

Lies, Damn Lies, and your Analysis: Practical Statistics for Neutrino Physics

Alex Himmel

DUNE Physics Week
November 14th, 2017

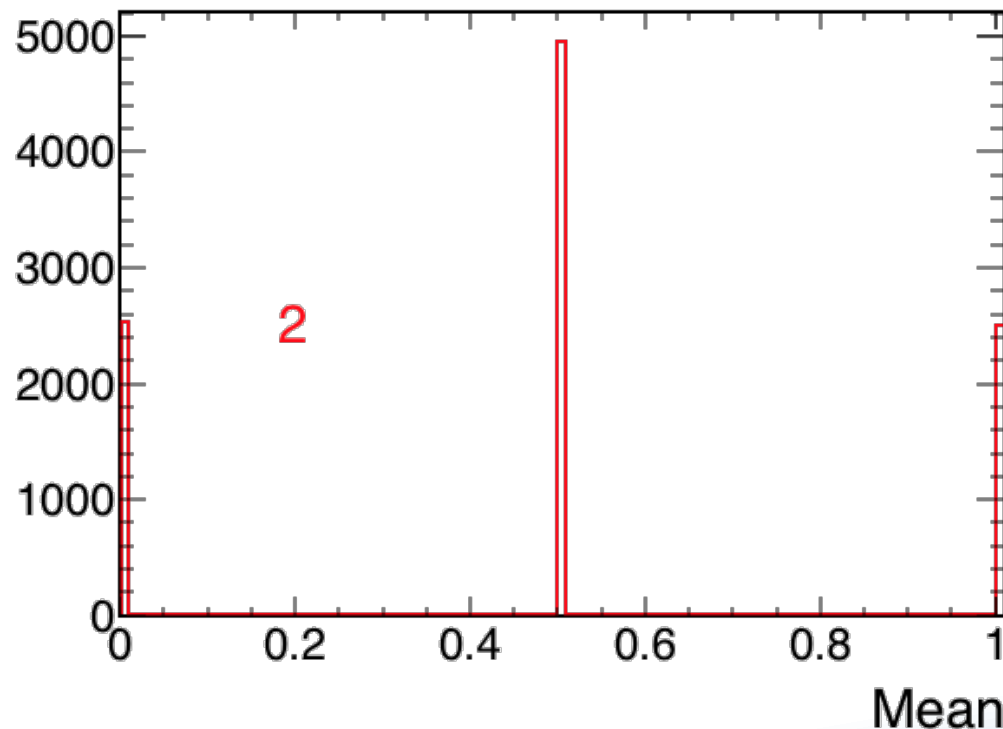


Statistics is a branch of **mathematics** dealing with the collection, analysis, interpretation, presentation, and organization of **data**.^{[1][2]}

- A collection of methods to extract meaning from data.
 - There are many, many methods.
 - The question you need to answer – is the method I’m using appropriate to my situation?
 - Make sure you’re clear about what you did, so others can interpret your results.
- **You are making an argument using data.**
- The answers are never simply “yes” or “no”
 - There is always a degree of uncertainty or level of agreement.

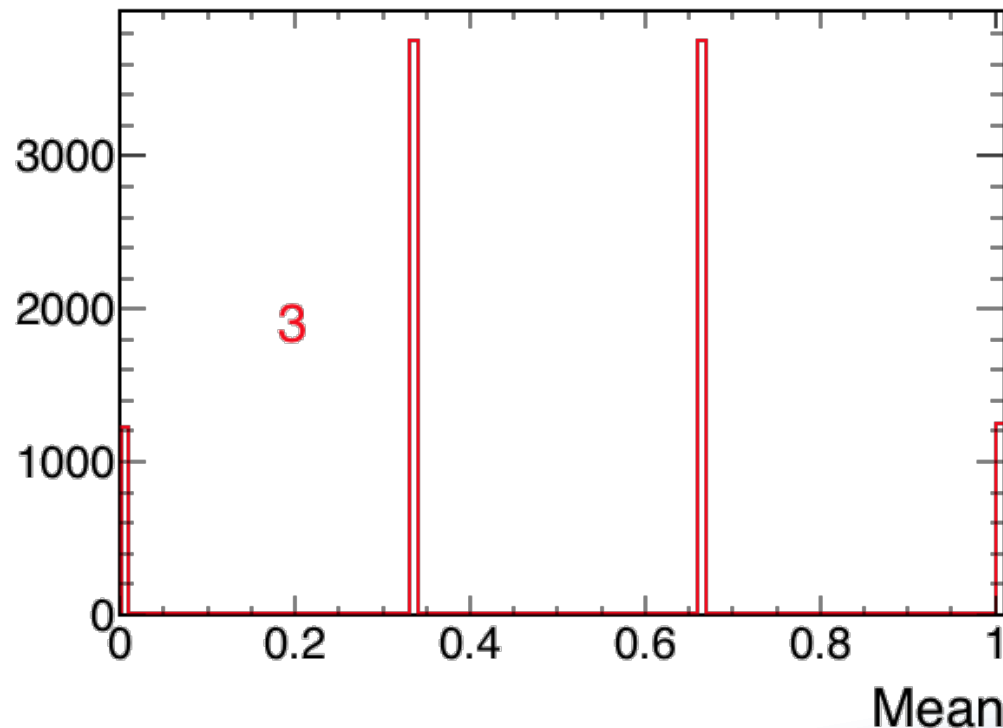
Central Limit Theorem

- The sum of a sufficiently large number of **independent random variables**.
 - It does not matter what distribution the underlying random variables come from.
- Example: coin flips. Heads = 0, tails = 1
 - Clearly not normally distributed.
- However, if we look at the distribution of the means:



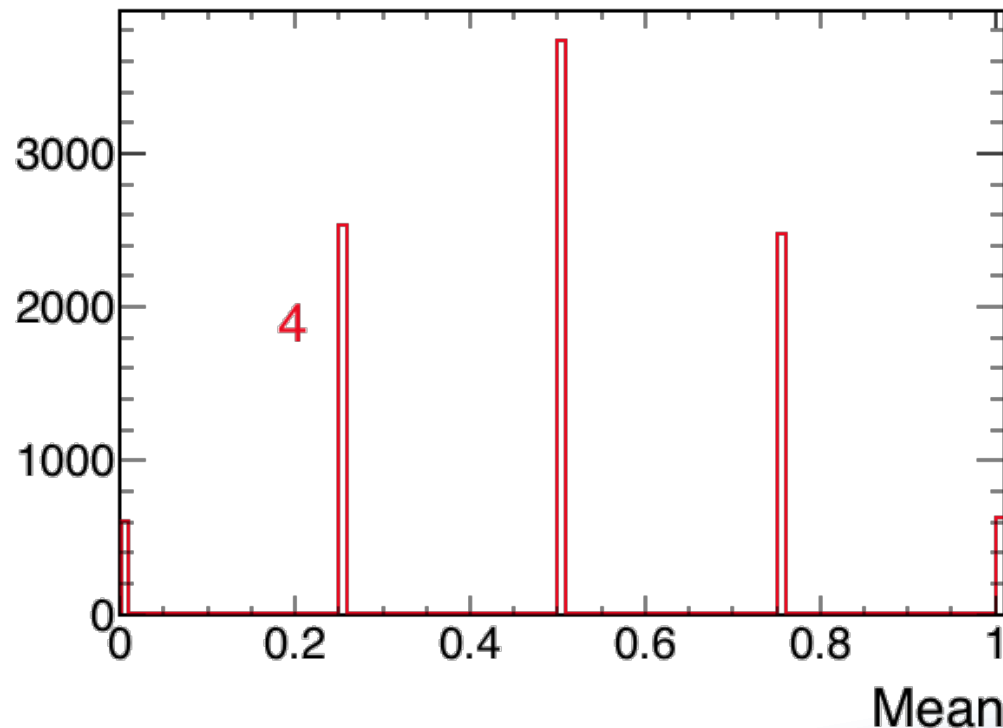
Central Limit Theorem

- The sum of a sufficiently large number of **independent random variables**.
 - It does not matter what distribution the underlying random variables come from.
- Example: coin flips. Heads = 0, tails = 1
 - Clearly not normally distributed.
- However, if we look at the distribution of the means:



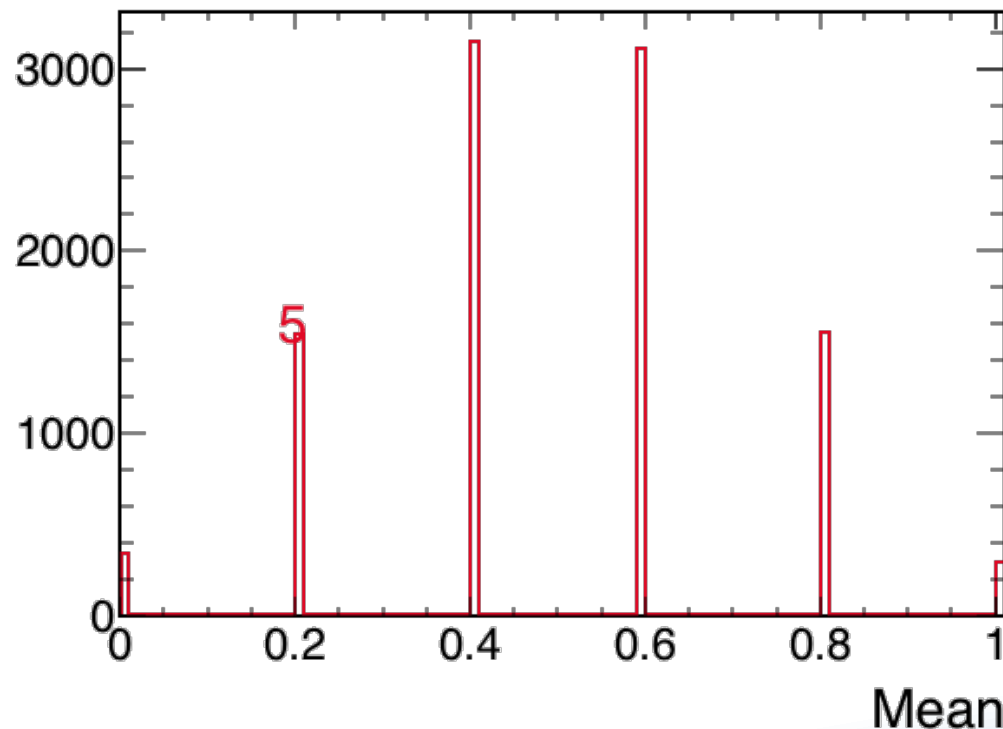
Central Limit Theorem

- The sum of a sufficiently large number of **independent random variables**.
 - It does not matter what distribution the underlying random variables come from.
- Example: coin flips. Heads = 0, tails = 1
 - Clearly not normally distributed.
- However, if we look at the distribution of the means:



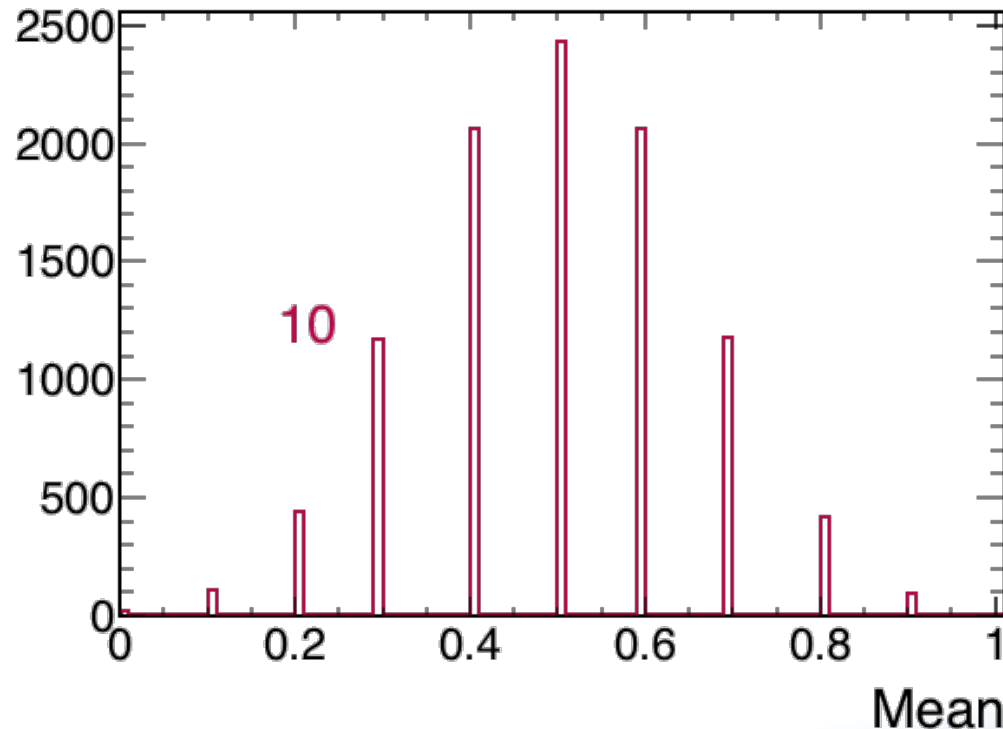
Central Limit Theorem

- The sum of a sufficiently large number of **independent random variables**.
 - It does not matter what distribution the underlying random variables come from.
- Example: coin flips. Heads = 0, tails = 1
 - Clearly not normally distributed.
- However, if we look at the distribution of the means:



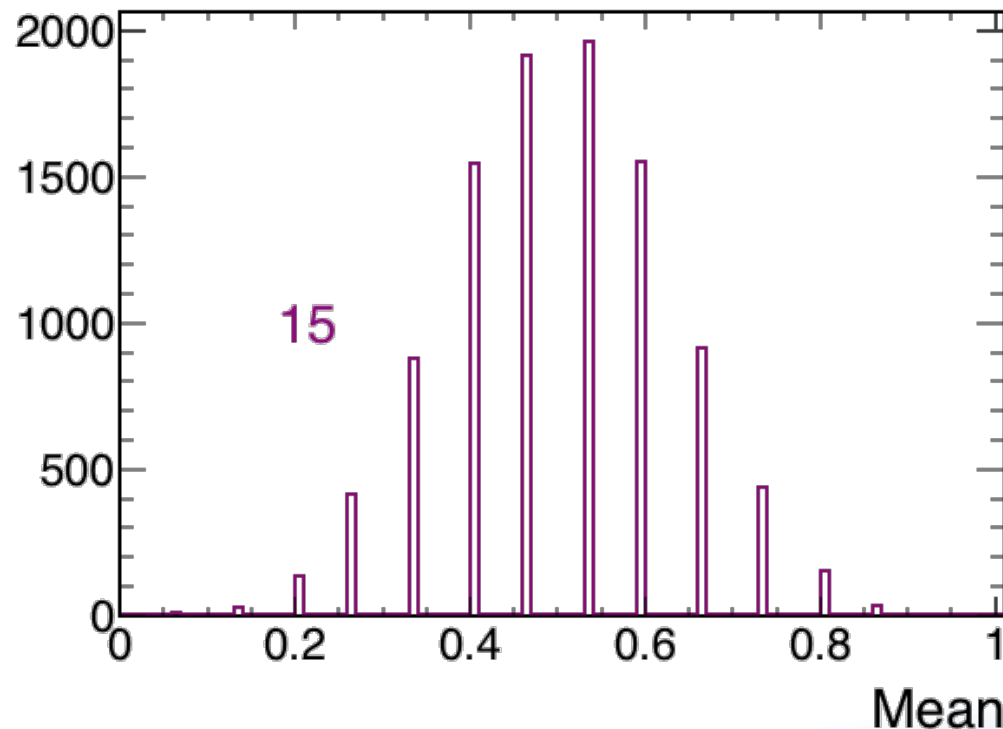
Central Limit Theorem

- The sum of a sufficiently large number of **independent random variables**.
 - It does not matter what distribution the underlying random variables come from.
- Example: coin flips. Heads = 0, tails = 1
 - Clearly not normally distributed.
- However, if we look at the distribution of the means:



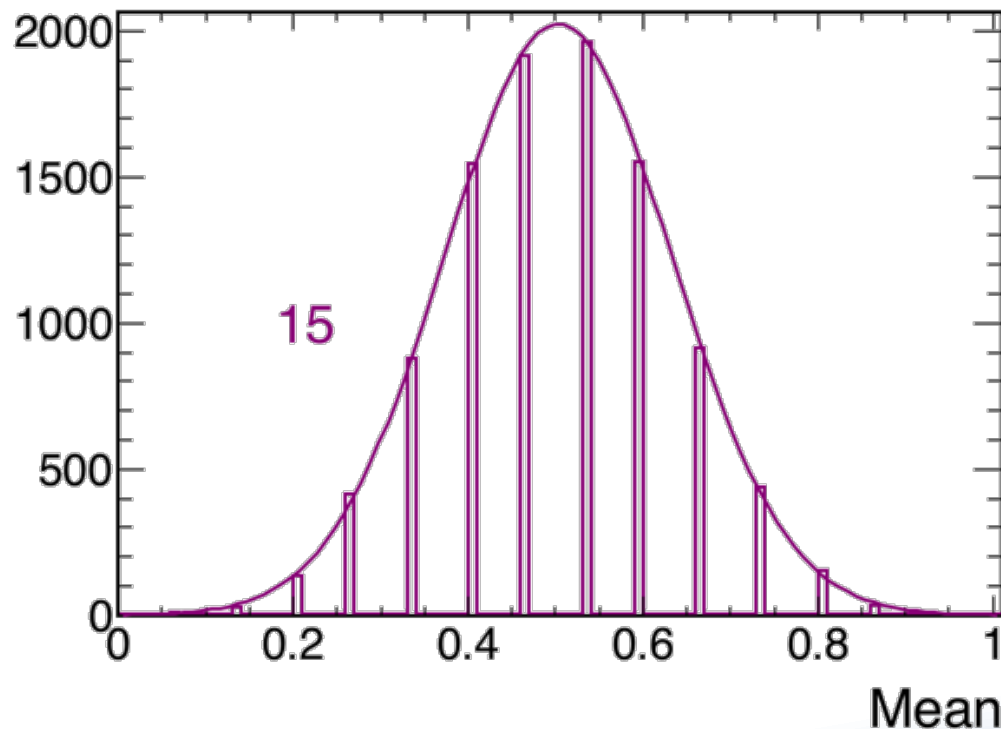
Central Limit Theorem

- The sum of a sufficiently large number of **independent random variables**.
 - It does not matter what distribution the underlying random variables come from.
- Example: coin flips. Heads = 0, tails = 1
 - Clearly not normally distributed.
- However, if we look at the distribution of the means:



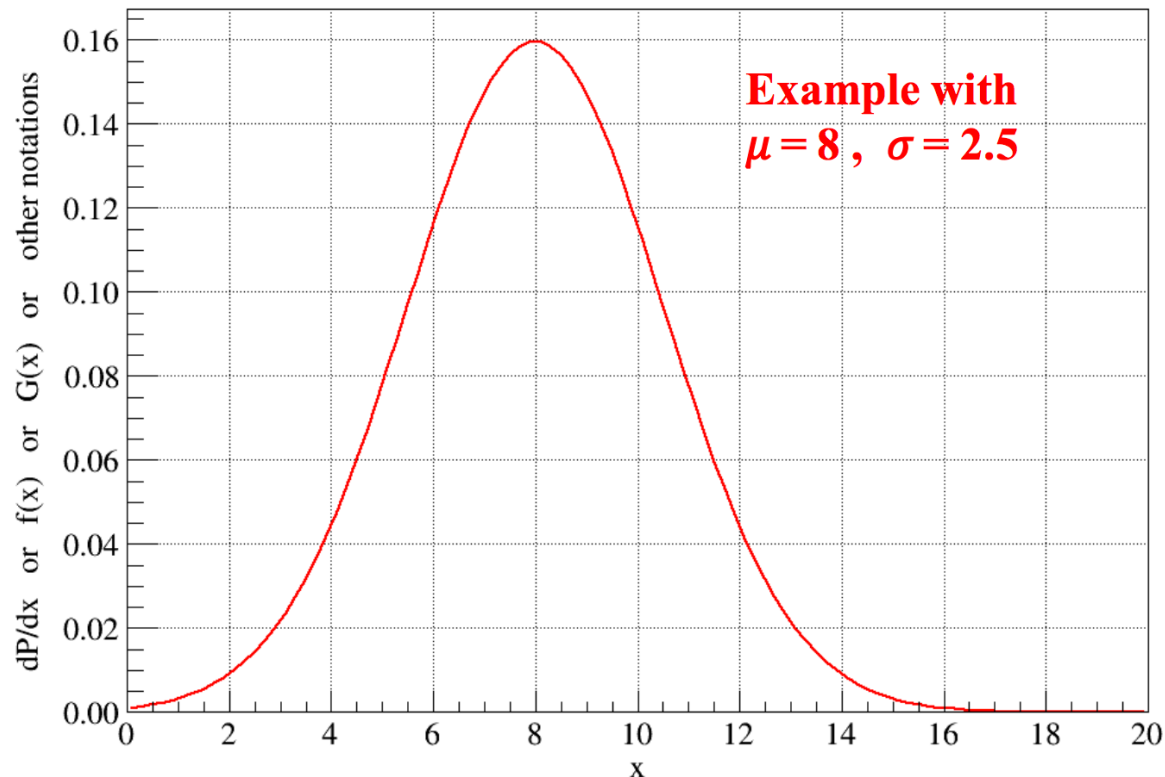
Central Limit Theorem

- The sum of a sufficiently large number of **independent random variables**.
 - It does not matter what distribution the underlying random variables come from.
- Example: coin flips. Heads = 0, tails = 1
 - Clearly not normally distributed.
- However, if we look at the distribution of the means:

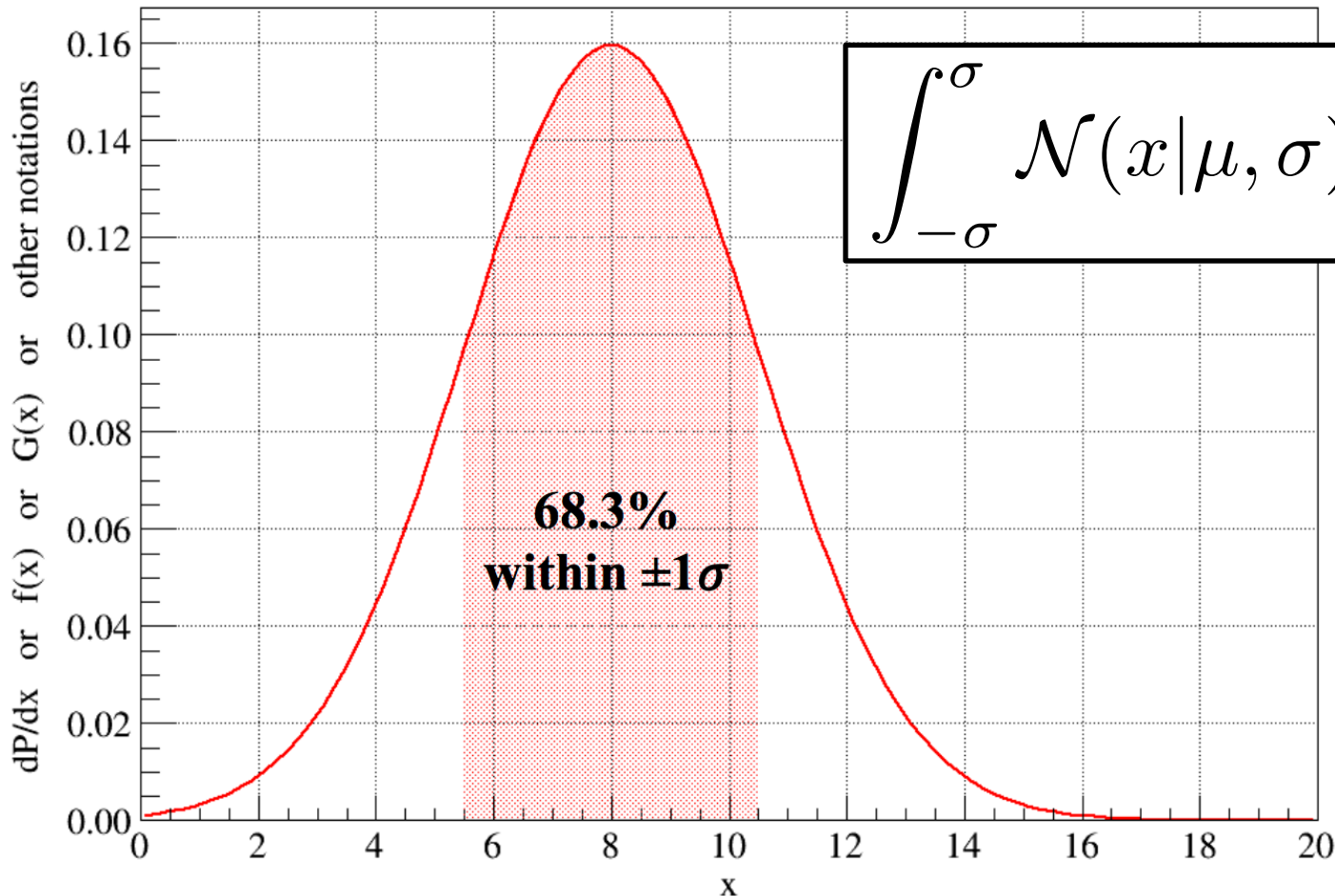


- This is why, under most circumstances, we treat errors as “Gaussian”...because most of the time it works.
- When doesn't it work?
 - Mostly when the stats are too low, plus a few other edge cases.

$$\mathcal{N}(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

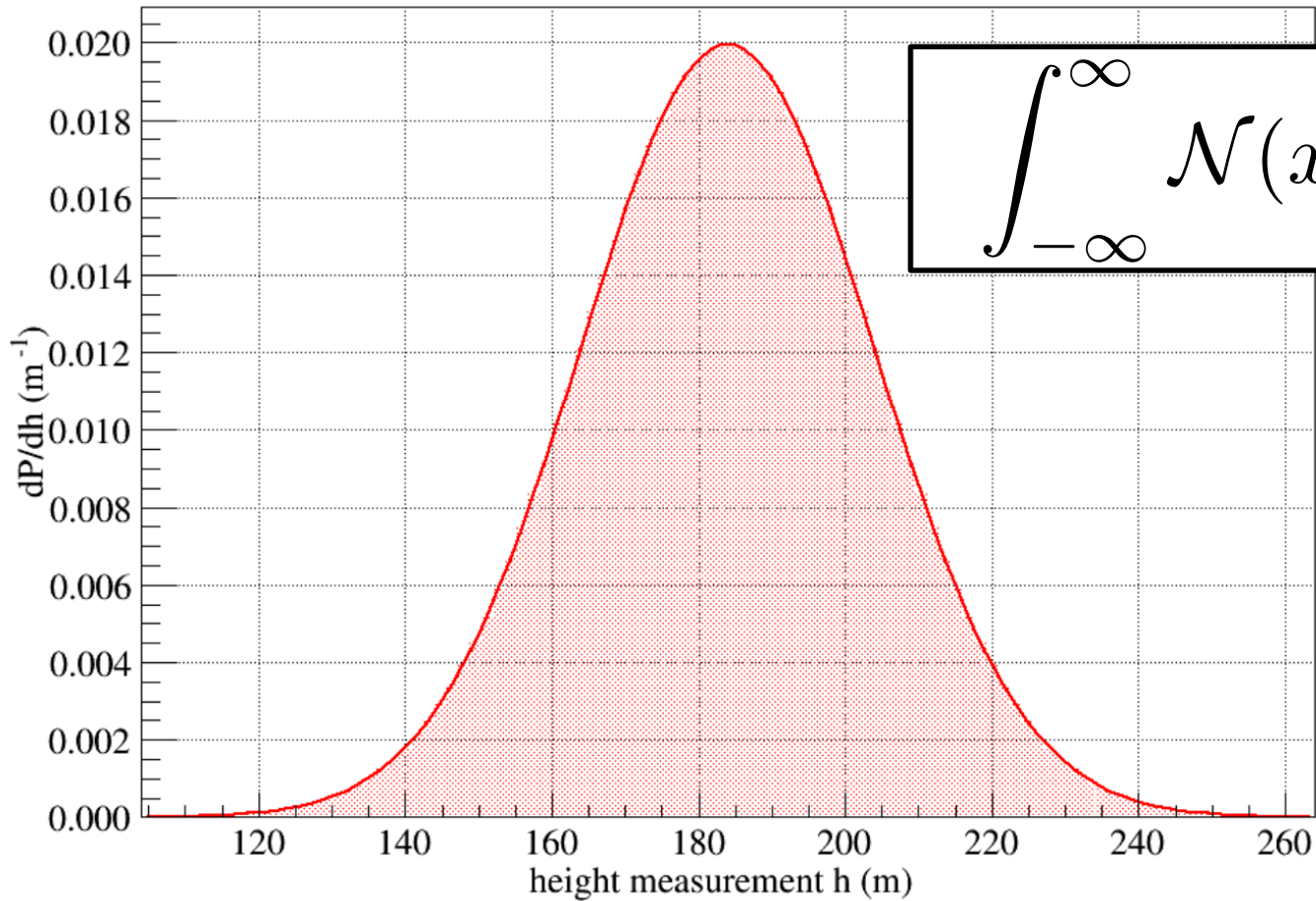


$$\int_a^b \mathcal{N}(x|\mu, \sigma) dx = \text{probability that } x \text{ is between } a \text{ and } b$$



$$\int_{-\sigma}^{\sigma} \mathcal{N}(x|\mu, \sigma) dx = 0.683$$

$$\int_a^b \mathcal{N}(x|\mu, \sigma) dx = \text{probability that } x \text{ is between } a \text{ and } b$$



$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma) dx = 1$$

How to Ask a Statistical Question

- The term for this is a “hypothesis test.”
- H_0 : Null hypothesis
 - The specific case, such as A and B are the same
- H_1 : Alternative hypothesis
 - The alternative to the null – A and B are different
- Significance level
 - How high a rate of false positives (rejecting the null, even if it is true) can you tolerate.
 - $\alpha = 0.05$ is common, but often not sufficient for physics.

Are two means the same?

$$\mu_1 = 2.5 \pm 0.1 \quad \mu_2 = 3.1 \pm 0.3$$

- H_0 : The difference between the means is 0
 - $\mu_2 - \mu_1 = 0$
- H_1 : The means are different
 - $\mu_2 - \mu_1 \neq 0$
- I can tolerate a 5% chance of saying they are different, even if they really are the same.
- Now, let's do the test.

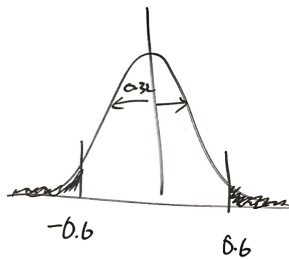
$$X = \mu_2 - \mu_1 = 0.6$$

$$\sigma_{\mu_2 - \mu_1} = \sqrt{\sigma_{\mu_1}^2 + \sigma_{\mu_2}^2} = 0.32$$

Are these different
at $p = 0.05$?

How often do we get 0.6 or more extreme assuming

$$X \sim N(0, 0.32)^2$$



$$X \rightarrow Z = \frac{X - \mu}{\sigma} = \frac{0.6 - 0}{0.32} = 1.88$$

$$Z \sim N(0, 1)$$

$$\int_{-\infty}^{-z} N(x, 1) dx + \int_z^{\infty} N(x, 1) dx < p$$

$$1 - \int_{-z}^z N(x, 1) dx < p$$

$$\text{Erf}(X/\sqrt{2})$$

$$0.06 > 0.05$$

"Fail to reject H_0 "

$$p = 1 - \int_{-Z}^Z \mathcal{N}(x, 1) dx$$

- This integral doesn't have an analytical solution, but we need it all the time, so its results are readily available as the “error function”

```
// Z-score (sigmas) -> p-value  
  
root [4] 1 - TMath::Erf(1.88 / TMath::Sqrt(2))  
(Double_t) 0.0601081
```


A Little Vocabulary

- Z is our **test statistic**
 - A single number we calculate as a “summary” of our data.
- You want to know how the test statistic is supposed to be distributed under the null hypothesis.
 - You need to know the distribution to calculate a p -value.
- Generally, there are **assumptions that must be met** for this to be true.
- If the conditions are not met, or there is no simple test statistic, all is not lost.
 - There are “non-parametric” techniques.

Is there signal above the background?

- Let's say we're members of a neutrino experiment called SUNE
 - The Statistical Underground Neutrino Experiment
- Thanks to our powerful off-axis design we expect only **1 background event**.
 - And since this is SOvA we have no systematic errors!
- We open the box and **observe 6 events**.
- Did we observe ν_e appearance?

- Let's translate into a hypothesis test:

- H_0 : Our observation is consistent with the background.

- $X = B$

- H_1 : There is a signal above our background estimate.

- $X > B$

- We are making an important claim, so we require

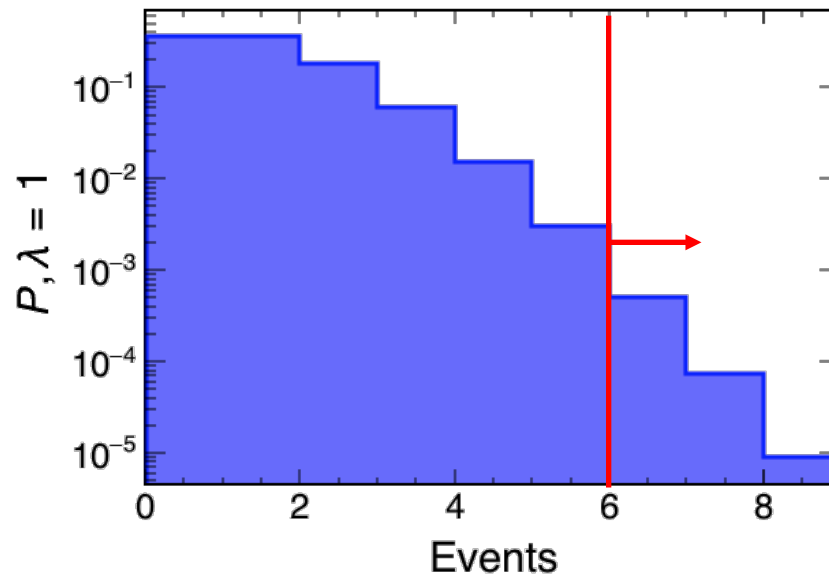
- $\alpha = 0.0027 (3 \sigma)$

```
// p-value -> sigmas  
root [6] TMath::NormQuantile(1 - 0.0027/2)  
(Double_t)3.0
```

- How is this different from the mean test?
 - The numbers involved are *small*.
 - This test is 1-sided instead of 2-sided
 - The distribution is not Gaussian, it is Poisson.
- How do we know it's Poisson?
 - This distribution describes the number of **independent events** (neutrinos in the FD)
 - occurring within a **fixed time interval** (periods 1&2).
 - This almost always describes neutrino physics data.
- But, if you have many events, then the Poisson just becomes...

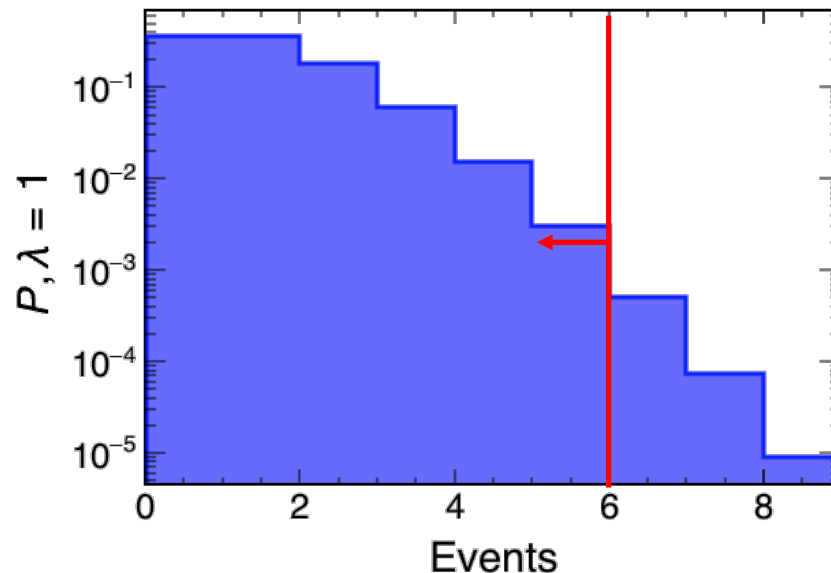
- First – what is our test statistic?
 - Just the number of observed events.
 - We know, under the null hypothesis, how that should be distributed – Poisson, rate 1
- We need to calculate a p -value to compare to our α .
 - To do that, we again need to integrate a distribution.

$$p = \sum_{X}^{\infty} \mathcal{P}(x, 1)$$



- First – what is our test statistic?
 - Just the number of observed events.
 - We know, under the null hypothesis, how that should be distributed – Poisson, rate 1
- We need to calculate a p -value to compare to our α .
 - To do that, we again need to integrate a distribution.

$$p = \sum_{X}^{\infty} \mathcal{P}(x, 1) = 1 - \sum_{0}^{X-1} \mathcal{P}(x, 1)$$



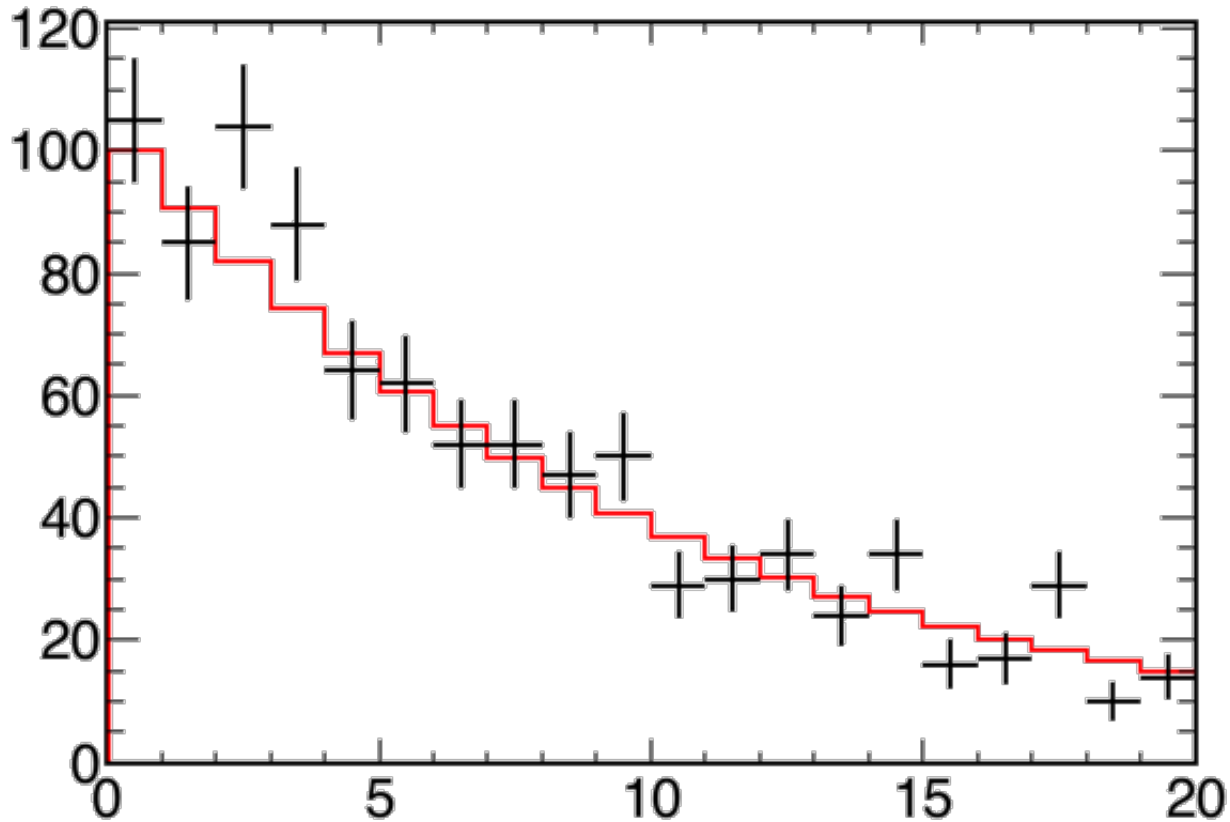
- Again, let's take advantage of built-in functions which already have the integral of the Poisson distribution.

```
root [14] 1 - ROOT::Math::poisson_cdf(5,1)
(double) 0.00059418
```

- $p(0.000594) < \alpha(0.0027)$
 - We reject the null hypothesis.
 - We have evidence of something other than background at the 3σ -level.

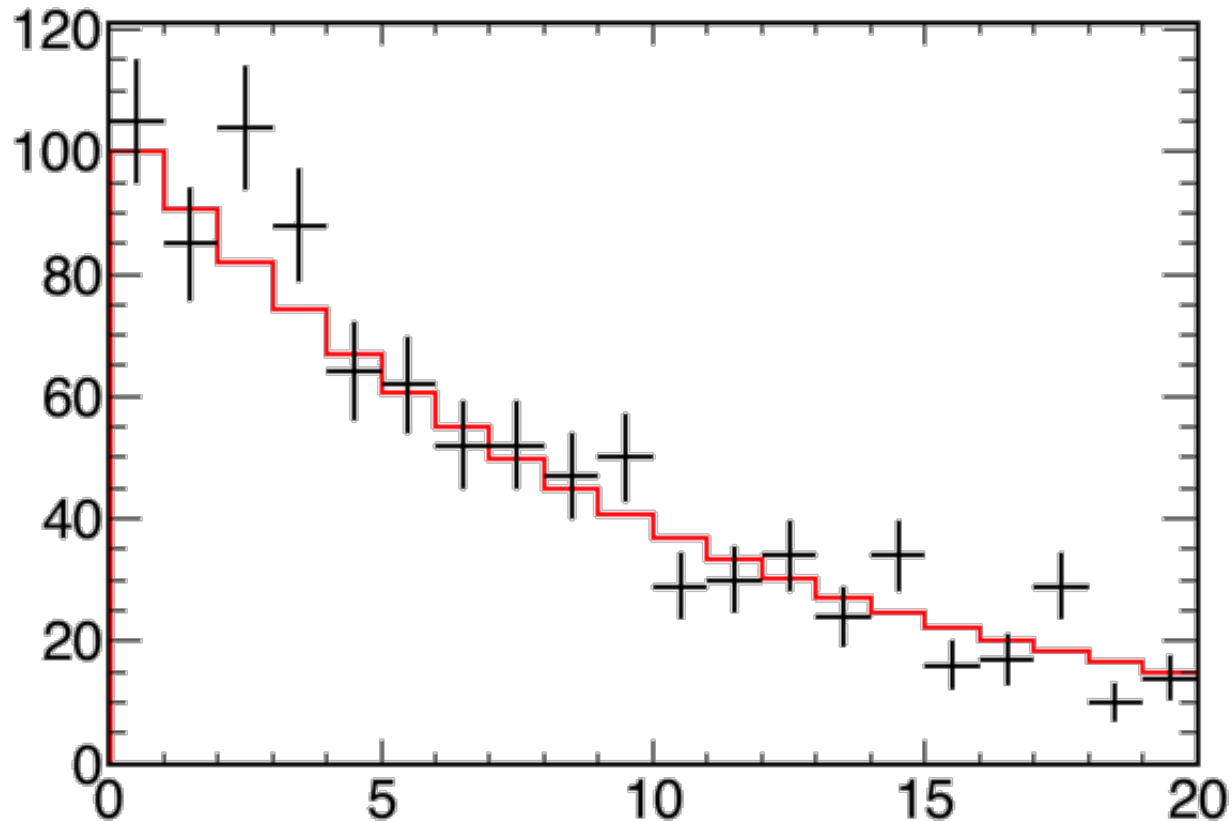
Data/MC Agreement

- Does the model (red) describe the data (black)?



Data/MC Agreement

- Is the data consistent with having been drawn from the model, given its uncertainties?

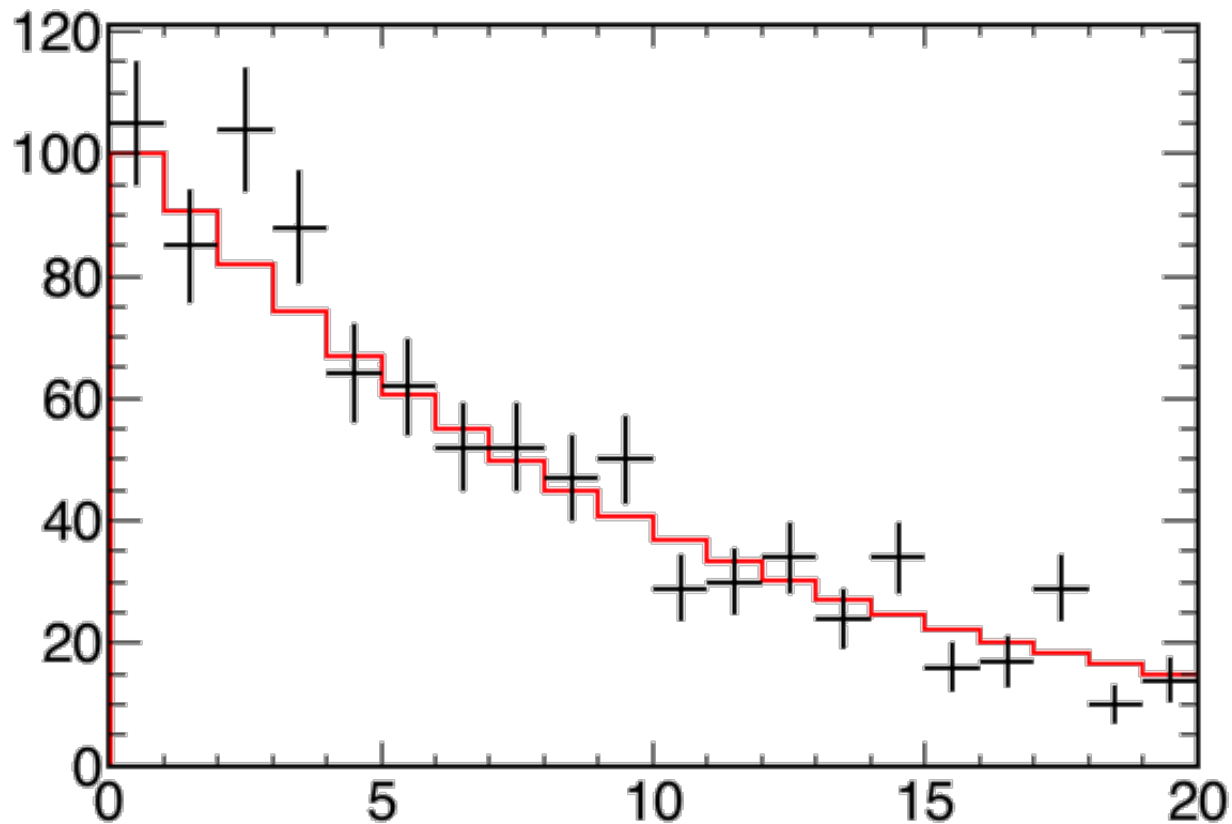


Data/MC Agreement

- Hypothesis test:

- H_0 : The data was drawn from the model in red.
- H_1 : The data is not consistent with the model.

- $\alpha = 0.05$



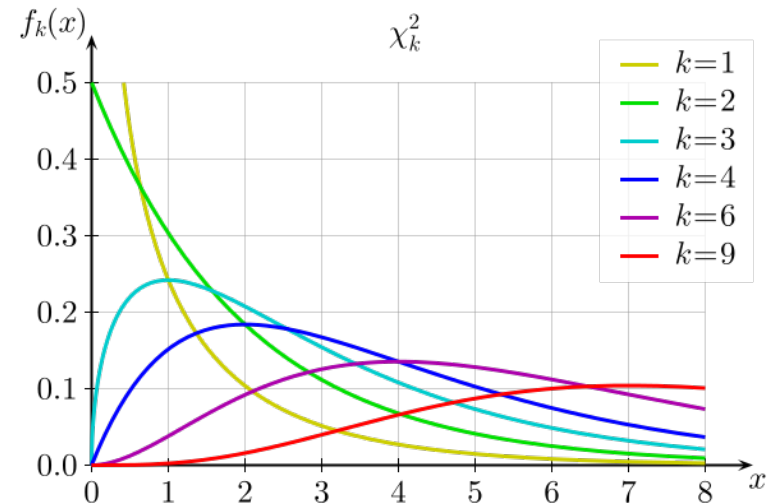
$$T = \sum_i^N \frac{(O_i - E_i)^2}{\sigma^2} \rightarrow \sum^N z^2 \sim \chi^2(N)$$

$$x \rightarrow O_i \sim N(E_i, \sigma)$$

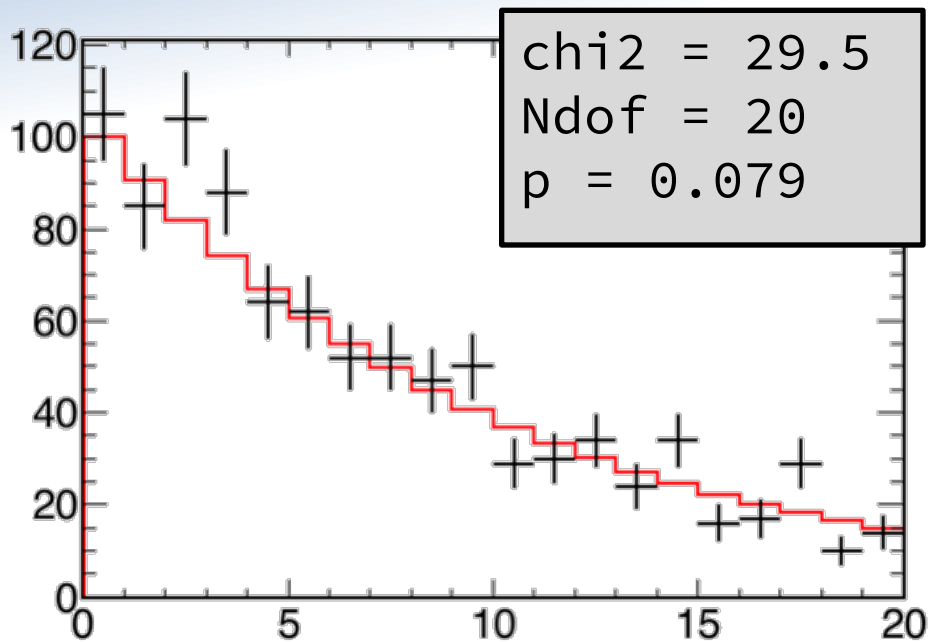
$$\frac{x - \mu}{\sigma} \rightarrow z$$

- This means that, assuming the null is true, we know what T 's distribution should be: the chisquared.

$$T = \sum^N Z^2 \sim \chi^2(N)$$



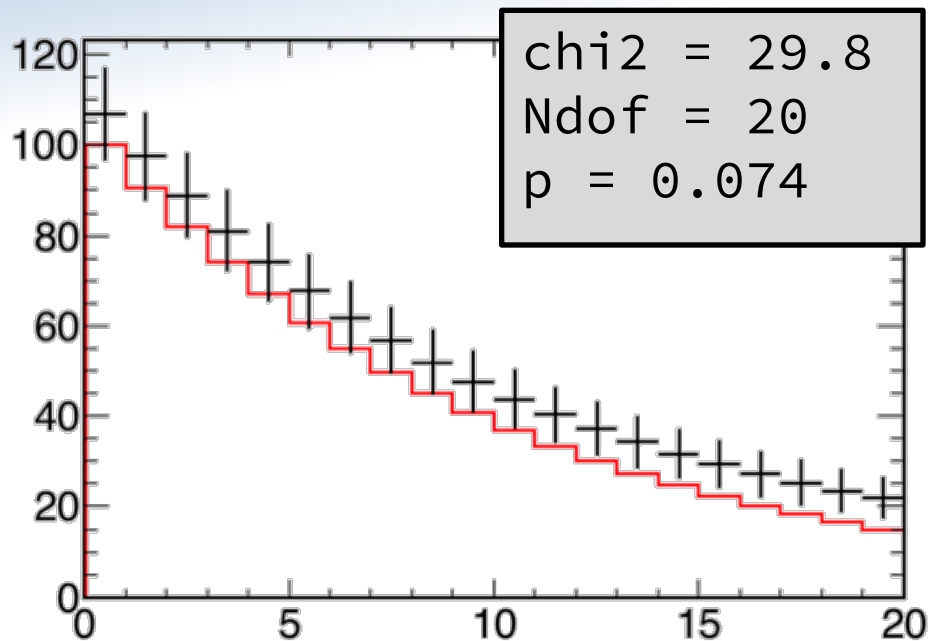
- This means that we can calculate T for our histograms, and then look up that value in this distribution to get a p -value.
 - Note that χ^2 depends on the number of “degrees of freedom”
 - For a histogram, N dof = number of bins.



```
def CalcChi2(hmc, hdata):
    chi2 = 0
    for i in range(1, hmc.GetNbinsX()+1):
        ei = hmc.GetBinContent(i)
        oi = hdata.GetBinContent(i)
        sigma = sqrt(ei)
        chi2 += (ei - oi)**2 / sigma**2
    return chi2

chi2 = CalcChi2(hpred, hrand)
Ndof = hpred.GetNbinsX()
p = TMath.Prob(chi2, Ndof)
```

- With p of 0.08, we fail to reject the H_0 .

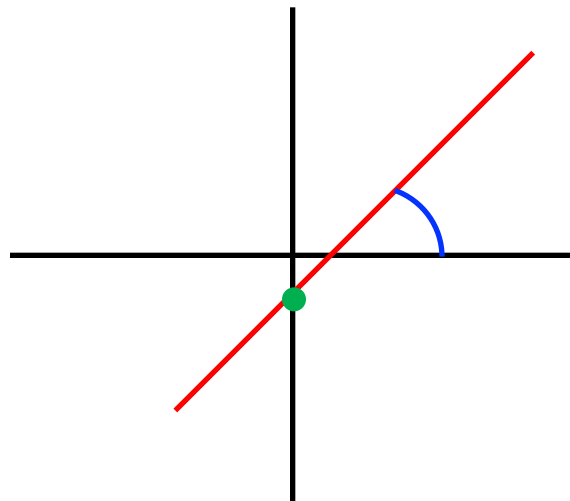


```
def CalcChi2(hmc, hdata):  
    chi2 = 0  
    for i in range(1, hmc.GetNbinsX()+1):  
        ei = hmc.GetBinContent(i)  
        oi = hdata.GetBinContent(i)  
        sigma = sqrt(ei)  
        chi2 += (ei - oi)**2 / sigma**2  
    return chi2  
  
chi2 = CalcChi2(hpred, hrand)  
Ndof = hpred.GetNbinsX()  
p = TMath.Prob(chi2, Ndof)
```

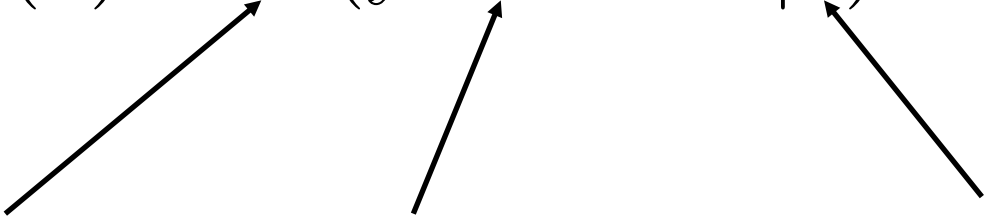
- Statistical tests are not a substitute for looking at the data!
- The results from a test are piece of the argument – they are not an answer themselves.

Parameter Estimation

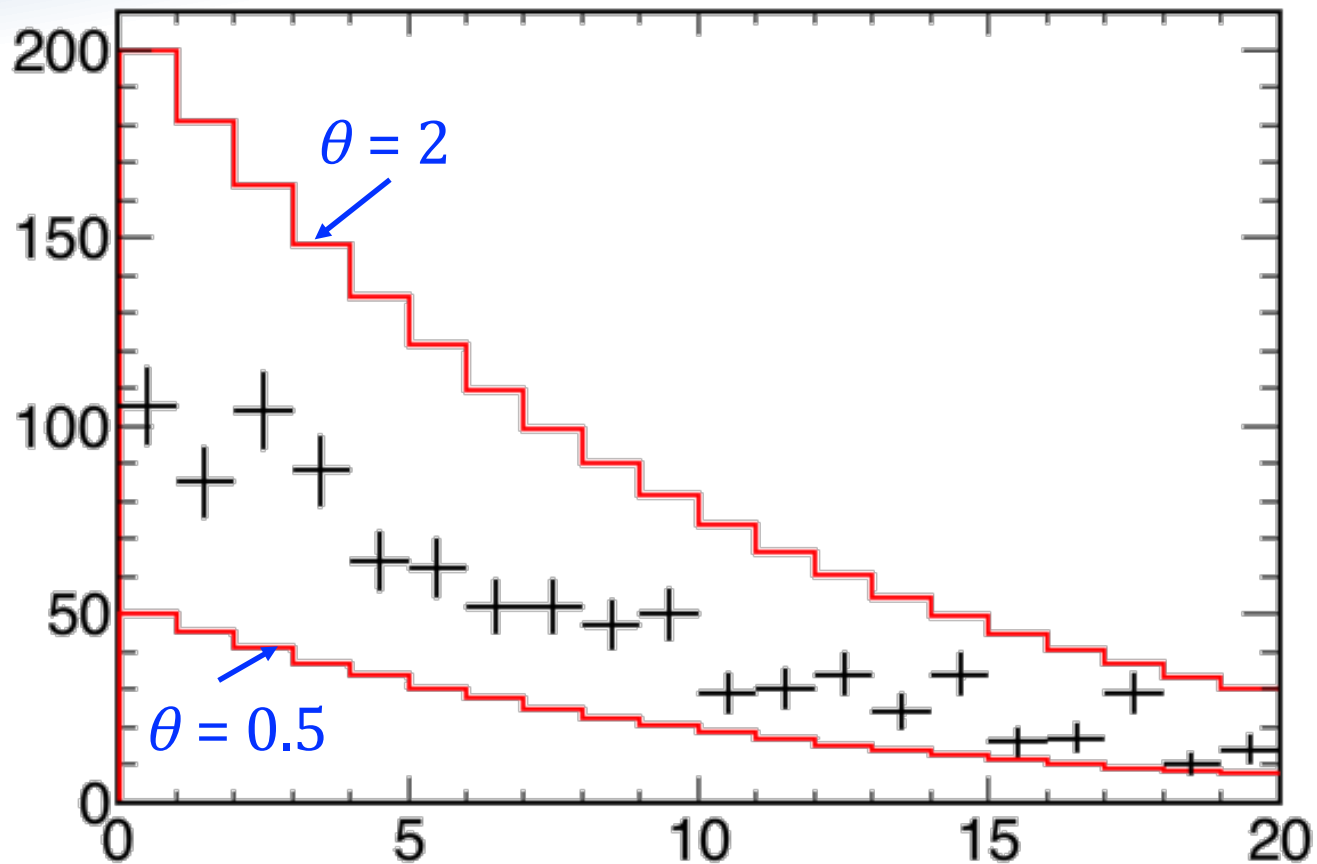
- Up until now, we've been asking yes-or-no questions.
- Often, what we want is to measure a value – this is **parameter estimation**.
 - Also sometimes called “regression”
 - In addition to data, this requires a model.
 - The parameters are the values which describe that model.
 - For example, a line is described by its **slope** and **y-intercept**.



- So, how do we estimate parameters given a model and data?
- We use a method called **maximum likelihood**
 - The key to which is the likelihood function:

$$\mathcal{L}(\vec{\theta}) = P(\text{your data} | \vec{\theta})$$
The diagram consists of three black arrows pointing upwards from the text below to the equation above. One arrow points from the word 'probability' to the function symbol \mathcal{L} . Another arrow points from the word 'data' to the probability symbol P . A third arrow points from the word 'parameters' to the parameter vector $\vec{\theta}$.

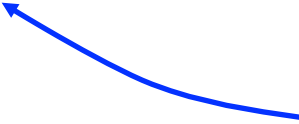
- The probability of your data assuming these parameters are true.



- Let's extend a familiar example.
- Now, we have a model, with a single parameter θ .

- Now, we need a likelihood function.
- To start, let's assume Gaussian errors.

$$\mathcal{L} = P(\vec{O}|\theta) = \prod^N e^{-(O_i - E_i(\theta))^2 / \sigma^2}$$



P of each bin,
assuming each is a
normal distribution.

- Now, we need a likelihood function.
- To start, let's assume Gaussian errors.

$$\mathcal{L} = P(\vec{O}|\theta) = \prod^N e^{-(O_i - E_i(\theta))^2 / \sigma^2}$$

P of each bin,
assuming each is a
normal distribution.

- In practice, instead of maximizing likelihood, we minimize $-2 \ln L$
 - Because addition is easier than multiplication.

$$-2 \ln \mathcal{L}(\theta) = \sum^N \frac{(O_i - E_i(\theta))^2}{\sigma^2}$$

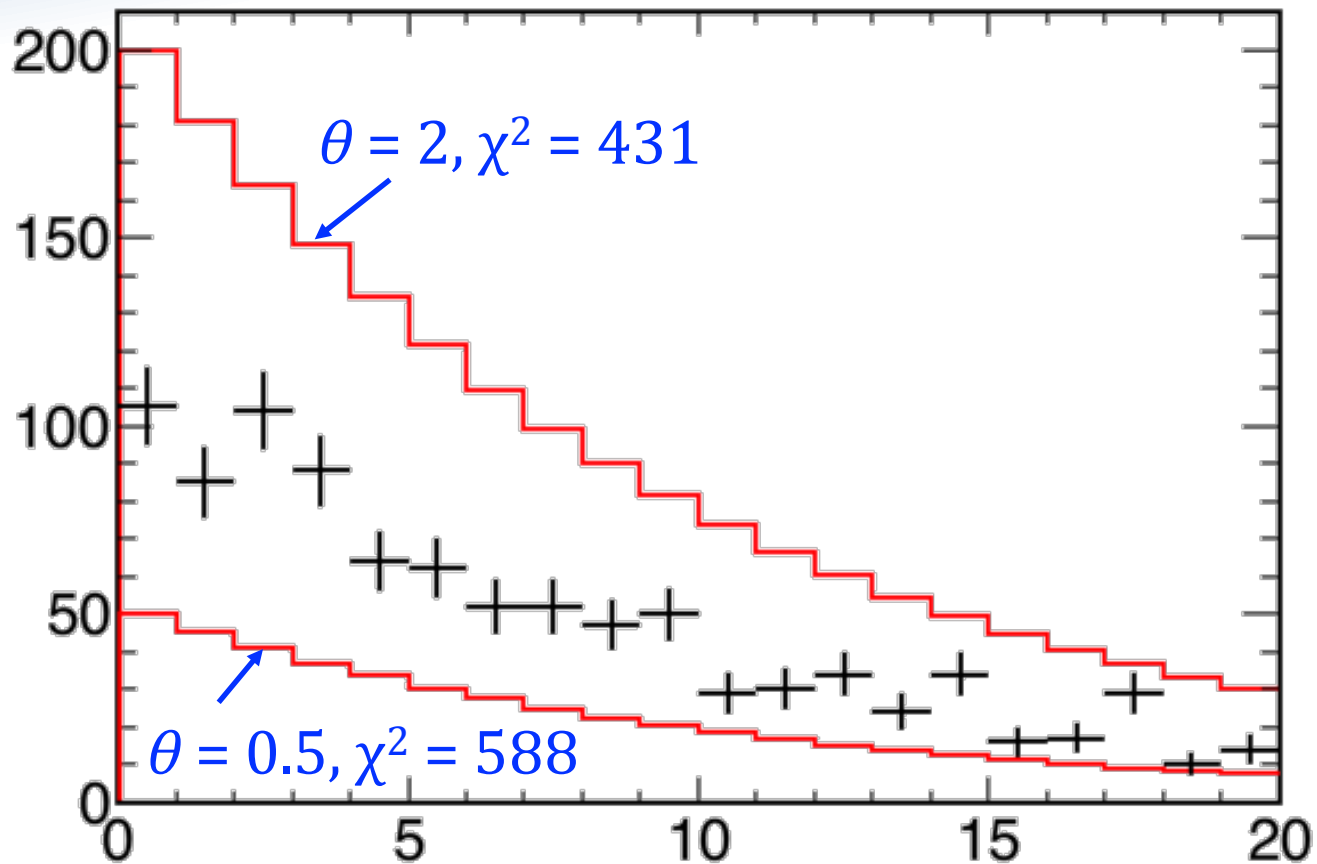
- Now, we need a likelihood function.
- To start, let's assume Gaussian errors.

$$\mathcal{L} = P(\vec{O}|\theta) = \prod^N e^{-(O_i - E_i(\theta))^2 / \sigma^2}$$

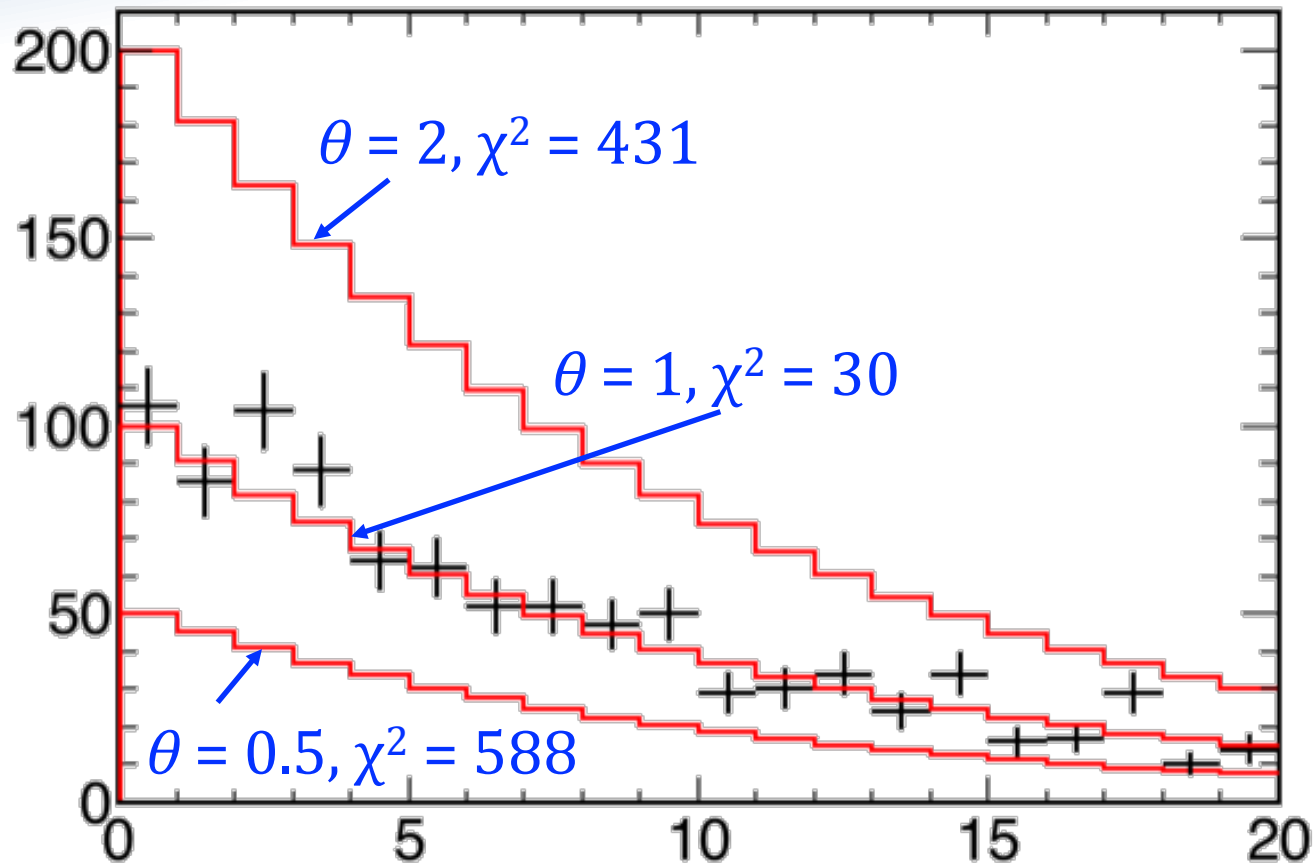
P of each bin,
assuming each is a
normal distribution.

- In practice, instead of maximizing likelihood, we minimize $-2 \ln L$
 - Because addition is easier than multiplication.

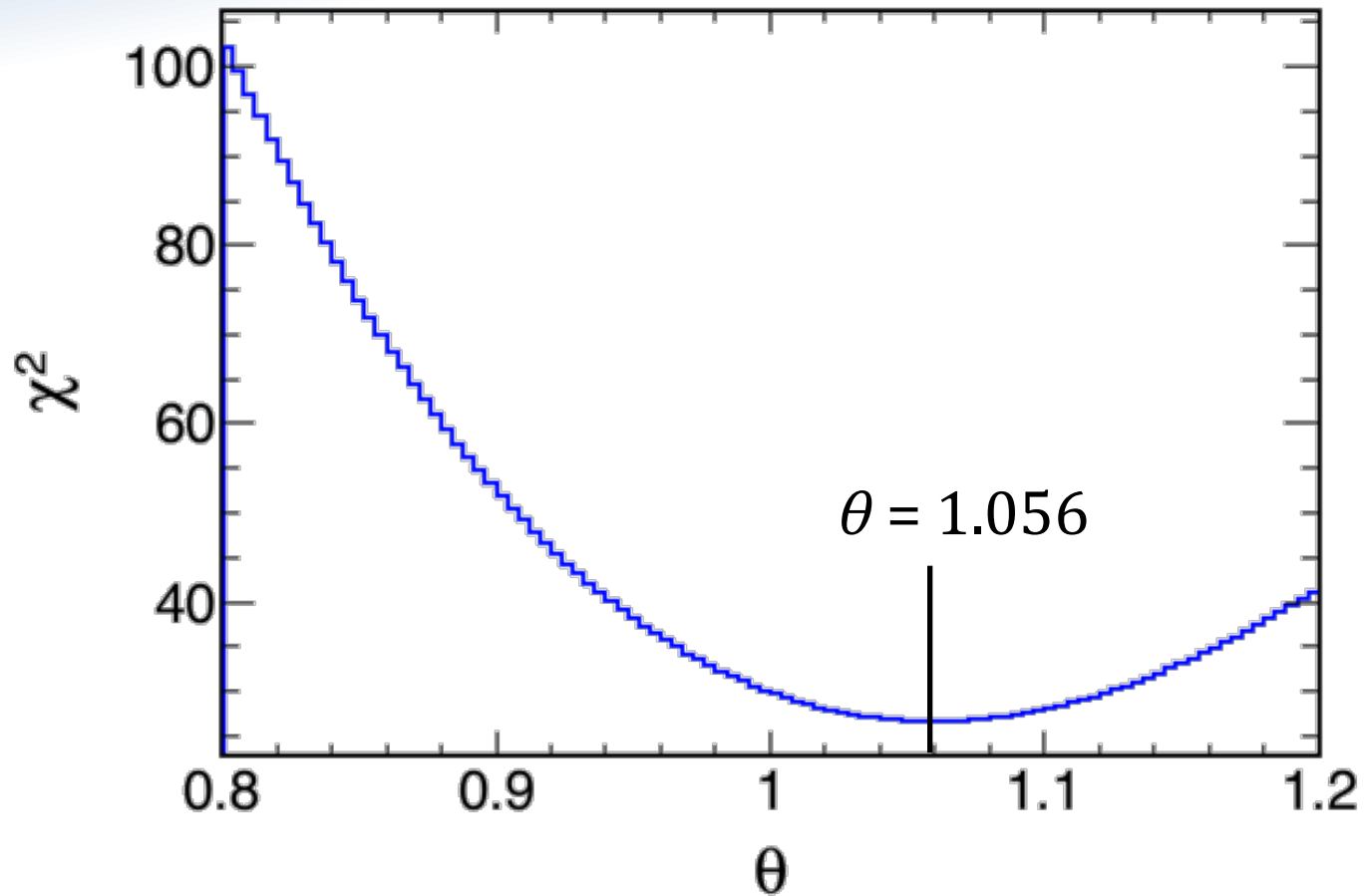
$$-2 \ln \mathcal{L}(\theta) = \sum^N \frac{(O_i - E_i(\theta))^2}{\sigma^2} = \chi^2$$



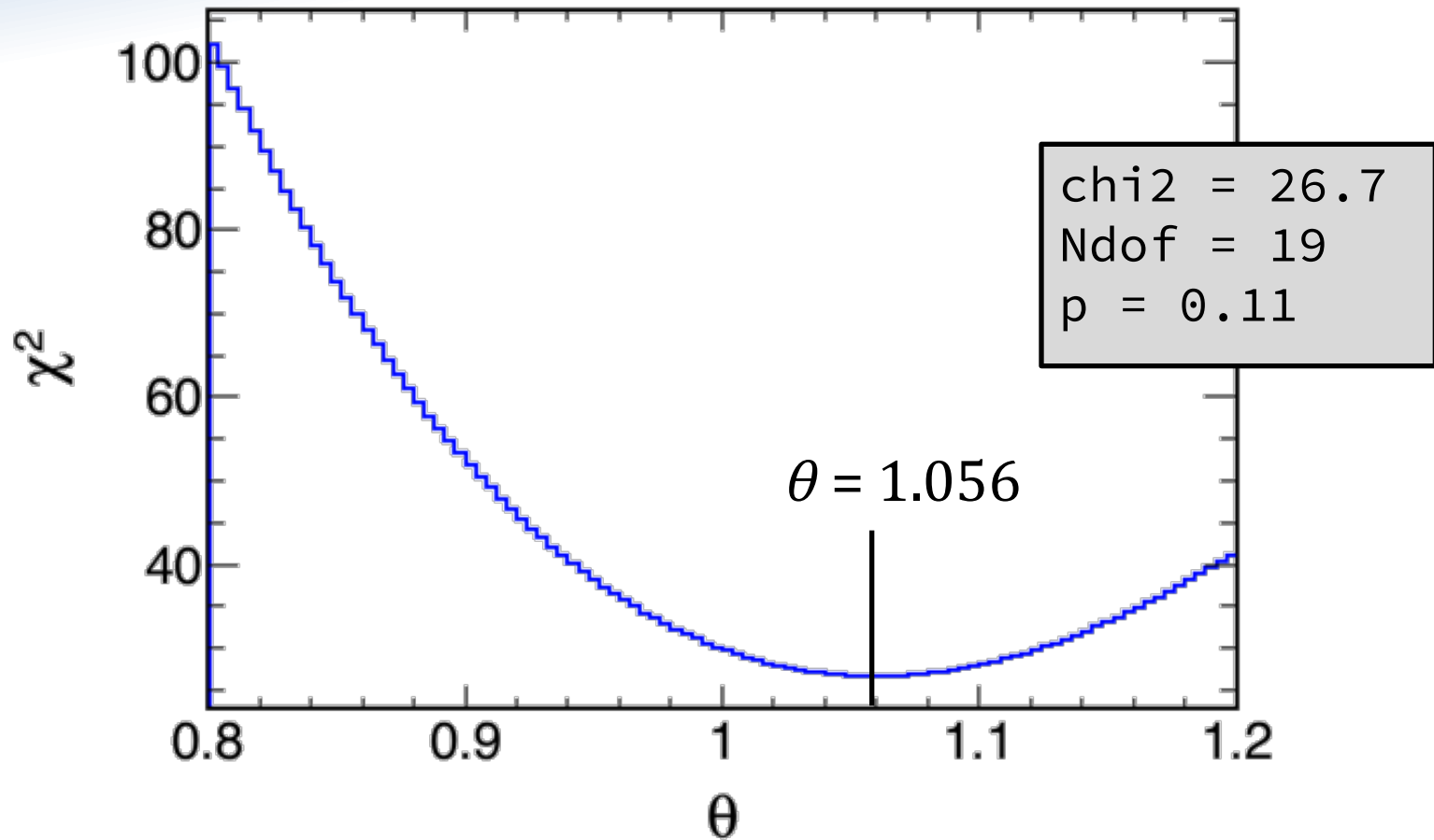
- We can then calculate χ^2 for each possible value of θ .
 - Both of these are pretty bad.



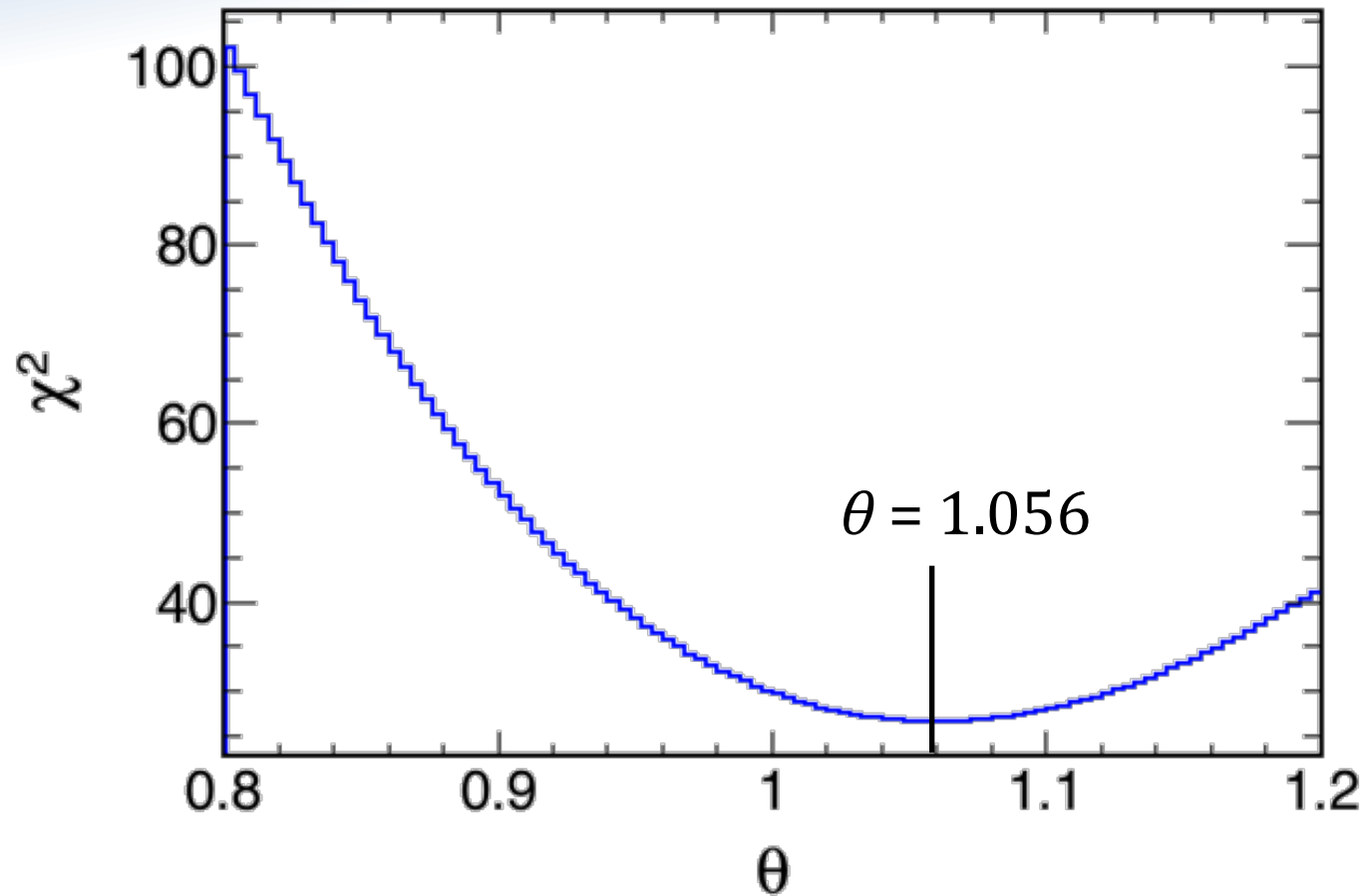
- We can then calculate χ^2 for each possible value of θ .
 - But 30 is pretty good.



- We find the minimum χ^2 (maximum L) when $\theta = 1.054$
- This is our *maximum likelihood estimate*, or “best fit”



- We can also ask, “how good a fit is this?”
 - Is this a reasonable model of this data?
- That is just the hypothesis test we did before.
 - But – you need to subtract 1 for each free parameter in the fit,



- An even better question – what is our uncertainty on our estimate?

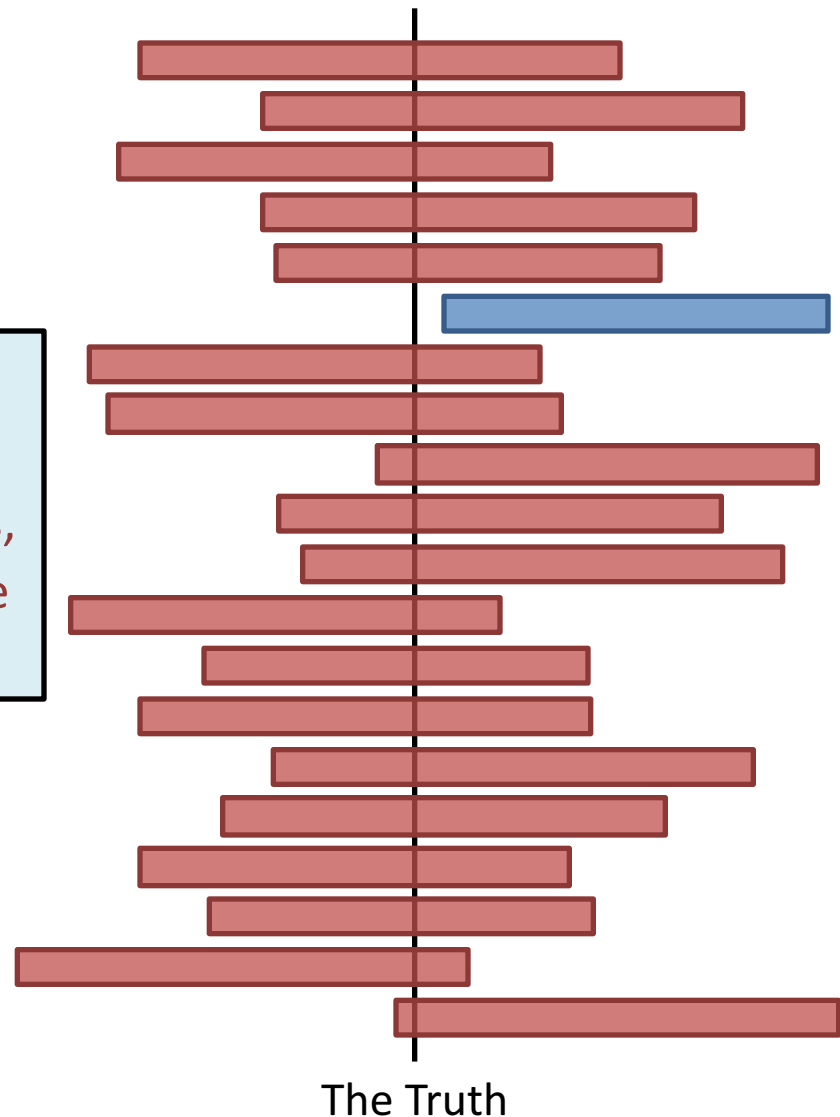
Building Confidence Intervals

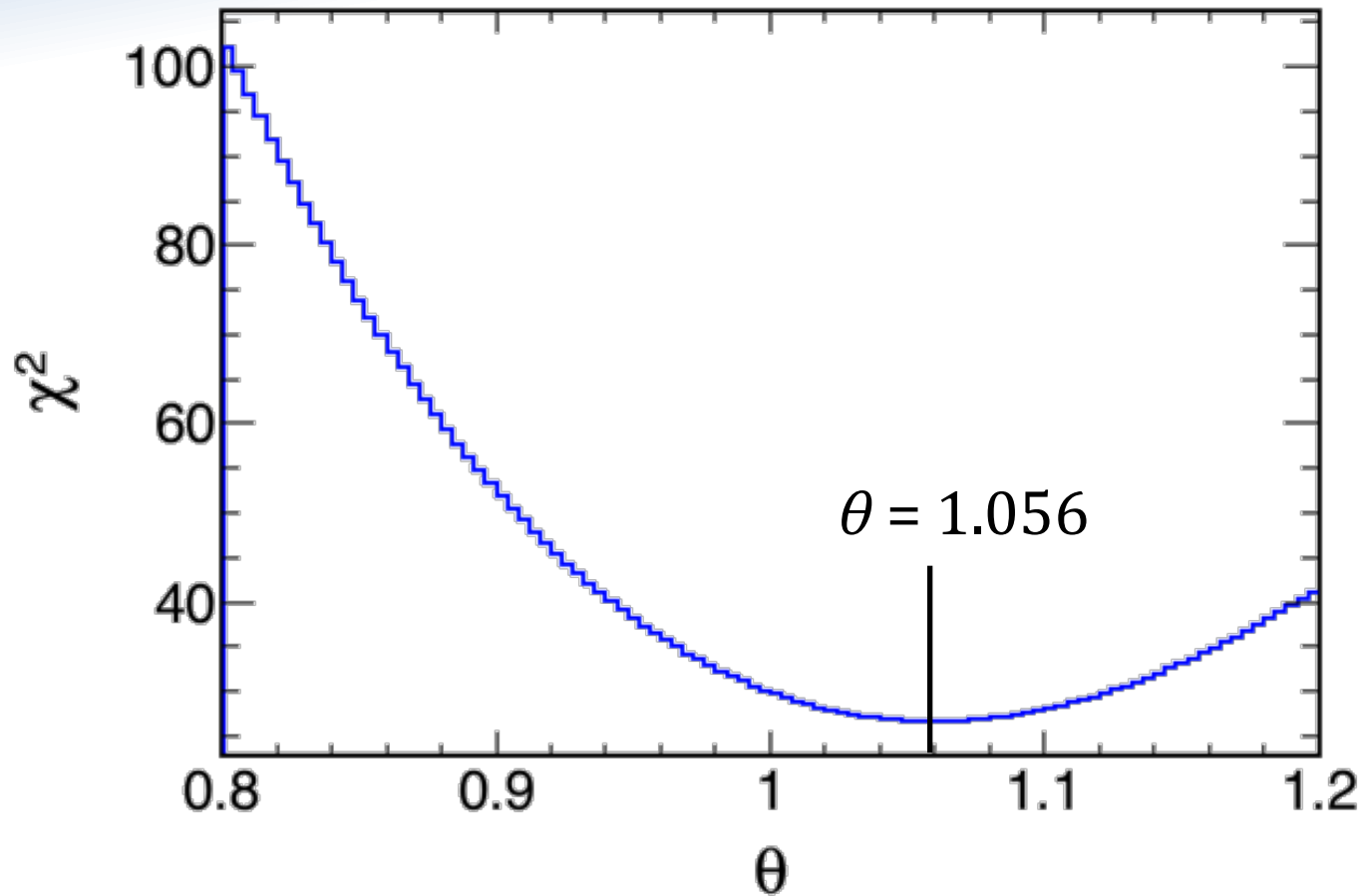
- Here we'll discuss “frequentist” confidence intervals, because that's what you will most often see.

Definition of an Confidence Interval
at level α :

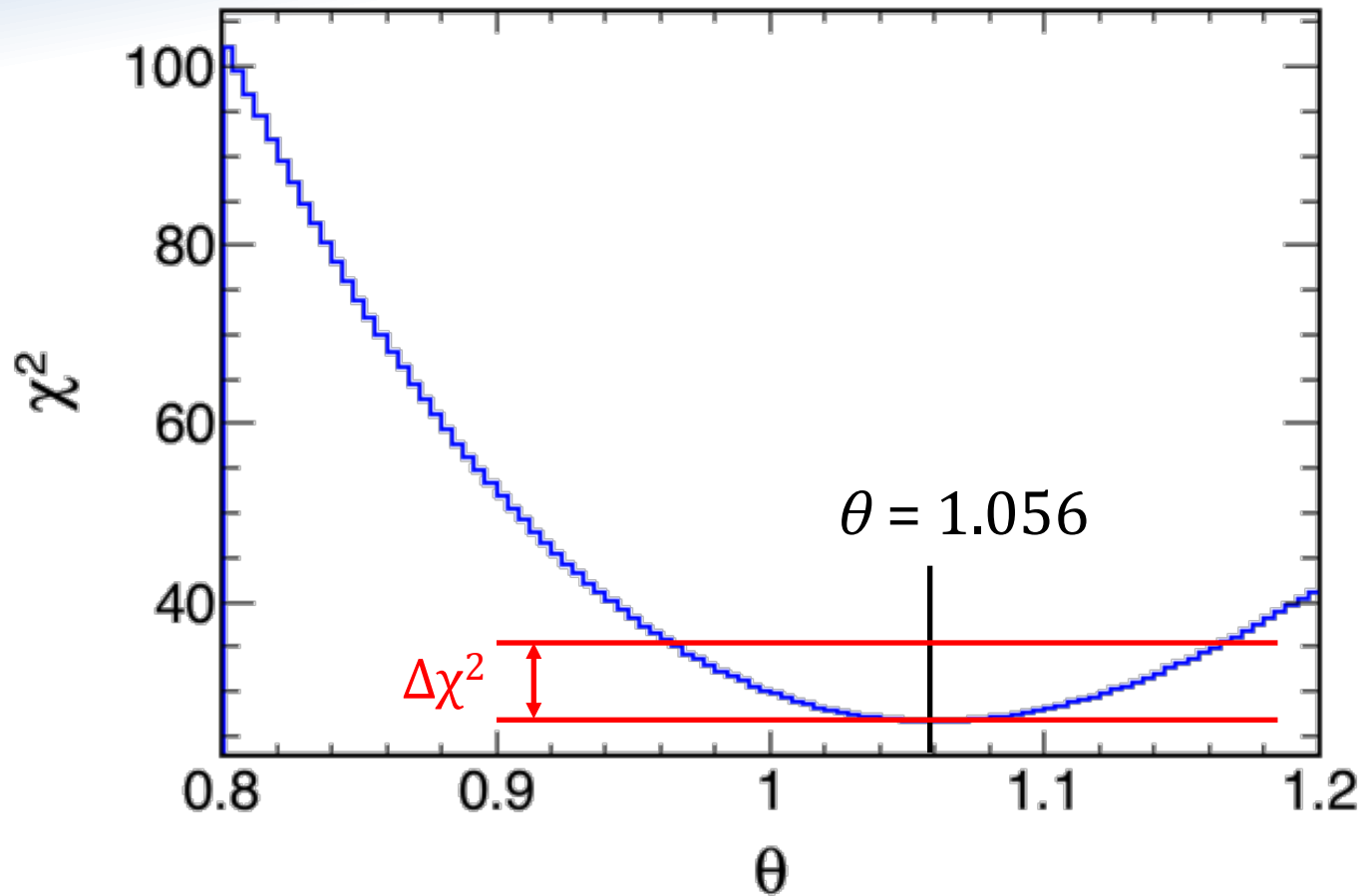
If we repeat the experiment numerous times,
 α of the intervals we draw will cover the true
value.

- This isn't really what you wanted to know, but it has been rigorously defined.
- There are many ways to construct CI's depending on the circumstance.

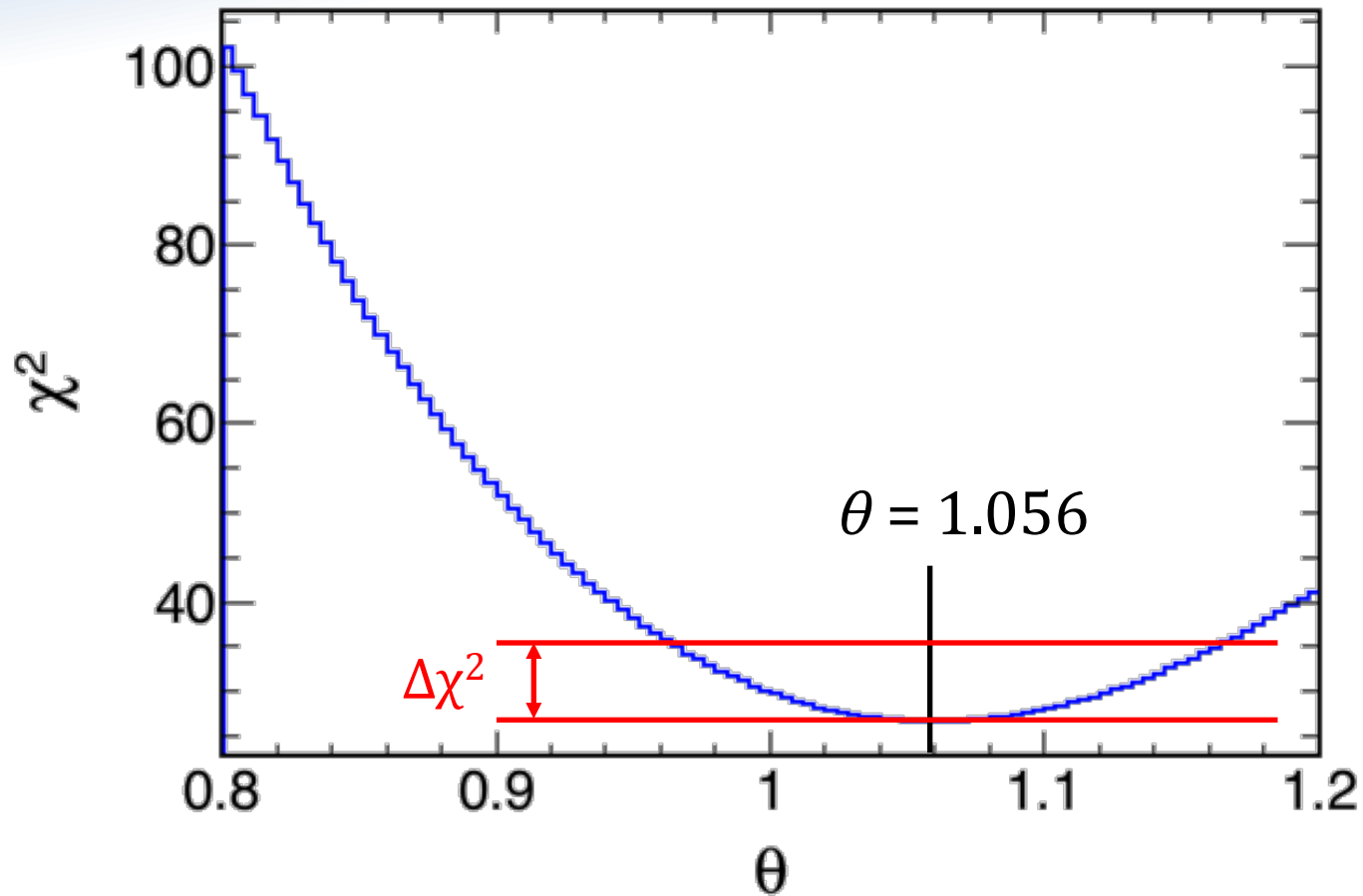




- If your problem has all Gaussian errors, then the distribution of the estimator of the parameter is *also* Gaussian.
 - Presented without proof, since that's what the PDG does, too.
 - This is the case for our example, too.



- We will use the likelihood distribution to draw the CI.
- We allow inside our CI any values of θ with small values $\Delta\chi^2$ relative to the best fit, and we exclude values of θ with larger values of $\Delta\chi^2$.



- The question you should be asking:
- How do I know what “up value” to choose to know which θ 's are in and which are out?

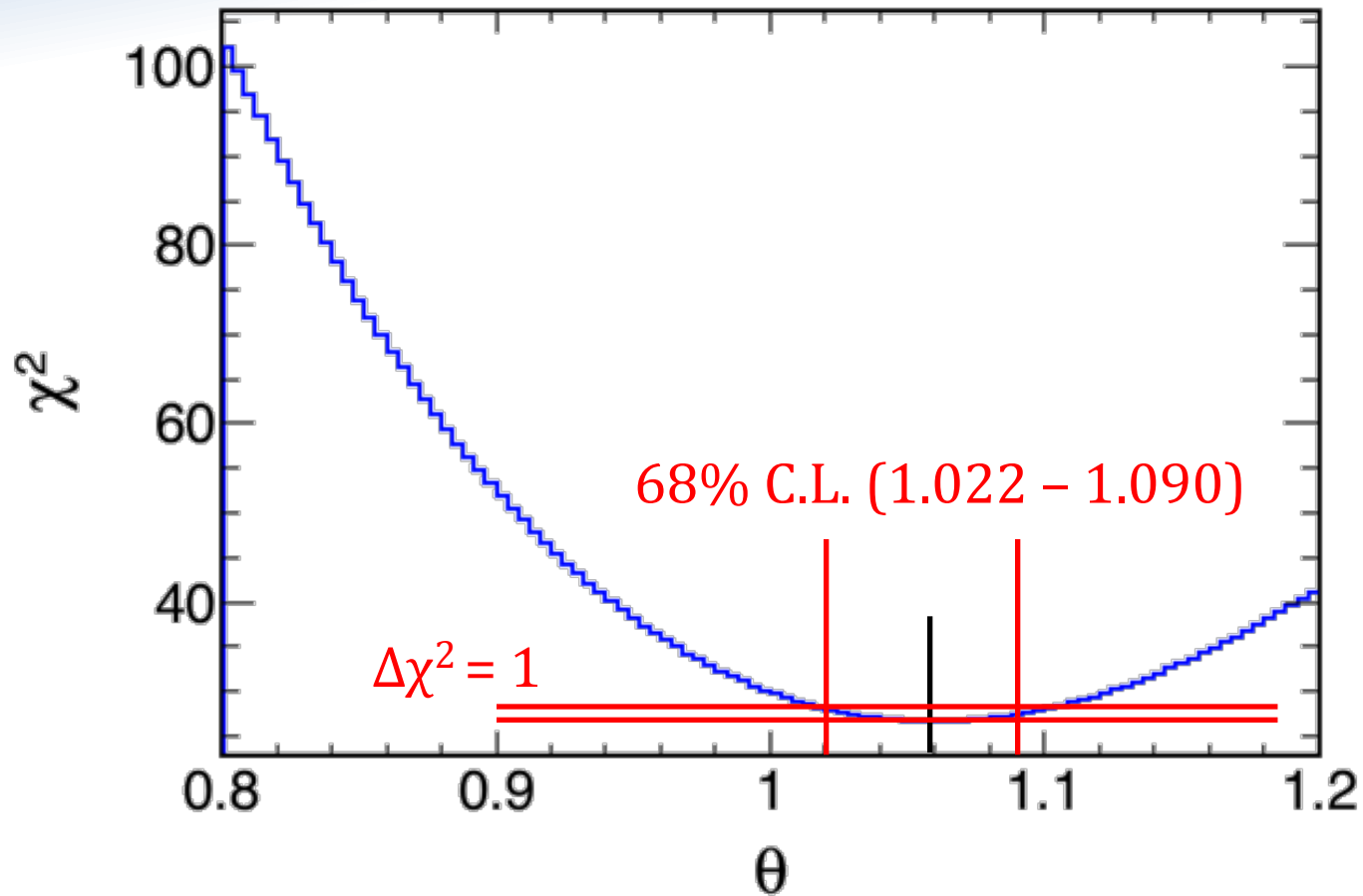
- Here is where we take advantage of everything being Gaussian.
- As with the hypothesis tests, we know what distribution $\Delta\chi^2$ should have, so we can look it up.
- This table comes from the PDG:

Table 37.2: Values of $\Delta\chi^2$ or $2\Delta\ln L$ corresponding to a coverage probability $1 - \alpha$ in the large data sample limit, for joint estimation of m parameters.

$(1 - \alpha)$ (%)	$m = 1$	$m = 2$	$m = 3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

The level of the CI you want to draw.

The number of dimensions.



- This 68% (e.g. 1σ) C.L. is what we generally report as an error band.
- So, in Stats-ese: ML Estimate 1.056 with 68% CL 1.022-1.090
- In Physic-ese: 1.056 ± 0.034

Choose your own adventure...

1. Confidence intervals in the real world
2. Multiple Trials
3. Nuisance Parameters
4. Frequentist vs. Bayesian Statistics
5. Feldman-Cousins

A little more realism

- Choice of likelihood function
 - It's rare in neutrino physics that we have so much data that χ^2 is valid.
 - Instead, we use an L which is based on bins with Poisson errors.

$$-2 \ln \mathcal{L}(\theta) = \sum^N \frac{(O_i - E_i(\theta))^2}{\sigma^2} = \chi^2$$

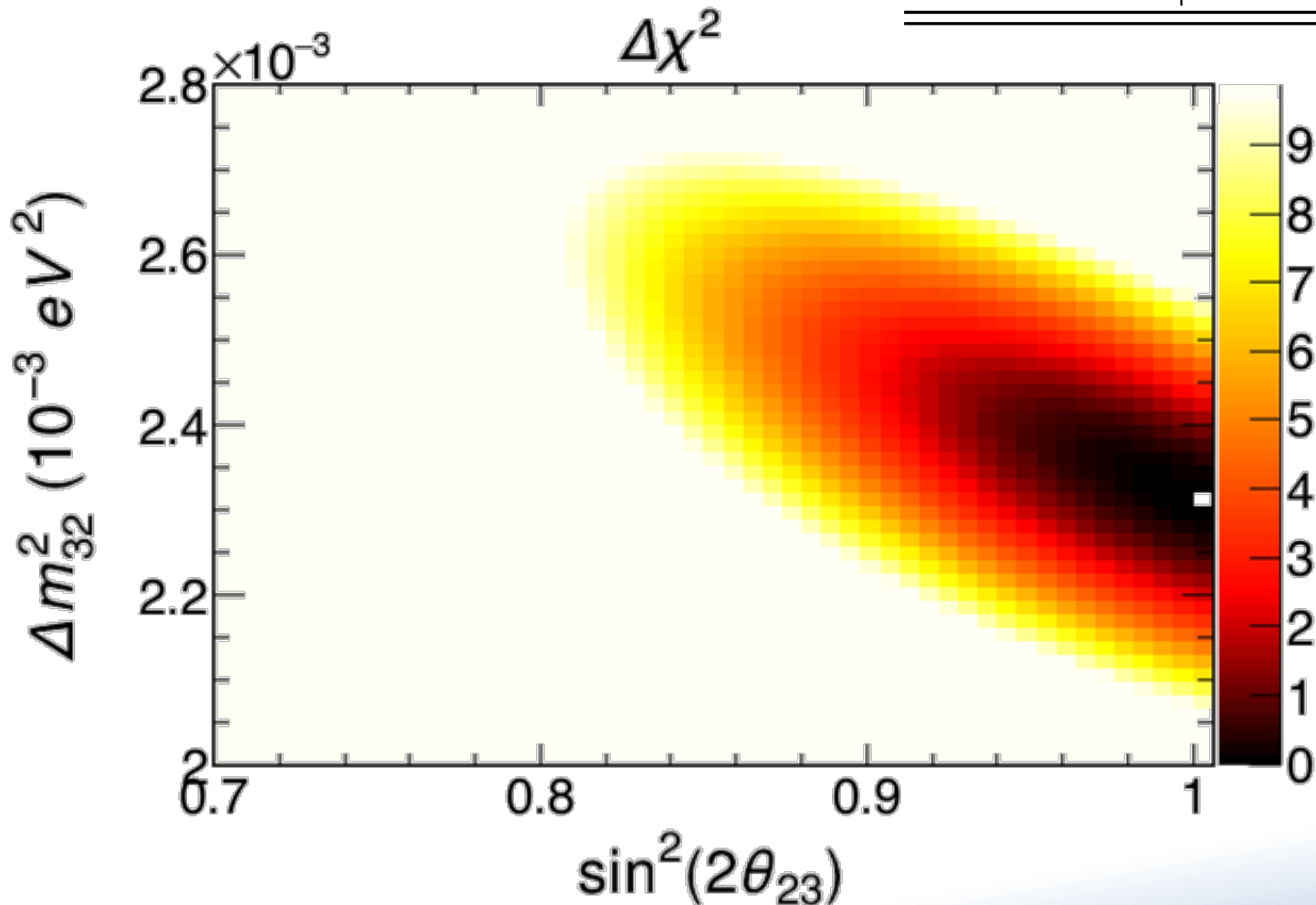


$$-2 \ln \lambda(\theta) = 2 \sum_{i=1}^N \left[\mu_i(\theta) - n_i + n_i \ln \frac{n_i}{\mu_i(\theta)} \right],$$

If you have bins with < 30 entries, you probably need this. Just look it up in the PDG.

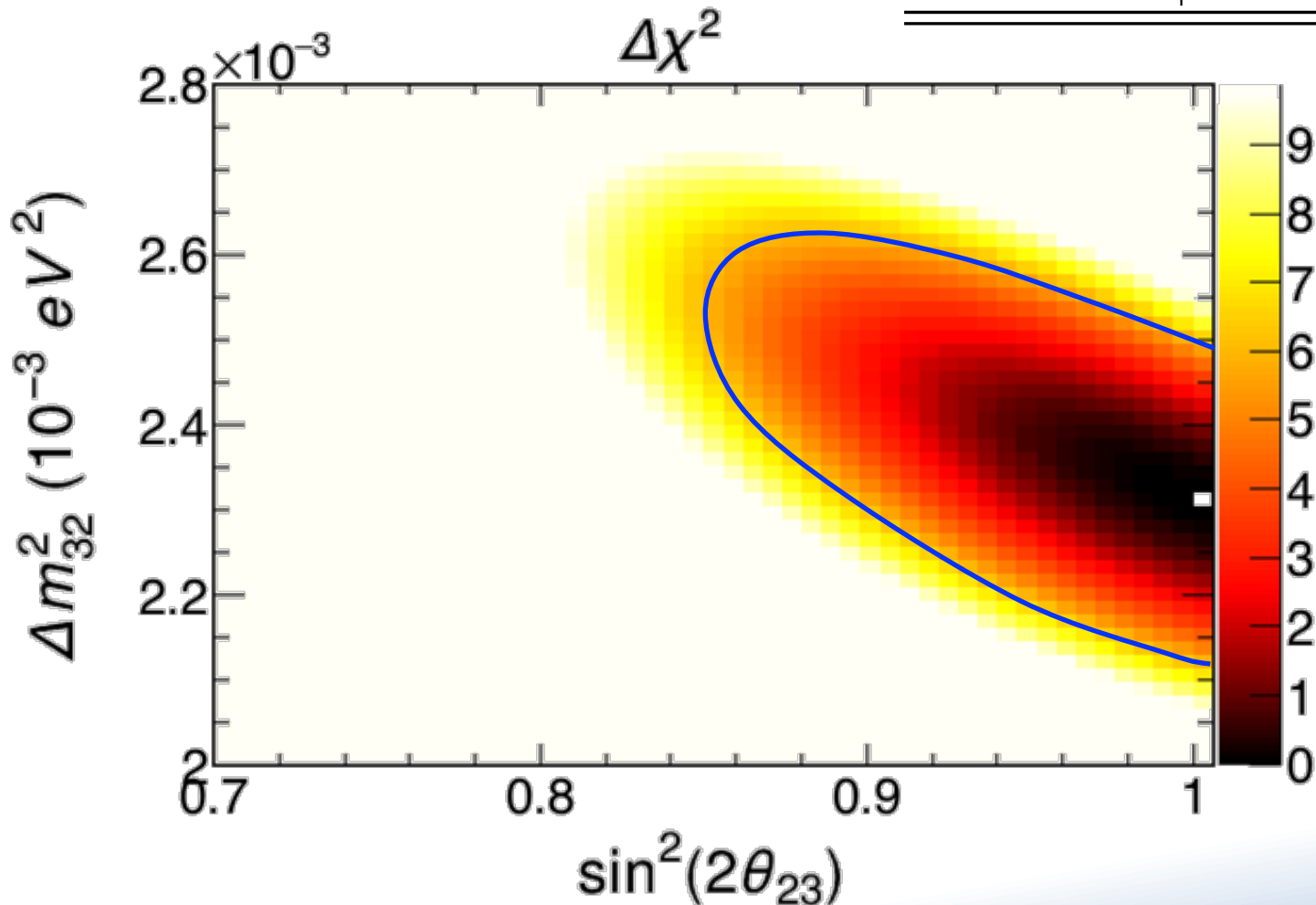
- More variables?
 - If you have 2 variables, and you want to show 2 variables, then it's straightforward.
 - Just pick the right up value, and points below it are in.

$(1 - \alpha)$ (%)	$m = 1$	$m = 2$	$m = 3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16



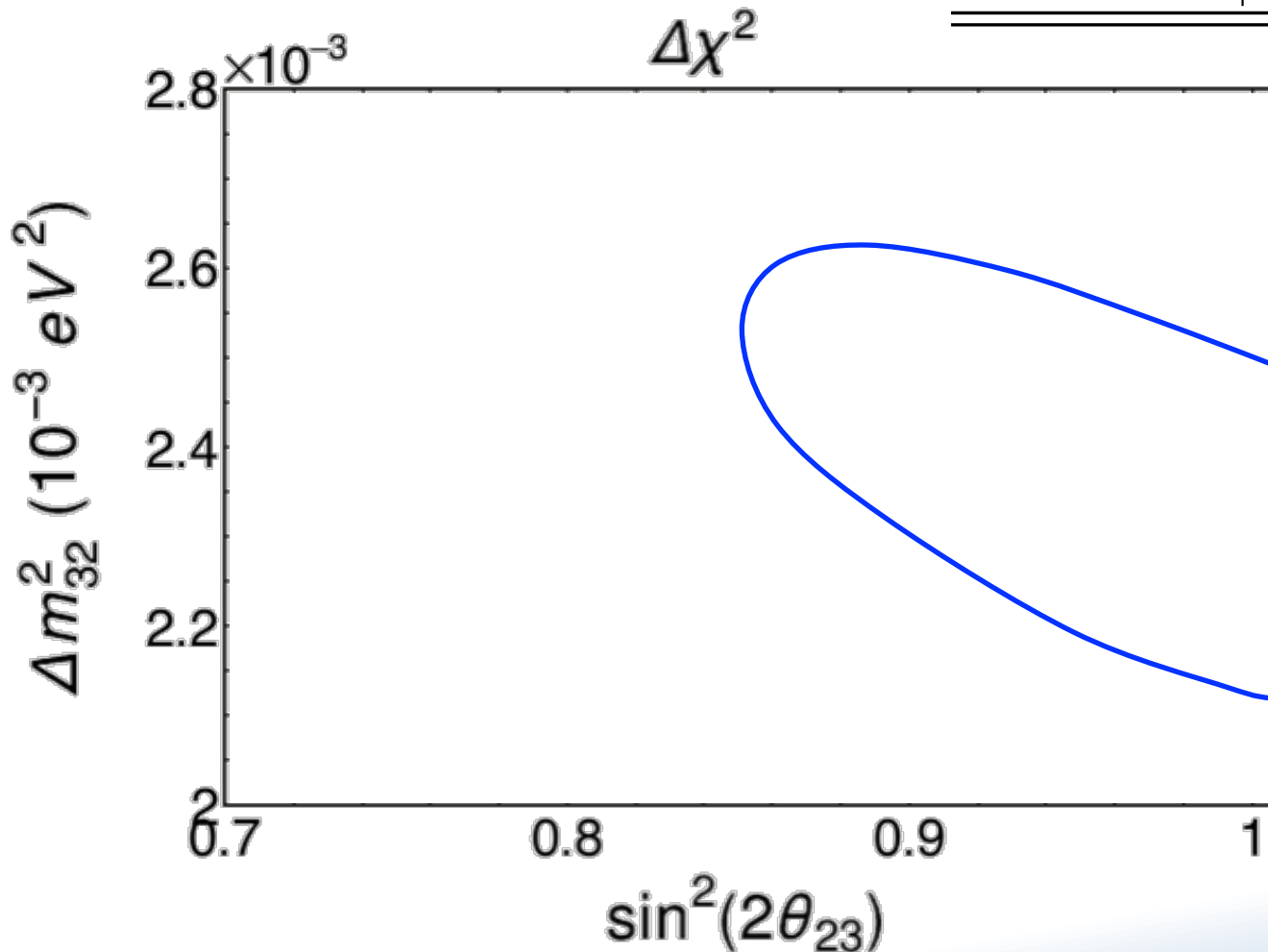
- More variables?
 - If you have 2 variables, and you want to show 2 variables, then it's straightforward.
 - Just pick the right up value, and points below it are in.

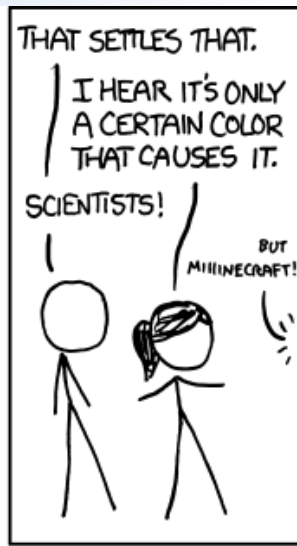
$(1 - \alpha)$ (%)	$m = 1$	$m = 2$	$m = 3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16



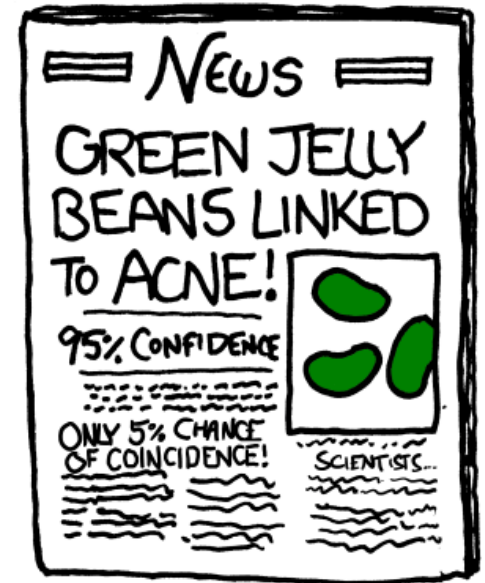
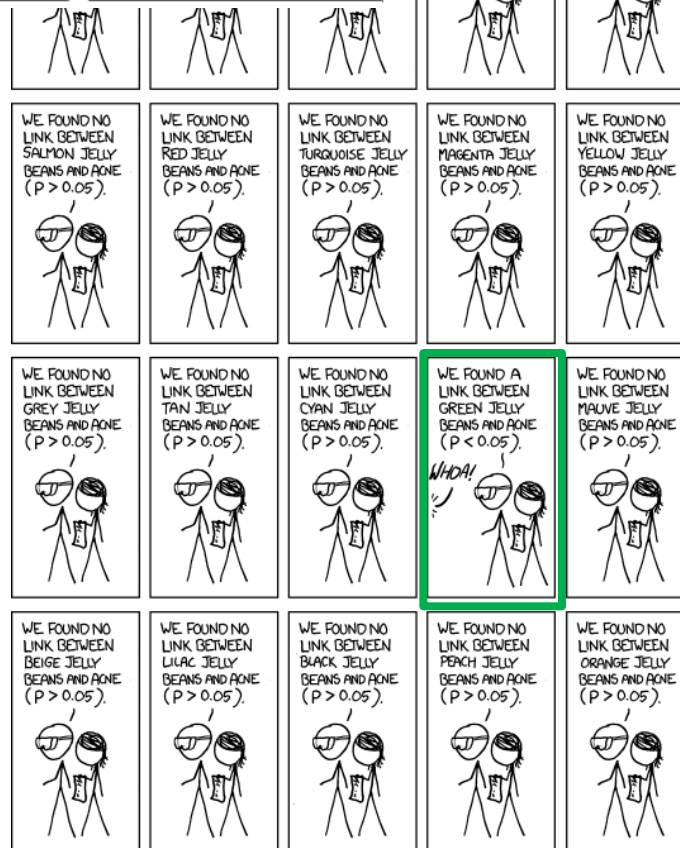
- More variables?
 - If you have 2 variables, and you want to show 2 variables, then it's straightforward.
 - Just pick the right up value, and points below it are in.

$(1 - \alpha)$ (%)	$m = 1$	$m = 2$	$m = 3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16





- Also – be wary of the “look elsewhere” effect.
 - More often a problem for us when looking at data-MC comparisons.



<https://xkcd.com/882/>

N tests, level α

want prob of at least 1
test failing threshold β to be α

$$P(p_i < \beta \text{ any } i) = \alpha$$

$$1 - P(p_i < \beta \forall i \dots N) = \alpha$$

$$1 - (1 - \beta)^N = \alpha$$

$$(1 - \beta)^N = (1 - \alpha)$$

$$1 - \beta = (1 - \alpha)^{1/N}$$

$$\beta = 1 - (1 - \alpha)^{1/N}$$

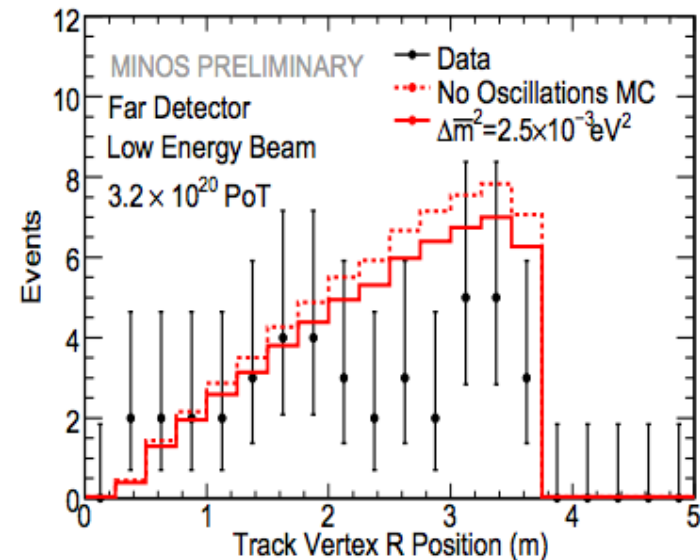
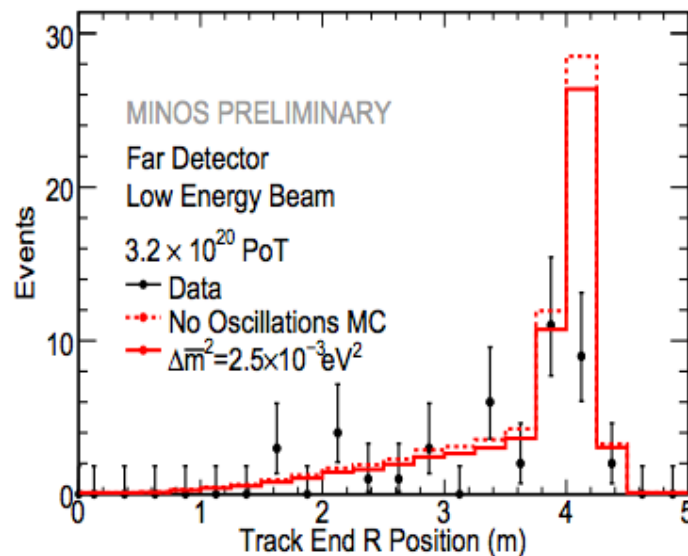
Say, $N = 7$ $\alpha = 0.05$

$$\beta = 1 - (0.95)^{1/7}$$

$$= 0.0073$$

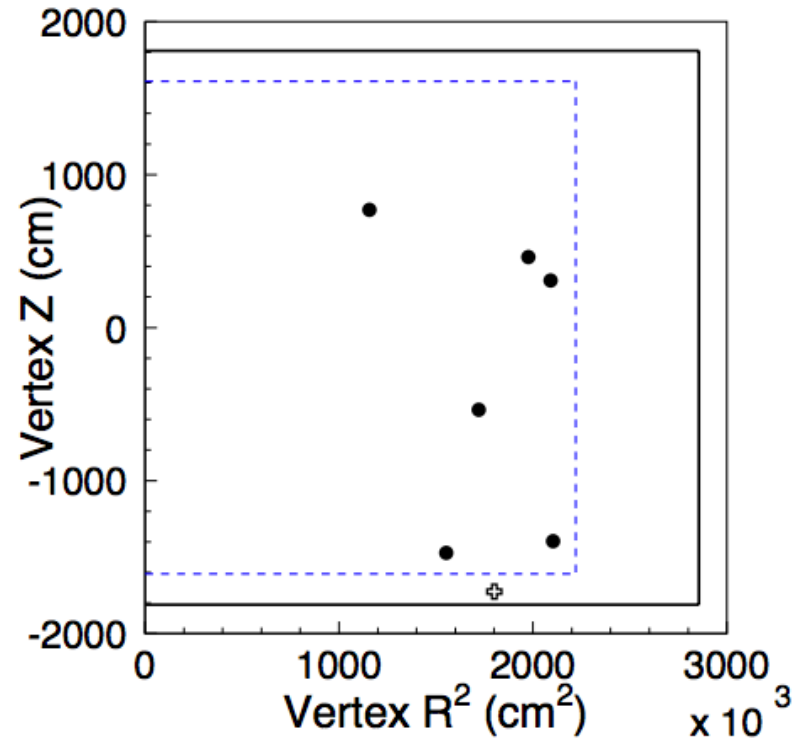
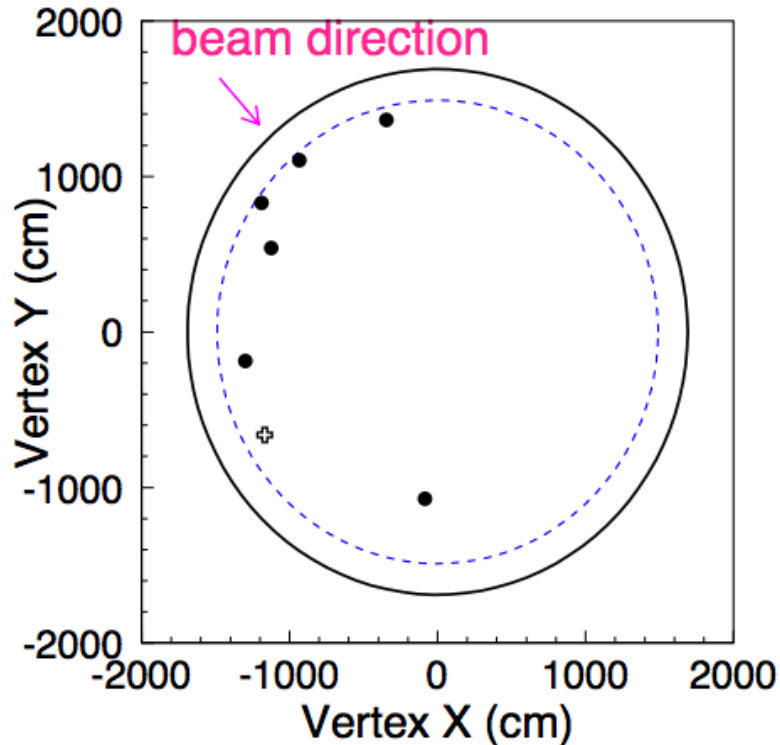
Anomalous FD Distributions

Z. Isvan



- ▶ Track End R has $\sim 3.3\sigma$ discrepancy at 4.1m (26 events expected 9 events seen discrepancy of $(26-9)/\sqrt{26}=3.3\sigma$)
 - ▷ Essentially all the missing events are in a single Track End R bin.
- ▶ Vertex R distribution also shows discrepancy in region $r > 2\text{m}$.

Vertex distribution of ν_e candidate events



These events are clustered at large R

→ Perform several checks. for example

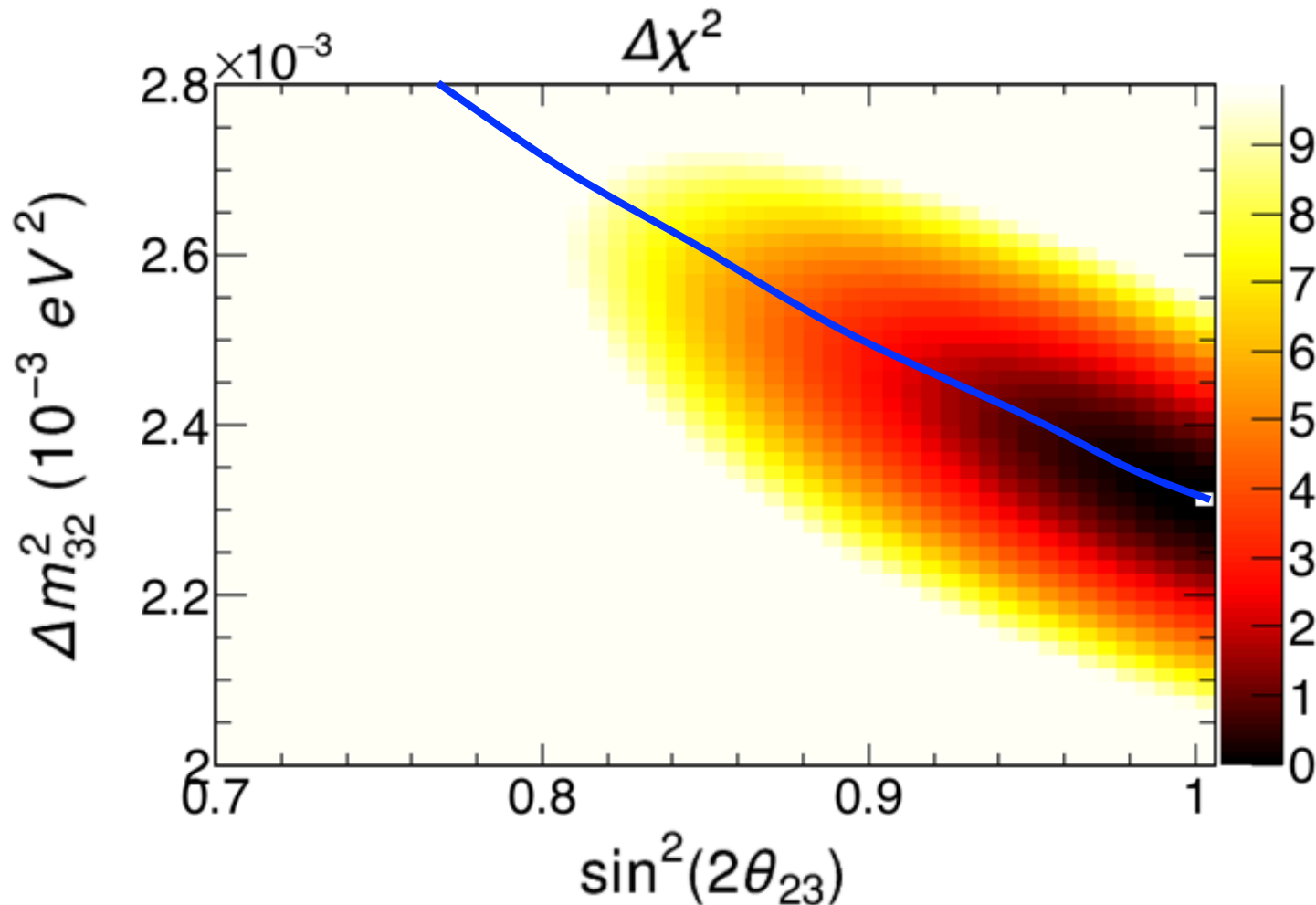
- * Check distribution of events outside FV → no indication of BG contamination
- * Check distribution of OD events → no indication of BG contamination
- * K.S. test on the R^2 distribution yields a p-value of 0.03

+ Event outside FV

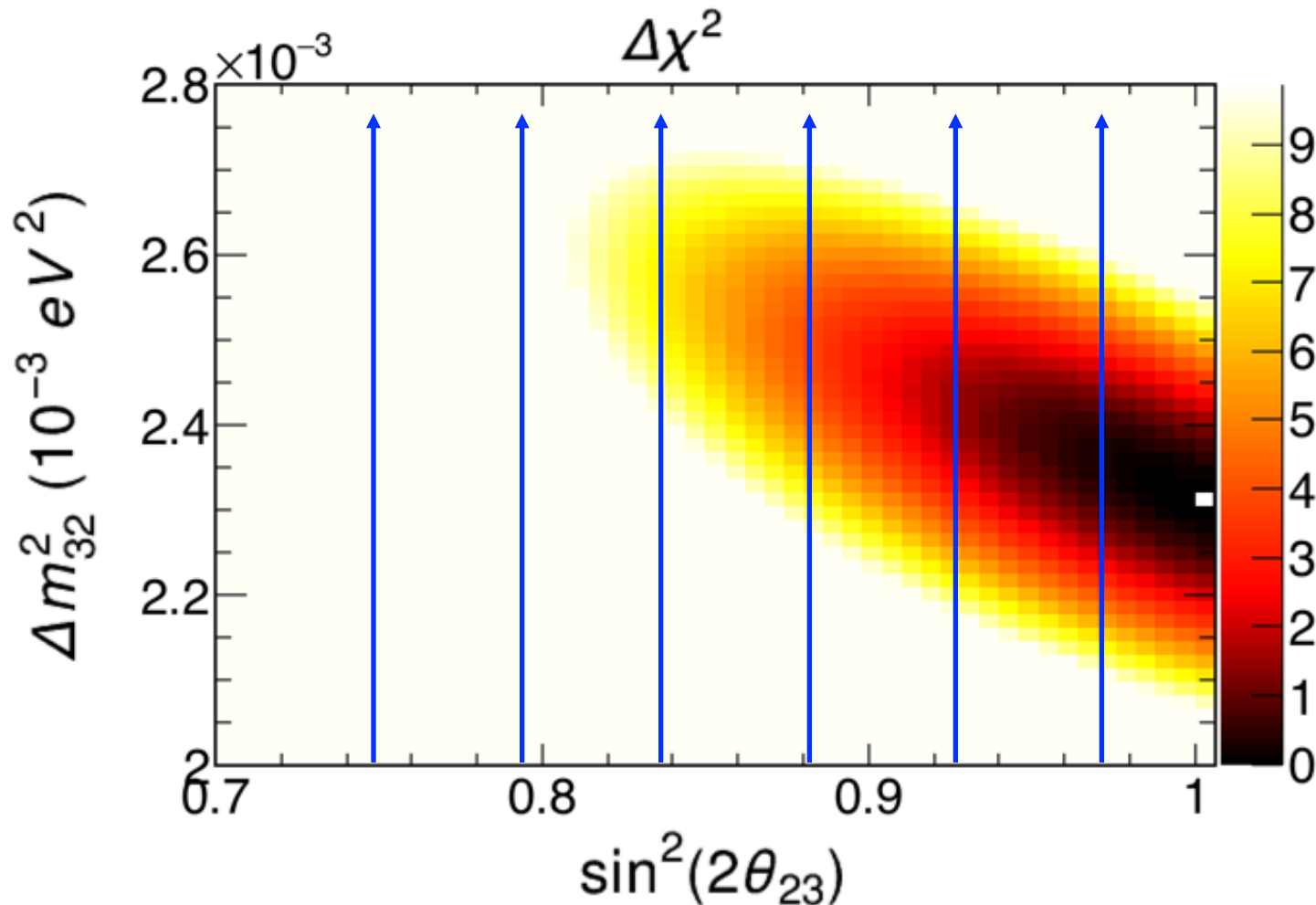
- Nuisance parameters
 - Often your likelihood depends on more parameters than you want to present.
 - Extra parameters can be physics or systematic uncertainties.
- For example, in the NOvA joint fit we do:
 $(\Delta m^2, \theta_{23}, \theta_{13}, \delta, \text{systematic errors}) \rightarrow (\theta_{23}, \delta)$
- Two different approaches:
 - Profiling
 - Marginalizing

- Profiling

- Take the best fit in all parameters you are not showing *at each point* you do show.
- More common, works under certain assumptions.



- Marginalizing
 - Integrate up all the values you are not showing.
 - Shows up more in Bayesian analyses.



Frequentist

- Apply solid mathematical rigor to answer a question that nobody cares about.

Bayesian

- Answers the question everyone is really interested in using assumptions no one believes.

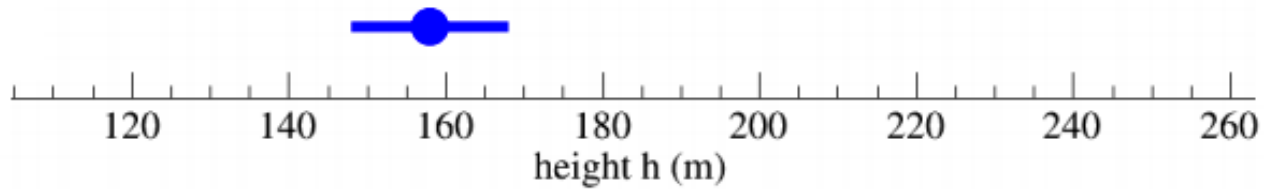
Frequentist

- The true value of a measurement is **an unknown constant**.
- Report the **probability of experimental outcomes**, given a value of that constant.
- Use that to construct a confidence interval which will contain the true value in α fraction of experiments.

Bayesian

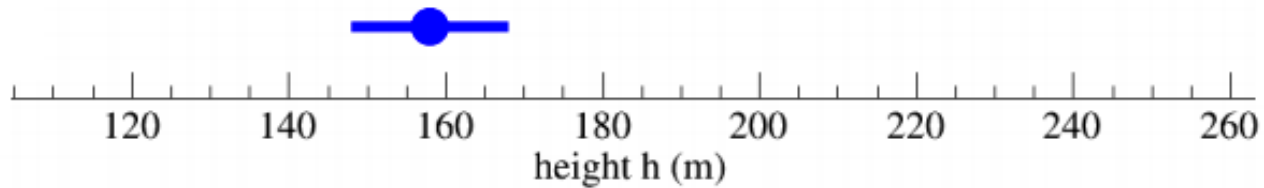
- The true value of a measurement is a **random variable**.
- **Before** the measurement, have a “prior” PDF of that variable.
- After the measurement, **update to a “posterior” PDF** using the data collected.

Frequentist



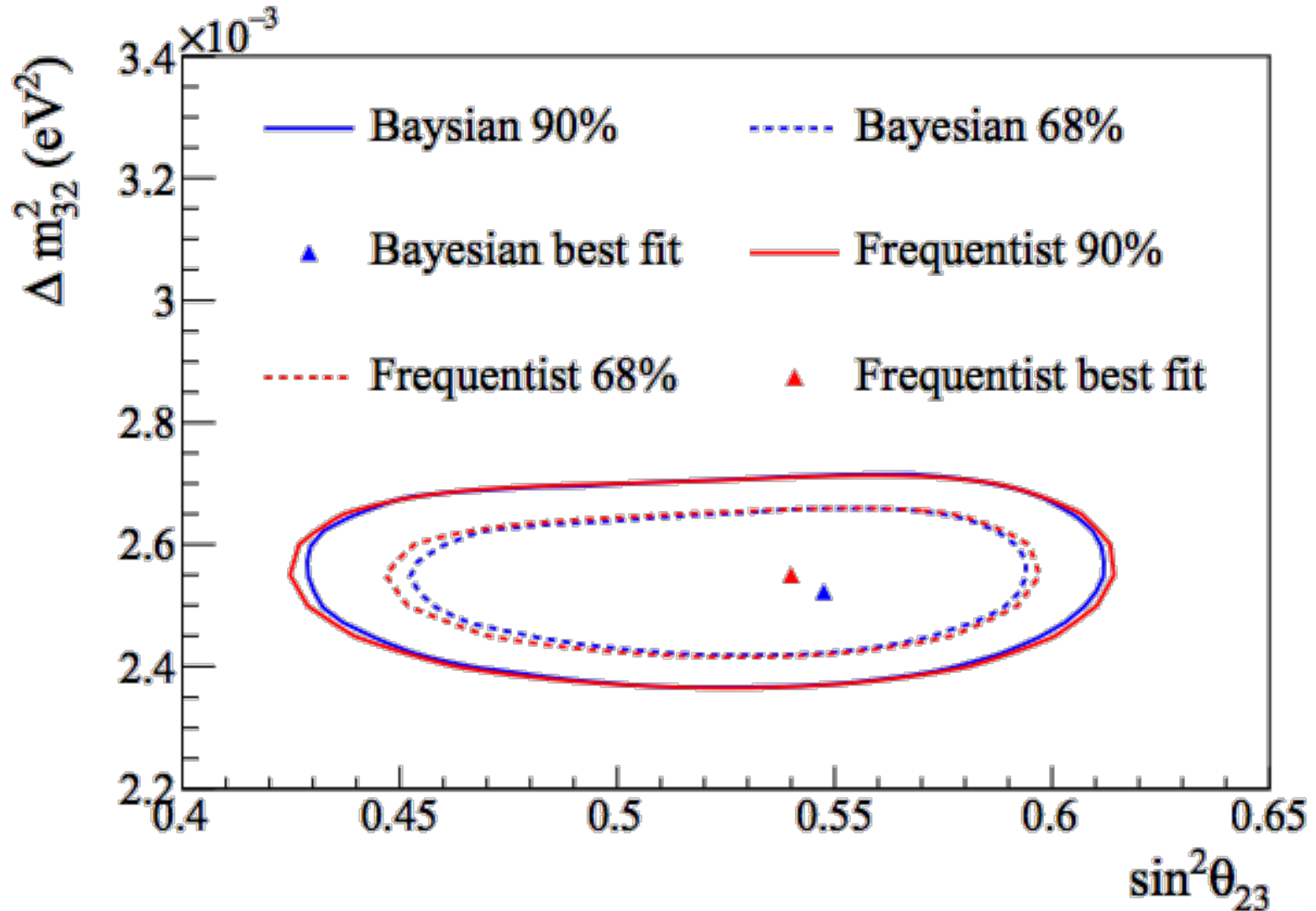
$$h = (158 \pm 20) \text{ m}$$

Bayesian



$$h = (158 \pm 20) \text{ m}$$

- In the real world:
 - This is from the latest T2K PRD, arXiv:1707.01048



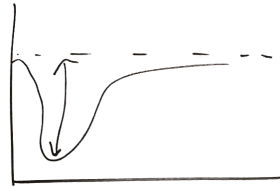
$(1 - \alpha)$ (%)	$m = 1$	$m = 2$	$m = 3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

- What if you can't trust the values from the PDG?
 - They don't have the right coverage:
a 90% C.L. is actually an 85% C.L.
- Commonly happens when statistics are low and the problem has a physical boundary.
 - Happens a lot in neutrino physics since $0 < \sin^2 2\theta < 1$

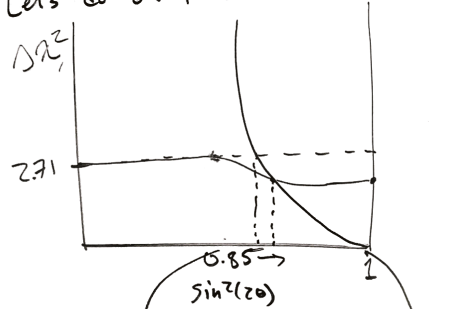
Feldman-Cousins

- The solution is a technique called **Feldman-Cousins**
- From a paper called “A Unified Approach to the Classical Statistical Analysis of Small Signals”
 - by Gary Feldman and Bob Cousins
 - Phys. Rev. **D57** (1998) 3873
- Let’s walk through an example.

1D: fit for $\sin^2(z\theta)$

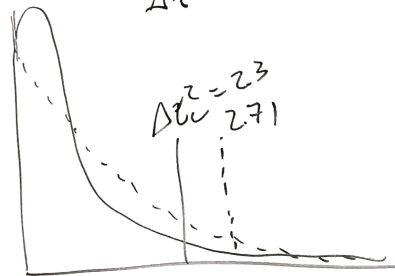
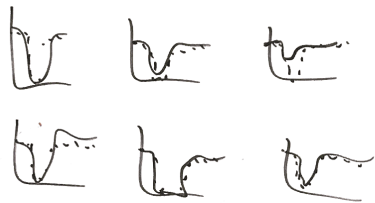
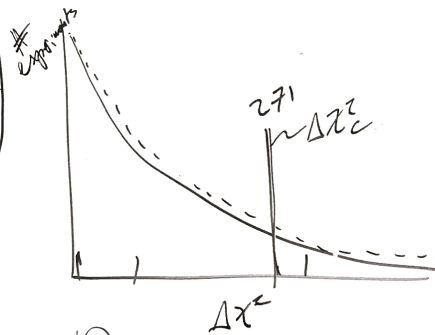
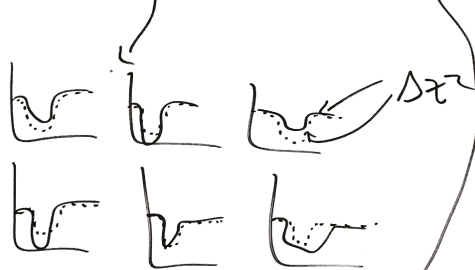


Let's do our fit:



Is $z.71 = 90\% \text{ C.L.}$?

"non-parametric
Statistics"



- A real-life example from the MINOS anti- ν_μ disappearance analysis circa 2010.

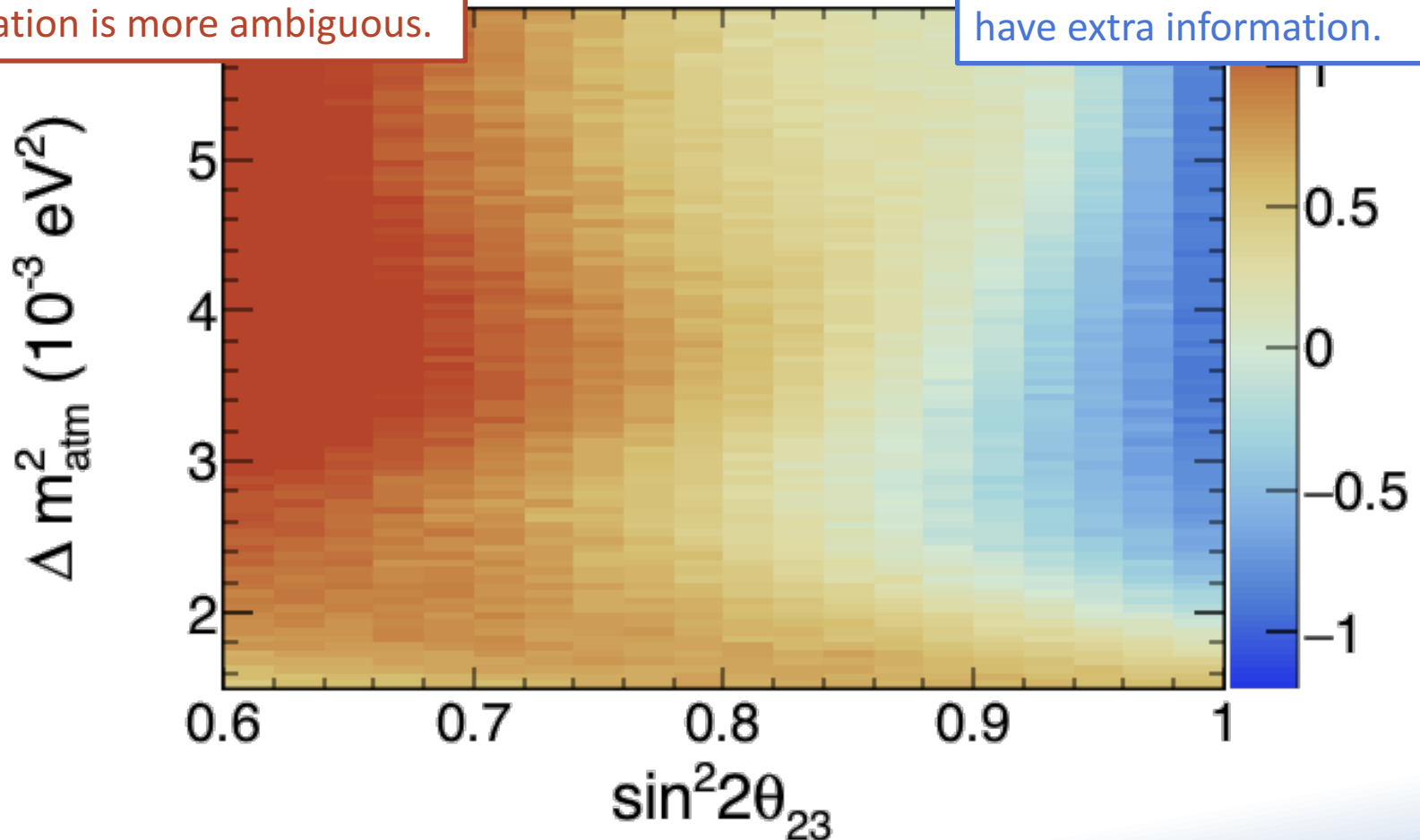
Up-values above nominal when there is a degeneracy...

Contours are looser since the information is more ambiguous.

$$\Delta\chi_c^2 - \Delta\chi_{\text{gaus}}^2$$

Up-values below nominal close to a physical boundary...

Contours are tighter since we have extra information.



Conclusion

- I've tried to show the statistical underpinnings of some of the most common statistical techniques we use.
 - But there are many, many more possible techniques.
 - There are numerous alternative ways to do everything I have presented here.
- Some general advice: **use the simplest method that is correct, but no simpler.**
 - If you use a technique that requires assumptions that you cannot meet, your results will be questioned.
 - But, if you use a more complicated technique, be prepared to explain how it works and why you chose it.
- I highly recommend the PDG statistics section as a place to find statistical techniques which are “commonly accepted” in physics.

Backups

$$\mu_1 = 2.5 \pm 0.1 \quad \mu_2 = 3.1 \pm 0.3$$

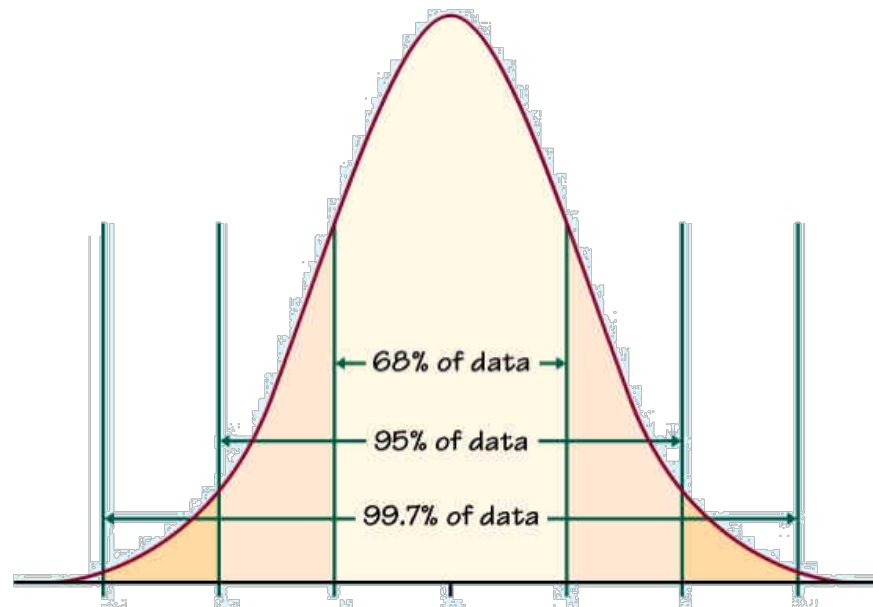
$$X = \mu_2 - \mu_1 = 0.6$$

- We know, from the central limit theorem, that means are normally distributed.
 - The difference between means is, too.
 - The standard deviation of that difference is:

$$\sigma_{\mu_2 - \mu_1} = \sqrt{\sigma_{\mu_2}^2 + \sigma_{\mu_1}^2} = 0.32$$

- Now the question is: Is 0.6 significantly different from 0 if it comes from a normal distribution with $\sigma = 0.32$?

- What we are asking is: **how likely** is it that we would get our result, or something more extreme **assuming the null hypothesis is true**?
 - This is the definition of p -value, which we compare to our α .
- “different from” means we are making a “two-sided” test:
 - If we set an $\alpha = 0.05$, we want to know if our value falls into the central $1-\alpha$ or 95% of the distribution.



- To start, we calculate a “Z-score,” which effectively converts from our specific normal distribution to the canonical $\mathcal{N}(0,1)$:

$$Z = \frac{X - \mu_0}{\sigma} = \frac{0.6 - 0}{0.32} = 1.88$$

This is what we mean when we say “1.88 σ ”

- But, what does that Z-score mean?
 - In other words, what is its p -value we can compare to α ?
 - What fraction of values in the distribution are **more extreme** than ours?

$$p = \int_{-\infty}^{-Z} \mathcal{N}(x, 1) dx + \int_Z^{\infty} \mathcal{N}(x, 1) dx$$

- But infinity is hard, so we can take advantage of the fact that probabilities all add up to 1 to do the inverse:

$$p = 1 - \int_{-Z}^Z \mathcal{N}(x, 1) dx$$

$$p = 1 - \int_{-Z}^Z \mathcal{N}(x, 1) dx$$

- This integral doesn't have an analytical solution, but we need it all the time, so its results are readily available as the "error function"

```
// Z-score (sigmas) -> p-value  
root [4] 1 - TMath::Erf(1.88 / TMath::Sqrt(2))  
(Double_t) 0.0601081
```

$$\mu_1 = 2.5 \pm 0.1 \quad \mu_2 = 3.1 \pm 0.3$$

- With $p = \mathbf{0.06}$, we have failed to reject the null hypothesis at $\alpha = \mathbf{0.05}$.
- “These two means are consistent at the 95% level.”
- Or, we might say:
 - “These means differ by 1.88σ ” or
 - “They are consistent at the 94% level”

- Now, we need to choose a test statistic.
 - There are several choices for this problem.
 - Which one is the right one depends on the circumstance.
- A good first guess: try a chisquare (χ^2) test.
 - This is what the test statistic looks like:

$$T = \sum^N \frac{(O_i - E_i)^2}{\sigma^2}$$

- Squared difference between the histograms, normalized by the expected uncertainty.

- Why this test statistic?
 - Let's see how it behaves assuming H_0 .
 - The data is drawn randomly from the model, so each bin, O_i , should be drawn randomly from the model:

$$O_i \sim \mathcal{N}(E_i, \sigma)$$

- Given that, the argument in the sum of the chisquare should look familiar – it is a Z-score, squared.

$$T = \sum^N \frac{(O_i - E_i)^2}{\sigma^2} \quad Z = \frac{X - \mu_0}{\sigma}$$