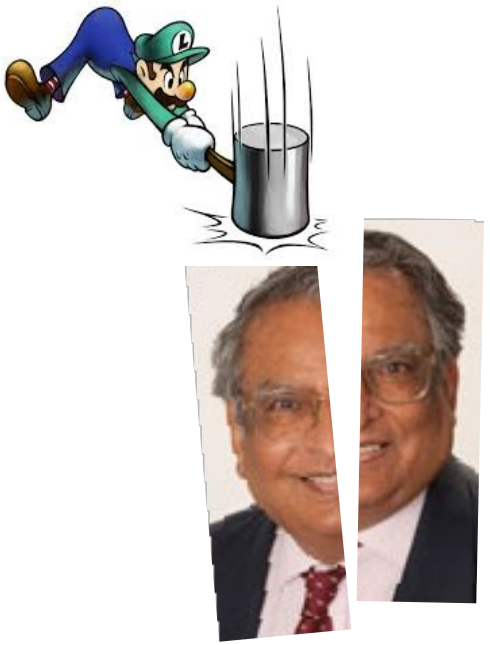
Advanced Statistical Methods – Lecture 3



Confidence intervals
Maximum log likelihood ratio
Wilks' theorem
Breaking Wilks' theorem
Minimizers

Confidence belt construction I

Goal: We seek a way to build, based on an experimental result with some random component, an interval in the model parameter space that is in, say, 95% of the cases including (or *covering*) the true model parameter.

General construction of the confidence belt

- We consider a *class of hypothesis with one free parameter*. The PDF is given by

$$P(x|\theta)$$

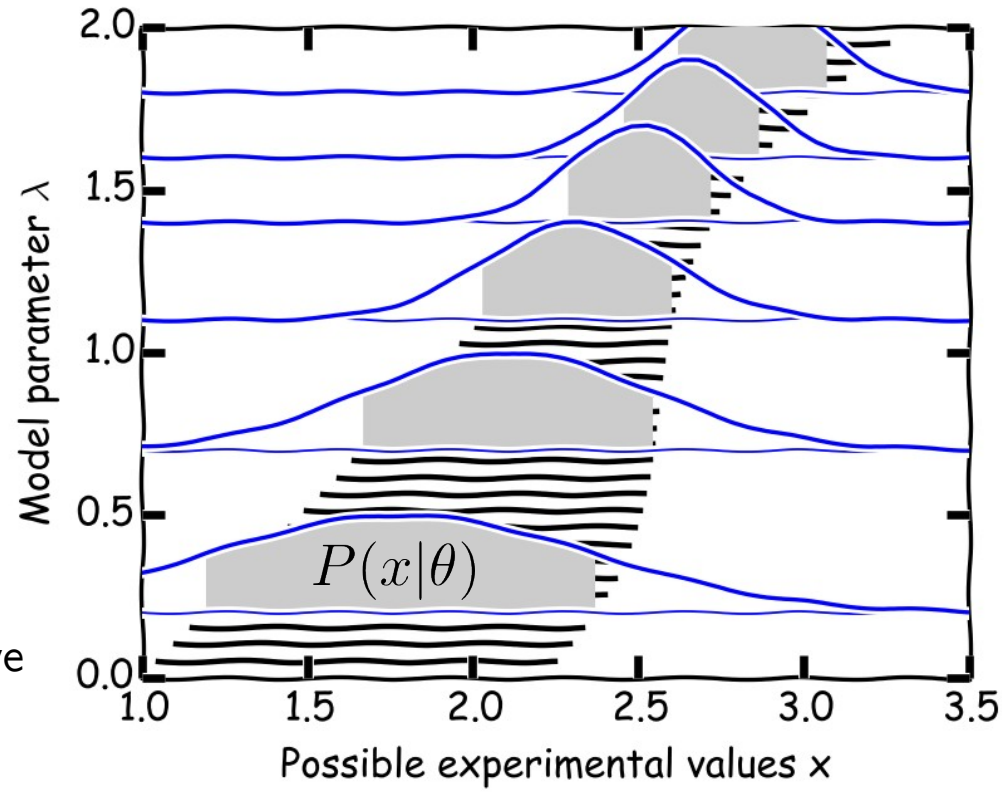
- An *acceptance interval*

$$[x_0(\theta), x_1(\theta)]$$

for a given (true) model parameter θ and coverage 1-α is given by any interval that satisfies the condition

$$\int_{x_0(\theta)}^{x_1(\theta)} P(x|\theta)dx = 1 - \alpha$$

- Any choice of boundaries that fulfill the above requirement define the ***confidence belt*** with a significance of 1-α

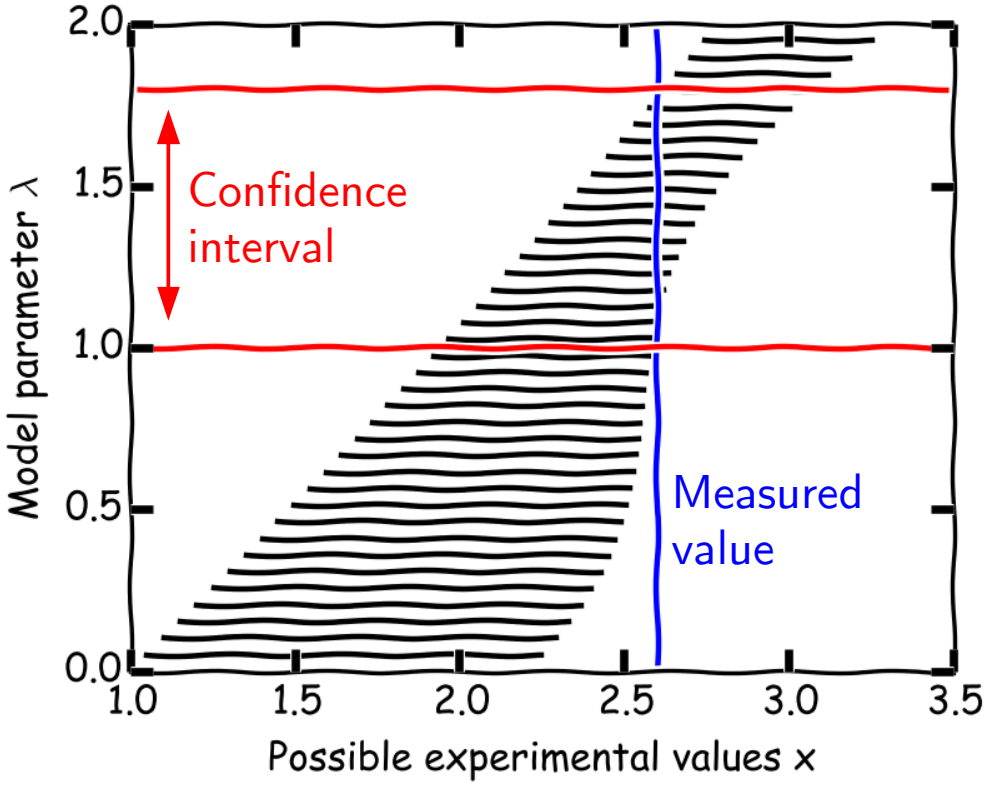


Confidence belt construction II

Definition of Confidence interval

- For a given observation x_{obs}, the *confidence interval* is given by the values of theta for which the acceptance interval contains x_{obs}.

$$I(x_{\rm obs}) = \{\theta | x_0(\theta) \leq x_{\rm obs} \leq x_1(\theta)\}$$

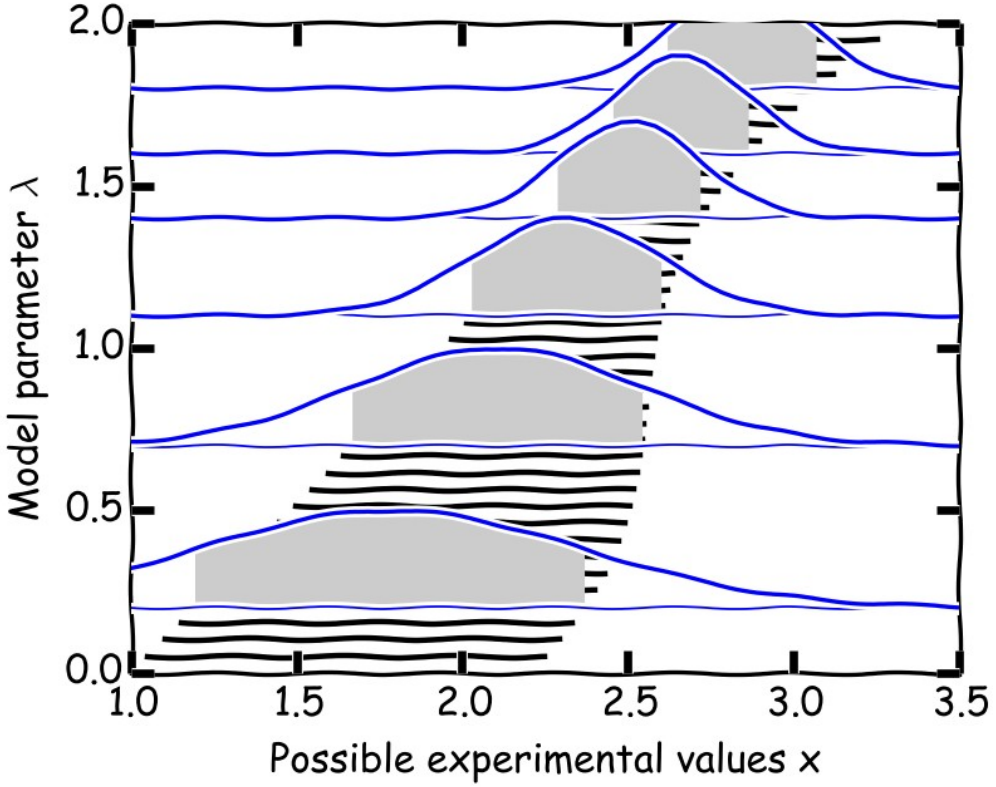


Result:

- By construction, *and independently of the true value of* θ, the confidence interval will *cover* (i.e. contain) the true value in exactly 1-α of the cases.

There are many possible confidence belts

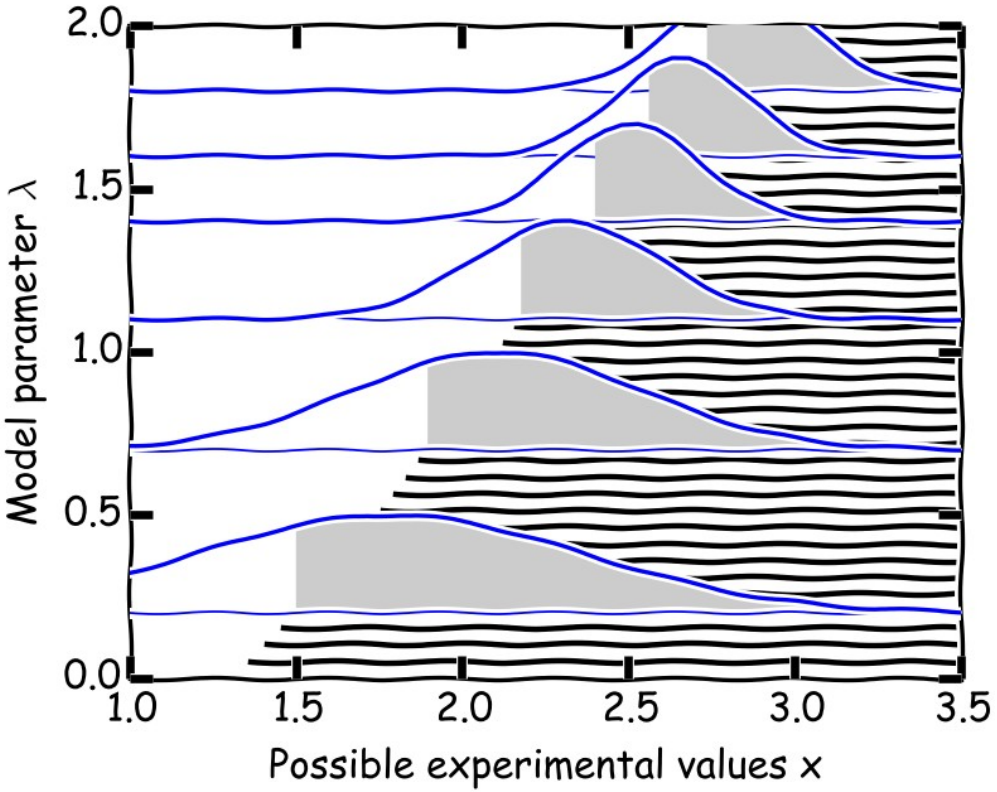
Two-sided confidence region:



Often written as (for 68% CL)

$$\theta = 3.5^{+0.4}_{-0.2}$$

One-sided confidence region:



$$\theta \leq 4.3\ (95\%\mathrm{CL})$$

The maximum likelihood construction

One extremely common definition of the interval is based on a specific test statistic: the **maximum likelihood ratio:**

$$\Lambda(\theta|x_{\rm obs}) = -2\ln\frac{\mathcal{L}(\theta|x_{\rm obs})}{\max_{\theta'}\mathcal{L}(\theta'|x_{\rm obs})}$$

The **confidence interval** is then simply the set of all model parameters for which the test statistic is small (i.e. good) enough.

$$I(x_{\rm obs}) = \{\theta|\Lambda(\theta|x_{\rm obs}) < c(\theta)\}$$

The **threshold value** *c(θ)* has to be defined such that the interval has the right *coverage*:

$$P(\ \Lambda(\theta|x) < c(\theta)\ \ |\ \ \theta) = 1-\alpha$$

Question: How to obtain *c(θ)*?
Answer: In the vast majority of the cases, *Λ* follows a chi-squared distribution!

$$\Lambda(\theta|x) \sim \chi^2_k$$

Note that *Λ* is

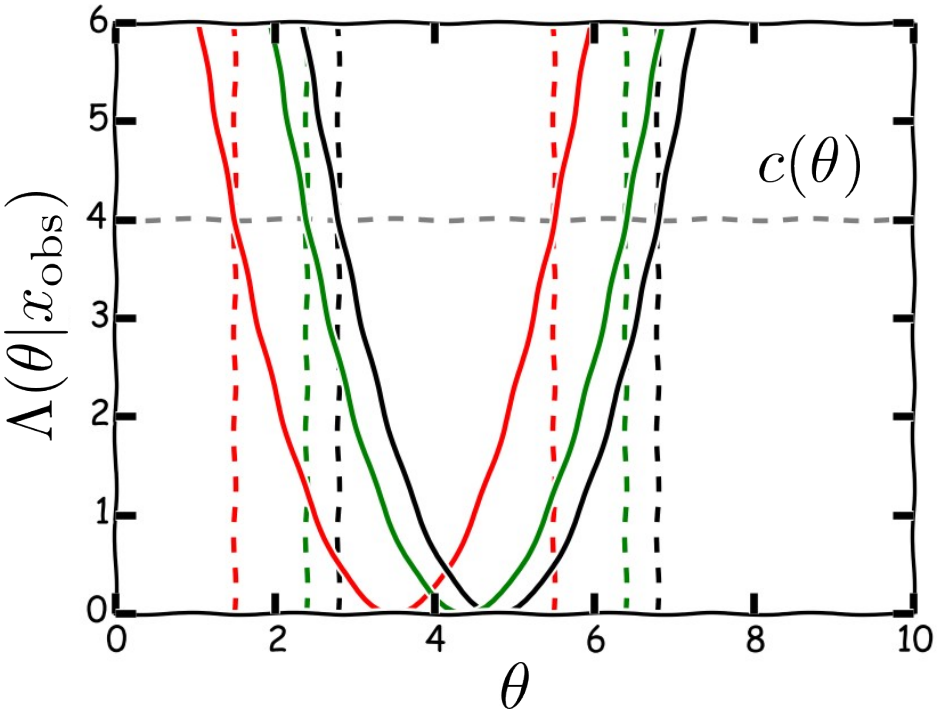
- non-negative

$$\Lambda(\theta|x) \geq 0$$

- zero in the case of the MLE

$$\Lambda(\hat{\theta}(x_{\rm obs})|x_{\rm obs}) = 0$$

Different observed values x_{obs} lead to different confidence intervals.



The maximum likelihood construction

Example:

- The normal distribution with unknown mean: $P(x|\theta) = N(x|\theta, \sigma)$

  (we consider σ as known, for simplicity)

- We can the calculate the likelihood ratio

$$\mathcal{L}(\theta|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\theta)^2}$$

$$\Lambda(\theta|x) = -2\ln \frac{\mathcal{L}(\theta|x)}{\max_{\theta'} \mathcal{L}(\theta'|x)} = \frac{(x-\theta)^2}{\sigma^2}$$

$$\max_{\theta'} \mathcal{L}(\theta'|x) = \frac{1}{\sqrt{2\pi}\sigma}$$

This follows a chi-squared distribution with one degree of freedom!

- In the simple case of a normal-distributed variable, we find that the maximum likelihood ratio Λ follows a chi-squared distribution.

Wilks' theorem

If the data *x* is distributed according to the likelihood function *L* for the true model parameters $\theta_1, \ldots, \theta_n$, then the maximum ln likelihood-ratio defined as

$$\Lambda(\theta_1, \ldots, \theta_k | \vec{x}) \equiv -2 \ln \frac{\mathcal{L}(\theta_1, \ldots, \theta_k, \hat{\theta}_{k+1}, \ldots, \hat{\theta}_n | \vec{x})}{\mathcal{L}(\hat{\theta}_1, \ldots, \hat{\theta}_n | \vec{x})}$$

where the $\hat{\theta}_i$ are MLEs for the likelihood function *L*, follows – in the large *N* limit – a chi-squared distribution with k degrees of freedom.

$$\Lambda(\theta_1, \ldots, \theta_k | \vec{x}) \sim \chi^2_k$$

Samuel S. Wilks
1937

Wilks' theorem (if it applies) makes it relatively simple to construct confidence intervals in a multi-dimensional model-parameter space.

Remember that, e.g.:

$$\mathcal{L}(\theta_1, \ldots, \theta_k, \hat{\theta}_{k+1}, \ldots, \hat{\theta}_n | \vec{x}) = \max_{\theta'_{k+1}, \ldots, \theta'_n} \mathcal{L}(\theta_1, \ldots, \theta_k, \theta'_{k+1}, \ldots, \theta'_n | \vec{x})$$

"Profiling" over nuisance parameters

Imagine you are interested in only one model parameter, whereas all remaining model parameters are irrelevant *nuisance* parameters (e.g. describing not the signal but the background).
You can then consider the one-dimensional maximum log likelihood ratio

$$\Lambda(\theta_1|\vec{x}) \equiv -2\ln\frac{\mathcal{L}(\theta_1,\hat{\theta}_2,\ldots,\hat{\theta}_n|\vec{x})}{\mathcal{L}(\hat{\theta}_1,\ldots,\hat{\theta}_n|\vec{x})}$$

which defines the confidence interval as

$$I(\vec{x}_{\rm obs}) = \{\theta_1|\Lambda(\theta_1|\vec{x}_{\rm obs}) < c(\theta_1)\}$$

Practically, the procedure is as follows:

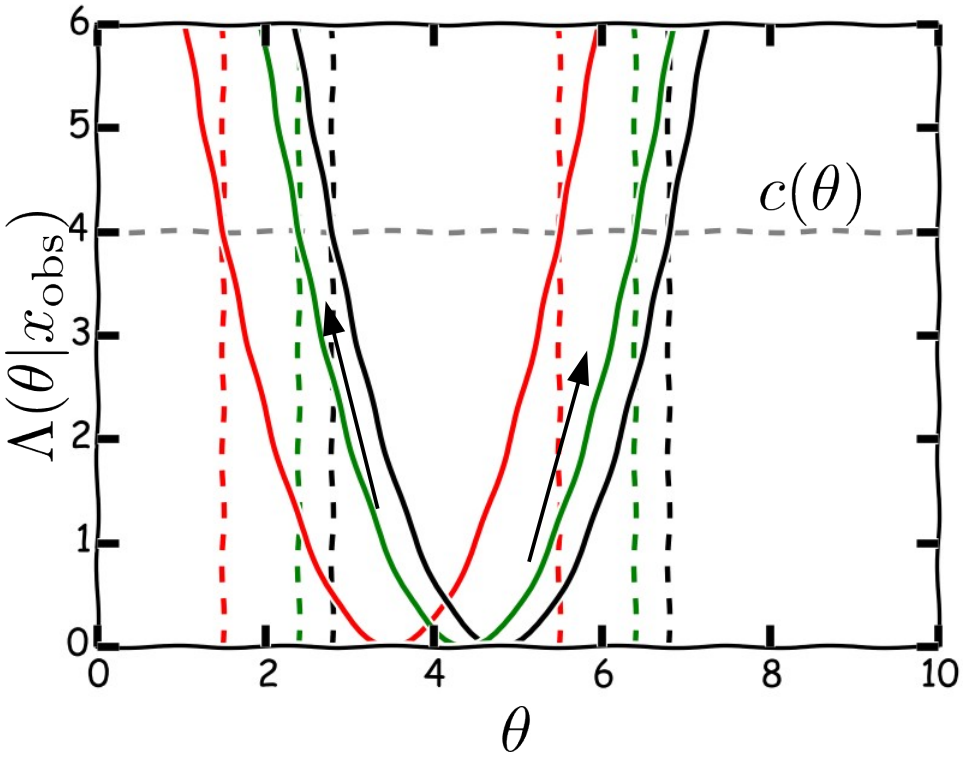
- *Minimize*

  $$-2\ln\mathcal{L}(\theta_1,\ldots,\theta_n|\vec{x}_{\rm obs})$$

  with respect to *all* model parameters.

- Decrease and increase the model parameter of interest (here theta1) until -2lnL increased by a value of *c, while refitting the other parameters.*

- The values of theta at which -2lnL changed by *c* define the boundaries of the confidence interval.



1 and 2-dimensional confidence intervals

The procedure can be easily generalized to two and more dimensional confidence intervals.

$$I(\vec{x}_{\rm obs}) = \left\{ \vec{\theta} \in \mathbb{R}^k | \Lambda(\theta_1, \theta_2, \ldots | \vec{x}_{\rm obs}) < c \right\}$$

Note:

- The threshold values *c* for different confidence level depend on the dimensionality of the confidence interval!

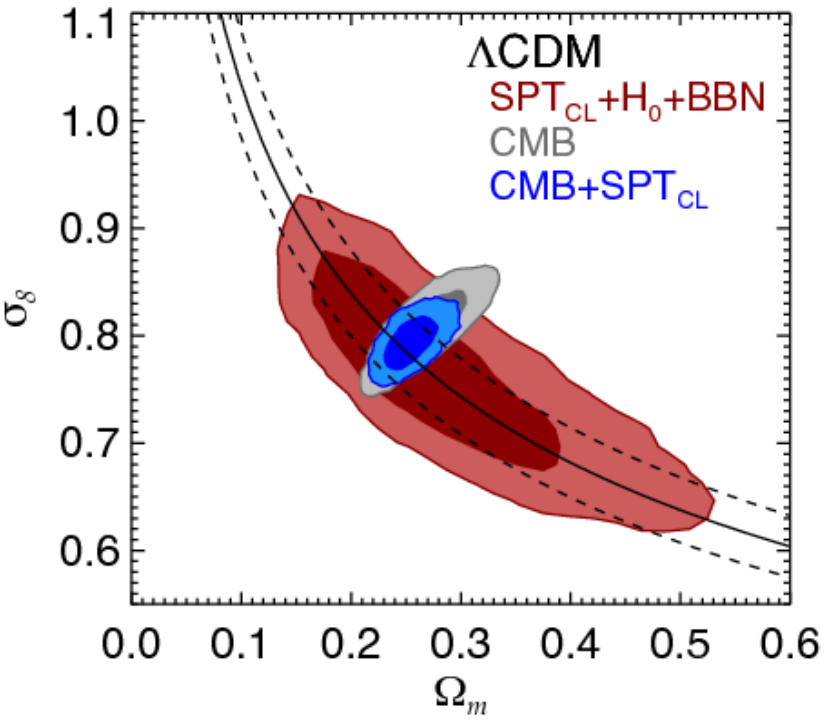
- **One dimensions** $\Lambda(\theta_1 | \vec{x}_{\rm obs}) \sim \chi^2_{k=1}$

$$c = 1 \quad (68.3\%\,{\rm CL})$$
$$c = 4 \quad (95.4\%\,{\rm CL})$$
$$c = 9 \quad (99.7\%\,{\rm CL})$$

- **Two dimensions** $\Lambda(\theta_1, \theta_2 | \vec{x}_{\rm obs}) \sim \chi^2_{k=2}$

$$c = 2.3 \quad (68.3\%\,{\rm CL})$$
$$c = 6.2 \quad (95.4\%\,{\rm CL})$$
$$c = 11.8 \quad (99.7\%\,{\rm CL})$$

The resulting regions can have complex shapes:



Upper limits on model parameters

Upper limits (of course, the same arguments also hold for lower limits)

- *Upper limits* on a parameter (in contrast to two-sided intervals) can be obtained by setting the -2lnL to zero below the best-fit value.

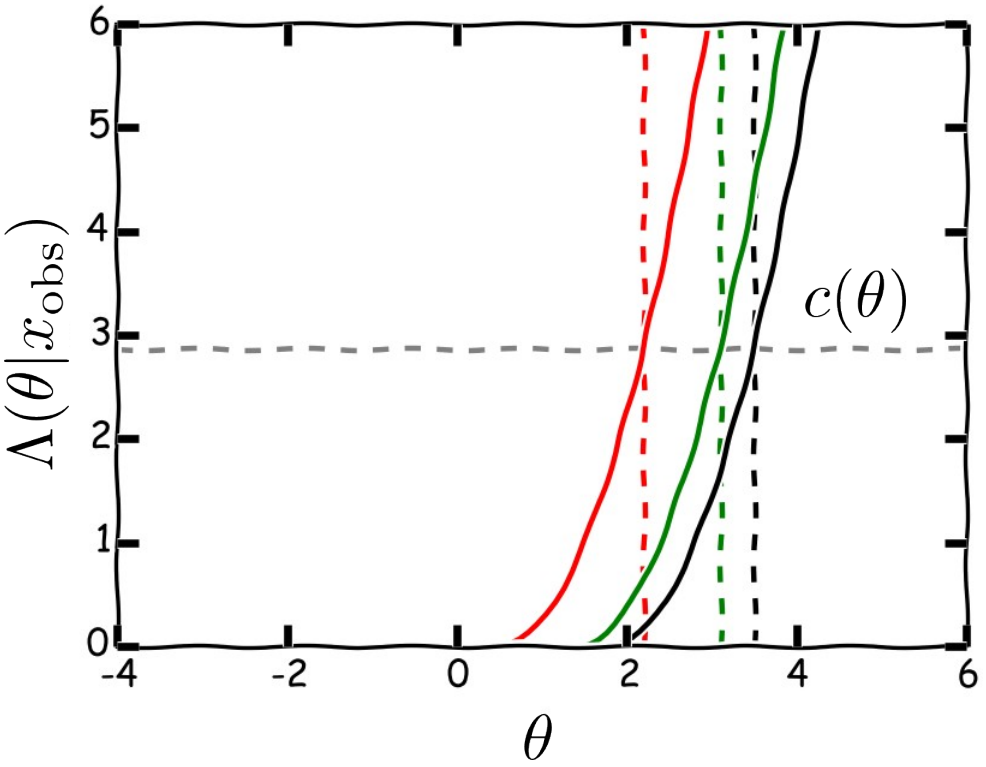
$$\Lambda(\theta|x) = -2\,\theta_H(\theta - \hat{\theta})\,\ln\frac{\mathcal{L}(\theta|x)}{\max_\theta \mathcal{L}(\theta|x)}$$

- The corresponding confidence interval is again defined via

$$I(\vec{x}_{\rm obs}) = \{\theta \in \mathbb{R}|\Lambda(\theta|x) < c(\theta)\}$$

- However, the threshold *c* is a bit more tricky in this case, since Lambda follows a chi-squared distribution only "half of the times". In the other 50% of the cases, it is just zero. Hence *c* is somewhat smaller, and one can show that

$$c = 2.86 \quad (95.4\%\,\mathrm{CL})$$



Significance

Calculating *p*-values

- Detecting a signal over a background usually means to show that the signal strength is different from zero with high significance.
- It is useful to take the ***test statistics*** based on the maximum log likelihood ratio, where we keep the signal flux at zero.

$$TS = \Lambda(\theta_1 = 0|\vec{x}) \equiv -2\ln\frac{\mathcal{L}(\theta_1, \hat{\theta}_2, \ldots, \hat{\theta}_n|\vec{x})}{\mathcal{L}(\hat{\theta}_1, \ldots, \hat{\theta}_n|\vec{x})}$$

- In absence of a signal, we have $TS \sim \chi^2_{k=1}$

- A large *TS* indicates the detection of a signal at high significance. The 5-sigma threshold corresponds to

$$TS \geq 25 \quad (5\sigma) \qquad\qquad p = 5.7 \times 10^{-7}$$

Notes:

- Often significances are quoted in terms of standard deviations.
- Significances depend strongly on the *tail of the distribution*, which is very sensitive to systematic uncertainties

Wilks' theorem: Sketch of the proof

1) Replacing data with maximum likelihood estimators

$$P(x_1,\ldots,x_N|\theta_1,\ldots,\theta_n) \longrightarrow P(\hat{\theta}_1,\ldots,\hat{\theta}_n|\theta_1,\ldots,\theta_n)$$

For a given PDF, $P(x_1,\ldots,x_N|\theta_1,\ldots,\theta_n)$, one can define the **maximum likelihood estimator**, $\hat{\theta}_j(x_1,\ldots,x_N)$. The MLE itself is also a random variable, and distributed according to some PDF $P(\hat{\theta}_1,\ldots,\hat{\theta}_n|\theta_1,\ldots,\theta_n)$.

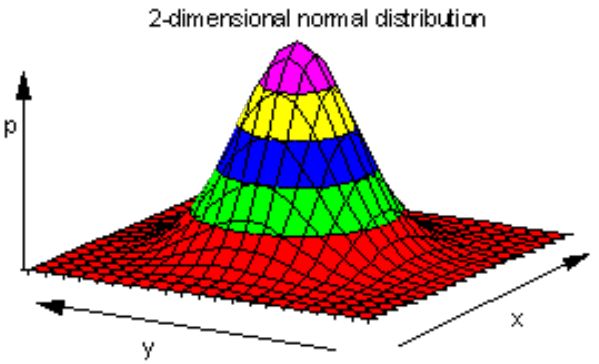
The trick is now to use *this* new PDF to study the properties of the maximum likelihood analysis.

2) Assuming that MLEs are normal distributed

We are interested in the case where **the maximum likelihood estimator are normal distributed.** This does not mean that the data has to be normal distributed.

$$P(\hat{\theta}_1,\ldots,\hat{\theta}_n|\theta_1,\ldots,\theta_n) \propto e^{-\frac{1}{2}\sum_{ij}(\hat{\theta}_i-\theta_i)\Sigma^{-1}_{ij}(\hat{\theta}_j-\theta_j)}$$

Very loosely speaking, this requirement means that "the data should be good enough to constrain the model parameters well".



Wilks' theorem: Sketch of the proof

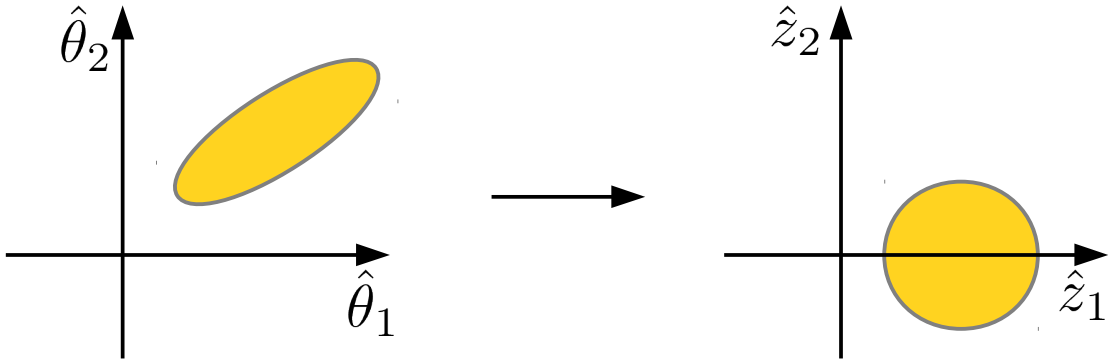
3) Rotating and rescaling to distribution with variance one

For simplicity, we rotate and rescale the model parameters such that the correlation matrix becomes the identity matrix. We introduce new model parameters

$$z_k = \sum_{j,i=1}^{n} R_{kj} U_{ji} \theta_i$$

such that

$$P(\hat{z}_1, \ldots, \hat{z}_n | z_1, \ldots, z_n) \propto e^{-\frac{1}{2}\sum_{i=1}^{n}(\hat{z}_i - z_i)^2}$$



4) Evaluate maximum log likelihood ratio

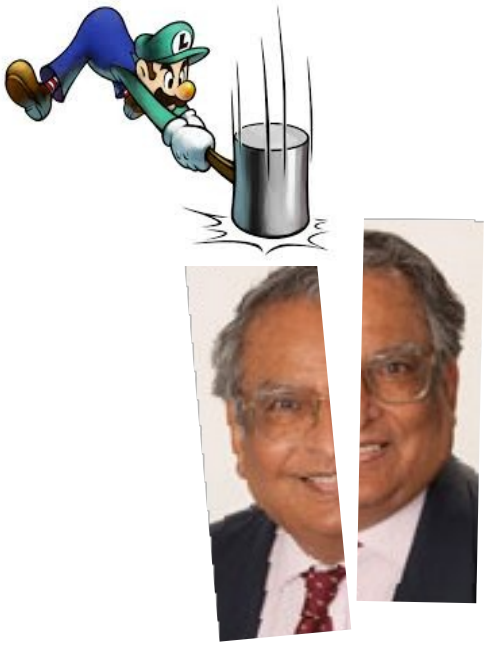
We now consider the maximum likelihood ratio, keeping *k* of the values fixed. The normalization factors cancel, and only *k* quadratic terms survive.

$$-2 \ln \frac{\max_{z'_{k+1},\ldots,z'_n} P(\hat{z}_1, \ldots, \hat{z}_n | z_1, \ldots, z_k, z'_{k+1}, \ldots, z'_n)}{\max_{z'_1,\ldots,z'_n} P(\hat{z}_1, \ldots, \hat{z}_n | z'_1, \ldots, z'_n)} = \sum_{i=1}^{k} (\hat{z}_i - z_i)^2$$

Per definition, this follows a chi-squared distribution with *k* degrees of freedom, $\sim \chi^2_k$!

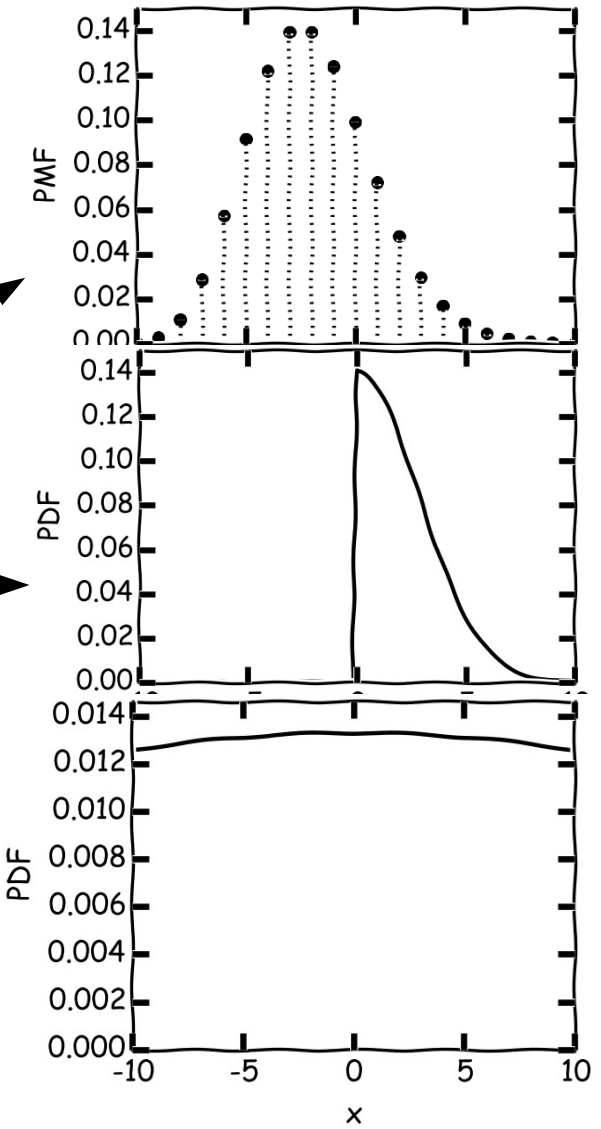
Breaking Wilks' theorem

Situation:
Maximum likelihood ratios can be approximated by a chi-squared distribution **if the MLEs of the model parameters are normal distributed.** Normal distribution of the data is often helpful, but not necessary.

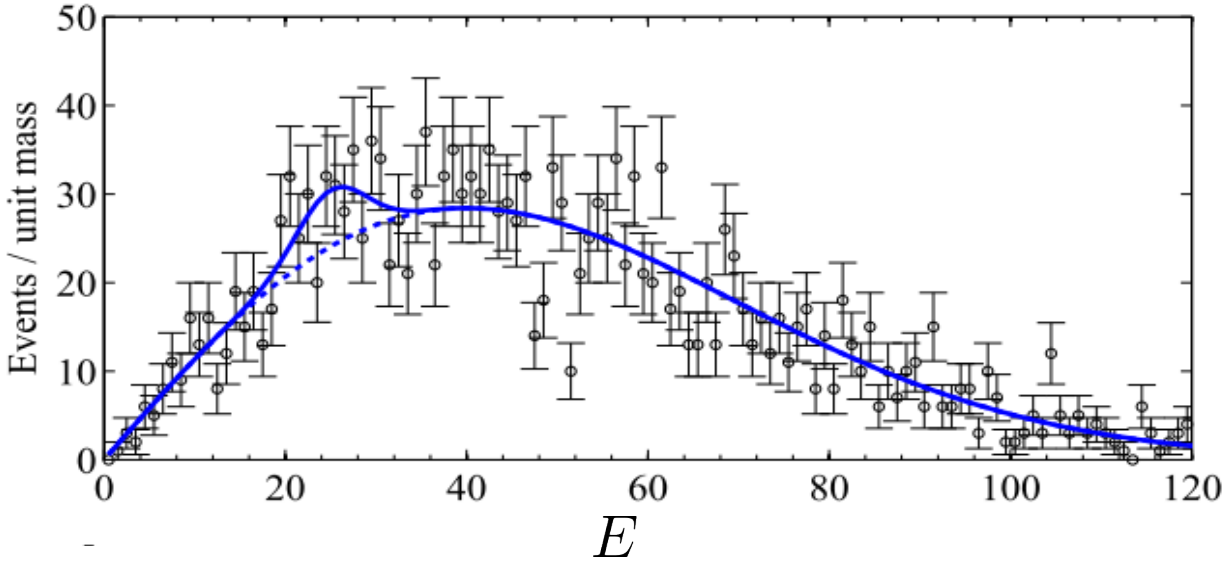


Ways to break Wilks' theorem

- Low number of events, Poisson noise
- Bounds on the model parameters (e.g. fluxes or masses can only be positive)
- Parameters that are irrelevant in the null hypothesis and have an (approximately) infinite variance



Example: Line searches



A very common scenario (X-ray and gamma-ray lines, Higgs bosons etc)

- Signal is "line" (narrow normal distribution with fixed width) with unknown strength and position
- Model for the BG depends on parameters

$$\frac{dN}{dE} = A_s \cdot N(E|\bar{E}, \Delta E) + bg(E|\zeta)$$

- The correct likelihood function is usually build up from Poisson distributions

$$\mathcal{L}(A_s, \bar{E}, \zeta|\vec{c}) = \prod_{i=1}^{n_{\mathrm{bins}}} P(c_i|\mu_i(A_s, \bar{E}, \zeta))$$

with expectation values $\mu_i(A_s, \bar{E}, \zeta) = \int_{E_i^-}^{E_i^+} dE' \frac{dN}{dE'}$ in energy range $[E_i^-, E_i^+]$

Unbinned analysis

Potential problems with a "binned" analysis (like the above Poisson likelihood)

- Result depends on binning of data
- Information loss due to binning

This can be avoided in an **unbinned likelihood analysis.** It is obtained from the Poisson likelihood in the limit of zero bin size, $n_{\mathrm{bins}} \to \infty$. In that limit, we find that

$$\mathcal{L} = \prod_{i=1}^{n_{\mathrm{bins}}} \frac{\mu_i^{k_i} e^{-\mu_i}}{k_i!} \qquad \text{with} \qquad \mu_i \ll 1 \quad \text{and} \quad k_i = 0, 1$$

This can be rewritten as a **unbinned likelihood function**

$$\mathcal{L} = P(n_{\mathrm{ev}}|\mu_{\mathrm{tot}}) \prod_{j=1}^{n_{\mathrm{ev}}} pdf(E_j)$$

where

$$\mu_{\mathrm{tot}} = \int_{E_{\mathrm{min}}}^{E_{\mathrm{max}}} \frac{dN}{dE}$$

(Total number of events)

$$pdf(E) = \frac{1}{\mu_{\mathrm{tot}}} \frac{dN}{dE}(E)$$

(PDF of *one* event)

Notes:

- The unbinned likelihood function does not depend on binning
- The total number of events enters as Poisson distribution
- Individual events are *not even close to a normal distribution,* ***but the maximum log likelihood ratio is likely still chi-squared distributed.***

Parameter boundaries

In many cases, we are interested in parameters that are bounded to be non-negative (fluxes, masses, etc). The confidence interval is then constructed according to

$$I(\vec{x}_{\rm obs}) = \{\theta | \Lambda(\theta|x_{\rm obs}) < c(\theta) \ \wedge \ \theta \geq 0\}$$

with the maximum log likelihood ratio given by

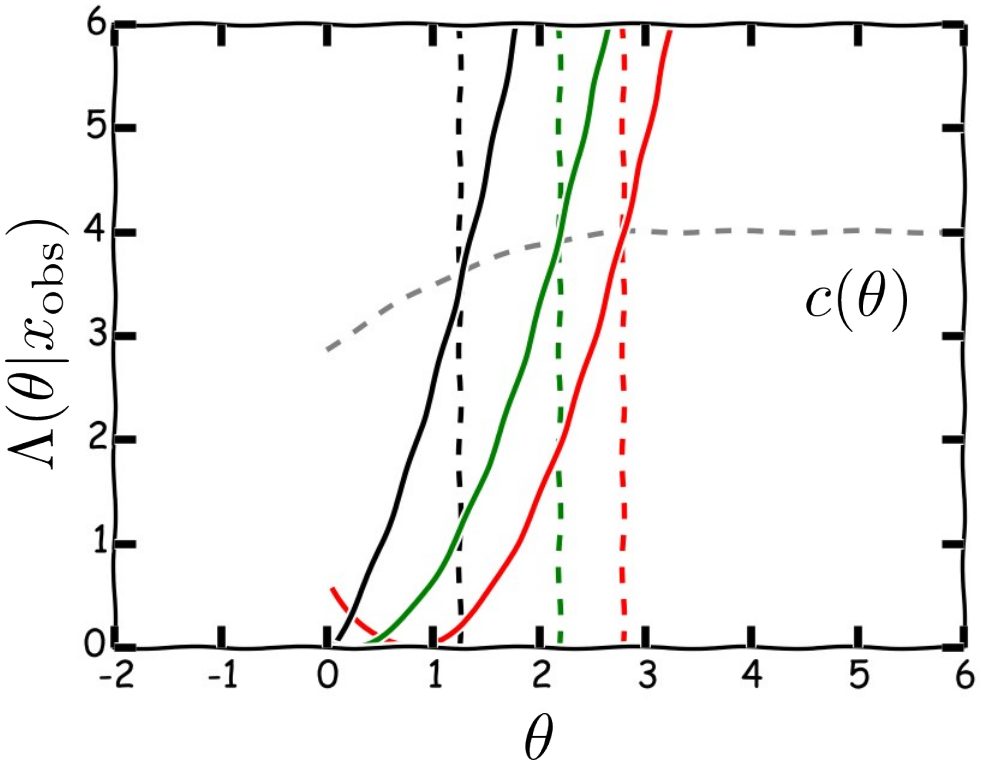
$$\Lambda(\theta|x) = -2 \ln \frac{\mathcal{L}(\theta|x)}{\max_{\theta \geq 0} \mathcal{L}(\theta|x)}$$

If the true value is theta=0, we obtain that the maximum likelihood ratio is zero in half of the cases, and follows a chi squared with one degree of freedom in the other half. Hence, the correct threshold value is

$$c(\theta = 0) = 2.86 \quad (95.4\%\,\mathrm{CL})$$

However, far away from the boundary, the value is again

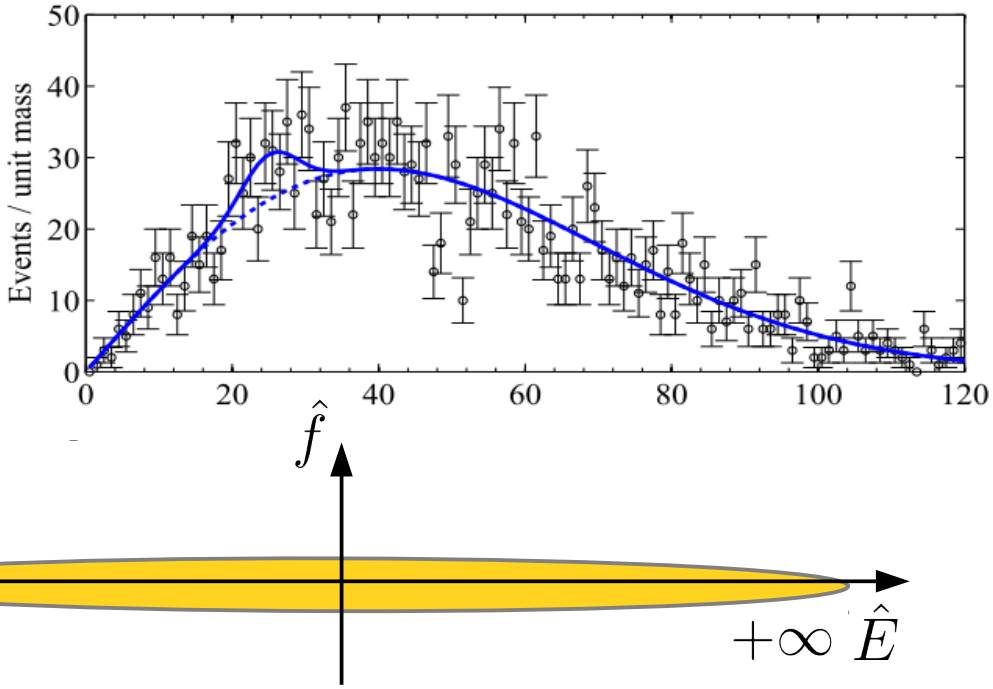
$$c(\theta \gg \Delta\theta) = 4 \quad (95.4\%\,\mathrm{CL})$$



Irrelevant parameters

Irrelevant parameters

- In case of a line search, the signal model has two free parameters
  - Normalization of the signal (flux)
  - Position of the signal (energy)
- In the case of the null hypothesis of zero signal (flux = 0), *the position of the signal is irrelevant* and does not affect the fit.



Consequences

- The variance in *E* direction is *infinite*.
- Hence, the rescaling of the parameters into parameters with variance one fails (step 3), and **Wilks' theorem breaks down.**
- The infinite variance of *E* indicates a *large number of trials*, which means that the TS does not follow a simple chi-squared distribution.

$$\Lambda(A_s, \bar{E}|\vec{x}) \nsim \chi^2_{k=2}$$

Solution

- Perform a Monte Carlo with mock data that does not contain a signal
- Naive estimate: two degrees of freedom, but several times

Trial corrections of an excess significance

"Look elsewhere effect"

- Often, one has to perform many *trials* in a search for a signal (e.g. because one does not know the position of a line, the position of a point source etc).
- Performing multiple trials increases the chance of seeing an upward fluctuation in at least one of the measurements.
- This effect must be corrected for when stating the significance of a signal that was found in an analysis with multiple trials.

Global vs local p-values

- *Local p-value*: Describes the probability of observing an upward fluctuation as large as the signal in one trial (e.g. for a known source or line position). This often follows the usual chi-squared distribution.
- *Global p-value*: Describes the probability of observing an upward fluctuation as large as the signal in *at least one of multiple trials.*

In general: $p_{\mathrm{global}} = (1 - p_{\mathrm{local}})^{n_{\mathrm{trials}}} \simeq n_{\mathrm{trials}}\, p_{\mathrm{local}}$

Examples:

$$\int_s^\infty N(x|0,1)\, dx = p$$

$5.0\sigma \quad p = 2.87 \times 10^{-7}$

	$n_{\mathrm{trials}} = 10$	$n_{\mathrm{trials}} = 100$
	$5.0\sigma \to 4.5\sigma$	$5.0\sigma \to 4.0\sigma$
	$4.0\sigma \to 3.4\sigma$	$4.0\sigma \to 2.7\sigma$
	$3.0\sigma \to 2.2\sigma$	$3.0\sigma \to 1.2\sigma$