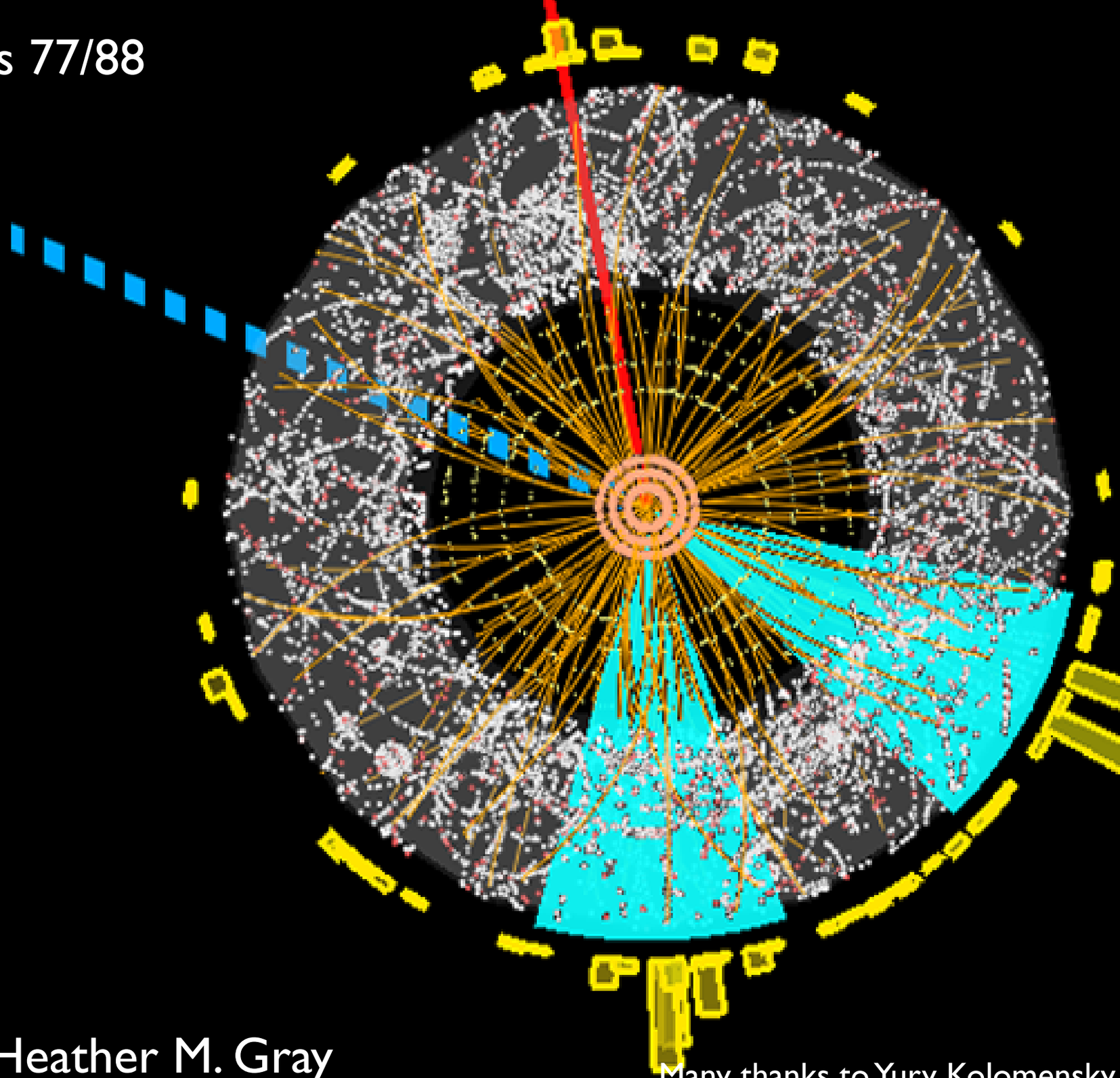
Introduction to Computational Techniques in Physics/Data Science Applications in Physics

Physics 77/88



Statistics and Probability, Interpreting Measurements

There are three kinds of lies: lies, damned lies and statistics
- Mark Twain

Prof. Heather M. Gray

Many thanks to Yury Kolomensky for material used in this course

The Statistics Boot Camp

- Introduction
- Definitions: results of experiments
  - Random variables, probability, PDFs
- Interpreting results
  - Point estimators
  - Max likelihood, least squares fits
  - Hypothesis testing, confidence limits
- Simulation (Monte Carlo techniques) — in two weeks

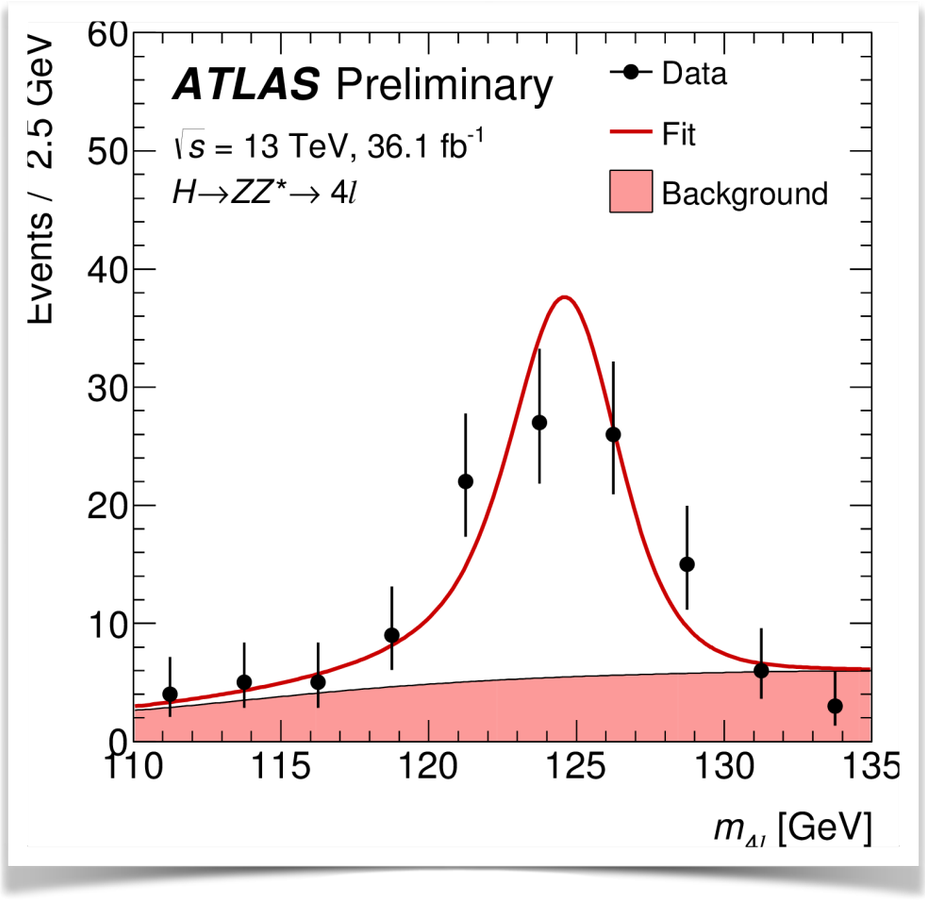
Statistics

This may seem rather dry, but if you turn out to be an experimentalist it's really important and useful!

- Statistics is a vital tool
  - Physics is an experimental science
  - Requires both and understanding
- From to
  - Make a set of
  - Summarise the
  - Most conclusions are drawn with some degree of
- Example
  -
  -
  -
- Many measurements are a priori uncertain and have to be interpreted in probabilistic terms
- **Classical statistics**: estimate given a and test whether a given is with the data

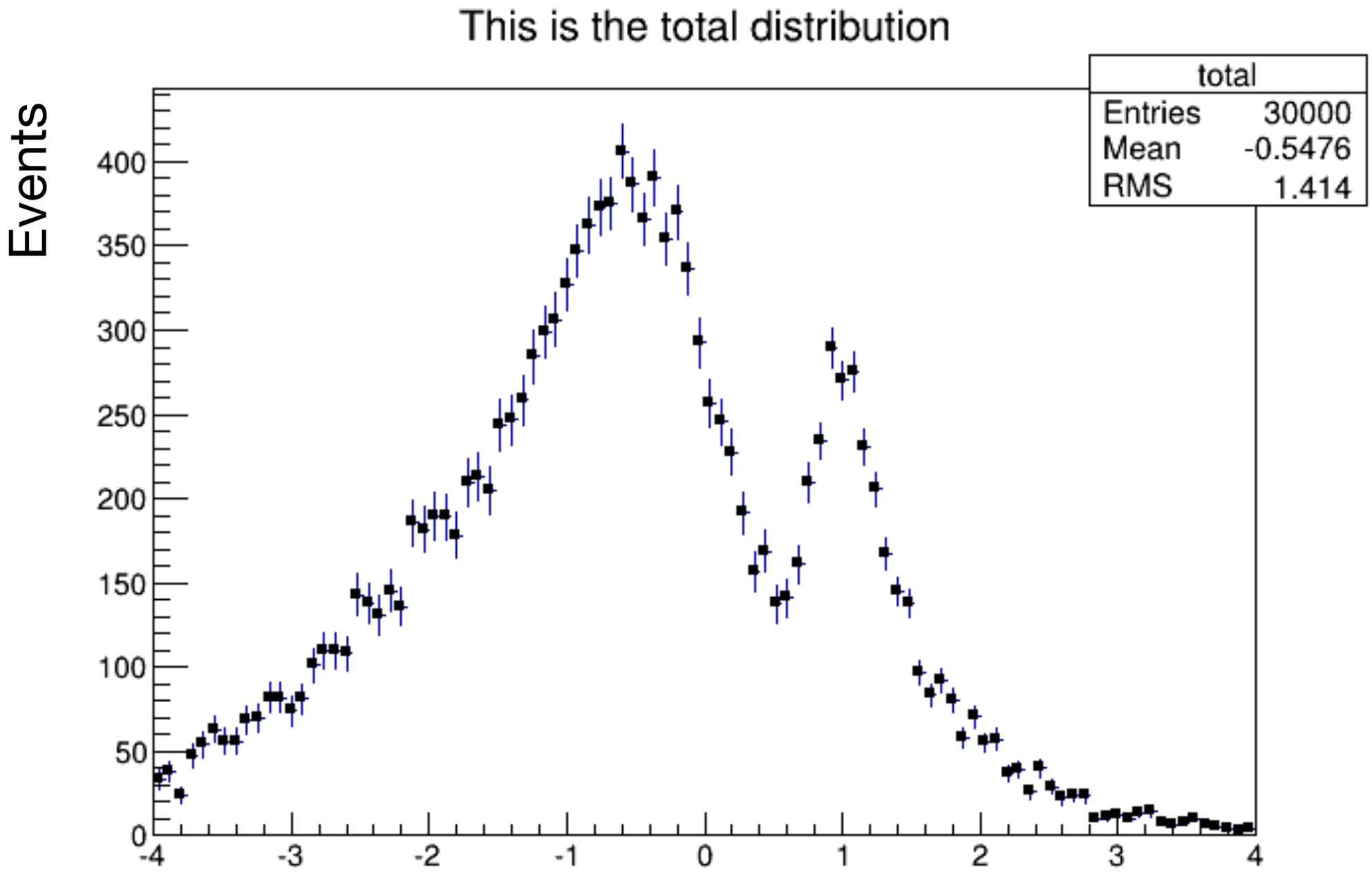
Describing the Data

- **Data**:
  - Physics:
    -
  - Other fields deal with
- **Numbers** are easier to handle mathematically; statistics will deal with quantitative measurements
  - data, e.g. integers (counts)
  - data, e.g. energies, momenta
    - Measure with some , set by the measuring apparatus or other external conditions

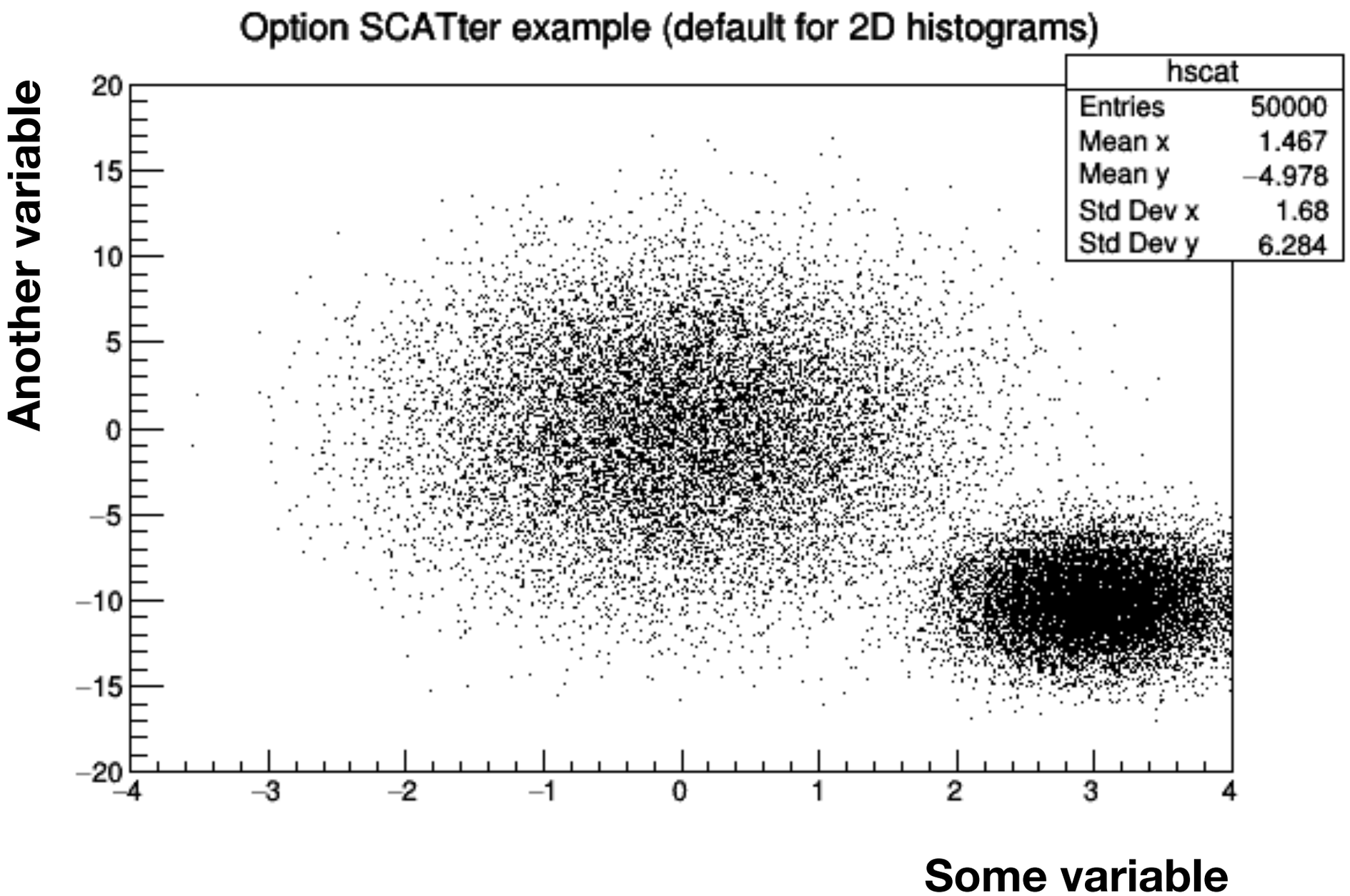


Why are there error bars on the data?

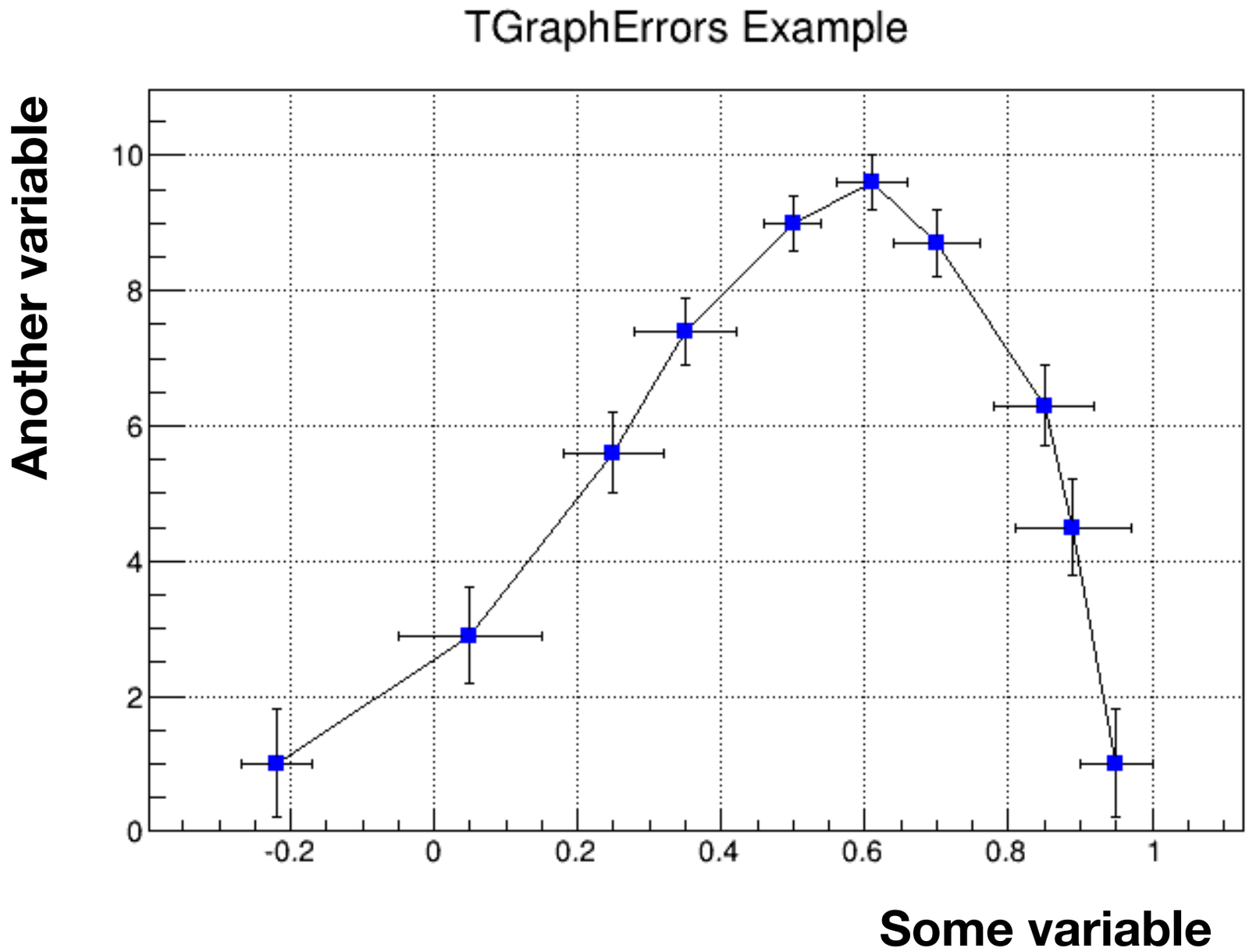
Histogram


This is the total distribution
total
Entries 30000
Mean -0.5476
RMS 1.414
Events
400
350
300
250
200
150
100
50
0
-4
-3
-2
-1
0
1
2
3
4
Some variable

Scatterplots



Graph


TGraphErrors Example
Another variable
Some variable
10
8
6
4
2
0
-0.2
0
0.2
0.4
0.6
0.8
1

Uncertainty and Error

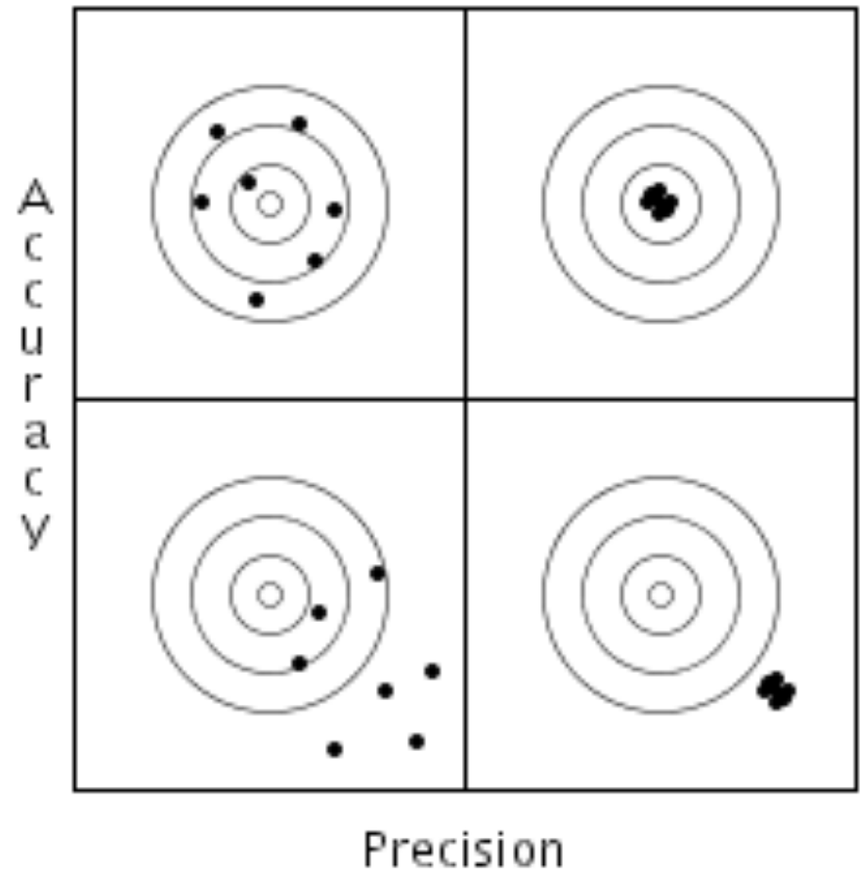
- In physics, the words and are used interchangeably to describe how far a is expected to deviate from the — typically
  - Use symbol for the
  - Formal definition is probabilistic: to find the within of the value
  - Often interpreted as a true values
  - We'll come back to the difference between and

Uncertainty and Error

- How do we what is ?
  - Underlying assumption: our experiment is of a of
    - Derive the from the of the
  - Implicit assumption: our experiment is
    - i.e. all experiments would return results

Precision vs Accuracy

- Precision: of the data around the value
  - Typically associated with uncertainty
- Accuracy: of the value from value
  -
  - Typically associated with uncertainty
- Bad data:
  - Data with distribution (e.g.)



http://anomaly.org/wade/blog/2006/01/accuracy_and_precision.html

Golden Rules

- When reporting of a , report its
  - Round off values to of
    - Rule of thumb: if the last digit is , digits otherwise
    - x = ± 0.024
    - y = ± 0.6
- Uncertainty can come from the in the and/or of the
    - Rule of thumb:
    - Statistically correct:

Probability: Definitions

- For numerical data, is often most convenient (and quantitative)
- Let's define probability now
  - Formally, it is a that is defined by
  - 1.
  - 2.
  - 3.

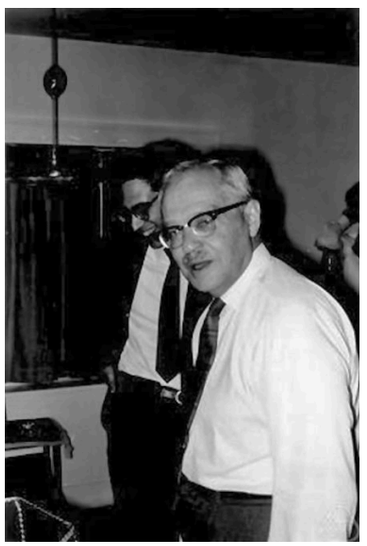
Two Interpretations

- "Frequentist" interpretation:
  - Probability is a a given is reported
    when experiments are an of times
      - Measurable parameters are represented by with
        assigned
        - CL measures a an estimator would fall in a
          , given a of a parameter
          - No probability is assigned to
- "Bayesian" interpretation:
  - More general: define probability as a that a given
    statement is
      - E.g. that the of parameter is in interval
      - This is somewhat , but follows how

Frequentist Probability

- Definitions
  - Let be of all of a
  - Any subset with only (
) is
  - Define $P(A) =$

John von Neumann



Jerzy Neyman

- Assume are (in principle)
- Confidence in a grows with
- Frequentist statistics is (and often argued for) in situations where can be

  so that validity of can be

  (e.g. particle physics)

Bayes Theorem

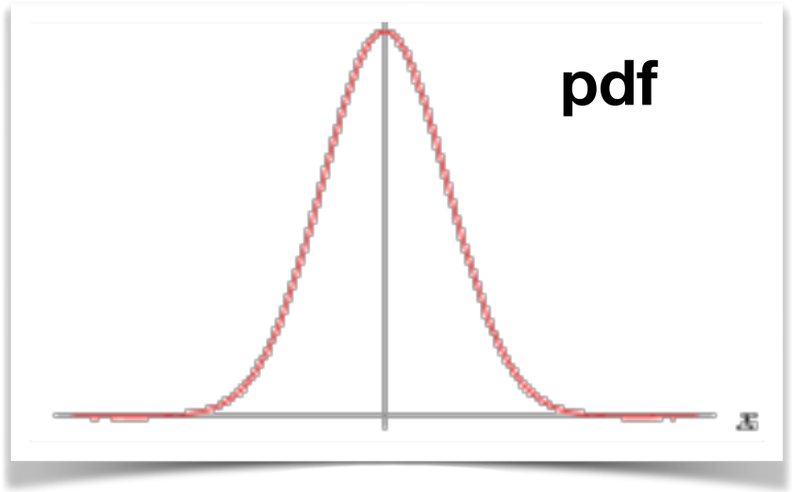
Thomas Bayes

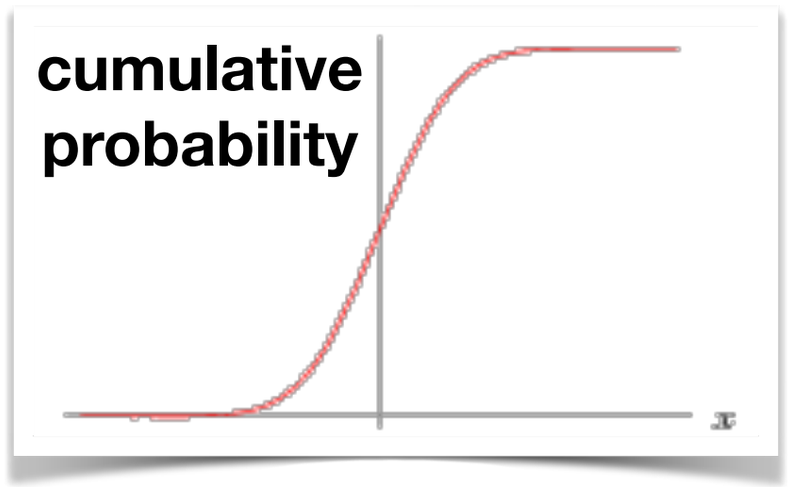
- probability of given
  -
- Interpreted within as
  - $P(\text{theory}\,|\,\text{data}) \propto$

- Allows one to a as a measure of that a given is correct (e.g. that some is in some)
- Requires assigning some to prior knowledge
  - That's where comes in

Probability: Random Variables and PDFs

- For a continuous variable, x, we define the
  - $f(x, \theta) =$
  - θ
- Integrate to obtain the
  - $F(a) = \int$
    - Probability that $x < a$
- Discrete variables:
- Expectation value
  - $E[u(x)] =$





Expectation Values

- Expectation value of a
  - $E[u(x)] =$
- Moments of a
  - $\alpha_n \equiv$
  - $m_n \equiv$

Mean and Variance

- Mean
  - $\mu = \int$
- Variance
  - $\sigma^2 =$

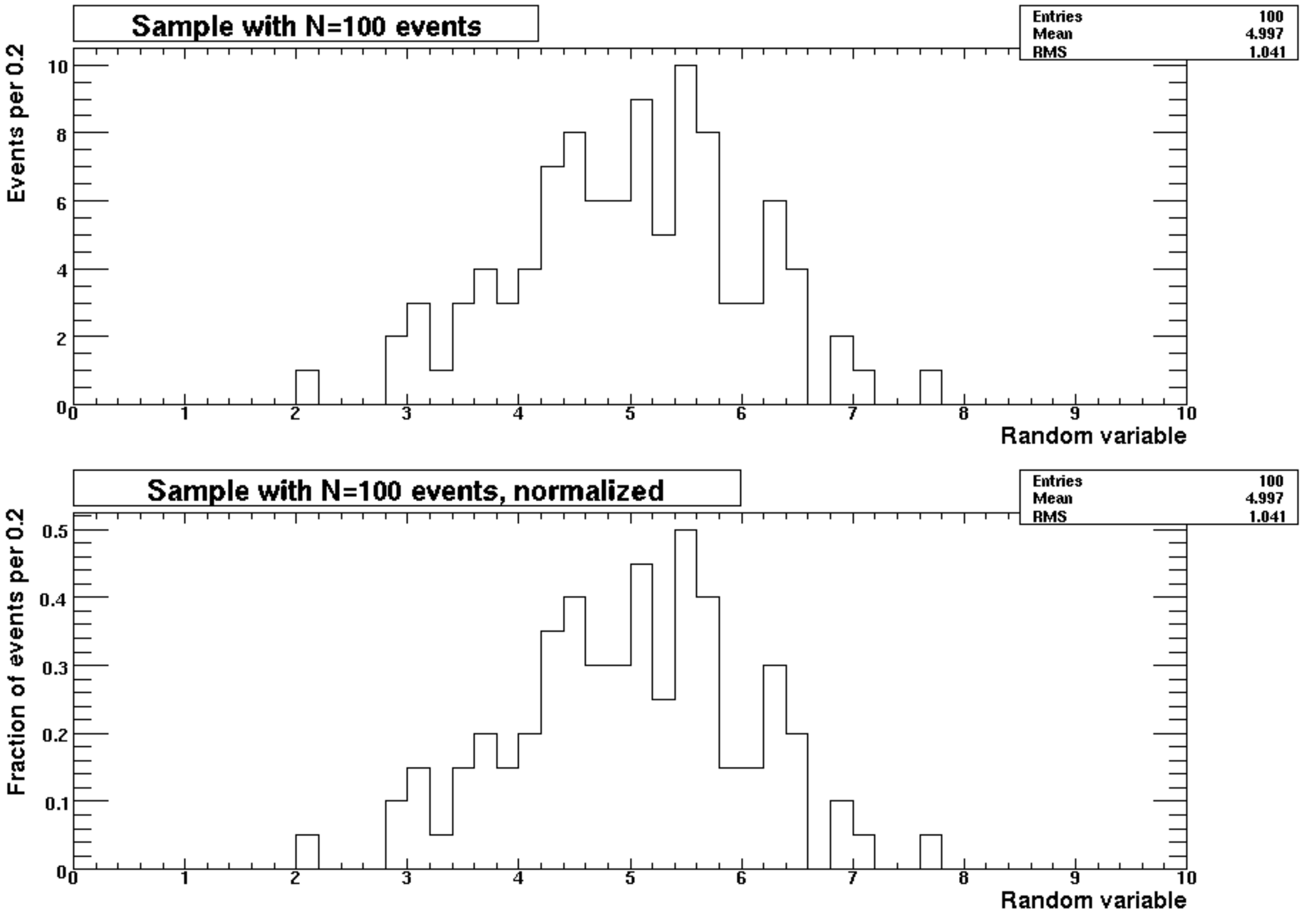
- σ is the standard deviation

If you only remember one thing today, remember this

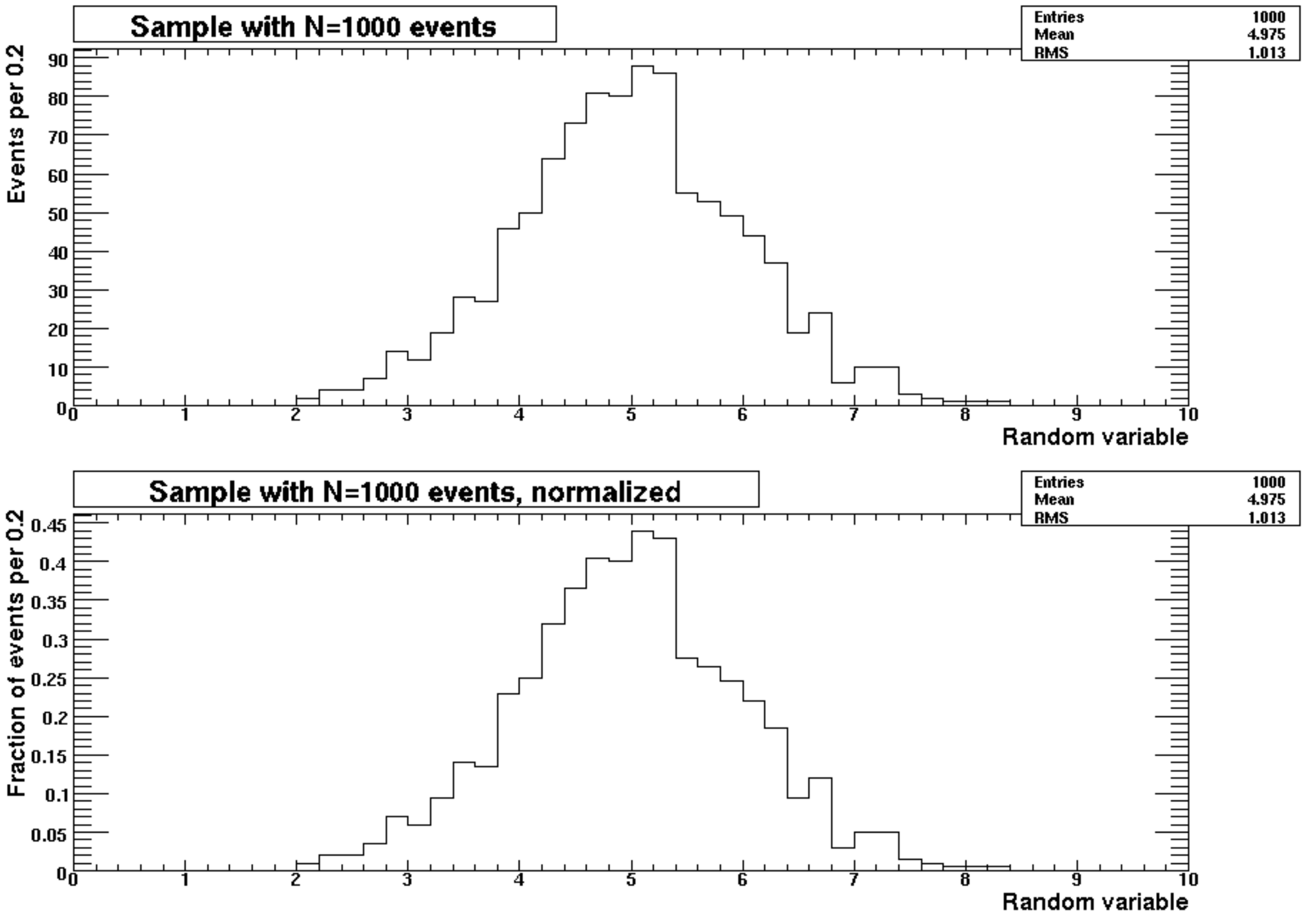
If we know the PDF, we know how to determine μ and σ

Example in jupyter notebook

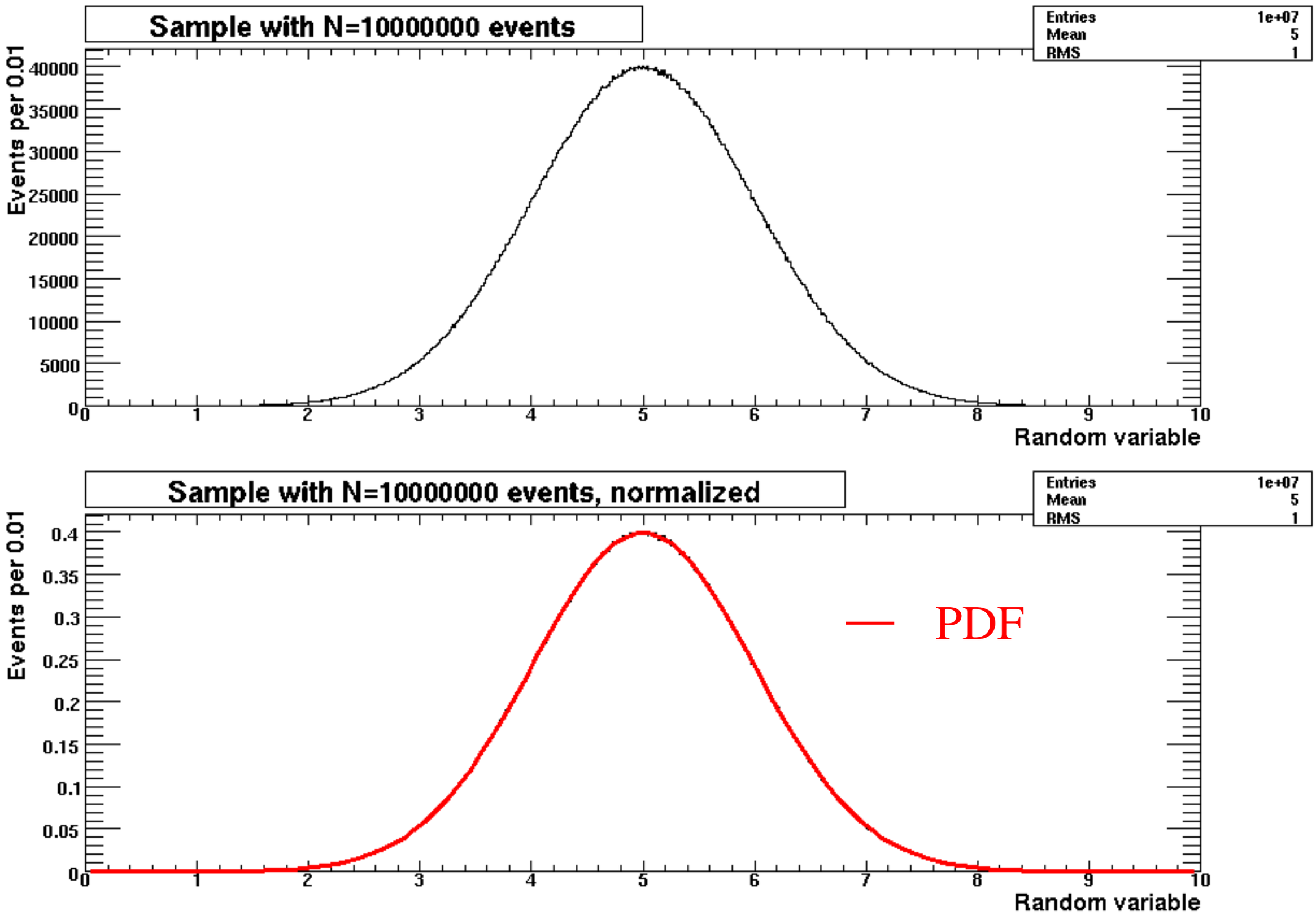
Sample from a Continuous PDF



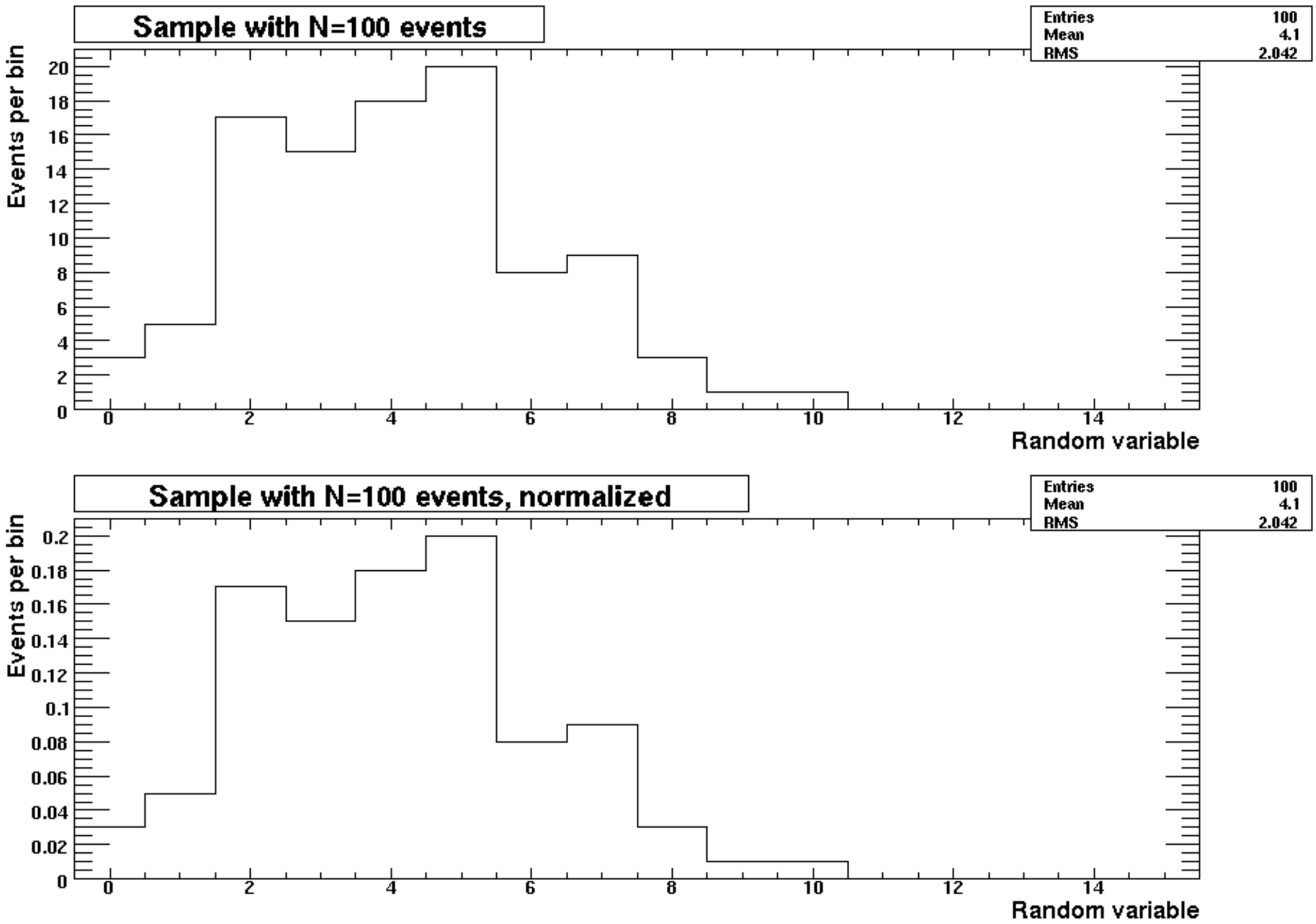
Sample From a Continuous PDF



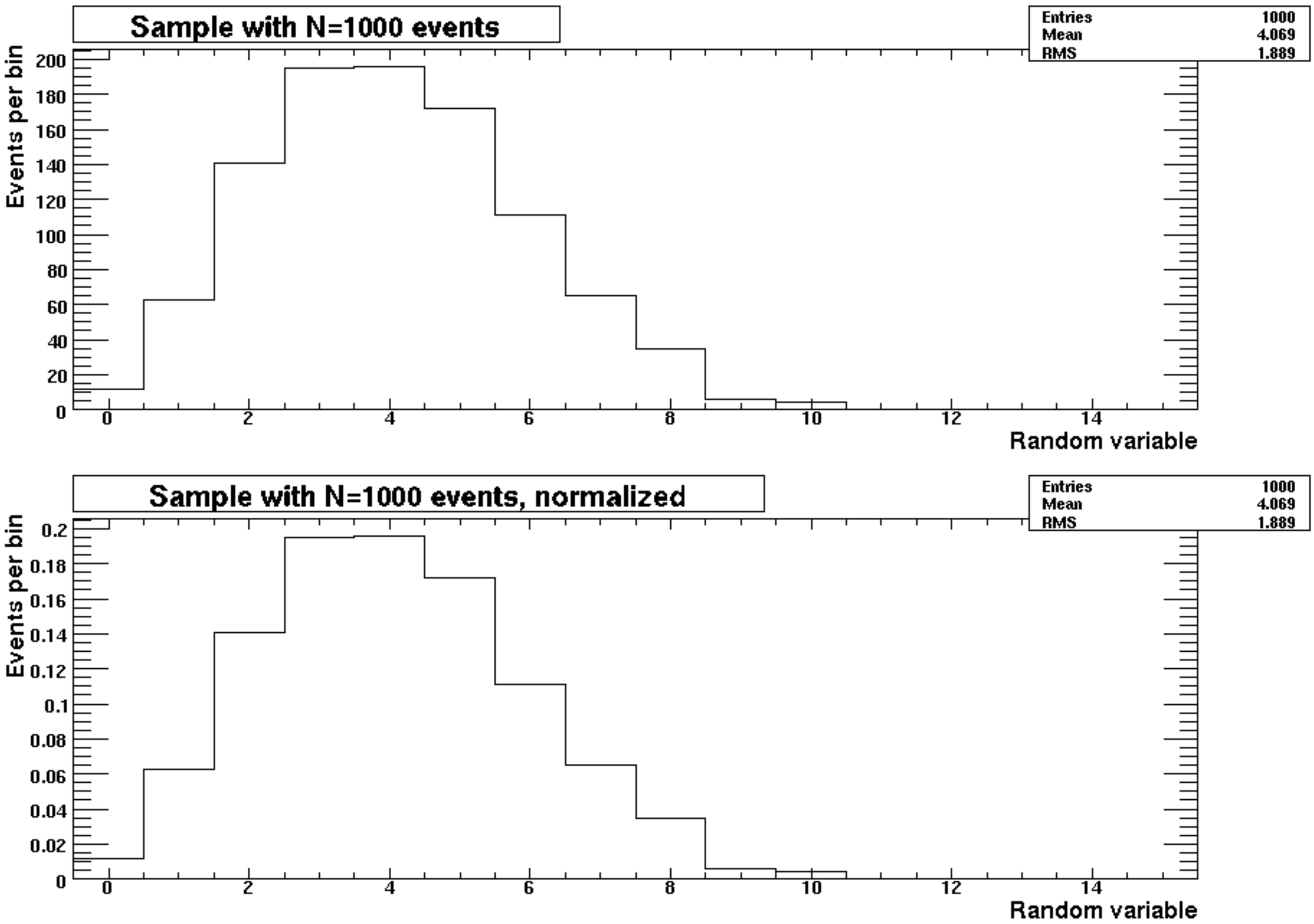
Sample From a Continuous PDF



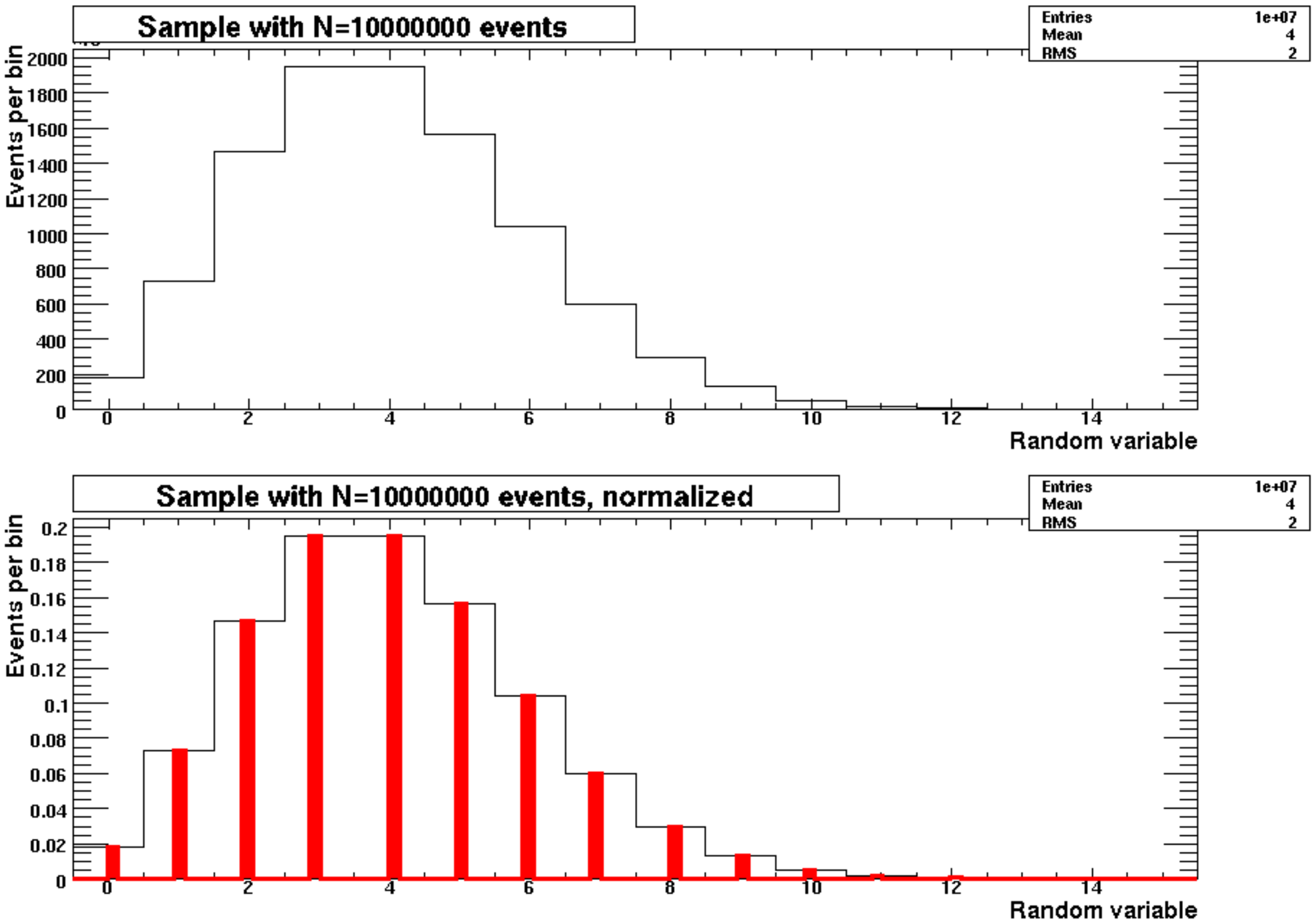
Sample From a Discrete PDF



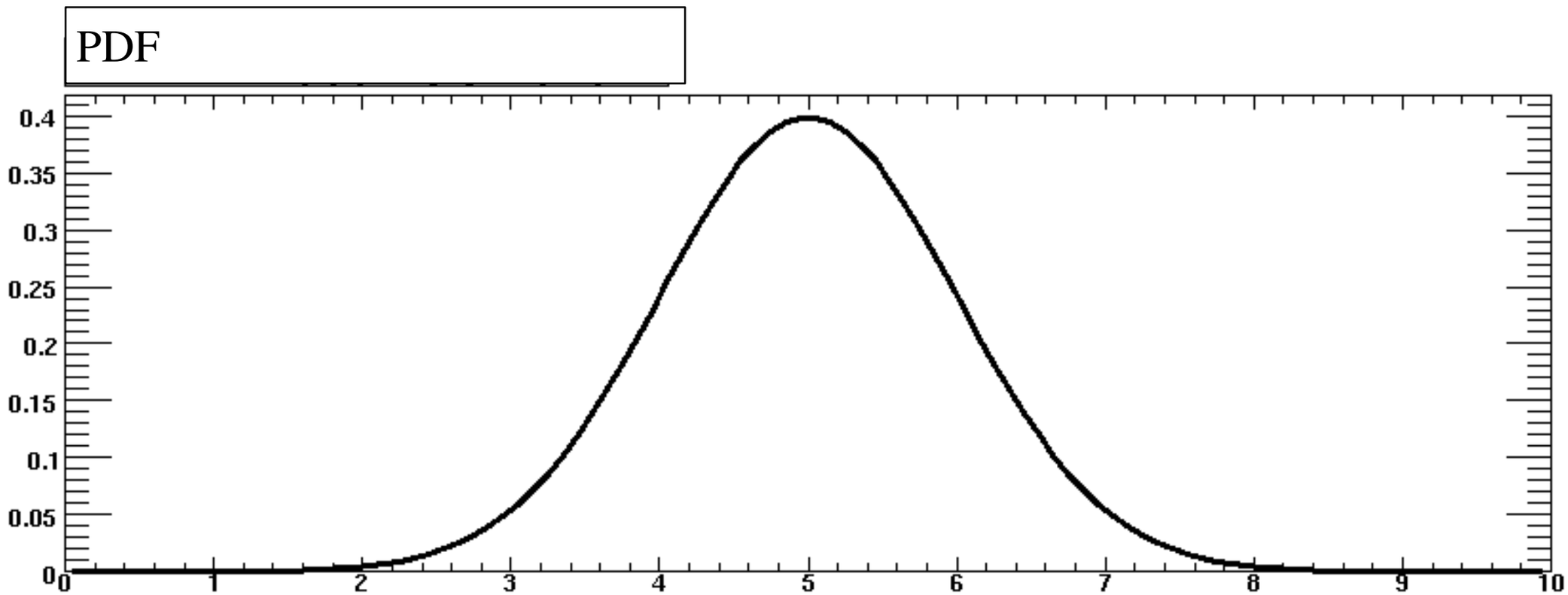
Sample From a Discrete PDF

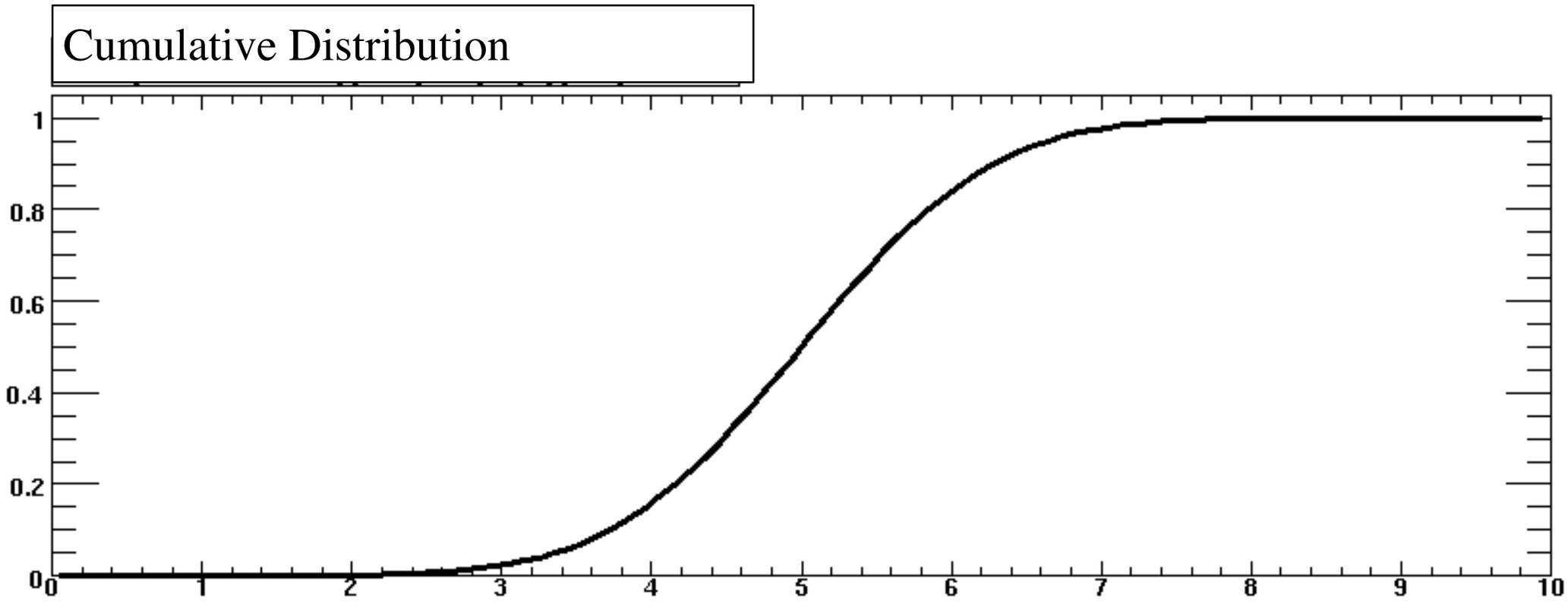


Sample From a Discrete PDF

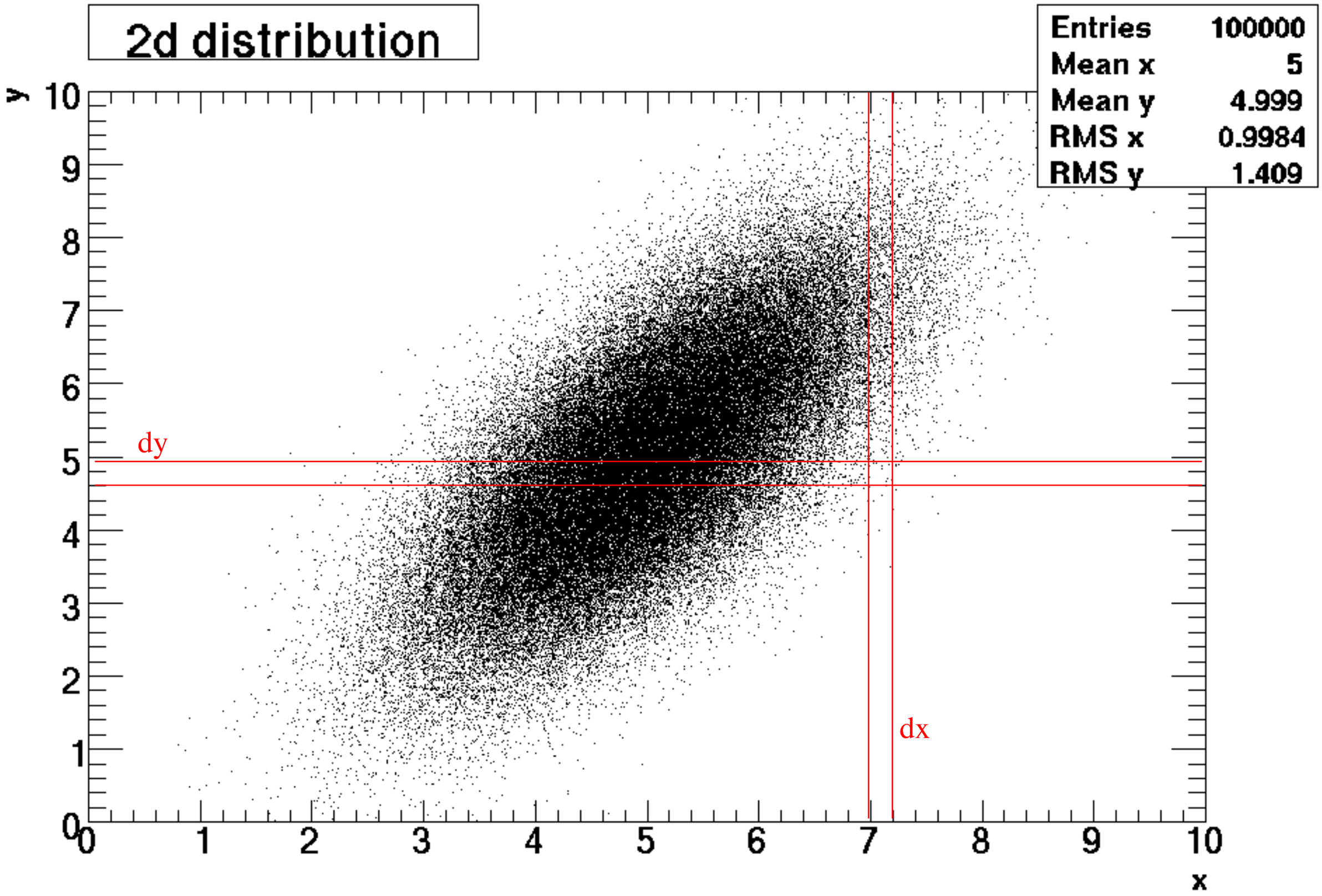


Cumulative Distribution

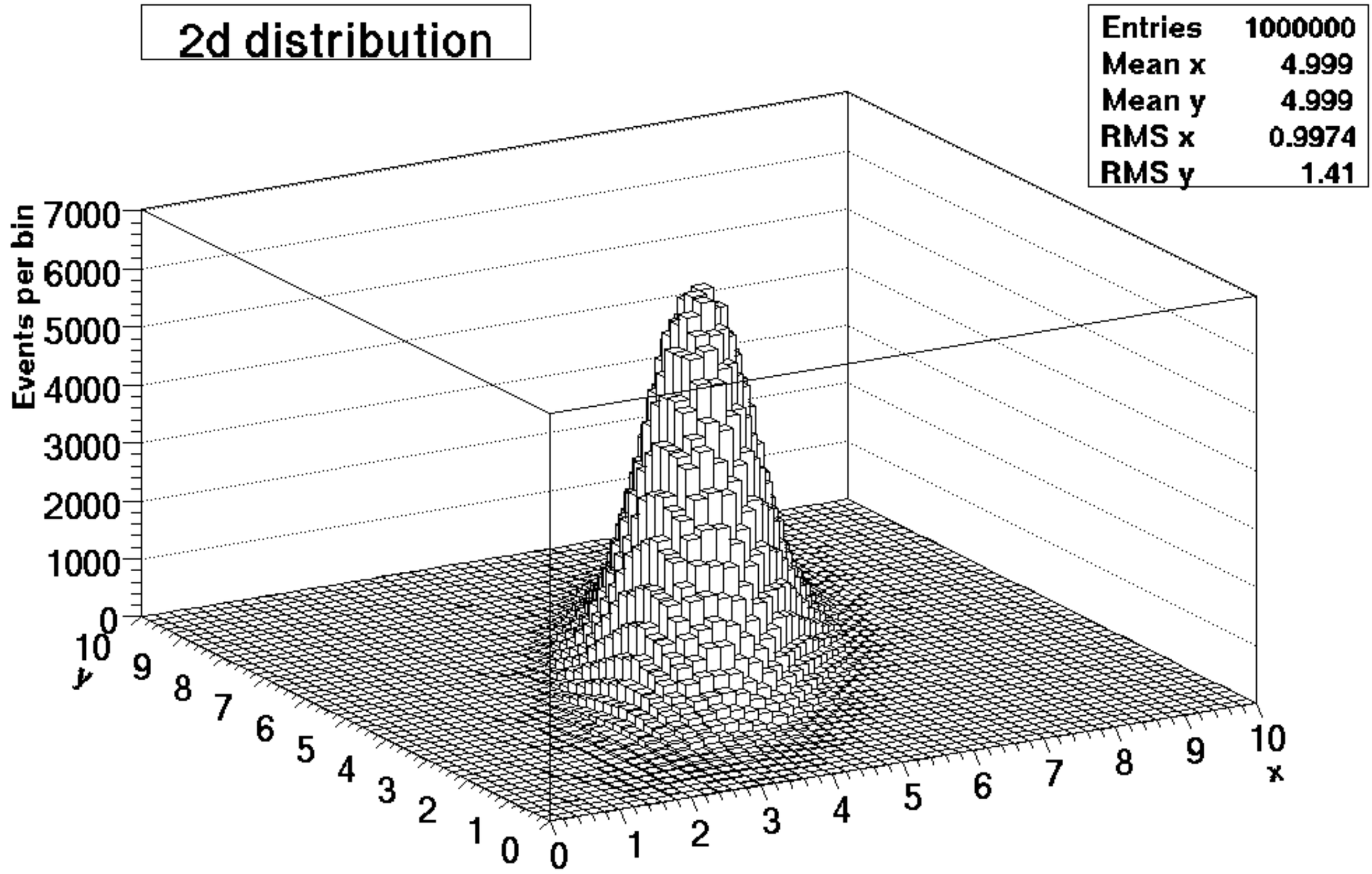




2d Distribution



2d Distribution



Most important distributions

- or distribution
- distribution
  -
  -
- distribution
  -

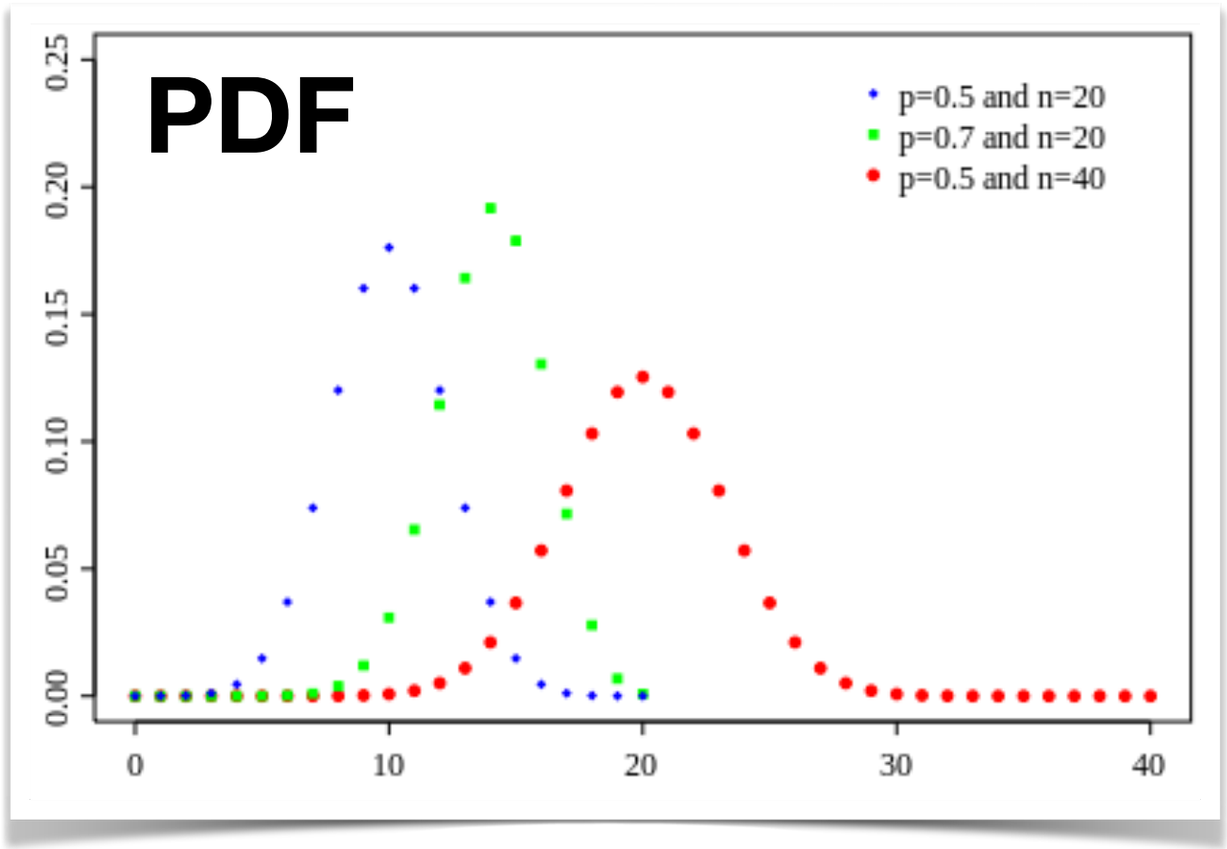
- There are many others
  -
  -
  -
  -

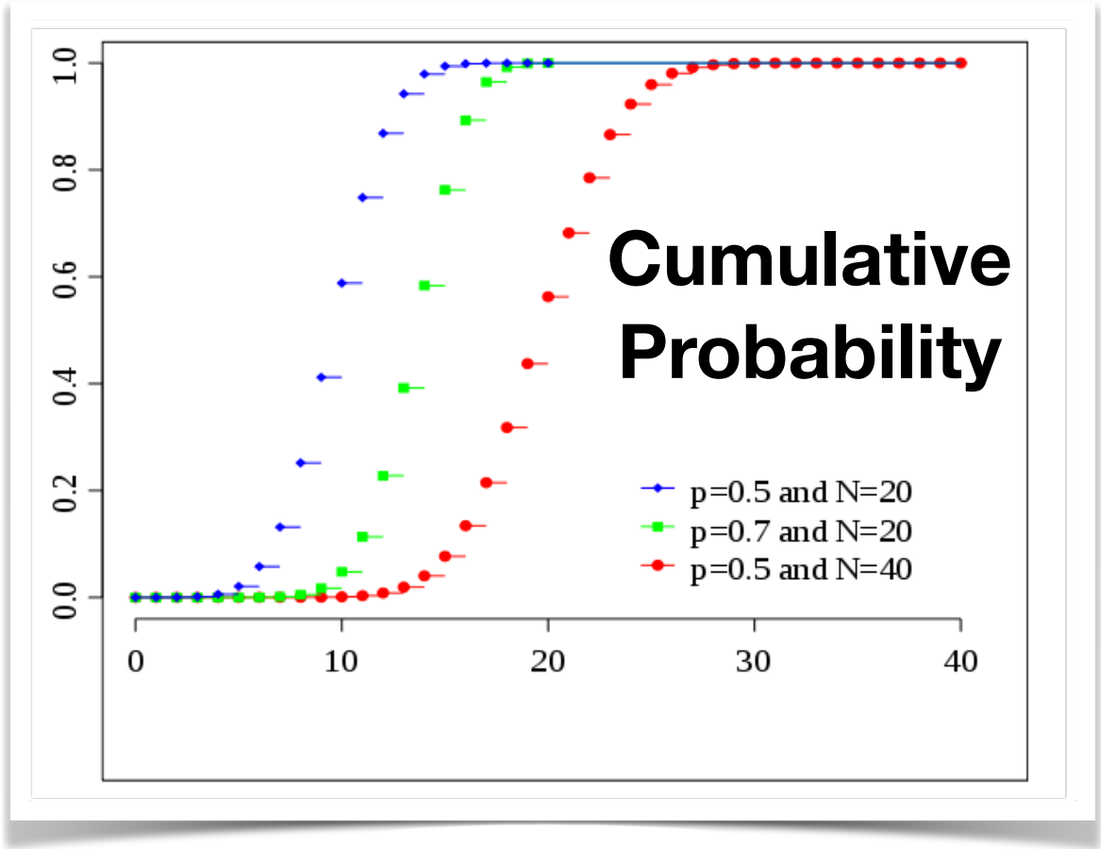
Binomial Distribution

- Random process with two possible outcomes
- $p =$ $q =$
- After n trials, prob of getting outcome #1 exactly k times is

$$f(k,p) = \qquad \text{where } \binom{n}{k} =$$

$\mu =$ $\sigma =$

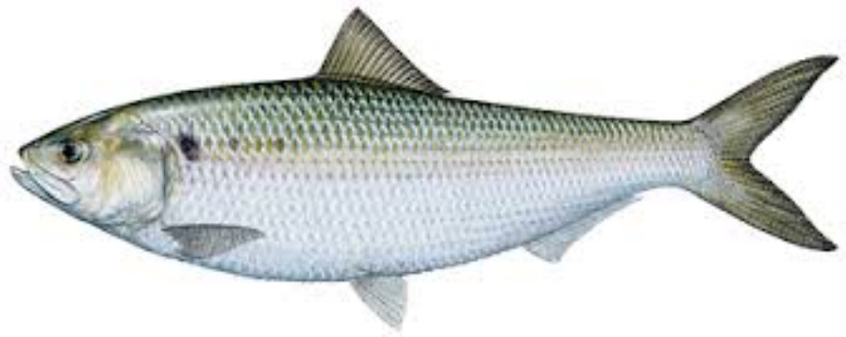




Poisson Distribution

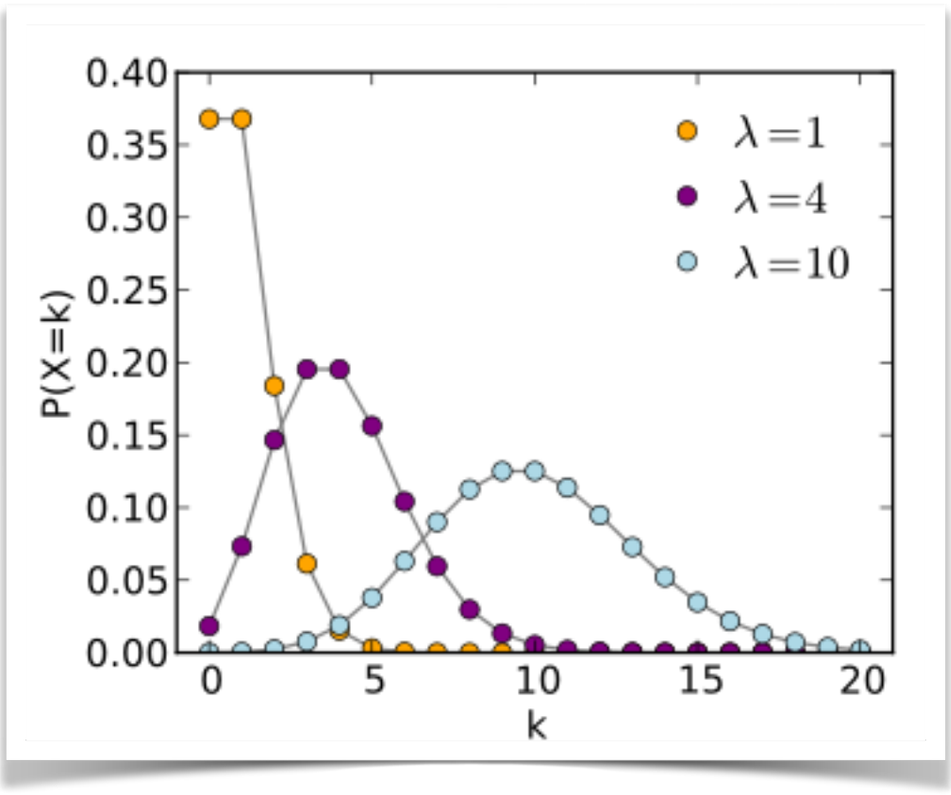
- Prob of finding exactly in the interval if the events occur with an

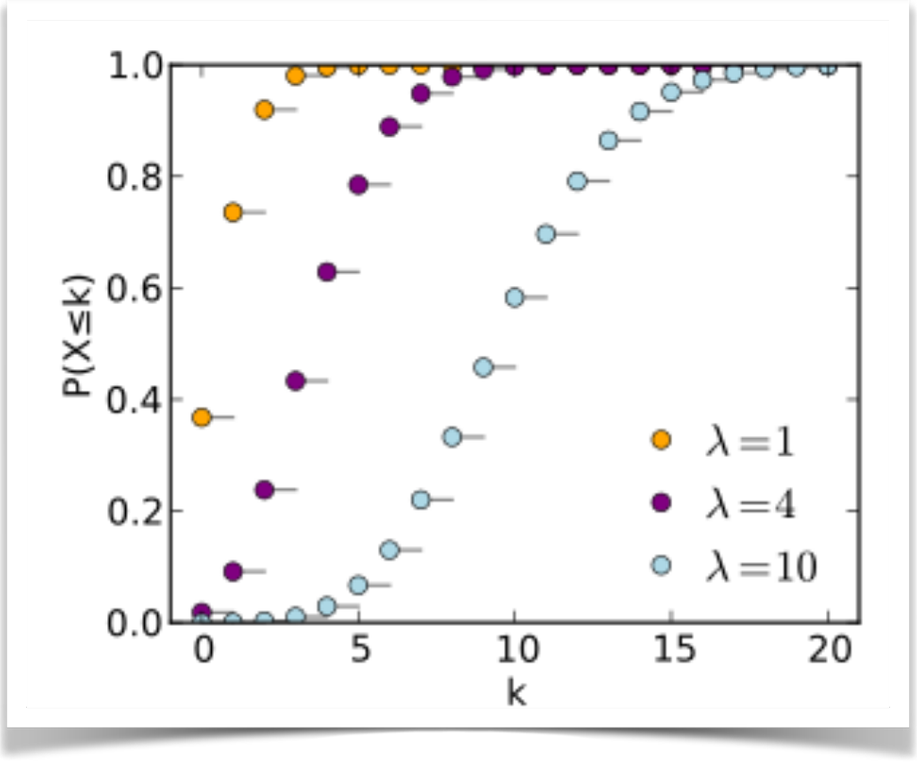
  - $f(k;\lambda) =$



- For large λ approaches a Gaussian

$\mu =$ $\sigma =$

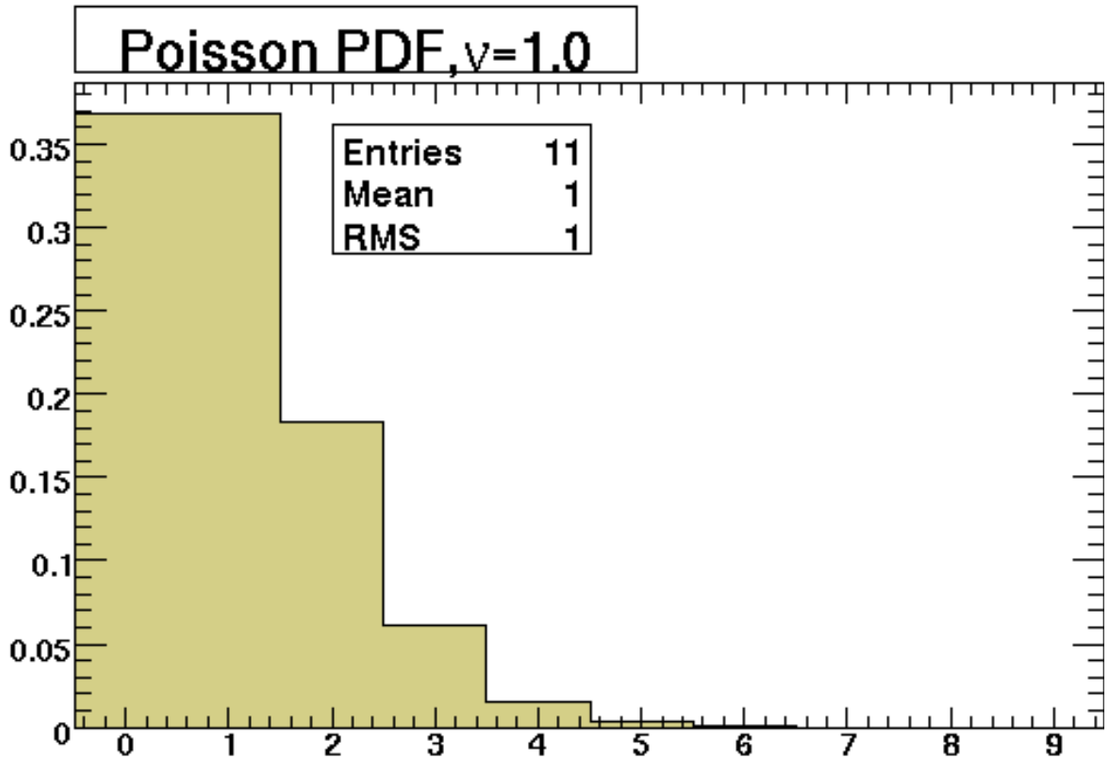


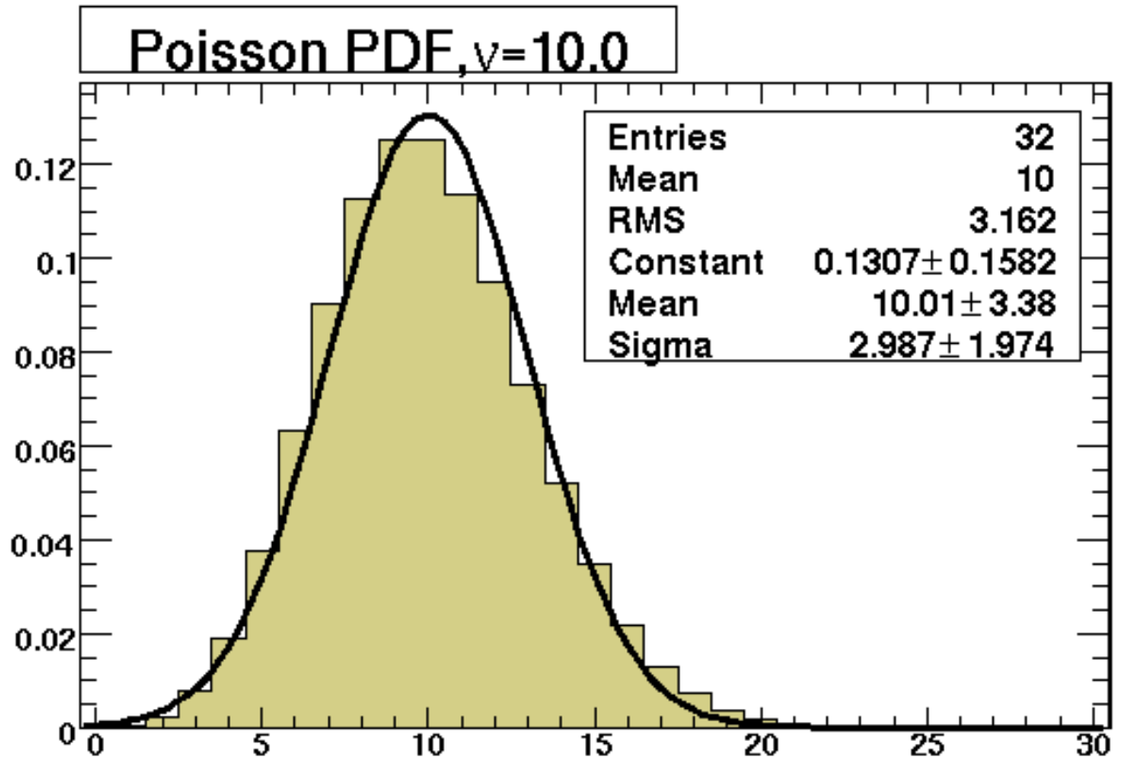


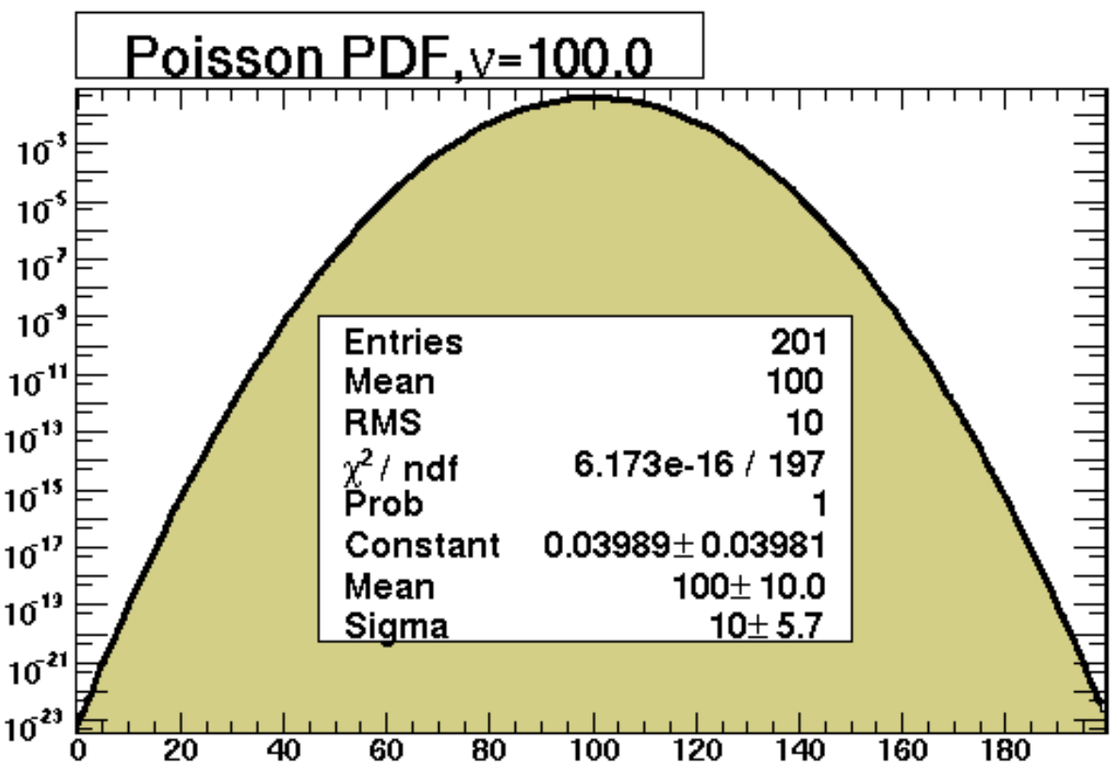
Example: Measure Efficiency

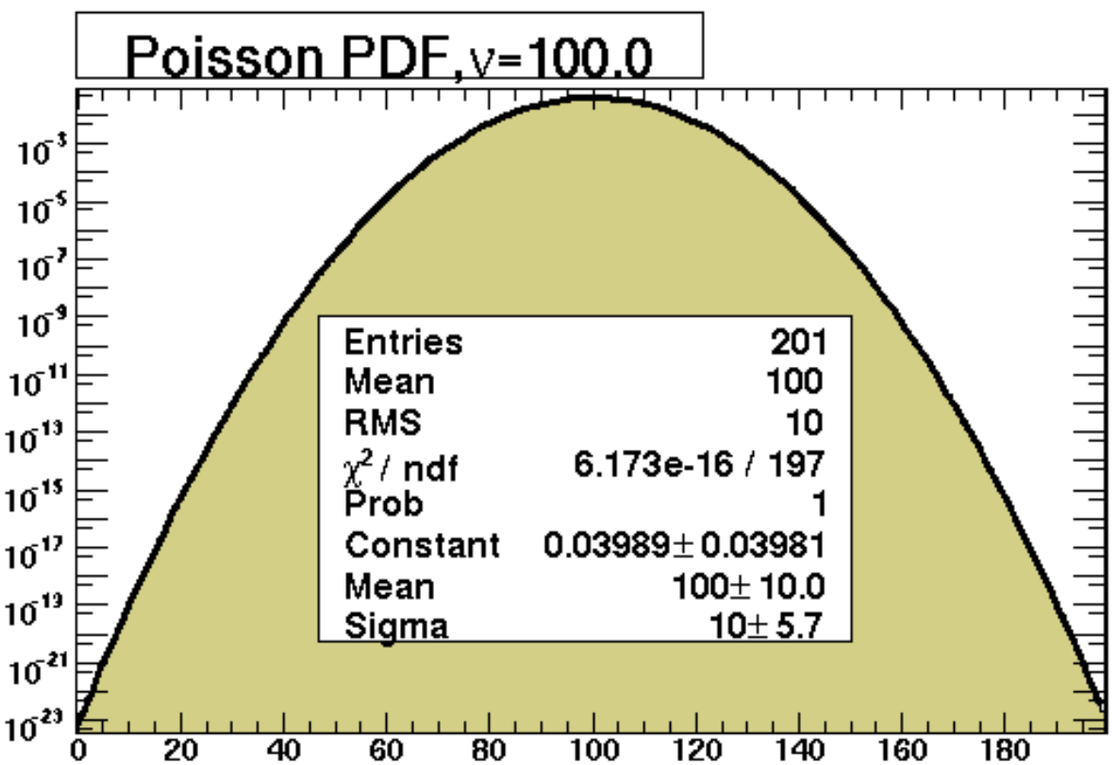
- Generate a sample of
- Apply selection; suppose events passed
- Estimate
  - $\hat{\epsilon} =$
  - $\sigma(\hat{\epsilon}) =$
  - $\sigma(\hat{\epsilon}) \neq$

Poisson Distribution



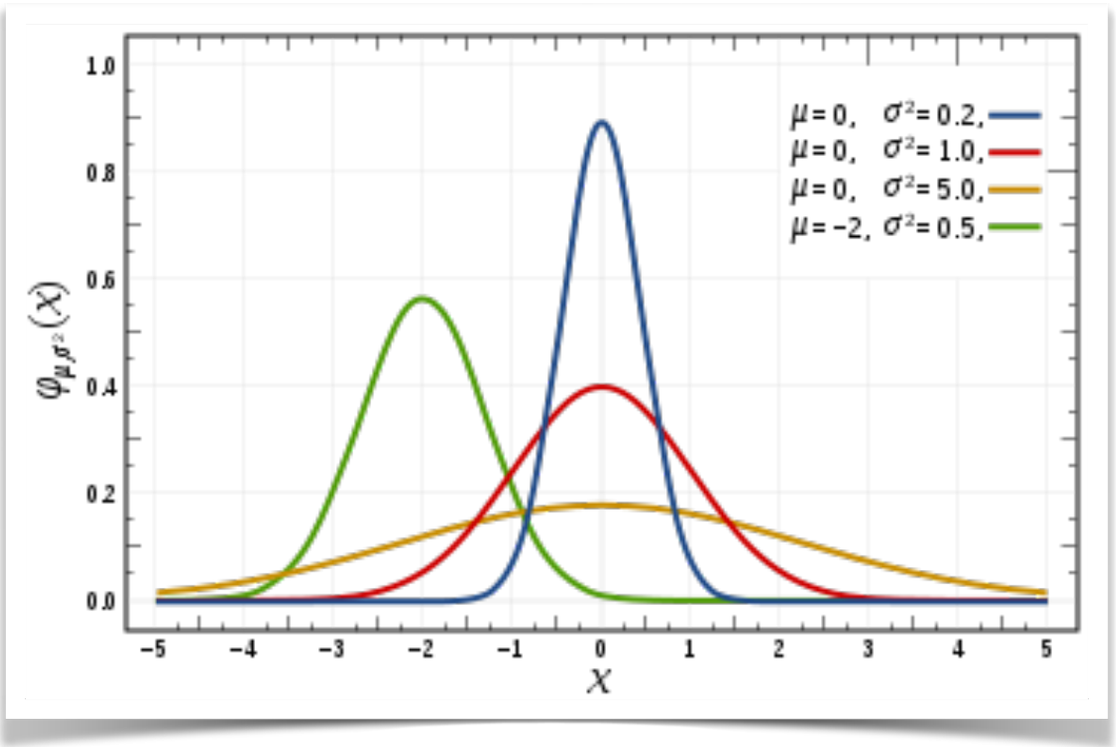


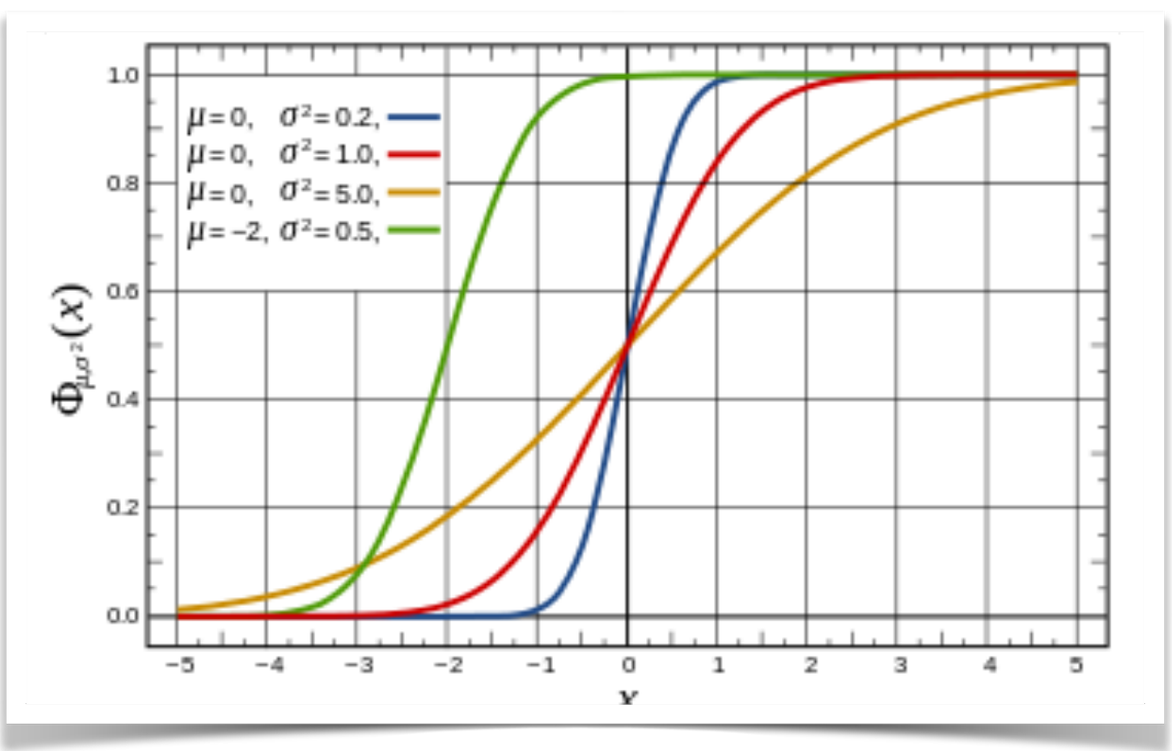




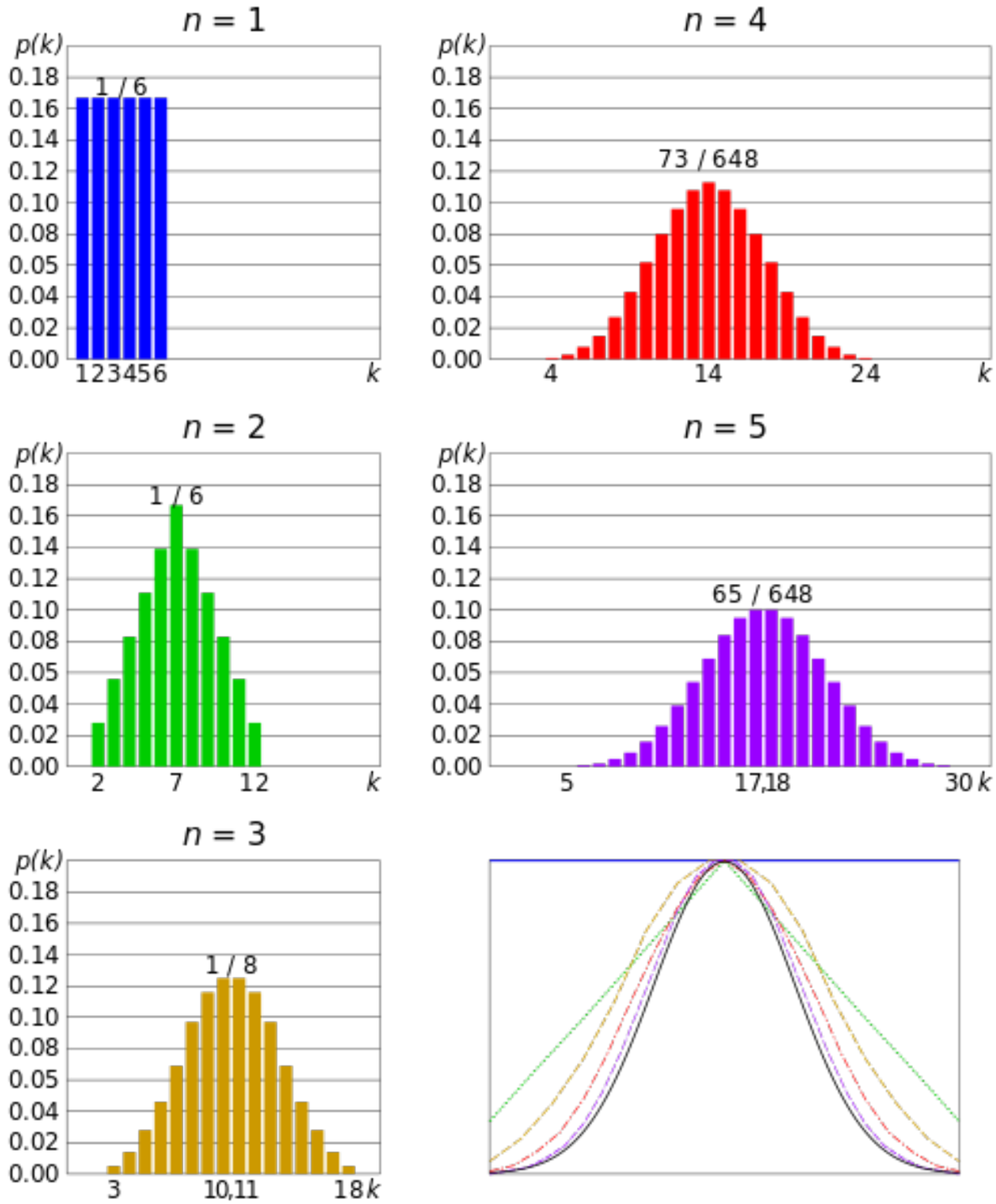
Gaussian Distribution

- **Central Limit Theorem**
  - Given a random sample $(x_1, x_2, \ldots, x_n)$ drawn from a pdf with mean and variance , if the mean is
    - $S/n =$
  - the distribution of S/n approaches the distribution as
    - $f(x; \mu, \sigma) =$





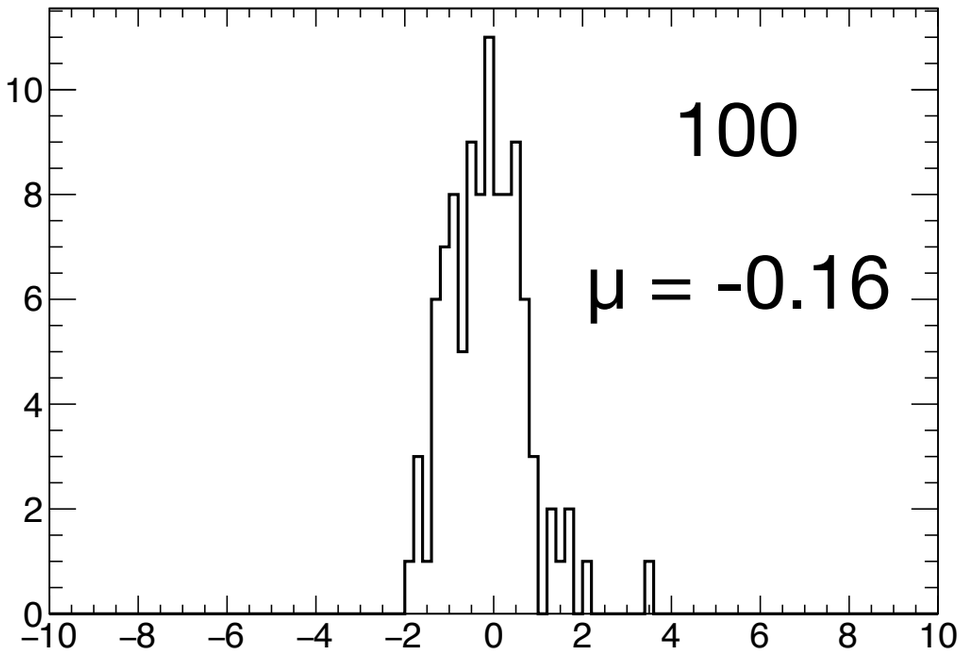
Central Limit Theorem

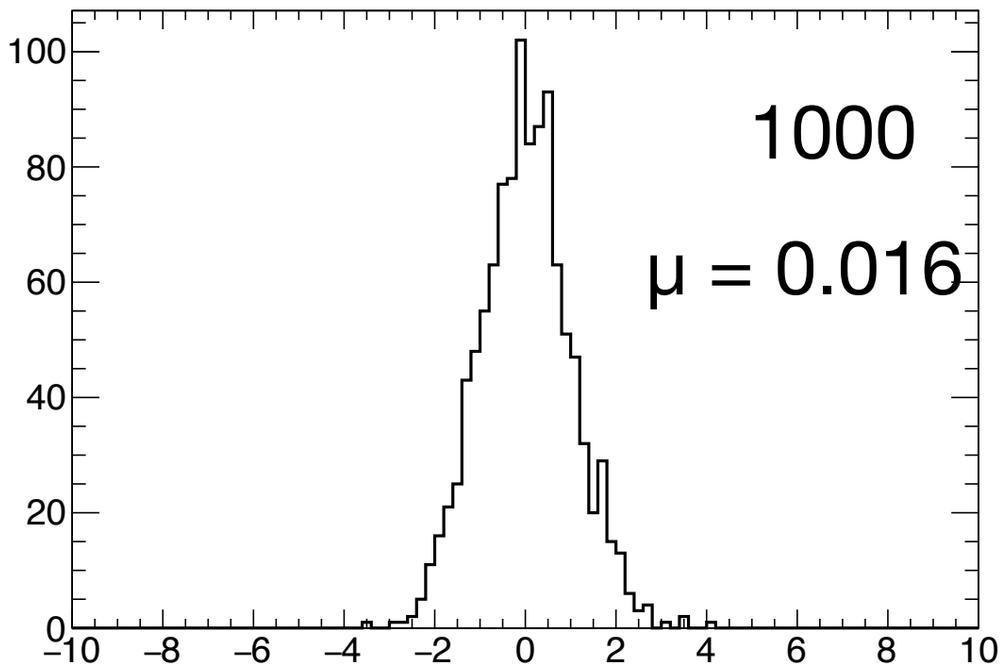


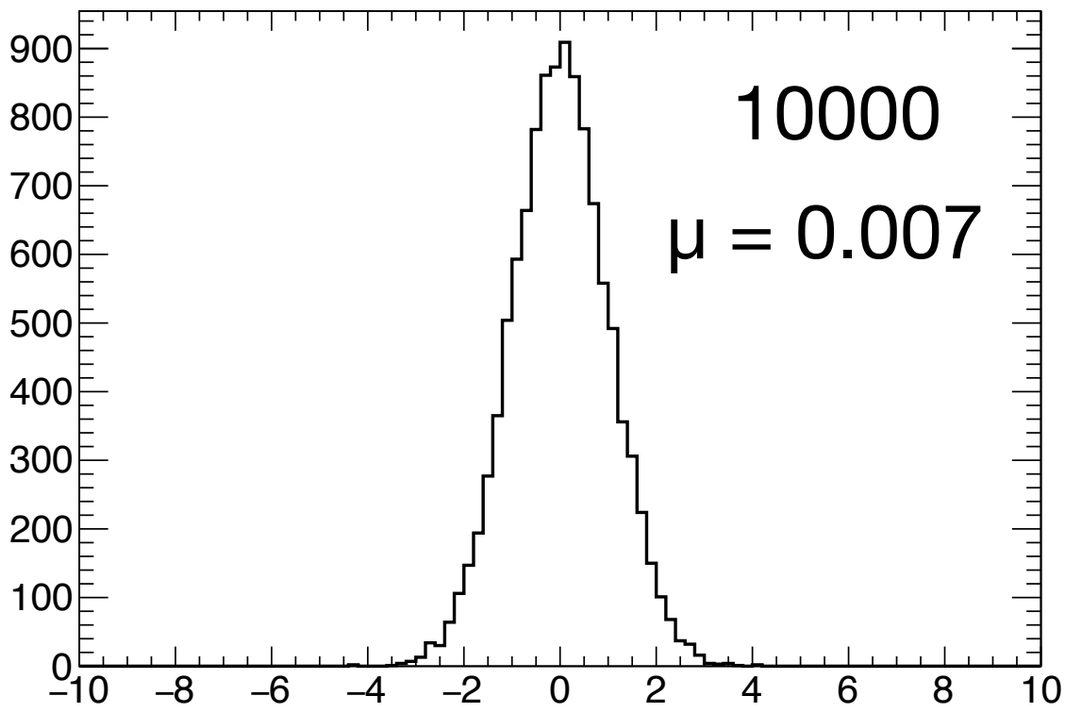
http://en.wikipedia.org/wiki/File:Dice_sum_central_limit_theorem.svg

Example

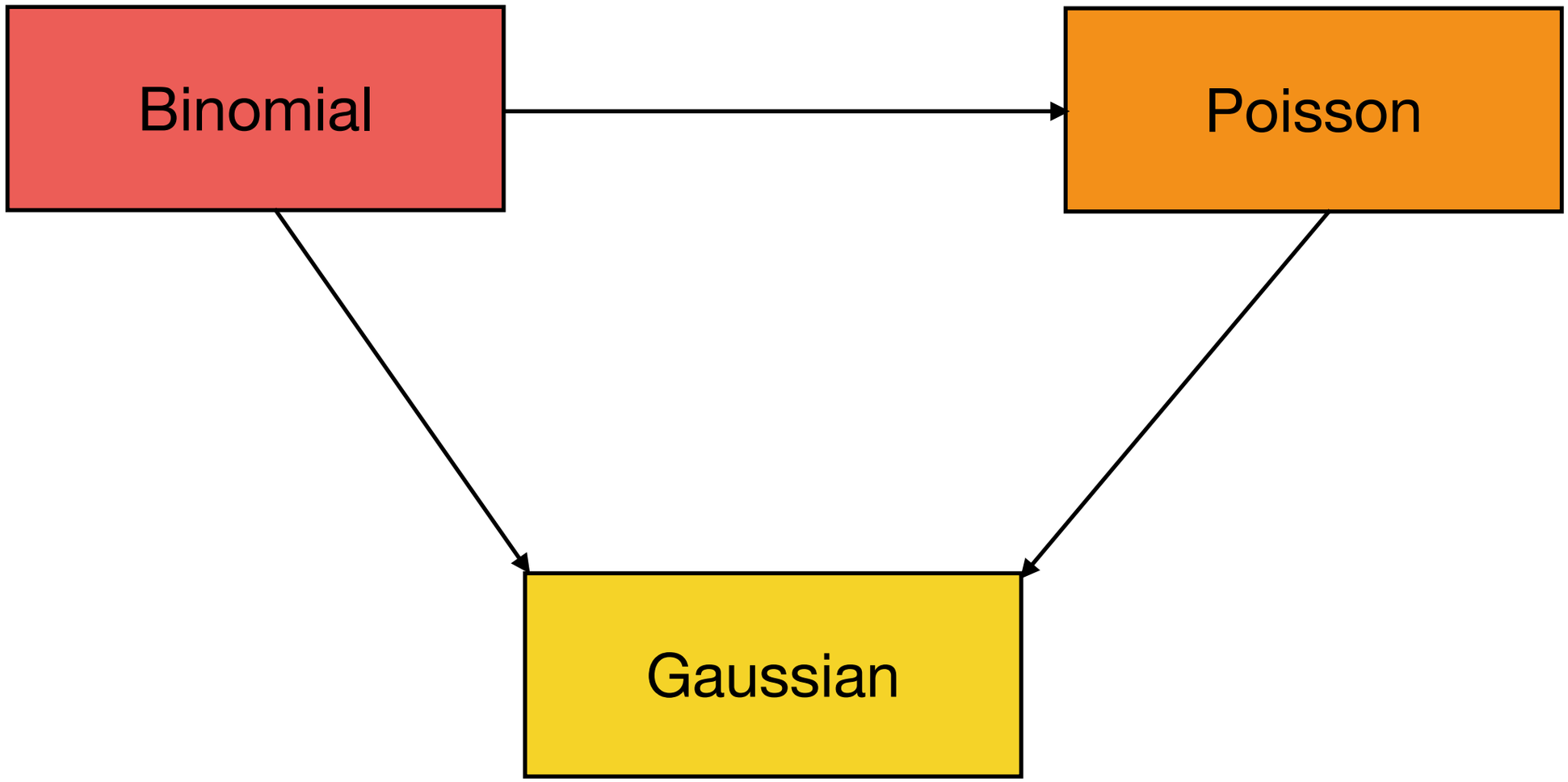
Example in jupyter notebook







Relationships between pdfs



Point Estimation

- Standard problem: set of described by
  - $f(x) \equiv$
- Point estimation:
    - $\hat{\theta} =$
    - Estimator of

Estimators

- Typical goal: estimate from experimental data and understand the uncertainty on that measurement
- **Characteristics** of an estimator
  -
  -
  -
  -
  -
  -
- **Uncertainty**: how far the might be from our estimate due to statistical fluctuations in the

Basic Estimators

- Estimators for and
- Shape of the
  -
    - Most , but may be
    - is not readily available
  -
    - for data
    - Convenient for , for linear functions
    - Automatic measure
    - Be careful of
      - (e.g. when becomes)

Mean and Variance from a Sample

- Estimators (equally data)
  - $\hat{\mu} =$
  - $\hat{\sigma^2} =$
- Variances of these
  - $V[\hat{\mu}] =$
  - $V[\hat{\sigma^2}] =$
  - $\sigma[\hat{\sigma}] =$

Likelihood Function

- Likelihood $\mathscr{L}(x;\theta)$: that a of will yield a for a
  - Need to both the and the value for in that theory
- With an ensemble of measurements, overall likelihood is obtained from the of the measurements

- Here θ represents one or more parameters

Log Likelihood

- To estimate the parameter(s), maximise the likelihood
  - Set derivative to zero
- Typically easier to maximise the
  - $\frac{\partial \mathscr{L}}{\partial \theta} =$

- If there are several we can minimise with respect to each of them

Likelihood Example: Poisson

- independent trials with results
- Likelihood function for observing if true mean is
  - $\mathscr{L}(n_i; \mu) =$
- Product over N measurements
  - $\mathscr{L}(\text{data}, \mu) =$
  - $\Rightarrow \ln \mathscr{L} =$

Best estimator is the mean value

Likelihood Example: Gaussian

- $G(x \mid \mu, \sigma) =$

- Take the derivative of the log likelihood
  - $\dfrac{\partial}{\partial \mu}(\ln \mathscr{L}) \mid_{\hat{\mu}=\mu} =$

- The unbiased estimator for σ is
  - $\hat{\sigma} =$

Binned vs unbinned likelihood functions

- Likelihood formalism works for any

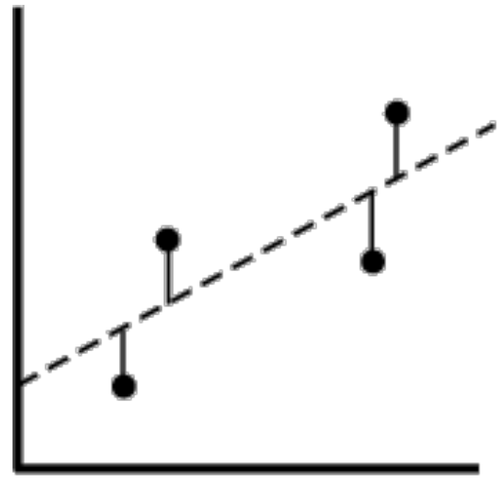
- Product of the is a
- **Example measurement**: Measure the of a particle of a given species for an of such particles produced at such that the at time :
  - $f(t) = \frac{1}{\tau} e^{-t/\tau}$
- Consider two ways to construct the likelihood
  - For decay measure and take the of all times ( )
  - Make a of the in bins of
    - Measurement is the in each bin ( )

The Likelihood and χ^2

- If the data is Gaussian, we have
  - $\ln \mathscr{L} =$
- Compare to
  - $\chi^2 = \sum_i^N \frac{(x_i - \mu)^2}{\sigma^2}$
- By inspection
  - $\chi^2 =$
- Note: the likelihood formulation works for not just

Method of Least Squares

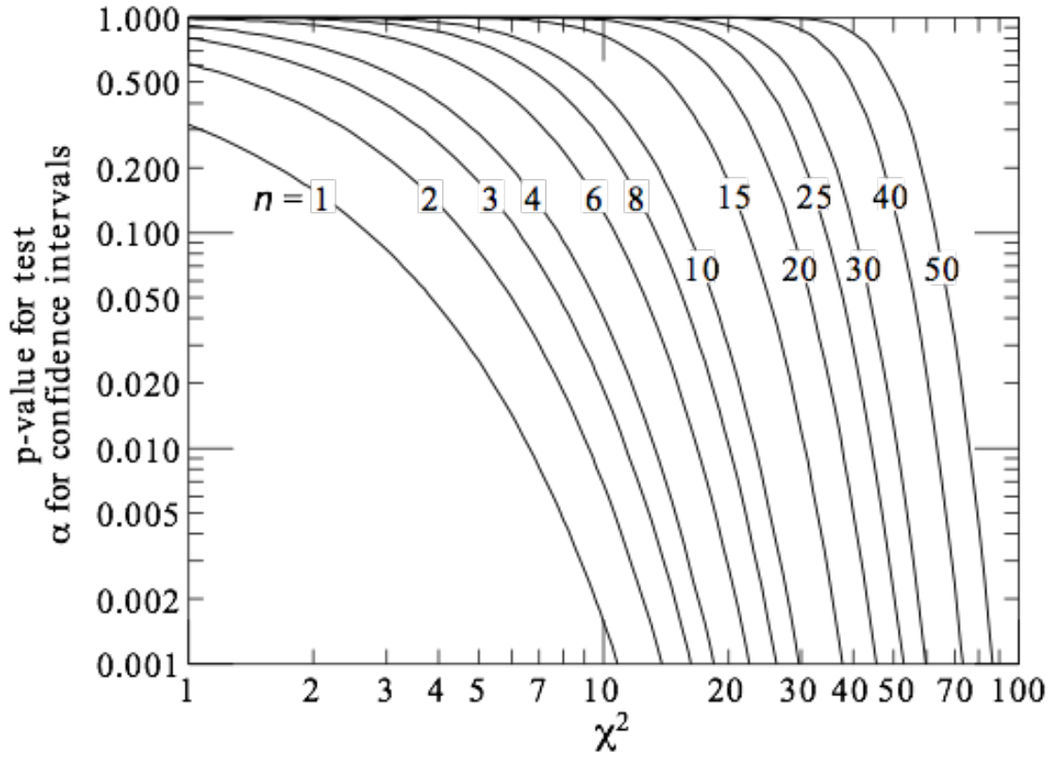
- Assume we have enough statistics for our measurements such that we can assume we are in the Gaussian regime
- Goal: Find the of of a that describes the
- Minimize the of the from the
  - Account for the
- Scatter defined by χ^2
  - $\chi 2 =$
- Can write the χ^2 in terms of our observables
  - $\chi 2 =$



- Minimise χ^2 with respect to θ
- Useful when minimising $\ln \mathscr{L}$ is slow (high statistics samples)

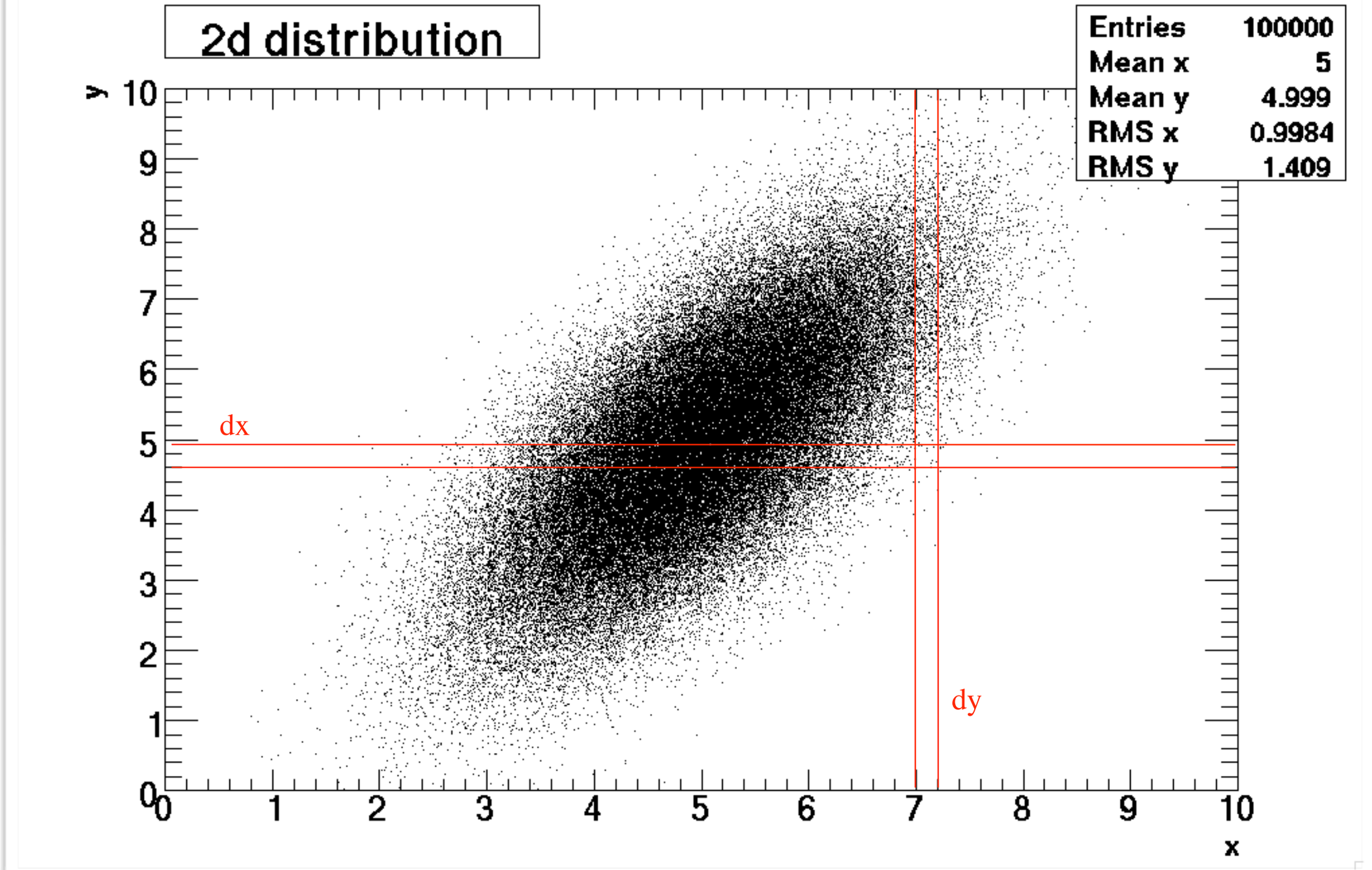
Example: chi-squared p-values

- One advantage of a is that the value of the can be interpreted as a
  - iff on each data point are and
  - the (distribution of around their) are
- In the plot below
  - $n =$
- For a , expect to be close to

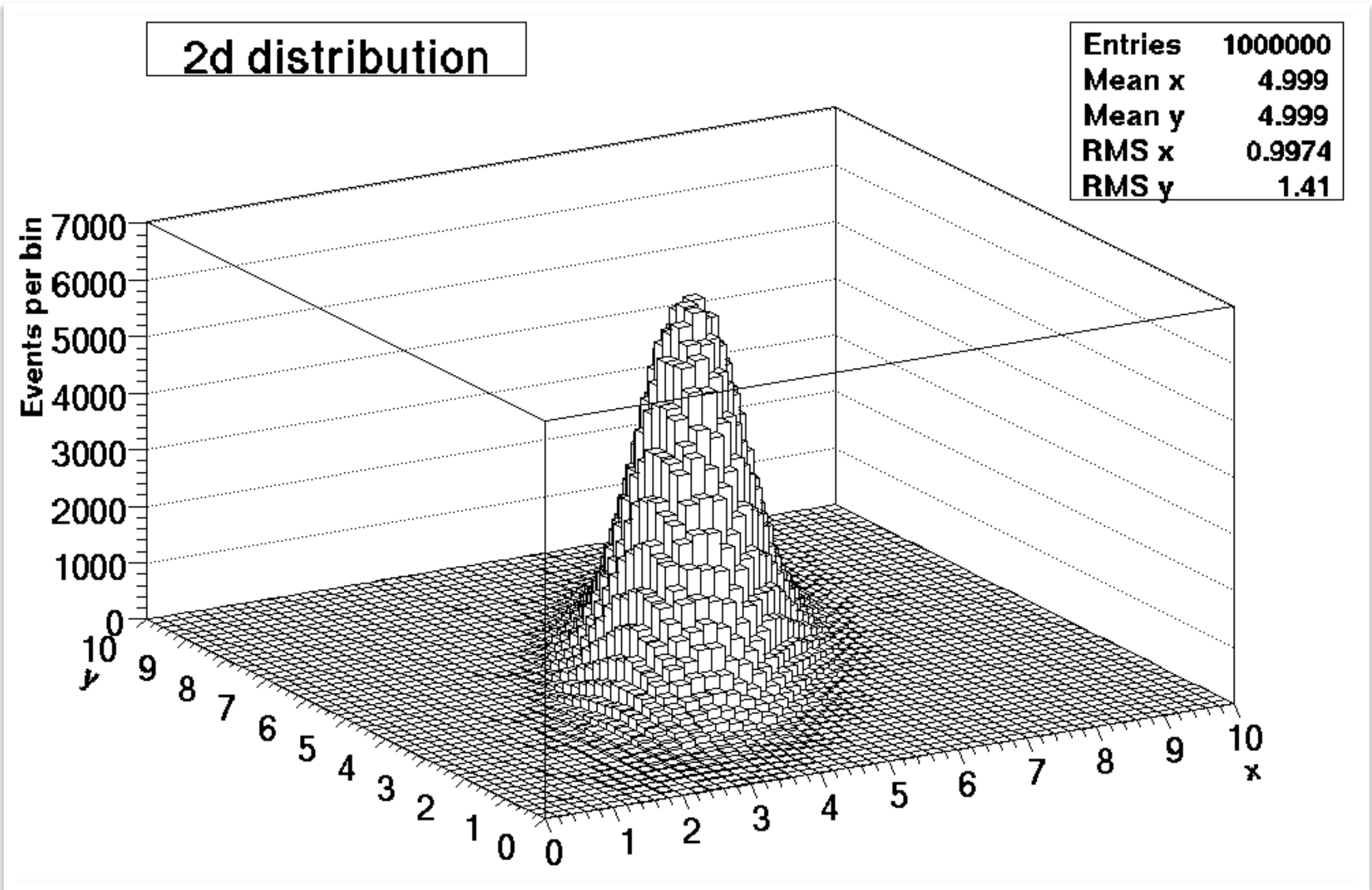


Example in jupyter notebook

2D distribution



2d distribution



Covariance and Correlation

Covariance Matrix

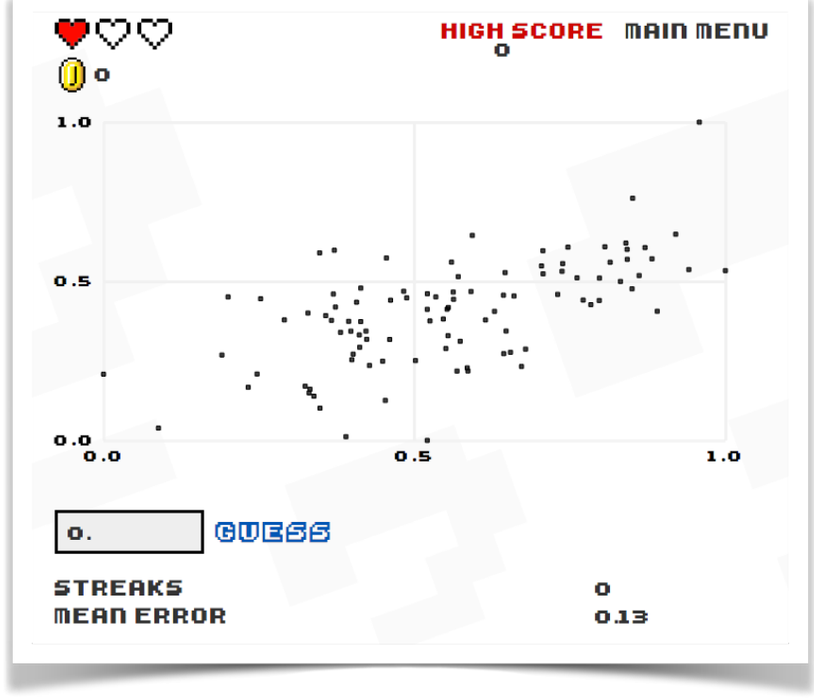
$$cov[x, y] =$$

- A representation of the N-dimensional parameter space as a covariance matrix
  - Diagonal elements:
  - Off-diagonal:

Correlation (normalised covariance)

If two variables are uncorrelated, independent variables, then $cov[x, y] = 0$ for $x \neq y$

$$\rho_{xy} =$$



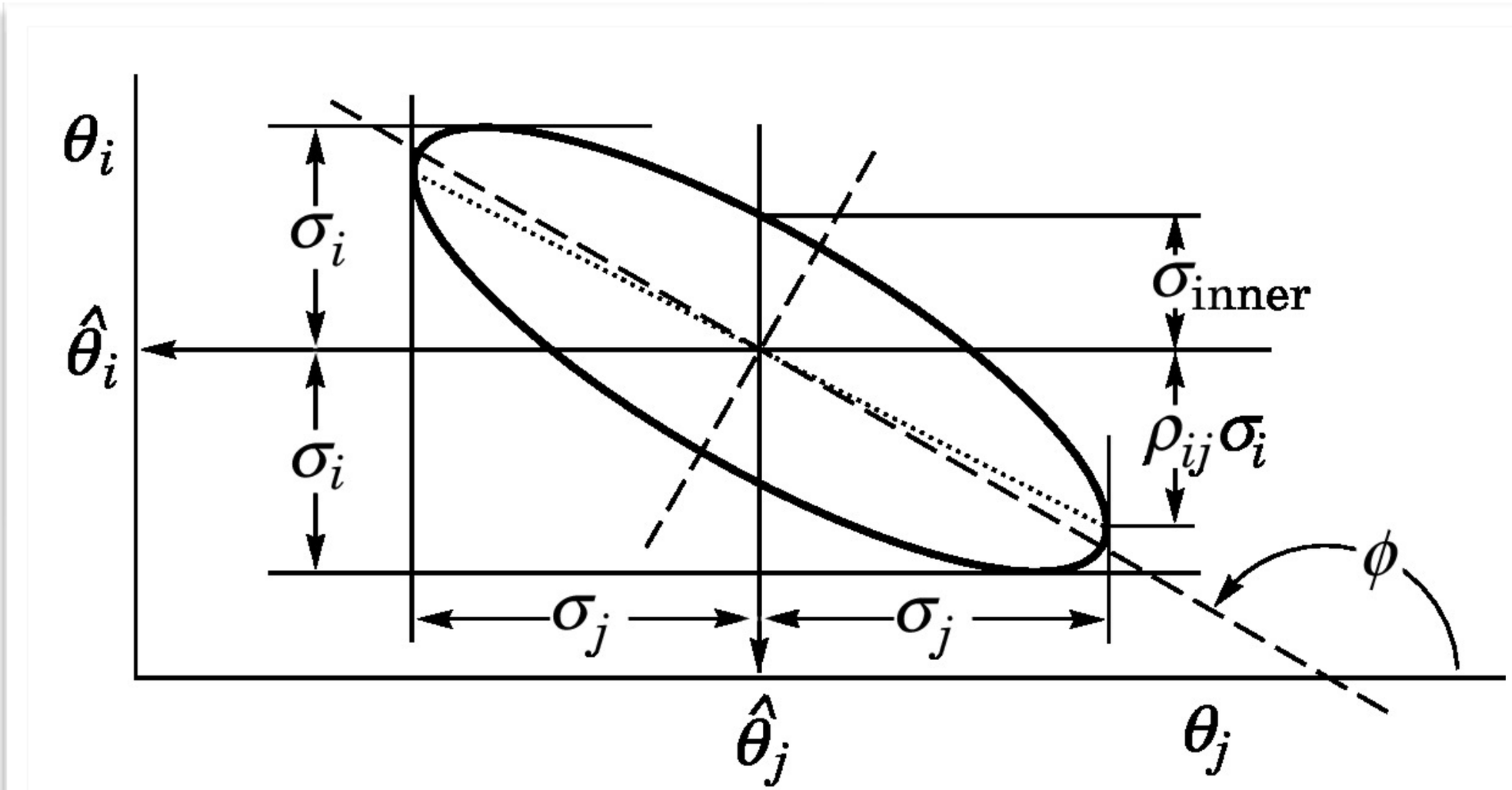
http://guessthecorrelation.com/

Covariance Matrix for a Gaussian

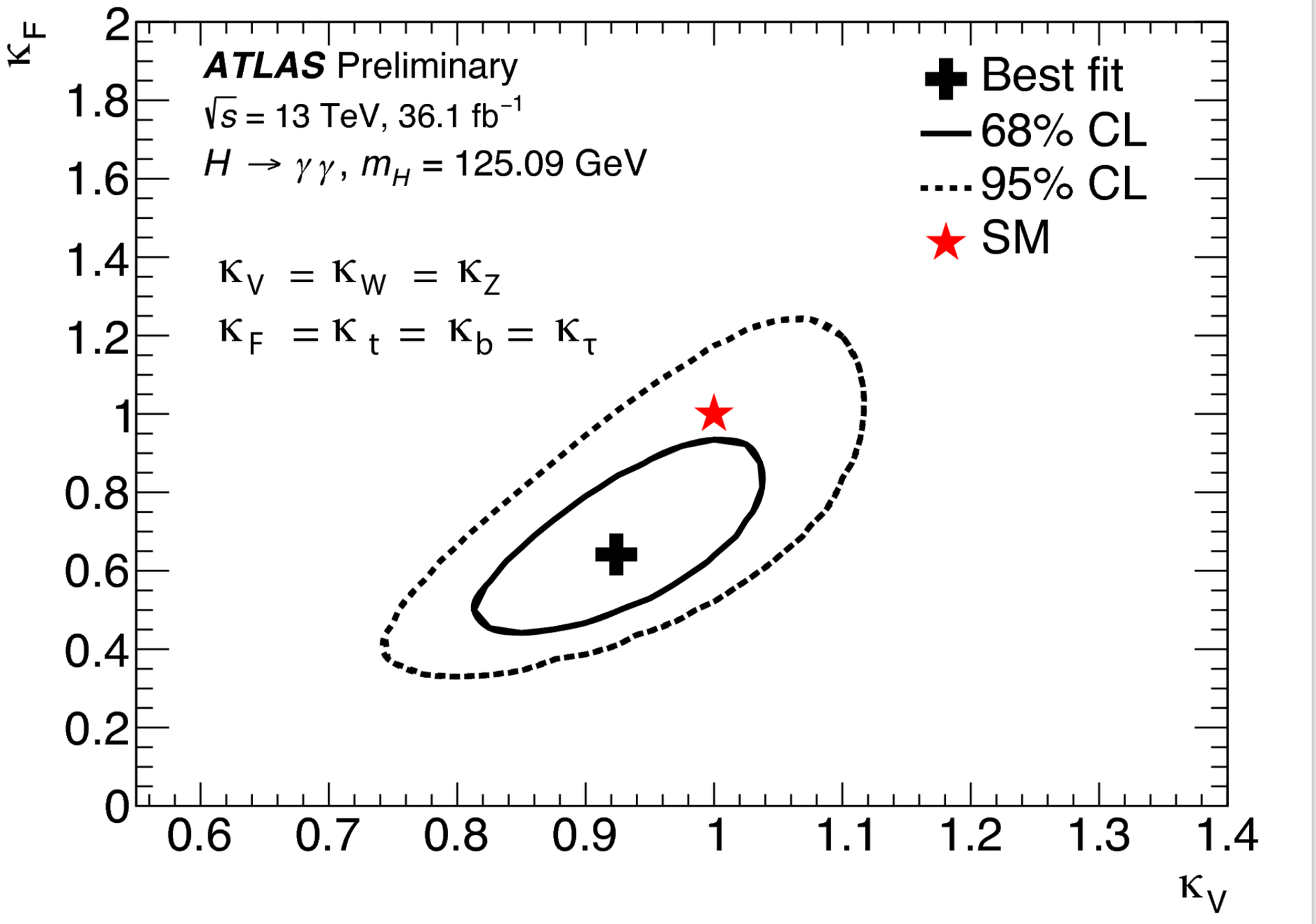
- If x and y are independent variables
  - $G(x, y \mid \mu_x, \sigma_x, \mu_y, \sigma_y) =$
  -
- Now, assume that x and $y$are correlated
- Covariance matrix is defined by
  - $\langle \hat{V}^{-1} \rangle_{ij} =$
- For a binned likelihood, where N is large and the likelihood can be reduced to a χ^2
  - $\langle \hat{V}^{-1} \rangle_{ij} =$

Correlated Uncertainties

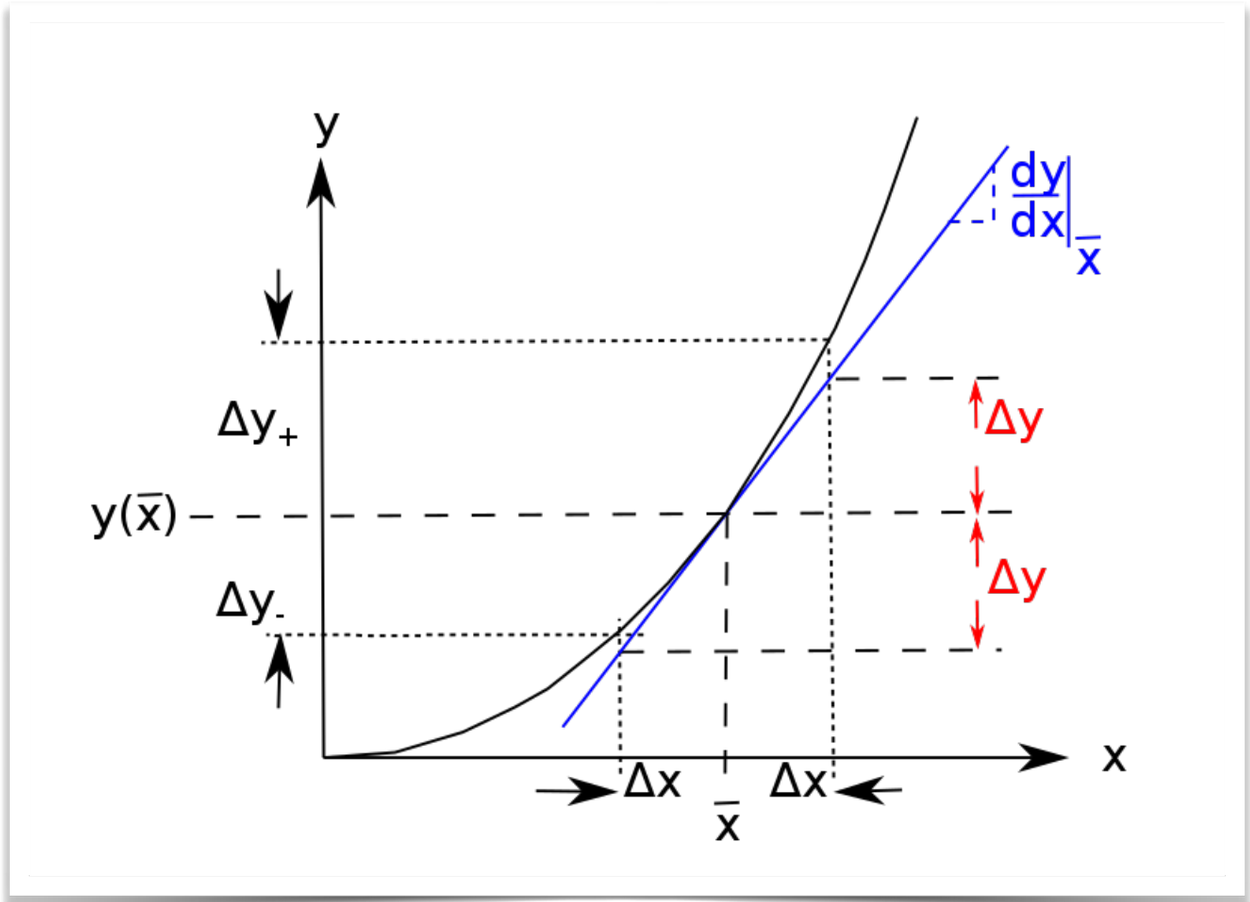
- Standard for two parameters with a correlation
- Slope related to correlation coefficient
- Correlation matrix typically determined from data numerically during fitting procedure



Correlation Example from Higgs



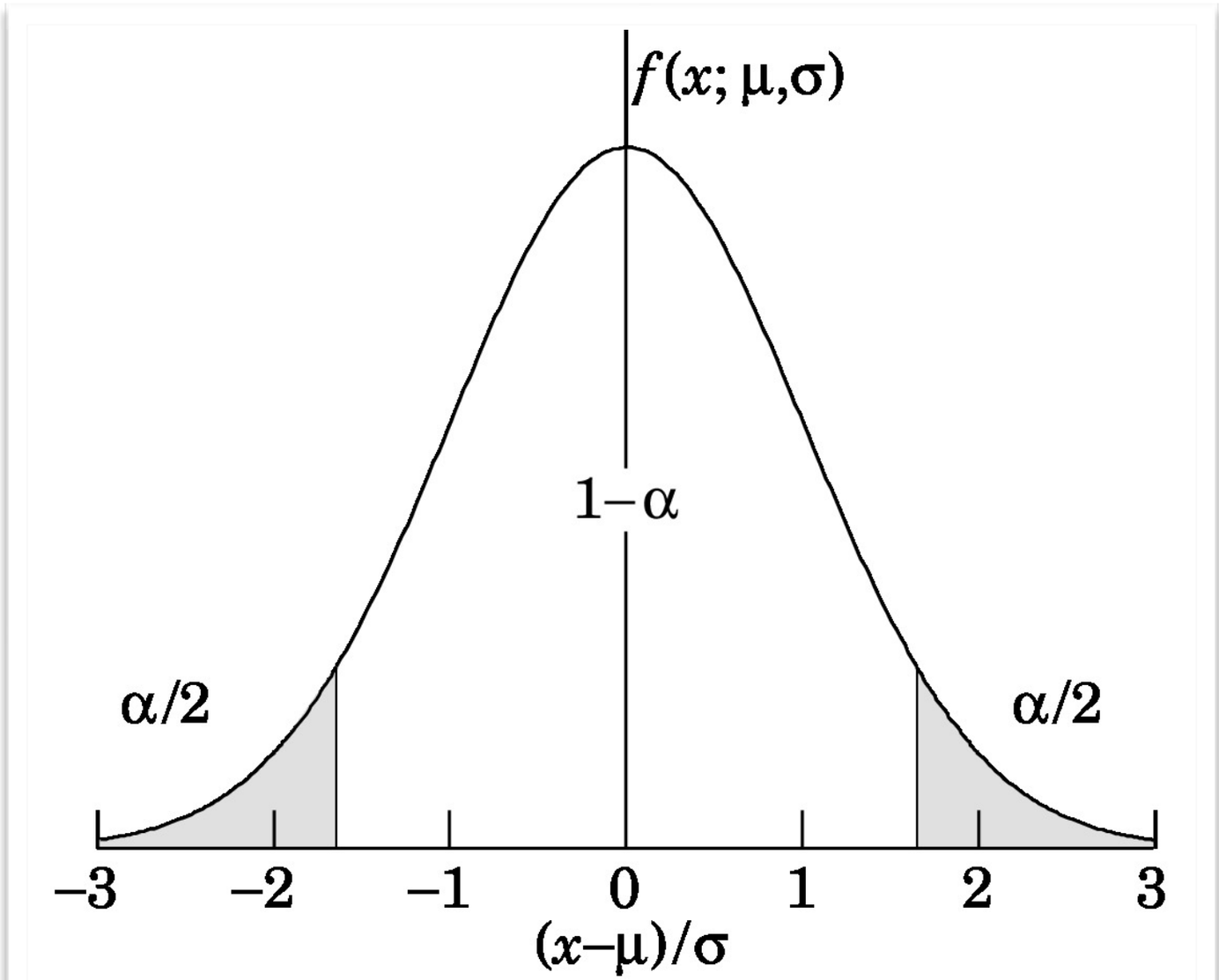
Propagation of Errors



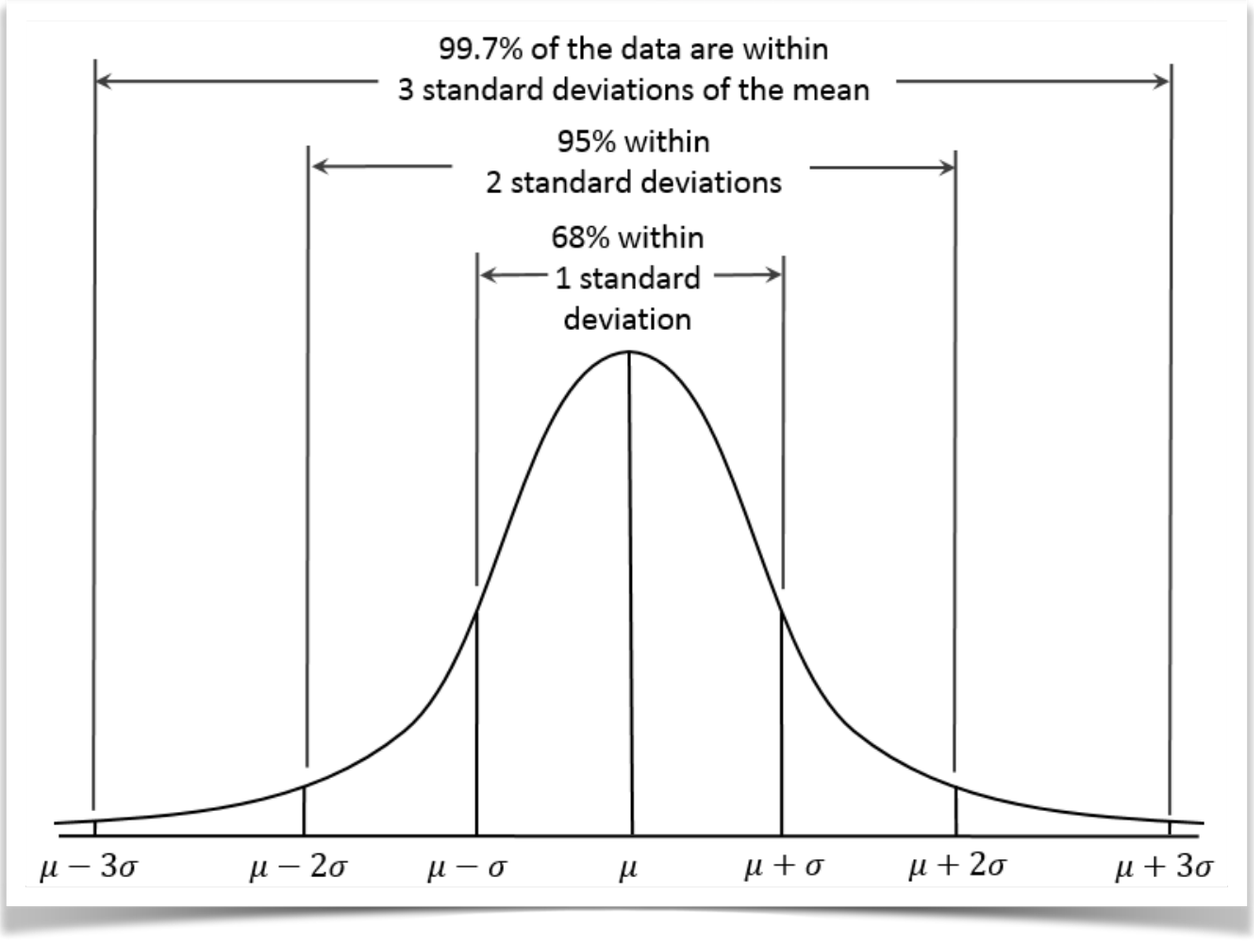
- Determine error on final measurement from known errors on input measurements
  - $\sigma_f^2 =$
- More dimensions are usually expressed as a matrix
- Useful reference: https://en.wikipedia.org/wiki/Propagation_of_uncertainty

Confidence Interval

- Fraction of the result not between and is
  - $1 - \alpha = \int$



Confidence Levels for a Gaussian



α	δ
0.3173	
4.55×10^{-2}	
2.7×10^{-3}	
5.7×10^{-7}	
2.0×10^{-9}	

Confidence Levels: Higgs