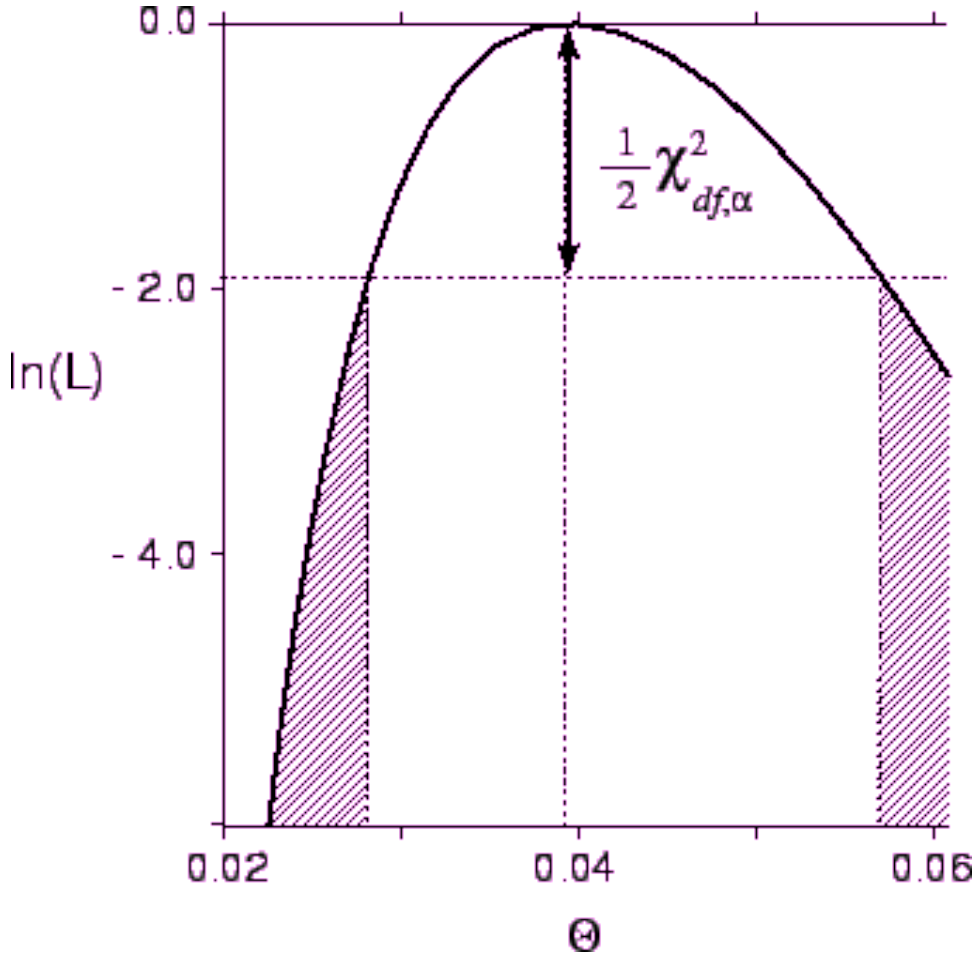
Physics 509: Hypothesis tests and goodness of fit

Scott Oser
Lecture #14



Remember the Neyman-Pearson lemma?

The acceptance region giving the highest power (and hence the highest signal purity) for a given significance level α (or selection efficiency 1-α) is a region of the test statistic space **t** such that:

$$\frac{g(\boldsymbol{t}|H_0)}{g(\boldsymbol{t}|H_1)} > c$$

Here $g(\mathbf{t}|H_i)$ is the probability distribution for the test statistic (which may be multi-dimensional) given hypothesis H_i, and *c* is a cut value that you can choose so as to get any significance level α you want.

This ratio is called the likelihood ratio.

It's actually possible to make a stronger statement.

Not only is the likelihood ratio the most powerful test for any given significance level, it also has some special properties ...

Often we can parameterize the probability distribution as a function of some parameter space Θ. The null hypothesis is often stated as saying that θ lies in a specified lower dimensional subspace Θ_0 of the total parameter space.

The likelihood ratio is given by

$$\Lambda(\vec{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta|\vec{x})}{\sup_{\theta \in \Theta} L(\theta|\vec{x})}$$

Likelihood ratio theorem

If the null hypothesis is true and the data consist of N independent, identically distributed samples from the probability distribution, then as N $\rightarrow \infty$ the statistic -2 ln Λ will be asymptotically distributed as a χ^2 with the number of degrees of freedom equal to the difference in the dimensionality of Θ and Θ_0.

$$\Lambda(\vec{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta|\vec{x})}{\sup_{\theta \in \Theta} L(\theta|\vec{x})}$$

Put another way, if the null hypothesis is a subspace (not just a subset) of the larger parameter space, then you can use a χ^2 to model the likelihood ratio!

An application of the likelihood ratio test

You flip a coin 1000 times and get 550 heads. Is this a fair coin?

H_0: p=0.5

H_1: 0≤p≤1

$$\ln L = \sum \ln P(D|P) = N_H \ln p + N_T \ln(1-p)$$

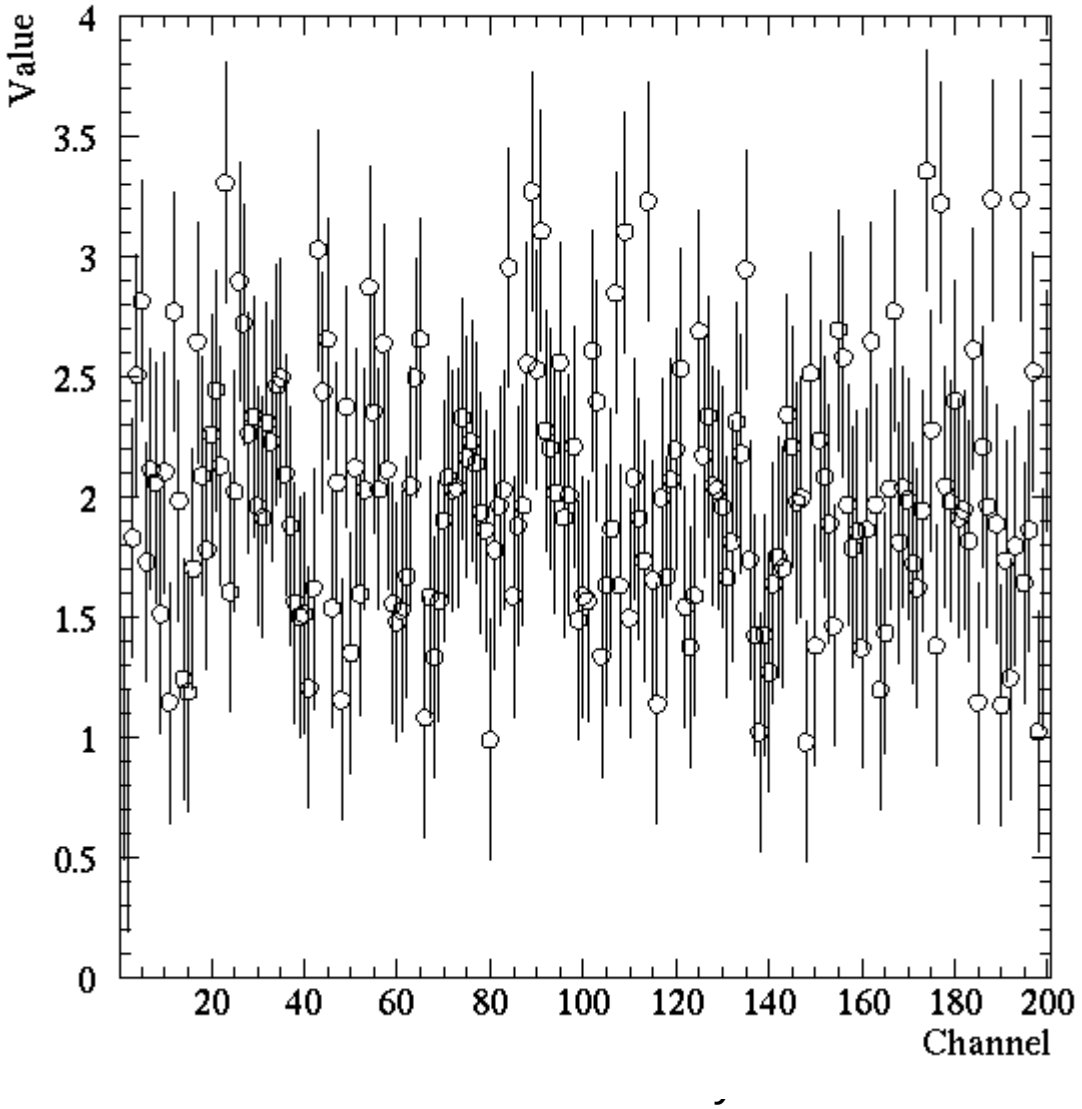
ML estimate for p under H_1 is p=0.55. The likelihood ratio is:

$-2\,\Delta(\ln L) = -2[(550 \ln 0.5 + 450 \ln 0.5) - (550 \ln 0.55 + 450 \ln 0.45)]$

$= 10.02$

This statistic should follow a chi-squared with one degree of freedom. P=0.0017

Hypothesis test: is there a bump?



Is there a spectral feature near channel 90 in this noisy spectrum?

Data has Gaussian errors with σ=0.5.

Background level is unknown but assumed to be flat.

Predicted form of peak is:

$$A \exp[-0.15 \cdot (x-90)^2]$$

Try a likelihood ratio test

Null hypothesis: Data distributed with normal errors around a flat line at some unknown level B. Negative log likelihood has the form:

$$-\ln L = \frac{1}{2}\sum_{i=1}^{200}\left|\frac{(y_i - B)}{\sigma}\right|^2$$

Peak hypothesis: Data distributed around flat line at B + peak of amplitude A

$$-\ln L = \frac{1}{2}\sum_{i=1}^{200}\left|\frac{(y_i - B - A\exp[-0.15\cdot(x_i - 90)^2])}{\sigma}\right|^2$$

Result of likelihood ratio test for bump-hunting

The null hypothesis has one free parameter, B. The peak hypothesis has two free parameters: B and A.

For this particular data set and alternate hypotheses:

Flat: best value is B=1.9985, and of course A≡0
Peak: best values are B=1.9759, A=0.9847

$2\,\Delta\,(\ln L) = 12.15$

How improbable is this?

Likelihood ratio theorem doesn't apply!

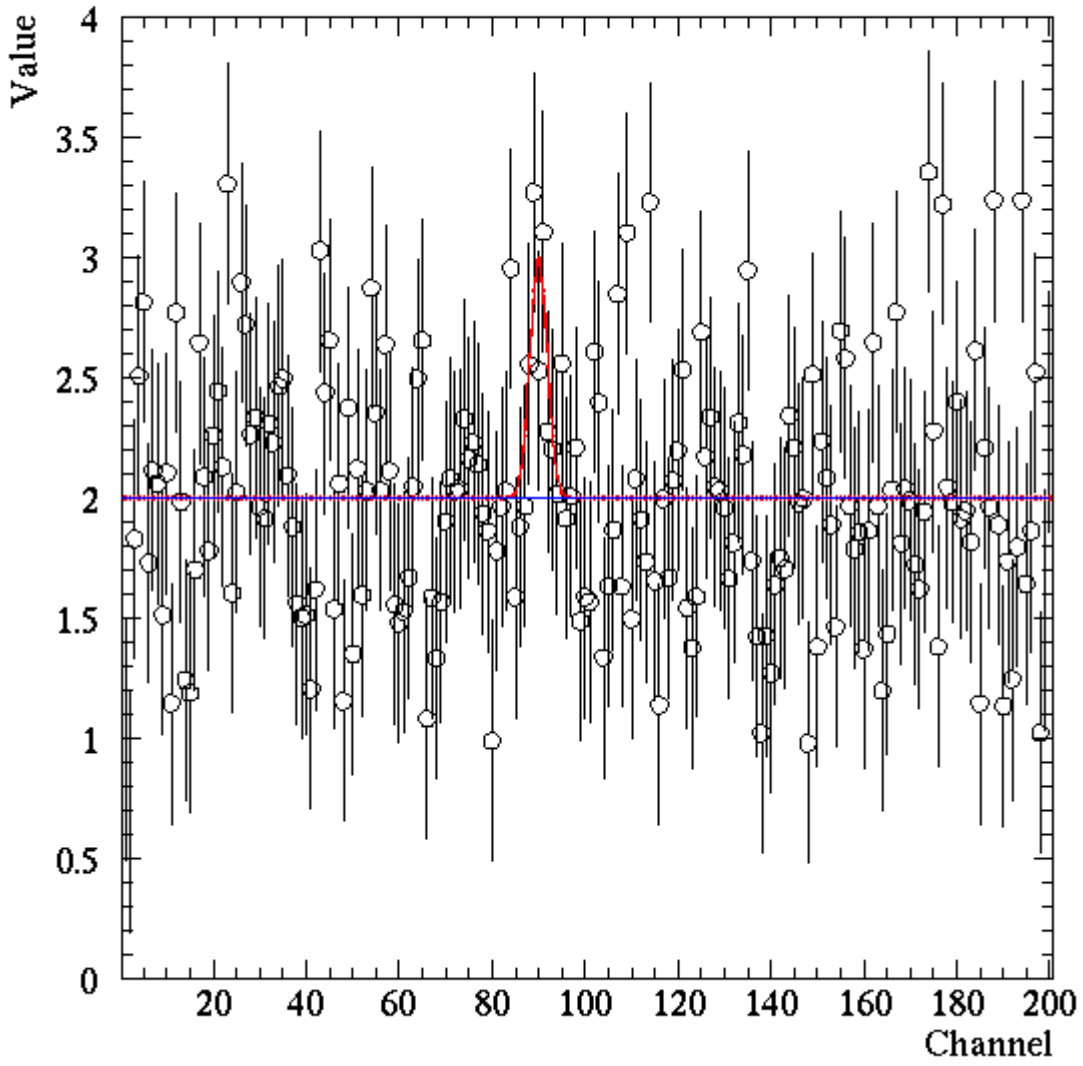
It's very tempting to say that the null hypothesis is a subspace of the more general hypothesis that there is a peak. You'd then conclude that $-2\,\Delta \ln L$ should have a χ^2 distribution with one degree of freedom, since more general hypothesis introduces one more free parameter. However, there's a catch: the data are not identically distributed samples from a parent distribution (at least not for the peak hypothesis)!

So the likelihood ratio theorem doesn't hold. We can still use the likelihood ratio test, but cannot assume that the LR will have an asymptotically χ^2 distribution.

Solution: use Monte Carlo. Generate several thousand fake data sets drawn from the null hypothesis, calculate $-2\,\Delta \ln L$, and see how often you get a value larger than the observed value of 12.15.

Answer: P(>12.15) = 0.0023.

The bump revealed!



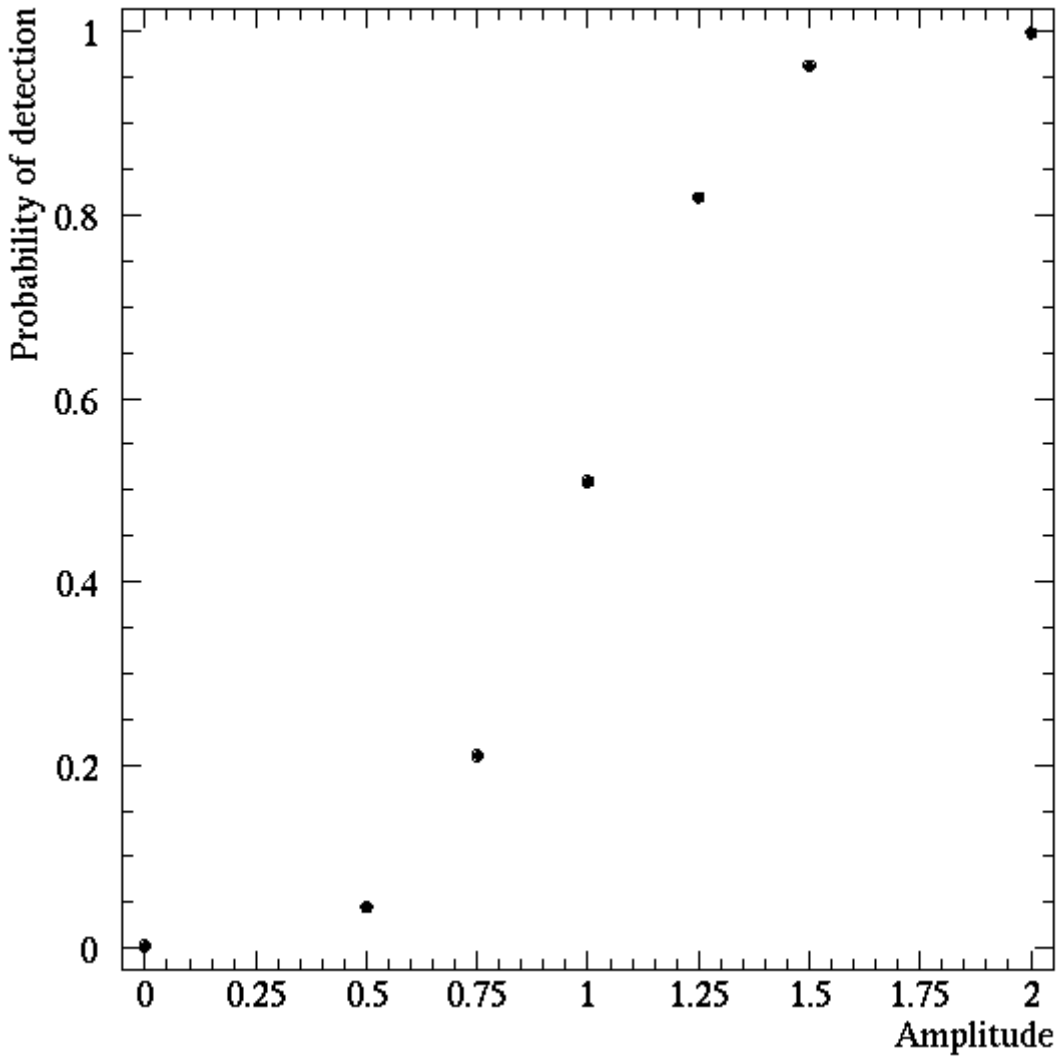
This particular hypothesis test is two-sided: we would also reject the null hypothesis if there were a dip.

If you wanted to do a one-sided test (reject flat only for a positive deviation), try fitting for the amplitude, and seeing by how many sigma the amplitude is positive.

How powerful is the test?

Let's suppose we will test for bumps at the 99.9% significance level (chance probability of null hypothesis being falsely rejected is 0.1%). Monte Carlo says to reject if $-2 \Delta (\ln L) > 13.82$

How sensitive is the test? To determine this, apply it to simulated data sets containing real peaks of varying amplitudes, and see how often they are detected.



Goodness of fit tests

A goodness of fit test is a special kind of hypothesis test. You're simply trying to answer the question “Does the model describe the data?”

To a large extent a goodness of fit test doesn't require you to specify an alternate hypothesis. You're not trying to decide “Is model A better than model B?” Rather, model A is being considered on its own.

There exist a number of goodness of fit tests. We'll study a number of them, and discuss their strengths and weaknesses.

χ^2 goodness of fit tests

The most familiar of all goodness of fit tests is the χ^2 test. The χ^2 statistic is defined as:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(n_i - \nu_i)^2}{\nu_i}$$ binned data, with ν_i = the expected number of events

Be very careful to distinguish the χ^2 statistic, as defined above, from the χ^2 distribution. The *statistic* is a number calculated from the data, while the latter is a family of probability distributions.

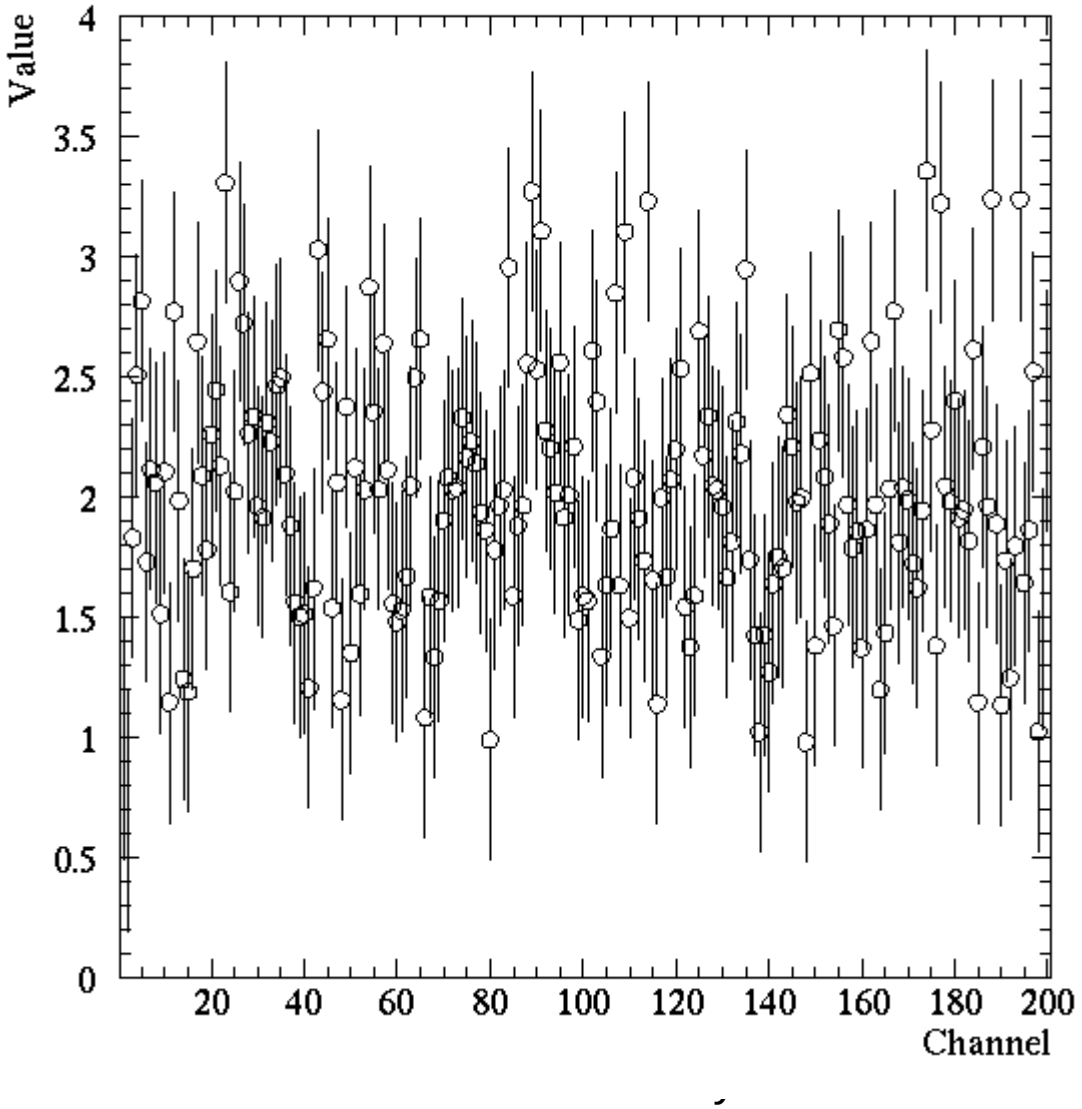
χ^2 goodness of fit tests

Of course the whole point is that often the χ^2 *statistic* will be distributed according to a χ^2 *distribution* with N-m degrees of freedom, where N is the number of data points and m is the number of free parameters fitted out of the data.

This only happens if the errors are distributed according to a Gaussian. (Recall that a χ^2 distribution with N degrees of freedom is what you get if you generate N Gaussian variables with μ=0 and σ=1, then square each one and add up the sum.)

Warning: It is often not the case that the χ^2 statistic follows a χ^2 distribution. Use Monte Carlo to be sure.

Relative vs. absolute χ^2 tests



Remember the bump-hunting data set? It has χ^2=225 for 199 d.o.f.

Probability of getting χ^2 this large or larger is 9.8%.

Reminder: fastest way to calculate this is to remember that

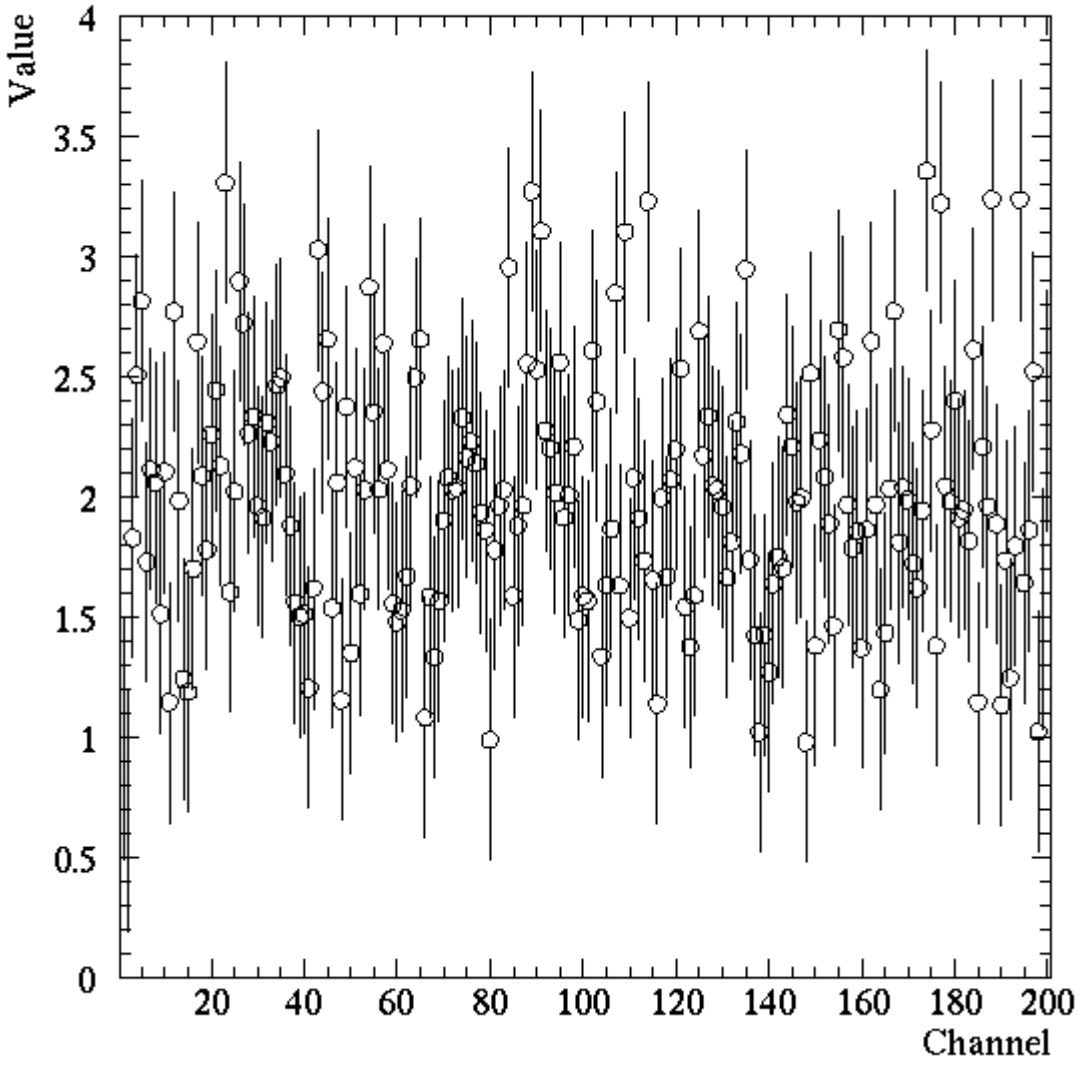
$$\sqrt{2\chi^2}$$

is a Gaussian with mean

$$\sqrt{2N-1}$$

and width=1, when N>30.

Relative vs. absolute χ^2 tests



A standard χ^2 goodness-of-fit test barely rejects the null hypothesis test at the 90% level, but the likelihood ratio test we used to compare the flat to peak models earlier had a much smaller P value of 0.0023.

There's a lesson here: the χ^2 test is a relatively weak test for testing a model.

Alternate hypothesis space for χ^2

There's a relatively easy way to understand why the χ^2 test is pretty weak. Although we say that a goodness of fit test doesn't require you to specify an alternate set of hypotheses, there is an implicit set —the set of all possible variations in the data! This is a very wide set of alternate hypotheses.

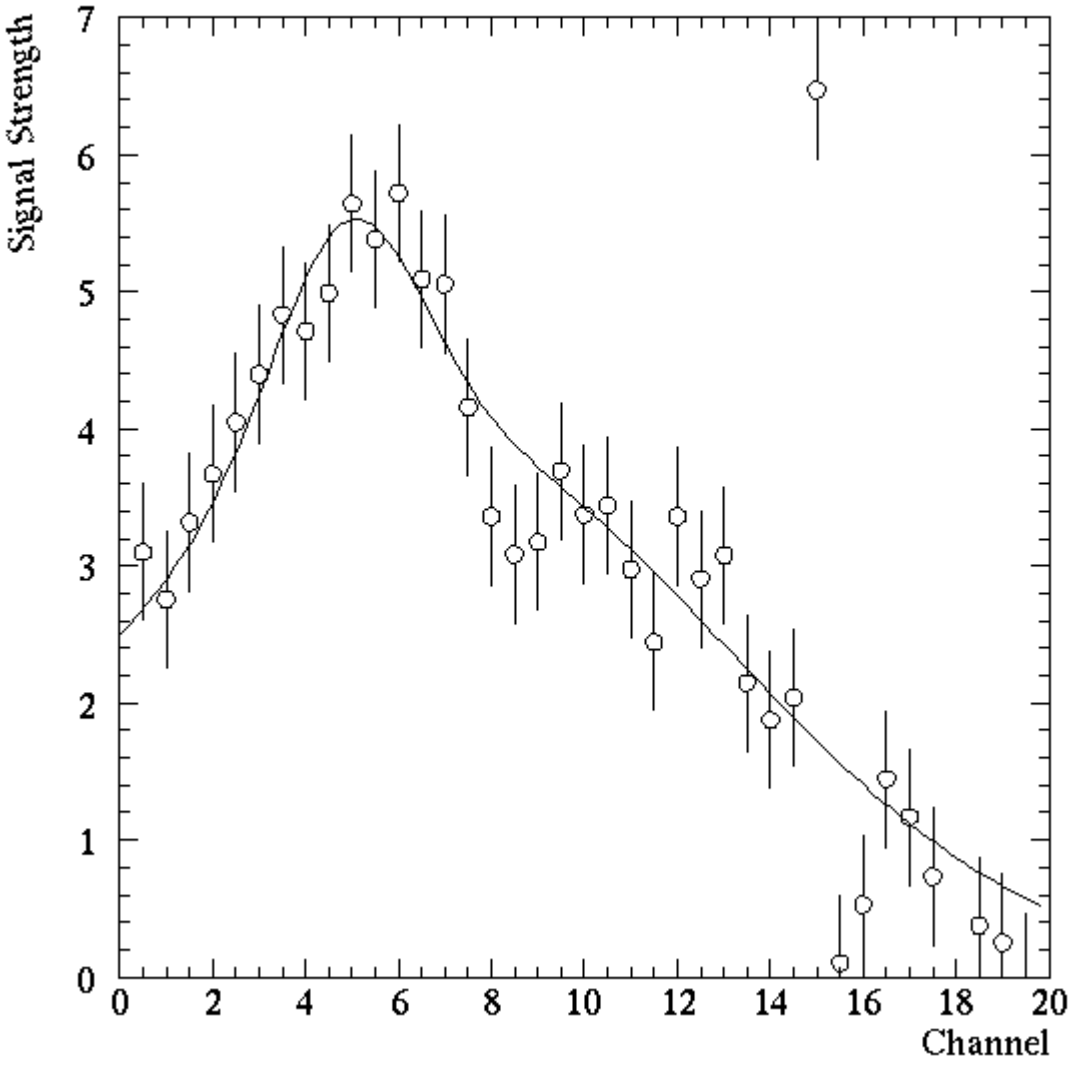
The more you can limit the set of alternate hypotheses under consideration, the stronger your statistical tests will be. It's easier to confirm or rule out the hypothesis that the luminosity of a star varies sinusoidally with a period of 3 weeks than it is to test the statement “the luminosity of the star is periodic with some undetermined period, phase, and light curve!”

What does a bad χ^2 tell you about your model?

Imagine measuring a spectrum using an instrument with poor resolution (σ=3 channels). What you observe is the underlying spectrum convolved with the instrument's resolution, which smears things out.

Your model is that there are two narrow lines in the data. You fit two Gaussians. Best-fit χ^2 = 132 for 35 degrees of freedom.

Do you reject the model?

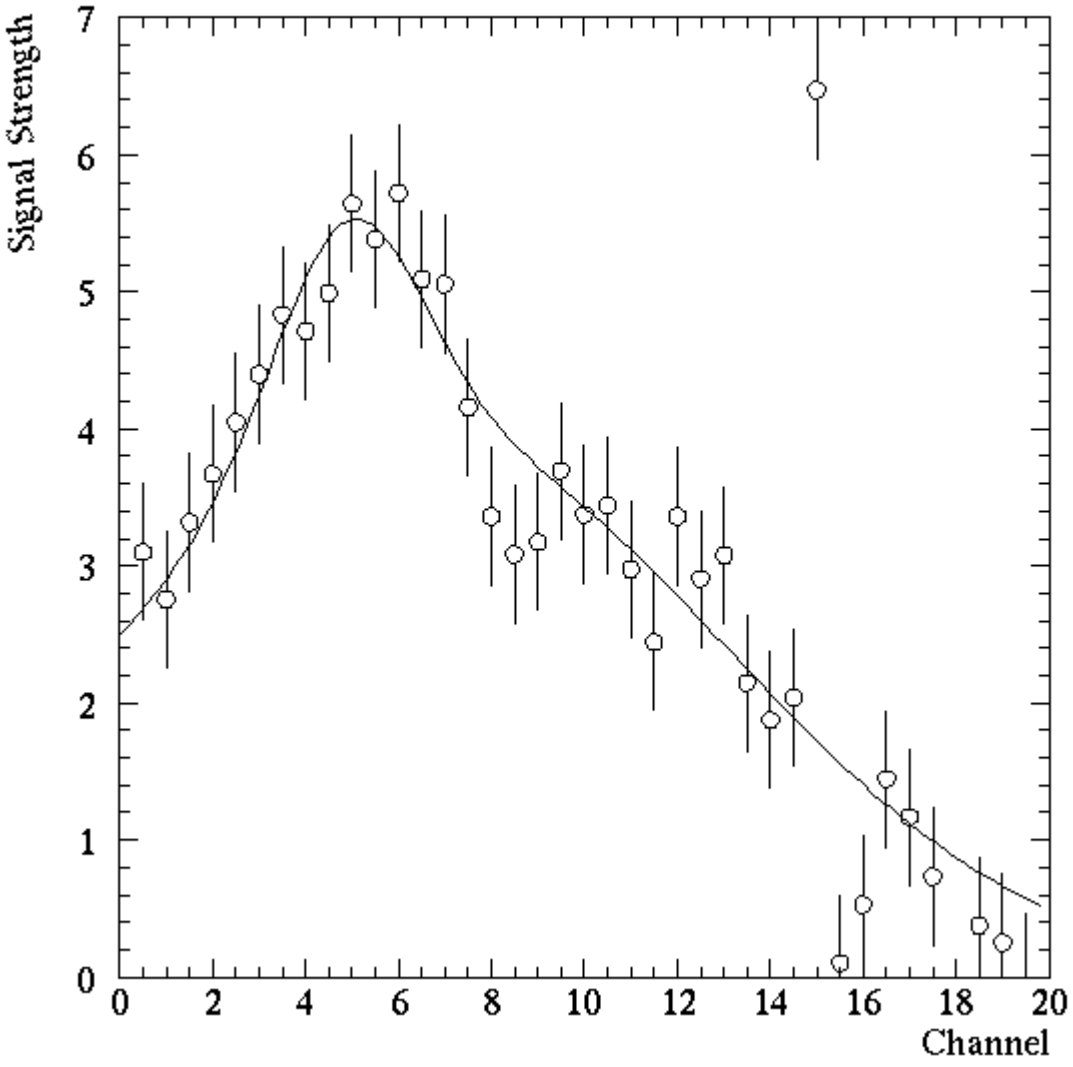


What about your model is wrong?

It's clear in this case that the long point at channel 15 is causing the huge χ^2

This cannot be a real feature in the underlying spectrum, given what you know about the instrument's resolution.

Only two possibilities: statistical fluctuation or instrumental error. The physics model itself must be blameless!



Kolomogorov-Smirnov test

The χ^2 goodness of fit test requires binned data, and enough statistics that the statistic follows a χ^2 distribution.

The Kolomogorov-Smirnov test is an alternate test that works for small numbers of events with no binning.

The basic method is to compare the cumulative distributions of the data to the model.

Visual representation of Kolomogorov-Smirnov test

1. Sort the data
2. Calculate the fraction of data points equal to or less than each point.
3. Calculate the cumulative distribution of the expected model (analytically or numerically).
4. Find whatever data point is farthest from the expected curve. The distance is D.
5. Calculate d=D√N. Look up in a table of KS values to figure out how uncommon such a value of d is.

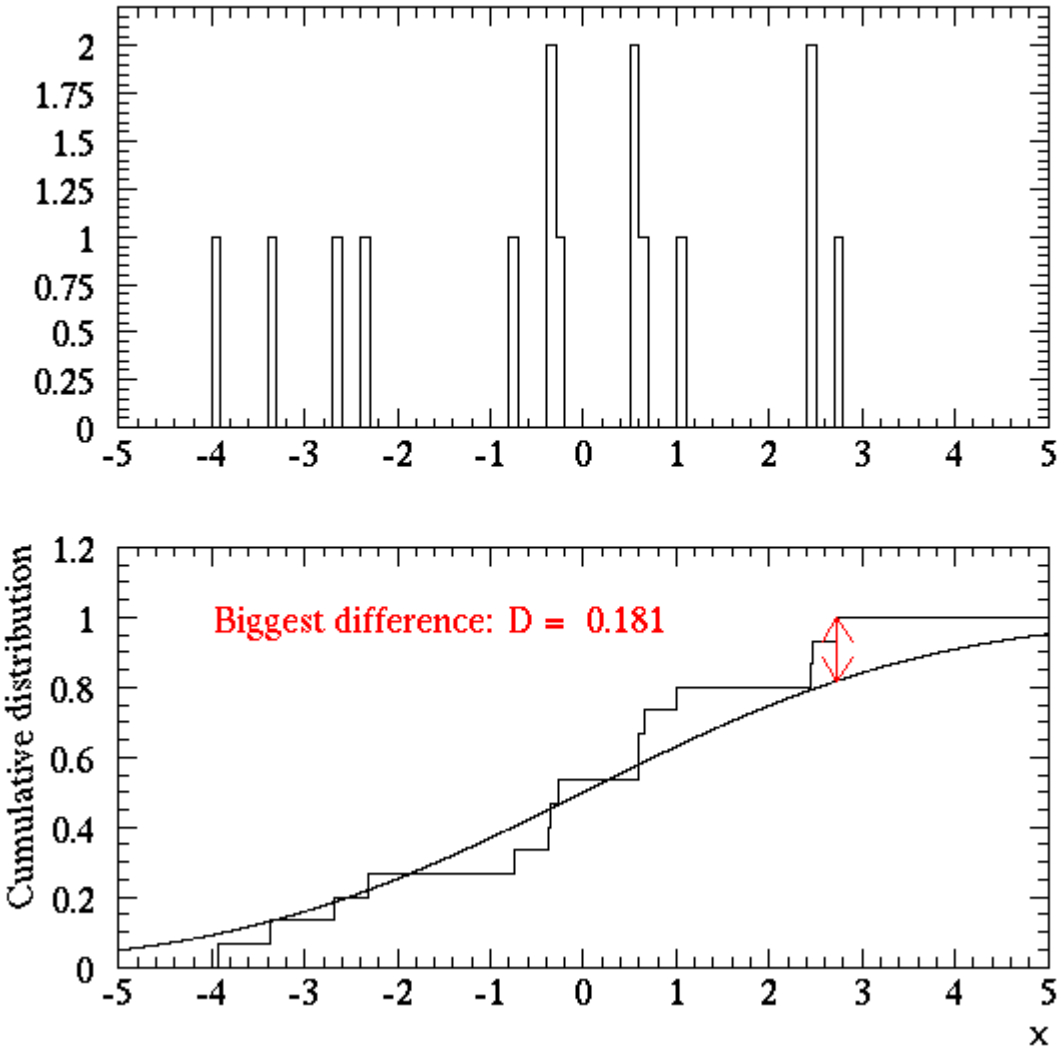
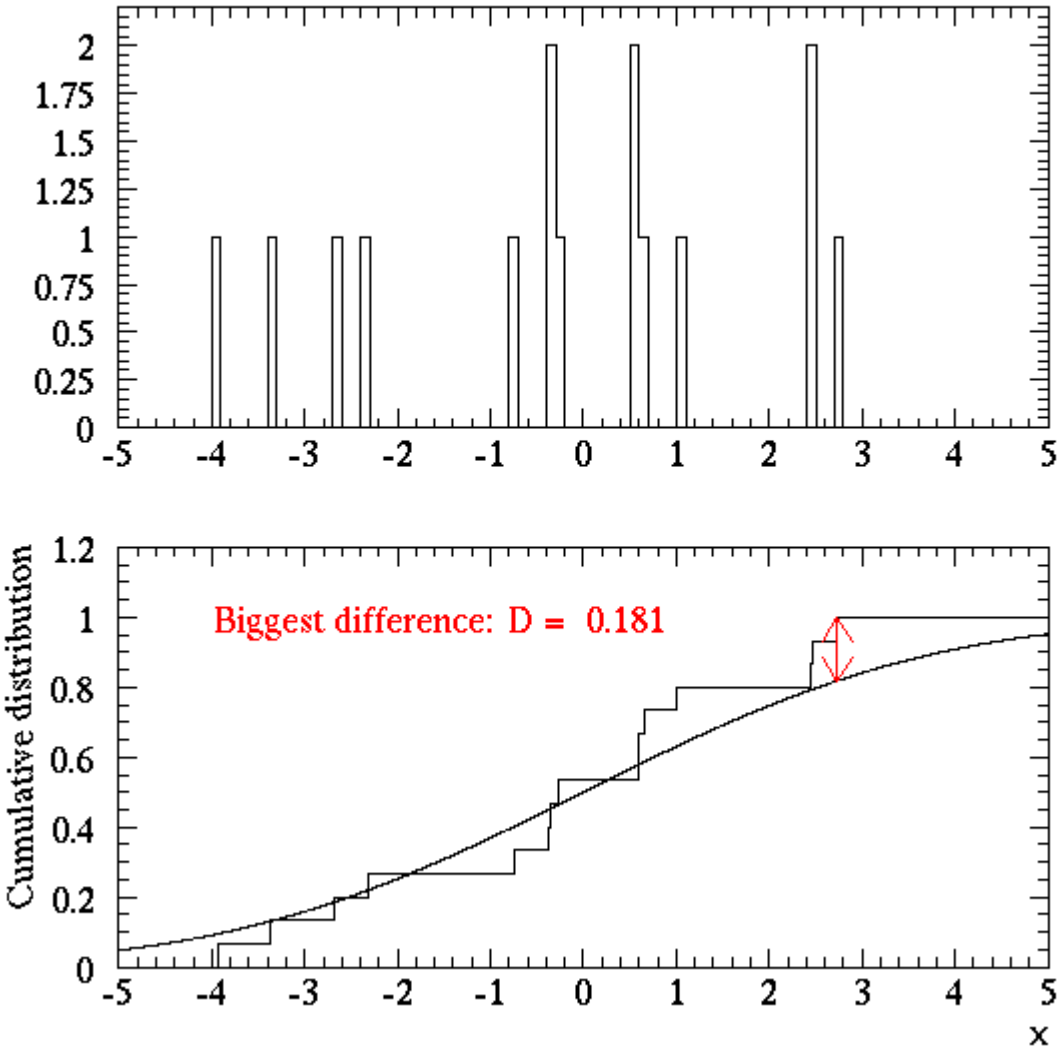


Table of KS values

d	Significance
1.63	1%
1.36	5%
1.22	10%
1.07	20%

(Note: these are approximate values. For a given number of data points, you can look up exact values from a table of significance vs. D.

For D=0.181 and N=15, I get d=0.701, which is in the meat of the distribution. The model (here a Gaussian) is a good fit!



KS test and the χ^2 test compared

KS test:

Requires no binning

Works for small statistics

Requires the distribution to be fixed beforehand: there's no way to account for the number of fitted parameters.

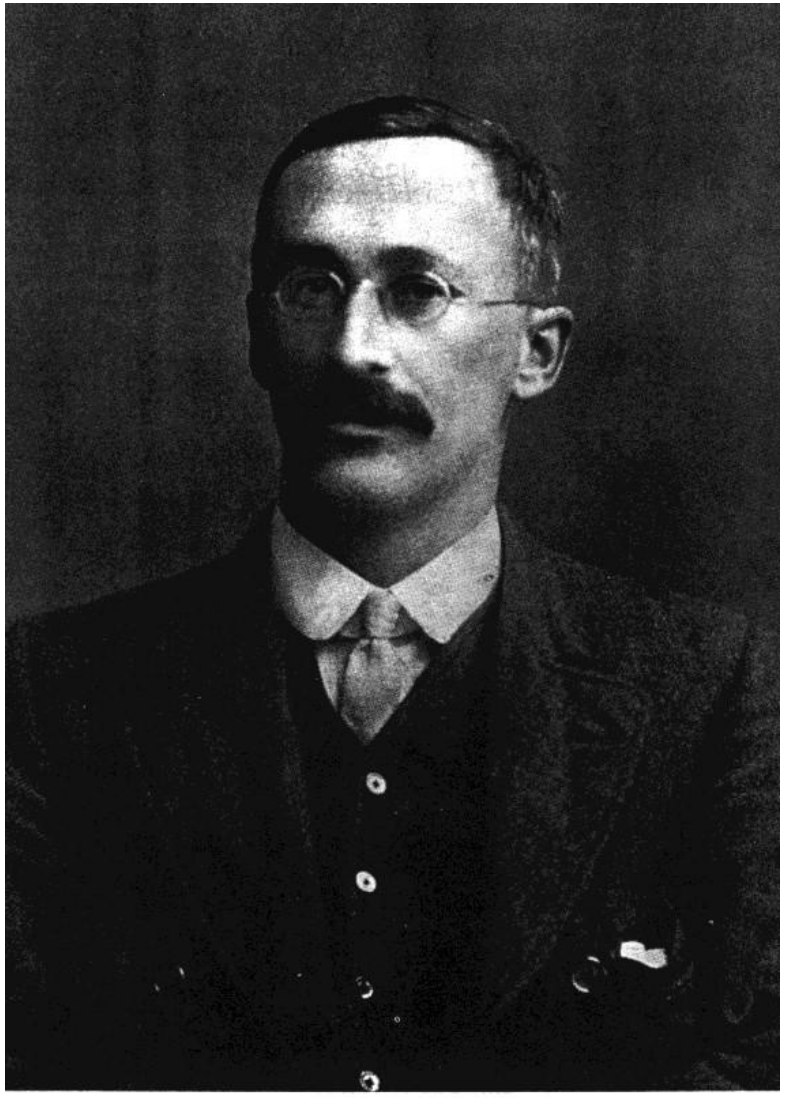
χ^2 test:

Data must be binned

Statistic has a simple distribution only when there are lots of events (Gaussian errors)

Can easily account for the fact that fitting for free parameters improves the fit by reducing the number of degrees of freedom accordingly

Cheers to William Sealy Gosset



'Student' in 1908

The “two sample” problem:known uncertainties

Suppose we have two samples of beer, and want to test the hypothesis that they have the same density.

Suppose as well that we know the uncertainties on our density measurements, and that they are Gaussian. What hypothesis test do we do?

The "two sample" problem: known uncertainties

Suppose we have two samples of beer, and want to test the hypothesis that they have the same density.

Suppose as well that we know the uncertainties on our density measurements, and that they are Gaussian. What hypothesis test do we do?

Simple ... test on the difference between the measured densities. This will have a Gaussian distribution with width:

$$\sigma_{X-Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

Example: X=1.04±0.05, Y=1.12±0.03, so X-Y=-0.08±0.058
(Can easily turn this into a one-sided test if you want to test X>Y.)

What if you don't know the uncertainties?

Often you won't know the uncertainties ahead of time. Instead you have to estimate them from the scatter in the data. (Obviously you need multiple measurements of both samples to do this.)

Consider the case that the underlying populations have identical but unknown standard deviations. First estimate standard deviations of each sample:

$$\hat{\sigma}_x = s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N_x - 1}} \qquad \hat{\sigma}_y = s_y = \sqrt{\frac{\sum_i (y_i - \bar{y})^2}{N_y - 1}}$$

Then our best weighted average for the true standard deviation is formed by:

$$S^2 = \frac{(N_x - 1)s_x^2 + (N_y - 1)s_y^2}{N_x + N_y - 2}$$

Two sample problem: equal but unknown uncertainties

Form the following test statistic, by approximating σ with S:

$$\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma^2}{N_x}+\frac{\sigma^2}{N_y}}} \approx \frac{(\bar{x}-\bar{y})}{S\sqrt{\frac{1}{N_x}+\frac{1}{N_y}}} \equiv t$$

If the sample measurements really follow a Gaussian distribution, then the numerator is also Gaussian while the denominator is basically the square root of a χ^2-distributed quantity with N_x+N_y-2 degrees of freedom.

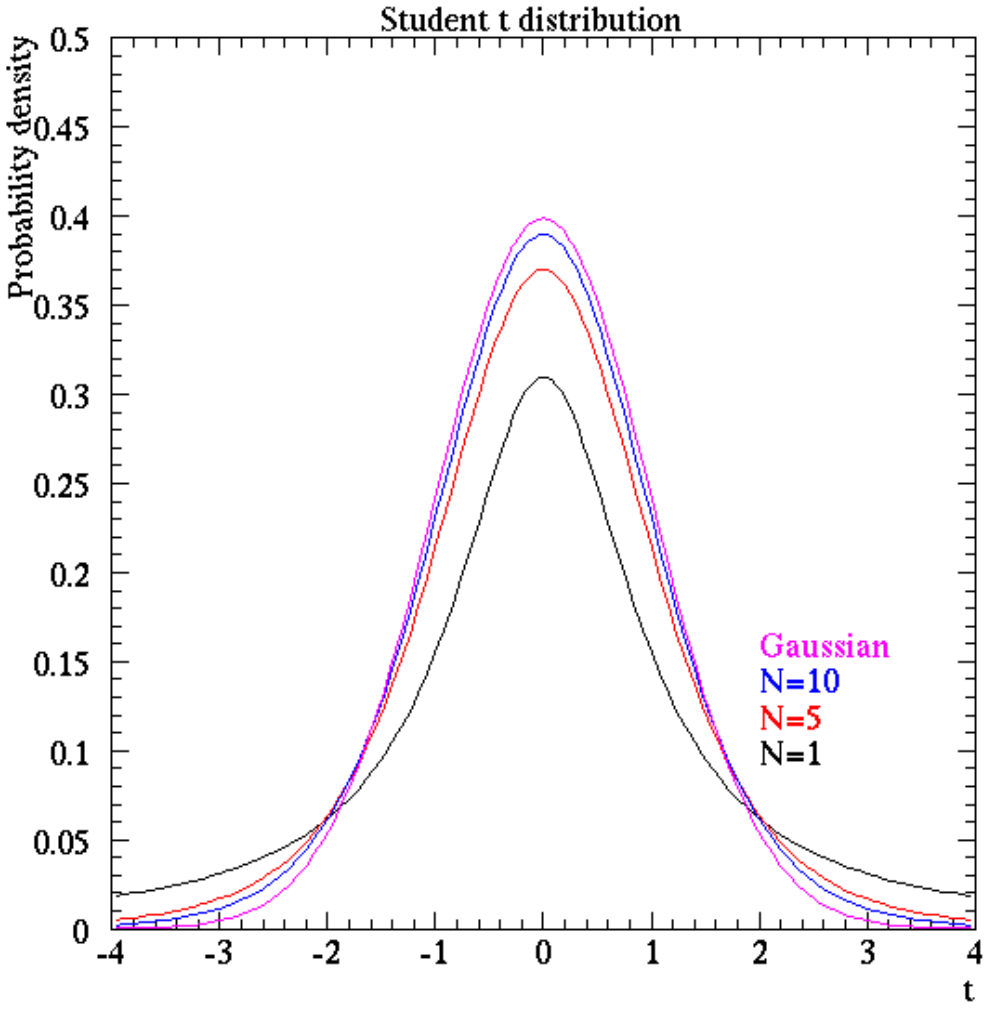
We call this statistic *t*.

Student's t-test

This peculiar test statistic has a very special distribution called Student's *t* distribution:

$$f(t|N) = \frac{\Gamma((N+1)/2)}{\sqrt{N\pi}\,\Gamma(N/2)} \left(\frac{1}{\left(1+\frac{t^2}{N}\right)} \right)^{\frac{(N+1)}{2}}$$

It's like a Gaussian with wider tails. As N gets big, it approaches a Gaussian.



What test should you use?

1) Always decide which tests you're going to use in advance of looking at the data. ALWAYS.

2) Use the most restrictive test you can get away with---in other words, limit the set of alternate hypotheses under consideration to the minimum reasonable list in order to increase the sensitivity of your test.

3) Avoid trials factors like the plague. *Whatever you do, don't keep trying different tests until you get an answer you like.*

4) Specify the significance level to which you'll work in advance of looking at the data.

5) Consider a Bayesian analysis, especially if you need to incorporate prior information.