

A brief, incomplete, and possibly incorrect introduction to statistics or A guide to physicists' jargon

The terms I hope to define

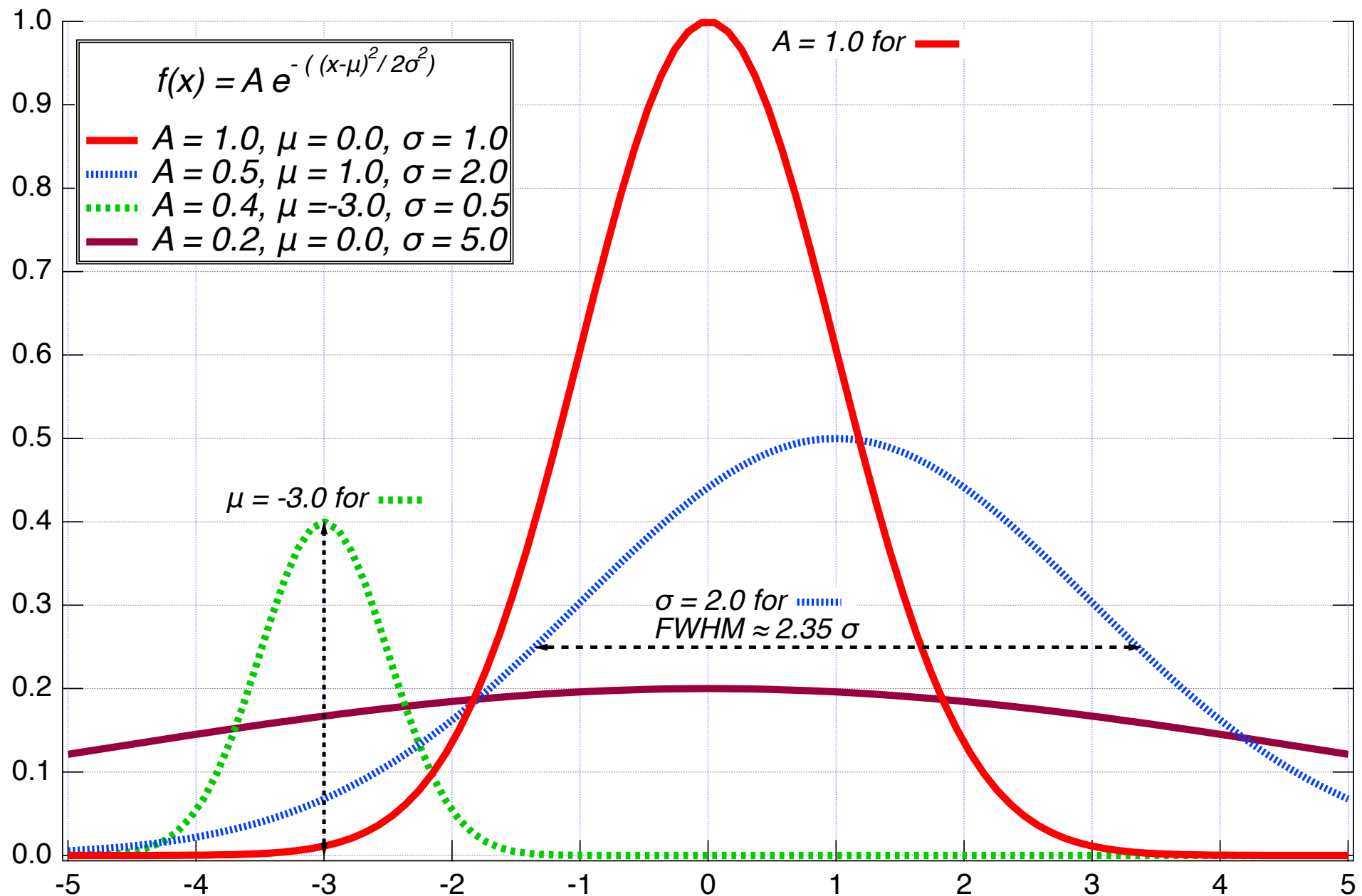
- gaussian
- chi-squared
- chi-squared per degrees of freedom
- sigma (as in “we’re looking for a five-sigma effect”)
- systematic error

Professor Michael Shaevitz, Director of Nevis Labs, is our expert on statistics. I’m half-remembering what he taught me, and making up the rest.

gaussian

$$f(x) = e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- μ is the mean of the distribution.
- σ = the standard deviation; it's related to the “full width at half maximum” (FWHM) of the curve by $\text{FWHM} = 2\sqrt{2\ln 2}\sigma \approx 2.35\sigma$.
- e = Euler's constant, a transcendental number that occurs often in calculations that relate to growth and increase. It's formally defined as $\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$.



caveats

- If you want to work with the normal distribution as a “probability density function” then you’ll want to include a normalization so the integral $\int_{-\infty}^{\infty} \mathcal{N}(x) dx = 1$

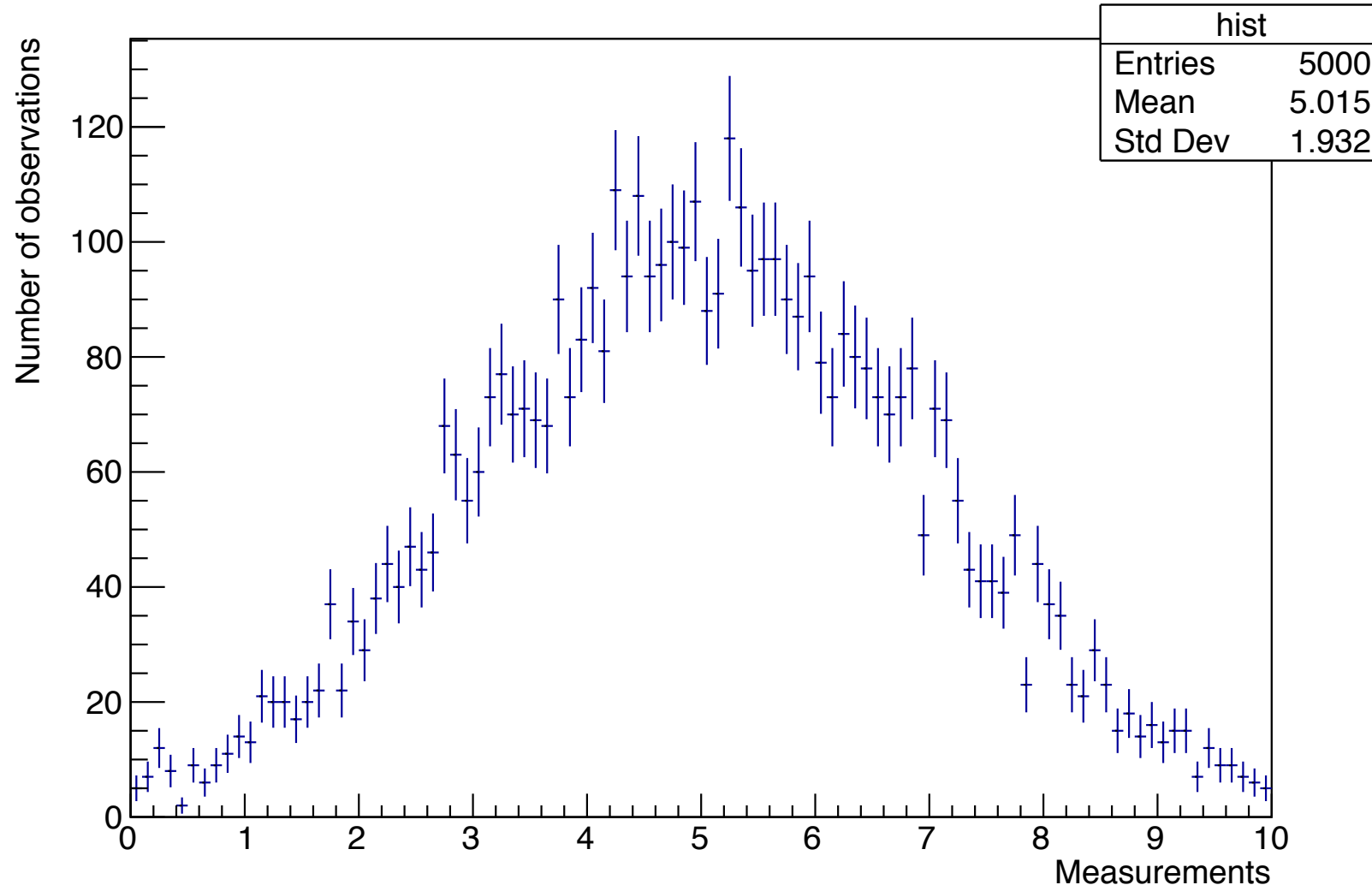
$$N(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

However, if you work with this form you can’t fiddle with the amplitude of the distribution. You don’t usually see it in physics.

- There are other functional forms in physics than the gaussian! However, I’m lazy, so that’s the only one I use in the tutorial.
- It makes some sense to stick with gaussians, since the sum of many random processes (even non-gaussian ones) tends towards a gaussian.

chi-squared

Measurements from my experiment



What's the probability that the underlying distribution of this histogram is a gaussian?

chi-square for a 1D histogram

$$\chi^2 = \sum_i \frac{(y_i - f(x_i; p_j))^2}{e_i^2}$$

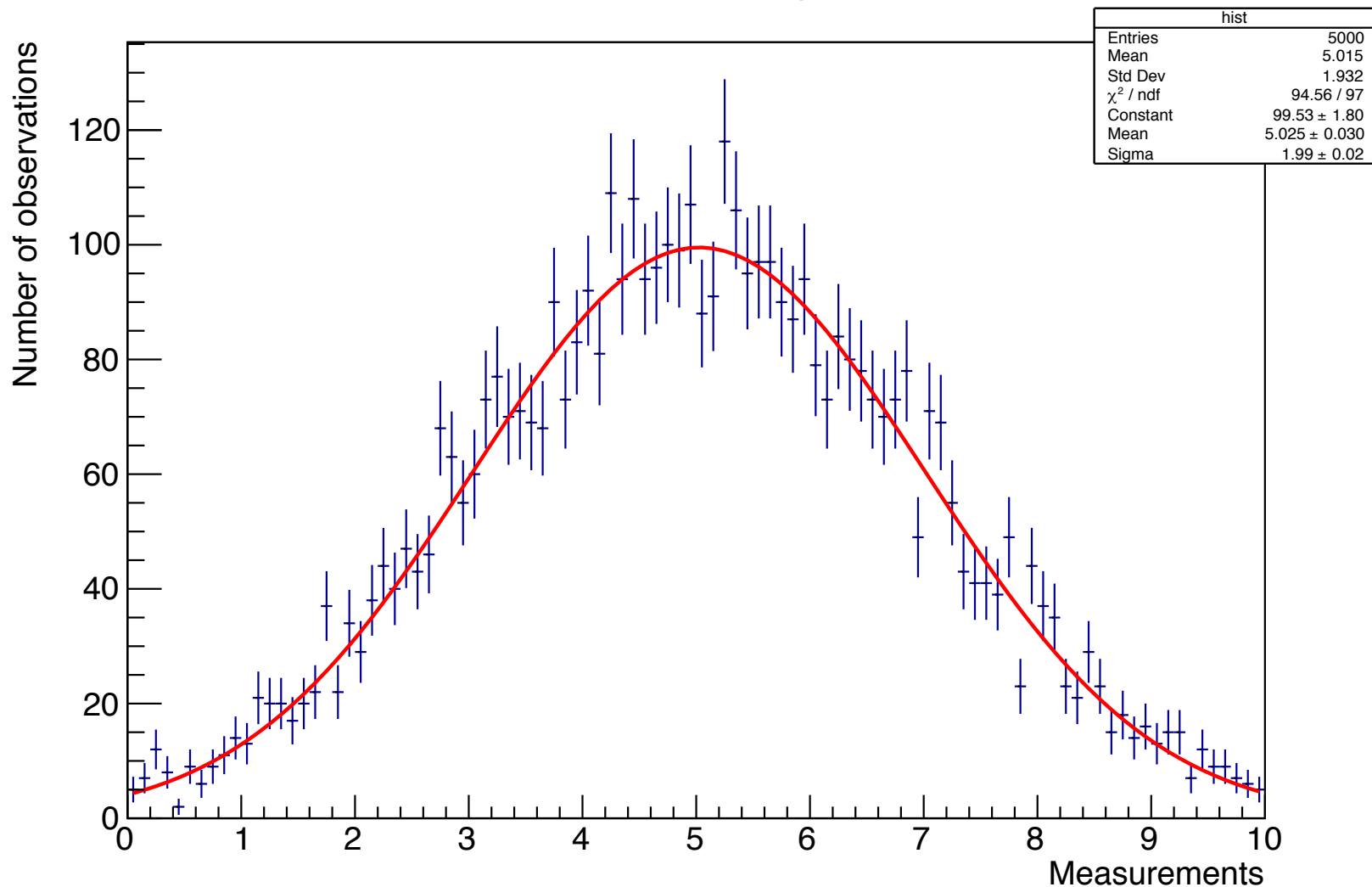
where:

- i means the i -th bin of the histogram (more generally, the i -th data point you've gathered).
- y_i means the data (or value of) the i -th bin of the histogram.
- e_i means the error in the i -th bin of the histogram (i.e., the size of the error bars).
- $f(x_i; p_j)$ means to compute the value of the function at x_i (the value on the x -axis of the center of bin i) given some assumed values of the parameters $p_0, p_1, p_2 \dots p_j$.

The process of “fitting” means to test different values of the parameters until you find those that minimize the value of χ^2 .

Fit $\equiv$ Test different values of A , μ , and σ until you find those that minimize the value of χ^2 .

Measurements from my experiment

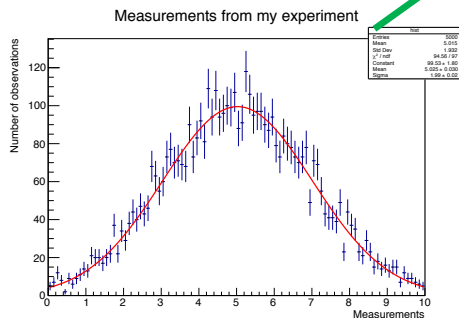


The underlying program that does this is Minuit. It's been a standard program for finding function minima for decades.

chi-squared per (number of) degrees of freedom

experiment

hist	
Entries	5000
Mean	5.015
Std Dev	1.932
χ^2 / ndf	94.56 / 97
Constant	99.53 ± 1.80
Mean	5.025 ± 0.030
Sigma	1.99 ± 0.02



What value of chi-square do you expect from a fit?

$$\chi^2 = \sum_i \frac{(y_i - f(x_i; p_j))^2}{e_i^2}$$

- For each individual bin i the data forms a little gaussian distribution of its own with a mean of y_i .
- The e_i acts as a scale of the difference between y_i and the function $f(x)$.
So if $f(x)$ is a reasonable approximation to y_i , $(y_i - f(x))/e_i$ will be around ± 1 .
- You add up those “1”s for each of the bins, and you might anticipate that χ^2 will be roughly equal to i , the number of bins.

But we have to adjust that...

There are three “free parameters” in the fit: A, μ, σ .

They’re going to be varied to make the chi-squared smaller. The net effect is that total number of “degrees of freedom” is:

$$\begin{aligned} \text{DOF} &= \text{number of data points} \\ &\quad - \text{number of free parameters in the function} \end{aligned}$$

In that fit a couple of pages ago, $\chi^2 / \text{ndf} = 94.56 / 97$.

What does this tell us?

Look it up on
the web

Upper-tail critical values of chi-square distribution with ν degrees of freedom

ν	Probability less than the critical value				
	0.90	0.95	0.975	0.99	0.999
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458

For the typical fits that we do in physics (lots of data points), it's sufficient that χ^2 / nfd is roughly 1.

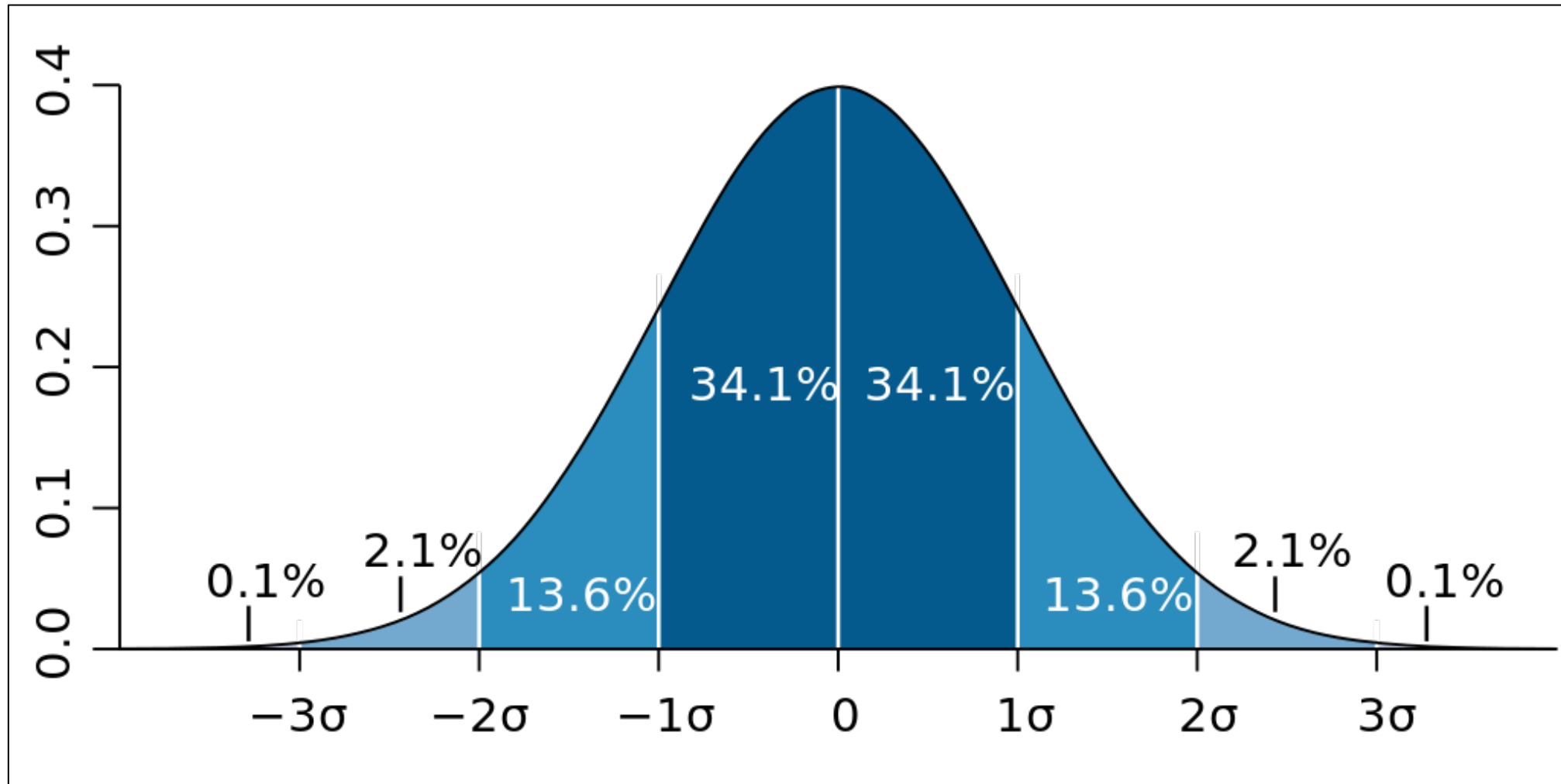
Why might χ^2 / ndf be much greater than 1?

- There's something wrong in the routine that's calculating χ^2
- The model that's being assumed for the function does not have enough parameters.
- The error bars for your data are too small
- Function-minimization programs can get “stuck” in a local minimum that's not the actual true minimum

Why might χ^2 / ndf be much less than 1?

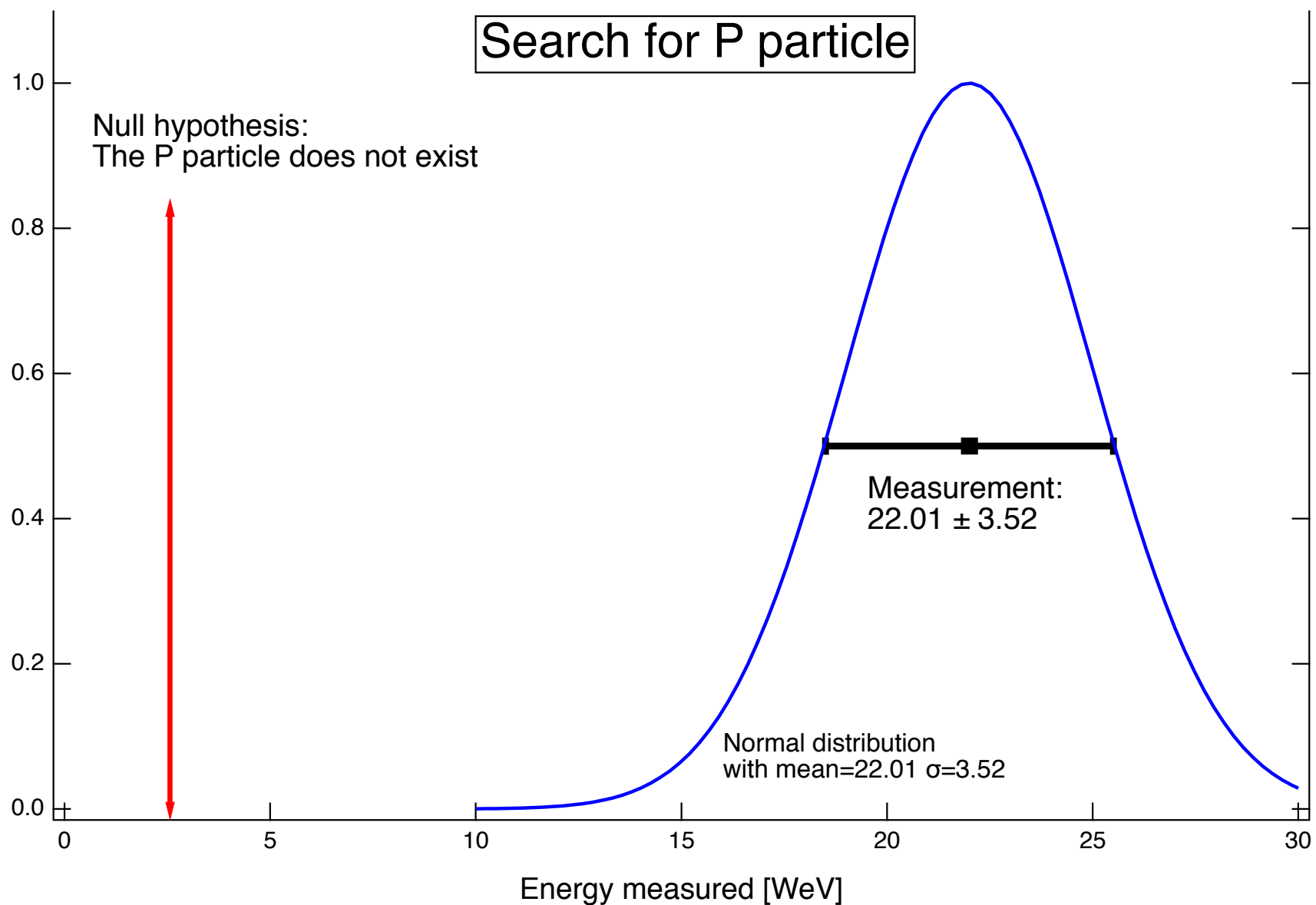
- Again, something wrong in the χ^2 calculation.
- Too many free parameters in the function you're using to fit.
- The errors on your data are too large.
- Someone has gone wrong in your data-analysis process and you're “tuning” the data to the model you want to fit.

“We’re looking for a five sigma effect”



3σ = “evidence”
 5σ = “discovery”

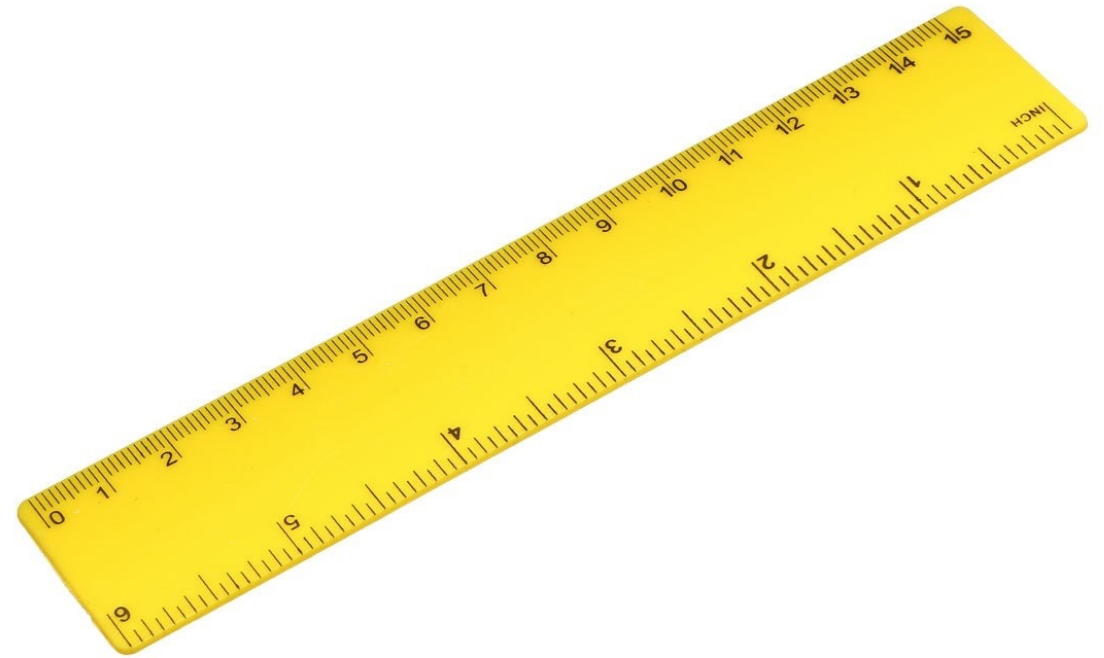
This how I think of it



Systematic error versus statistical error

Statistical error = random errors
in the measurement.

Systematic error = a bias in the
measurement, but you don't
know how much that bias is.



Example of an experiment's systematic errors

ATLAS PUB Note

ATL-PHYS-PUB-2018-001

31st January 2018

*Investigation of systematic
uncertainties on the measurement
of the top-quark mass using
lepton transverse momenta*

Uncertainty	Δm_{top} [GeV]
Statistics	0.94
Method calibration	0.40
Signal MC generator	0.62
Single-top Wt generator	0.28
Hadronisation and parton shower	0.55
ISR and FSR	1.39
Underlying Event	0.67
Colour Reconnection	0.23
Parton distribution function	0.42
Single-top contribution	0.10
Leptons	0.50
$E_{\text{T}}^{\text{miss}}$	0.12
b -tagging	0.08
Jet energy scale	0.60
Jet energy resolution	0.32
Jet vertex fraction	0.05
Total	2.27