Opinionated Lessons in Statistics

by Bill Press

#24 Goodness of Fit

Good time now to review the universal rule-of-thumb (meta-theorem):

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

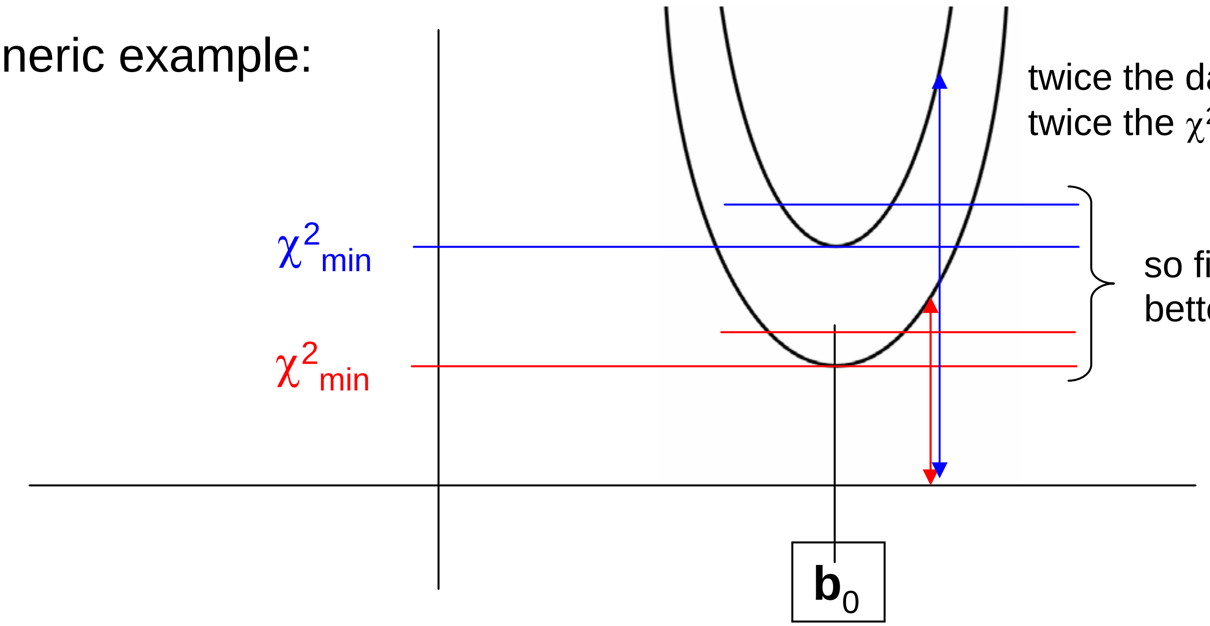
Simple example:

"measurement precision" = "accuracy of a fitted parameter"

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\mathrm{Var}(\mu) = \frac{1}{N^2}\mathrm{Var}\left(\sum_{i=1}^{N} x_i\right) = \frac{1}{N^2}\left[N\,\mathrm{Var}(x)\right] = \frac{1}{N}\mathrm{Var}(x)$$

Generic example:



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

so fixed $\Delta\chi^2$ implies $\sqrt{2}$ better precision

$\Delta\chi^2$ 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

Let's discuss Goodness of Fit (at last!)

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

That is a very Bayesian thing to do, since Bayesians start with an EME set of hypotheses. It also makes it difficult for Bayesians to deal with the notion of a model's goodness of fit.

So we must now again become frequentists for a while!

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

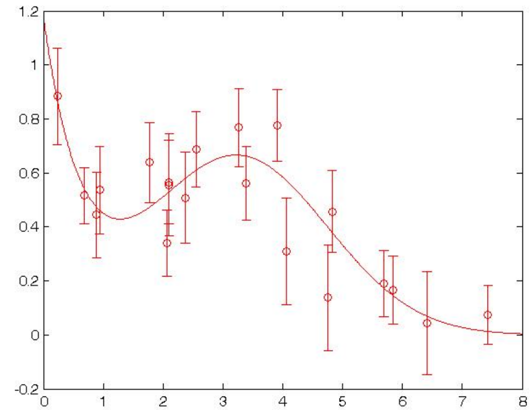
Then the "statistic" $\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i} \right)^2$ is the sum of N t^2-values.

(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 <u>very unlikely</u> to be from this distribution, we consider the model to be disproved. In other words, <u>it is a p-value test</u>.

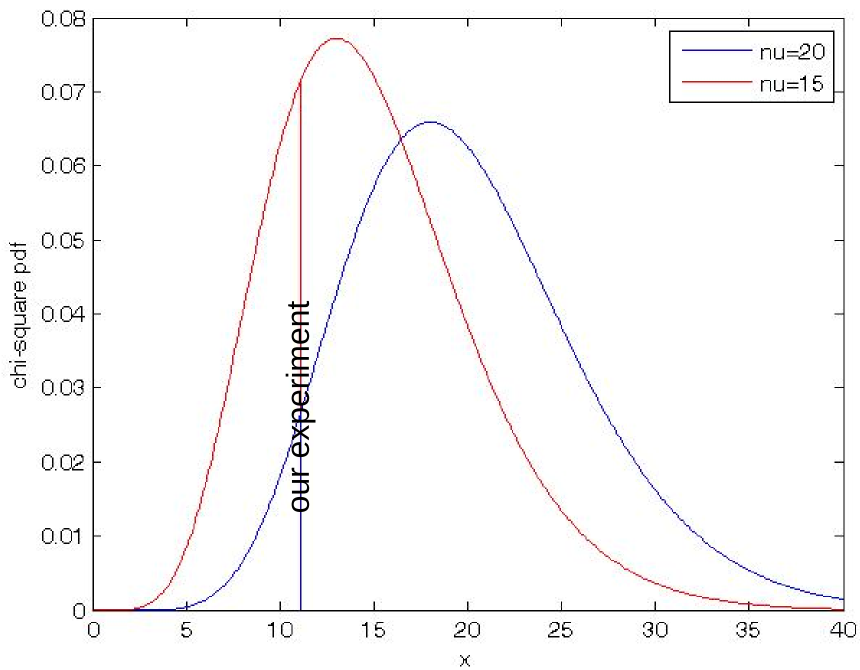
How is our fit by this test?



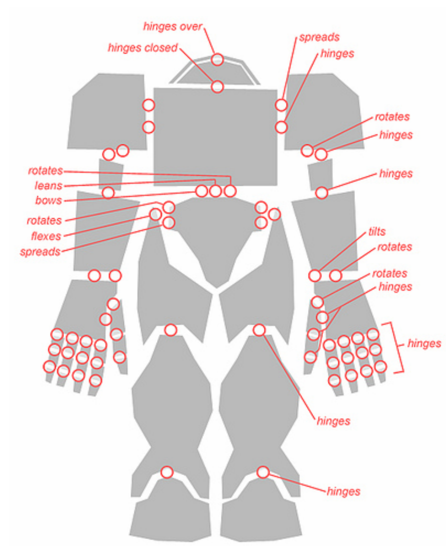
In our example, $\chi^2(\mathbf{b}_0) = 11.13$

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 <u>not</u> distributed as Chisquare(20), but rather as Chisquare(15)! Why?



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



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:

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

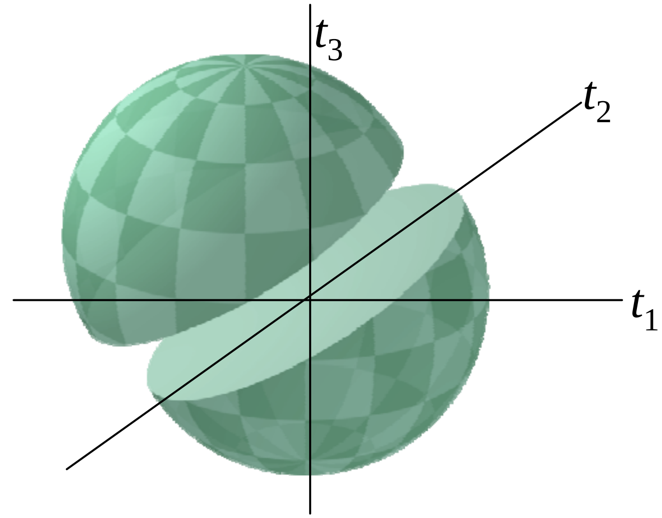
$$C = \sum_i \alpha_i(\sigma_i t_i + \mu_i)$$

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

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

a hyper plane through the origin in t space!

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

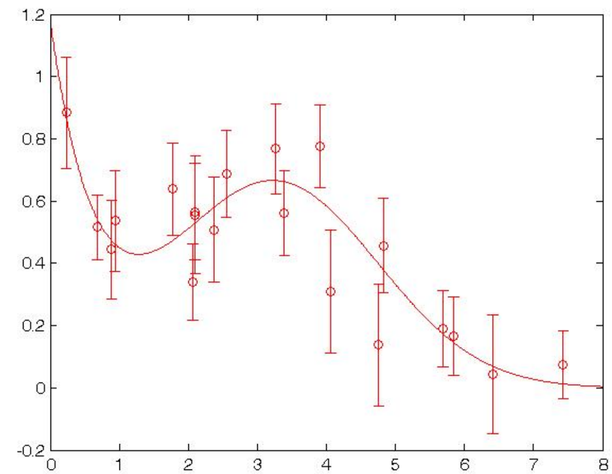


So the distribution of distance from origin is the same as a multivariate normal "ball" in the lower number of dimensions. Thus, each linear constraint reduces ν by exactly 1.

We don't have explicit constraints on the y_i's. But as we let the y_i's wiggle around (within the distribution of each) we want to keep the MLE estimate $\mathbf{b}_0$ (the parameters) fixed so as to see how χ^2 is distributed for this MLE – not for all possible **b**'s. (20 wiggling y_i's, 5 b_i's kept 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.).

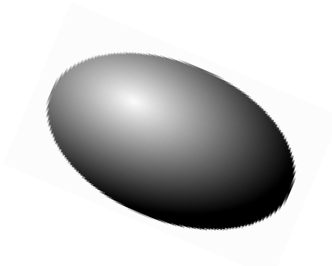
Review:



1. Fit for parameters by minimizing

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

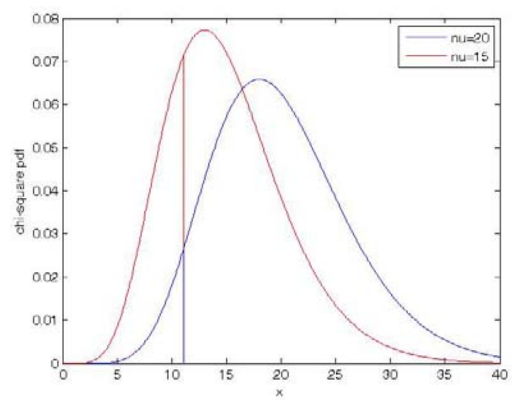
2. (Co)variances of parameters, or confidence regions, by the change in χ^2 (i.e., $\Delta\chi^2$) from its minimum value χ^2_{min}.



$\Delta\chi^2$ 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.



Don't confuse typical values of χ^2 with typical values of $\Delta\chi^2$!

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