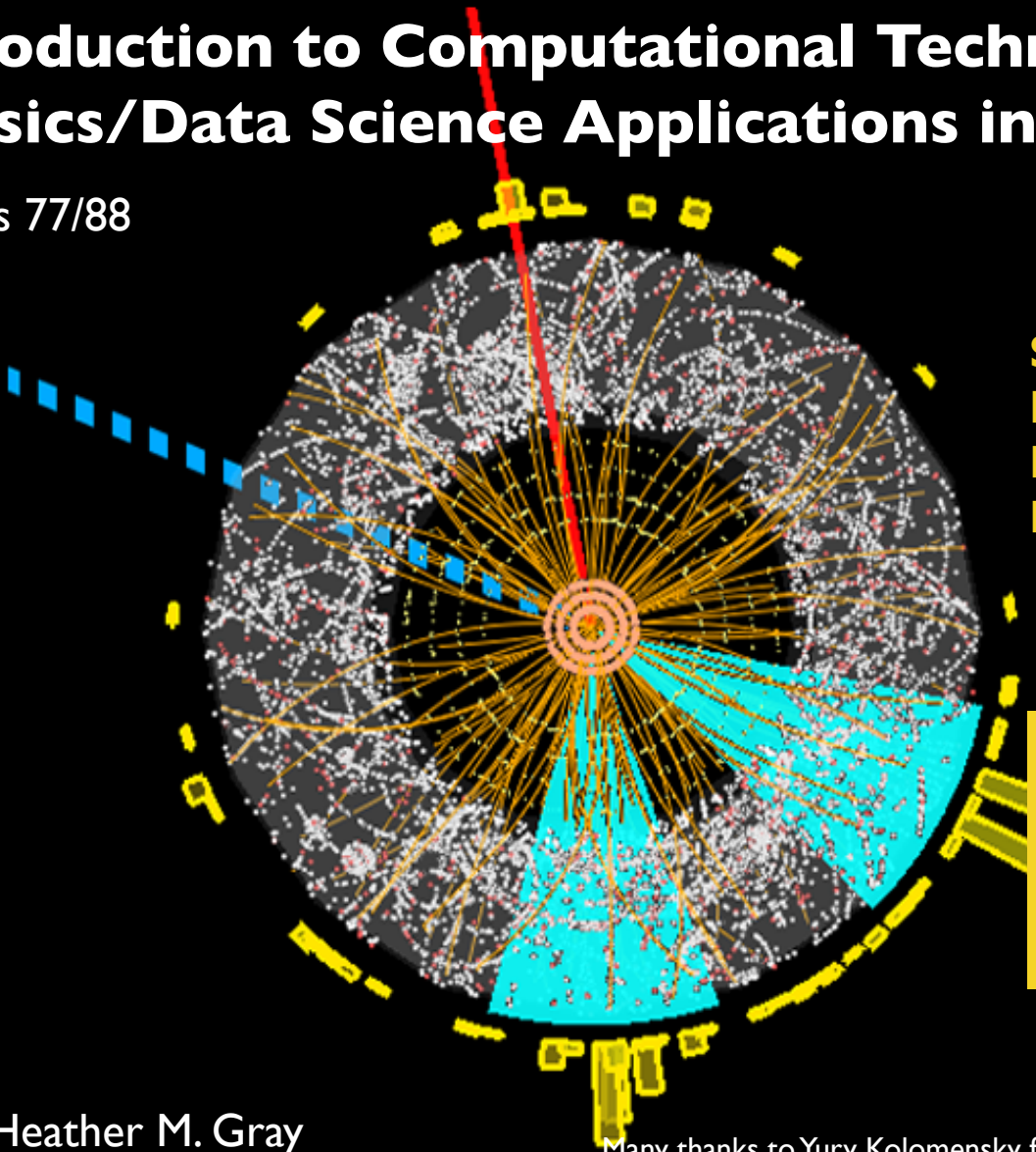
Introduction to Computational Techniques in Physics/Data Science Applications in Physics

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

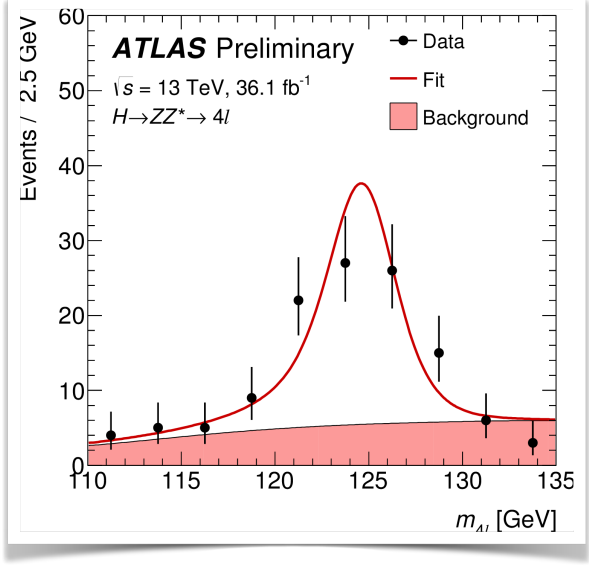
Statistics

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 qualitative and quantitative understanding
- From observation to a set of laws
  - Make a set of measurements
  - Summarise the result
  - Most conclusions are drawn with some degree of (un)certainty
- Example
  - Grarity exists
  - $F = G_N \frac{m_1 m_2}{r^2}$
  - $G_N = 6.67708 \pm 0.0031 \times 10^{-11}\ m^3\ kg^{-1}\ s^{-2}$
- Many measurements are a priori uncertain and have to be interpreted in probabilistic terms
- **Classical statistics**: estimate probabilities given a finite amount of data and test whether a given model is consistent with the data

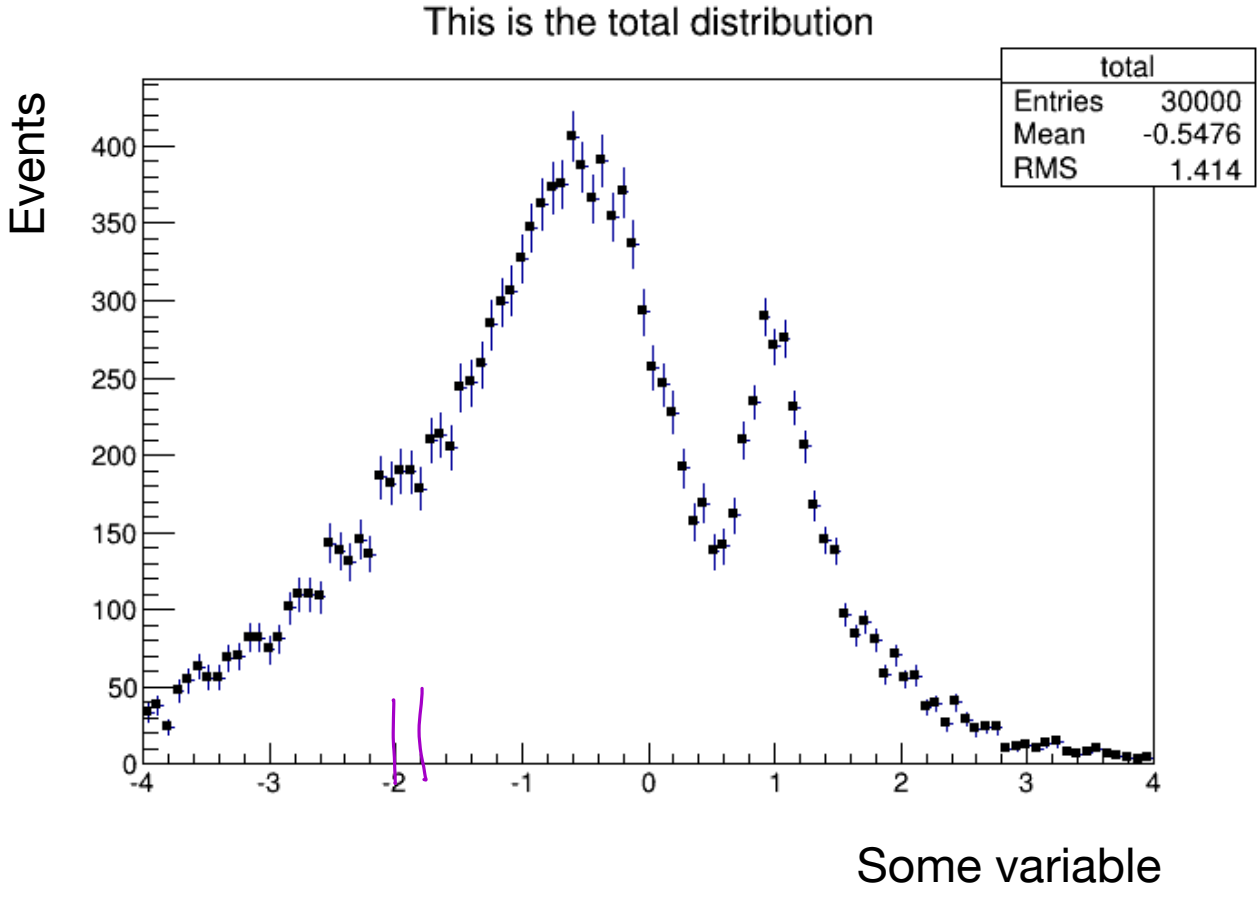
Describing the Data

- **Data**: result of a measurement
  - Physics: typically have quantitative data
    - number
  - Other fields deal with qualitative data
- **Numbers** are easier to handle mathematically; statistics will deal with quantitative measurements
  - discrete data, e.g. integers (counts)
  - continuous data, e.g. energies, momenta
    - Measure with some precision, set by the measuring apparatus or other external conditions

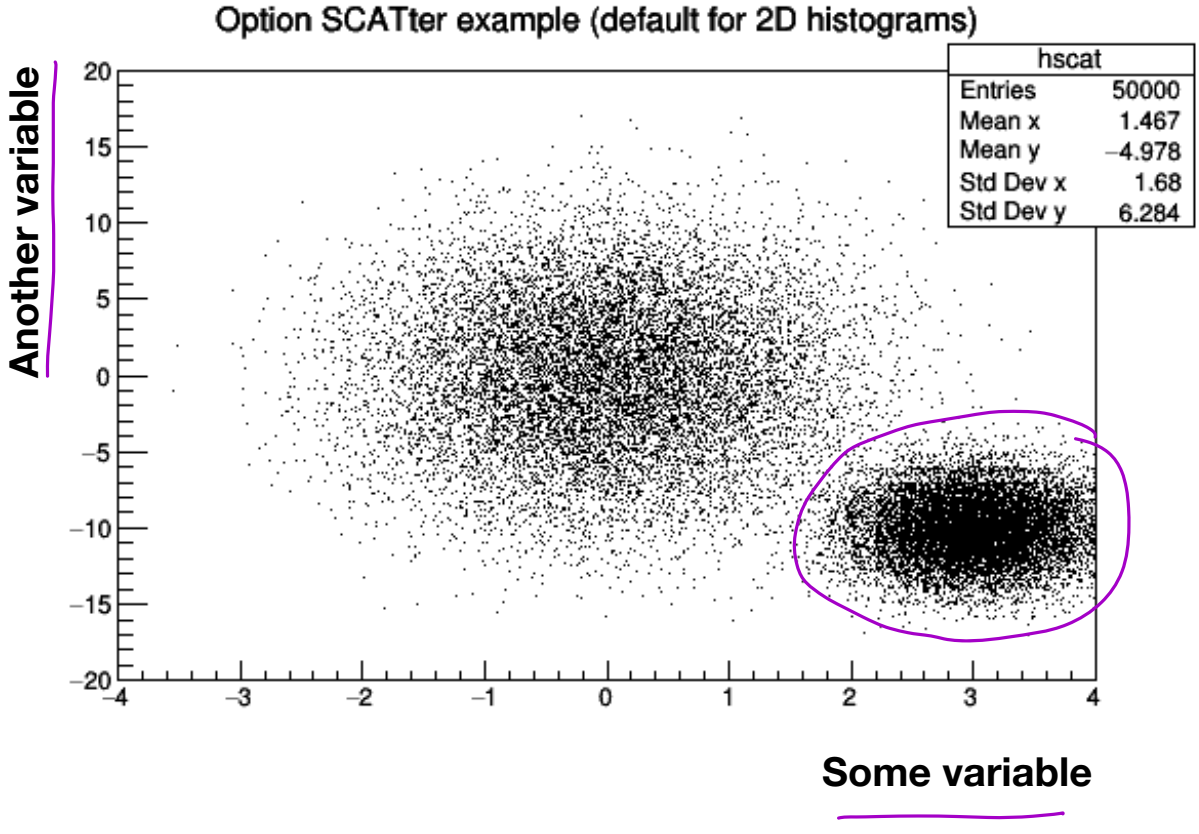


Why are there error bars on the data?

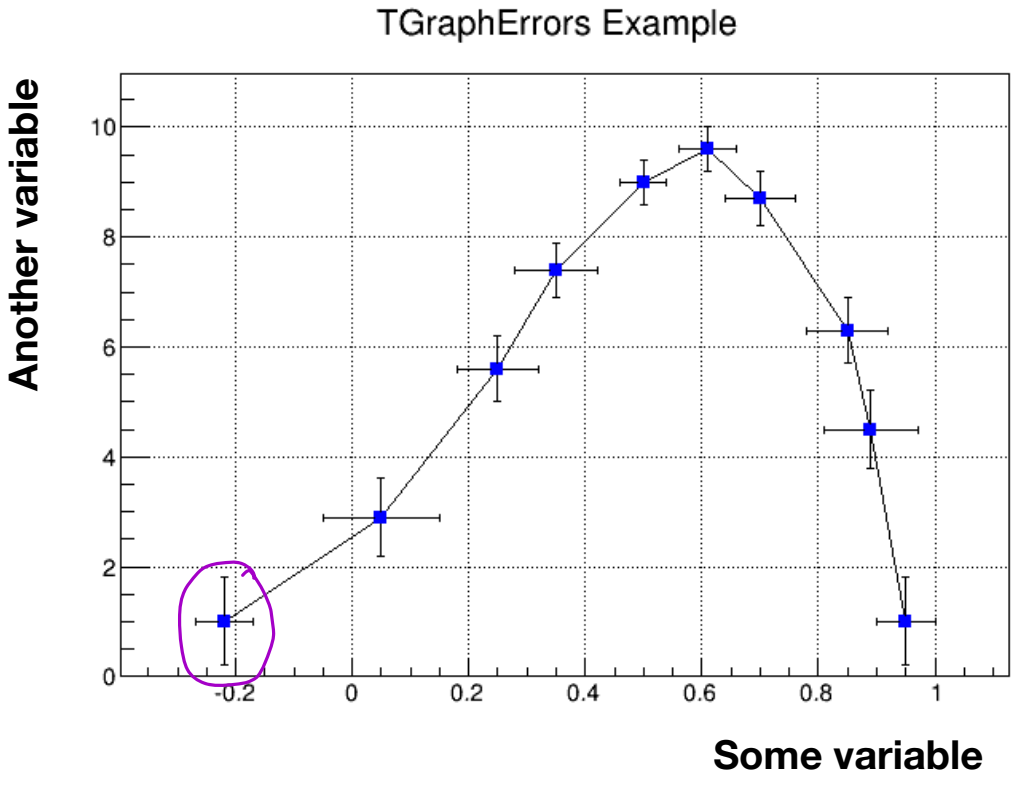
Histogram


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
0
2
4
6
8
10
-0.2
0
0.2
0.4
0.6
0.8
1

Uncertainty and Error

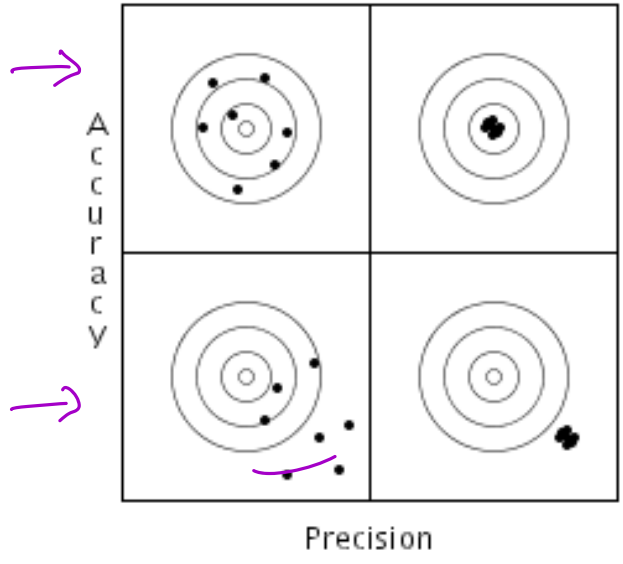
- In physics, the words uncertainty and error are used interchangeably to describe how far a particular measurement is expected to deviate from the true value — typically
  - Use symbol σ for the error
  - Formal definition is probabilistic: 68% chance to find the experimental result within $\pm 1\sigma$ of the true value
    (frequentist interpretation)
  - Often interpreted as a range of possible true values
    (Bayesian interpretation)
  - We'll come back to the difference between Bayesian and frequentist

Uncertainty and Error

- How do we define what is typical ?
  - Underlying assumption: our experiment is one sample of a popular of similar measurements
    - Derive the value of σ from the properties of the population
  - Implicit assumption: our experiment is mistake-free
    - i.e. all similar experiments would return similar results

Precision vs Accuracy

- Precision: spread of the data around the average value
  - Typically associated with statistical uncertainty
- Accuracy: deviation of the average value from true value
  - Bias
  - Typically associated with systematic uncertainty
- Bad data: outliers
  - Data inconsistent with distribution (e.g. noise, mistakes)



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

Golden Rules

- When reporting results of a measurement, always report its uncertainty
  - Round off values to 1-2 digits of uncertainty
    - Rule of thumb: 1 digit if the last digit is >4, 2 digits otherwise
    - x = 3.142 ± 0.024
    - y = 3.1 ± 0.6
- Uncertainty can come from the spread in the data and/or precision of the instrument
  - Rule of thumb: "half of the last digit"
  - Statistically correct: $\sigma_{inst} = \frac{\text{last digit}}{\sqrt{12}}$

Probability: Definitions

- For numerical data, probabilistic definition is often most convenient (and quantitative)
- Let's define probability now
  - Formally, it is a quantity that is defined by Kolmogorov axioms

  - 1. For every subset A in S $P(A) \geq 0$
  - 2. For disjoint subsets ($A \cap B = \emptyset$)
    $P(A \cup B) = P(A) + P(B)$
  - 3. $P(S) = 1$

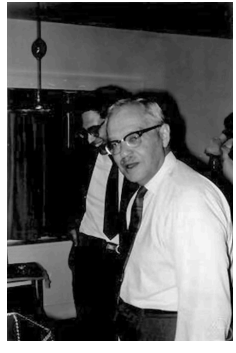
Two Interpretations

- "Frequentist" interpretation:
  - Probability is a limiting frequency a given outcome is reported when experiments are repeated an infinite number of times
    - Measurable parameters are represented by estimater with assigned confidence levels (CL)
      - CL measures a probability an estimator would fall in a certain range, given a true value of a parameter
    - No probability is assigned to constants of nature
- "Bayesian" interpretation:
  - More general: define probability as a degree of belief that a given statement is true
    - E.g. that the true value of parameter x is in interval $[a,b]$
    - This is somewhat subjective, but follows how most people think

Frequentist Probability

- Definitions
  - Let S be set of all possible outcomes of a measurement
  - Any subset A with only one element (single outcome) is the elementary outcome
  - Define $P(A) = \lim_{N \to \infty} (\#\text{occ of } A \text{ in } N \text{ trials})$
- Assume outcomes are (in principle) repeatable
- Confidence in a measurement grows with N
- Frequentist statistics is appropriate (and often argued for) in situations where measurements can be reproducibly repeated so that validity of approach can be tested (e.g. particle physics)

John von Neumann



Jerzy Neyman

Bayes Theorem

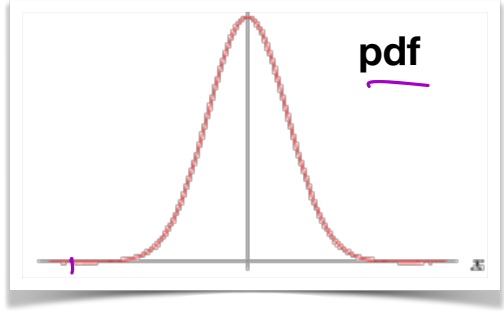
Thomas Bayes

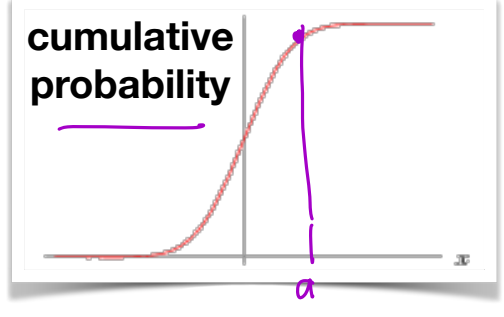
- Conditional probability of A given B
  - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Interpreted within Bayesian statistics as
  - $P(\text{theory}|\text{data}) \propto P(\text{data}|\text{theory})\,P(\text{theory})$

  posterior probability ↑ likelihood/result ↖ prior probability
- Allows one to interpret a single experiment as a measure of (subjective) probability that a given hypothesis is correct (e.g. that some fundamental const. is in some range)
- Requires assigning some probability to prior knowledge
  - That's where subjectivity comes in

Probability: Random Variables and PDFs

- For a continuous variable, x, we define the probability density function (pdf)
  - $f(x, \theta) =$ probability that x lies between x and $x+dx$
  - θ = parameters of the function
- Integrate to obtain the probability
  - $F(a) = \int_{-\infty}^{a} f(x)\, d(x)$
  - Probability that $x < a$
- Discrete variables: integral → sum $\sum_{x_i}^{q} f(x_i)$
- Expectation value
  - $E[u(x)] = \langle u(x) \rangle = \int_{-\infty}^{\infty} u(x) f(x)\, dx$





Expectation Values

- Expectation value of a function
  - $E[u(x)] = \int_{-\infty}^{\infty} u(x)\, f(x)\, dx$
- Moments of a function
  - $\alpha_n \equiv E(x^n) = \int_{-\infty}^{\infty} x^n f(x)\, dx$ nth moment
  - $m_n \equiv E[(x-\alpha)^n] = \int_{-\infty}^{\infty} (x-\alpha)^n f(x)\, dx$ nth central moment

Mean and Variance

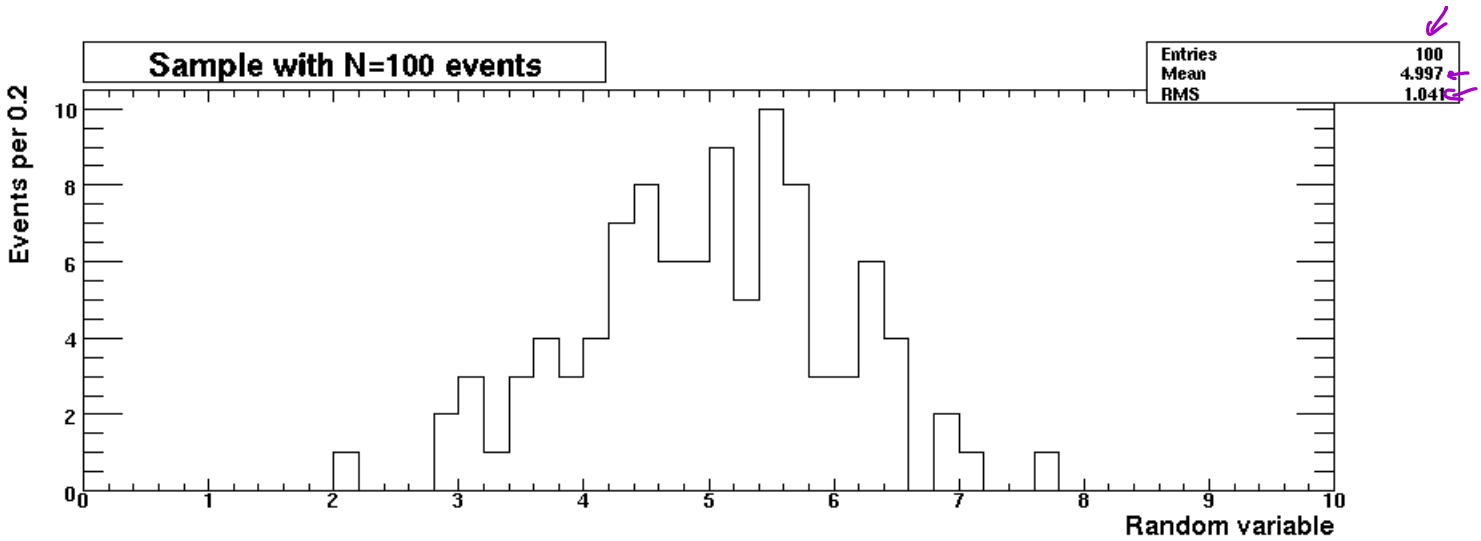
- Mean
  - $\mu = \int_{-\infty}^{\infty} x\, f(x)\, dx$
- Variance
  - $\sigma^2 = Var(x) = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2$
- σ is the standard deviation

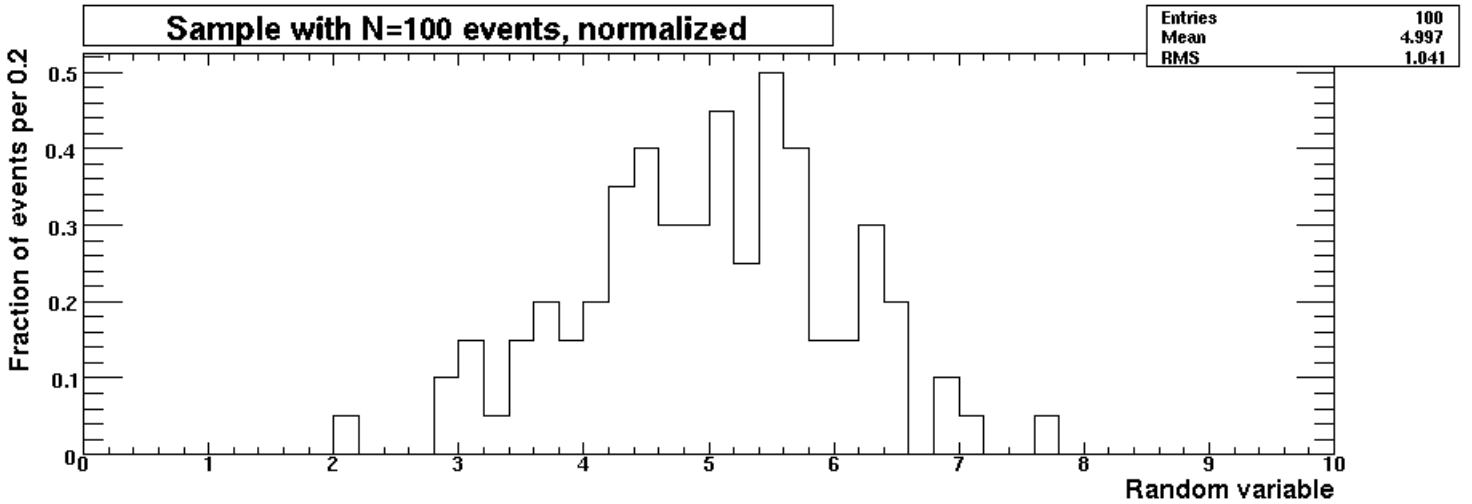
If you only remember one thing today, remember this

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

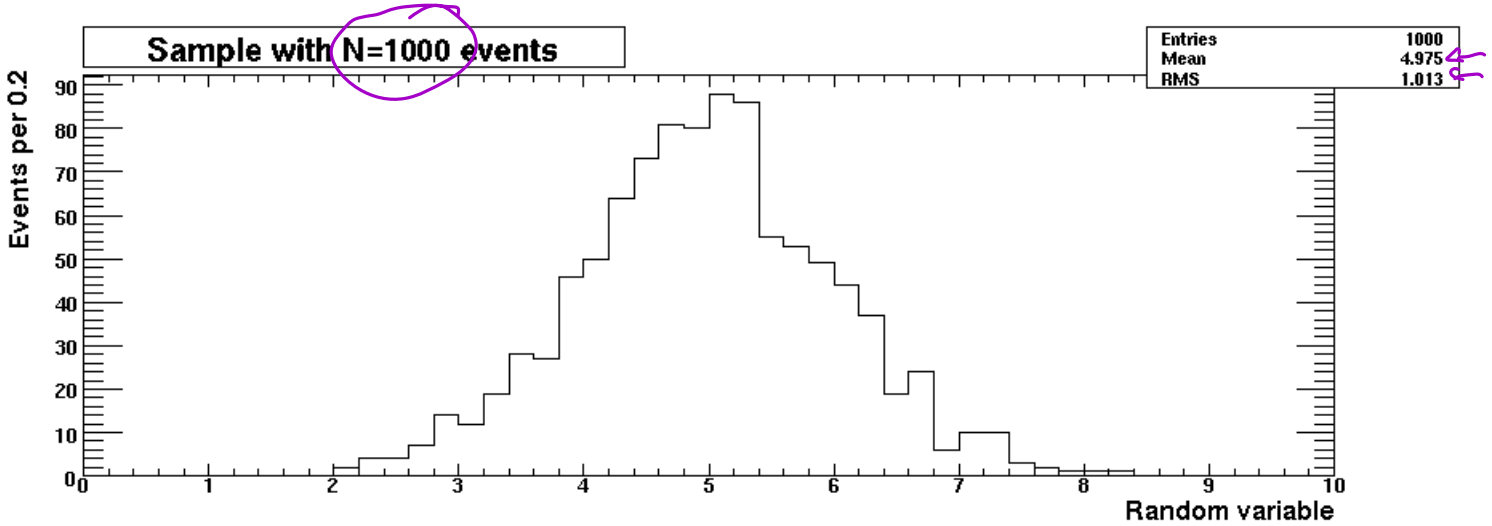
Example in jupyter notebook

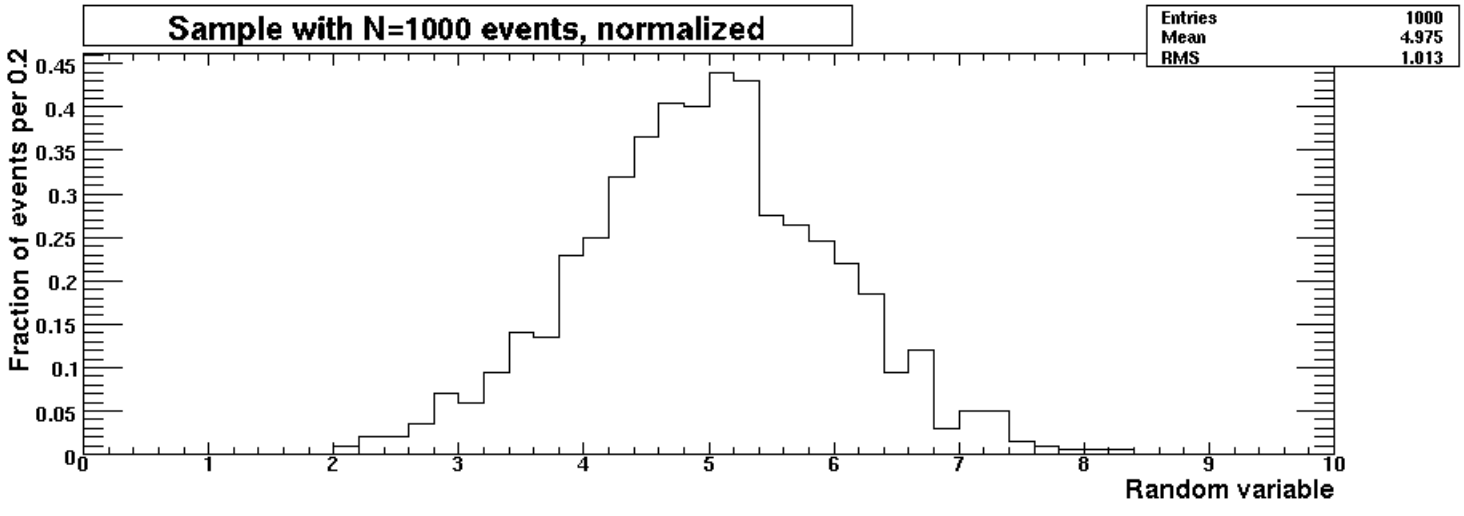
Sample from a Continuous PDF



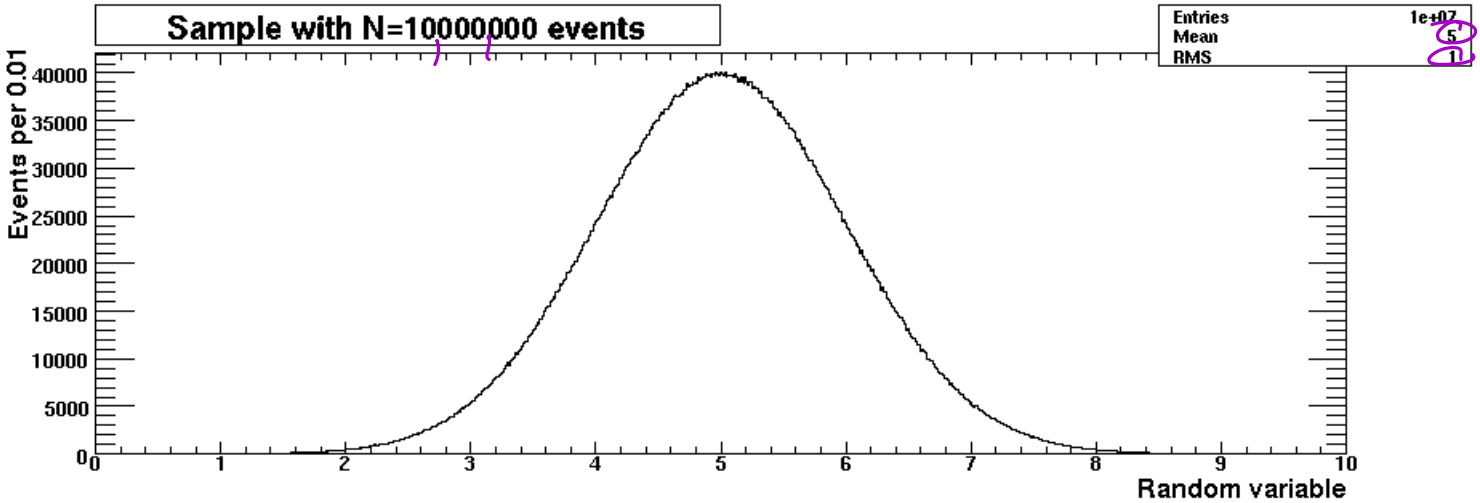


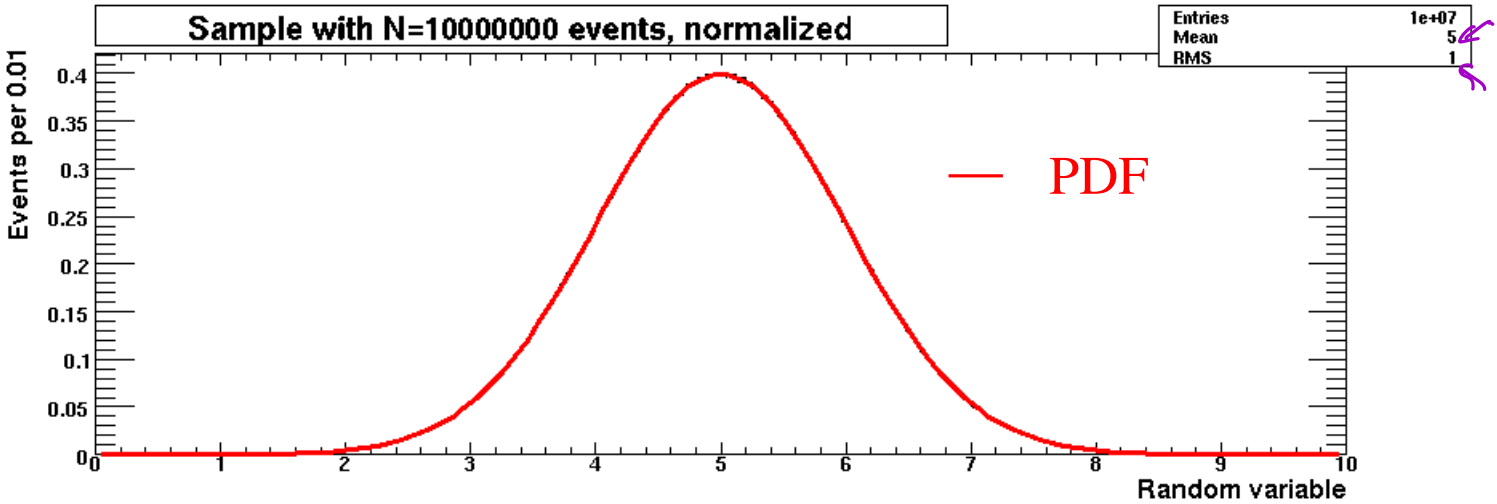
Sample From a Continuous PDF



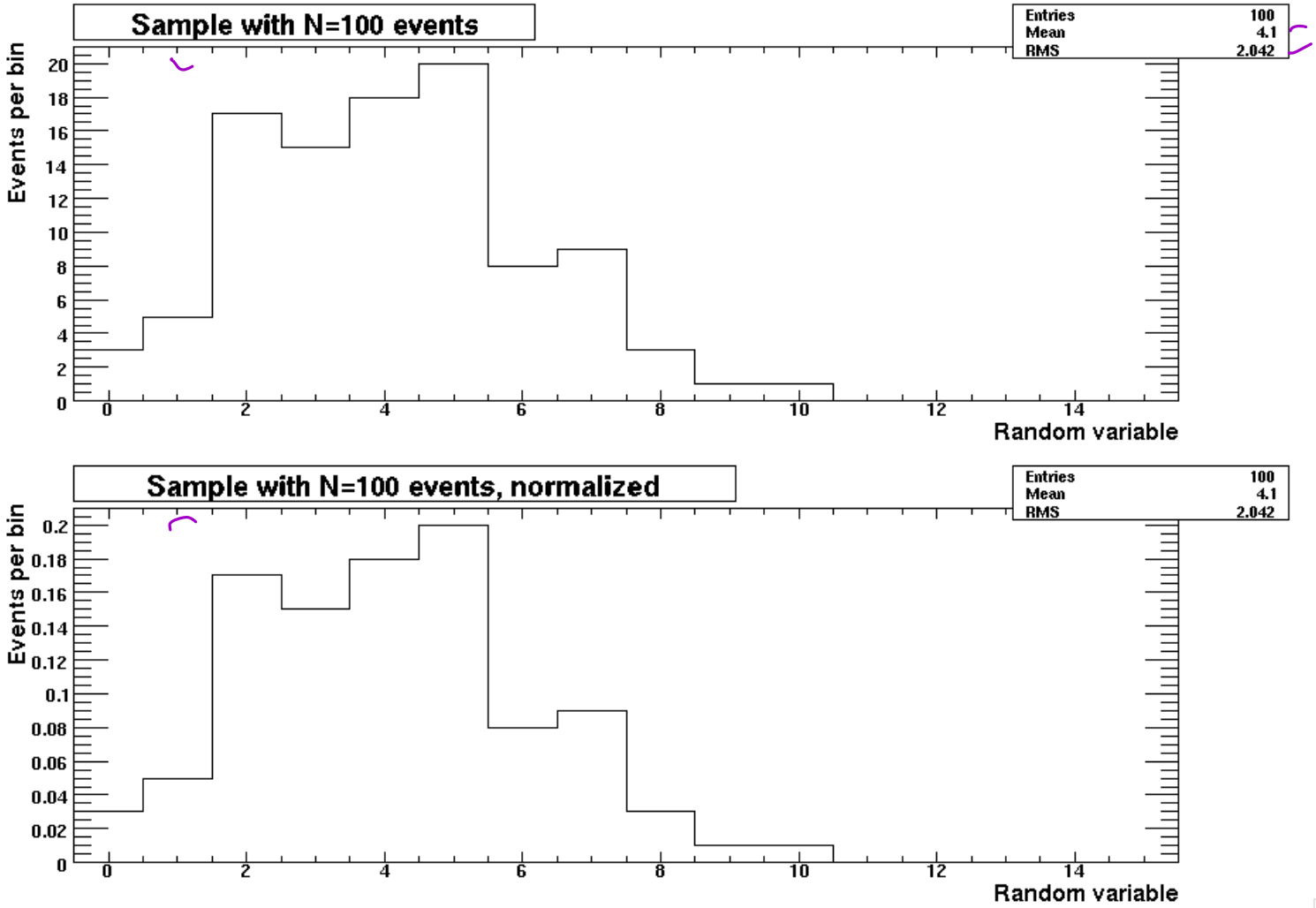


Sample From a Continuous PDF

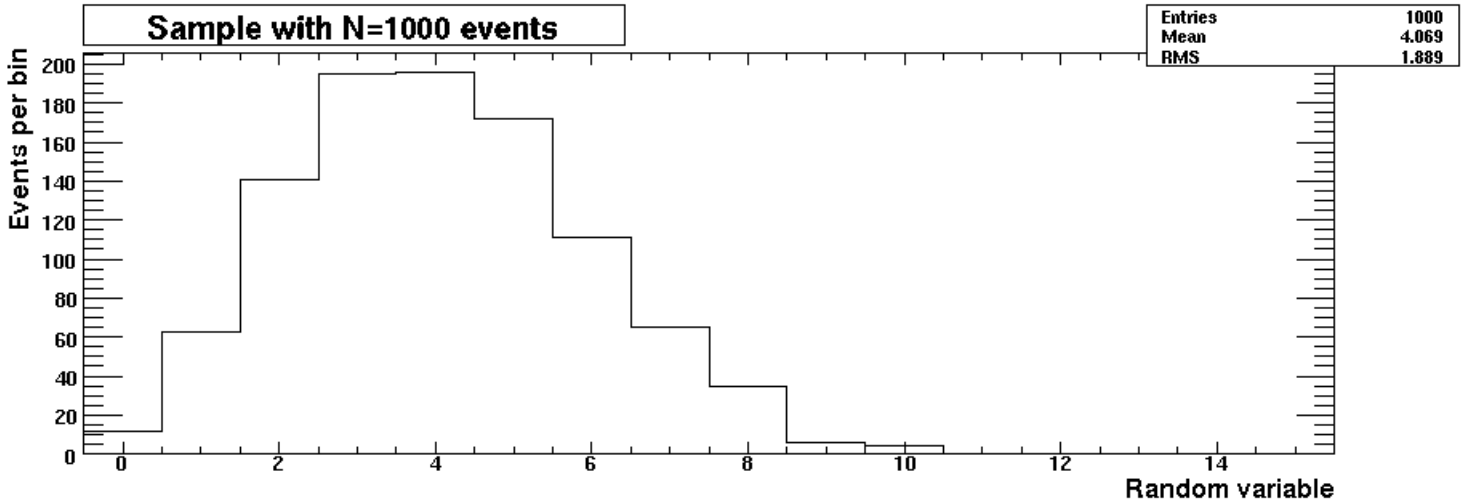


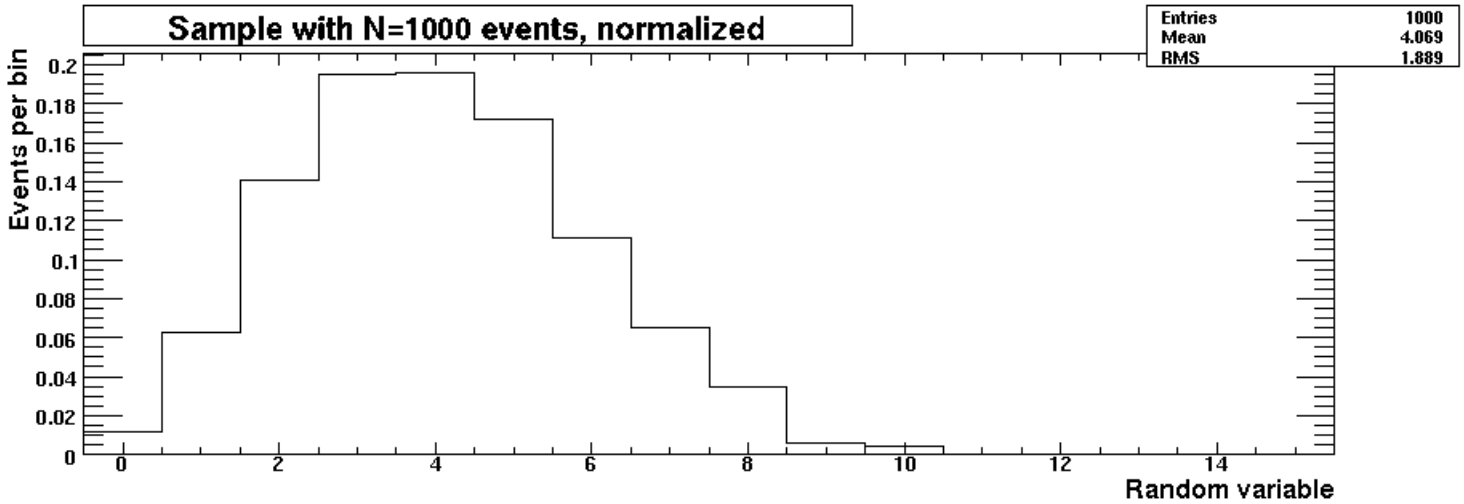


Sample From a Discrete PDF

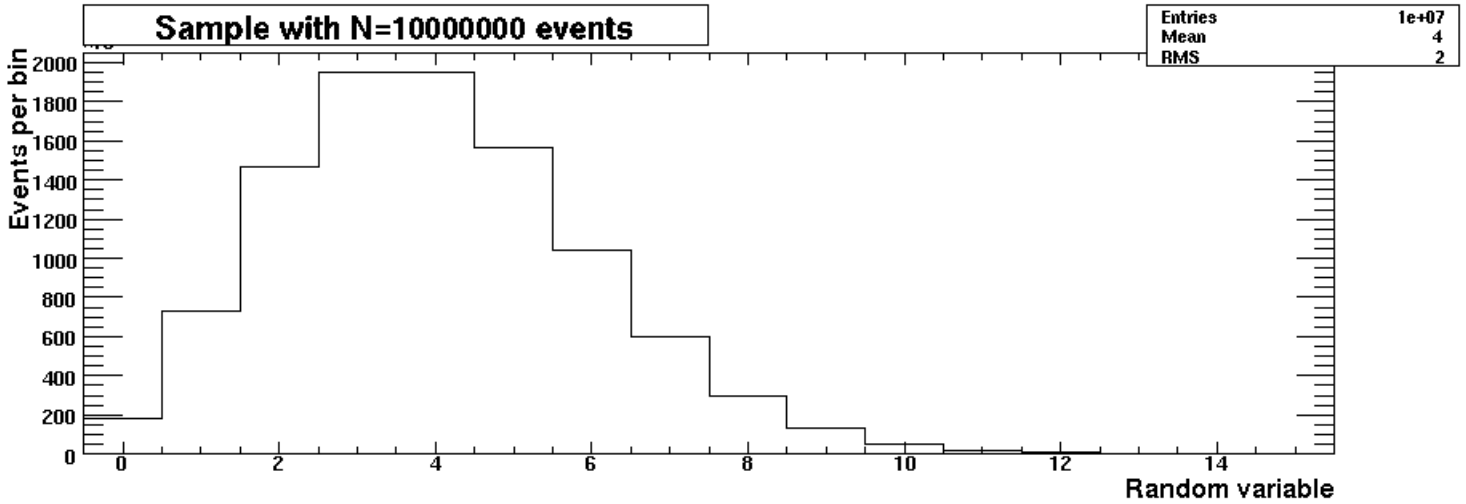


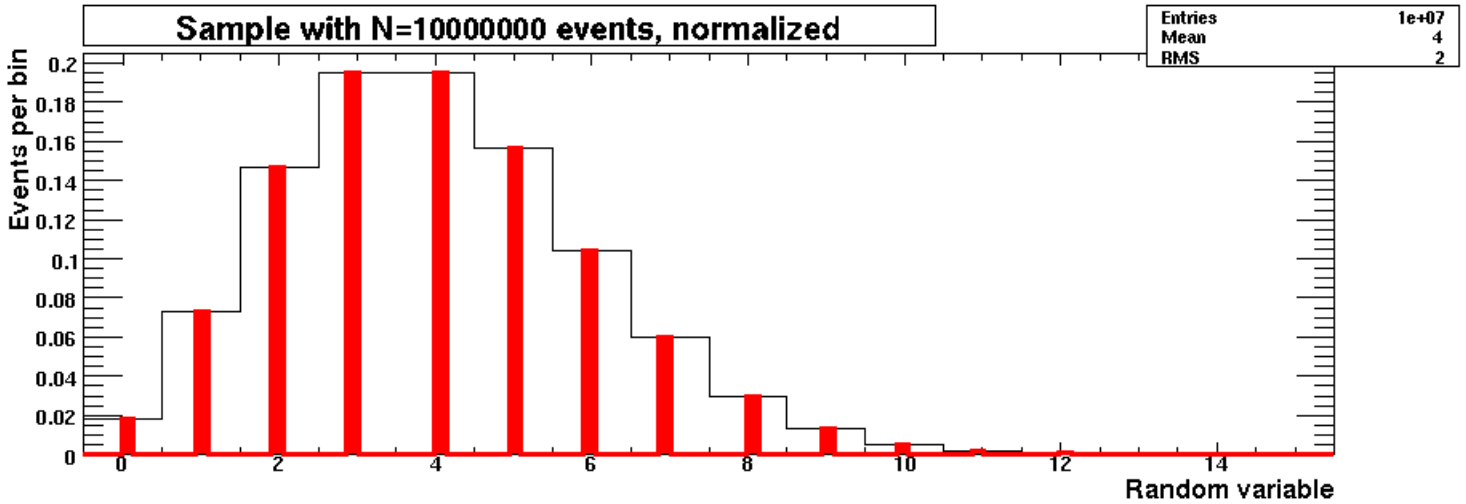
Sample From a Discrete PDF



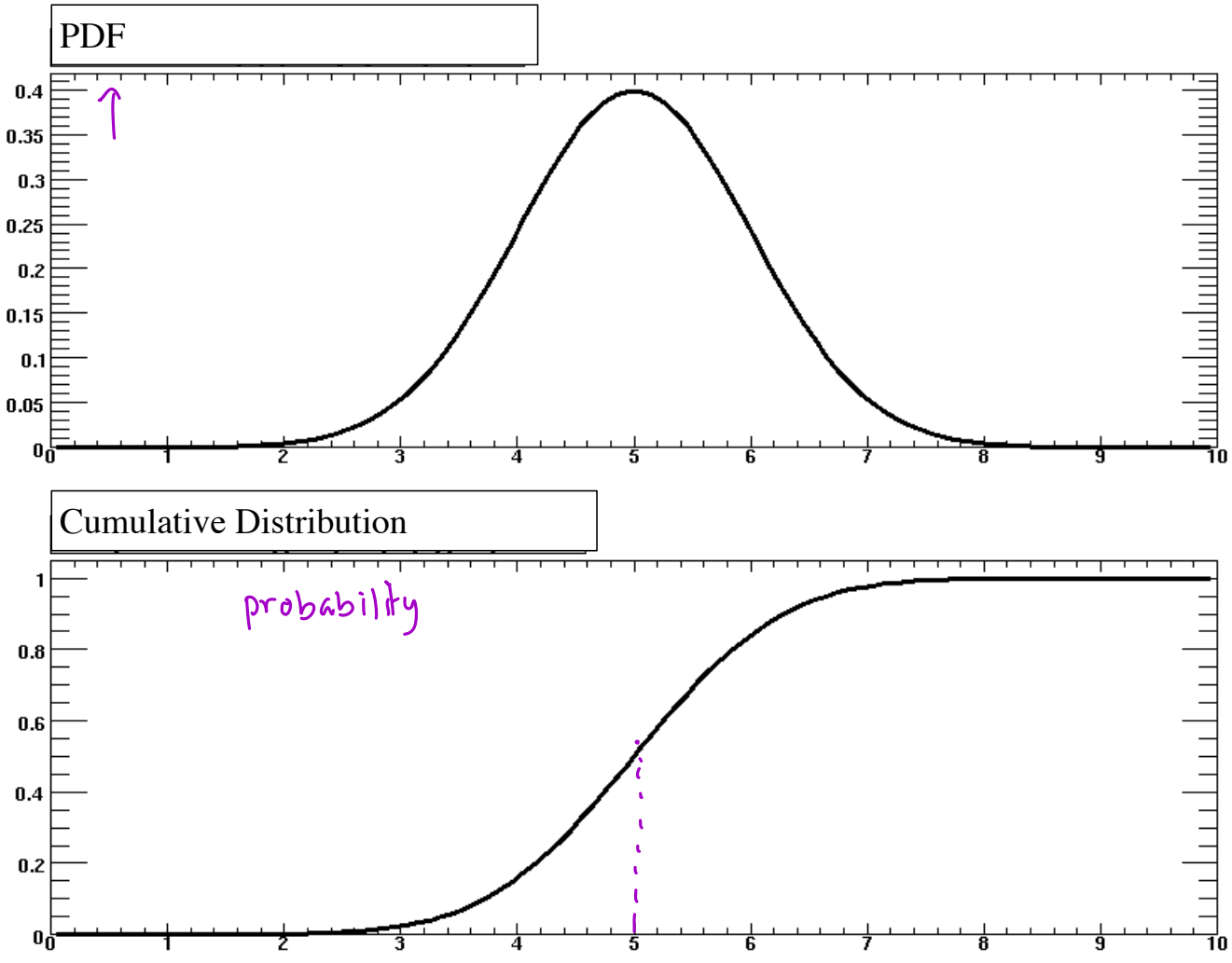


Sample From a Discrete PDF

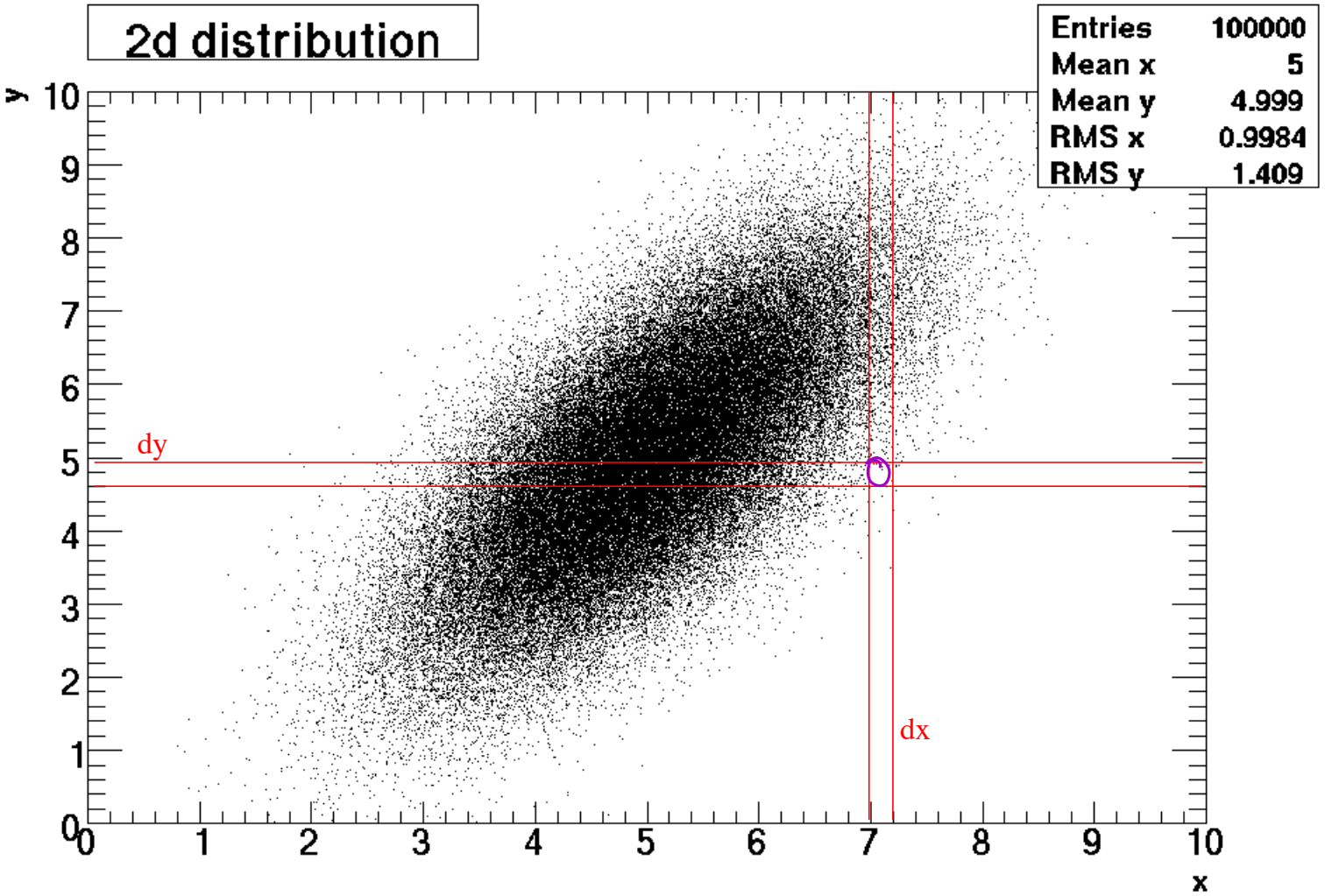




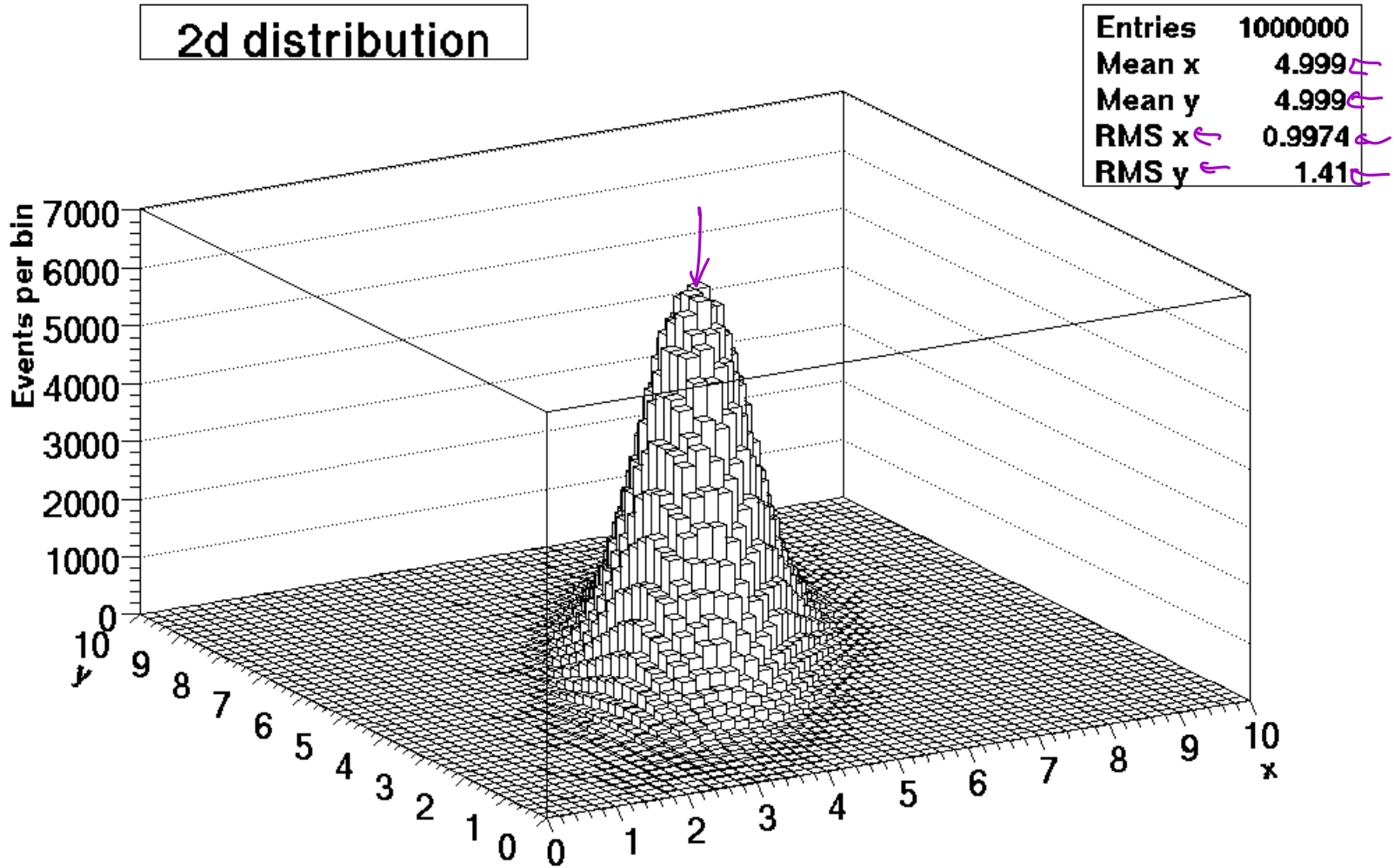
Cumulative Distribution



2d Distribution



2d Distribution



Most important distributions

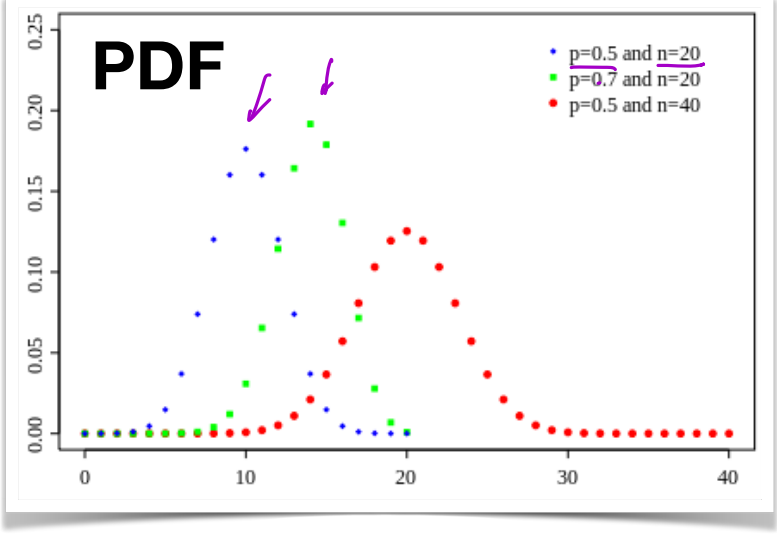
- Gaussian or normal distribution
  - central limit theorem → use all the time
- Poisson distribution
  - Number of events in data, e.g. rare event
  - Geiger counter clicks
- Binomial distribution
  - Efficiencies, # hits in detectors
- There are many others
  - uniform
  - exponential
  - Gamma
  - Chi2 or χ^2

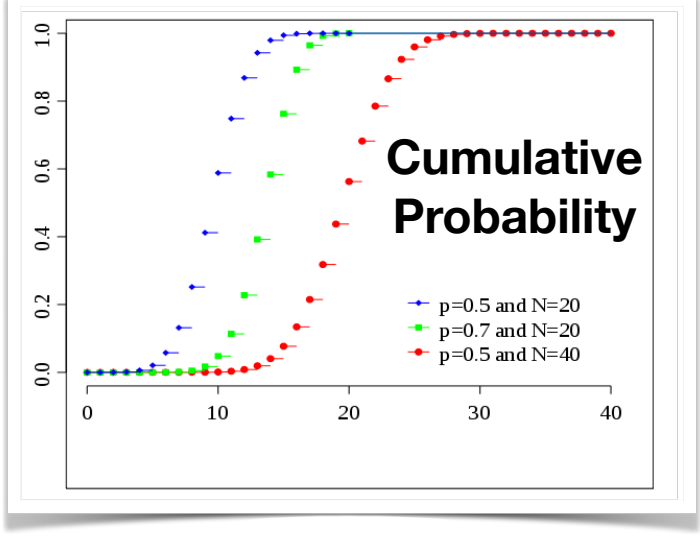
Binomial Distribution

- Random process with two possible outcomes
- $p =$ Prob of outcome 1 $q =$ Prob of outcome 2 $q = 1 - p$
- After n trials, prob of getting outcome #1 exactly k times is

$$f(k,p) = \binom{n}{k} p^k q^{n-k} \quad \text{where} \quad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

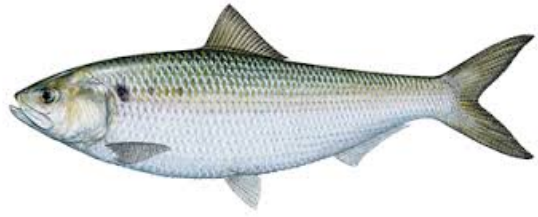
$$\mu = np \qquad \sigma = np(1-p)$$





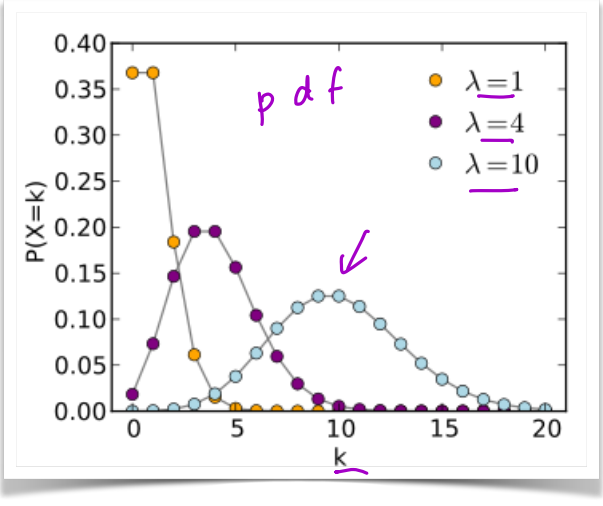
Poisson Distribution

- Prob of finding exactly k events in the interval x and $x+dx$ if the events occur with an rate of λ
  - $f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$

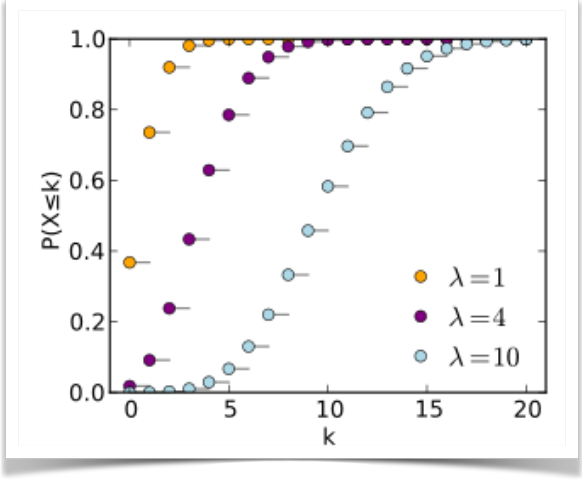


- For large λ approaches a Gaussian

$\mu = \lambda \qquad \sigma = \lambda$



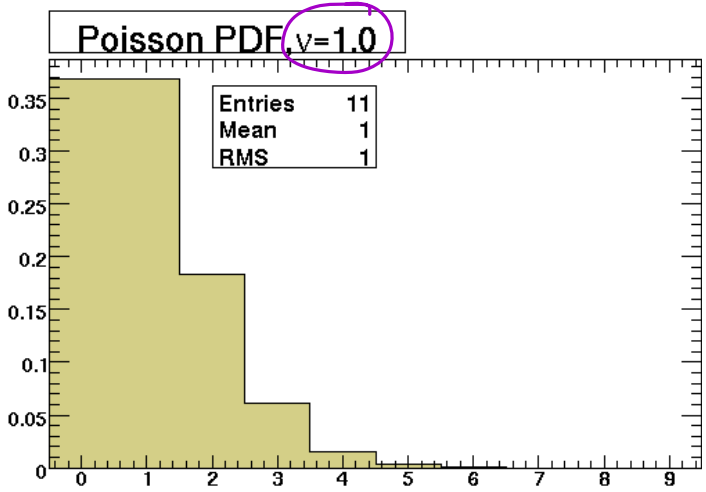
Cumulative probability

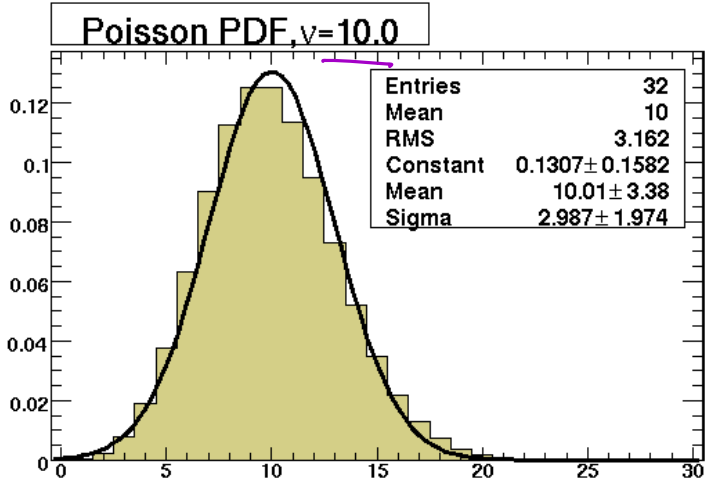


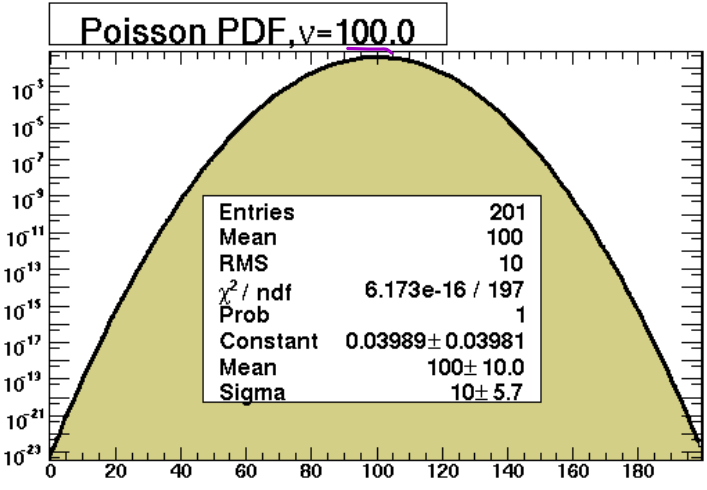
Example: Measure Efficiency

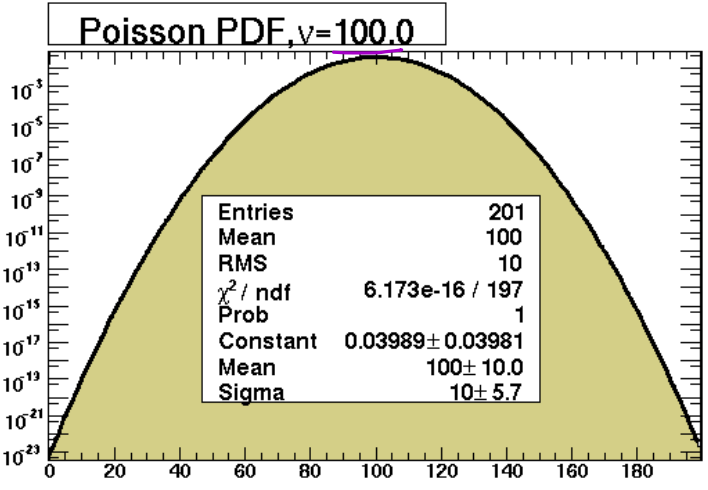
- Generate a sample of N events
- Apply selection; suppose n_{pass} events passed
- Estimate
  - $\hat{\epsilon} = \frac{n_{pass}}{N}$
  - $\sigma(\hat{\epsilon}) = \sqrt{V(\hat{\epsilon})} = \sqrt{\frac{V(n_{pass})}{N^2}} \simeq \sqrt{\frac{\hat{\epsilon}(1-\hat{\epsilon})}{N}}$
  - $\sigma(\hat{\epsilon}) \neq \frac{\sqrt{n_{pass}}}{N} = \sqrt{\frac{\hat{\epsilon}}{N}}$

Poisson Distribution



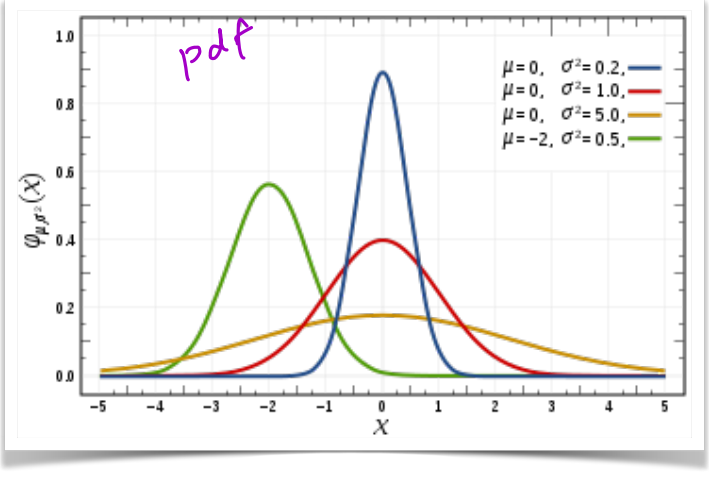


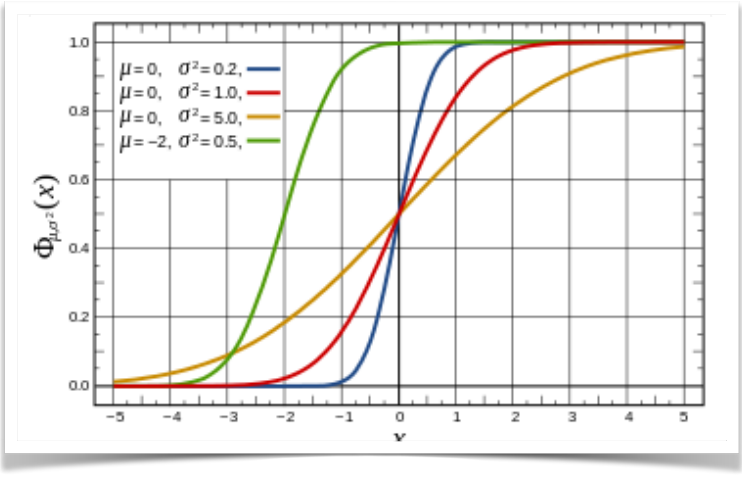




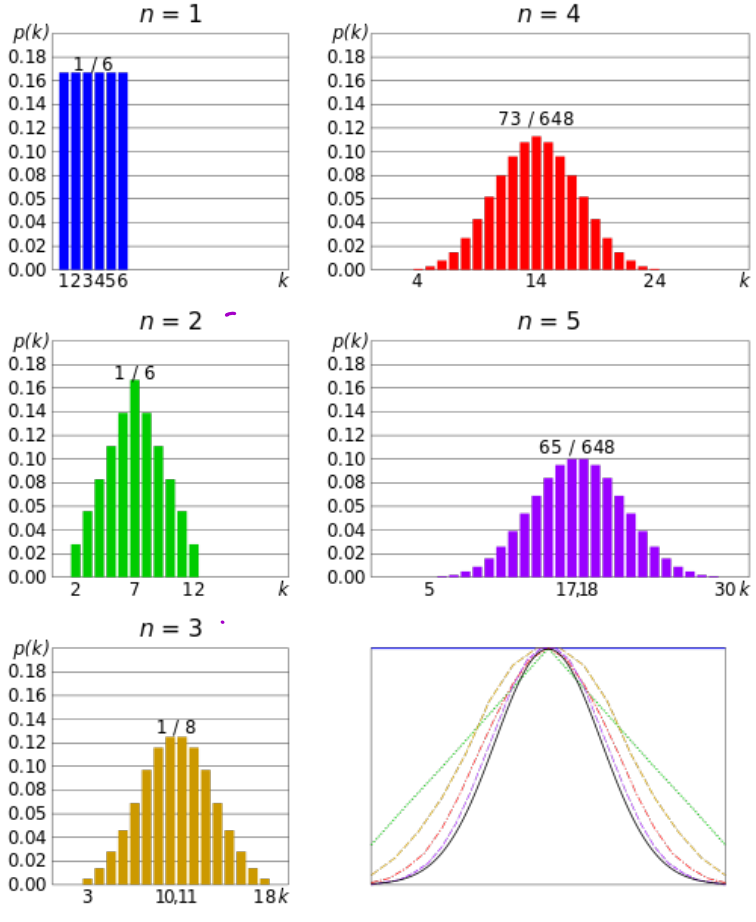
Gaussian Distribution

- **Central Limit Theorem**
  - Given a random sample $(x_1, x_2, \ldots, x_n)$ drawn from a pdf with mean μ and variance σ^2, if the mean is
    - $S/n = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - the distribution of S/n approaches the normal distribution as $n \to \infty$
    - $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$





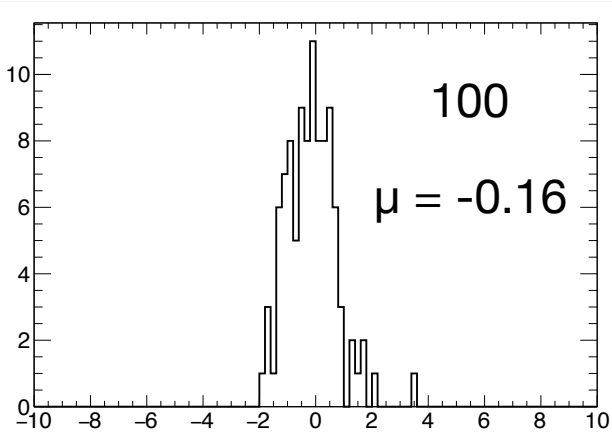
Central Limit Theorem

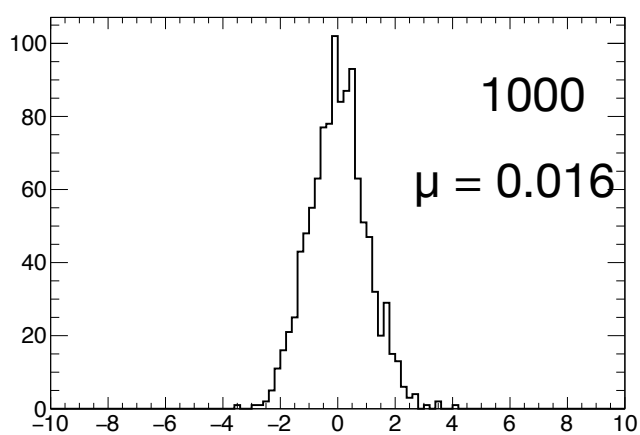


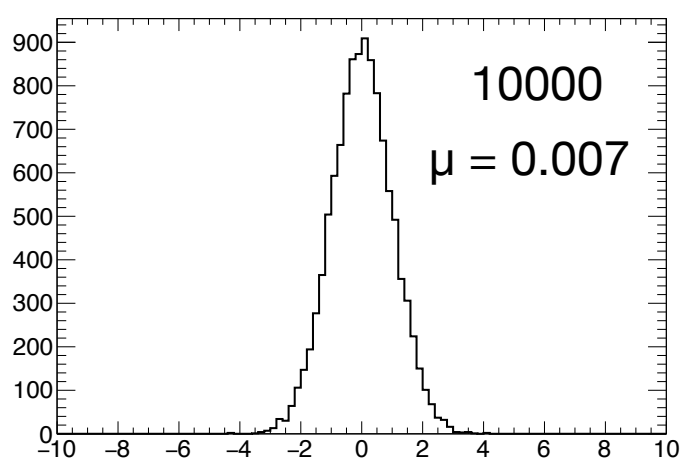
http://en.wikipedia.org/wiki/File:Dice_sum_central_limit_theorem.svg

Example

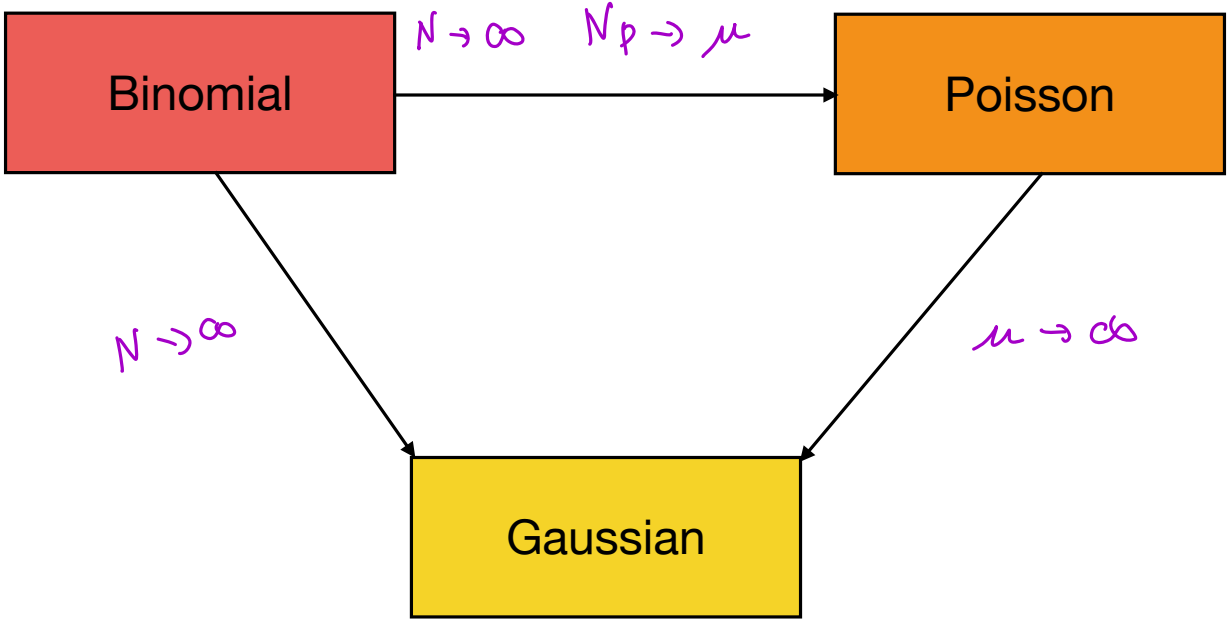
Example in jupyter notebook







Relationships between pdfs



Point Estimation

- Standard problem: set of values described by $x_1, x_2, \ldots, x_n$
  - $f(x) \equiv f(x_n; \theta)$
    - data
    - parameters
- Point estimation:
  - $\hat{\theta} = \theta(x_1, x_2, \ldots, x_n)$
    - Estimator of θ

Estimators

- Typical goal: estimate true values of one or more parameters from experimental data and understand the uncertainty on that measurement
- **Characteristics** of an estimator
  - consistency
    → approaches true value asymptotically for infinite data
  - bias
    → difference wrt true value for finite dataset
  - efficiency
    → variance of the estimator
  - sufficiency
    → depends on true value
  - robustness
    → sensitivity to data, e.g. outliers
- **Uncertainty**: how far the true parameter might be from our estimate due to statistical fluctuations in the ensemble of the measurements

Basic Estimators

- Estimators for mean and variance
- Shape of the PDF (fitting)
  - Maximum likelihood
    - Most efficient, but may be biased
    - Goodness-of-fit is not readily available
  - Least Chi-squared
    - maximum likelihood for Gaussian distributed data
    - Convenient for binned dat, analytic solns for linear functions
    - Automatic goodness-of-fit measure
    - Be careful of Gaussian approximation
      - (e.g. when Poisson becomes Gaussian)

Mean and Variance from a Sample

- Estimators (equally weighted data)
  - $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$ $\quad N > 0$
  - $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \hat{\mu})^2$ $\quad N > 0$
- Variances of these
  - $V[\hat{\mu}] = \frac{\sigma^2}{N}$ $\quad \sigma(\hat{\mu}) = \frac{\sigma}{\sqrt{N}}$
  - $V[\hat{\sigma}^2] = \frac{1}{N}\left(m_4 - \frac{N-3}{N-1}\sigma^4\right)$
  - $\sigma[\hat{\sigma}] = \frac{\sigma}{\sqrt{2N}}$ Gaussian data of x and large N

Likelihood Function

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

$$\mathscr{L}(x;\theta) = \prod_{i=1}^{n} \mathscr{L}_i$$

- Here θ represents one or more parameters

Log Likelihood

- To estimate the parameter(s), θ maximise the likelihood
  - Set derivative to zero
- Typically easier to maximise the $\ln \mathcal{L}$
  - $\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial}{\partial \theta} \ln \prod_{i=1}^{n} \mathcal{L}_i$

$$= \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln \mathcal{L}_i$$

$$= 0$$

$\ln(ab) = \ln a + \ln b$

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

$$\frac{\partial \mathcal{L}}{\partial \theta_1} = 0 \quad \frac{\partial \mathcal{L}}{\partial \theta_2} = 0 \quad \dots \quad \frac{\partial \mathcal{L}}{\partial \theta_i} = 0$$

Likelihood Example: Poisson

- N independent trials with results n_i
- Likelihood function for observing n_i if true mean is μ
  - $\mathscr{L}(n_i;\mu) = \dfrac{e^{-\mu}\mu^{n_i}}{n_i!}$
- Product over N measurements
  - $\mathscr{L}(\text{data},\mu) = \prod_{i=1}^{N} \dfrac{e^{-\mu}\mu^{n_i}}{n_i!}$
  - $\Rightarrow \ln\mathscr{L} = \sum_{i=1}^{N} \left(-\mu + n_i \ln\mu - \ln(n_i!)\right)$

$$= -N\mu + \sum_i^N n_i \ln\mu + \text{const}$$

$$\frac{\partial \ln\mathscr{L}}{\partial \mu} = -N + \sum_i^N \frac{n_i}{\mu} = 0$$

$$N = \frac{\sum_i^N n_i}{\mu}$$

$$\Rightarrow \mu = \frac{1}{N}\sum_{i=1}^{N} n_i$$

Best estimator is the mean value

Likelihood Example: Gaussian

- $G(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- Take the derivative of the log likelihood
  - $\frac{\partial}{\partial \mu}(\ln \mathcal{L})\big|_{\hat{\mu}=\mu} = \frac{\partial}{\partial \mu}\left(-\sum_i \frac{(x_i-\mu)^2}{2\sigma^2} + \text{const}\right)$

$$= \sum_i \frac{x_i - \mu}{\sigma^2}$$

$$\Rightarrow \hat{\mu} = \frac{1}{N}\sum_i x_i$$

- The unbiased estimator for σ is
  - $\hat{\sigma} = \frac{1}{N-1}\sum_i (x_i - \mu)^2$

Binned vs unbinned likelihood functions

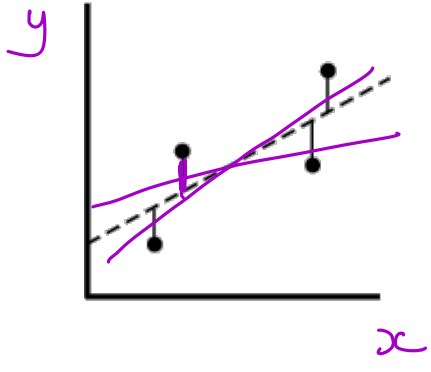
- Likelihood formalism works for any well-behaved probability density function (pdf)
- Product of the likelihood is a product over measurements
- **Example measurement**: Measure the lifetime of a particle of a given species for an ensemble of such particles produced at $t = 0$ such that the # p at time t :
  - $f(t) = \frac{1}{\tau} e^{-t/\tau}$
- Consider two ways to construct the likelihood
  - For decay i measure t_i and take the product of all measured times $\mathcal{L}$ (unbinned likelihood)
  - Make a histogram of the number of decays in bins of time
    - Measurement is the number of decays in each time bin (binned likelihood)

The Likelihood and χ^2

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

Method of Least Squares

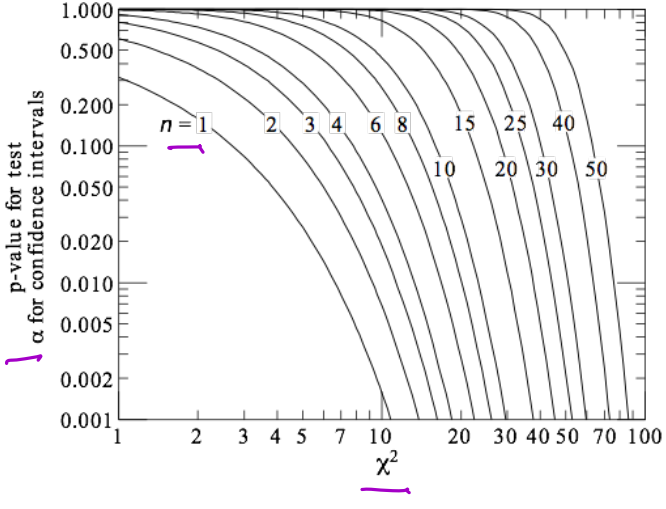
- Assume we have enough statistics for our measurements such that we can assume we are in the Gaussian regime
- Goal: Find the best estimater of parameter of a function that describes the data
- Minimize the distance of the data from the fit
  - Account for the uncertainties
- Scatter defined by χ^2
  - $\chi 2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2}$
- Can write the χ^2 in terms of our observables
  - $\chi 2 = \sum_{i=1}^{N} \frac{(y_i - F(x_i, \theta))^2}{\sigma^2}$
- Minimise χ^2 with respect to θ
- Useful when minimising $\ln \mathscr{L}$ is slow (high statistics samples)



Example: chi-squared p-values

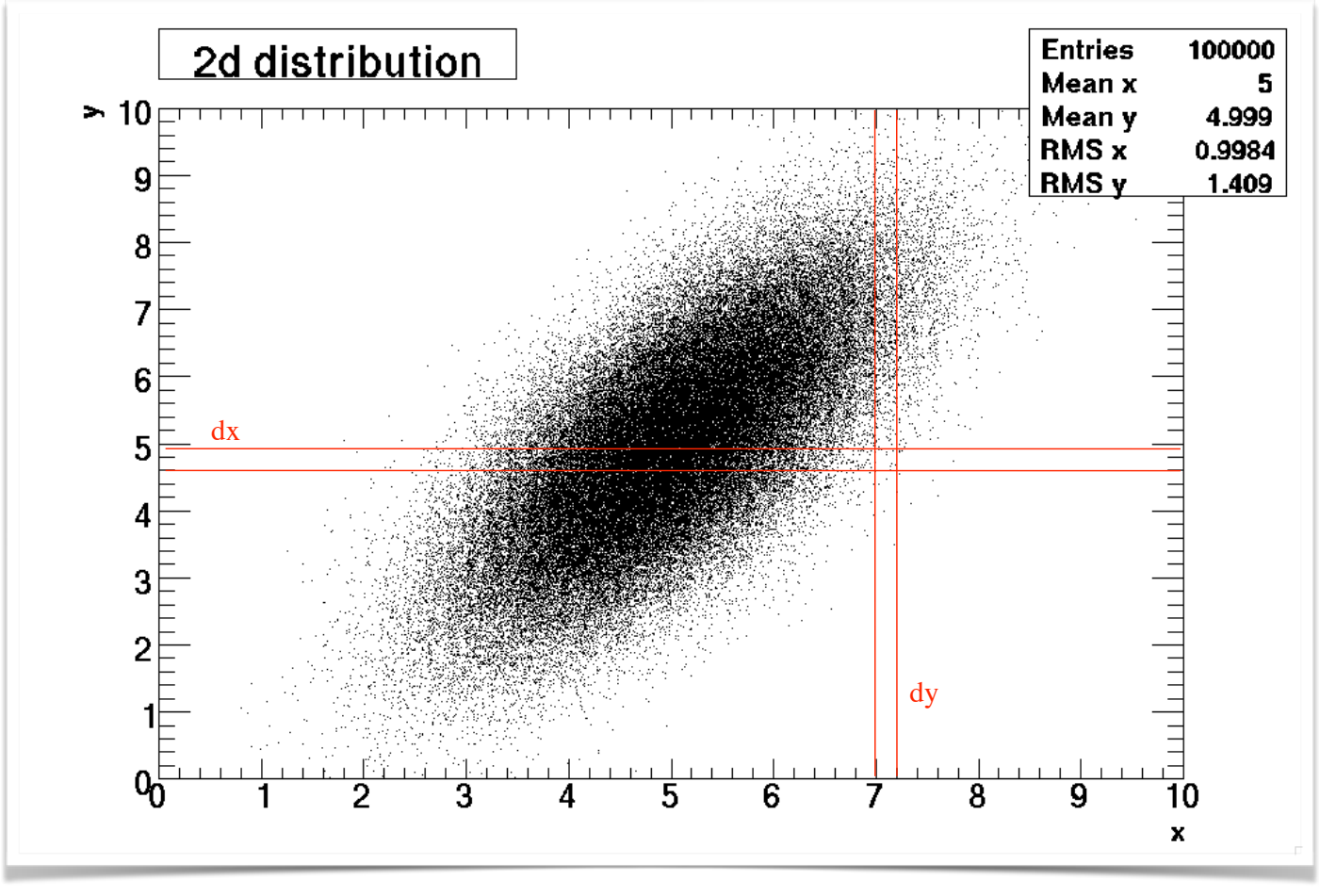
- One advantage of a χ^2 is that the value of the $\min \chi^2$ can be interpreted as a goodness-of-fit
  - iff errors on each data point are known and
  - the noise (distribution of data around their expected value) are Gaussian
- In the plot below
  - $n =$ number of degrees of freedom n dof $= N_{data} - N_{parameters}$
- For a good fit, expect χ^2 close to be close to dof

$\chi^2/dof \sim 1$

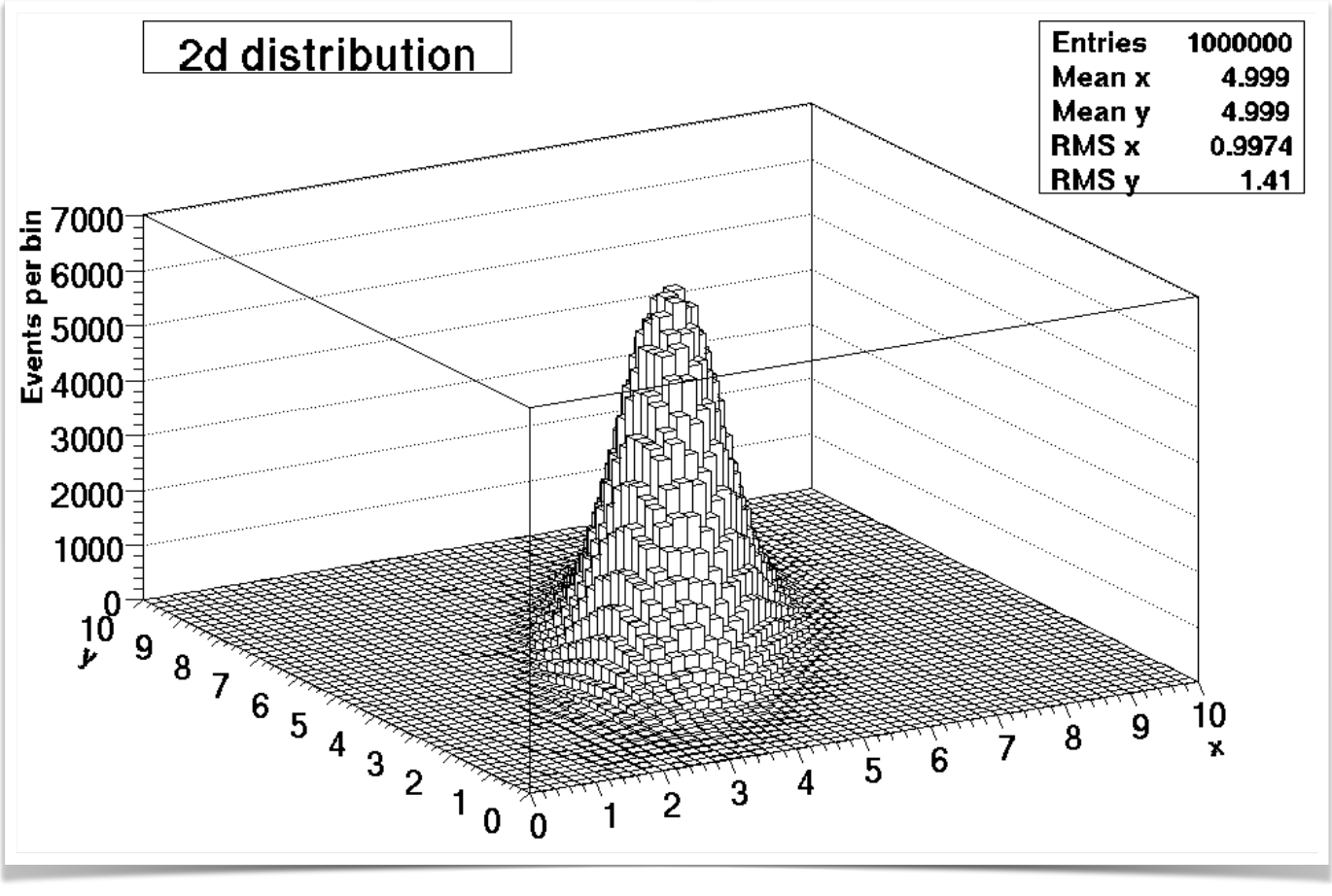


Example in jupyter notebook

2D distribution



2d distribution



Covariance and Correlation

Covariance Matrix

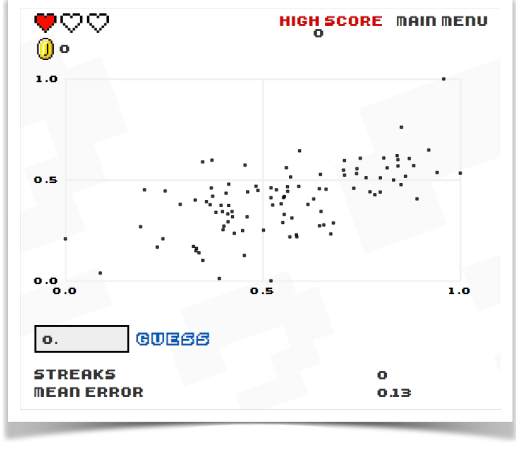
$$cov[x, y] = E\left((x - \mu_x)(y - \mu_y)\right] = E[xy] - \mu_x \mu_y$$

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

Correlation (normalised covariance)

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

$$\rho_{xy} = \frac{cov[x, y]}{\sigma_x \sigma_y}$$



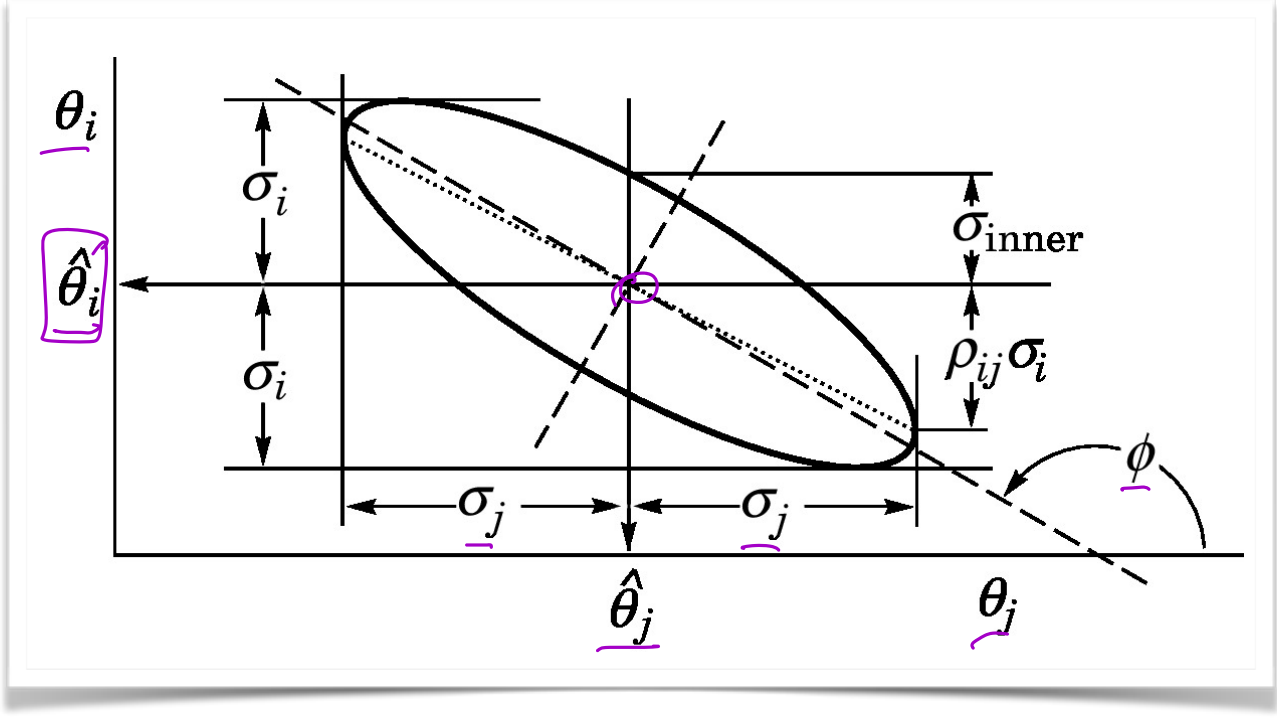
http://guessthecorrelation.com/

Covariance Matrix for a Gaussian

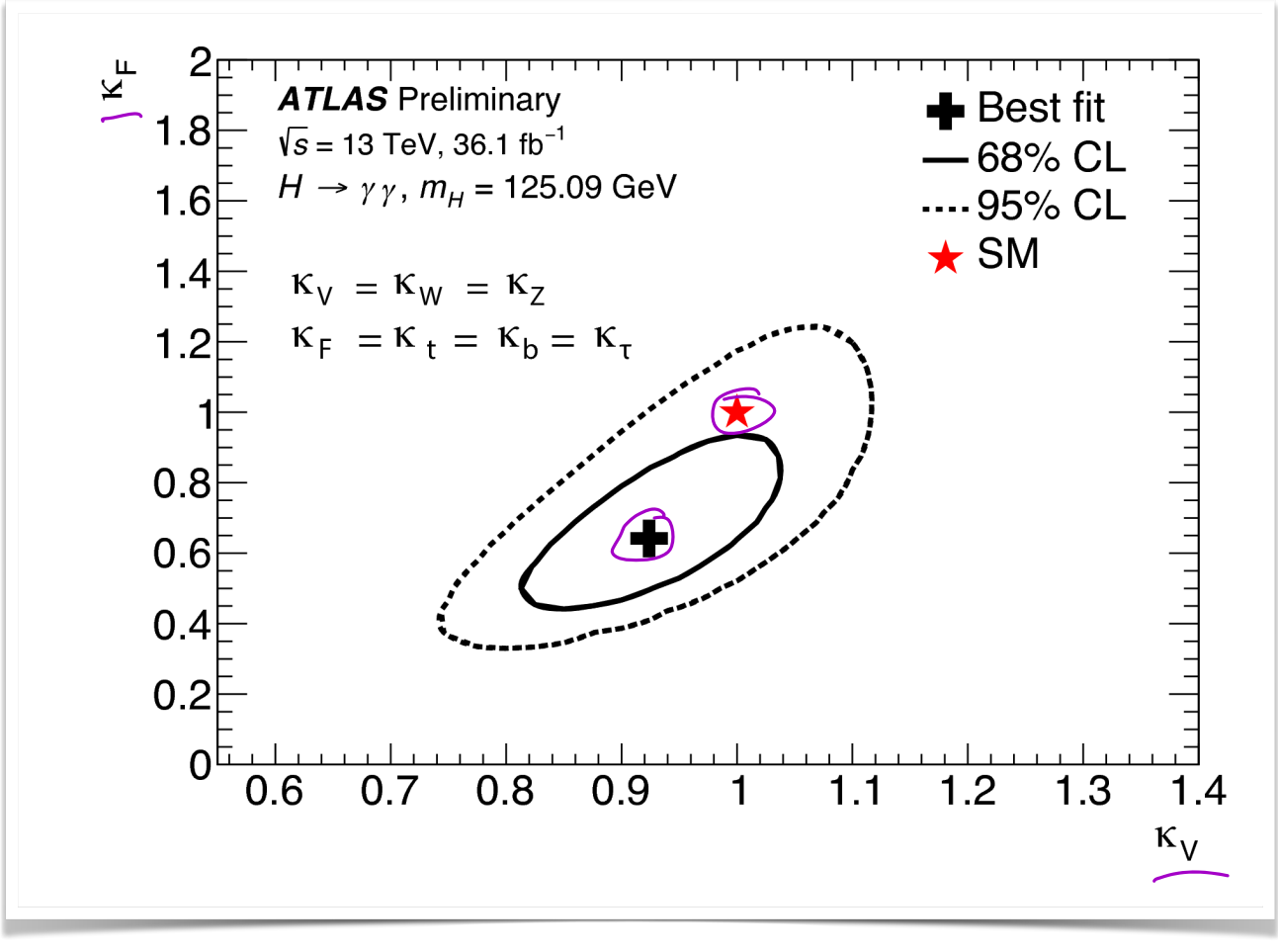
- If x and y are independent variables
  - $G(x, y \,|\, \mu_x, \sigma_x, \mu_y, \sigma_y) = \frac{1}{\sqrt{2\pi}\,\sigma_x} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \frac{1}{\sqrt{2\pi}\,\sigma_y} e^{-\frac{(y-\mu_y)^2}{2\sigma_y}}$
  - $\frac{\partial}{\partial \mu_x} \ln \mathcal{L} = \frac{1}{\sigma_x^2}$
- Now, assume that x and y are correlated
- Covariance matrix is defined by
  - $\langle \hat{V}^{-1} \rangle_{ij} = -\frac{\partial^2 \ln \mathcal{L}}{\partial \mu_i \partial \mu_j}$
- For a binned likelihood, where N is large and the likelihood can be reduced to a χ^2
  - $\langle \hat{V}^{-1} \rangle_{ij} = -\frac{\partial \chi^2}{\partial \mu_i \partial \mu_j}$

Correlated Uncertainties

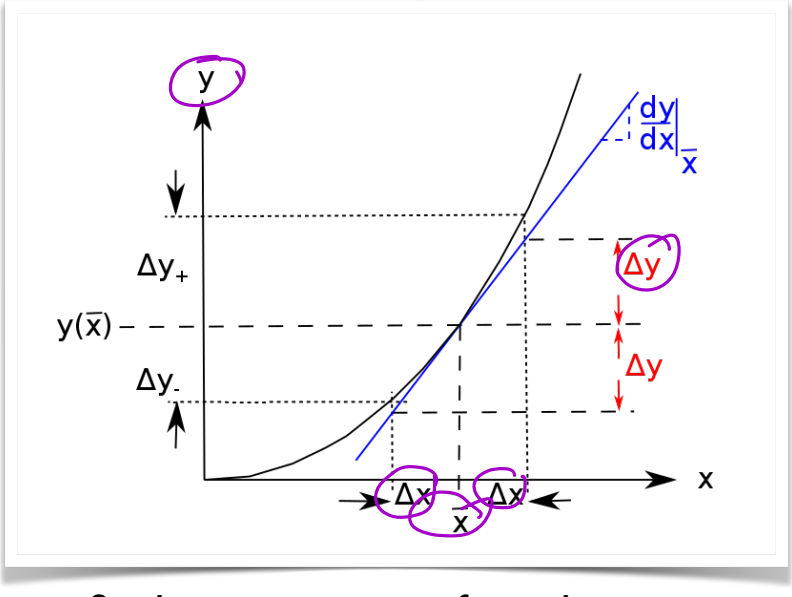
- Standard error ellipse for two parameters with a negative correlation
- Slope related to correlation coefficient $\frac{\partial \theta_i}{\partial \theta_j}$
- Correlation matrix typically determined from data numerically during fitting procedure



Correlation Example from Higgs



Propagation of Errors



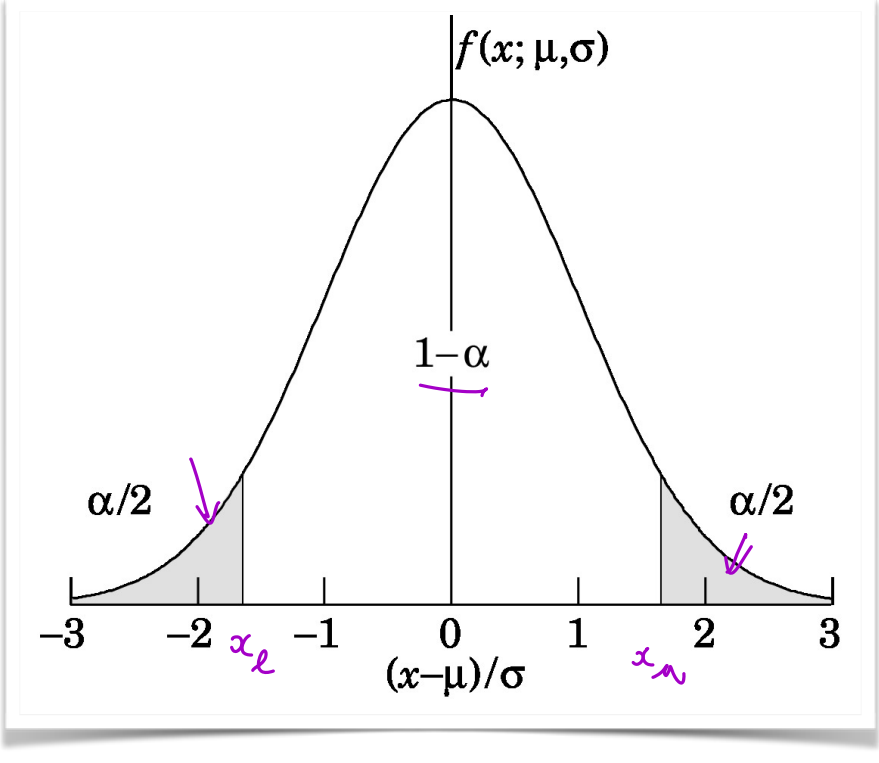
- Determine error on final measurement from known errors on input measurements
- $\sigma_f^2 = \left(\frac{\partial f}{\partial \alpha}\right)^2 + \left(\frac{\partial f}{\partial \beta}\right)^2 + 2\frac{\partial f}{\partial \alpha}\frac{\partial f}{\partial \beta} \, cov\,\alpha\beta$
- More dimensions are usually expressed as a matrix
- Useful reference: https://en.wikipedia.org/wiki/Propagation_of_uncertainty

Confidence Interval

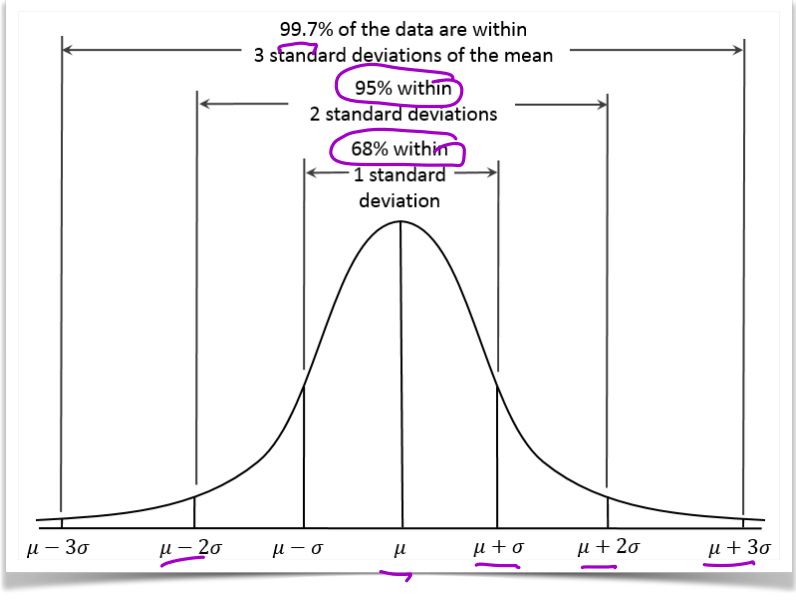
$\alpha = 5\%$. 95%.

$\alpha = 10\%$. 90%.

- Fraction of the result not between x_ℓ and x_u is α
  - $1 - \alpha = \int_{x_\ell}^{x_u} f(x;\theta)\, dx$ ← PDF



Confidence Levels for a Gaussian



p-value

α	δ
0.3173	1σ
4.55×10^{-2}	2σ
2.7×10^{-3}	3σ
5.7×10^{-7}	4σ
2.0×10^{-9}	5σ

Confidence Levels: Higgs