Lectures 11: Maximum likelihood IV.
(nonlinear least square fits)

χ^2 fitting procedure!

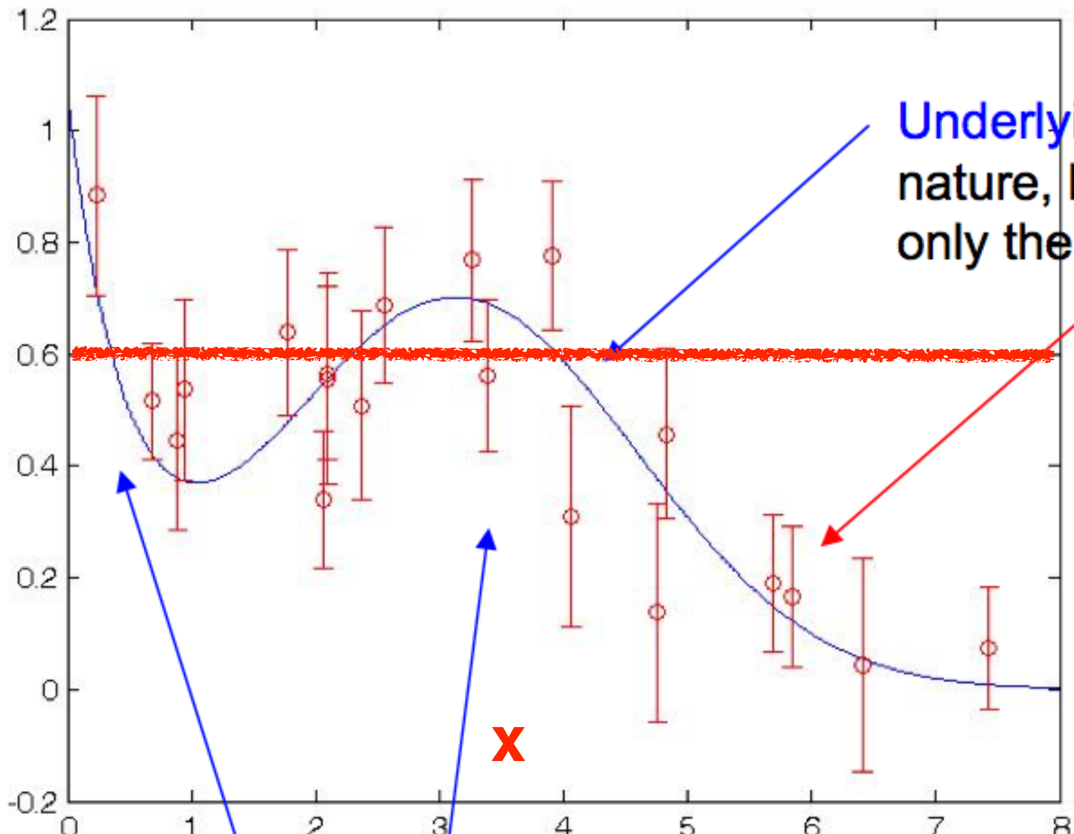
from Lecture 9:

An example might be something like fitting a known functional form to data

$$f(x) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2}\frac{(x-b_4)^2}{b_5^2}\right)$$

$= 2 \cdot p(x) - 0.4$
$= y(x|b)$

measured value of 2p-0.4 as a function of x



Underlying curve is known to nature, but not to us! We see only the red data points.

Fit 5 parameters from 20 irregularly space points, with normal errors of known standard deviations.

Can we do it? How well?

increasing temperature x in some arbitrary units

for example, this rise might be an instrumental or noise effect, while this bump might be what you are really interested in

from Lecture 9: Maximum Likelihood discussion

Fitting is usually presented in frequentist, MLE language.
But one can equally well think of it as Bayesian:

$$
\begin{aligned}
P(\mathbf{b}|\{y_i\}) &\propto P(\{y_i\}|\mathbf{b})P(\mathbf{b}) \\
&\propto \prod_i \exp\left[-\tfrac{1}{2}\left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i}\right)^2\right] P(\mathbf{b}) \\
&\propto \exp\left[-\tfrac{1}{2}\sum_i \left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i}\right)^2\right] P(\mathbf{b}) \\
&\propto \exp[-\tfrac{1}{2}\chi^2(\mathbf{b})]P(\mathbf{b})
\end{aligned}
$$

frequentist: $P(\mathbf{b}) \sim \delta(\mathbf{b}-\mathbf{b}_0)$ **$\mathbf{b}_0$?**

Bayesian: $P(\mathbf{b}) \sim$ const simplest, leads to same $\mathbf{b}_0$ determination

repeating the experiment with y_i and σ_i we also test f(x) as a hypothesis

Now the idea is: Find (somehow!) the parameter value $\mathbf{b}_0$ that minimizes χ^2 .

For linear models, you can solve linear “normal equations” or, better, use Singular Value Decomposition. See NR3 section 15.4

In the general nonlinear case, you have a general minimization problem, for which there are various algorithms, none perfect.

Those parameters are the MLE. (So it is Bayes with uniform prior.)

from Lecture 9: Maximum Likelihood discussion

Nonlinear fits are often easy in MATLAB (or other high-level languages) **if** you can make a reasonable starting guess for the parameters:

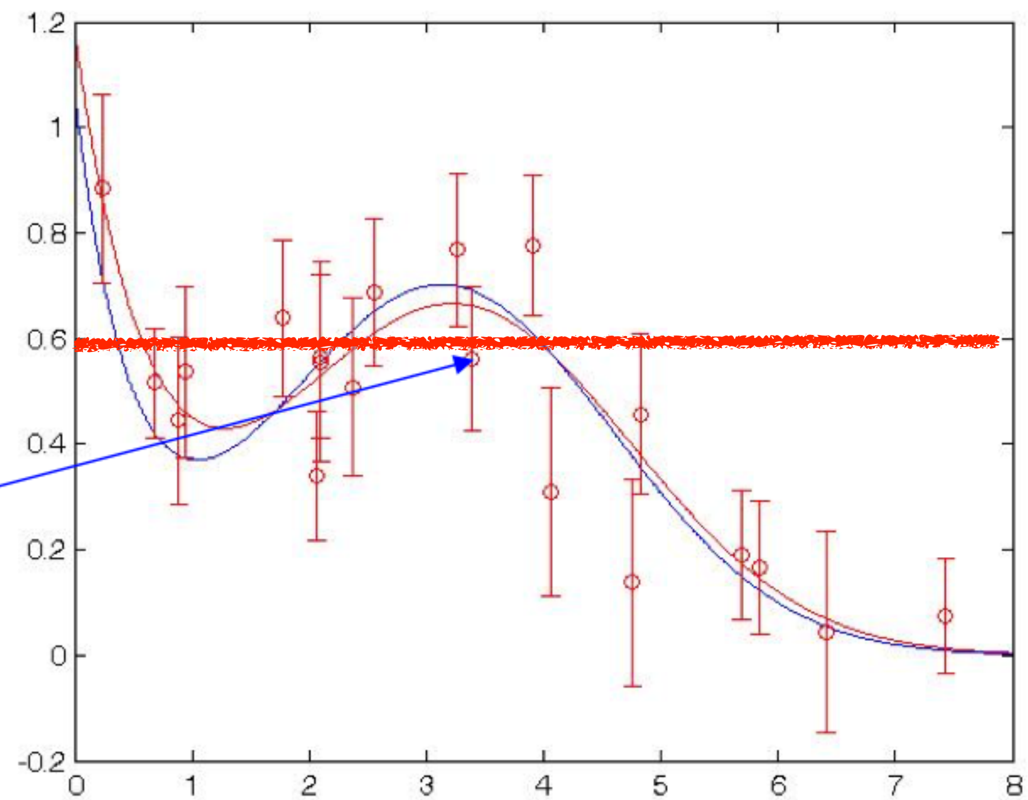
$$y(x|\mathbf{b}) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2}\frac{(x-b_4)^2}{b_5^2}\right)$$

$$\chi^2 = \sum_i \left(\frac{y_i - y(x_i|\mathbf{b})}{\sigma_i}\right)^2$$

```
ymodel = @(x,b) b(1)*exp(-b(2)*x)+b(3)*exp(-(1/2)*((x-b(4))/b(5)).^2)
chisqfun = @(b) sum(((ymodel(x,b)-y)./sig).^2)
```

```
bguess = [1 2 .5 3 1.5]
bfit = fminsearch(chisqfun,bguess)
xfit = (0:0.01:8);
yfit = ymodel(xfit,bfit);
```

```
bfit = 1.1235    1.5210    0.6582
       3.2654    1.4832
```



Suppose that what we really care about is the area of the bump, and that the other parameters are "nuisance parameters".

increasing temperature x in some arbitrary units

Maximum Likelihood parameter errors?

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

Taylor series:

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

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

Then

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

with

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

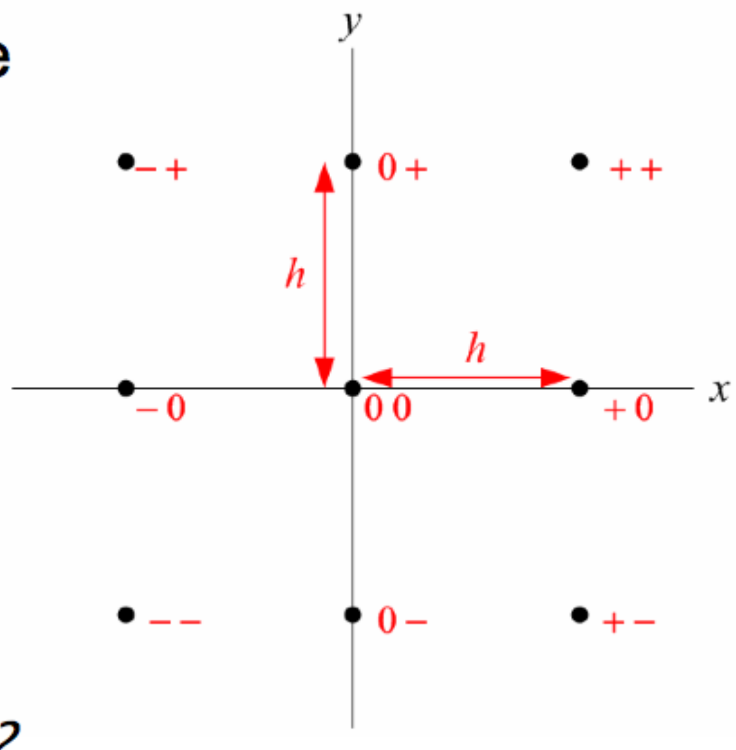
covariance (or "standard error") matrix of the fitted parameters

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!

Maximum Likelihood parameter errors?

Numerical calculation of the Hessian by finite difference

$$\frac{\partial^2 f}{\partial x \partial y} \approx \frac{1}{2h}\left(\frac{f_{++} - f_{-+}}{2h} - \frac{f_{+-} - f_{--}}{2h}\right)$$
$$= \frac{1}{4h^2}\left(f_{++} + f_{--} - f_{+-} - f_{-+}\right)$$



bfit = 1.1235 1.5210 0.6582 3.2654 1.4832

```
chisqfun = @(b) sum(((ymodel(x,b)-y)./sig).^2)
h = 0.1;
unit = @(i) (1:5) == i;
hess = zeros(5,5);
for i=1:5, for j=1:5,
        bpp = bfit + h*(unit(i)+unit(j));
        bmm = bfit + h*(-unit(i)-unit(j));
        bpm = bfit + h*(unit(i)-unit(j));
        bmp = bfit + h*(-unit(i)+unit(j));
        hess(i,j) = (chisqfun(bpp)+chisqfun(bmm)...
          -chisqfun(bpm)-chisqfun(bmp))./(2*h)^2;
    end
end
covar = inv(0.5*hess)
```

This also works for the diagonal components. Can you see how?

Maximum Likelihood parameter errors?

For our example, $y(x|\mathbf{b}) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2}\frac{(x-b_4)^2}{b_5^2}\right)$

```
bfit =
    1.1235    1.5210    0.6582    3.2654    1.4832
hess =
   64.3290  -38.3070   47.9973  -29.0683   46.0495
  -38.3070   31.8759  -67.3453   29.7140  -40.5978
   47.9973  -67.3453  723.8271  -47.5666  154.9772
  -29.0683   29.7140  -47.5666   68.6956  -18.0945
   46.0495  -40.5978  154.9772  -18.0945   89.2739
covar =
    0.1349    0.2224    0.0068   -0.0309    0.0135
    0.2224    0.6918    0.0052   -0.1598    0.1585
    0.0068    0.0052    0.0049    0.0016   -0.0094
   -0.0309   -0.1598    0.0016    0.0746   -0.0444
    0.0135    0.1585   -0.0094   -0.0444    0.0948
```

This is the covariance structure of all the parameters, and indeed (at least in CLT normal approximation) gives their entire joint distribution!

The standard errors on each parameter separately are $\sigma_i = \sqrt{C_{ii}}$

```
sigs =
    0.3672    0.8317    0.0700    0.2731    0.3079
```

But why is this, and what about two or more parameters at a time (e.g. b_3 and b_5)?

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.

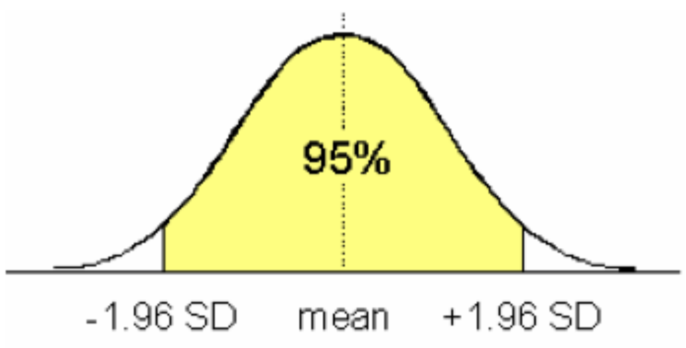
(not quite)

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.

confidence intervals

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

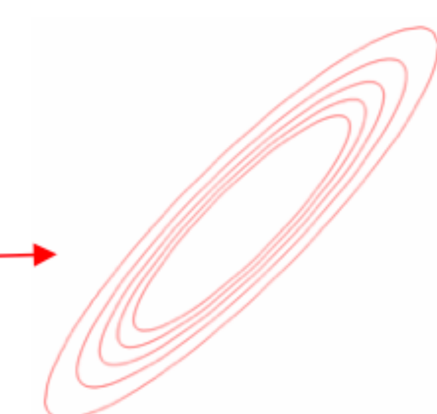


(Remember to take the square root of the variances to get the standard deviations!)

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 (from Lecture 10)

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

$$y = x^2$$

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

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

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

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

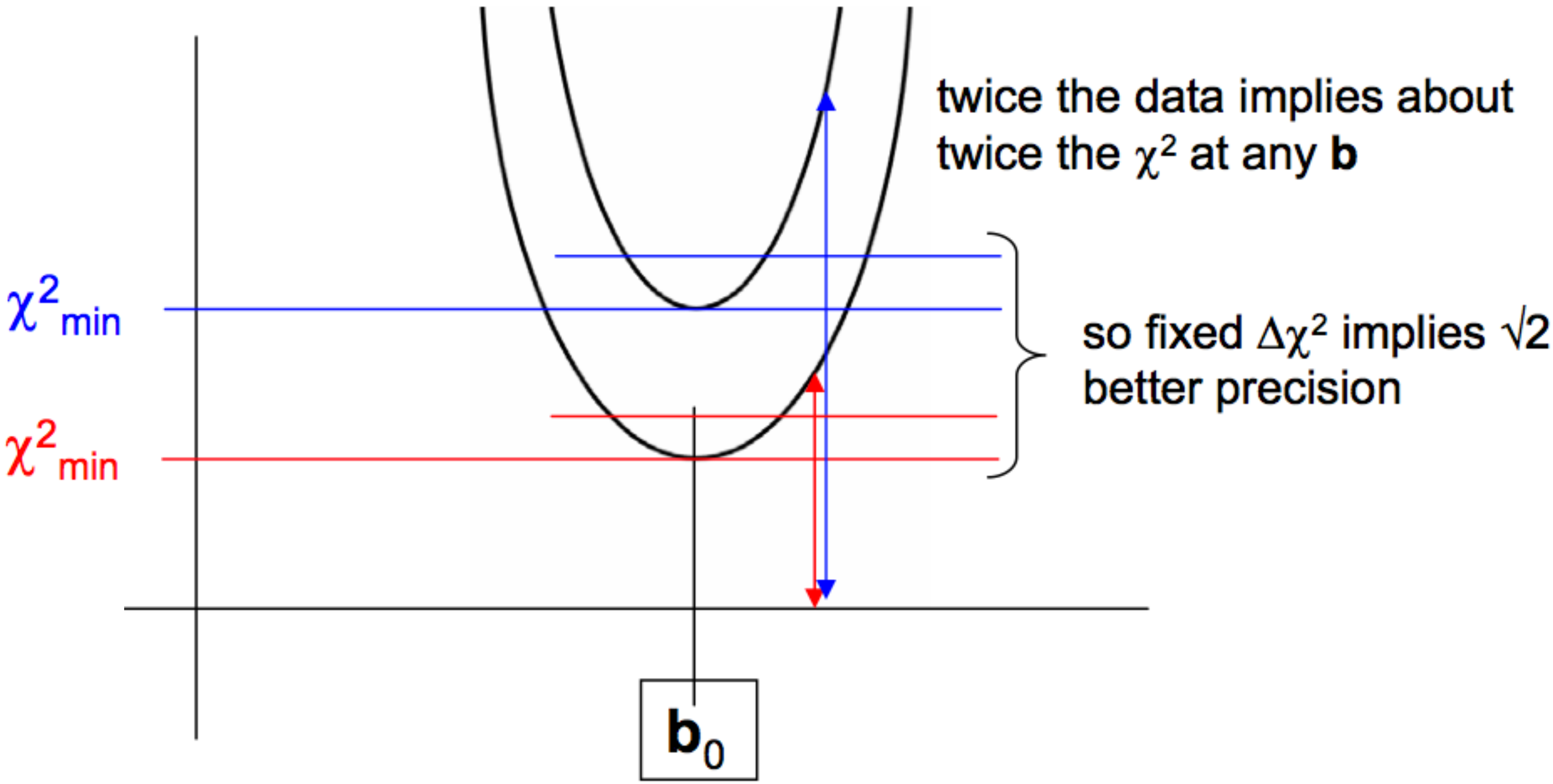
$\mathrm{Chisquare}(\nu)$ is a distribution (special case of Gamma), defined as

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

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

χ^2 distribution

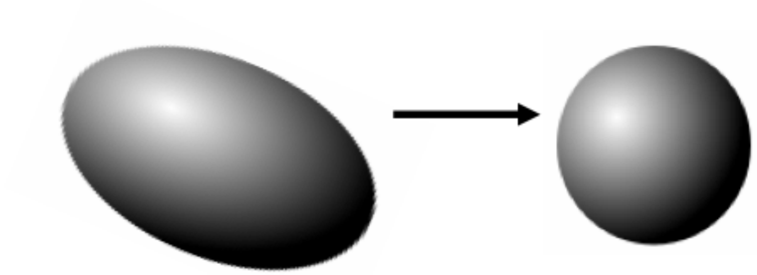
Measurement precision improves with the amount of data *N* as *N*$^{-1/2}$



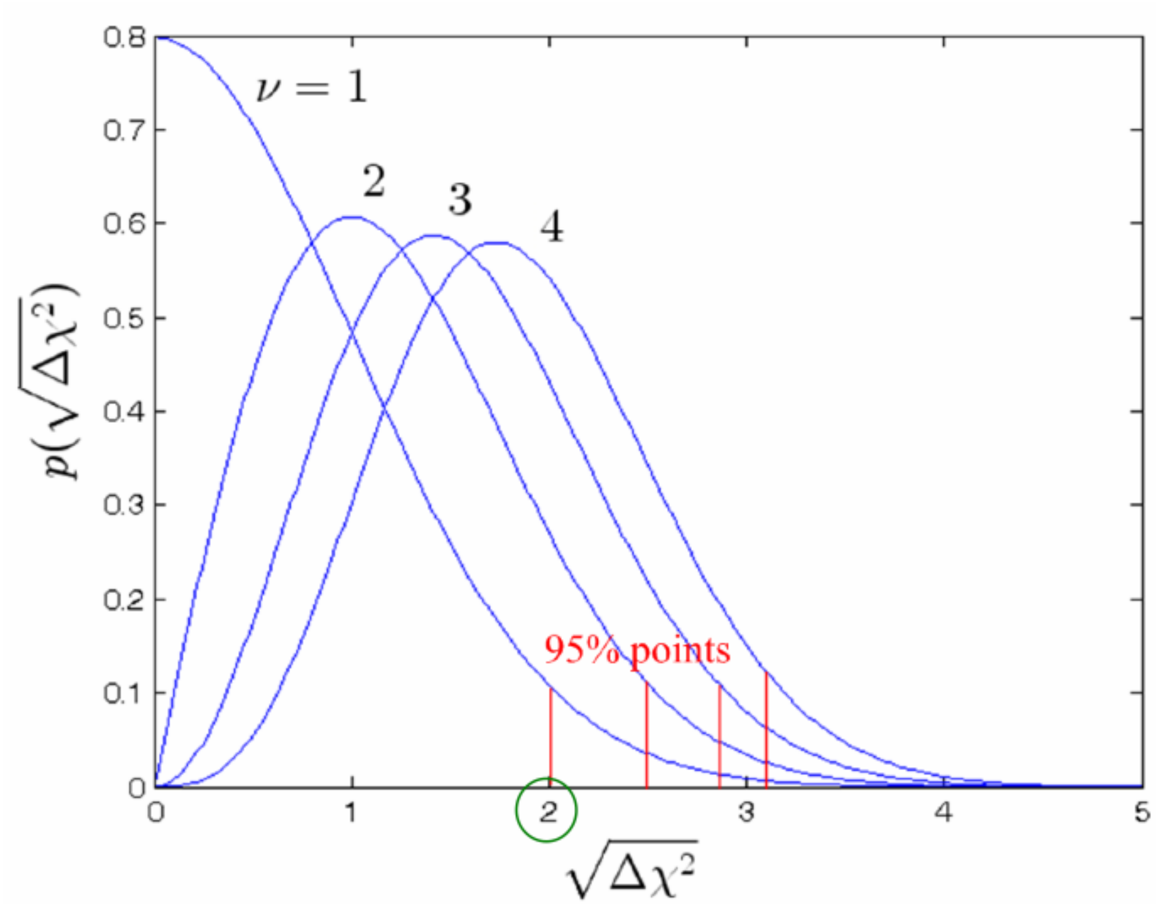
confidence intervals

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

p	ν = 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

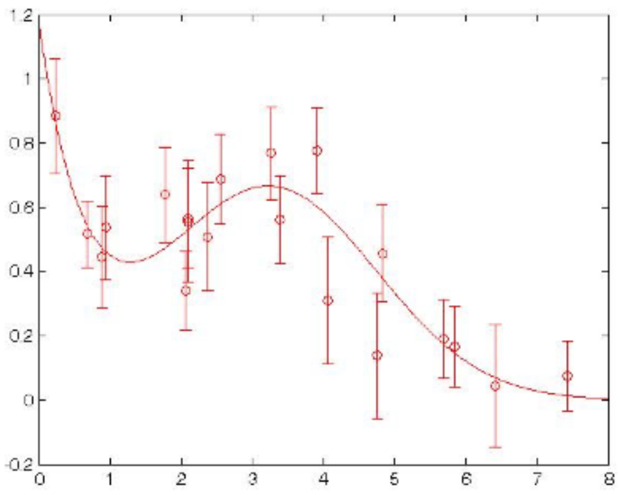
You sometimes learn "facts" like: "delta chi-square of 1 is the 68% confidence level". We now see that this is true only for one parameter at a time.

what is the Degree of Freedom?

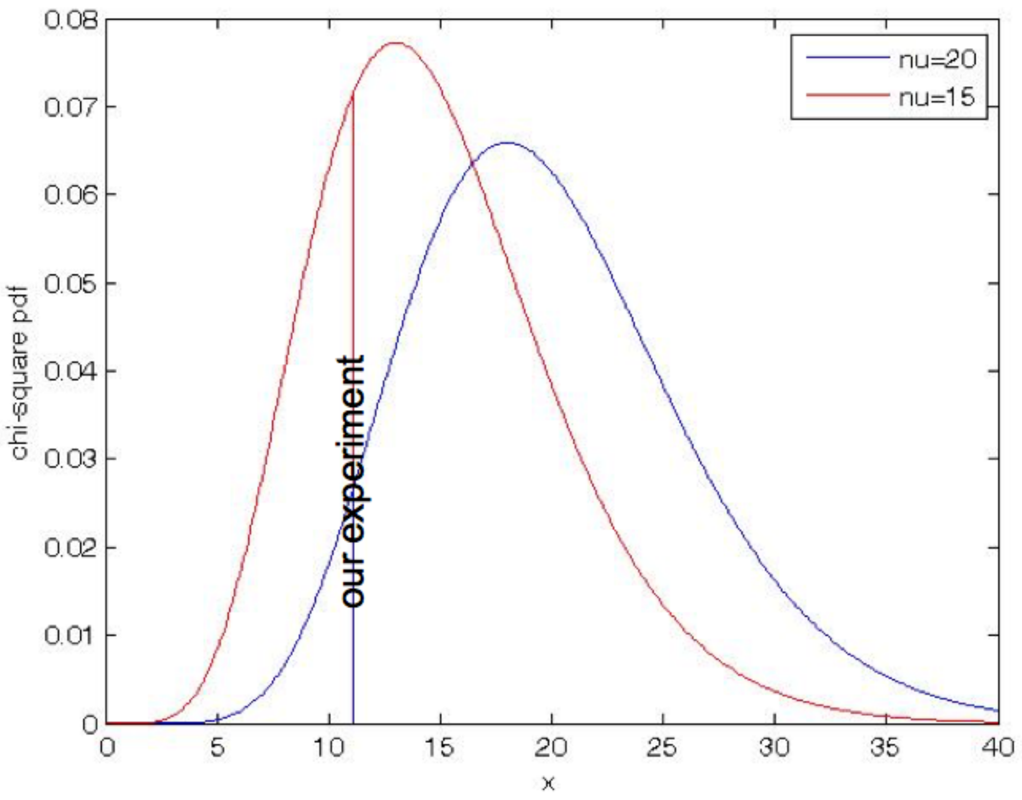
How is our fit by this test?

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 <u>linear constraints</u>:

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

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

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

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

Linear constraint:

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

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

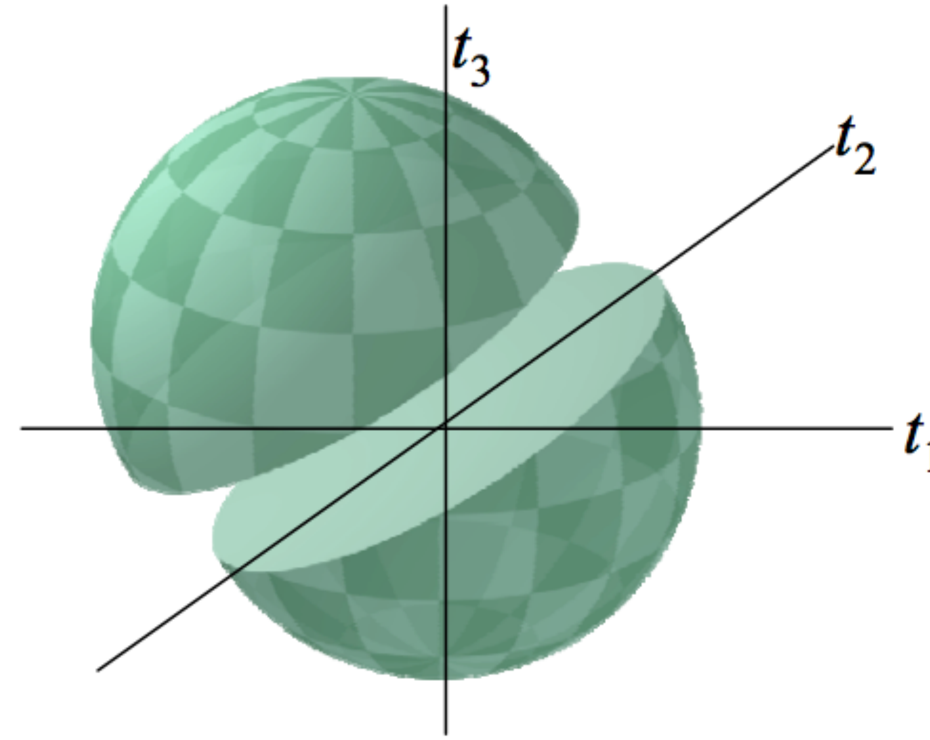
$$= \sum_i \alpha_i \sigma_i t_i + C$$

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

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.



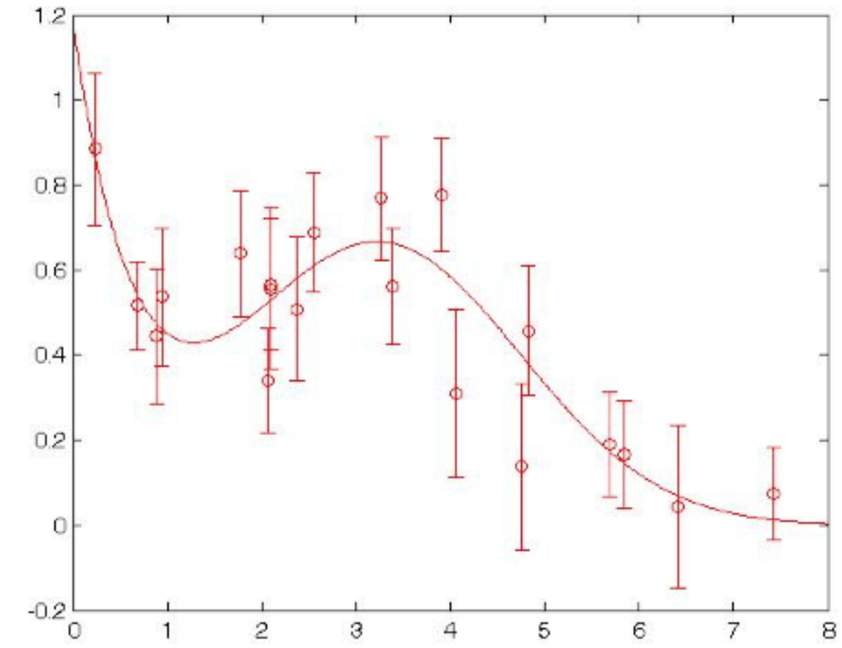
So the distribution of distance from origin is the same as a multivariate normal "ball" in the lower number of dimensions. Thus, each linear constraint reduces ν by exactly 1.

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?

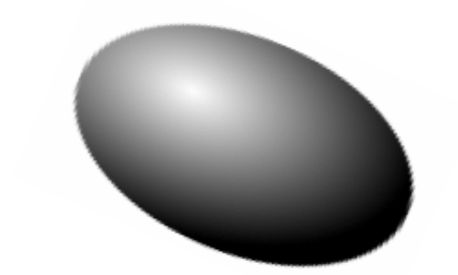
Review:



1. Fit for parameters by minimizing

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

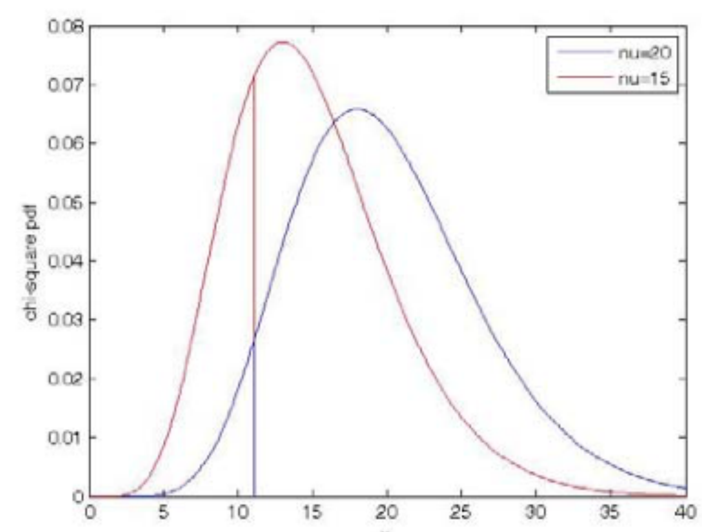
2. (Co)variances of parameters, or confidence regions, by the change in χ^2 (i.e., $\Delta\chi^2$) from its minimum value χ^2_{min}.



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

p	ν = 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 particular data set, there is no 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