Distributions and Parameter Estimation

Roger Barlow
The University of Huddersfield

LHC physics school, NCP Islamabad

20^{th} August 2025

Probability Distributions

Data values can be: integer (discrete) or real (continuous)
(They may also be ranked or categorical but let's not go there)

Discrete values are described by *probability distributions*: P_r, pure dimensionless numbers

Real values are described by *probability density functions* or pdfs: $P(x)$
$P(x)$ has dimensions $[x]^{-1}$. $\int P(x)\,dx$ or $P(x)\,\Delta x$ are pure numbers

Parton Distribution Functions are also Probability Density Functions so no problem in calling them pdf.

You will also (sometimes) meet the Cumulative Density Function (cdf).
$C(x) = \int_{-\infty}^{x} P(x')\,dx'$

Expectation Values

Unitarity (something has got to happen)
Expressed by $\sum_r P_r = 1$ or $\int_{-\infty}^{\infty} P(x)\,dx = 1$ as appropriate

The average result, the expectation value
Expressed by $\langle r \rangle = \sum_r rP_r$ or $\langle x \rangle = \int_{-\infty}^{\infty} xP(x)\,dx$ as appropriate
Often denoted by μ

Higher Moments: $\mu_n = \langle r^n \rangle$ or $\langle x^n \rangle$
Central moments: $\mu'_n = \langle (r-\mu)^n \rangle$ or $\langle (x-\mu)^n \rangle$
Variance V is just the second central moment..$V = \left\langle (x-\mu)^2 \right\rangle$
Notice $V = \left\langle (x-\mu)^2 \right\rangle = \left\langle x^2 \right\rangle - 2\mu \langle x \rangle + \mu^2 = \left\langle x^2 \right\rangle - \langle x \rangle^2$
V is often written as σ^2. (Physicists prefer σ, statisticians prefer V)
Skew $\gamma = \left\langle (x-\mu)^3 \right\rangle / \sigma^3$ Curtosis $K = \left\langle (x-\mu)^4 \right\rangle / \sigma^4 - 3$
Generally for any $f(x)$: $\langle f(x) \rangle = \sum_r f(r)P_r$ or $\int_{-\infty}^{\infty} f(x)P(x)\,dx$

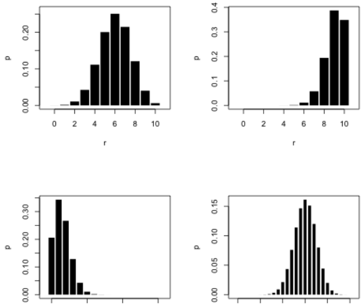
Some people use $E(f)$ rather than $\langle f \rangle$. Be prepared to meet either.

The Binomial Distribution

Binomial: Number of successes in N trials, each with probability p of success

$$P(r; p, N) = \frac{N!}{r!(N-r)!} p^r q^{N-r} \qquad (q \equiv 1 - p)$$

Binomial distributions for
(1) $N = 10, p = 0.6$
(2) $N = 10, p = 0.9$
(3) $N = 15, p = 0.1$
(4) $N = 25, p = 0.6$



Mean $\mu = Np$, Variance $V = Npq$, Standard Deviation $\sigma = \sqrt{Npq}$

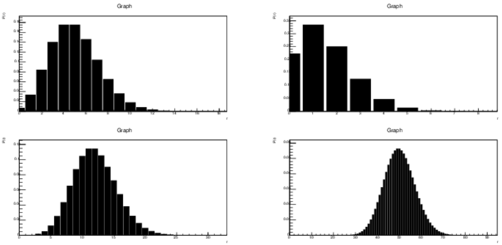
The Poisson Distribution

Number of events occurring at random rate λ

$$P(r;\lambda) = e^{-\lambda}\frac{\lambda^r}{r!}$$

Limit of binomial as $N \to \infty$, $p \to 0$ with $Np = \lambda = constant$:

$\frac{N^r p^r}{r!}\left(1 - \frac{\lambda}{N}\right)^N$



Poisson distributions for
(1) $\lambda = 5$
(2) $\lambda = 1.5$
(3) $\lambda = 12$
(4) $\lambda = 50$

Mean $\mu = \lambda$, Variance $V = \lambda$, Standard Deviation $\sigma = \sqrt{\lambda} = \sqrt{\mu}$

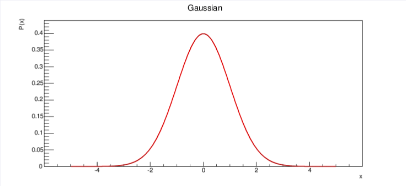
Meet this a lot as it applies to event counts - on their own or in histogram bins

The Gaussian

The Formula

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

The Curve



Only 1 Gaussian curve, as μ and σ are just location and scale parameters

Properties

Mean is μ and standard deviation σ. Skew and kurtosis are 0.

The Central Limit Theorem

Why the Gaussian is so important

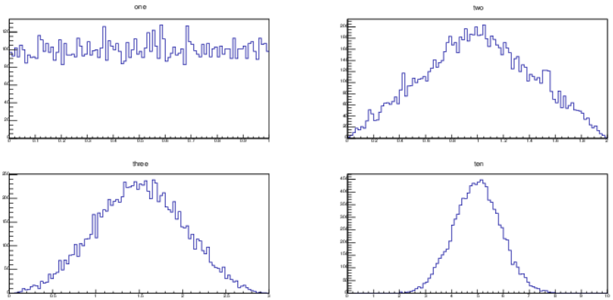
If the variable X is the sum of N independent variables $x_1, x_2 \ldots x_N$ then

1. Means add: $< X >=< x_1 > + < x_2 > + \cdots < x_N >$
2. Variances add: $V_X = V_1 + V_2 + \ldots V_N$
3. If the variables x_i are independent and identically distributed (i.i.d.) then $P(X)$ tends to a Gaussian for large N

(1) is obvious
(2) is pretty obvious, and means that standard deviations add in quadrature, and that the standard deviation of an average falls like $\frac{1}{\sqrt{N}}$
(3) applies whatever the form of the original $p(x)$

Demonstration

Take a uniform distribution from 0 to 1. It is flat. Add two such numbers and the distribution is triangular, between 0 and 2.



With 3 numbers, it gets curved. With 10 numbers it looks pretty Gaussian

Proof

Introduce the Characteristic Function $< e^{ikx} >= \int e^{ikx} P(x)\, dx = \tilde{P}(k)$
Expand the exponential as a series
$< e^{ikx} >=< 1+ikx+\frac{(ikx)^2}{2!}+\frac{(ikx)^3}{3!} \cdots >= 1+ik < x > +(ik)^2\frac{<x^2>}{2!}+(ik^3)\frac{<x^3>}{3!} \dots$
Take logarithm and use expansion $\ln(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} \dots$
this gives power series in (ik), where coefficient $\frac{\kappa_r}{r!}$ of $(ik)^r$ is made up of expectation values of x of total power r
$\kappa_1 =< x >, \kappa_2 =< x^2 > - < x >^2, \kappa_3 =< x^3 > -3 < x^2 >< x > +2 < x >^3$
$\dots$ These are called the Semi-invariant cumulants of Thièle . Under a change of scale α, $\kappa_r \to \alpha^r \kappa_r$. Under a change in location only κ_1 changes.

If X is the sum of i.i.d. random variables: $x_1 + x_2 + x_3 ...$ then $P(X)$ is the convolution of $P(x)$ with itself N times
The FT of convolution is the product of the individual FTs
The logarithm of a product is the sum of the logarithms
So $P(X)$ has cumulants $K_r = N\kappa_r$
To make graphs commensurate, need to scale X axis by standard deviation, which grows like $\sqrt{N}$. Cumulants of scaled graph $K'_r = N^{1-r/2}\kappa_r$
As $N \to \infty$ these vanish for $r > 2$. Leaving a quadratic.
If the log is a quadratic, the exponential is a Gaussian. So $\tilde{P}(X)$ is Gausian.
The FT of a Gaussian is a Gaussian. QED.

Gaussian or Normal?

Even if the distributions are not identical, the CLT tends to apply, unless one (or two) dominates.

Most ‘errors’ fit this, being compounded of many different sources.

Statisticians call it the ‘Normal’ distribution. Physicists don’t. But be prepareZZd.

Summary on Distributions

You will use the Gaussian constantly, and the Poisson very often.
There are other distributions, but you can look up their properties: pdf, cdf, variance[1], mean[2] etc.

[1] Except for the Cauchy/Lorentz/Breit Wigner
[2] Except for the Landau

Fitting
or
Estmation

What's happening

You have a dataset $\{x_1, x_2, \ldots x_N\}$
and a pdf $P(x, a)$ with unknown parameter(s) a

You want to know:

1. What is the value for a according to the data?
2. What is the error on that value?
3. Does the resulting $P(x, a)$ actually describe the data?

This is called 'estimation' by statisticians and 'fitting' by physicists

Also applies when finding a property rather than a parameter, and then sometimes when one has a parent population rather than a pdf

General considerations

An Estimator is a function of all the x_i which returns some value for a
Write $\hat{a}(x_1, x_2, \ldots x_N)$
There is no 'correct' estimator. You would like an estimator to be

- Consistent: $\hat{a}(x) \to a$ for $N \to \infty$
- Unbiassed: $\langle \hat{a} \rangle = a$
- Efficient: $V(\hat{a}) = \left\langle \hat{a}^2 \right\rangle - \langle \hat{a} \rangle^2$ is small
- Invariant under reparameterisation: $\widehat{f(a)} = f(\hat{a})$
- Convenient

But no estimator is perfect, and these requirements are self-contradictory

Bias: a simple example

Suppose you want to estimate the mean $\mu \equiv \langle x \rangle$ for some pdf, and you choose $\hat{\mu} = \overline{x} = \frac{1}{N}\sum_i x_i$
Then $\langle \hat{\mu} \rangle = \frac{1}{N}\sum_i \langle x_i \rangle = \frac{1}{N}\sum_i \langle x \rangle = \langle x \rangle$. Zero bias.

Suppose you want to estimate the variance $V \equiv \left\langle x^2 \right\rangle - \langle x \rangle^2$ for some pdf, and you choose $\hat{V} = \overline{x^2} - \overline{x}^2 = \frac{1}{N}\sum_i x_i^2 - \left(\frac{1}{N}\sum_i x_i\right)^2$
$\hat{V} = \frac{N-1}{N^2}\sum_i x_i^2 - \frac{1}{N^2}\sum_i \sum_{j \neq i} x_i x_j$
Take expectation values. $\left\langle \hat{V} \right\rangle = \frac{N-1}{N}\left\langle x^2 \right\rangle - \frac{N(N-1)}{N^2}\langle x \rangle^2 = \frac{N-1}{N} V$
The 'obvious' $\hat{V}$ underestimates the true V.

- This is understandable: a fluctuation drags the mean with it, so variations are less
- This can be corrected for (Bessel's correction) by an $N/(N-1)$. Many statistical calculators offer σ_n and σ_{n-1}
- This correction cures the bias for V. Actually σ is still biassed. But V is more useful.
- Biasses are typically small and correctable

Efficiency is limited

The Minimum Variance Bound

If $\hat{a}$ is unbiassed

$$V(\hat{a}) \geq \left\langle \left(\frac{\partial \ln L}{\partial a}\right)^2 \right\rangle^{-1} = \left\langle -\frac{\partial^2 \ln L}{\partial a^2} \right\rangle^{-1}$$

also named for Cramér, Rao, Fréchet, Darmois, Aitken and Silverstone (equivalent form exists if there is a bias)

$$L(x_1, x_2 ... x_n; a) = P(x_1; a) \times P(x_2; a) ... \times P(x_n; a)$$

Same as likelihood featuring in Bayes' theorem, though emphasis here is that L is Likelihood for *all* measurements of sample

Fun algebra with the likelihood function

Writing. $\int ... \int dx_1 ... dx_n$ as just $\int dx$

	Unitarity	No bias: $\langle \hat{a} \rangle = a$
Start with	$\int L(x;a)\,dx = 1$	$\int \hat{a}(x) L(x;a)\,dx = a$
Differentiate	$\int \frac{\partial L}{\partial a}\,dx = 0$	$\int \hat{a}(x) \frac{\partial L}{\partial a}\,dx = 1$
Chain rule	$\int L \frac{\partial \ln L}{\partial a}\,dx = 0^*$	$\int \hat{a}(x) L \frac{\partial \ln L}{\partial a}\,dx = 1$

Multiply column 1 by a and subtract from column 2: $\int (\hat{a} - a) \frac{\partial \ln L}{\partial a} L\,dx = 1$

Invoke Schwarz' lemma $\int u^2\,dx \times \int v^2\,dx \geq \left(\int uv\,dx\right)^2$

with $u \equiv (\hat{a} - a)\sqrt{L},\ v \equiv \ \frac{\partial \ln L}{\partial a}\sqrt{L}$

$\int (\hat{a} - a)^2 L\,dx. \times \int \left(\frac{\partial \ln L}{\partial a}\right)^2 L\,dx \geq 1$

or $\left\langle (\hat{a} - a)^2 \right\rangle \left\langle \left(\frac{\partial \ln L}{\partial a}\right)^2 \right\rangle \geq 1$

$V_{\hat{a}} \geq \frac{1}{\left\langle \left(\frac{\partial \ln L}{\partial a}\right)^2 \right\rangle}$

Finally, differentiate Eq. *: $\left\langle \left(\frac{\partial \ln L}{\partial a}\right)^2 \right\rangle + \left\langle \frac{\partial^2 \ln L}{\partial a^2} \right\rangle = 0$ (Fisher information)

Maximum likelihood estimation

If $\{x_i\}$ are independent then the likelihood $L(x_1, x_2 \ldots x_N|a) = \prod_i P(x_i|a)$

The ML estimator

To estimate a using data $\{x_1, x_2 \ldots x_N\}$, find the value of a for which the total log likelihood $\sum \ln P(x_i|a)$ is maximum.

3 types of problem

1. Differentiate, set to zero, solve the equation(s) algebraically
2. Differentiate, set to zero, solve the equation(s) numerically
3. Maximise numerically

Things to note

- No deep justification for ML estimation, except that it works well
- These are not 'the most likely values' of a. They are the values of a for which the values of x are most likely
- The logs make the total a sum, easier to handle than a product
- Remember a minus sign if you use a minimiser

Maximum likelihood estimation

- Consistent: Almost always
- Unbiassed; It is biassed. But the bias usually falls like $1/N$
- Efficient: In the large N limit ML saturates the MVB, and you can't do better than that
- Invariant under reparameterisation: clearly.
- Convenient. Usually

It is used for

- Basic unbinned raw-measurement data
- Data with structure, such as binned histograms

but the same principles apply to both.
Start with the first

Simple Examples(1): Gaussian mean and standard deviation

$\{x_i\}$ have been gathered from a Gaussian. What are the ML estimates for μ and σ?

$\ln L = \sum -\frac{1}{2}((x_i - \mu)/\sigma)^2 - N \ln(\sqrt{2\pi}\sigma)$

Differentiating wrt μ and σ and setting to zero gives 2 equations
$\sum_i (x_i - \hat{\mu})/\hat{\sigma}^2 = 0 \qquad \sum (x_i - \hat{\mu})^2/\hat{\sigma}^3 - N/\hat{\sigma} = 0$

which are happily decoupled and give
$\hat{\mu} = \frac{1}{N}\sum_i x_i, \qquad \hat{\sigma}^2 = \frac{1}{N}\sum (x_i - \hat{\mu})^2$ (!)

Simple Examples(2): Lifetime

Measuring a lifetime from a sample of decay times t_i: $P(t|\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$
Log likelihood. $\sum_{i=1}^{N} -\frac{t_i}{\tau} - N \ln \tau$
Differentiate and set to zero $\sum \frac{t_i}{\hat{\tau}^2} - \frac{N}{\hat{\tau}} = 0$ so $\hat{\tau} = \frac{1}{N} \sum t_i$

Exercise for the student

Working instead with the decay rate λ, so $P(t|\lambda) = \lambda e^{-\lambda t}$, you can show that $\hat{\lambda} = N / \sum t_i = 1/\hat{\tau}$..... Invariance under reparameterisation

Make it a bit more realistic (and complicated). Suppose that your apparatus does not record decays with $t > T$.

$P(t|\tau) = \frac{1}{\tau} \frac{e^{-t/\tau}}{1 - e^{-T/\tau}}$

$\ln L = -\sum \frac{t_i}{\tau} - N \ln \tau - N ln(1 - e^{-T/\tau})$

$0 = \frac{\partial \ln L}{\partial \tau}|_{\hat{\tau}} = \sum \frac{t_i}{\hat{\tau}^2} - \frac{N}{\hat{\tau}} + \frac{N T e^{-T/\hat{\tau}}}{\hat{\tau}^2 (1 - e^{-T/\hat{\tau}})}$

$\hat{\tau} = \frac{1}{N} \sum_i t_i + \frac{T}{e^{T/\hat{\tau}} - 1}$

Solve numerically - perhaps by repeated iteration

Simple Examples(3). Signal plus background

Suppose x_i have been gathered from $P(x; a) = aS(x) + (1 - a)B(x)$

$$\ln L = \sum_i \ln(aS(x_i) + (1 - a)B(x_i))$$

Shows typical pdf for Signal+Background

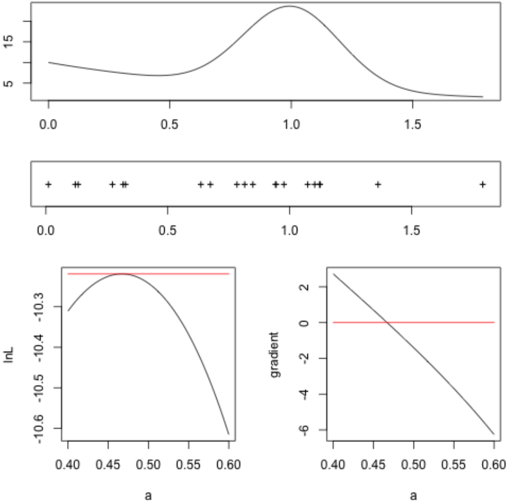
And typical data

And $\ln L$ with its maximum

Quicker and cleaner (if possible) to differentiate and set to zero

$$\sum \frac{S(x_i) - B(x_i)}{\hat{a}S(x_i) + (1 - \hat{a})B(x_i)} = 0$$

but still needs numerical solution



Errors from ML

To first order, looking at the difference between the true a_0 and the estimated $\hat{a}$

$0 = \frac{\partial \ln L}{\partial a}\Big|_{a=\hat{a}} = \frac{\partial \ln L}{\partial a}\Big|_{a=a_0} + (\hat{a} - a_0)\frac{\partial^2 \ln L}{\partial a^2}\Big|_{a=a_0}$

Deviations of $\hat{a}$ from a_0 are due to deviations of $\frac{\partial \ln L}{\partial a}\Big|_{a=a_0}$ from zero, divided by the second derivative

$V(\hat{a}) = V(\frac{\partial \ln L}{\partial a}\Big|_{a=a_0})/\left(\frac{\partial^2 \ln L}{\partial a^2}\Big|_{a=a_0}\right)^2 = \left\langle \left(\frac{\partial lnL}{\partial a}\right)^2 \right\rangle\Big|_{a=a_0}/\left(\frac{\partial^2 \ln L}{\partial a^2}\Big|_{a=a_0}\right)^2$

Which is all very well, but we don't know what a_0 is...

Approximate by using the actual value of our $\hat{a}$: $V(\hat{a}) = -\left(\frac{\partial^2 \ln L}{\partial a^2}\right)^{-1}$

Noter that this is the MVB (in this approximation). ML is efficient

So the error is given by the second derivative of the log likelihood

---

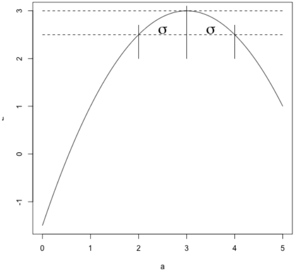
How to find the second derivative (one way anyway)

$\ln L(a) = \ln L(\hat{a}) + \frac{1}{2}(a - \hat{a})^2 \frac{\partial^2 \ln L}{\partial a^2}$.... (first derivative is zero)

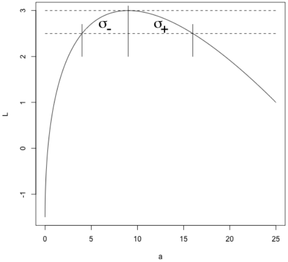
$\quad = \ln L(\hat{a}) - \frac{1}{2}\left(\frac{a-\hat{a}}{\sigma_a}\right)^2$

At $a = \hat{a} \pm \sigma_a$, $\ln L = \ln L(\hat{a}) - \frac{1}{2}$. $\Delta \ln L = -\frac{1}{2}$ gives the error

ML errors



Basic lnL errors
The interval $[\hat{a} - \sigma_a, \hat{a} + \sigma_a]$ from the $\Delta \ln L = -\frac{1}{2}$ points is a 68% central confidence interval

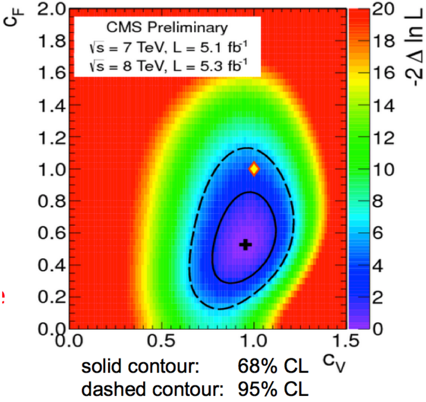


Asymmetric errors (messy!)
If a monotonically reparameterised as $f(a)$, the ML estimate is $\hat{f} = f(\hat{a})$.
$[f(\hat{a} - \sigma_a), f(\hat{a} + \sigma_a)] = [\hat{f} - \sigma_f^-, \hat{f} + \sigma_f^+]$ is the 68% central confidence region.
If $\ln L(a)$ not symmetric, assume this is what is happening and quote separate σ^+, σ^-.
For details see preprint to appear in N.I.M.
`https://arxiv.org/abs/2411.15499`

ML errors

More than one parameter

For 2 (or more) unknown parameters use same technique to map out 68% (or whatever) confidence egions
Only difference is that $\Delta \ln L$ is different.
Given by cumulative probability for χ^2 distribution with 2 (or whatever) degrees of freedom
(χ^2 details coming up)



Fitting data points

Suppose your data is a set of x_i, y_i pairs with predictions $y_i = f_i = f(x_i; a)$
x_i known precisely, y_i measured with Gaussian errors σ_i

- Usually one quantity can be precisely specified
- The σ_i may all be the same. If so, the algebra is easier
- The likelihood is the product of Gaussians $\frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}((y_i - f(x_i, a))/\sigma_i)^2}$

$\ln L = -\frac{1}{2} \sum_i (\frac{y_i - f_i}{\sigma_i})^2$+boring constants
Introduce $\chi^2 = \sum_i (\frac{y_i - f(x_i, a)}{\sigma_i})^2$
Maximum Likelihood $\to$ minimum χ^2. ('Method of Least Squares')
If f is linear in a (e.g. $f(x) = a_1 + a_2 x + a_3 \sqrt{x}$) then this gives a set of equations soluble in one step. If more complicated, need to iterate.

Very simple example: the straight line fit

$$f(x) = a_1 + a_2 x$$

Simple case: all σ_i the same

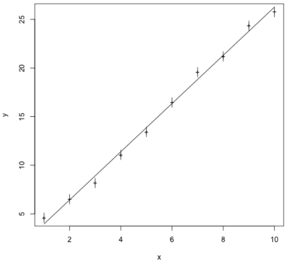
$\chi^2 = \sum \left(\frac{y_i - a_1 - a_2 x_i}{\sigma}\right)^2$

Differentiate and set to zero.

2 Equations

$\sum y_i - a_1 - a_2 x_i = 0$

$\sum x_i (y_i - a_1 - a_2 x_i) = 0$



Simple to unscramble by hand

First is $a_1 = \overline{y} - a_2 \overline{x}$

Substitute in 2nd and get $a_2 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}$

In more general cases, write these as matrices

Linear Regression

Such straight line fits are linked to the statistical modelling technique of 'linear regression' . The formulæ are the same.

But there are subtle differences

Goodness of fit

Does the model $f(x; a)$ provide a good description of the y_i?

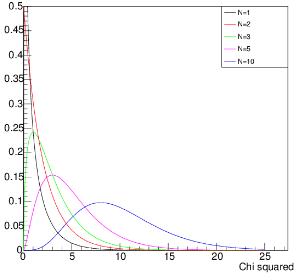
Naïvely each term in χ^2 sum ≈ 1

More precisely:

$p(\chi^2, N) = \frac{1}{2^{N/2}\Gamma(N/2)}\chi^{N/2-1}e^{-\chi^2/2}$

Distribution as N dimensional Gaussian, integrated over hypersphere

Quantify by p-value: probability that, if the model is true, χ^2 would be this large, or larger



Each fitted parameter reduces the effective number by 1. (A linear constraint reduces the dimensionality of the hyperspace by 1).

Degrees of freedom $N_D = N - N_f$

Goodness of fit

Reasons for large χ^2:

- Bad theory
- Bad data
- Errors underestimated
- Unsuspected negative correlation between data points (unlikely)
- Bad luck

Reasons for small χ^2:

- Errors overestimated
- Unsuspected positive correlation between data points (more likely)
- Good luck

Although $-\frac{1}{2}\chi^2$ is a log likelihood, $-2 \ln L$ is not a χ^2. It tells you nothing about goodness of fit.
(Wilks' theorem says it does for differences in similar models. Useful for comparisons but not absolute.)

Using Toy Monte-Carlo for Likelihood and goodness of fit

Obvious suggestion: Take the fitted model, run many simulations, plot the spread of fitted likelihoods and use to get $p-$value
This is wrong - J G Heinrich, CDF/MEMO/BOTTOM.CDFR/5630[3]
Test case: model simple exponential $P(t) = \frac{1}{\tau} e^{-t/\tau}$
Then whatever the original sample looks like you get
Log Likelihood $= \sum(-t_i/\tau - \ln\tau) = -N(\overline{t}/\tau + \ln\tau)$
ML gives $\hat{\tau} = \overline{t} = \frac{1}{N}\sum_i t_i$
and this max log likelihood is $\ln L(\hat{\tau}; x) = -N(1 + \ln\overline{t})$
Any distribution with the same $\overline{t}$ has the same likelihood, after fitting.

What you can do: Histogram the $p(x_i; \hat{a})$ values. This should be flat (almost- the fitting will distort it).
If not enough data - cumulative plot should be straight line. Use max deviation as test statistic. Apply K-S test or use toy Monte Carlo.

[3] Many thanks to Jonas Rademacker for pointing this out

4 ways of fitting data

- Full ML. Write down the likelihood and maximise $\sum_j \ln P(x_j, a)$ where j runs over all events. Slow for large data samples, and no goodness of fit.
- Binned ML. Put it in a histogram and maximise the log of the Poisson probabilities $\sum_i n_i \ln f_i - f_i$ where i runs over all bins $f_i = NP(x_i)w$: don't forget the bin width w. Quicker - but lose info from any structure smaller than bin size
- Put it in a histogram and minimise $\chi^2 = \sum_i (n_i - f_i)^2/f_i$ (Pearson's χ^2). This assumes the Poisson distributions are approximated by Gaussians so do not use if bin contents small . But you do get a goodness of fit.
- Put it in a histogram and minimise $\chi^2 = \sum_i (n_i - f_i)^2/n_i$ (Neyman's χ^2). This makes the algebra and fitting a lot easier. But introduces bias as downward fluctuations get more weight. And disaster if any $n_i = 0$

So there are many ways and they are not all equivalent: choose carefully!