Advanced Statistical Methods

Lecture 2

Hypotheses and parameters

A) Testing of *simple* hypotheses

- p-values, significance, power

Null

OR

Alternative

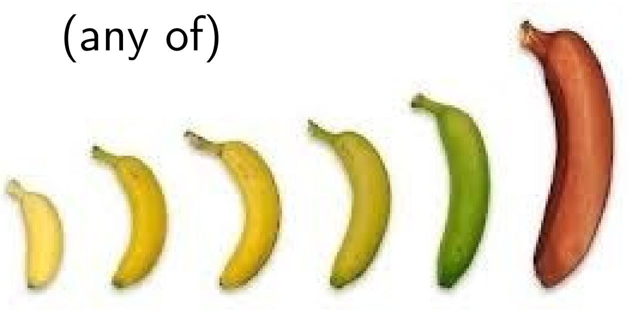
B) General Parameter estimation

- bias, variance, maximum likelihood

C) Testing of *composite* (and nested) hypotheses

- confidence regions
- Wicks' theorem

(any of)



Null

OR

(any of)

Alternative

Testing of simple hypothesis

Null

OR

Alternative

Statistical significance

If the outcome of a measurement under a given null hypothesis *H* is sufficiently unlikely, that hypothesis can be rejected.

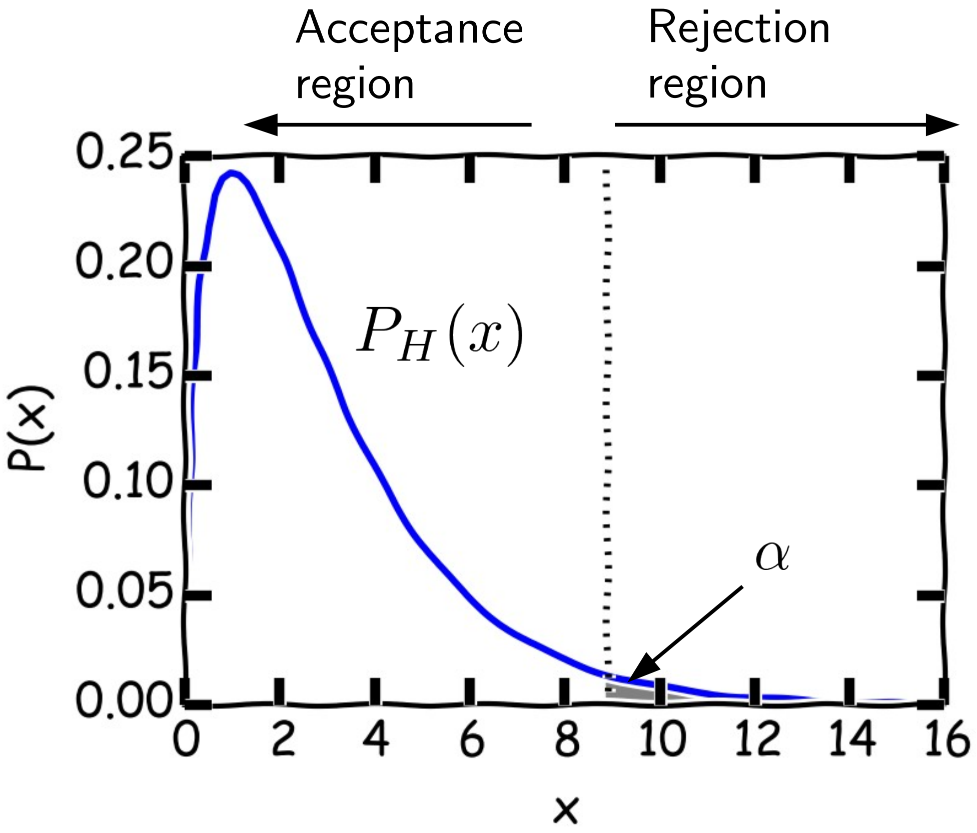
The **statistical significance** of rejection is given by the *p-value*. It gives the probability for the given or a more extreme observation to occur provided the null hypothesis is true.

$$\int_{x_{\mathrm{obs}}}^{\infty} P_H(x)dx = p$$

Here, *x* denotes a *test statistic*.

One says that the hypothesis is rejected when a certain predefined threshold or *alpha level* is reached.

$$p \leq \alpha$$



Notes:

- *p* follows by construction a uniform distribution between 0 and 1
- It is often equivalently expressed in units of Gaussian *sigma*

$$\int_{s}^{\infty} N(x|0,1)dx = p$$

- Typical values for *α* in particle physics: 3.0σ ($p = 1.35\text{x}10^{-3}$) or 5.0σ ($p = 2.87\text{x}10^{-7}$)
  "hint" "discovery"

Statistical power

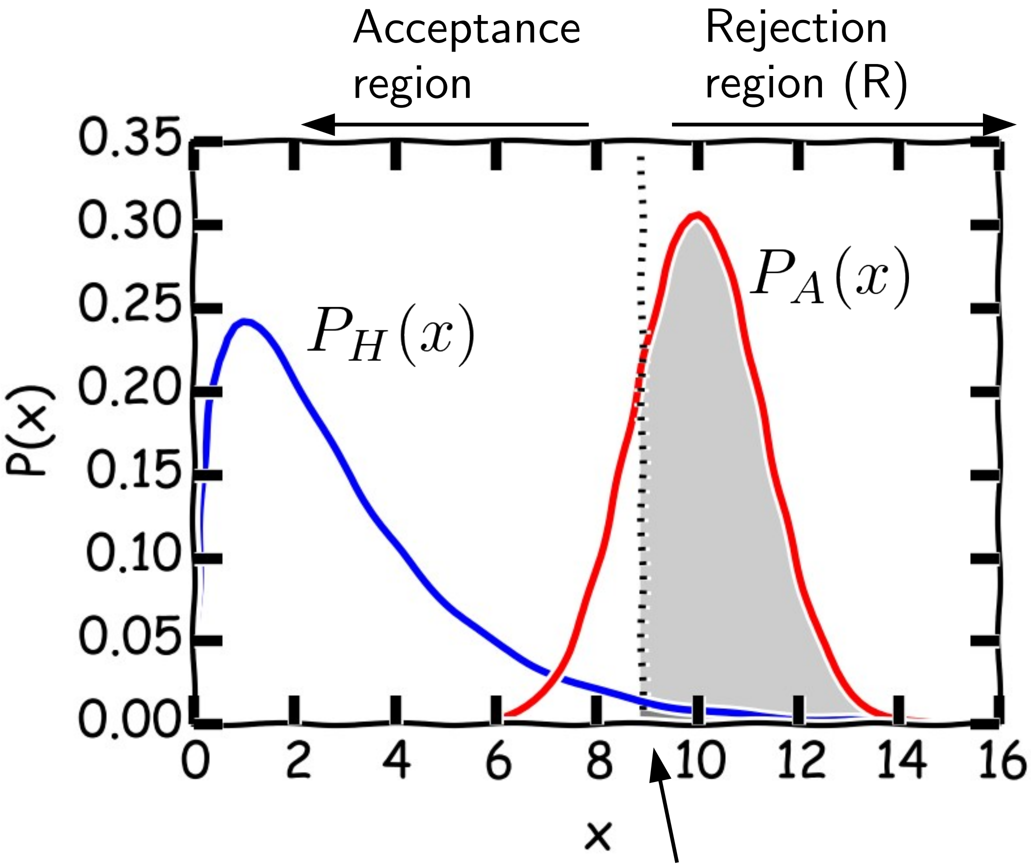
Statistical significance of the rejection is given by

$$\int_R P_H(x)dx = \alpha$$

“The observation has a *p*-value smaller than α.”

Statistical power of the test is given by:

$$\int_R P_A(x)dx = 1 - \beta$$



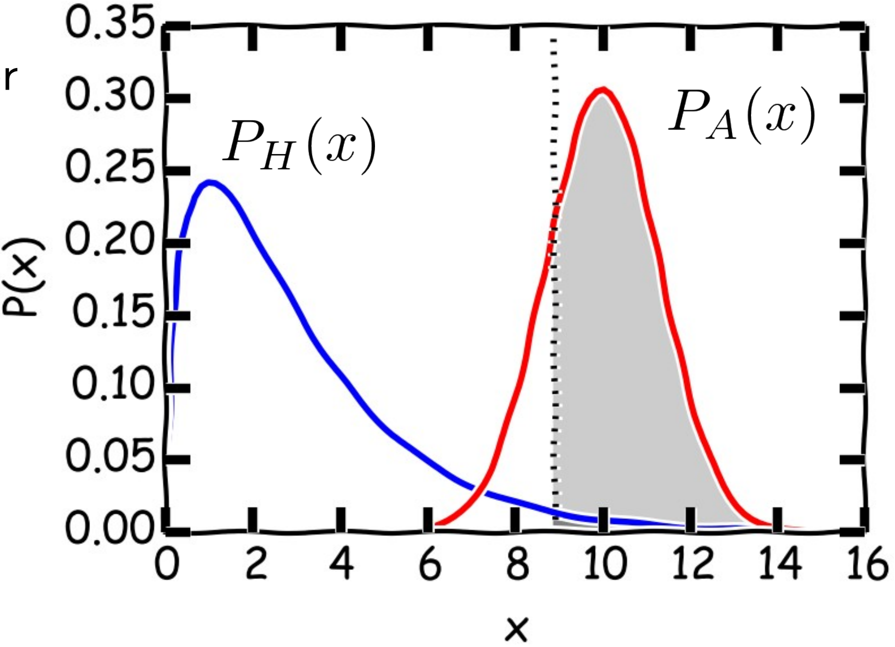
Note: The rejection region is here simply defined by a threshold for *x*.

A good test minimizes the chance for the following failure modes:

- Type I error: *Reject* a *true* null hypothesis (with probability α)
- Type II error: *Accept* a *false* null hypothesis (with probability β)

Maximizing statistical power

Q: Given a desired significance-level α, what is the rejection region that maximizes the statistical power of a test?



Neyman-Pearson Lemma

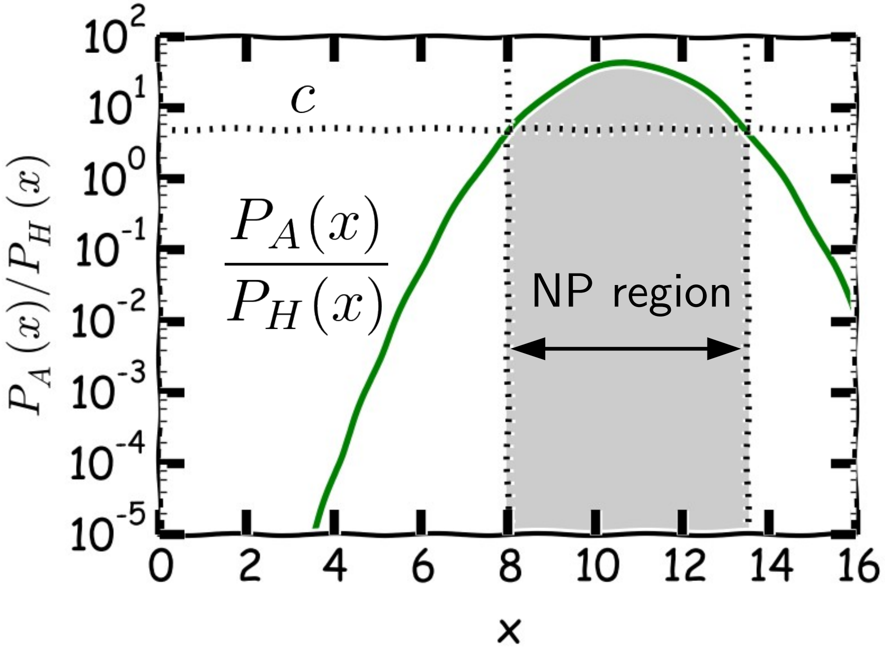
- The rejection region that maximizes the statistical power is given by all *x* that have a large enough ***likelihood ratio***:

$$\frac{P_A(x)}{P_H(x)} > c$$

- Here, *c* is fixed such that the test has the desired significance

$$\int P_H(x)\theta_H\left(\frac{P_A(x)}{P_H(x)} - c\right) dx = \alpha$$

$$(\theta_H : \text{Heavisidestep} - \text{function})$$



Notes:

- *The likelihood ratio is omnipresent in all of statistics.*
- *The rejection region is defined by threshold on likelihood ratio → it can have complex boundaries.*

Generalization to many observables

The Neyman Pearson lemma can be easily applied to cases with many observables. One example is a **large number of samples** of the same observable:

$$P_H(\vec{x}) = \prod_i P_H(x_i) \qquad\qquad P_A(\vec{x}) = \prod_i P_A(x_i)$$

It is convenient to define the "log-likelihood ratio":

$$T \equiv -2\ln\frac{P_H(\vec{x})}{P_A(\vec{x})} = -2\sum_i \ln\frac{P_H(x_i)}{P_A(x_i)}$$

The threshold for rejecting the null, *c*, is obtained from

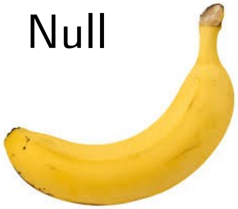
$$\int_c^\infty P(T|H)dT = \alpha$$

Notes:

- In the large number of samples limit, the CLT ensures that T follows a normal distribution
- In complicated cases, *P(T|H)* is best estimated by a MC simulation

Goodness-of-fit: Pearson's chi-squared test

Null



OR

Something completely different

Pearson's chi-squared test

- Test statistic is defined as

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{\Delta E_i^2}$$

O_i: observed value in bin *i*
E_i: expected value in bin *i*
ΔE_i: Standard deviation bin i

- If data is drawn from the null hypothesis with the indicated errors

$$O_i = E_i \pm \Delta E_i$$

the test statistic follows a chi-squared distribution with *k*=*N* degrees of freedom.

Goodness-of-fit: The K-S test

The Kolmogorov-Smirnov Test

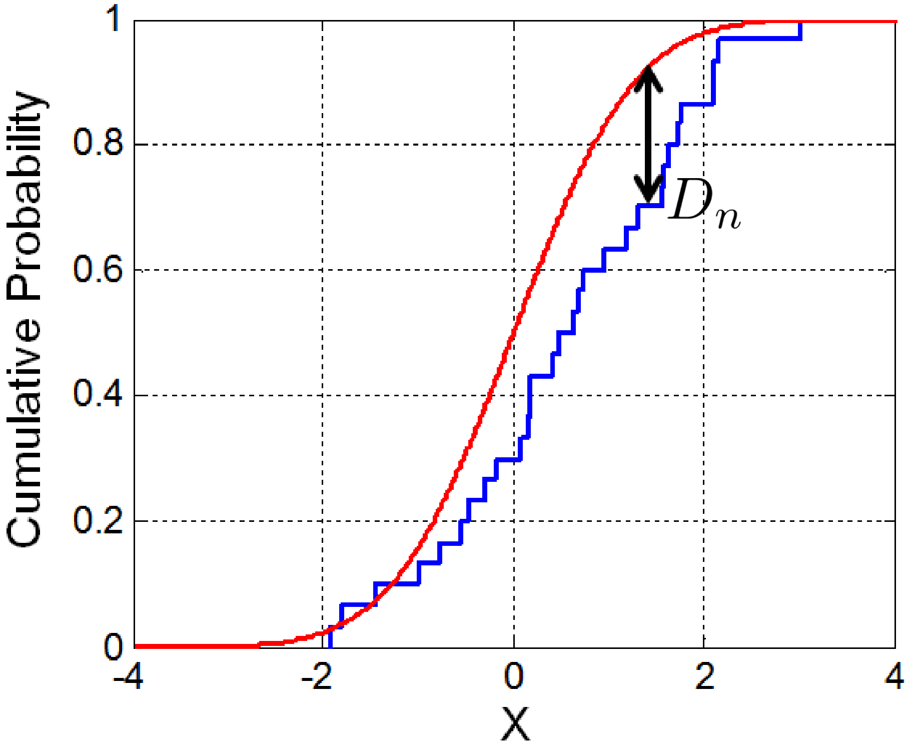
- First, construct *empirical distribution function*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\theta_H(x - x_i)$$

- Second, calculate maximal distance between expected distribution and constructed CDF

$$D_n = \sup_x |F_n(x) - F(x)|$$

- $D_n\sqrt{n}$ follows a Kolmogorov distribution in the large *n* limit. If the value is too large, the null hypothesis can be rejected.



Notes:

- This test is sensitive to *any* deviation from the null hypothesis. Use it with care!
- There is a similar test for comparing two measured distributions instead of a distribution and the expectation.
- See also: Cramer-von Mises test, Anderson-Darling test, Shapiro-Wilk test

General parameter estimation

This is common to both Frequentist and Bayesian approaches!

Basic quantities

Situation

- We have a model that describes the data, but the precise model parameters are unknown
- An ***estimator*** is a map from the experimental data onto the model parameter space. It is a random variable.

Model parameters: $\vec{\theta}$

Estimator: $\vec{t} = \vec{t}(data)$

Relevant properties:

Bias: $\vec{\beta} = \langle \vec{t} \rangle - \vec{\theta}$

Variance: $var(t_k) = \langle t_k^2 \rangle - \langle t_k \rangle^2$

Mean squared error: $mse(t_k) = \langle (t_k - \theta_k)^2 \rangle$

For an biased estimator, the MSE is larger than the variance!

$$mse(t_k) = \sigma_k^2 + \beta_k^2$$

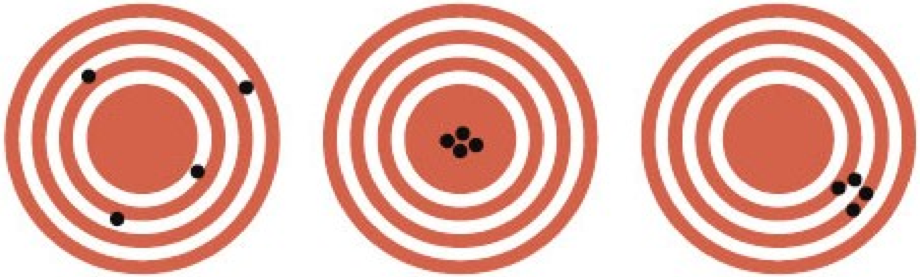
"Unbiased estimator"

$$\beta = 0$$

"Minimum variance estimator"

$$var(t_k) \quad \text{minimal}$$

Good accuracy



Good precision

"Consistent" estimator: unbiased in the limit of a large number of data points

"Efficient" estimator: minimum variance in the limit of a large number of data points

Estimating directly observed quantities

In the case of a large number of measurements, there are obvious estimators for the mean and variance of the *measured parameter:*

Estimator for the **mean** of the underlying distribution:

$$\hat{\mu} = \frac{1}{N} \sum x_i$$

Note: this estimator is per definition unbiased, but it does not automatically have minimum variance.

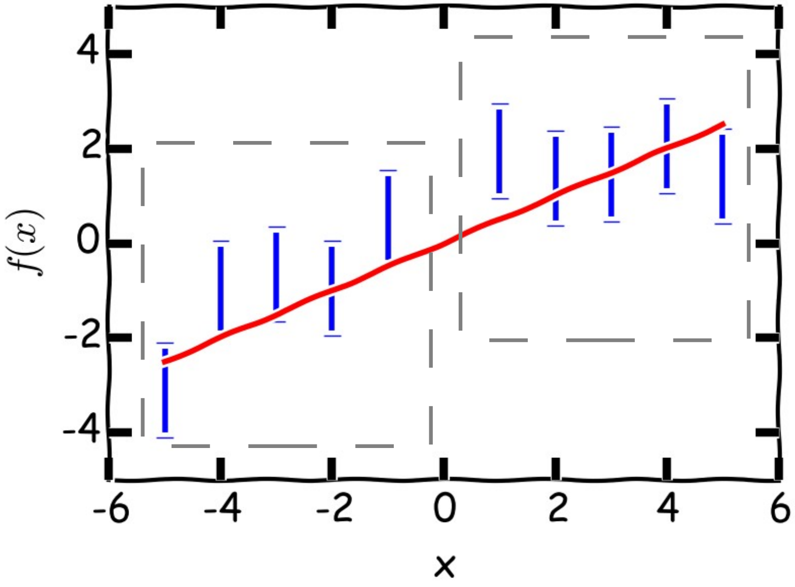
Estimator for **variance**:

$$\widehat{var(x)} = \frac{1}{N-1} \sum (x_i - \hat{\mu})^2$$

Correction factor (since we use data to estimate the mean)

An example for a sub-optimal estimator

Model: A linear relation with unknown slope.



Mean value:

$$\langle y_k \rangle = \alpha x_k$$

Variance:

$$var(y_k) = \sigma^2$$

A simple estimator

- Average all points at x>0 and at x<0 independently.
- Calculate slope from these two resulting average points.

$$\hat{\alpha} = \frac{\langle y \rangle_{II} - \langle y \rangle_I}{\langle x \rangle_{II} - \langle x \rangle_I}$$

Variance of the estimator

$$var(\hat{\alpha}) = \frac{4\sigma^2}{N(\langle x \rangle_{II} - \langle x \rangle_I)^2}$$

For a specific configuration * this yields: $var(\hat{\alpha}) = \frac{\sigma^2}{90}$

* $x_1, \ldots, x_{10} = -5, -4, -3, -2, -1, 1, 2, 3, 4, 5$

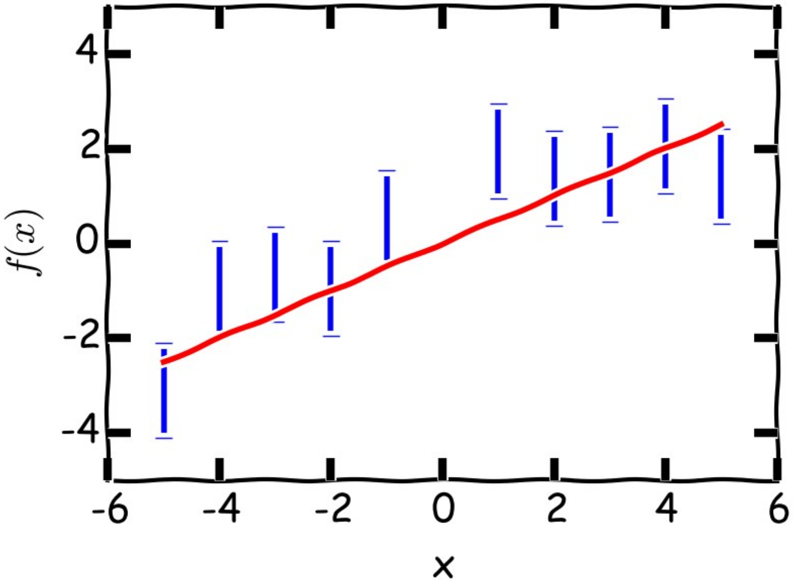
The better estimator: chi-squared

A **more efficient estimator** is obtained by least square fitting:

$$\chi^2 = \sum (y_i - \alpha x_i)^2$$

Minimizing requires:

$$\left.\frac{d\chi^2}{d\alpha}\right|_{\alpha=\hat{\alpha}} = 0$$



- The estimator can be shown to be given by the analytic expression

$$\hat{\alpha} = \frac{\langle xy \rangle}{\langle x^2 \rangle}$$

- The variance reads $$var(\hat{\alpha}) = \frac{\sigma^2}{N\langle x^2 \rangle}$$

For the previous example* this becomes: $$var(\hat{\alpha}) = \frac{\sigma^2}{110}$$

* $x_1, \ldots, x_{10} = -5, -4, -3, -2, -1, 1, 2, 3, 4, 5$

Maximum likelihood estimator

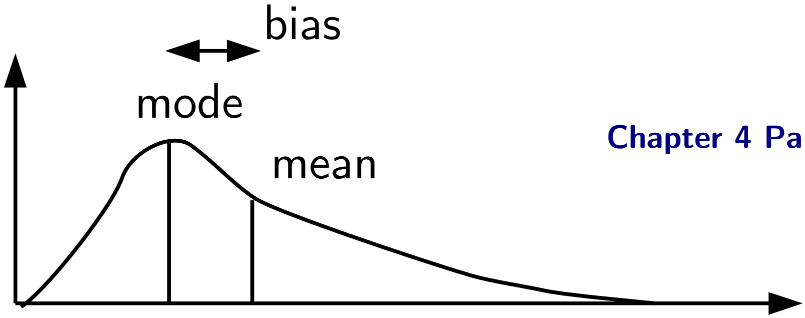
The MLE maximizes the likelihood function for a give set of data

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta|x)$$

Edgeworth

Properties of the MLE:

- The MLE is in general **biased**



Chapter 4 Parameter Estimation

- Thanks to the CTL, it is however in most cases *consistent*
- An *unbiased* MLE has <u>minimum</u> variance

$$var(\hat{\theta}) = -\frac{1}{\left\langle \frac{d^2 \ln \mathcal{L}}{d\theta^2} \right\rangle}$$

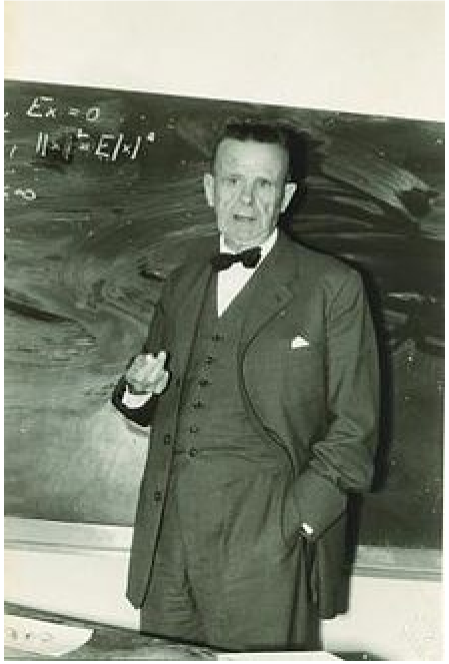
- A consistent MLE is also *efficient*

The Optimum: The Cramér-Rao bound

Cramer-Rao bound:
For *any* estimator, there exists *a lower bound on the variance* that is given by the inverse of the "Fisher information" (for proof see e.g. Barlow):

$$var(\hat{\theta}) = -\frac{1}{\left\langle \frac{d^2 \ln \mathcal{L}}{d\theta^2} \right\rangle} \equiv \frac{1}{\mathcal{I}(\theta)}$$

Definitions: An estimator that saturates this bound is called *minimum variance estimator* (MVE). If the CRB is only saturated in the limit of a large number of measurements, it is called *efficient estimator.*



Harald Cramér

In case of a biased estimator, the lower limit is

$$var(\hat{\theta}) \geq \frac{(1 + d\beta/d\theta)^2}{\mathcal{I}(\theta)}$$

This can be both *larger* and *smaller* than the unbiased bound.

Quantifying information gain

The ***score*** of a likelihood function parametrizes the sensitivity towards parameter change.

$$s(\theta|x) = \frac{\partial}{\partial\theta} \ln \mathcal{L}(\theta|x)$$

- The first moment of the score is zero $\langle s(\theta|x) \rangle = 0$
- The second moment is the ***Fisher information*** that was mentioned above

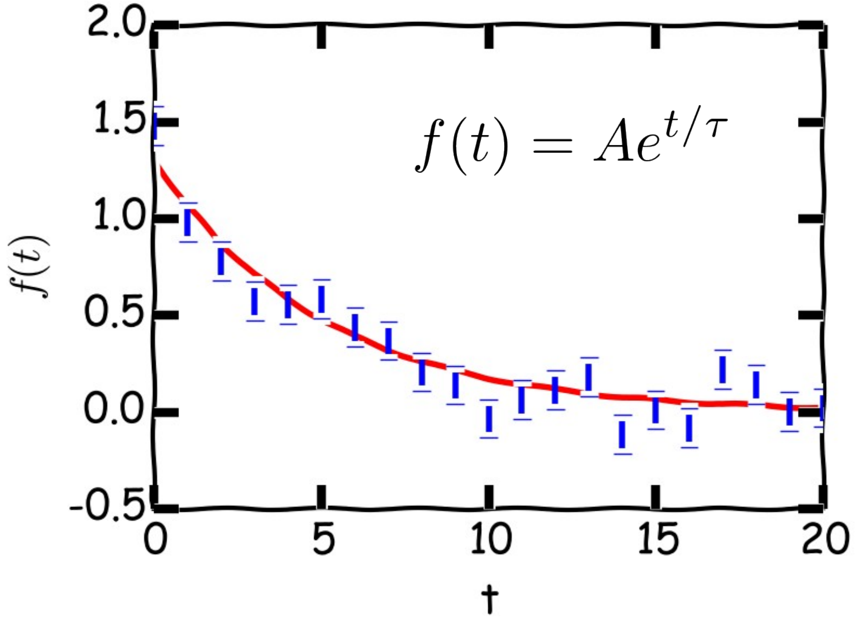
$$\mathcal{I}(\theta) = \left\langle \left(\frac{\partial}{\partial\theta} \ln \mathcal{L}(\theta|x) \right)^2 \right\rangle = - \left\langle \frac{\partial^2}{\partial\theta^2} \ln \mathcal{L}(\theta|x) \right\rangle$$

Fisher information

- Parametrizes the information gain from a measurement
- In case of multivariate normal distributions, it corresponds to covariance matrix

Fisher information and experimental design

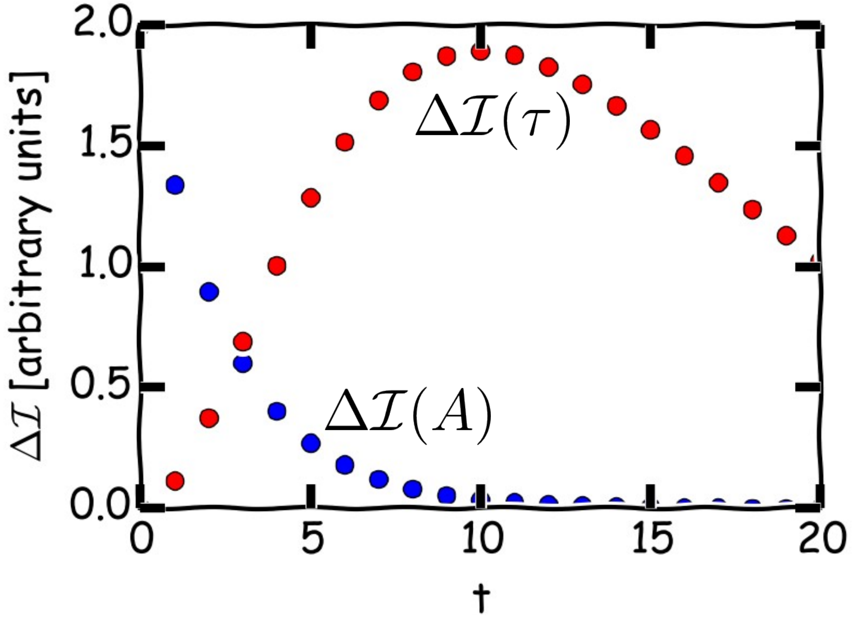
Fisher information quantifies how much information is gained by a given measurement. It is additive in case of multiple measurements, and can guide experimental design.



Scenario:

- Consider some exponential decay with unknown amplitude and lifetime:
- The quantity *f* is measured at discrete time steps with identical errors *Δf*
- The MLE estimator can be obtained from

$$-2\ln\mathcal{L} = \frac{1}{\Delta f^2}\sum_{i\geq 0}(Ae^{-t_i/\tau} - \underbrace{A_0 e^{-t_i/\tau_0}}_{\text{“Asimov data”}})^2$$



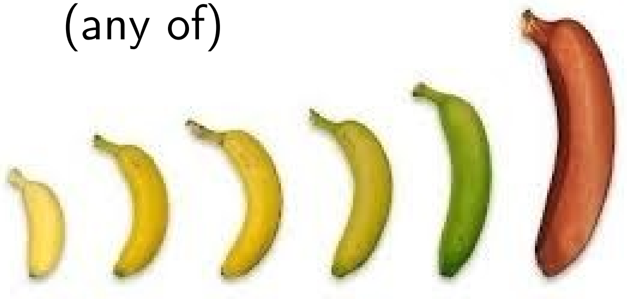
The implied Fisher information for the two free parameters is

$$\mathcal{I}(A) = \frac{1}{\Delta f^2}\sum_{i\geq 0} e^{-2t_i/\tau_0}$$

$$\mathcal{I}(\tau) = \sum_{i\geq 0}\frac{A_0^2 t_i^2}{\Delta f^2\,\tau_0^4}e^{-2t_i/\tau_0}$$

Composite hypotheses & confidence regions

(any of)



Null

OR

(any of)

Alternative

Errors of estimators

Statistical errors

- Thanks to the CLT, errors are often normal distributed, such that estimator and variance are a full description of the situation.

$$\theta = 4.3 \pm 1.2$$

estimate

standard deviation

- Connection to Frequentist statistics: the error range *covers* the true value in 68.3% of the cases
- A function of estimators is itself an estimator, with a total variance that is the weighted sum of the individual variances

$$\sigma_f^2 = \sum_i \left(\frac{\partial f}{\partial x_i} \right)^2 \sigma_{x_i}^2$$

Remark:

Systematic errors

- Systematic errors enter the measurements *as bias*, which is often unknown. This is sometimes written as

$$\theta = 4.3 \pm 1.2_{\text{stat.}} \pm 0.4_{\text{syst.}}$$

- Systematic errors *do not propagate using the above sum rule.*

Exact error bars: Confidence belt

Construction of the confidence belt

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

$$P(x|\theta)$$

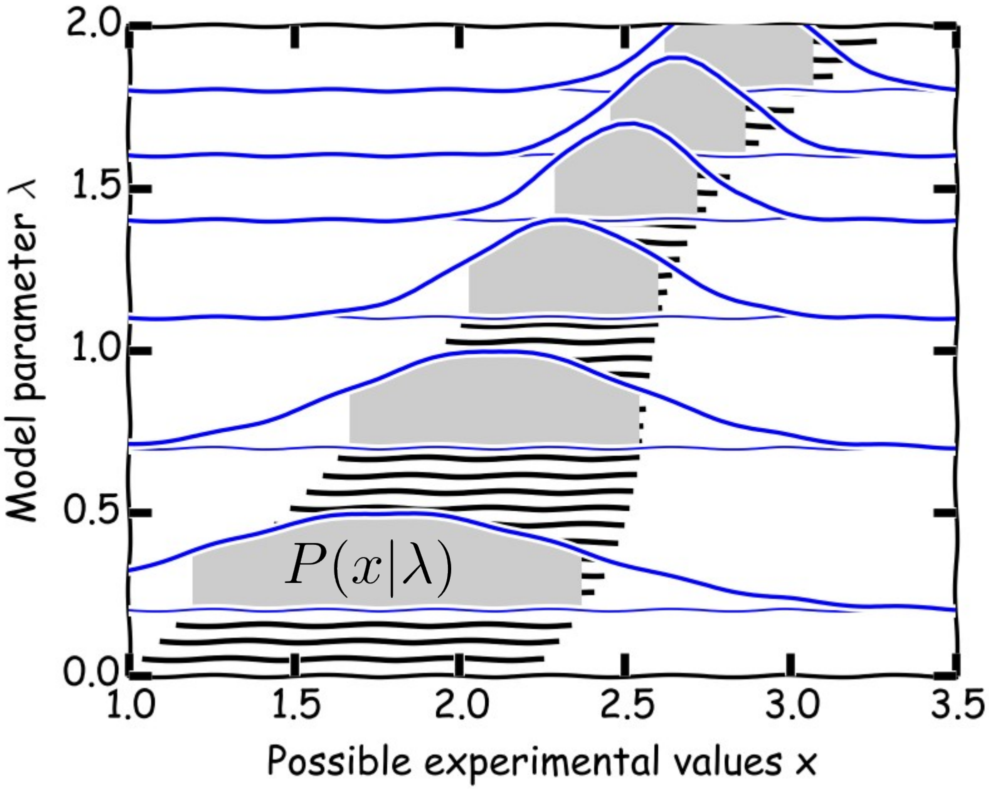
- An *acceptance interval*

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

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

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

- This defines the *confidence belt.*



Exact error bars: Confidence belt

Construction of the confidence belt

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

$$P(x|\theta)$$

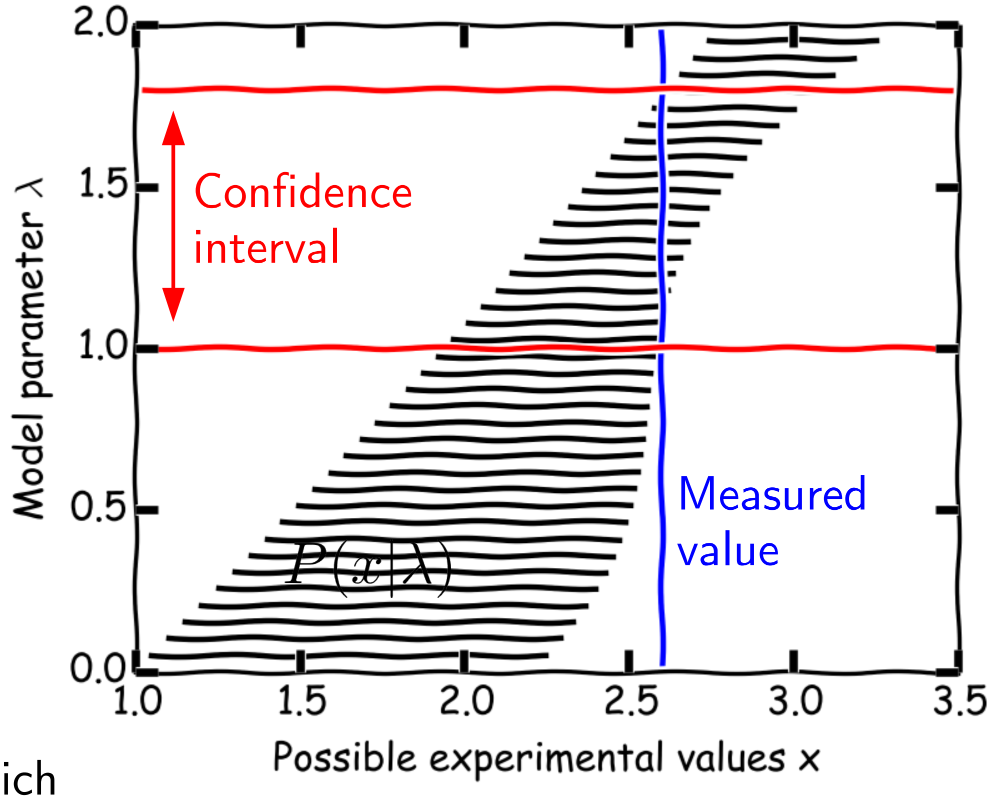
- An *acceptance interval*

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

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

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

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



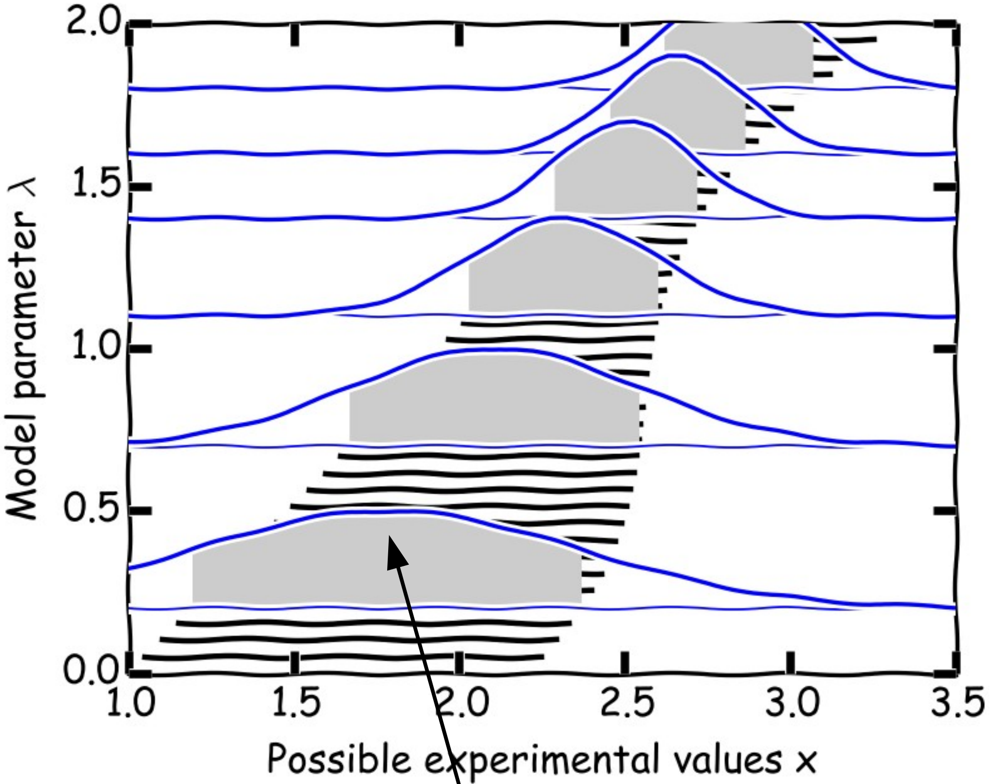
$$I(x_{\text{obs}}) = \{x_0(\theta) \leq x_{\text{obs}} \leq x_1(\theta) \,|\, \theta \in \mathbb{R}\}$$

Note:

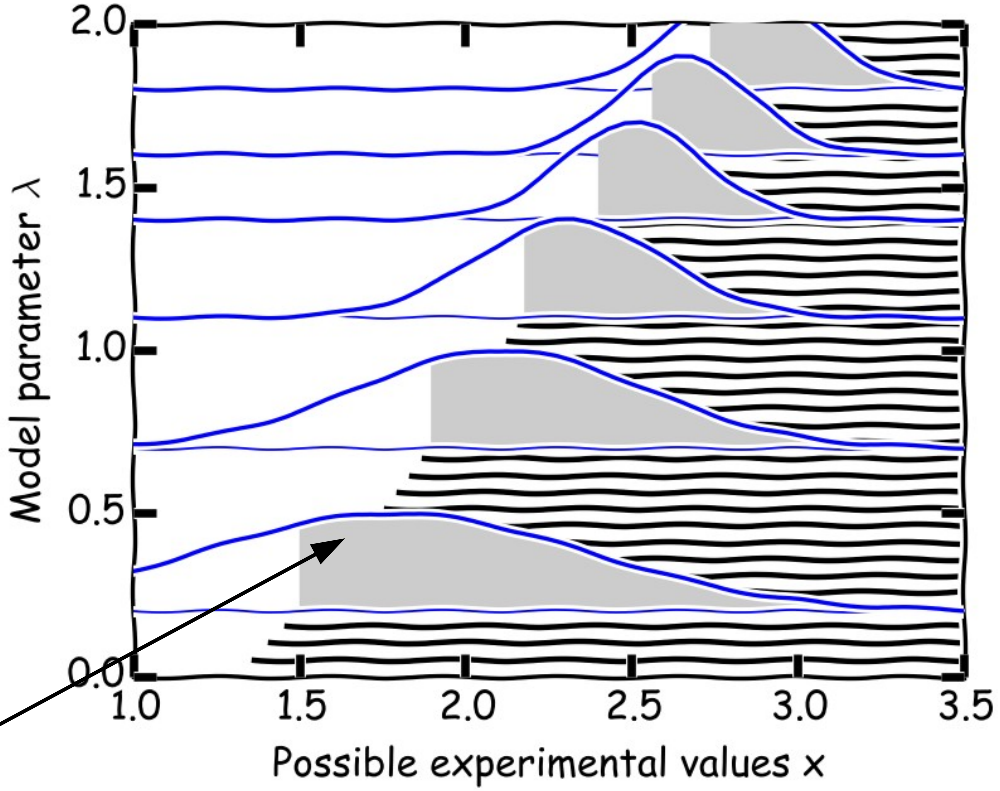
- By construction, and independently of the true value of θ, the confidence region will *cover* the true value in exactly 1-α of the cases.

One sided and two-sided limits

Two-sided confidence region:



One-sided confidence region:



The PDFs for a given model parameter
λ are identical!

Nested composite hypothesis

Previous scenario is special case of **composite nested hypotheses**

- *Null hypothesis*: Model parameter θ is fixed to certain value
- *Alternative hypothesis*: Model describes data, but θ is unconstrained
- *Confidence interval:* All values of θ for which the null hypothesis is *not* rejected

In general

- Alternative hypothesis: Composite model with *n* free parameters

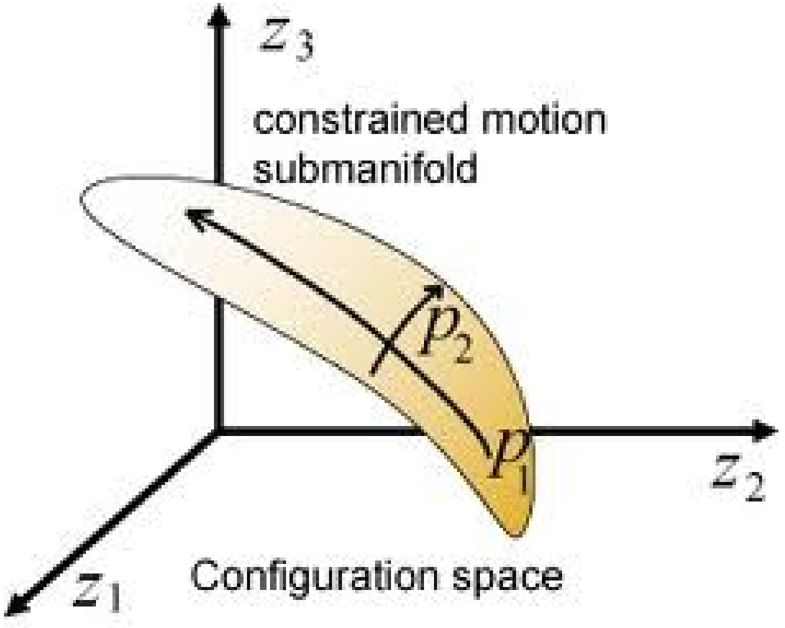
$$P(\vec{x}|\theta_1, \theta_2, \ldots, \theta_n)$$

- Null hypothesis: Composite model with *n-k* free parameters, and *k* constraints

$$f_i(\theta_1, \theta_2, \ldots, \theta_n) = 0, \quad i = 1, 2, \ldots, k$$

Here simply:

$$\begin{array}{ll} \theta_1, \theta_2, \ldots, \theta_k & \text{fixed} \\ \theta_{k+1}, \ldots, \theta_n & \text{free} \end{array}$$



Notes:

- The null hypothesis in two nested composite models typically lives on a submanifold of the parameter space of the alternative model.

Likelihood ratio construction of conf. belt

Confidence regions can be readily constructed by applying the above Neyman Pearson Lemma to the ratio of the *maximum likelihoods* of the composite nested hypotheses.

$$I(x_{\mathrm{obs}}) = \left\{ \underbrace{2 \ln \frac{P(x_{\mathrm{obs}}|\hat{\theta}_1, \dots, \hat{\theta}_n)}{P(x_{\mathrm{obs}}|\theta_1, \dots, \theta_k, \hat{\theta}_{k+1}, \dots, \hat{\theta}_n)}}_{\equiv \Lambda(x_{\mathrm{obs}}, \theta_1, \dots, \theta_k)} < c(\vec{\theta}) \,|\, \vec{\theta} \in \mathbb{R}^k \right\}$$

$$\hat{\theta}_i \,:\, \mathrm{MLE}$$

such that:

$$\int P(\vec{x}|\theta_1, \dots, \theta_n) \theta_{\mathrm{H}} \left(\underbrace{\tilde{c}(\theta_1, \dots, \theta_n)}_{\simeq c(\theta_1, \dots, \theta_k)} - \Lambda(x, \theta_1, \dots, \theta_k) \right) dx = \alpha$$

Problem:

- How to determine value of *c* for different significance level α? This can again be done by a MC, but should be repeated for *all* regions in *n*-dim parameter space
- In general, the threshold *c* will depend not only on the *k* parameters of interest, but also on the remaining *n-k* nuisance parameters.

Remedy:

- If θ_1, ..., θ_k are true values, and in the large-sample limit, assuming certain regularity conditions, **Wilks' theorem** states that: $\Lambda \sim \chi^2_k$

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.:

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