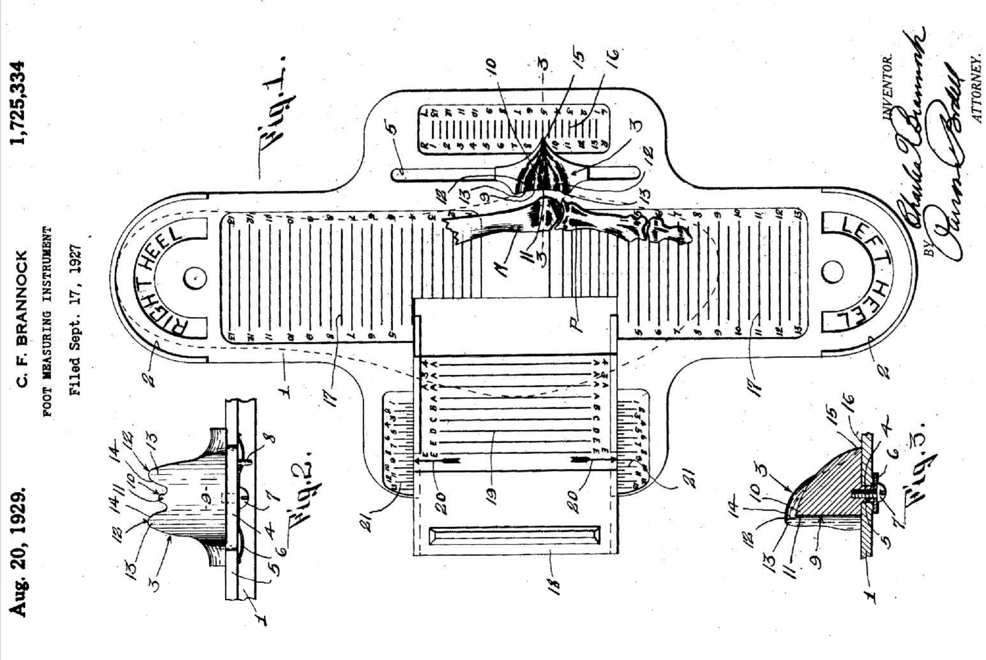
Physics 2660

Lecture 9

Today
- Agreement of data with theory (or other data)
- Fitting a model to your data

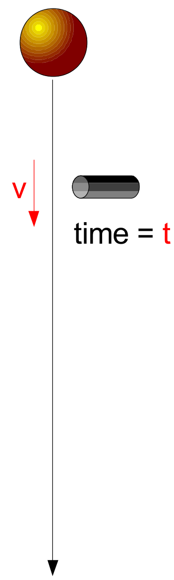
Part 1: Goodness of Fit



Today we'll talk about a couple of related topics: how to objectively evaluate the quality of a theory, given a set of data it claims to describe, and how to adjust the parameters of a theory to make the theory fit the data as well as it can.

We'll start with measuring how good a given theory is.

A Gravity Experiment:



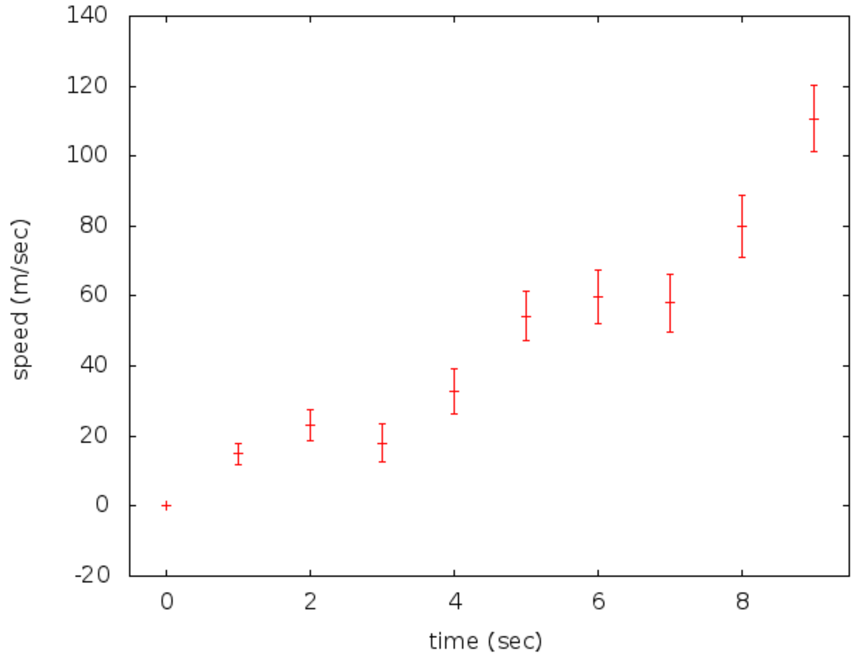
Let's say we're doing an experiment that measures the velocity, v, of a falling object at some time, t.

We collect some data (values of v at various times) and then try to come up with a general theory (a “model”) that is consistent with our observed data.

This is the process of “inductive reasoning”, whereby we look at specific data and try to develop generalizations from them.

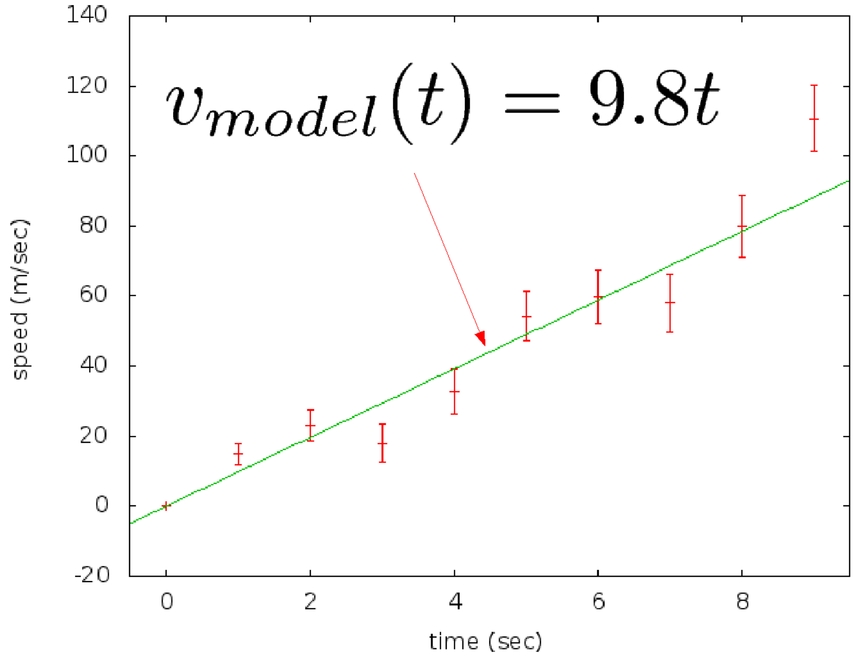
This is a big part of how science progresses!

Experimental Results:



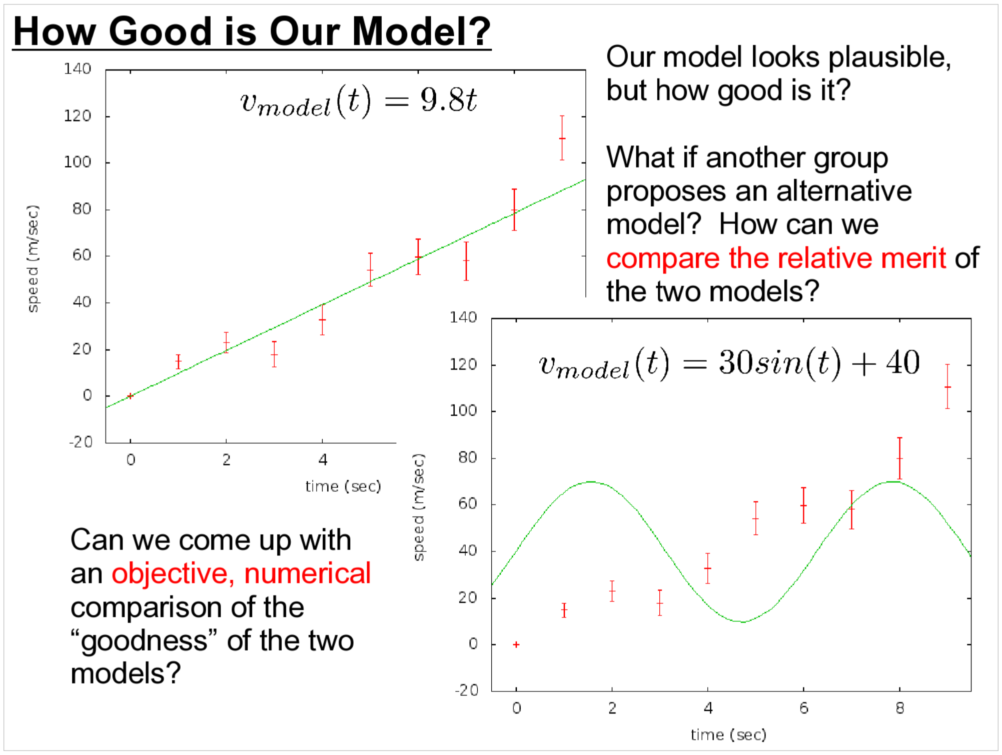
Here's the data we gathered. It shows velocity values for various values of time. After looking at it, we might wonder if a simple theory could explain why the velocities have the values we observed.

Our Model:



Here's our theory. We think it's plausible that the data could be explained by a simple linear relationship between velocity and time.

But how well does our theory fit the data?



What of somebody else has a different theory, like the one on the bottom. We might look at it and think it doesn't fit the data very well, but how can we quantify that?

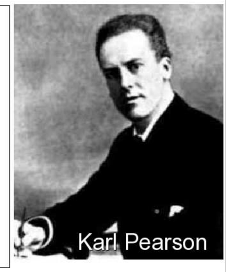
Is there some numerical way to compare these two theories to the data, so we can say which one is more likely?

The Chi-squared Statistic:

One way to objectively compare the quality of various models is by calculating a number called the "χ^2 statistic" for each model. As we'll see, models that are more likely to produce the observed set of measurements will have a smaller value for χ^2.

"The facts once classified, once understood, the judgment based upon them ought to be independent of the individual mind which examines them."

- Karl Pearson, The Grammar of Science, (1900)



$$\chi^2 = \sum_{i=1}^{k} \frac{(v_i - v_{model}(t_i))^2}{\sigma_i^2}$$

Where "k" is the number of independent measurements. k is called the number of "degrees of freedom".

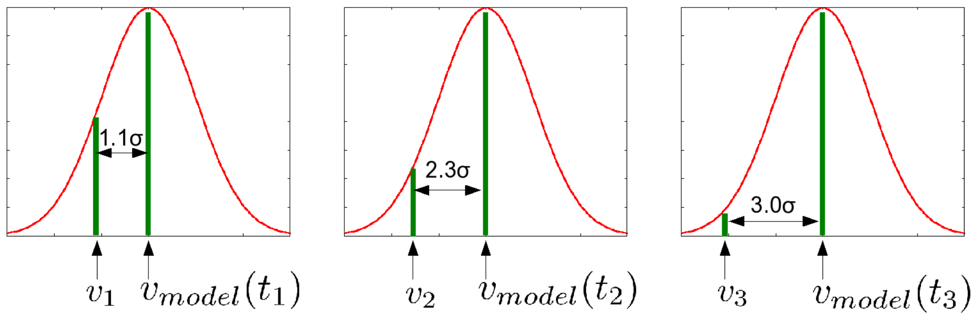
Pearson invented Chi-squared around 1900. He was a "science evangelist" who passionately believed in the power of objectivity and the scientific method (taking this to extremes in some cases, such as embracing eugenics). His book "The Grammar of Science" was a big influence on the young Einstein, and inspired some of his ideas about relativity.

There are several different formulations for chi-squared. For this lecture I've chosen the one that I think is most intuitive. No matter how chi-squared is calculated, it will always be a measure of the overall deviation of your data points from some model, as we'll see a little later.

I'll also stick to a consistent set of variables here, but you'll find that every book (or web site) has its own conventions, so don't be surprised if you see a different set of variables names elsewhere.

Calculating Chi-squared:

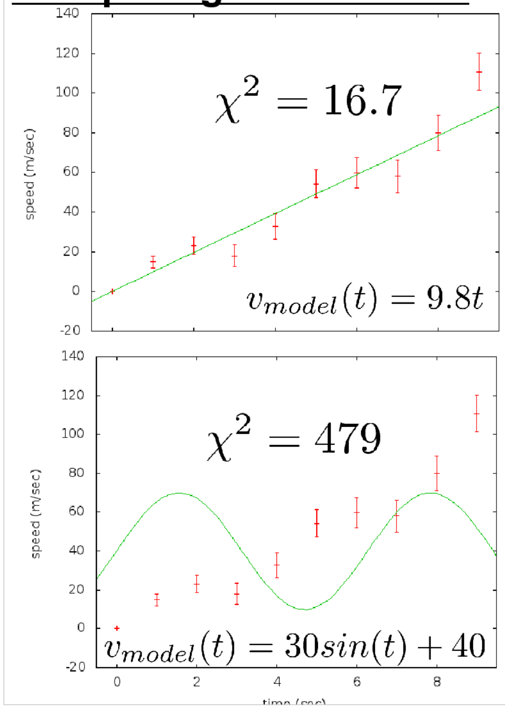
$$\chi^2 = \sum_{i=1}^{k} \frac{(v_i - v_{model}(t_i))^2}{\sigma_i^2}$$



$$\chi^2 = (1.1)^2 + (2.3)^2 + (3.0)^2 + \ldots$$

As you can see, if all of our observed values are close to the predicted values, the calculated chi-squared will be small. If our observed values are far away from the predicted values, chi-squared will be large.

Comparing Our Models:



With χ^2, we can quantitatively compare our two models, to see which is more likely to have produced the data we observed.

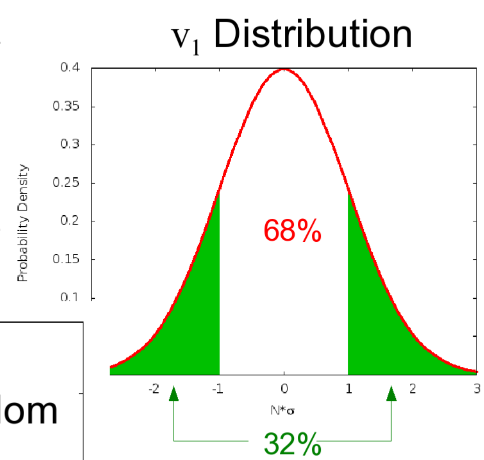
As you can see at left, our linear model wins the χ^2 contest!

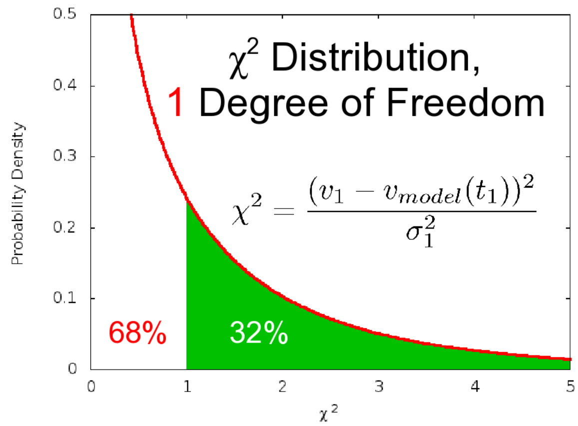
Since χ^2 measures the collective deviation of the data points from the model, the model that produces the smallest χ^2 is most consistent with the data.

We'll prove this a little more rigorously in the next section.

Chi-squared Distribution:

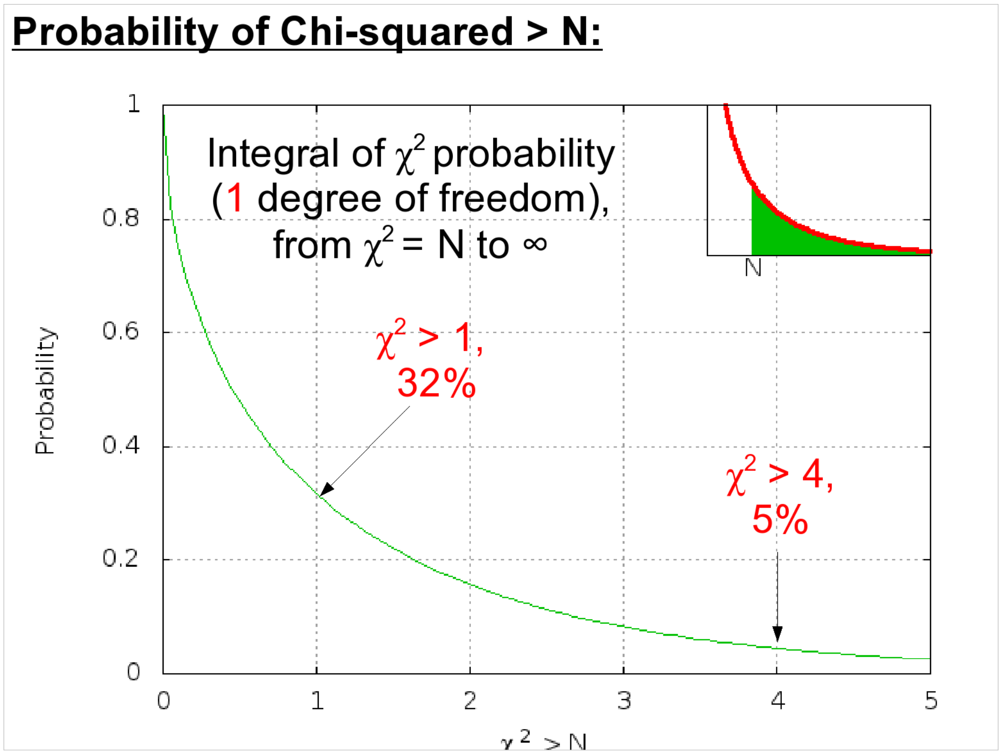
If we just picked one time (say, 5 sec) and repeatedly measured the speed at that time, we'd probably find that our measurements fell into a Normal distribution, like the one at right.





If we generated a χ^2 value for each of these speed measurements, the χ^2 values would fall in a distribution like the one at the left.

Notice that, while the Gaussian distribution covers values both postive and negative, the value of chi-squared will always be positive.

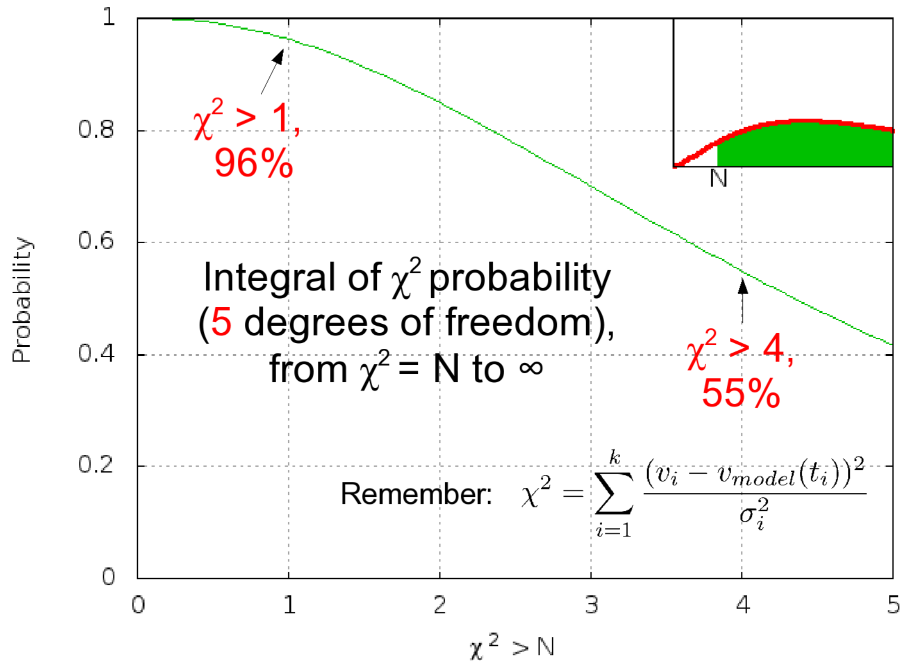


This shows the probability of seeing a chi-squared value in excess of some given number, for 1 degree of freedom (one data point). For example, there's a 32% chance of seeing a chi-squared greater than 1, and a 5% chance of seeing a chi-squared greater than 4.

What good does this do? Couldn't we get the same information by just looking at the integral of the Gaussian (Normal) distribution as we've done before? Why bother looking at this new chi-squared number?

The reason is that chi-squared can be calculated for more than just one data point.

More Degrees of Freedom:



This is like the previous plot, but this time it shows us the probability of seeing a chi-squared value in excess of some given number for 5 degrees of freedom (five data points).

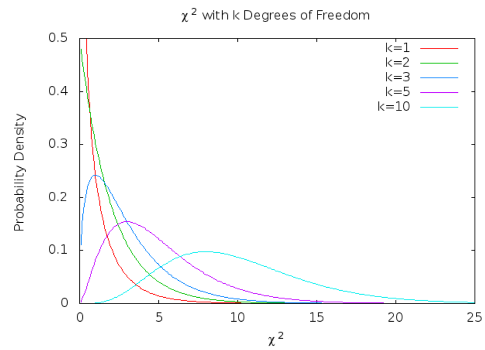
It tells us, for example, that if we have five data points there's a 96% chance we'll see a chi-squared value greater than 1, and a 55% chance we'll see a value greater than 4. There's a smaller and smaller chance of seeing larger and larger chi-squared values, and an extended chart like this would let us quantify those probabilities.

As you can see by looking at the way chi-squared is calculated, you'd expect to see larger values of chi-squared for more degrees of freedom (data points).

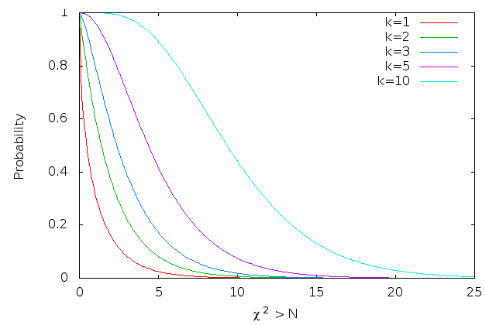
Chi-squared with k Degrees of Freedom:

We can extend this to any number of degrees of freedom (i.e., number of data points).

The top graph shows the chi-squared distribution for k=1, 2, 3, 5 and 10 data points.

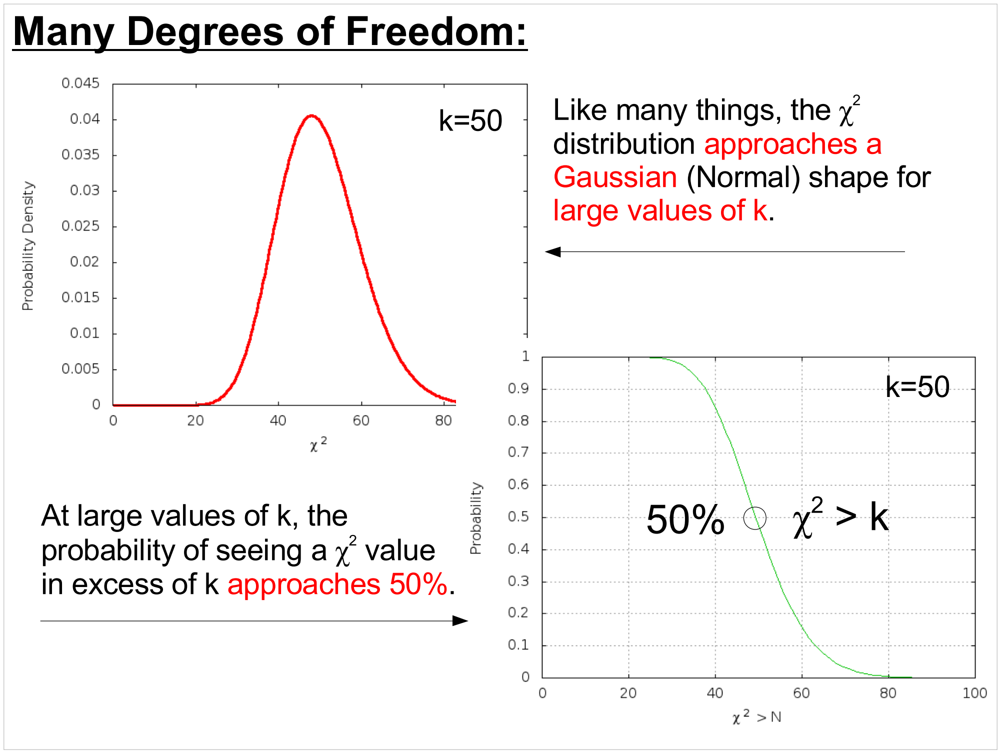


The bottom graph shows the integral probability of observing a chi-squared value in excess of some given number, N.



Notice that, after the first couple of values of k, the curves in the top graph develop a peak, which moves to the right as k increases. This peak is at the most probable value of chi-squared for that value of k.

In the bottom graph, the curves also tend to move toward the right as k increases. As with the top graph, this is due to the fact that chi-squared tends to get bigger as we increas the number of data points (degrees of freedom).



As the number of data points (k) gets bigger, we see that chi-squared tends to be bigger than k about half the time, and smaller than k about half the time. For large values of k, the most probable value of chi-squared is equal to k.

Reduced Chi-squared:

When comparing χ^2 values with different degrees of freedom, it's often useful to look at what's called the "reduced χ^2" value. This is just χ^2 divided by the number of degrees of freedom, k:

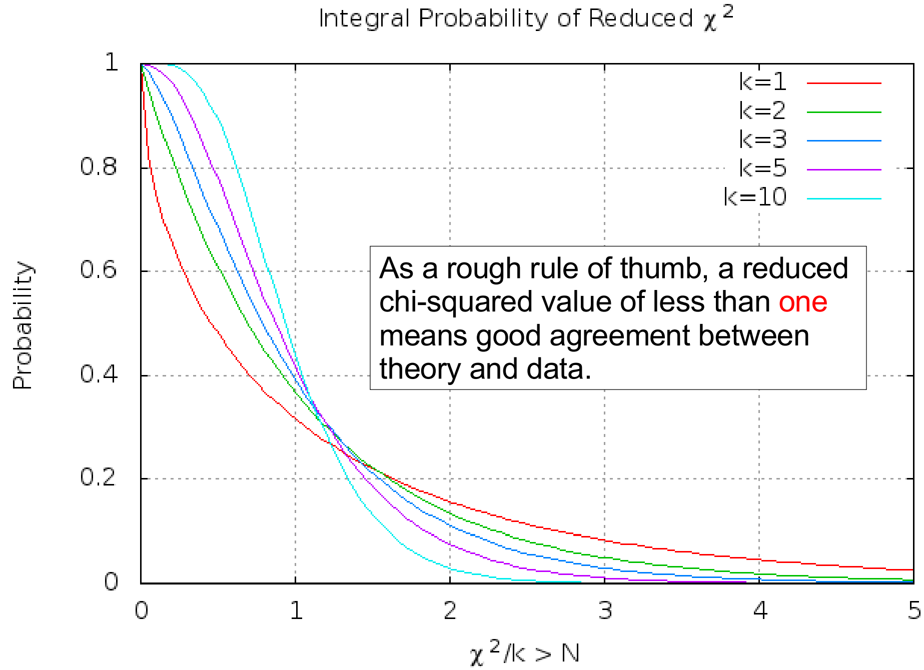
$$\boxed{\chi^2_{red} = \frac{\chi^2}{k}}$$

If all of our data points deviated from the model's predictions by 1 standard deviation, the value of the reduced chi-squared would be 1:

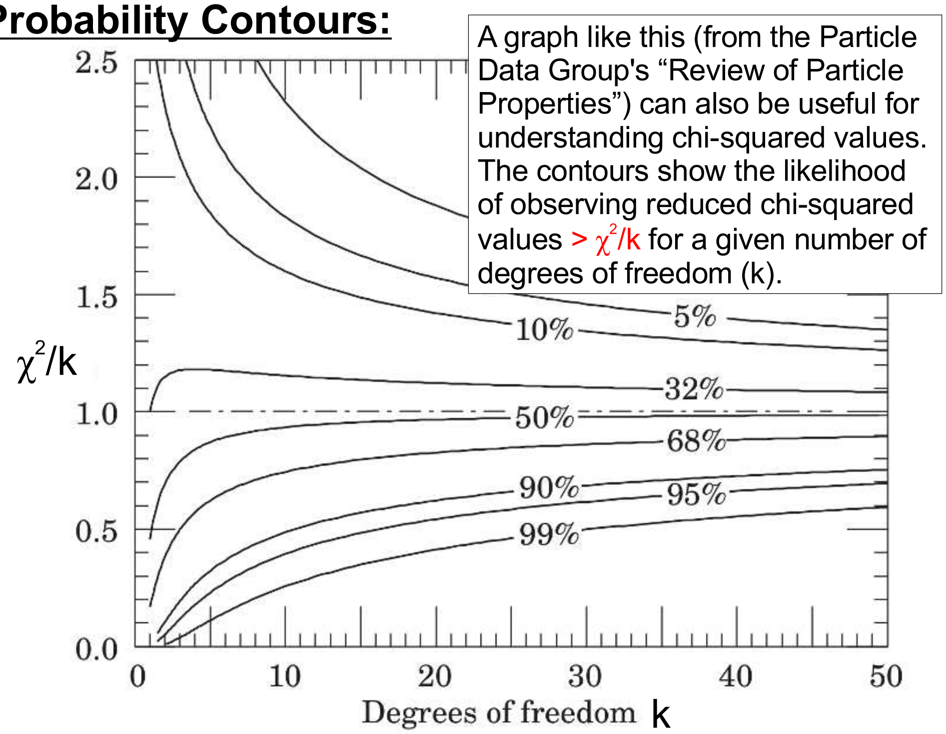
$$\chi^2_{red} = \frac{1}{k}\sum_{i=1}^{k}\frac{(v_i - v_{model}(t_i))^2}{\sigma_i^2} = \frac{1}{k}\sum_{i=1}^{k}\frac{\sigma_i^2}{\sigma_i^2} = \frac{1}{k}k = 1$$

When we report a value for reduced chi-squared, we still need to state the value of k, too, since the shapes of the chi-squared distributions change depending on k.

Probabilities for Reduced Chi-squared Values:

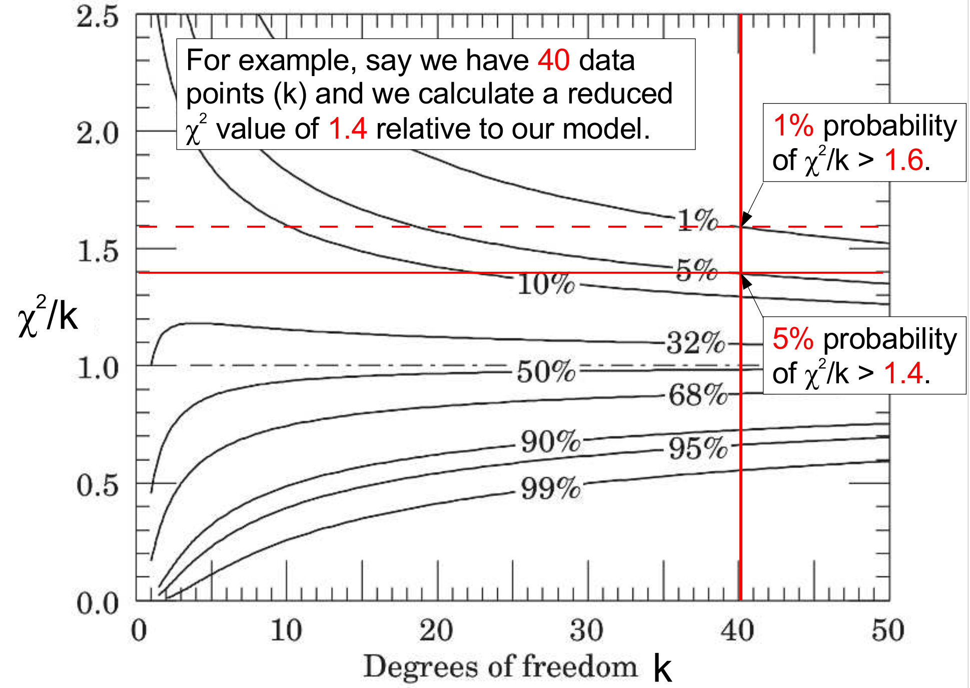


Probability Contours:



We'll see how to use this in the next slide.

Using Probability Contours:



What Questions Can We Answer with χ^2 ?

- ✗ Is my theory correct?
- ✗ Based on my data, what is the probability that my theory is correct?
- ✗ What is the probability that my theory produced the physics we observe?
- ✓ Given a particular theory, what is the probability that data this different or more different could have occurred?
- ✓ Are the data inconsistent with the model?

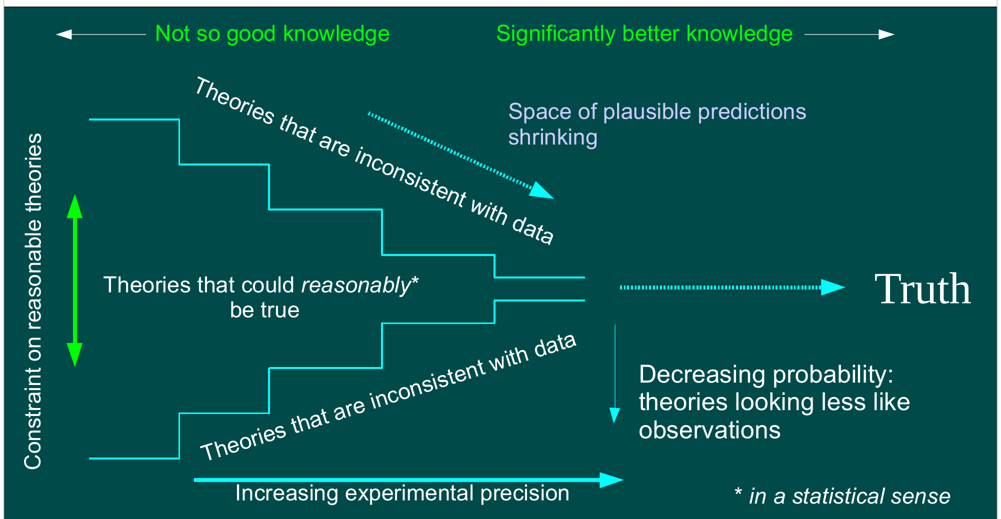
More data might show flaws in your theory. One small data set can't prove that your theory is correct.

There are infinitely many possible theories. And what does "correct" mean? Maybe more than one theory is "correct", in that it makes accurate predictions.

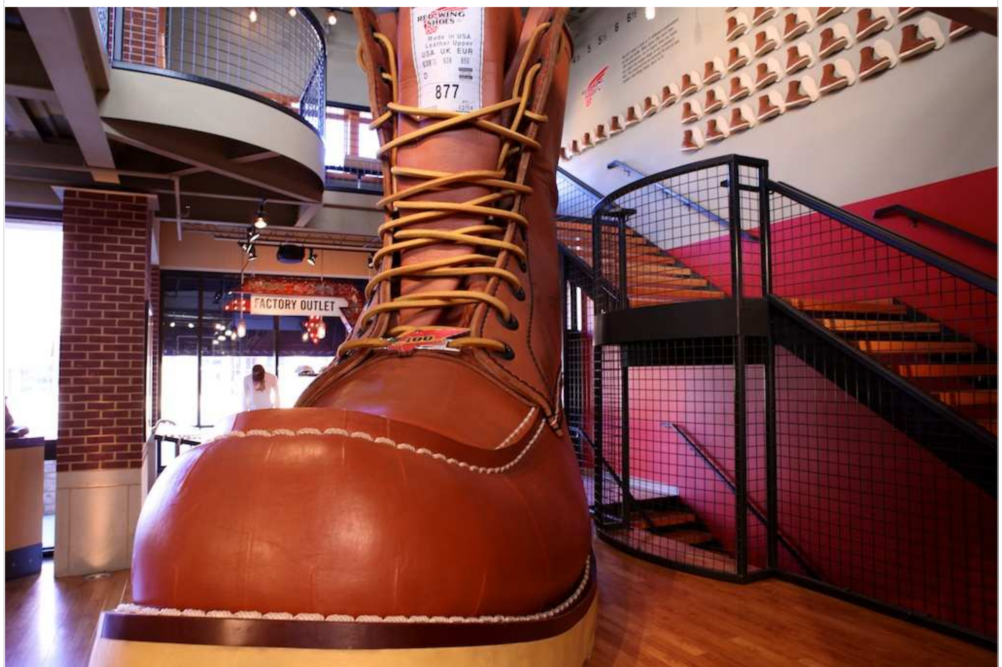
These are the kinds of statements that can be supported by chi-squared values.

The March of Science:

Science progresses by starting with a set of alternative theories, making tests to eliminate some of them, then formulating a new set of refined theories and testing them.



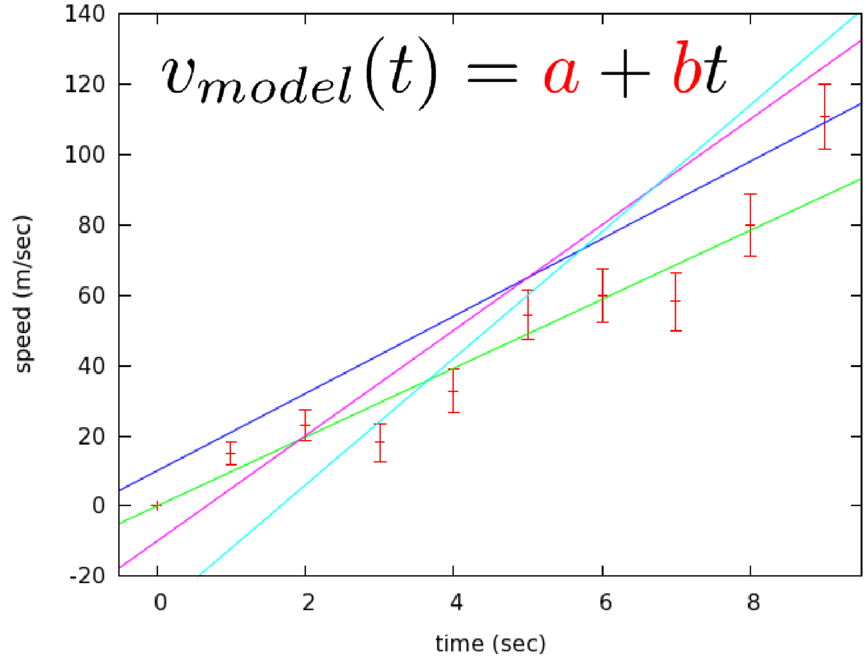
<u>Part 2: Finding the Best Fit:</u>



What if we have a theory that's more general. Say, for example, we think that our data can be explained by a linear relationship between v and t, but we don't know the slope and y-intercept of that line (two adjustable parameters of our theory).

How can we determine the best set of parameters (the ones that make our theory most likely to explain the data)?

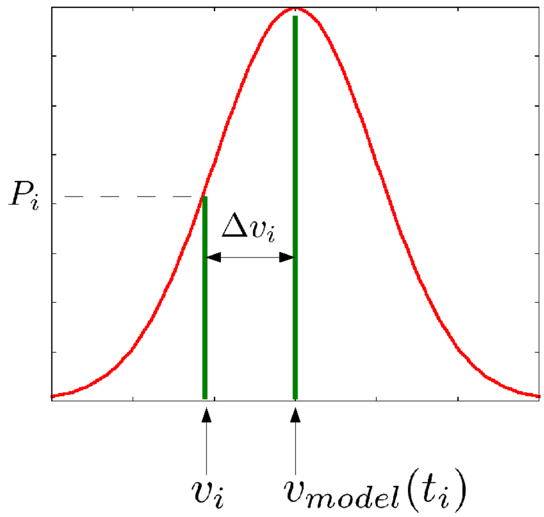
A Parametrized Model:



For example, say that we think our velocity data can be explained by the relationship above, with some (as yet unknown) values for the adjustable parameters *a* and *b*.

How can we find the best values of *a* and *b*?

Probability of Observing a Given Value:



Given a particular choice of model parameters, what's the probability that we'll observe a given velocity value?

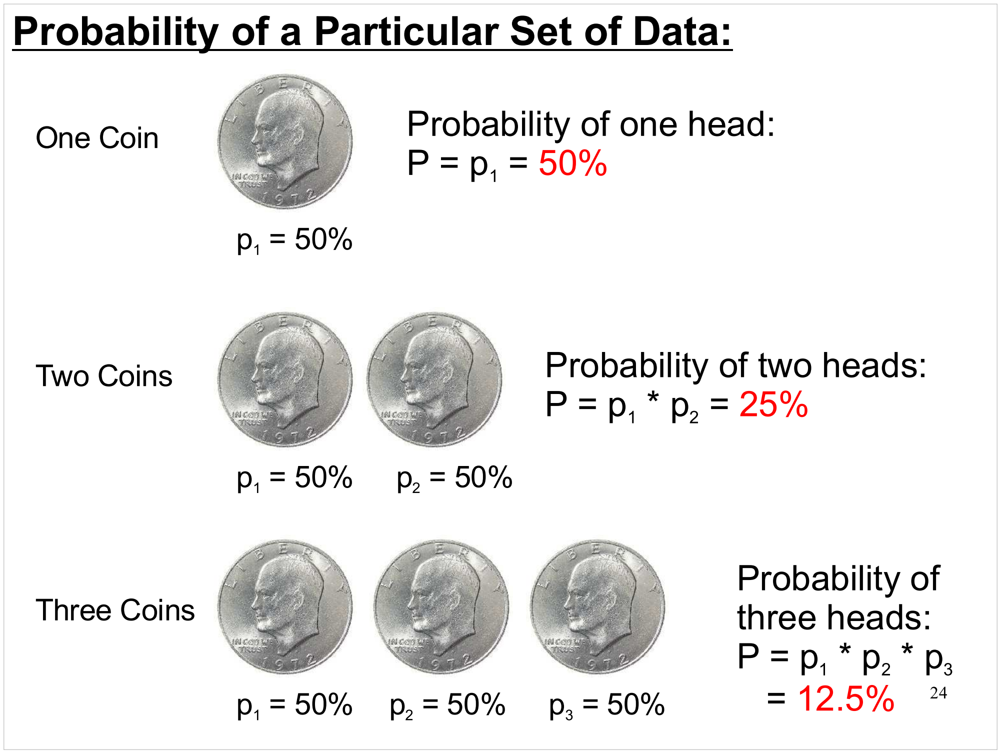
If we assume the observed values are drawn from a Gaussian (Normal) distribution, we can calculate the probability like this:

$$P_i = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}\frac{(v_i - v_{model}(t_i))^2}{\sigma_i^2}}$$

23

The above assumes that the velocity values vary in a Gaussian distribution. This isn't always the case, but it's often good enough.

The equation at the bottom just describes the shape of the Gaussian curve. P_i is the probability of observing a value vi that deviates by Δv_i from the predicted value. The Gaussian (or Normal) curve is just a probability distribution.



So we can compute the probability of one point deviating by a given amount from the predicted value. Now what's the probability of seeing the whole set of data that we observed, with each data point's deviation from the prediction?

As in the example above, it's just the product of the individual probabilities.

Probability of Getting this Set of Values :

So, what's the probability of getting the whole set of values we actually observed? This is just the product of the probabilities of observing the individual values.

$$P = \prod_{i=1}^{k} P_i$$

Substituting in the expression for P_i, we get:

$$P = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma_i^2}}\, e^{-\frac{1}{2}\frac{(v_i - v_{model}(t_i))^2}{\sigma_i^2}}$$

We can profitably rewrite this as two terms, like this:

$$P = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot e^{-\frac{1}{2}\sum\frac{(v_i - v_{model}(t_i))^2}{\sigma_i^2}}$$

Back to Chi-squared:

To find the best set of parameters, *a* and *b*, for our model

$$v(t) = a + b\,t$$

we want to maximize the probability, P. This will tell us which values of a and b are most likely to produce our observed data.

$$P = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot e^{-\frac{1}{2}\sum \frac{(v_i - v_{model}(t_i))^2}{\sigma_i^2}}$$

Since the first term doesn't depend on a and b, the problem reduces to maximizing the second, exponential, term. This is just equivalent to minimizing the sum in the exponent. But wait! This sum is just:

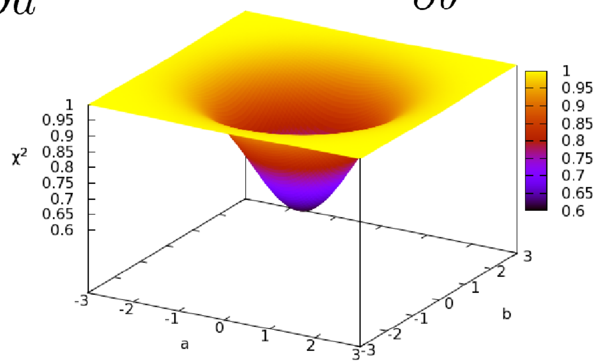
$$\chi^2 \equiv \sum_{i=1}^{k} \frac{(v_i - v_{model}(t_i))^2}{\sigma_i^2}$$

The product “P” is called a likelihood function. We could construct a likelihood function for any set of data, no matter whether the data points vary in a Gaussian way or some other way. We would just need to insert the appropriate probability expression into the product.

Minimizing Chi-squared:

So, to find the best values for *a* and *b* in our model, we must find the values of *a* and *b* that minimize χ^2. We can find the minimum by looking for the values of *a* and *b* that satisfy the following conditions:

$$\frac{\partial \chi^2}{\partial a} = 0 \qquad \text{and} \qquad \frac{\partial \chi^2}{\partial b} = 0$$



You can think of the possible values of chi-squared as the surface above. We get a different value of chi-squared for each choice of values for our adjustable parameters, a and b.

Minimizing chi-squared means finding the values of a and b that produce the smallest chi-squared value.

Minimum, Assuming all Sigmas are Equal:

Let's start out by assuming that the standard deviation is the same for all of our data points. In other words, set all the σ_i values to the same number, σ.

Our theory predicts that $v_{model}(t) = a + bt$, so:

$$\begin{aligned}\frac{\partial \chi^2}{\partial a} &= \frac{\partial}{\partial a}\left[\frac{1}{\sigma^2}\sum (v_i - a - bt_i)^2\right] \\ &= \frac{-2}{\sigma^2}\sum (v_i - a - bt_i) = 0\end{aligned}$$

$$\begin{aligned}\frac{\partial \chi^2}{\partial b} &= \frac{\partial}{\partial b}\left[\frac{1}{\sigma^2}\sum (v_i - a - bt_i)^2\right] \\ &= \frac{-2}{\sigma^2}\sum [t_i(v_i - a - bt_i)] = 0\end{aligned}$$

We'll look at the case where all sigmas aren't equal in a minute.

Solution for Uniform Sigma:

We can solve the preceding equations for *a* and *b*:

$$\boxed{a = \frac{1}{\Delta}(\sum t_i^2 \sum v_i - \sum t_i \sum t_i v_i)}$$

$$\boxed{b = \frac{1}{\Delta}(k \sum t_i v_i - \sum t_i \sum v_i)}$$

Where, for convenience, we've defined the quantity Δ as:

$$\Delta = k \sum t_i^2 - (\sum t_i)^2$$

(Notice that σ doesn't appear anywhere in these equations, because we've assumed that all the σ_i values are the same.)

We can plug numbers into these equations and get the values of *a* and *b* that maximize the probability of getting our observed data.

These look like long, complicated expressions to do by hand, but you can see that they'd be pretty easy to do with a computer program. We just loop over all of the data points, adding things up.

Solution for Non-uniform Sigma:

If we don't assume that all the σ_i values are the same, we can still work through the algebra and come up with a (slightly more complicated) solution for the best values of *a* and *b*:

$$\boxed{a = \frac{1}{\Delta}\left(\sum \frac{t_i^2}{\sigma_i^2} \sum \frac{v_i}{\sigma_i^2} - \sum \frac{t_i}{\sigma_i^2} \sum \frac{t_i v_i}{\sigma_i^2}\right)}$$

$$\boxed{b = \frac{1}{\Delta}\left(\sum \frac{1}{\sigma_i^2} \sum \frac{t_i v_i}{\sigma_i^2} - \sum \frac{t_i}{\sigma_i^2} \sum \frac{v_i}{\sigma_i^2}\right)}$$

Where, this time, we've defined the quantity Δ as:

$$\Delta = \sum \frac{1}{\sigma_i^2} \sum \frac{t_i^2}{\sigma_i^2} - \left(\sum \frac{t_i}{\sigma_i^2}\right)^2$$

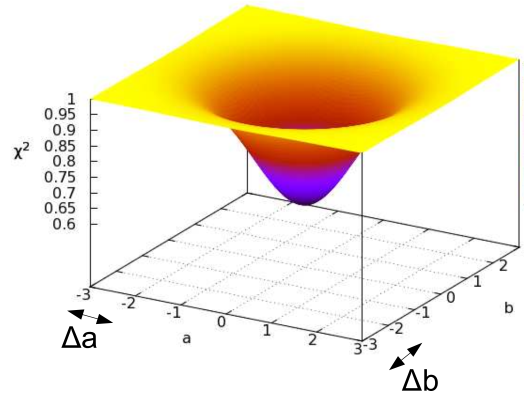
It would be tedious to work through these calculations by hand, but it's easy to write a computer program to do them for us.

Again, these calculations could easily be done in a program by just looping over the data points.

Fitting Arbitrary Functions:

We've been talking about fitting a linear function, $v_{model}(t) = a + b\,t$, to a set of data. What if we want to fit a more complex function?

In some cases, we could follow a similar procedure and come up with an analytical solution giving the best-fit values for the parameters in our model.

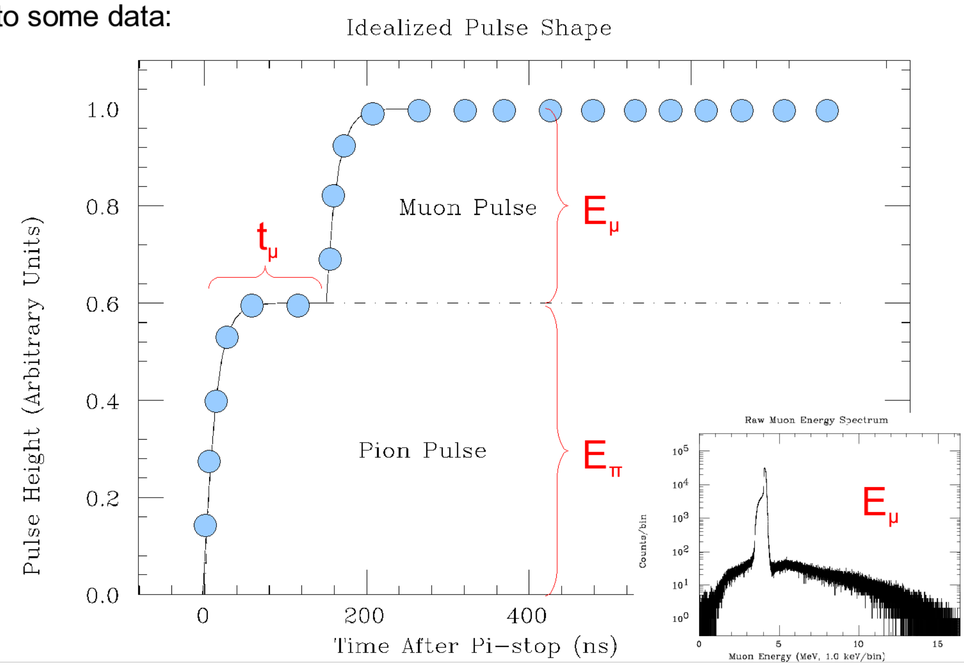


For complicated functions, we can just try different parameter values (*a* and *b*, in our example), calculating χ^2 for each one until we find the minimum. We can do this by brute force, stepping through a grid of values, or we can use root-finding algorithms like Newton's method to help us find the minimum quickly.

We can do this for any model, with any number of parameters.

Example of Fitting an Arbitrary Function:

Here's a real-world example of fitting a complicated 3-parameter function to some data:



This data is from an experiment in which pions were stopped in a detector, where they decayed into muons and neutrinos. For each pion, we got a signal out of the detector like the one in the large graph. The first part of the signal shows the pion's energy. Then, when the pion decays, we get another bump of energy.

For each of these signals, a program needed to find the best-fit values of t_μ, E_μ and E_π. We were interested in determining the muon energy.

The program processed many millions of events in this way, finding the muon energy for each.

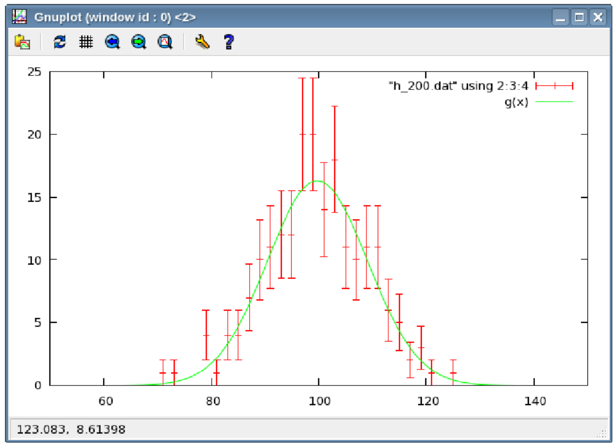
Fitting with Gnuplot:

```
After 5 iterations the fit converged.
final sum of squares of residuals : 11.2835
rel. change during last iteration : -9.9567e-06

degrees of freedom    (FIT_NDF)                        : 22
rms of residuals      (FIT_STDFIT) = sqrt(WSSR/ndf)    : 0.716162
variance of residuals (reduced chisquare) = WSSR/ndf   : 0.512888

Final set of parameters
=======================

a               = 16.3073
m               = 99.7576
s               = 9.3664
```



Fortunately, programs like gnuplot are there to help us with simple fitting jobs, so we don't have to do it by hand!

We'll experiment with this in lab this week.

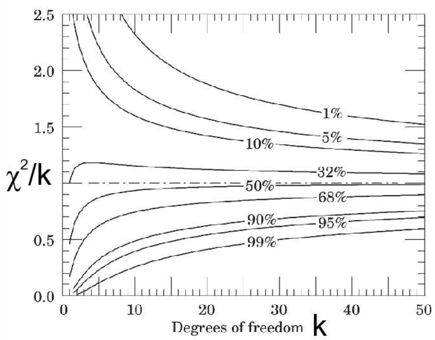
Gnuplot really isn't great at this kind of thing, but it's useful for some simple tasks.

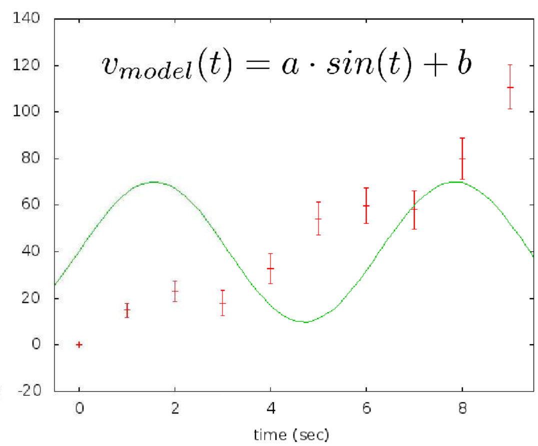
How Good is Our Best Fit?

So, we've found find a set of parameters that minimizes χ^2. Does that mean that our model, when we use these parameters, is a good one? Not necessarily.

A bad model may not fit our data very well even with the best possible choice of parameters. Consider the model below. No matter what values we choose for *a* and *b*, it still won't fit the data very well.

After we've fit our model to the data, we need to look at the minimum value of χ^2 to see how good our best fit really is.





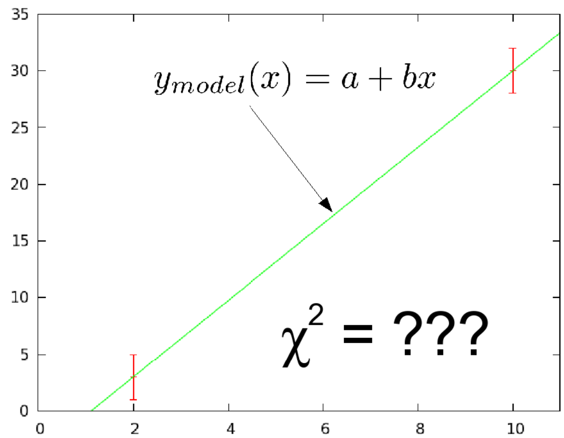
Degrees of Freedom After Fitting:

Consider a case where we have only two data points, and we're fitting a straight line to them. We can make the line go exactly through the points. In this case, we can make χ^2 equal to exactly zero!

This doesn't tell us anything useful. If the number of points equals the number of fitting parameters, we can make any arbitrary function fit the data perfectly.

Because of this, after fitting a function to some data we report the resulting χ^2 as though the number of degrees of freedom have been reduced by an amount equal to the number of fitting parameters:

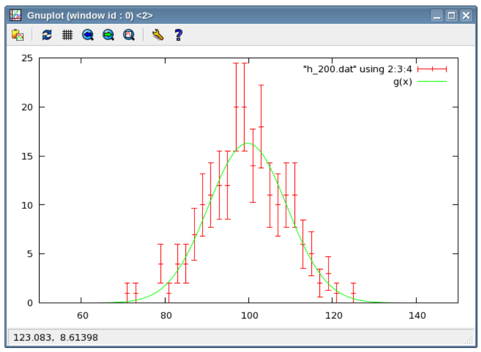
$$k_{fit} = k - n_{parameters}$$

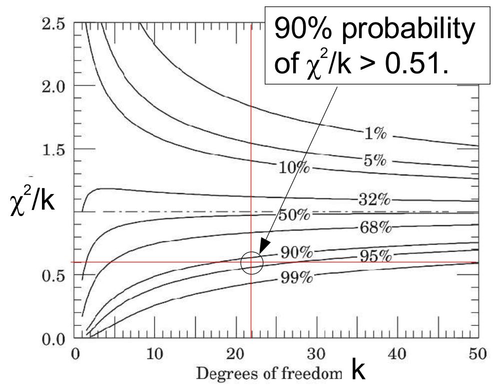


If the number of adjustable parameters is greater than or equal to the number of data points, then chi-squared is undefined.

Checking Goodness of Fit:

Number of data points = 25
Number of fitting parameters = 3
χ^2/k = 0.512888
k = 22





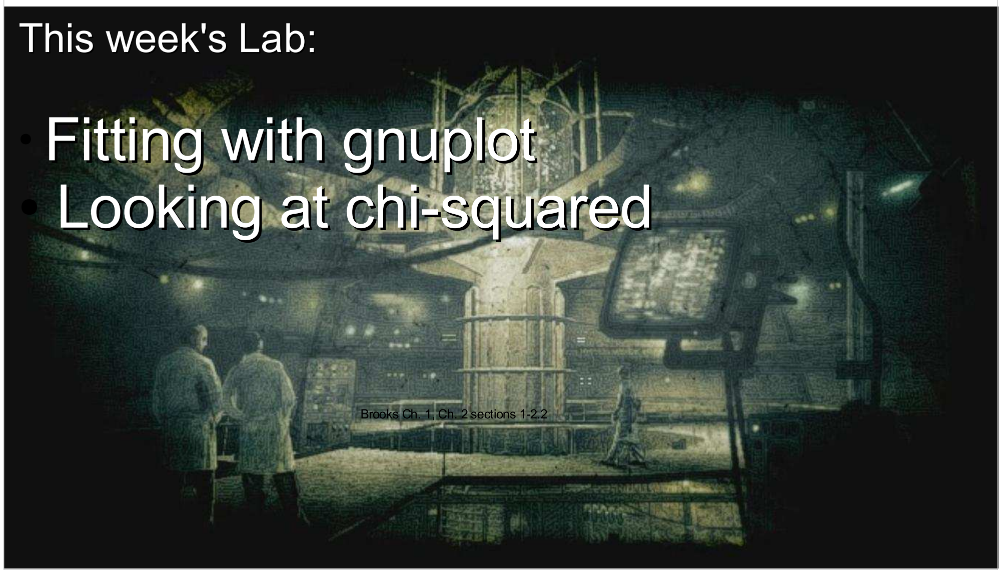
So, we see that there's a 90% probability that data points would have a greater chi-squared value relative to this model. That sounds a little fishy. Why are the points so close to the model? Did somebody cheat? Were the values of sigma over-estimated? Did gnuplot do something wrong?

Next Time:

- More on Fitting
- Bitwise operators and binary files

This week's Lab:

- Fitting with gnuplot
- Looking at chi-squared



Brooks Ch. 1, Ch. 2 sections 1-2.2

The End

Thanks!