変数の数 : 自由度  $\nu$

- この値はカイ二乗分布と呼ばれる確率密度関数に従う

# $\chi^2$ を最小にする理由

- 最も尤もらしいモデル関数は、測定されたデータ値の分布が最も生じやすい関数のはずでる

- ▶ 各データ点の誤差（ばらつき）は正規分布に従うとする
- ▶ 各データ点の値が出る確率の積が、手元の標本になる確率になると見なす

$$\text{Prob.} \propto \prod_{i=1}^N \frac{1}{\sqrt{2\pi\delta y_i^2}} \exp \left[ -\frac{(y_i - f(x_i))^2}{2\delta y_i^2} \right]$$

$$\propto \exp \left[ -\sum_{i=1}^N \frac{(y_i - f(x_i))^2}{2\delta y_i^2} \right]$$

$$= \exp(-\chi^2)$$

- 結局、 $\chi^2$  を最小にするのが、確率最大になる

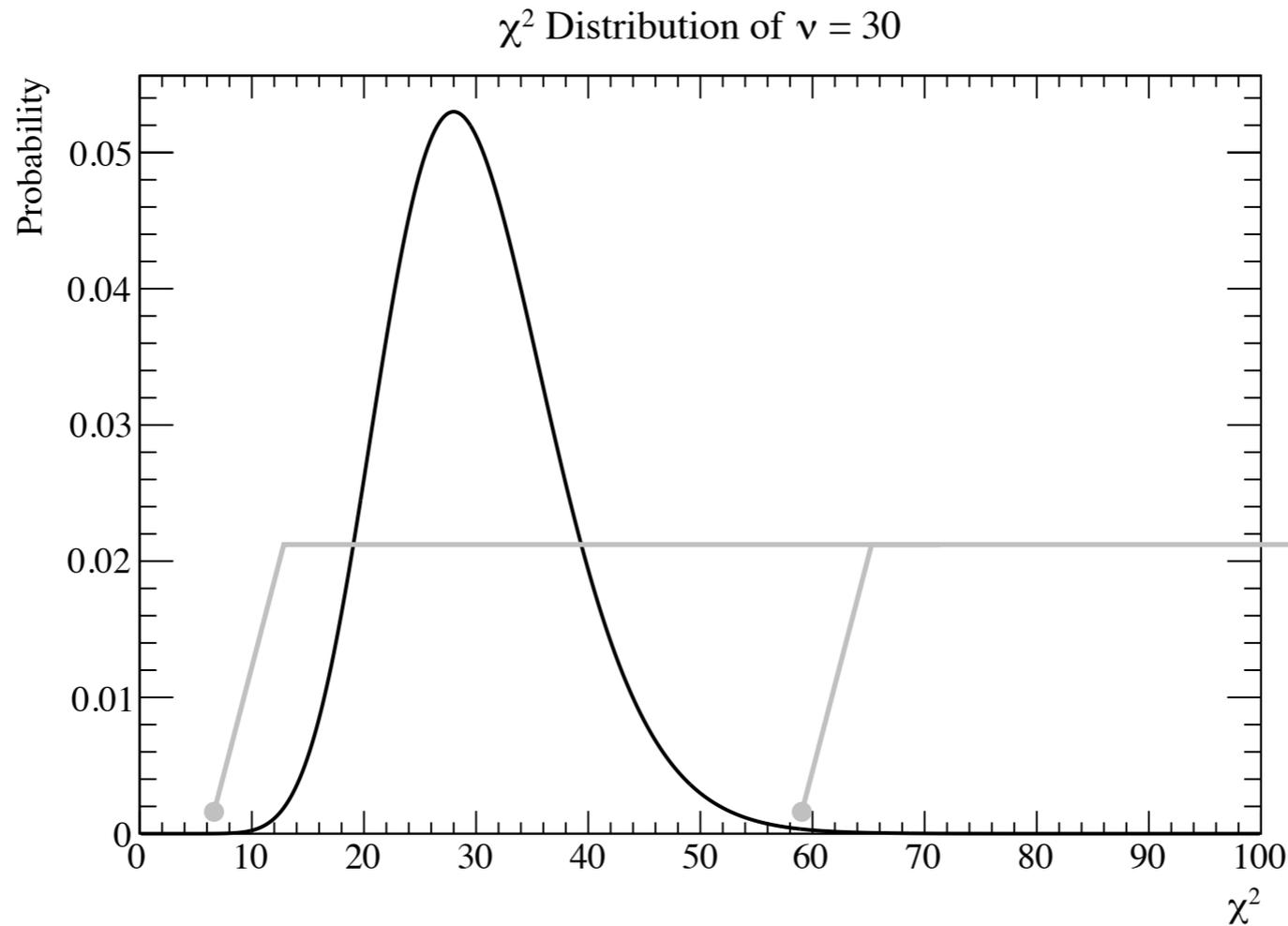
# カイ二乗分布

---

- 自由度  $\nu$  のカイ二乗の値は、カイ二乗分布に従う

$$P_{\nu}(\chi^2) = \frac{(\chi^2)^{\nu/2-1} e^{-\chi^2/2}}{\Gamma(\nu/2) 2^{\nu/2}}$$

# カイ二乗分布と p 値

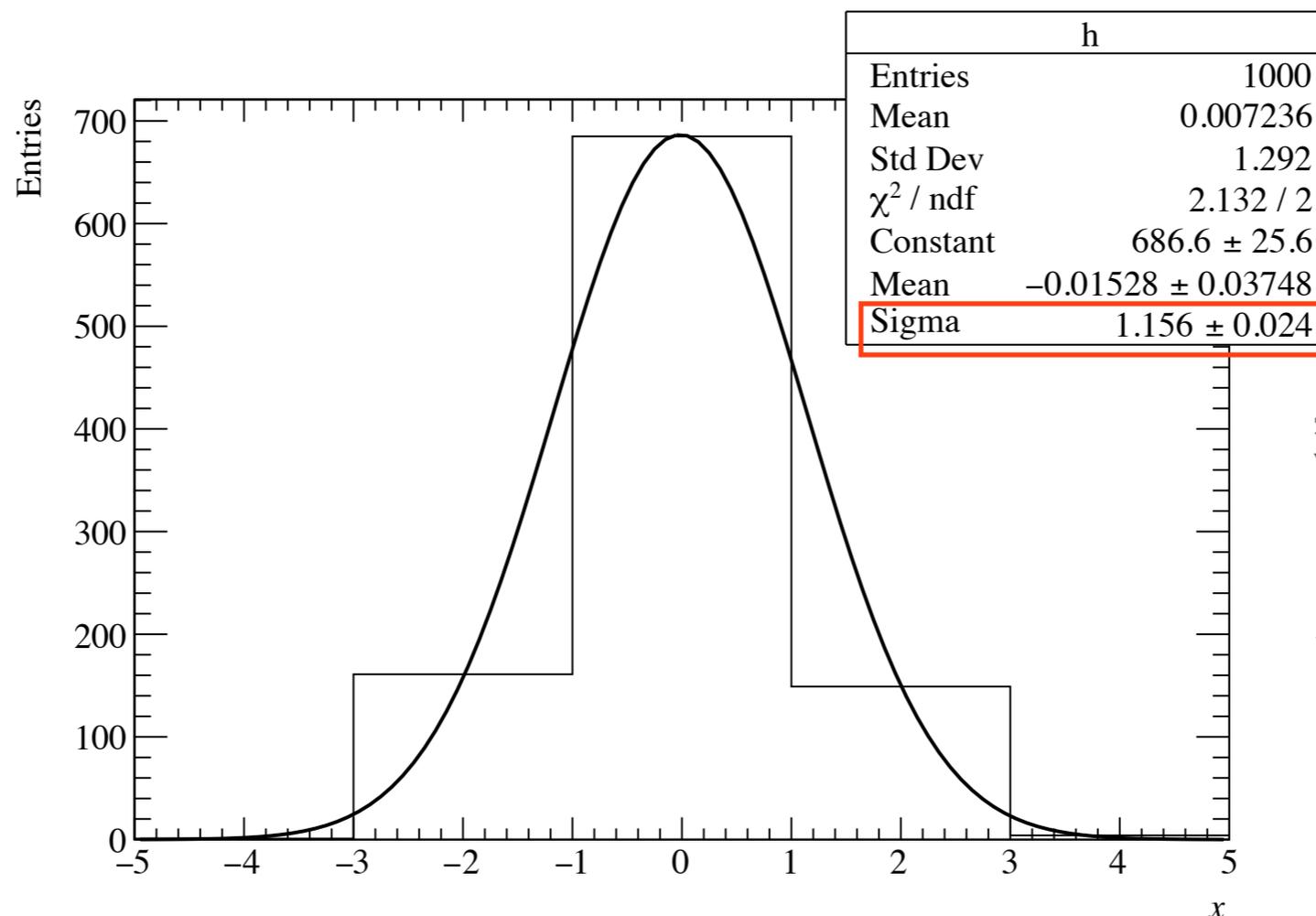


このあたりに来ると  
確率としてありえない  
 $p < 0.01$  や  $p > 0.99$   
くらいの場合、誤差の  
評価が正しいか要確認

```
$ root
root [0] TF1* pdf = new TF1("pdf", "ROOT::Math::chisquared_pdf(x, [0], 0)", 0,
100)
root [1] pdf->SetTitle("#chi^{2} Distribution of #nu = 30;#chi^{2};Probability")
root [2] pdf->SetParameter(0, 30)
root [3] pdf->SetNpx(500)
root [4] pdf->Draw()
root [5] TMath::Prob(27.25, 30)
(Double_t) 0.610115
```

- ① カイ二乗分布の 1 次元関数 TF1 を作る
- ② 自由度  $\nu = 30$  に設定
- ③ TF1 の点数を増やし表示を滑らかに (本質的でない)
- ④ 確率の計算  
 $\nu = 30$ 、 $\chi^2 = 27.25$  の場合、 $p = 0.61$

# モデル関数に比べてビン幅が広過ぎる場合

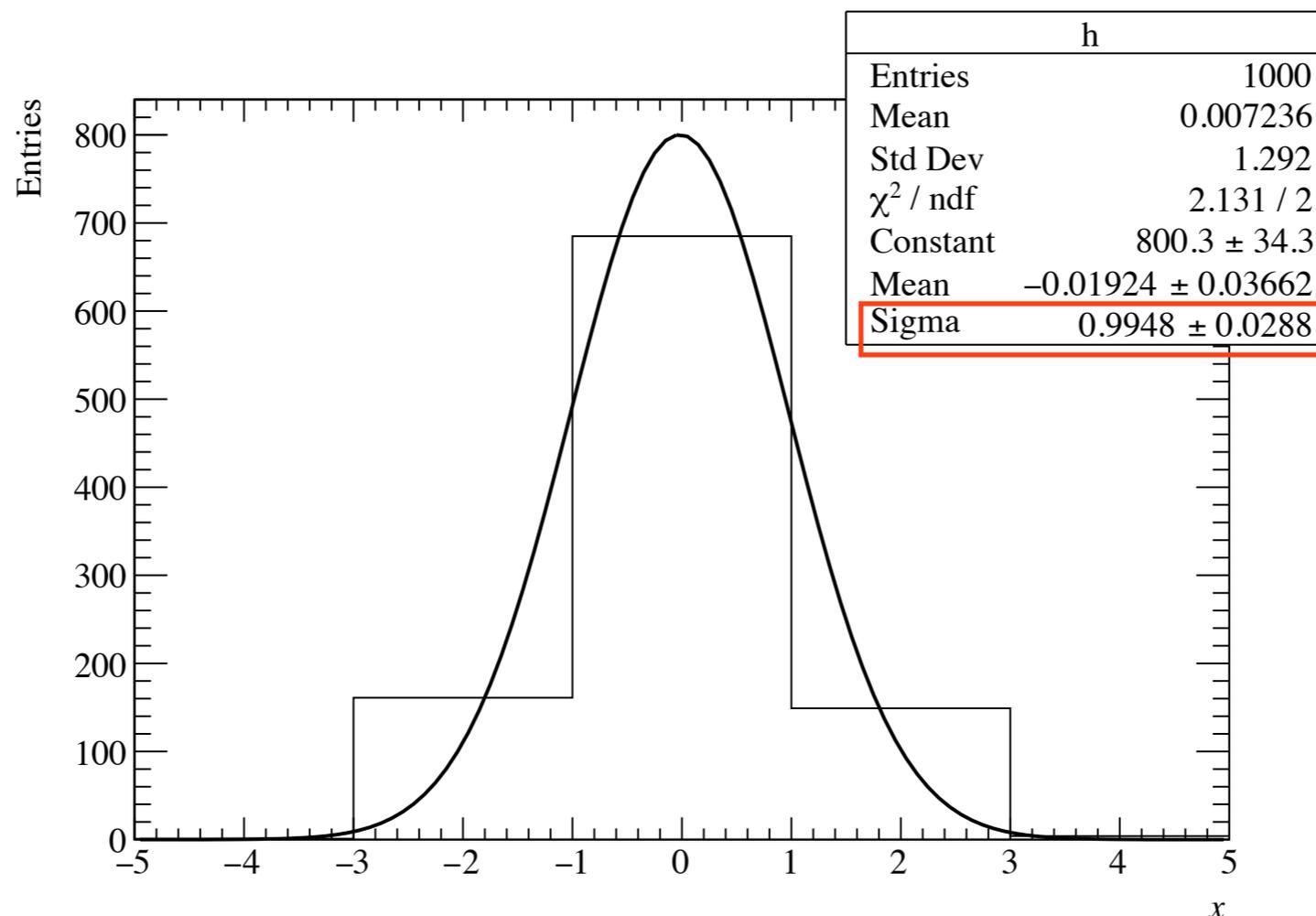


変数の推定を誤る！

ROOT がビンの中心値で  
カイ二乗を計算するため

```
root [0] TH1D* hist = new TH1D("h", ";#it{x};Entries", 5, -5, 5)
root [1] hist->FillRandom("gaus", 1000)
root [2] hist->Fit("gaus")
FCN=2.13212 FROM MIGRAD      STATUS=CONVERGED      52 CALLS      53 TOTAL
          EDM=2.37573e-07    STRATEGY= 1      ERROR MATRIX ACCURATE
EXT  PARAMETER
NO.  NAME      VALUE      ERROR      STEP      FIRST
   1  Constant  6.86581e+02  2.55989e+01  1.87885e-02  -1.43368e-05
   2  Mean      -1.52834e-02  3.74843e-02  3.22360e-05  8.88105e-03
   3  Sigma      1.15649e+00  2.36229e-02  4.99586e-06  -1.09181e-01
```

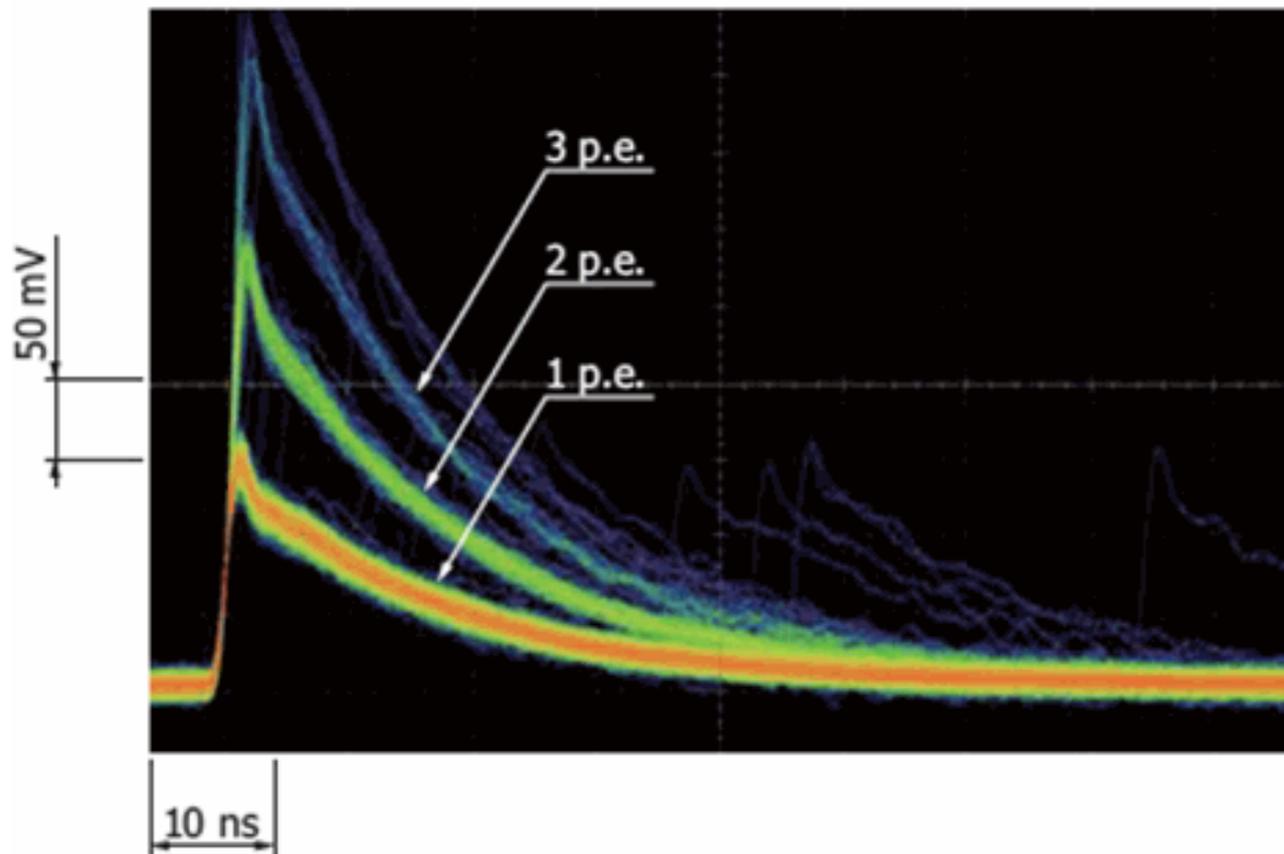
# “i” (integral) オプションを使う



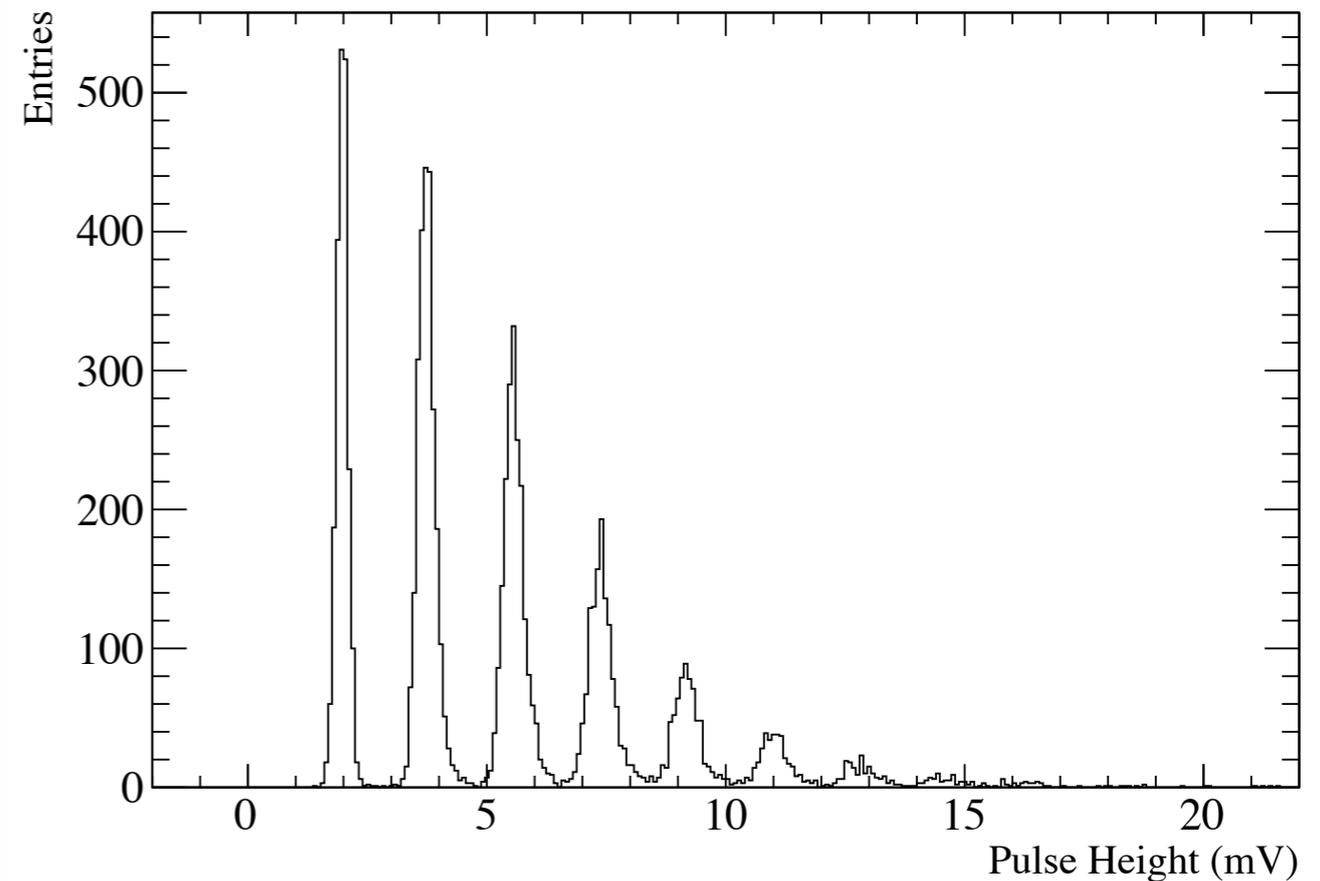
```
root [0] TH1D* hist = new TH1D("h", ";#it{x};Entries", 5, -5, 5)
root [1] hist->FillRandom("gaus", 1000)
root [2] hist->Fit("gaus", "i")      “i”を追加
FCN=2.13123 FROM MIGRAD      STATUS=CONVERGED      104 CALLS      105 TOTAL
EDM=2.22157e-07      STRATEGY= 1      ERROR MATRIX ACCURATE
EXT PARAMETER      STEP      FIRST
NO.  NAME      VALUE      ERROR      SIZE      DERIVATIVE
1  Constant      8.00322e+02      3.43377e+01      2.18847e-02      -1.06589e-05
2  Mean      -1.92391e-02      3.66159e-02      3.15299e-05      -1.19143e-03
3  Sigma      9.94826e-01      2.88088e-02      5.67273e-06      -9.54913e-02
```

# 実験室におけるデータ例

# 正規分布でのフィット例



半導体光検出器の出力波形例  
(浜松ホトニクス)

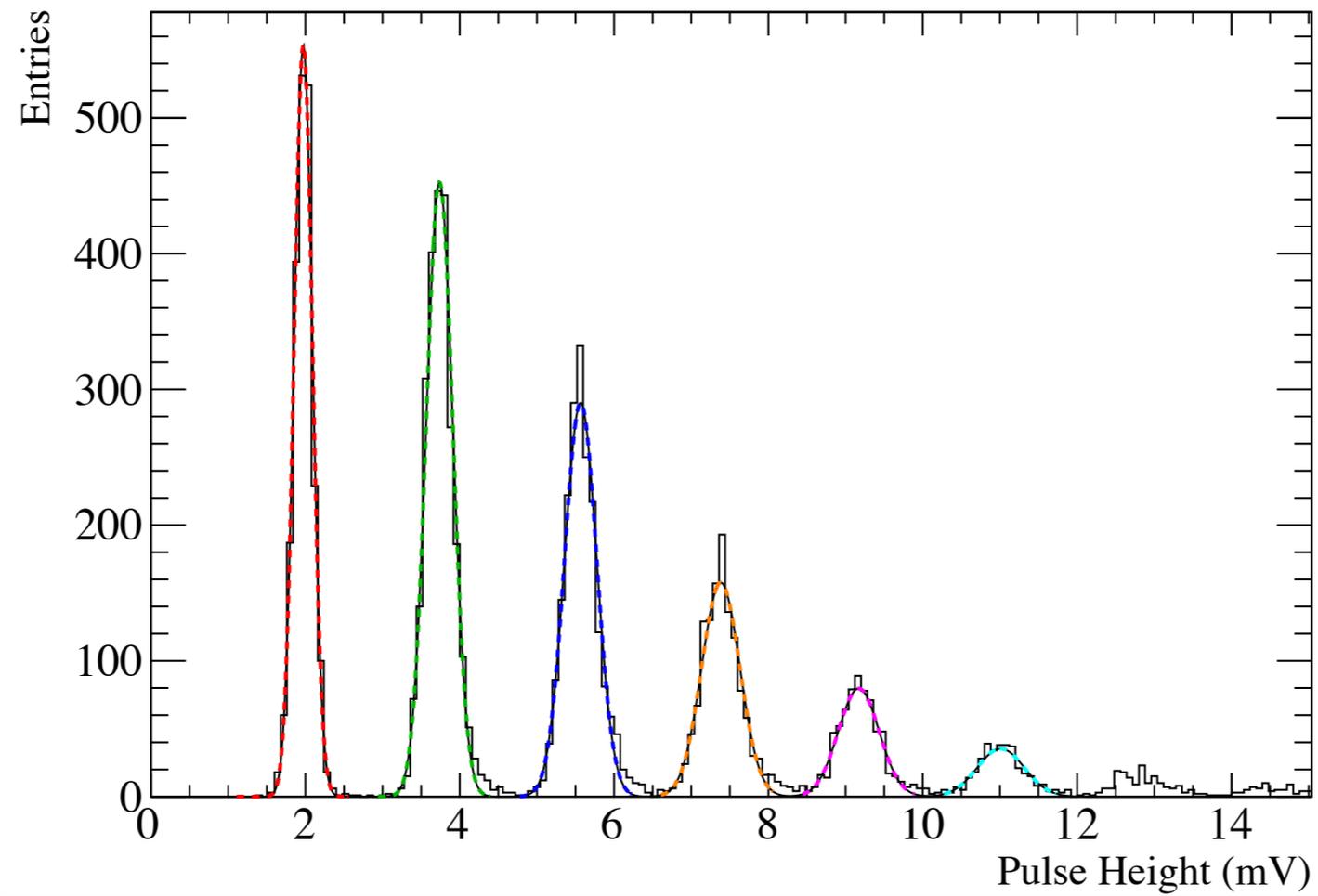


半導体光検出器の出力波高分布例  
(データ提供：日高直哉)

[http://www.hamamatsu.com/us/en/community/optical\\_sensors/sipm/physics\\_of\\_mppc/index.html](http://www.hamamatsu.com/us/en/community/optical_sensors/sipm/physics_of_mppc/index.html)

- 光検出器の出力電荷や波高分布は、正規分布でよく近似できる場合が多い
- 半導体光検出器の場合、光電変換された光電子数に比例して波高が綺麗に分かれる
- 光電子数分布や利得 (gain) の評価に正規分布でのフィット

# 複数の正規分布によるフィットの例



```
$ root
root [0] .x MppcFit.C
```

## 第2回のまとめ

---

- ヒストグラムとは何か
- TH1 を使った ROOT でのヒストグラムの例
- 正規分布
- カイ二乗分布と確率
- ROOT での 1 次元ヒストグラムのフィッティング
  
- 分からなかった箇所は、各自おさらいしてください