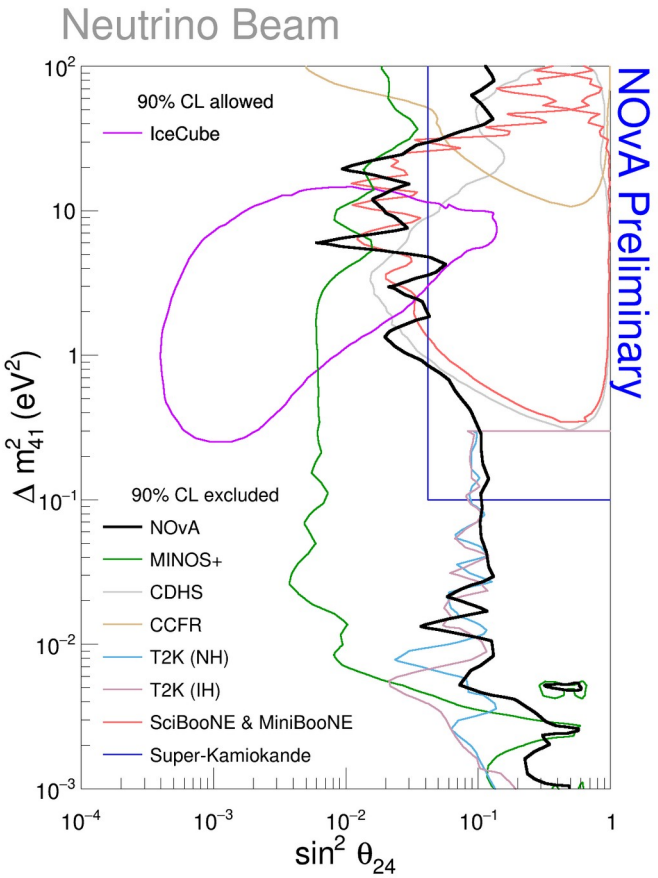
UNIVERSITY OF Cincinnati

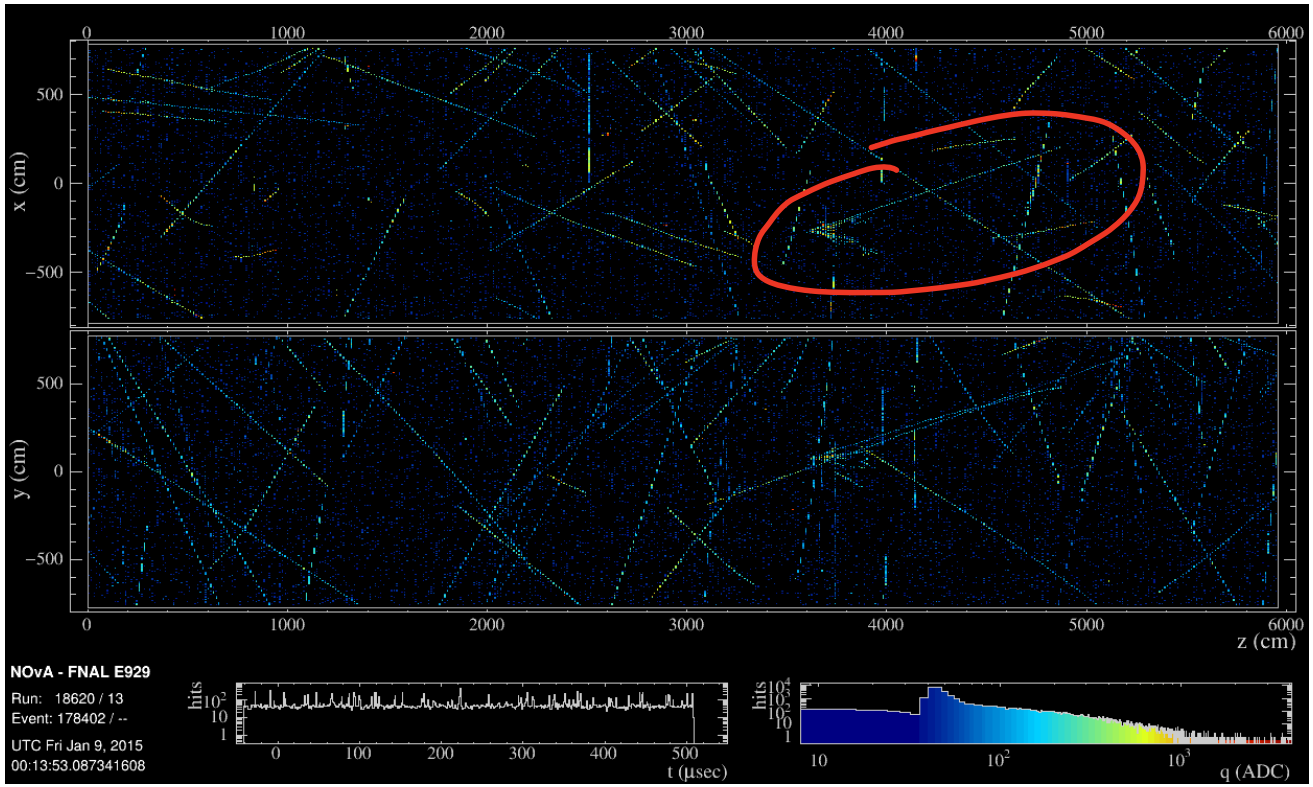
Making Sense of Your Data: Statistics and Machine Learning

Adam Aurisano
University of Cincinnati

Understanding the Universe Through Neutrinos
29 April 2024



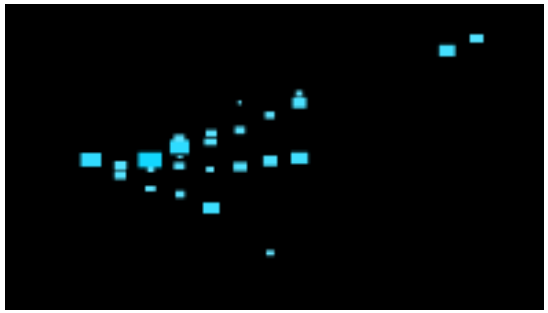
Introduction



- At a high level, we are interested in understanding the role neutrinos play in the universe
- High level quantities, like oscillation parameters, can help constrain physics in the early universe
- Before we can do that, we need to be able to make sense of experimental data

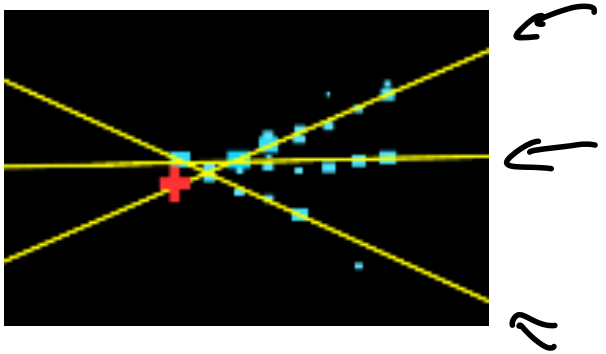
Reconstruction

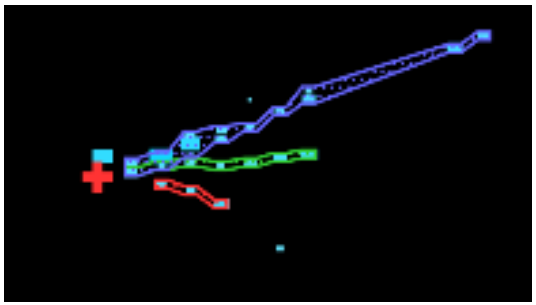
In general, detectors produce hits in time and space, with the details depending on the detector technology in question. These hits need to be interpreted in the context of neutrino interactions before we can do anything else.



Event Separation: Coarse event-level time-space clustering (slicing) using the DBSCAN algorithm.

Vertexing: Find lines of energy deposition using the Hough transform. Find the best point line radiate from.

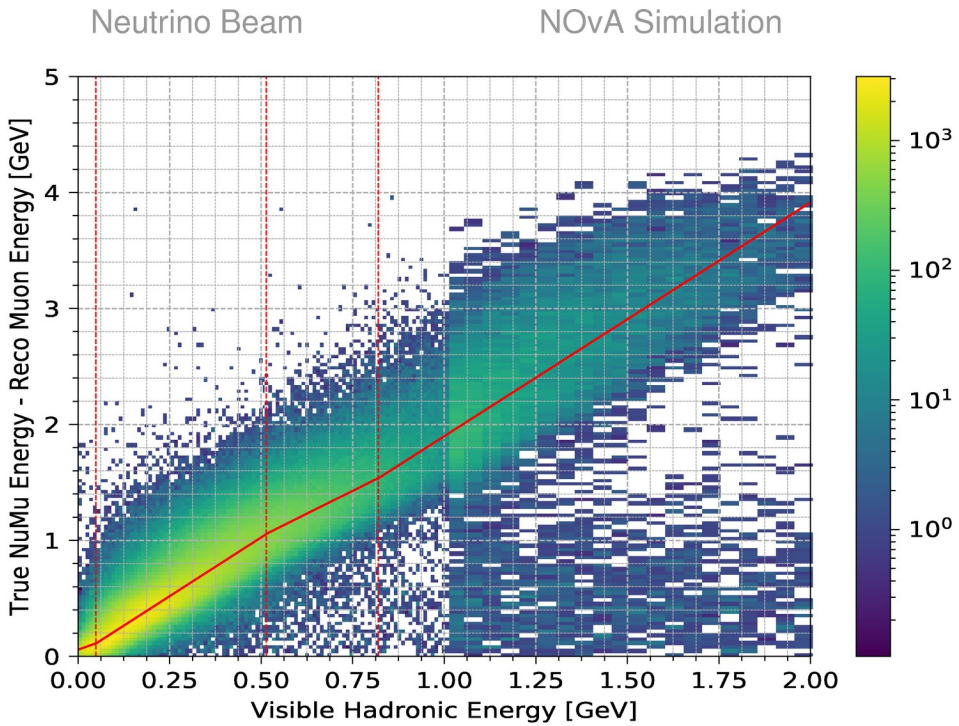


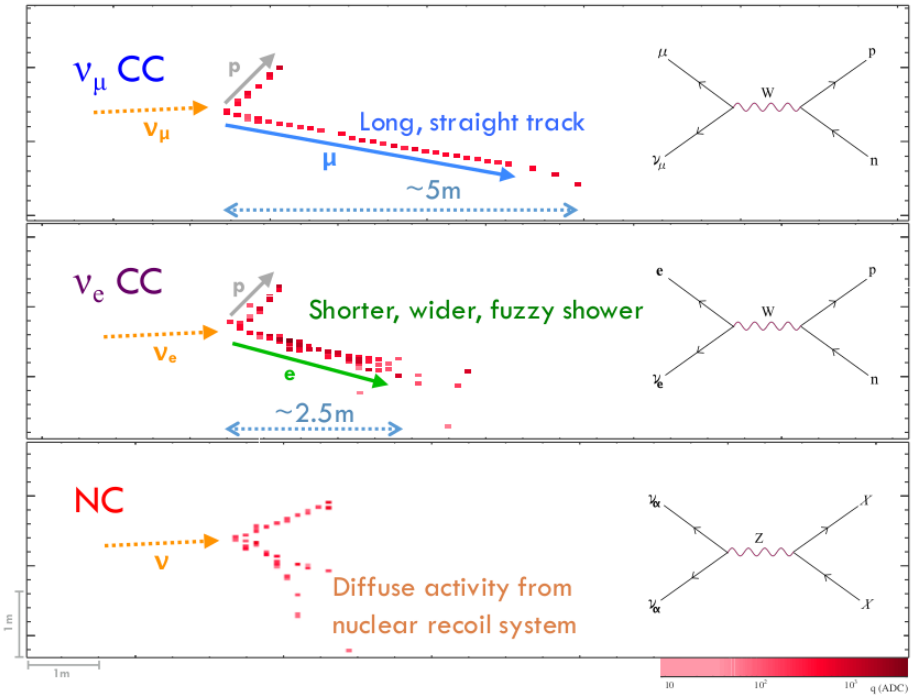


Prong Clustering: Find clusters in angular space around the vertex in each view. Merge views using topology and prong dE/dx.

Event Selection and Energy Estimation

Event Selection:
Once events have been reconstructed, selection algorithms determine the flavor of the neutrino that produced it

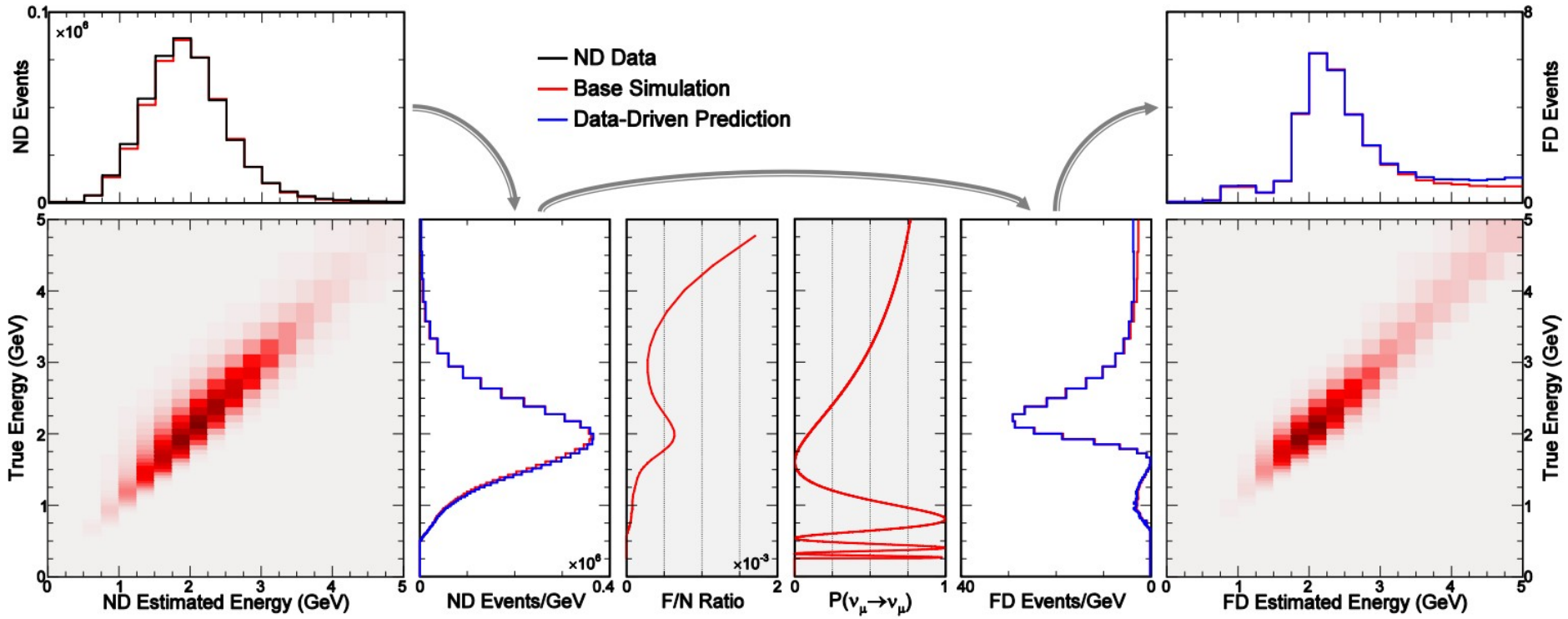




Energy Estimation:
Since neutrinos oscillate according to true energy, we use observable quantities to estimate the true energy as well as possible

Parameter Extraction

- The neutrino spectrum is measured before oscillations at the ND
  - Combination of flux, cross section, and efficiency
- The measured spectrum is used to correct the raw FD MC predictions using the Far/Near ratio
  - Each component oscillates differently, so they each are extrapolated separately
  - ND data/MC disagreements are allocated to different component either proportionally to MC predictions or using data-driven decomposition techniques
- Since the detectors are functionally similar, the combined flux and cross-section uncertainties largely cancel
- A fit is performed to find a set of parameters which produce a prediction which best matches the observed data



Probability and Statistics

- Every step of the analysis process is uncertain/stochastic
- Large uncertainties in our understanding of our neutrino beams and of neutrino cross sections
- Due to detector and electronic noise, measured energies are known imprecisely
- Particles interact with the detector stochastically
  - Even if the detector measured energy deposits perfectly, they would still vary from particle to particle
- To make any measurements requires probability and statistics
  - Probability: the study of how likely a given event is to occur
    - Model $\longrightarrow$ Predict outcome
  - Statistics: the study of how to interpret data
    - Data $\longrightarrow$ infer parameters of model
- Statistics is essentially the inverse problem of probability, and inverse problems are always difficult
- We'll start by looking at the fundamentals of probability

Fundamentals of Probability

Kolomogorov's Axioms

- Probability, colloquially, is how often you can expect a given event to happen
- Can make this mathematically rigorous using set theory
- Consider a set S, which represents the sample space, and a subset A
- For each subset A, we assign a real number P(A), which is the probability of the set
- P(A) follows our intuitive understanding of probability if:

$$P(A) \geq 0 \, \forall A \in S$$

$$\text{If } A \cap B = \emptyset, \text{then } P(A \cup B) = P(A) + P(B)$$

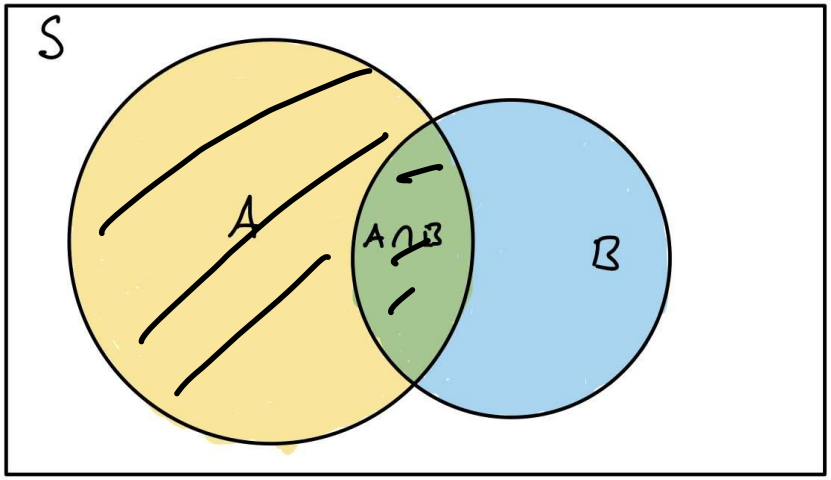
$$P(S) = 1$$

Sample Space

- What S is is somewhat ambiguous and varies based on interpretation
  - Will come back to this shortly
- Can be
  - discrete
  - continuous
- A variable that takes a specific value for each element of S is called a random variable
- Random variables can be multicomponent vectors if each element is associated with several quantities

Joint Probability

- Suppose S is a vector labeled space
  - Possible dice rolls of a pair of dice
  - A: die 1 yields odd number, die 2 yields any number
  - B: die 1 yields any number, die 2 yields even number
  - Probability of die 1 being odd and die 2 being even is the intersection of A and B

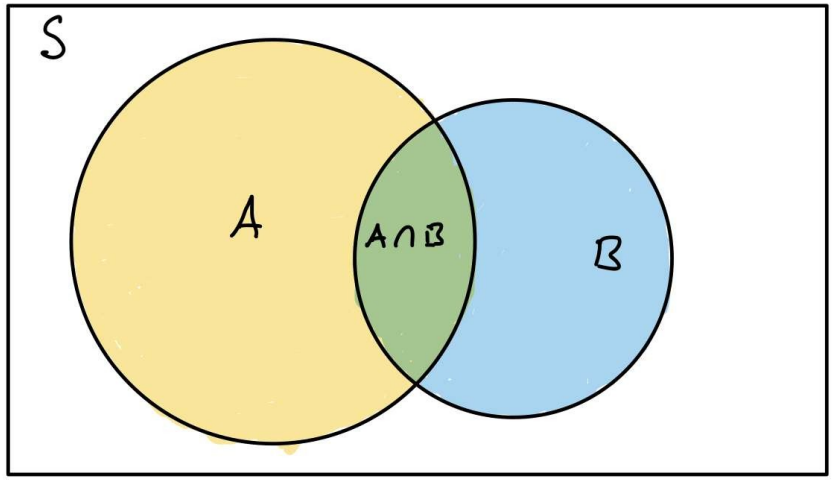


The probability of the intersection is the joint probability.

We usually write it as P(A,B)

Marginal Probabilities

- Some times we only care about one of the components of the vector
- If we compute the probability of one component without reference to the other component, the result is the marginal probability
- If we only care about the value of die 2, we can add up the probabilities of all possible values of die 1

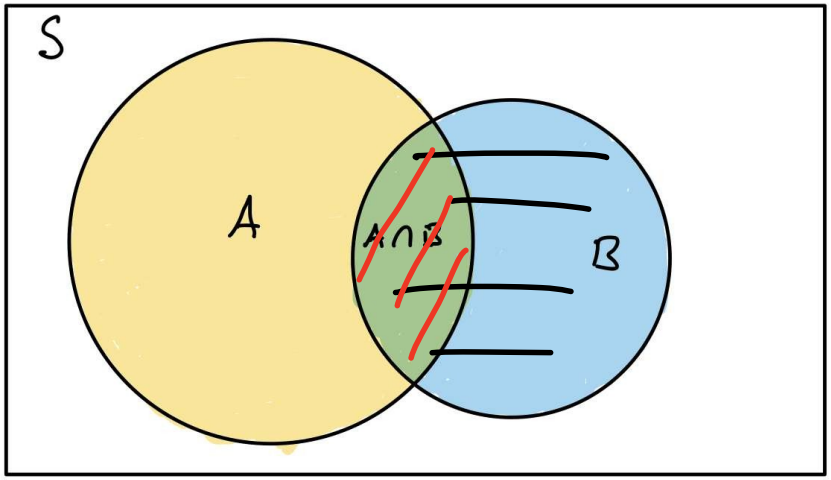


$$P(B) = \sum_i P(A_i, B)$$

Marginalizing is "integrating out" the component we aren't interested in

Conditional Probability

- The joint probability, P(A,B), is the size of $A \cap B$ within S
- The conditional probability is the probability of $A \cap B$ given that B is true
- That means that we need to divide the joint probability by the probability of any outcome in the set B



$$P(A|B) = \frac{P(A,B)}{P(B)}$$

or

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

Bayes Theorem

- We can put these two expressions together to get Bayes's Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- This tells us how we can change the order of the conditioning
- This is important because we often know one conditional probability, but we really want to know the other

Example: Triple Screen Test

- The importance of this often comes up in medical cases
- Let's consider the triple screen test, a blood test to help determine if a fetus has Down syndrome
  - 70% sensitivity (probability of positive result given Down syndrome is present)
  - 5% false positive rate (probability of positive result given Down syndrome is not present)
  - 0.1% prevalent rate of Down syndrome in general populace
- Suppose a triple screen test comes back positive. What is the probability that the fetus has Down syndrome?

Example: Triple Screen Test

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(A|B) = P(Down|+)$

Posterior probability that Down syndrome is present

$P(B|A) = P(+|Down)$

Sensitivity of test

$P(A) = P(Down)$

Prior probability that Down syndrome is present

$P(B) = P(+)$

Probability of getting a positive result, regardless of the presence of Down syndrome

Example: Triple Screen Test

Marginal probability:

$$P(+) = P(Down, +) + P(!Down, +)$$

Or by the rules of conditional probability:

$$P(+) = P(+|Down)P(Down) + P(+|!Down)P(!Down)$$

Therefore

$$P(Down|+) = \frac{0.70 * 0.001}{0.70 * 0.001 + 0.05 * 0.999} = 1.4\%$$

After a positive test, risk is 14x higher, but still a very low probability

Example: Triple Screen Test

$$P(Down|+) = \frac{0.70 * 0.001}{0.70 * 0.001 + 0.05 * 0.999} = 1.4\%$$

Suppose a second test existed with similar sensitivity and false positive rates. What would be the probability of Down syndrome being present if both tests are positive?

Example: Triple Screen Test

$$P(Down|+) = \frac{0.70 * 0.001}{0.70 * 0.001 + 0.05 * 0.999} = 1.4\%$$

Suppose a second test existed with similar sensitivity and false positive rates. What would be the probability of Down syndrome being present if both tests are positive?

New prior becomes posterior of the first test:

$$P(Down|+) = \frac{0.70 * 0.014}{0.70 * 0.014 + 0.05 * 0.986} = 16.6\%$$

Bayes theorem allows us to update our knowledge as new information becomes available

Interpretations of Probability

- Rules of probability tell how probability is manipulated, but what it is
- Two main interpretations
  - Frequentist
  - Bayesian

Frequentist Interpretation

- Probability is the relative frequency of an event's occurrence
- Set S corresponds to all possible outcomes of a measurement
- The probability of a particular event to occur is:

$$P(A) = \lim_{n \to \infty} \frac{\text{instances of outcome A in } n \text{ measurements}}{n}$$

- Only outcome of measurements have probabilities
  - Can only talk about consistency of models with observed data, not probability of a model being true

Bayesian Interpretation

- Probability is subjective
- S consists of hypotheses
  - Constructed such that only one hypothesis can be true
- P(A) = degree of belief that hypothesis A is true
- Bayes theorem provides machinery to update our degree of belief in any particular hypothesis

$$P(\text{theory}|\text{data}) \propto P(\text{data}|\text{theory})P(\text{theory})$$

- P(theory) is our prior belief in the theory
- P(data|theory) is the likelihood of observing particular data if a given model were true

Why Does it Matter?

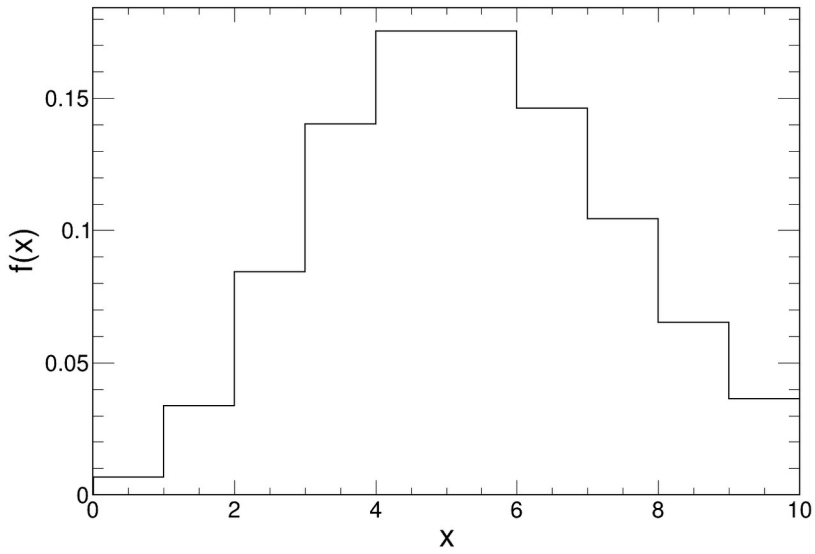
- Frequentist
  - definition of probability is objective
  - measurement of probability only possible in the limit of infinite statistics
  - can not answer the question everyone actually wants to know
- Bayesian
  - definition of probability is subjective
  - answers can be highly dependent on choice of prior
  - answers the question everyone actually wants to know
- Everyone is a little of both, and some questions are more naturally approached under one framework than another
- Generally speaking, in the limit of infinite data, both interpretations give the same answers (i.e. the explanatory power of the data overpowers reasonable prior belief)

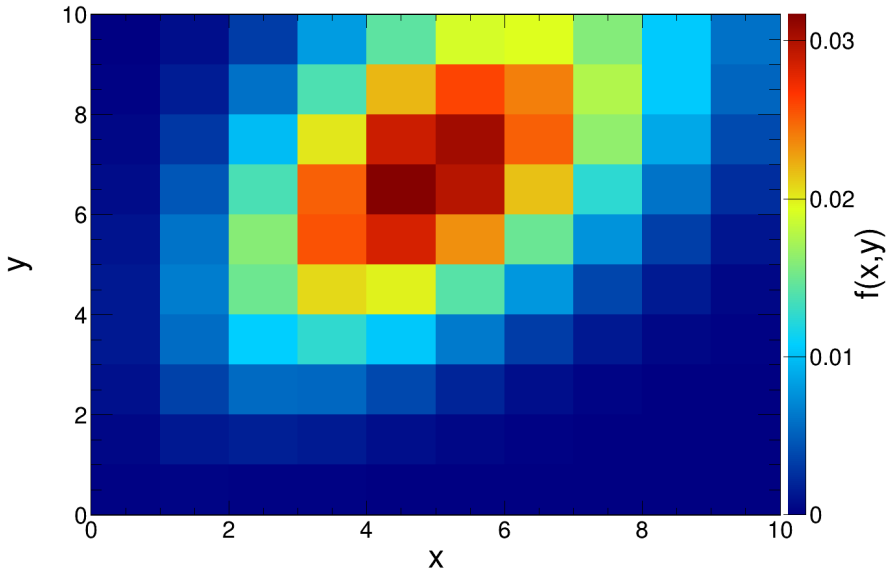
Probability Density Functions

- Discrete distribution:
  - Probability of experiment yielding value x: P(x)
- Continuous distribution:
  - Probability of x having a value in an infinitesimal interval (x, x+dx): f(x)dx
    - Note: probability of any particular outcome is infinitesimal in continuous case, but value of f(x) gives the relative frequency of occurrence
- Normalization $\sum_{i}^{N} P(x_i) = 1 \qquad \int_S f(x)\,dx = 1$
- In general, continuous distributions require integrals instead of sums

Histograms

- Histograms, when normalized, approximate a pdf
- Each bin contains a approximation of f(x) integrated over the bin width
- Multidimensional histograms are an approximation of joint pdfs





Moments

Can characterize distributions using moments

Mean: a measure of the central point in the distribution

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx$$

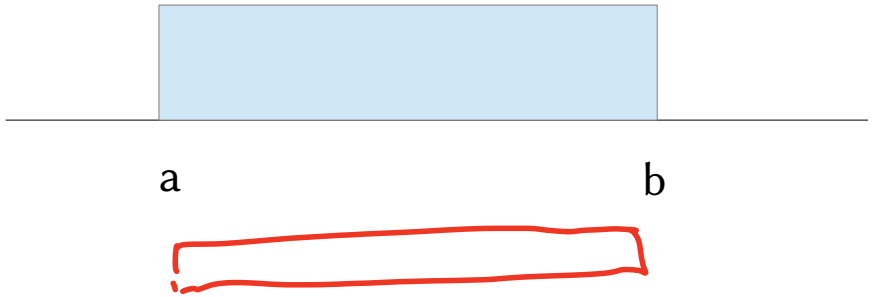
Variance: a measure of the width of the distribution

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx$$

Square root of variance is the standard deviation – we usually take this as the standard error

Example:

A digital pulse has a constant value between times a and b, and it is zero otherwise.



What is the standard error on the time of the pulse?

$$f(x) = \begin{cases} c & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \qquad c = \frac{1}{b-a}$$

$$\mu = \frac{1}{b-a}\int_a^b x\,dx = \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}$$

$$\sigma^2 = \frac{1}{b-a}\int_a^b (x-\mu)^2\,dx$$

$$\Rightarrow \sigma = \frac{|b-a|}{\sqrt{12}}$$

Covariance

Can generalize to multivariate distributions

$$\vec{\mu} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \vec{x} f(\vec{x}) \, dx_1 \, dx_2$$

Variance generalized to covariance matrix: diagonals are 1D variances and off-diagonals are related to correlation

$$\Sigma_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) f(x, y) \, dx \, dy$$

Or more compactly

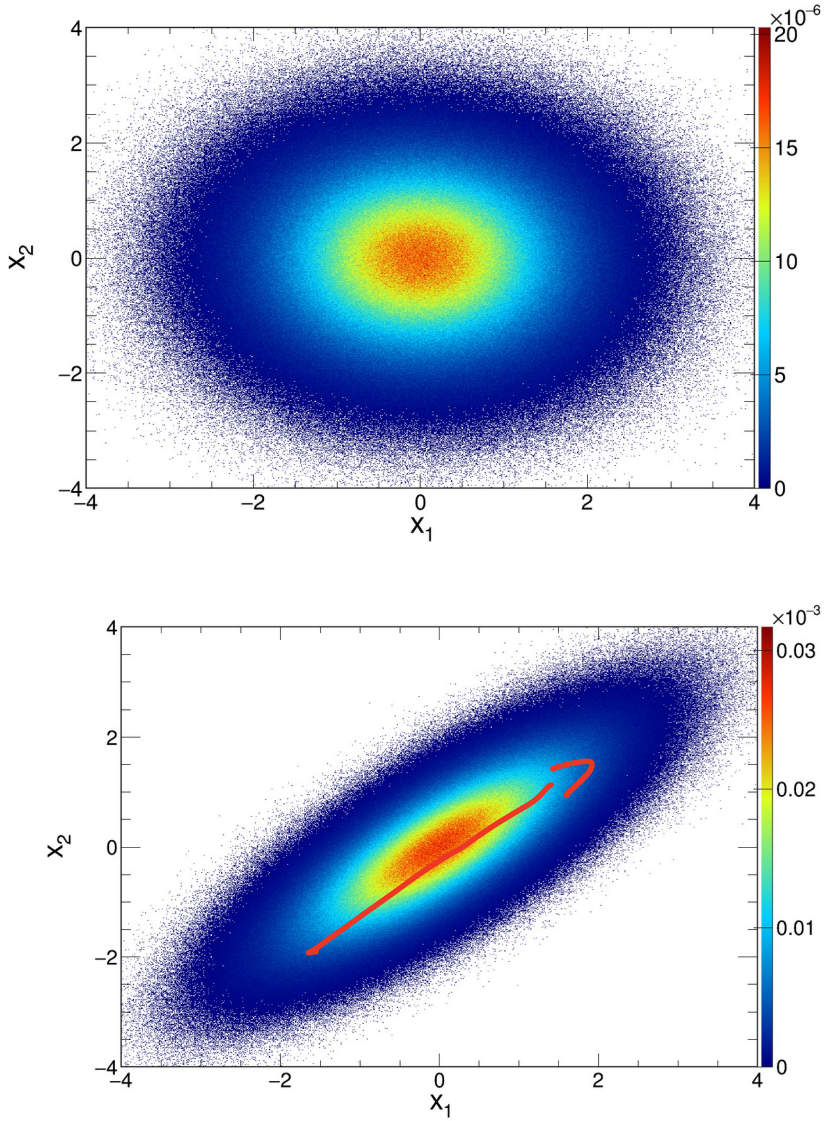
$$\Sigma = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\vec{x} - \vec{\mu})(\vec{x} - \vec{\mu})^T f(x_1, x_2) \, dx_1 \, dx_2$$

Correlation

- The correlation coefficient quantifies the linear relationship between two random variables
- Correlation is related to covariance simply: $\rho_{xy} = \sigma_{xy}/\sigma_x\sigma_y$
- In the first, x_1 and x_2 are independent Gaussian ⟶ correlation = 0
- In the second x_1 and x_2 are also Gaussian, but

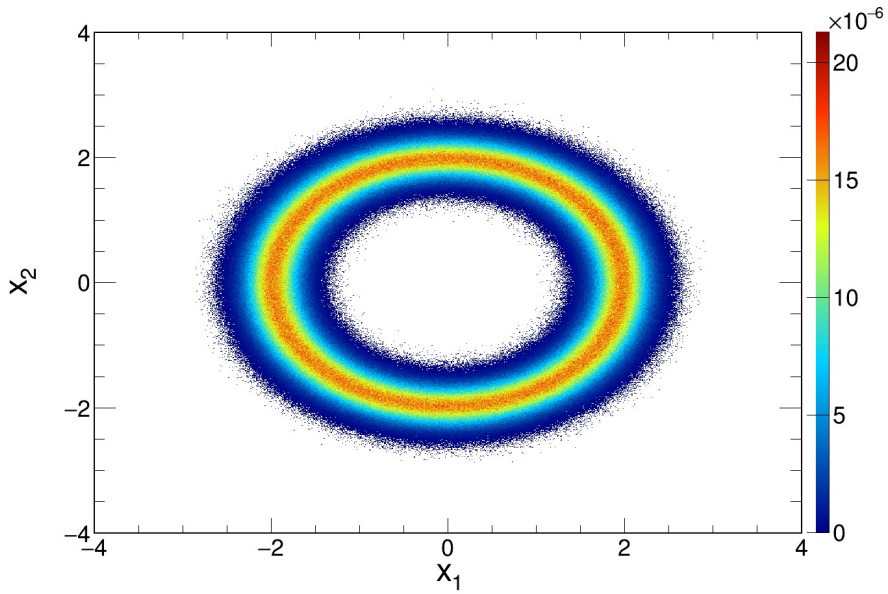
$$x_2 = \rho\xi_1 + \sqrt{1-\rho^2}\xi_2$$

  - Correlation = 0.8



Lack of Correlation != No Relationship

- Correlation coefficient only quantifies linear relationships
- Random distribution here has a correlation coefficient of zero
- x_1 and x_2 are clearly related to each other



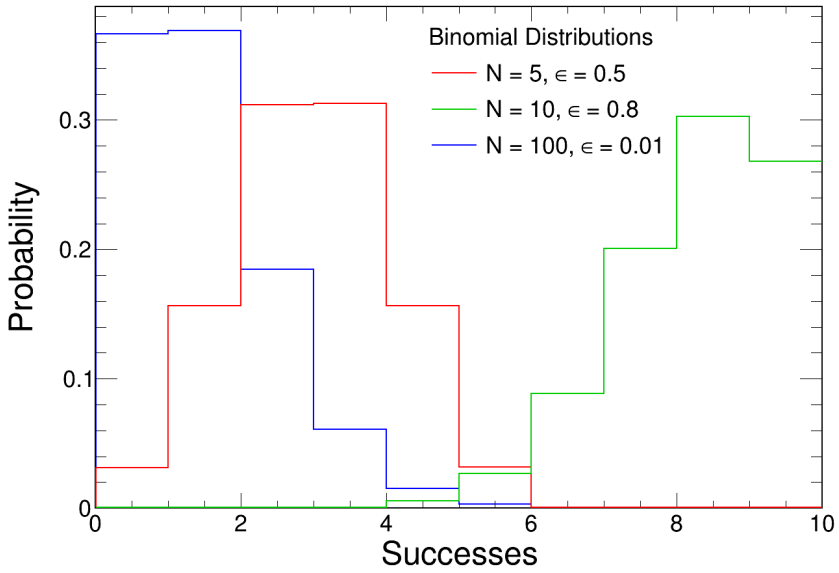
Binomial Distribution

- Represents probability of n successes given N trials, each with a probability ε of success
- Strictly speaking, all bin counts are drawn from a binomial distribution
  - Number of neutrinos of a given energy is finite
  - A small fraction interact in the detector
- Generally, the number of successes is very low compared to the number of trials
  - In this case, binomial distribution → Poisson

$$f(n|N,\epsilon) = \frac{N!}{n!(N-n)!}\epsilon^n(1-\epsilon)^{N-n}$$

$$\mu = N\epsilon$$

$$\sigma^2 = N\epsilon(1-\epsilon)$$



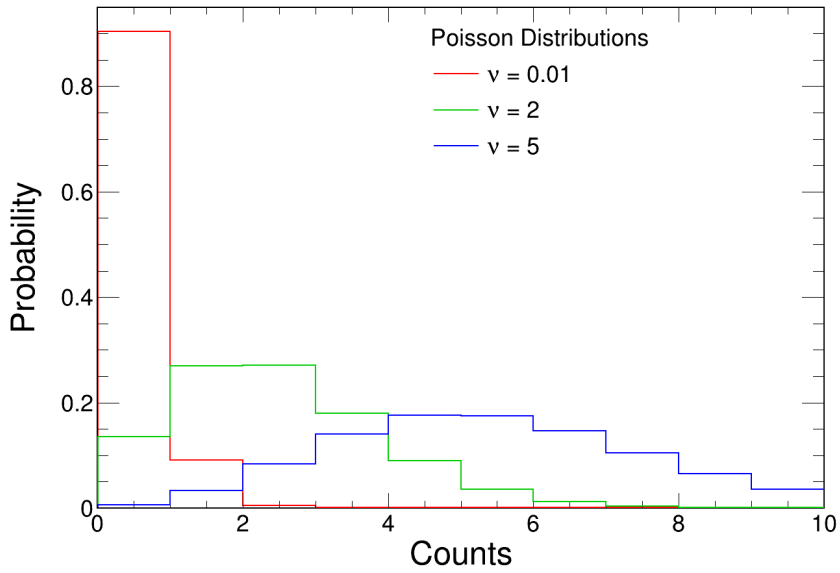
Poisson Distribution

- The limiting distribution where
  - $N \longrightarrow \infty$
  - $\varepsilon \longrightarrow 0$
  - $N\varepsilon \longrightarrow \nu$
- Predicted number of events in a bin is ν
- Actual number of events observed, n, is Poisson distributed

$$f(n|\nu) = \frac{\nu^n e^{-\nu}}{n!}$$

$$\mu = \nu$$

$$\sigma^2 = \nu$$



Gaussian Distribution

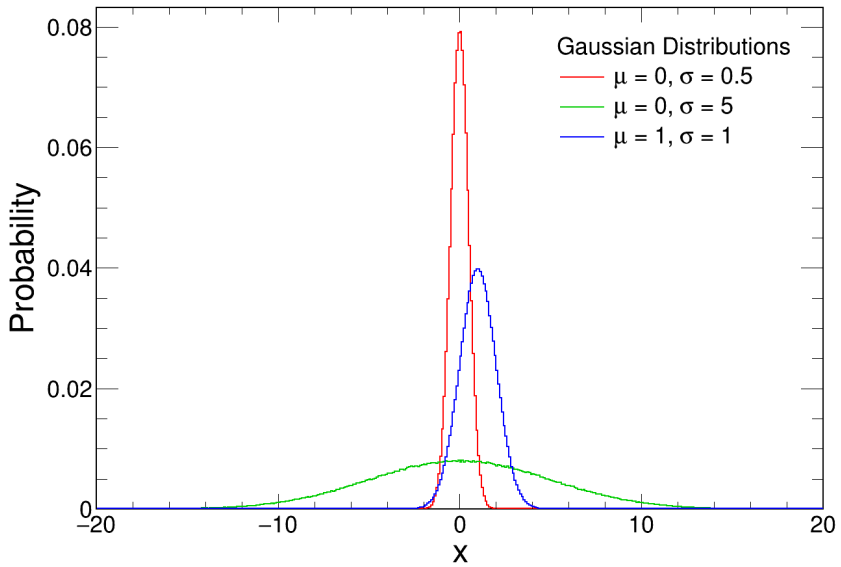
- If the expected number of events is large, the distribution can be approximated as Gaussian
  - Origin of $\sqrt{N}$ error bars
- Continuous distribution
- 68.3% of probability between μ–σ and μ+σ
- Sum of two Gaussian is also a Gaussian
  - Means add
  - Standard deviations add in quadrature

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

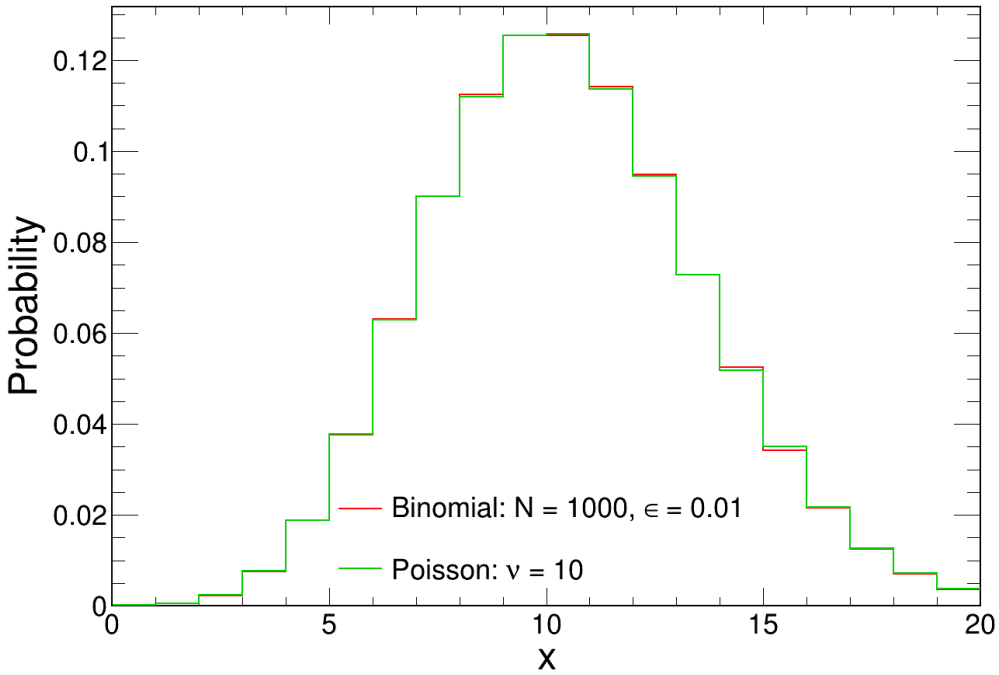
$$\mu = \mu$$

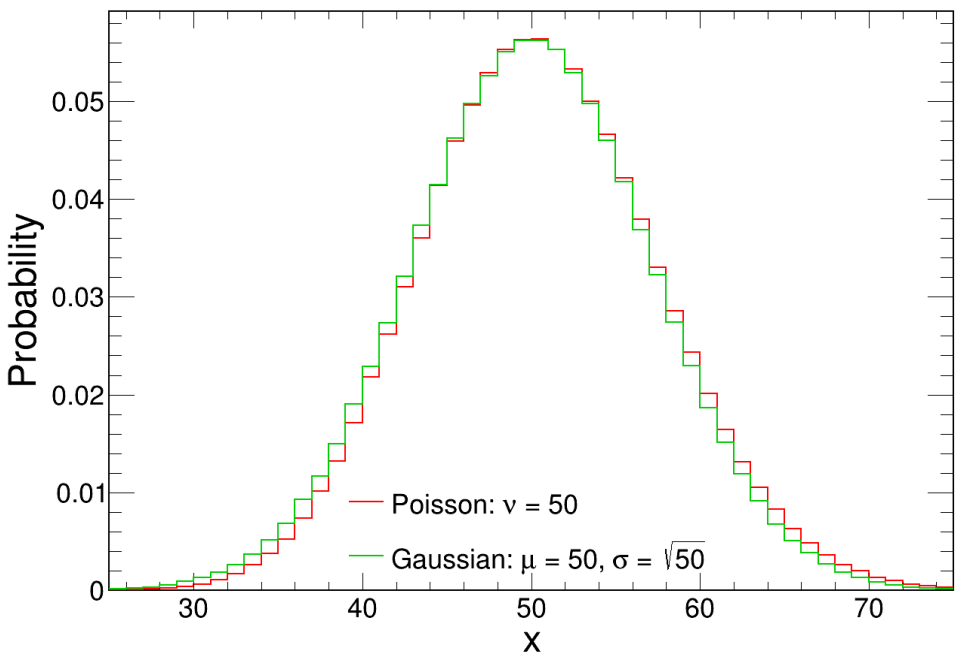
$$\sigma^2 = \sigma^2$$

(That is, the moments are the parameters of the pdf)



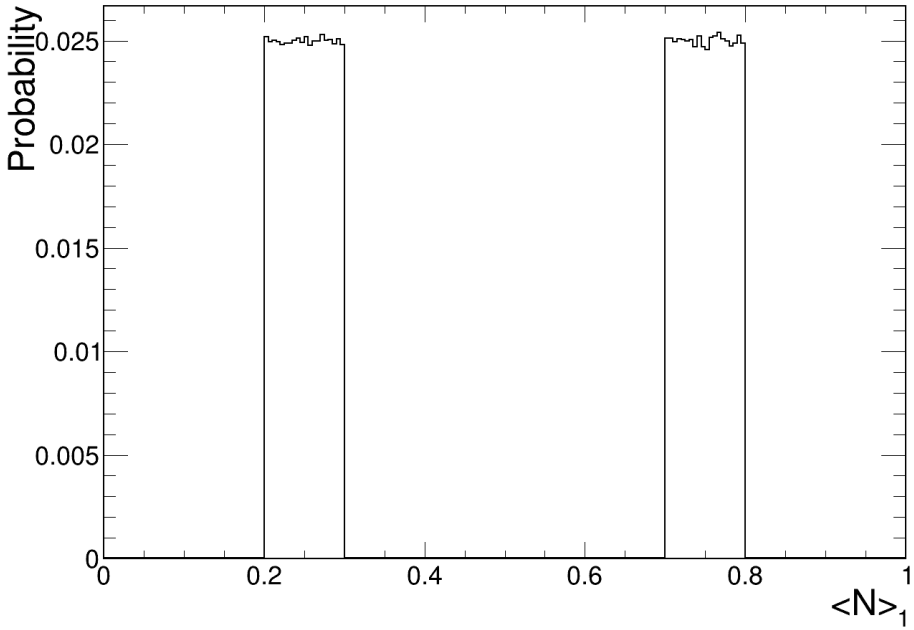
Relationship Between Distributions





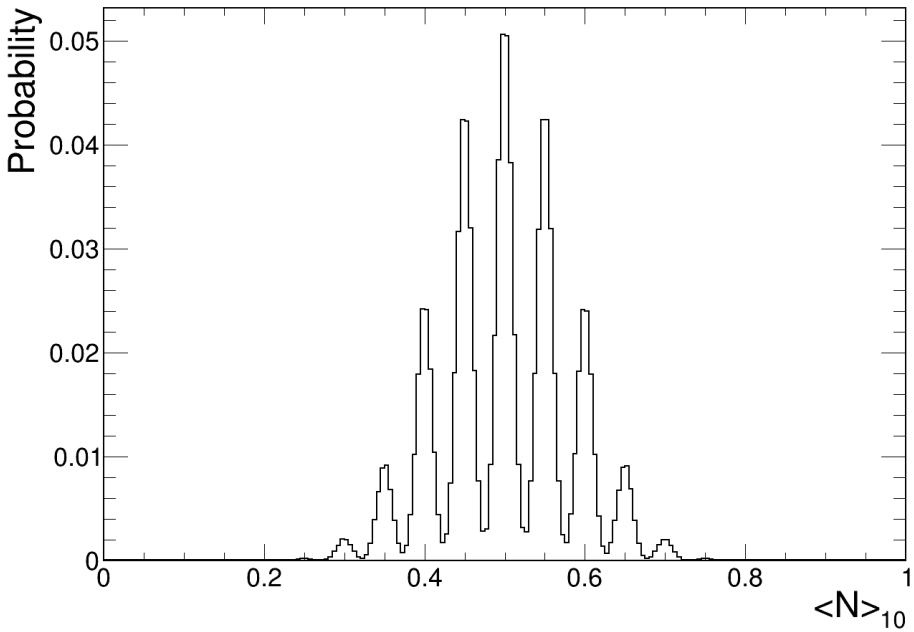
Central Limit Theorem

- The sum of small, uncorrelated random numbers is asymptotically Gaussian distributed
- This is true even for very non-Gaussian underlying distributions
- This is the reason why Gaussian uncertainties are so common in statistical analyses



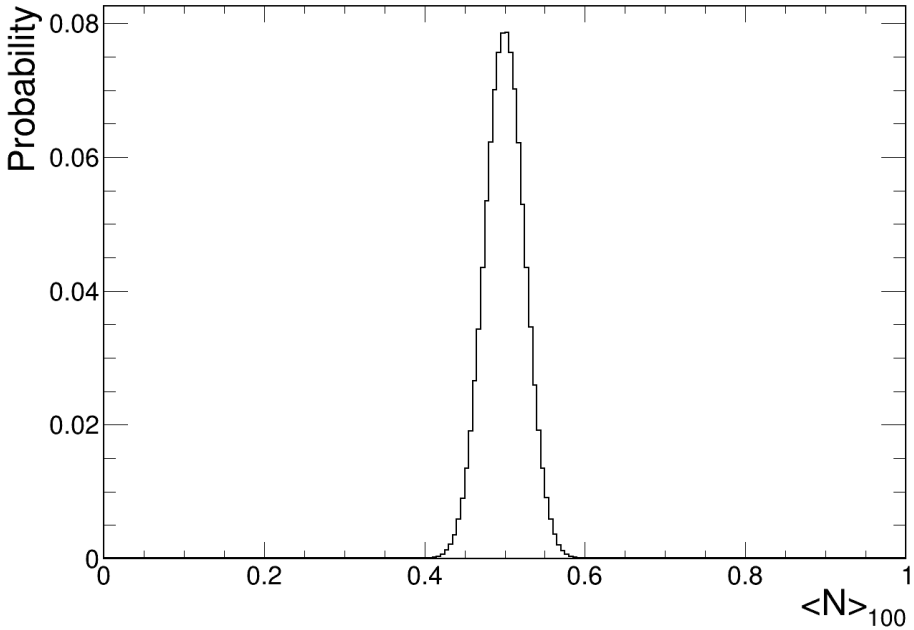
Central Limit Theorem

- The sum of small, uncorrelated random numbers is asymptotically Gaussian distributed
- This is true even for very non-Gaussian underlying distributions
- This is the reason why Gaussian uncertainties are so common in statistical analyses



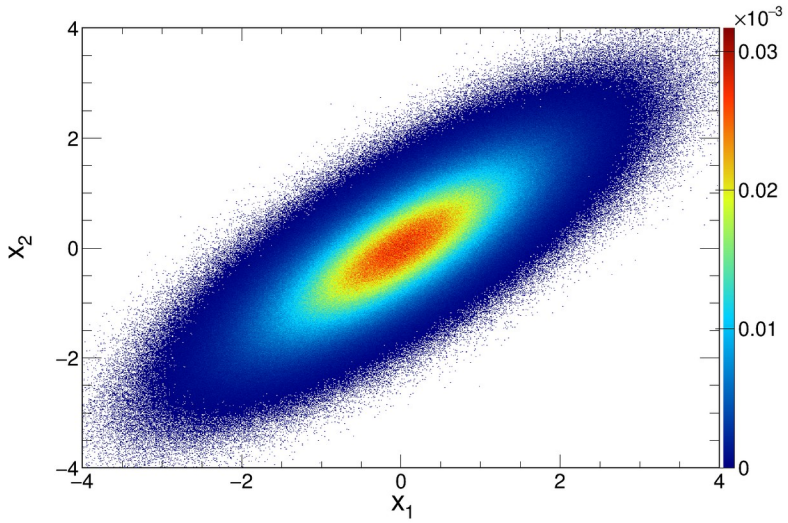
Central Limit Theorem

- The sum of small, uncorrelated random numbers is asymptotically Gaussian distributed
- This is true even for very non-Gaussian underlying distributions
- This is the reason why Gaussian uncertainties are so common in statistical analyses



Multivariate Gaussian

- Multivariate Gaussian is the N-dimensional generalization of the Gaussian
- Standard deviation is replaced by the covariance matrix
- Can be a good model for histogram bin contents if contents are relatively large, and systematic uncertainties have correlations between bins



Can draw numbers from multivariate Gaussian easily

- Cholesky decompose covariance matrix: $AA^T = \Sigma$
- Draw z, an N-dimensional vector of unit Gaussian random numbers
- $x = \mu + Az$

$$f(\vec{x}|\vec{\mu},\Sigma) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right]$$

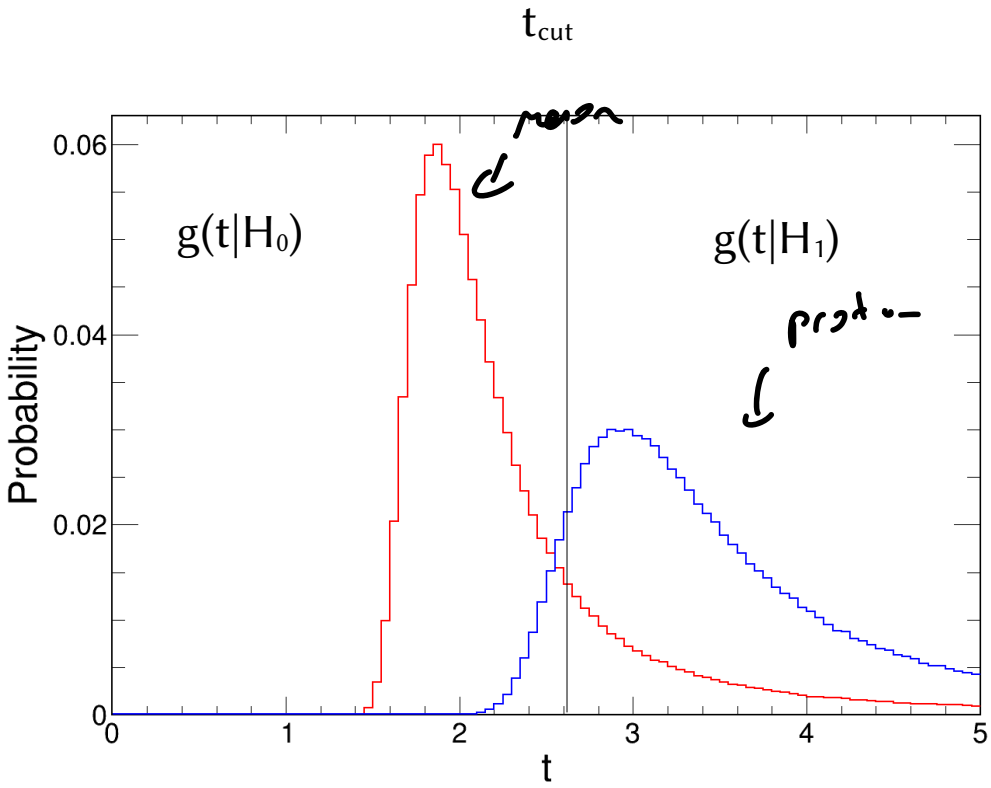
Hypothesis Testing

Hypothesis Testing

- Language of statistics requires us to pick a default hypothesis
  - Call $H_0 \longrightarrow$ null hypothesis
  - If H_0 has no free parameters, it is a simple hypothesis, otherwise it is a composite hypothesis
- Can have many alternate models
  - Call one of them we are interested in testing against H_1

Particle Separation

- Suppose we have identified a track in our detector, and we want to know if it is a muon or a proton
- Consider muon hypothesis to be our null hypothesis
- Construct a test statistics t(x) which is a function only of the data
- Distribution of t under different hypothesis tells us whether we can reject our null hypothesis
- Define a critical region for t such that we reject H_0 if t_{Obs} is in the critical region

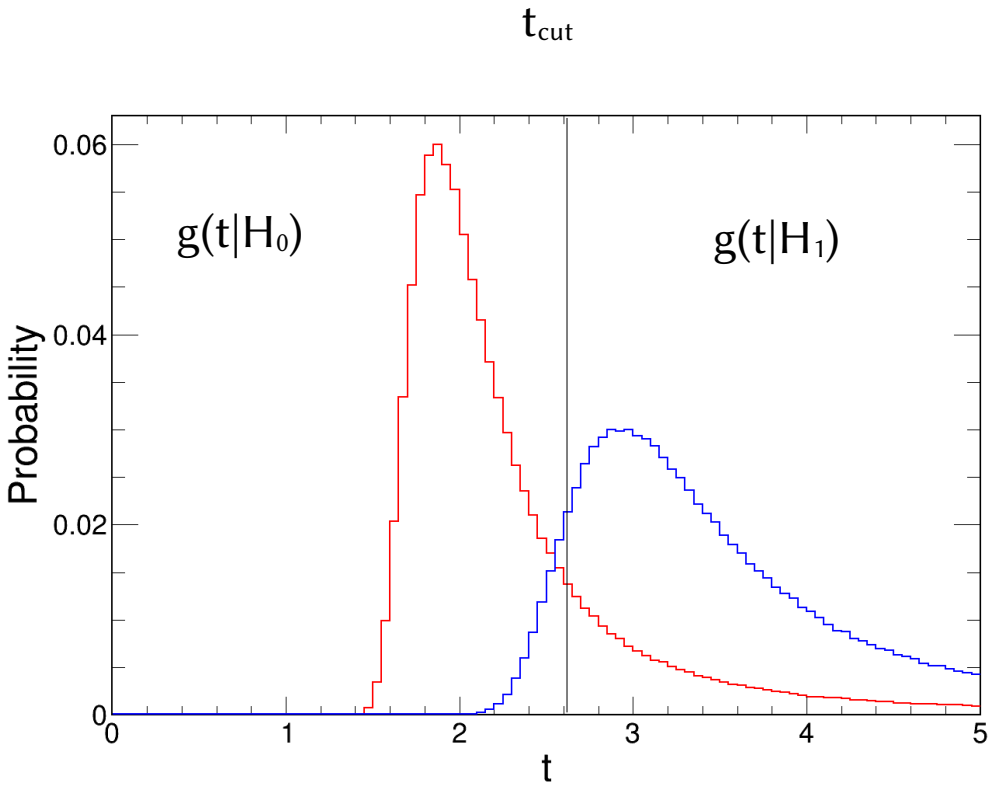


t_{cut} is chosen so that probability of rejecting H_0 when it is true is α

$$\alpha = \int_{t_{cut}}^{\infty} g(t|H_0)\, dt$$

Particle Separation

- Suppose we have identified a track in our detector, and we want to know if it is a muon or a proton
- Consider muon hypothesis to be our null hypothesis
- Construct a test statistics t(x) which is a function only of the data
- Distribution of t under different hypothesis tells us whether we can reject our null hypothesis



For a given α, the probability of accepting H_0 when H_1 is true is β

$$\beta = \int_{-\infty}^{t_{cut}} g(t|H_1)\, dt$$

Neyman-Pearson Lemma

- If the test statistic, t, is one dimensional, β is completely specified by picking α
- If the test statistic is a vector, there are many critical regions for which the significance level is α
- Neyman-Pearson lemma says that the likelihood ratio produces the acceptance region with the highest power (that is, highest signal purity for a chosen selection efficiency)

$$\frac{g(\vec{t}|H_0)}{g(\vec{t}|H_1)} > c$$

c is determined by fixing 1 – α (selection efficiency)

Likelihoods can often be constructed as multi-dimensional histograms from simulation