Statistics and Probability, Interpreting Measurements

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

Announcements: Course moving Online

- From next week onwards, **classes will be online only**
  - I will broadcast the class on zoom at the usual time
  - Please connect because this will give you the possibility to ask questions during the lecture and it is a good technique for accountability
  - I will record and post those lectures on the class page
- We will not be taking attendance at workshops until further notice, however, workshops will still be due on Mondays before lectures as before
- Like the lectures, the workshops will be virtual and held over zoom
  - Please set up a group chat in slack with the people that you generally work together for the workshop
    - You can use that chat to communicate with each other while connected on zoom to the class
  - It will continue to be part of your grade to submit workshops

Announcements: Course moving Online

- Things to think about
  - In person lectures and workshops are an important accountability mechanism - for your learning please think of alternate methods of accountability
    - Ideas could include:
      - holding yourself accountable to attend the lecture online at the usual time
      - forming a virtual online study group with members of the class
  - Asking questions in a video lecture is more challenging
    - I will continue to call on people to answer questions
    - I will try to pause for questions often, but please post a message on zoom if you'd like to ask something and I'll let you ask
    - Also, post any questions (either during or afterwards) on slack

Announcements: Course moving Online

- I will continue to call on people for questions in class, however attendance will no longer form part of your class participation grade
- Some things that count towards participation
  - Asking or answering questions in lectures, on zoom, on slack
  - Attending office hours (online or in person), submitting workshops
- Zoom connections will be available to all office hours -- i.e. for Nick's in addition to mine. If you feel unwell, but still want to attend office hours, please connect via zoom. Here is the zoom link: https://berkeley.zoom.us/j/170666271
  - Please can you try to connect on zoom this week and report on slack any problems that you experience
- Please continue to use slack as usual for online communication.
  - In case you aren't able to attend a lecture or a workshop, please post any questions that you might have there.

Announcements: Final Project

- The final project will make up 40% of your final grade
- The projects will be performed in assigned groups of 3 or 4 students
  - Working in teams is an important skill in being a scientist
  - I will announce the teams next week and provide some suggested topics
- Next week, your homework assignment (HW7) will be the **project proposal**
  - Each member of the group will submit their own proposal, but you will need to agree on the topic with the other members of the group
- The main deliverable of your project will be a **piece of code** and associated data (if any)
- During the workshop sessions of the last week of class and on Monday of RRR week, each group will provide a 10 minute **presentation** and demonstration of their code (via zoom if needed)
- Each student will be required to prepare a **short report** (3-4 pages) that will be due during the examination period

Announcements: Final Project

- The breakdown for your grade of the final project will be as follows
  - Project Proposal: 10%
  - Project Implementation (code): 30%
  - Project Presentation: 30%
  - Project Report: 30%
- Note that the grade for the implementation and the presentation will be common between members of the group, but your proposals and reports will be graded independently

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 $\mathscr{L}(x;\theta)$:

    -

- 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
  - $\dfrac{\partial \mathscr{L}}{\partial \theta} =$

- 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

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

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

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

Binned vs unbinned likelihood functions

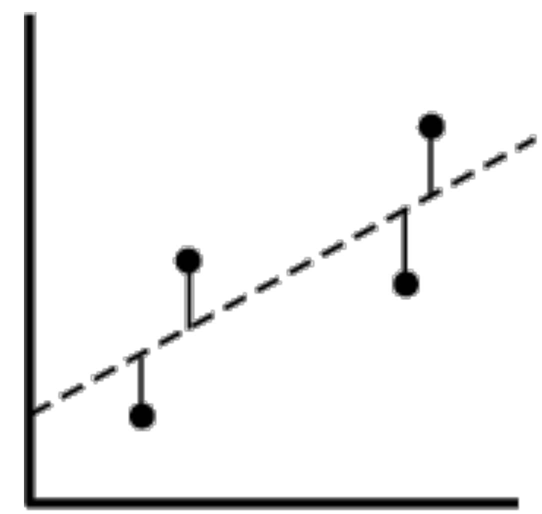
- Likelihood formalism works for any
- Product of the is a
- **Example measurement**: Measure the of a particle of a given species for an ensemble of such particle produced at $t = 0$ such that the decay at time t:
  - $f(t) = \frac{1}{\tau} e^{-t/\tau}$
- Consider two ways to construct the likelihood
  -
  -
    -

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 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:
-
- Scatter defined by χ^2
  - $\chi 2 =$
- Can write the χ^2 in terms of our observables
  - $\chi 2 =$

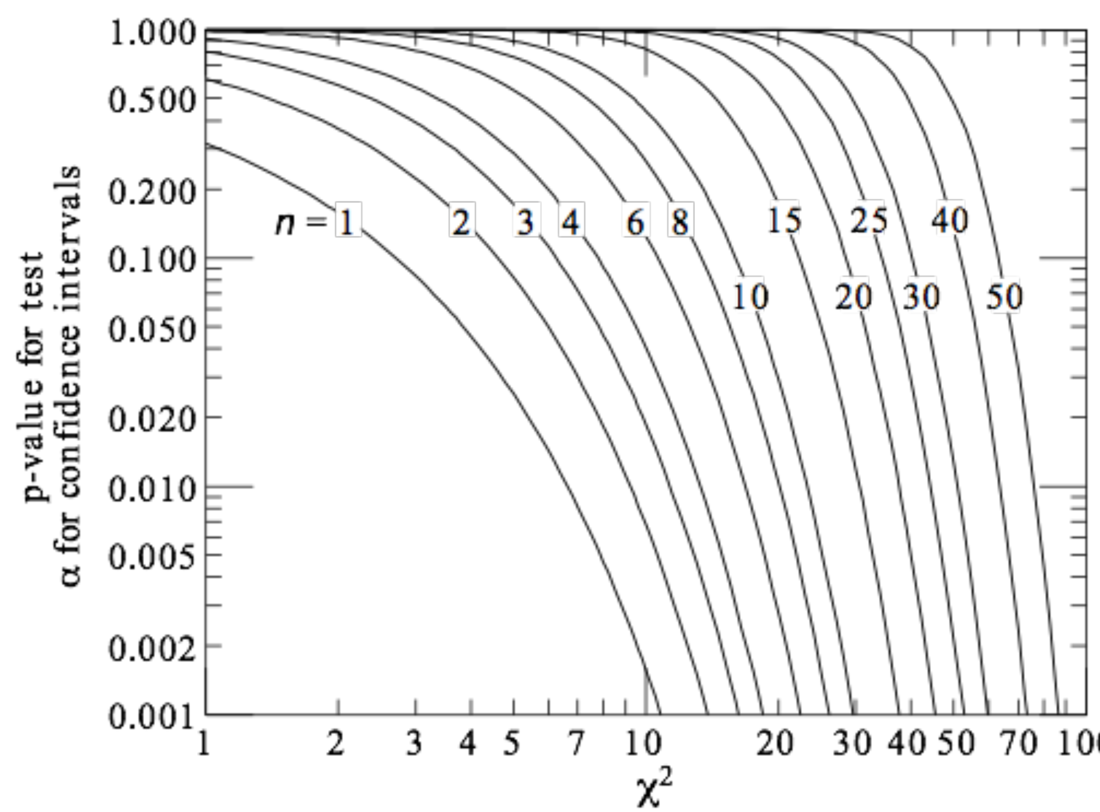


- Minimise χ^2 with respect to θ
- Useful when minimising $\ln \mathscr{L}$ is slow (high statistics samples)

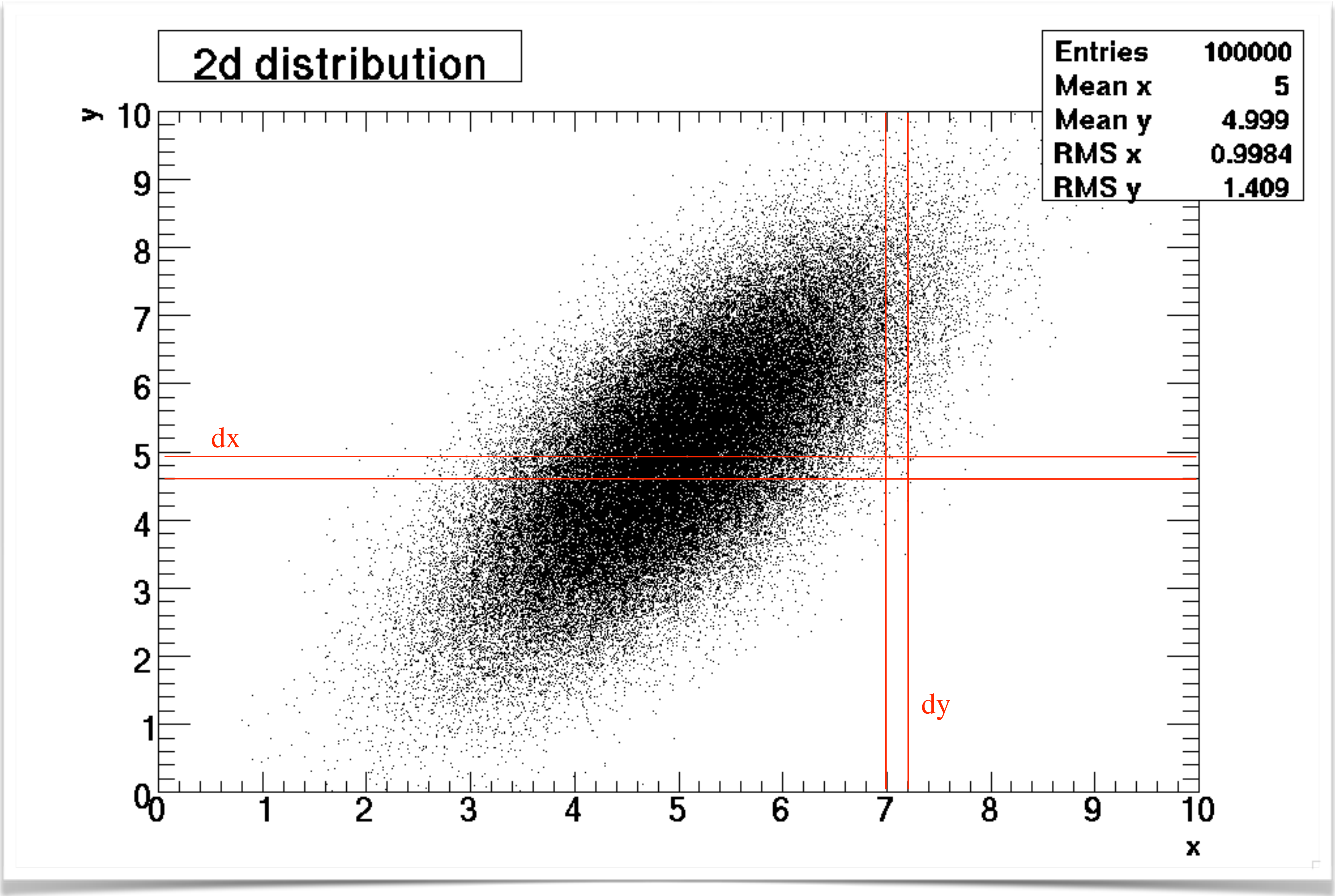
Example: chi-squared p-values

- 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

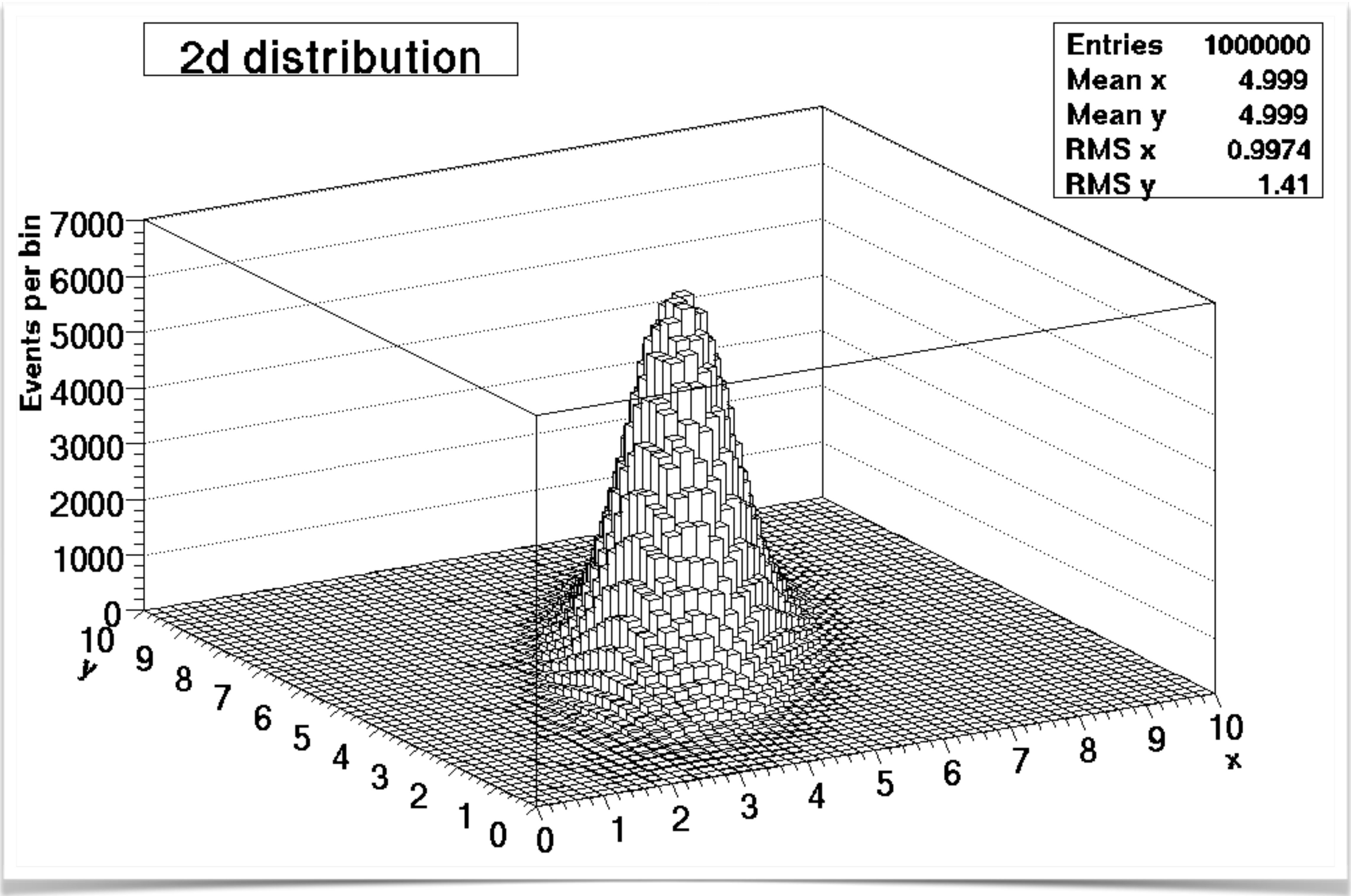
    to be close to



2D distribution



2d distribution

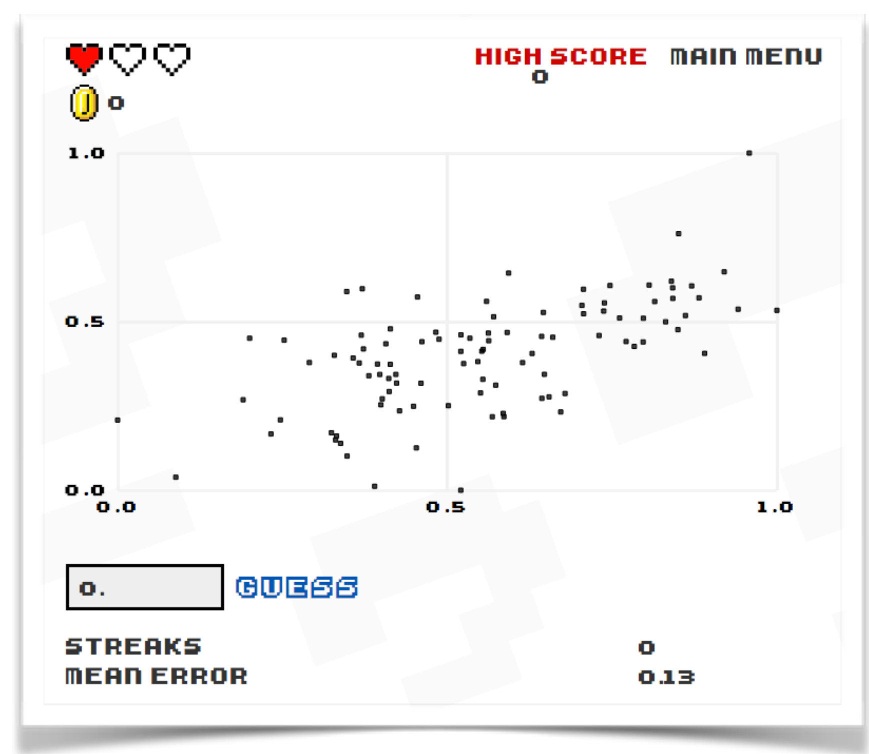


Covariance and Correlation

Covariance Matrix

$$cov[x, y] =$$

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



http://guessthecorrelation.com/

Correlation (normalised covariance)

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

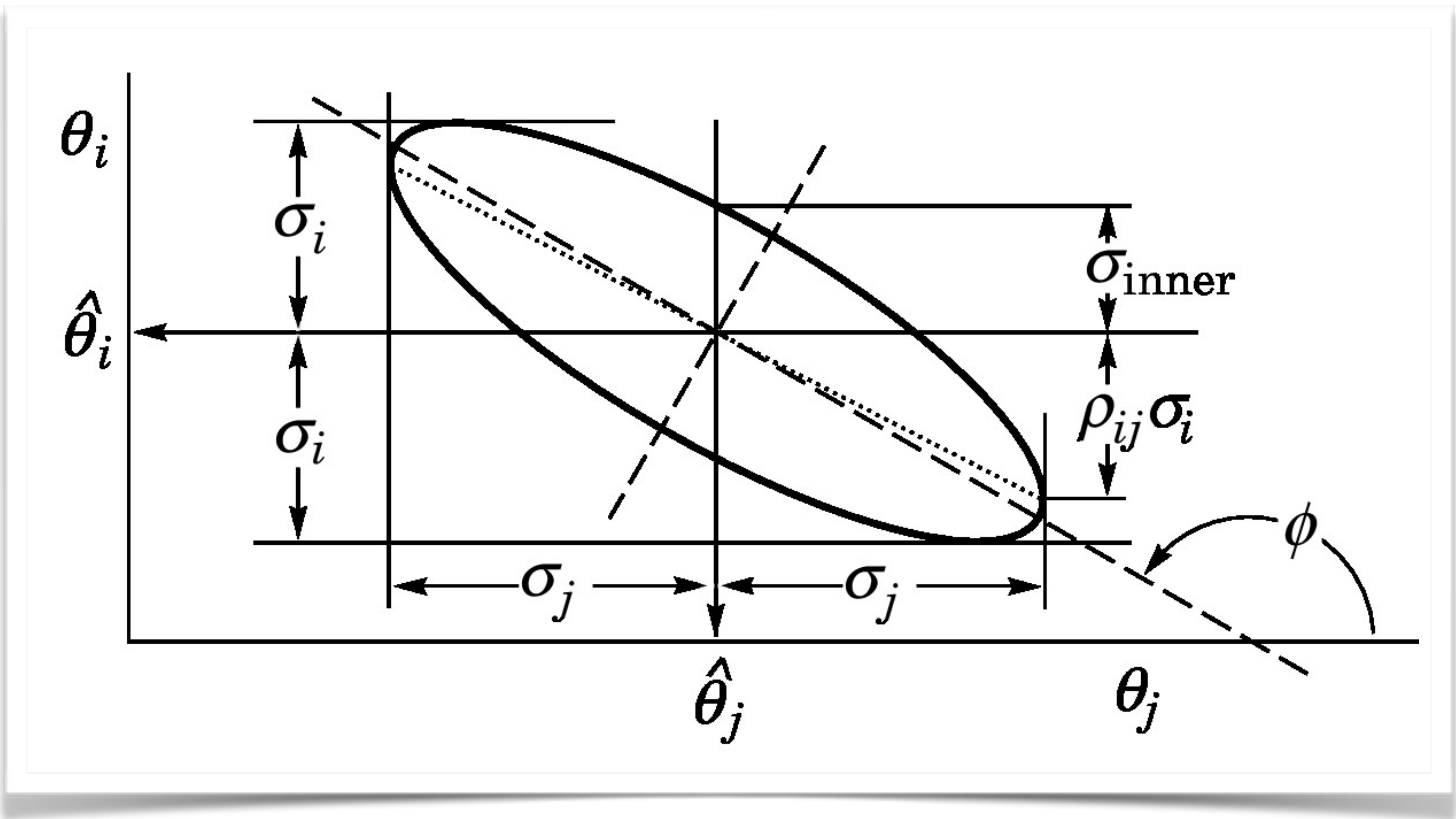
$$\rho_{xy} =$$

Covariance Matrix for a Gaussian

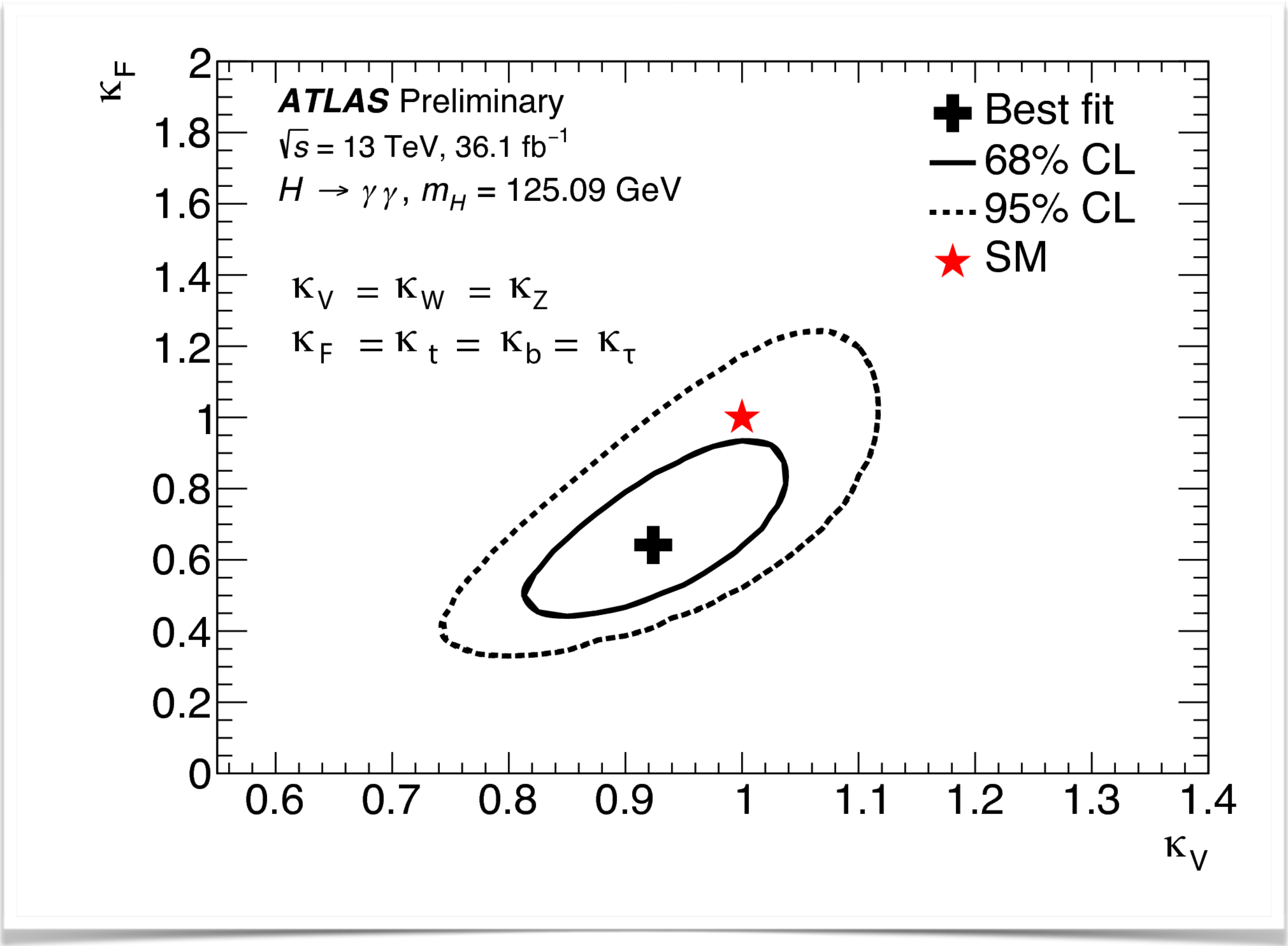
- If x and y are independent variables
  - $G(x, y \,|\, \mu_x, \sigma_x, \mu_y, \sigma_y) =$
  -
- Now, assume that x and $y$are correlated
- Covariance matrix is defined by
  - $\langle \hat{V}^{-1} \rangle_{ij} =$
- For a binned likelihood, where N is large and the likelihood can be reduced to a χ^2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

Correlated Uncertainties

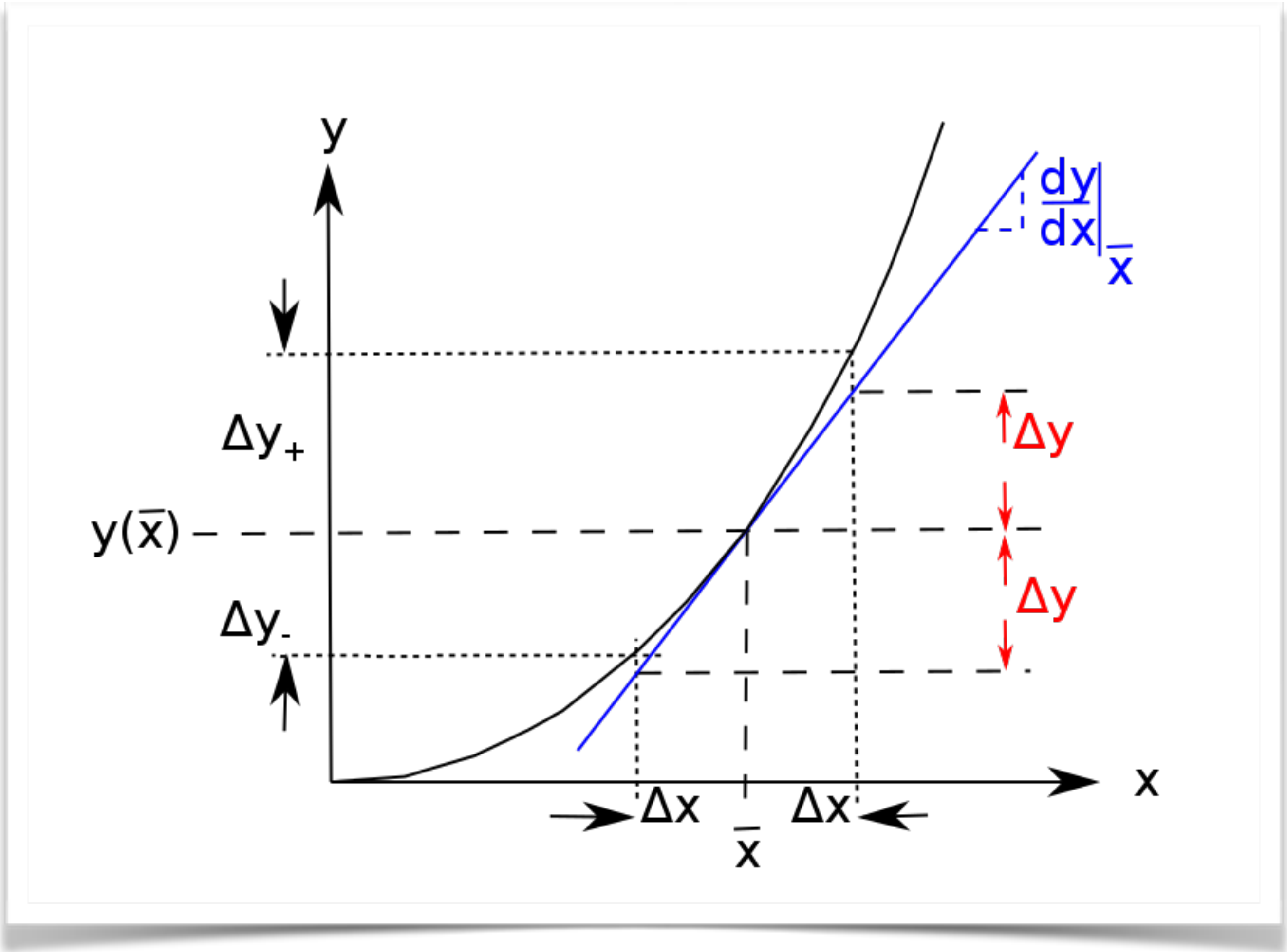
- Standard for two parameters with a correlation
- Slope related to correlation coefficient
- Correlation matrix typically determined from data numerically during fitting procedure



Correlation Example from Higgs



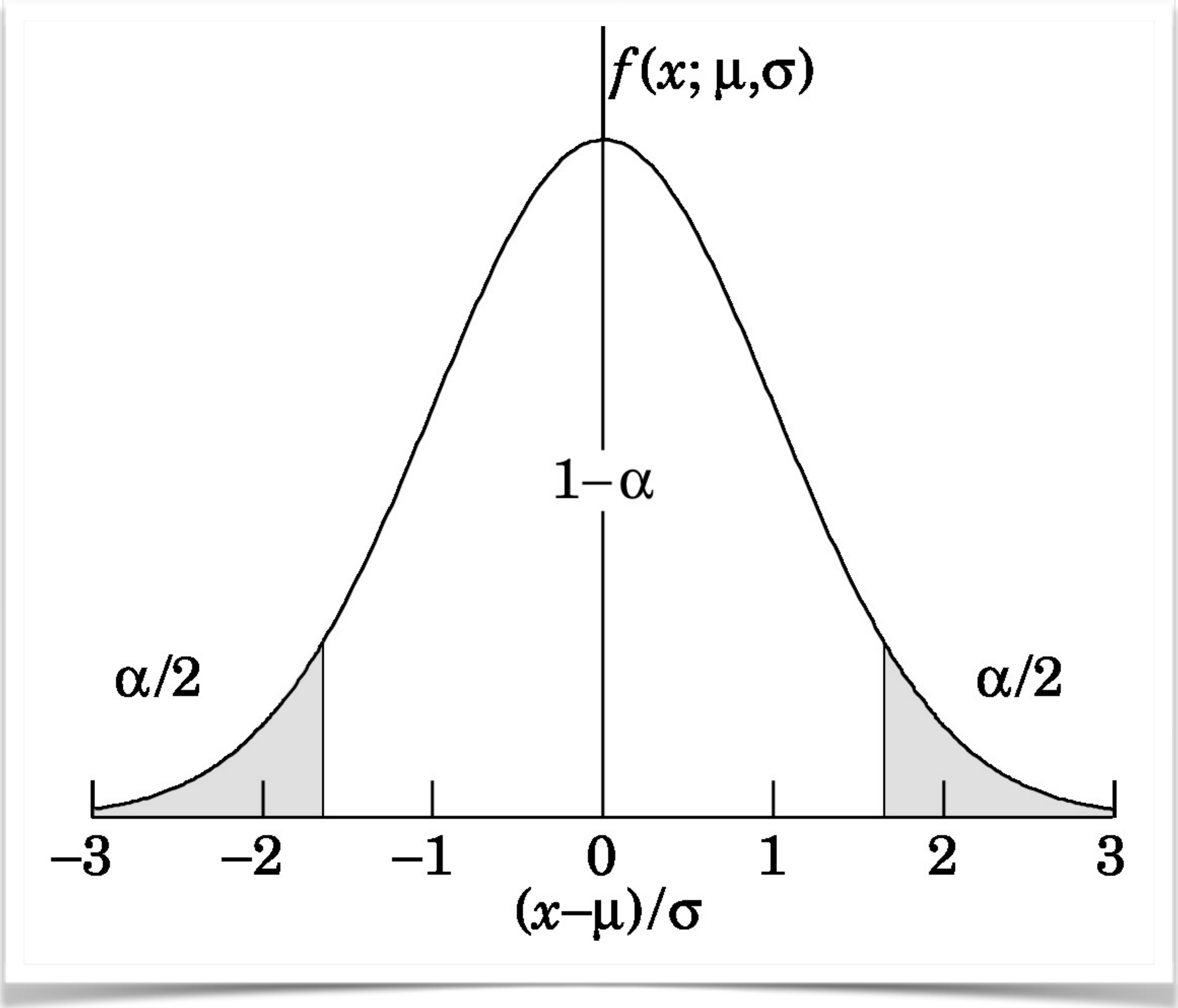
Propagation of Errors



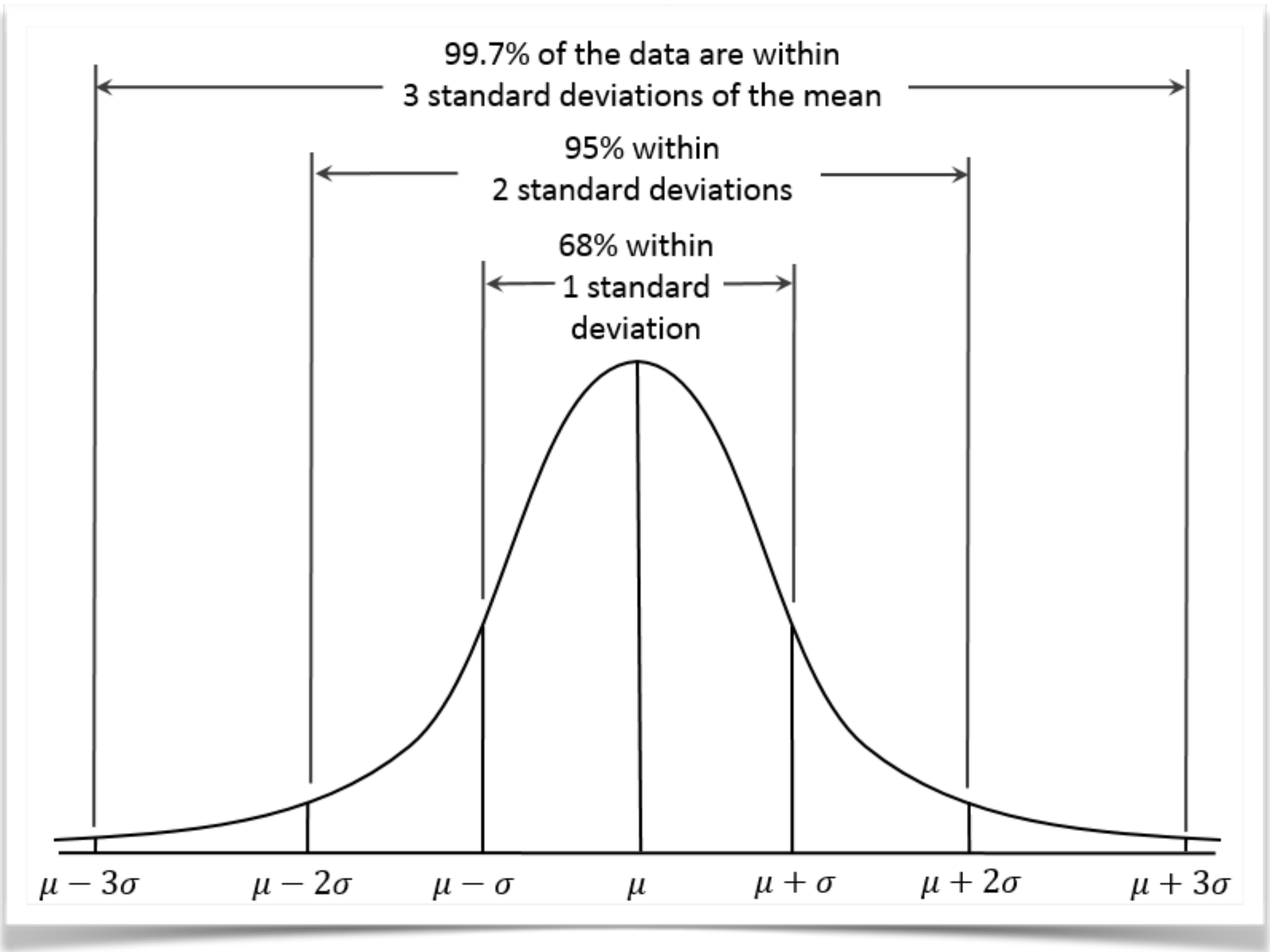
- Determine error on final measurement from known errors on input measurements
  - $\sigma_f^2 =$
- More dimensions are usually expressed as a matrix
- Useful reference: https://en.wikipedia.org/wiki/Propagation_of_uncertainty

Confidence Interval

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

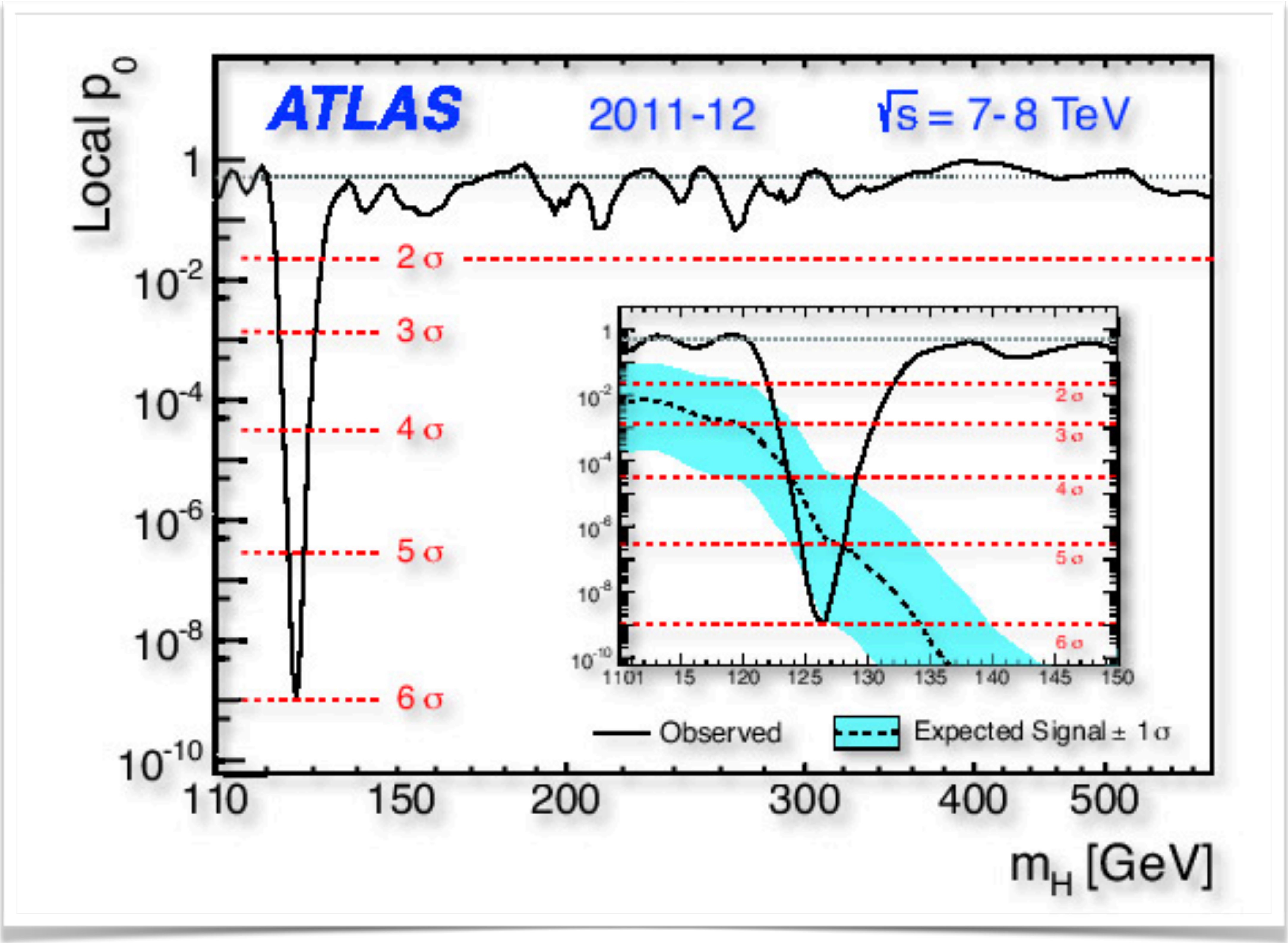


Confidence Levels for a Gaussian



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

Confidence Levels: Higgs



Example in jupyter notebook