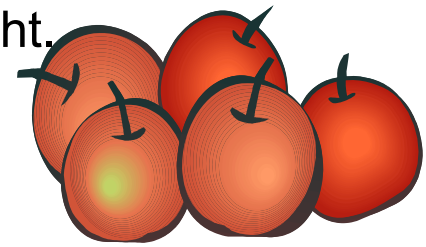
Confidence intervals

Example: take apples of different weight.
Need something to describe their distribution.

$\mu = 68\ g \qquad \sigma = 17\ g$



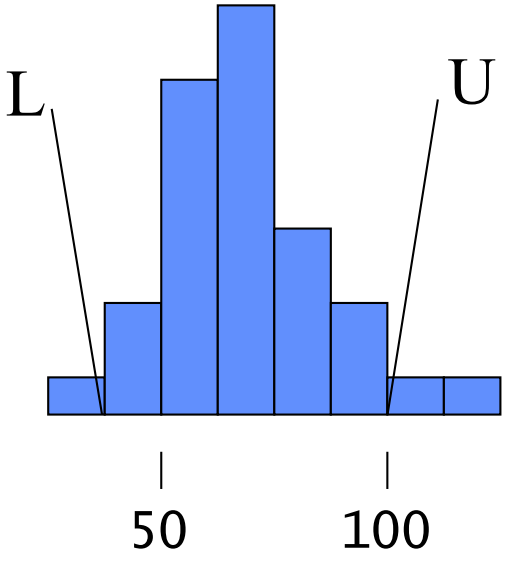
All weights between 24 and 167 g (tolerance)

90% lie between 50 and 100 g

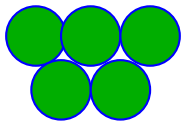
94% are less than 100 g

96% are more than 50 g

} confidence level statements



- can quote any level (68 %, 95 %, 99 % ...)
- upper or lower or two-sided ($x < \mathrm{U}$ $x > \mathrm{L}$ $\mathrm{L} < x < \mathrm{U}$)
- two-sided has even further choice (central, shortest …)

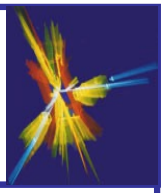


Particles with same weight, where ”distribution” is due to measurement uncertainty
What can one say about the mass m?

$\mu = 68 \quad \sigma = 17$

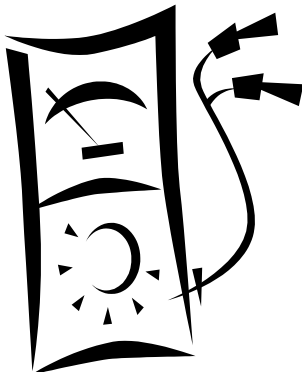
”m < 90” or ”m > 46” or ”40 < m < 96” @ 90 % CL

Each statement either always true or always false
Solution: refer to ensemble of statements



Example: a measurement without bias, gaussian error 0.1.

For a value x_{true}, measurement gives a value x_{meas} according to a Gaussian.



x_{meas} within 0.1 of x_{true} 68 % of the time $\Rightarrow$
x_{true} within 0.1 of x_{meas} 68 % of the time

$\Rightarrow x_{meas} - 0.1 < x_{true} < x_{meas} + 0.1$ @ 68 % CL

L < x < U @ 95 % confidence level (CL)

Statement belongs to an ensemble of statements of which 95 % are true. 95 % is the coverage.

NB! the statement not about x but about L &/or U.

Often best strategy is a pragmatic one

Ideal approach:

1. choose strategy
2. examine data
3. quote result

Pragmatic approach:

1. examine data
2. choose strategy
3. quote result

Example:

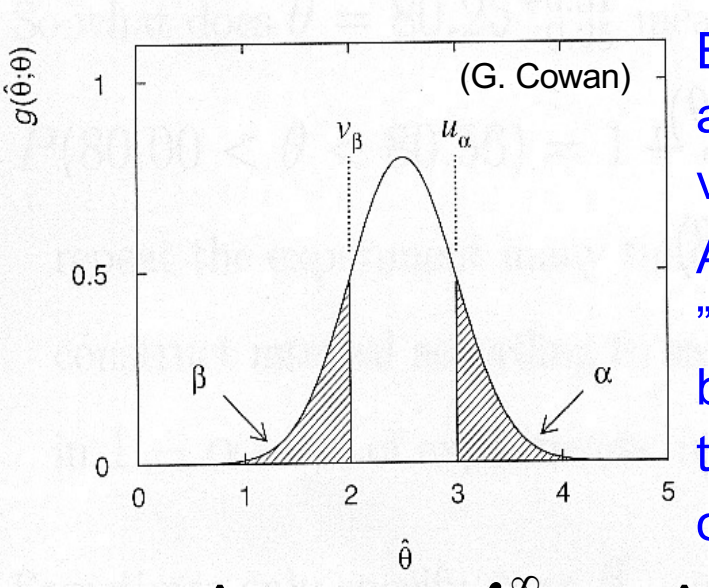
You have a background of 3.2

Observe 5 events?
Quote one-sided upper limit for signal
(9.27-3.2 = 6.07 at 90 % CL)

Observe 25 events?
Quote two-sided signal interval

Classical confidence intervals

n measurements of x that are used to evaluate estimator for a parameter θ. Obtained value $\hat{\theta}_{\text{obs}}$. Assume that by some means (e.g. analytic calculation or MC study) pdf $g(\hat{\theta};\theta)$ known, i.e. for given θ, know $g(\hat{\theta})$–distribution.



Example of pdf $g(\hat{\theta};\theta)$ for an estimator $\hat{\theta}$ for a given value of true parameter, θ.

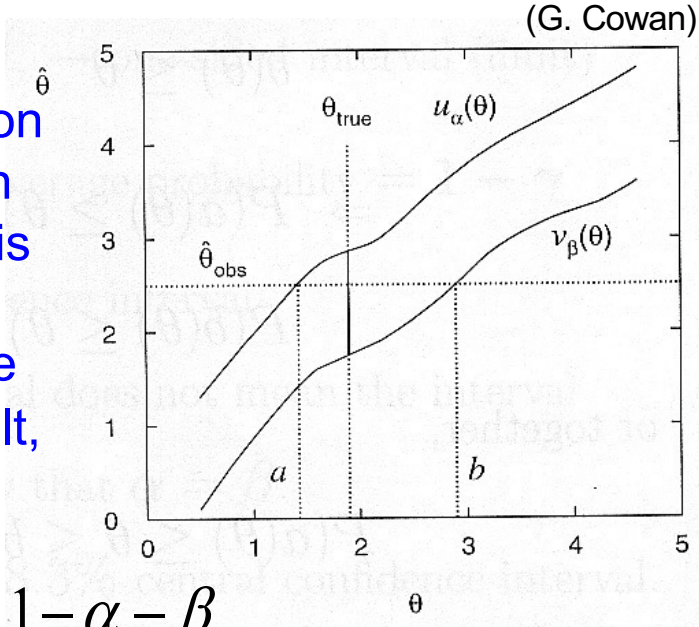
Able from $g(\hat{\theta};\theta)$ to specify "upper & lower tail probabilities", that corresponds to fixed probabilities α & β of observing $\hat{\theta} \geq u_\alpha$ & $\leq v_\beta$.

$$\alpha = P(\hat{\theta} \geq u_\alpha(\theta)) = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)\, d\hat{\theta} = 1 - G(u_\alpha(\theta);\theta)$$

$$\beta = P(\hat{\theta} \leq v_\beta(\theta)) = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)\, d\hat{\theta} = G(v_\beta(\theta);\theta)$$

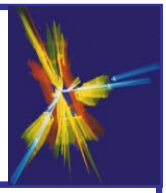
(G. Cowan)

Example of functions $u_\alpha(\theta)$ & $v_\beta(\theta)$ as a function of true value of θ. Region between the two curves is the "confidence belt".
The probability to find the estimator $\hat{\theta}$ inside the belt, regardless of value of θ:



$$P(v_\beta(\theta) \leq \hat{\theta} \leq u_\alpha(\theta)) = 1 - \alpha - \beta$$

Classical confidence intervals (cont.)



If $u_\alpha(\theta)$ & $v_\beta(\theta)$ monotonic functions of θ, which should be if $\hat{\theta}$ is a good estimator for θ, can determine their inverse:

$$\text{if } a(\hat{\theta}) \equiv u_\alpha^{-1}(\hat{\theta}) \text{ then } \hat{\theta} \geq u_\alpha(\theta) \rightarrow a(\hat{\theta}) \geq \theta$$

$$\text{if } b(\hat{\theta}) \equiv v_\beta^{-1}(\hat{\theta}) \text{ then } \hat{\theta} \leq v_\beta(\theta) \rightarrow b(\hat{\theta}) \leq \theta$$

The equations for the confidence belt can then be written

$$\left.\begin{aligned} P(a(\hat{\theta}) \geq \theta) &= \alpha \\ P(b(\hat{\theta}) \leq \theta) &= \beta \end{aligned}\right\} \Leftrightarrow \quad P(a(\hat{\theta}) \leq \theta \leq b(\hat{\theta})) = 1 - \alpha - \beta$$

Interval $[a(\hat{\theta}), b(\hat{\theta})]$ the confidence interval for θ at $1-\alpha-\beta$ confidence level (or coverage probability). Constructed intervals $[a,b]$ from several repetitive experiments (with similar sample size) include true value of parameter θ in a $1-\alpha-\beta$ fraction of the experiments. NB! a & b random values & defined regardless of true value of θ, by definition unknown. $\theta \in [a,b]$ either true or false in a frequentist view.

Often report interval $[a,b]$ as $\hat{\theta}^{+d}_{-c}$, i.e. $c = \hat{\theta} - a, d = b - \hat{\theta}$.

If only $\alpha(\beta)$ specified $\rightarrow$ one-sided confidence interval/limit.

Often choose $\alpha = \beta = \gamma/2 \rightarrow$ central confidence interval with $1-\gamma$ confidence level. NB! central confidence interval doesn't mean symmetric interval around $\hat{\theta}$, but only $\alpha = \beta$. Convention in physics: 68.3 % central confidence intervals.

Usually don't reconstruct confidence belt, but rather solve e.g. numerically for interval limits a & b. An equivalent procedure since $\hat{\theta}_{obs} = u_\alpha(\theta) = v_\beta(\theta)$

$$\alpha = \int_{\hat{\theta}_{\text{obs}}}^{\infty} g(\hat{\theta}; a)\, d\hat{\theta} = 1 - G(\hat{\theta}_{\text{obs}}; a)$$

$$\beta = \int_{-\infty}^{\hat{\theta}_{\text{obs}}} g(\hat{\theta}; b)\, d\hat{\theta} = G(\hat{\theta}_{\text{obs}}; b)$$

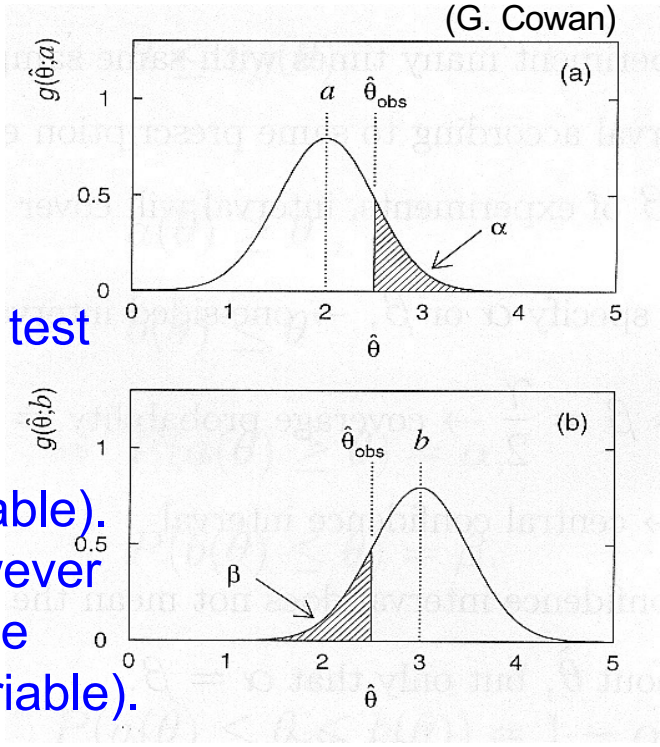
Classical confidence intervals (cont.)

a hypothetical value of θ such that $P(\hat{\theta} > \hat{\theta}_{\text{obs}}) = \alpha$

b hypothetical value of θ such that $P(\hat{\theta} < \hat{\theta}_{\text{obs}}) = \beta$

In case of a goodness-of-fit test hypothesis e.g. $\theta = a$ first specified & then e.g. α determined (α random variable). For confidence interval however α first specified & then value a determined (a random variable).

(G. Cowan)



Confidence interval for a Gaussian distributed estimator:

Important application of confidence intervals: estimators with Gaussian distributed mean θ & standard deviation $\sigma_{\hat{\theta}}$

Then the cumulative distribution function

$$G(\hat{\theta};\theta,\sigma_{\hat{\theta}}) = \int_{-\infty}^{\hat{\theta}} (2\pi\sigma_{\hat{\theta}}^2)^{-1/2} \exp\left(-(\hat{\theta}'-\theta)^2 / 2\sigma_{\hat{\theta}}^2\right) d\hat{\theta}'$$

Confidence interval $[a,b]$ for θ obtain by solving for a & b:

$$\alpha = 1 - G(\hat{\theta}_{\text{obs}};a,\sigma_{\hat{\theta}}) = 1 - \Phi\left((\hat{\theta}_{\text{obs}} - a)/\sigma_{\hat{\theta}}\right)$$

$$\beta = G(\hat{\theta}_{\text{obs}};b,\sigma_{\hat{\theta}}) = \Phi\left((\hat{\theta}_{\text{obs}} - b)/\sigma_{\hat{\theta}}\right)$$

$$\Phi(x) = \int_{-\infty}^{x} (2\pi)^{-1/2} e^{-x'^2/2} dx'$$ cumulative function of "standard" Gaussian.

$$\begin{cases} a = \hat{\theta}_{\text{obs}} - \sigma_{\hat{\theta}} \Phi^{-1}(1-\alpha) \\ b = \hat{\theta}_{\text{obs}} + \sigma_{\hat{\theta}} \Phi^{-1}(1-\beta) \end{cases}$$

where Φ^{-1} is the inverse of function Φ, i.e. the **quantile** of the standard Gaussian.

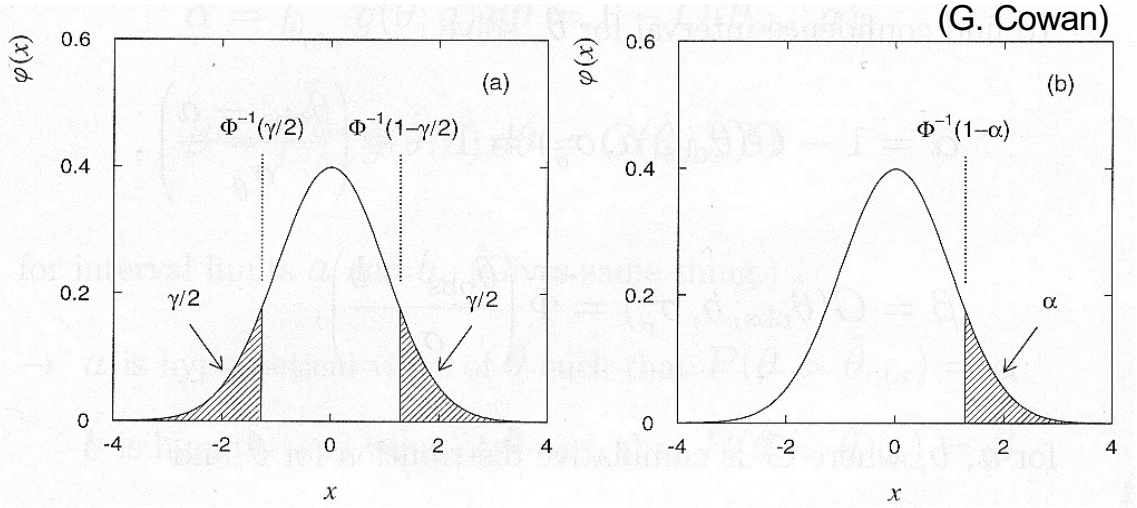
Confidence intervals for Gaussian distributed estimators

$$a = \hat{\theta}_{\text{obs}} - \sigma_{\hat{\theta}}\Phi^{-1}(1-\alpha)$$
$$b = \hat{\theta}_{\text{obs}} + \sigma_{\hat{\theta}}\Phi^{-1}(1-\beta)$$

$\Leftrightarrow$ $\Phi^{-1}(1-\alpha)$ & $\Phi^{-1}(1-\beta)$ represent how far a & b are from $\hat{\theta}_{obs}$ in units of standard deviations.

Standard Gaussian pdf $\Phi(x)$ shown for central & one-sided confidence intervals in terms of quantiles:

(G. Cowan)



Often take an integer number for the quantile.... but also sometimes take a round number for the confidence level.

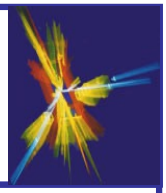
central		one-sided	
$\Phi^{-1}(1-\gamma/2)$	$1-\gamma$	$\Phi^{-1}(1-\alpha)$	$1-\alpha$
1	0.6827	1	0.8413
2	0.9544	2	0.9772
3	0.9973	3	0.9987

central		one-sided	
$1-\gamma$	$\Phi^{-1}(1-\gamma/2)$	$1-\alpha$	$\Phi^{-1}(1-\alpha)$
0.90	1.645	0.90	1.282
0.95	1.960	0.95	1.645
0.99	2.576	0.99	2.326

More values given by standard tables / statistics books or by using computer routines (e.g. `cdf/icdf in mathlab`).
For conventional 68.3 % central confidence interval one has $\alpha = \beta = \gamma/2$, with $\Phi^{-1}(1-\gamma/2) = 1$, i.e. one standard deviation. Hence for a Gaussian distributed estimator:

$$[a,b] = [\hat{\theta}_{\text{obs}} - \sigma_{\hat{\theta}}, \hat{\theta}_{\text{obs}} + \sigma_{\hat{\theta}}] \quad \Leftrightarrow \quad \hat{\theta}_{\text{obs}} \pm \sigma_{\hat{\theta}}$$

Even if $\sigma_{\hat{\theta}}$ not known, can use $\hat{\sigma}_{\hat{\theta}}$ instead. $\hat{\sigma}_{\hat{\theta}}$ generally acceptable if $\hat{\theta}$ a good estimator, (at least) if sample large.



Confidence interval for Poisson distributed estimator

Another common estimator is when outcome is a Poisson distributed variable n, then $\hat{\nu} = n$ & estimate $\hat{\nu}_{\text{obs}} = n_{\text{obs}}$

The pdf is then $f(n;\nu) = \nu^n e^{-\nu}/n!$

A complication for discrete variables is that functions $u_\alpha(\nu)$ & $v_\beta(\nu)$, defining the confidence belt, doesn't exist for all ν values, since $\hat{\nu}$ only take on discrete values. However the confidence interval can still be determined using our 2nd receipe by numerically solving for a & b once $\hat{\nu}_{\text{obs}}$ known.

$$\alpha = P(\hat{\nu} \geq \hat{\nu}_{\text{obs}}; a) = \sum_{n=n_{\text{obs}}}^{\infty} f(n;a) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} a^n e^{-a}/n!$$

$$\beta = P(\hat{\nu} \leq \hat{\nu}_{\text{obs}}; b) = \sum_{n=0}^{n_{\text{obs}}} f(n;a) = \sum_{n=0}^{n_{\text{obs}}} b^n e^{-b}/n!$$

Solution simplified if Poisson distribution can be related to the chi-square one (approximately true when ν > ~ 10):

$$\sum_{n=0}^{m} \frac{\nu^n}{n!} e^{-\nu} = \int_{2\nu}^{\infty} f_{\chi^2}(z; n_d = 2(m+1))\, dz = 1 - F_{\chi^2}(2\nu; n_d = 2(m+1)),$$

where F_{χ^2} is cumulative chi-square distribution for n_d d.o.f.

$$a = \tfrac{1}{2} F^{-1}_{\chi^2}\left(\alpha; n_d = 2n_{\text{obs}}\right) \quad b = \tfrac{1}{2} F^{-1}_{\chi^2}\left(1-\beta; n_d = 2(n_{\text{obs}}+1)\right)$$

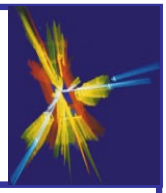
where $F^{-1}_{\chi^2}$ is the quantile (= inverse of cumulative distribution) of the chi-square distribution obtained either from standard tables or computer routines (e.g. `chi2inv` in `Mathlab`).

An important special case is when the estimate n_{obs} = 0:

$$\beta = \sum_{n=0}^{0} b^n e^{-b}/n! = e^{-b} \quad \rightarrow \quad b = -\ln\beta$$

for an upper limit at confidence level of $1-\beta$ = 95(90) % → $b \approx$ 3.0 (2.3) i.e. if ν = 3 (2.3), then probability to observe 0 events = 5 (10) %. No lower limit defined when n_{obs} = 0.

Confidence intervals for ML & LS estimators



Some useful Poisson lower & upper limits for measurement n_{obs}:

n_{obs}	lower limit a			upper limit b		
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\beta = 0.1$	$\beta = 0.05$	$\beta = 0.01$
0	–	–	–	2.30	3.00	4.61
1	0.105	0.051	0.010	3.89	4.74	6.64
2	0.532	0.355	0.149	5.32	6.30	8.41
3	1.10	0.818	0.436	6.68	7.75	10.04
4	1.74	1.37	0.823	7.99	9.15	11.60
5	2.43	1.97	1.28	9.27	10.51	13.11
6	3.15	2.61	1.79	10.53	11.84	14.57
7	3.89	3.29	2.33	11.77	13.15	16.00
8	4.66	3.98	2.91	12.99	14.43	17.40
9	5.43	4.70	3.51	14.21	15.71	18.78
10	6.22	5.43	4.13	15.41	16.96	20.14

Approximative confidence intervals for ML & LS estimators

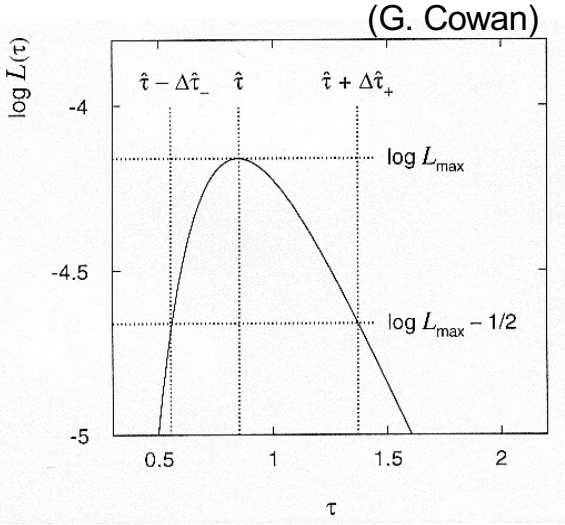
In case of non-Gaussian ML & LS estimators, an alternative way of extracting the confidence interval exist, when making use of the change of the functional value close to the ML/LS estimate. Works approximately even when the ln L is not parabolic i.e. the estimator non-Gaussian. Confidence interval then defined using:

$$\ln L(\hat{\theta}^{+d}_{-c}) = \ln L_{\max} - N^2/2 \qquad \text{or}$$

$$\chi^2(\hat{\theta}^{+d}_{-c}) = \chi^2_{\min} + N^2, \quad \text{since } \chi^2 = -2\ln L,$$

where $N = \Phi^{-1}(1-\gamma/2)$ is the standard Gaussian quantile corresponding to a central confidence level of $1-\gamma$, e.g. $N = 1 \Rightarrow 1-\gamma = 0.683$.

(G. Cowan)



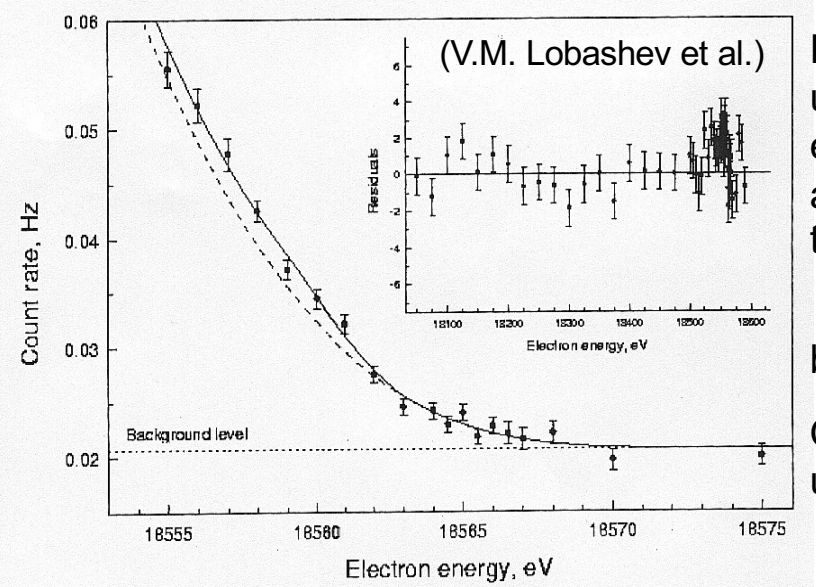
Our exponential ML example: ln L as a function of the lifetime τ for 50 observations. Estimate & confidence interval:

$$[\hat{\tau} - \Delta\hat{\tau}_-, \hat{\tau} + \Delta\hat{\tau}_+] \quad \text{or} \quad \hat{\tau} = 0.85^{+0.52}_{-0.30}$$

Here ihe interval sufficiently asymmetric to lead to different uncertainties for the + direction and the – direction.

Limits near a physical boundary

A complication occurs when an estimator can take unphysical values. Can happen if the estimator is of the form: $\hat{\theta} = x - y$, where x & y are random variables having measurement errors.
A typical example: the measurement of neutrino mass in β-decay. From relativistic kinematics: the mass (m) squared, $m^2 = E^2 - p^2$ so a non-zero mass will alter shape of high-energy end of the electron energy spectrum.



However any distorsion in upper end of the electron energy spectrum may cause a serious problem. A fit to the spectrum on the left give:

$$m_\nu^2 = -22 \pm 5\ \mathrm{eV}^2$$

but one knows that $m_\nu \geq 0$.

Question: how to quote an upper limit in such a case?

Essentially 2 possible ways: either simply shift any negative estimates to zero i.e. $\theta_{\text{upper limit}} = \max(\hat{\theta}_{\text{obs}}, 0) + \sigma_{\hat{\theta}} \Phi^{-1}(1-\beta)$

or to apply Bayes' theorem and impose a prior pdf $\pi(\theta)$ for the estimator that require all estimates to be physical (here ≥ 0):

$$1-\beta = \int_{-\infty}^{\theta_{\text{upper limit}}} p(\theta \mid \bar{x})\, d\theta = \int_{-\infty}^{\theta_{\text{upper limit}}} L(\bar{x} \mid \theta)\pi(\theta)\, d\theta \Big/ \int_{-\infty}^{+\infty} L(\bar{x} \mid \theta)\pi(\theta)\, d\theta$$

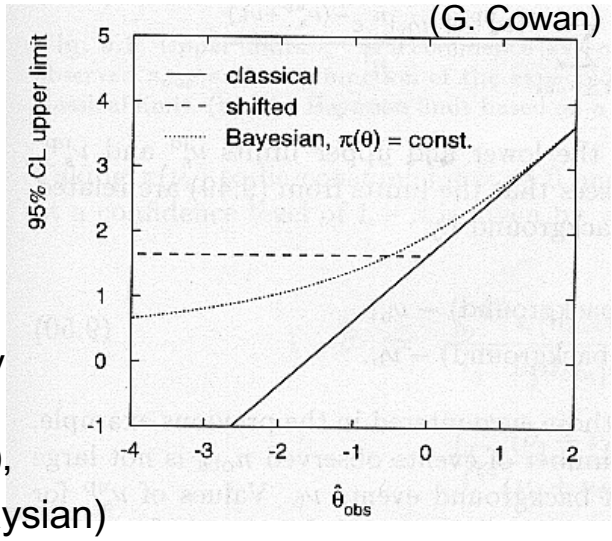
$$\text{with} \quad \pi(\theta) = \begin{cases} 0, & \theta < 0 \\ 1, & \theta \geq 0 \end{cases}$$

where x represents observed data, $p(\theta|x)$ the posterior pdf for θ & $L(x|\theta)$ the likelihood function of observed data x for a given θ. This prescription means effectively that one first normalizes the physical region to a unit area & then one integrates up to value $\theta_{\text{upper limit}}$ such that the fraction of physical region covered is $1-\beta$.

Limits near a physical boundary

Figure to the right shows upper limits at 95 % confidence level as a function of estimate for a Gaussian-distributed estimator with $\sigma_{\hat{\theta}} = 1.0$. The curves shown are for the classical (full line), shifted classical (dashed) & Bayesian method (dotted). In the neutrino mass example $m_\nu^2 = -22 \pm 5\ \mathrm{eV}^2$ would imply a one-sided 95 % CL upper limit of −14 eV² (classical, here non-physical!), +8 eV² (shifted classical) & +3 eV² (Baysian)

(G. Cowan)



($F_{gauss}(4.4\sigma) \approx 5.4 \cdot 10^{-6}$, $F^{-1}{}_{gauss}(0.05 \times 5.4 \cdot 10^{-6}) \approx 5.0\sigma \rightarrow -22+5.0\times5\ \mathrm{eV}^2 \approx +3\ \mathrm{eV}^2$)

Upper limit on the mean of a Poisson variable with background: Suppose the observed number of events n is a sum of desired signal events n_s & background events n_b; $n = n_s + n_b$. The sum of 2 Poisson variables is a Poisson variable as well, i.e. $\nu = \nu_s + \nu_b$.

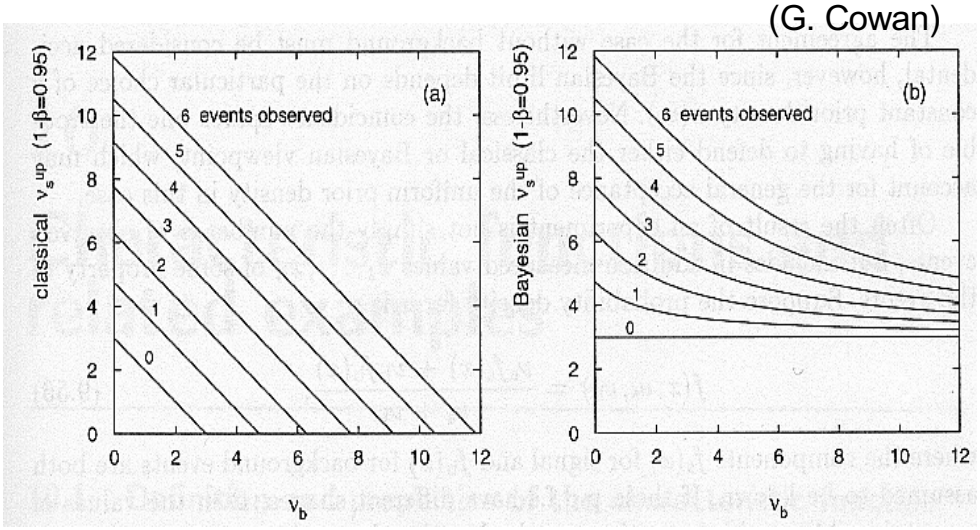
$$f(n;\nu_s,\nu_b) = \frac{(\nu_s+\nu_b)^n}{n!} e^{-(\nu_s+\nu_b)} \quad \rightarrow \quad L(n_{obs}\,|\,\nu_s) = \frac{(\nu_s+\nu_b)^{n_{obs}}}{n_{obs}!} e^{-(\nu_s+\nu_b)}$$

here can a similar Bayesian approach be used for the upper limit to avoid negative ν_s solutions ($\pi(\nu_s) = 1$ for $\nu_s \geq 0$ & $\pi(\nu_s) = 0$ for $\nu_s < 0$).

$$\beta = e^{-(\nu_s^{up}+\nu_b)} \sum_{n=0}^{n_{obs}} \frac{(\nu_s^{up}+\nu_b)^n}{n!} \Big/ e^{-\nu_b} \sum_{n=0}^{n_{obs}} \frac{\nu_b^n}{n!}$$

numerically solve for ν_s^{up} in this equation

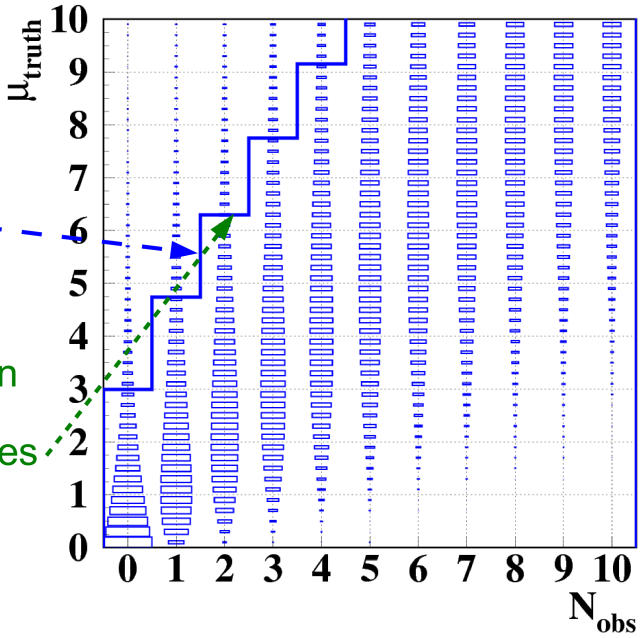
(G. Cowan)



Upper limit ν_s^{up} at 95 % confidence level for different number of observed events n_{obs} as a function expected number of background events ν_b. figures: classical (left) & Bayesian (right).

Frequentist upper limit, Poisson data

- Neyman construction: scan μ and find for each μ the max. N_{obs} with $P(N \in [N_{obs}, \infty]) \geq 95\%$ → vertical lines
- For given N_{obs} find largest μ, where N_{obs} is just contained in the interval → horizontal lines → μ_{limit}



- Note: discrete N_{obs} but continuous m → limit steps

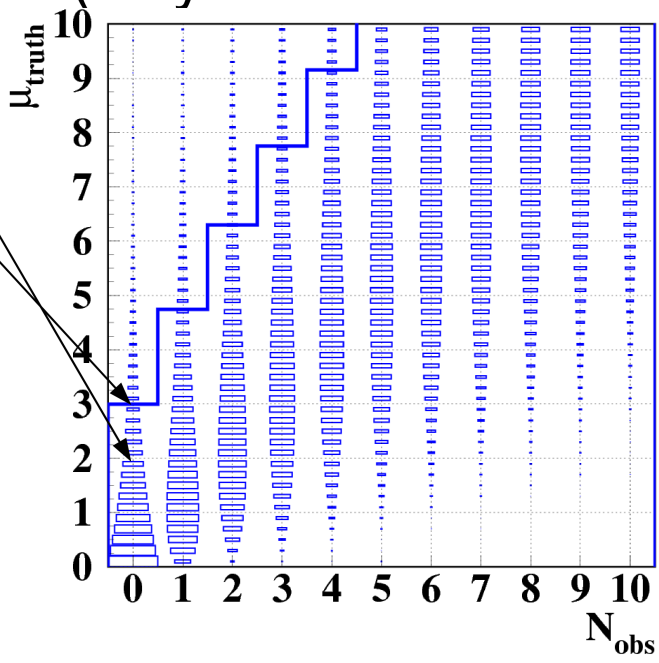
Exercise 1 (Neyman construction)

(a)

μ	N_{obs}	α = P(N<N_{obs})
2	1	0.14
3	1	0.05
5	2	0.04
10	5	0.03

(b)

N_{obs}	μ_{limit}
0	3.0
2	6.3
10	17.0
100	118.1



Probability: $\sum_{N=0}^{N_{obs}-1} \frac{e^{-\mu}(\mu)^N}{N!} = \alpha =$ TMath::Prob$(2*\mu, 2*N_{obs})$

Inverse function: $\mu_{limit} =$ TMath::ChisquareQuantile$(1-\alpha, 2*(N_{obs}+1))/2$

Bayesian upper limits

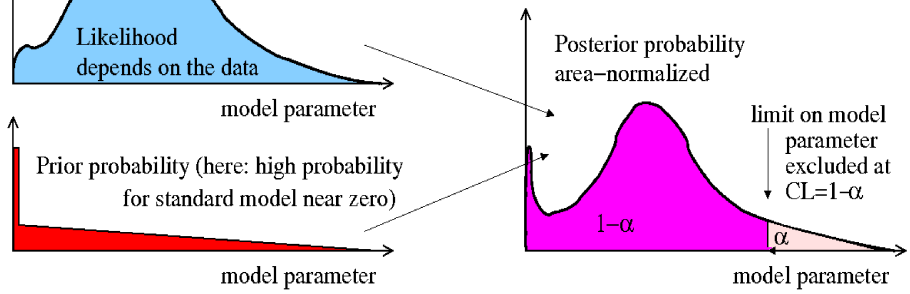
- Bayesian limit: exclude a set of theories, such that the posterior probability of the excluded theories is 1-CL

Enumerator: integrate over allowed theories

$$CL = P(\mu \le \mu_{limit} | N_{obs}) = 1 - \alpha = \frac{\int_0^{\mu_{limit}} L(N_{obs}|\mu)\pi(\mu)\,d\mu}{\int_0^{\infty} L(N_{obs}|\mu)\pi(\mu)\,d\mu}$$

$\pi(\mu)$: prior probability of the model μ
$L(N_{obs}|\mu)$: Likelihood

Denominator: integrate all theories (normalisation)



Bayesian limit: integrate over models, fixed N_{obs}

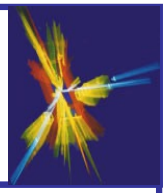
Frequentist limit: integrate over N_{obs}, test each model

Bayesian limit exercise

- Exercise 2a: Bayesian limit for N_{obs}=0,2,10,100 (flat prior)
  (use Root macro)
- Exercise 2b: use a prior P(μ)=μ, N_{obs}={0,2,10,100}

- For this example: Bayes flat=Frequentist
- Prior P(μ)=μ gives more conservative limit

	frequentist	Bayes flat	Bayes P(μ)=μ
N_{obs}	μ_{limit}	μ_{limit}	μ_{limit}
0	3.0	3.0	4.7
2	6.3	6.3	7.8
10	17.0	17.0	18.2
100	118.1	118.2	119.3



Limits with background

- Expected number of events:

  $\mu = s + b$, s, b: signal and background event yield, respectively

  - s=0: standard model
  - s>0: new physics
- Assume background known. What is the limit on the signal?
- Frequentist: set limit on μ, then subtract b
- Bayesian: use prior probability which is zero for s<0

Exercise 3 (limit with background)

- Calculate Frequentist and Bayesian limits for N_{obs}={0,2} and b={0.5,2.0,3.5}

Poisson parameter: $\mu = s + b$

	b=0.5		b=2.0		b=3.5	
	N_{obs}=0	N_{obs}=2	N_{obs}=0	N_{obs}=2	N_{obs}=0	N_{obs}=2
Bayesian	3.0	5.8	3.0	4.8	3.0	4.3
Frequentist	2.5	5.8	1.0	4.3	-0.5	2.8

- Problem for Frequentist limit, N_{obs}=0 and b=3.5:

  limit excludes all signal above s=-0.5.

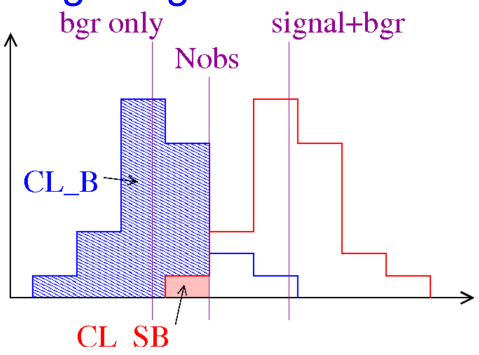
  Even the “standard model” s=0 is excluded

The CL_s (modified frequentist) method

- Frequentist limit: $1-CL \geq \alpha = \mathrm{CL}_{SB} = P(N \leq N_{\mathrm{obs}}; \mu = s+b)$
- CL_s limit: $1-CL \geq \mathrm{CL}_S = \frac{\mathrm{CL}_{SB}}{\mathrm{CL}_B} = \frac{P(N \leq N_{\mathrm{obs}}; \mu = s+b)}{P(N \leq N_{\mathrm{obs}}; \mu = b)}$
- Probability is normalized to background probability
- $CL_B \leq 1 \rightarrow CL_S \geq CL_{SB}$: same α requires larger signal

  Limit is "conservative"
- For zero signal: $CL_s = 1$

  → zero signal is never excluded

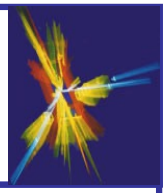


Exercise 5 (CL_s method)

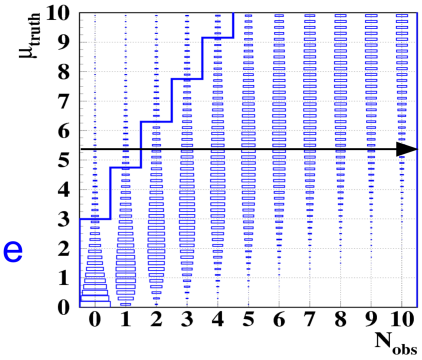
- Frequentist limit: $1-CL \geq \alpha = \mathrm{CL}_{SB} = P(N \leq N_{\mathrm{obs}}; \mu = s+b)$
- CL_s limit: $1-CL \geq \mathrm{CL}_S = \frac{\mathrm{CL}_{SB}}{\mathrm{CL}_B} = \frac{P(N \leq N_{\mathrm{obs}}; \mu = s+b)}{P(N \leq N_{\mathrm{obs}}; \mu = b)}$

	b=0.5		b=2.0		b=3.5	
	$N_{obs}=0$	$N_{obs}=2$	$N_{obs}=0$	$N_{obs}=2$	$N_{obs}=0$	$N_{obs}=2$
Bayesian	3.0	5.8	3.0	4.8	3.0	4.3
Frequentist	2.5	5.8	1.0	4.3	-0.5	2.8
CL_s	3.0	5.8	3.0	4.8	3.0	4.3
Expected	3.3		4.2		4.9	

- For this example, CL_s is identical to Bayesian (with flat prior)



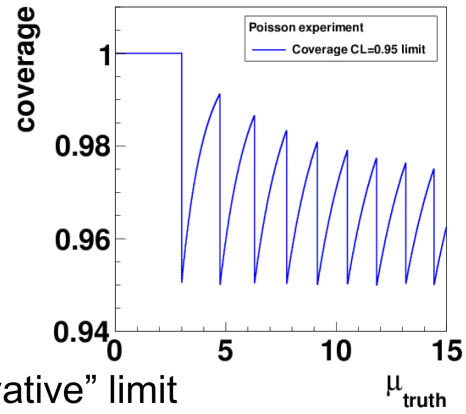
Coverage



- Coverage: given the limit procedure, calculate for each μ_{truth} probability to include the true value in the Confidence interval
- Poisson example (exercise 2)

$$P_{incl}(\mu_{truth}) = \sum_N P_{\mu,truth}(N)\Theta(\mu_{truth} \le \mu_{limit}(N))$$

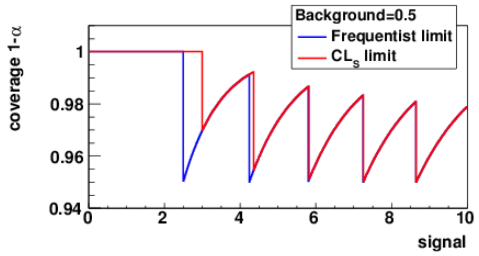
$$\text{where } \Theta(\mu_{truth} \le \mu_{limit}) = \begin{cases} 1 \text{ if } \mu_{truth} \le \mu_{limit} \\ 0 \text{ otherwise} \end{cases}$$

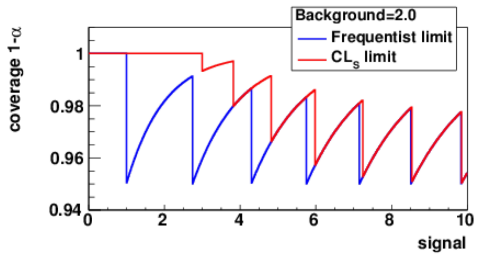


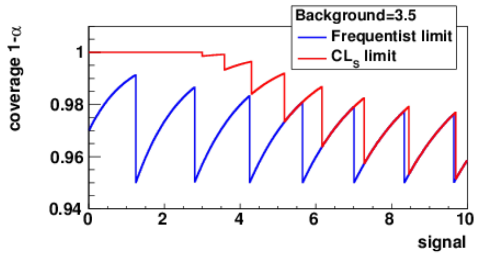
- coverage=0.95: exact coverage
- coverage<0.95: undercoverage
- coverage>0.95: overcoverage, “conservative” limit
- “Simple” Poisson case: overcoverage (discrete measurement)

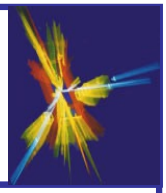
Summary of CLs pros & cons

- CL_s method avoids problem with limits better than the experiments sensitivity
- Limits on s always > 0
- Disadvantage: CL_s method is conservative, in particular for small signals









Treatment of nuisance parameters

"Coverage" key thing for confidence intervals and levels, "undercoverage" i.e.

$$P(x \in [x_1, x_2]) < 1-\alpha \quad \text{for some } x$$
$$\text{even if } P(x \in [x_1, x_2]) = 1-\alpha \text{ required}$$

should be avoided, at least if significant. "overcoverage"

$$P(x \in [x_1, x_2]) > 1-\alpha \quad \text{for some } x$$
$$\text{even if } P(x \in [x_1, x_2]) = 1-\alpha \text{ required}$$

"conversative" and annoying (result appear less sensitive than measurement really is) but not problematic. Often under- or overcoverage originate from "approximative" methods used to obtain confidence interval or level.

A nuisance parameter ("haittaparametri"): a parameter of no interest to measure but that still affects the resulting measurement, e.g. background, model parameters etc...

How include nuisance parameters and especially their uncertainties in confidence interval and levels?

Use the likelihood ratio (λ) & its correspondance with χ^2 distribution ($-2\log\lambda \approx \chi^2$ with k d.o.f. = parameters fixed in λ) in the large sample limit, may cause undercoverage in certain cases for small numbers. Advantages: can be generalized to arbritrary number of nuisance parameters and also the uncertainties on them can easily be included.

Then either integrate over nuisance parameters see e.g. R.D. Cousins and V.L. Highland: Nucl. Instr. Meth. A 320 (1992) 331 or take the nuisance parameters corresponding to maximal value of likelihood function see e.g. W.A. Rolke, A.M. Lopez and J. Conrad: Nucl. Instr. Meth. A 551 (2005) 493–503.

Limits and confidence intervals in presence of nuisance parameters (W.A. Rolke, A.M. Lopez and J. Conrad: Nucl. Instr. Meth. A 551 (2005) 493.)

Likelihood ratio: $$\lambda(\bar{\theta}_0 \mid \bar{x}) = \frac{\max\{L(\bar{\theta}_0, \bar{\pi} \mid \bar{x}); \bar{\pi}\}}{\max\{L(\bar{\theta}, \bar{\pi} \mid \bar{x}); \bar{\theta}, \bar{\pi}\}}$$

$\bar{\theta}$ = parameters of interest & $\bar{\pi}$ = nuisance parameters

Example: assume measurement of events in 2 region; signal region: x events; background region: y events.

$x \sim \text{Poisson}(\mu + b)$ $\quad y \sim \text{Poisson}(\tau b)$ $\quad \Leftrightarrow$

$$f(x, y \mid \mu, b) = \frac{(\mu+b)^x}{x!} e^{-(\mu+b)} \frac{(\tau b)^y}{y!} e^{-\tau b} \quad \Rightarrow \quad \begin{aligned} \hat{\mu} &= x - y/\tau \\ \hat{b} &= y/\tau \end{aligned}$$

Fix μ & maximize over b (local maximum $df/db = 0$) $\Longrightarrow$

$$\hat{b}(\mu) = \left(x + y - (1+\tau)\mu + \sqrt{(x + y - (1+\tau)\mu)^2 + 4(1+\tau)y\mu}\right) \Big/ 2(1+\tau)$$

"Profile likelihood":

$$\lambda(\mu \mid x, y) = \frac{L(\mu, \hat{b}(\mu) \mid x, y)}{L(\hat{\mu}, \hat{b} \mid x, y)}$$

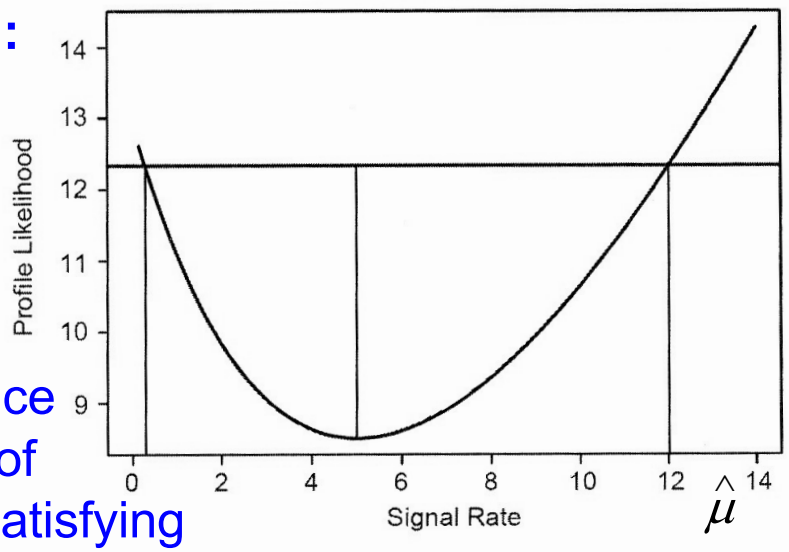


Fig. 1. Example of the $-2 \log \lambda$ curve. This is the case $x = 8$, $y = 15$ and $\tau = 5.0$. We find the 95% confidence interval to be (0.28,12.02).

$100(1-\alpha)$ % confidence interval $\Leftrightarrow$ increase of $-2\log\lambda$ by χ^2_0 value satisfying $P(x \geq \chi^2_0) = \alpha$ for a χ^2 distribution with number d.o.f. = 1.

95 % CL $\Leftrightarrow \chi^2_0 = 3.8414 = (1.96)^2$

Treatment of nuisance parameters

How treat nuisance parameters and their uncertainties?
Example: treatment of efficiency & background. Assume efficiency z as binomial obtained from m MC events.

$x \sim \text{Poisson}(e\mu + b)$ $y \sim \text{Poisson}(\tau b)$ $z \sim \text{Binomial}(m,e)$ $\Leftrightarrow$

Fix μ and differentiate $\log L$ to find maximum $\Rightarrow$

$$\partial \log L(\mu, b, e \,|\, x, y, z) / \partial b = x/(e\mu + b) - 1 + y/b - \tau = 0$$

$$\partial \log L / \partial e = x\mu/(e\mu + b) - \mu + z/e - (m - z)/(1 - e) = 0$$

system of equations to be solved numerically to obtain profile likelihood curve as function of signal rate μ only.

Example 2: Background & efficiency assumed Gaussian

$x \sim \text{Poisson}(e\mu + b)$ $y \sim \text{Gaussian}(b, \sigma_b)$ $z \sim \text{Gaussian}(e, \sigma_e)$ $\Leftrightarrow$

$$\frac{\partial \log L}{\partial b} = \frac{x}{e\mu + b} - 1 + \frac{(y - b)}{\sigma_b^2} = 0$$

$$\frac{\partial \log L}{\partial e} = \frac{x\mu}{e\mu + b} + \frac{(z - e)}{\sigma_e^2} - \mu = 0$$

Solvable analytically but numerical solution easier. Aim: get profile likelihood as function of μ only.

Note: profile likelihood method sometimes unable to return limits e.g. when value of $-2\log\lambda$ below value defining confidence level for all signal rate μ values.

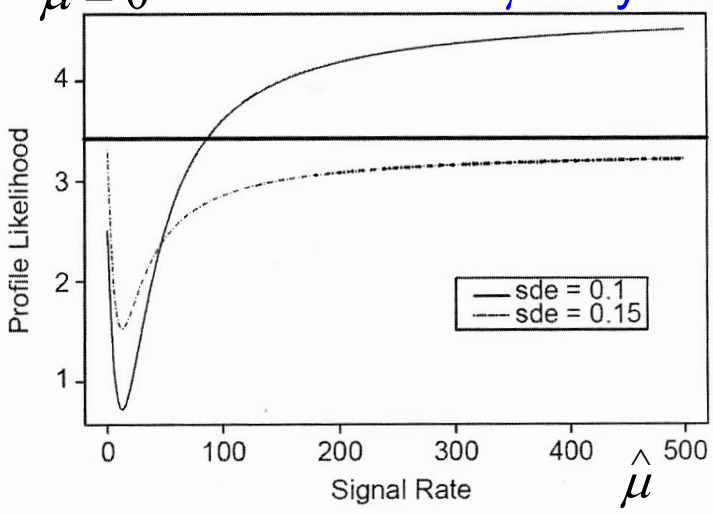
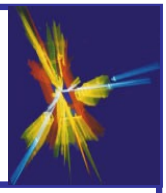


Fig. 3. An illustration of why the profile likelihood method does sometimes not return a limit. Here $x = 5$, $y = 2.5$, $\sigma_b = 0.4$ and $z = 0.2$. If $\sigma_e = 0.1$ the curve moves above the required level (for a 90% confidence interval) and the upper limit is 85.9 but if $\sigma_e = 0.15$ the curve stays below the level and no limit is found.



Prototype search analysis

Search for signal in a region of phase space; result is histogram of some variable x giving numbers:

$$\mathbf{n} = (n_1, \ldots, n_N)$$

Assume the n_i are Poisson distributed with expectation values

$$E[n_i] = \mu s_i + b_i$$

strength parameter

where

$$s_i = s_{\text{tot}} \int_{\text{bin } i} f_s(x; \boldsymbol{\theta}_s)\, dx\,, \qquad b_i = b_{\text{tot}} \int_{\text{bin } i} f_b(x; \boldsymbol{\theta}_b)\, dx\,.$$

signal background

Often also have a subsidiary measurement that constrains some of the background and/or shape parameters:

$$\mathbf{m} = (m_1, \ldots, m_M)$$

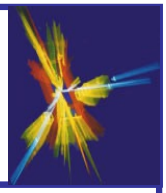
Assume the m_i are Poisson distributed with expectation values

$$E[m_i] = u_i(\boldsymbol{\theta})$$

nuisance parameters ($\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_b$,b_{tot})

Likelihood function is

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$



The profile likelihood ratio

Base significance test on the profile likelihood ratio:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

maximizes L for specified μ

maximize L

Define critical region of test of μ by the region of data space that gives the lowest values of $\lambda(\mu)$.

Important advantage of profile LR is that its distribution becomes independent of nuisance parameters in large sample limit.

Test statistic for discovery

Suppose relevant alternative to background-only ($\mu = 0$) is $\mu \geq 0$.

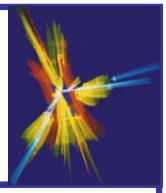
So take critical region for test of $\mu = 0$ corresponding to high q_0 and $\hat{\mu}$ > 0 (data characteristic for $\mu \geq 0$).

That is, to test background-only hypothesis define statistic

$$q_0 = \begin{cases} -2 \ln \lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

i.e. here only large (positive) observed signal strength is evidence against the background-only hypothesis.

Note that even though here physically $\mu \geq 0$, we allow $\hat{\mu}$ to be negative. In large sample limit its distribution becomes Gaussian, and this will allow us to write down simple expressions for distributions of our test statistics.



Intervals/limits using profile likelihood

Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554

Distribution of q_0 in large-sample limit

Assuming approximations valid in the large sample (asymptotic) limit, we can write down the full distribution of q_0 as

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]$$

The special case $\mu' = 0$ is a "half chi-square" distribution:

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2}$$

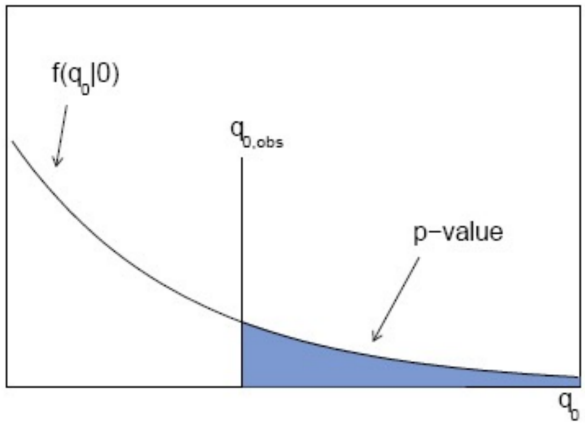
In large sample limit, $f(q_0|0)$ independent of nuisance parameters; $f(q_0|\mu')$ depends on nuisance parameters through σ.

p-value for discovery

Large q_0 means increasing incompatibility between the data and hypothesis, therefore p-value for an observed $q_{0,\text{obs}}$ is

$$p_0 = \int_{q_{0,\text{obs}}}^{\infty} f(q_0|0)\, dq_0$$

use e.g. asymptotic formula



From p-value get equivalent significance,

$$Z = \Phi^{-1}(1 - p)$$

Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554

Cumulative distribution of q_0, significance

From the pdf, the cumulative distribution of q_0 is found to be

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)$$

The special case $\mu' = 0$ is

$$F(q_0|0) = \Phi\left(\sqrt{q_0}\right)$$

The *p*-value of the $\mu = 0$ hypothesis is

$$p_0 = 1 - F(q_0|0)$$

Therefore the discovery significance Z is simply

$$Z = \Phi^{-1}(1 - p_0) = \sqrt{q_0}$$

Monte Carlo test of asymptotic formula

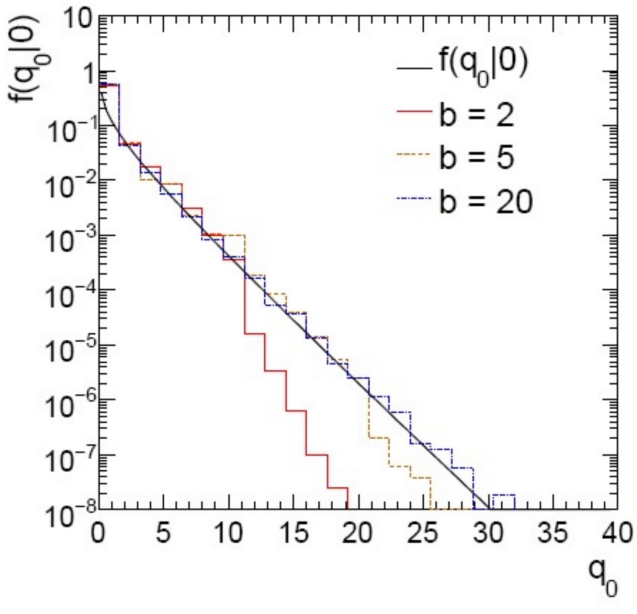
$$n \sim \text{Poisson}(\mu s + b)$$
$$m \sim \text{Poisson}(\tau b)$$

μ = param. of interest

b = nuisance parameter

Here take s known, $\tau = 1$.

Asymptotic formula is good approximation to 5σ level ($q_0 = 25$) already for $b \sim 20$.



Intervals/limits using profile likelihood

Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554

Test statistic for upper limits

For purposes of setting an upper limit on μ use

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \le \mu \\ 0 & \hat{\mu} > \mu \end{cases} \qquad \text{where} \qquad \lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

I.e. when setting an upper limit, an upwards fluctuation of the data is not taken to mean incompatibility with the hypothesized μ :

From observed q_μ find p-value: $$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu$$

Large sample approximation: $$p_\mu = 1 - \Phi\left(\sqrt{q_\mu}\right)$$

To find upper limit at CL = 1–α, set $p_\mu = \alpha$ and solve for μ.

Monte Carlo test of asymptotic formulae

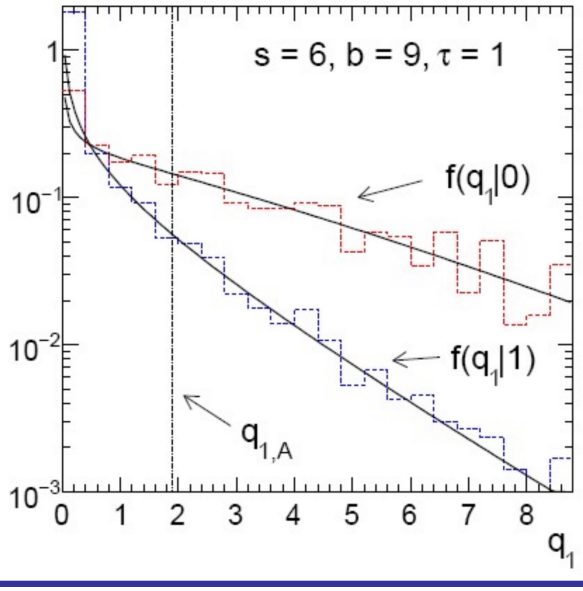
Consider again $n \sim$ Poisson($\mu s + b$), $m \sim$ Poisson(τb)
Use q_μ to find p-value of hypothesized μ values.

E.g. $f(q_1|1)$ for p-value of $\mu = 1$.

Typically interested in 95% CL, i.e., p-value threshold = 0.05, i.e., $q_1 = 2.69$ or $Z_1 = \sqrt{q_1} = 1.64$.

Median[$q_1|0$] gives "exclusion sensitivity".

Here asymptotic formulae good for $s = 6$, $b = 9$.

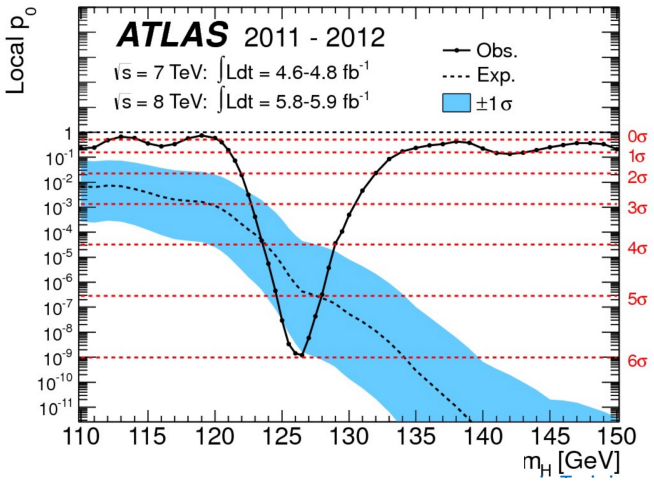


How to read the p_0 plot

The "local" p_0 means the p-value of the background-only hypothesis obtained from the test of $\mu = 0$ at each individual m_H, without any correct for the Look-Elsewhere Effect.

The "Expected" (dashed) curve gives the median p_0 under assumption of the SM Higgs (μ = 1) at each m_H.

ATLAS, Phys. Lett. B 716 (2012) 1-29



The blue band gives the width of the distribution ($\pm 1\sigma$) of significances under assumption of the SM Higgs.

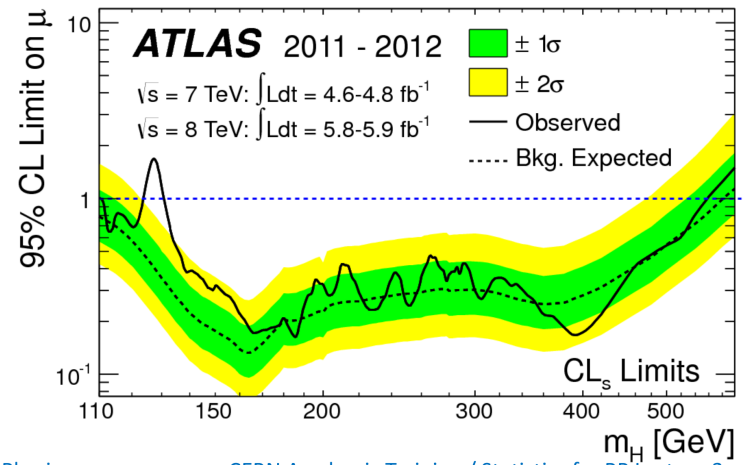
How to read the green and yellow limit plots

For every value of m_H, find the upper limit on μ.

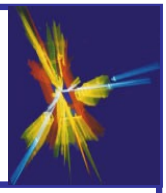
Also for each m_H, determine the distribution of upper limits μ_{up} one would obtain under the hypothesis of μ = 0.

The dashed curve is the median μ_{up}, and the green (yellow) bands give the $\pm$ 1σ (2σ) regions of this distribution.

ATLAS, Phys. Lett. B 716 (2012) 1-29



Unfolding



Consider random variable y. Goal: determine pdf $f(y)$
Measured pdf $f_{\text{meas}}(y)$ distorted from true pdf $f(y)$ due to:

- ”limited” resolution, i.e. measured y –value not identical to the true y –value due to experimental effects ($f(y)$ ”smeared out”).
- efficiency, i.e. the fraction of true events really detected. Often depends either directly or indirectly on the y–value.
- ”background/noise”, i.e. events not being true events (under study). Background/noise amount depends often on y–value.
- ....

The procedure of correcting measured distribution to get ”true” distribution is unfolding (”deconvolution”/”unsmearing”).

NB! In many cases no need to unfold measured distribution, if prediction can be distorted for experimental effects with e.g. MC.

Good reasons for unfolding a measurement:

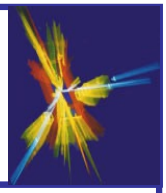
- comparisons with other experiments & (”new”) predictions (possible even after many years has passed from actual measurement).
- true distribution needed for the extraction of some parameter(s).
- ”image” reconstruction (in medicin, astronomy, crystallography ...)

Unfolding may seem a triviality but for many real distributions making a correct unfolding can be a rather complicated issue.

Methods of unfolding:

- unfolding by inversion of ”response matrix”.
- unfolding using correction factors.
- regularized unfolding.

Unfolding



Unfolding problem more easily viewed in terms of histograms.
Works irrespectively of the specific functional form of pdf $f(y)$.

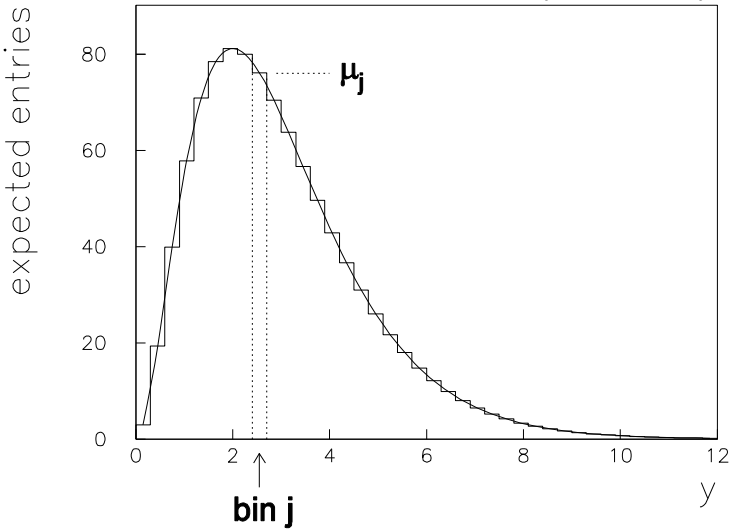
(G. Cowan)

Probability to find y in bin j, p_j

$$p_j = \int_{\text{bin } j} f_{\text{true}}(y)dy \quad j = 1,\ldots,M$$

"True histogram":

$$\mu_j = \mu_{\text{tot}} p_j$$



Goal: construct estimators for μ_j (or p_j) $\rightarrow$ # of parameters = M.
Problem: measured y has an uncertainty $\rightarrow$ migration btwn bins.

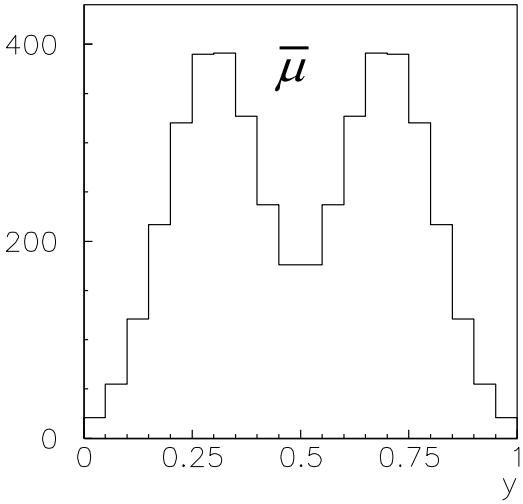
Effect of measurement errors: y = true & x = measured value.

$$f_{\text{meas}}(x) = \int R(x \mid y) f_{\text{true}}(y)dy \quad \Rightarrow \quad \nu_i = \sum_{j=1}^{M} R_{ij}\mu_j, i = 1,..,N$$

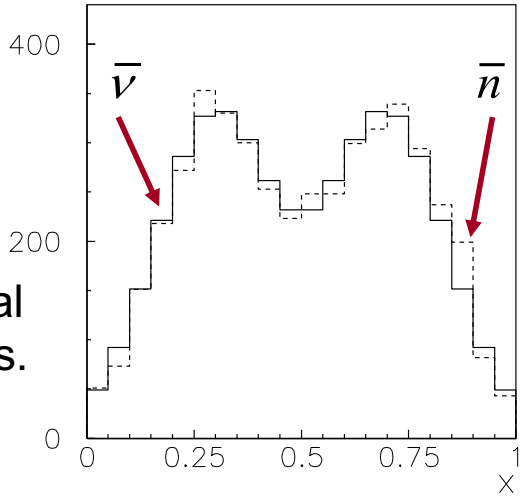
R_{ij} = P(observed in bin i | true value in bin j) = **response matrix**,
ν_i = expectation value for content of bin i for observed histogram.

$$\text{Data} : \vec{n} = (n_1,\ldots,n_N), \quad \text{where } \nu_i = E[n_i]$$

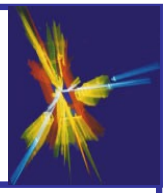
(G. Cowan)



NB! μ_i & ν_i constants, n_i subject to statistical fluctuations.



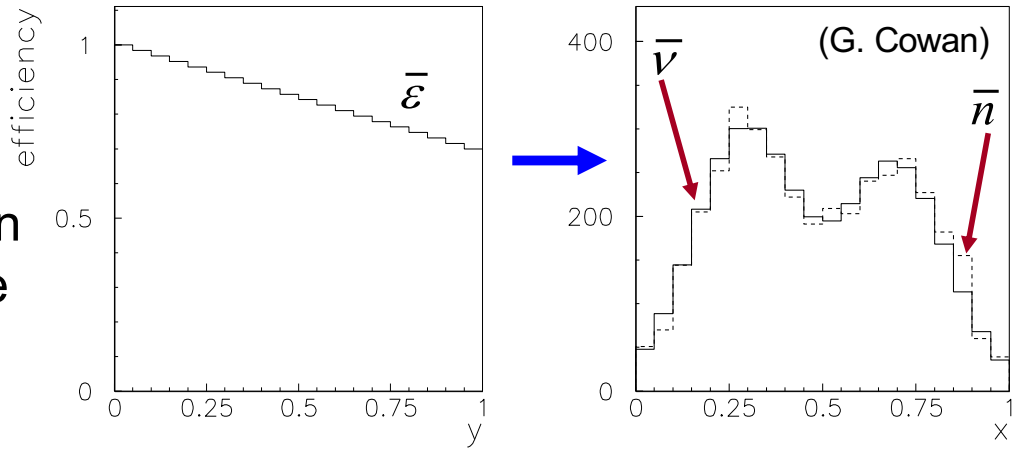
Unfolding



Distribution can also be distorted due to efficiency & background.

P(observed anywhere|true value in bin i) =
$\sum_{j=1}^{M} R_{ij}$= ε_i(efficiency of bin i)

N.B. ε_i depends on bin i of true histogram.



Occusionally observe event even though no true (i.e. those studied) event occured (”background/noise”).

$\Longrightarrow$

$$\nu_i = \sum_{j=1}^{M} R_{ij}\mu_j + \beta_i$$

β_i = expected # of background/noise events in bin i.
Obtained from e.g. calibration runs or MC simulation.
NB! Any uncertainties in R_{ij}, ε_i or β_i lead to systematics.

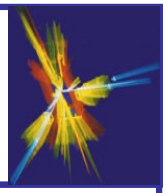
Unfolding problem can be summarized in one equation:

$$E[\bar{n}] = \bar{\nu} = \mathrm{R}\,\bar{\mu} + \bar{\beta}$$

To find estimators for $\bar{\mu}$ need either to know probability distribution the observed data or its covariance matrix:

$$e.g. \quad P(n_i, \nu_i) = \nu_i^{n_i} e^{-\nu_i} / n_i! \,(\mathrm{Poisson}) \quad \text{or} \quad V_{ij} = \mathrm{cov}[n_i, n_j]$$

to construct the likelihood function or the χ^2, respectively.



Comment on R matrix:

$$E[\bar{n}] = \bar{\nu} = \mathrm{R}\,\bar{\mu} + \bar{\beta}$$

If we define:

$s(x|y)$ = conditional probability for observing x in case of truth being y ("resolution function").

$$\int s(x \mid y)dx = 1$$

$$\nu_i = \mu_{TOT} \int_{bin\ i} dx \int s(x \mid y)\varepsilon(y) f_{\text{true}}(y)dy$$

$$\nu_i = \sum_{J=1}^{J=M} \frac{\int_{bin\ i} dx \int_{bin\ j} s(x \mid y)\varepsilon(y) f_{\text{true}}(y)dy}{(\mu_J / \mu_{TOT})} \mu_J$$

$$\nu_i = R_{ij}\mu_j$$

NB! Response matrix R depends on $f_{\text{true}}(y)$.

Dependence cancels if $s(x|y)$ & $\varepsilon(y)$ ~ constant over any bin of y:

$$R_{ij} = \int_{bin\ i} dx \int_{\text{bin } j} s(x \mid y)\varepsilon(y) f_{\text{true}}(y)dy \Big/ \int_{\text{bin } j} f_{\text{true}}(y)dy$$

Response matrix element R_{ij} conditional probability that an event found with measured value x in bin i given that the true value y was in bin j. The off-diagonal elements responsible of smearing true spectrum. R determined by calibration where true value y known a priori or from MC simulations where experimental effects are introduced.

Inverting response matrix

Most obvious (but often not optimal) method of unfolding consists of directly inverting response matrix (if possible).

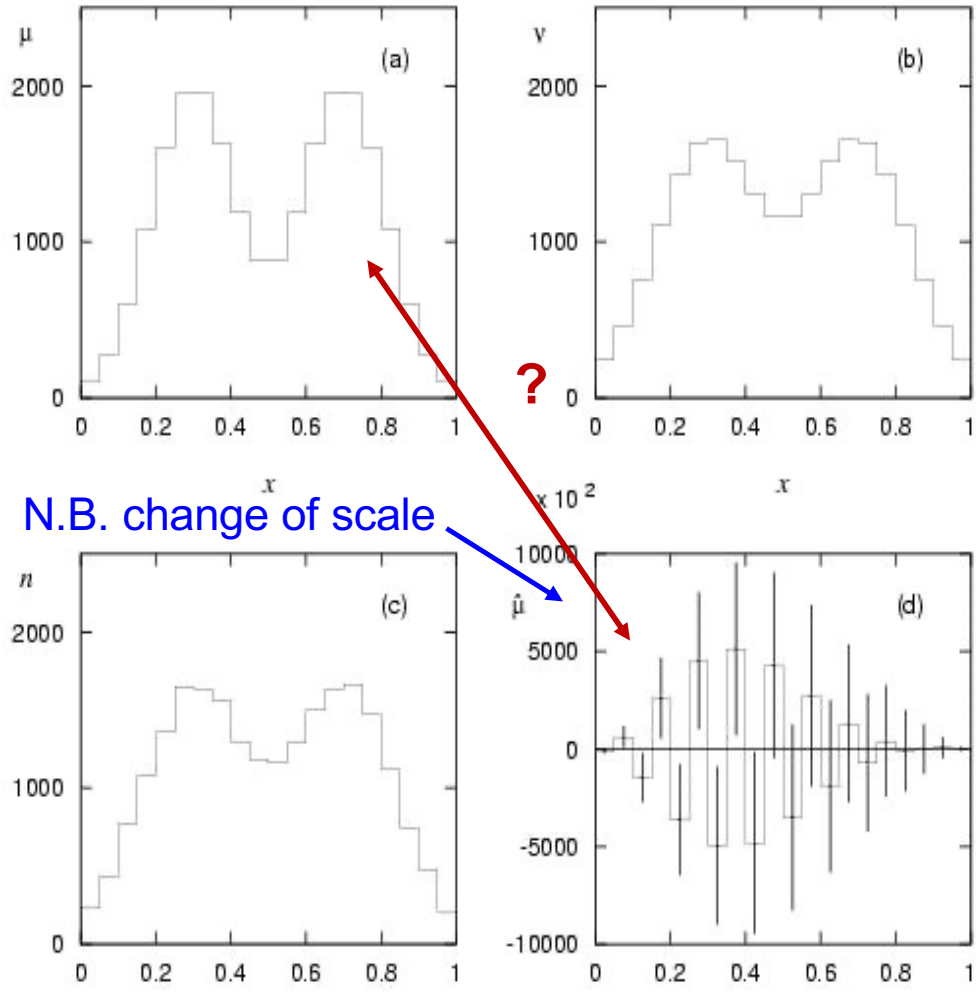
$$\text{Assume} \quad \bar{\nu} = \mathrm{R}\,\bar{\mu} + \bar{\beta} \quad \Rightarrow \quad \bar{\mu} = \mathrm{R}^{-1}(\bar{\nu} - \bar{\beta})$$

Suppose further that data Poissonian: $P(n_i, \nu_i) = \nu_i^{n_i} e^{-\nu_i} / n_i!$

$$\Rightarrow \quad \ln L(\bar{\mu}) = \sum_{i=1}^{N} (n_i \ln \nu_i - \nu_i) \qquad \text{ML estimator}: \hat{\bar{\nu}} = \bar{n}$$

Estimator for the true distribution: $\hat{\bar{\mu}} = \mathrm{R}^{-1}(\bar{n} - \bar{\beta})$

Simple example where such a method gives unacceptable result. Assume R based on a Gaussian resolution ~1.5 × bin width, all ε_i = 1 & β_i = 0. Estimates oscillate & uncertainties as large as estimates. Correlation btwn neighbour bins ~ –1.

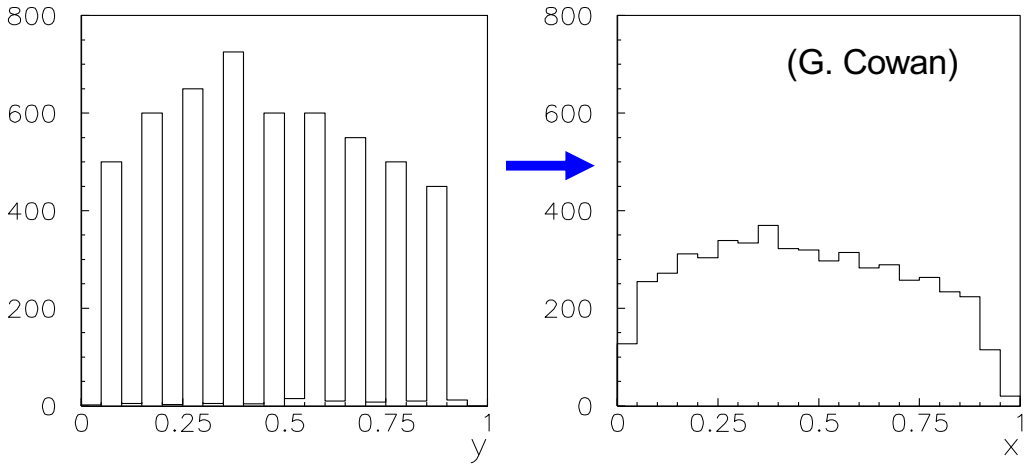


(G. Cowan)

Note that choosing wider bins would improve estimate signifi-cantly (resulting in smaller values for R_{ij} off-diagonal elements).

Inverting response matrix

What went wrong? Suppose $\bar{\mu}$ really had fine structure then R will tend to wash such a structure out but leave some residual.



Fine structure restored by applying

$$\bar{\mu} = \mathrm{R}^{-1}\bar{\nu}$$

However $\bar{n}$ includes statistical fluctuations that R^{-1} interprets as original fine structure. Estimator $\hat{\bar{\mu}}$ has "huge" fine structure.

Let's examine the ML solution a bit in detail:

$$E[\hat{\bar{\mu}}] = \mathrm{R}^{-1}(E[\bar{n}] - \bar{\beta}) = \mathrm{R}^{-1}(\bar{\nu} - \bar{\beta}) = \bar{\mu} \quad \rightarrow \quad \text{unbiased!}$$

Compute the covariance for the estimators:

$$U_{ij} = \mathrm{cov}(\hat{\mu}_i, \hat{\mu}_j) = \sum_{k,l=1}^{N} (R^{-1})_{ik}(R^{-1})_{jl}\,\mathrm{cov}[n_k, n_l] = \sum_{k=1}^{N} (R^{-1})_{ik}(R^{-1})_{jk}\nu_k$$

assume n_i & n_j independent Poisson variables, cov[n_k,n_l] = $\delta_{kl}\nu_k$.

Recall the RCF bound for an unbiased estimator:

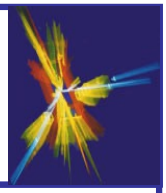
$$(U^{-1})_{ij} = -E\left[\frac{\partial^2 \ln L}{\partial\mu_i \partial\mu_j}\right] = E\left[\sum_{k=1}^{N} \frac{n_k R_{ki} R_{kj}}{\nu_k^2}\right] = \sum_{k=1}^{N} \frac{R_{ki} R_{kj}}{\nu_k} \quad \Rightarrow$$

ML estimator saturates the RCF bound (even though variance huge).

To reduce variance will have to introduce some bias!!

Strategy: accept small biases (systematic error) in exchange for large reductions in variances of the estimates (statistical error).

Correction factor method



In case of equal bining for $\vec{\mu}$ & $\vec{\nu}$, a commonly used estimator is

$$\hat{\mu}_i = C_i(n_i - \beta_i), \quad \text{where} \quad C_i = \mu_i^{\text{MC}} / \nu_i^{\text{MC}} \ (\text{correction factor})$$

where μ_i^{MC} & ν_i^{MC} obtained from MC (that contains no backgrounds). The covariances for the estimators:

$$U_{ij} = \text{cov}[\hat{\mu}_i, \hat{\mu}_j] = C_i C_j \,\text{cov}[n_i, n_j] = C_i^2 \delta_{ij} \nu_i$$

(independent Poissonians)

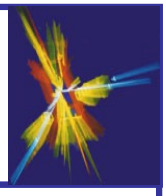
Usually the C_i's $\approx O(1)$ so the variances doesn't became large and in addition method is simple to implement (no matrix inversion). However estimates are biased (in accordance to the "RCF bound"):

$$E[\hat{\mu}_i] = C_i E[n_i - \beta_i] = C_i(\nu_i - \beta_i) \quad = \frac{\mu_i^{\text{MC}}}{\nu_i^{\text{MC}}} \nu_i^{\text{sig}}$$

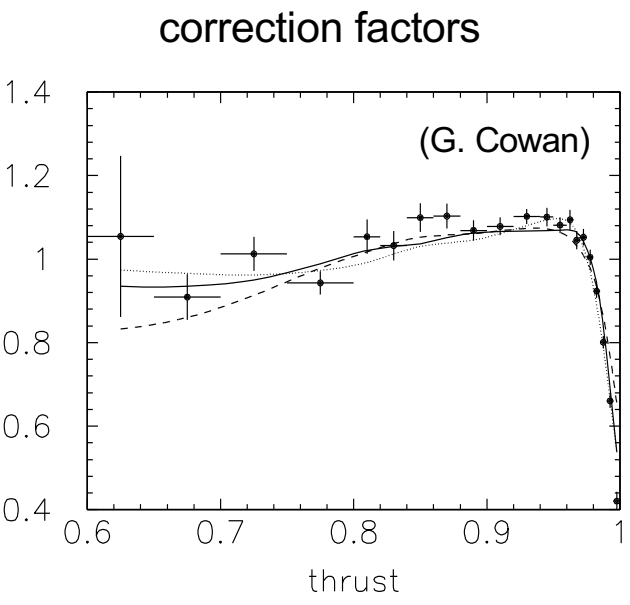
$$\Rightarrow \quad b_i = \left(\frac{\mu_i^{\text{MC}}}{\nu_i^{\text{MC}}} - \frac{\mu_i}{\nu_i^{\text{signal}}} \right) \nu_i^{\text{signal}}, \quad \text{where} \quad \nu_i^{\text{signal}} = \nu_i - \beta_i$$

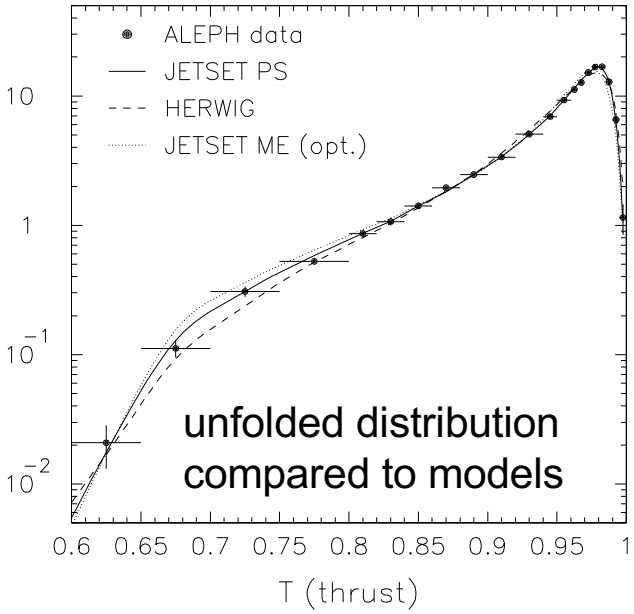
NB! Bias tends to pull the estimates toward the MC value $\rightarrow$ complication of testing predictions with correction factor method. Solution: iterative determination of $\hat{\mu}_i$ or systematic error due to MC dependence (estimated e.g. by computing correction factors for different predictions). Iterative: use estimated distribution $\hat{\vec{\mu}}$ "as" MC distribution (better agreement prediction vs estimate).

Works well when bin width $\geq$ few times resolution i.e. the off-diagonal entries in the response matrix R_{ij} are small.



Example: correction factors for the thrust ("back-to-backness") distribution for $Z^0 \to q\bar{q}$ (quark-antiquark)





Thrust

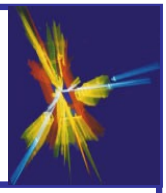
- *Thrust* measures the distribution of jets in a event.

$$T = \max\left(\frac{\sum_{i=1,N} \vec{p}_i \cdot \hat{n}_T}{\sum_{i=1,N} |\vec{p}_i|}\right)$$

- The unit vector $\hat{n}$ is where *T* is maximized is known at the *thrust axis*

- The range of T is: ½<*T*<1
  - T≈½ for an isotropic event
  - T=1 for an event with 2 back-to-back jets

Regularized unfolding



Regularization: impose "smoothness" on estimators $\hat{\bar{\mu}}$.
Consider only estimators "reasonable" having L/χ^2 values close to ML/LS solution (deviating at most by fixed $\Delta\ln L/\Delta\chi^2$).

$$\ln L(\bar{\mu}) \geq \ln L_{\max} - \Delta \ln L \quad \text{or} \quad \chi^2(\bar{\mu}) \leq \chi^2_{\min} + \Delta\chi^2$$

Size of $\Delta\ln L/\Delta\chi^2$ determine trade-off btwn bias & variance.
Choose "smoothest" estimator from above by maximizing:

$$\varphi(\bar{\mu}) = \alpha \ln L(\bar{\mu}) + S(\bar{\mu}),$$

where $S(\bar{\mu})$ = regularization function ("smoothness" value) & α = regularization parameter (to satisfy given $\Delta\ln L/\Delta\chi^2$).

α = 0 gives smoothest distribution possible (ignores data!) & $\alpha \to \infty$ gives ML/LS solution ($\equiv$ response matrix inversion).

For method to work, surfaces of constant $\ln L(\bar{\mu})$ & $S(\bar{\mu})$ must behave "nicely" (i.e. no changes convex $\to$ concave or multiple local maxima). Note also $\bar{\nu}$ can be used as variable in $\ln L$ & S since $\bar{\nu} = R\bar{\mu} + \bar{\beta}$. Now only $\hat{\bar{\nu}} \neq \bar{n}$.

If total # of entries also left free get additional requirement:

$$\hat{\nu}_{\text{tot}} = \sum_{i=1}^{N} \hat{\nu}_i = \sum_{i=1}^{N} \left(\sum_{j=1}^{M} R_{ij} \hat{\mu}_j + \beta_i \right) = n_{\text{tot}} \quad \Rightarrow$$

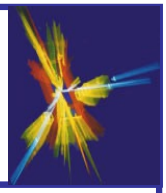
$$\varphi(\bar{\mu}, \lambda) = \alpha \ln L(\bar{\mu}) + S(\bar{\mu}) + \lambda \left(n_{\text{tot}} - \sum_{i=1}^{N} \nu_i \right)$$

$$\partial\varphi/\partial\lambda = 0 \quad \Rightarrow \quad \sum_{i=1}^{N} \nu_i = n_{\text{tot}}$$

λ is a Lagrange multiplier.

To apply method, still need:
– a regularization function $S(\bar{\mu})$ (describe 2 most common).
– a prescription for setting regularization parameter α.
Goodness of estimators judged by their bias & variance.

Tikhonov regularization



Take mean square of k^{th} derivative as smoothness value

$$S[f_{\text{true}}(y)] = -\int \left(d^k f_{\text{true}}(y)/dy^k\right)^2 dy, \quad \text{where } k = 1,2,...$$

Often take $k = 2$ since then $S \approx$ mean square curvature.
For histograms derivatives replaced by finite differencies.
For histograms with equal bin widths, S can be given as

$$S(\vec{\mu}) = -\sum_{i=1}^{M-1} (\mu_i - \mu_{i+1})^2 \quad (k=1)$$

$$S(\vec{\mu}) = -\sum_{i=1}^{M-2} (-\mu_i + 2\mu_{i+1} - \mu_{i+2})^2 \quad (k=2)$$

$$S(\vec{\mu}) = -\sum_{i=1}^{M-3} (-\mu_i + 3\mu_{i+1} - 3\mu_{i+2} + \mu_{i+3})^3 \quad (k=3)$$

NB! 2nd derivate not well defined for 1st & last histogram bin.

General expression: for any k & regardless of bin widths

$$S(\vec{\mu}) = -\sum_{i,j=1}^{M} G_{ij}\mu_i\mu_j = -\vec{\mu}^T G \vec{\mu},$$

where G symmetric matrix of constants, that can easily be calculated from choice of regularization function S. e.g. for **Tikhonov regularization** with $k = 2$ & equal bin width:

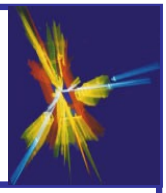
$$\left.\begin{array}{c} G_{11} = G_{MM} = 1 \quad G_{22} = G_{M-1,M-1} = 5 \quad G_{ii} = 6 \\ G_{12} = G_{21} = G_{M,M-1} = G_{M-1,M} = -2 \\ G_{i,i\pm1} = G_{i\pm1,i} = -4 \quad G_{i,i\pm2} = G_{i\pm2,i} = 1 \end{array}\right\} \begin{array}{l} 3 \le i \le M-2 \\ \text{all other } G_{ij} = 0 \end{array}$$

1st & 2nd $S(\mu)$ derivates (need for estimators & covariances):

$$\partial S(\vec{\mu})/\partial\mu_i = -2\sum_{j=1}^{M} G_{ij}\mu_j \quad \& \quad \partial^2 S(\vec{\mu})/\partial\mu_i\partial\mu_j = -2G_{ij}$$

Tikhonov $k = 2$ regularization used e.g. in particle physics.

Regularization with entropy



Another common regularization function is **entropy** H

$$H = -\sum_{i=1}^{M} p_i \ln p_i, \text{with probability distribution } \bar{p} = (p_1,\ldots,p_M)$$

All p_i equal (one p_i =1, others = 0) → max (min) entropy ("smoothness") = uniform distribution (all entries in 1 bin).
Use entropy as regularization function (smoothness value)

$$S(\bar{\mu}) = H(\bar{\mu}) = -\sum_{i=1}^{M} \frac{\mu_i}{\mu_{\text{tot}}} \ln \frac{\mu_i}{\mu_{\text{tot}}}$$

Such estimator based on **principle of maximum entropy**.
Orgin: number of ways a histogram can be constructed:

$$\Omega(\bar{\mu}) = \frac{\mu_{\text{tot}}!}{\mu_1!\mu_2!\ldots\mu_M!} \quad \text{using } \ln n! \approx n(\ln n - 1) \;\&\; \sum \mu_i = \mu_{\text{tot}}$$

$$\Rightarrow \ln\Omega(\bar{\mu}) = -\sum_{i=1}^{M} \mu_i \ln(\mu_i / \mu_{\text{tot}}) = \mu_{\text{tot}} H(\bar{\mu})$$

1st & 2nd $S(\mu)$ derivates (need for estimators & covariances):

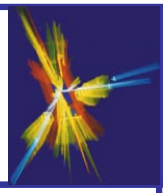
$$\frac{\partial S}{\partial \mu_i} = -\frac{1}{\mu_{\text{tot}}} \ln\frac{\mu_i}{\mu_{\text{tot}}} - \frac{S}{\mu_{\text{tot}}} \;\&\; \frac{\partial^2 S}{\partial \mu_i \partial \mu_j} = \frac{1}{\mu_{\text{tot}}^2}\left[1 - \frac{\delta_{ij}\mu_{\text{tot}}}{\mu_i} + \ln\left(\frac{\mu_i\mu_j}{\mu_{\text{tot}}^2}\right) + 2S\right]$$

Easily generalized to many dimensions & therefore widely used for image reconstruction in i.e. medicin & astronomy.
NB! entropy value not dependent on the order of the bins.

Often motived by Bayesian statistics but need some prior π, if entropy itself used $\pi(\bar{\mu}) = \Omega(\bar{\mu}) = \exp(\mu_{\text{tot}} H(\bar{\mu})) \Rightarrow$

$$\varphi(\bar{\mu}) = \ln f(\bar{\mu}\,|\,\bar{n}) = \ln L(\bar{\mu}\,|\,\bar{n}) + \mu_{\text{tot}} H(\bar{\mu}) \;\;\&\;\; \alpha = 1/\mu_{\text{tot}}$$

where φ now joint probability density & α fixed parameter.
Problems for large μ_{tot} → weight of entropy term too large leading to unreasonably large bias to uniform distribution.



Variance & bias of the estimators

Variance & bias of $\bar{\mu}$:

In general equations determining $\hat{\bar{\mu}}(\bar{n})$ non-linear so therefore expand $\hat{\bar{\mu}}(\bar{n})$ about $\bar{n}_{\text{obs}}$ (i.e. the observed data):

$$\hat{\bar{\mu}}(\bar{n}) \approx \hat{\bar{\mu}}_{\text{obs}} - \mathrm{A}^{-1}\mathrm{B}(\bar{n} - \bar{n}_{\text{obs}}),$$

μ has M +1 components (1 additional from λ fixing ν_{tot})

$$A_{ij} = \begin{cases} \dfrac{\partial^2 \varphi}{\partial \mu_i \partial \mu_j}, & i,j = 1 \ldots M \\ \dfrac{\partial^2 \varphi}{\partial \mu_i \partial \lambda} = -1, & i = 1 \ldots M,\ j = M+1 \\ \dfrac{\partial^2 \varphi}{\partial \lambda^2} = 0, & i = M+1,\ j = M+1 \end{cases} \qquad B_{ij} = \begin{cases} \dfrac{\partial^2 \varphi}{\partial \mu_i \partial n_j}, & i = 1 \ldots M,\ j = 1 \ldots N \\ \dfrac{\partial^2 \varphi}{\partial \lambda \partial n_j} = 1, & i = M+1,\ j = 1 \ldots N \end{cases}$$

A symmetric $(M+1) \times (M+1)$ matrix, B $(M+1) \times N$ matrix.

Use error propagation to get covariance $U_{ij} = cov[\hat{\mu}_i, \hat{\mu}_j]$

$$\mathrm{U} = \mathrm{CVC}^T, \text{where } \mathrm{C} = \mathrm{A}^{-1}\mathrm{B} \ \ \& \ \ C_{ik} = \partial \hat{\mu}_i / \partial n_k$$

Estimator for bias:

$$\hat{b}_i = E[\hat{\mu}_i] - \mu_i \approx \sum_{j=1}^{N} C_{ij}(\hat{\nu}_j - n_j) = \sum_{j=1}^{N} \frac{\partial \hat{\mu}_i}{\partial n_j}(\hat{\nu}_j - n_j),$$

$$\text{where} \quad \hat{\nu}_j = \sum_{k=1}^{M} R_{jk}\hat{\mu}_k + \beta_j \quad (\text{NB! } \hat{\nu}_j \neq n_j)$$

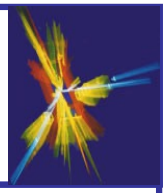
Approx. valid only for small $(\hat{\nu}_j - n_j)$ i.e. small $\Delta \ln L$ or $\Delta\chi^2$.

Covariance $W_{ij} = cov[\hat{b}_i, \hat{b}_j]$ from propagation of errors:

$$\mathrm{W} = (\mathrm{CRC} - \mathrm{C})\,\mathrm{V}(\mathrm{CRC} - \mathrm{C})^T = (CR - I)U(CR - I)^T$$

Variances of estimated biases W_{ii} used to tell whether estimated biases are significantly different from zero.

Choosing regularization parameter α, can make use of the variance of estimate U_{ii}, the bias $\hat{b}_i$ & its variance W_{ii}.



Choosing regularization parameter

Choosing regularization parameter α: Various definitions of optimal trade-off btn bias & variance exists. Here a few:

• minimize the mean squared error averaged over all bins

$$\mathrm{MSE} = \frac{1}{M}\sum_{i=1}^{M}(U_{ii} + \hat{b}_i^2) \quad (M = \#\,\text{of bins for true histogram})$$

iterative produce: try α–value & for that maximize $\varphi(\bar{\mu}, \lambda)$. Repeat procedure until minimal MSE solution found.

• minimize the weighted mean squared error

$$\mathrm{MSE}' = \frac{1}{M}\sum_{i=1}^{M}\frac{(U_{ii} + \hat{b}_i^2)}{\hat{\mu}_i}$$

weights each bin by the precision its estimate determined (Poisson variance). Minimisation similar as in MSE.

• average change of the χ^2 of each bin by one

$$\Delta\chi^2 = 2\Delta\ln L = N \quad (\#\,\text{of bins for the measured data})$$

works if migration btwn bins not large, popular in imaging

• change effective χ^2 by one (account for bin correlations)

$$\Delta\chi^2_{\mathrm{eff}} = (\hat{\nu} - \bar{n})^T\,\mathrm{RCV}^{-1}(\mathrm{RC})^T(\hat{\nu} - \bar{n}) = 1$$

estimate ν_i gets contributions also from neighbouring bins to n_i. reduced ν_i–n_i coupling taken into account (matrix RC).

• allowing the biases to vary within their own errors

$$\chi^2_b = \sum_{i=1}^{M}\frac{\hat{b}_i^2}{W_{ii}} = M \quad (\#\,\text{of bins for true histogram})$$

In practice, final estimates often stable w.r.t. $\Delta\ln L$ until a certain point, where variances increases rapidly so choice of α often roughly method independent. But numerical implementations of unfolding methods a nontrivial task.

Choosing regularization parameter

Maximal entropy Tikhonov regularization

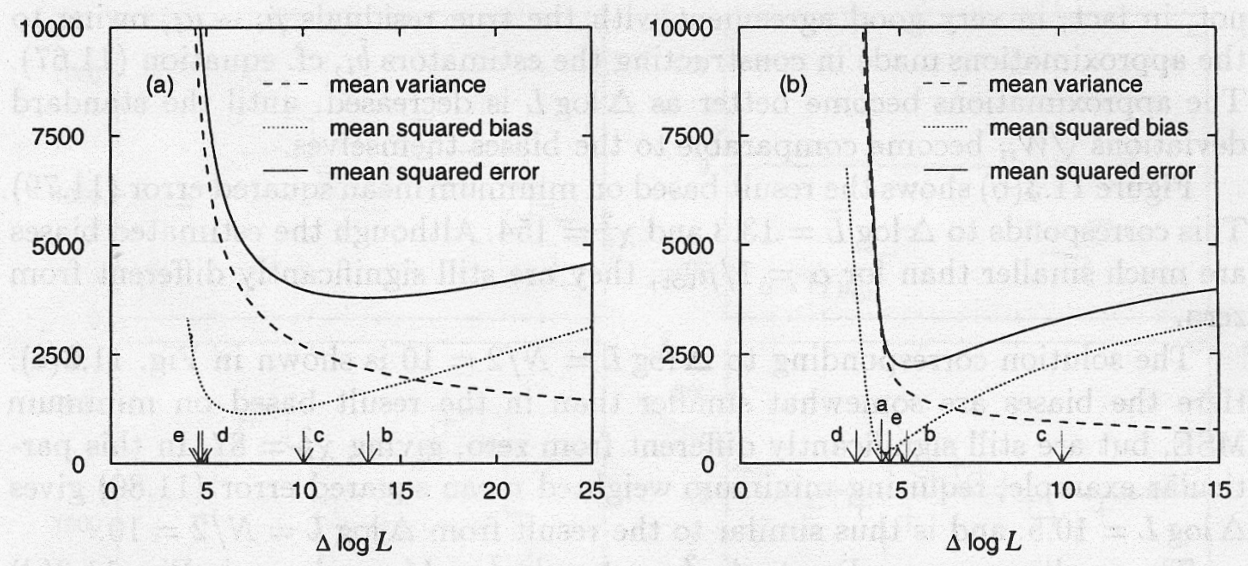


Fig. 11.2 The estimated mean variance, mean squared bias, and their sum, the mean squared error, as a function of $\Delta \log L$ for (a) MaxEnt and (b) Tikhonov regularization ($k = 2$). The arrows indicate the solutions from Figs 11.3 and 11.4: (b) is minimum MSE, (c) is $\Delta \log L = N/2$, (d) is $\Delta\chi^2_{\text{eff}} = 1$, and (e) is $\chi^2 b = M$. For the MaxEnt case, the Bayesian solution $\Delta \log L = 970$ is not shown. For Tikhonov regularization, (a) gives the solution for minimum weighted MSE.

(a) minimum weighted MSE

(b) minimum MSE

(c) $\Delta \log L = N/2$

(d) $\chi^2_{\text{eff}} = 1$

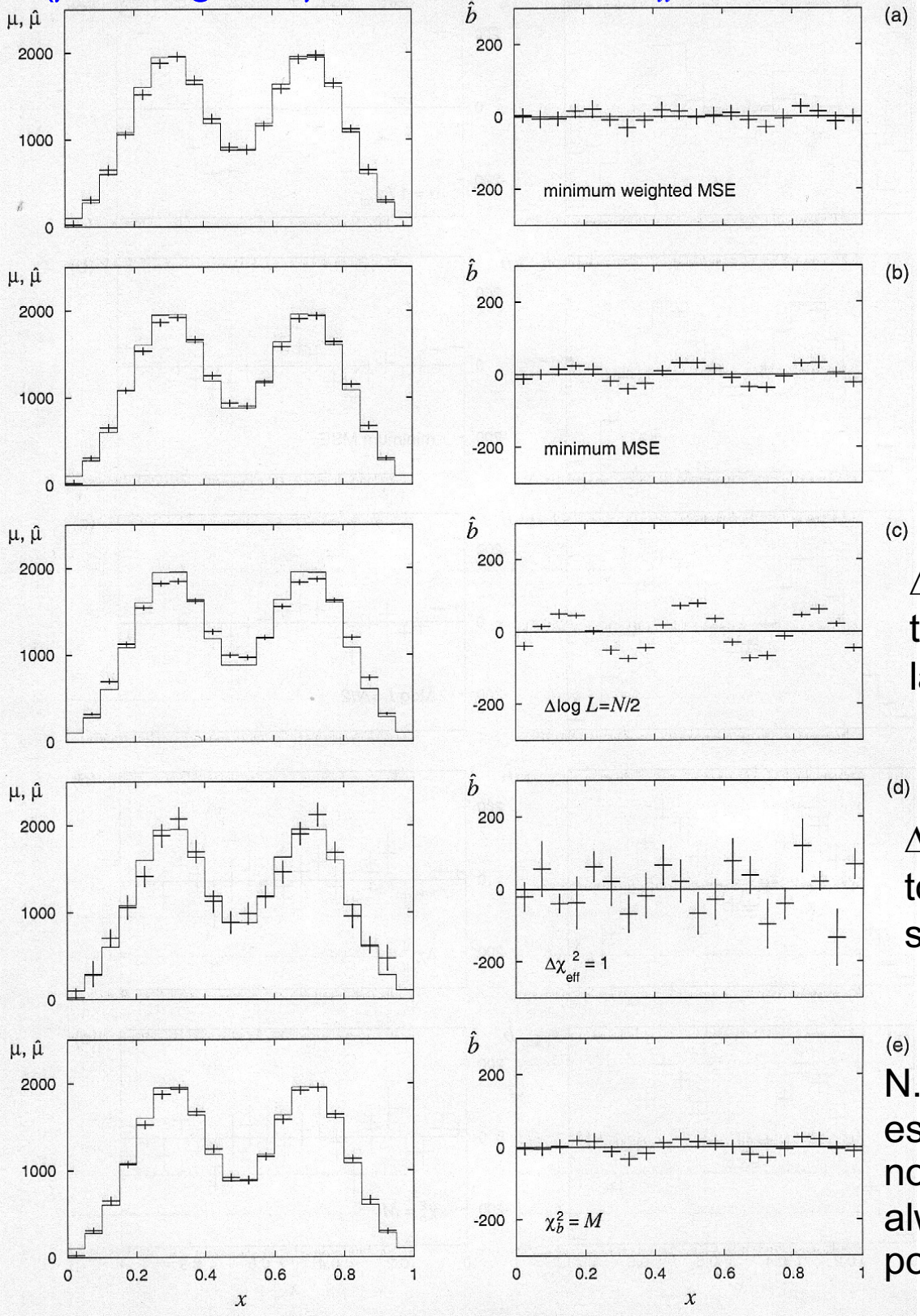
(e) $\chi^2_{\text{b}} = \text{M}$

Above MSE behaviour as function of $\Delta \log L$ typical. Slow decrease to minimum, then rapid increase at small $\Delta \log \text{L}$. In practice ”best” solutions usually close to position where MSE start to rapidly rise. N.B. Each combination of regularization function & regularization parameter unique. Which combination optimal depends on the application.

Tikhonov regularization with k = 2

(G. Cowan)

(μ = original, $\hat{\mu}$ = estimate $\pm \sqrt{U_{ii}}$, $\hat{b}$ = bias $\pm \sqrt{W_{ii}}$)



$\Delta \ln L$ too large !!

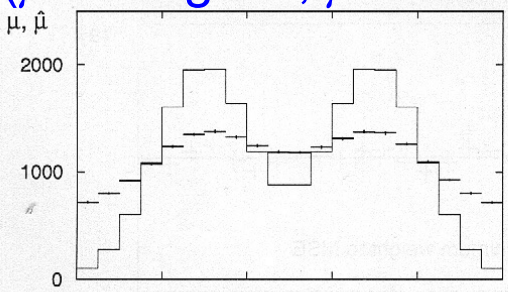
$\Delta \ln L$ too small !!

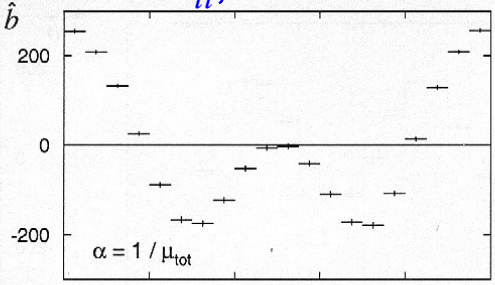
N.B. μ estimate not always positive.

Regularization with maximum entropy

(G. Cowan)

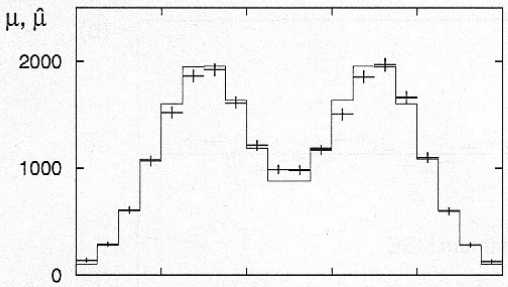
(μ = original, $\hat{\mu}$ = estimate $\pm \sqrt{U_{ii}}$, $\hat{b}$ = bias $\pm \sqrt{W_{ii}}$)

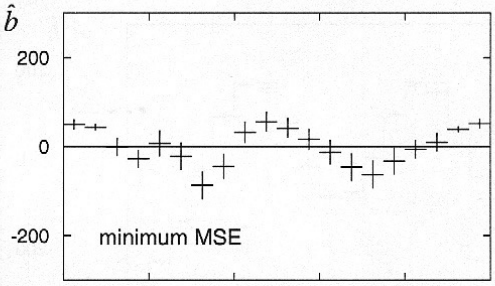




(a)

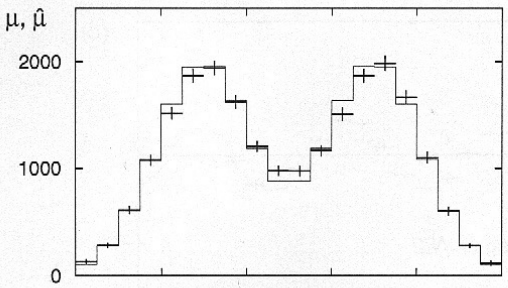
Baysian approach fails totally.

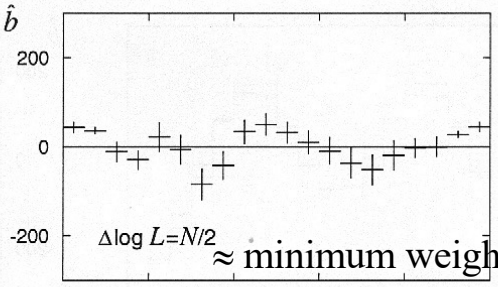




(b)

$\Delta \ln L$ a bit too large !!

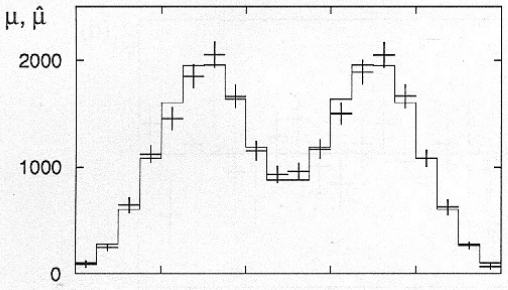


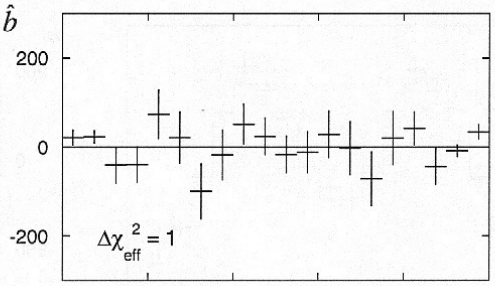


(c)

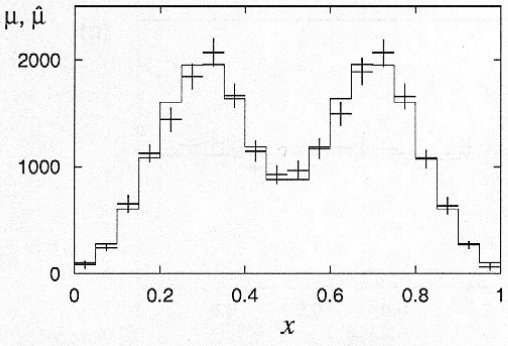
$\Delta \ln L$ a bit too large !!

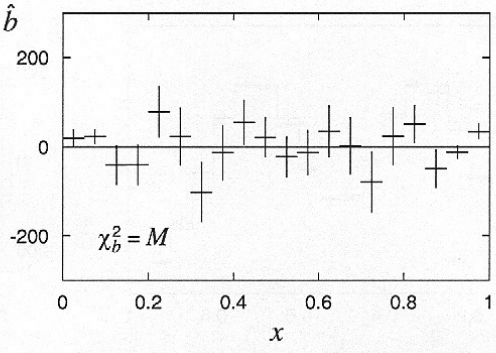
≈ minimum weighted MSE





(d)





(e)

N.B. μ estimate always positive.

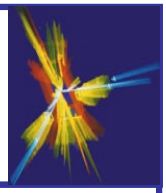
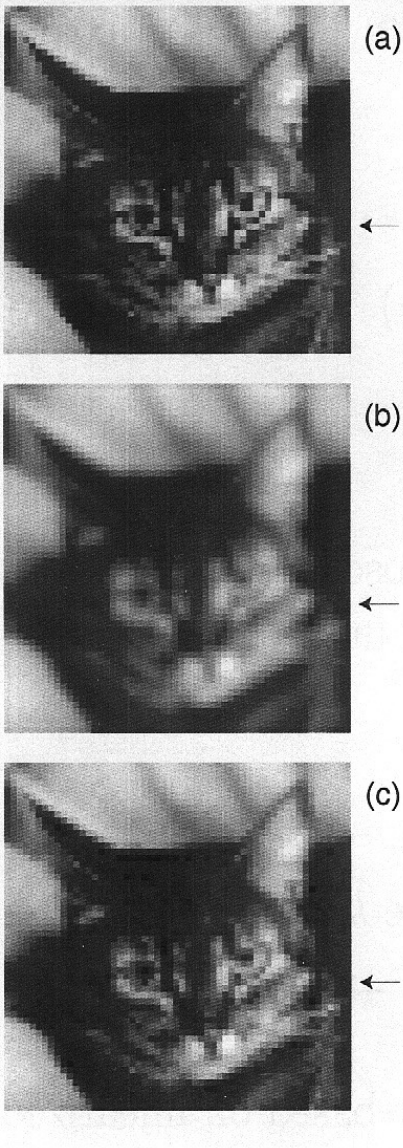
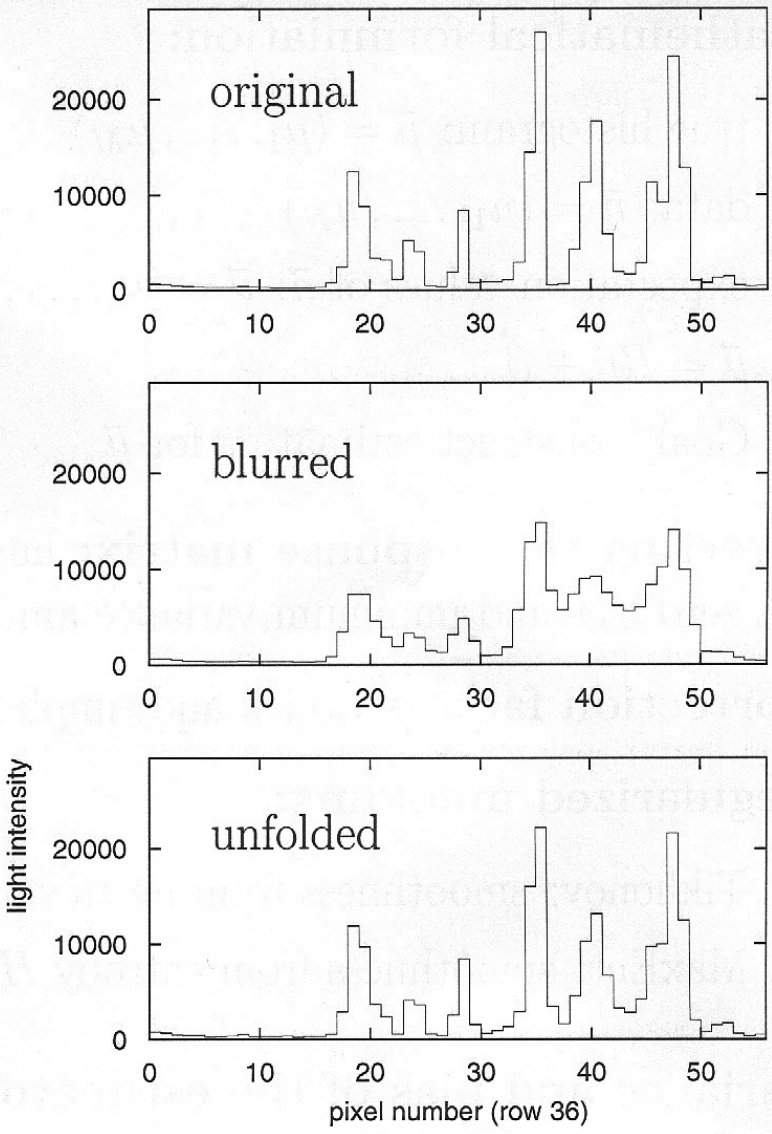


Image reconstruction with maximum entropy method

(G. Cowan)





Light intensity at pixel row 36

Original image blurred with Gaussian resolution function, image unfolded with maximum entropy method with $\Delta \ln L = N/2$ ($N = 3136$). Often used for reconstruction of images in astronomy since method results only in a small bias against point sources (”peaks”) and is easy to generalize to two (or more) dimensional problems.

THAT WAS ALL FOR THIS COURSE !!