Statistics for Physicists

J. Hardin

IceCube Boot Camp 6/14/2022

Intro

- Hi, I’m John Hardin - I’m a postdoc in Janet Conrad’s group, and I am in the MEOWS working group
- I’ve been asked to talk about statistics - I’ve tried to avoid overlap with yesterday, but please treat any overlap as a refresher.
- I’m going to start pretty basic, I’m sorry to those of you that have seen some or all of this. I hope it’s still useful.
- Please feel free to interrupt and ask questions - it works better if you do.

Outline

1) Probability (What questions are we asking?)
2) The Loglikelihood and χ^2 (How do we answer them?)
3) Test statistics and confidence regions (What does that answer mean?)
4) Real Experimental Considerations (Systematics, etc)
5) Physics statistics jargon (“Brazil Plots”, “Feldman Cousins”, “Look Elsewhere”, “Sigma”)

Probability
Or
“What Questions are We Asking?”

Probability

- Probability of A
- Probability of A given B

- Bayes Theorem

$$P(A)$$

$$P(A|B)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(A) = \frac{P(A|B)P(B)}{P(B|A)}$$

Notation

- Model (parameter set) $\vec{\theta}$
- Data x_i
- Probability of Data given Model $P(\vec{x}|\vec{\theta})$

Ultimately, what questions are we asking

- Strict Frequentist: What is the probability of the data given out model (“rule out”)

$$P(\vec{x}|\vec{\theta})$$

$$P(\vec{\theta}|\vec{x})$$

- Bayesian: What is the probability of our model (Given our data and previous belief)

$$P(\vec{\theta}|\vec{x}) = \frac{P(\vec{x}|\vec{\theta})P(\vec{\theta})}{P(\vec{x})}$$

Ultimately, what questions are we asking

Easy

Hard/Debatable

Not as problematic
as it seems

$$\boxed{P(\vec{x}|\vec{\theta})}$$

$$P(\vec{\theta}|\vec{x})$$

$$P(\vec{\theta}|\vec{x}) = \frac{\boxed{P(\vec{x}|\vec{\theta})}\,\boxed{P(\vec{\theta})}}{\boxed{P(\vec{x})}}$$

As Physicists, we are lazy

- We will focus on the easy problem, and talk about the “harder” bits later
- It’s relevant to both kinds of interpretation, so you can ignore the statistical philosophical wars for now
- We will be focusing on 2 things:
    - Exclusion
    - Estimation

$$P(\vec{x}|\vec{\theta})$$

There are no frequentists in foxholes

The Loglikelihood and χ^2

Or

"How Do We Answer Them?"

Start with the easy thing

- The probability of independent statistical events is just the product of their probabilities
- A model, by definition, provides a probability that a given point is observed
- So, we just multiply the probabilities (And sweep the infinitesimals under the rug)

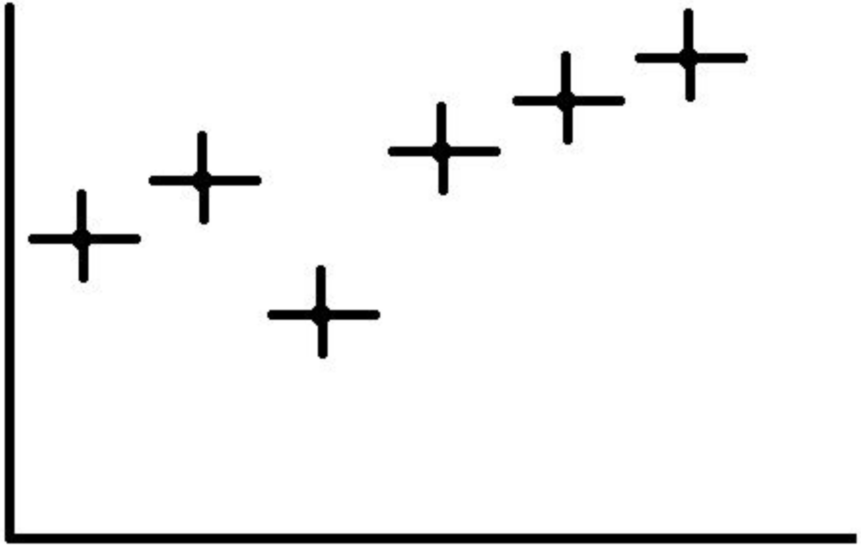
$$L(\vec{\theta}) = \prod_i pdf(x_i|\vec{\theta})$$

Computers, therefore:

$$LL(\vec{\theta}) = ln(\prod_i pdf(x_i|\vec{\theta})) = \sum_i ln(pdf(x_i|\vec{\theta}))$$

What about Bins?

- "Unbinned" has strictly more information, but we may wish to bin things
- Doing complicated things to various bins for simplicity of models (Ratios, primarily, but be careful)
- Faster



But How?

- Bin at location b_i, with count n_i - basic binned likelihood

$$LL(\vec{\theta}) = \sum_i n_i * ln(pdf(b_i|\vec{\theta}))$$

- Assume each bin has gaussian error $\boldsymbol{\sigma}_i$, expectation $\boldsymbol{\mu}_i$, and observed value y_i

$$LL(\vec{\theta}) = \sum_i ln(e^{-\frac{(y_i - \mu_i(\vec{\theta}))^2}{2\sigma_i^2}})$$

$$= -\frac{(y_i - \mu_i(\vec{\theta}))^2}{2\sigma_i^2}$$

$$= -\frac{\chi^2(\vec{\theta})}{2}$$

- We like the χ^2 (for reasons to come)

What do we do with this

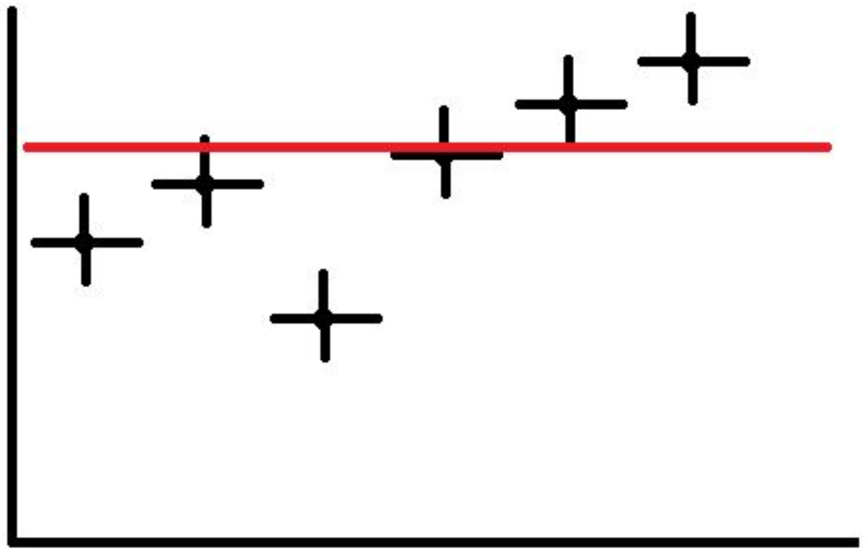
- Optimize!
- We find the model that is most likely to produce our data
- Produce deltas
- The LL is the "best" information you can get
    - χ^2 is nice too I guess

$$\Delta LL = Max_{\vec{\theta}}(LL(\vec{\theta})) - LL(\vec{\theta}_0)$$

$$\Delta\chi^2 = \chi^2(\vec{\theta}_0) - Min_{\vec{\theta}}(\chi^2(\vec{\theta}))$$

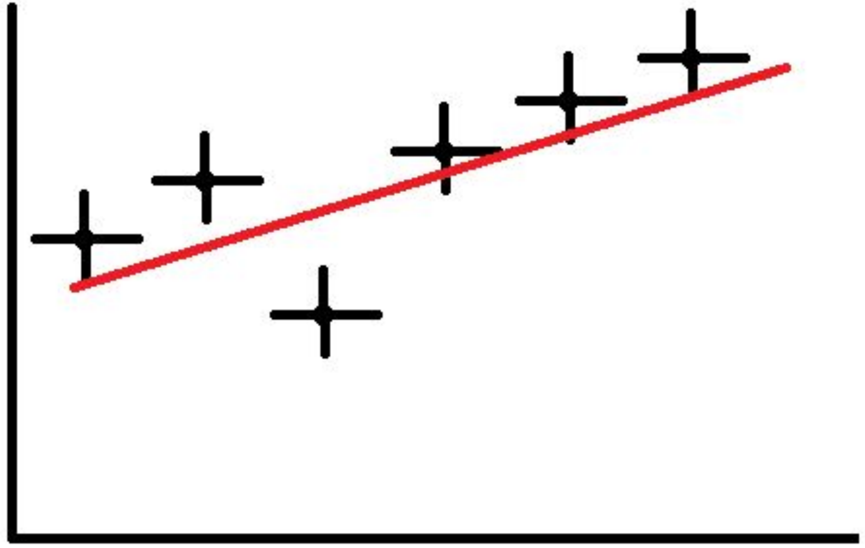
The null model

- Just a boring model with 0 or no parameters (No slope, for instance)



The Actual Model

- This has some parameters and dynamism
- You can see the χ^2 going down as it fits better
- So now what



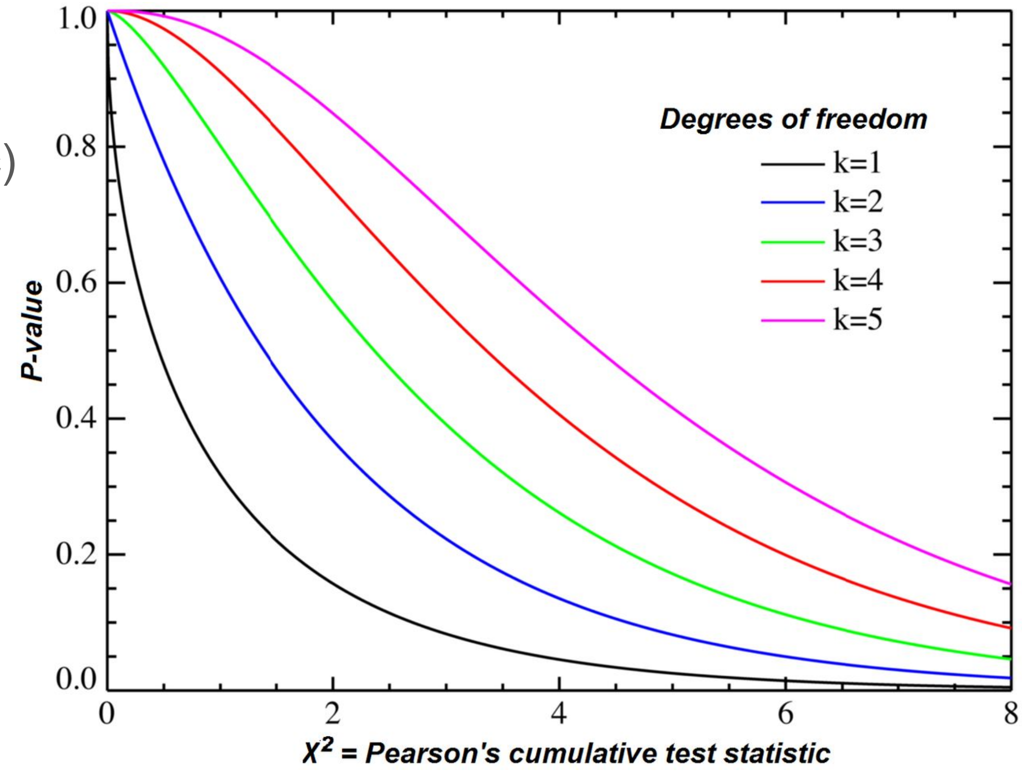
Test Statistics and Confidence Regions

Or

“What Does the Answer Mean?”

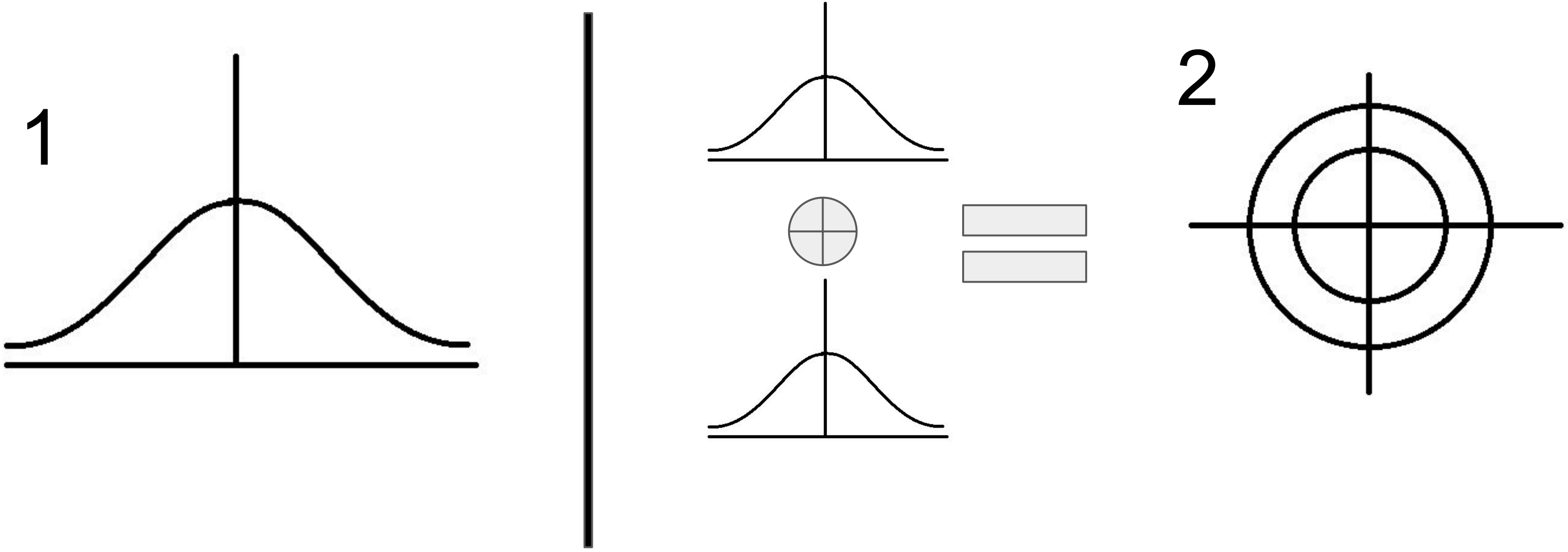
These are what we call “Test Statistics”

- If you take a stats class, you will see others (t-tests, k-tests, Rs, etc)
- We don’t care about those - you can usually derive them from the LL (but you shouldn’t - just look them up when needed)
- But how do we interpret them?
- Let’s start with the χ^2 - its the easiest because it’s a bunch of gaussians



By Mikael Häggström - File:Chi-square distributionCDF.png, Public Domain, https://commons.wikimedia.org/w/index.php?curid=10633630

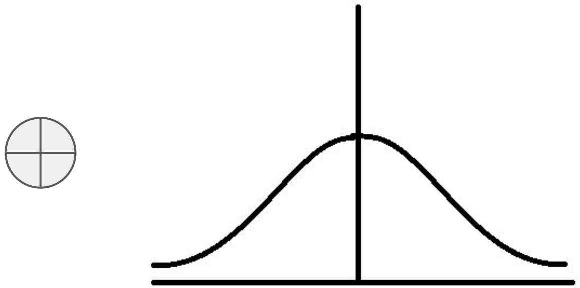
“Degrees of Freedom”



“Degrees of Freedom”

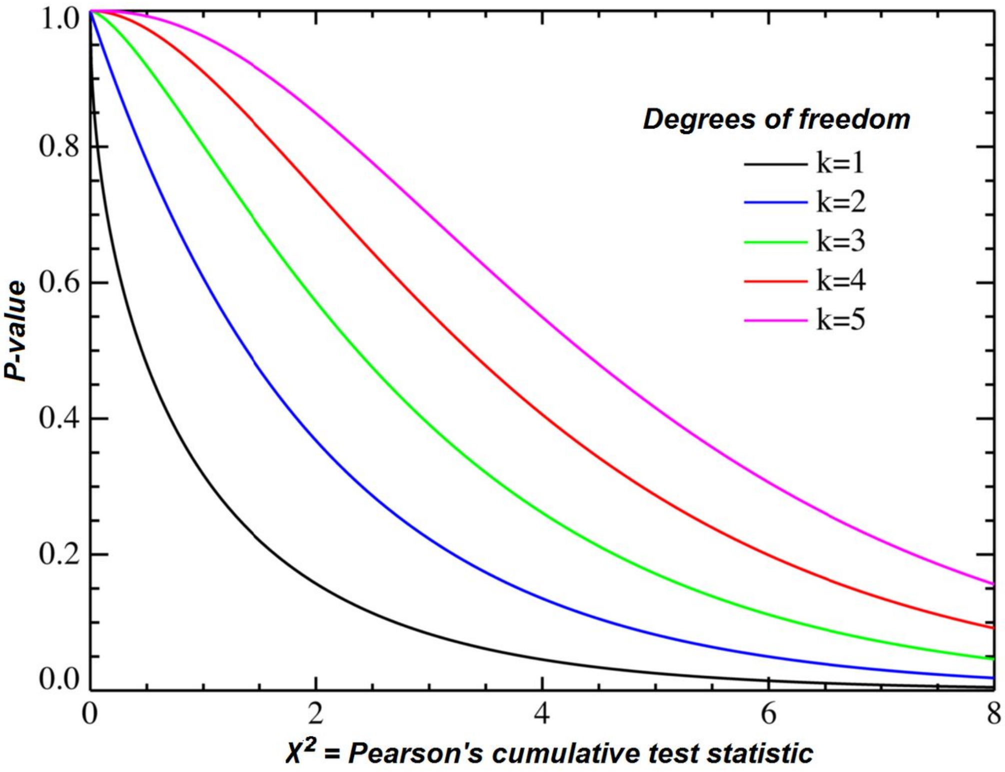
- N degrees of freedom
- Pretty easy to calculate analytically
- Rests on the assumption that everything underlying is gaussian

N



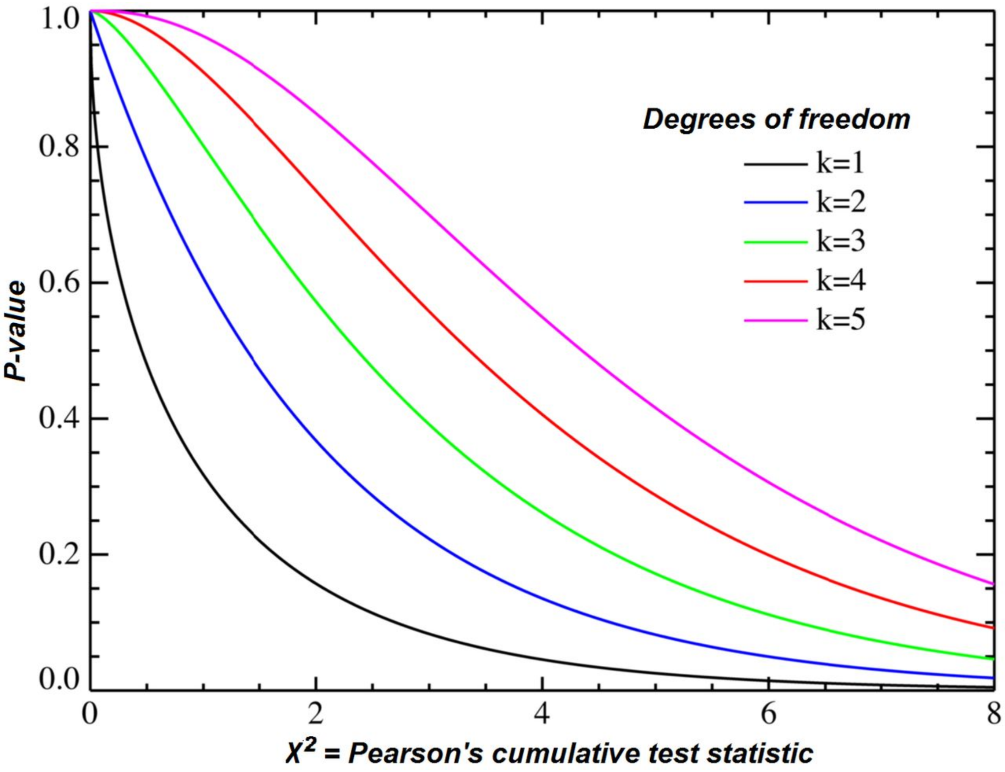
The “P-value”

- We get a “p-value”
- A p-value is the probability that our test statistic will be this weird or weirder
- In physics, we often convert this p-value to an equivalent **σ** - how far one would have you be from a standard gaussian to be just as weird
- We mark 3**σ** (1/300) as “evidence” and 5**σ** (3e-7) as “discovery”
    - I have many opinions on many things, but especially on these



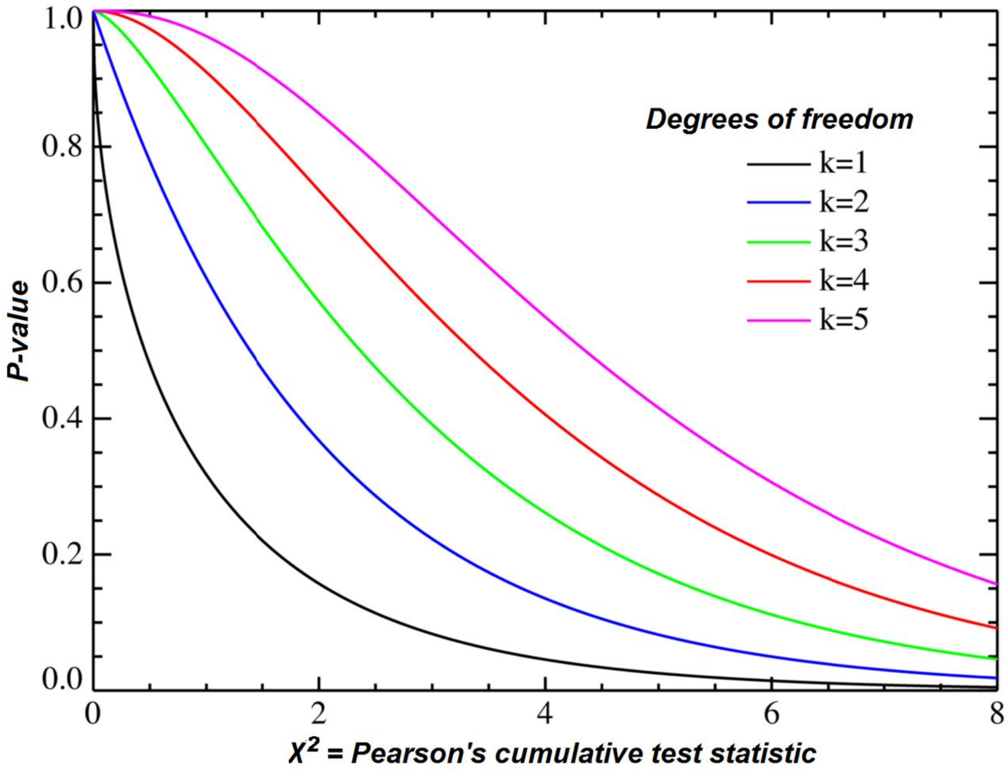
Why are we talking about χ^2

- I've pulled a fast one on you
- We talk about the χ^2 distribution differently from the $\Delta\chi^2$ - The former is a theoretical distribution from a combination of gaussians, and the latter is a common test statistic
- It's important to keep those straight



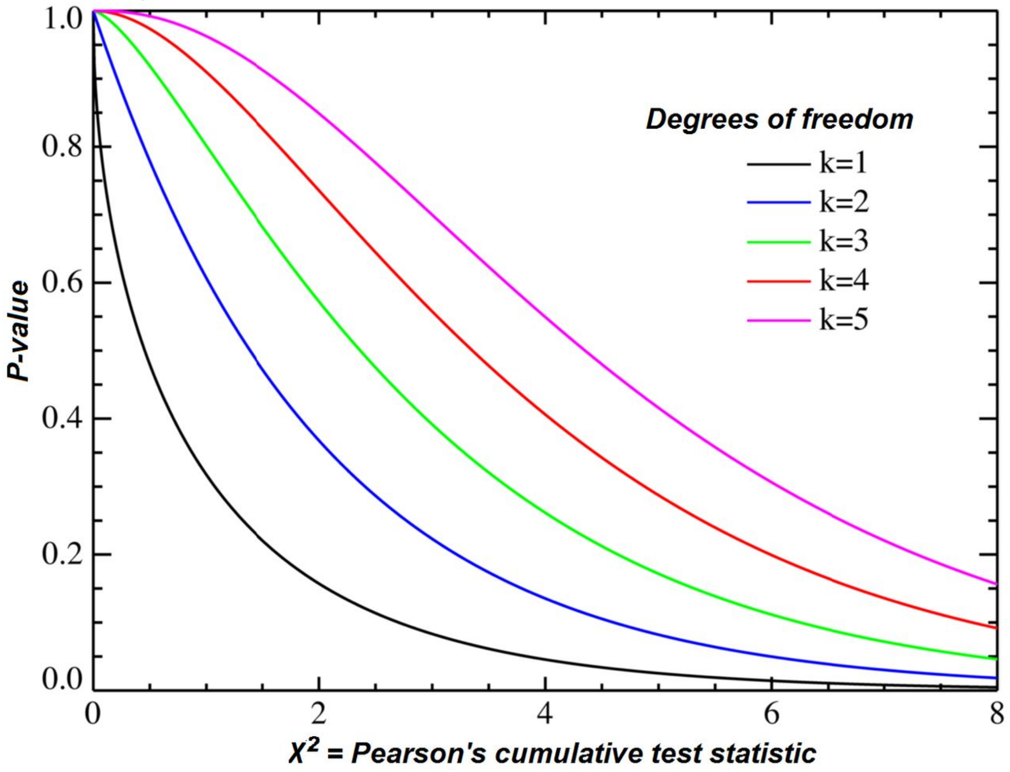
Why are we talking about χ^2

- Under certain assumptions, according to a theorem known as "Wilks Theorem", the $\Delta\chi^2$ will follow a χ^2 distribution with n degrees of freedom where n is the number of parameters in your model
- Not only that, but a 2*ΔLL will ALSO follow the χ^2 distribution with n degrees of freedom



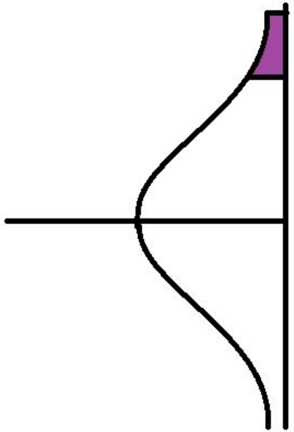
This plot is useful, but a warning

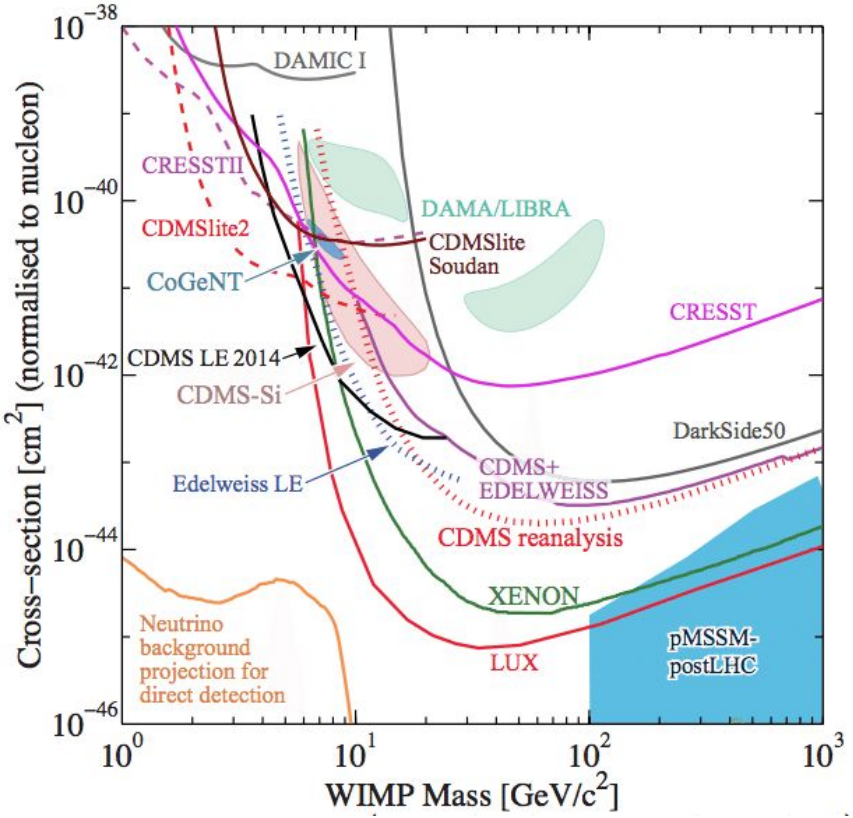
- We OFTEN violate those conditions in physics making us more and less “conservative” - in those cases, the $\Delta\chi^2$ will NOT follow a χ^2 distribution
- When this is the case, it is important to build up your own test statistic distribution by throwing from your null mode and applying your fitting procedure. You will have to throw O(1/(desired pval)) to do it properly



Back to exclusion

- This is very traditional science
- We reject the null at XX%
- This is the mode we use for discovery
- I like to think about it as a 1-D gaussian, but this is often done in many parameters
- You can even do exclusion differentially for some parameters

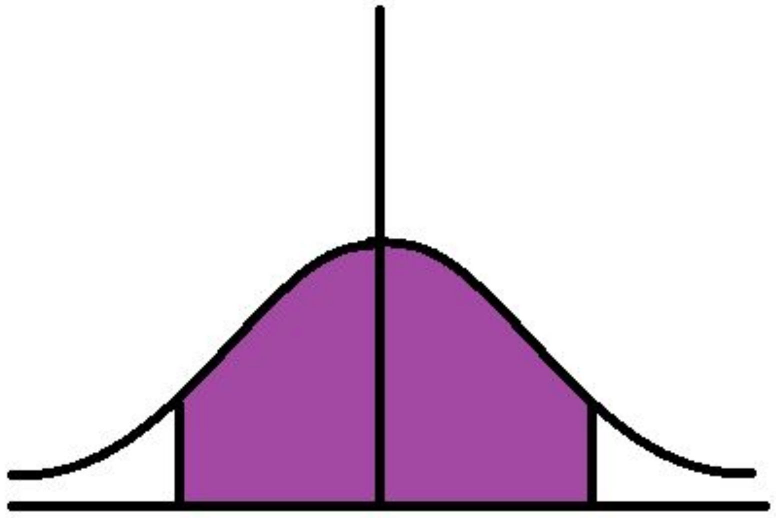




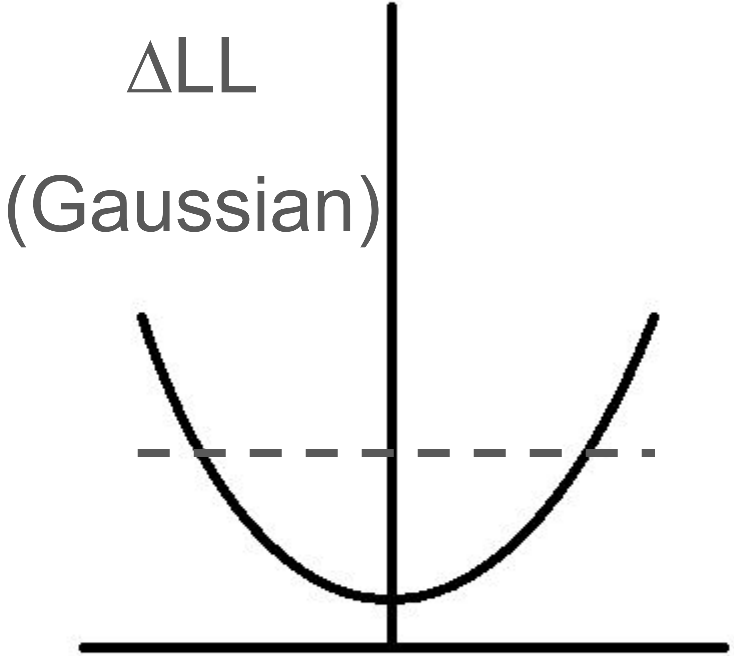
https://www.researchgate.net/figure/Current-exclusion-limits-and-regions-of-interest-of-dark-matter-searches-for_fig1_320890690 (2017)

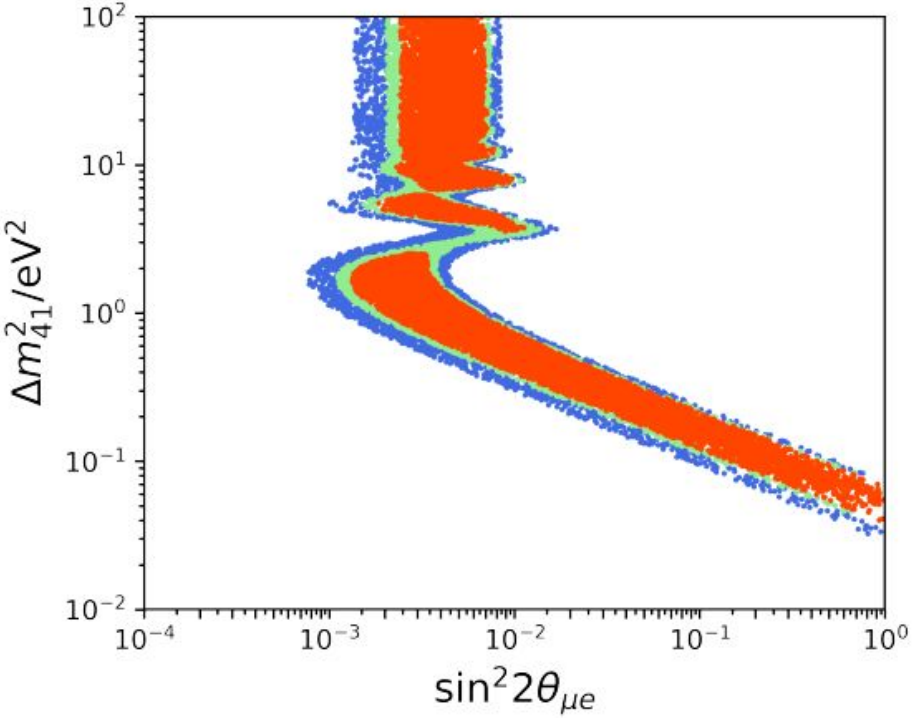
If we find something

- Parameter estimation!
- As good frequentists, we construct confidence intervals
- Formally, a 90% confidence interval is constructed to contain the true parameter 90% of the time
- Many degenerate ways of doing this, but we normally pick density
- We can again use Wilks theorem to assume everything is gaussian OR we should check with realizations
- Often we have to only check some of the model space



Other views

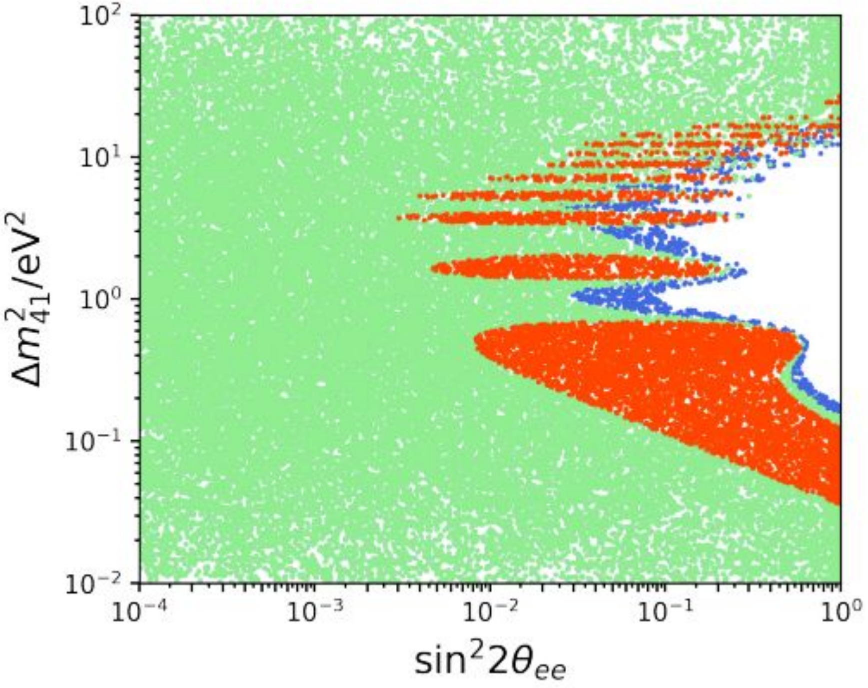




https://arxiv.org/pdf/1906.00045.pdf

Summary for Exclusions and Estimation

- The ΔLL is the gold standard for information about your space
- Wilks theorem is useful as a shortcut
- You should check it anyway



https://arxiv.org/pdf/1906.00045.pdf

Real Experimental Considerations

Or

“What is Everyone Actually Arguing about at Working Group Meetings?”

First, a bonus tip

- You may have noticed something squirelly about the normalization in the LL
- If you have the overall normalization as a parameter (as is often convenient), it will want to go to infinity
- You can fix this with a normalized LL
- The math comes from a poisson distribution, but I won't bore you with it

$$LL(\vec{\theta}) \rightarrow LL(\vec{\theta}) - \int_x F(\vec{\theta})$$

Systematics

- The Loglikelihood version
- “Pull terms”
- We assume we know something, but not everything, about where the non physically relevant params ought to be
- So we penalize the likelihood for moving away
- Assume we have two physics params, and θ_2, is, say, DOM efficiency

$$LL(\vec{\theta}) \to LL(\vec{\theta}) - \frac{(\theta_2 - \mu_{\theta_2})^2}{2\sigma^2_{\theta_2}}$$

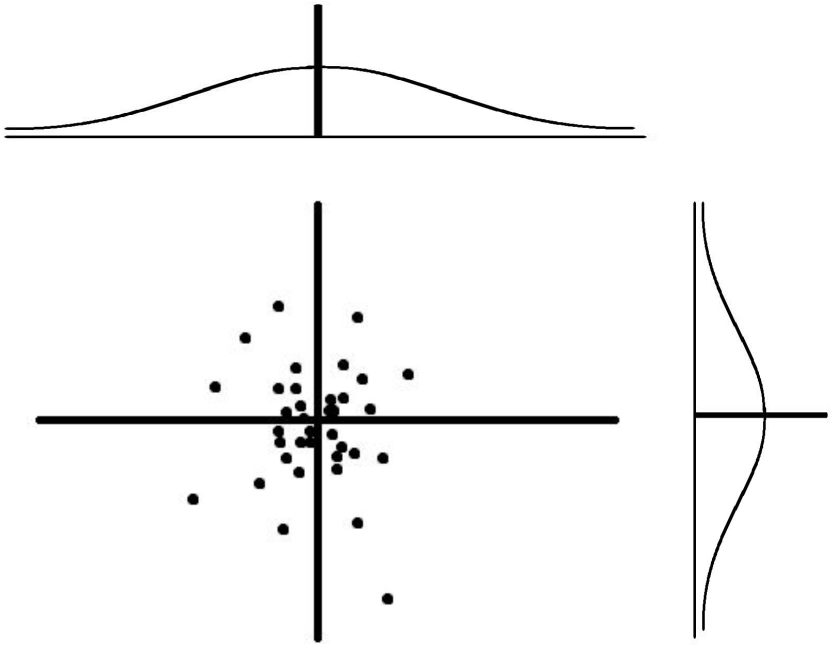
Aside: regularization

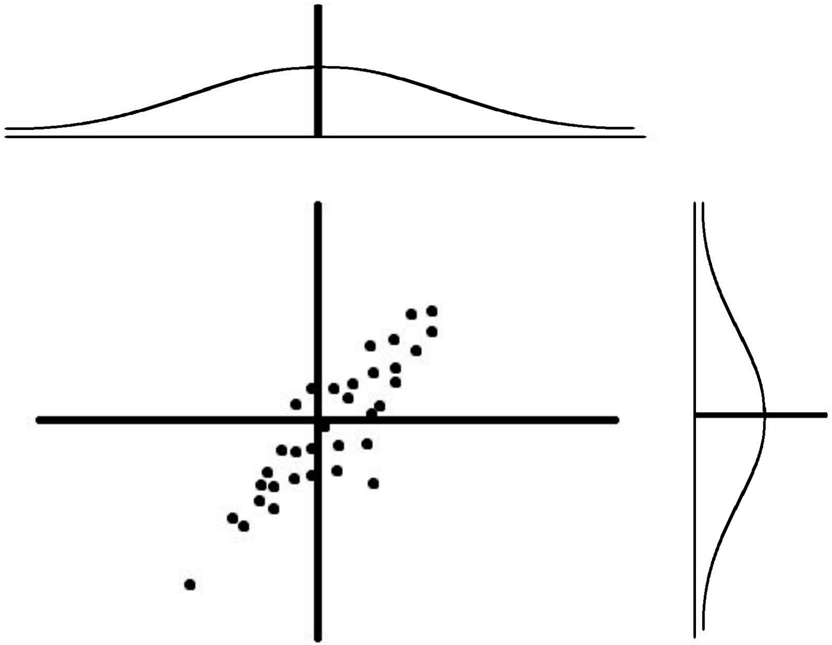
- Occasionally, you will have a very general model
- You might want to avoid overfitting by penalizing amplitudes
- These things are sort of Bayesian (the math tends to be the same)

$$F(\vec{A}) = \sum_{j}^{n} A_j x^j$$

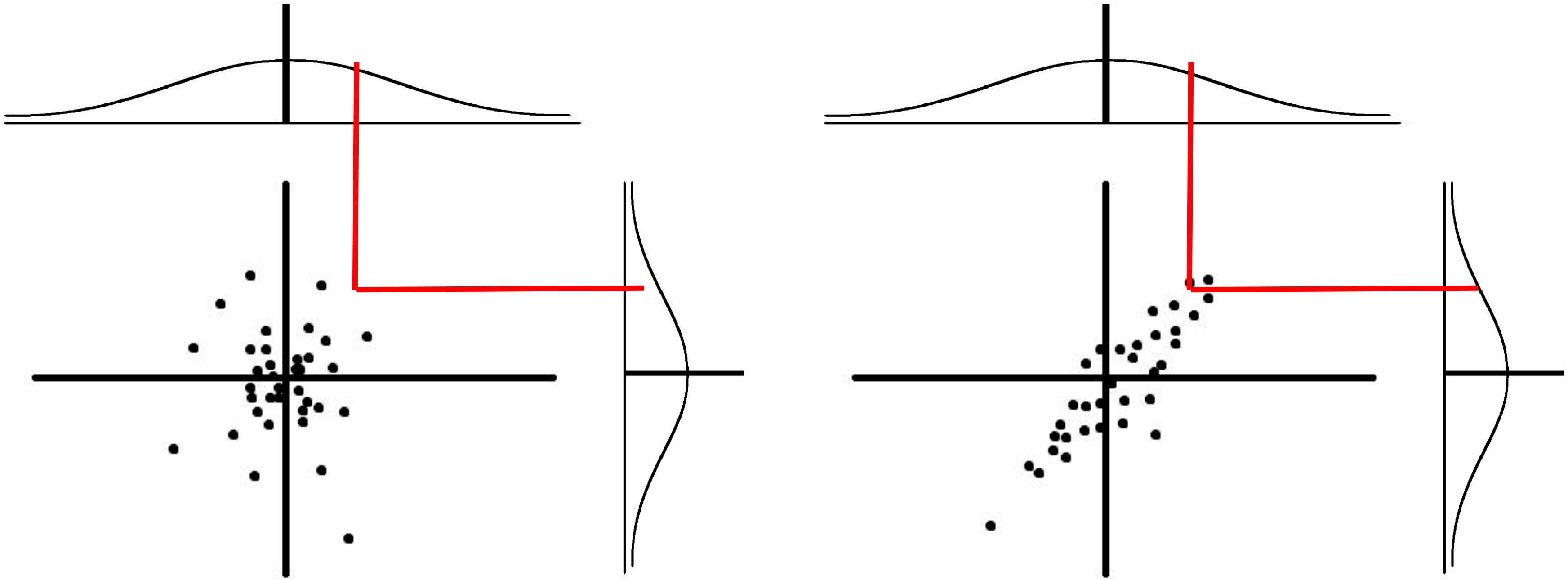
$$LL(\vec{A}) \to LL(\vec{A}) - \sum_{j}^{n} |A_j|^m$$

Systematics part 2

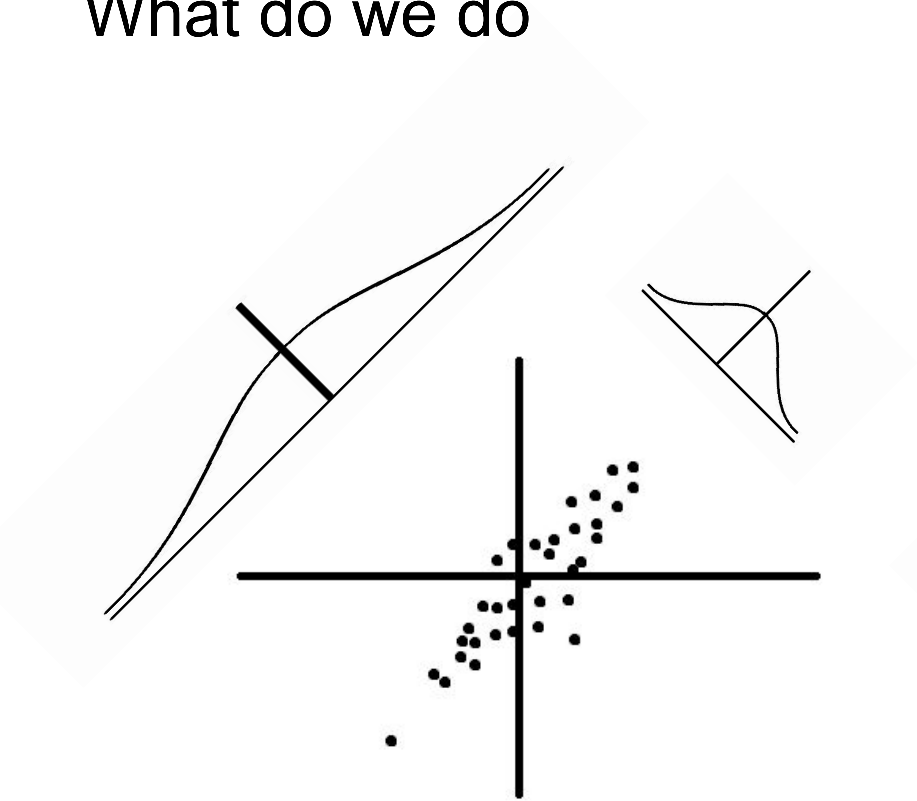


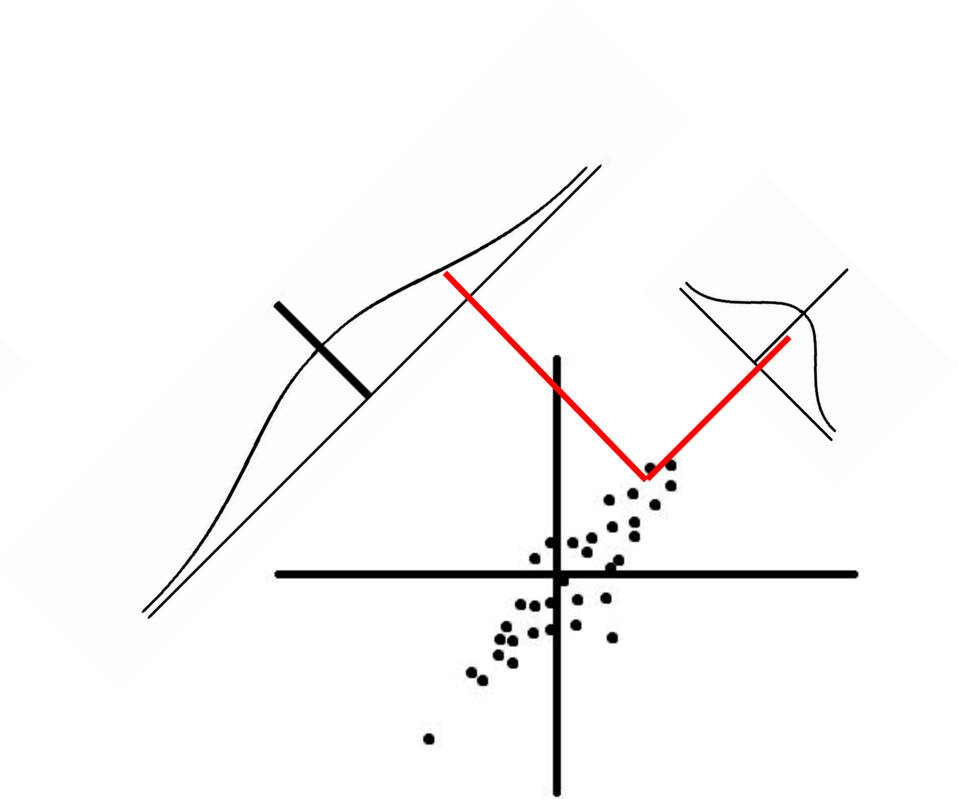


How weird is the same point



What do we do





The Correlation Matrix

- When there is no correlation, it is diagonal
- Can be “block diagonal”
- Makes it easy to convert a set of correlated residuals (r) to a χ^2
- This is useful, but relies on a gaussian assumption
- In principle, the correlation matrix is a function of model parameters - be careful

$$r_i(\vec{\theta}) = n_i - \mu_i(\vec{\theta})$$

$$\chi^2(\vec{\theta}) = \vec{r}(\vec{\theta})^T C(\vec{\theta}) \vec{r}(\vec{\theta})$$

General Systematics Warnings

- Systematics can be as much art as science
- You will mostly see them as pull terms or correlation matrices
- In principle, you are worried about things that look like your signal, even a little bit
- Check carefully - trials are your best friend

https://huggingface.co/spaces/dalle-mini/dalle-mini

Physics Jargon
Or
“What Did They Mean by That Barely Googlable Term?”

Sigma

- We talked about this already
- My pet peeve: “**σ**” >~8 - that is not a real p-val that you can check
- Look at arxiv.org/ftp/arxiv/papers/1103/1103.5672

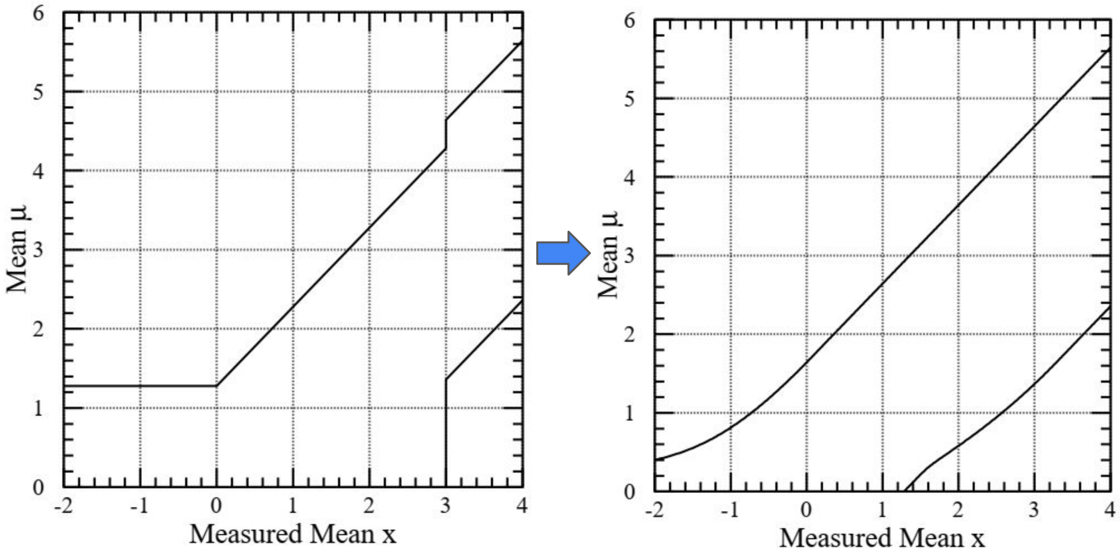
- 7 **σ** is 1/7.8e11
- 10 **σ** is 1/1.3e21
- 25 **σ** is ⅓.3e135

- Tails don’t stay that gaussian that long

- “We were seeing things that were 25-standard deviation moves, several days in a row,”
    - David Viniar, just before 2008
- “This fits to 11 Sigma”
    - A talk on the (real, but not that real) pentaquark

Feldman-Cousins

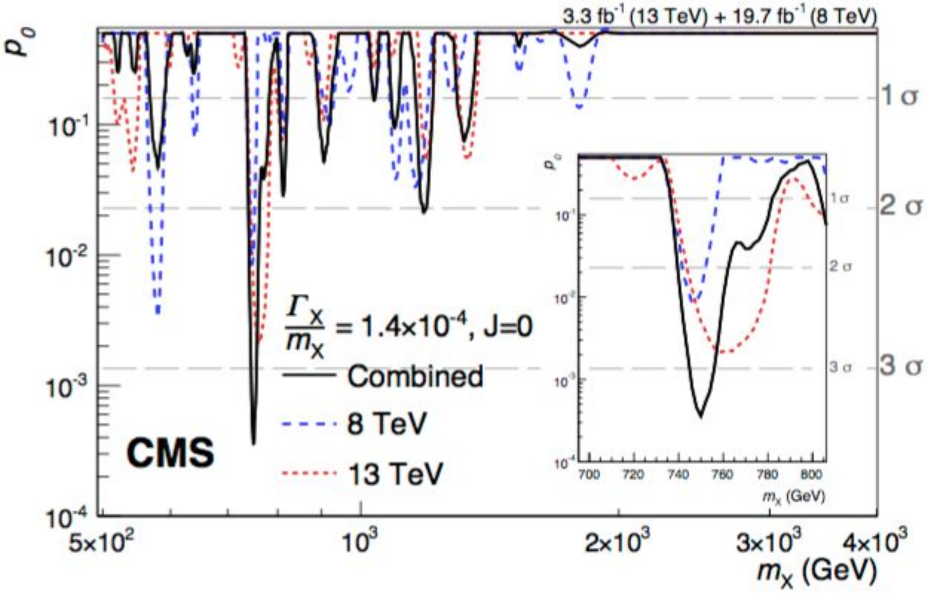
- You should go read the paper
- Essentially, it aims to unify exclusion and estimation
- Other meanings: Trials



https://arxiv.org/abs/physics/9711021

Look-Elsewhere

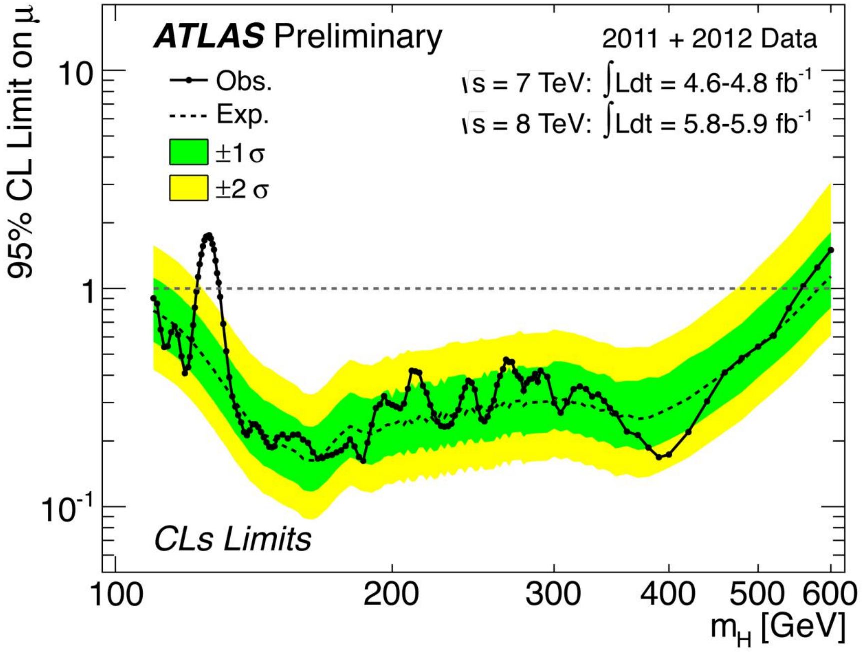
- One of the ways that Wilks fails
- If you are looking for a bump on noise, you can focus on each slice of the plot individually
- This gives you more bites at the apple than Wilks assumes
- ALWAYS run trials



https://arxiv.org/pdf/1606.04093.pdf

Brazil Plots

- Popularized by the Higgs Boson
- Run null trials, then general exclusion bands
- Put the middle 1 **σ** of bands in green and the middle 2**σ** in yellow
- Then overlay the actual data
- Where the data fails to exclude as strong as you expect - that's where you expect your new physics



https://home.cern/resources/image/physics/infographics-gallery

Last Summary

1) Probability (Bayesian? Frequentist? Strange notation?)
2) The Loglikelihood and χ^2 (How I learned to stop worrying and love LL)
3) Test statistics and confidence regions (How I learned to stop worrying and love running null realizations)
4) Real Experimental Considerations (What is really going on)
5) Physics statistics jargon (How we talk about things)

Thank you

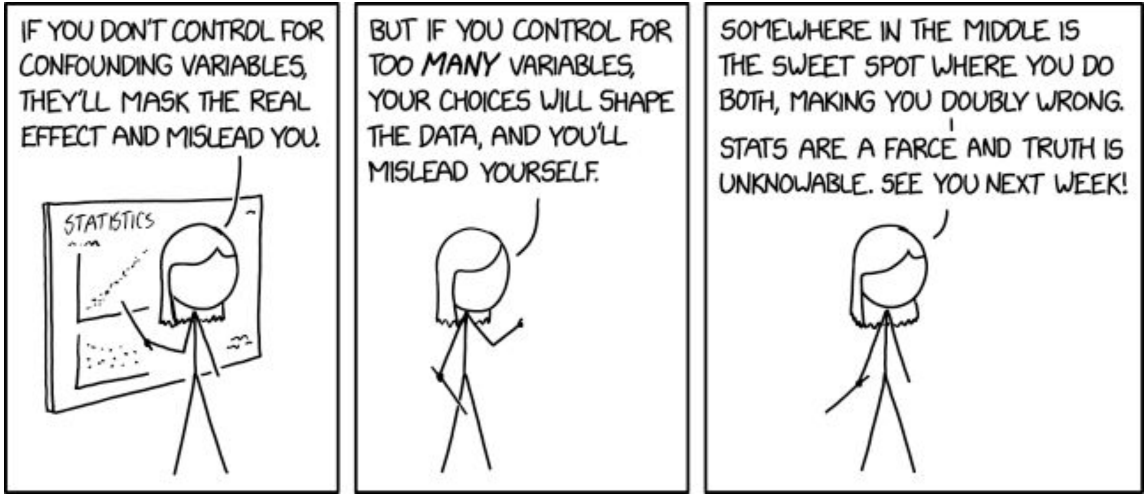
Or

“Any Questions?”

Rejected General Systematics Warnings

- Systematics can be as much art as science
- You will mostly see them as pull terms or correlation matrices
- In principle, you are worried about things that look like your signal, even a little bit
- Check carefully - trials are your best friend

https://xkcd.com/2560/



https://huggingface.co/spaces/dalle-mini/dalle-mini