CS395T
Computational Statistics with Application to Bioinformatics

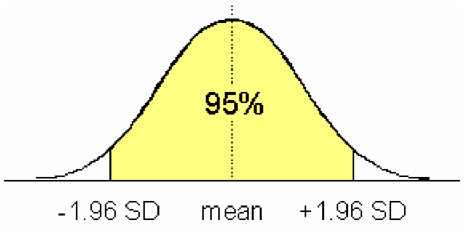
Prof. William H. Press
Spring Term, 2011
The University of Texas at Austin

Lecture 13

Small digression:

You can give confidence intervals or regions, instead of (co-)variances

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

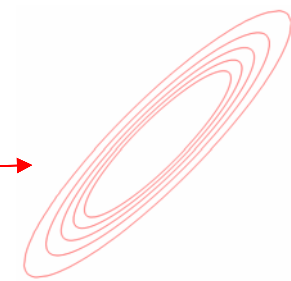


(Remember to take the square root of the variances to get the standard deviations!)

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!

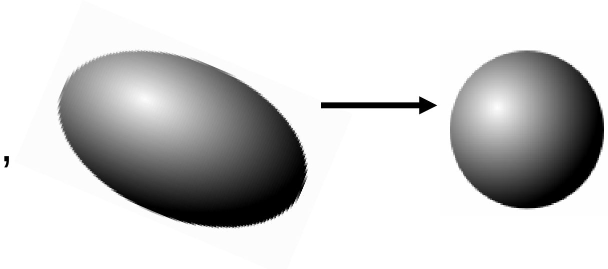
It is *conventional* to use contours of probability density as the shapes (= contours of $\Delta\chi^2$) since these are maximally compact.

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

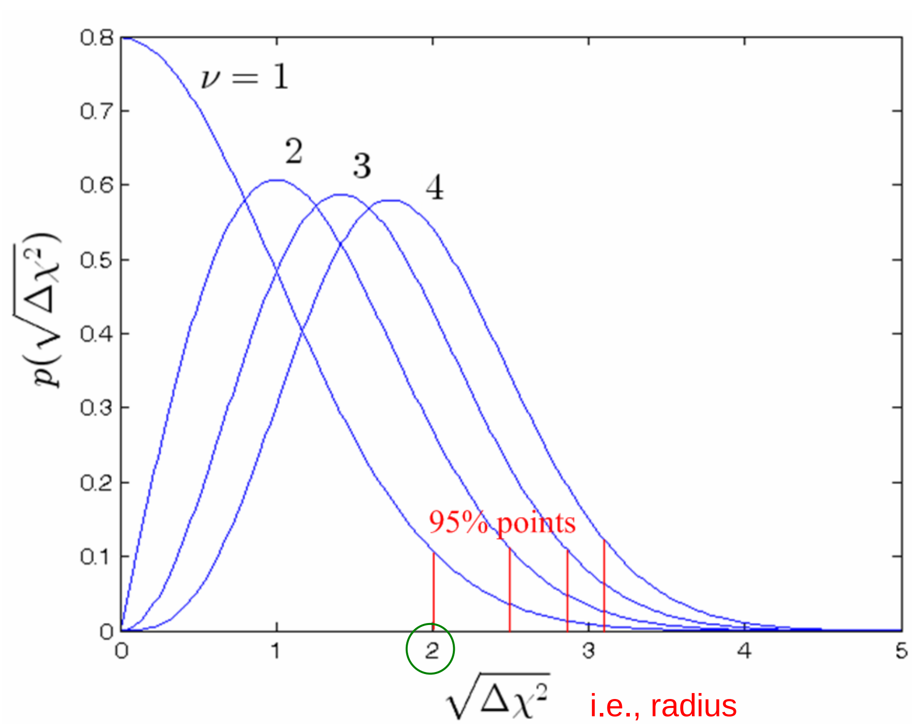


What $\Delta\chi^2$ contour in ν dimensions contains some percentile probability?

Rotate and scale the covariance to make it spherical. Contours still contain same probability. (In equations, this would be another "Cholesky thing".)



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.

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

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

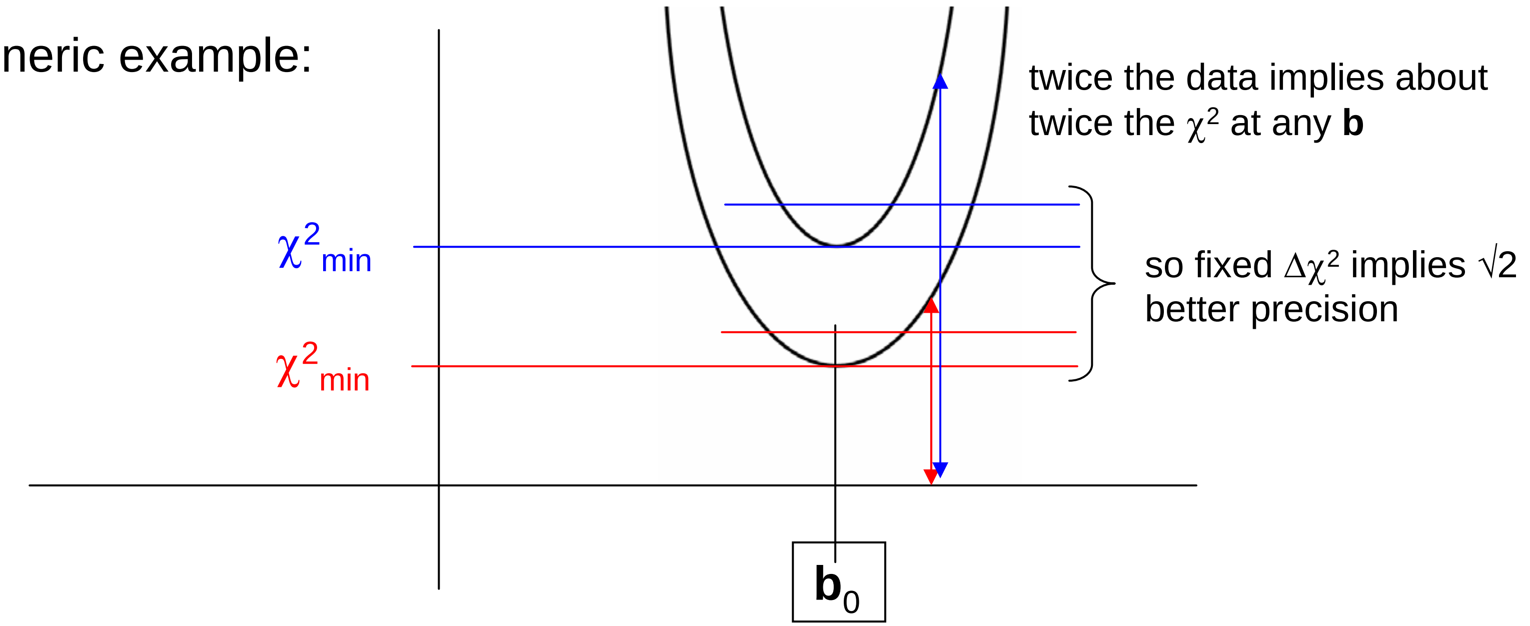
Simple example:

"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:



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 become frequentists for a bit!

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

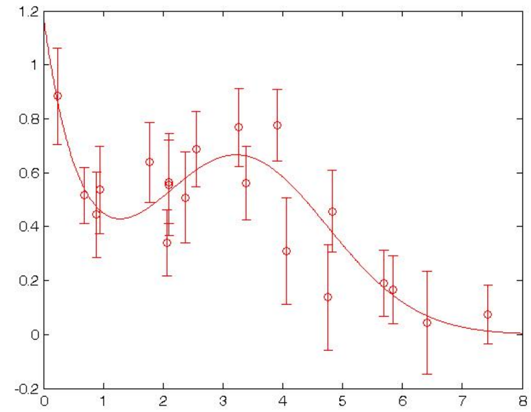
Then the "statistic" $\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i} \right)^2$ is the sum of N t^2-values.

(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.

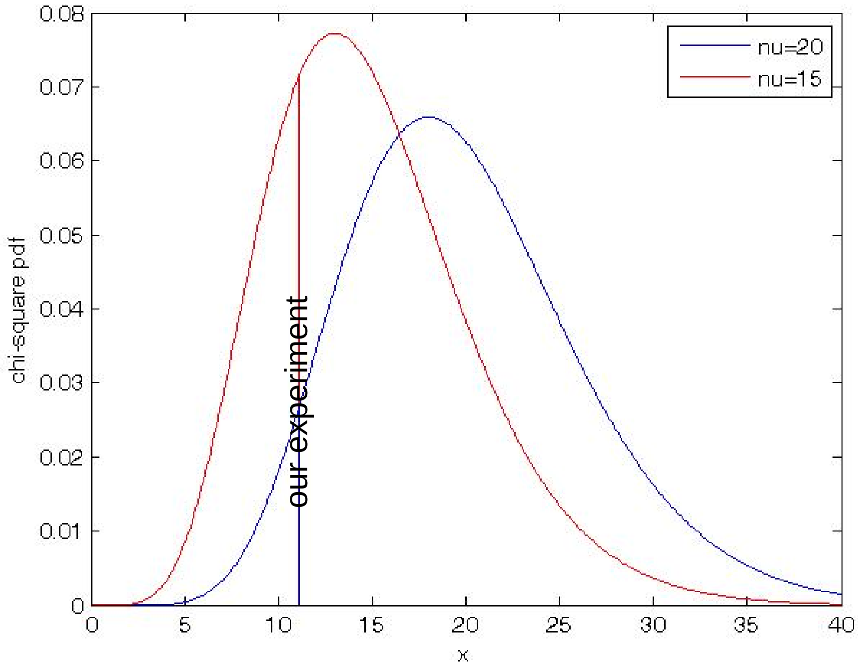
How is our fit by this test?



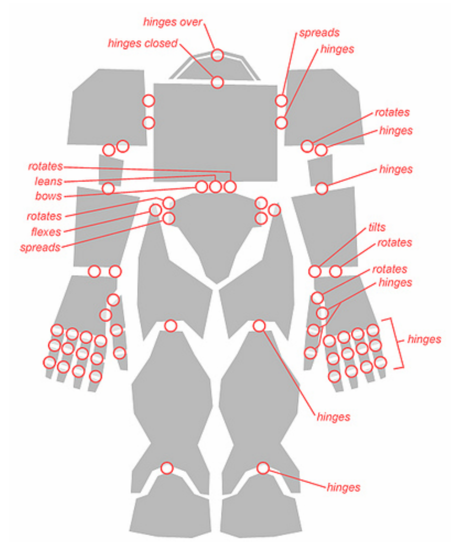
In our example, $\chi^2(\mathbf{b}_0) = 11.13$

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



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:

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

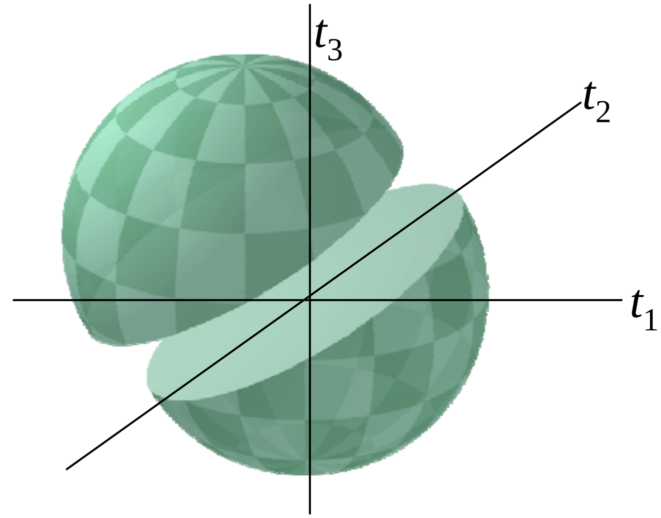
$$C = \sum_i \alpha_i(\sigma_i t_i + \mu_i)$$

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

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

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.

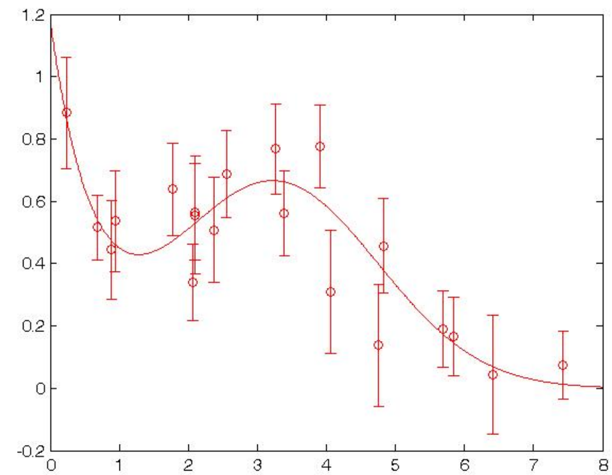
We don't have explicit constraints on the y_i's. But if we wiggle the y_i's around (within the distribution of each) we want to keep the MLE estimate $\mathbf{b_0}$ (i.e., the curve) fixed so as to see how χ^2 is distributed for this MLE – not for all possible **b**'s. (20 wiggling y_i's, 5 b_i's kept 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.).

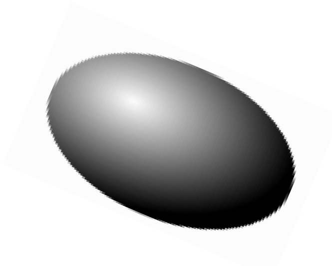
Review:

1. Fit for parameters by minimizing

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



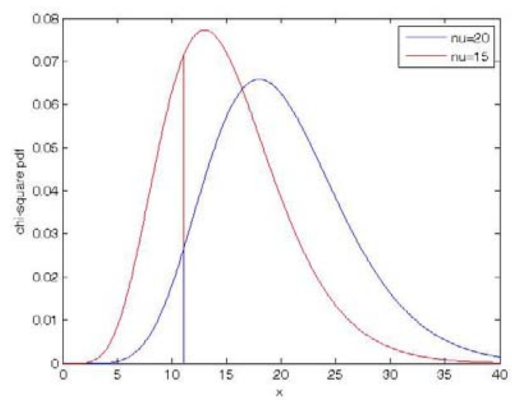
2. (Co)variances of parameters, or confidence regions, by the change in χ^2 (i.e., $\Delta\chi^2$) from its minimum value χ^2_{min}.



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

	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.



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 <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 :

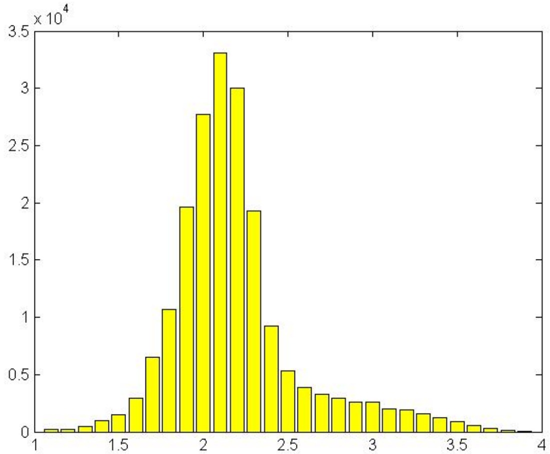
$$\Delta\chi^2 \propto N(\delta b)^2, \quad \delta b \propto N^{-1/2} \quad \Rightarrow \quad \Delta\chi^2 \propto \text{const}$$

quadratic, because at minimum

universal rule of thumb

Let's turn from (x,y,σ) data to data that comes as counts of things.

Two common examples are "binned values" (histograms) and contingency tables.



	C_0	C_1
f_0	n_{00}	n_{01}
f_1	n_{10}	n_{11}

Counts are distributed according to (in general, unknown) probabilities p_i or p_{ij} across the bins or table entries. The model (with parameters maybe) predicts the p's.

$n_i \sim \mathrm{Binomial}(N, p_i)$ or more precisely, $\{n_i\} \sim \mathrm{Multinomial}(N, \{p_i\})$

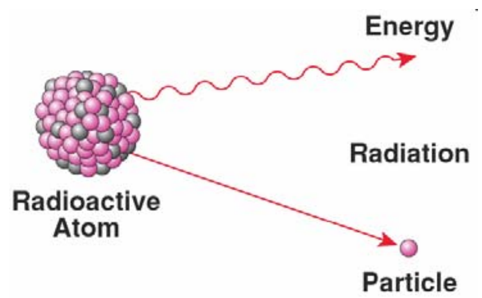
For histograms (but not necessarily contingency tables) one commonly has

$n_i \ll N \;\Rightarrow p_i \ll 1$ for all i

$n_i \ll N \;\Rightarrow p_i \ll 1$ for all i implies that counts are (close to) Poisson distributed

$$\text{Binomial}(n, N, p) \Rightarrow$$

$$\begin{aligned} P(n) &= \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \\ &= \frac{1}{n!} \frac{N!}{(N-n)!} p^n e^{(N-n)\ln(1-p)} \\ &\approx \frac{1}{n!} (Np)^n e^{-(Np)} \\ &\sim \text{Poisson}(Np) \end{aligned}$$



Sometimes this is not even an approximation, but exact because of how the data is gathered. Everyone's favorite example: radioactive decays.

It depends on whether N was a constraint, or "just happened". We will return to this issue when we discuss contingency tables: details of the exact protocol can subtly affect the statistics of the result.

Also recall, $x \sim \text{Poisson}(\lambda) \Rightarrow \mu(x) = \lambda,\ \text{Var}(x) = \lambda$