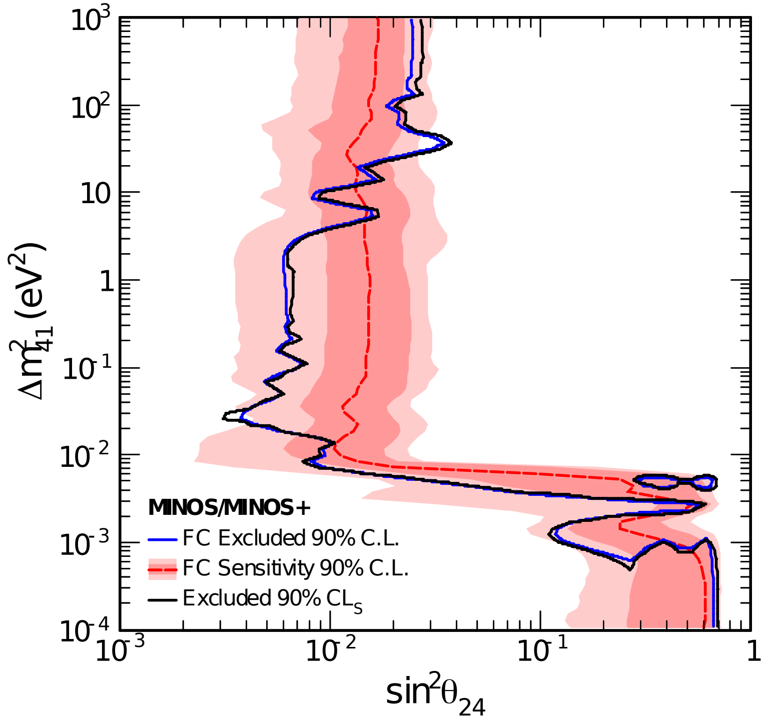
Making Sense of Your Data: Statistics and Machine Learning

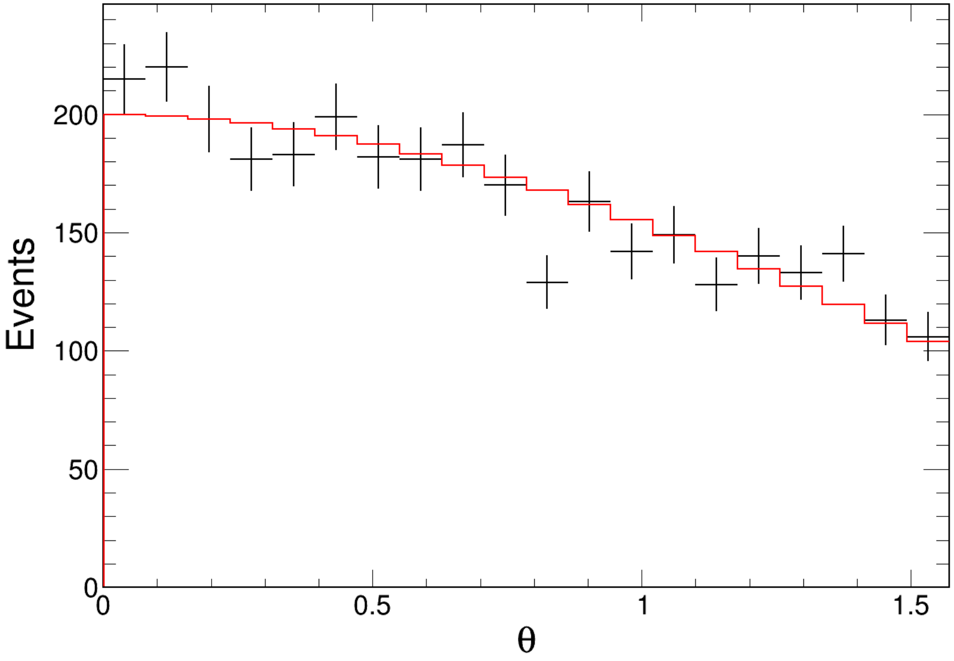
Adam Aurisano
University of Cincinnati

Understanding the Universe
Through Neutrinos
30 April 2024



Goodness of Fit

- Suppose the red line corresponds to a simulated prediction and the black crosses correspond to data
- Is the data consistent with having been drawn from the distribution in red?
- Phrase as a hypothesis test
  - H_0 = data is consistent with simulation
  - $\alpha = 0.05$
    - This is a common choice



Goodness of Fit

Since the number of events per bin are large, we can treat each points as Gaussian distributed

We can write our likelihood like this, assuming each bin has uncorrelated uncertainties

$$\mathcal{L}(\vec{x}|\vec{\mu}(\vec{\theta})) = \prod_i^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right)$$

Since sums are generally easier to deal with than products, we take the natural log:

$$-2\ln\mathcal{L} = \underbrace{\sum_i^N \frac{(x_i-\mu_i)^2}{\sigma_i^2}} + C$$

This quantity is the same as χ^2 up to a constant

Now the question is, what is the probability of finding a χ^2 at least as large as what we observe in data, under the assumption that H_0 is true?

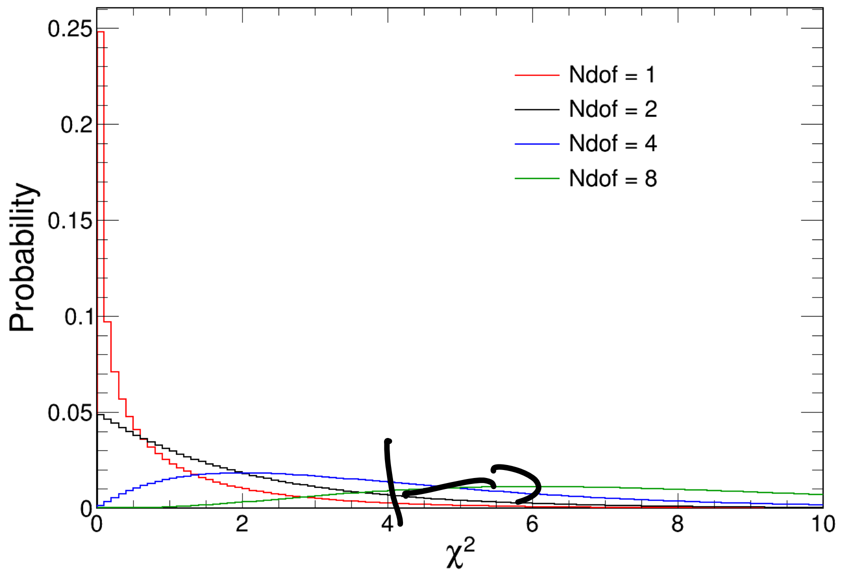
χ^2 Distribution

$$z = \frac{x-\mu}{\sigma} \qquad x \sim \text{Gaus}(\mu,\sigma)$$

$$z \sim \text{Gaus}(0,1)$$

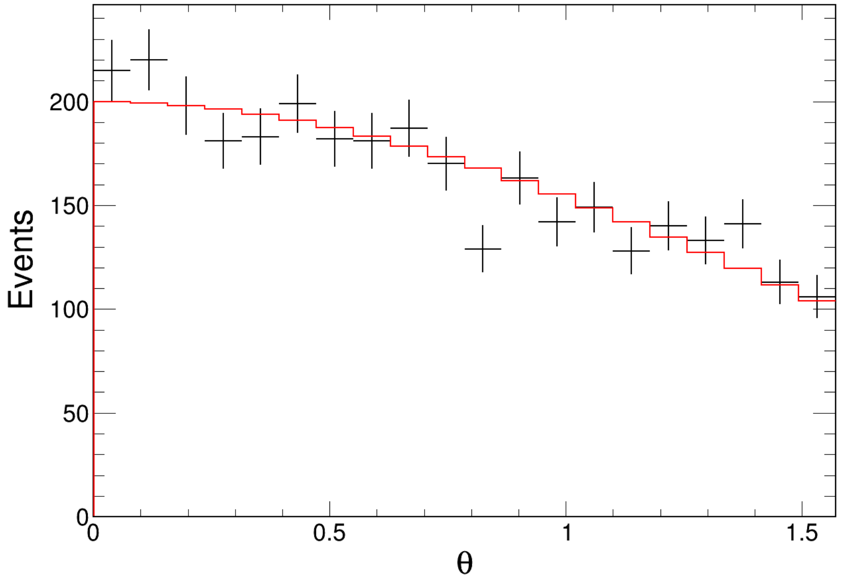
$$\chi^2 = \sum_i \left(\frac{x-\mu}{\sigma}\right)^2 = \sum_i z_i^2$$

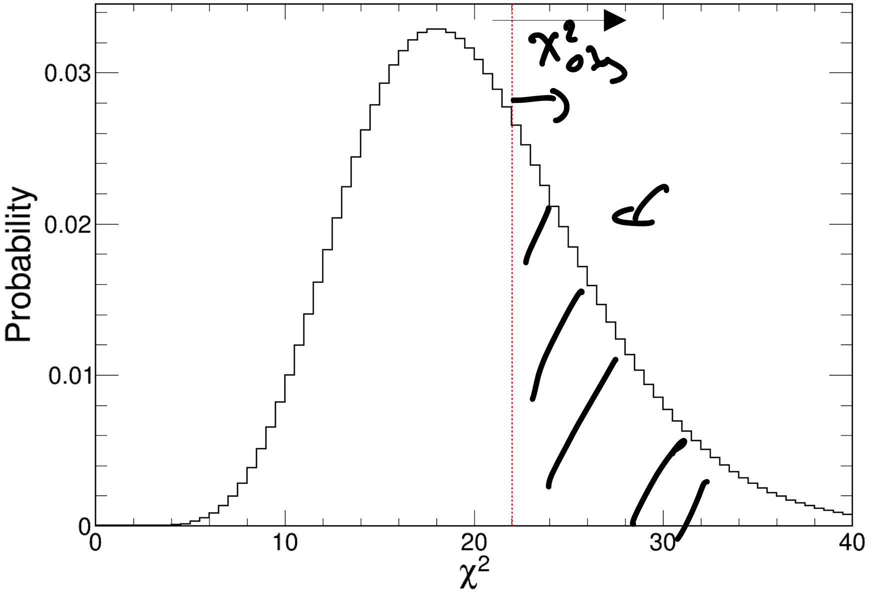
- Since our data are Gaussian distributed, χ^2 is a good test statistic
- For a given number of degrees of freedom, we can use TMath::Prob(chi2,ndof) to determine the tail integral of the χ^2 distribution
- Ndof is roughly the number of bins – number of free parameters
  - In this case, there are no free parameters



Goodness of Fit

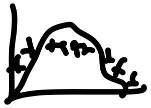
- Back to our example
- There are 20 bins and no parameters ⟶ 20 dof
- We can plot the probability distribution of a χ^2 with dof = 20 and put the observed χ^2 on it
- p-value is integral form red line to infinity
  - $\chi^2_{obs} = 22.02$
  - p-value = 0.34
- We cannot reject the H_0 at the 5% level
  - Note: if the bin counts are small, the test statistic may not be χ^2 distributed. In that case, you can use Monte Carlo methods to compute the distribution of the test statistic

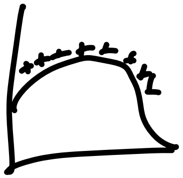




Goodness of Fit and Error Bars

- In the absence of any other information, the observed value in a bin is the best estimator of the expected value in that bin
- Since counts are Poisson distributed, we draw errors equal to sqrt(N)
  - Expectation is that if we reran the experiment many times, the data would fall within the error bars 68.3% of the time
- Strictly speaking, data has no error (we observe what we observed), but it is conventional to attribute statistical uncertainties to data points rather than predictions
- Roughly 2/3 of bins should have the prediction touching the error bars
  - Too many or too few could mean over- or under-estimating uncertainties
- If many bins in a row are above or below the prediction, mismodeling is possible, even if χ^2 is fine
  - All above or below → normalization problem?
  - Above on one side and below on the other → calibration problem?
  - Above on two sides and below in the middle → resolution problem?



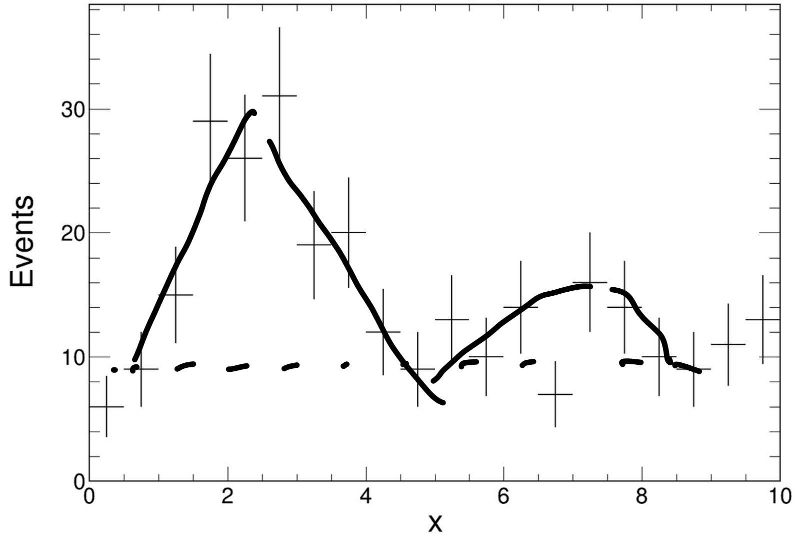


P-Values

- The tail probabilities we have been calculating are known as p-values
- The statement to remember is:
  - "What is the probability of obtaining at least as discrepant a result as we observed in data if the null hypothesis is true?"
- Although there are many asymptotic results telling what p-values should be, you will often find yourself needing to do a pseudo-experiment study to empirically determine the distribution of your test statistic
- In ROOT, use can use the TRandom3 class to draw high quality random number to produce your pseudo-experiment ensemble
- p-values should be uniformly distributed if H_0 is true

Signal Significance

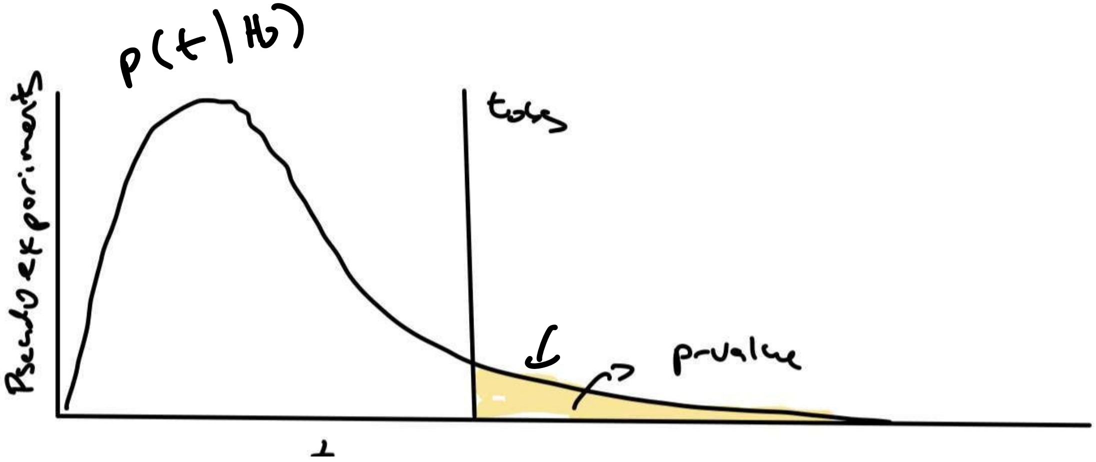
- Suppose we observed the following data distribution
- We want to know if there are significant peaks above background
- Null hypothesis
  - Data is drawn from the background distribution (uniform here)
  - What is the probability that our test statistic is at least as large as we observe in data?

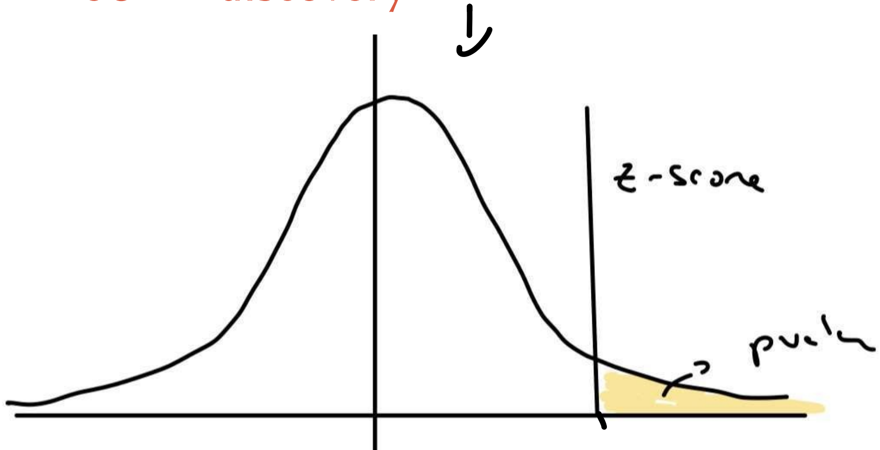


- Generate pseudoexperiments
  - Fluctuate predicted background counts according to any systematic uncertainties
  - Poisson fluctuate those counts get a single pseudoexperiment
- Fit for your peak
- Place the resulting test statistic in a histogram
- Determine what fraction of pseudoexperiments has a test statistic larger or equal to the test statistic observed in data

Signal Significance

- The tail probability corresponds to the p-value
- In HEP we usually say if the z-score (significance) of the observation is
  - $> 3\sigma$ ⟶ hint, evidence
  - $> 5\sigma$ ⟶ discovery





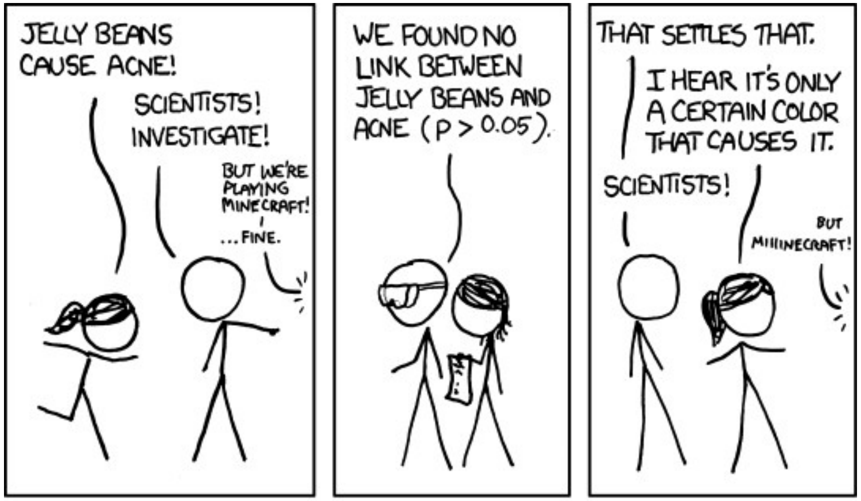
$$p = \frac{1}{2}\left(1 - \mathrm{Erf}(z/\sqrt{2})\right)$$

$$z = \mathrm{NormQuantile}(1 - p)$$

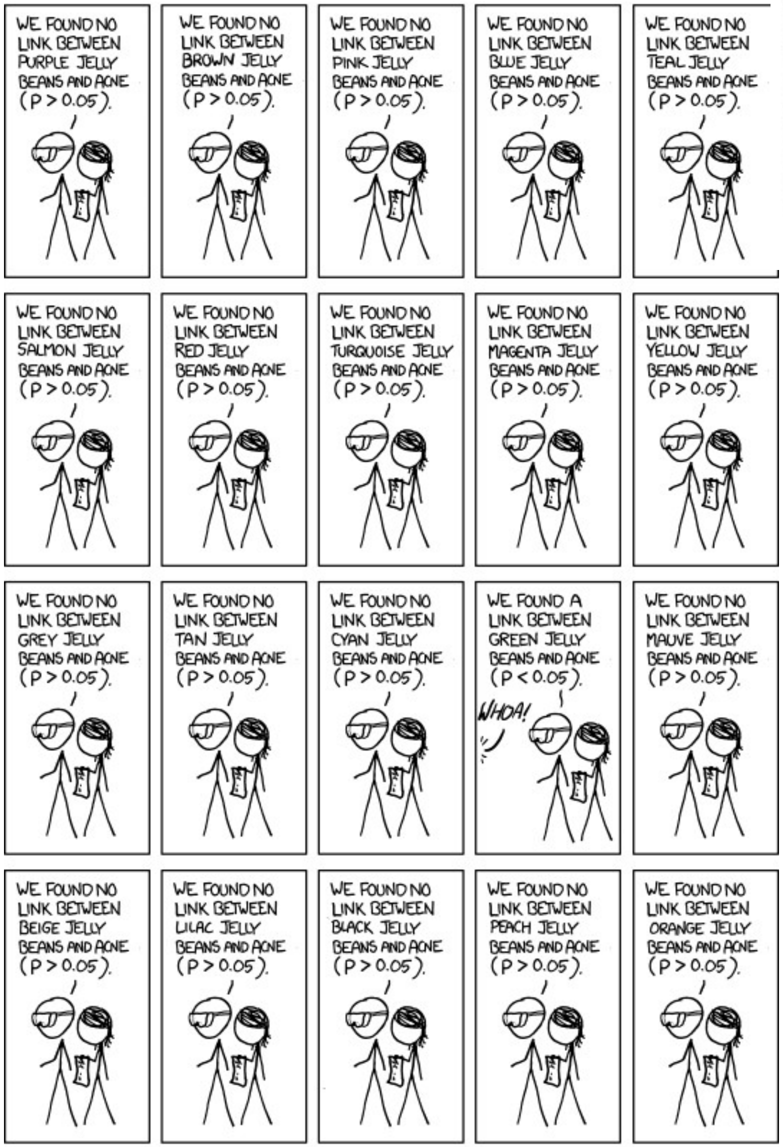
- Z-score refers to the point on the x-axis for unit Gaussian such that the tail probability matches the p-value in question
- We can convert easily between p-values and z-scores

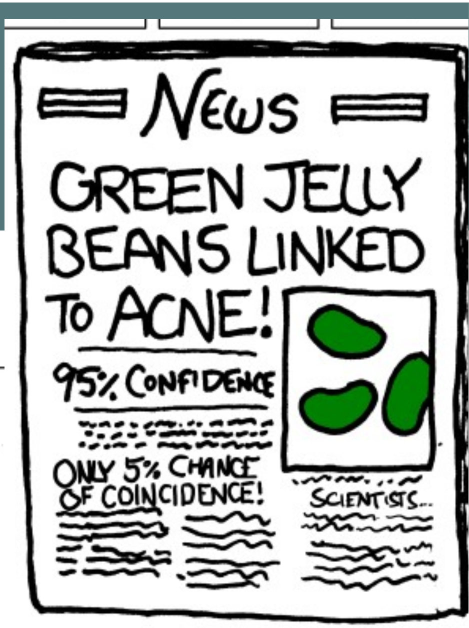
ROOT contains both functions:
TMath::Erf and TMath::NormQuantile

Look-Elsewhere Effect


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

- Warning!
  - If you test for a signal at multiple locations, you run the risk of falsely rejecting the null hypothesis
  - Need to divide p-value by number of independent measurements
  - If you allowed the position of the peak to float in the previous example, this would be automatically accounted for


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Point Estimation

Maximum Likelihood Principle

- The likelihood is the probability of having observed the data given a model
- If the model has parameters, intuitively, we can see that the parameters that maximize the likelihood are best parameters
- Parameter estimators of this form can be shown to be
  - unbiased
  - efficient
  - asymptotically normal
  - invariant under transformations

Likelihoods

$$\mathcal{L} = \prod_{i}^{n} f(x_i|\vec{\theta})$$

- If we know the probability distribution for each bin, we can compute the likelihood across all bins
- θ̄ is the set of parameters that specify the model being fit
- We want to maximize the likelihood:

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = 0$$

Likelihoods

- Products are generally hard to work with, so we typically take the natural log of the likelihood
  - This does not affect the location of the extrema
- Optimization problems are usually frames as minimizations, so we make it negative
- Finally, if we write: $$\left. -2\ln\lambda = -2\ln\frac{\mathcal{L}(\vec{x}|\vec{\mu})}{\boxed{\mathcal{L}(\vec{x}|\vec{x})}} \dashrightarrow \right\}$$
  - which is asymptotically χ^2 distributed

Gaussian Likelihood

$$\mathcal{L}(\vec{x}|\vec{\mu}(\vec{\theta})) = \prod_i^n \frac{1}{\underline{\sqrt{2\pi\sigma_i^2}}} \exp\left(-\frac{(x_i - \overset{\rightarrow}{\mu_i})^2}{2\sigma_i^2}\right)$$

$$-2\ln\frac{\mathcal{L}}{\underline{\mathcal{L}_0}} = \sum_i^n \frac{(x_i - \mu_i)^2}{\sigma_i^2} = \chi^2$$

The Gaussian likelihood, with uncorrelated uncertainties, reproduces the χ^2 test statistic

$$\mathcal{L}(\vec{x}|\vec{x}) = \prod_i^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\overset{0}{\cancel{(x_i - x_i)^2}}}{2\sigma_i^2}\right) = \prod_i^n \frac{1}{\sqrt{2\pi\sigma_i^2}}$$

Example: Linear Fit

- Suppose we have four data points
  - $Y = \{2.3, 2.5, 4.25, 4.75\}$
  - $X = \{0, 1, 2, 3\}$
  - $\sigma = 0.5$
- We believe that these data points derived from a linear model
- What is the best fit slope and intercept?
- Note: this can be done by hand

Example: Linear Fit

$$\chi^2 = \sum_i \left(\frac{y_i - \mu_i}{\sigma_i}\right)^2$$

$$\mu_i(m,b) = m x_i + b$$

$$\chi^2 = \sum \left(\frac{y_i - m x_i - b}{\sigma_i}\right)^2$$

$$\frac{\partial \chi^2}{\partial m} = 0 \qquad \frac{\partial \chi^2}{\partial b} = 0$$

Connection to Ordinary Least Squares

Ordinary least squares is very closely connected to χ^2 fit: compare your homework to this result

Suppose we measure N data points, and we believe they are modeled by a line. Lines have two parameters, so this is an overconstrained system. How can we solve it?

$$\begin{aligned} y_0 &= m x_0 + b \\ y_1 &= m x_1 + b \\ &\vdots \\ y_n &= m x_n + b \end{aligned}$$

$$\begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_0 & 1 \\ \vdots & \\ x_n & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix}$$

$$\begin{pmatrix} x_0 & \cdots & x_n \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} y_0 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_0 & \cdots & x_n \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_0 & 1 \\ \vdots & \\ x_n & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix}$$

$$\begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix} = \begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & \underset{\nwarrow}{\sum 1} \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix}$$

Poisson Likelihood

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

$$\mathcal{L}(\vec{n}|\vec{\nu}(\vec{\theta})) = \prod_i^N \frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!}$$

$$-2\ln\lambda = 2\sum_i^N \left(n_i \ln\frac{n_i}{\nu_i} + \nu_i - n_i \right)$$

This should be used any time you have a doubt about whether or not your bins have enough events for the Gaussian approximation

This method correctly handles bins with zero observed counts

Non-Linear Fitting

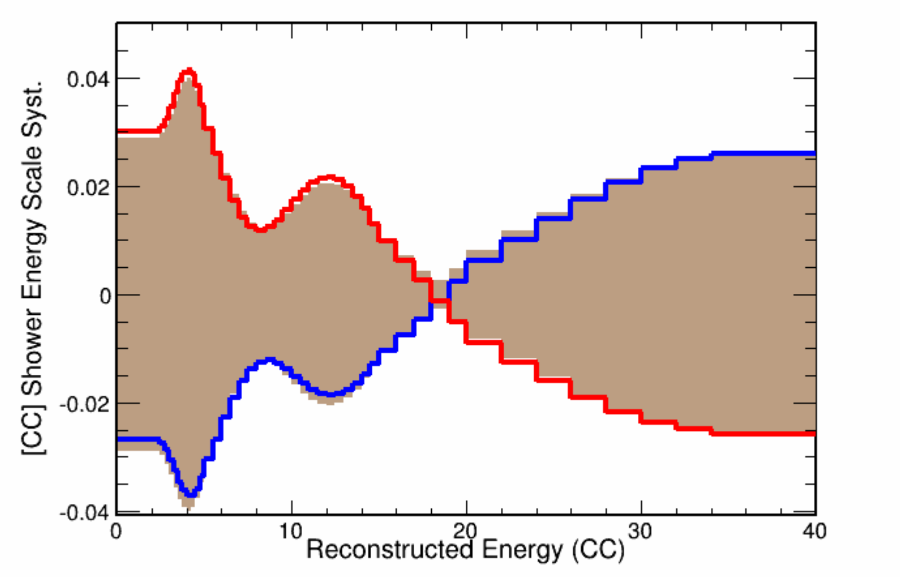
- Most of the time, you will not get a system of linear equations if you work with a Poisson log likelihood
- Many optimizers exist which can minimize the objective function
  - Minuit is a very commonly used package in HEP
    - MIGRAD: uses Hessian matrix to construct 2nd approximations of the objective function
      - Very fast if the function is well behaved
      - Has an internal estimate of the distance to minimum to let you know you converged
      - Sometimes goes in very wrong directions
    - Simplex: uses geometrical objects to find the minimum
      - Very slow, but sometimes helpful to refine your search
- There is no surefire way to minimize a function with multiple minima
  - Do parameter scans to make surfaces of the log likelihood as a function of different parameters
    - Does it look lumpy? Maybe the minimizer got stuck in a false minimum?
- If you want to do an analysis with precision, you have to make sure that you can robustly minimize the log likelihood, even in the presence of reasonable flucutations
  - Do pseudoexperiment studies to check

Evaluating Systematic Uncertainties

- Systematic uncertainties are when the prediction you are trying to fit to data depends on parameters that are imperfectly known
- For instance
  - Final state interactions
  - Cross sections
  - Hadron production
  - Detector properties
  - etc, etc...
- Many things can affect our ability to accurately predict what we will see, and it is critical to think about this early in the analysis process
  - A result with systematic uncertainties quantified is not a result

Evaluating Systematic Uncertainties

- In general, we need to identify a potential source of mismodeling and determine how the prediction changes as we change it
- The amount you change the parameter could be guided by an auxiliary measurement
  - Often it is an informed guess
  - Systematic uncertainties are usually at least a little bit Bayesian at their core



The fractional change in the MINOS charged current sample due to +1σ and -1σ changes in the shower energy scale

Incorporating Systematic Uncertainties

To incorporate systematic uncertainties into our fit, we have to think about what the combined statistical and systematic likelihood is:

Imagine a parameter β that controls the size of a particular systematic uncertainty

Based on how we constructed our error bands, β =1 corresponds to the upper band and β = -1 corresponds to the lower band. Therefore, we can consider β to be Gaussian distributed with mean = 0 and σ = 1.

The joint likelihood is then the product of the statistical and systematic probabilities:

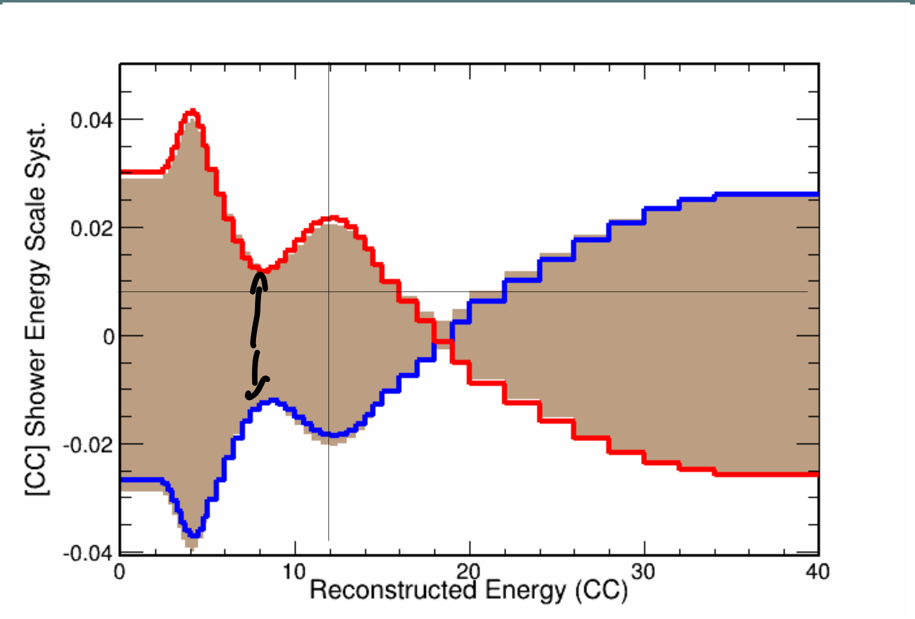
$$-2\ln\lambda = \underbrace{2\sum_i^N \left(n_i \ln\frac{n_i}{\mu_i(\vec{\beta})} + \mu_i(\vec{\beta}) - n_i\right)} + \overbrace{\sum_j^M \beta_j^2}$$

penalty terms

See arXiv:1103.0354 for more in-depth discussions

Incorporating Systematic Uncertainties

$$\mu_i(\vec{\beta}) = \prod_j^M \underline{(1+\alpha_j(\beta_j))}\,\underline{\nu_i}$$



- α is an interpolated scale factor between the two error bands
  - Lets us evaluate the systematic shift at other σ than ±1
  - Can use higher order Lagrange interpolation if you have ±2σ, ±3σ error bands as well
  - Often good enough to just linearly extrapolate at the ±1σ end points
    - Try this!
- Product of all systematic scale factors lets us scale our prediction as a function of many systematic parameters

$$\alpha_i = \frac{\beta_i(\beta_i-1)}{2}\alpha_i^- + \frac{\beta_i(\beta_i+1)}{2}\alpha_i^+$$

-1σ shift

+1σ shift

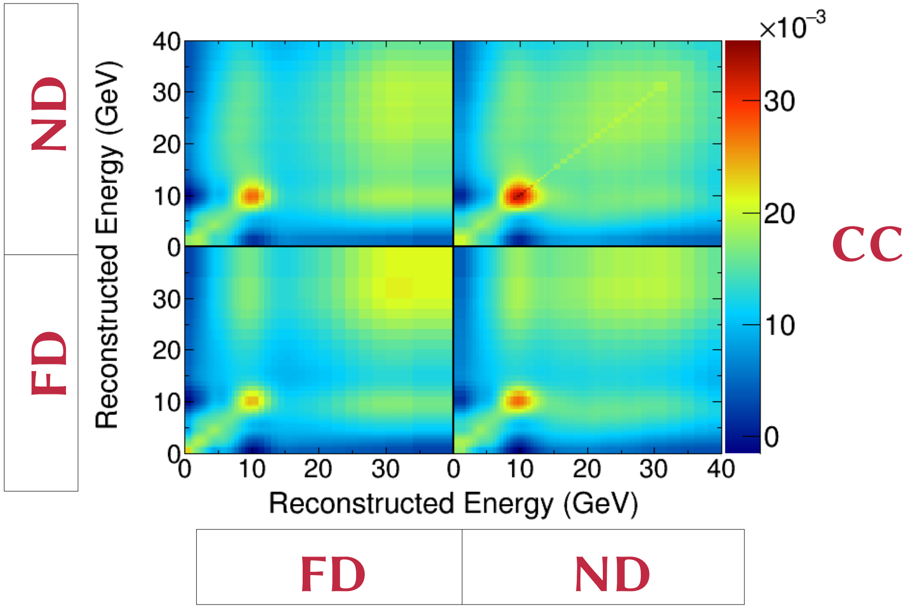
shift at $\beta_i\sigma$

The "pull method" lets us include systematic uncertainties that are fully correlated across bins

Multivariate Gaussian Likelihood

- The Poisson log likelihood with pulls is very useful, but it has some issues
  - Each systematic is a new fit parameter for Minuit to handle
  - Only fully correlated uncertainties are easy to include
    - Some simulation-based uncertainties cannot be reduced to a single degree of freedom
- Multivariate Gaussian likelihood replaces pulls with a covariance matrix
  - Covariance matrices from each uncertainties are added together
    - Equivalent to adding in quadrature
  - Since the uncertainties are not optimized during the fitting process, post-fit agreement cannot be judged by eye

$$\chi^2_{CC,NC} = \sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - \mu_i)[\mathbf{V}^{-1}]_{ij}(x_j - \mu_j) \quad \leftarrow$$



Full joint FD/ND covariance matrix from 2018 MINOS+ sterile search

Hybrid Poisson-Gaussian Likelihood

- It is possible to combine the strengths of the Poisson + pulls and the multivariate Gaussian likelihood
- Idea:
  - Use statistical part of Poisson likelihood
  - Treat systematic shifts in each bin (and each prediction component) as free parameters constrained by a multivariate Gaussian
  - Solve for best fit systematic pulls using Newton’s method at fixed oscillation parameters

$$\chi^2 = \underbrace{2\sum_{i}^{N}\left[\left(\sum_{\alpha}^{M}\mu_{\alpha i}s_{\alpha i}\right) - x_i + x_i\log\left(\frac{x_i}{\sum_\alpha^M \underbrace{\mu_{\alpha i}s_{\alpha i}}}\right)\right]} + \overbrace{\sum_{ij}^{N}\sum_{\alpha\beta}^{M}(s_{\alpha i}-1)V^{-1}_{\alpha i\beta j}(s_{\beta j}-1)},$$

$$\frac{\partial\chi^2}{\partial s_{\gamma k}} = 2\left(\mu_{\gamma k} - \frac{\mu_{\gamma k}x_k}{\sum_\alpha^M \mu_{\alpha k}s_{\alpha k}} + \sum_{\alpha i}^{MN}\left(s_{\alpha i}-1\right)V^{-1}_{\alpha i\gamma k}\right), \leftarrow$$

$$\frac{\partial^2\chi^2}{\partial s_{\gamma k}\partial s_{\delta l}} = 2\left(\frac{\mu_{\gamma k}\mu_{\delta l}x_k}{\left(\sum_\alpha^M \mu_{\alpha k}s_{\alpha k}\right)^2}\delta_{kl} + V^{-1}_{\gamma k\delta l}\right). \leftarrow$$

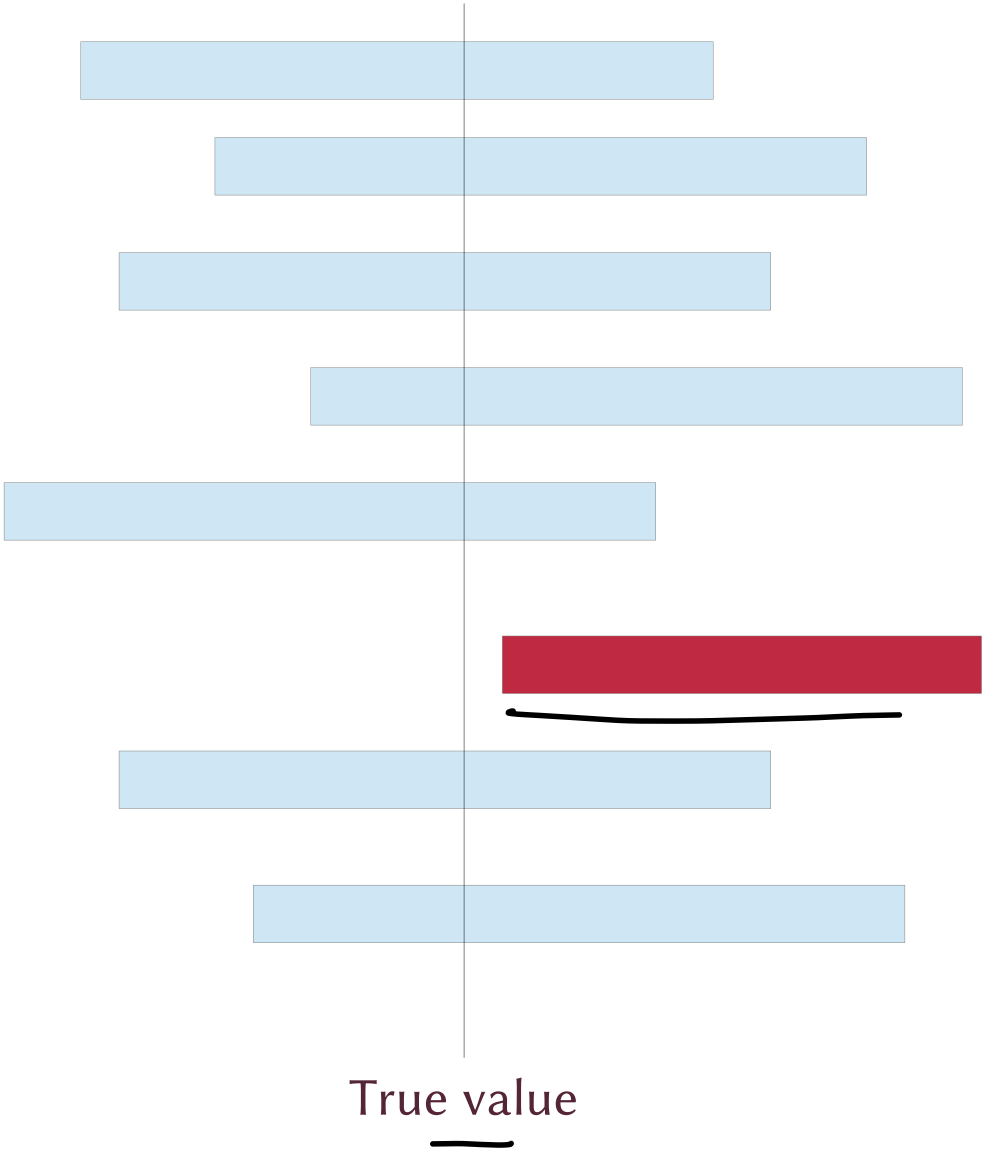
This method was used in the 2022 NOvA sterile analysis

$$H_{\chi^2}(\vec{s})\Delta\vec{s} = -\nabla\chi^2(\vec{s})$$

Interval Estimation

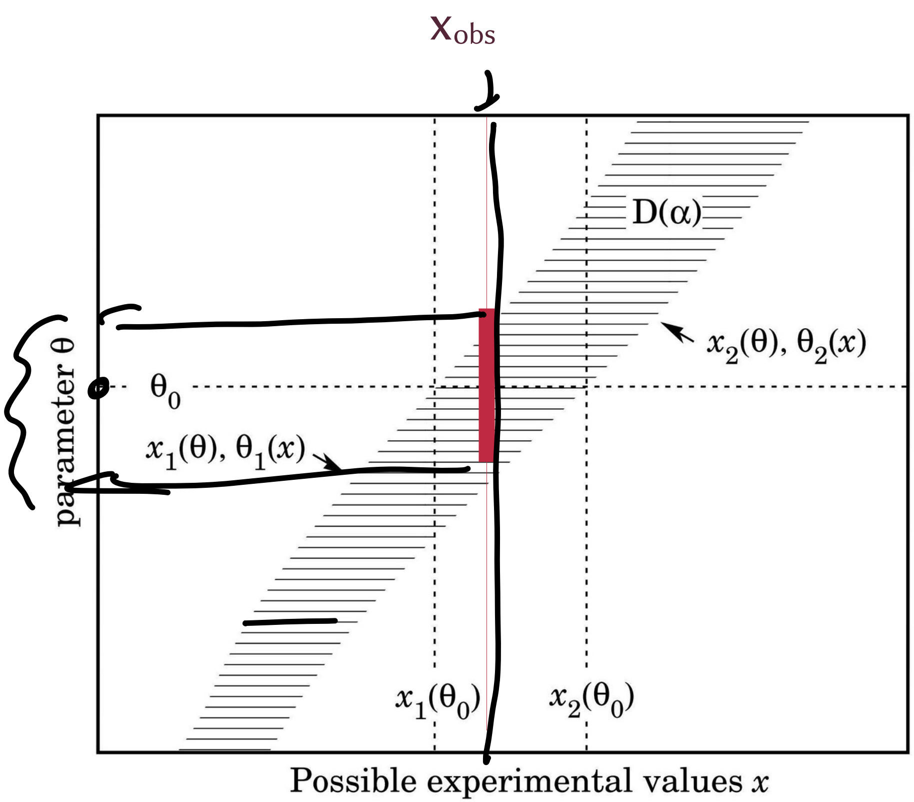
Interval Estimation

- We know know how to find best fit points, but we really need to have an estimate of the uncertainty of this result
- Frequentist idea of a confidence interval at $\alpha\%$ confidence level
  - If we repeat the experiment many times, $\alpha\%$ of the confidence intervals we draw will include the true value of the parameter
- 68.3% confidence level bands correspond to our conception of error bars



Neyman Construction

- For each possible value of the parameter of interest, we draw pseudoexperiments to find the possible range of experimental value it could produce
- We construct an interval which contains α% of the pseudoexperiments for each possible true value
  - Confidence belt
- Draw a vertical line at x_{obs}
  - Any value of θ for which the line intersects part of the confidence belt is part of your confidence interval



PDG statistics review