

# Statistical Data Analysis

and Anomalies in Particle Physics and Astrophysics

## Lecture 2



4<sup>th</sup> KMI School

Nagoya University

15-17 December 2022

<https://www.kmi.nagoya-u.ac.jp/workshop/kmi-school-2022/>



Glen Cowan

Physics Department

Royal Holloway, University of London

[g.cowan@rhul.ac.uk](mailto:g.cowan@rhul.ac.uk)

[www.pp.rhul.ac.uk/~cowan](http://www.pp.rhul.ac.uk/~cowan)

# Outline

Lecture 1: Probability, Bayes vs. Frequentist  
Frequentist parameter estimation  
Hypothesis tests

→ Lecture 2:  $p$ -values  
Confidence limits  
Systematic uncertainties  
Bayesian parameter estimation

Lecture 3: Significance, sensitivity  
Bayes factors  
Models for anomalies

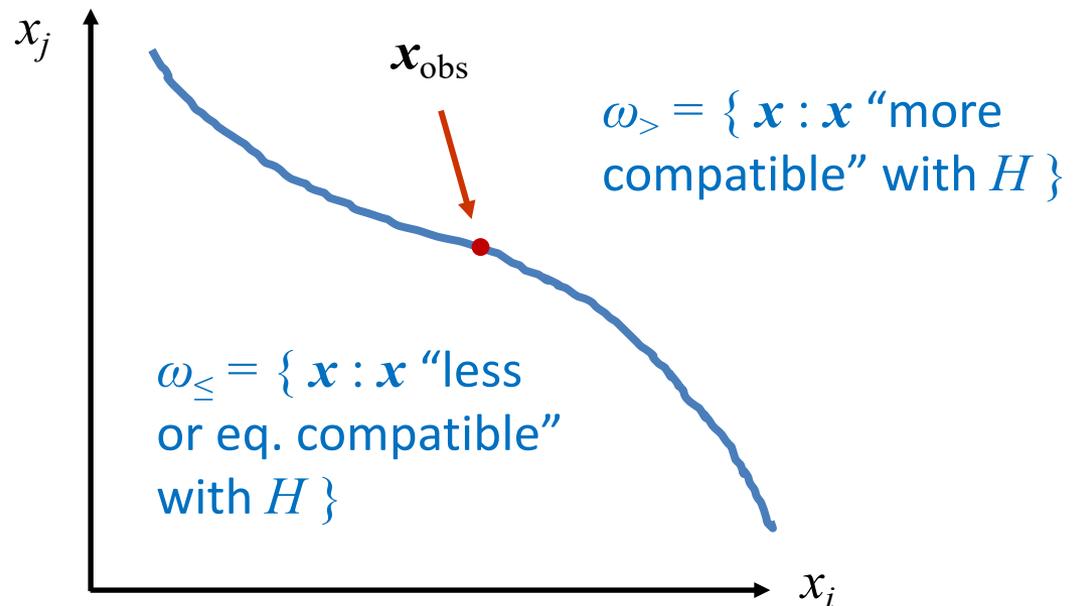
# Testing significance / goodness-of-fit

Suppose hypothesis  $H$  predicts pdf  $f(\mathbf{x}|H)$  for a set of observations  $\mathbf{x} = (x_1, \dots, x_n)$ .

We observe a single point in this space:  $\mathbf{x}_{\text{obs}}$ .

How can we quantify the level of compatibility between the data and the predictions of  $H$ ?

Decide what part of the data space represents equal or less compatibility with  $H$  than does the point  $\mathbf{x}_{\text{obs}}$ . (Not unique!)



# $p$ -values

Express level of compatibility between data and hypothesis (sometimes ‘goodness-of-fit’) by giving the  $p$ -value for  $H$ :

$$p = P(\mathbf{x} \in \omega_{\leq}(\mathbf{x}_{\text{obs}}) | H)$$

- = probability, under assumption of  $H$ , to observe data with equal or lesser compatibility with  $H$  relative to the data we got.
- = probability, under assumption of  $H$ , to observe data as discrepant with  $H$  as the data we got or more so.

Basic idea: if there is only a very small probability to find data with even worse (or equal) compatibility, then  $H$  is “disfavoured by the data”.

If the  $p$ -value is below a user-defined threshold  $\alpha$  (e.g. 0.05) then  $H$  is rejected (equivalent to hypothesis test of size  $\alpha$  as seen earlier).



## $p$ -value of $H$ is not $P(H)$

The  $p$ -value of  $H$  is not the probability that  $H$  is true!

In frequentist statistics we don't talk about  $P(H)$  (unless  $H$  represents a repeatable observation).

If we do define  $P(H)$ , e.g., in Bayesian statistics as a degree of belief, then we need to use Bayes' theorem to obtain

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H) dH}$$

where  $\pi(H)$  is the prior probability for  $H$ .

For now stick with the frequentist approach;  
result is  $p$ -value, regrettably easy to misinterpret as  $P(H)$ .

# The Poisson counting experiment

Suppose we do a counting experiment and observe  $n$  events.

Events could be from *signal* process or from *background* – we only count the total number.

Poisson model:

$$P(n|s, b) = \frac{(s + b)^n}{n!} e^{-(s+b)}$$

$s$  = mean (i.e., expected) # of signal events

$b$  = mean # of background events

Goal is to make inference about  $s$ , e.g.,

test  $s = 0$  (rejecting  $H_0 \approx$  “discovery of signal process”)

test all non-zero  $s$  (values not rejected = confidence interval)

In both cases need to ask what is relevant alternative hypothesis.

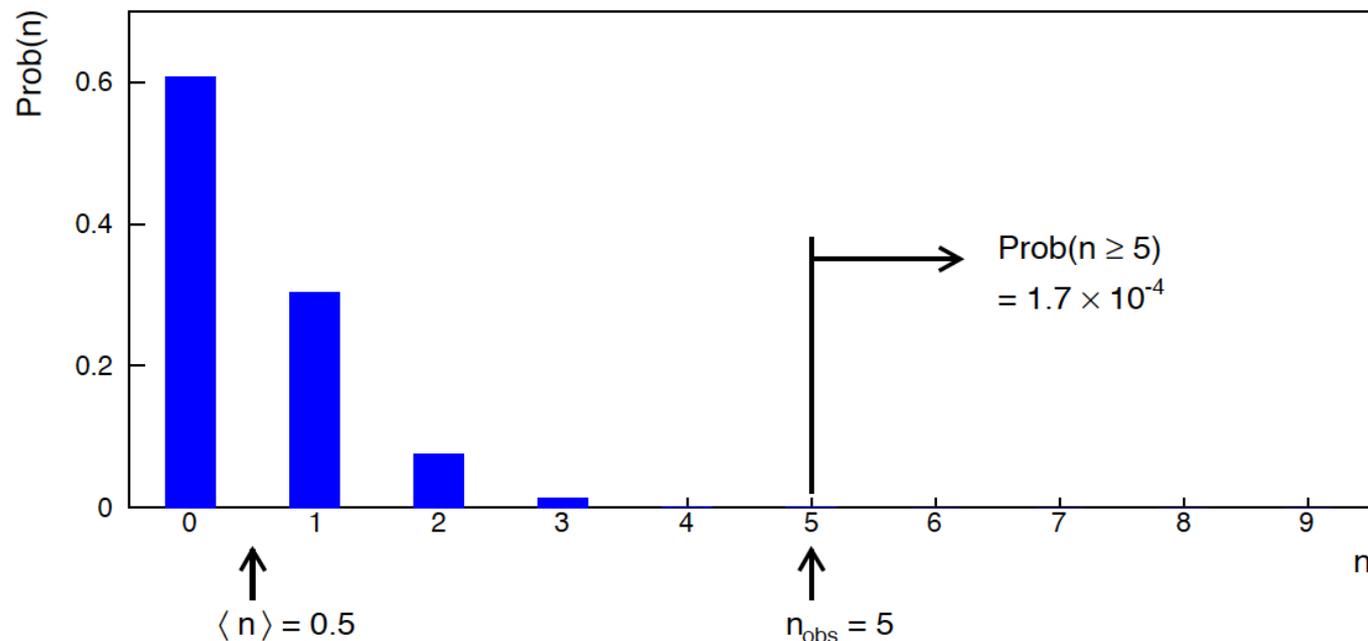
# Poisson counting experiment: discovery $p$ -value

Suppose  $b = 0.5$  (known), and we observe  $n_{\text{obs}} = 5$ .

Should we claim evidence for a new discovery?

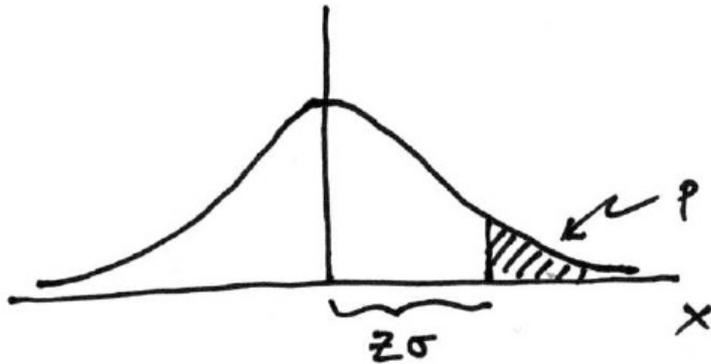
Give  $p$ -value for hypothesis  $s = 0$ , suppose relevant alt. is  $s > 0$ .

$$\begin{aligned} p\text{-value} &= P(n \geq 5; b = 0.5, s = 0) \\ &= 1.7 \times 10^{-4} \neq P(s = 0)! \end{aligned}$$



# Significance from $p$ -value

Often define significance  $Z$  as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same  $p$ -value.



$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z)$$

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

in ROOT:

```
p = 1 - TMath::Freq(Z)
```

```
Z = TMath::NormQuantile(1-p)
```

in python (scipy.stats):

```
p = 1 - norm.cdf(Z) = norm.sf(Z)
```

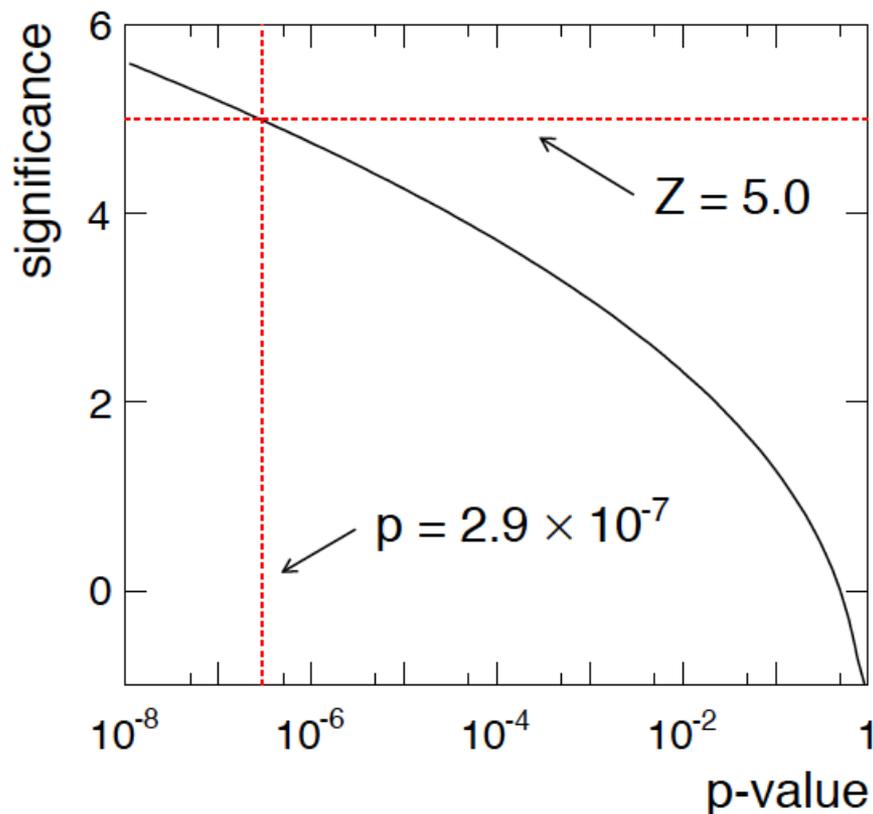
```
Z = norm.ppf(1-p)
```

Result  $Z$  is a “number of sigmas”. Note this does not mean that the original data was Gaussian distributed.

# Poisson counting experiment: discovery significance

Equivalent significance for  $p = 1.7 \times 10^{-4}$ :  $Z = \Phi^{-1}(1 - p) = 3.6$

Often claim discovery if  $Z > 5$  ( $p < 2.9 \times 10^{-7}$ , i.e., a “5-sigma effect”)



In fact this tradition should be revisited:  $p$ -value intended to quantify probability of a signal-like fluctuation assuming background only; not intended to cover, e.g., hidden systematics, plausibility signal model, compatibility of data with signal, “look-elsewhere effect” ( $\sim$ multiple testing), etc.

# Confidence intervals by inverting a test

In addition to a ‘point estimate’ of a parameter we should report an interval reflecting its statistical uncertainty.

**Confidence intervals** for a parameter  $\theta$  can be found by defining a test of the hypothesized value  $\theta$  (do this for all  $\theta$ ):

Specify values of the data that are ‘disfavoured’ by  $\theta$  (critical region) such that  $P(\text{data in critical region} | \theta) \leq \alpha$  for a prespecified  $\alpha$ , e.g., 0.05 or 0.1.

If data observed in the critical region, reject the value  $\theta$ .

Now invert the test to define a confidence interval as:

set of  $\theta$  values that are not rejected in a test of size  $\alpha$  (confidence level CL is  $1 - \alpha$ ).

# Relation between confidence interval and $p$ -value

Equivalently we can consider a significance test for each hypothesized value of  $\theta$ , resulting in a  $p$ -value,  $p_\theta$ .

If  $p_\theta \leq \alpha$ , then we reject  $\theta$ .

The confidence interval at  $CL = 1 - \alpha$  consists of those values of  $\theta$  that are not rejected.

E.g. an upper limit on  $\theta$  is the greatest value for which  $p_\theta > \alpha$ .

In practice find by setting  $p_\theta = \alpha$  and solve for  $\theta$ .

For a multidimensional parameter space  $\theta = (\theta_1, \dots, \theta_M)$  use same idea – result is a confidence “region” with boundary determined by  $p_\theta = \alpha$ .

# Coverage probability of confidence interval

If the true value of  $\theta$  is rejected, then it's not in the confidence interval. The probability for this is by construction (equality for continuous data):

$$P(\text{reject } \theta | \theta) \leq \alpha = \text{type-I error rate}$$

Therefore, the probability for the interval to contain or “cover”  $\theta$  is

$$P(\text{conf. interval “covers” } \theta | \theta) \geq 1 - \alpha$$

This assumes that the set of  $\theta$  values considered includes the true value, i.e., it assumes the composite hypothesis  $P(x|H, \theta)$ .

# Frequentist upper limit on Poisson parameter

Consider again the case of observing  $n \sim \text{Poisson}(s + b)$ .

Suppose  $b = 4.5$ ,  $n_{\text{obs}} = 5$ . Find upper limit on  $s$  at 95% CL.

Relevant alternative is  $s = 0$  (critical region at low  $n$ )

$p$ -value of hypothesized  $s$  is  $P(n \leq n_{\text{obs}}; s, b)$

Upper limit  $s_{\text{up}}$  at  $\text{CL} = 1 - \alpha$  found from

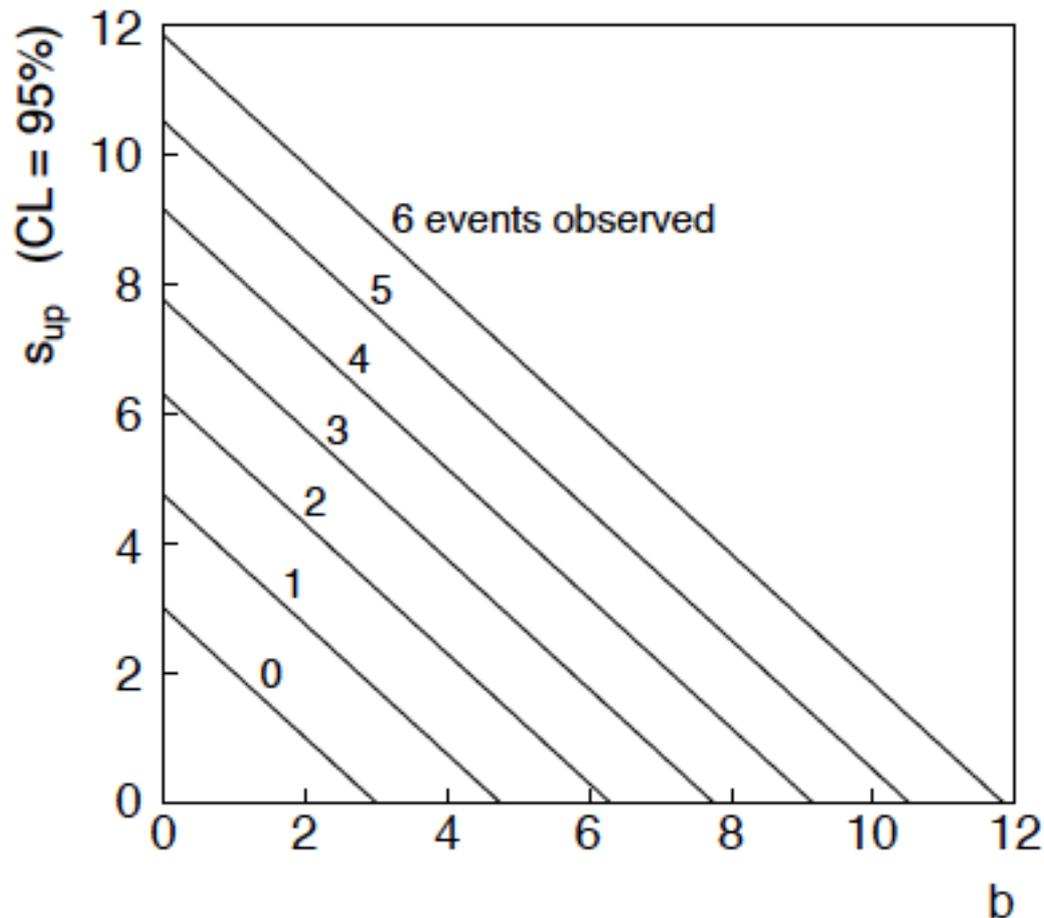
$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}} + b)}$$

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1}(1 - \alpha; 2(n_{\text{obs}} + 1)) - b$$

$$= \frac{1}{2} F_{\chi^2}^{-1}(0.95; 2(5 + 1)) - 4.5 = 6.0$$

# $n \sim \text{Poisson}(s+b)$ : frequentist upper limit on $s$

For low fluctuation of  $n$ , formula can give negative result for  $s_{\text{up}}$ ; i.e. confidence interval is empty; all values of  $s \geq 0$  have  $p_s \leq \alpha$ .



# Limits near a boundary of the parameter space

Suppose e.g.  $b = 2.5$  and we observe  $n = 0$ .

If we choose  $CL = 0.9$ , we find from the formula for  $s_{\text{up}}$

$$s_{\text{up}} = -0.197 \quad (CL = 0.90)$$

## Physicist:

We already knew  $s \geq 0$  before we started; can't use negative upper limit to report result of expensive experiment!

## Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma when testing parameter values for which one has very little experimental sensitivity, e.g., very small  $s$ .

# Expected limit for $s = 0$

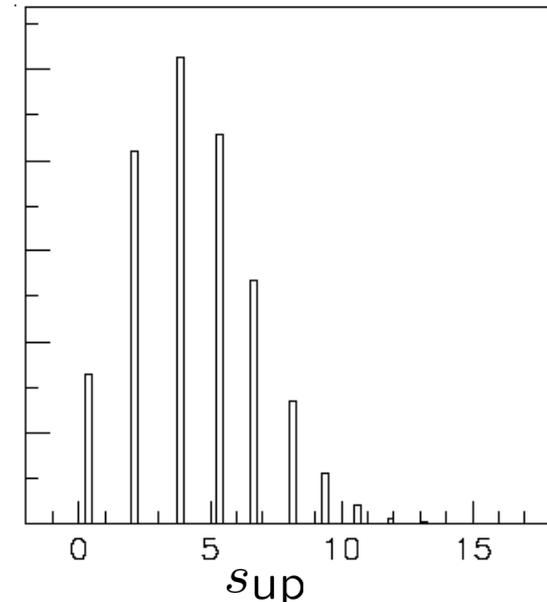
Physicist: I should have used  $CL = 0.95$  — then  $s_{up} = 0.496$

Even better: for  $CL = 0.917923$  we get  $s_{up} = 10^{-4}$  !

Reality check: with  $b = 2.5$ , typical Poisson fluctuation in  $n$  is at least  $\sqrt{2.5} = 1.6$ . How can the limit be so low?

Look at the mean limit for the no-signal hypothesis ( $s = 0$ ) (sensitivity).

Distribution of 95% CL limits with  $b = 2.5$ ,  $s = 0$ .  
Mean upper limit = 4.44



# The Bayesian approach to limits

In Bayesian statistics need to start with ‘prior pdf’  $\pi(\theta)$ , this reflects degree of belief about  $\theta$  before doing the experiment.

Bayes’ theorem tells how our beliefs should be updated in light of the data  $x$ :

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta) d\theta} \propto p(x|\theta)\pi(\theta)$$

Integrate posterior pdf  $p(\theta|x)$  to give interval with any desired probability content.

For e.g.  $n \sim \text{Poisson}(s+b)$ , 95% CL upper limit on  $s$  from

$$0.95 = \int_{-\infty}^{s_{\text{sup}}} p(s|n) ds$$

# Bayesian prior for Poisson parameter

Include knowledge that  $s \geq 0$  by setting prior  $\pi(s) = 0$  for  $s < 0$ .

Could try to reflect 'prior ignorance' with e.g.

$$\pi(s) = \begin{cases} 1 & s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Not normalized; can be OK provided  $p(n|s)$  dies off quickly for large  $s$ .

Not invariant under change of parameter — if we had used instead a flat prior for a nonlinear function of  $s$ , then this would imply a non-flat prior for  $s$ .

Doesn't really reflect a reasonable degree of belief, but often used as a point of reference; or viewed as a recipe for producing an interval whose frequentist properties can be studied (e.g., coverage probability, which will depend on true  $s$ ).

# Bayesian upper limit with flat prior for $s$

Put Poisson likelihood and flat prior into Bayes' theorem:

$$p(s|n) \propto p(n|s)\pi(s) = \frac{(s+b)^n}{n!} e^{-(s+b)} \times 1, \quad s \geq 0$$

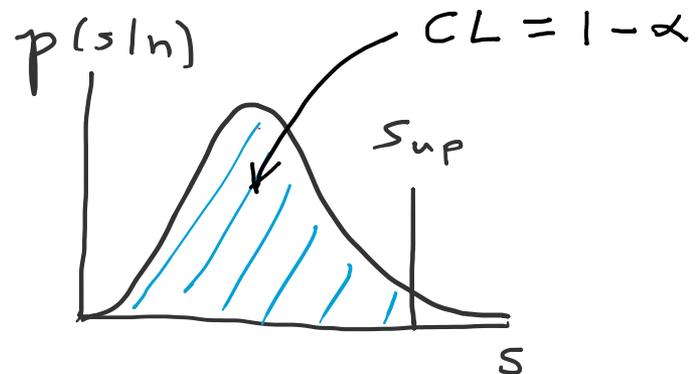
Normalize to unit area:

$$p(s|n) = \frac{(s+b)^n e^{-(s+b)}}{\Gamma(b, n+1)}$$

upper incomplete  
gamma function

Upper limit  $s_{\text{up}}$  determined by

$$1 - \alpha = \int_0^{s_{\text{up}}} p(s|n) ds$$



# Bayesian interval with flat prior for $s$

Solve to find limit  $s_{\text{up}}$ :

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1} [p, 2(n+1)] - b$$

where

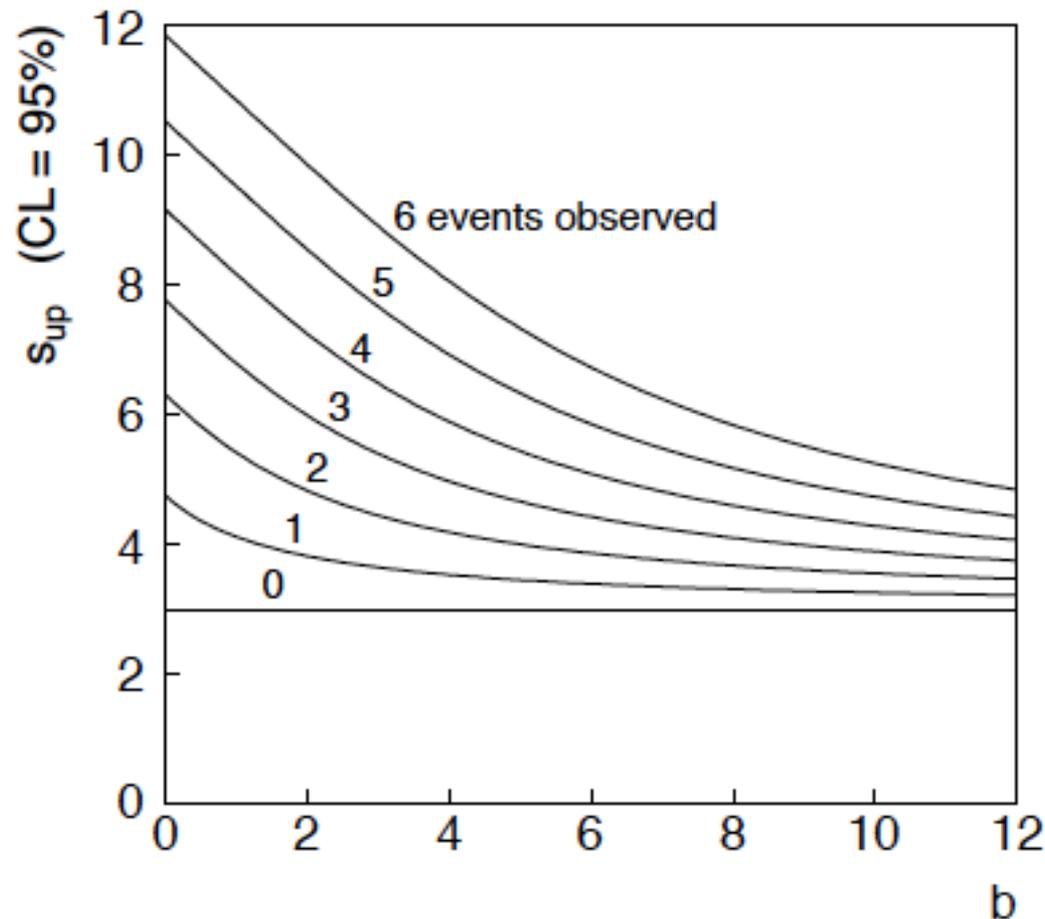
$$p = 1 - \alpha \left( 1 - F_{\chi^2} [2b, 2(n+1)] \right)$$

For special case  $b = 0$ , Bayesian upper limit with flat prior numerically same as one-sided frequentist case ('coincidence').

# Bayesian interval with flat prior for $s$

For  $b > 0$  Bayesian limit is everywhere greater than the (one sided) frequentist upper limit.

Never goes negative. Doesn't depend on  $b$  if  $n = 0$ .



# Approximate confidence intervals/regions from the likelihood function

Suppose we test parameter value(s)  $\theta = (\theta_1, \dots, \theta_N)$  using the ratio

$$\lambda(\theta) = \frac{L(\theta)}{L(\hat{\theta})} \quad 0 \leq \lambda(\theta) \leq 1$$

Lower  $\lambda(\theta)$  means worse agreement between data and hypothesized  $\theta$ . Equivalently, usually define

$$t_\theta = -2 \ln \lambda(\theta)$$

so higher  $t_\theta$  means worse agreement between  $\theta$  and the data.

$p$ -value of  $\theta$  therefore

$$p_\theta = \int_{t_{\theta, \text{obs}}}^{\infty} f(t_\theta | \theta) dt_\theta$$

need pdf

# Confidence region from Wilks' theorem

Wilks' theorem says (in large-sample limit and provided certain conditions hold...)

$$f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta}) \sim \chi_N^2$$

chi-square dist. with # d.o.f. =  
# of components in  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)$ .

Assuming this holds, the  $p$ -value is

$$p_{\boldsymbol{\theta}} = 1 - F_{\chi_N^2}(t_{\boldsymbol{\theta}}|\boldsymbol{\theta}) \quad \leftarrow \text{set equal to } \alpha$$

To find boundary of confidence region set  $p_{\boldsymbol{\theta}} = \alpha$  and solve for  $t_{\boldsymbol{\theta}}$ :

$$t_{\boldsymbol{\theta}} = F_{\chi_N^2}^{-1}(1 - \alpha)$$

Recall also

$$t_{\boldsymbol{\theta}} = -2 \ln \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}$$

# Confidence region from Wilks' theorem (cont.)

i.e., boundary of confidence region in  $\theta$  space is where

$$\ln L(\boldsymbol{\theta}) = \ln L(\hat{\boldsymbol{\theta}}) - \frac{1}{2} F_{\chi_N^2}^{-1}(1 - \alpha)$$

For example, for  $1 - \alpha = 68.3\%$  and  $n = 1$  parameter,

$$F_{\chi_1^2}^{-1}(0.683) = 1$$

and so the 68.3% confidence level interval is determined by

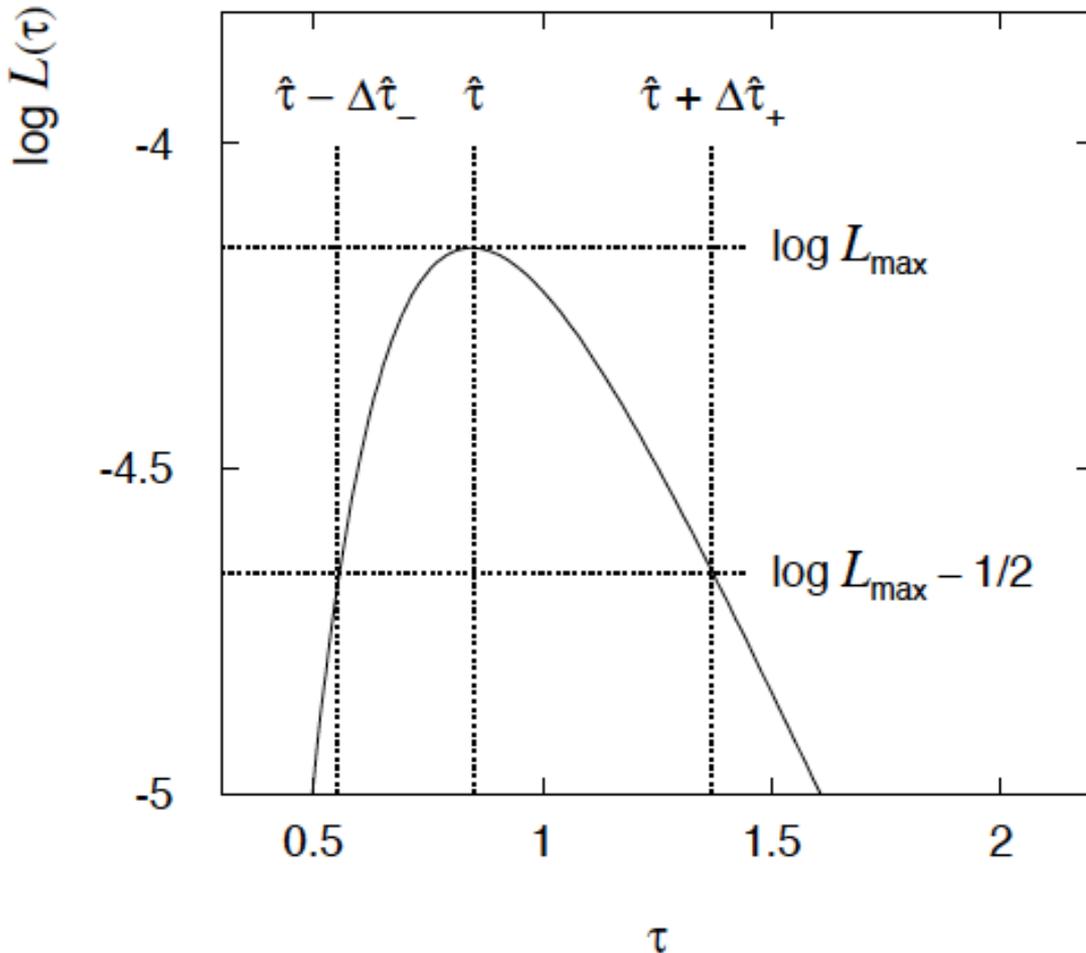
$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

Same as recipe for finding the estimator's standard deviation, i.e.,

$[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$  is a 68.3% CL confidence interval.

# Example of interval from $\ln L(\theta)$

For  $N=1$  parameter,  $CL = 0.683$ ,  $Q_\alpha = 1$ .



Our exponential example, now with only  $n = 5$  events.

Can report ML estimate with approx. confidence interval from  $\ln L_{\max} - 1/2$  as “asymmetric error bar”:

$$\hat{\tau} = 0.85_{-0.30}^{+0.52}$$

# Multiparameter case

For increasing number of parameters,  $CL = 1 - \alpha$  decreases for confidence region determined by a given

$$Q_\alpha = F_{\chi_n^2}^{-1}(1 - \alpha)$$

$Q_\alpha$	$1 - \alpha$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

← # of par.

# Multiparameter case (cont.)

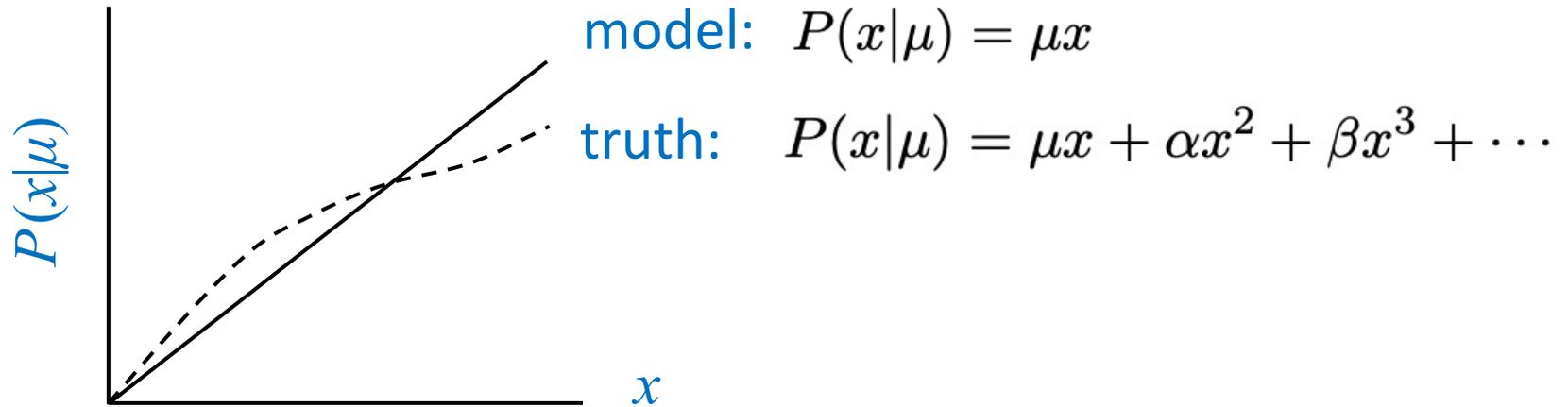
Equivalently,  $Q_\alpha$  increases with  $n$  for a given  $CL = 1 - \alpha$ .

$1 - \alpha$	$\bar{Q}_\alpha$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

← # of par.

# Systematic uncertainties and nuisance parameters

In general, our model of the data is not perfect:



Can improve model by including additional adjustable parameters.

$$P(x|\mu) \rightarrow P(x|\mu, \boldsymbol{\theta})$$

Nuisance parameter  $\leftrightarrow$  systematic uncertainty. Some point in the parameter space of the enlarged model should be “true”.

Presence of nuisance parameter decreases sensitivity of analysis to the parameter of interest (e.g., increases variance of estimate).

# Profile Likelihood

Suppose we have a likelihood  $L(\boldsymbol{\mu}, \boldsymbol{\theta}) = P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\theta})$  with  $N$  parameters of interest  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$  and  $M$  nuisance parameters  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)$ . The “profiled” (or “constrained”) values of  $\boldsymbol{\theta}$  are:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\mu}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} L(\boldsymbol{\mu}, \boldsymbol{\theta})$$

and the profile likelihood is:  $L_p(\boldsymbol{\mu}) = L(\boldsymbol{\mu}, \hat{\boldsymbol{\theta}})$

The profile likelihood depends only on the parameters of interest; the nuisance parameters are replaced by their profiled values.

The profile likelihood can be used to obtain confidence intervals/regions for the parameters of interest in the same way as one would for all of the parameters from the full likelihood.

# Example: fitting a straight line

Data:  $(x_i, y_i, \sigma_i)$ ,  $i = 1, \dots, n$ .

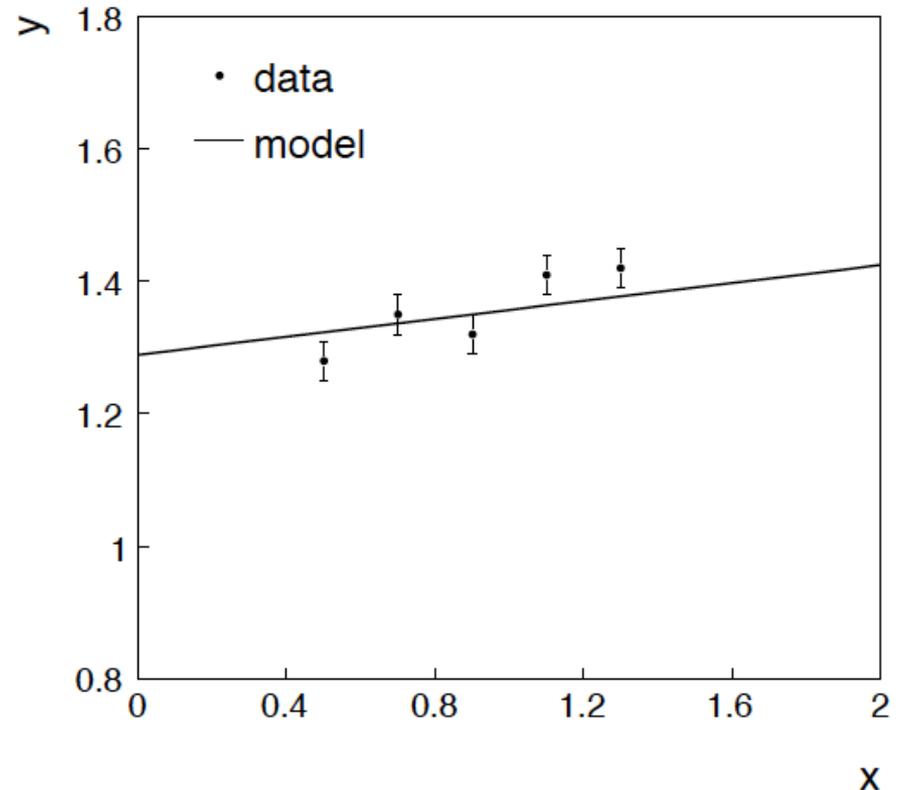
Model:  $y_i$  independent and all follow  $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

$$\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x,$$

assume  $x_i$  and  $\sigma_i$  known.

Goal: estimate  $\theta_0$

Here suppose we don't care about  $\theta_1$  (example of a "nuisance parameter")



# Maximum likelihood fit with Gaussian data

In this example, the  $y_i$  are assumed independent, so the likelihood function is a product of Gaussians:

$$L(\theta_0, \theta_1) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[ -\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right],$$

Maximizing the likelihood is here equivalent to minimizing

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

i.e., for Gaussian data, ML same as Method of Least Squares (LS)

# $\theta_1$ known a priori

$$L(\theta_0) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[ -\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right].$$

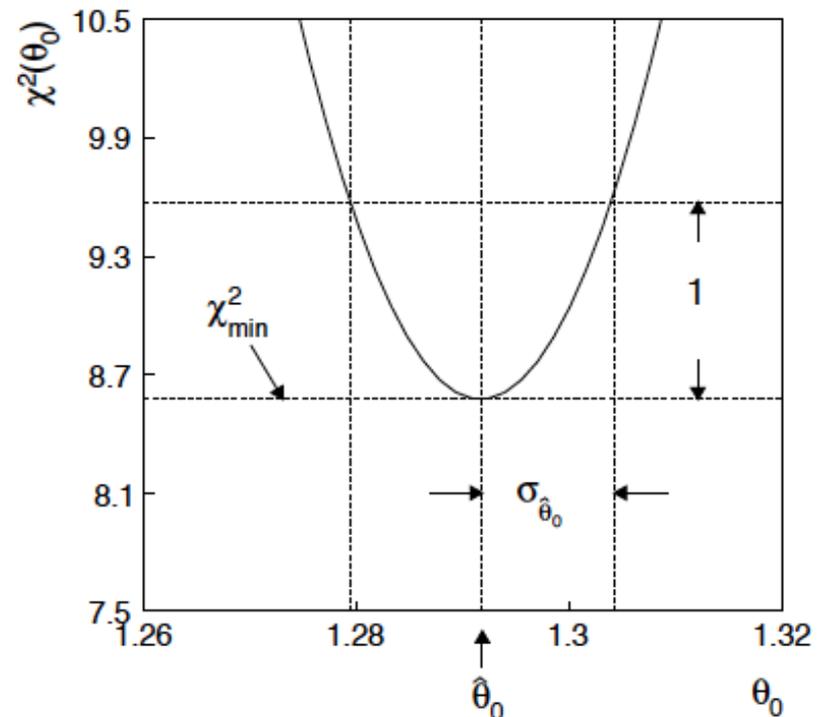
$$\chi^2(\theta_0) = -2 \ln L(\theta_0) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

For Gaussian  $y_i$ , ML same as LS

Minimize  $\chi^2 \rightarrow$  estimator  $\hat{\theta}_0$ .

Come up one unit from  $\chi_{\min}^2$

to find  $\sigma_{\hat{\theta}_0}$ .



# ML (or LS) fit of $\theta_0$ and $\theta_1$

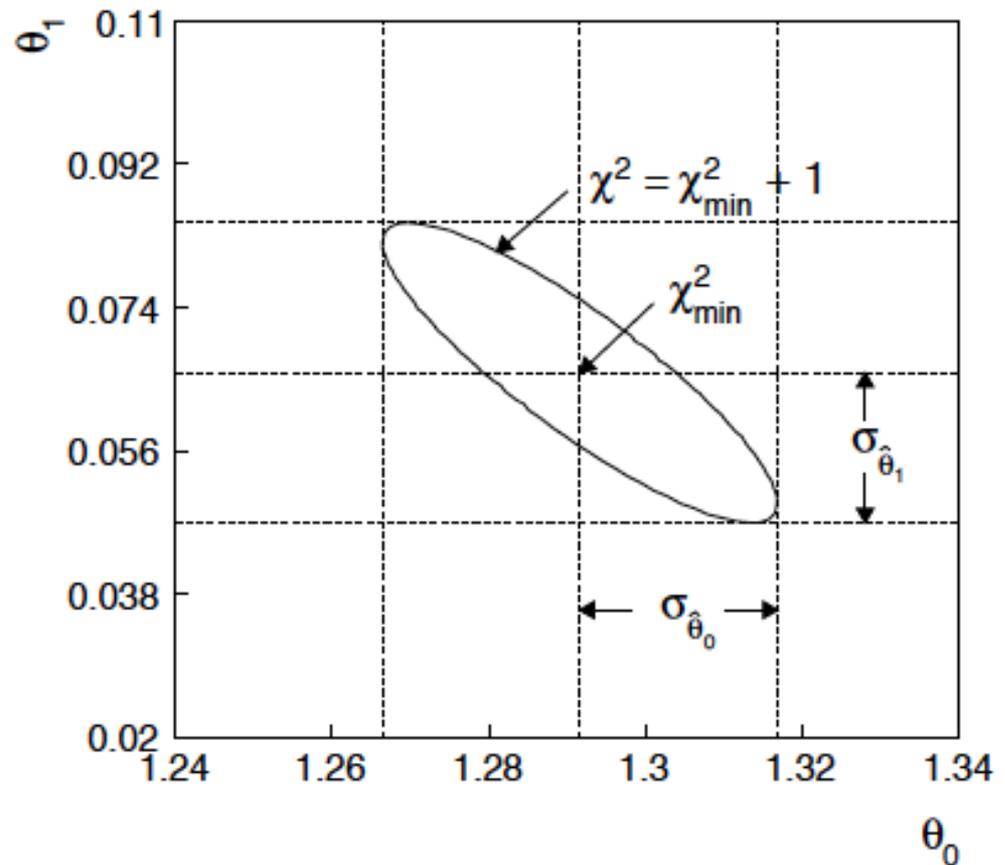
$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

Standard deviations from  
tangent lines to contour

$$\chi^2 = \chi_{\min}^2 + 1 .$$

Correlation between

$\hat{\theta}_0$ ,  $\hat{\theta}_1$  causes errors  
to increase.

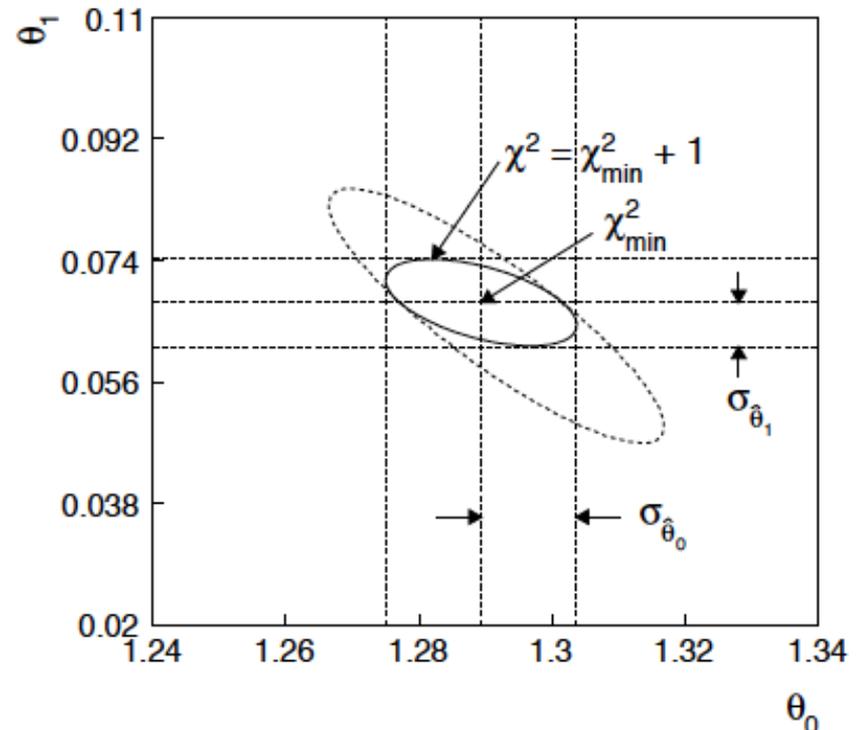


If we have a measurement  $t_1 \sim \text{Gauss}(\theta_1, \sigma_{t_1})$

$$L(\theta_0, \theta_1) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-(t_1 - \theta_1)^2 / 2\sigma_{t_1}^2} \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[ -\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right]$$

$$\chi^2(\theta_0, \theta_1) = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} + \frac{(t_1 - \theta_1)^2}{\sigma_{t_1}^2}$$

The information on  $\theta_1$   
improves accuracy of  $\hat{\theta}_0$ .



# Reminder of Bayesian approach

In Bayesian statistics we can associate a probability with a hypothesis, e.g., a parameter value  $\theta$ .

Interpret probability of  $\theta$  as ‘degree of belief’ (subjective).

Need to start with ‘prior pdf’  $\pi(\theta)$ , this reflects degree of belief about  $\theta$  before doing the experiment.

Our experiment has data  $x$ ,  $\rightarrow$  likelihood  $L(x|\theta)$ .

Bayes’ theorem tells how our beliefs should be updated in light of the data  $x$ :

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta') d\theta'} \propto L(x|\theta)\pi(\theta)$$

Posterior pdf  $p(\theta|x)$  contains all our knowledge about  $\theta$ .

# Bayesian approach: $y_i \sim \text{Gauss}(\mu(x_i; \theta_0, \theta_1), \sigma_i)$

We need to associate prior probabilities with  $\theta_0$  and  $\theta_1$ , e.g.,

$$\pi(\theta_0, \theta_1) = \pi_0(\theta_0)\pi_1(\theta_1) \quad \leftarrow \text{suppose knowledge of } \theta_0 \text{ has no influence on knowledge of } \theta_1$$

$$\pi_0(\theta_0) = \text{const.} \quad \leftarrow \text{'non-informative', in any case much broader than } L(\theta_0)$$

$$\pi_1(\theta_1) = p(\theta_1|t_1) \propto p(t_1|\theta_1)\pi_{\text{Ur}}(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-(t_1-\theta_1)^2/2\sigma_t^2} \times \text{const.}$$

prior after  $t_1$ ,  
before  $\mathbf{y}$

Ur = "primordial"  
prior

Likelihood for control  
measurement  $t_1$

# Bayesian example: $y_i \sim \text{Gauss}(\mu(x_i; \theta_0, \theta_1), \sigma_i)$

Putting the ingredients into Bayes' theorem gives:

$$p(\theta_0, \theta_1 | \vec{y}) \propto \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu(x_i; \theta_0, \theta_1))^2 / 2\sigma_i^2} \pi_0 \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2 / 2\sigma_{t_1}^2}$$

posterior  $\propto$  likelihood  $\times$  prior



Note here the likelihood only reflects the measurements  $\mathbf{y}$ .

The information from the control measurement  $t_1$  has been put into the prior for  $\theta_1$ .

We would get the same result using the likelihood  $P(\mathbf{y}, t | \theta_0, \theta_1)$  and the constant “Ur-prior” for  $\theta_1$ .

# Marginalizing the posterior pdf

We then integrate (marginalize)  $p(\theta_0, \theta_1 | \mathbf{y})$  to find  $p(\theta_0 | \mathbf{y})$ :

$$p(\theta_0 | \mathbf{y}) = \int p(\theta_0, \theta_1 | \mathbf{y}) d\theta_1$$

In this example we can do the integral (rare). We find

$$p(\theta_0 | \mathbf{y}) = \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2 / 2\sigma_{\theta_0}^2}$$

$$\hat{\theta}_0 = \text{same as MLE}$$

$$\sigma_{\theta_0} = \sigma_{\hat{\theta}_0} \quad (\text{same as for MLE})$$

For this example, numbers come out same as in frequentist approach, but interpretation different.

# Marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x) d\theta_1 .$$

often high dimensionality and impossible in closed form,  
also impossible with ‘normal’ acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized  
Bayesian computation.

MCMC (e.g., Metropolis-Hastings algorithm) generates  
correlated sequence of random numbers:

cannot use for many applications, e.g., detector MC;  
effective stat. error greater than if all values independent .

Basic idea: sample multidimensional  $\theta$  but look only at  
distribution of parameters of interest.

# MCMC basics: Metropolis-Hastings algorithm

Goal: given an  $n$ -dimensional pdf  $p(\theta)$  up to a proportionality constant, generate a sequence of points  $\theta_1, \theta_2, \theta_3, \dots$

1) Start at some point  $\vec{\theta}_0$

2) Generate  $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$

Proposal density  $q(\theta; \theta_0)$   
e.g. Gaussian centred about  $\theta_0$

3) Form test ratio  $\alpha = \min \left[ 1, \frac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)} \right]$

4) Generate  $u \sim \text{Uniform}[0, 1]$

5) If  $u \leq \alpha$ ,  $\vec{\theta}_1 = \vec{\theta}$ , ← move to proposed point

else  $\vec{\theta}_1 = \vec{\theta}_0$  ← old point repeated

6) Iterate

# Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

Still works if  $p(\theta)$  is known only as a proportionality, which is usually what we have from Bayes' theorem:  $p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)\pi(\theta)$ .

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric:  $q(\theta; \theta_0) = q(\theta_0; \theta)$

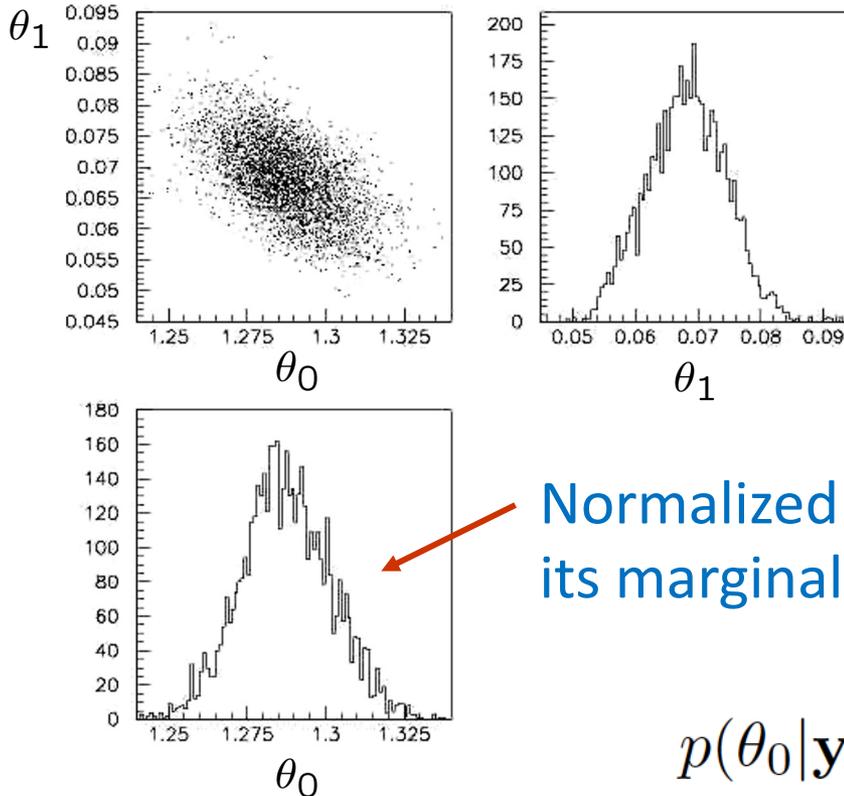
Test ratio is (*Metropolis-Hastings*):  $\alpha = \min \left[ 1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)} \right]$

I.e. if the proposed step is to a point of higher  $p(\theta)$ , take it; if not, only take the step with probability  $p(\theta)/p(\theta_0)$ .

If proposed step rejected, repeat the current point.

# Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:



Normalized histogram of  $\theta_0$  gives its marginal posterior pdf:

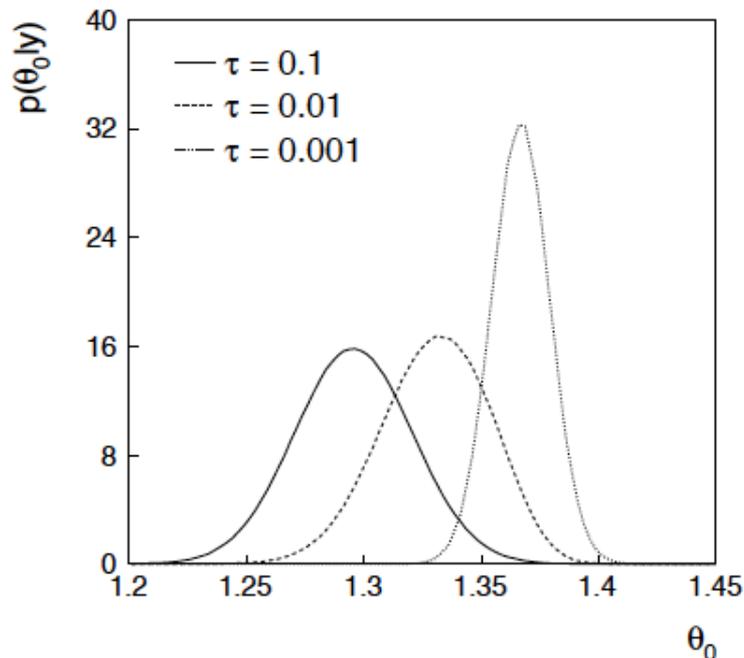
$$p(\theta_0|\mathbf{y}) = \int p(\theta_0, \theta_1|\mathbf{y}) d\theta_1$$

# Bayesian method with alternative priors

Suppose we don't have a previous measurement of  $\theta_1$  but rather, an “expert” says it should be positive and not too much greater than 0.1 or so, i.e., something like

$$\pi_1(\theta_1) = \frac{1}{\tau} e^{-\theta_1/\tau}, \quad \theta_1 \geq 0, \quad \tau = 0.1.$$

From this we obtain (numerically) the posterior pdf for  $\theta_0$ :



This summarizes all knowledge about  $\theta_0$ .

Look also at result from variety of priors.

# Extra slides

# Priors from formal rules

Last time we took the prior for a Poisson mean to be constant to reflect a lack of prior knowledge; we noted this was not invariant under change of parameter.

Because of difficulties in encoding a vague degree of belief in a prior, one often attempts to derive the prior from formal rules, e.g., to satisfy certain invariance principles or to provide maximum information gain for a certain set of measurements.

Often called “objective priors”

Form basis of Objective Bayesian Statistics

The priors do not reflect a degree of belief (but might represent possible extreme cases).

In Objective Bayesian analysis, can use the intervals in a frequentist way, i.e., regard Bayes’ theorem as a recipe to produce an interval with a given coverage probability.

# Priors from formal rules (cont.)

For a review of priors obtained by formal rules see, e.g.,

Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996).

Formal priors have not been widely used in Particle Physics, but there has been interest in this direction, especially the reference priors of Bernardo and Berger; see e.g.

L. Demortier, S. Jain and H. Prosper, *Reference priors for high energy physics*, Phys. Rev. D 82 (2010) 034002, arXiv:1002.1111.

D. Casadei, *Reference analysis of the signal + background model in counting experiments*, JINST 7 (2012) 01012; arXiv:1108.4270.

# Jeffreys prior

According to *Jeffreys' rule*, take prior according to

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}$$

where

$$I_{ij}(\boldsymbol{\theta}) = -E \left[ \frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right] = - \int \frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} P(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}$$

is the Fisher information matrix.

One can show that this leads to inference that is invariant under a transformation of parameters in the following sense:

Start with the Jeffreys prior for  $\theta$ :  $\pi_\theta(\theta) \sim \sqrt{(\det I(\theta))}$

Use it in Bayes' theorem to find:

$$P(\theta|\mathbf{x}) \propto P(\mathbf{x}|\theta)\pi_\theta(\theta)$$

## Jeffreys prior (2)

Now consider a function  $\eta(\theta)$ . The posterior for  $\eta$  is

$$P(\eta|\mathbf{x}) = P(\theta|\mathbf{x}) \left| \frac{d\theta}{d\eta} \right|$$

Alternatively, start with  $\eta$  and use its Jeffreys' prior:

$$\pi_{\eta}(\eta) \propto \sqrt{\det I(\eta)}$$

Use this in Bayes' theorem:  $P(\eta|\mathbf{x}) \propto P(\mathbf{x}|\eta)\pi_{\eta}(\eta)$

One can show that Jeffreys' prior results in the same  $P(\eta|\mathbf{x})$  in both cases. For details (single-parameter case) see:

<http://www.pp.rhul.ac.uk/~cowan/stat/notes/JeffreysInvariance.pdf>

# Jeffreys prior for Poisson mean

Suppose  $n \sim \text{Poisson}(\mu)$ . To find the Jeffreys' prior for  $\mu$ ,

$$P(n|\mu) = \frac{\mu^n}{n!} e^{-\mu} \qquad \frac{\partial^2 \ln P(n|\mu)}{\partial \mu^2} = -\frac{n}{\mu^2}$$

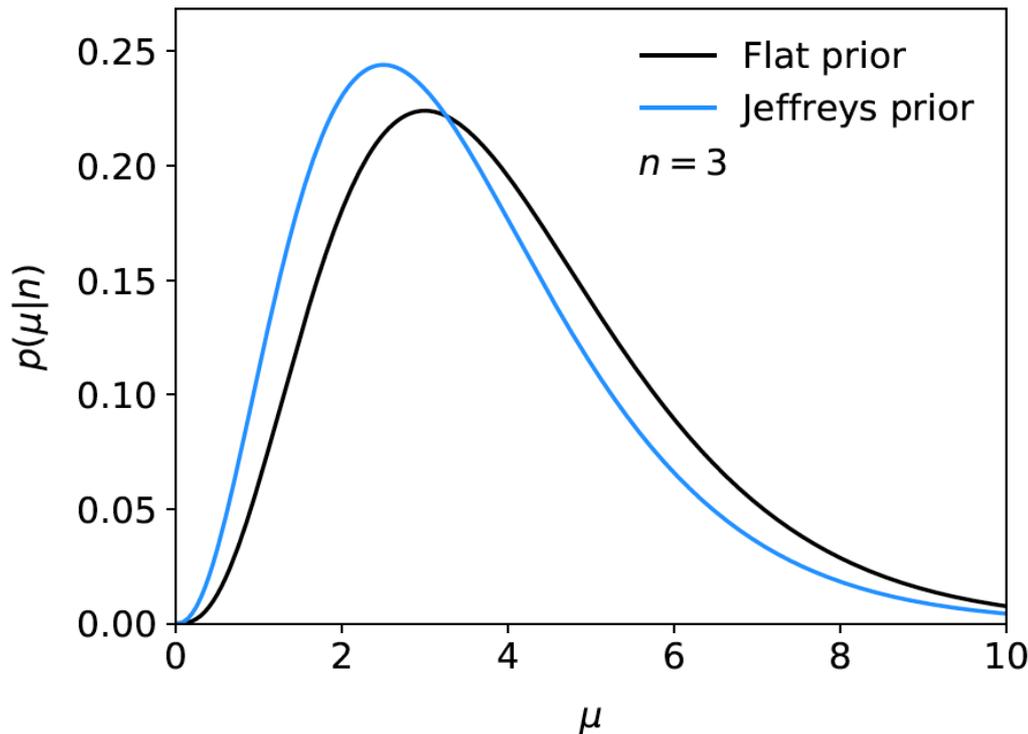
$$I = -E \left[ \frac{\partial^2 \ln P(n|\mu)}{\partial \mu^2} \right] = \frac{E[n]}{\mu^2} = \frac{1}{\mu}$$

$$\pi(\mu) \propto \sqrt{I(\mu)} = \frac{1}{\sqrt{\mu}}$$

So e.g. for  $\mu = s + b$ , this means the prior  $\pi(s) \sim 1/\sqrt{(s + b)}$ , which depends on  $b$ . But this is not designed as a degree of belief about  $s$ .

# Posterior pdf for Poisson mean

From Bayes' theorem,  $p(\mu|n) \propto \mu^n e^{-\mu} \pi(\mu)$



Flat,  $\pi(\mu) = \text{const.}$

$$p(\mu|n) = \frac{\mu^n e^{-\mu}}{\Gamma(n+1)}$$

Jeffreys,  $\pi(\mu) \sim 1/\sqrt{\mu}$

$$p(\mu|n) = \frac{\mu^{n-\frac{1}{2}} e^{-\mu}}{\Gamma(n+\frac{1}{2})}$$

In both cases, posterior is special case of gamma distribution.

# Upper limit for Poisson mean

To find upper limit at  $CL = 1 - \alpha$ , solve

$$1 - \alpha = \int_0^{\mu_{\text{up}}} p(\mu|n) d\mu$$

Jeffreys prior:  $\mu_{\text{up}} = P^{-1}(n + \frac{1}{2}, 1 - \alpha) = 7.03$

Flat prior:  $\mu_{\text{up}} = P^{-1}(n + 1, 1 - \alpha) = 7.75$

$n=3,$   
 $CL=0.95$

where  $P^{-1}$  is the inverse of the normalized lower incomplete gamma function (see `scipy.special`)

$$P(a, \mu_{\text{up}}) = \frac{1}{\Gamma(a)} \int_0^{\mu_{\text{up}}} \mu^{a-1} e^{-\mu} d\mu$$

# Profiling

The  $\ln L = \ln L_{\max} - 1/2$  contour in the  $(\theta_0, \theta_1)$  plane is a confidence region at CL = 39.3%.

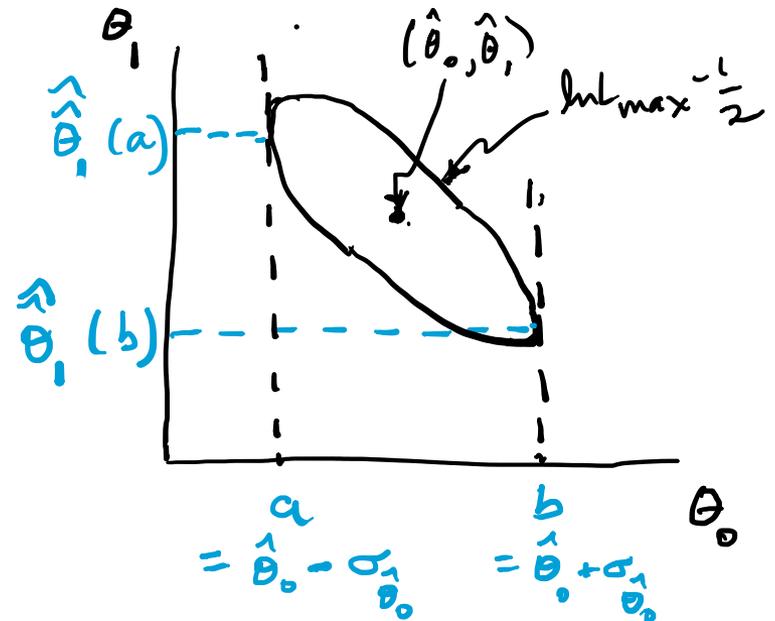
Furthermore if one wants to know only about, say,  $\theta_0$ , then the interval in  $\theta_0$  corresponding to  $\ln L = \ln L_{\max} - 1/2$  is a confidence interval at CL = 68.3% (i.e.,  $\pm 1$  std. dev.).

I.e., form the interval for  $\theta_0$  using

$$\ln L(\theta_0, \hat{\theta}_1(\theta_0)) = \ln L_{\max} - 1/2$$

where  $\theta_1$  is replaced by its “profiled” value

$$\hat{\theta}_1(\theta_0) = \operatorname{argmax}_{\theta_1} L(\theta_0, \theta_1)$$



# Profile Likelihood

Suppose we have a likelihood  $L(\boldsymbol{\mu}, \boldsymbol{\theta}) = P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\theta})$  with  $N$  parameters of interest  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$  and  $M$  nuisance parameters  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M)$ . The “profiled” (or “constrained”) values of  $\boldsymbol{\theta}$  are:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\mu}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} L(\boldsymbol{\mu}, \boldsymbol{\theta})$$

and the profile likelihood is:  $L_p(\boldsymbol{\mu}) = L(\boldsymbol{\mu}, \hat{\boldsymbol{\theta}})$

The profile likelihood depends only on the parameters of interest; the nuisance parameters are replaced by their profiled values.

The profile likelihood can be used to obtain confidence intervals/regions for the parameters of interest in the same way as one would for all of the parameters from the full likelihood.

# Profile Likelihood Ratio – Wilks theorem

Goal is to test/reject regions of  $\mu$  space (param. of interest).

Rejecting a point  $\mu$  should mean  $p_\mu \leq \alpha$  for all possible values of the nuisance parameters  $\theta$ .

Test  $\mu$  using the “profile likelihood ratio”: 
$$\lambda(\mu) = \frac{L(\mu, \hat{\theta})}{L(\hat{\mu}, \hat{\theta})}$$

Let  $t_\mu = -2 \ln \lambda(\mu)$ . Wilks' theorem says in large-sample limit:

$$t_\mu \sim \text{chi-square}(N)$$

where the number of degrees of freedom is the number of parameters of interest (components of  $\mu$ ). So  $p$ -value for  $\mu$  is

$$p_\mu = \int_{t_{\mu, \text{obs}}}^{\infty} f(t_\mu | \mu, \theta) dt_\mu = 1 - F_{\chi_N^2}(t_{\mu, \text{obs}})$$

## Profile Likelihood Ratio – Wilks theorem (2)

If we have a large enough data sample to justify use of the asymptotic chi-square pdf, then if  $\mu$  is rejected, it is rejected for any values of the nuisance parameters.

The recipe to get confidence regions/intervals for the parameters of interest at  $CL = 1 - \alpha$  is thus the same as before, simply use the profile likelihood:

$$\ln L_p(\mu) = \ln L_{\max} - \frac{1}{2} F_{\chi_N^2}^{-1}(1 - \alpha)$$

where the number of degrees of freedom  $N$  for the chi-square quantile is equal to the number of parameters of interest.

If the large-sample limit is not justified, then use e.g. Monte Carlo to get distribution of  $t_\mu$ .

# Profile construction (“hybrid resampling”)

K. Cranmer, PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics, 2008.  
oai:cds.cern.ch:1021125, cdsweb.cern.ch/record/1099969.

Approximate procedure is to reject  $\theta$  if  $p_\theta \leq \alpha$  where the  $p$ -value is computed assuming the value of the nuisance parameter that best fits the data for the specified  $\theta$  :

$$\hat{\hat{\nu}}(\theta)$$

“double hat” notation means value of parameter that maximizes likelihood for the given  $\theta$ .

The resulting confidence interval will have the correct coverage for the points  $(\theta, \hat{\hat{\nu}}(\theta))$  .

Elsewhere it may under- or overcover, but this is usually as good as we can do (check with MC if crucial or small sample problem).

# “Hybrid frequentist-Bayesian” method

Alternatively, suppose uncertainty in  $\nu$  is characterized by a Bayesian prior  $\pi(\nu)$ .

Can use the marginal likelihood to model the data:

$$L_{\text{m}}(x|\theta) = \int L(x|\theta, \nu)\pi(\nu) d\nu$$

This does not represent what the data distribution would be if we “really” repeated the experiment, since then  $\nu$  would not change.

But the procedure has the desired effect. The marginal likelihood effectively builds the uncertainty due to  $\nu$  into the model.

Use this now to compute (frequentist)  $p$ -values  $\rightarrow$  the model being tested is in effect a weighted average of models.