Lecture 10: Maximum likelihood IV.
(nonlinear least square fits)

χ^2 fitting procedure!

Multivariate normal distribution

correlated data

Multivariate Normal Distributions

components x_i of vector **x** are correlated random variables

Generalizes Normal (Gaussian) to M-dimensions
Like 1-d Gaussian, completely defined by its mean and (co-)variance
Mean is a M-vector, covariance is a M x M matrix

$$N(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}\det(\boldsymbol{\Sigma})^{1/2}}\exp[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]$$

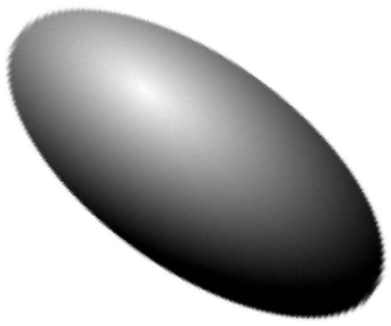
normalize distribution in all components x_i

The mean and covariance of r.v.'s from this distribution are*

$$\boldsymbol{\mu} = \langle\mathbf{x}\rangle \qquad \boldsymbol{\Sigma} = \left\langle(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right\rangle$$

$\Sigma_{ij} = <(x_i-\mu_i)(x_j-\mu_j)>$ matrix notation

In the one-dimensional case σ is the standard deviation, which can be visualized as "error bars" around the mean.



In more than one dimension Σ can be visualized as an error ellipsoid around the mean in a similar way.

$$1 = (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

Multivariate normal distribution

correlated data

Question: What is the generalization of

$$\chi^2 = \sum_i \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2, \qquad x_i \sim \mathrm{N}(\mu_i, \sigma_i)$$

to the case where the x_i's are normal, **but not independent**?
I.e., **x** comes from a multivariate Normal distribution?

How accurately are the fitted parameters determined?
As Bayesians, we would **instead** say, what is their posterior distribution?

Taylor series:

$$-\tfrac{1}{2}\chi^2(\mathbf{b}) \approx -\tfrac{1}{2}\chi^2_{\text{min}} - \tfrac{1}{2}(\mathbf{b}-\mathbf{b}_0)^T \left[\frac{1}{2}\frac{\partial^2 \chi^2}{\partial \mathbf{b} \partial \mathbf{b}}\right] (\mathbf{b}-\mathbf{b}_0)$$

x → b

So, while exploring the χ^2 surface to find its minimum, we must also calculate the Hessian (2nd derivative) matrix at the minimum.

Then

$$P(\mathbf{b}|\{y_i\}) \propto \exp\left[-\tfrac{1}{2}(\mathbf{b}-\mathbf{b}_0)^T \boldsymbol{\Sigma}_b^{-1} (\mathbf{b}-\mathbf{b}_0)\right] \cancel{P(\mathbf{b})}$$

with

$$\boldsymbol{\Sigma}_b = \left[\frac{1}{2}\frac{\partial^2 \chi^2}{\partial \mathbf{b} \partial \mathbf{b}}\right]^{-1}$$

covariance (or "standard error") matrix of the fitted parameters

Notice that if (i) the Taylor series converges rapidly and (ii) the prior is uniform, then the posterior distribution of the **b**'s is multivariate Normal, a very useful CLT-ish result!

correlated data in general multivariate distribution:

The covariance matrix is a more general idea than just for multivariate Normal.
You can compute the covariances of any set of random variables.
It’s the generalizaton to M-dimensions of the (centered) second moment Var.

$$\mathrm{Cov}\ (x, y) = \langle (x - \overline{x})(y - \overline{y}) \rangle$$

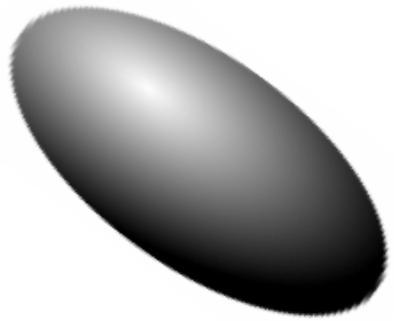
For multiple r.v.’s, all the possible covariances form a (symmetric) matrix:

$$\mathbf{C} = C_{ij} = \mathrm{Cov}\ (x_i, x_j) = \langle (x_i - \overline{x_i})(x_j - \overline{x_j}) \rangle$$

Notice that the diagonal elements are the variances of the individual variables.

The variance of any linear combination of r.v.’s is a quadratic form in $\mathbf{C}$:

$$\begin{aligned} \mathrm{Var}\ \left(\sum \alpha_i x_i\right) &= \left\langle \sum_i \alpha_i (x_i - \overline{x_i}) \sum_j \alpha_j (x_j - \overline{x_j}) \right\rangle \\ &= \sum_{ij} \alpha_i \left\langle (x_i - \overline{x_i})(x_j - \overline{x_j}) \right\rangle \alpha_j \\ &= \boldsymbol{\alpha}^T \mathbf{C}\, \boldsymbol{\alpha} \end{aligned}$$



This also shows that $\mathbf{C}$ is positive definite, so it can still be visualized as an ellipsoid in the space of the r.v.’s., where the directions are the different linear combinations.

correlated data in general multivariate distribution:

The covariance matrix is closely related to the linear correlation matrix.

$$r_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}$$

more often seen written out as

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}}$$

When the null hypothesis is that X and Y are independent r.v.'s, then r is useful as a p-value statistic ("test for correlation"), because

1. For large numbers of data points *N*, it is normally distributed,

   $r \sim \mathrm{N}(0, N^{-1/2})$

   so $r\sqrt{N}$ is a normal t-value

2. Even with small numbers of data points, if the underlying distribution is multivariate normal, there is a simple form for the p-value (comes from a Student t distribution).

χ^2 distribution goodness of fit

we have assumed that, for some value of the parameters $\mathbf{b}$ the model $y(\mathbf{x}_i|\mathbf{b})$ is correct

Suppose that the model $y(\mathbf{x}_i|\mathbf{b})$ does fit. This is the null hypothesis.

Then the "statistic" $\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i} \right)^2$ is the sum of N t^2-values.

(not quite)

So, if we imagine repeated experiments (which Bayesians refuse to do), the statistic should be distributed as Chisquare(N).

If our experiment is very unlikely to be from this distribution, we consider the model to be disproved. In other words, it is a p-value test.

χ^2 distribution (from Lecture 9)

$$p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad \Rightarrow \qquad x \sim \mathrm{N}(0,1)$$

$$y = x^2$$

$$p_Y(y)\,dy = 2p_X(x)\,dx$$

$$p_Y(y) = y^{-1/2} p_X(y^{1/2}) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y}$$

χ^2 is a “statistic” defined as the sum of the squares of n independent t-values.

$$\chi^2 = \sum_i \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2, \qquad x_i \sim \mathrm{N}(\mu_i, \sigma_i)$$

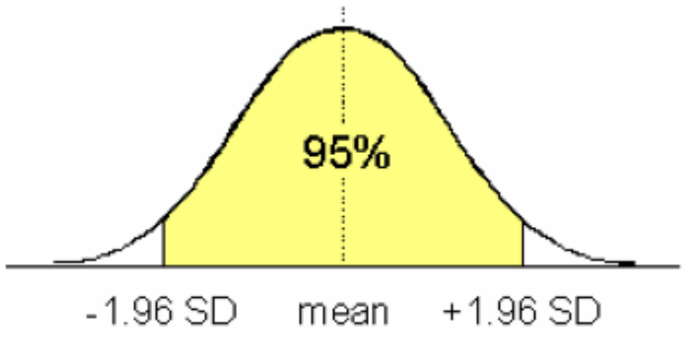
$\mathrm{Chisquare}(\nu)$ is a distribution (special case of Gamma), defined as

$$\chi^2 \sim \mathrm{Chisquare}(\nu), \qquad \nu > 0$$

$$p(\chi^2) d\chi^2 = \frac{1}{2^{\frac{1}{2}\nu} \Gamma(\frac{1}{2}\nu)} (\chi^2)^{\frac{1}{2}\nu - 1} \exp\left(-\tfrac{1}{2}\chi^2\right) d\chi^2, \qquad \chi^2 > 0$$

confidence intervals

The variances of *one parameter* at a time imply confidence intervals as for an ordinary 1-dimensional normal distribution:

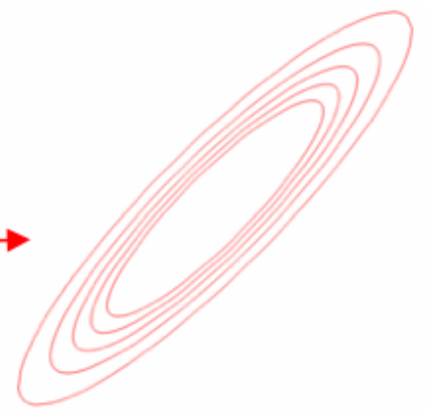


(Remember to take the square root of the variances to get the standard deviations!)

If you want to give confidence regions for *more than one parameter* at a time, you have to decide on a shape, since any shape containing 95% (or whatever) of the probability is a 95% confidence region!

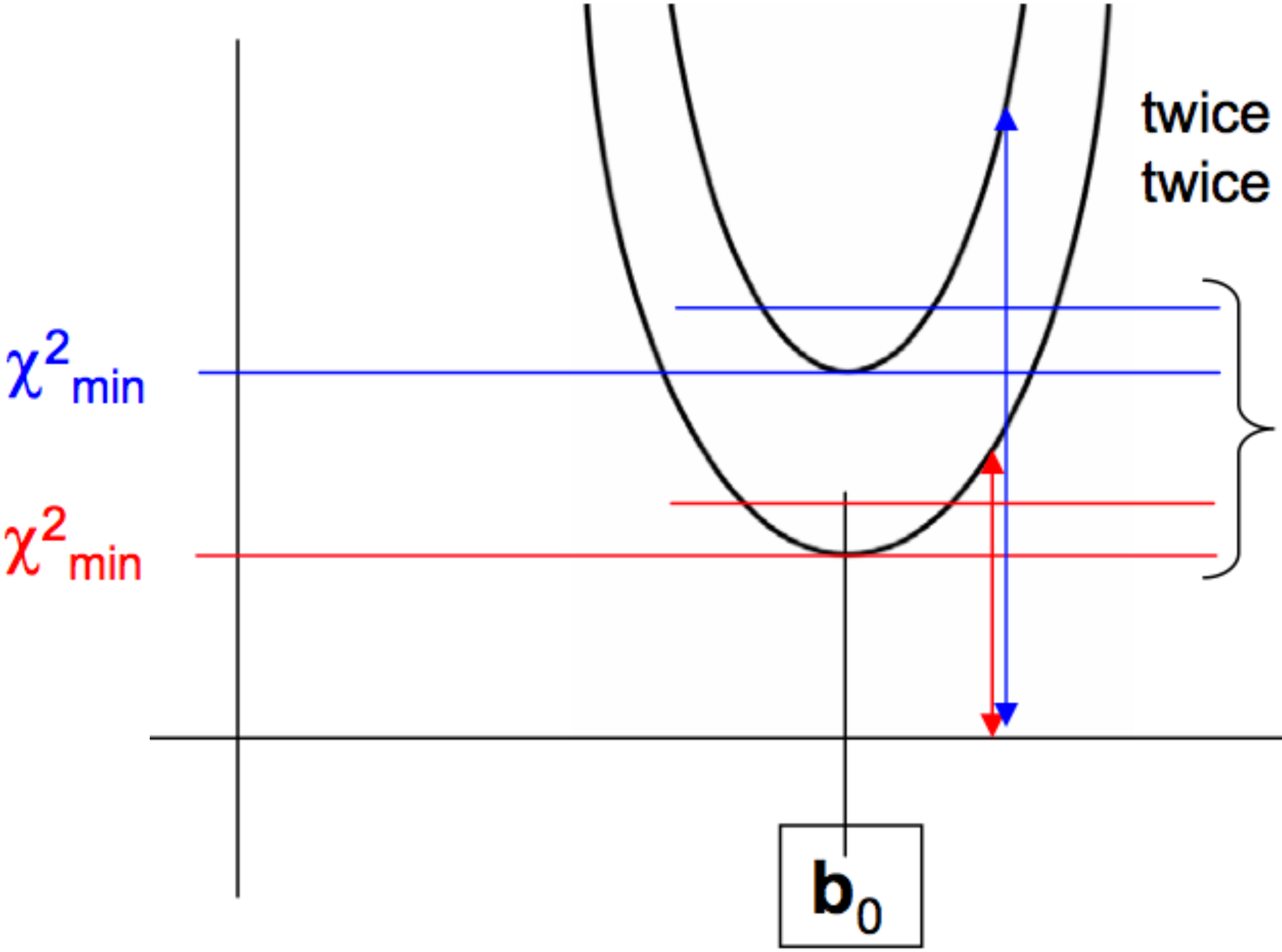
It is *conventional* to use contours of probability density as the shapes (= contours of $\Delta\chi^2$) since these are maximally compact.

But which $\Delta\chi^2$ contour contains 95% of the probability? →



χ^2 distribution

Measurement precision improves with the amount of data *N* as *N*$^{-1/2}$



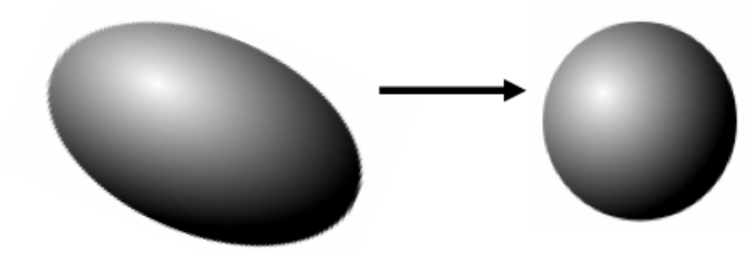
twice the data implies about twice the χ^2 at any **b**

so fixed $\Delta\chi^2$ implies √2 better precision

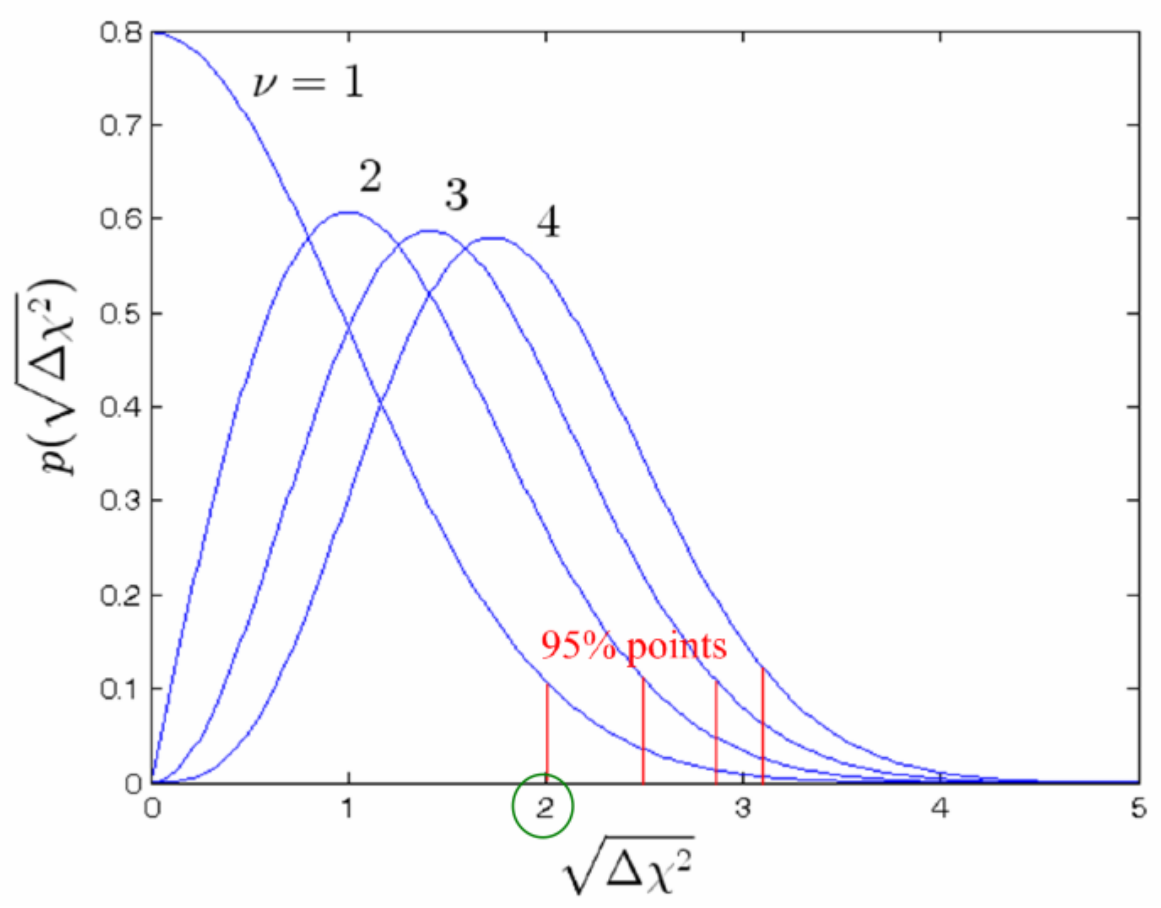
confidence intervals

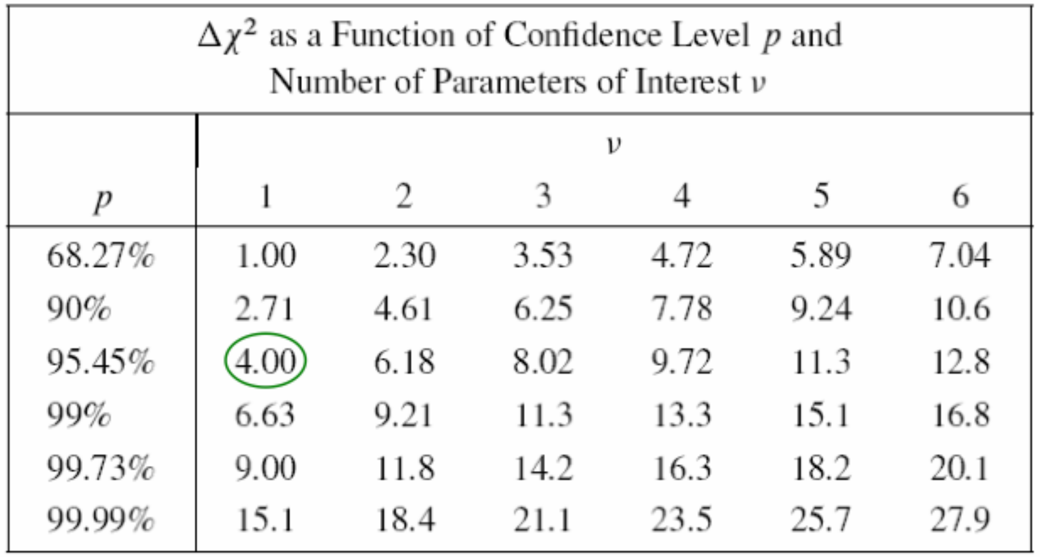
What $\Delta\chi^2$ contour in ν dimensions contains some percentile probability?

Rotate and scale the covariance to make it spherical.
(Linear, so contours still contain same probability.)



Now, each dimension is an independent Normal, and contours are labeled by radius squared (sum of ν individual t^2 values), so $\Delta\chi^2 \sim$ Chisquare(ν)





$\Delta\chi^2$ as a Function of Confidence Level p and Number of Parameters of Interest ν

	ν					
p	1	2	3	4	5	6
68.27%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.45%	4.00	6.18	8.02	9.72	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.9

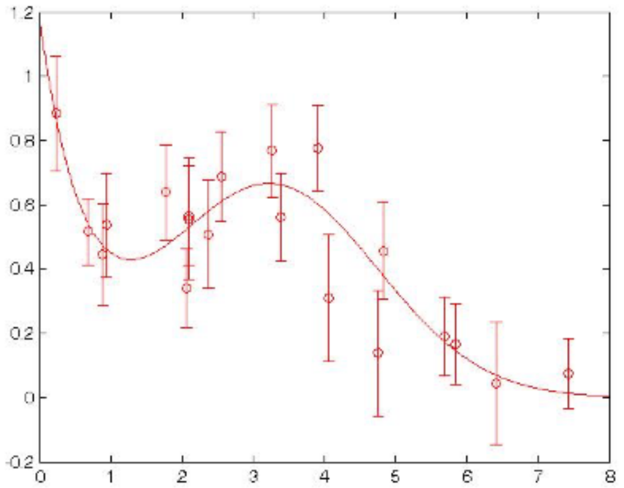
You sometimes learn "facts" like: "delta chi-square of 1 is the 68% confidence level". We now see that this is true only for one parameter at a time.

what is the Degree of Freedom?

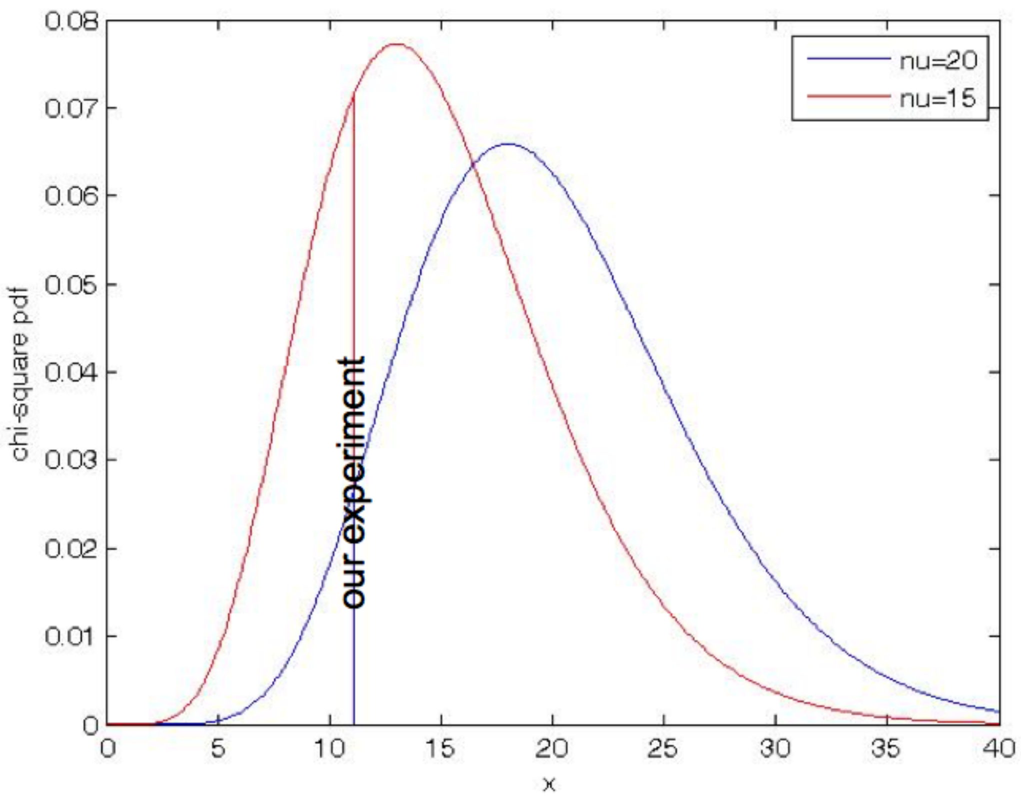
How is our fit by this test?

In our example, $\chi^2(\mathbf{b}_0) = 11.13$

This is a bit unlikely in Chisquare(20), with (left tail) p=0.0569.



In fact, if you had many repetitions of the experiment, you would find that their χ^2 is not distributed as Chisquare(20), but rather as Chisquare(15)! Why?



the magic word is:
“degrees of freedom” or DOF

what is the Degree of Freedom?

Degrees of Freedom: Why is χ^2 with N data points "not quite" the sum of N t^2-values? Because DOFs are reduced by constraints.

First consider a hypothetical situation where the data has linear constraints:

$$t_i = \frac{y_i - \mu_i}{\sigma_i} \sim \mathrm{N}(0,1)$$

joint distribution on all the t's, if they are independent

$$p(\mathbf{t}) = \prod_i p(t_i) \propto \exp\left(-\tfrac{1}{2}\sum_i t_i^2\right)$$

$$\chi^2 \text{ is squared distance from origin } \sum t_i^2$$

Linear constraint:

$$\sum_i \alpha_i y_i = C = \langle C \rangle = \sum_i \alpha_i \mu_i$$

$$C = \sum_i \alpha_i(\sigma_i t_i + \mu_i)$$

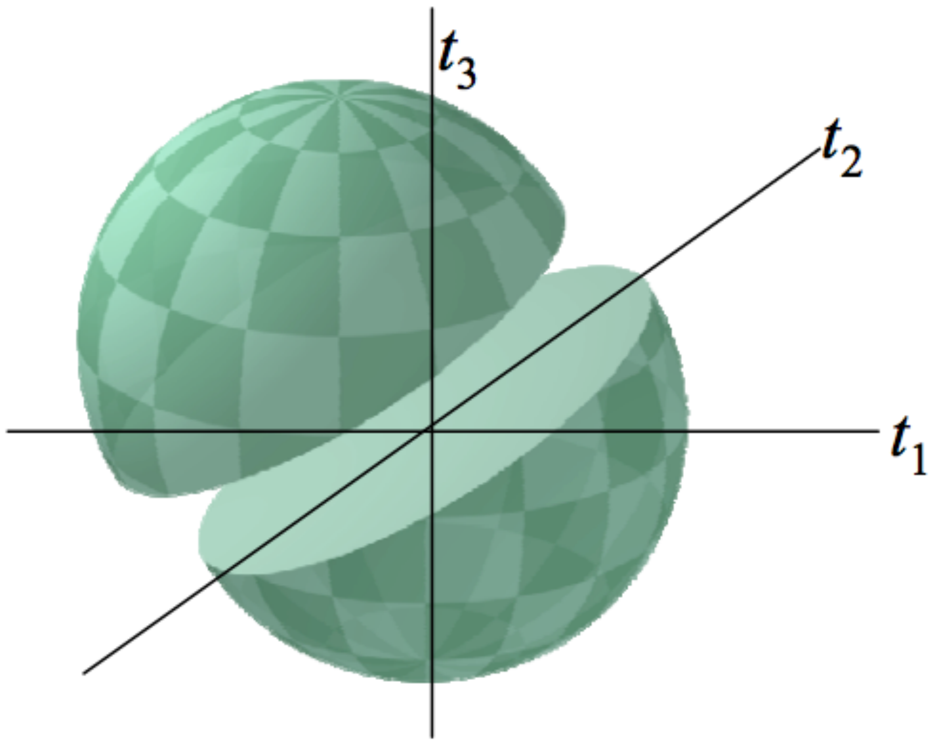
$$= \sum_i \alpha_i \sigma_i t_i + C$$

$$\text{So, } \sum_i \alpha_i \sigma_i t_i = 0$$

a hyper plane through the origin in t space!

what is the Degree of Freedom?

Constraint is a plane cut through the origin. Any cut through the origin of a sphere is a circle.



So the distribution of distance from origin is the same as a multivariate normal "ball" in the lower number of dimensions. Thus, each linear constraint reduces ν by exactly 1.

We don't have explicit constraints on the y_i's. But as the y_i's wiggle around (within their errors) we do have the constraint that we want to keep the MLE estimate $\mathbf{b_0}$ fixed. (E.g., we have 20 wiggling y_i's and only 5 b_i's to keep fixed.)

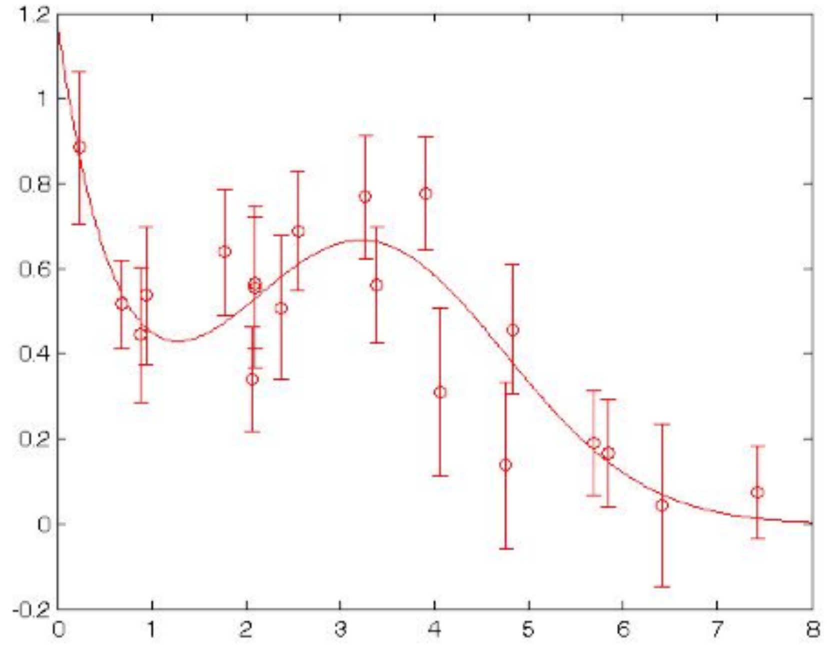
So by the implicit function theorem, there are M (number of parameters) approximately linear constraints on the y_i 's. So $\nu = N - M$, the so-called number of degrees of freedom (d.o.f.).

what is the Degree of Freedom?

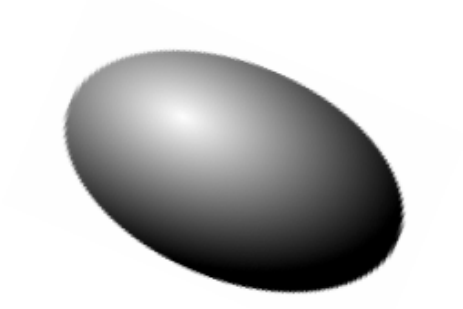
Review:

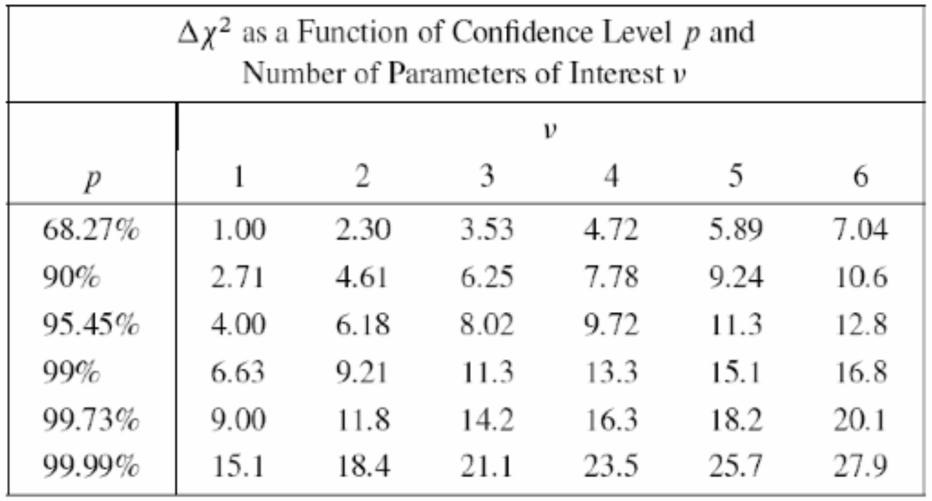
1. Fit for parameters by minimizing

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i}\right)^2$$



2. (Co)variances of parameters, or confidence regions, by the change in χ^2 (i.e., $\Delta\chi^2$) from its minimum value χ^2_{min}.

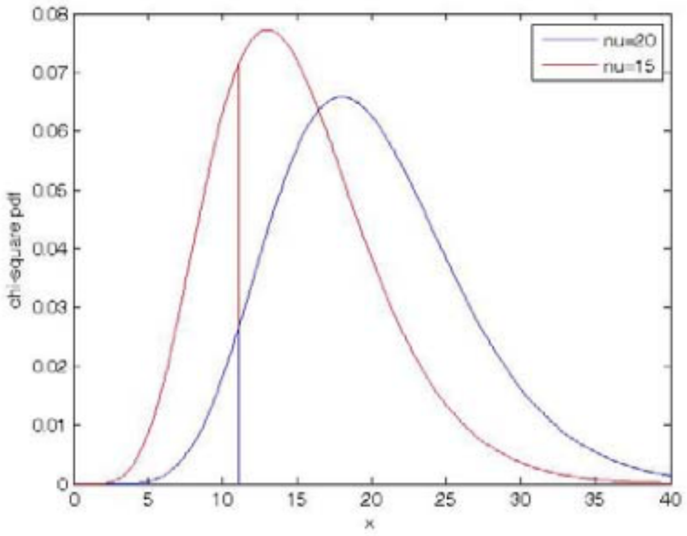




$\Delta\chi^2$ as a Function of Confidence Level p and Number of Parameters of Interest ν

	ν					
p	1	2	3	4	5	6
68.27%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.45%	4.00	6.18	8.02	9.72	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.9

3. Goodness-of-fit (accept or reject model) by the p-value of χ^2_{min} using the correct number of DOF.



Goodness-of-fit

Goodness-of-fit with $\nu = N - M$ degrees of freedom:

we expect $\chi^2_{\min} \approx \nu \pm \sqrt{2\nu}$

this is an RV over the population of different data sets (a frequentist concept allowing a p-value)

Confidence intervals for parameters **b**:

we expect $\chi^2 \approx \chi^2_{\min} \pm O(1)$

this is an RV over the population of possible model parameters for a single data set, a concept shared by Bayesians and frequentists

How can $\pm O(1)$ be significant when the uncertainty is $\pm\sqrt{2\nu}$?

Answer: Once you have a <u>particular</u> data set, there is <u>no</u> uncertainty about what its $\chi^2_{\min}$ is. Let's see how this works out in scaling with N:

χ^2 increases linearly with $\nu = N - M$

$\Delta\chi^2$ increases as N (number of terms in sum), but also decreases as $(N^{-1/2})^2$, since **b** becomes more accurate with increasing N :

$$\Delta\chi^2 \propto N(\delta b)^2, \quad \delta b \propto N^{-1/2} \quad \Rightarrow \quad \Delta\chi^2 \propto \text{const}$$

quadratic, because at minimum

universal rule of thumb