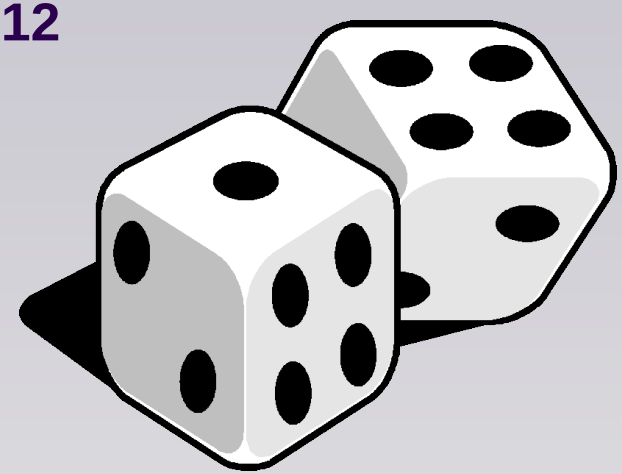
Introduction to Statistics

CERN Summer Student Lecture Program 2012



Helge Voss

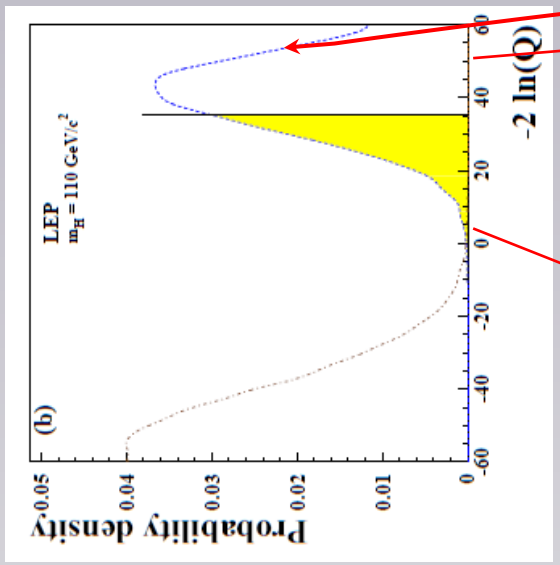
… and Machine Learning

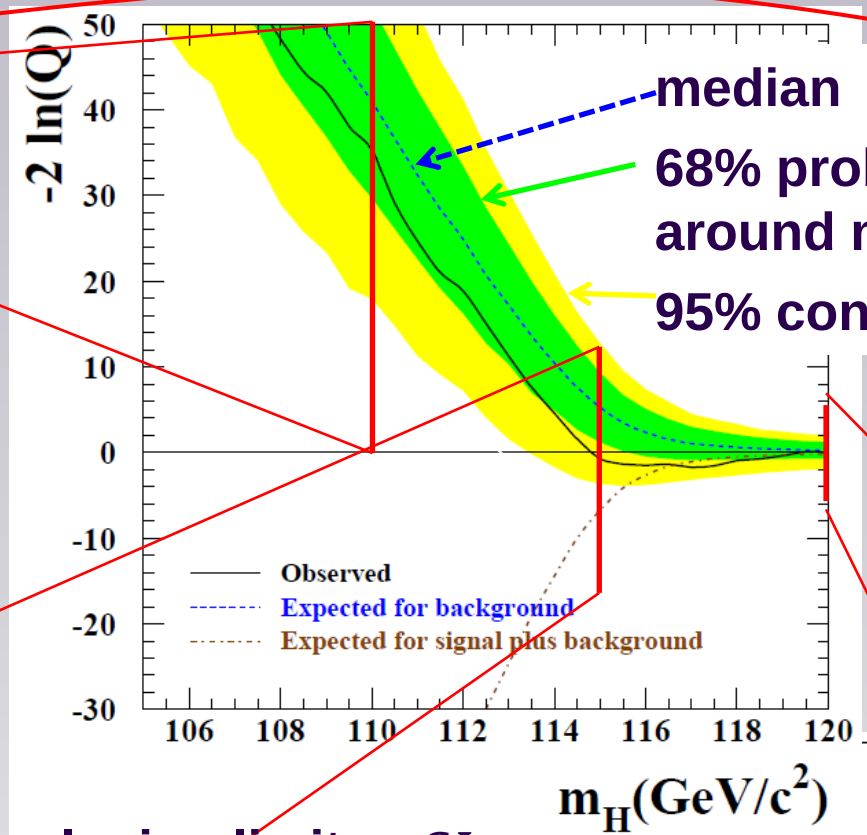
(in the last lecture)

Outline

- Why Statistics
- What is Probability :
  - frequentist / Bayesian interpretation
  - Hypothesis testing
    - error types and Neyman-Pearson Lemma, confidence level α and p-value
    - new particle searches – example: Higgs
- **Lecture 3**
  - **Parameter estimation**
    - Maximum Likelihood fit
    - χ^2-fit
  - **Neyman Confidence belts**
  - **(Monte Carlo Methods (Random numbers/Integration) → see slides)**
- Lecture 4
  - Machine Learning / Pattern Recognition

Example: LEP SM Higgs Limit

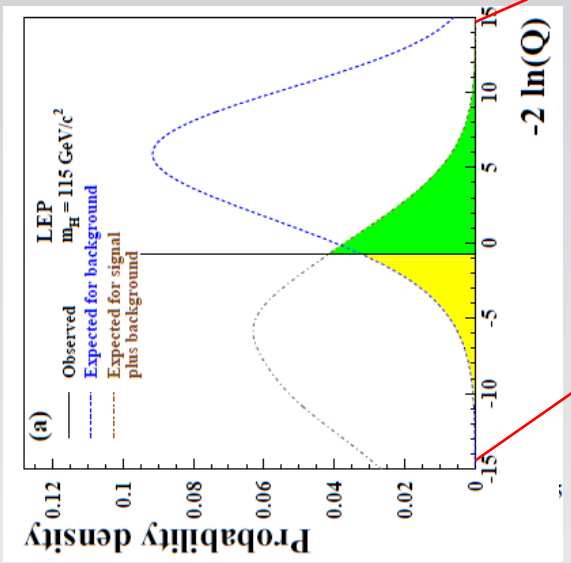




median

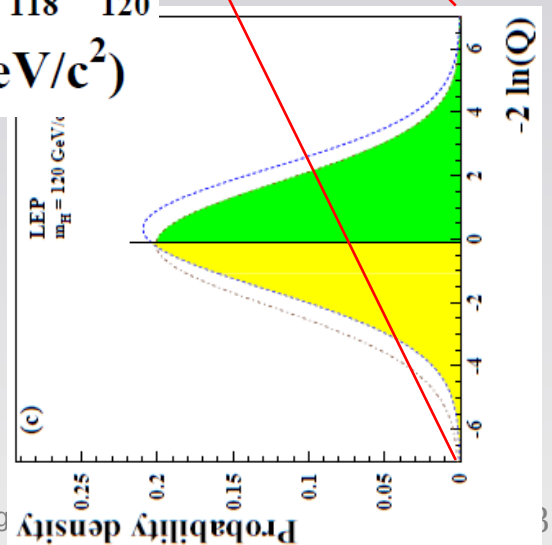
68% prob. content around median

95% content of



Exclusion limit → CL_{s+b}

- draw "bands" around expectation for signal+bkg
- excluded at 95%CL where "observed" lies outside 95% CL band



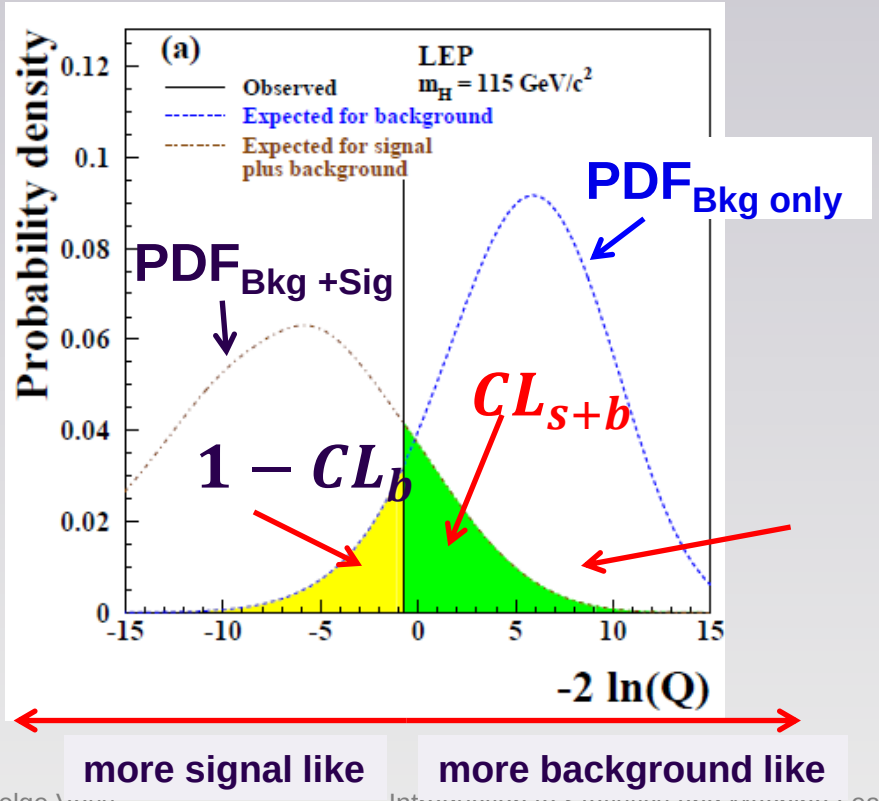
Example LEP Higgs Search

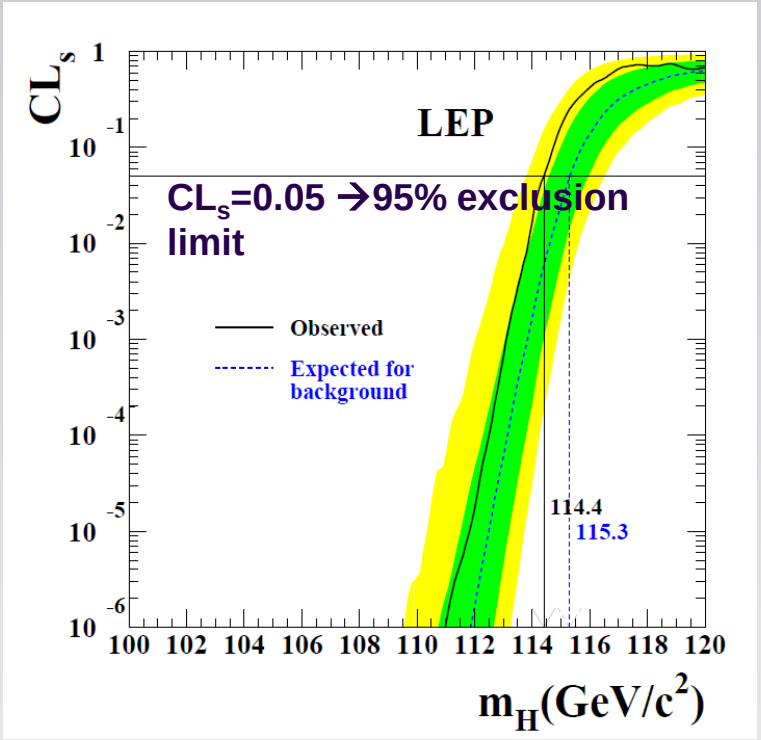
- "avoid" Being Lucky when setting the limit and observing an event count less than the expected background (we'll come back to that, later…)

→ rather than "quoting" in addition the expected sensitivity

→ weight your CL_{s+b} by it:

$$CL_s = \frac{p_{s+b}}{1-p_b} = \frac{CL_{s+b}}{1-CL_b} = \frac{P(LLR \geq LLR_{obs}|H_1)}{P(LLR \leq LLR_{obs}|H_0)}$$

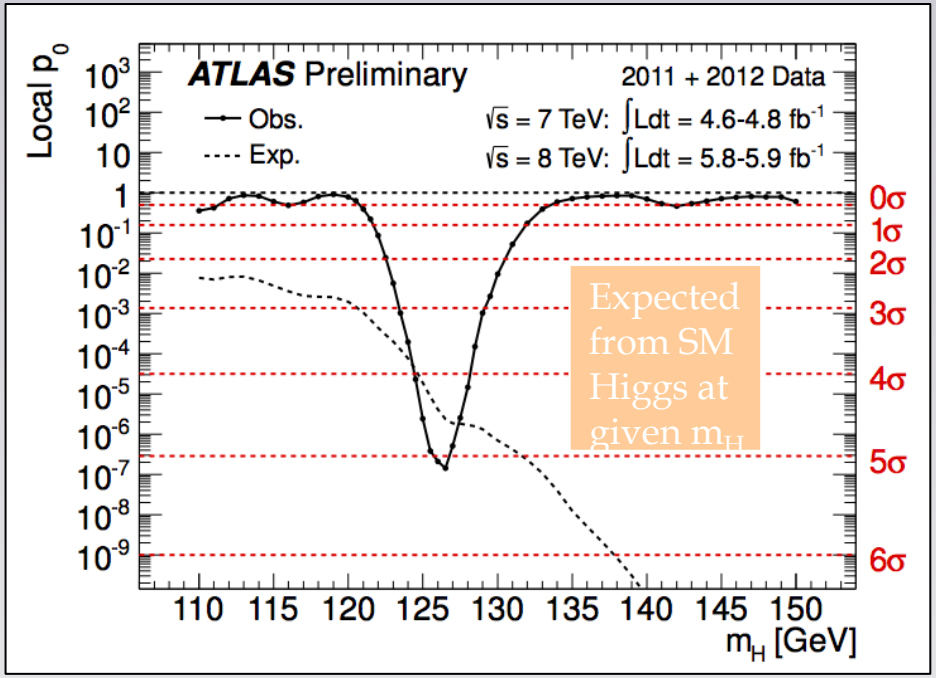




ATLAS/CMS Higgs Search

- Aim for DISCOVERY → disprove H_0 = background ONLY
  - somewhat different test statistic: "profile Likelihood ratio" of Likelihood function $L(\mu, \theta), with\ \mu = \frac{\sigma}{\sigma_{SM}}, \theta: nuisance\ parameters$
  - p-value for discovery: Bkg only hypothesis ($\mu = 0$)

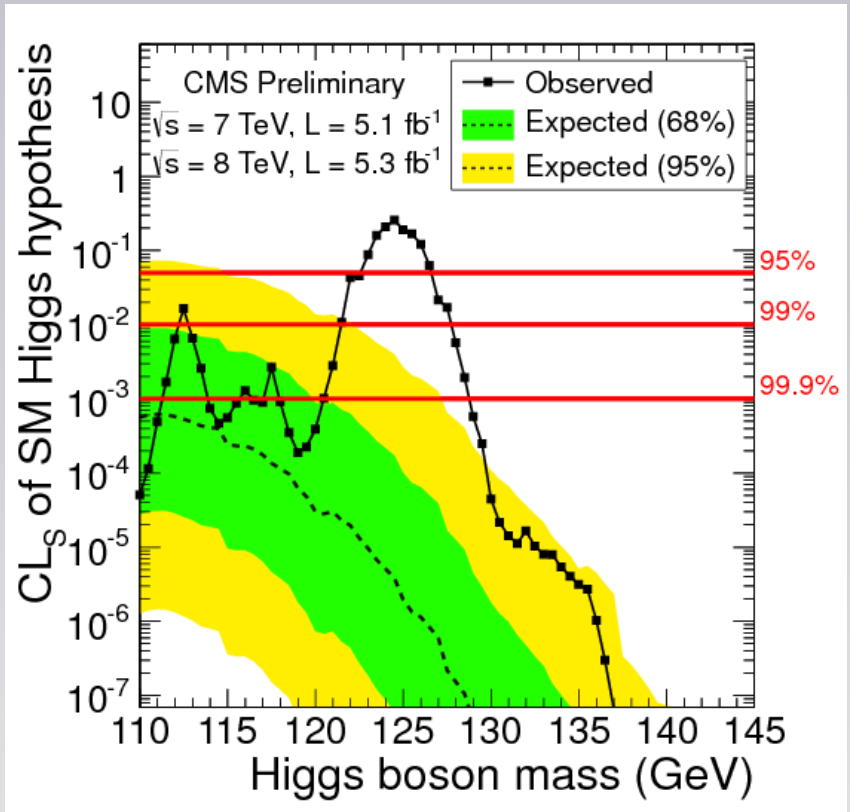
- p-value calculated "locally" for every Higgs mass
- Look at any "dip" in p-values over whole mass range
  - think as "binned" in Higgs mass resolution
- Random samples of a distribution, histogram it → 1 out of 20 bins (5%) will deviate 2σ from expectation.. e.t.c.



- LOOK-ELSEWHERE-EFFECT
  - not taken into account → local p-value

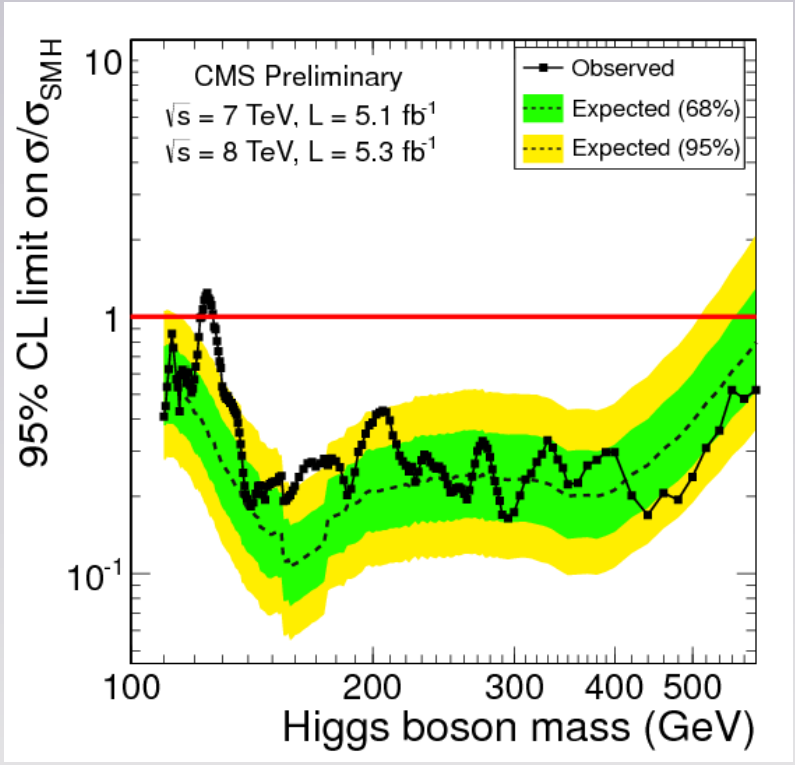
CL_s and Excluded Cross Section

- $CL_s = \frac{p_{s+b}}{1-p_b}$



- adjust $\mu = \frac{\sigma}{\sigma_{SM}}$ such that $CL_s = 95\%$

→ limit on $\mu = \frac{\sigma}{\sigma_{SM}}$



Message:

They can nicely exclude everything at "high Confidence levels" apart from where they see the signal

Systmatic Uncertainties

- **standard popular way: (Cousin/Highland)**
  - **integrate over all systematic errors and their "Probability distribution)**
    - →marginalisation of the "joint probability density of measurement parameters and systematic error)

! Bayesian ! (probability of the systematic parameter)

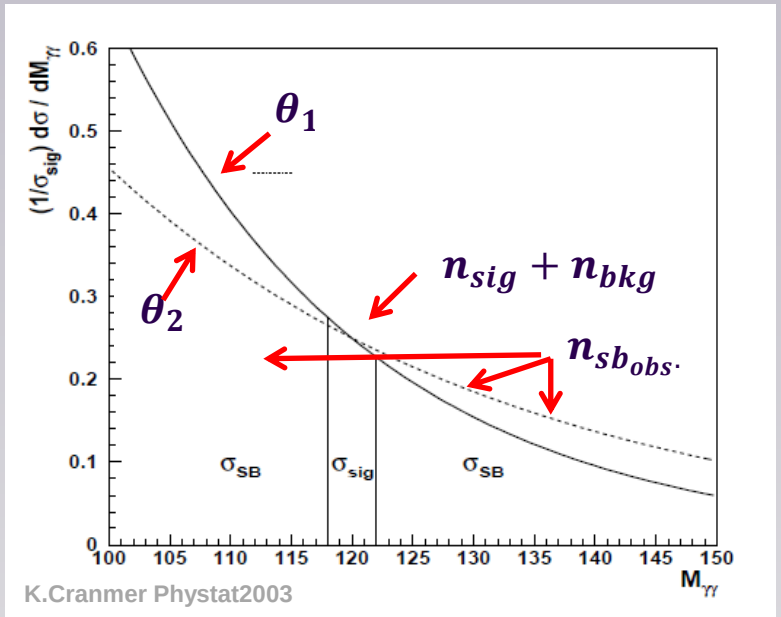
- →"hybrid" : frequentist intervals and Bayesian systematic
- →has been shown to have possible large "undercoverage" for very small p-values /large significances (i.e. underestimate the chance of "false discovery" !!

- **LEP-Higgs: generated MC to get the PDFs with "varying" param. within systematic uncertainty**

→ essentially the same as "integrating over" → need probability density for "how these parameters vary"

Systematic Uncertainties

- We can do better: systematic uncertainly as "free parameter" in the fit



- eg. background → sidebands
- parametrise $f_{sb}(n_{sideband}; \theta) = f_{sb}(n_{sb}; \theta)$
- uncertainty: scale/shape ?
  - free parameter θ in parametrisation

→ $f_{sb}(n_{sb}) \rightarrow f_{sb}(n_{sb}; \theta)$

  - extrapolate to signal region

→ bkg expecation:

$$b = b(n_{sb}; \theta) = f_{sig}(n_{sb}; \theta)$$

Note: no need to specify prior probability

- Likelihood function includes:
  - parameters of interest
  - parameters describing the influence of the sys. uncertainty

→ the latter are called: nuisance parameters

$$\underbrace{P(n_{sig}, n_{bkg}, n_{sb_{obs.}} | s, \theta)}_{\text{joint model}} = \underbrace{P(n_{sig} + n_{bkg} | s + b(n_{sb}; \theta))}_{\text{measurement of interest}} \underbrace{P(n_{sb_{obs.}} | f_{sb}(n_{sb}; \theta))}_{\text{sideband}}$$

Nuisance Parameters and Profile Likelihood

- Likelihood function includes:
  - parameters of interest $\mu = \frac{\sigma}{\sigma_{SM}}$
  - parameters of the sys. uncertainty (nuisance parameters θ)

→ $L = L(\mu, \theta)$:

- "most likely parameters μ and θ are found where the Likelihood is maximised

- used in test statistic: $L(\mu, \hat{\theta})$ … i.e. Likelihood "maxmized" w.r.t. θ

But: let's now talk about

- "Maximum Likelihood" fitting
- Parameter fitting in general

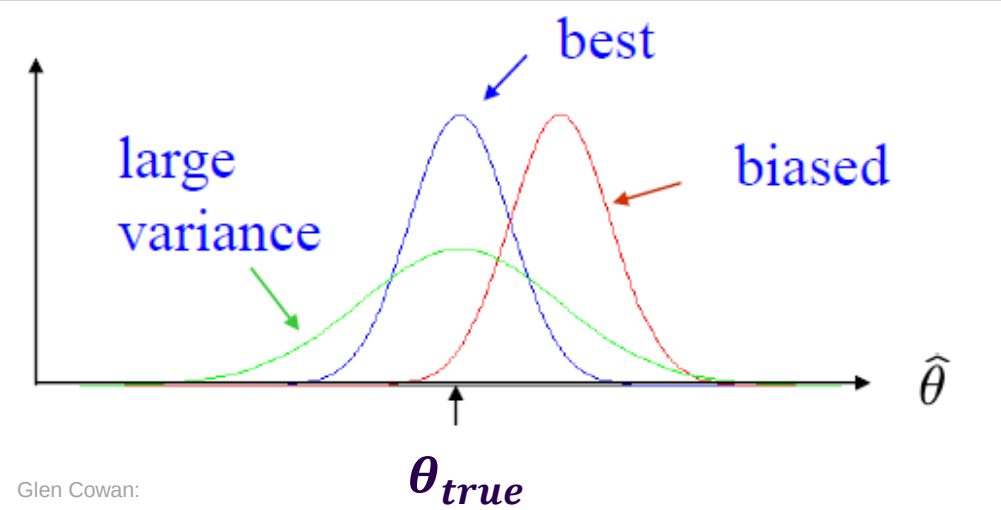
Parameter Estimation

- "estimator" → estimate characteristic parameter of parent distribution using a limited "sample" from the distribution. e.g.:
  - mean value: "estimator": $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$
    - there are others: $\hat{\mu} = \frac{1}{2}(x_{i_{min}} + x_{i_{max}})$
  - variance: "estimator" : $\hat{V} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$
  - median: …..
  - …. polarisation in your differential cross section

properties of estimators

- biased or unbiased
- large or small variance

→ distribution of $\hat{\theta}$ on many measurements ?

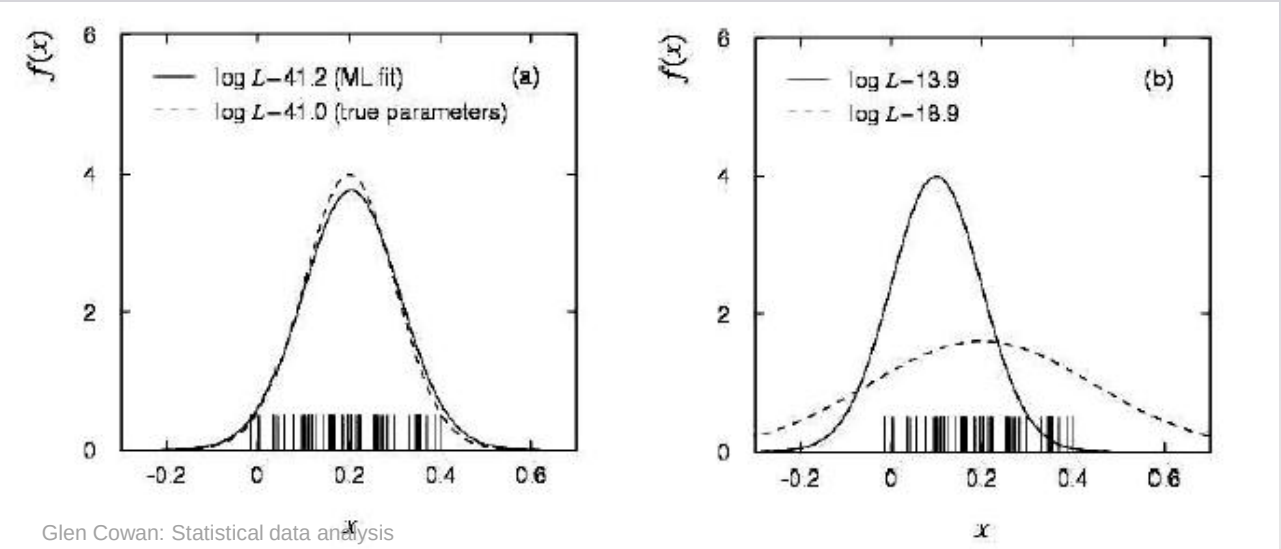


- Small bias and small variance are typically "in conflict"

Maximum Likelihood Estimatior

- want to measure/estimate some parameter θ
  - e.g. mass, polarisation, etc..
- observe: $\vec{x}^i = (x_1, \ldots . x_n)_i \quad i = 1, K$
  - e.g n observables for K events
- "hypothesis" i.e. PDF $\mathrm{P}(\vec{x}; \theta)$ - distribution $\vec{x}$ for given θ
  - e.g. diff. cross section

  → K independent events: $\mathrm{P}(\vec{x}^1, \ldots \vec{x}^K; \theta) = \prod_i^K \mathrm{P}(\vec{x}^i; \theta)$
- for fixed $\vec{x}$ regard $\mathrm{P}(\vec{x}; \theta)$ as function of θ (i.e. Likelihood! $L(\theta)$)
  - θ close to θ_{true} → Likelihood $L(\theta)$ will be large



Glen Cowan: Statistical data analysis

→ try to maximise $L(\theta)$

→ typically:

→ -2Log($L(\theta)$)

→ minimise

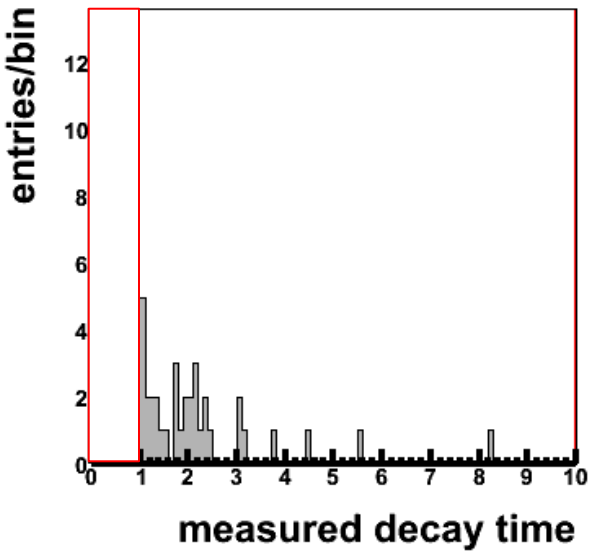
→ $\hat{\theta}$

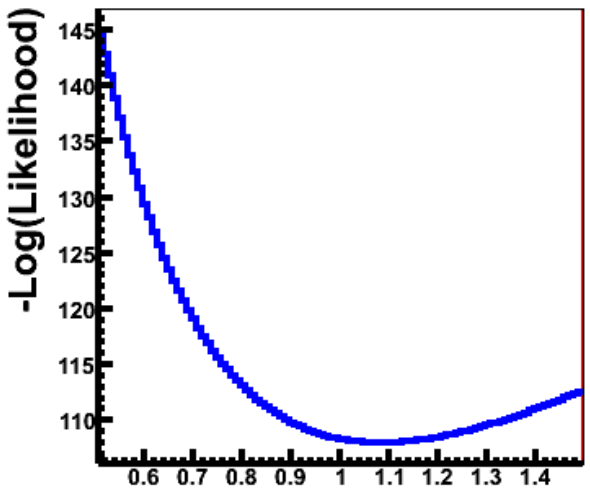
→ Maximum Likelihood estimator

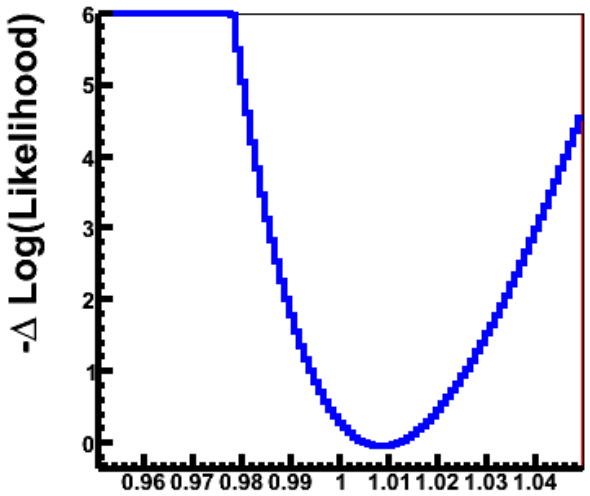
Maximum Likelihood Estimator

Lifetime τ =? : decay times are exponentially distributed $P(t) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$

- **can observe decay times > t_0 only →** $P(t) = e^{\frac{t_0}{\tau}} \frac{1}{\tau} e^{-\frac{t}{\tau}}$
- **data sample:** $\{t_1, t_2, \ldots . t_N\} \rightarrow L(\tau) = \prod_i^N P(t_i)$
- $-ln(L(\tau)) = \sum_i^N ln(P(t_i))$
- $= -N\left(\frac{t_0}{\tau} - ln(\tau)\right) + \frac{1}{\tau}\sum_i^N t_i$





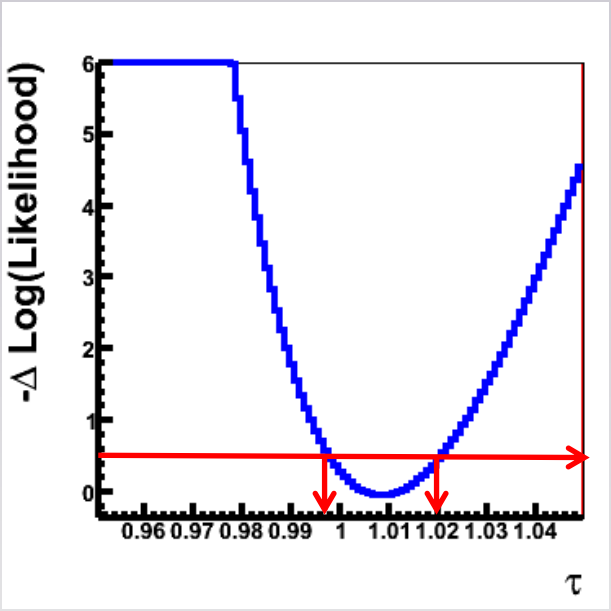


zoom and subract offset

Maximum Likelihood Estimator

- **Error on estimated parameter $\hat{\theta}$ $(\hat{\tau})$?**
- **Taylor expansion:**

$$-ln\big(L(\theta)\big) \approx -ln\Big(L\big(\hat{\theta}\big)\Big) - \underbrace{\left[\frac{dln(L\theta)}{d\theta}\right]_{\hat{\theta}} \big(\theta-\hat{\theta}\big)}_{=0 \text{ (minimum)}} - \frac{1}{2}\left[\frac{d^2ln(L\theta)}{d\theta^2}\right]_{\hat{\theta}} \big(\theta-\hat{\theta}\big)^2 + \ldots$$



→ $L(\theta) \approx L\big(\hat{\theta}\big) e^{\frac{1}{2}\left[\frac{d^2ln(L\theta)}{d\theta^2}\right]_{\hat{\theta}} \big(\theta-\hat{\theta}\big)^2} = L\big(\hat{\theta}\big) e^{-\frac{\big(\theta-\hat{\theta}\big)^2}{2\sigma^2}}$

→ $\frac{1}{\sigma^2} = -\left[\frac{d^2ln(L\theta)}{d\theta^2}\right]_{\hat{\theta}}$

→ $-ln\big(L(\theta)\big) \approx -ln\Big(L\big(\hat{\theta}\big)\Big) + \frac{1}{2\sigma^2}\big(\theta-\hat{\theta}\big)^2$

→ **read off parabolic $ln(L)$ curve:** $-ln\Big(L\big(\hat{\theta}\pm\sigma_\theta\big)\Big) = -ln\Big(L\big(\hat{\theta}\big)\Big) + \frac{1}{2}$

Maximum Likelihood Estimator

example: PDF(x) = Gauss(x,μ,σ) → $\mathrm{L}(\mu|x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

→ estimator for μ_{true} from the data measured in an experiment $x_1, \ldots . x_N$

→ full Likelihood $\mathrm{L}(\mu|x) = \prod_i^N \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$

→ typically: $-2\ln\left(\mathrm{L}(\mu|x)\right) = \sum_i^N \left(\frac{(x_i-\mu)^2}{2\sigma^2}\right) + N\frac{1}{\sqrt{2\pi\sigma^2}}$ Note: It's a function of μ !

→ $-2\Delta\ln\left(\mathrm{L}(\mu)\right) = \sum_i^N \left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)$

→ χ^2 , least squares

Maximum Likelihood Estimator

- **Maximum Likelihood is typically unbiased only in the limit $K \to \infty$**
  - **If Likelihood function is "Gaussian" (often the case for large N → central limit theorem)**

→ get "error" estimate from or -2$\Delta log(L) = 1$

→ if (very) none Gaussian

→ revert typically to (classical) Neyman confidence intervals

Binned Maxmimum Likelihood

- rather than having $\mathrm{P}(\vec{x}^1,\ldots \vec{x}^K;\theta) = \prod_i^K \mathrm{P}(\vec{x}^i;\theta)$ for each event I

→ use binned events (i.e. a histogram)

→ e.g. if $\mathrm{P}(\vec{x}^i;\theta)$ is not analytically available

→ in each bin i there are n_i events, Poisson distributed around μ_i

→ get prediction $\mu_i = \mu_i(\theta)$ from "Monte Carlo" or analytical model

$$L(\theta) = P\left(n_1, \ldots n_{n_{bins}}; \theta\right) = \prod_i \frac{{\mu_i}^{n_i}}{n_i!} e^{-\mu_i}$$

$$-2ln(L)(\theta) = 2\sum_i^{n_{bins}} (ln(n_i!) - n_i ln(\mu_i) + \mu_i)$$

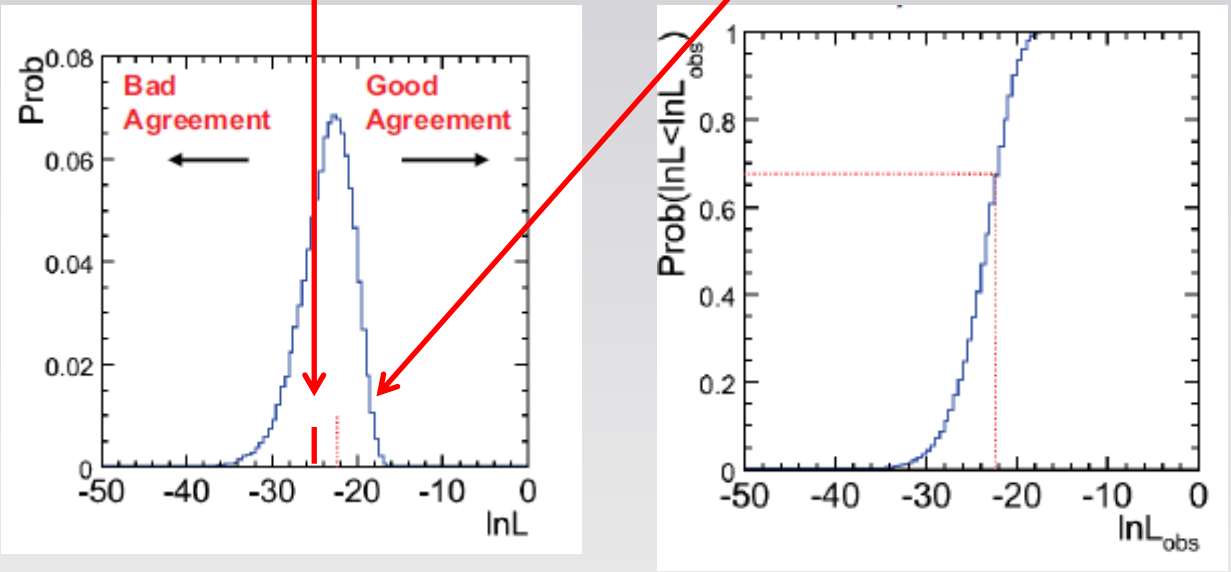
Goodness-of-Fit

- So far we know the "uncertainty" on the fitted value of θ, but…
- did the fitted model "really" describe the data?

- The value of the lnL (log Likelihood) at the minimum does not "mean anything" → **calibrate!**

→ determine the distribution of lnL fit results with Monte Carlo toys!

→ check your "data"-fit



Goodness-of-Fit

- Easier with Gaussian distributed variables

→ least square fit

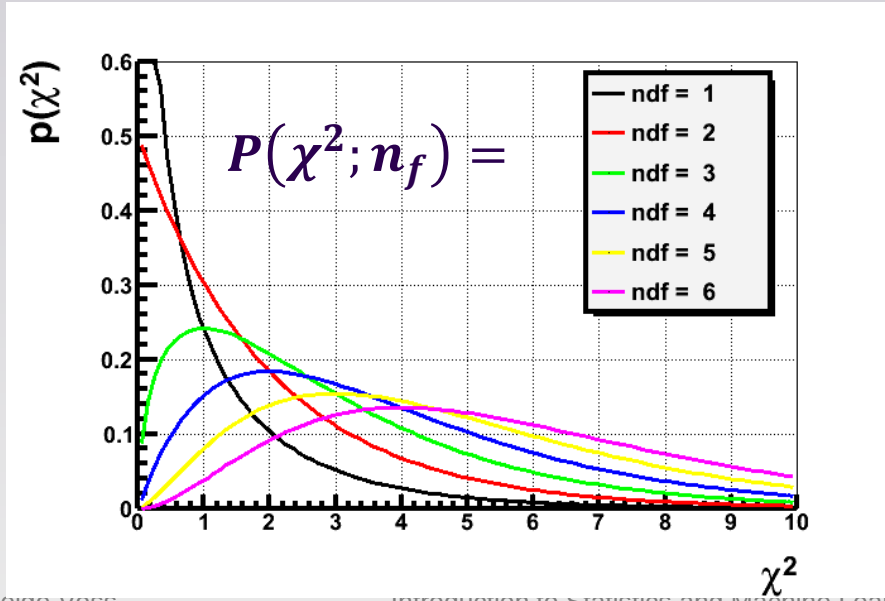
→ χ^2

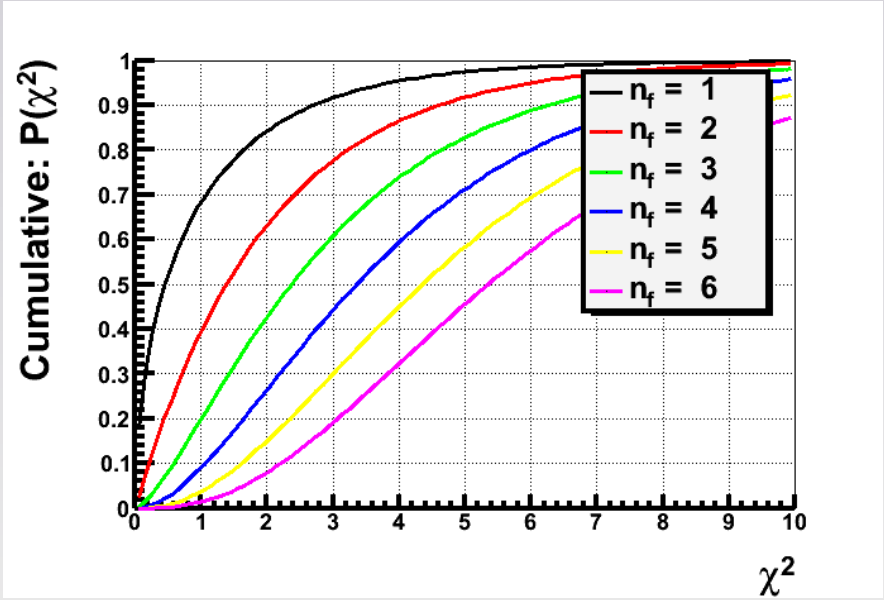
$$\chi^2 = \sum_i^n \left(\frac{(\widehat{\mu_i} - \mu_i(\theta))^2}{2\sigma^2} \right)$$

has known distribution: $E[\chi^2] = n_f$: #number of "degres of freedom"
i.e. $n - \#fitted\ parameters$

Chi2 Probability: The 1- cumulative distr. of $P(\chi^2, n_f)$ −distribution

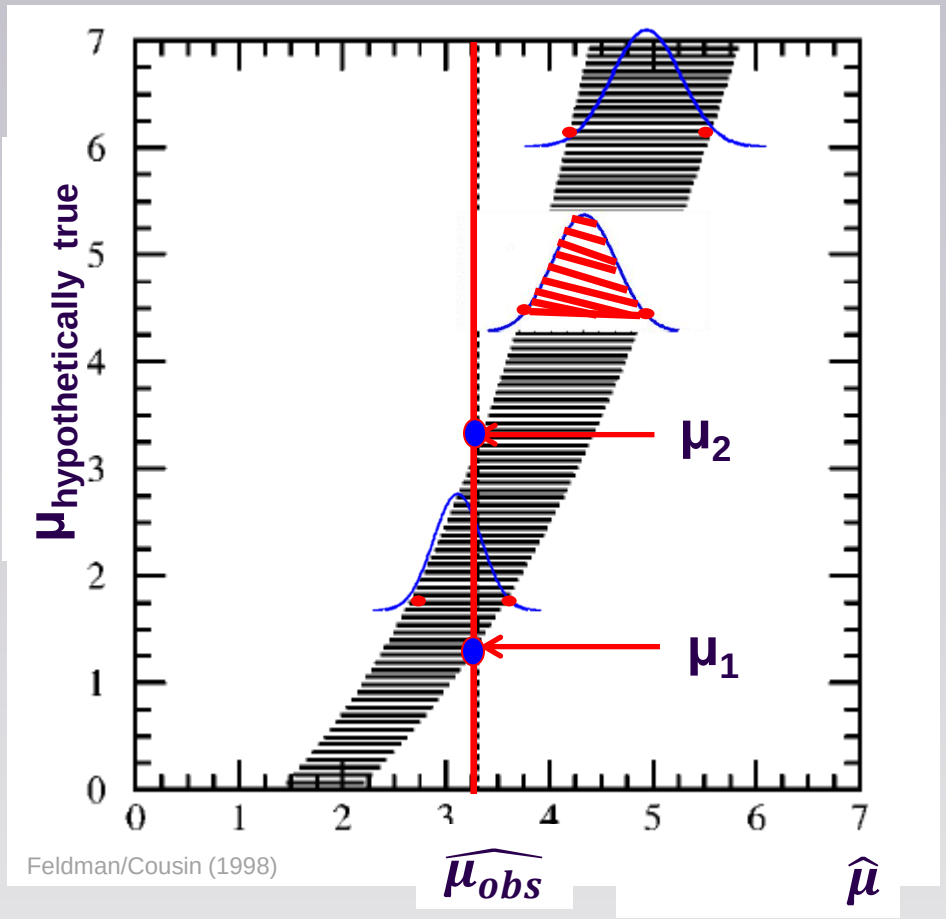
→ how often to expect "worse" fit result (i.e. with larger χ^2 $value$ at min.)





Classical Confidence Intervals

- **Neymans Confidence belt for CL α (e.g. 90%)**



- each $\mu_{hypothetically\ true}$ has a PDF of how the measured values will be distributed
- determine the (central) intervals ("acceptance region") in these PDFs such that they contain α
- do this for ALL $\mu_{hyp.true}$
- connect all the "red dots" → confidence belt

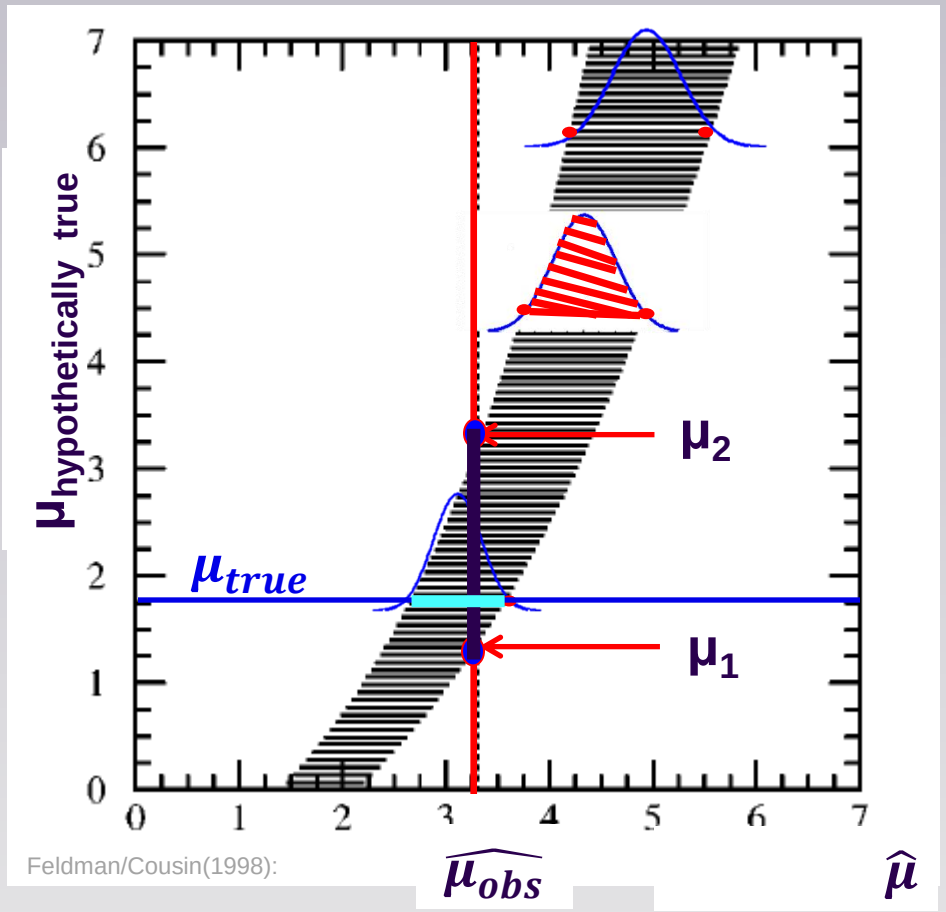
- measure $\widehat{\mu_{obs}}$:

→ conf. interval =[μ_1, μ_2] given by vertical line intersecting the belt.

- by construction: for each $\widehat{\mu_{obs}}$. (taken according PDF(μ_{true}) the confidence interval [μ_1, μ_2] contains μ_{true} in α = 90% cases

Classical Confidence Intervals

- **Neymans Confidence belt for CL α (e.g. 90%)**



→conf.interval =[μ_1, μ_2] given by **vertical** line intersecting the belt.

- by construction:
  - $P(\mu < \widehat{\mu_{obs}}; \mu_2) = \frac{1-\alpha}{2}$
  - $P(\mu > \widehat{\mu_{obs}}; \mu_1) = \frac{1-\alpha}{2}$
- if the true value were μ_{true}

→ lies in $[\mu_1, \mu_2]$ if it intersects ▮

→ $\widehat{\mu_{obs}}$ intersects ▬ in 90%
(that's how it was constructed)

→ only those x_{obs} give $[\mu_1, \mu_2]$'s that intersect with the ▬

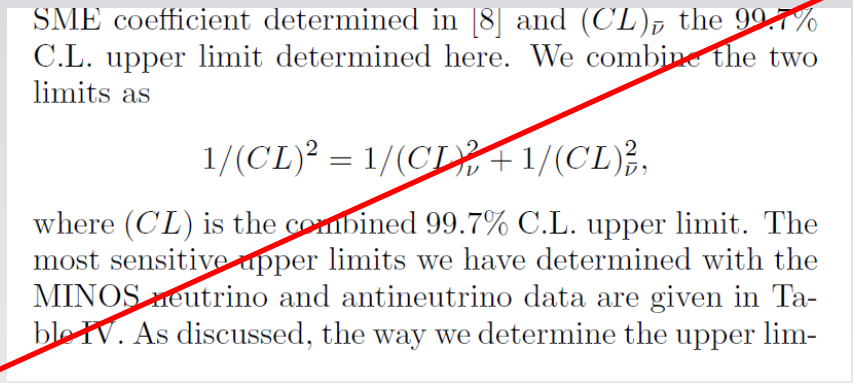
→ 90% of intervals cover μ_{true}

Combine Confidence Intervals

- $\hat{\mu}$: Gaussian PDF: Neyman CL → Maximum Likelihood (ML)
- In the limit of ML approximation (Gaussian PDF's) → combine "as usual"
  - But: don't be fooled to believe you are combining statements about where the 'true parameters' are likely to be !
- Otherwise:
  - Combine individual measurements (not the derived confidence intervals)

→ construct "confidence belt" of combined measurement

→ obviously you cannot combine directly "upper limits" this way:



SME coefficient determined in [8] and $(CL)_{\bar{\nu}}$ the 99.7% C.L. upper limit determined here. We combine the two limits as

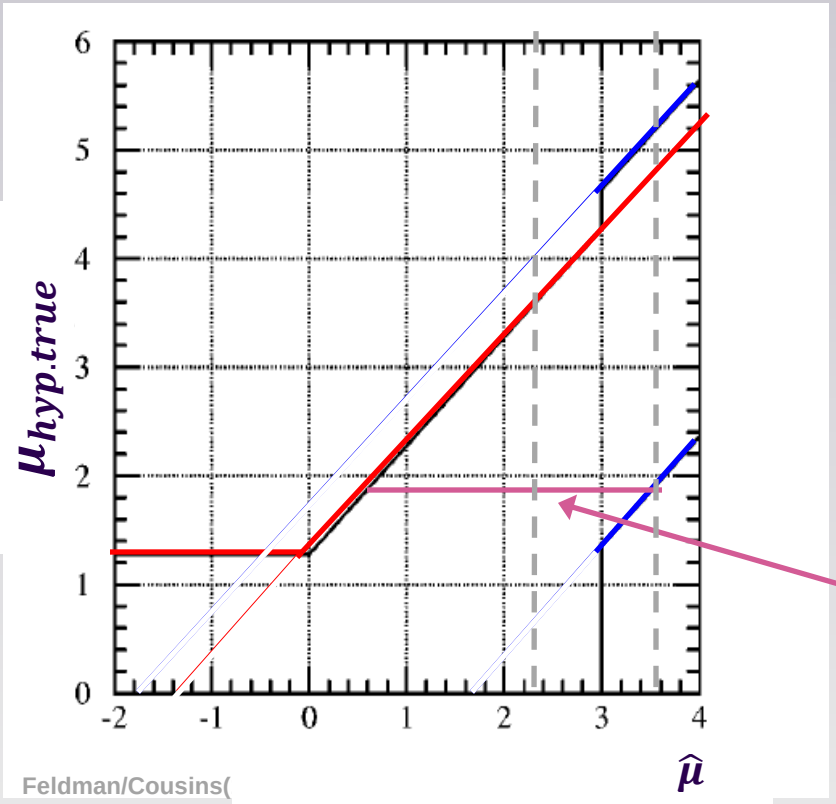
$$1/(CL)^2 = 1/(CL)_{\nu}^2 + 1/(CL)_{\bar{\nu}}^2,$$

where (CL) is the combined 99.7% C.L. upper limit. The most sensitive upper limits we have determined with the MINOS neutrino and antineutrino data are given in Table IV. As discussed, the way we determine the upper lim-

Flip-Flop

When to quote measuremt or a limit!

- estimate Gaussian distributed quantity $\hat{\mu}$ that cannot be < 0 (e.g. mass)
- same Neyman confidence belt construction as before with 90%CL:
  - once for measurement (two sided, each tail contains 5%)
  - once for limit (one sided tails contains 10%)



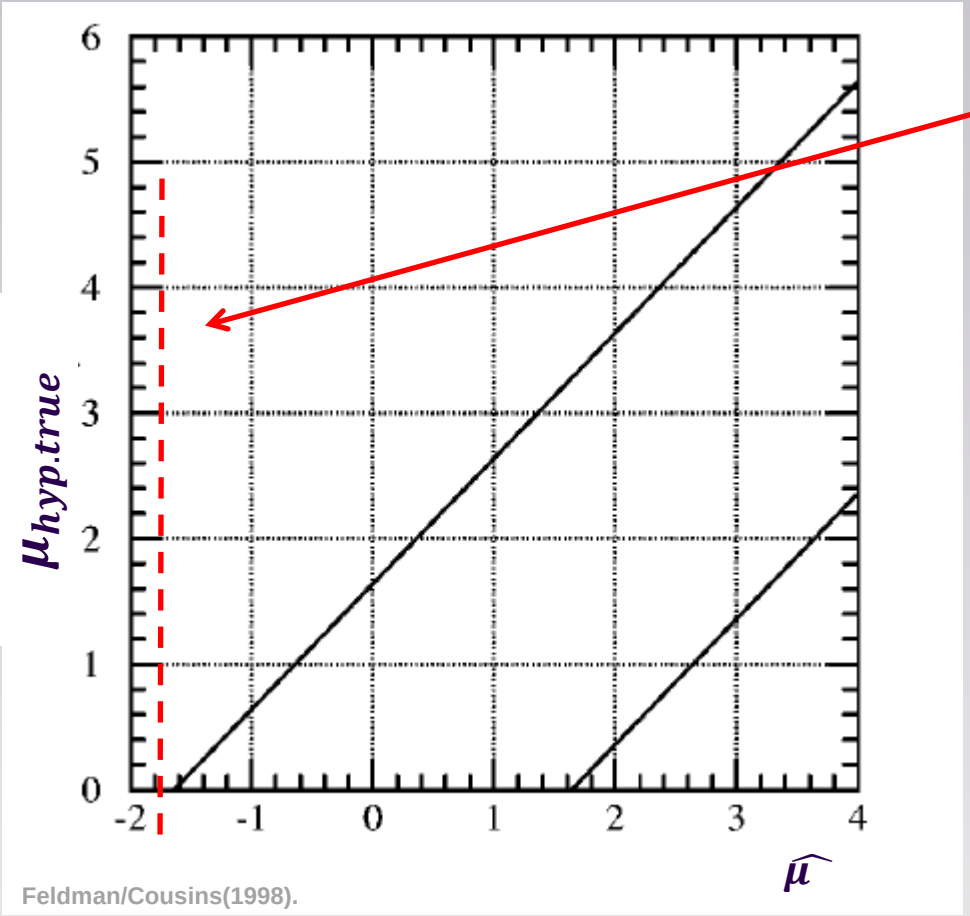
- decide: if x_{obs}<0 assume you =0
  → conservative
- if you observe $\widehat{\mu_{obs}}$<3
  → quote upper limit only
- if you observe$\widehat{\mu_{obs}}$>3
  → quote a measurement

→ induces "undercovering" as this acceptance region contains only 85% !!

Some things people don't like..

same example:

- **estimate Gaussian distributed quantity $\hat{\mu}$ that cannot be < 0 (e.g. mass)**



Feldman/Cousins(1998).

- **using proper confidence belt**
- **assume: $\widehat{\mu_{obs}} = -1.8$**
  → confidence interval is EMPTY!

- **Note: that's OK from the frequentist interpretation $\mu_{true} \in [conf.interv.]$ in 90% of (hypothetical) measurements.**

Obviously we were 'unlucky' to pick one out of the remaining 10%

- **nonetheless: tempted to "flip-flop" ??? tsz .. tsz.. tsz..**

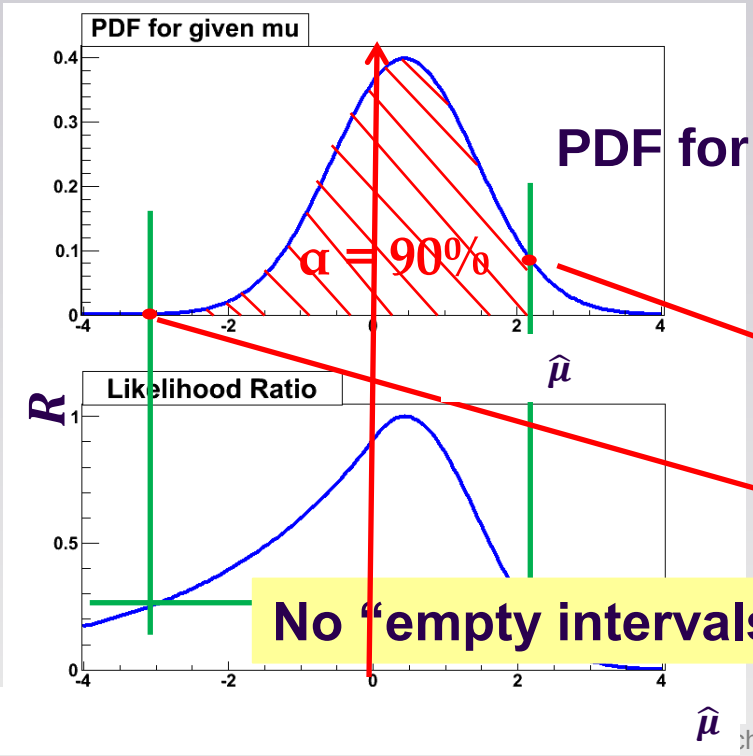
Feldman Cousins: a Unified Approach

- How we determine the "acceptance" region for each $\mu_{hyp.true}$ is up to us, as long as it covers the desired integral of size α (e.g. 90%)
- → standard: conf. central (for measurement) or one sided (for limits)

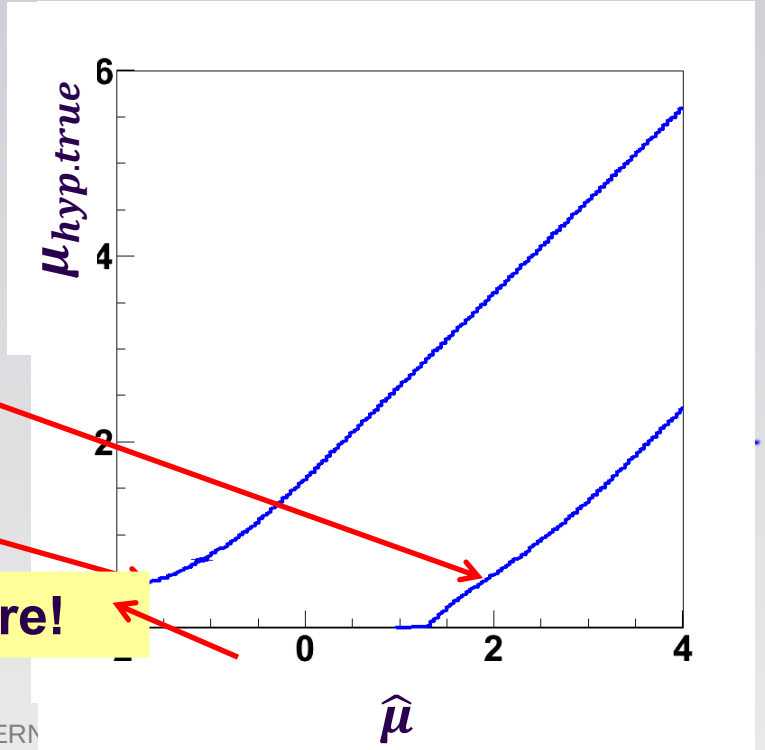
→ include those "$\hat{\mu}$", for which the likelihood ratio R is large, first:

$$R = \frac{L(\hat{\mu}|\mu_{hyp.true})}{L(\hat{\mu}|\widehat{\mu_{best}})}$$

- $\widehat{\mu_{best}}$ of μ given the estimator $\hat{\mu}$
- → $\widehat{\mu_{best}} = \hat{\mu}$ if in ALLOWED region
- → $\widehat{\mu_{best}} = \mu_{min_{allowed}}$ otherwise



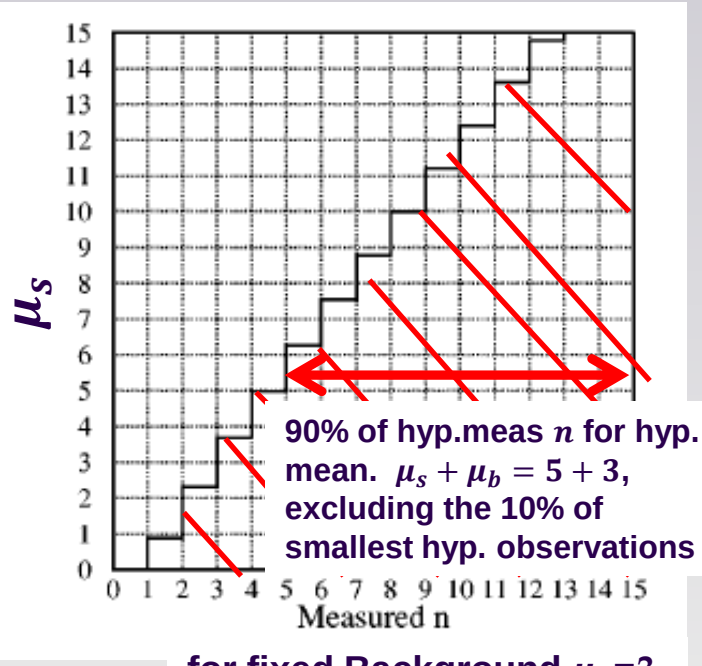
PDF for μ_{true}



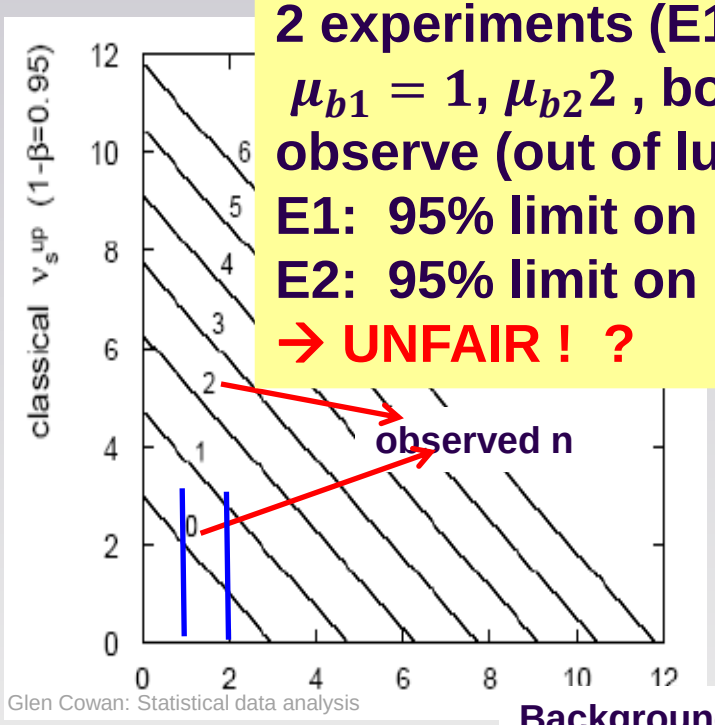
No "empty intervals anymore!

Being Lucky…

- upper limit on signal μ_s on top of known (mean) background μ_b
  - measure n, (n_s+n_b) events → Poisson distribute
  - $P(n) = Poisson(n, \mu_s + \mu_b)$
- Neyman: draw confidence belt with
  - "μ_s" on the "y-axis" (the possible true values of μ_s)



sorry… the plots don't match: one is of 90%CL the other for 95%CL



2 experiments (E1 and E2)
$\mu_{b1} = 1, \mu_{b2} 2$, both observe (out of luck) 0 ev.
E1: 95% limit on $\mu_s \sim 2$
E2: 95% limit on $\mu_s \sim 1$
→ UNFAIR ! ?

Being Lucky …

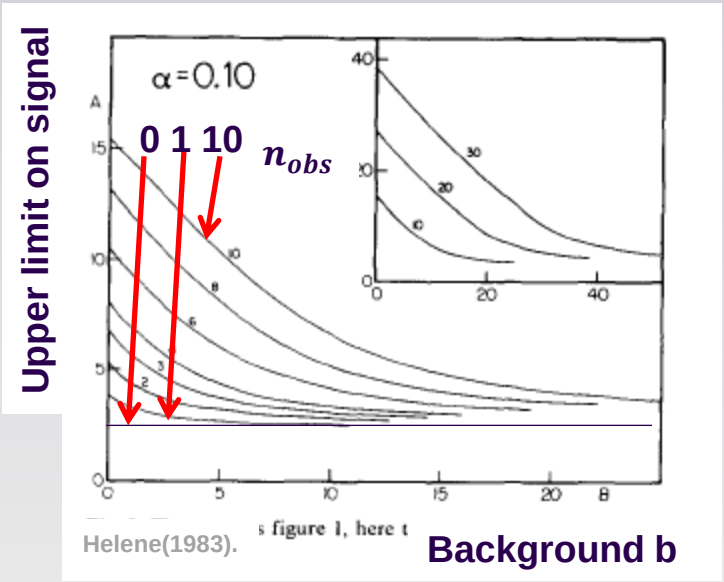
Bayesian: rather than constructing Confidence belts:

- turn Likelihood for μ_s (for given n_{obs}) into Posterior probability for μ_s

  $i.e\ Poisson(n_{obs}; \mu_s + \mu_b) \rightarrow L(n_{obs}; \mu_s)$

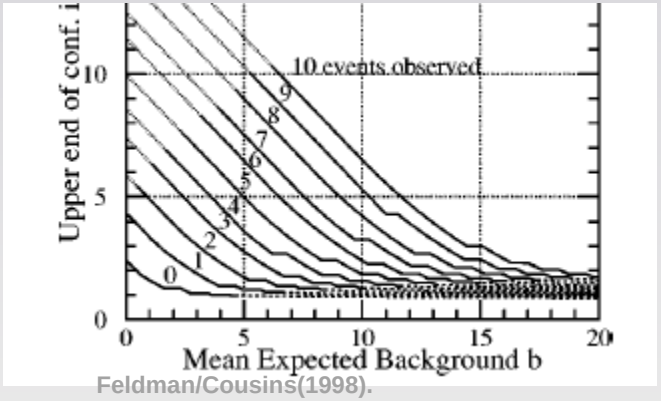
- $p(\mu_s|n_{obs}) = L(n_{obs}; \mu_s) * \pi(\mu_s)$ add prior probability on "s":

- $\pi(\mu_s) = \begin{cases} 0 & \mu_s < 0 \\ uniform & \mu_s > 0 \end{cases}$



Feldman/Cousins

- no empty intervals, but still "unfairness" (better limits if 0 observed and larger b-expected)
- perfectly "fine" in frequentist interpretation:
- should quote "limit+sensitivity"



Being Lucky -- Exclusion limits when 0 events observed

- Cousin-Feldman→ Likelihood ordering (1998)
- Roe Woodroofe → Constraint Likelihood ordering (1999)
- Mandelkern Schultz → Maximum likelihood estimator (2000)
- Cousins → why one should stick to likelihood ordering (2001)

- You see… all still very recent ! There's always debate going on, and its all not simply "textbook"

Summary

- **CLs … the HEP limit;**
  - **CLs … ratio of "p-values" … statisticians don't like that**
  - **new idea: Power Constrained limits**
    - rather than specifying "sensitivity" and "Neyman conf. interval"
    - "accept" limits only within experimental "sensitivity !
    - →lots of "different" ideas floating around how to "set limits"
    - →Hey! We don't need that anymore …well at least not for the Higgs.. ☺
- **.. a bit about Profile Likelihood, systematic error.**
- **Parameter estimation**
  - **Maximum Likelihood fit**
  - **χ^2 -fit (least squares)**
- **what to do if estimator is non-Gaussian:**
  - **Neyman – confidence intervals**
  - **what "bothers" people with them**
- **Feldmans/Cousins confidence belts/intervals**
  - **unifies "limit" or "measurement" confidence belts**

Monte Carlo Methods

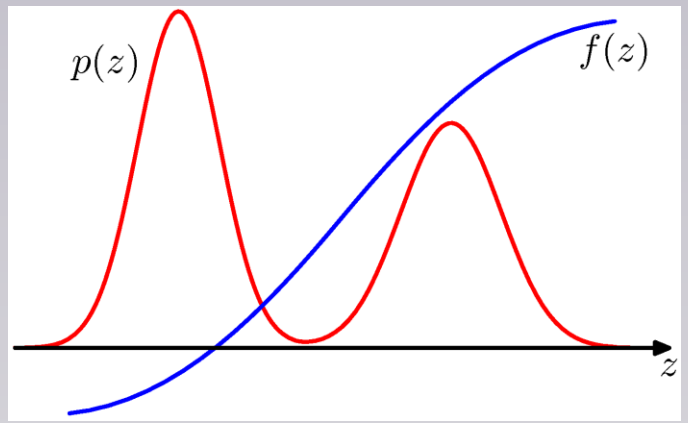
- Monte Carlo Integration

- Bootstrap (Monte Carlo re-sampling)

- Jackknife

Monte Carlo Integration

$$E[f(z)] \rightarrow \int f(z)p(z)dz$$



- simple n equidistant step-wise summation?

→ o.k. in 1 or "very few" dimensions D

→ n-steps grows exponentially with D

random sampling converges faster for large D

→ go to Monte Carlo

sorry, no prove..

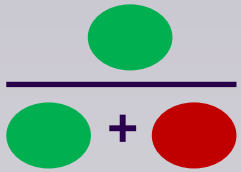
- often: the distribution $p(z)$ is not even fully known analytically:
- often: the normalisation of the distr. $p(z)$ is not known: e.g. calculation of Bayesian expectation values
- NOTE: for $E[f(z)]$ the normalisation of $p(z)$ is irrelevant !

→ not solvable by analytic integration

→ numeric integration

Hit-And-Miss Rejection Sampling

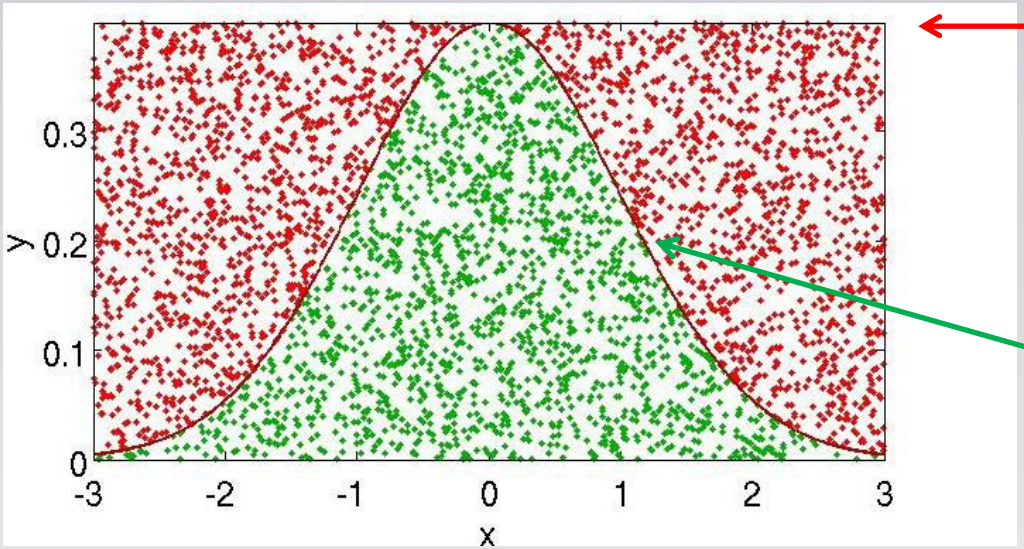
- generate random numbers with distribution $p(z)$
  - generate uniform random numbers in "enclosing space"
  - for each such random number, accept it with probability



$= \frac{p(z)}{q(z)}$

e.g. generate 2nd random number uniform in $[0; q(z)]$ and accept if it is > $p(z)$

→ accepted events follow $p(z)$ distribution



enclosing function : $q(z)$ defines proposal distribution.
→ some function that you can easily sample from

$p(z)$ the function defining the distribution we want to sample from

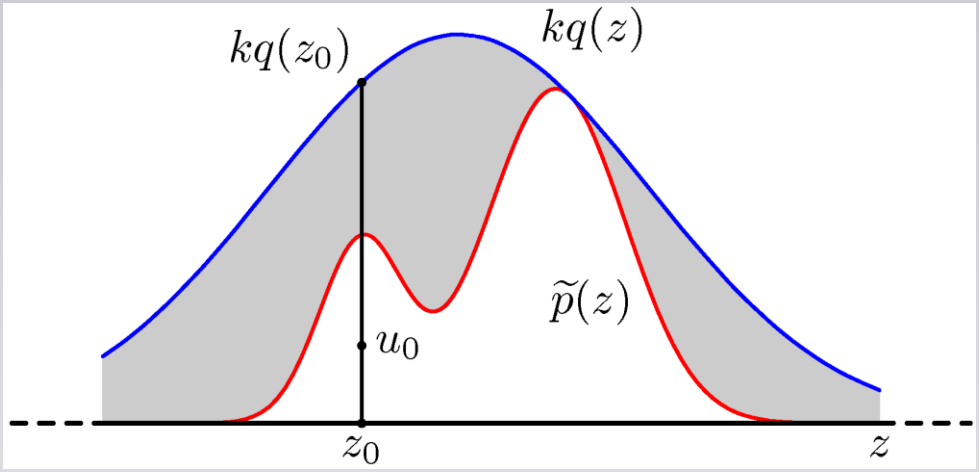
- Note: $fraction\ of\ accepted\ ev. \times \int q(z)dz = \int p(z)dz$

Rejection sampling

- one can get a bit more effective (less rejection)

→ none "square"/uniform proposal function $q(z)$

→ still sample "uniformly" in area under $q(z)$ and do as before



there are also techniques that automatically adapt the proposal distribution iteratively

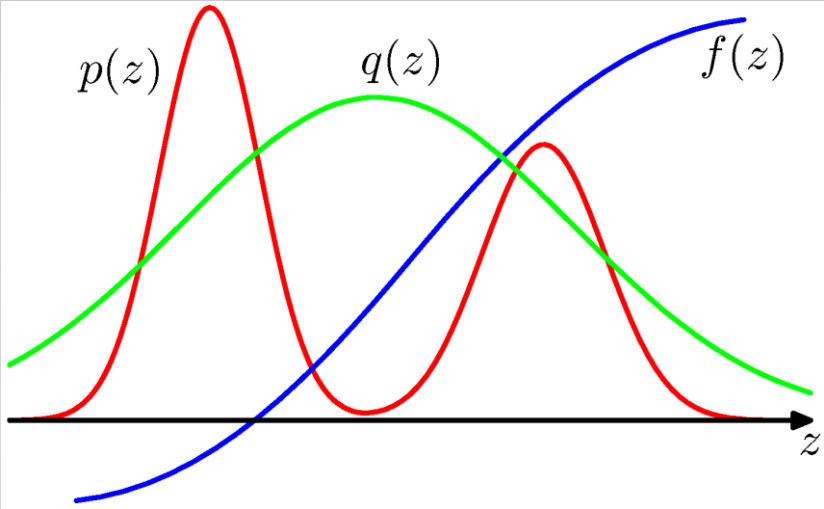
$$fraction\ of\ accepted\ ev. \times \int q(z)dz = \int p(z)dz$$

if only integration, not $p(z)$ random event generation: even more clever: →

Importance Sampling

- Rather than 'rejecting' events with p= $\frac{p(z)}{q(z)}$

→ weigh them by factor $\frac{p(z)}{q(z)}$: "importance weights"



- Note: in this way, the proposal $q(z)$ does not even have to "enclose" $p(z)$, as weight can also be > 1
- $E[f(x)]$ also with unknown normalisations of $p(z)$ and $q(z)$

→$p(z) = \frac{\tilde{p}(z)}{Z_p}$ and $q(z) = \frac{\tilde{q}(z)}{Z_q}$

$$E[f(z)] = \int f(z)\frac{\tilde{p}(z)}{Z_p}dz = \frac{Z_q}{Z_p}\int f(z)\frac{\tilde{p}(z)}{\tilde{q}(z)}q(z)dz \simeq \frac{Z_q}{Z_p}\frac{1}{N}\sum^{N}\frac{\tilde{p}(z^{(i)})}{\tilde{q}(z^{(i)})}f(z^{(i)})$$

$$\text{with } \frac{Z_p}{Z_q} = \frac{1}{Z_q}\int \tilde{p}(z)dz = \int \frac{\tilde{p}(z)}{\tilde{q}(z)}q(z)dz \simeq \frac{1}{N}\sum^{N}\frac{\tilde{p}(z^{(i)})}{\tilde{q}(z^{(i)})}$$

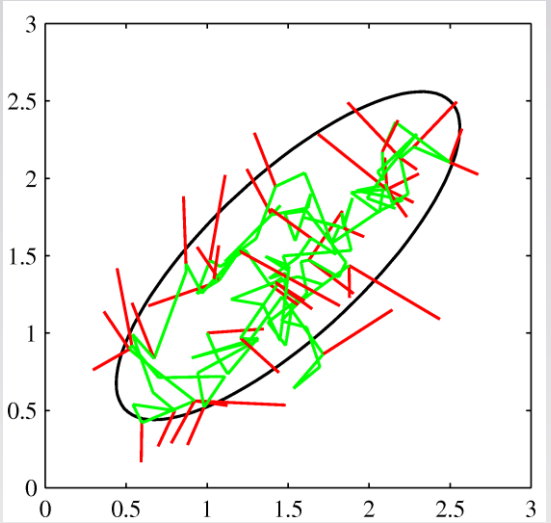
Markov Chain Monte Carlo

- previous techniques
  - → accuracy depends on how closely $q(z)$ follows $p(z)$
  - → problem for "sparse" , "unknown" $p(z)$
    - every "random point" chosen independent of previous one
- Markov chain: (e.g. random walk)
  - → consecutive random steps depend on previous location in random variable space
  - → allows to favor stepping into regions where $p(z)$ large

Metropolis Algorithm

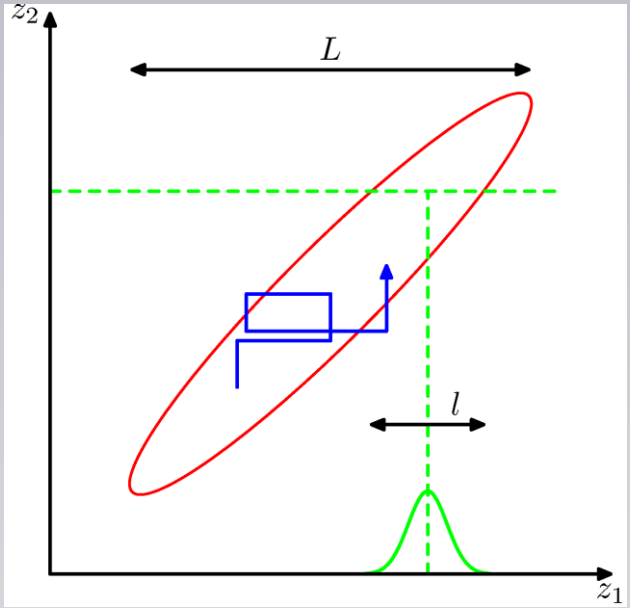
- Start somewhere in z −space at random
  - sample this point
- provide "proposal distribution" $q(z'|z)$ to jump from $z \rightarrow z'$
  - e.g. Gaussian with some "metric" in z −space , symmetric in $z \leftrightarrow z'$
  - accept z' if:
    - $p(z') > p(z)$
    - or with probability $\frac{p(z')}{p(z)}$ only if $p(z') < p(z)$
  - sample either the new point (if accepted) or old point (again)
  - iterate

- Sample points z will → wander closer and closer to the "center", still jumping enough from time to time to sample the "whole space".
- → samples of will follow the distribution $p(z)$ (although consecutive samples are correlated)
- → normalisation of $p(z)$ not necessary for
  - sampling algorithm
  - determination of $E[f(z)]$



Gibbs Sampling

- just like the Metropolis algorithm apart from:
  - propose to jump only in 1-coordinate at the time
  - cycle through the coordinates.



- Note there are (few) conditions for arbitrary Markov chains to really sample the distribution. i.e. each point has to be “reachable” … which I’m not going to elaborate on ☺