Probability, Statistics, and Maximum Likelihood

Fermi Summer School

J. Patrick Harding
Los Alamos National Laboratory

29 May 2018

Overview

1) Statistics and Probability
2) Likelihood Analysis
3) Maximum Likelihood Ratio Test
4) Advanced Likelihood Topics
   - Trials Factors
   - Uncertainty in Background Measurement
   - Tools for Maximizing Likelihood

1) Statistics

Notation:

- $P(x;p)$ is probability of measuring x given inputs p
- Probability Density Function (pdf) $f(x;p)$
  - $dP(x;p) = f(x;p)dx$ is differential probability for continuous variables x given inputs p
- Cumulative distribution function (cdf)

$$F(x;p)=\int_{-\infty}^{x} f(x;p)\,dx$$

What do you need to calculate parameter values and uncertainties?

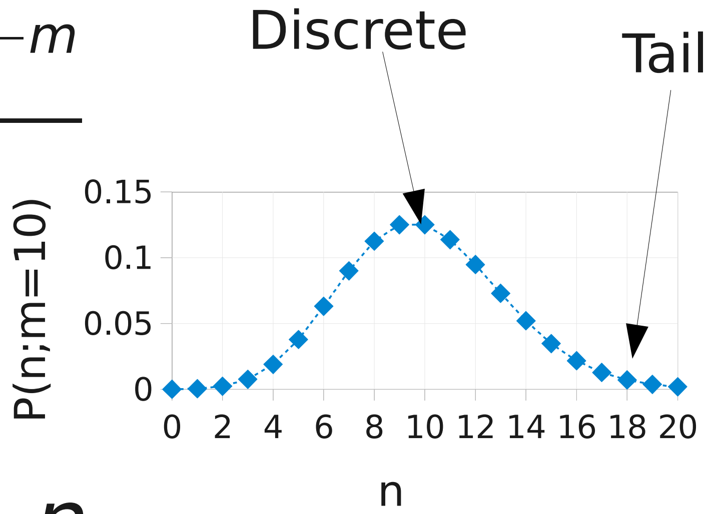
1) Statistical distribution of the data
2) Pick your statistical framework
    1) Bayesian or Frequentist?
3) Calculate
    1) Parameter value
    2) Confidence Interval
    3) Significance of the parameter

What do you need to calculate parameter values and uncertainties?

1) **Statistical distribution** of the data
2) Pick your statistical framework
    1) Bayesian or Frequentist?
3) Calculate
    1) Parameter value
    2) Confidence Interval
    3) Significance of the parameter

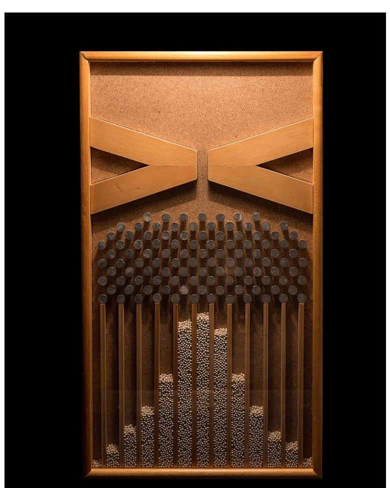
Poisson Distribution

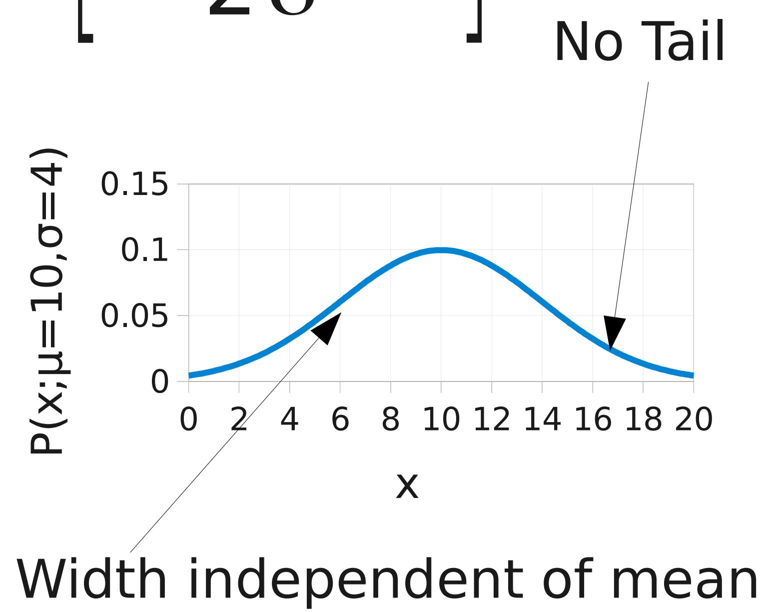
- "Counting Statistics"
- Measured **discrete** variable *n* with expected value *m*
- Probability $P(m;n)=\frac{m^n e^{-m}}{n!}$
- Mean *m*
- Standard deviation $\sqrt{m}$
- For large numbers *m* and *n*, approximates to a Gaussian (via Sterling's approximation)



Gaussian Distribution

- Measured **continuous** variable *x* with expected value *μ* and width *σ*
- pdf $f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$
- Mean *μ*
- Standard Deviation *σ*





What do you need to calculate parameter values and uncertainties?

1)Statistical distribution of the data

2)Pick your statistical framework

- **1)Bayesian or Frequentist?**

3)Calculate

- **1)Parameter value**
- **2)Confidence Interval**
- 3)Significance of the parameter

Confidence Intervals

- Bayesian
  1) Assumes answer follows a distribution
  2) Assume a “prior distribution” based on pre-existing knowledge/prejudice
  3) Take some data
  4) Update prior distribution to “posterior distribution” based on data
  5) Get confidence interval from posterior distribution
  6) Is fraction of time distribution is in interval
- Answer depends on your initial prior

Confidence Intervals

- Frequentist
  1) Assumes there is a single correct answer but that sampled size is finite
  2) Take some data
  3) Assume data is representative of the full distribution
  4) Get confidence interval from data distribution
  5) Is fraction of time a random data point lies in that interval
- Prior knowledge goes into model interpretation of confidence interval

Bayesian vs. Frequentist

- Bayesian
  - Answer is the distribution of true values
  - Confidence interval is a range in the distribution
  - Depends on prior knowledge
- Frequentist
  - Answer is a guess at the “right value”
  - Confidence interval is how likely you are to guess the right value is in that interval once you've taken data

Bayesian vs. Frequentist

Bayesian

Max of the distribution

Width of the distribution

Frequentist

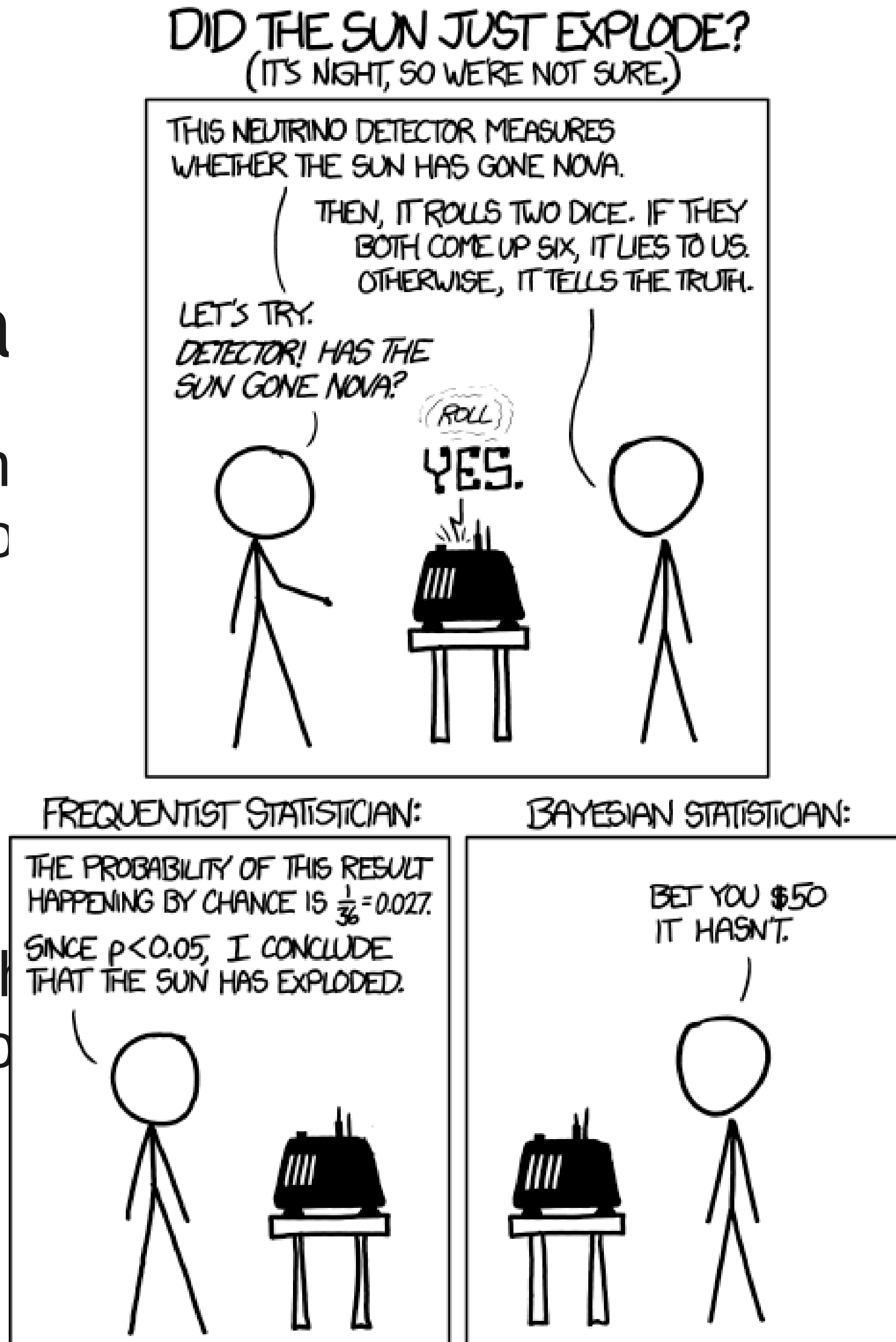
Guess at right answer

Range the right answer may be in

$$10 \pm 3$$

Bayesia … equentist

Max of th … ess at right
distributio … answer



Width of th … nge the right
distributio … ver may be in

What do you need to calculate parameter values and uncertainties?

1) Statistical distribution of the data
2) Pick your statistical framework
    1) Bayesian or Frequentist?
3) Calculate
    1) Parameter value
    2) Confidence Interval
    3) **Significance of the parameter**

Significance and "p-values"

- Statistical significance is a measure of how well the null result can reproduce the data
- The p-value says how probable it is that the data is due to background fluctuations:
  - $p \sim 1$ is likely to be from background
  - $p \sim 0$ is likely not due to background
  - value itself depends on the underlying null model assumed (examples later)
- Significance (σ) is often interpreted as the number of standard deviations in the background which would be necessary to reproduce the data

2) Likelihood Analysis

- We are going to use a pseudo-Bayesian analysis for this discussion
- Using methods of a Bayesian analysis, but without using priors
  - Prior knowledge goes into our model instead of our statistics
- Interpreting results as a Frequentist
  - e.g. - What is our PSF?
  - e.g. - What is the flux from the Crab?

Source Model

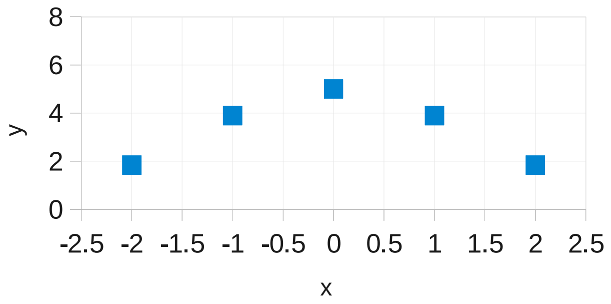
- Model includes everything you know (or wish to know) about a source
- Known parameters and free parameters are both included
- Qualitative behavior is assumed from prior knowledge
  - PSF is Gaussian (or double-Gaussian)
  - Extended source vs point source
  - Steady in time vs decaying signal in time
  - ...

Your Significance is Only as Good as Your Model

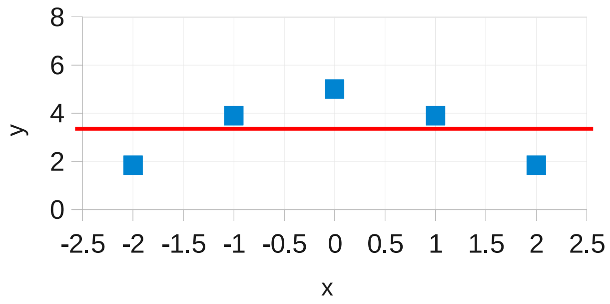
Example:

(Significances assuming one free parameter)

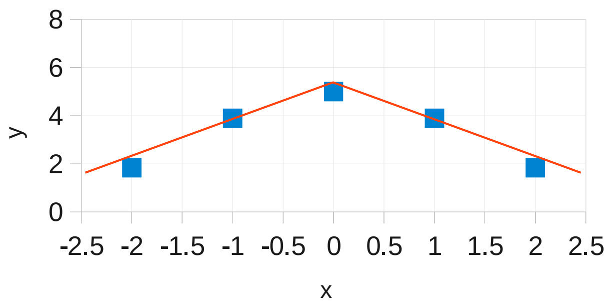
Data



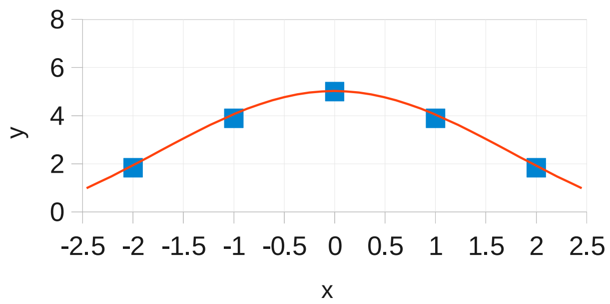
Model 1: 3.7σ



Model 2: 4.05σ



Model 3: 4.06σ



Likelihood

- Likelihood $\mathcal{L}$ is the product of the probabilities of each bin
- For Poisson probabilities:

$$\mathcal{L}=\prod_k P_k=\prod_k \frac{m_k^{n_k} e^{-m_k}}{n_k!}$$

$$ln(\mathcal{L})=\sum_k ln(P_k)=\sum_k n_k ln(m_k)-m_k-ln(n_k!)$$

- Likelihood by itself ***does not*** tell you any information about the goodness/badness of your model fit to the data
- *Note*: the n! term does not affect the outcome of the significance calculations (as we will see)

Binned vs Unbinned Analysis

Unbinned

- Loop over each event
- Don't lose information
  - Makes the most out of each piece of data
- For large data sets, is computationally cumbersome
  - Doesn't necessarily gain much information

Binned

- If too much data is cumbersome, we combine it into bins
- Bins can be in anything:
  - Time
  - Space/solid angle
  - Energy
  - Reconstruction quality
  - ...

Maximizing Likelihood

- Your model M has parameters p
- For any set of parameters p*, this model gives a prediction for each bin, m_k
- Given your measured bin data n_k, find the set of parameters p that gives the model that maximizes $\mathcal{L}$ (or $\ln\mathcal{L}$)
- Can use fitting techniques (discussed later) to find $\mathcal{L}_{max}$
- Note: Often easier to minimize $-\mathcal{L}$

3) Maximum Likelihood Ratio Test

- To determine the significance of your model M_1, you must define a null hypothesis M_0 to compare to
- Separately maximize the likelihood of M_0 and M_1 with respect to their parameters p
- The ratio of $\mathcal{L}_{max,1}/\mathcal{L}_{max,0}$ is used to determine the significance of M_1 over M_0

Test Statistic

- Likelihood ratio test uses a test statistic *TS* to determine significance:

$$TS = 2\ln\left(\mathcal{L}_{max,1}\right) - 2\ln\left(\mathcal{L}_{max,0}\right)$$

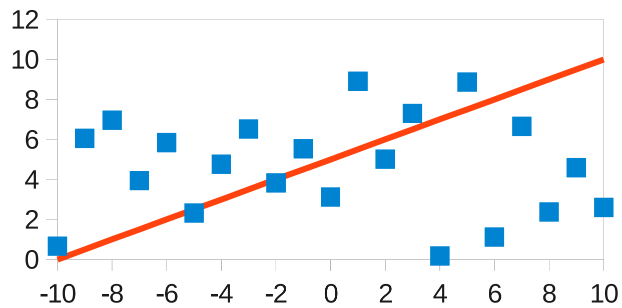
- **Note:** $\mathrm{TS}_{\mathrm{max}} \neq 2\ln(\mathcal{L}_{\mathrm{max},1}) - 2\ln(\mathcal{L}_{\mathrm{max},0})$
  - Need to maximize each separately

Nested Models

- If model M_0 is *identical* to with some of its parameters set to fixed values, then these models are "nested"
- For nested models M_0 with ν_0 free parameters and M_1 with ν_1 free parameters, and a *large* number of counts, *TS* follows a χ^2-distribution with $(\nu_1$-$\nu_0)$ degrees of freedom (Wilks' Thm)

Nested Models Examples

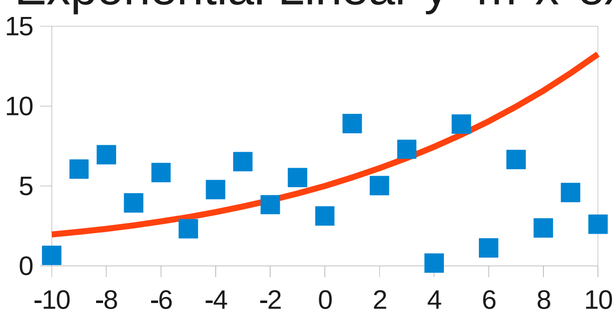
Linear y=m*x+b



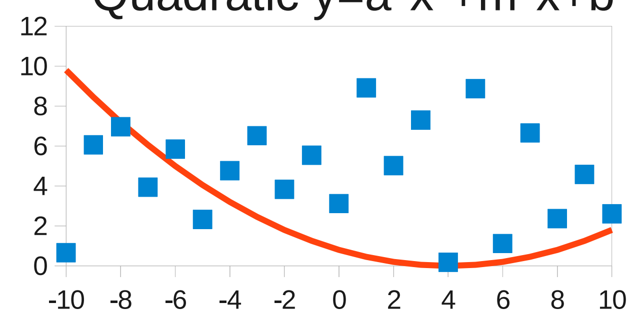
Nested with c=0

Nested with a=0

Exponential Linear y=m*x*exp(c*x)+b



Quadratic $y=a*x^2+m*x+b$



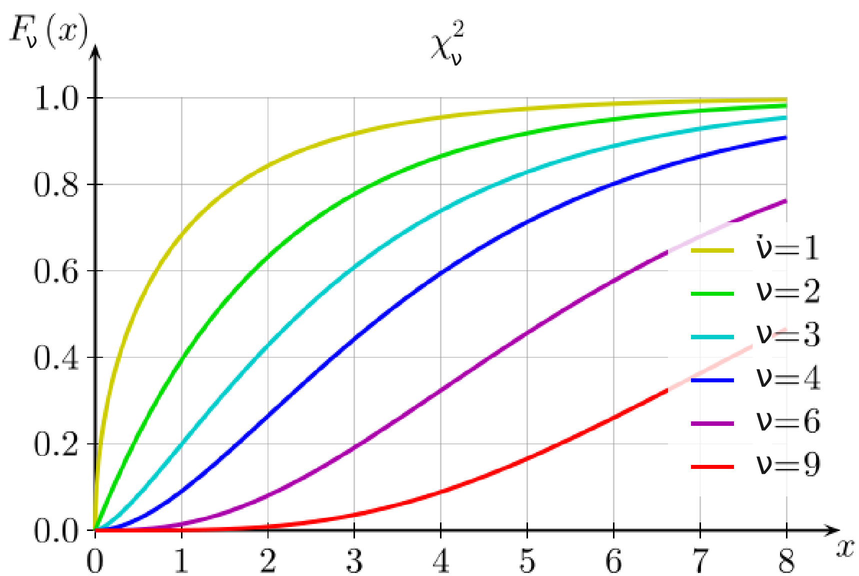
Un-nested

The χ²-Distribution

- The χ^2-distribution with ν degrees of freedom has the pdf:

$$f(x;\nu)=\frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$$

- The cumulative distribution function is:

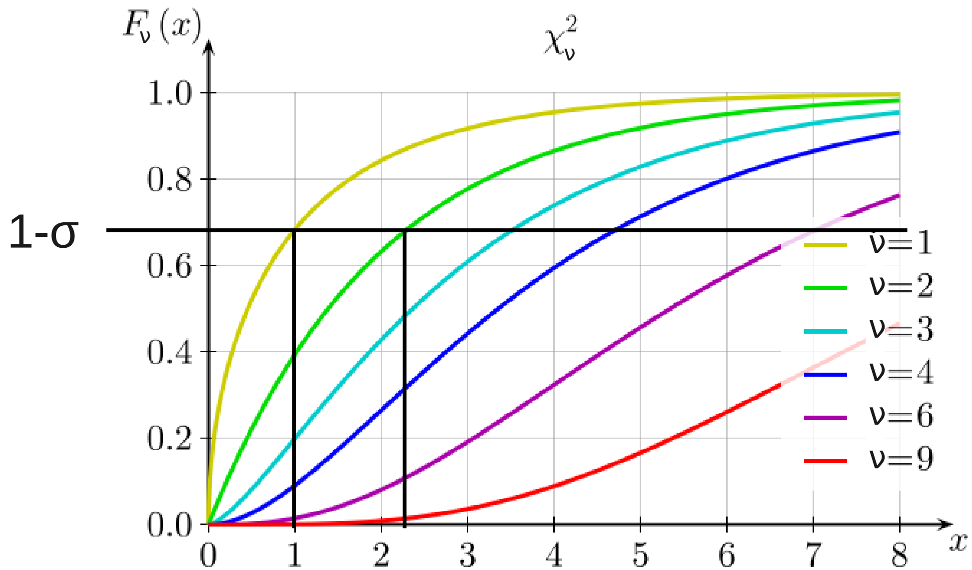


The χ^2-Distribution

- The χ^2-distribution with ν degrees of freedom has the pdf:

$$f(x;\nu)=\frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$$

- The cumulative distribution function is:

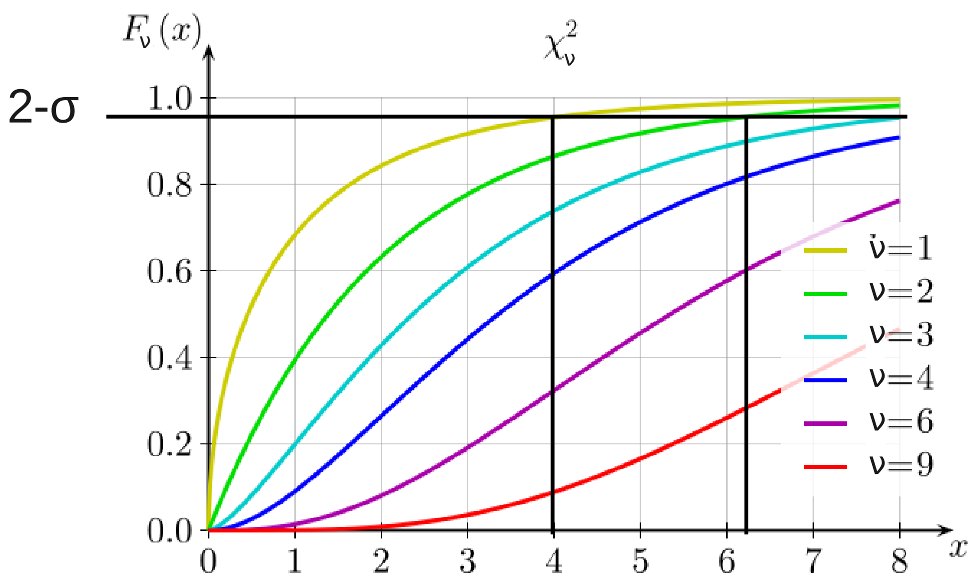


The χ^2-Distribution

- The χ^2-distribution with ν degrees of freedom has the pdf:

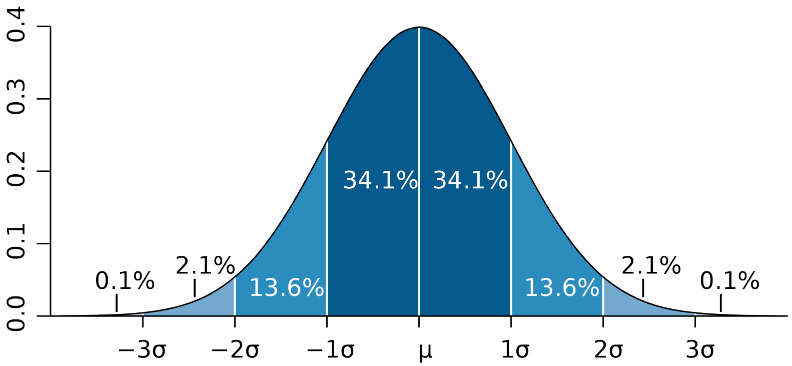
$$f(x;\nu)=\frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$$

- The cumulative distribution function is:



Confidence Intervals

- Q: Why are 1-σ at 68% and 2-σ at 95%?
- A: Because we like Gaussian p-values.
  - For a Gaussian of width σ peaked at $x=\mu$,
    - 68.27% of the curve lies within $-\sigma \leq x-\mu \leq \sigma$
    - 95.45% of the curve lies within $-2\sigma \leq x-\mu \leq 2\sigma$
    - $\mathrm{erf}(q/\sqrt{2})$ of the curve lies within $-q\sigma \leq x-\mu \leq q\sigma$



χ^2 Confidence Intervals

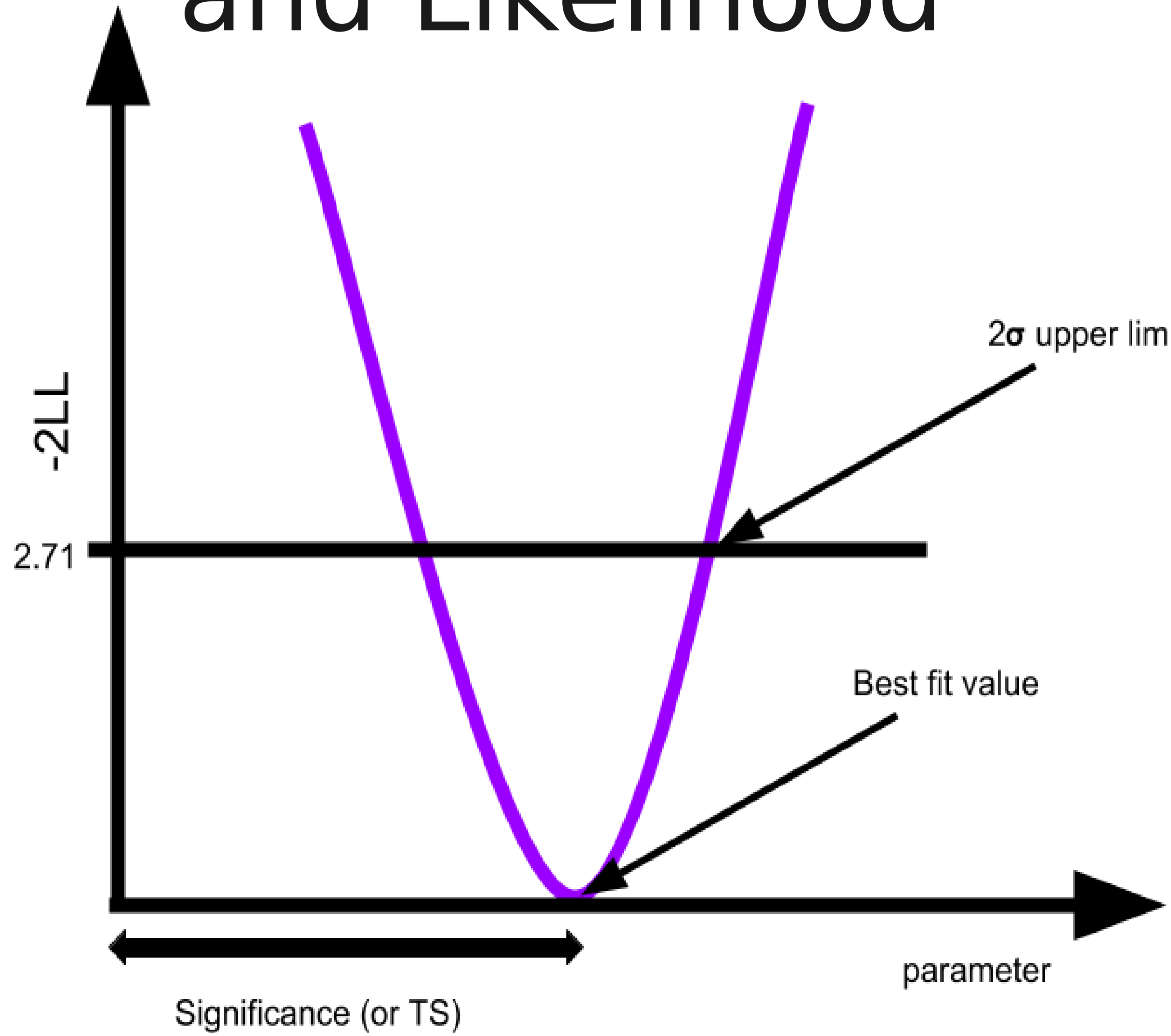
- For a *TS* which follows the χ^2-distribution with ν degrees of freedom, the corresponding confidence level and significance are:

C.L.	σ	d.o.f.=1	d.o.f.=2	d.o.f.=3	d.o.f.=4	d.o.f.=5	d.o.f.=6
68.27%	1	1.00	2.30	3.53	4.72	5.89	7.04
95.45%	2	4.00	6.17	8.02	9.70	11.3	12.8
99.73%	3	9.00	11.8	14.2	16.3	18.2	20.1
99.994%	4	16.0	19.3	22.1	24.5	26.8	28.9
99.99994%	5	25.0	28.7	31.8	34.6	37.1	39.5

- For error bars on a parameters *p*, find the parameter values *p** for which $\Delta TS = TS_{max} - TS(p^*)$=this value

Confidence Intervals and Likelihood

Note: Beware 1-sided vs 2-sided confidence intervals



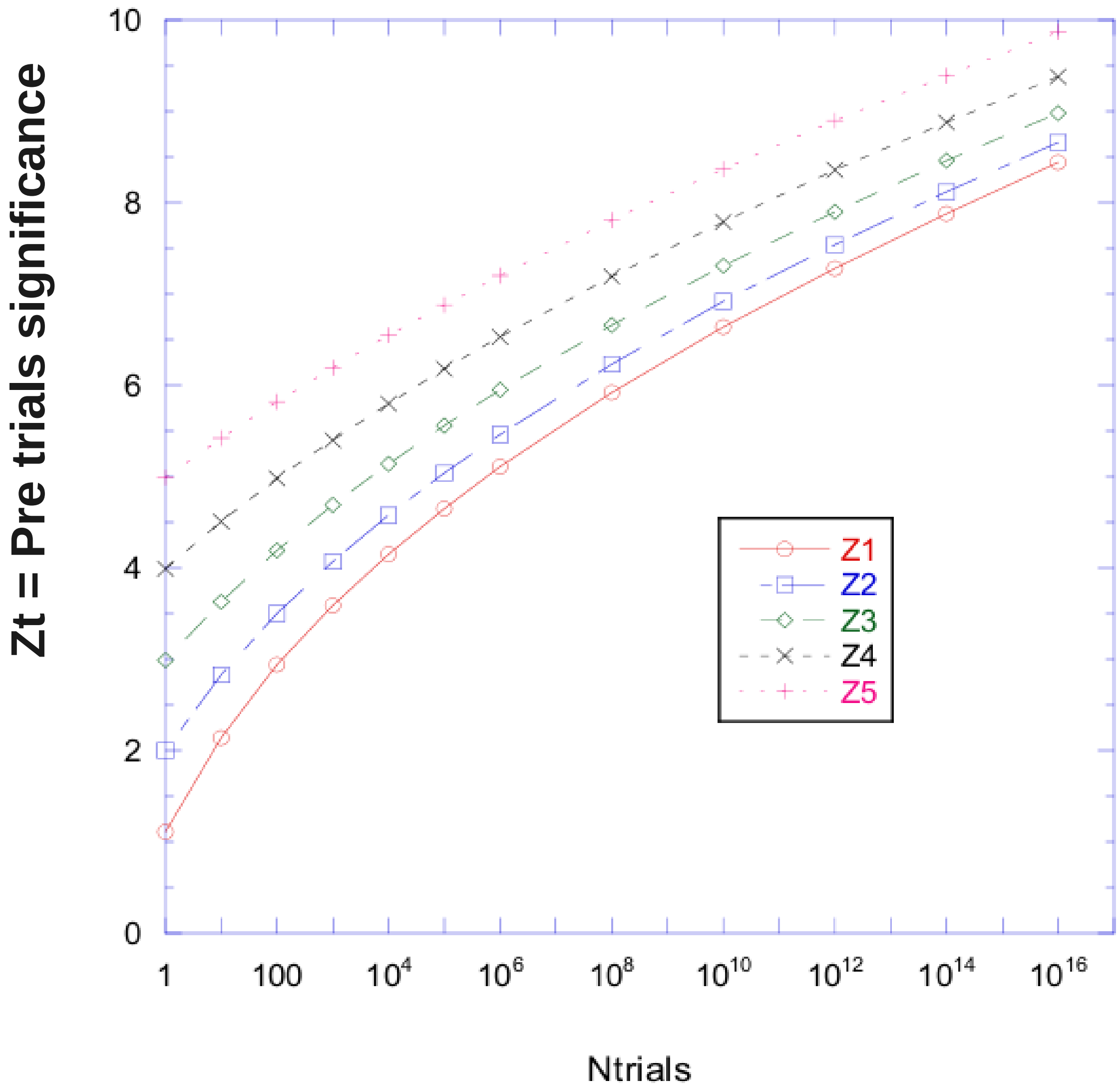
4a) Trials Factors

- If you observe 400 random points in space and see one at 3-σ (99.73% CL), should you get excited?
  - No, you'd expect ~1 3-σ background fluctuation for each ~370 points (0.27/100*370=1)
- The mathematical way to account for this is a “trials factor” which reduces the significance based on the probability of the fluctuation coming from background when you try multiple points at random

When You Need a Trials Factor

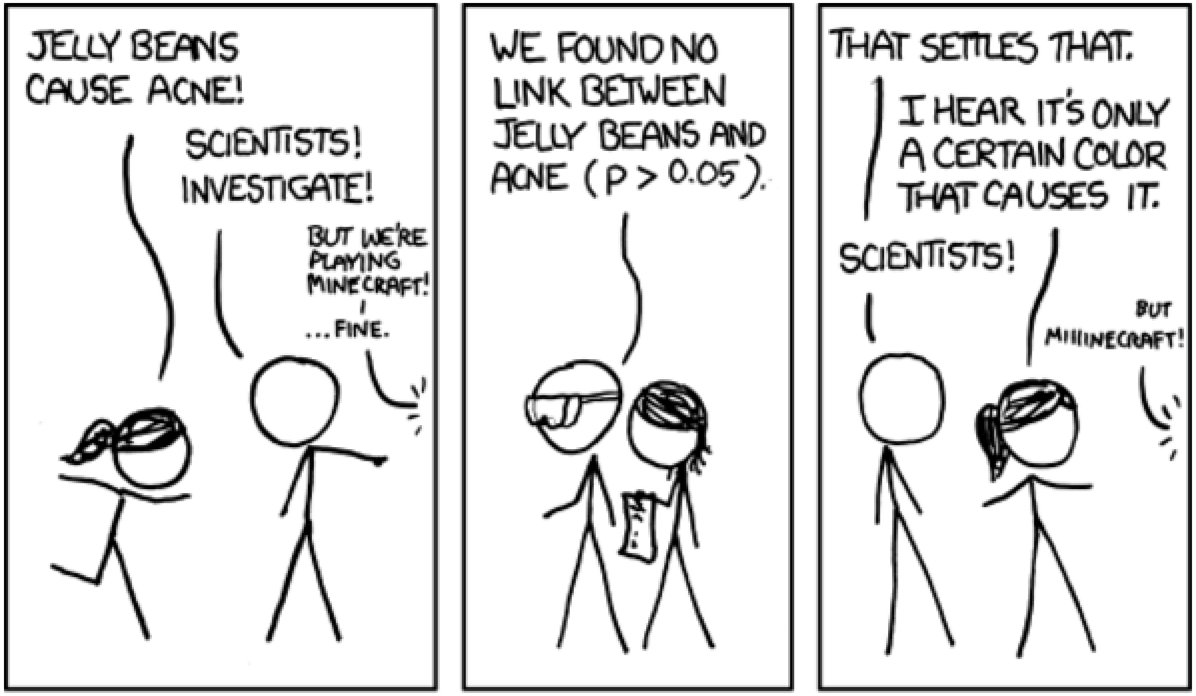
- You need a trials factor if you are:
  - doing an unbiased search of multiple locations looking for any sources
  - looking at a single source location with multiple spectral assumptions
  - looking at a single source but letting the location float
- If you are looking for a source with a fixed, known location and spectrum (e.g. the Crab), then you do not need trials. That's about it.

How Trials Hurt Significance

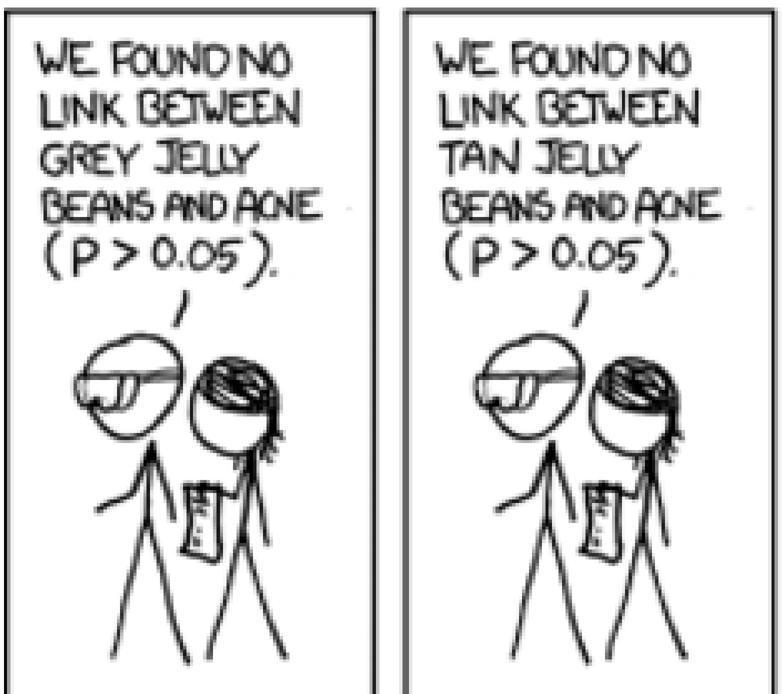


- Small effect if large significance, small number of trials
- Z1-Z5 are a “post-trials significance” of 1-5
- Note: Definition based on Bonferonni Method

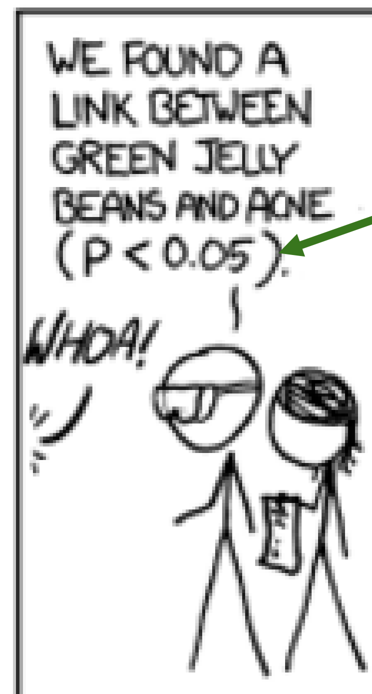
Figure courtesy of J. Linnemann



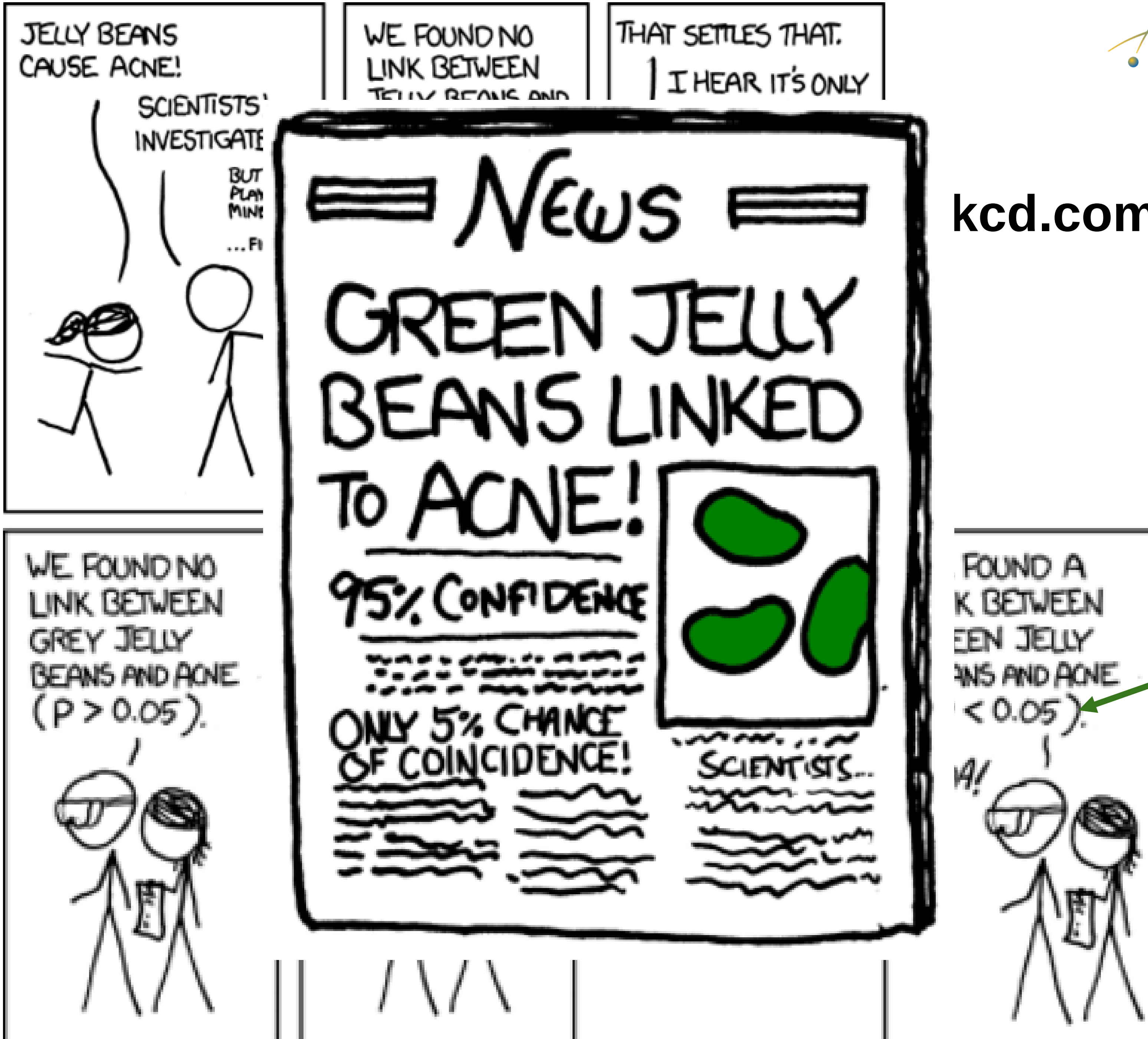
xkcd.com/882



17 colors later...



2σ



kcd.com/882

2σ

4b) Uncertainty in the Background

- So far, we have assumed that the background is perfectly measured and has no uncertainty
- In truth, we get our background from data, and data has uncertainties
  - Lots of time spent looking at nothingto quantify backgrounds
- Example: the background N_B is oversampled, from a large data set N_{off} of size $N_B=\alpha N_{off}$

Li & Ma Single-bin Analysis

- Li & Ma (1983) treats the single-bin version of an uncertain background in the case of large number of statistics
- In that case, the significance is:

$$\sigma = \sqrt{2}\left\{ N_{on} \ln\left[\frac{1+\alpha}{\alpha}\left(\frac{N_{on}}{N_{on}+N_{off}}\right)\right] + N_{off} \ln\left[(1+\alpha)\left(\frac{N_{off}}{N_{on}+N_{off}}\right)\right]\right\}^{1/2}$$

- Used in many particle and astrophysics calculations
- Available at adsabs.harvard.edu/full/1983ApJ...272..317L

Including Uncertainty in the Background

- To completely include uncertainty in the background, you need to include a model of your full oversampled background in the calculation
- For a true number of background events b, we need to know not just the probability of observing at least N_{on} events given αb but also know the probability of the unknown quantity b given our measured number of total background events N_{off} (Alexandreas et al 1993)

$$P(\geq N_{on};\alpha,N_{off}) = \sum_{n_{on}=N_{on}}^{\infty} \int_{b=0}^{\infty} db\, P(N_{off};b)\, P(n_{on};\alpha b)$$

or an equivalent analytic expression therein

4c) Tools for Maximizing Likelihood

- Analytic
- Grid
- Minuit
- Markov-Chain Monte Carlo

Analytic Maximization

- Calculus for the win!
- Solving $\partial\mathcal{L}/\partial p=0$ (or $\nabla_p\mathcal{L}=0$) analytically makes the maximization quick
  - Sometimes, you can maximize w.r.t. p_1 analytically but need to do p_2 numerically
  - Sometimes, you need to re-parameterize in terms of $p_1'=f(p_1,p_2)$ to analytically do it
- **Warning:** $\frac{\partial}{\partial p}\sum_k ln(P_k(p))=\sum_k \frac{\partial}{\partial p} ln(P_k(p))=0$ does not mean $\frac{\partial}{\partial p} ln(P_k(p))=0$

Grid Search

- Try a few sample points in each parameter to find the set of parameters with the maximum likelihood
- Can be computationally very expensive
  - For 3 parameters, 100 sample points each you need to calculate the likelihood 1,000,000 times
- If parameter space has peaks and valleys, easy to miss global maximum

Minuit

- Fitting tool available in ROOT software
- Typically use MIGRAD routine
  - Uses derivatives to find its way through parameter space to maximum likelihood
- Can be fast, even for a few parameters
- Still need to be careful about peaks and valleys in parameter space giving you false best-fits

Markov-Chain Monte Carlo (MCMC)

- One useful method to get the full distribution of *TS* (or anything) for a model is MCMC
- You basically wander around in parameter space from regions of lower probability to regions of higher probability, stopping when you find the maximum
- Using the Metropolis-Hastings algorithm, you don't always move to the higher-probability point
  - You stay where you are sometime, based on the ratio of P(your location)/P(new location)
- Doing this, you map out the full probability space while simultaneously finding the maximum

For More Info

- Particle Data Group Statistics and Probability
  - pdg.lbl.gov/2015/reviews/rpp2015-rev-statistics.pdf
- Last Year's Lecture (by Liz Hayes)
  - confluence.slac.stanford.edu/display/LSP/Fermi+Summer+School+2017