INTERVAL ESTIMATION

- Goal. Confidence interval.

In interval estimation, we want to find the range $\theta_a \le \theta \le \theta_b$ which contains the true value θ_0 with probability β.

Such interval is called "confidence interval" with probability content β.

Typically, we choose $\beta = 68.3\%$ and call it 1 standard deviation error.
However, the 68% interval corresponds to ± 1 STD only for a Gaussian distr.

Given an observation x from a PDF $f(x|\theta)$, the probability content β of the region $[a, b]$ in x-space is:

$$\beta = P(a \le x \le b) = \int_a^b f(x|\theta)\,dx$$

If $f(x|\theta)$ and the parameter θ are known, one can always compute β given a and b.

If the parameter θ is unknown, we need to find another variable

$$z = z(x, \theta)$$

such that the PDF of z is independent of θ: ~~$f(z|\theta) = f($~~

$$f(z|x,\theta) = f(z|x)$$

If this can be found, we can find the optimal range $[\theta_a, \theta_b]$ in θ space such that:

$$P(\theta_a < \theta < \theta_b) = \beta$$

This interval $[\theta_a, \theta_b]$ is called "confidence interval".

A method which yields such an interval $[\theta_a, \theta_b)$ is said to possess the property of coverage.

Notice that: → θ_0 is an unknown constant
→ θ_a and θ_b are functions of x, not of θ.

- Normally distributed data

Let $f(x|\theta)$ be a normal distribution: $f(x|\theta) = \frac{1}{\sqrt{2\pi}\,\sigma^2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

→ If μ and σ are known:

$$\beta = P(a \le x \le b) = \cancel{\int_a^b N(\mu,\sigma)} = \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right) dx$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

with Φ = cumulative of Gaussian

→ If μ is unknown, we can't compute the probability content of the interval $[a,b]$. But we can compute the probability ~~that~~ β that x lies in some interval relative to the unknown mean, say $[\mu+c, \mu+d]$.

Let's define: $y = \frac{x-\mu}{\sigma}$

We have: $$\beta = P(\mu+c \le x \le \mu+d) = \int_{\mu+c}^{\mu+d} N(\mu,\sigma^2)\, dx'$$

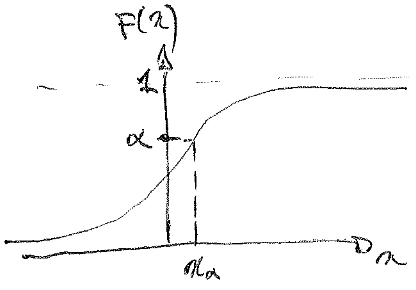
$$= \int_{c/\sigma}^{d/\sigma} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy = \Phi\left(\frac{d}{\sigma}\right) - \Phi\left(\frac{c}{\sigma}\right)$$

We can invert to obtain: $\beta = P(x-d \le \mu \le x-c)$

↳ This is still a probability statement about x. μ is an unknown constant.

- α - point

Given a random variable x with PDF $f(x)$ and CDF $F(x)$, The α point is defined as: $$\int_{-\infty}^{x_\alpha} f(x)dx = F(x_\alpha) = \alpha$$

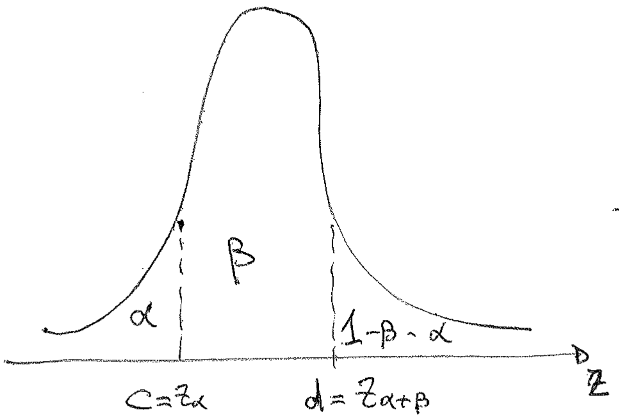


• Confidence intervals for the mean of a Gaussian

For any Gaussian, we can re-define: $z = \frac{x-\mu}{\sigma}$

~~and ol~~ which is a standard-normal variable: $f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$

Estimating the interval $[c, d]$ so that $P(c \leq z \leq d) = \beta$ is equivalent to finding $[z_\alpha, z_{\alpha+\beta}]$:



→ There's infinite choices of the interval!

→ A standard choice is the central interval symmetric around zero, so that:

$$\alpha = \frac{1-\beta}{2}$$

$\beta = \frac{1-\alpha}{2}$	z_α	$z_{\alpha+\beta}$	
0.6827	-1	$+1$	→ ±1σ
0.9	-1.65	$+1.65$	
0.95	-1.96	$+1.96$	
0.9545	-2	$+2$	→ ±2σ
0.9973	-3	$+3$	→ ±3σ

• Confidence intervals for several parameters

Suppose we have an n-dimensional Gaussian:

$$f(\vec{x} | \vec{\theta}) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{|C|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\theta})^T C^{-1} (\vec{x}-\vec{\theta})\right)$$

Each x_i is normal, therefore $Q(\vec{x}, \vec{\theta}) = (\vec{x}-\vec{\theta})^T C^{-1} (\vec{x}-\vec{\theta})$ is a $\chi^2(n)$ distribution, and does not depend on $\vec{\theta}$:

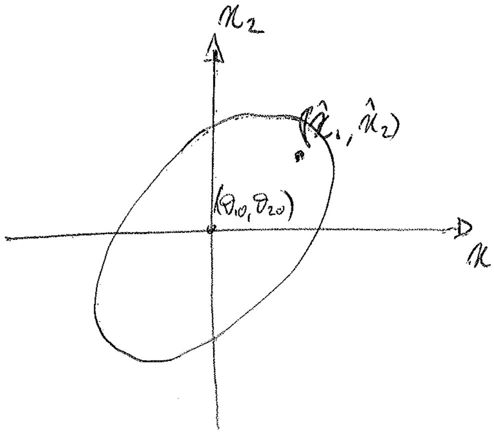
$$Q(\vec{x}, \vec{\theta}) = Q(\vec{x})$$

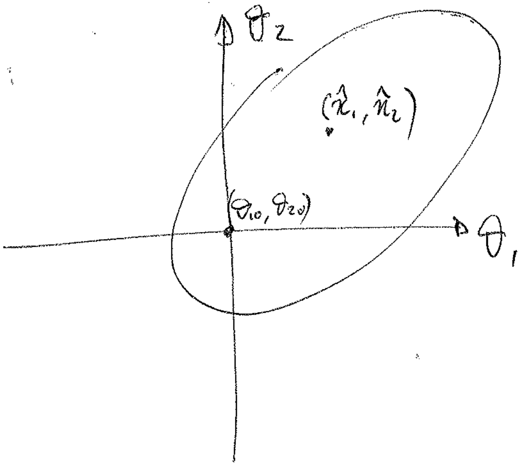
We can write: $P[Q(\vec{x},\vec{\theta}) \leq K_\beta^2] = \beta$

where K_β^2 is the β-point of the $\chi^2(n)$ distribution

This region is defined in $\vec{x}$-space as: $Q(\vec{x},\vec{\theta}) \leq K_\beta^2$

and is a hyperellipsoid of constant probability:





This holds for any dimension.

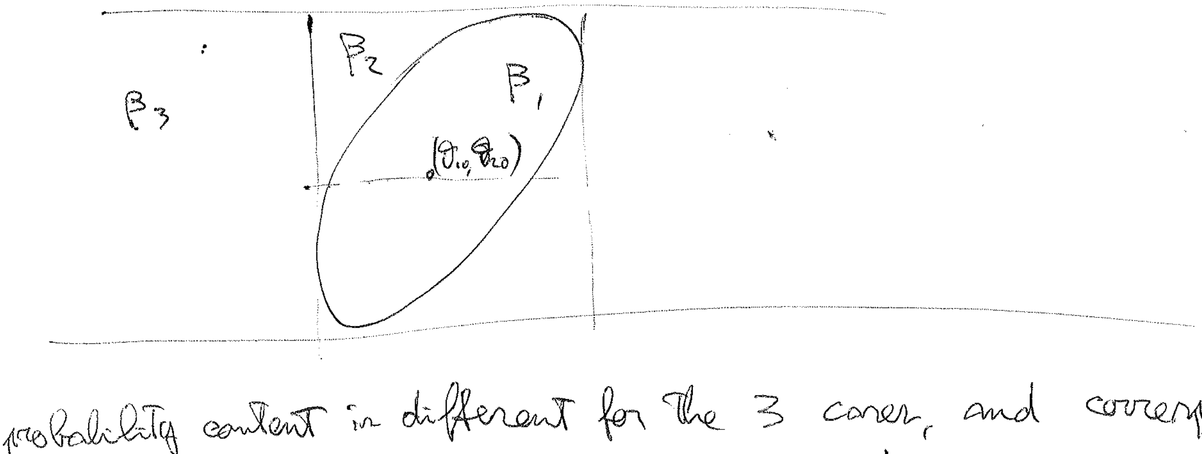
Notice that we have 3 ways to quote an interval for e.g. x_2:

~~a) $P[(\theta_1-\sigma_1 \leq x_1 \leq \theta_1+\sigma_1) \wedge (\theta_2-\sigma_2 \leq x_2 \leq \theta_2+\sigma_2)] = \beta_2$~~

a) $P[Q(\vec{x},\vec{\theta}) \leq \vec{K}_\beta] = \beta_a$ ⇒ ellipse

b) $P[(\theta_1-\sigma_1 \leq x_1 \leq \theta_1+\sigma_1) \wedge (\theta_2-\sigma_2 \leq x_2 \leq \theta_2+\sigma_2)] = \beta_b$ ⇒ square

c) $x_2 = \theta_{20} \pm \sigma_2 \Rightarrow P(\theta_2-\sigma_2 \leq x_2 \leq \theta_2+\sigma_2) = \beta_c$ ⇒ band



The probability content is different for the 3 cases, and corresponds to the 1-dim case just for case ⓒ!

- Second derivative matrix

Assume $\vec{x}$ has an n-dim Gaussian PDF.
One can prove that the n-dim covariance matrix C can be obtained frome the inverse of the 2nd order partial derivative matrix of $-\ln\mathcal{L}$:

$$C^{-1}_{ij} = -\frac{\partial \ln\mathcal{L}(\vec{x}\,|\,\vec{\theta})}{\partial\theta_i\,\partial\theta_j}$$

This covariance matrix gives an n-dim elliptic contour with the correct coverage only if the PDF is exactly Gaussian!
This is the "standard" classical method used to compute uncertainties in common fitting algorithms, e.g. Migrad/Hesse of Minuit/ROOT.

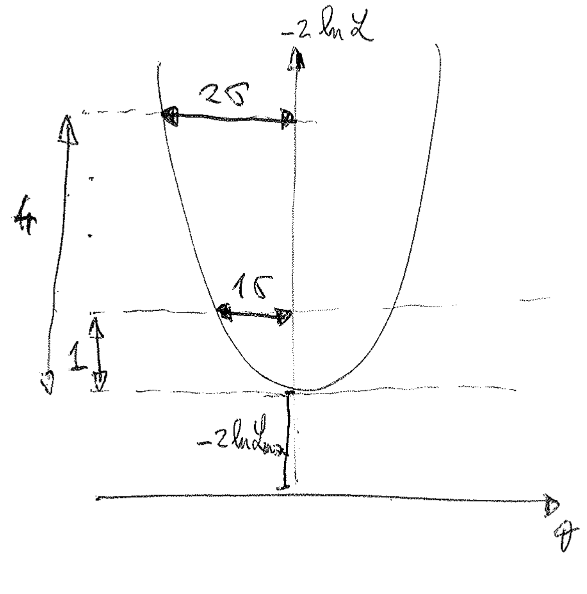
- Log-Likelihood scan

Another common method consists in taking a scan of $-2\ln\mathcal{L}$ around its minimum value, $-2\ln\mathcal{L}_{max}$.

→ For a Gaussian 1-dim distribution:

$$\ln\mathcal{L}(x|\mu) = \ln c - \frac{(\mu-x)^2}{2\sigma^2} \quad \Rightarrow \text{parabola in } \mu$$

$$-2\ln\mathcal{L} = -2\ln\mathcal{L}_{max} + \frac{(\mu-x)^2}{\sigma^2}$$

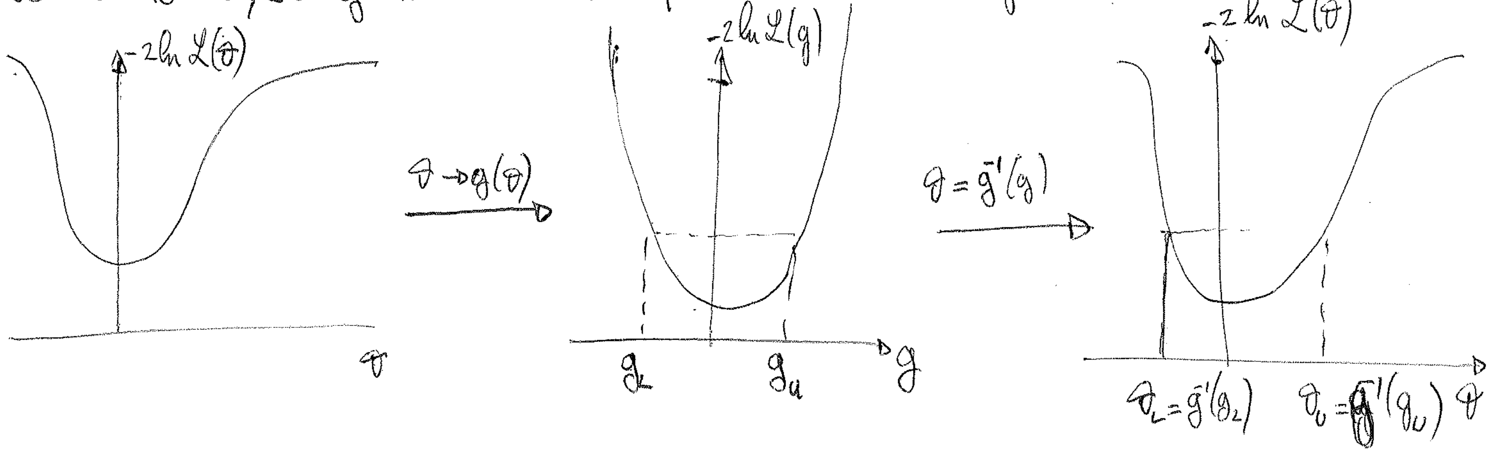


The intercept at $-2\ln\mathcal{L} = -2\ln\mathcal{L}_{max}+1$ provides the $\pm1\sigma$ interval.

The intercept at $+4$ provides the $\pm2\sigma$, and so on

→ For a non-Gaussian PDF, the $\ln \mathcal{L}$ is not parabolic.
However, the maximum $\mathcal{L}$ estimate is invariant, so instead of estimating ϑ, one can estimate any (monotonic) function of ϑ.

So one can find (in principle) a transformation $\vartheta \rightarrow g(\vartheta)$ that makes the $\ln \mathcal{L}$ parabolic, compute the confidence interval on g, and transform back to ϑ, being sure that we preserve the coverage.


$-2\ln\mathcal{L}(\vartheta)$
ϑ
$\vartheta \rightarrow g(\vartheta)$
$-2\ln\mathcal{L}(g)$
g_L
g_u
g
$\vartheta = g^{-1}(g)$
$-2\ln\mathcal{L}(\vartheta)$
$\vartheta_L = g^{-1}(g_L)$
$\vartheta_U = g^{-1}(g_U)$
ϑ

In practice, we can ~~use the same metho~~ find the interval $[\vartheta_L, \vartheta_U]$ *without* actually doing the transformations g and g^{-1}, but simply finding the intercepts at $-2\ln\mathcal{L}_{max} + 1$, as we did for the Gaussian case.

Note however that the uncertainties are now asymmetric.

This is the method used by MINOS in ROOT/Minuit.

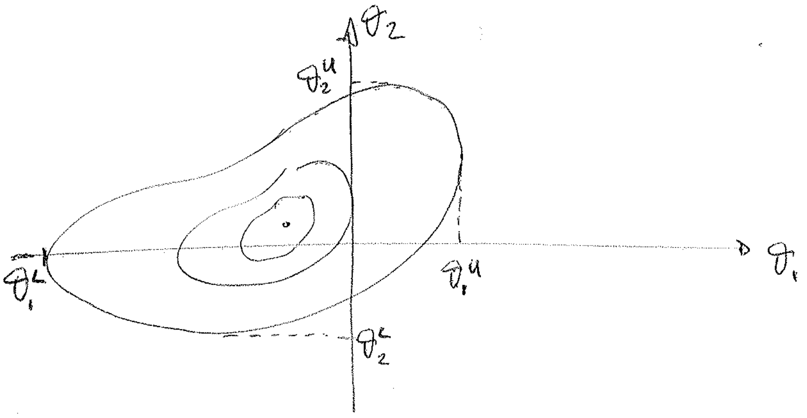
Notice that we have made the *experimental* $\mathcal{L}$ distribution Gaussian in the parameter instead of finding a transformation that would make the *theoretical* distribution Gaussian (i.e. using the true value ϑ_0). To order $1/N$ this is the same, but for small samples it is not exact!

- $\mathcal{L}$ scan in dim > 1 : Profile likelihood

For dim > 1 we have: $\ln \mathcal{L}_{\vec\theta}(\vec x|\vec\theta) = \ln \mathcal{L}_{max} - \frac{1}{2}\chi^2_\beta(\text{[illegible]})$

$$\hookrightarrow -2\ln \mathcal{L}(\vec x|\vec\theta) = -2\ln \mathcal{L}_{max} + \chi^2_\beta(\text{[illegible]})$$

In principle, we can compute the contours:



Notice that: → The inner contour is more nearly elliptical than the outer ones

→ The coverage is improved with respect to the Gaussian approximation

→ This is still an approximation valid for large n.

- Variance of transformed variables, aka Error propagation

Suppose we have a variable $\vec x$ with a given PDF$_x$ and mean $\vec\mu$ an variance $V(\vec x)$

Suppose we want to compute the variance of the transformed variable $\vec y = \vec y(\vec x)$, and that the function can be expanded in Taylor series around $\vec\mu$:

$$\vec y(\vec x) = \vec y(\vec\mu) + \sum_i (x_i - \mu_i)\frac{\partial y}{\partial \mu_i} + \dots$$

The expectation value of y is: $\bar y = y(\vec\mu)$

The variance is:
$$V(y) = E[y - E(y)]^2$$
$$\simeq E\left[\sum_i (x_i-\mu_i)\frac{\partial y}{\partial x_i}\right]^2$$
$$\simeq \sum_i \sum_j \frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j} E[(x_i-\mu_i)(x_j-\mu_j)]^{\cancel{2}}$$
$$\simeq \sum_{i,j} \frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j} \mathrm{Cov}(x_i, x_j)$$

If the variables x_i are independent: $\sigma_y^2 \simeq \sum_i \left(\frac{\partial y}{\partial x_i}\right)^2 \sigma_i^2$ with $\sigma_i^2 = V(x_i)$

Example: linear combination of variables:

$$z = ax + by$$

$$\sigma_z^2 = \left(\frac{\partial z}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial z}{\partial y}\right)^2 \sigma_y^2 + \cancel{2\rho\sigma_x\sigma_y}\, 2\frac{\partial z}{\partial x}\frac{\partial z}{\partial y}\mathrm{cov}(x,y)$$

$$= a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\,\rho\,\sigma_x\sigma_y$$

Example → Product of variables: $\left(\frac{\sigma_{xy}}{xy}\right)^2 = \frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2} + \frac{2\rho\sigma_x\sigma_y}{xy}$

→ Product and ratio of uncorrelated variables:

$$\left(\frac{\sigma_{xy}}{xy}\right)^2 = \left(\frac{\sigma_{x/y}}{x/y}\right)^2 = \frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2}$$

→ Logarithm of variable: $\sigma_{\ln x} = \frac{\sigma_x}{x}$

Caveats: → Valid only if the transformation of variable can be expanded in Taylor series

Alternatively, one can use numerical methods to extract the PDF of the transformed variable, and compute the confidence interval on it.

- Confidence intervals for any PDF: Ordering rules

The approximation of $-2\ln\mathcal{L}$ with a Gaussian or its excursion around its minimum guarantee an exact coverage only for a small set of cases, and in particular for large n.

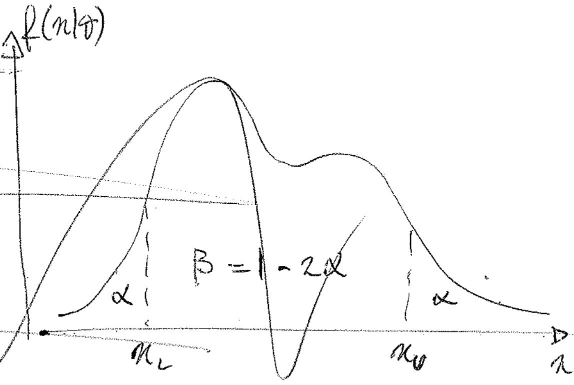
Here we'll see a general approach.

Suppose we have a variable x with PDF $f(x|\theta)$.
In general, f is not symmetric, so we need to decide how to compute an interval corresponding to some predefined probability content β.

→ ~~Central interval~~

$$[x_L(\theta), x_U(\theta)] = [\bar{x}(\theta)-\delta, \bar{x}(\theta)+\delta]$$

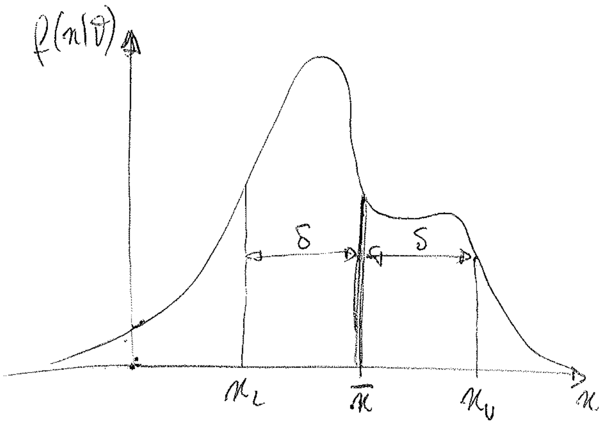
$$\int_{-\infty}^{x} f(x|\theta)\,dx = \int_{x_U}^{+\infty} f(x|\theta)\,dx = \frac{1-\beta}{2}$$



→ Central interval

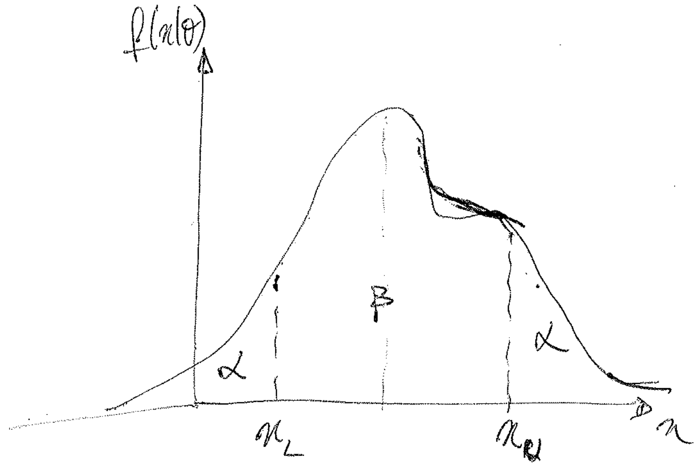
$$[x_L(\theta), x_U(\theta)] = [\bar{x}(\theta)-\delta, \bar{x}(\theta+\delta)]$$

$\bar{x}$ could be the mean or the mode



→ Equal areas

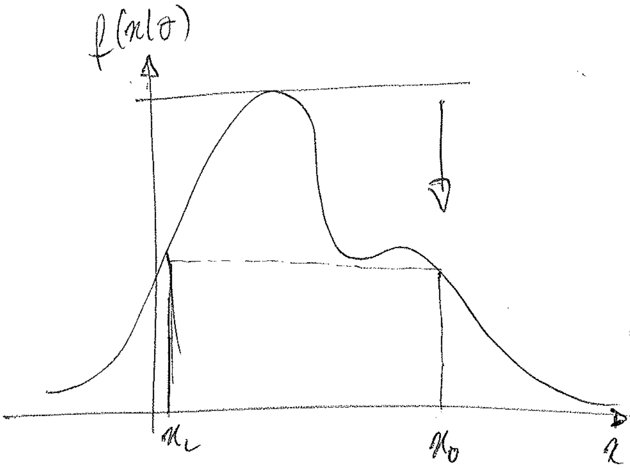
$$\int_{-\infty}^{x_L} f(x|\theta)\,dx = \int_{x_U}^{+\infty} f(x|\theta)\,dx = \frac{1-\beta}{2}$$



→ Shortest interval

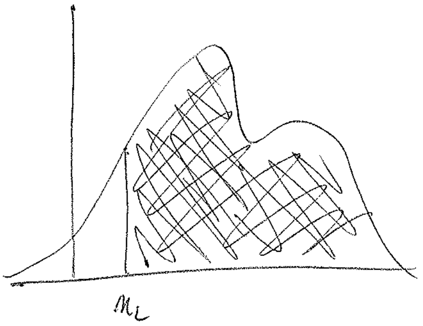
$$f(x_L|\theta) = f(x_U|\theta) \wedge \int_{x_L}^{x_R} f(x|\theta)\,dx = \beta$$

↳ Start with a horizontal line $y = f_{max}(x|\theta)$, and lower until you satisfy such conditions.



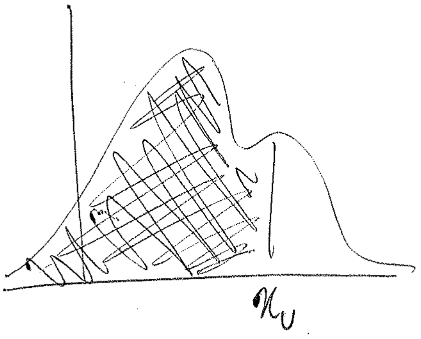
→ Lower limit

$$\int_{-\infty}^{x_L} f(x|\theta)\,dx = 1-\beta$$



→ Upper limit

$$\int_{x_U}^{+\infty} f(x|\theta)\,dx = 1-\beta$$



Example: Ordering rules

- Neyman confidence belt

Take a variable x with PDF $f(x|\theta)$ and θ unknown.

~~Assume x could be an estimator of the parameter θ.~~

Suppose $t(x)$ is some function of the data.

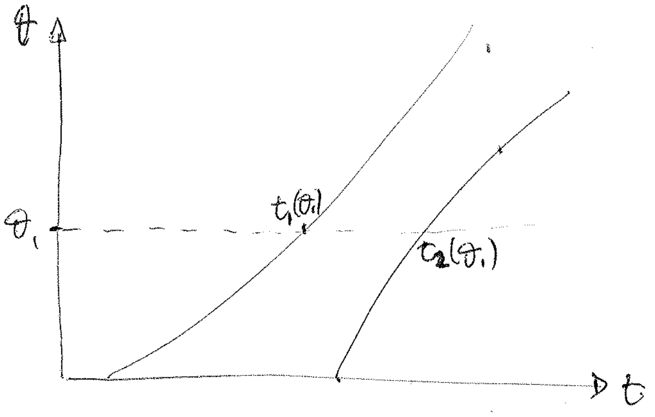
We can write: $\beta = P(t_1 \le t \le t_2) = P(t_1(\theta) \le t \le t_2(\theta))$.

$$\overset{!}{=} \int_{t_1}^{t_2} f(t|\theta)\, dt$$

Assume that we have a way to determine t_1 and t_2 for each value of θ.
$\downarrow$

From the PDF of t: $f(t(x)|\theta)$ we get t_1 and t_2

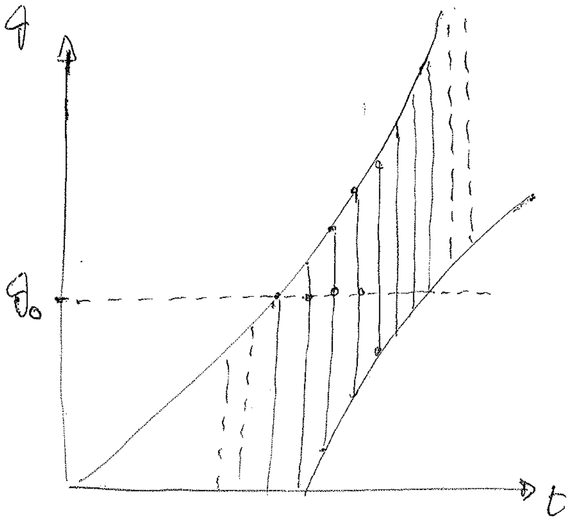
Such values form two curves in the (t,θ) space: ~~Then~~



The ~~sp~~ space between these curves is called the "Neyman confidence belt".

From this belt, ~~we want to~~ given a specific measured value t_0, we want to ~~x~~ extract an interval on the parameter θ.

Suppose θ_0 is the true, unknown value of θ.
If we repeat the measurement many times, a fraction β of the measurements will fall ~~[illegible]~~ in $[t_1(\theta_0), t_2(\theta_0)]$ by ~~[illegible]~~ definition:



By construction, if I ~~[illegible]~~ "invert" the belt and compute $\theta_1(t)$ and $\theta_2(t)$ for any of the measured values of t, the true value θ_0 will fall in $[\theta_1, \theta_2]$ exactly a fraction β of the times:

$$P(\theta_1 \leq \theta \leq \theta_2) = \beta$$

Notice that: → θ_1 and θ_2 ~~are~~ depend only on the data.

→ If the observable t is discrete, we need to construct the belt so that:

$$P(t_1 \leq t \leq t_2) \geq \beta$$

Therefore $P(\theta_1 \leq \theta \leq \theta_2) \geq \beta$

In such a case, we say that the metod "overcovers".

<u>Examples</u>: → Neyman belt for Gaussian case

→ Neyman belt for binomial interval

- Flip-flopping problem

Suppose a variable x has a PDF $f(x|\mu) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ with known σ.

Suppose we have a physics theory that limits μ to be non-negative: $\mu \geq 0$.
This is for example the case of μ representing a mass.

Suppose we do some measurement of x, and that our instrument ~~can~~ is subject to fluctuations so that x can be negative.
This is the case of an analogic scale.

The quoted ~~central interval must~~ central value for μ must always be ≥ 0.

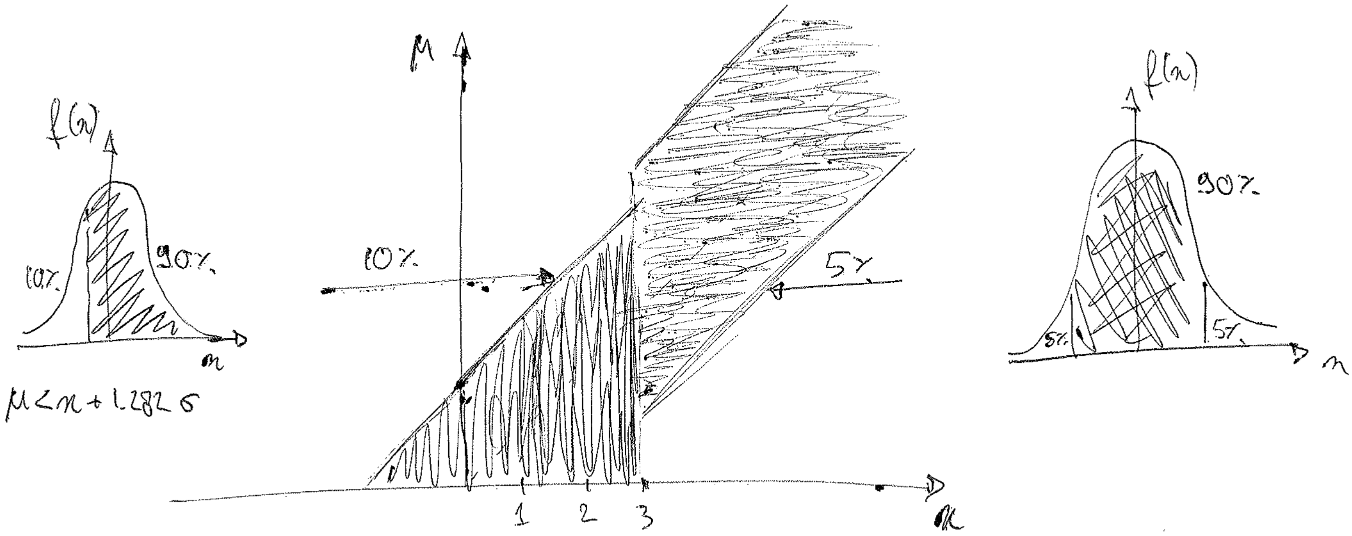
Imagine we decide to quote:

$$\hat{\mu}(x) = \begin{cases} x & \text{if } \frac{x}{\sigma} \geq 3 \\ 0 & \text{if } \frac{x}{\sigma} < 3 \end{cases}$$

The corresponding 90% interval will be:

$$[\mu_1, \mu_2] = \begin{cases} [\hat{\mu} - 1.65\sigma, \hat{\mu} + 1.65\sigma] & \text{if } \frac{x}{\sigma} \geq 3 \\ [0, \hat{\mu} + 1.282\sigma] & \text{if } \frac{x}{\sigma} < 3 \end{cases}$$

The corresponding belt would be:



For some values of μ, e.g. $\mu = 2.5$, we have a coverage of 85% only!

Notice that: → The coverage is a property of the method, not of ~~the~~ a particular interval

→ The flip-flopping issue arises from the fact that our ordering rule depends on the outcome of the measurement. Feldman and Cousins showed that this should not be done!

<u>Example: Flip-flopping</u>

- Unified Feldman-Cousins approach

Recall that we can build a method with exact coverage in an infinite number of ways, just by changing the ordering rule.

So we can look for an ordering rule with a smooth transition from a central interval to an upper limit to avoid the flip-flopping issue.

The Feldman-Cousins approach uses the $\mathcal{L}$-ratio as an ordering rule:

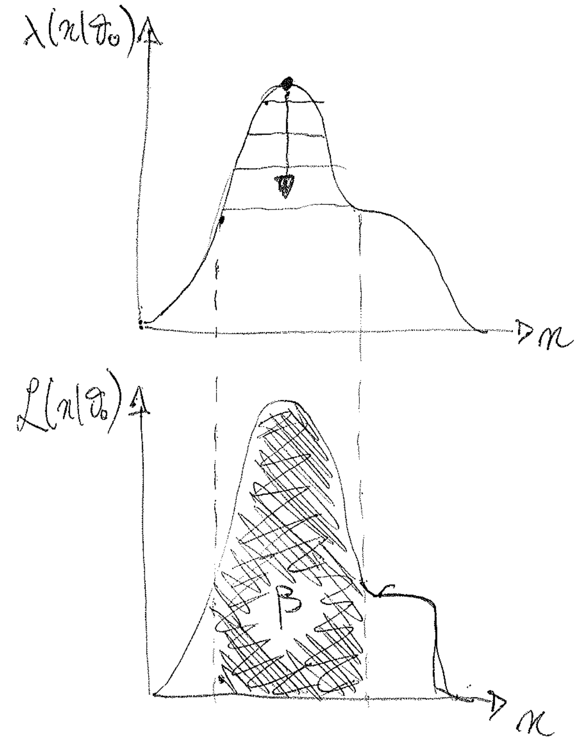
$$\lambda(x|\vartheta_0) = \frac{\mathcal{L}(x|\vartheta_0)}{\mathcal{L}(x|\hat{\vartheta})}$$, where $\hat{\vartheta}$ is the value of ϑ that maximizes $\mathcal{L}$.

To build the FC confidence belt we proceed as follows:

1) Find $\hat{\vartheta}$

~~2) Start from For each~~

2) Fix ϑ to some value ϑ_0.
Then start from $\hat{x} = \max\limits_{x} \lambda(x|\vartheta_0)$ and move to the left and right until we get a coverage β on $\mathcal{L}(x|\vartheta_0)$



3) Repeat for all values of ϑ (in the physical range).

- Examples → FC Belt for Gaussian

  → FC belt for electron neutrino mass

- ~~Numerical~~ Analytical FC belt calculation for Gaussian

Recall the flip-flopping case, where we had x with a PDF:

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}\cancel{\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\cancel{\sigma^2}}\right)$$

The value ~~[illegible]~~ $\hat{\mu}$ that maximizes $f(x|\mu)$ given some measured x is:

$$\hat{\mu}(x) = \max\{x, 0\}$$

The PDF for x, using the max-$\mathcal{L}$ estimate for μ is:

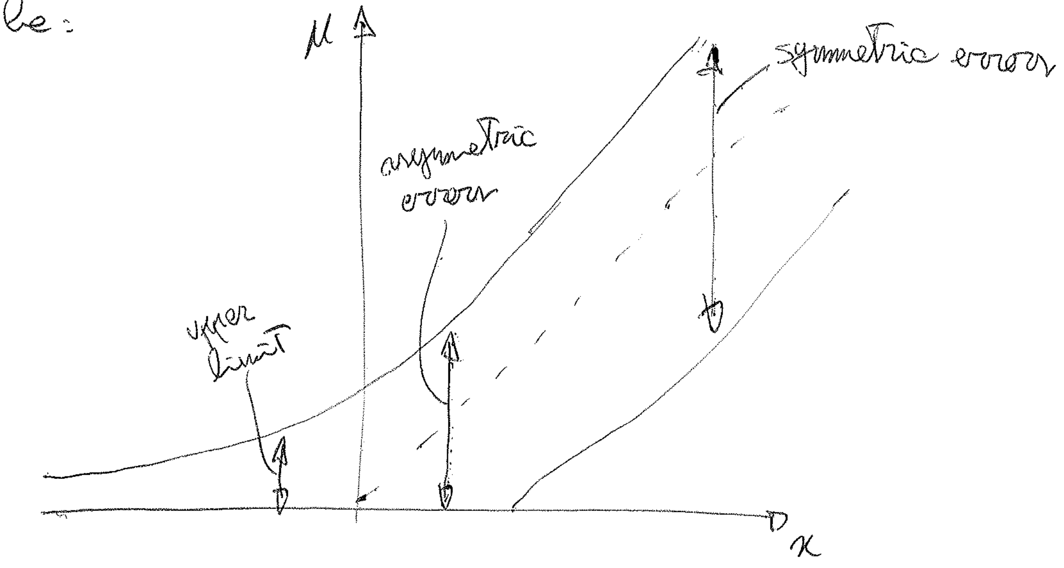
$$f(x|\hat{\mu}(x)) = \begin{cases} \frac{1}{\sqrt{2\pi}} & \text{if } x \geq 0 \\ \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) & \text{if } x < 0 \end{cases}$$

The likelihood ratio becomes:

$$\lambda(x|\mu) = \frac{f(x|\mu)}{f(x|\hat{\mu}(x))} = \begin{cases} \exp\left(-\frac{(x-\mu)^2}{2}\right) & \text{if } x \geq 0 \\ \exp\left(x\mu - \frac{\mu^2}{2}\right) & \text{if } x < 0 \end{cases}$$

At this point, we can find the interval $[\mu_1, \mu_2]$ numerically for any value of μ.

The result will be:



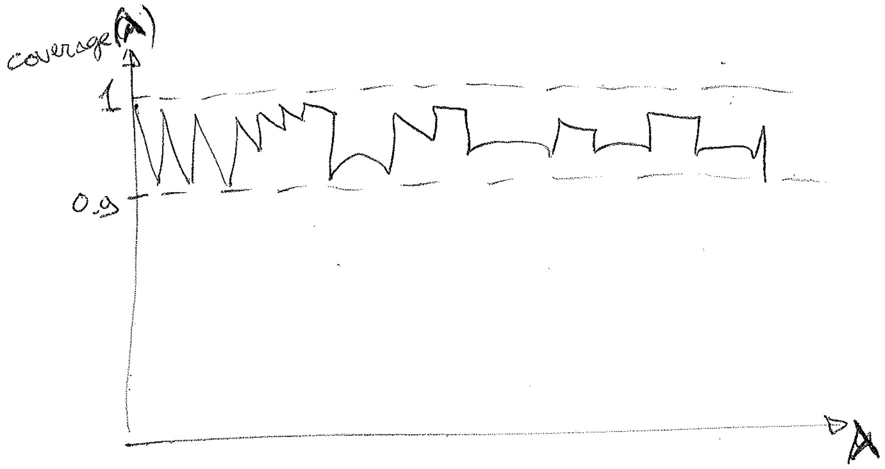
- FC for discrete data

For discrete data, the requirement on the coverage must be changed:

$$\int_{x_1}^{x_2} f(x|\theta)\,dx = \beta \longrightarrow \cancel{\int_{x_1}^{x_2} f(} \sum_{i=L}^{U} P(x_i|\theta) \geq \beta .$$

This can lead to overcoverage, but never to under coverage.

For example, if λ is the mean of a Poisson distribution, ~~looks~~ the coverage will look like:



- Summary: Which method ~~d~~ should I use?

→ If the uncertainties in the parameter estimates are small compared to

a) the non-linearities of the model and

b) the distance to the nearest boundary of the physics region

⇒ The Normal Theory is sufficient, and all methods will provide the same numerical results for the confidence interval.

→ Empirical method to test if Normal Theory is sufficient:

1) Compute confidence interval with Normal Theory method

2) Compute confidence interval with profile likelihood (available in many programs)

If (1) and (2) give the same results, the Normal Theory is sufficient.

If (1) and (2) give different results:

→ If the number of parameters is small (~ 2), Feldman-Cousins is easy to implement.

→ If the number of parameters is $\gtrsim 3$, then FC might become very complicated or CPU intensive, but in this situations typically the profile-$\mathcal{L}$ method provides good coverage.

→ Coverage calculation

In any case, one should make sure the method provides the desired coverage!

This can be done as follows:

1) Assume values for each parameter, generate $\gtrsim 10^4$ Toy-MC experiments, count how many times the confidence interval covers the true value of each parameter.

2) Repeat for different values of the parameters.