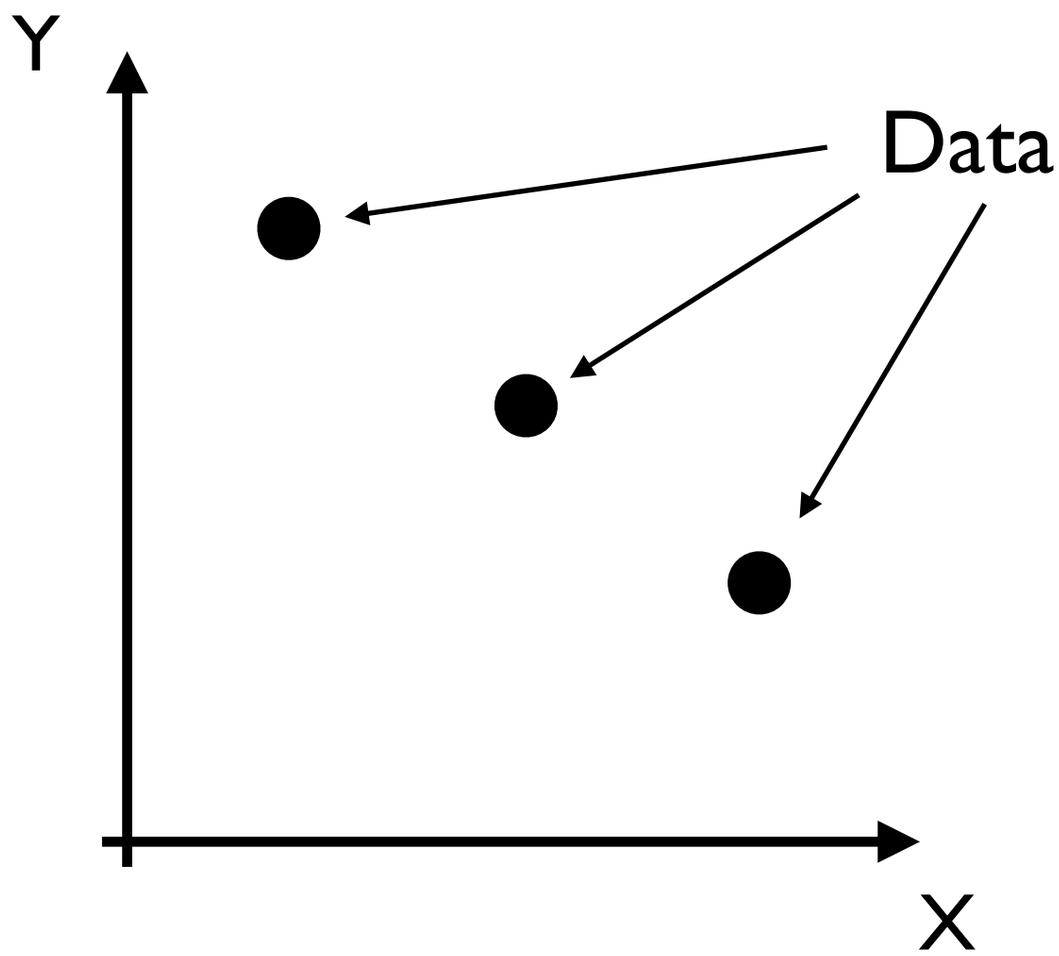
Confidence Interval Estimation
Part I

Roman Kogler (roman.kogler@desy.de)

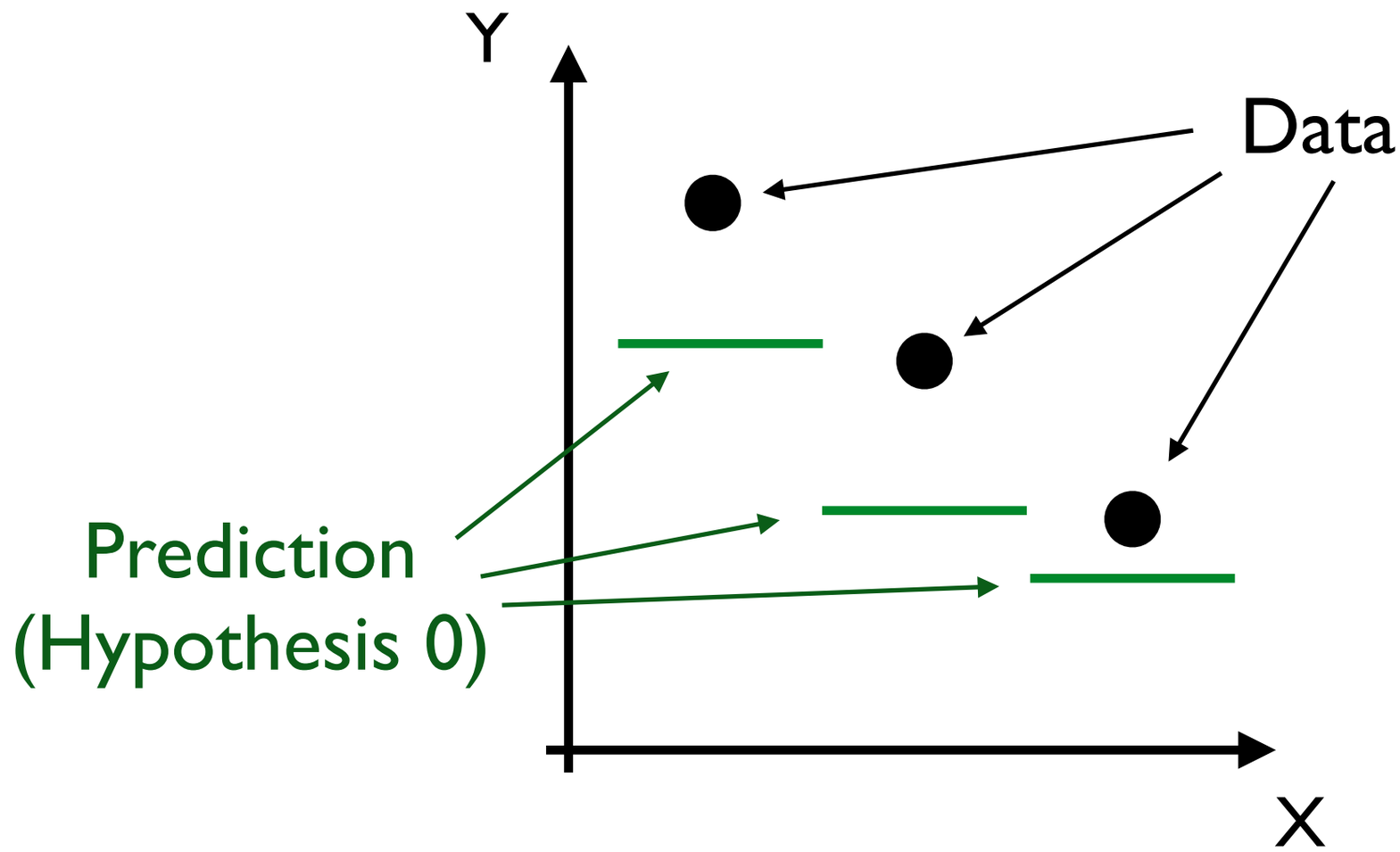
Terascale Statistics School 2023
DESY Hamburg, July 2023

Interpretation of Data



What do we learn?

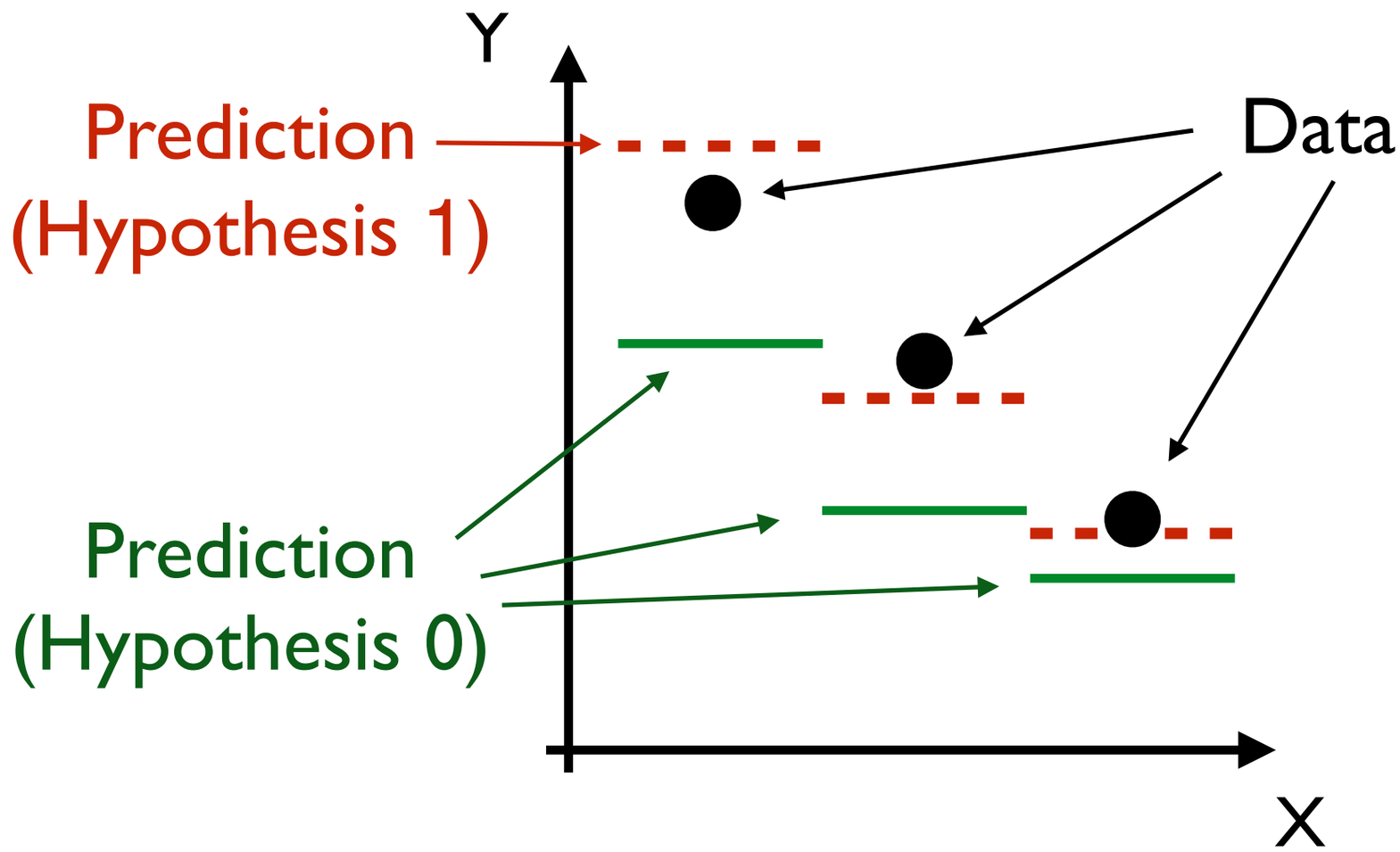
Interpretation of Data



What have we learned?

Now?

Interpretation of Data

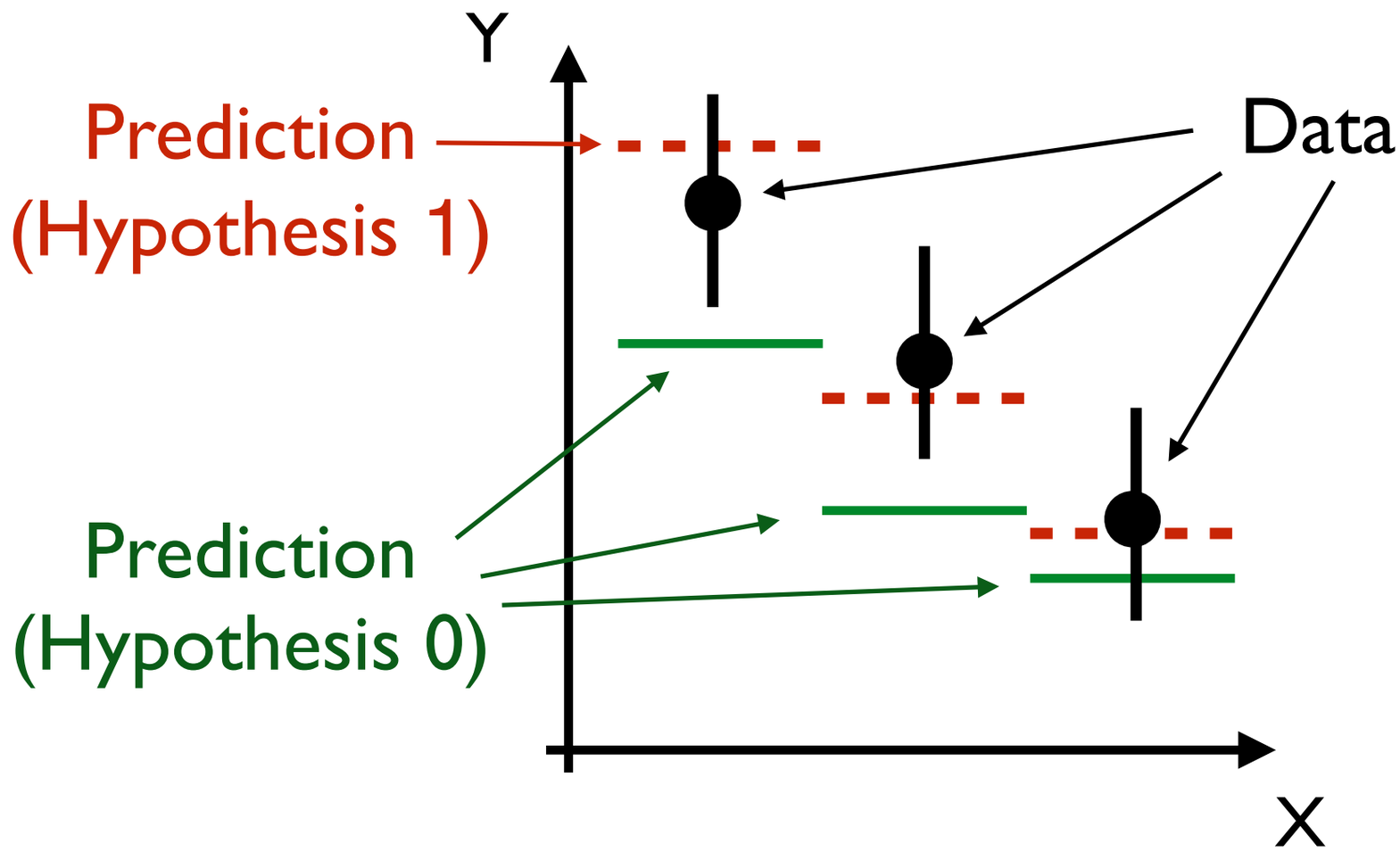


What have we learned?

Now?

Does the data favour H1?

Interpretation of Data



What have we learned?

Now?
Does the data favour H1?

We can't say unless we know the uncertainty (confidence intervals)!

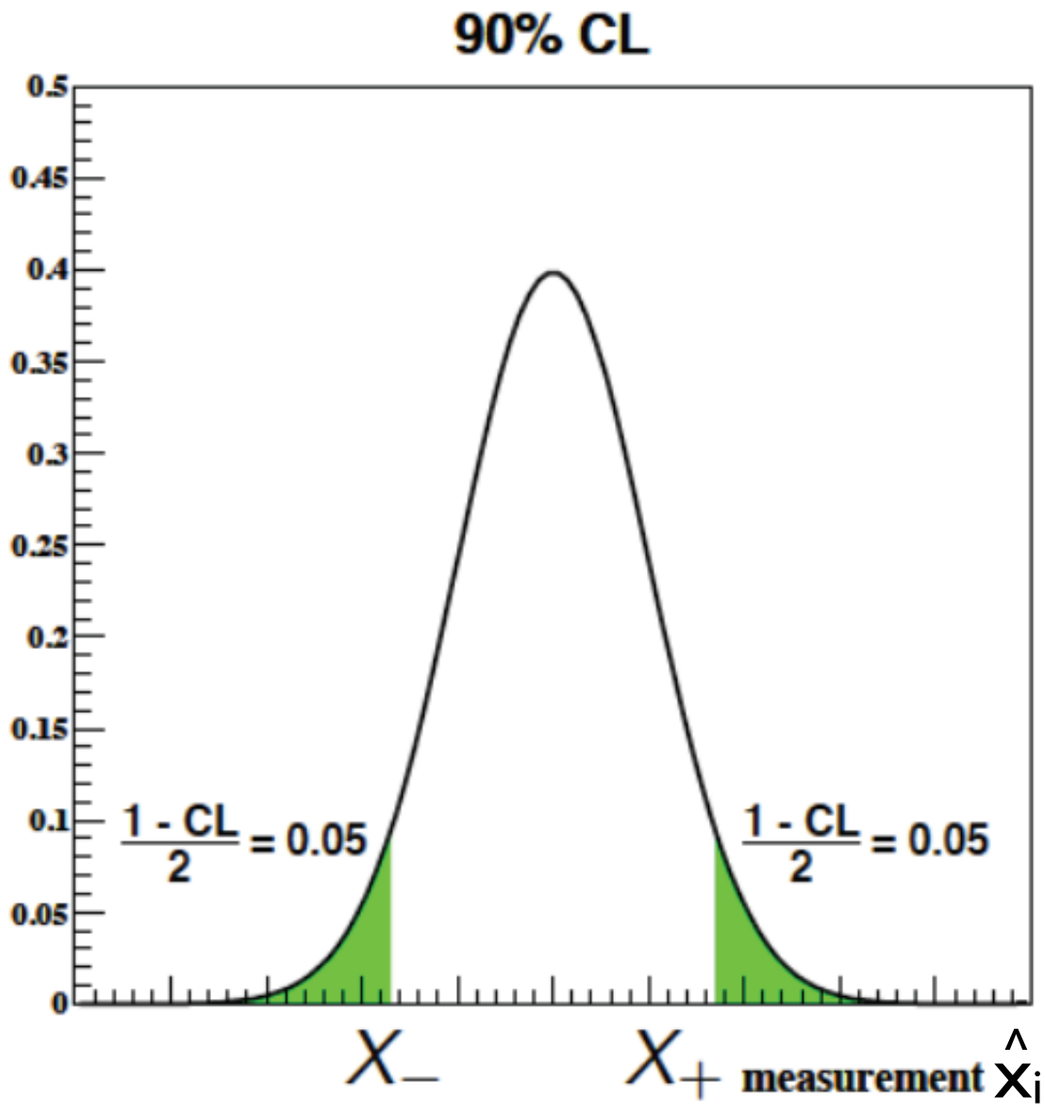
Confidence Intervals

Start with a simple question

We know a true value x and the corresponding pdf, and we want to know in which interval a certain amount of measurements $\hat{x}_i$ will fall.

The value of the measurement $\hat{x}_i$ lies in the interval $[X_-, X_+]$ in "CL" % of the time.

Corresponds to the statement "the interval $[X_-, X_+]$ has CL% confidence level"



$$p(X_- \geq x_i \geq X_+) = \int_{X_-}^{X_+} f(x)dx = \mathrm{CL}$$

More Realistic Problem - Gaussian Case

In a real experiment, x is unknown - need to estimate it

- Gaussian distributed estimator, cumulative probability (CLT):

$$G(\hat{x}, x, \sigma_{\hat{x}}) = \int_{-\infty}^{\hat{x}} \frac{1}{\sqrt{2\pi}\sigma_{\hat{x}}} \exp\left(-\frac{(x'-x)^2}{2\sigma_{\hat{x}}}\right) \mathrm{d}x'$$

with $\hat{x}$ being the measured value and σ_x the standard deviation (resolution) of the measurement

- Confidence interval [a,b] can be obtained through solving

$$\alpha = 1 - G(\hat{x}, a, \sigma_{\hat{x}}) = 1 - \phi\left(\frac{\hat{x}-a}{\sigma_{\hat{x}}}\right) \qquad \text{lower CL is } 1-\alpha$$

$$\beta = G(\hat{x}, b, \sigma_{\hat{x}}) = \phi\left(\frac{\hat{x}-a}{\sigma_{\hat{x}}}\right) \qquad \text{upper CL is } 1-\beta$$

Gaussian Case

We obtain

$$a = \hat{x} - \sigma_{\hat{x}} \phi^{-1}(1-\alpha)$$

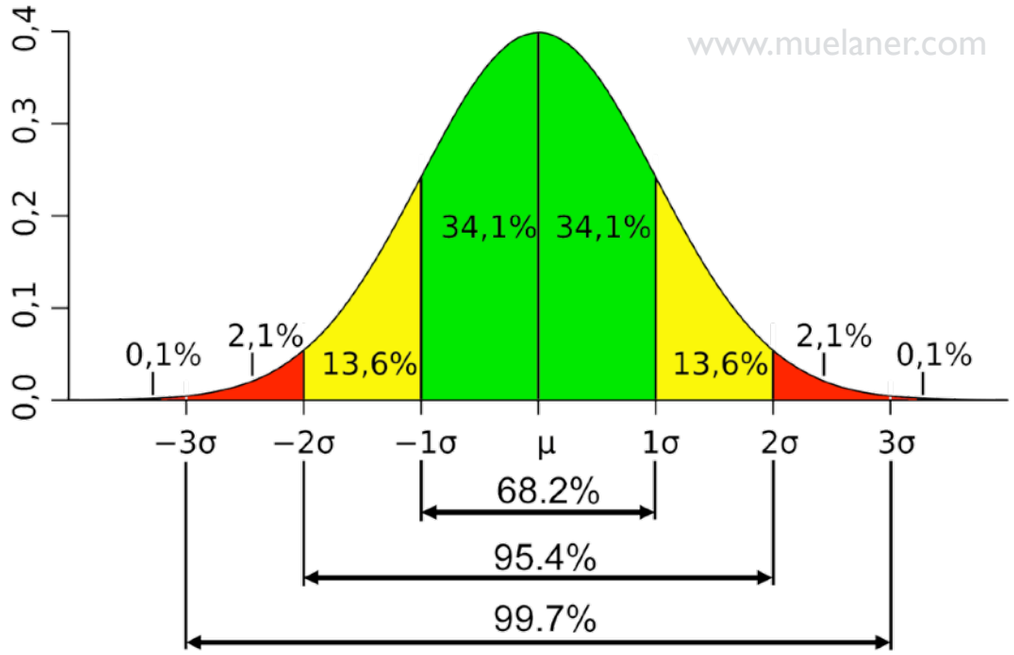
$$b = \hat{x} + \sigma_{\hat{x}} \phi^{-1}(1-\beta)$$

Φ^{-1} are quantiles:

$1-\gamma$	$\Phi^{-1}(1-\gamma)$
0.6827	1
0.9544	2
0.9973	3

Results in the typical "1σ error bar":

$$[a, b] = [\hat{x} - \sigma_{\hat{x}}, \hat{x} + \sigma_{\hat{x}}]$$



- Note:
  - More complicated if $\sigma_{\hat{x}}$ is unknown, have to rely on $\hat{\sigma}_{\hat{x}}$
  - If $\hat{x}$ does not follow a Gaussian pdf

Coverage of Uncertainties

- Parameter μ_m estimated from data sample **x**, errors σ_1 and σ_2 (example: fitting procedure)
- The data follow a probability density $p(x|\mu)$ with given value of parameter μ (fixed, true)
- The result $\mu_{m-\sigma 1}{}^{+\sigma 2}$ determines an interval $[\mu_m-\sigma_1, \mu_m+\sigma_2]$ (the region between error bars).
- What ist the coverage probability of the interval? (short: coverage)

Coverage of Uncertainties

- Parameter μ_m estimated from data sample **x**, errors σ_1 and σ_2 (example: fitting procedure)
- The data follow a probability density $p(x|\mu)$ with given value of parameter μ (fixed, true)
- The result $\mu_{m-\sigma 1}{}^{+\sigma 2}$ determines an interval $[\mu_m - \sigma_1, \mu_m + \sigma_2]$ (the region between error bars).
- What ist the coverage probability of the interval? (short: coverage)

- Coverage $C(\mu)$ is a function of the (unknown) parameter μ, defined as the probability that, with $\mu_m \equiv \mu(n)$, $\sigma_1 \equiv \sigma_1(n)$ and $\sigma_2 \equiv \sigma_2(n)$

$$\mu_m - \sigma_1 \leq \mu \leq \mu_m + \sigma_2$$

$C(\mu)$ is the probability that an experiment will obtain an interval that includes, or "covers", the true value μ.

Frequentist require $C(\mu) = C_0$ or at least $C(\mu) \geq C_0$ with $C_0 = 0.6827$, corresponding to 1σ of a Gaussian distribution

Poisson Distributed Data

- The origin of a statistical error in particle physics is most often the observation of Poisson distributed data.
- The probability of observing n events, if the mean value is μ, is given by

$$p(n, \mu) = \frac{\mu^n e^{-\mu}}{n!}$$

- n is a random variable and μ is fixed.

- There are several different methods for the error-bar scheme.

Joel G. Heinrich, Coverage of error bars for Poisson data, CDF/MEMO/STATISTICS/PUBLIC/6438 (2003)

Choice of σ = √n

- Best estimate of μ: Observation n

- Calculate the uncertainty as √n:

$$n - \sqrt{n} < \mu < n + \sqrt{n}$$

- Common choice, motivated by Poisson distribution,

$$\sqrt{V[x]} = \sqrt{\mu} = \sqrt{n}$$

- Study the results and implications in the exercise

- Serious undercoverage

Pearson's χ^2 Interval

- Pearson's χ^2 is given by

$$\chi^2(\mu, n) = \frac{(n-\mu)^2}{\mu}$$

Pearson's χ^2 Interval

- Pearson's χ^2 is given by

$$\chi^2(\mu, n) = \frac{(n-\mu)^2}{\mu}$$

- After observing n events, the estimate is obtained by minimising χ^2,

  i.e. $\mu_m = n$ with $\chi^2 = 0$

- Uncertainties are obtained from the interval such that

$$\chi^2(\mu_m, n) < \Delta$$

  and we get

$$\sigma_1 = \sqrt{n\Delta + \Delta^2/4} - \Delta/2$$

$$\sigma_2 = \sqrt{n\Delta + \Delta^2/4} + \Delta/2$$

Pearson's χ^2 Interval

- Pearson's χ^2 is given by

$$\chi^2(\mu, n) = \frac{(n-\mu)^2}{\mu}$$

- After observing n events, the estimate is obtained by minimising χ^2,

  i.e. $\mu_m = n$ with $\chi^2 = 0$

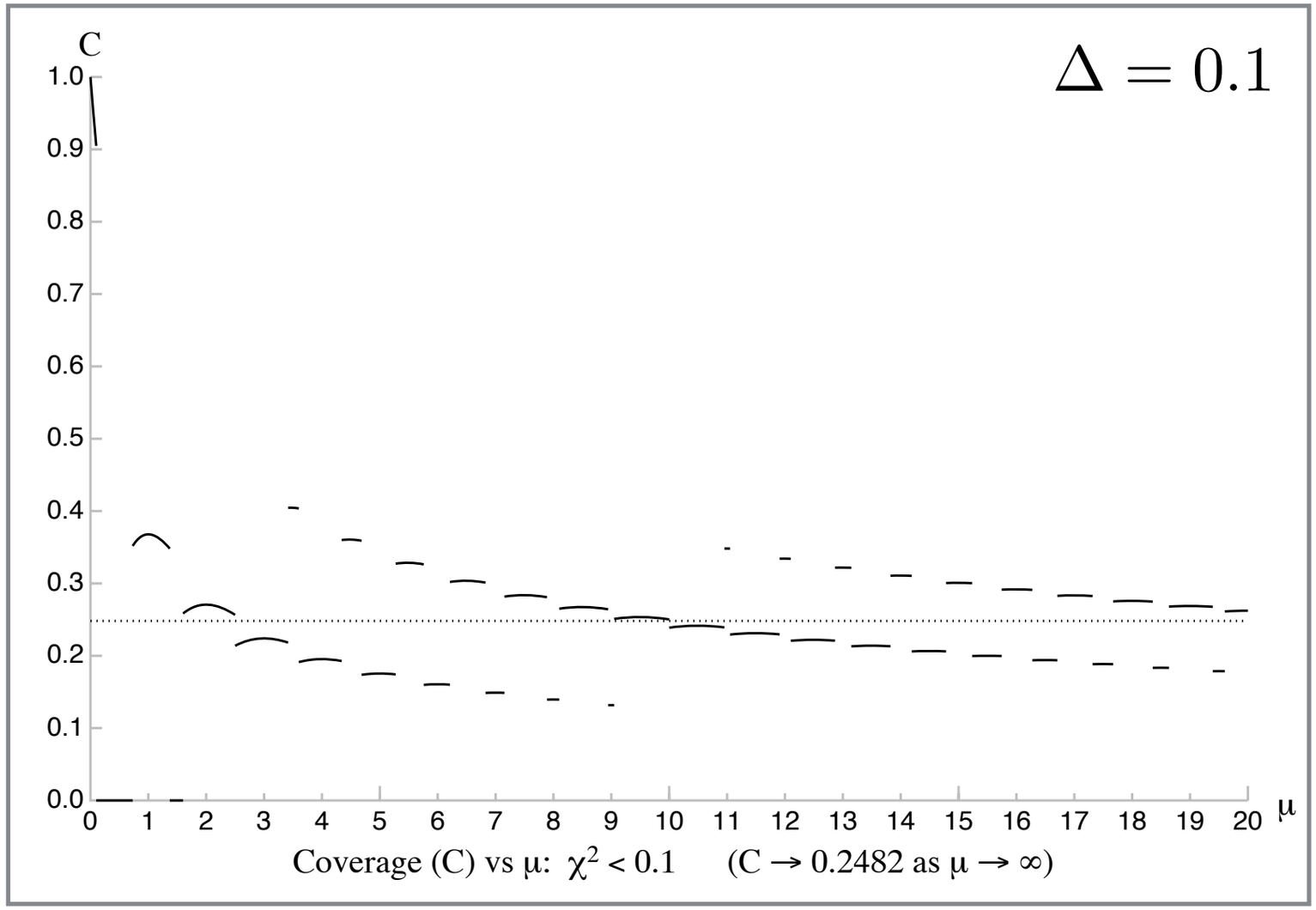
- Uncertainties are obtained from the interval such that

$$\chi^2(\mu_m, n) < \Delta$$

and we get

$$\sigma_1 = \sqrt{n\Delta + \Delta^2/4} - \Delta/2$$

$$\sigma_2 = \sqrt{n\Delta + \Delta^2/4} + \Delta/2$$



many discontinuities, values between 0 (no coverage) and 1 (overcoverage)

Pearson's χ^2 Interval

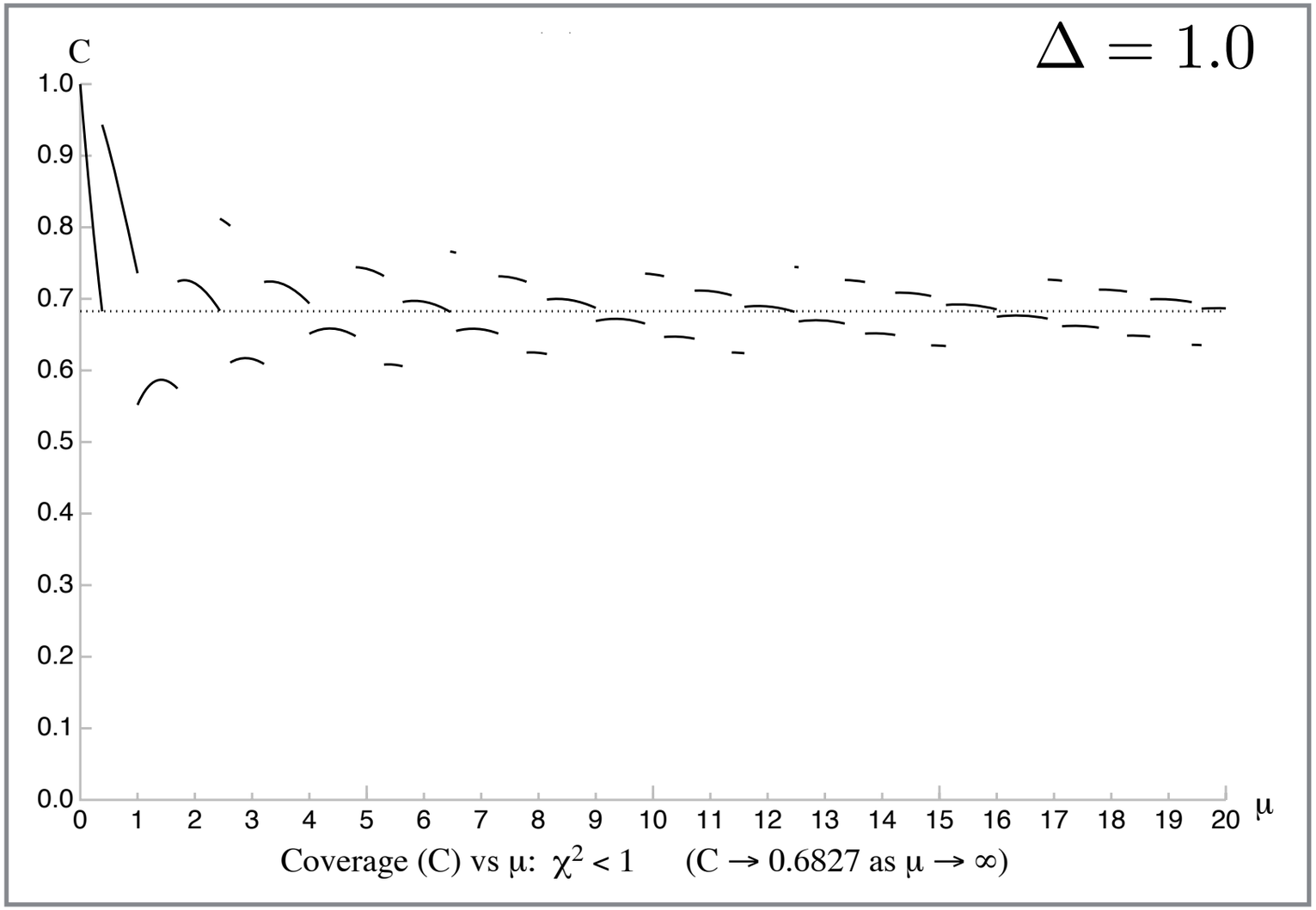
- Standard choice of

$$\Delta = 1.0$$

- minimum coverage is 1/1.5 = 0.5518 obtained at μ ≈ 1
- mean value around 0.68
- example: n = 6 then we would say

$$\mu = 6^{+3}_{-2}$$

but μ is still unknown, it could be close to one we only know with certainty C ≥ 0.5518



Pearson's χ^2 Interval

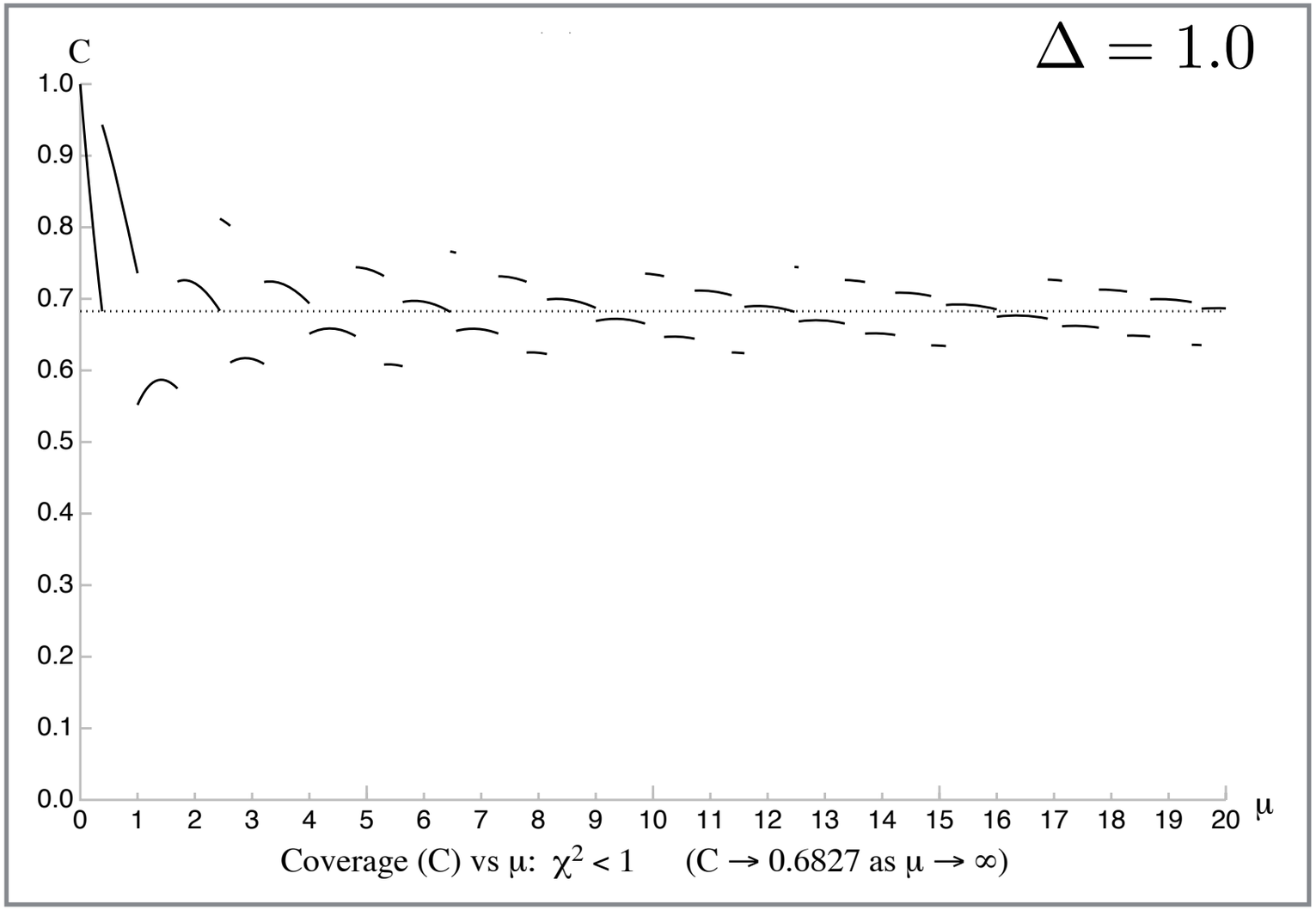
- Standard choice of

$$\Delta = 1.0$$

- minimum coverage is 1/1.5 = 0.5518 obtained at μ ≈ 1
- mean value around 0.68
- example: n = 6 then we would say

$$\mu = 6^{+3}_{-2}$$

but μ is still unknown, it could be close to one we only know with certainty C ≥ 0.5518



undercoverage can be fixed by choosing Δ = 1.5, but then the mean value is around C = 0.78 (overestimation of uncertainties)

Likelihood Intervals

- Error estimate based on the value of the likelihood (case of a maximum likelihood fit)

$$-2\ln\lambda(\mu, n) =$$

$$2[(\mu - n) + n\ln(n/\mu)]$$

- With the uncertainty defined by

$$-2\ln\lambda(\mu, n) \leq \Delta$$

Likelihood Intervals

- Error estimate based on the value of the likelihood (case of a maximum likelihood fit)

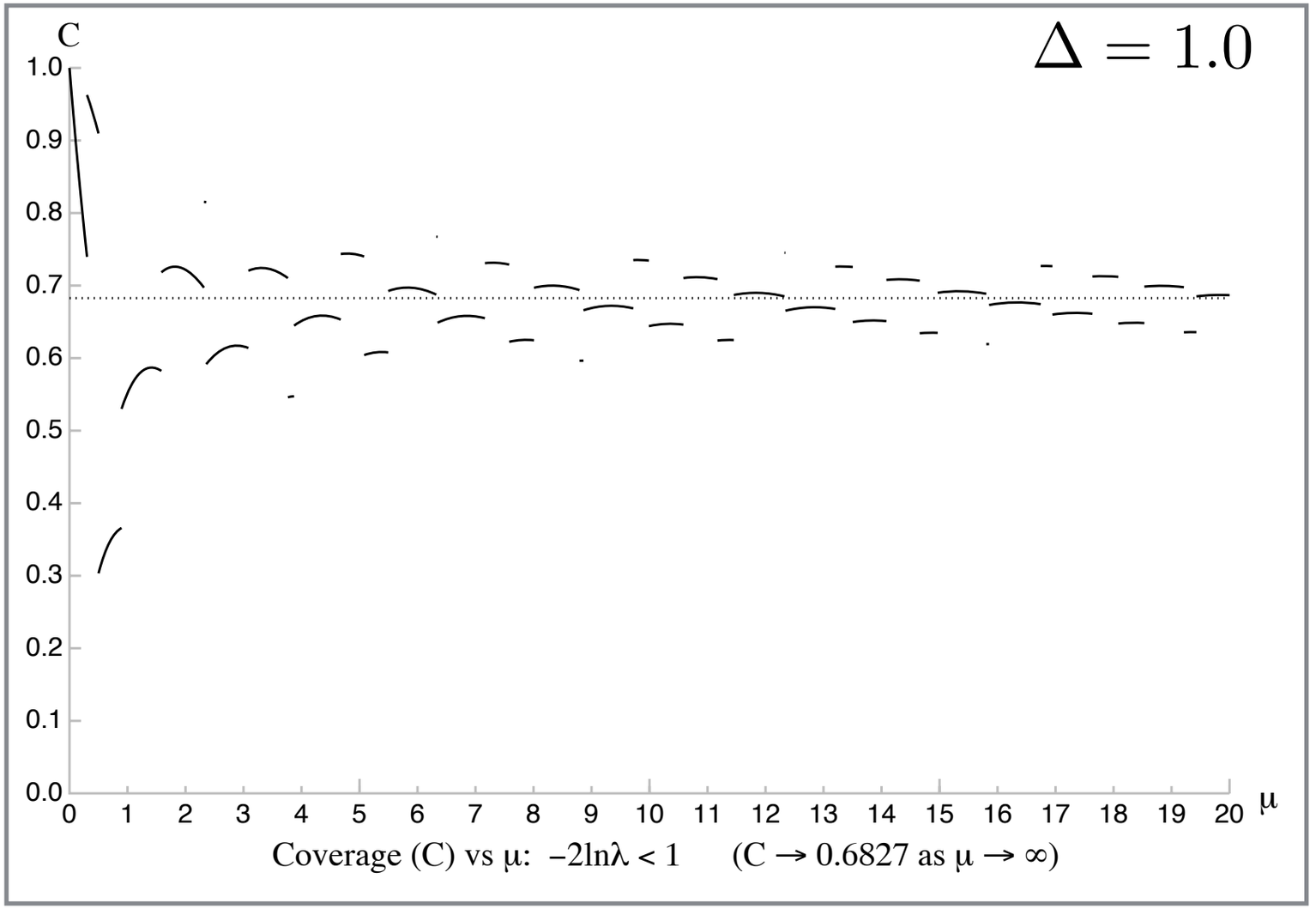
$$-2\ln\lambda(\mu, n) =$$

$$2[(\mu - n) + n\ln(n/\mu)]$$

- With the uncertainty defined by

$$-2\ln\lambda(\mu, n) \leq \Delta$$

- Minimum is 0.3033 in the vicinity of μ = 0.5



undercoverage can be fixed by choosing Δ = 2.581, but then the mean value is around C = 0.89 (even worse overestimation of uncertainties)

Frequentist's Central Intervals

- Coverage achieved by the frequentist's 68.27% interval choice:

$$\sum_{k=0}^{n-1} \frac{e^{-\mu}\mu^k}{k!} \geq \frac{1-C_0}{2}$$
$$\sum_{n+1}^{\infty} \frac{e^{-\mu}\mu^k}{k!} \geq \frac{1-C_0}{2}$$

- Minimum coverage guaranteed to be larger than $C_0 = 0.6827$
- Coverage larger on average
- Special case for n = 0:

$$\mu \leq \ln \frac{2}{1-C_0}$$

Frequentist's Central Intervals

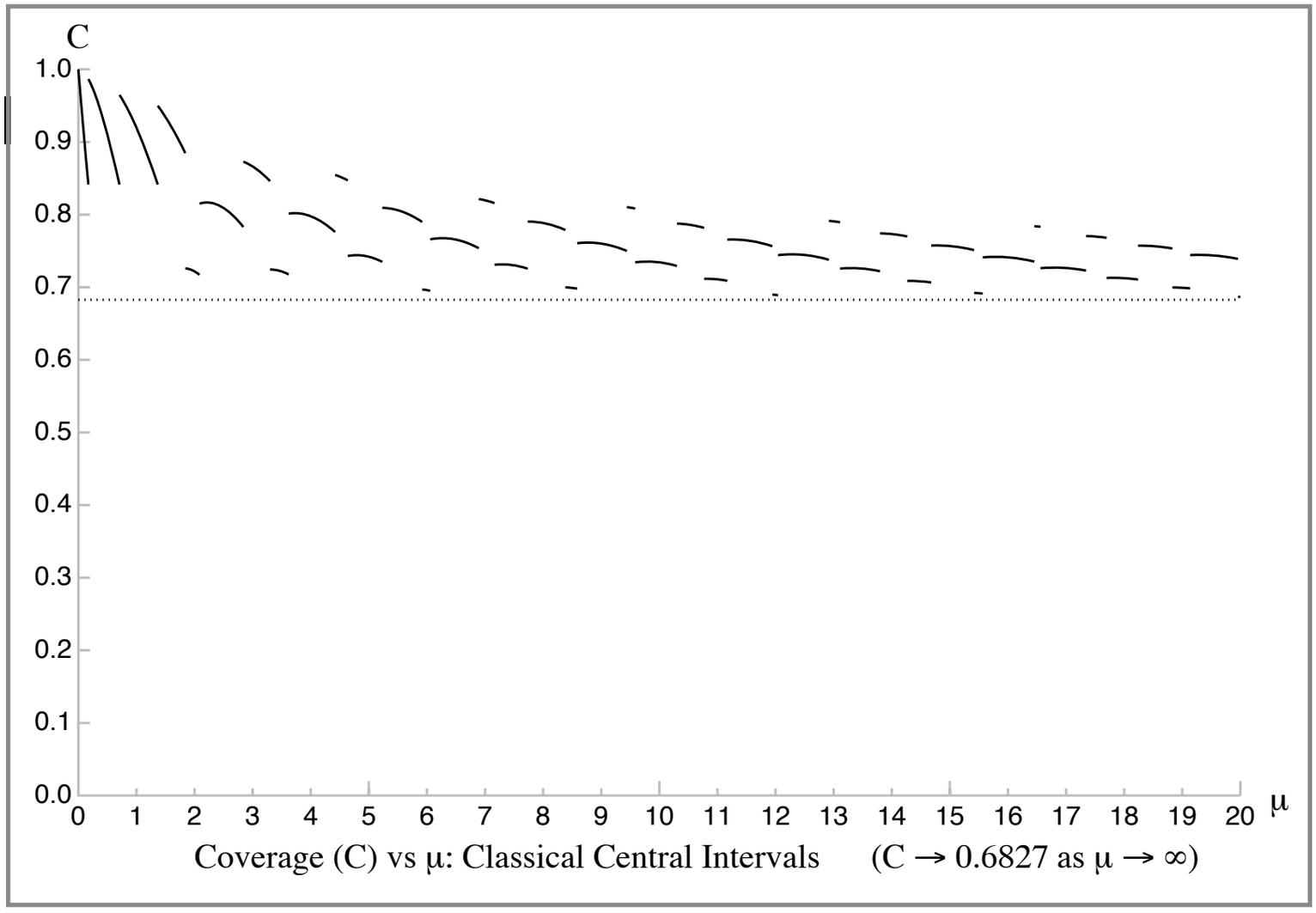
- Coverage achieved by the frequentist's 68.27% interval choice:

$$\sum_{k=0}^{n-1} \frac{e^{-\mu}\mu^k}{k!} \geq \frac{1-C_0}{2}$$

$$\sum_{n+1}^{\infty} \frac{e^{-\mu}\mu^k}{k!} \geq \frac{1-C_0}{2}$$

- Minimum coverage guaranteed to be larger than $C_0 = 0.6827$
- Coverage larger on average
- Special case for n = 0:

$$\mu \leq \ln \frac{2}{1-C_0}$$



more complicated constructions possible, but not widely used

exact coverage $C(\mu) = C_0$ never possible

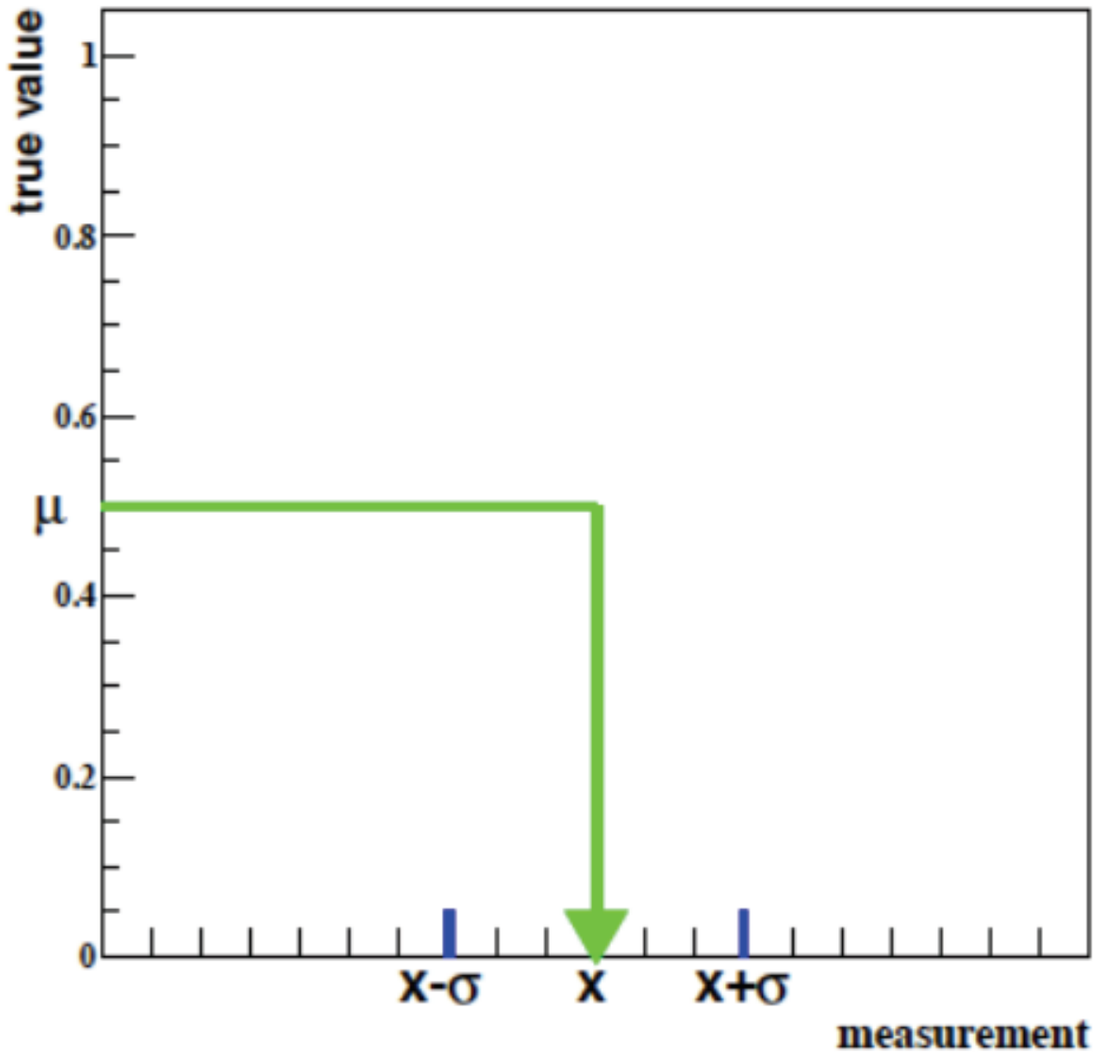
Neyman Construction

Back to the original problem: Quote confidence interval for true parameter μ. Get confidence belt via **Neyman construction:**

1. For a given true parameter μ the distribution of x is known

$$P(X \in [x-\sigma, x+\sigma]) = CL$$

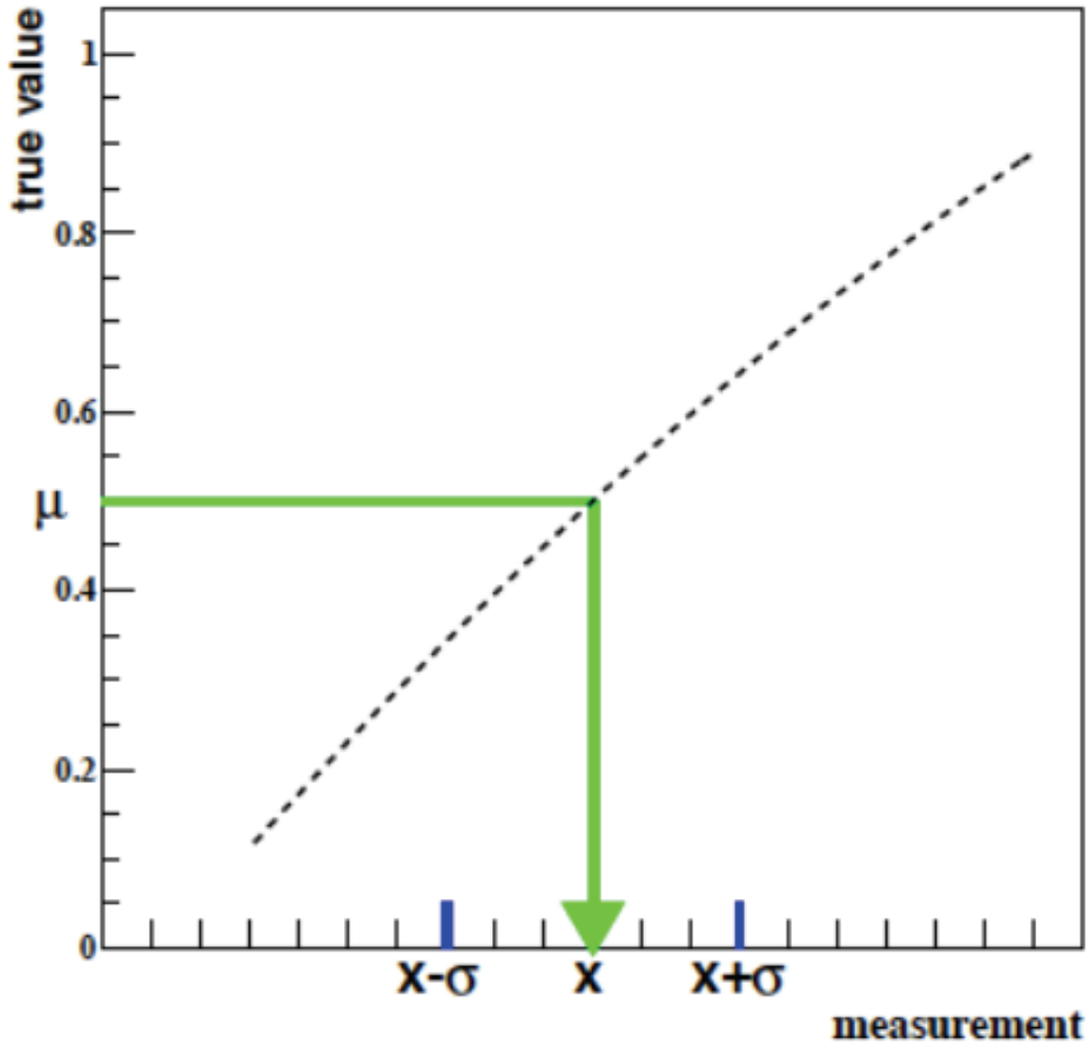
so called "acceptance interval"



Neyman Construction

Back to the original problem: Quote confidence interval for true parameter μ. Get confidence belt via **Neyman construction:**

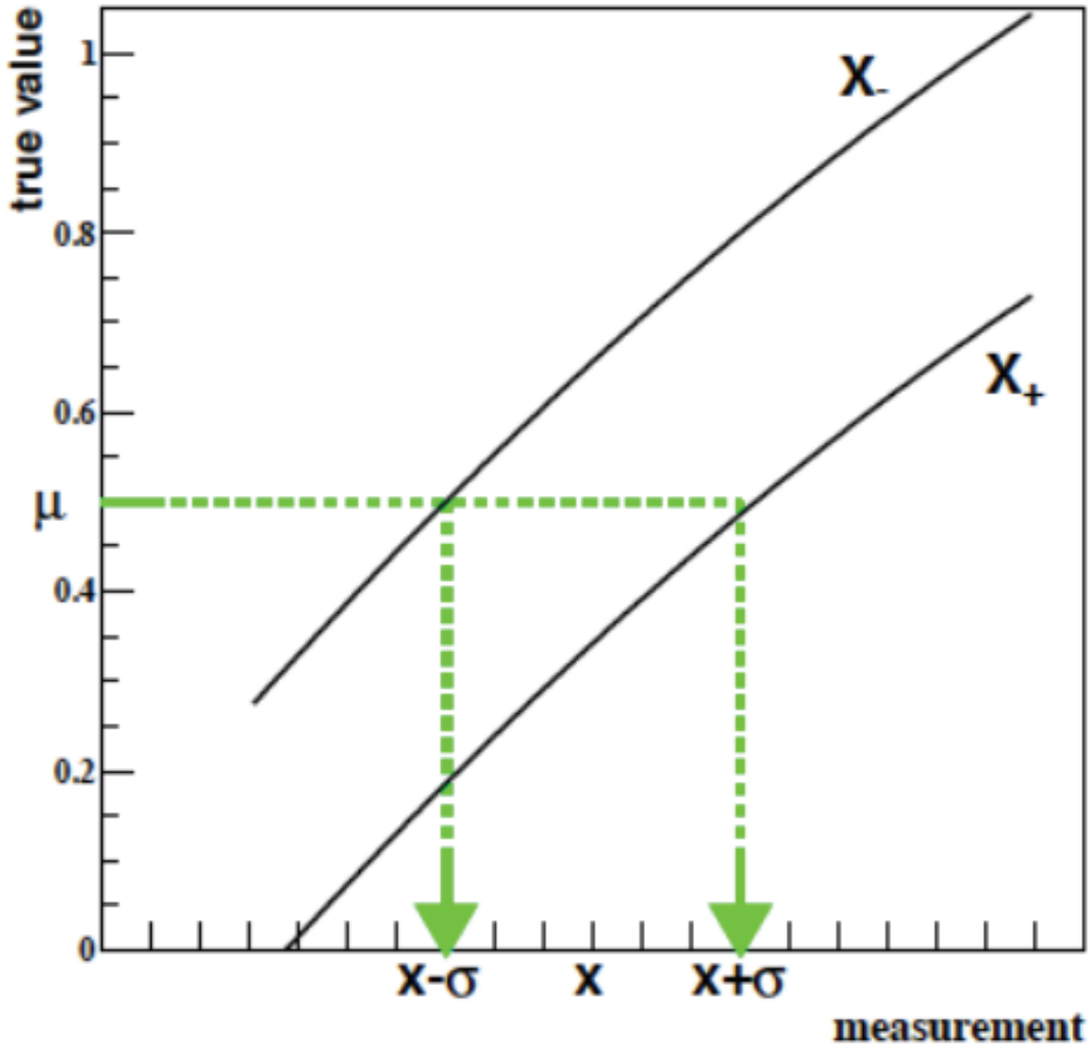
1. For a given true parameter μ the distribution of x is known
2. For each μ we calculate the mean x



Neyman Construction

Back to the original problem: Quote confidence interval for true parameter μ. Get confidence belt via **Neyman construction:**

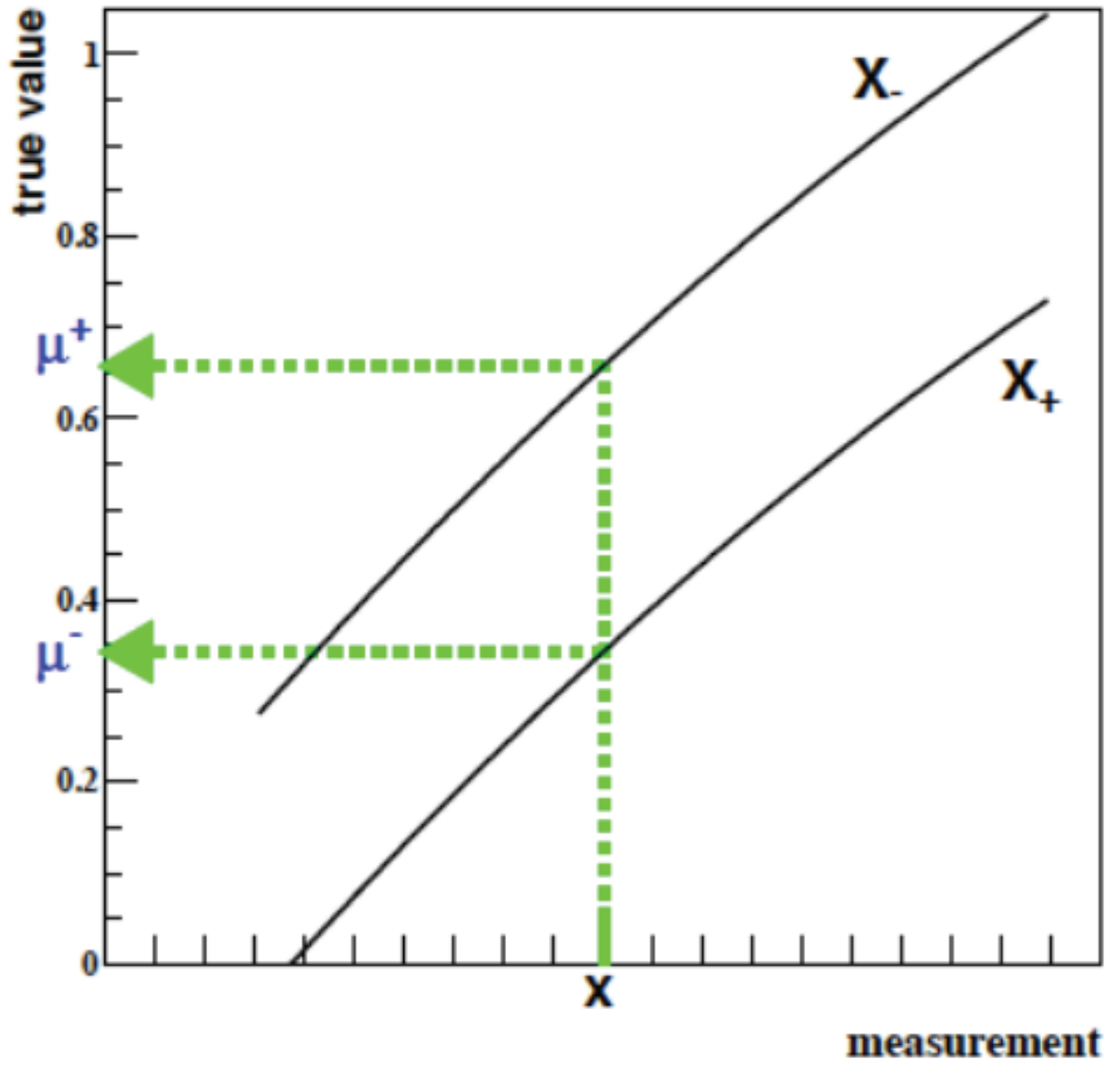
1. For a given true parameter μ the distribution of x is known
2. For each μ we calculate the mean x
3. For each μ and a given CL we can calculate X_- and X_+



Neyman Construction

Back to the original problem: Quote confidence interval for true parameter μ. Get confidence belt via **Neyman construction:**

1. For a given true parameter μ the distribution of x is known
2. For each μ we calculate the mean x
3. For each μ and a given CL we can calculate X_- and X_+
4. For a measured x we can get the confidence interval $[\mu_+, \mu_-]$



The confidence belt contains CL % of the expected measurements.

Meaning: If μ would be μ_+ (μ_-) than the probability to observe x or less (x or more) is CL/2 %. **This is not a statement about μ, but about X_- and X_+.**

(wrong: construct the confidence interval by a Gaussian around x)

Neyman Construction: Poisson Distribution

We observed k events, what are the CL limits on μ?

95% poisson upper limit (solve for μ_+)

$$0.95 = \sum_{r=k+1}^{\infty} P(r;\mu_+)$$

(or equivalent):

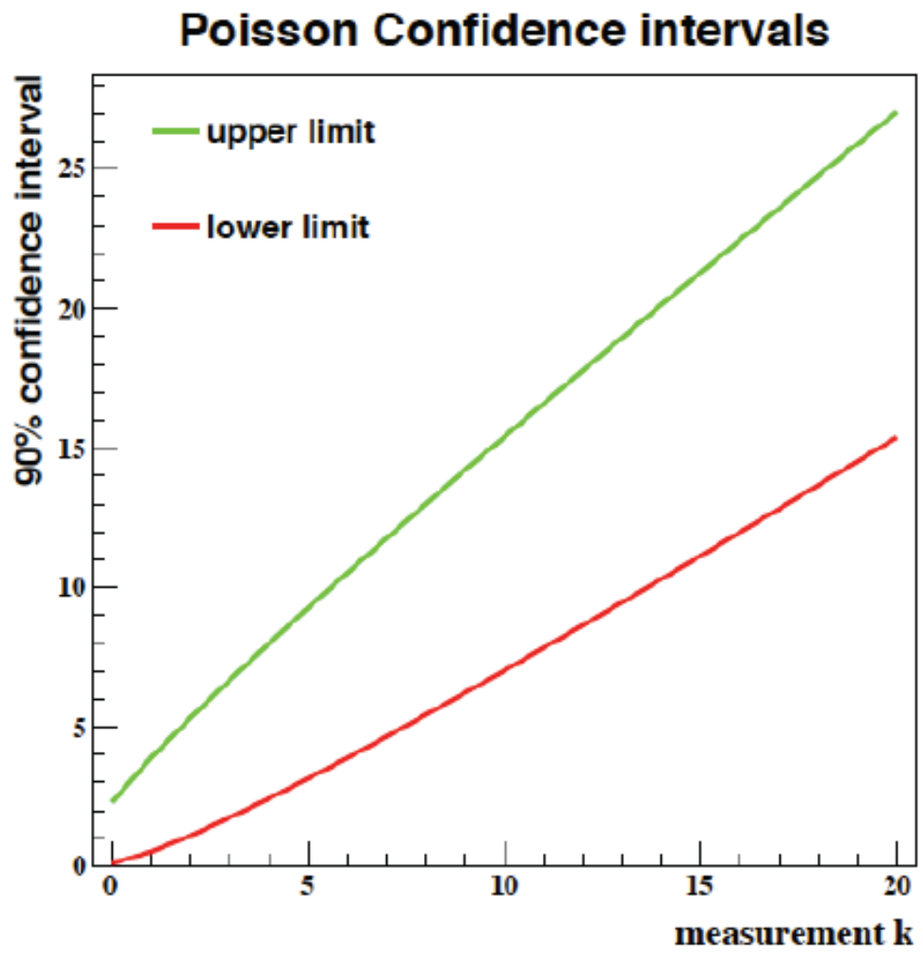
$$0.05 = \sum_{r=0}^{k} P(r;\mu_+)$$

95% poisson **lower** limit

$$0.95 = \sum_{r=0}^{k-1} P(r;\mu_-)$$

(or equivalent)

$$0.05 = \sum_{r=k}^{\infty} P(r;\mu_-)$$



95% Confidence intervals

observation	lower limit	upper limit
0	0	2.99
1	0.05	4.74
2	0.35	6.29
3	0.82	7.75
4	1.37	9.15
5	1.97	10.5
6	2.61	11.84
7	3.29	13.15
8	3.98	14.43
9	4.69	15.7
10	5.43	16.96