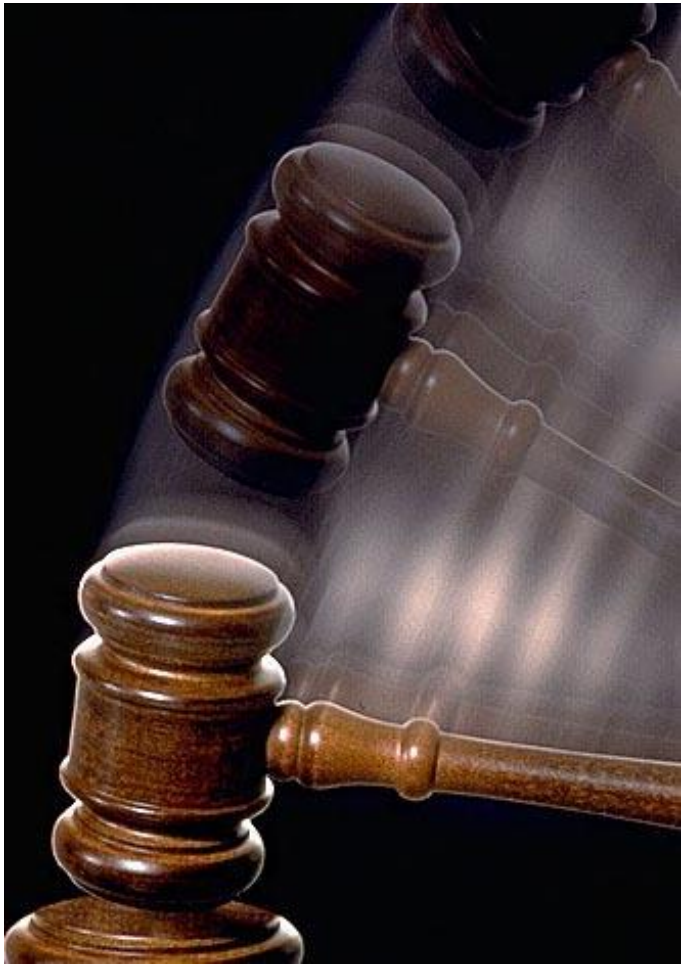
Physics 509: Intro to Hypothesis Testing

Scott Oser
Lecture #13



What is a hypothesis test?

Most of what we have been doing until now has been parameter estimation---within the context of a specific model, what parameter values are most consistent with our data?

Hypothesis testing addresses the model itself: is your model for the data even correct?

Some terminology:

- simple hypothesis: a hypothesis with no free parameters
  Example: the funny smell in my coffee is cyanide
- compound hypothesis: a hypothesis with one or more free parameters
  Example: "the mass of the star is less than $1M_{\odot}$
  "there exists a mass peak indicative of a new particle"
  (without specifying the mass)
- test statistic: some function of the measured data that provides a means of discriminating between alternate hypotheses

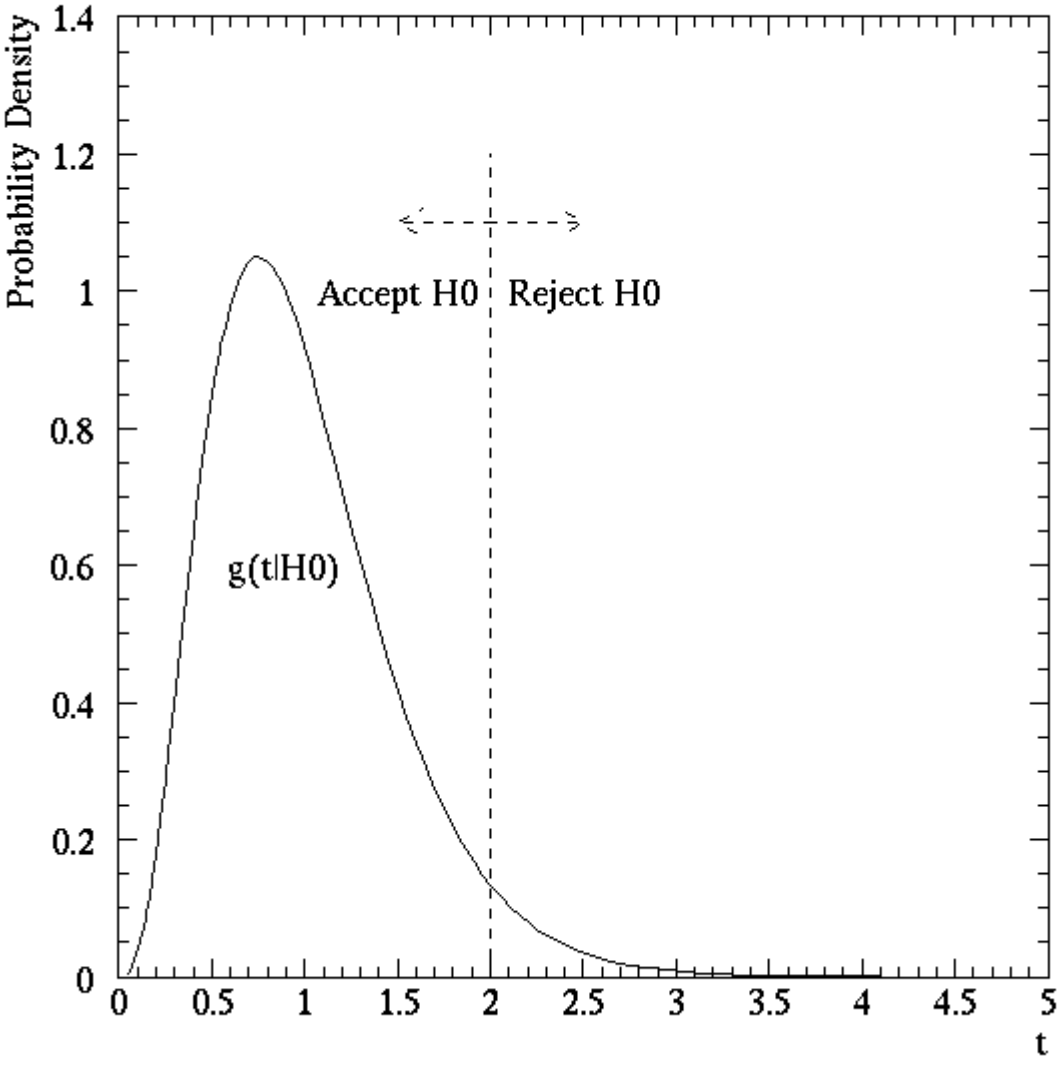
Bayesian hypothesis testing

As usual, Bayesian analysis is far more straightforward than the frequentist version: in Bayesian language, all problems are hypothesis tests! Even parameter estimation amounts to assigning a degree of credibility to the proposition "μ is between 5 and 5.01".

$$P(H|D,I)=\frac{P(H|I)\,P(D|H,I)}{P(D|I)}$$

- Bayesian hypothesis testing requires you to explicitly specify the alternative hypotheses. This comes about when calculating $P(D|I)=P(D|H_1,I)+P(D|H_2,I)+P(D|H_3,I)$ ...
- Hypothesis testing is more sensitive to priors than parameter estimation. For example, hypothesis testing may involve Occam factors, whose values depend on the range and choice of prior. (Occam factors do not arise in parameter estimation.) For parameter estimation you can sometimes get away with improper (unnormalizable) priors, but not for hypothesis testing.

Classical frequentist testing: Type I errors



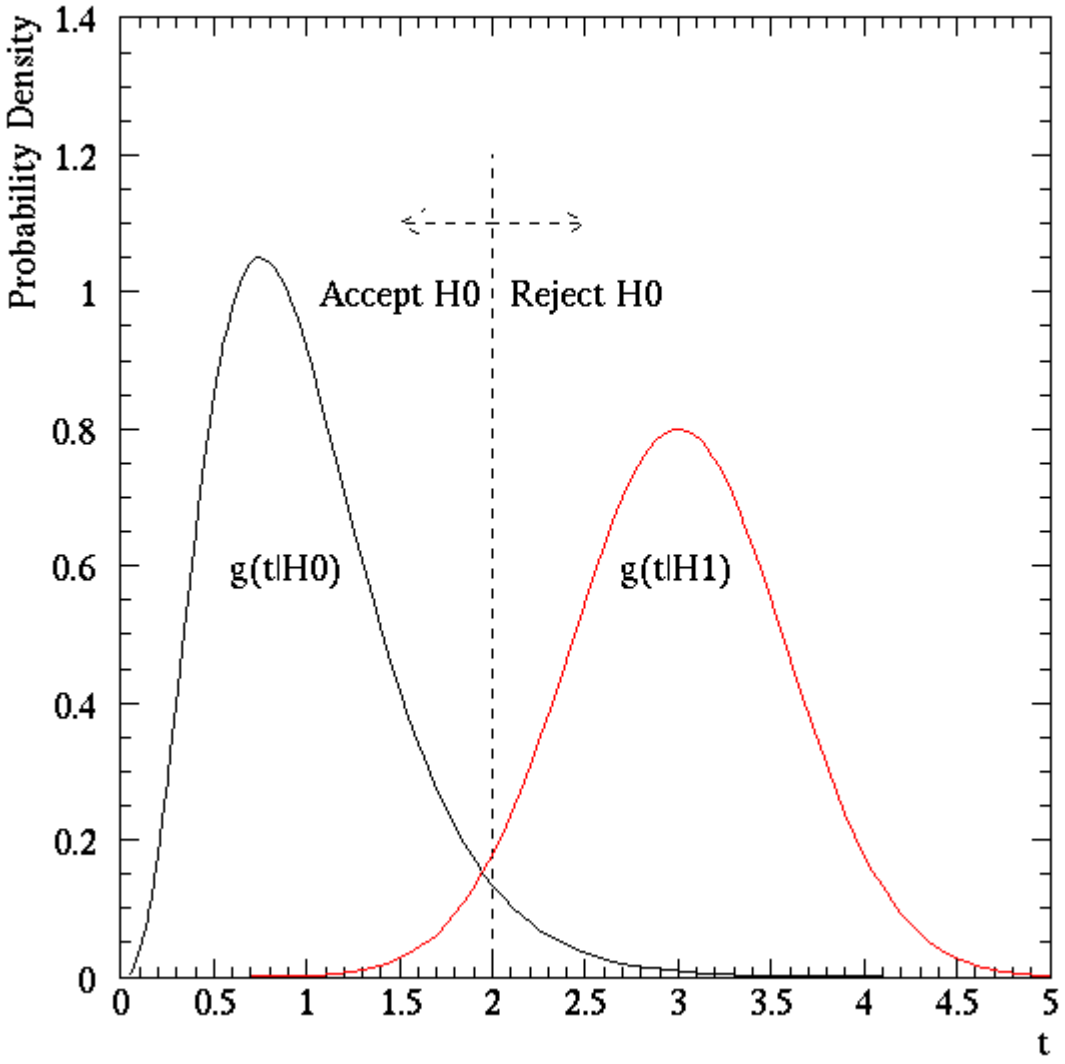
Many other cuts possible---two-sided, non-contiguous, etc.

In frequentist hypothesis testing, we construct a test statistic from the measured data, and use the value of that statistic to decide whether to accept or reject the hypothesis. The test statistic is a lower dimensional summary of the data that still maintains discriminatory power.

We choose some cut value on the test statistics t.

Type I error: We reject the hypothesis H0 even though the hypothesis is true.
Probability = area on tail = α

Classical frequentist testing: Type II error



Type II error: We accept the hypothesis H0 even though it is false, and instead H1 is really true.

Probability = area on tail of g(t|H1)= β

$$\beta = \int_{-\infty}^{t_{cut}} dt\, g(t|H1)$$

Often you choose what probability you're willing to accept for Type I errors (falsely rejecting your hypothesis), and then choose your cut region to minimize β.

You have to specify the alternate hypothesis if you want to determine β.

Significance vs. power

α (the probability of a type I error) gives the **significance** of a test. We say we see a significant effect when the probability is small that the default hypothesis (usually called the "null hypothesis") would produce the observed value of the test statistic. You should always specify the significance at which we intend to test the hypothesis before taking data.

$1-\beta$ (one minus the probability of a type II error) is called the **power** of the test. A very powerful test has a small chance of wrongly accepting the default hypothesis. If you think of H0, the default hypothesis, as the status quo and H1 as a potential new discovery, then a good test will have high significance (low chance of incorrectly claiming a new discovery) and high power (small chance of missing an important discovery).

There is usually a trade-off between significance and power.

The balance of significance and power

CM colleague: “I saw something in the newspaper about the g-2 experiment being inconsistent with the Standard Model. Should I take this seriously?”
Me: “I wouldn't get too excited yet. First of all, it's only a 3σ effect ...”
CM colleague: “Three sigma?!? In my field, that's considered *proven beyond any doubt*!”

What do you consider a significant result? And how much should care about power? Where's the correct balance?

Cost/benefits analysis

Medicine: As long as our medical treatment doesn't do any harm, we'd rather approve some useless treatments than to miss a potential cure. So we choose to work at the 95% C.L. when evaluating treatments.
<Corollary: 5% of whatever your doctor prescribes doesn't work.>

Condensed matter experimentalist: I'll write 100 papers over my career. Having a result proven wrong looks bad, and shouldn't happen more than once in my lifetime. Since my experiments don't cost a lot of time or money to reproduce, I'll work at the 99% C.L.

Particle physicist: If I claim discovery of a new particle, my field is going to propose spending $10 billion and 15 years to build a new accelerator. There really isn't any easy way for another group to check my result. The standards of my field therefore require me to work at the "5σ" C.L.

Legal system analogy

Cost/benefits: it's better to acquit guilty people than to put innocent people in jail. “Innocent until proven guilty”. (Of course many legal systems around the world work on opposite polarity!)

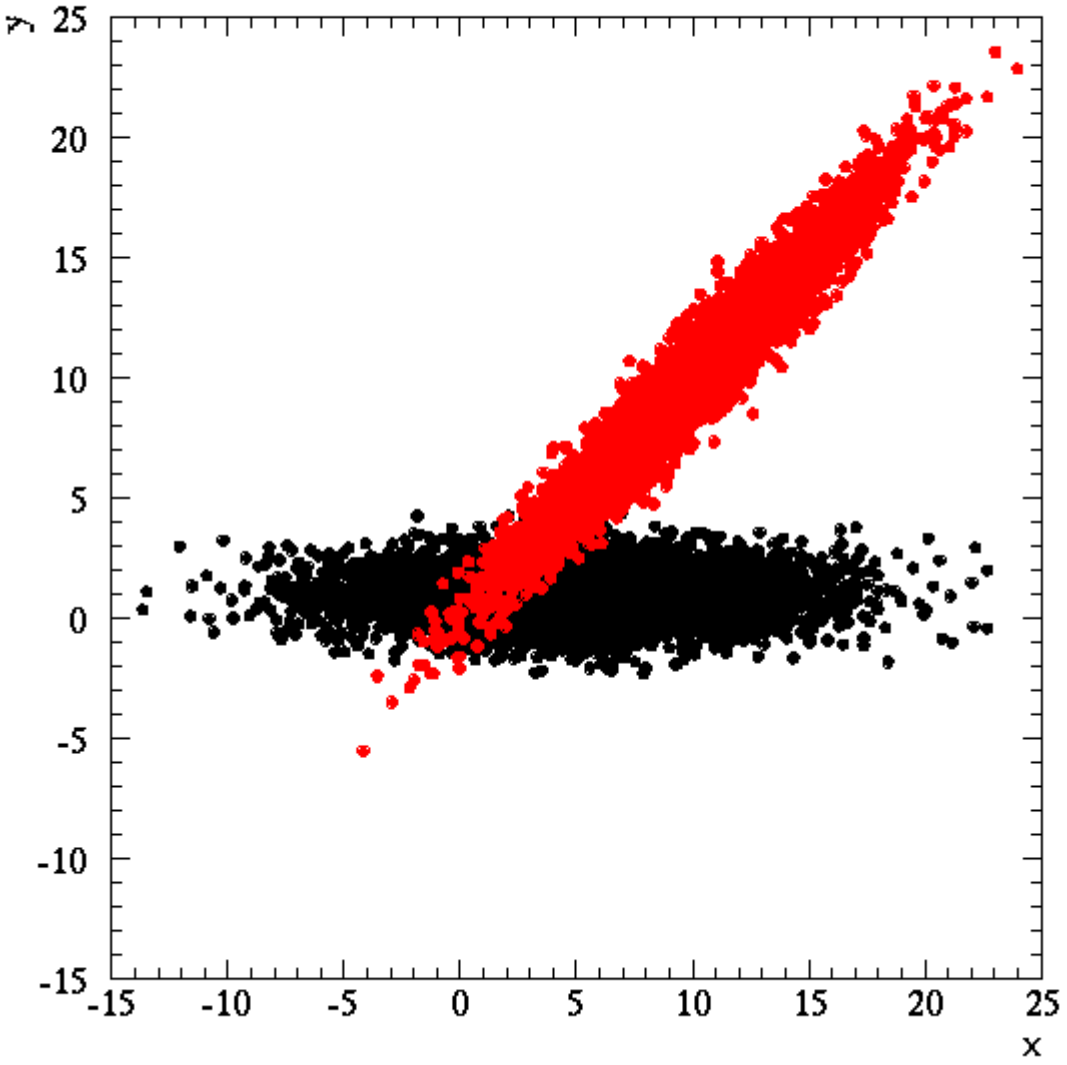
Type I error: we reject the default hypothesis that the defendant is guilty and send her to jail, even though in reality she didn't do it.

Type II error: we let the defendant off the hook, even though she really is a crook.

US/Canadian systems are supposedly set up to minimize Type I errors, but more criminals go free.

In Japan, the conviction rate in criminal trials is 99.8%, so Type II errors are very rare.

Care to draw a cut boundary?



Consider this scatter plot. There are two classes of events, and you have two test statistics x and y that you measure for each.

How would you draw a cut boundary to optimally distinguish between the two kinds of events?

The Neyman-Pearson lemma to the rescue

There is a powerful lemma that can answer this question for you:

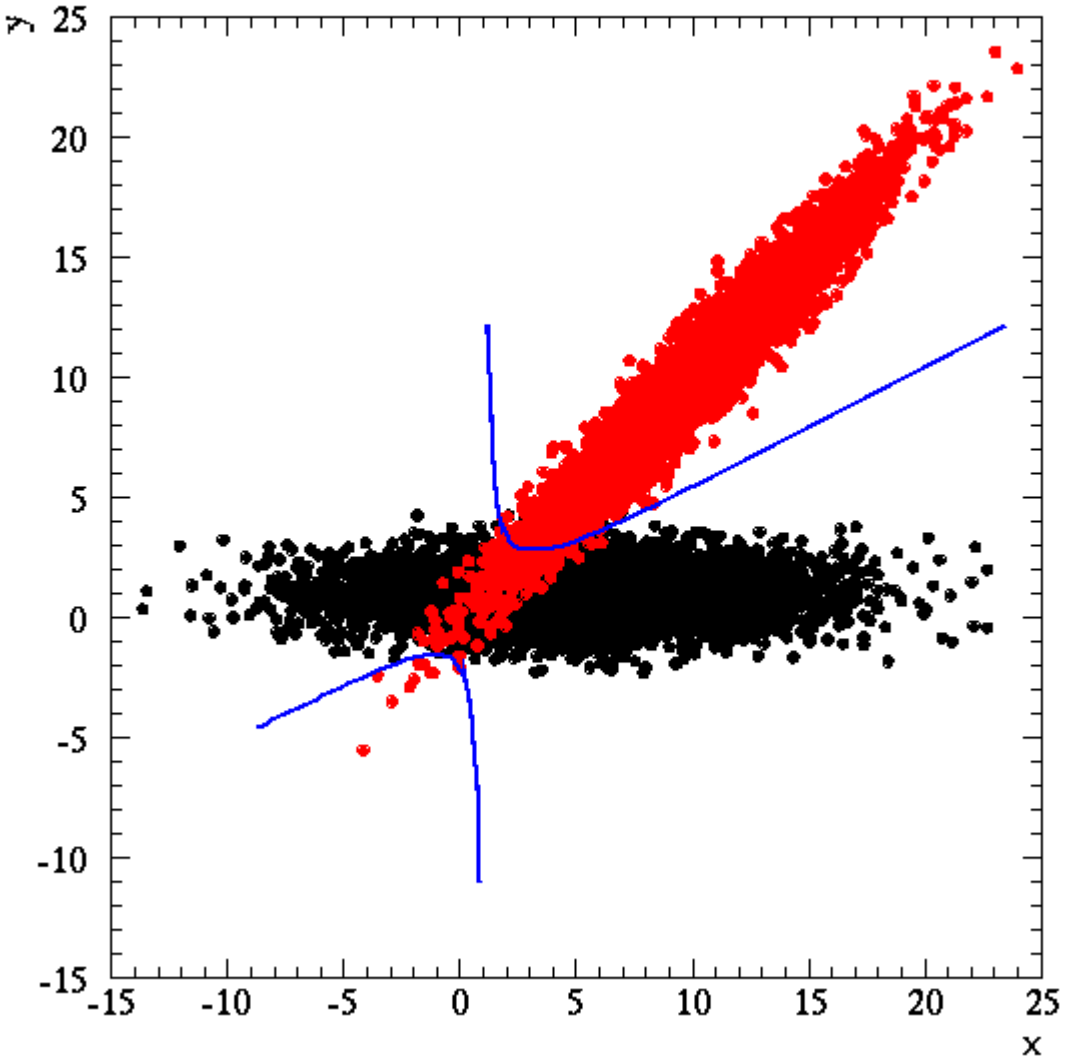
"The acceptance region giving the highest power (and hence the highest signal purity) for a given significance level α (or selection efficiency 1-α) is a region of the test statistic space **t** such that:

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

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.

A Neyman-Pearson cut boundary



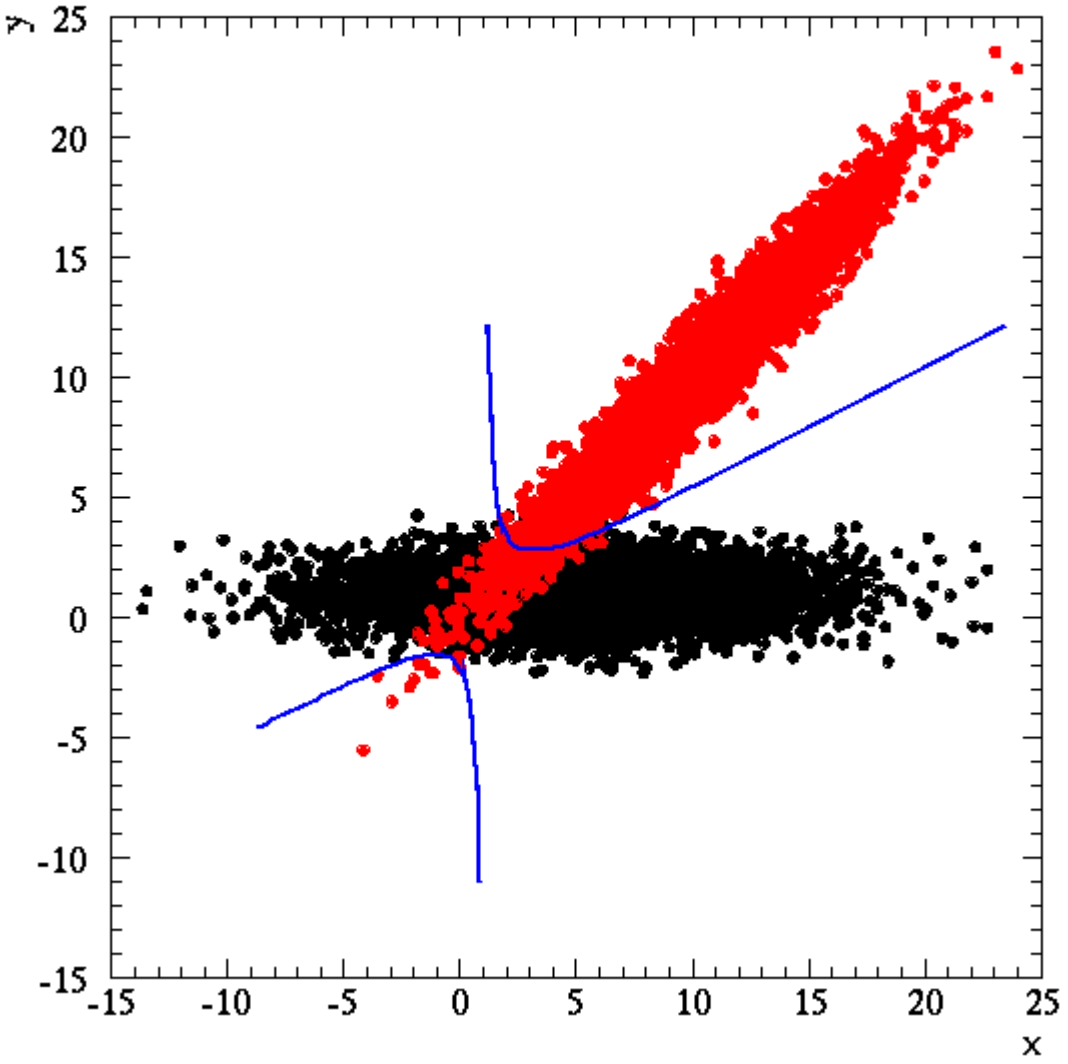
Using the shape of the two probability distributions:

$$g(x,y|\text{black}) \propto \exp\left[\frac{-(x-5)^2}{2\cdot 5^2} - \frac{(y-1)^2}{2\cdot 1^2}\right]$$

$$g(x,y|\text{red}) \propto \exp\left[\frac{-(x-10)^2}{2\cdot 4^2} - \frac{(y-x)^2}{2\cdot 1^2}\right]$$

I form a ratio, and draw a cut contour at a particular value of that ratio. In this case it's a cool-looking hyperbola.

Don't be so quick to place a cut



Even though there's an optimal cut, be careful ... a cut may not be the right way to approach the analysis. For example, if you want to estimate the total rate of red events, you could count the number of red events that survive the cuts, and then correct for acceptance, but that throws away information.

A better approach would be to do an extended ML fit for the number of events using the known probability distributions!

Interpretation of hypothesis tests

“Comparison of SNO's CC flux with Super-Kamiokande's measurement of the ES flux yields a 3.3σ excess, providing evidence at the 99.96% C.L. that there is a non-electron flavor active neutrino component in the solar flux.”

What do you think of this wording, which is only slightly adapted from the SNO collaboration's first publication?

Interpretation of hypothesis tests

“Comparison of SNO's CC flux with Super-Kamiokande's measurement of the ES flux yields a 3.3σ excess, providing evidence at the 99.96% C.L. that there is a non-electron flavor active neutrino component in the solar flux.”

In revision 2 of the paper, this was changed to:

“The probability that a downward fluctuation of the Super-Kamiokande result would produce a SNO result >3.3σ is 0.04%.”

Can you explain to me why it was changed?

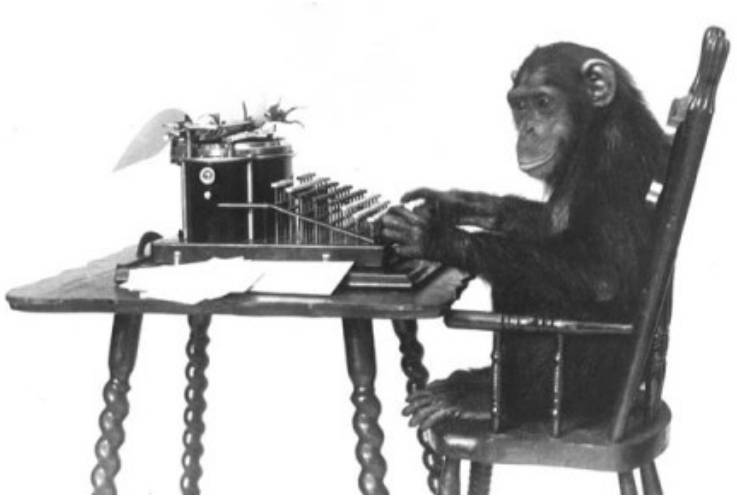
What to do when you get a significant effect?

Suppose your colleague comes to you and says “I found this interesting 4σ effect in our data!” You check the data and see the same thing. Should you call a press conference?

What to do when you get a significant effect?

Suppose your colleague comes to you and says “I found this interesting 4σ effect in our data!” You check the data and see the same thing. Should you call a press conference?

This depends not only on what your colleague has been up to, but also on how the data has been handled!



A trillion monkeys typing on a trillion typewriters will, sooner or later, reproduce the works of William Shakespeare.

Don't be a monkey.

Trials factors

Did your colleague look at just one data distribution, or did he look at 1000?

Was he the only person analyzing the data, or have lots of people been mining the same data?

How many tunable parameters were twiddled (choice of which data sets to use, which cuts to apply, which data to throw out) before he got a significant result?

The underlying issue is called the “trials penalty”. If you keep looking for anomalies, sooner or later you're guaranteed to find them, even by accident.

Failure to account for trials penalties is one of the most common causes of bad but statistically significant results.

Why trials factors are hard

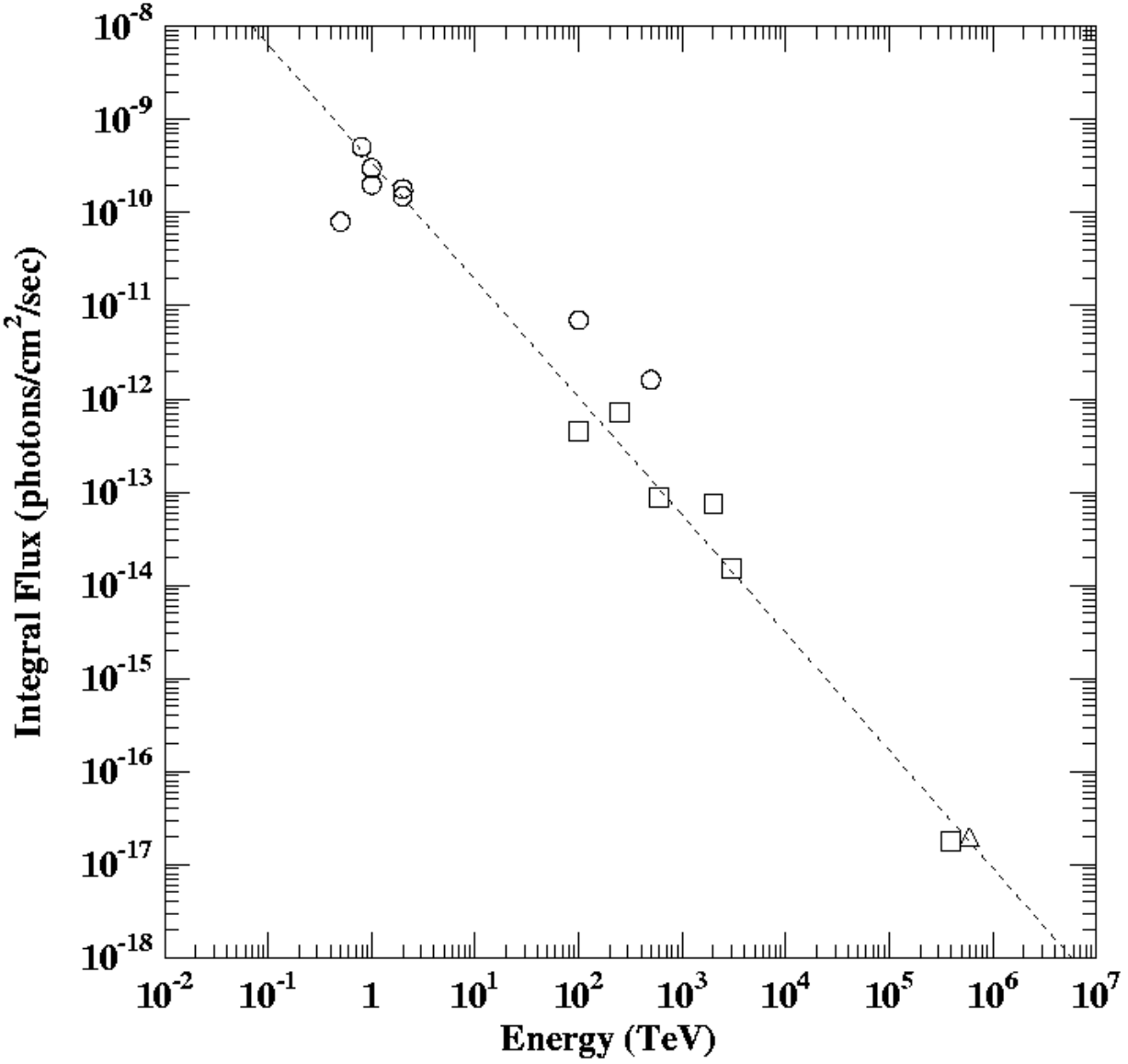
It can be really difficult to account for trials factors. For one thing, do you even know how many trials were involved?

Example: 200 medical researchers test 200 drugs. One researcher finds a statistically significant effect at the 99.9% C.L., and publishes. The other 199 find nothing, and publish nothing. You never hear of the existence of these other studies.

Chance of one drug giving a false signal: 0.1%.
Chance that at least one of 199 drugs will give a significant result at this level: 18%

Failing to publish null results is not only stupid (publish or perish, people!), but downright unethical. (Next time your advisor tells you that your analysis isn't worth publishing, argue back!)

An aside: gamma-ray astronomy history

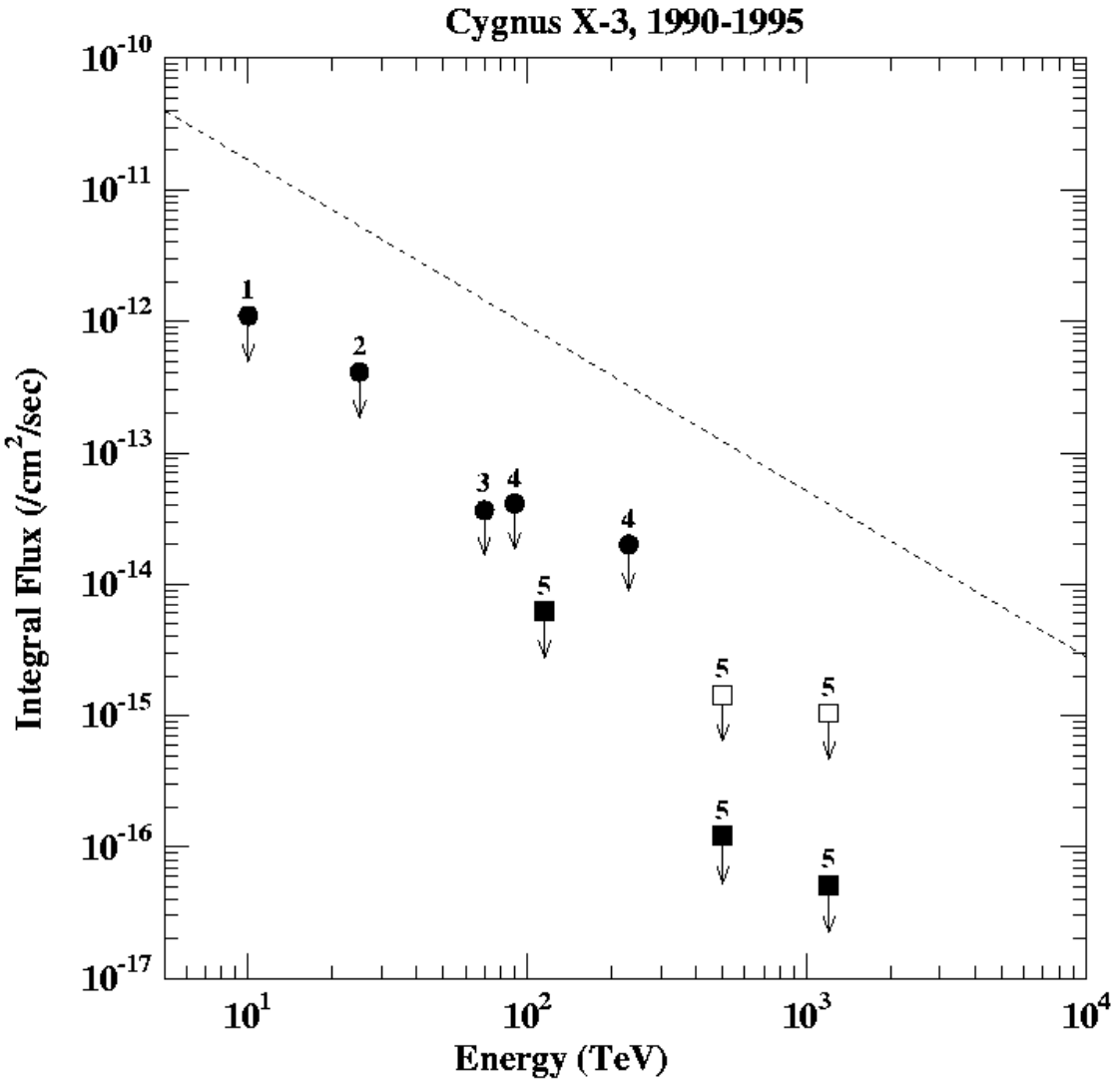


In the 1980's, many experiments operating at very different energy ranges detected high-energy gamma-rays from Cygnus X-3. Typical statistical significance was 3-4σ, and signals were hard to pull out---lot of data massaging.

But multiple independent measurements all claimed something, and the collective data was nicely fit by a consistent power law!

So much better detectors were built.

Gamma-ray astronomy: the next generation



New detectors were orders of magnitude more sensitive. Yet they saw nothing!

It's possible, but highly conspiratorial, to imagine that Cygnus X-3 “turned off” just as the new experiments came on line.

A likelier interpretation of the earlier results is that they were a combination of statistical fluctuations and trial factors--- maybe people were so convinced that Cygnus was there that they kept manipulating their data until they “found something”.

Since sensitivity of experiments also follows a power law, this explains seemingly convincing energy spectrum.

Moral

Science is littered with many examples of statistically significant, but wrong, results. Some advice:

- Be wary of data of marginal significance. Multiple measurements at 3σ are not worth a single measurement at 6σ.
- Distrust analyses that aren't blind.
- Consider trials factors carefully, and quiz others about their own trials factors.
- Remember the following aphorism: “You get a 3σ result about half the time.”

My favourite PhD thesis

SEARCHES FOR NEW PHYSICS IN DIPHOTON EVENTS IN p-pbar COLLISIONS AT √s=1.8 TEV
David A. Toback
University of Chicago, 1997

"We have searched a sample of 85 pb^{-1} of p-pbar collisions for events with two central photons and anomalous production of missing transverse energy, jets, charged leptons (*e*, μ, and τ), *b*-quarks and photons. We find good agreement with Standard Model expectations, with the possible exception of one event that sits on the tail of the missing E_T distribution as well has having a high-E_T central electron and a high-E_T electromagnetic cluster."

The infamous event

The event in question was: $ee\gamma\gamma$ + missing transverse energy

The expected number of such events in the data set from Standard Model processes is 1×10^{-6}

Supersymmetry could produce such events through the decay of heavy supersymmetric particles decaying into electrons, photons, and undetected neutral supersymmetric particles.

This is a one in a million event!

Dave's conclusion: “The candidate event is tantalizing. Perhaps it is a hint of physics beyond the Standard Model. Then again it may just be one of the rare Standard Model events that could show up in 10^{12} interactions. Only more data will tell.”

It was never seen again.

The optional stopping problem

An example from Gregory, Section 7.4:

Theory prediction: The fraction of nearby stars that are of the same spectral class as the Sun (G class) is f=0.1

Observation: Out of N=102 stars, 5 were G class

How unlikely is this?

One way to view this is as a binomial outcome. Gregory argues:

$$\text{P-value} = 2 \times \sum_{m=0}^{5} p(m|N,f) = 2 \times \sum_{m=0}^{5} \frac{N!}{m!(N-m)!} f^m (1-f)^{N-m} = 0.10$$

(Factor of two supposedly because it's a 2-sided test: theory could be either too high or too low.)

The optional stopping problem as a binomial

This is actually not quite correct. The correct way to calculate this is to calculate P(m|N,f), sort them from most probable to least probable, then add up the probabilities of all outcomes which are more probable than m=5.

m	Probability	Cumulative Prob.
10	0.131	0.131
9	0.127	0.259
11	0.122	0.381
8	0.110	0.490
12	0.103	0.593
7	0.083	0.676
13	0.079	0.756
14	0.056	0.811
6	0.055	0.866
15	0.036	0.902
5	0.030	0.933
16	0.022	0.955
4	0.014	0.969

We expected to see 10.2 G-type stars on average, but saw only 5. How unlikely is that?

P=90.2% to get a value more likely that m=5. In other words, ~10% chance to get a result as unlikely as m=5.

This is not strong evidence against the theory.

The optional stopping problem as neg. binom.

The uppity observer comes along and says: “You got it all wrong! My observing plan was to keep observing until I saw m=5 G-type stars, then to stop. The random variable isn't m, it's N---the total number of stars observed. You should be using a negative binomial distribution to model this!”

$$\text{P-value} = 2\times \sum_{m=102}^{\infty} p(N|m,f) = 2\times \sum_{m=102}^{\infty} \binom{N-1}{m-1} f^m (1-f)^{N-m} = 0.043$$

This is actually not a good way to calculate P: it assumes that the probability of observing too many stars is equal to the probability of observing too few. In reality the negative binomial distribution is not that symmetric.

The optional stopping problem as neg. binom.

A correct calculation of a negative binomial distribution gives:

N	Probability	Cumulative Prob.
41	0.0206	0.021
40	0.0206	0.041
42	0.0205	0.062
39	0.0205	0.082
43	0.0204	0.103
38	0.0204	0.123
44	0.0203	0.143
37	0.0202	0.164
45	0.0201	0.184
...	...	...
12	0.0016	0.975
102	0.0015	0.977

The probability of getting a value of N as unlikely as N=102, or more unlikely, is 2.5%.

This is rather different than just adding up the probability of N≥102 and multiplying by 2, which was 4.3%.

In any case, the chance probability is less than 5%---data seems to rule out theory at the 95% C.L.

Optional stopping: a paradox?

Everyone agrees we saw 5 of 102 G-type stars, but we get different answers as for the probability of this outcome depending on which model for data collection (binomial or negative binomial) is assumed.

The interpretation depends not just on the data, but on the observer's intent while taking the data!

What if the observer had started out planning to observe 200 stars, but after observing 3 of 64 suddenly ran out of money, and decided to instead observe until she had seen 5 G-type stars? Which model should you use?

Paradox doesn't arise in Bayesian analysis, which gives the same answer for either the binomial or the negative binomial assumption.