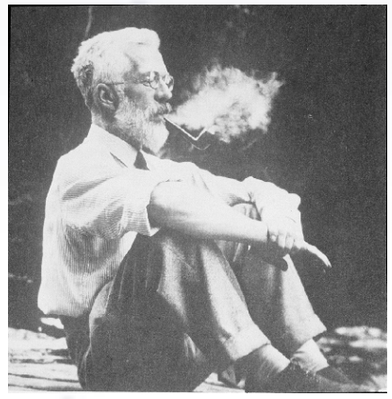
Applied Statistics

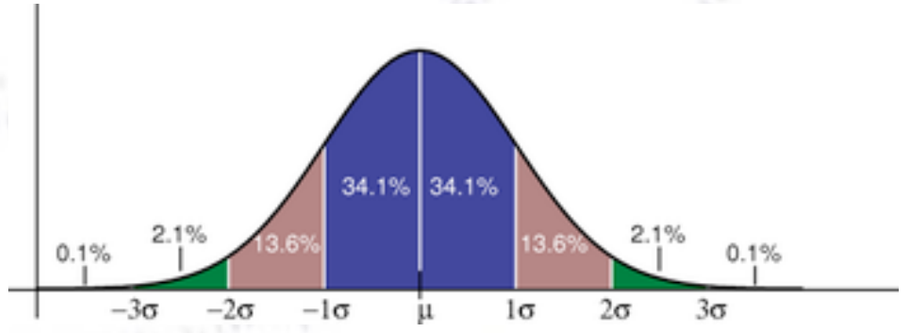
Principle of maximum likelihood





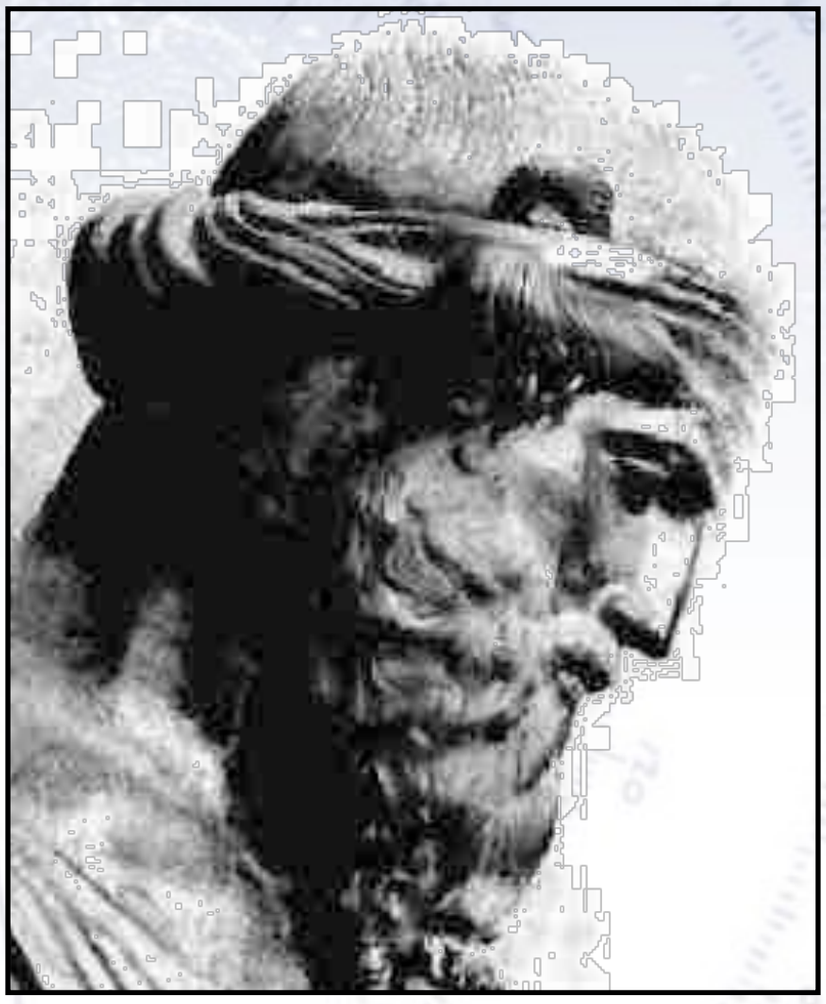



Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense"

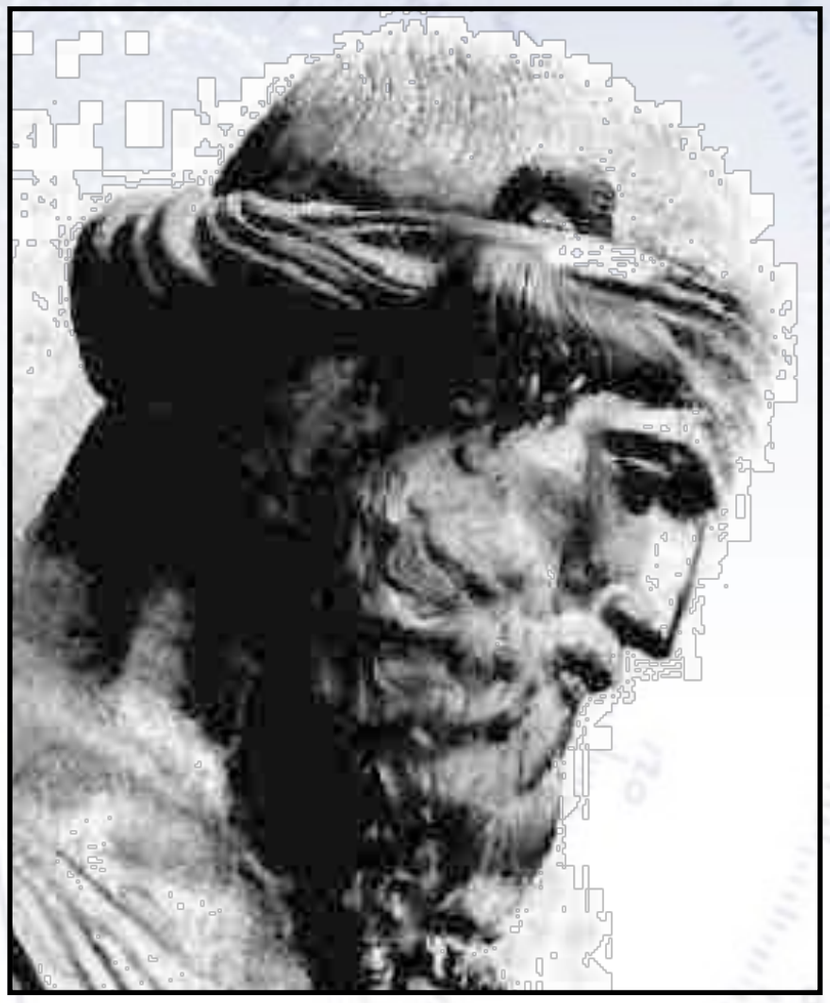
Likelihood function



"I shall stick to the principle of likelihood…"

[Plato, in Timaeus]

Likelihood function



Given a PDF f(x) with parameter(s) θ, what is the chance that with N observations, x_i falls in the intervals $[x_i, x_i + dx_i]$?

$$\mathcal{L}(\theta) = \prod_i f(x_i, \theta) dx_i$$

Likelihood function

Given a set of measurements **x**, and parameter(s) **θ**, the likelihood function is defined as:

$$\mathcal{L}(x_1, x_2, \ldots, x_N; \theta) = \prod_i p(x_i, \theta)$$

The **principle of maximum likelihood** for parameter estimation consist of maximising the likelihood of parameter(s) (here **θ**) given some data (here **x**).
There is nothing strange about this - it is exactly the same we do for the ChiSquare!

The likelihood function plays a central role in statistics, as it can be shown to be:

- Consistent (converges to the right value).
- Asymptotically normal (converges with Gaussian errors).
- Efficient (reaches the Minimum Variance Bound (MVB, Cramer-Rao) for large N).

$$V(\hat{a}) \geq \frac{1}{< (d \ln L / da)^2 >}$$

To some extend, this means that the likelihood function is “optimal”, that is, if it can be applied in practice.

Likelihood vs. Chi-Square

For computational reasons, it is often much easier to minimise the logarithm of the likelihood function:

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \theta}\right|_{\theta=\bar{\theta}} = 0$$

In problems with Gaussian errors, it turns out that the **log likelihood function** boils down to the **Chi-Square** with a constant offset and a factor -2 in difference.

See Barlow 5.6

The likelihood function for fits comes in two versions:

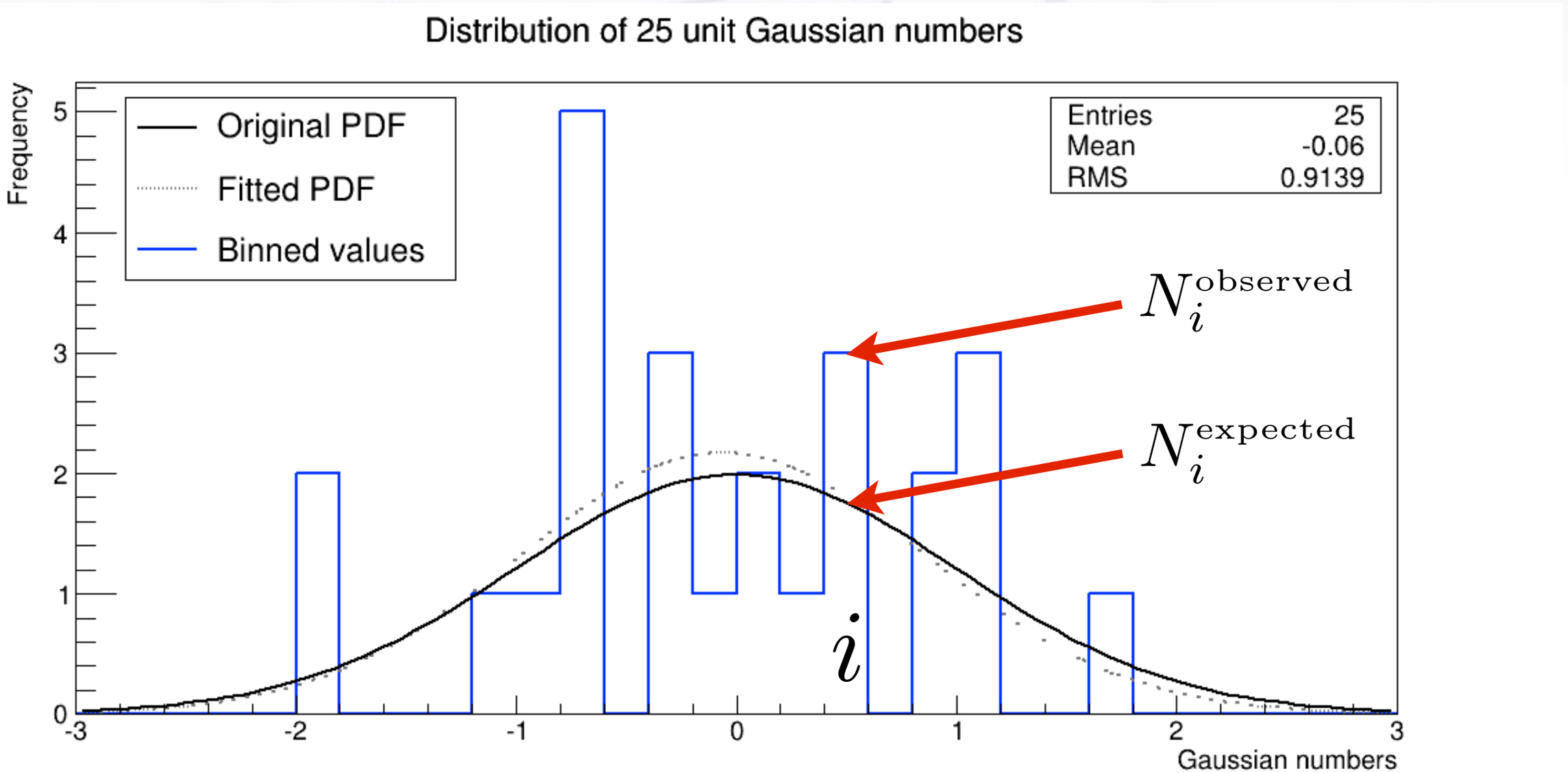
- Binned likelihood (using Poisson) for histograms.
- Unbinned likelihood (using PDF) for single values.

The “trouble” with the likelihood is, that it is unlike the Chi-Square, there is NO simple way to obtain a probability of obtaining certain likelihood value!

ChiSquare

Recall, the ChiSquare is a sum over bins in a histogram:

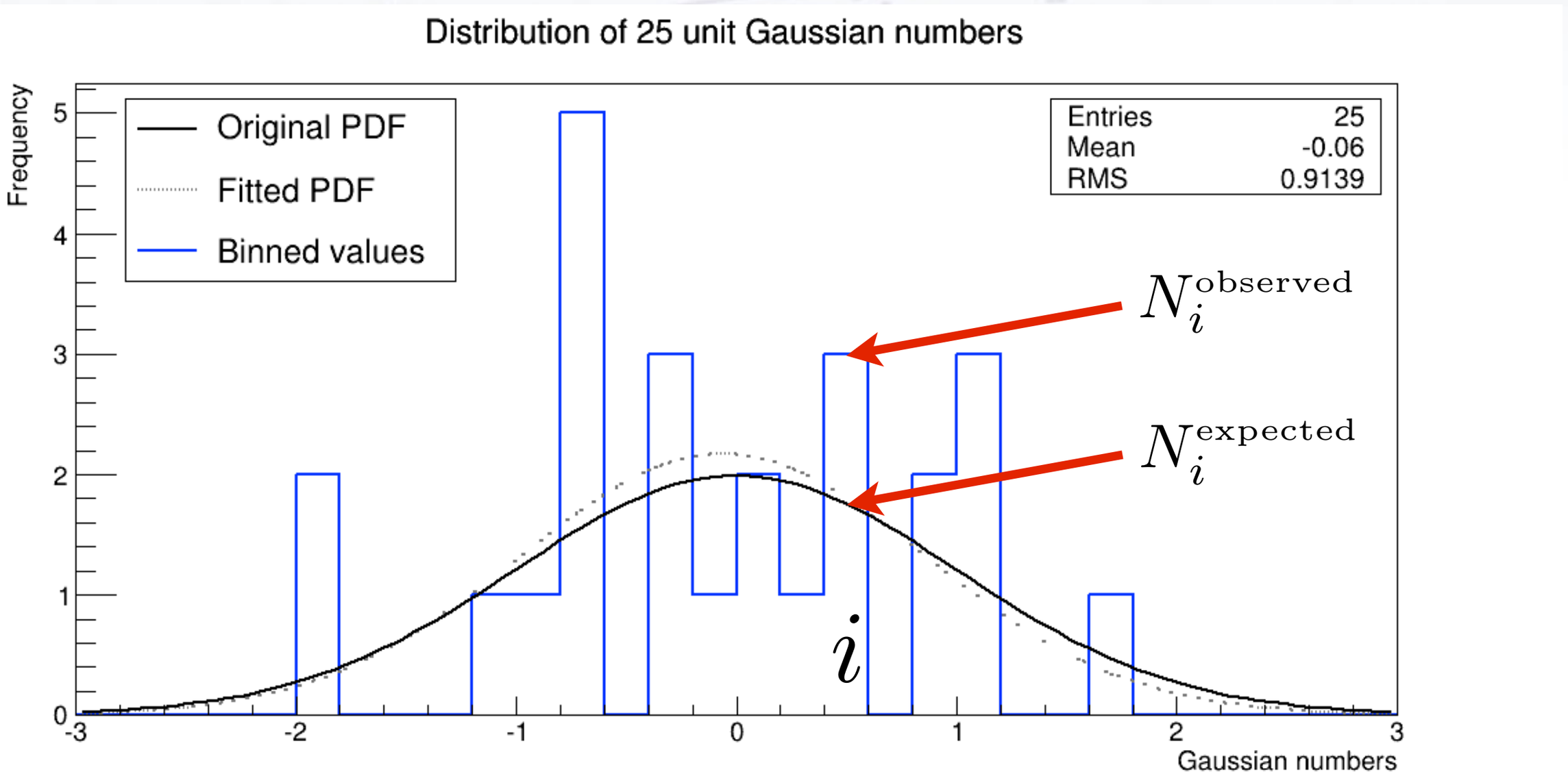
$$\chi^2(\theta) = \sum_i^{N_{\text{bins}}} \left(\frac{N_i^{\text{observed}} - N_i^{\text{expected}}}{\sigma(N_i^{\text{observed}})} \right)^2$$



ChiSquare

Recall, the ChiSquare is a sum over bins in a histogram:

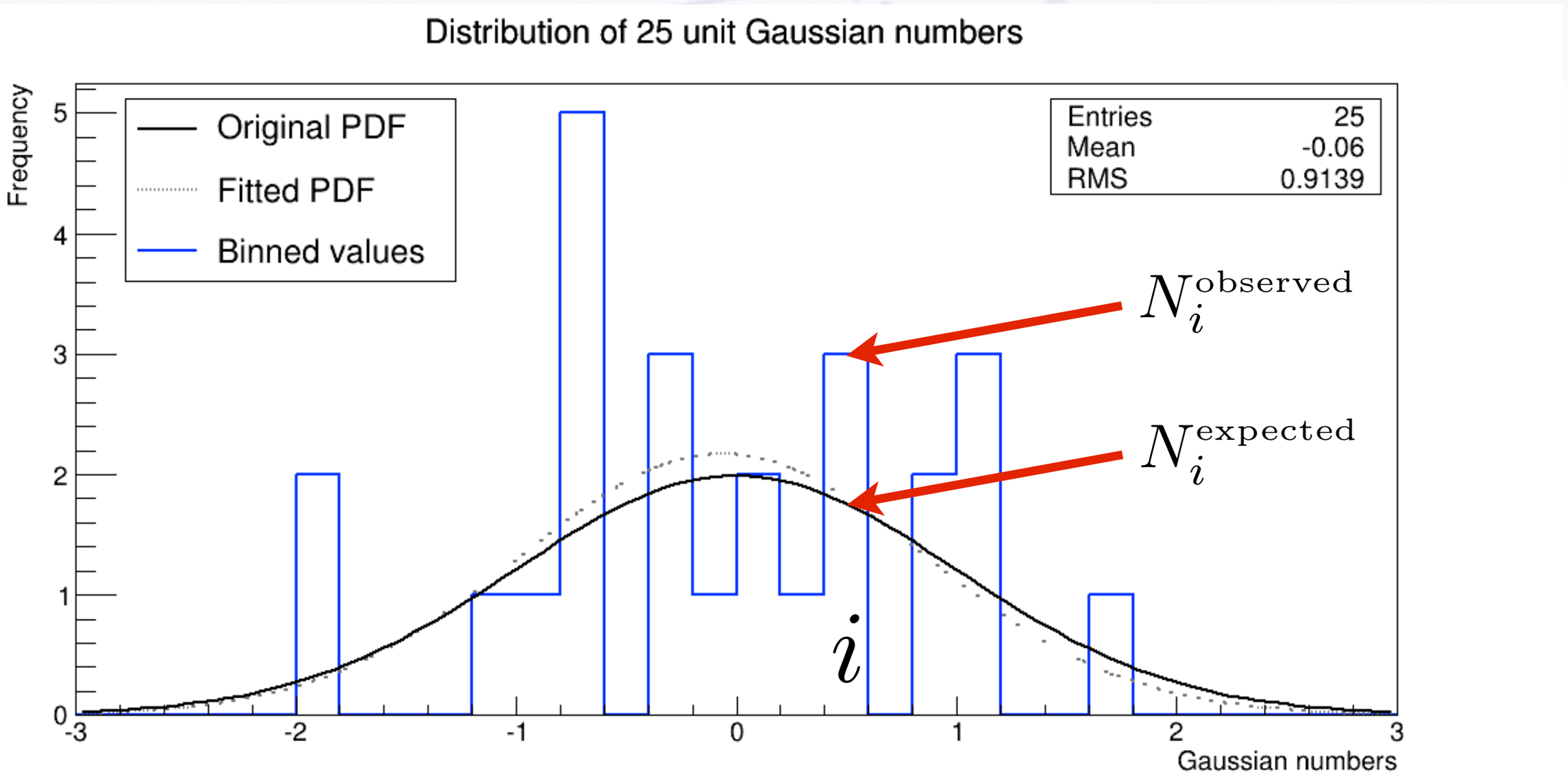
$$\chi^2(\theta) = \sum_i^{N_{\text{bins}}} \frac{(N_i^{\text{observed}} - N_i^{\text{expected}})^2}{N_i^{\text{observed}}}$$



ChiSquare

Recall, the ChiSquare is a sum over bins in a histogram:

$$\chi^2(\theta) = \sum_i^{N_{\text{bins}}} \frac{(N_i^{\text{observed}} - N_i^{\text{expected}})^2}{N_i^{\text{expected}}}$$

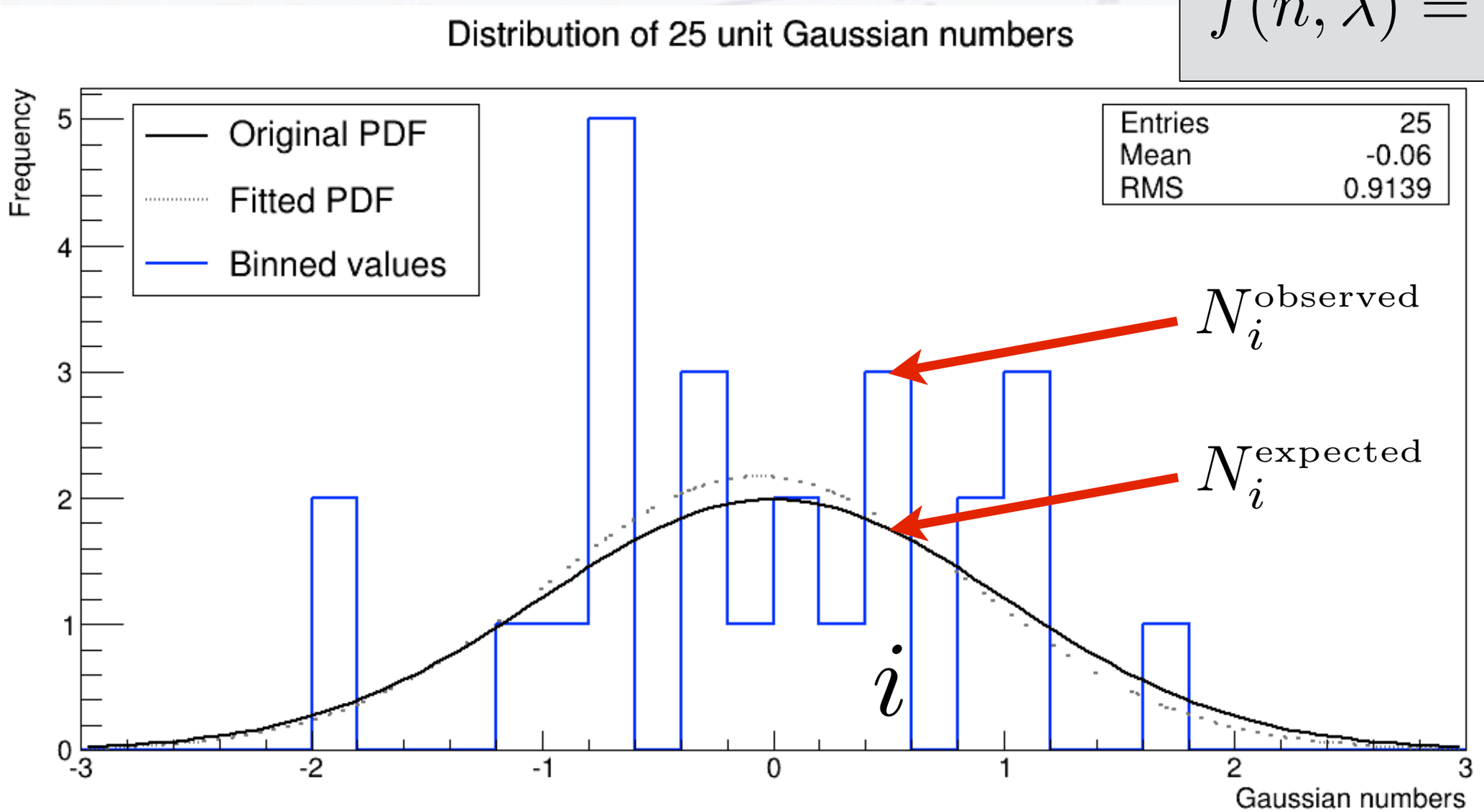


Binned Likelihood

The binned likelihood is a sum over bins in a histogram:

$$\mathcal{L}(\theta)_{\text{binned}} = \prod_{i}^{N_{\text{bins}}} \text{Poisson}(N_i^{\text{expected}}, N_i^{\text{observed}})$$

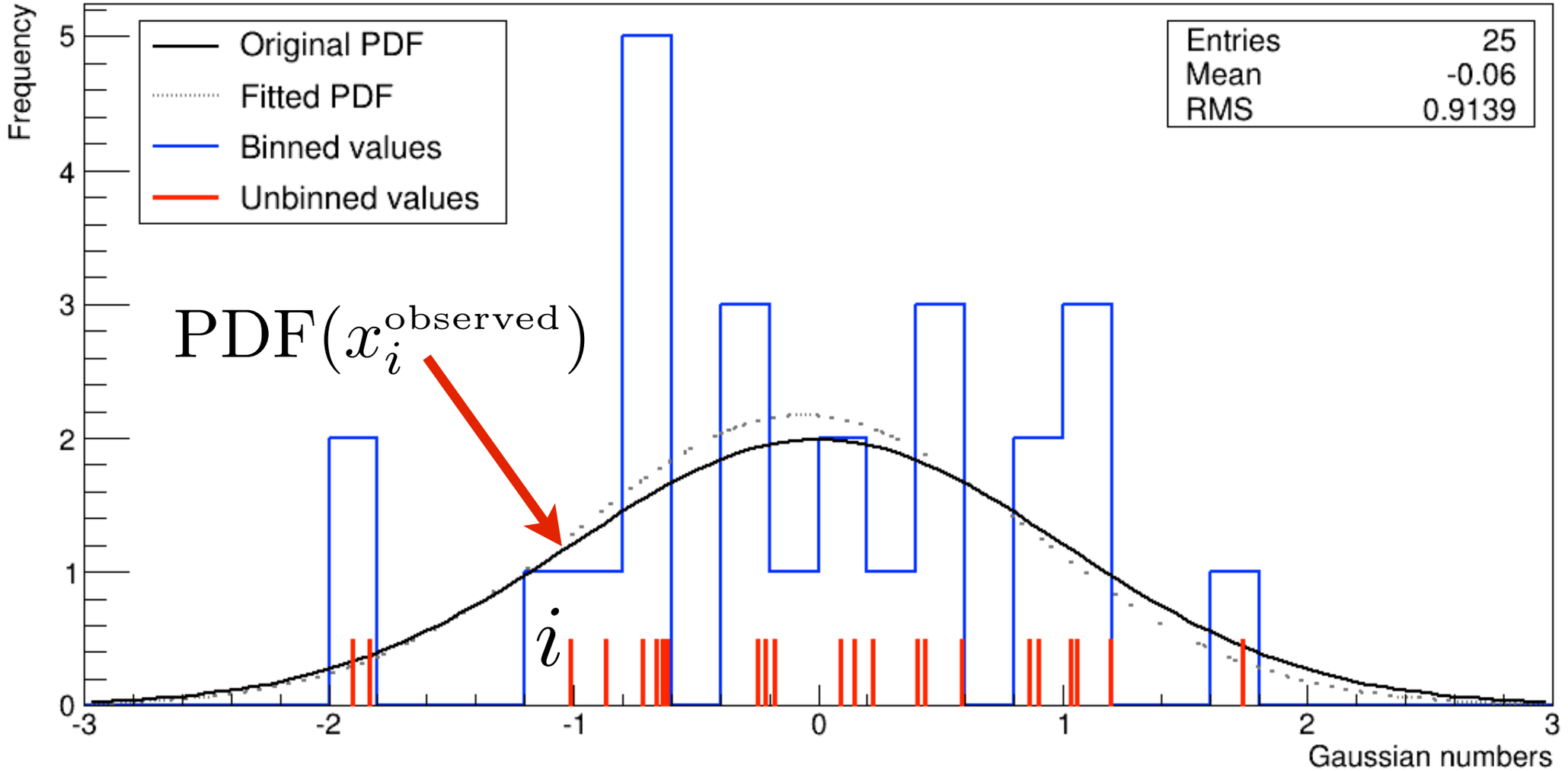
$$f(n,\lambda) = \frac{\lambda^n}{n!} e^{\lambda}$$



Unbinned Likelihood

The binned likelihood is a sum over single measurements:

$$\mathcal{L}(\theta)_{\text{unbinned}} = \prod_{i}^{N_{\text{meas.}}} \text{PDF}(x_i^{\text{observed}})$$



Methods of fitting

In summary, there are four methods of fitting histograms with parameters θ, in order of increasing accuracy, but decreasing speed and convergence:

1. **Minimise the ("Neyman") Chi-Square:**
   Problem: Breaks in empty bins ($N^{obs} = 0$).
   Note: Minuit disregards empty bins!

$$\chi^2(\theta) = \sum_i^{N_{\text{bins}}} \frac{(N_i^{\text{Obs.}} - N_i^{\text{Exp.}}(\theta))^2}{N_i^{\text{Obs.}}}$$

2. **Minimise the ("Pearson") Chi-Square:**
   Minor problem: What range to include?
   Note on 1+2: Applies only to histograms.
   If errors are provided, these are used directly.

$$\chi^2(\theta) = \sum_i^{N_{\text{bins}}} \frac{(N_i^{\text{Obs.}} - N_i^{\text{Exp.}}(\theta))^2}{N_i^{\text{Exp.}(\theta)}}$$

3. **Minimise -2Ln(LLH) of each bin (Poisson):**
   Note: This can be used for low statistics binned data, avoiding the Gaussian approx.

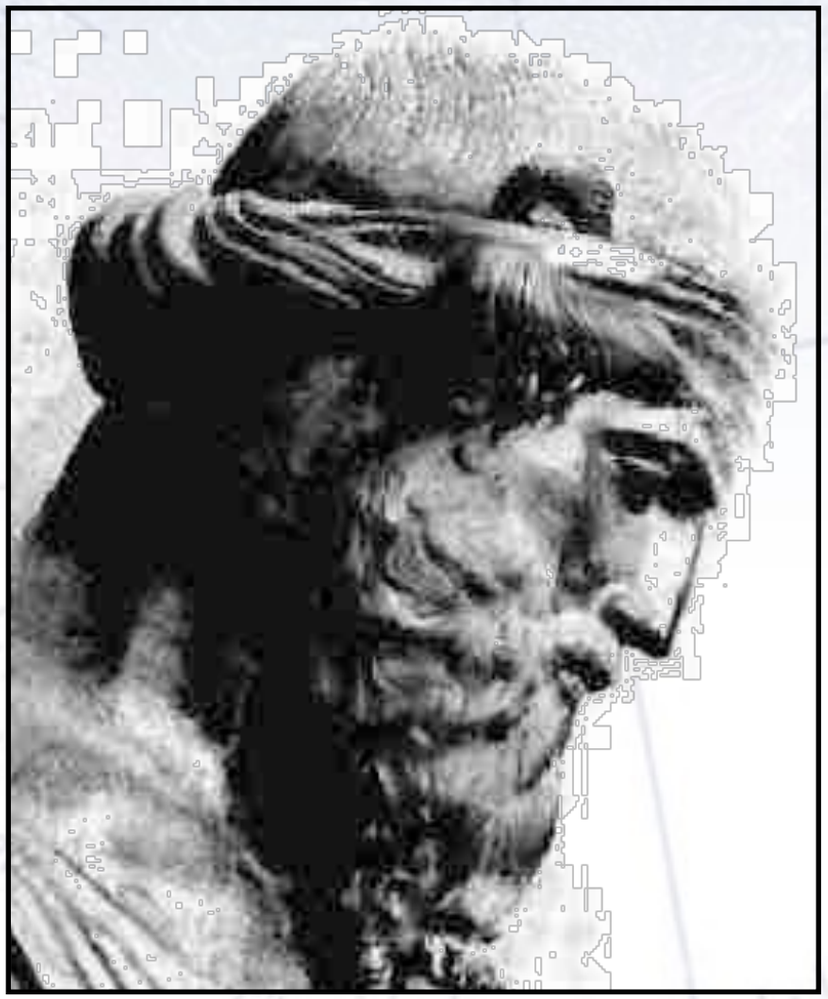
$$-2\ln\mathcal{L}(\theta)_{\text{binned}} = -2\sum_{i\in N_{\text{bins}}} \ln \text{Pois}(N_i^{\text{Obs.}}, N_i^{\text{Exp.}}(\theta))$$

4. **Drop binning and minimise the unbinned -2Ln(LLH) likelihood.**
   Note: Sum runs over events not bins!
   Note: Fit parameters in PDF.

$$-2\ln\mathcal{L}(\theta)_{\text{unbinned}} = -2\sum_{i\in N_{\text{events}}} \ln \text{PDF}(N_i^{\text{Obs.}})$$

The unbinned likelihood is generally the best method in case of low statistics.

Notes on the likelihood



For a large sample, the likelihood is indeed unbiased and has the minimum variance - that is hard to beat! Also, the binned LLH approaches the unbinned version. However...

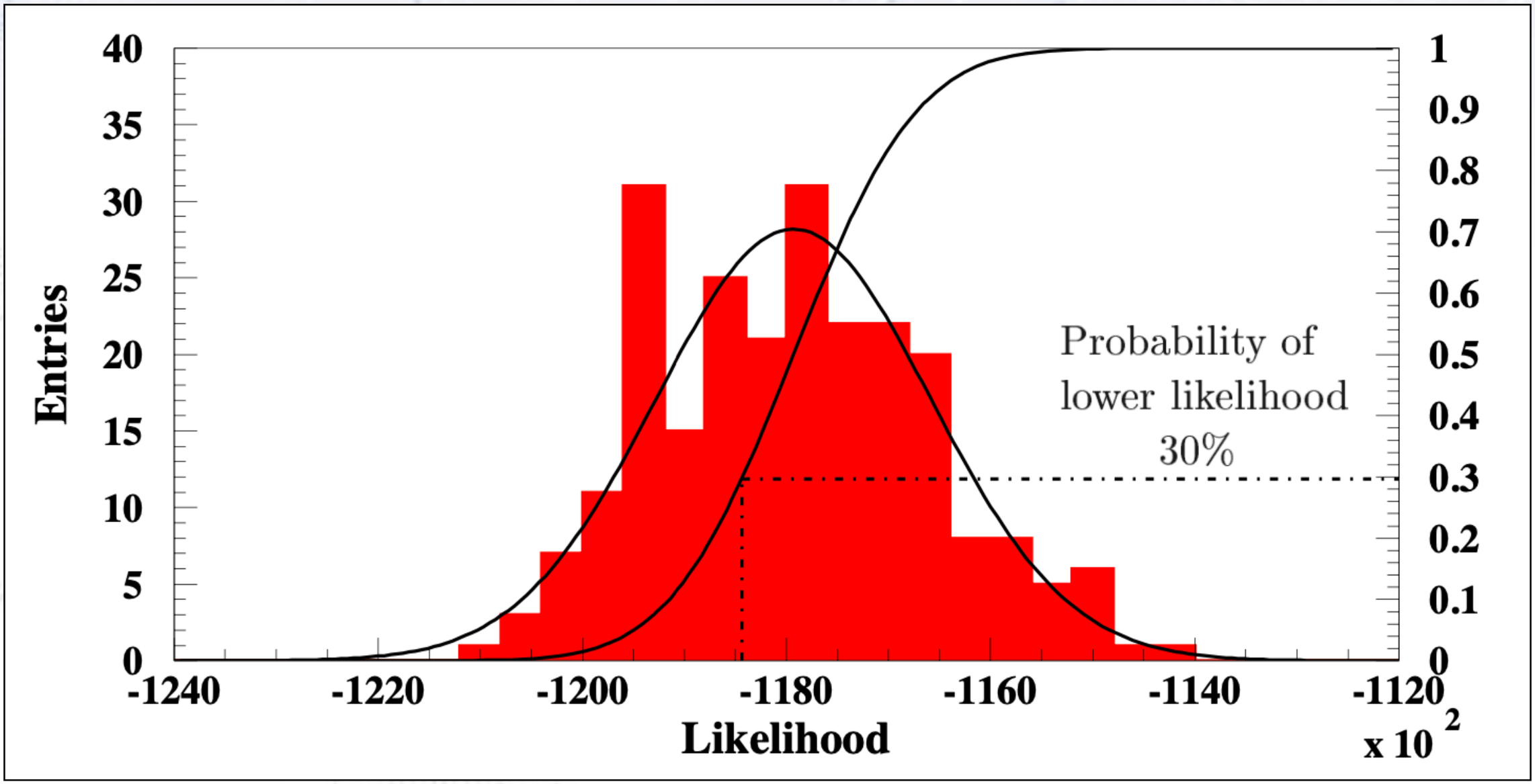
For the likelihood, you have to know your PDF. This is also true for the Chi-Square, but unlike for the Chi-Square, you get no goodness-of-fit measure to check it!

This can be obtained through simulation!

Also, for small statistics, the likelihood is not necessarily unbiased, but still fares much better than the ChiSquare! But be careful with small statistics. The way to avoid this problem is also simulation.

Evaluating the likelihood

As mentioned, it is possible to evaluate a likelihood fit (i.e. get a p-value). This is done by first fitting the data, and then (given the fit parameters) one simulates data according to these parameters, and fit these many times.

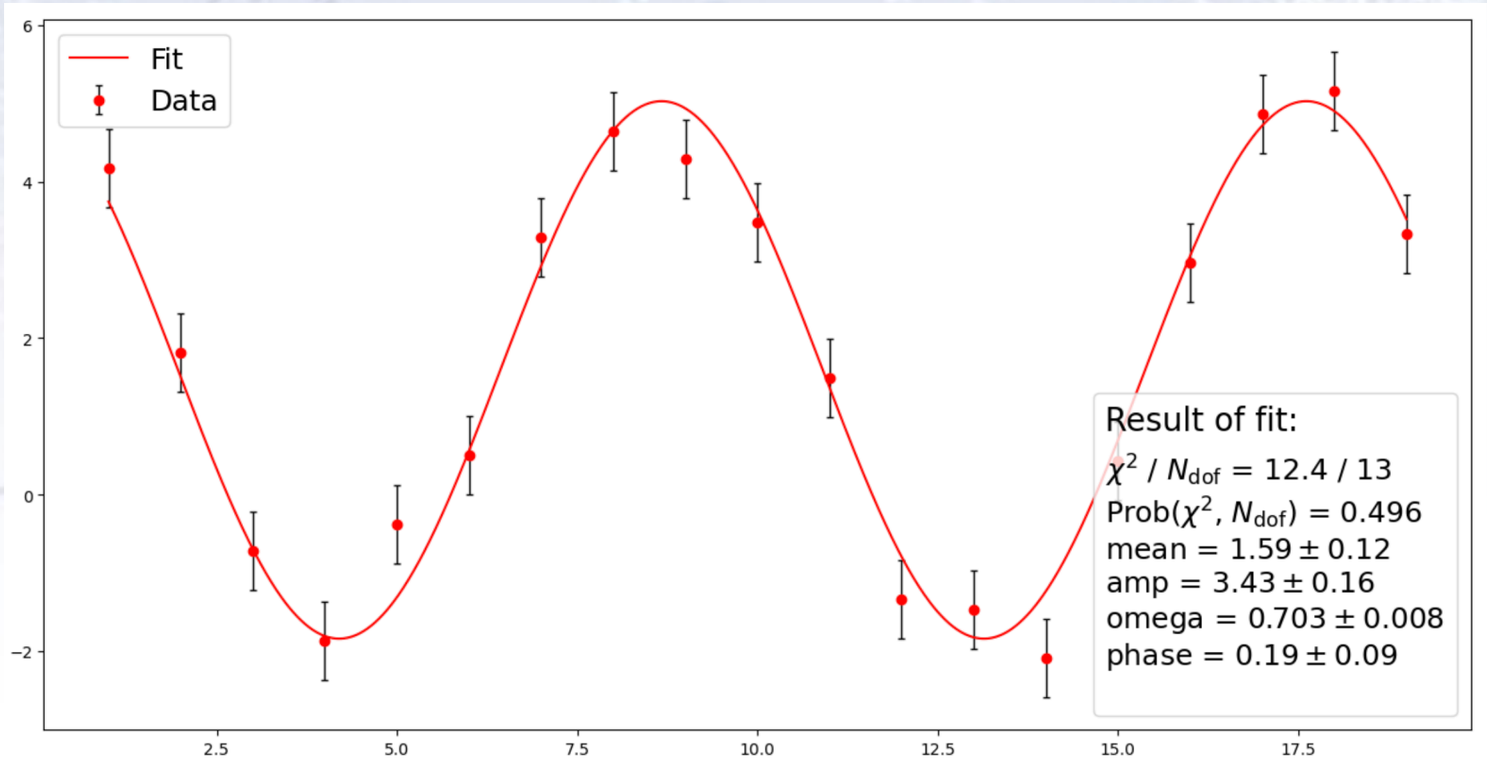


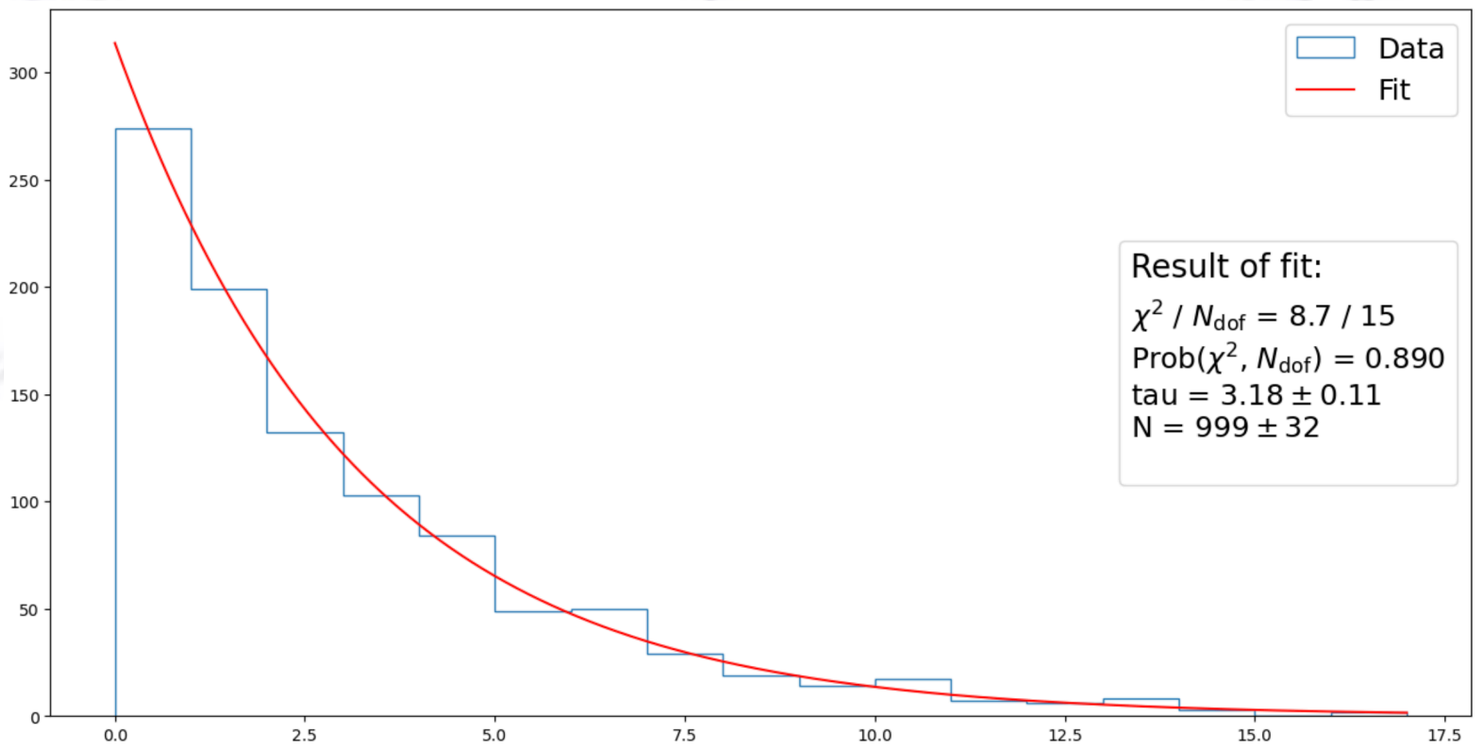
Finally, you see what fraction of the fits have a worse LLH. Note the Gaussian shape!

Application of Likelihood fit

For graphs of data points where the uncertainties are Gaussian, one should fit with a ChiSquare.

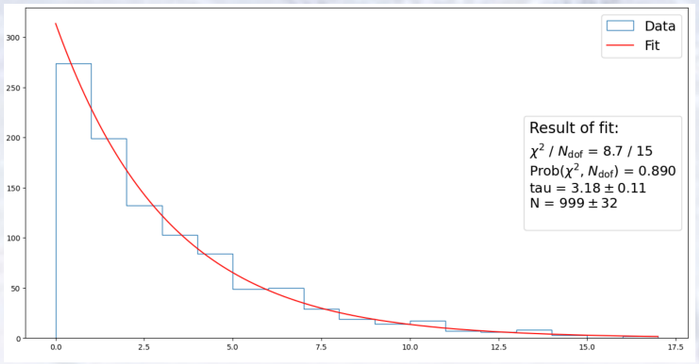
There is no “low statistics” in these cases, as few data points does not make it non-Gaussian.



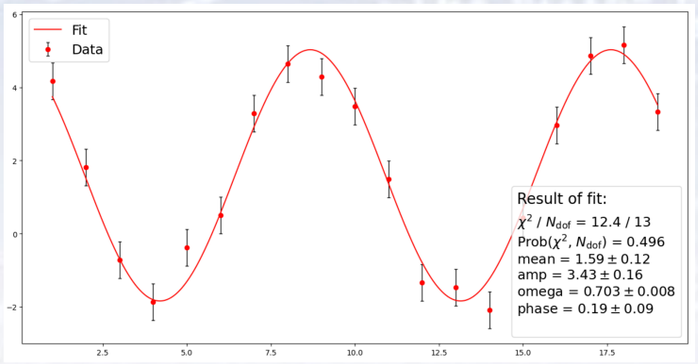


The likelihood fit applies to fitting histograms. It is here, that “low statistics” (in bins) is an issues, since the uncertainties are then not Gaussian.
This makes the likelihood superior in getting the best parameter estimates.

What fitting method?



Data has (Gaussian) uncertainties?



No

Yes

Standard ChiSquare Fit!

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

Put data in histogram. High statistics? (80% bins has 5+ entries?)

Yes

Fixed range? Empty bins?

No

Minuit ChiSquare Fit!

$$\chi^2 = \sum_{i, y_i \neq 0} \frac{(O_i - E_i)^2}{O_i}$$

Yes

Pearson ChiSquare Fit!

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

No

Original data binned?

No

Unbinned LLH fit

$$\mathcal{L} = -2 \sum_i \ln(PDF(x_i))$$

Yes

Binned LLH fit

$$\mathcal{L} = -2 \sum_i \ln(Pois(O_i, E_i))$$

The likelihood ratio test

Not unlike the Chi-Square, were one can compare χ^2 values, the likelihood between two competing hypothesis can be compared (SAME offset constant/factor!).

While their individual LLH values do not say much, their RATIO says everything!

As with the likelihood, one often takes the logarithm and multiplies by -2 to match the Chi-Square, thus the "test statistic" becomes:

$$D = -2 \ln \left(\frac{\text{likelihood for null model}}{\text{likelihood for alternative model}} \right)$$
$$= -2 \ln(\text{likelihood for null model}) + 2 \ln(\text{likelihood for alternative model})$$

If the two hypothesis are simple (i.e. no free parameters) then the **Neyman-Pearson Lemma** states that this is the best possible test one can make.

If the alternative model is not simple but nested (i.e. contains the null hypothesis), this difference approximately behaves like a Chi-Square distribution with N_{dof} = N_{dof}(alternative) - N_{dof}(null). This is called **Wilk's Theorem**.

The likelihood ratio test

Not unlike the Chi-Square, were one can compare χ^2 values, the likelihood between two competing hypothesis can be compared (SAME offset constant/factor!).

While their individual LLH values do not say much, their RATIO says everything!

As with the likelihood, one often takes the logarithm and multiplies by -2 to match the Chi-Square, thus the "test statistic" becomes:

$$D = -2 \ln \left(\frac{\text{likelihood for null model}}{\text{likelihood for alternative model}} \right)$$
$$= -2 \ln(\text{likelihood for null model}) + 2 \ln(\text{likelihood for alternative model})$$

If the two hypothesis are simple (i.e. no free parameters) then the **Neyman-Pearson Lemma** states that this is the best possible test one can make.

If the alternative model is not simple but nested (i.e. contains the null hypothesis), this difference approximately behaves like a Chi-Square distribution with N_{dof} = N_{dof}(alternative) - N_{dof}(null). This is called **Wilk's Theorem**.

Nested Models?

What does this "nested" thing mean in Wilk's Theorem, and why is it essential?

<u>In short:</u>
"Nested" means that the Null hypothesis f(x) is "included" in the Alternative hypothesis g(x), and thus that g(x) can do anything f(x) can… and more!

<u>In terms of function spaces:</u>
Let F and G be the function spaces for all versions of f(x) and g(x), respectively.
Then f(x) is nested in g(x) if F is a subspace of G, i.e. G "contains" F: $F \subset G$

<u>Example:</u>
- Null hypothesis: $f(x) = N_{bkg} * \exp(-x/\tau)$
- Alt. hypothesis: $g(x) = N_{bkg} * \exp(-x/\tau) + N_{sig} * \text{Gaussian}(\mu, \sigma)$

Here, g(x) can do everything that f(x) can do… and more!

The nested requirement ensures that the LLH value of the Alternative Hypothesis is always better (i.e. lower for -2LLH) than for the Null Hypothesis, and hence that the likelihood ratio (which should follow a ChiSquare distribution) is positive!

Wilk's Theorem: Example

Consider two models (hypotheses) for data:
Null: Data is only background (exponential).
Alternative: Data is both signal and background (exponential with Gaussian peak).

The models are not simple but nested: **The alternative contains the null hypothesis**.
Therefore, Wilk's theorem applies.

To do a **hypothesis test** of which model best matches the data, we do two likelihood fits, and calculate (-2 ln of) the ratio of the obtained likelihood values, which behaves like a Chi-Square distribution with $N_{dof} = N_{dof}(\text{alternative}) - N_{dof}(\text{null}) = 3$.

- If the fits give e.g. -2 ln(LLH_alt / LLH_null) = 2.3.
  Then this corresponds to a (null) p-value = ProbChi2(Chi2=2.3, N_{dof}=3) = 0.51
  Thus the null hypothesis can not be rejected (no certainty of signal).

- If the fits give e.g. -2 ln(LLH_alt / LLH_null) = 20.9
  Then this corresponds to a (null) p-value = ProbChi2(Chi2=2.3, N_{dof}=3) = 0.00011
  Thus the null hypothesis can be rejected at 99.989% significance level.