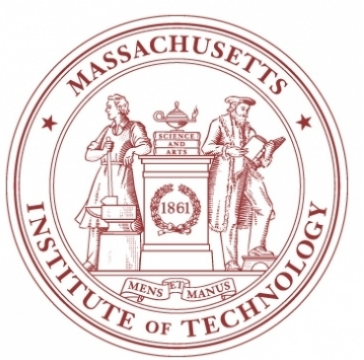
Numerical Methods & Data Analysis in HEP

Mike Williams

Department of Physics

Massachusetts Institute of Technology



IDPASC Lectures

January 29 – 31, 2013

Lecture 3

So far we have talked about using machine learning techniques to classify your data; how to use 2-sample tests to compare data samples; and about regression and multivariate GOF tests.

Today's lecture will cover the following:

- frequentists vs Bayesians;
- limits;
- non-parametric density estimation;
- non-parametric regression.

Tomorrow there will be a short practical session. It's only an hour long so it'll be fairly simple but hopefully you'll get a feel for a few of these topics.

Frequentist vs Bayesian

It's now time to confront the epic division in statistics: Frequentists *vs* Bayesians.

Frequentists

Objective probability. Always talk about fraction from an ensemble that have some property. Can only talk about things for which there can be an ensemble, but for these things total agreement (in theory).

Bayesians

Subjective probability. Given prior belief about something, can use observations (data) to update belief. Can talk about anything, but room for disagreement (your prior belief may not agree with mine).

Bayes' Theorem

Bayes' Theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

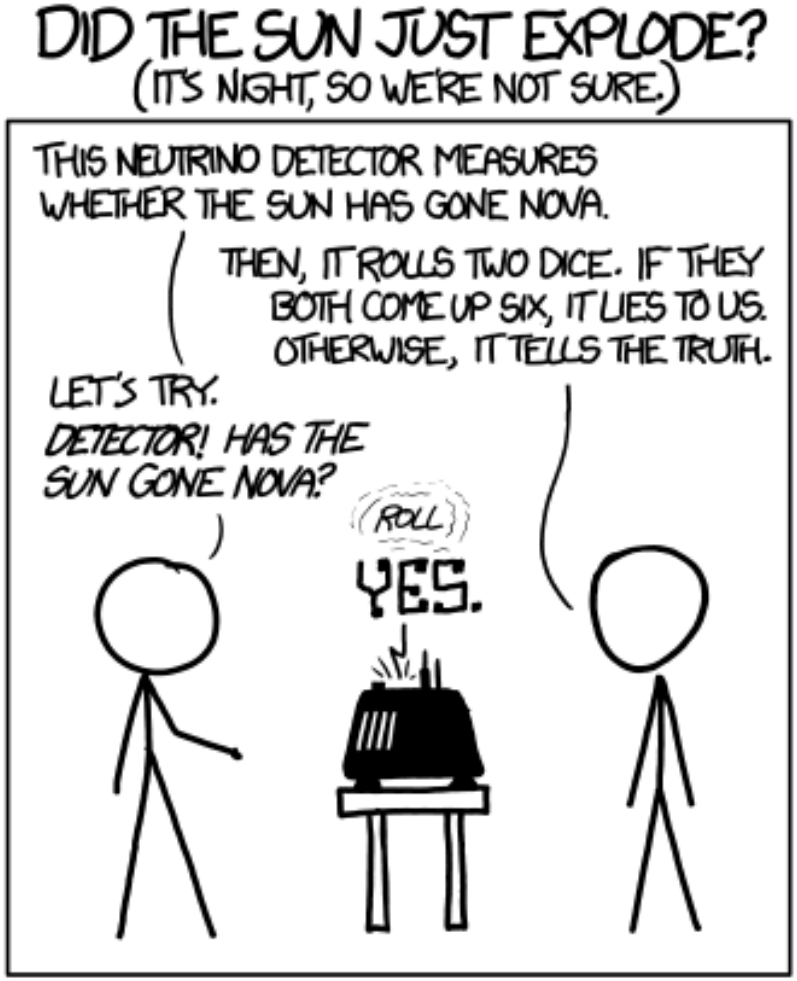
Simple example: My friend belongs to a cult that thinks the world will end on Feb. 1, 2013. I am skeptical. I think there's a 10% chance he's right. One of the predicted signs of the coming doom is that it will rain in Santiago de Compostela on Jan. 31. I check historical data and find that the probability of rain on this date is 50% (in the absence of doom).

$$P(\text{doom}|\text{rain}) = \frac{P(\text{rain}|\text{doom})P(\text{doom})}{P(\text{rain})} = \frac{1\times 0.1}{1\times 0.1 + 0.5\times 0.9} = 0.18$$

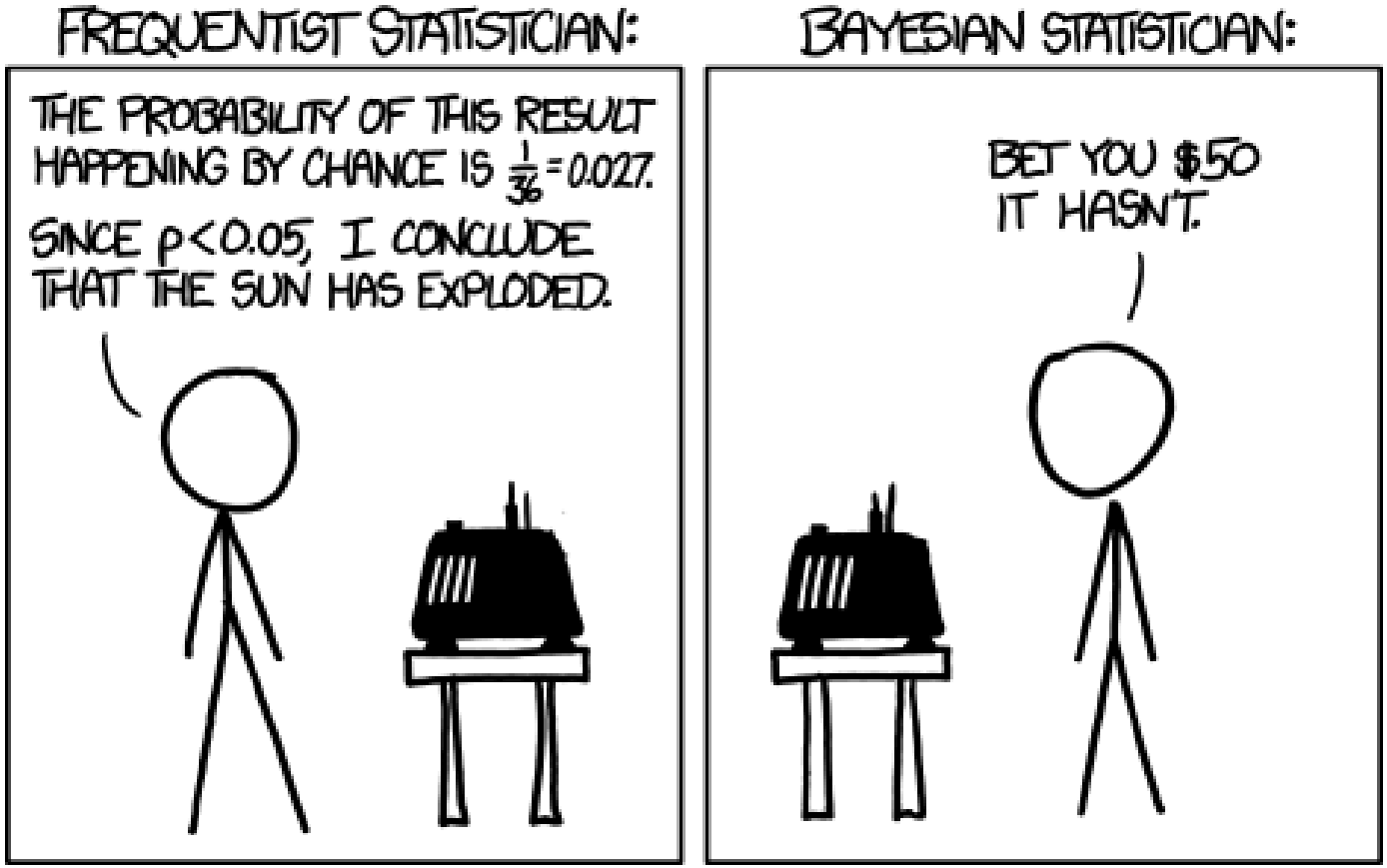
I.e., if it rains today I will think there's an 18% he's right. If it doesn't rain then 0% since his cult assigned 100% probability to it raining today.

OK, this was a stupid example, but how about "Was the universe created by a Big Bang?"

Frequentist vs Bayesian



Frequentist vs Bayesian



In HEP we require $p < 5\sigma \sim 1/1.7\text{M}$ for discovery.

Frequentist vs Bayesian

Let's change the cartoon problem slightly: Let's say instead of dice it has a random number generator that throws numbers between 1 and 1.7M. If it gets 66, it lies; otherwise, it tells the truth. We try and it says "yes".

Frequentist Physicist

$p < 5\sigma$; thus, I reject the "sun has not exploded" hypothesis (I should accept that there's a 1 in 1.7M chance I'm wrong, but I might not). A physicist might claim that the sun has exploded; a statistician wouldn't.

Bayesian Physicist

Bayes' theorem says $P(\mathrm{explode}|\mathrm{yes}) = \frac{P(\mathrm{yes}|\mathrm{explode})P(\mathrm{explode})}{P(\mathrm{yes})}$. Our prior belief for the probability of the sun exploding during the required time window is ϵ, while the probability of the machine lying is ℓ; thus, $P(\mathrm{explode}|\mathrm{yes}) = (1-\ell)\epsilon/[(1-\ell)\epsilon + \ell(1-\epsilon)] \approx \epsilon/(\epsilon+\ell)$. If $\epsilon << \ell$, then $P(\mathrm{explode}|\mathrm{yes}) \approx \epsilon/\ell$; *i.e.*, my belief that the sun has exploded will be 1.7M times bigger but still very very small.

Frequentist vs Bayesian

Bayesians can tackle any problem. *E.g.*, does God exist? All you need is prior belief in the existence of God (some probability) and then after an "event" you can update your belief (one way or the other). Of course, no reason for any two people to agree!

The advantage of frequentist statistics is that it's objective; however, the trade off is that there are many questions you can't answer. In fact, you can only really answer one question: if a given hypothesis is true, what is the probability of obtaining the observed result.

In physics, if $p < 5\sigma$, we claim "discovery"; however, nothing special actually happens at 5σ. In frequentist statistics, all this means is that the null hypothesis is rejected with a confidence level $> 1 - 6 \times 10^{-7}$. Had the significance been 4.9σ, then the probability would've been 10^{-6}.

Frequentist vs Bayesian

Frequentism is objective, but the "claims" made by physicists based on its results are subjective. Most physicists claim to be frequentists; however, frequentists can't say whether we've seen the Higgs boson; they can only reject the null hypothesis at a certain confidence level.

How many believe that we've discovered the Higgs boson? Ask one and you'll probably get a "yes". You may get a "we have to measure the couplings first". The fact that they are even willing to give an answer other than "such questions have no meaning" implies that, at heart, they are Bayesians (they've defined $P(\mathrm{Higgs})$).

For better or worse, humans use Bayesian inference . . . and physicists are human.

Frequentist vs Bayesian

You truly avoid Bayesism completely if you take this view:

Scientific theories say nothing about truth. Scientific theories are simply useful tools for assigning probabilities to future events.

The correct answer for you to the Higgs question is: The existence or not of the Higgs boson has no meaning. The SM $-$ Higgs is rejected at the $1-\epsilon$ CL. The SM $+$ Higgs is now the simplest model which cannot be rejected at any appreciable CL. The use of Occam's razor is for pure utilitarian reasons, and since the only use for this theory is utility, it is OK to take this approach.

Would you like to change your answer to the Big Bang question now?

Frequentist vs Bayesian

So, most physicists are really (implicit) Bayesians; however, that doesn't mean that frequentist reporting isn't valid.

Here's my (current) opinion on the matter:

- If a frequentist result is possible, then it's preferred; it's objective and everybody (Bayesians included) will know exactly what it means and can freely use your result however they want.
- If a certain measurement can only be made by making some assumptions (*e.g.*, on the shape of the background), then report a Bayesian measurement. I see no reason not to make a measurement because it can only be done by inputing a bit of *prior knowledge*. Clearly state your measurement is Bayesian and what your subjective assumptions are. People are free to disagree, but provided you haven't done anything "crazy" the disagreement will be at a $\sim 20\%$ level. You can kill a lot of theories with an "order of magnitude" limit.

Limits

Setting limits is very important in physics since, unfortunately, most of the things we look for we don't find (probably they don't exist).

Poisson Limits

n observed	low: $\sum P(n,\lambda) = 0.05$	high: $\sum P(n,\lambda) = 0.95$
0	–	3.00
1	0.05	4.74
2	0.36	6.30
3	0.82	7.75
⋮	⋮	⋮

Unfortunately, it's not always (ever?) this easy . . .

WTF?

We expect 3.0 background events and observe $n = 0$; thus, at the 95% CL $\lambda = s + b < 3.0$. So, we get an upper limit on the signal at $s < 0$?

- On one had, we know that as many as 5% of our claims at 95% CL should be wrong. So, fine, this is one of them. Publish (to avoid bias) and move on.
- On the other hand, we know that $s \geq 0$. Can't we use this knowledge somehow?

Feldman-Cousins

Let's go back to our WTF example. Our background expectation is $b = 3.0$, we observe $n = 0$. What do we learn about s? Let's start in the frequentist paradigm ...

First, what sounds like a different problem. Prior to looking at your data, you need to decide if you're going to quote a CI or a limit. If you don't, you will bias your result ... but how do you know what to quote without looking at the data?

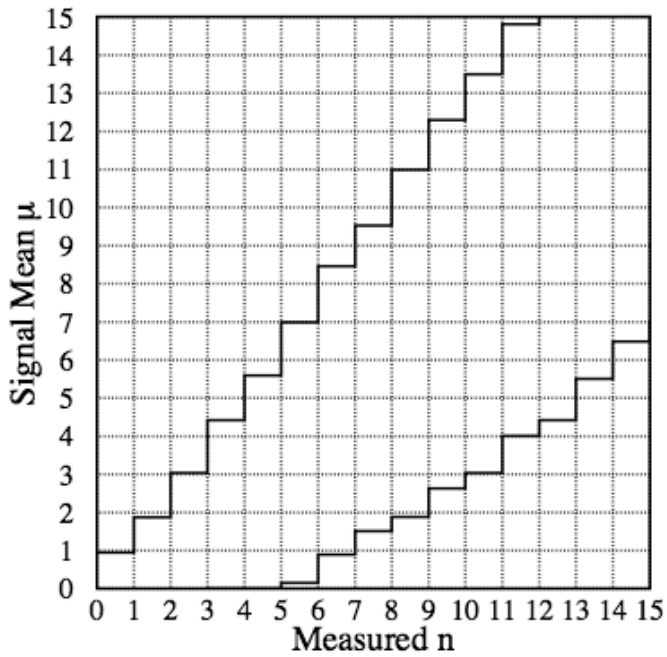
See G.Feldman and R.D.Cousins, *A unified approach to the classical statistical analysis of small signals*, PRD **57**, 3873-3889 (1998).

Feldman-Cousins

For our example, for each value of s, construct a "belt" that contains the desired probability by *ranking* on $P(n|s) = (s+b)^n \exp(-(s+b))/n!$, where $s = n - b$ or 0.

- For each s, start with the highest $P(n|s)$ and keep adding until you reach the CL.
- Do this for all s.
- For any measured n, read "vertically" (*e.g*, for $n = 0$ the interval is $[0, 1.1)$; for $n = 6$ its $(0.2, 8.5)$).
- Guarantees correct coverage; automatically returns limit or CI.
- Works for MVA problems too!
- In general, rank on likelihood.

90% CL for $b = 3.0$



Bayesian Limits

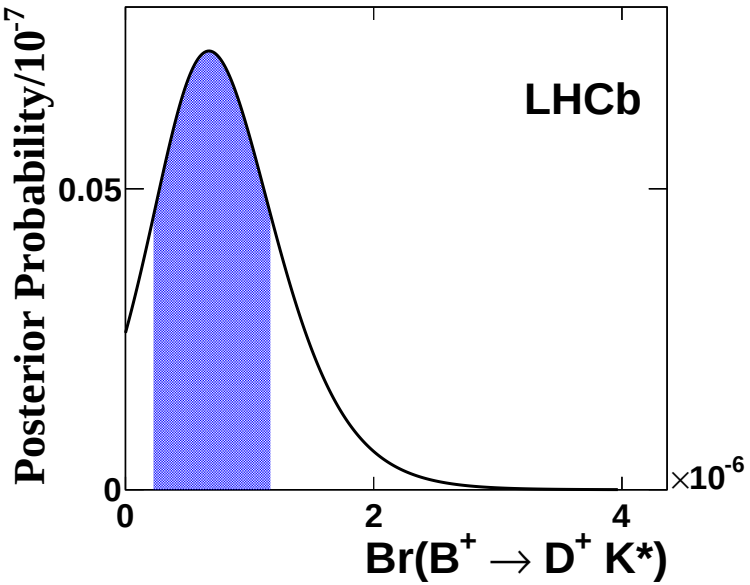
To calculate a Bayesian limit we need to first determine the *posterior* distribution for the signal

$$p(s|n) \propto \int \int P(n|s,b,\epsilon)\pi(s)\pi(b)\pi(\epsilon)dbd\epsilon,$$

where $\pi(x)$ is the prior for x and $P(n|s,b,\epsilon) = \exp(-(\epsilon s + b))(\epsilon s + b)^n/n!$ (ϵ is the efficiency/normalization).

The limit @ 90% *credibility level* is $\int_0^{\mathcal{B}_{90}} p(\mathcal{B}|n_{\rm obs})\mathrm{d}\mathcal{B} = 0.9 \int_0^{\infty} p(\mathcal{B}|n_{\rm obs})\mathrm{d}\mathcal{B}$.

Naturally get 1 or 2-sided interval; however, credibility based on priors used.



Priors, Priors, Priors

Bayesian methods seem really great but there are some *odd* features.

A common choice for the signal prior is uniform on $[0, \infty)$, but if you work with x^2 instead of x you get a different answer (something uniform in x is not uniform in x^2).

This means your "objective" *totally ignorant* of x declaration does not survive a change of variables.

Good practice is to try several priors (*e.g.*, uniform, Jefferys) and see how much your result changes. If the difference is small, great. If not, then your result is simply reflecting what you choose to believe about x prior to looking at your data (*i.e.*, it is meaningless).

Also, the more data you have the less dependence on priors you'll have.

Limits Summary

- Despite the very different philosophies, Bayesian and Frequentist methods give very similar results with enough data and even with small statistics tend to give results consistent at the 20% level (roughly).
- Other limit-setting options on the market (*e.g.*, CLs, profile likelihood) too. No single right way but plenty of wrong ones.
- If you set a Bayesian limit you **must** check that the choice of prior doesn't have an big effect ($>$ 20% or so).
- You must follow this approach: decide on a strategy; examine the data; publish your result. If you swap the first two you bias the result!
- If you have a big systematic uncertainty, more honest to use a Bayesian technique. Otherwise, use either but "standard" is to use (inspired) frequentist.

Why Non-Parametric?

Why use non-parametric methods? In many cases the model would be **very** complicated and we don't care what it is. We don't care what f is so why try and model it?

Some examples where non-parametric techniques are useful:

- The 2-sample tests from yesterday are good examples. We want to know if $f_A = f_B$; we don't care at all about f.
- Signal/background separation where we have MC that accounts for physics and detector efficiencies. We really don't care what f is just how much of the sample is signal.
- We want to *visualize* the data. Having a PDF here wouldn't help us.
- Unfolding.
- *etc.*

There are many examples. Just remember, if you don't care about f, why spend a lot of time determining f?

Non-Parametric Methods

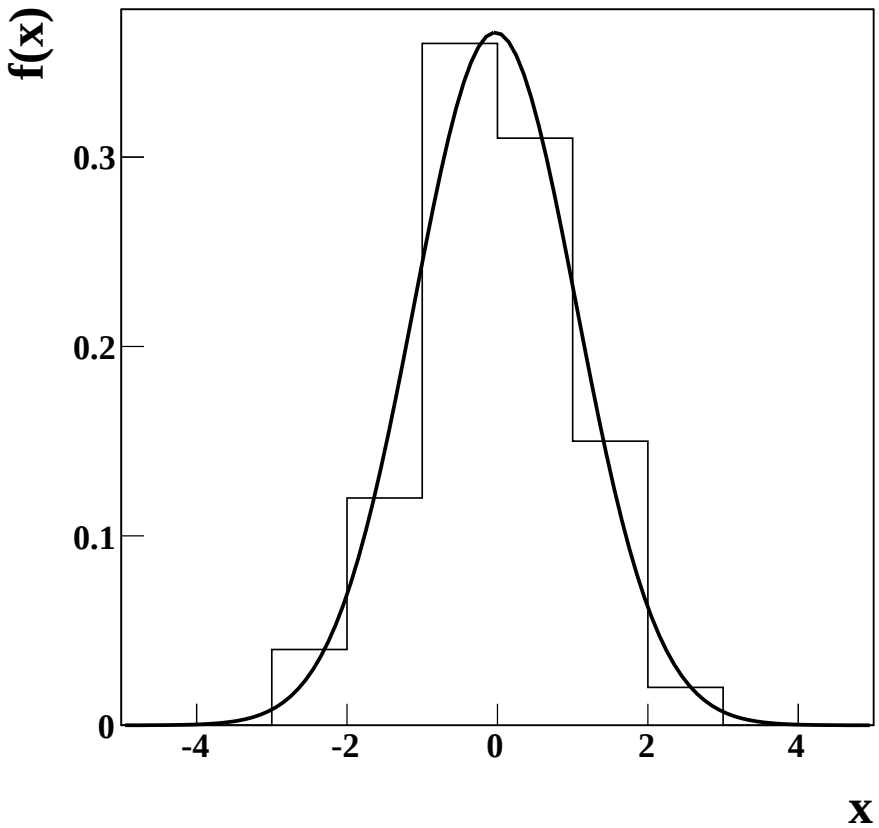
There are two main categories:

- No PDF: Some techniques, like the 2-sample tests from yesterday, do not require that we try and estimate f.
- Estimated PDF: Some times we do want to have an estimate of f, but we don't care about its functional form. In these cases we can use the data to estimate f.

Let's start with the second case . . .

Histograms

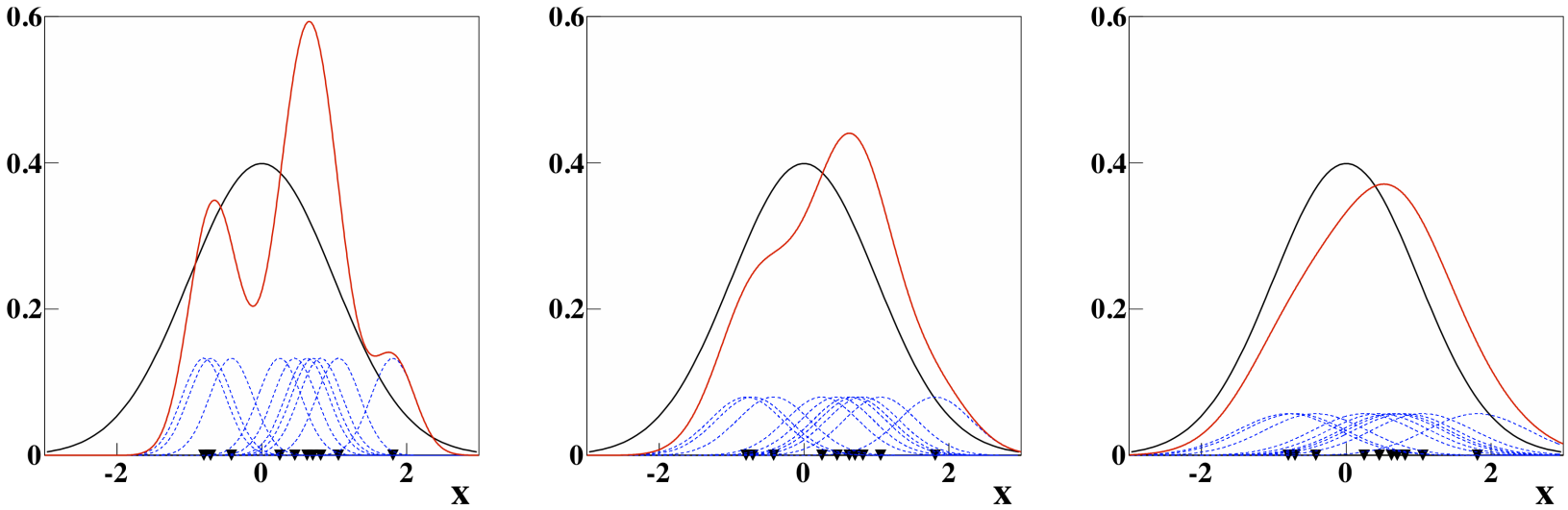
A histogram estimates the PDF by computing which bin $\vec{x}$ is in and then using the number of data observed in that bin (properly normalized).



Main criticism: information on location of each datum is ignored. Hit by curse of dimensionality very quickly.

Kernel Estimation

A better approach: build the PDF as $\sum f_i(\vec{x})/n$; *i.e.*, put a function at each datum and for any $\vec{x}$ sum up the contributions from each datum.

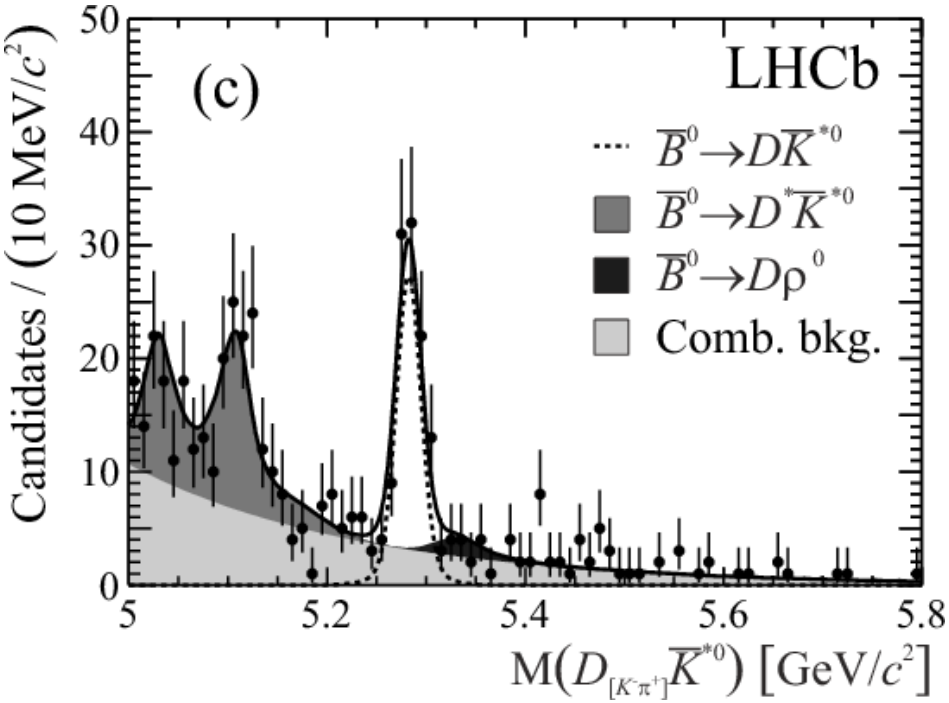


Typically put a Gaussian at each point. Quality of PDE depends on width chosen. Entire *industry* in choosing the width.

See K.Cranmer, *Kernel estimation in high-energy physics*, CPC **136**, 198-207 (2001).

Kernel PDFs in Regression

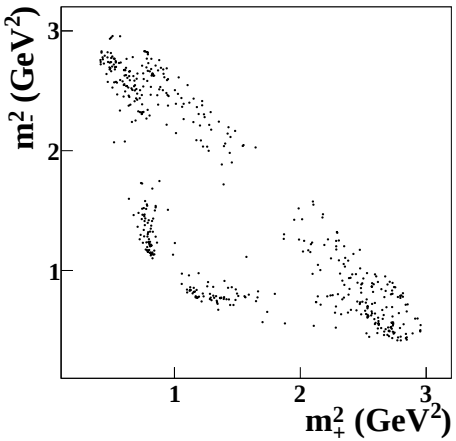
Can use kernel PDF as component of full PDF in regression; can combine with parametric PDF components.



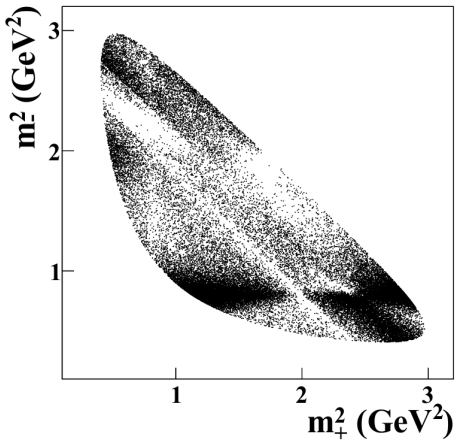
If you don't know a bkgd component's PDF, then guessing what functional form "kind of looks like the data" doesn't add any information.

Non-PDF Nonparametric Regression

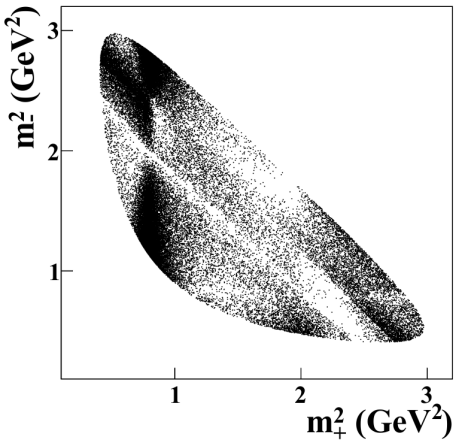
I have data sets from unknown f_i, and I know, *e.g.*, $f_0(\vec{x}) = \sum \alpha_i f_i(\vec{x})$. I want to know α_i but don't care about f_i. Here's an example:



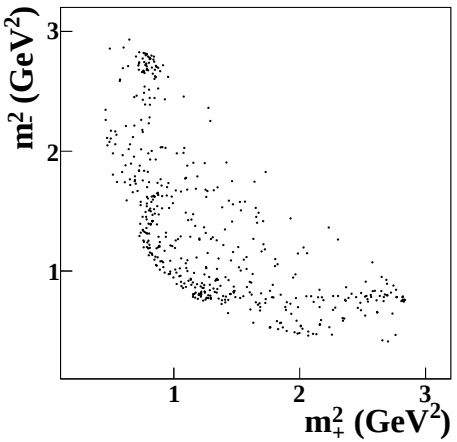
$= \beta_0$



$+ \beta_1$



$+ \beta_2$



I just want to know β_i. Clearly trying to model f_i is going to be very hard. Can I avoid it and avoid using kernels?

Non-PDF Nonparametric Regression

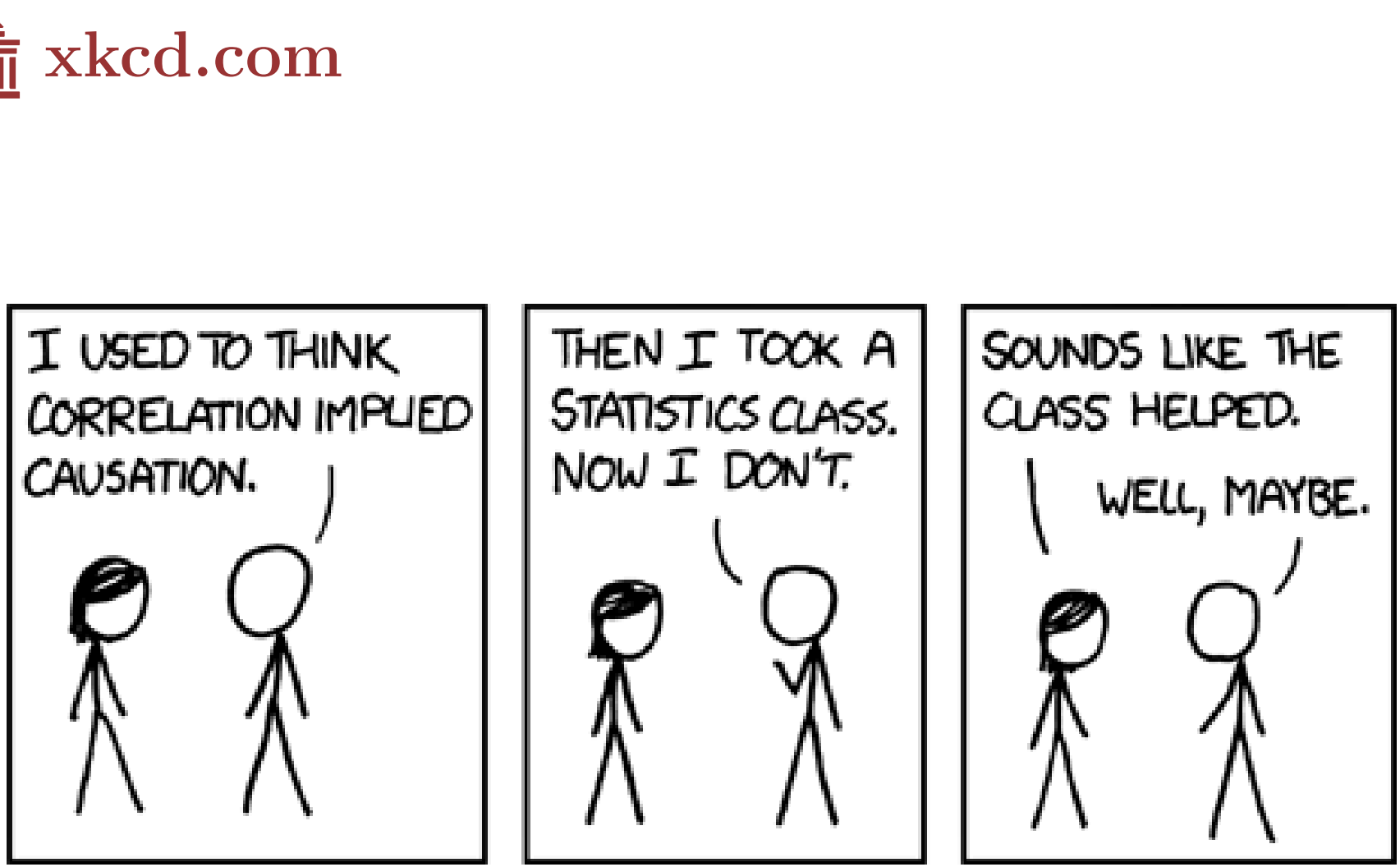
Remember the energy test? We can use it here for regression by allowing the "charges" of the 3 samples to float. These charges are then related to the coefficients we want to know.

In this example (see ref below), it correctly determines the parameters of interest with 7% relative uncertainties. It nevers tries to estimate the PDF, it circumvents it by using the fact that the data sets we have follow those PDFs.

See M.Williams, *Nonparametric regression using the concept of minimum energy*, JINST **5**, P10003 (2011).

Summary

- Bayesians and frequentists philosophies are very different; however, with enough data the data "speaks for itself" and you'll get the same result using either approach.
- Bayesians can answer any question but you're free to disagree with them.
- It's very important when setting limits to choose your strategy prior to looking at your data. If you have a large systematic uncertainty, use a Bayesian method. If not, I prefer frequentist but do what you want (is easiest to get past your referees).
- Why parametrize f if you don't care about f? Non-parametric methods work well. Plenty out there. Use them!


xkcd.com
I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.