Statistical Methods. Lecture 1

Claudio Campagnari
UCSB

U.S. DEPARTMENT OF ENERGY | Office of Science

UCSB

E. Rutherford:

If your experiment needs statistics, you ought to have done a better experiment.

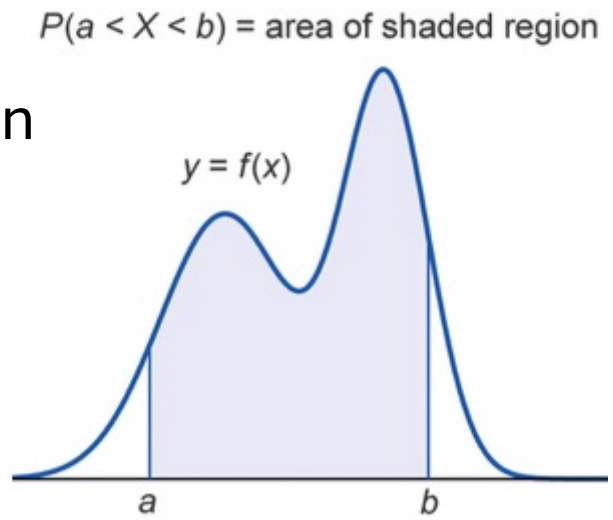
A paper says that the cross-section for some process is

$$\sigma = 100 \pm 10 \text{ nb}$$

What does it mean?

Bayesian:

- *Given info from experiment, the probability that the true value of σ is between 90 nb and 110 nb is 68.3%*
  - 68.3% is the $\pm 1\sigma$ content of the Normal or Gaussian distribution
  - Central **credible** interval
  - Bayesian analysis gives probability (pdf) of the truth.



Frequentist:

- *If I was to repeat the experiment an infinite number of times, I would get a distribution of experimental results that depends on the true value of σ. For true values between 90 and 110 nb, my experimental result of 100 nb would be within the central 68.3% of the distribution of results.*
  - 68.3%, central **confidence** interval
  - Probability of the truth does not enter
  - From frequentist "confidence belt"

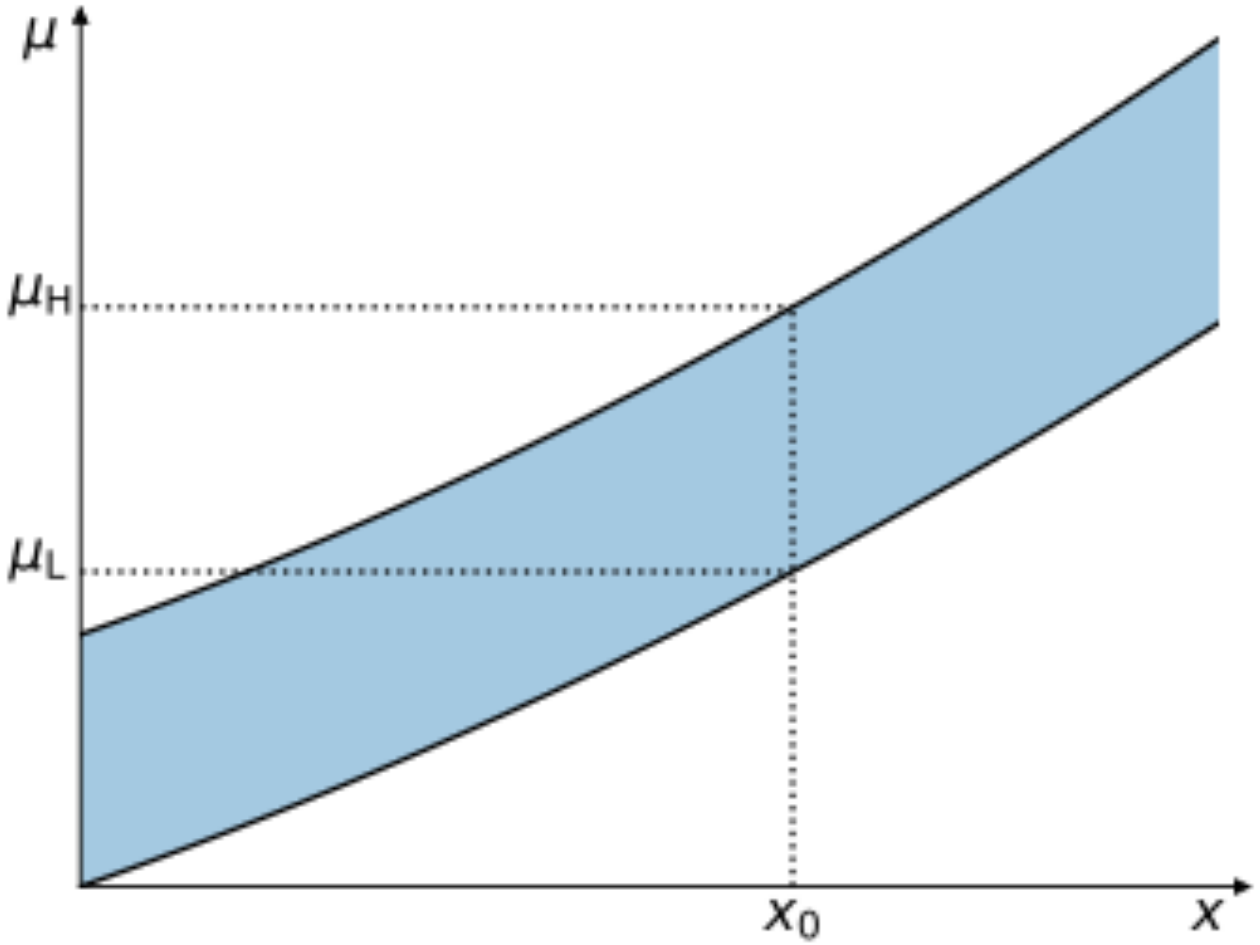
Confidence Belt

- x = possible experimental results
- x_0= result of this experiment

- μ = true value

At each μ, the shaded region "covers" the 68.3% central distribution of possible experimental outcomes.

$\mu_L < \mu < \mu_H$ is the 68.3% central confidence interval



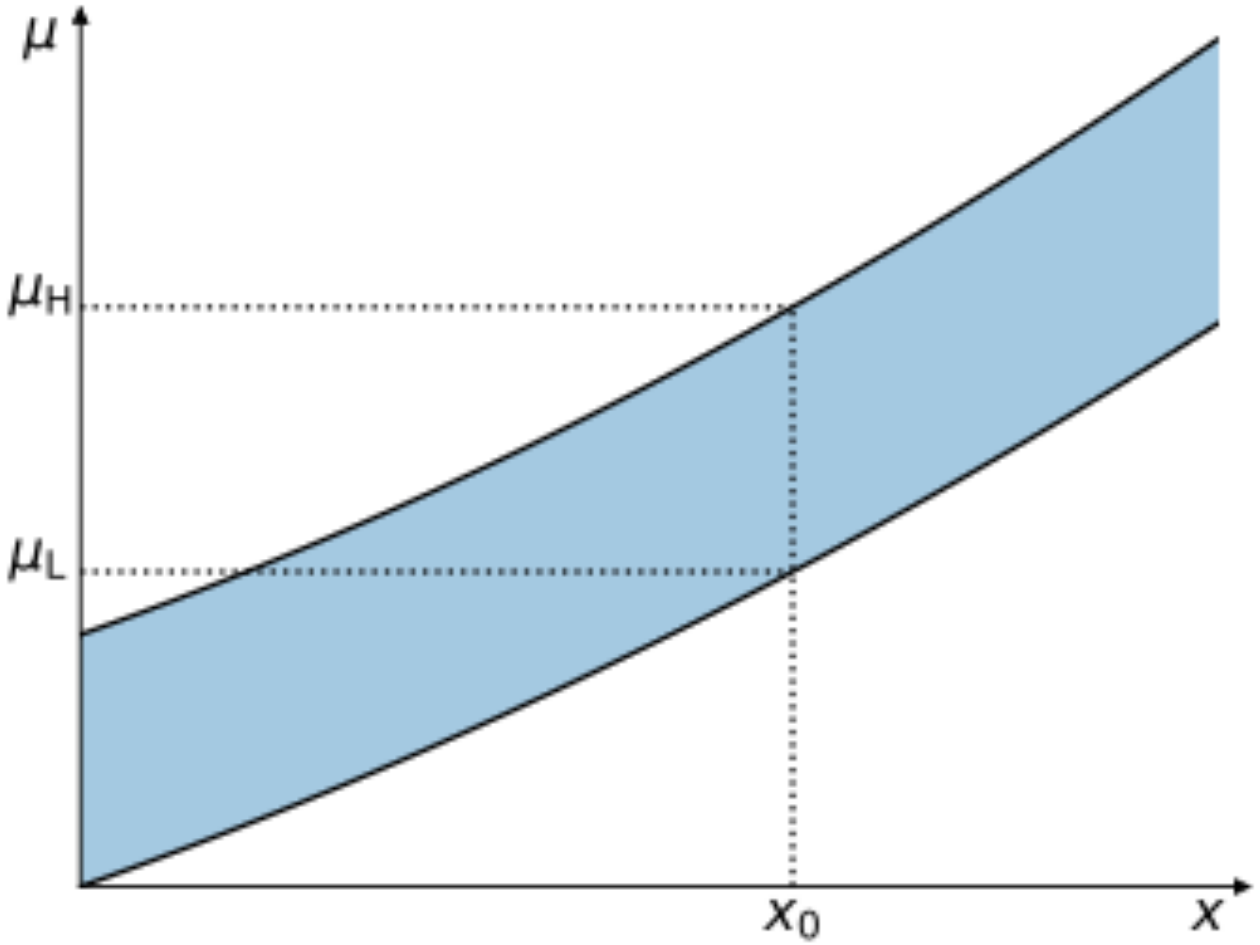
Confidence Belt

- x = possible experimental results
- x_0= result of this experiment

- μ = true value

At each m, the shaded region "covers" the 68.3% central distribution of possible experimental outcomes.

$\mu_L < \mu < \mu_H$ is the 68.3% central confidence interval



Of course it goes without saying that that as input you need p(x|μ), i.e., the probability of x for a given μ.

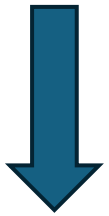
In the same way you can get 90%, 95%, … central confidence intervals as well as upper/lower confidence limits

Bayes Theorem

$$P(A \text{ and } B) = p(B|A) \cdot p(A)$$

$$P(A \text{ and } B) = p(A|B) \cdot p(B)$$

$$\boxed{p(B|A) = \frac{p(A|B) \cdot p(B)}{p(A)}}$$



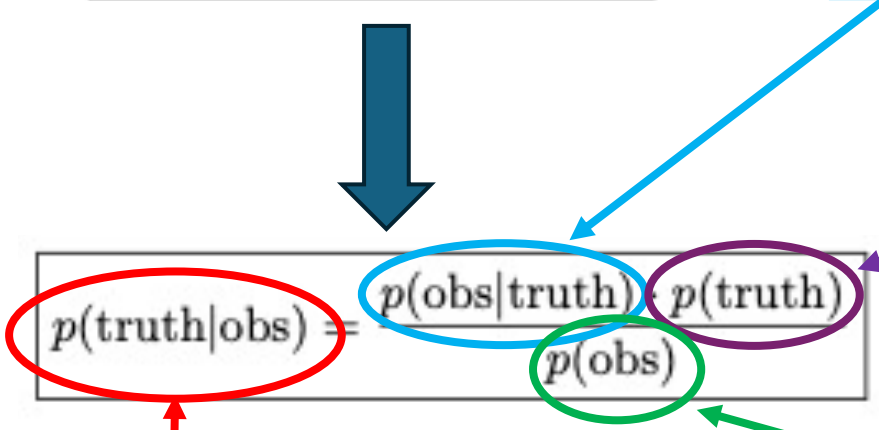
$$\boxed{p(\text{truth}|\text{obs}) = \frac{p(\text{obs}|\text{truth}) \cdot p(\text{truth})}{p(\text{obs})}}$$

Bayes Theorem

$$P(A \text{ and } B) = p(B|A) \cdot p(A)$$

$$P(A \text{ and } B) = p(A|B) \cdot p(B)$$

$$p(B|A) = \frac{p(A|B) \cdot p(B)}{p(A)}$$



$$p(\text{truth}|\text{obs}) = \frac{p(\text{obs}|\text{truth}) \cdot p(\text{truth})}{p(\text{obs})}$$

Comes from our understanding of the experiment.
The same info that is needed in a frequentist analysis

Input to the analysis.
The arbitrary "prior"

????????????
After the experiment is done, just a number
To be reabsorbed in the normalization of the posterior

What we want
The "posterior"

Louis Lyons

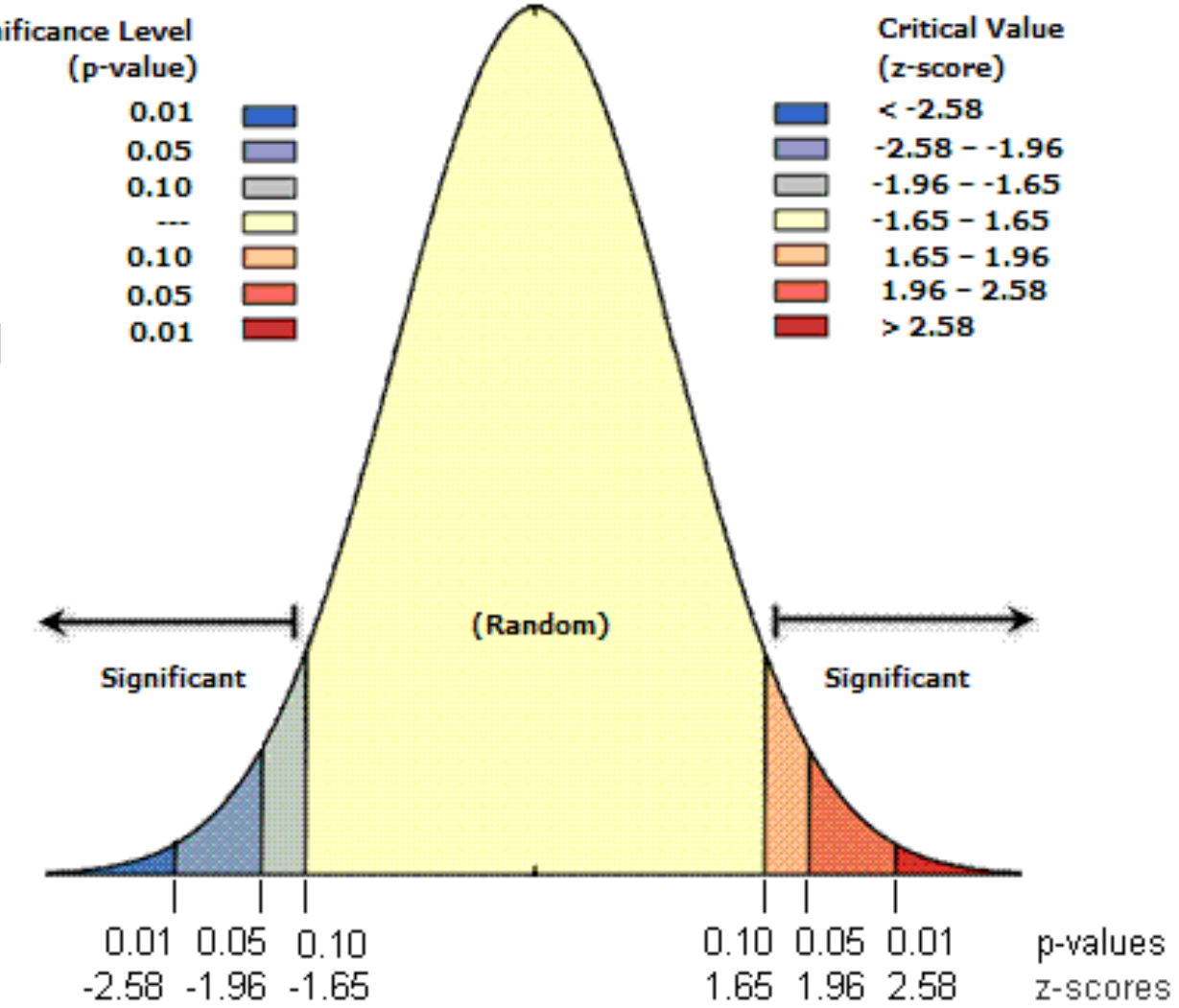
Bayesians address the question everyone is interested in using assumptions that no one believes.

Frequentists use impeccable logic to deal with an issue of no interest to anyone.

p-value, Z-score, Significance

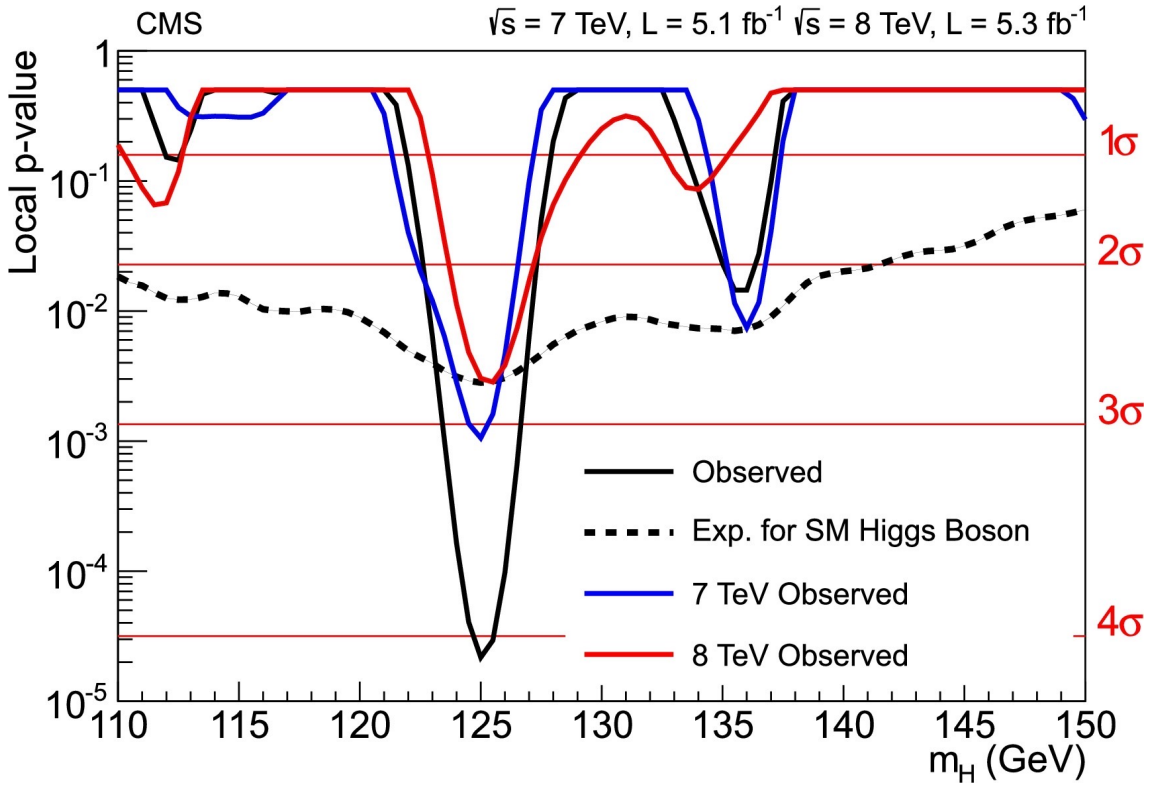
- Used in hypothesis testing
- Probability that under a null-hypothesis (e.g., the Standard Model) we would get a result as extreme or more extreme than the one we have.
  - e.g. for a counting experiment we expect N_{SM} events and we observe N. What is the probability under the SM hypothesis of observing at least N?
  - **A p-value is not the probability that the null-hypothesis is wrong**
- Careful: what does "extreme" mean.
  - In counting experiment, just $N >> N_{SM}$ or also $N << N_{SM}$???
    - In more complicated experiments?? We'll come back to that later.
  - One-sided vs. two-sided p-values. Excesses and deficits.
  - I think deficits are just as interesting as excesses (in a different way…)
- p-values can be turned into Z-scores (significances in number of sigmas) based on the integral of the Normal distribution
- By construction, p-values are such that repeating the experiment many times you expect a uniform distribution between 0 and 1.

These are two-sided p-values

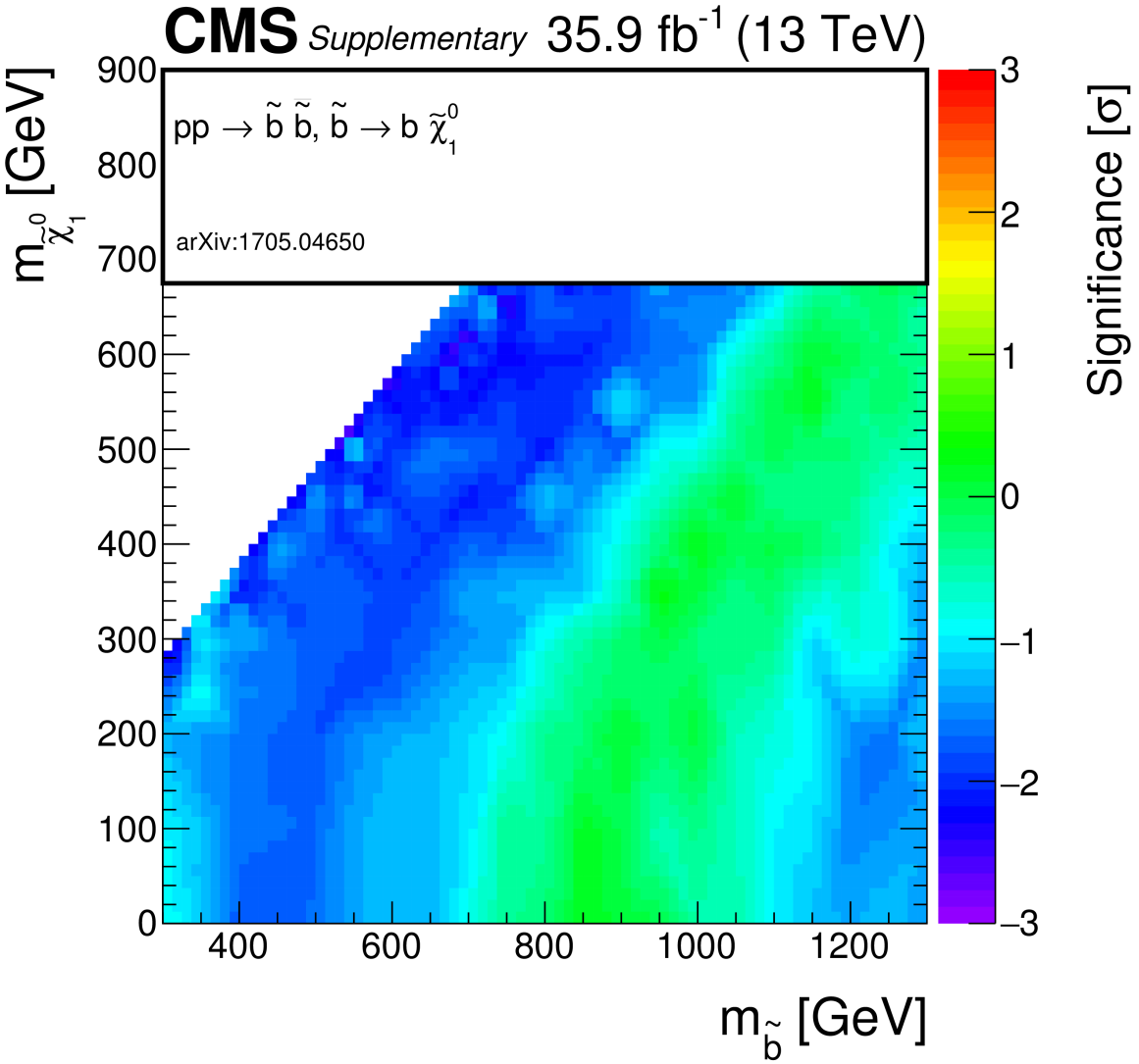


http://tinyurl.com/49kambrw

Examples of presenting both deficits and excesses, or only excesses



https://doi.org/10.1016/j.physletb.2012.08.021



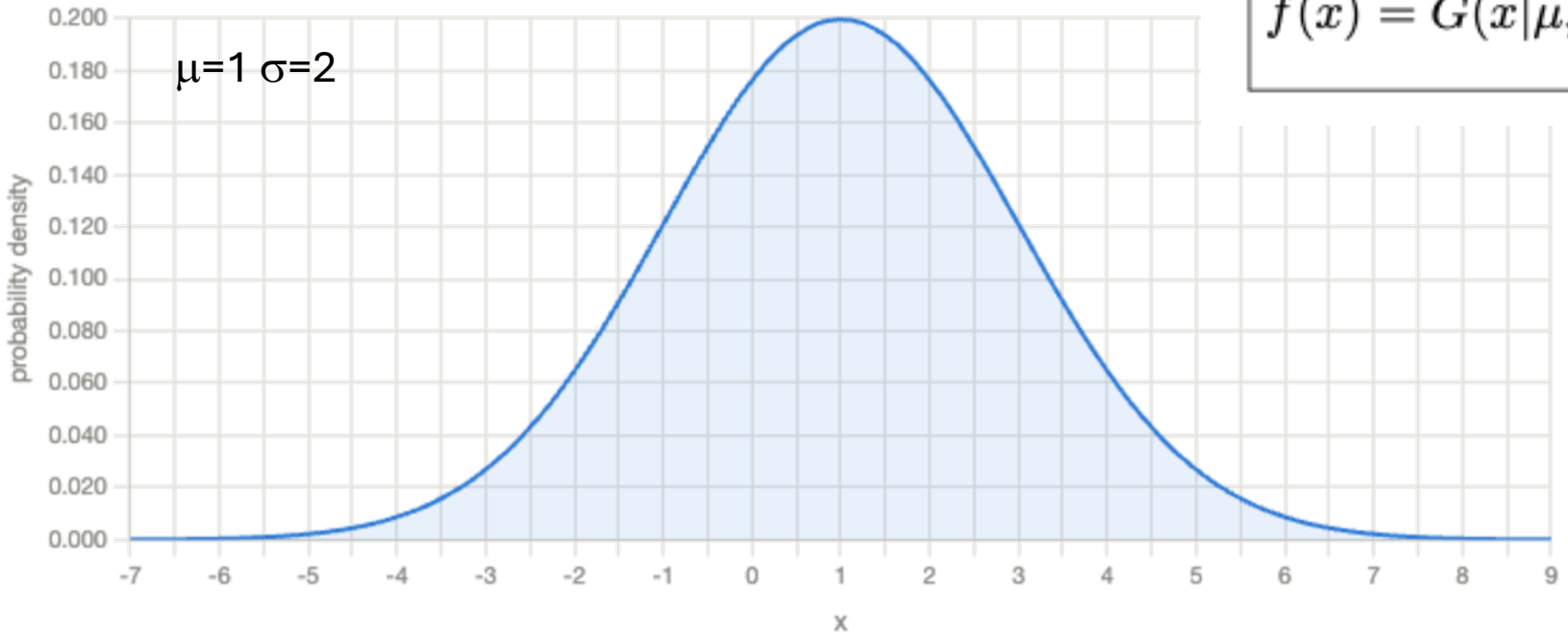
https://link.springer.com/article/10.1140/epjc/s10052-017-5267-x

Probability Distribution (Mass) Functions: PDF and PMFs

HEP makes heavy use of special PDFs and PMFs. I recommend that you look up and familiarize yourself with all of these

1. Gaussian/Normal
2. Multi-variate Gaussian
3. Double-Gaussian
4. Poisson
5. χ^2
6. Lognormal
7. Error Function and Complementary Error Function
8. Binomial
9. Gamma
10. Beta
11. (Double) Crystal Ball
12. Breit-Wigner (relativistic and non-relativistic), Voigtian
13. Cruijff
14. Landau
15. Argus

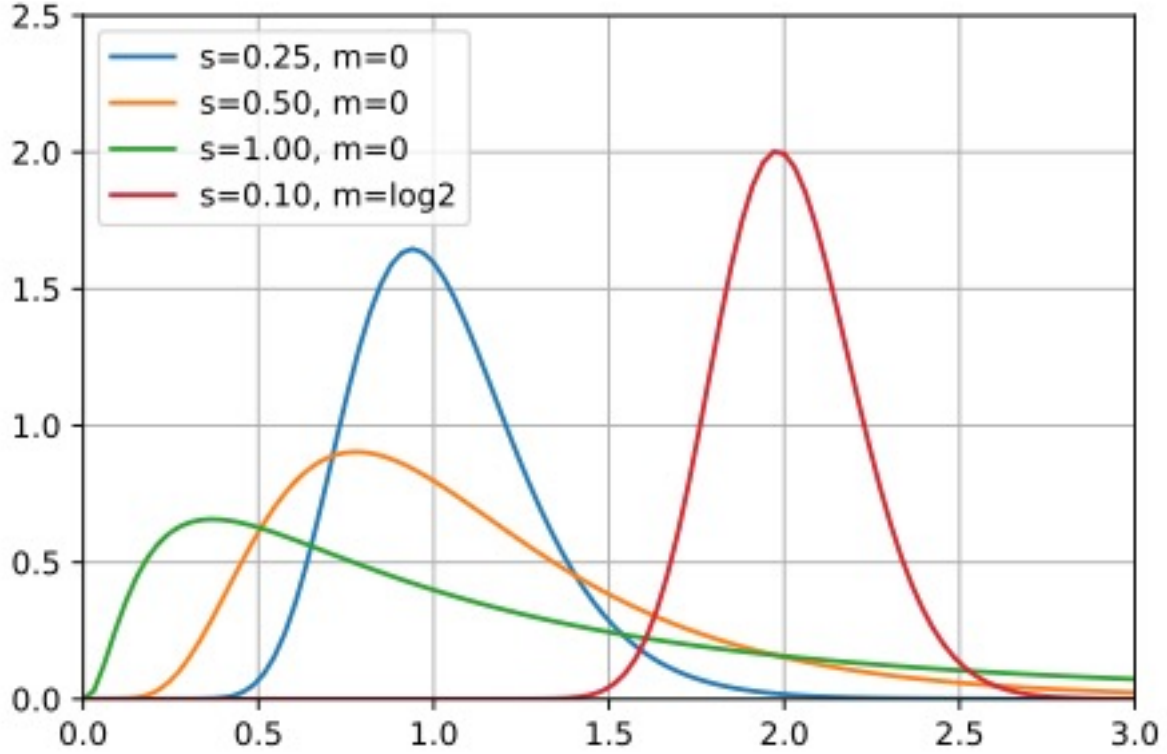
Gaussian/Normal



$$f(x) = G(x|\mu,\sigma) \equiv \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Normal has $\mu=0$ $\sigma=1$
- Central Limit Theorem: the **sum** of many random variable is asymptotically a Gaussian of mean equal to the sum of means and variance equal to the sum of variances

Lognormal



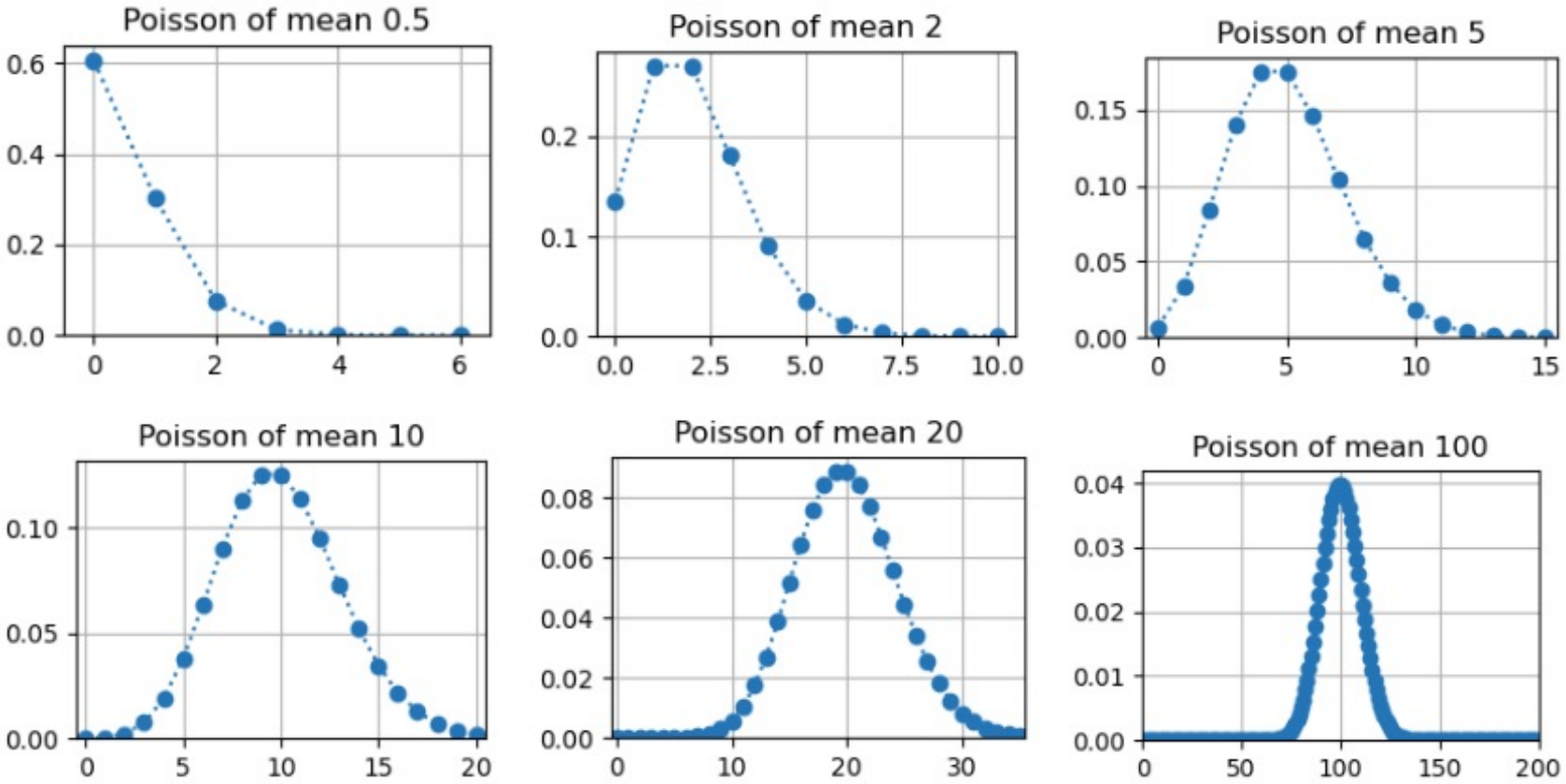
$$p(y) \equiv \frac{1}{\sqrt{2\pi}s}\ \frac{1}{y}\ \exp(-\frac{(\log y - m)^2}{2s^2})$$

Mean (μ)	$\exp(m + s^2/2)$
Variance (σ^2)	$(\exp(s^2) - 1)\exp(2m + s^2)$
Median	$\exp(m)$
Mode	$\exp(m - s^2)$
s^2	$\log\left(1 + (\frac{\sigma}{\mu})^2\right)$
$\exp(m)$	$\mu\ \exp\left(-s^2/2\right)$

- The pdf of the **product** of many random positive variables (consequence of CLT).
- Has become popular in HEP analyses because
  1. It is approximately Gaussian for small s
  2. It goes smoothly to zero as y→0, which is good to represent PDFs of positive quantities
  3. It can be argued that in some HEP cases it is more justified than Gaussian
     - e.g. the efficiency of electron ID is product of many efficiencies of the components of the algorithm

Poisson PMF: counting statistics

$$p(N|\mu) = \frac{\mu^N e^{-\mu}}{N!}$$



- Fun (or not-so-fun) fact: it gain notoriety in 1898 when it was shown to reproduce the number of deaths of Prussian army soldiers from horse kicks
- Mean = μ Variance = μ
- Approaches Gaussian at large μ.
  - It is a PMF not a PDF!!

Poisson Limits. No Background, no systematics yet.

$$p(\text{truth}|\text{obs}) = \frac{p(\text{obs}|\text{truth}) \cdot p(\text{truth})}{p(\text{obs})}$$

$$p(N|\mu) = \frac{\mu^N e^{-\mu}}{N!}$$

- The truth is μ
- The observation is N

I write $p(\mu) = \pi(\mu)$ to denote the prior

$$p(\mu|N) = p(N|\mu)\frac{\pi(\mu)}{p(N)}$$

The pdf for μ (the posterior) is

$$f(\mu) = A \cdot \mu^N e^{-\mu} \cdot \pi(\mu)$$

A is a normalization factor

The answer depends on the prior

Bayesian

Poisson Limits. No Background, no systematics yet.

$$p(\text{truth}|\text{obs}) = \frac{p(\text{obs}|\text{truth}) \cdot p(\text{truth})}{p(\text{obs})}$$

$$p(N|\mu) = \frac{\mu^N e^{-\mu}}{N!}$$

- The truth is μ
- The observation is N

I write $p(\mu) = \pi(\mu)$ to denote the prior

$$p(\mu|N) = p(N|\mu)\frac{\pi(\mu)}{p(N)}$$

The pdf for μ (the posterior) is

$$f(\mu) = A \cdot \mu^N e^{-\mu} \cdot \pi(\mu)$$

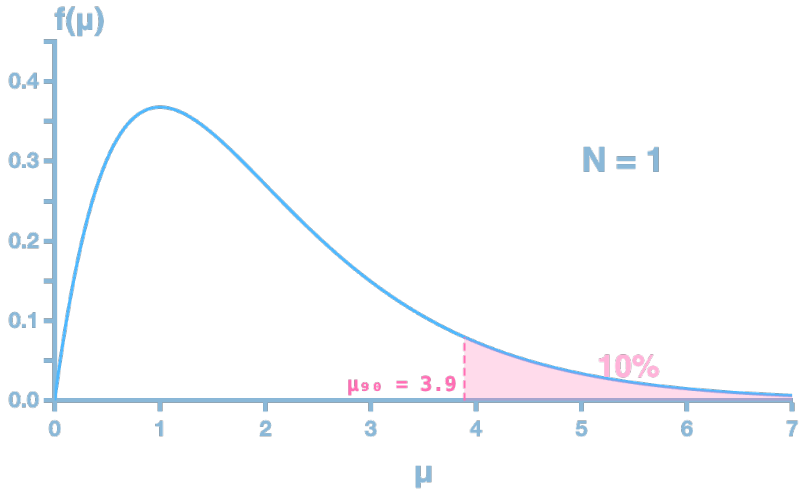
A is a normalization factor

The answer depends on the prior

Often (**arbitrary!)** people take the prior to be uniform (flat) between 0 and infinity. Then the 90% (say) CL on μ is given by

$$\frac{1}{N!}\int_{\mu_{90}}^{\infty} \mu^N e^{-\mu} d\mu = 0.1$$

$$e^{-\mu_{90}} \sum_{k=0}^{k=N} \frac{\mu_{90}^k}{k!} = 0.1$$



Bayesian

Poisson Limits. No Background, no systematics yet.

For a 90% limit we find μ_{90} such that:

- The truth is μ
- The observation is N

$$p(\text{counts} \leq N|\mu_{90}) = 1 - 0.9 = 0.1$$

Which gives:

$$e^{-\mu_{90}} \sum_{k=0}^{k=N} \frac{\mu_{90}^k}{k!} = 0.1$$

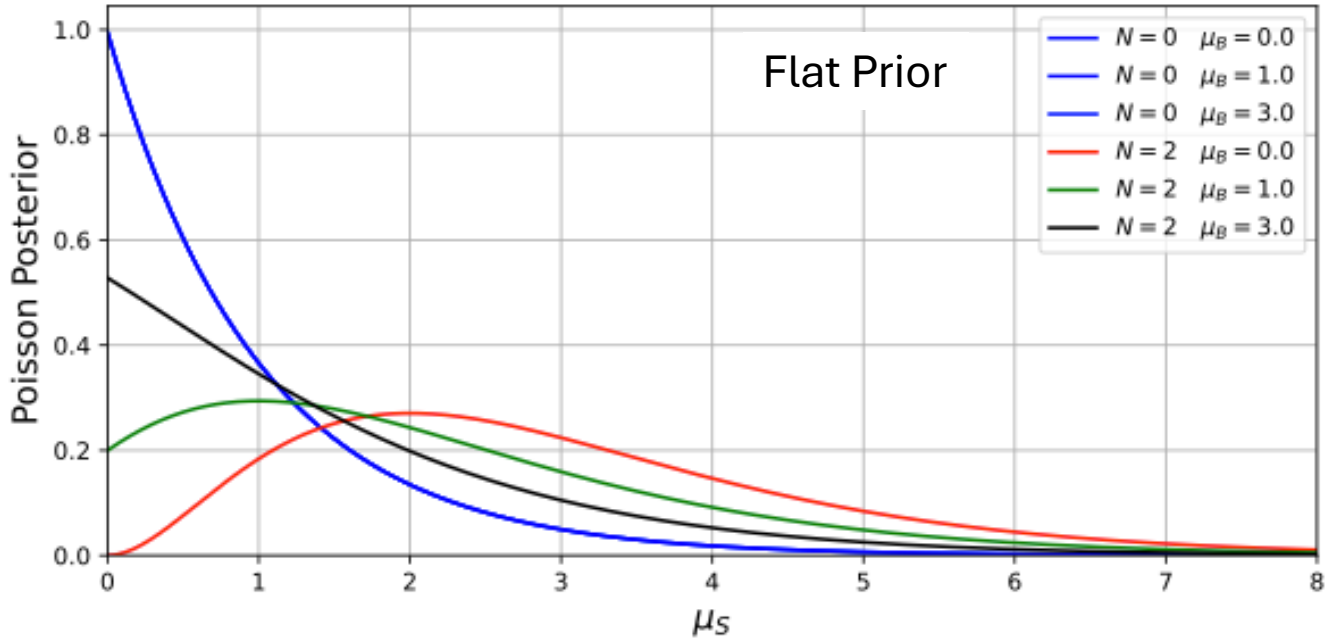
For no background and no systematic uncertainties Bayesian Poisson limits **with a flat prior** and Frequentist limits are exactly the same

Frequentist

Poisson limits with background. No systematics

$$p(N|\mu_S,\mu_B) = \frac{(\mu_S+\mu_B)^N e^{-(\mu_S+\mu_B)}}{N!}$$

$$p(\mu_S|N,\mu_B) = f(\mu_S) = C\cdot(\mu_S+\mu_B)^N e^{-(\mu_S+\mu_B)}\cdot\pi(\mu_S)$$



Note: for N=0 pdfs and therefore limits independent of background (!)

Bayesian

Poisson limits with background. No systematics

$$p(N|\mu_S,\mu_B) = \frac{(\mu_S+\mu_B)^N e^{-(\mu_S+\mu_B)}}{N!}$$

For 95% (say) limits we solve

$$\sum_{k=0}^{k=N} e^{-(\mu_S^{95}+\mu_B)} \frac{(\mu_S^{95}+\mu_B)^k}{k!} = 0.05$$

So far so good….but….

Frequentist

Poisson limits with background. No systematics

$$p(N|\mu_S,\mu_B) = \frac{(\mu_S+\mu_B)^N e^{-(\mu_S+\mu_B)}}{N!}$$

For 95% (say) limits we solve

$$\sum_{k=0}^{k=N} e^{-(\mu_S^{95}+\mu_B)} \frac{(\mu_S^{95}+\mu_B)^k}{k!} = 0.05$$

So far so good….but….

Imagine N=0. The solution is

$$e^{-(\mu_S^{95}+\mu_B)} = 0.05$$
$$\mu_S^{95} + \mu_B = -\log(0.05) = 2.996$$
$$\mu_S^{95} = 2.996 - \mu_B$$

What if μ_B=3. Or larger.

N=0 with μ_B=3 in the absence of a signal happens 4.97% of the times. Not very rare. Then the solution is negative.

It means that at 95% CL the experiment excludes any signal, no matter how small.
Formally correct, but "a problem!"

Frequentist

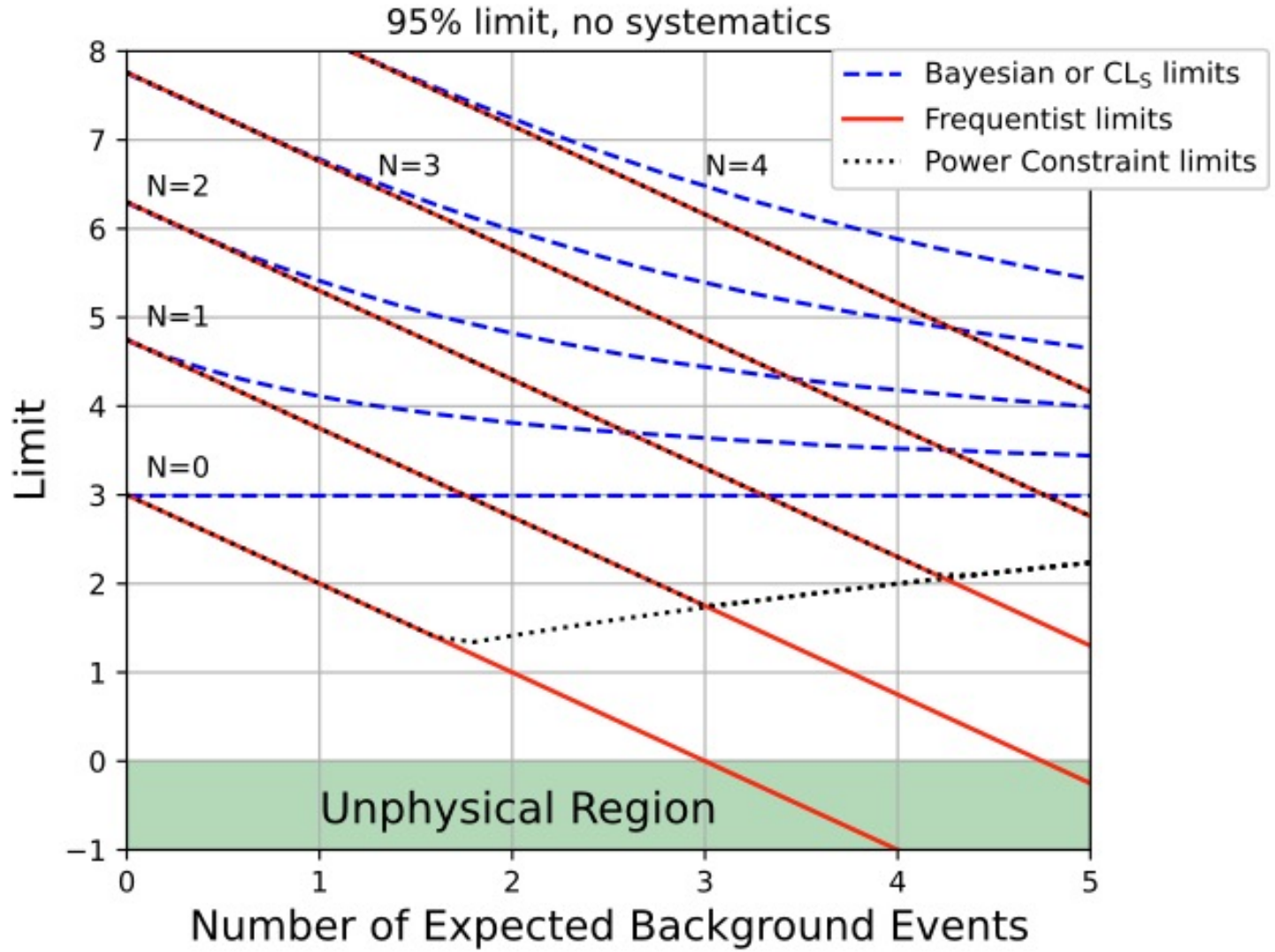
Two solutions for frequentist limits going "crazy"

In either case not strictly correct frequentist limits at the level claimed

1. **Power Constrained**: artificially not allow limit to get too strong
   - e.g.: never stronger than $\mu_B - \sqrt{\mu_B}$

2. **CLs**: freq. limit comes from the p-value of the signal. Do not allow limit to get too small by taking the ratio of p-values for the signal+background and the background-only hypotheses
   - In the absence of systematic uncertainties on the efficiencies, CLs limits are identical to Bayesian with a flat prior. https://arxiv.org/pdf/1404.1340.pdf

There is an additional method on the market (Feldman-Cousins) which we will cover tomorrow


95% limit, no systematics
Bayesian or CL$_S$ limits
Frequentist limits
Power Constraint limits
N=0
N=1
N=2
N=3
N=4
Unphysical Region
Limit
Number of Expected Background Events
8
7
6
5
4
3
2
1
0
−1
0
1
2
3
4
5

What about systematics, aka “nuisances”

Troubled waters……..

- Systematics, especially theo uncertainties, but also expt uncertainties, are mostly judgement calls.
- In the old days we would (mostly? often?) repeat the analysis changing analysis parameters within reasonable assumptions and then make another judgement call at the end as to the overall effect.
  - We would also concentrate on the most important effects (another judgement call).
- Now we assign pdfs (Lognormal? Gaussian?). And we feed them into complicated mathematical machineries.
  - The curse of too much computing power: “stick every effect in, no matter how small”
  - Also, BTW, pdfs of nuisances are pdfs of the true value of the nuisances, and the probability of the truth is not even a frequentist concept
    - When I made these objections, I was told to pretend that these are not pdfs but likelihoods of “auxiliary experiments” (which is complete nonsense for theoretical uncertainties)
- Keep this in mind. If your result is strongly dependent on the systematics… remember what Lord Rutherford said.
- Nevertheless, let’s proceed ☺

Aside: when I was a graduate student……

- My PhD thesis was on a search for a BSM kaon decay
- N=0 → published Bayesian limit on branching ratio in Phys. Rev. Lett.
- Did not include systematics. Asked my “elders” why
- They said “Do not worry”
  - Limits are squishy things, not precision measurements
  - Systematics have little effect in Poisson regime
- Did not bother estimating expected background for the paper
  - I did do it later “for fun” when writing my thesis. Was ~ 0.7 events.
- Nobody cared at the time about ”expected limit”, only “observed”

We have come a long way since….

Bayesian Nuisances

$$p(\mu, \vec{\theta}|\vec{x}) \propto p(\vec{x}|\mu, \vec{\theta})\, \pi(\mu)\, \pi(\vec{\theta})$$

- $\vec{x}$ is the data
- $\vec{\theta}$ is the set of nuisances
- μ is the truth we care about
- $\pi(\mu)$ is the prior on the truth
- $\pi(\vec{\theta})$ is the prior on $\vec{\theta}$, i.e. the $\vec{\theta}$ pdf

- What we care about is $p(\mu|\vec{x})$
- The $\vec{\theta}$ dependence is eliminated by averaging

$$p(\mu|\vec{x}) = C \int p(\vec{x}|\mu, \vec{\theta})\, \pi(\mu)\, \pi(\vec{\theta})\, d\vec{\theta}$$

Marginalization

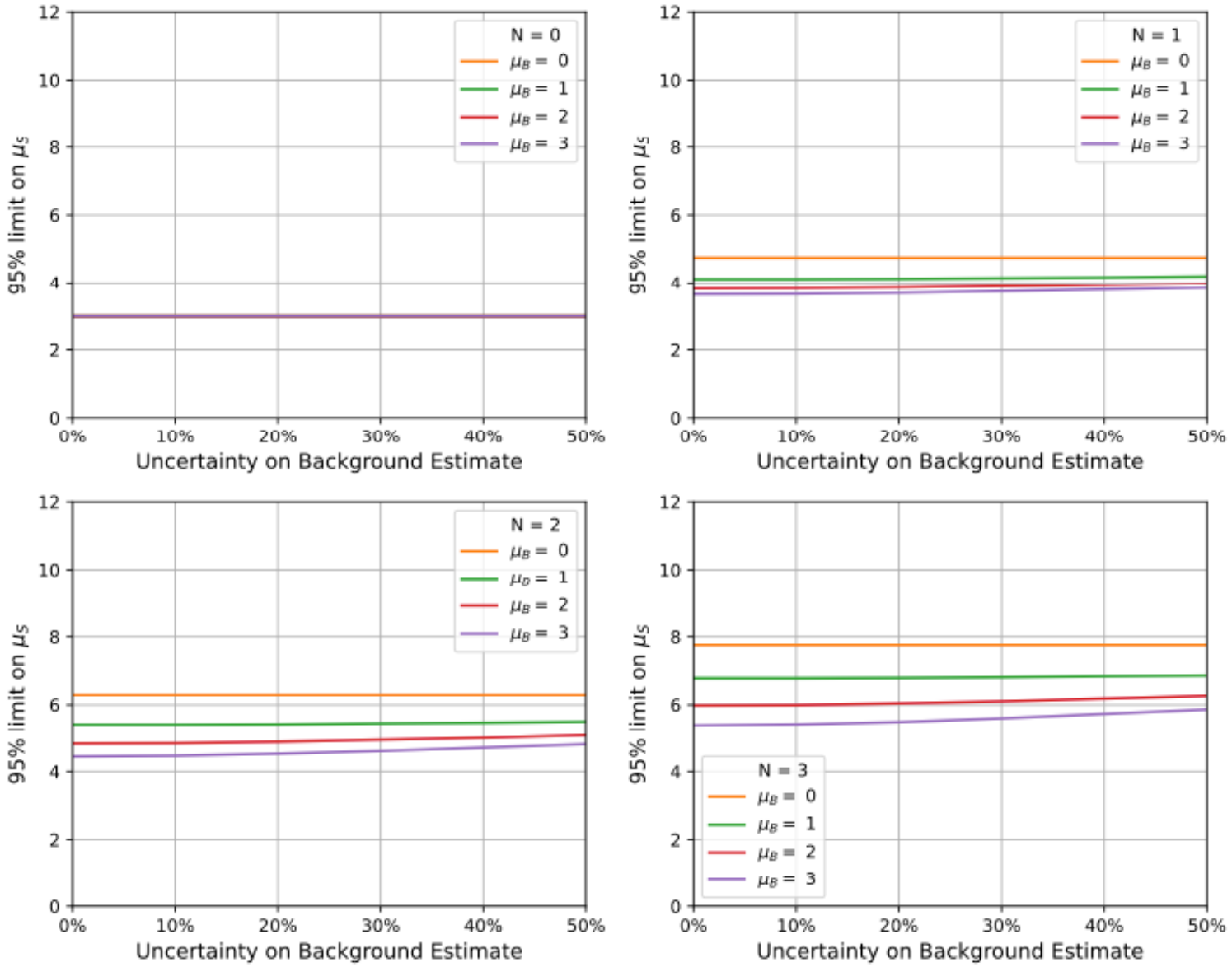
- Conceptually straightforward but integration can be “hard”
  - Markov-Chain integration methods help

Frequentist Nuisances

- In principle, should build multi-dimensional belt
- Not practical. Interpretation even more obscure.
- One approach is to build a “simple” confidence belt averaging over $\vec{\theta}$
  - Hybrid-frequentist/Bayesian method. https://tinyurl.com/csy7wenn
- The other approach is the profile-likelihood method
  - We will come to that later

Systematic effects on upper limits in the Poisson regime (1)

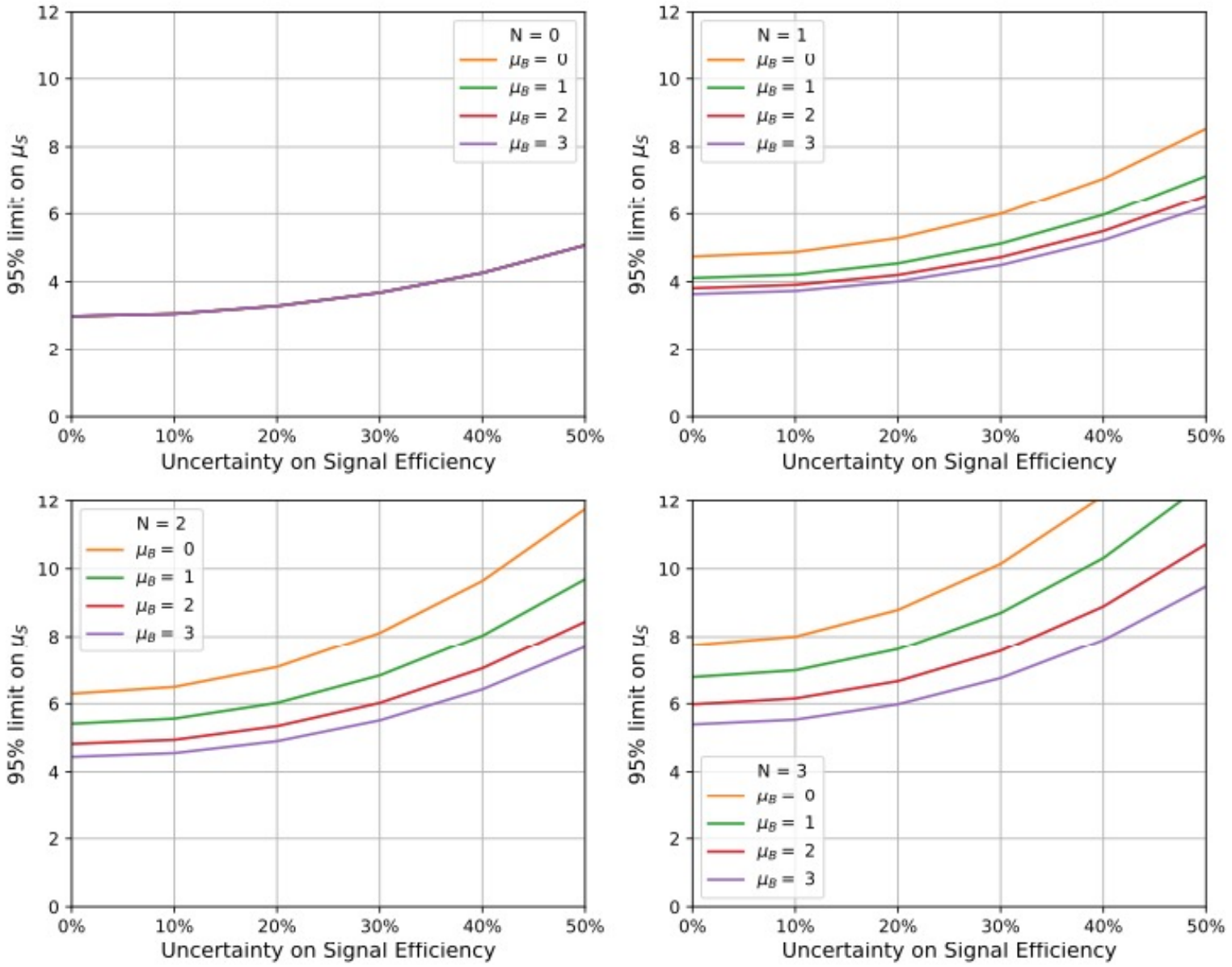
Bayesian with flat prior



Background uncertainty does not matter very much!

Systematic effects on upper limits in the Poisson regime (2)

Bayesian with flat prior



Signal efficiency uncertainty starts to matter only when it is $\gtrsim$ 20%

Estimators

- Procedures to “estimate” from the data the parameter(s) of interest
  - Notation: $\vec{\hat{a}}$ is the estimate of the true value(s) $\vec{a}$ from a dataset $\vec{x}$

- What do we want
  1. **Consistent**: $\vec{\hat{a}} \rightarrow \vec{a}$ with many measurements (in one experiment)
  2. **Unbiased**: Expectation value of $\vec{\hat{a}}$ should be $\vec{a}$ (over many equivalent experiments)
  3. **Efficient**: The variance of $\vec{\hat{a}}$ should be as small as possible

Likelihood and Maximum Likelihood Estimator (MLE)

The likelihood is the probability of the data $\vec{x}$ as a function of the parameters of interest $\vec{a}$ (and maybe also the nuisances $\vec{\theta}$).

The ML estimate $\hat{\vec{a}}$ is the value of $\vec{a}$ that maximizes the likelihood

(or minimizes the negative-log-likelihood, NLL)

Why use likelihood?

- Makes intuitive sense ☺
- MLE (usually) **consistent**
- Asymptotically the likelihood is Gaussian and is **maximally efficient**
- However: in low stats case MLE **can be biased** (Watch out!)

Example of biased MLE

Measuring an exponential distribution

$$p(t_i|\tau) = \frac{1}{\tau} e^{-t_i/\tau}$$

$$\mathrm{NLL} = \sum \frac{t_i}{\tau} + \sum \log \tau$$

$$\mathrm{NLL} = \frac{1}{\tau} \sum t_i + N \log \tau$$

Set derivative wrt to $\tau = 0$

$$\hat{\tau} = \frac{\sum t_i}{N}$$

$$< \hat{\tau} >= \frac{< \sum t_i >}{N} = \frac{\sum < t_i >}{N} = \frac{N\tau}{N} = \tau$$

Unbiased!!!!

Example of biased MLE

Measuring an exponential distribution

$$p(t_i|\tau) = \frac{1}{\tau} e^{-t_i/\tau}$$

$$\mathrm{NLL} = \sum \frac{t_i}{\tau} + \sum \log \tau$$

$$\mathrm{NLL} = \frac{1}{\tau} \sum t_i + N \log \tau$$

Set derivative wrt to $\tau = 0$

$$\hat{\tau} = \frac{\sum t_i}{N}$$

$$< \hat{\tau} >= \frac{< \sum t_i >}{N} = \frac{\sum < t_i >}{N} = \frac{N\tau}{N} = \tau$$

Unbiased!!!!

$$p(t_i|\lambda) = \lambda e^{-\lambda t_i} \quad . \lambda = 1/\tau.$$

$$\mathrm{NLL} = \lambda \sum t_i - \sum \log \lambda$$

$$\mathrm{NLL} = \lambda \sum t_i - N \log \lambda$$

Set derivative wrt to $\lambda = 0$

$$\hat{\lambda} = \frac{N}{\sum t_i}$$

$$\langle \hat{\lambda} \rangle = \left\langle \frac{N}{\sum t_i} \right\rangle = N \left\langle \frac{1}{\sum t_i} \right\rangle = \frac{N}{N-1} \lambda$$

Biased!!!!

Important Lesson!

Check your likelihood fits for biases, e.g., by toy MC

It is not just a bug check

Asymptotic Likelihood (1 parameter, for simplicity)

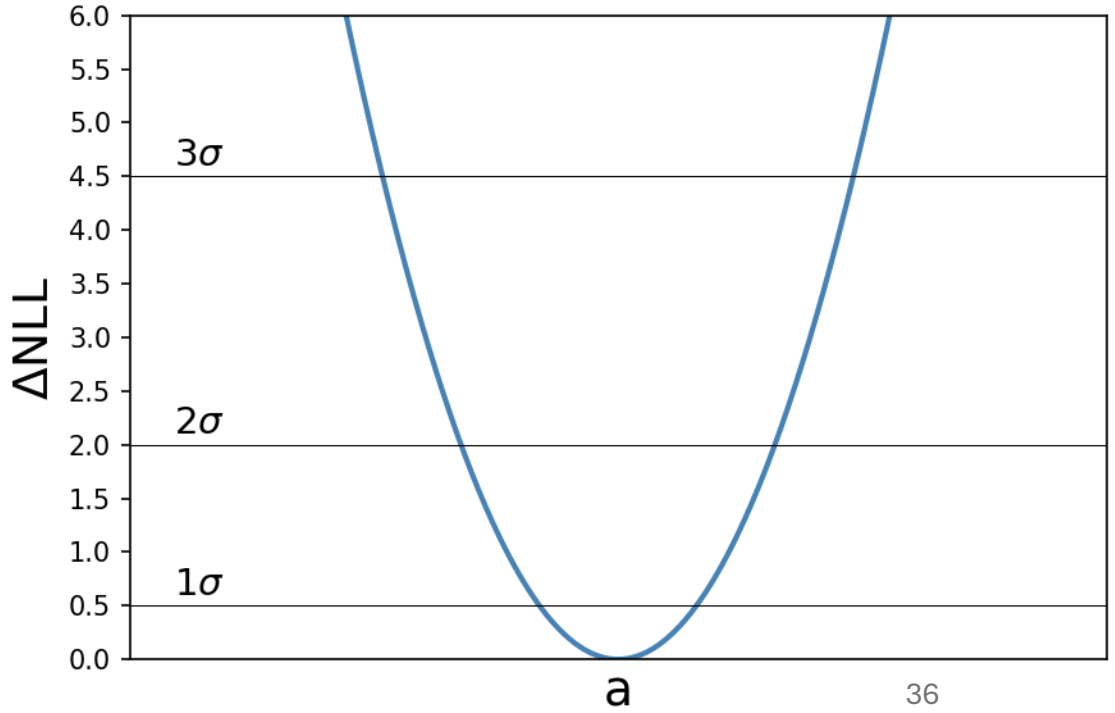
$$\mathcal{L}(a) = Ae^{-(a-\hat{a})^2/2\sigma^2}$$ **Gaussian**

$$\mathrm{NLL} = -\log \mathcal{L}(a) = \frac{(a-\hat{a})^2}{2\sigma^2} - \log A = \frac{(a-\hat{a})^2}{2\sigma^2} - \log \mathcal{L}(\hat{a})$$

$$\Delta\mathrm{NLL} = -(\log \mathcal{L}(a) - \log \mathcal{L}(\hat{a})) = \frac{(a-\hat{a})^2}{2\sigma^2}$$ **Parabola**

Crucial result: $$\sigma^2(\hat{a}) = -\frac{1}{\mathbf{E}\left(\frac{d^2 \log \mathcal{L}}{da^2}\right)}$$

Identifying the expectation value of the 2nd derivative with the 2nd derivative of the experimental result, we conclude that the 1σ, 2σ, 3σ,….can be read out from a plot of the ΔNLL



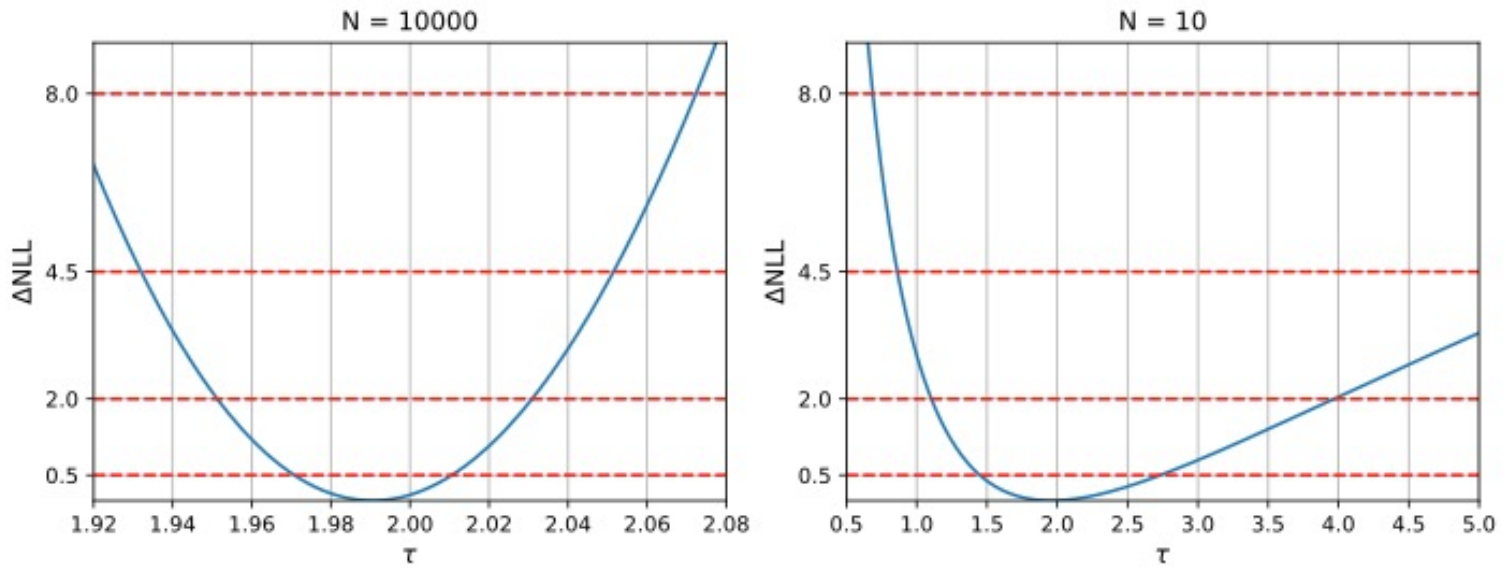
What if the ΔNLL is asymmetric (low stats)

Can still use ΔNLL(a) for intervals

Why?
Likelihood = prob of data as a function of truth (*a*)

Invariant under re-parametrization
$a \rightarrow \alpha$

Imagine re-parametrization so that
ΔNLL(α) parabolic



ΔNLL(τ) function from two Toy MC draws from PDF $p(x) \propto e^{-x/\tau}$ with $\tau = 2$. One of the two MC had $N = 10K$ draws, the other only $N = 10$.

Correspondence with χ^2

N independent measurements x_i with variance σ_i^2 and with mean that depends on $\vec{a}$

$$p(x_i|\vec{a}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp - \frac{(x_i - \mu_i(\vec{a}))^2}{2\sigma_i^2}$$

$$\text{NLL} = \sum \frac{(x_i - \mu_i(\vec{a}))^2}{2\sigma_i^2} + \sum \log \sqrt{2\pi}\sigma_i$$

$$\Delta\text{NLL} = \sum_i \frac{(x_i - \mu_i(\vec{a}))^2}{2\sigma_i^2} = \frac{1}{2}\chi^2$$

Minimizing χ^2 is the same as minimizing NLL, but careful about the factor of 2 for interpretation

Extension to more than one variable

$$\frac{1}{\sigma^2(\hat{a})} = -\mathbf{E}\left(\frac{d^2 \log \mathcal{L}}{da^2}\right) \quad \rightarrow \quad (V_{ij})^{-1} = -\mathbf{E}\left(\frac{\partial^2 \log \mathcal{L}}{\partial a_i \partial a_j}\right)$$

PDG

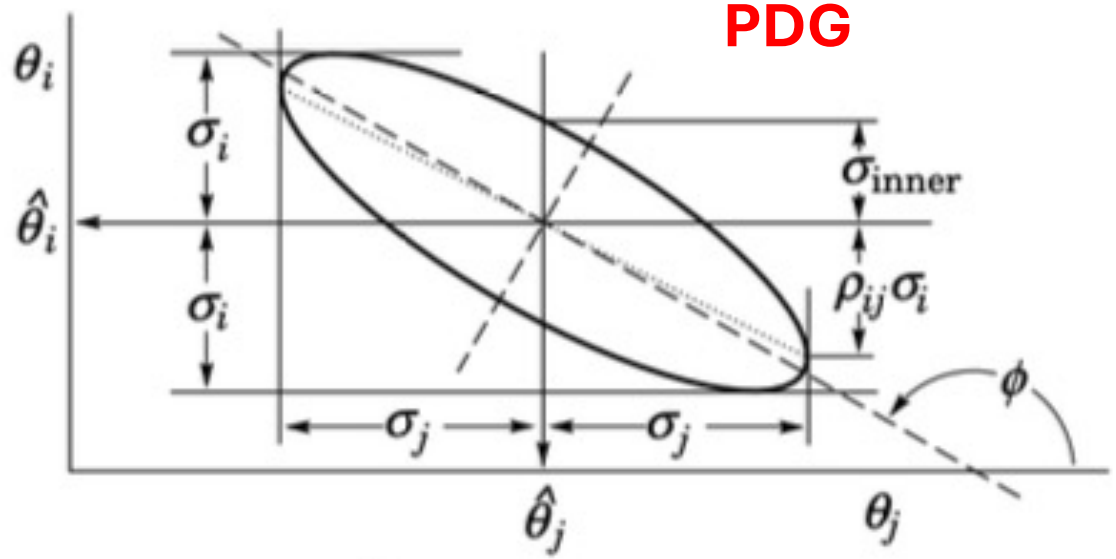


Figure 45: Contour to delimit the region of $\chi^2 \leq 1$ or $\Delta\text{NLL} \leq 0.5$ in the case of two parameters θ_i and θ_j.

$\chi^2_{\max}$	$N=1$	$N=2$	$N=3$	$N=4$
1	0.683	0.393	0.199	0.090
4	0.954	0.865	0.739	0.594
9	0.997	0.989	0.971	0.939

Coverage for the condition $\chi^2 \leq \chi^2_{\max}$ for different number (N) of parameters.

Coverage	$N=1$	$N=2$	$N=3$	$N=4$
0.683	1.00	2.30	3.53	4.72
0.90	2.71	4.61	6.25	7.78
0.95	3.84	5.99	7.82	9.49
0.99	6.63	9.21	11.3	13.2

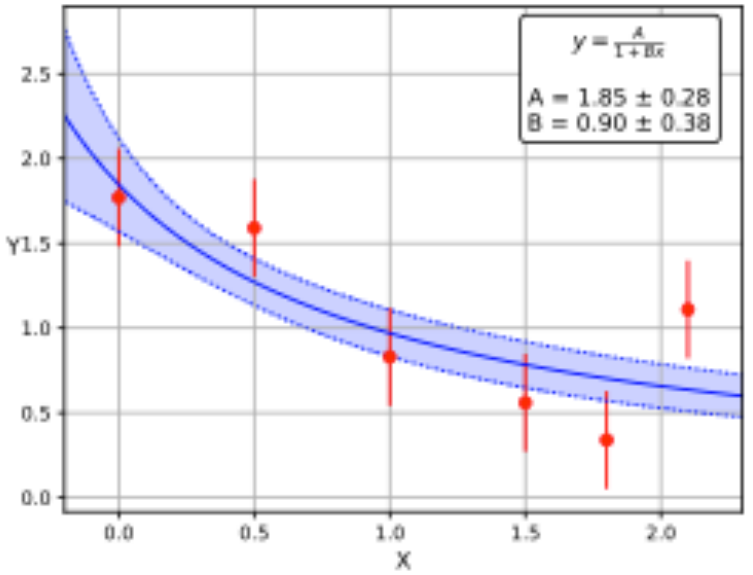
Values of $\chi^2_{\max}$ to achieve a given coverage for different number (N) of parameters.

Propagation of errors, how to construct uncertainty bands

- Interested in derived quantity $z = f(\vec{a})$.
- Clearly $\hat{z} = f(\hat{\vec{a}})$. What about its uncertainty?

$$\sigma^2(z) = \vec{dz}^T \; V \; \vec{dz}$$

$$\vec{dz} \equiv \begin{bmatrix} \frac{\partial z}{\partial a_1} \\ \frac{\partial z}{\partial a_2} \\ \vdots \\ \frac{\partial z}{\partial a_N} \end{bmatrix} \equiv \left[\frac{\partial z}{\partial \vec{a}}\right]$$



- Application to 1σ uncertainty bands for function $y(\vec{a}, x)$:

$$(\delta y(x))^2 = \left[\frac{\partial y}{\partial \vec{a}}\right]^T V \left[\frac{\partial y}{\partial \vec{a}}\right]$$

Partial derivatives computed at each *x*

Profile likelihood

- Multi-dimensional likelihood $\mathcal{L}(\vec{a}) = \mathcal{L}(\overrightarrow{a_1}, \overrightarrow{a_2})$ where we want to focus on the subset of parameters $\overrightarrow{a_1}$
- Profile likelihood ratio $\lambda(\overrightarrow{a_1}) = \frac{\mathcal{L}(\overrightarrow{a_1}, \widehat{\widehat{\overrightarrow{a_2}}})}{\mathcal{L}(\widehat{\overrightarrow{a_1}}, \widehat{\overrightarrow{a_2}})}$
  - The denominator, with single hats, gives the maximum of the likelihood
  - The numerator is a function of $\overrightarrow{a_1}$ and at each value of $\overrightarrow{a_1}$ the likelihood is maximized with respect to $\overrightarrow{a_2}$ (hence the weird double hat notation).
- Wilk's theorem: $-2\log\lambda(\overrightarrow{a_1})$ is asymptotically distributed as χ^2 with N_1 degrees of freedom. (N_1=dimensionality of $\overrightarrow{a_1}$)
- A **profile** ΔNLL scan gives a function with the same properties as described earlier with respect to $\overrightarrow{a_1}$ only

Profile likelihood (continued)

- The common use in frequentist statistics is to “profile away” the nuisances (ie: $\overrightarrow{a_2}$ in the previous page would be the set of nuisances)
- This means that the nuisances are fitted together with the quantity of interest.
  - The knowledge of the nuisances (systematics) is then improved in the analysis.
    - **Whether you like it or not**

Fitting Tools

- Minuit is the tool of choice for numerical minimization in HEP.
  - C++ and python implementations. (Fortran also, if still “alive”).
- Every HEP student should have some familiarity with this program
  - Minos function in Minuit does profile likelihood scan

- In some case it is convenient to write your own χ^2 fitter using linear algebra techniques. Big picture:
  - Start from 1st guess at parameters
  - Linearize the χ^2 by first order Taylor expansion around the guess
  - Gives set of linear equations that can be solved for the parameters
  - Put the solution into a new (better) guess and iterate until you are happy.

Fitting with Linear Algebra

We have M measurements: $\vec{y}^m = \begin{bmatrix} y_1^m \\ y_2^m \\ \vdots \\ y_M^m \end{bmatrix}$

with covariance $W = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1(M-1)}^2 & \sigma_{1M}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2(M-1)}^2 & \sigma_{2M}^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{(M-1)1}^2 & \sigma_{(M-1)2}^2 & \cdots & \sigma_{(M-1)(M-1)}^2 & \sigma_{(M-1)M}^2 \\ \sigma_{M1}^2 & \sigma_{M2}^2 & \cdots & \sigma_{M(M-1)}^2 & \sigma_{MM}^2 \end{bmatrix}$ (Note, of course $\sigma_{ij}^2 = \sigma_{ji}^2$).

We want to fit a function $\vec{y}(\vec{a}) = \begin{bmatrix} y_1(\vec{a}) \\ y_2(\vec{a}) \\ \vdots \\ y_M(\vec{a}) \end{bmatrix}$ where the N parameters to be determined are $\vec{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}$

20.1.1 The fit

The χ^2 can be written in terms of a matrix equation:

$$\chi^2 = [\vec{y}^m - \vec{y}(\vec{a})]^T \;\; W^{-1} \;\; [\vec{y}^m - \vec{y}(\vec{a})] \tag{141}$$

where:

- $[\vec{y}^m - \vec{y}(\vec{a})]^T$ is a $1 \times M$ matrix
- W^{-1} is a $M \times M$ matrix
- $[\vec{y}^m - \vec{y}(\vec{a})]$ is a $M \times 1$ matrix

We start by guessing a solution $\vec{a}_0$ and define a new vector $\delta\vec{a} \equiv \vec{a} - \vec{a_0}$.
Next we expand in a Taylor series:

$$y_i(\vec{a}) = y_i(\vec{a_0}) + \sum_{j=1}^{N} \frac{\partial y_i}{\partial a_j}\delta a_j = y_i(\vec{a_0}) + \sum_{j=1}^{N} \frac{\partial y(x_i, \vec{a})}{\partial a_j}\delta a_j$$

This expansion is exact for functions that are linear in $\vec{a}$, i.e., polynomials example $y(x, \vec{a}) = a_1 + a_2x^2 + a_3/x$ is also linear in $\vec{a}$

The expansion can be written compactly as $\vec{y} = \vec{y_0} + A\delta\vec{a}$, where:

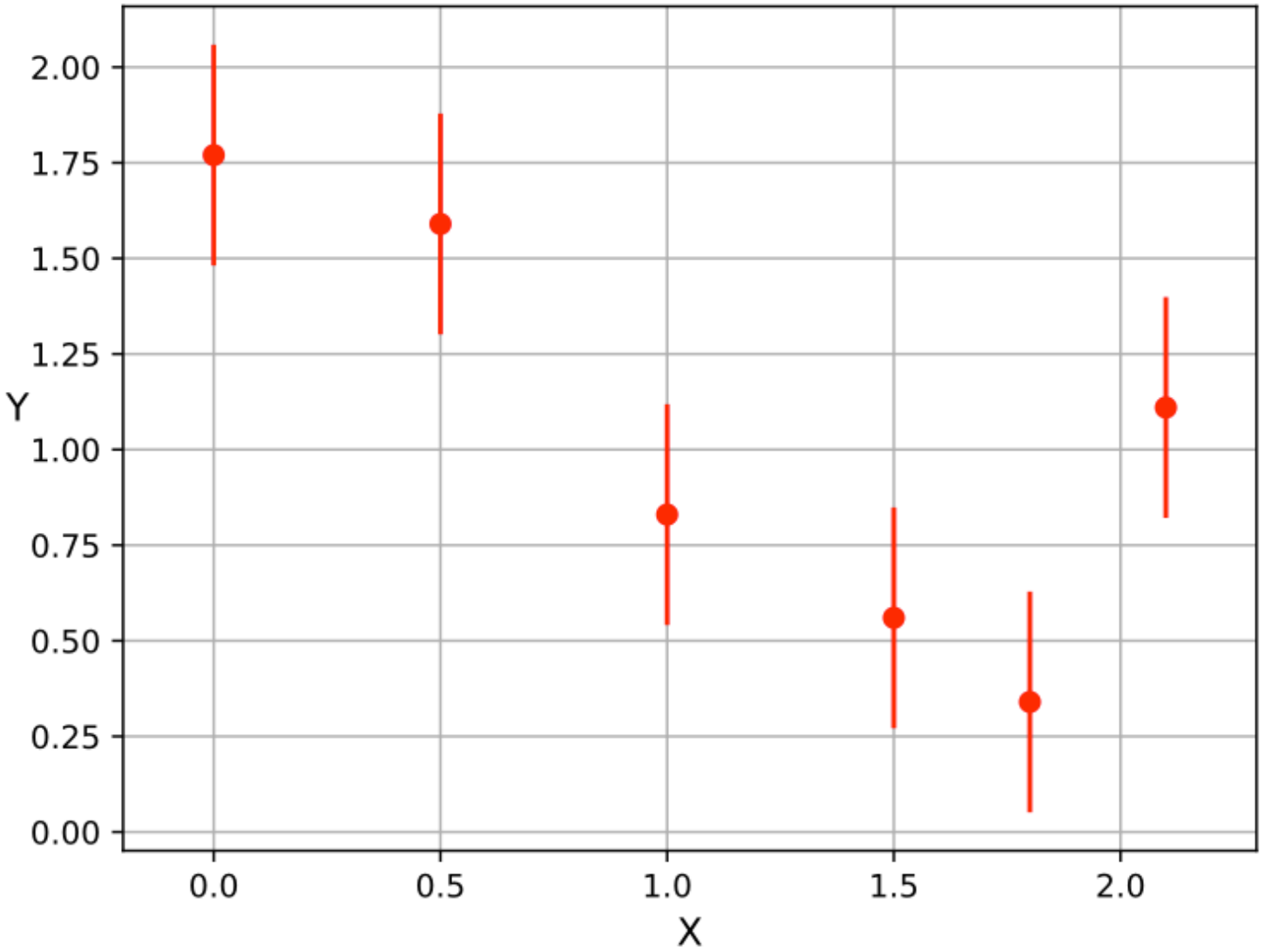
- $\vec{y} \equiv \vec{y}(\vec{a})$
- $\vec{y_0} \equiv \vec{y}(\vec{a_0})$
- $A_{ij} \equiv \frac{\partial y_i}{\partial a_j}$ is a $M \times N$ matrix.

Minimizing χ^2 yields $\boxed{\delta\vec{a} = (A^T W^{-1} A)^{-1} A^T W^{-1} \delta\vec{y}}$ $\delta\vec{y} \equiv \vec{y}^m - \vec{y_0}$.

$$\boxed{\vec{a} = \vec{a_0} + \delta\vec{a}}$$

With covariance matrix $\boxed{\boldsymbol{V = G = (A^T W^{-1} A)^{-1}}}$

Toy Example



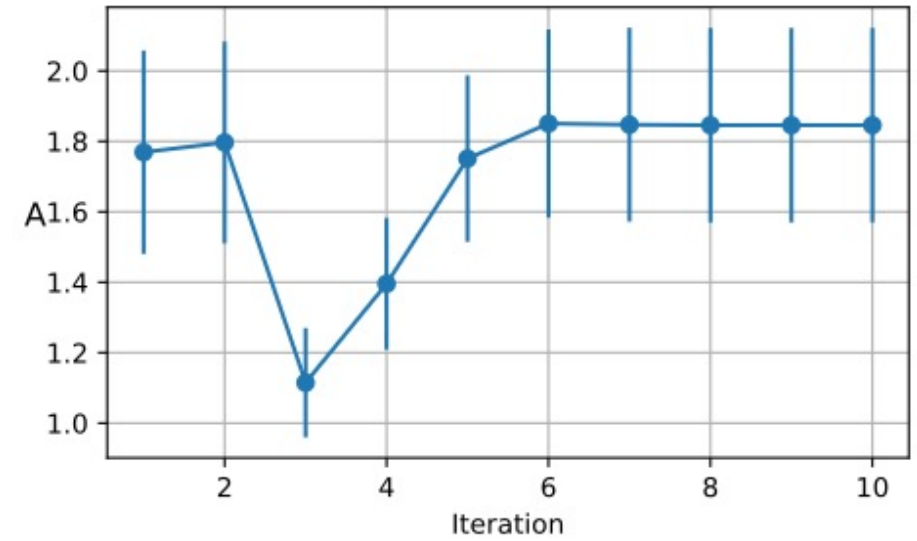
Six data points with

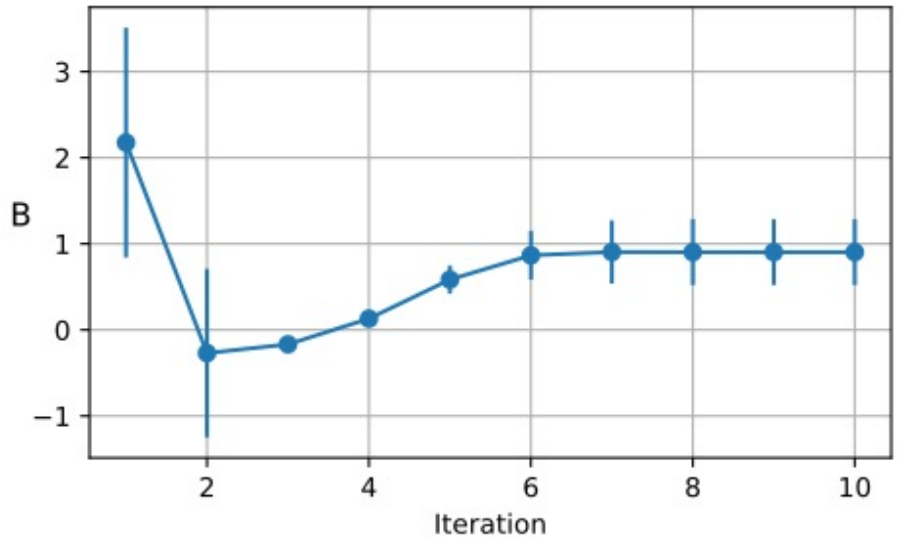
$$y(x) = \frac{2}{1+x} + \delta_y$$

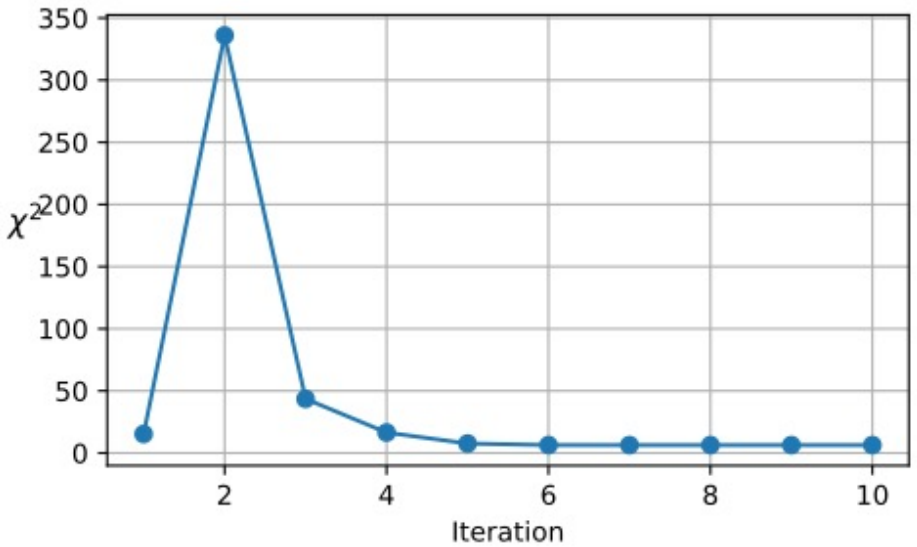
And δ_y= rand ± 0.5

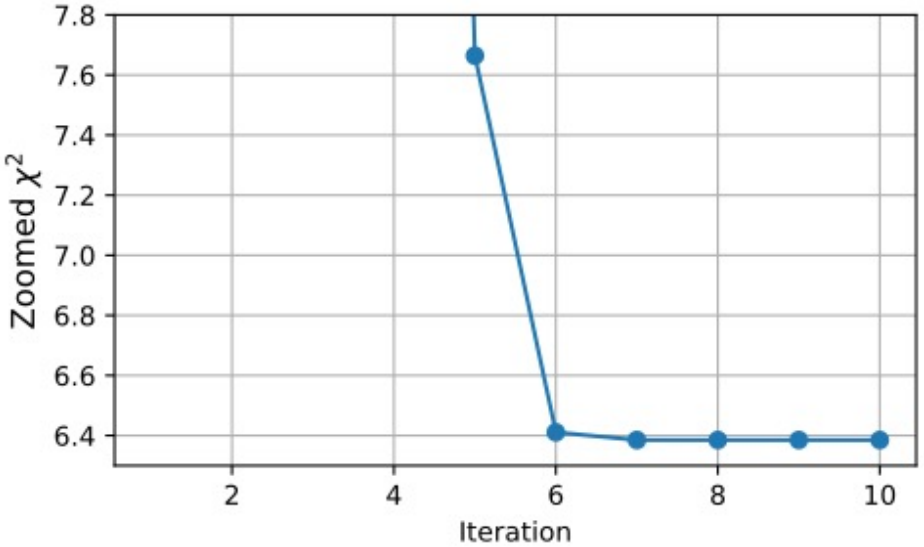
Fit $$y(x) = \frac{A}{1+Bx}$$

With starting point
A=12 and B=10









The evolution of the χ^2 and the fitted values of A and B after each iteration.

“True” values were A=2 and B=1